# Coulombe's Macroeconomic Random Forest let loose on the house price index

ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Bachelor Thesis BSc2 Econometrics/Economics

Reynier de Graaff 477004

Supervisor: dr. Anastasija Tetereva
Second assessor: Eoghan O'neill

3 July 2022

**Abstract**

In this thesis, I describe the macroeconomic random forest (MRF) machine learning algorithm developed by Coulombe (2020). I start with describing decision trees, building the theory to random forest and eventually end at the MRF. I investigate the ability of the MRF to predict the housing market. Different MRF models are used to forecast the US house price index and they are compared with their OLS counterpart. It is found that the MRF models often forecast better than their OLS counterparts. One of the MRF models used is based on the model created by Adams & Füss (2010). It is found that this model does not contribute to forecast accuracy. At last, the Generalized Time-Varying Parameters (GTVP), which are the time-varying coefficients the MRF outputs, and the Variable Importance measures (VI) of the ARRF model are investigated. The GTVPs seem to adjust to recessions and the variable importance measures show that housing starts is an important driver for house prices.

# Contents

# 1 Introduction

House prices were seen as quite stable, steadily rising and with no peaks and troughs. In 2008 the housing bubble in the US busted, which led to the Great Recession in the US. Trillions of dollars were lost, unemployment increased, people lost their homes and so on. Examining the All-Transactions House Price Index from the FRED-QD database (McCracken & Ng, 2020), it is easily seen that the first peak and troughs occurred during this period. In 2021 and 2022, the House Price Index is reporting its biggest surges ever. Many newspapers and magazines wrote about this and fears of the next housing bubble were rising. With the housing market being deeply rooted in the economy and correlated with macroeconomic variables/indicators, investigating its relations with macroeconomic variables/indicators could have great potential. It may grant insights and possibilities to prevent the next housing bubble.

With the innovation in computational power, new possibilities in the field of econometrics and data science arose. One of those is machine learning. With machine learning we are able to create more sophisticated and complex models and fit those to data. An example of these machine learning methods is Random Forests (RF). A RF is able to fit and predict data using a large set of decision trees. Coulombe (2020) developed a more advanced RF method called Macroeconomic Random Forest (MRF). This RF type should be able to capture the more linear trends, which often are present in macroeconomic variables.

Combining these two subjects we get to the purpose of this paper. This paper will investigate whether it is possible to estimate and predict the real house prices with a MRF. The variable this research focuses on is the All-Transaction House Price Index for the United States. The main research question is therefore:

> *Can we estimate and predict the US house prices with Macroeconomic Random Forests?*

Prediction and estimation are two different things and should be evaluated differently. For prediction, it is important that the predictions are close to the actual values at that time. We have to define close, therefore this thesis investigates whether an MRF can realize forecasting gains compared to OLS estimation. Therefore the first subquestion is:

> *Is it possible to realize significant forecasting gains with a MRF model compared to its OLS counterpart?*

For the estimation part, there are two parts. For the first part, I want to examine the Generalized Time-Varying parameters the MRF outputs of the best predicting MRFs. The second part consists of looking into the Variable importance measures for the out-of-box observations, out-of-sample observations and the generalized time-varying parameters. Therefore the next subquestion is:

> *How do the Generalized Time-Varying Parameters and the variable importance measures of the best predicting MRF model in this paper look?*

For the last part, the model of Adams & Füss (2010) will be implemented into the MRF. They created a model where the housing price index is determined by real construction price, real economic activity and long-term interest. The model will be built based on this, but construction price is exchanged for the personal consumption expenditures price index from the FRED-QD to focus on a more overall macroeconomic price index. Therefore the last research question is:

> *How does the model of Adams & Füss (2010) perform as a MRF in predicting and can we realize gains over other MRFs from this?*

The motivation for this research is connected with the current sharp increase in house prices and the development of machine learning models. Currently, there is a sharp increase in the house price index (McCracken & Ng, 2020). With the last housing crisis just a little more than a decade away, interests in the future of house prices are increasing. Gains in predictive accuracy could be of great value in preventing the next housing crisis. Furthermore, knowledge of the relation of certain macroeconomic variables with the house prices or the main drivers behind the house price could be useful for policies concerning housing.

While there are many econometric estimation methods which are quite well established and have proven to be accurate and predictive, machine learning models still provide new features and advantages. Certain machine learning methods are capable of handling large datasets, whereas the current econometric models are often not capable of that. Forecasting gains can be achieved with these methods. One of those machine learning methods is Random forests. While RFs are capable of handling large datasets, they are not really interpretable. As Coulombe (2020) states: "ML is currently of great use to macroeconomic forecasting, but of little help to macroeconomics". His MRF offers a solution, shifting the focus from the estimation of the dependent variable to the estimation of the coefficients of the independent variables driving the dependent variable. According to Coulombe (2020), MRF is a better estimator than most other machine learning algorithms and its output, the GTVPs, is interpretable.

My hypothesis is that the MRF will be able to estimate and predict real US house prices. I hypothesise that the MRF will be able to realize significant forecasting gains against its OLS counterpart. I expect the GTVPs to change over time, especially in periods of recession I expect to show different trends than usual. For the VIs I expect variables highly correlated with housing prices to be the most important, such as mortgage rates. For the model of Adams & Füss (2010), I expect the model not necessarily to realize gains over other MRF models as the housing market often depends on past trends.

The outline of the thesis will be as follows. In section 2, the relevant literature on RFs and macroeconomic influences on house prices is reviewed. In section 3, a description of the data used will be given. Furthermore, some summary statistics and graphs of factors will be shown. In section 4, a detailed explanation of the methods used in this paper will be given. In section 5, a discussion of the results will be given. Finally, section 6 will end with a conclusion and discussion of this thesis.

## 2 Literature

The concept of decision trees has been around for a long time. It is one of the most basic forms of a model. The first RF algorithm was developed by Ho (1995). She was the first to propose building multiple decision trees and averaging the predictions of those trees to decrease bias and variance. The trees were built on random subsets of the variables or, in other words, random subspaces. So for each tree, a different set of state variables was chosen to build a tree to predict the dependent variable.

Breiman (1996) proposed a new method to improve the accuracy of predictors, bagging. His concept was to create many bootstrap replicates of a dataset and, with those replications, create multiple versions of a predictors. He proved that averaging these predictors leads to a more accurate version of the original predictor. Later on, Breiman (2001) proposed to use bagging to create multiple trees. He also proposed to decorrelate the trees by at each split giving the algorithm a different randomly chosen subset of splitting variables to consider. The idea of decorrelating the trees was influenced by Amit & Geman (1997), who had already written about choosing from a random subset of splitting variables at each split. Finally, by combining the bagging and decorrelation of the trees, Breiman (2001) proposed his random forest proposition.

The linearization of trees was already proposed earlier by Wang & Witten (1997). They already considered using a linear regression plane in the leaves to model the data in that leaf. Similarly, Alexander & Grimshaw (1996) also considered using a linear regression in the leaf to model the data. Their algorithm did, however, differ from Wang & Witten (1997). Friedberg et al. (2021) were the ones to consider a group of them to create a local linear forest. This type of RF is able to exploit the smooth trends of the dependent variable, whereas a regular RF is not. They find that this type of RF improves on asymptotic rates of convergence compared to a regular if there are smooth trends present. They also find substantial gains in accuracy.

At last, there is Coulombe (2020), who improves the local linear forest algorithm with his MRF. This type of RF takes into consideration the smoothness of transitions of trends. As often visible in macroeconomics, the state of variables transitions smoothly from one state to another. With his MRF he also proposes to create Moving Average Factors (MAFs) for the full set of available variables and using Block Bayesian Bootstrap as bootstrapping method. In an earlier paper, Coulombe et al. (2021) already researched the use of MAFs and found that it could provide substantial forecasting gains. The Block Bayesian Bootstrap is a mix of the Bayesian Bootstrap of Rubin (1981) and the Block bootstrap of Mackinnon (2006). Cirillo & Muliere (2013) had already proposed an urn based version of Bayesian bootstrap similar to that of Coulombe (2020). According to Coulombe (2020), the MRF exhibited forecasting and accuracy gains.

There is lots of research on the effects of macroeconomic variables on the housing market. Sutton (2002) looked into the effects of GNP, interest rates and equity prices on the real house prices of different countries. They found a positive relationship for GNP and equity prices with

real house prices and a negative relationship between interest rates and house prices. Tsatsaronis & Zhu (2004) researched the effects of shocks of several macroeconomic variables on the house prices of multiple countries. Their main finding is that there is a strong and long-lasting link between inflation and interest rates with housing prices. Tripathi (2019) also researched the effect of several macroeconomic variables on house prices and found that several of them had a positive effect, including GDP, inflation, money supply and GDP growth rate. These articles all used models with cross-country relations. In all these papers, the US was included in the estimation.

Cohen & Karpavičiūtė (2017) researched the effects of macroeconomic variables on house prices in Lithuania. She found that GDP, unemployment and house prices in the previous period are of influence in the next period. Égert & Mihaljek (2007) looked into the effects of macroeconomic variables on the house price for several central and eastern European countries. They found a strong positive relationship between GDP per capita and house prices and a negative relationship between interest rates and house prices. Hossain & Latif (2009) investigated the house prices in Canada using a GARCH model. They found that GDP growth rate, housing price appreciation rate and inflation affect the volatility of housing prices. Sutton (2002) investigated the relationship between house prices and interest rates in the US and around the world. They found that especially the short-term interest rate plays a large role in the changes in house price. They also find that the effect is rather gradually than on impact.

Garriga et al. (2019) create a theoretical model for the movement of house prices and derives from that theoretically reduction in mortgage rates always has s positive effect on house prices. Goodhart & Hofmann (2008) find a multi-directional link between house prices, monetary variables and the macroeconomy. They find that money supply, credit and house prices all affect each other. They also find that shocks on these variables affect several macroeconomic variables and vice-versa.

Adams & Füss (2010) create a model from which they derive house prices depend on economic activity, construction costs and long-term interest rates. They also investigate multiple countries, including the US. They find positive effects for construction costs and economic activity and negative effects for long-term interest rates in the US. Their model is the model that also is going to be implemented in a MRF later on in this thesis. Construction costs will then be substituted for the regular consumer price index to investigate a more macroeconomic relation. Case & Shiller (1990) found that the ratio of construction cost to price, increases in population and changes in real income per capita are all positively related to price changes in the house market in the next year. Their research was based on data of 4 cities in the US, Atlanta, Chicago, Dallas and San Francisco.

# 3 Data

The database used in this thesis is the FRED-QD dataset of McCracken & Ng (2020). The dataset contains real-time updated quarterly data of macroeconomic variables from the FRED

database which is maintained by the Federal Bank of St. Louis. The dataset starts in 1959 quarter 1 and ends in 2022 quarter 2, so $N = 253$. The dataset is imported in R using the cykbennie/fbi package, which McCracken recommends on the website of the Federal Bank of St. Louis. The database contains 246 variables.

While some of the series are stationary ($I(0)$), most of them are not and require transformations to become stationary. McCracken & Ng (2020) provide benchmark transformation codes which link to a certain transformation for each series, such that each series is transformed to be stationary. The codes and their transformations are: (1) no transformation, (2) $\Delta x_t$, (3) $\Delta^2 x_t$, (4) $log(x_t)$, (5) $\Delta log(x_t)$, (6) $\Delta^2 log(x_t)$ and (7) $\Delta(x_t/x_{t-1} - 1.0)$.

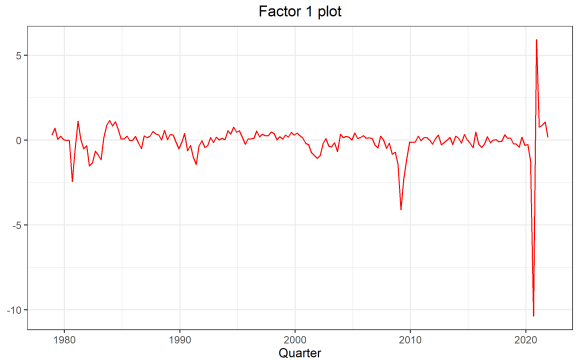


Figure 1: Factor 1 of FRED-QD.

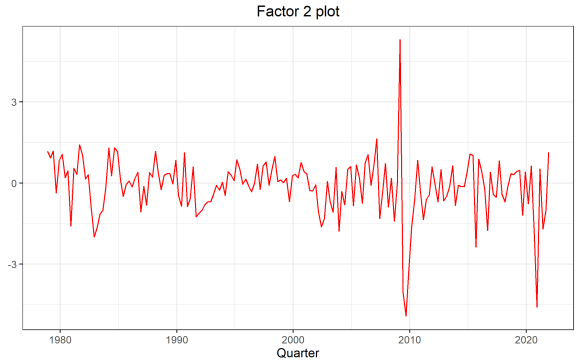


Figure 2: Factor 2 of FRED-QD

Furthermore, McCracken & Ng (2020) compute principal component analysis based factors of the whole dataset. The factors can summarize the whole dataset in just a few series. The principal component analysis is done on the already transformed variables. McCracken & Ng (2020) state that the first factor is the real activity factor, which is largely determined by the employment and industrial production variables. The second factor is the forward-looking factor, which is largely determined by interest rate term spreads as well as housing permits and starts. The third factor represents a pure consumer price index because it is heavily related to variables associated with price groups. The fourth and fifth factors are harder to interpret but according to McCracken & Ng (2020), they mostly correlate with earning and productivity variables. The graphs for the first two factors can be seen above. The factors did change a bit if you compare them to McCracken & Ng (2020), however similarities can still be found. This change is likely due to the addition of new observations in the dataset and differences in estimation.

One of the advantages of the MRF is that it can handle a lot of data. Therefore the whole dataset is going to be used for this research. On top of that, Coulombe (2020) uses the first five factors. As such, the same factors will also be used in this paper. The MRF does require an estimation in the leaves. Depending on the model, different variables will be used. The variables used are the first two factors, lags of the house price index and the variables from the model of Adams & Füss (2010). The model of Adams & Füss (2010) uses the variables long-term interest rate, construction costs and economic activity. They create the economic activity variable by doing principal component analysis on the matrix of the variables of real money supply, real

consumption, real industrial production, real GDP, and employment, and then taking the first factor. So the same method is used to create this variable. The principal component analysis is done on the already transformed variables. In this thesis construction costs sis replaced with the regular consumer price index to investigate this model on a more macroeconomic level. For the long-term interest, rate they take the ten year government bond yield.

Table 1: *Stationarity codes and summary statistics after transformation for stationarity for unemployment rate, house price index, personal consumption expenditures index, 10-year treasury maturity rate and economic activity.*

| | Stationarity code | Mean | St. Dev. | Min. | Max. |
|---|---|---|---|---|---|
| UNRATE (Percent) | 2 | -0.008069 | 0.7290 | -4.1334 | 9.1667 |
| USSTHPI (Index 1980Q1 = 100) | 5 | 0.004219 | 0.0127 | -0.0371 | 0.0431 |
| PCECTPI (Index 2012=100) | 6 | 0.000051 | 0.0038 | -0.0267 | 0.0131 |
| GS10 (Percent) | 2 | -0.008135 | 0.4594 | -2.4500 | 1.5400 |
| Economic activity | - | 0.000000 | 1.9226 | -9.3493 | 23.8805 |

The table above shows the stationarity codes and some summary statistics after transformation for the variables. The variables that are going to be used are unemployment rate (UNRATE), all-transaction house price index in the United States (USSTHPI), personal consumption expenditures chain-type index (PCECTPI), 10-year treasury constant maturity rate (GS10) and economic activity. The unemployment rate will be used as the dependent variable in the replication part. The house price index is the dependent variable in the extension, while the others are the independent variables in the model based on Adams & Füss (2010). The mean of economic activity is 0 because the factors were standardized and centered.

# 4 Methodology

There are several methods of fitting a model to data. One of the most famous methods is Ordinary Least Squares (OLS). This method tries to fit a linear model to data by minimizing the squared residuals. OLS is, under certain assumptions, able to model linear relations, but when non-linear relations are present in the data, OLS will not be able to fit an accurate model to the data. One method to model non-linear relations is a tree-based model. Basic trees partition the data into different smaller pieces and then fit a constant, often the mean, to these smaller pieces of data (Hastie et al., 2001).

### 4.1 Decision Trees

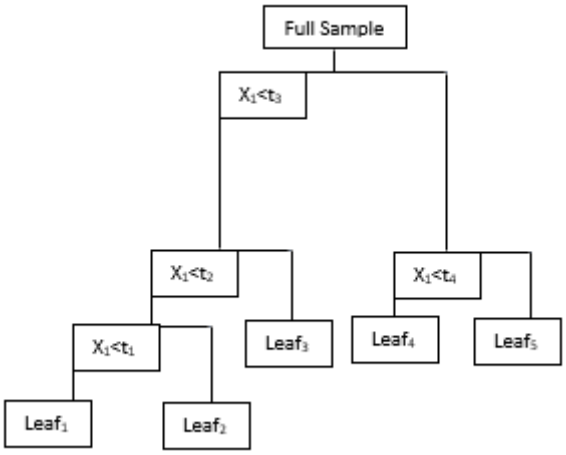


Figure 3: Example of a decision tree

Take a data set consisting of $T$ observations. The data set consists of one dependent variable $y_t$ and $p$ state variables $x_{t,p}$ for $i = 1, 2, ..., T$. We want to model this data with a tree that fits the smaller partitions of data with the mean. The tree splits the data into $M$ partitions $l_1, ..., l_M$, which are also called leaves. The partitions are based on conditions of the state variables. Take for example the figure above, here we base the partitions on only one state variable, $X_1$, which represents the time. Let $t_1 < t_2 < t_3 < t_4$, then if t is between $t_2$ and $t_3$ we end up in Leaf 3, $l_3$. Now each leaf contains the data ($y_t$ and $x_{t,p}$) of certain regions of $t$. Now in this case the data is only partitioned based on time, we could however include another variable in the partition, $X_2$. This variable represents not time but hot and cold weather for example and we use this variable to partition the data in leaf 3 further. This creates leaves 6 and 7. So when $t$ is between $t_2$ and $t_3$ and we have hot weather, we end up in leaf 6.

So now we partitioned the data into separate leaves, we need to link a result to those leaves. The most basic solution is to take the mean of the of the observations of the dependent variable in that leaf. This leads to the formula for the outcomes of the model similar as in Hastie et al. (2001):

$$y_t = \sum_{t=1}^{T} \mu_m I(t \in l_m). \tag{1}$$

In this formula, the $\mu_t$ represents the mean of the leaf $l_m$ and $I$ is the identity function. It is in this case for each observation not possible to be part of multiple leaves. So the sum of the product of the mean and identity function for a certain $t$ is a sum of 0s and one non-zero value. This results that if an observation with time t is part of a certain leaf $l_m$, it gets assigned the mean value of that leaf.

The next question is how do we choose our splitting variables and the values at which to split. The goal is to model the data as accurate as possible with the tree. Thus the tree needs some

kind of formula to determine which independent variable to use for splitting and at which value to split, the splitting rule. The goal is to split the data such that the difference between the mean of the data before the split and the data is decreased the most by splitting the data into two partitions and modelling those partitions with their respective means. This way splitting the data will always improve the accuracy of the model. Define the leaves $l_1$ and $l_2$ of a certain set of data as:

$$l_1(j,c) = \{t \in l | S_{t,j} \leq c\} \quad \& \quad l_2(j,c) = \{t \in l | S_{t,j} > c\}. \tag{2}$$

So $l_1$ contains all observations with time t for which $S_{t,j} \leq c$. The difference between the mean and the observation is calculated with the squared residual. The splitting rule for a certain set of data then is defined similar as in Coulombe (2020):

$$\min_{j \in \mathcal{J}, c \in \mathbb{R}} \left[ \sum_{t \in l_1} (y_t - \mu_1)^2 + \sum_{t \in l_2} (y_t - \mu_2)^2 \right], \tag{3}$$

where $j \in \mathcal{J}$ denotes the splitting variable, $\mathcal{J}$ is the set of all splitting variables we consider and $c$ is the splitting point. $\mu_1$ and $\mu_2$ are the means of the partitions $l_1$ and $l_2$ respectively and $y_t$ is the dependent variable. Now we can see the data in a certain leaf as a new set of data. We can again apply the splitting rule on that data to again split the data into two more leaves. So, we can repeat the splitting process on the two partitions of data we obtain from the split before such that we partition those partitions further, that is split $l_1$ and $l_2$ further into four new partitions. We repeat this process until certain stopping conditions are met, so $l_1$ and $l_2$ are recursively split until we have the M partitions $l_1, ..., l_M$. Often a minimum amount of observations left in a leaf is defined as the stopping condition. So when there are less than a certain amount of observations left in a leaf, we choose not to split the data anymore and it becomes a terminal leaf.

A decision tree does, however, have some issues. When the data has very linear characteristics, linear estimation models will exploit the linearity of the data to fit the model, which often gives a better and more efficient fit. Decision trees, however, are not able to recognize the linear characteristics and therefore will not exploit this in the estimation. Another problem is the bias-variance trade-off of trees. Running the tree deep such that it splits the data into many very small parts will give a very good fit to the data. Ultimately you can let a tree split the data until each terminal leaf contains its own single data point. However, using such a tree for prediction will often result in bad predictions. Namely, the tree seems to be perfectly fit to the data we used to estimate, but the data we want to predict is different. This is called overfitting. We could let the tree not run deep, but this then creates a bias in the estimation and prediction. Ultimately we want a model which fits well with the data but avoids overfitting. A tree is also not very robust. Small changes in the tree could impact its outcomes significantly. (Hastie et al., 2001)(James et al., 2013)

### 4.2 Random Forest

A RF (Breiman, 2001) is an improved version of the decision tree based model. A RF makes use of bagging (Breiman, 1996) and decorrelated trees (Breiman, 2001) to create a model with

low variance and low bias. With these procedures, we create multiple trees which run deep. The trees are required to run deep to get obtain a low bias. We take the average of the results from the trees to get the final result.

Bagging is the procedure of generating multiple different versions of our predictor and using those to calculate a mean of all those predictors (Breiman, 1996). In this case, the trees are the predictors. So, we need to generate many different versions of the tree. This is done by taking many bootstrap replications of the same size or a bit smaller than the original dataset from the original dataset and using these replications to create new trees. The size is determined by the subsampling rate, which is usually around 0.75. The original dataset in this case can also be called the training set, as it is used to train the trees.

To further explain, take a certain dataset, which we call the original dataset. Then we create new datasets by randomly picking observations, which are all the variables at a certain time $t$ ($Z_t = (Y_t, X_t, S_t)$), from the original dataset, these are called the bootstrapped replications. (Breiman, 1996) We are allowed to pick with replacement. So, the bootstrapped replications do not necessarily contain every observation from the original dataset and can contain some observations multiple times. On every new bootstrapped replication we build a new decision tree using the same methods described in the section before. The stopping condition is set such that the tree runs deep. Eventually, we have $B$ bootstrapped replications with each their own decision tree fit to that replication. So, we end up with $B$ different trees.

Now, these trees are used to predict the dependent variable. We can run the parameters through all the trees and get $B$ predictions from them. We then average these predictions to get our final prediction. Breiman (1996) proves in his paper that the mean square error of this prediction is always smaller or equal to the prediction from the original dataset. With bagging we can view the trees as identically distributed (i.d.) variables. (Hastie et al., 2001) This means that the expectation from each tree is the same, so there is no bias. Since the trees are quite noisy, they benefit greatly from averaging them, so the average of the trees is a better predictor. The variance decreases through averaging. The formula for the variance of the average of i.d. variables, with each variance $\sigma^2$ and correlation $\rho$, is:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \leq \sigma^2. \tag{4}$$

It is clear that the variance decreases if we use the average of several i.d variables. When $B$ becomes large the second term disappears, however the first term stays. Clearly, the lower the correlation, the lower the first term. As such, we can benefit greatly from decorrelating our i.d. variables which in this case are our trees. (Hastie et al., 2001) Breiman (2001) proposes to grow our trees semi-stochastically to keep the correlation between the trees low. His idea is to consider a different subset of parameters at every splitting point in the tree. This is achieved by randomly selecting a fraction $m$ of the variables, usually $\frac{1}{3}$. So in equation 3 we take $j \in \mathcal{J}^-$ instead of $j \in \mathcal{J}$. This means we take $j$ from a different subset than the subsets used before. This ensures that the trees are grown randomly. Without this, the trees will often choose the

strongest predictor for the split and as such the correlation increases. But now the trees are grown quite randomly and as a result this decreases the correlation and therefore the variance. It also improves the computational speed. Now the algorithm to grow the forest can be presented.

**Algorithm 1** Random Forest algorithm

**for** $b = 1, ..., B$ **do**
(a) Draw a bootstrap sample from the original dataset.
(b) Grow the tree on the bootstrapped data by recursively repeating the following process for each node in the tree until the minimum node size is reached for each node.
1) Randomly select $m$ of the variables from the total set of variables.
2) Pick the best variable and splitting point according to the splitting rule.
3) Split the node into two new nodes.

**end for**
Output random forest $\mathcal{F}$ with $B$ trees.

Now we can present a RF model. A basic time-dependent RF looks like this:

$$\begin{aligned} y_t &= \mu_t + \epsilon_i, \\ \mu_t &= \mathcal{F}(S_t). \end{aligned} \tag{5}$$

In this $y_t$ is our dependent variable. This variable is predicted by the scalar value $\mu_t$. So, our dependent variable is estimated and predicted with a time-varying mean. This time varying mean is obtained from the forest $\mathcal{F}(S_t)$ where $\mathcal{F}$ denotes the forest of $B$ trees, which is created using the algorithm above, and $S_t$ is the set of all available variables. Each tree $b$, $b = 1, ..., B$, will compute its own $\mu_{b,t}$ using $S_t$ and then $\mu_t$ is calculated with $\frac{1}{B}\sum_{b=1}^{B}\mu_{b,t}$. The splitting rule and outcome used for the trees is the same as in 2.1.

In contrast to decision trees RF do not overfit when they run deep because of the Law of Large Numbers (Breiman, 2001). This does mean that $B$ needs to be large enough. According to Hastie et al. (2001) 200 is enough. The randomness and the large number of trees that run deep, keep the bias and variance low.

### 4.3 Local Linear Forests

While RF is a good predictor with low bias and variance, its weakness is the inability to capture smooth trends. The RF uses a time-varying mean which as a result creates a kind of model which is close to a step function when smooth trends are present. Friedberg et al. (2021) improved the RF model by introducing a local linear trend with the RF. They named it Local Linear Forests. This model tries to model the smaller partitions of data with a linear model instead of the mean. To each partition a small linear model is fit with OLS. Instead of a time-varying mean, it calculates a time-varying trend with the smaller partitions. This improves the smoothness of a RF and this model is able to capture linear trends. The general model for a local linear forest

is:

$$\begin{aligned} y_t &= X_t\beta_t + \epsilon_t, \\ \beta_t &= \mathcal{F}(S_t). \end{aligned} \tag{6}$$

Here $y_t$ is the dependent variable, $X_t$ is a vector containing our independent variables and $\epsilon_t$ is the residual. The independent variables are also contained in $S_t$. The local linear forest estimates the Generalized Time-Varying Parameters (GTVPs) $\beta_t$. This is a vector of the same length as $X_t$ containing the parameter estimates which can vary over time for each independent variable. The $\beta_t$ is estimated using the local linear forest $\mathcal{F}(S_t)$, where $\mathcal{F}$ represents the forest and $S_t$ the set of all available variables.

This new RF model also comes with a new splitting rule which needs to be used to build the trees according to algorithm 1. The data should be split such that it minimizes the sum of squared residuals. This time the residual, which needs to be minimized, is defined differently. The splitting rule is defined similar as in Friedberg et al. (2021):

$$\min_{j\in\mathcal{J}^-,c\in\mathbb{R}}\left[\min_{\beta_1}(\sum_{t\in l_1}(y_t - X_t\beta_1)^2 + \lambda\|\beta_1\|_2) + \min_{\beta_2}(\sum_{t\in l_2}(y_t - X_t\beta_2)^2 + \lambda\|\beta_2\|_2)\right]. \tag{7}$$

Here the $\mathcal{J}^-$ is a random subset of variables from all the observed state variables. Here again, we choose the state variable $j$ from a random subset of state variables $\mathcal{J}^-$ such that we decorrelate the trees and we choose a value $c$, the splitting point, which minimises the 2 least squares equations within. $\lambda$ is the Ridge regularization parameter. $y_t$ is our dependent variable and $X_t$ our matrix of independent variables with their vector of estimated coefficients $\beta$. The least squares minimization is now formulated as with OLS. We choose the $\beta_1\&\beta_2$ such that they minimize the residuals. Eventually, each leaf contains its own $\beta_t$ vector. The data is split such that we minimize the squared residual by fitting different regressions to either side of the split.

This splitting formula also contains a Ridge penalty function. This function helps the model to regularize. The Ridge penalty shrinks the $\beta_t$s towards zero. The higher the $\lambda$, the faster the $\beta$s are shrunk. It helps avoid the trees from overfitting locally in the leaf as less useful coefficients are shrunk to 0. (Friedberg et al., 2021). It also helps in the case of correlated covariates in the regression in the leaf. When the independent variables in the regression are correlated, two correlated variables can, for example, cancel each other out. The coefficient of one variable can be highly positive, while the coefficient for the other variable can be highly negative. The Ridge shrinkage will ensure that this is penalized and that only the coefficient for, for example, one variable is calculated. If we set $X_t = 1$ we return to the standard RF model with a Ridge shrinkage.

### 4.4 Macroeconomic Random Forests

Coulombe (2020) proposed an improvement for the local linear forest of Friedberg et al. (2021). He argued that $\beta_t$ should smoothly transition to its neighbours $\beta_{t+1}$ & $\beta_{t+2}$. As Coulombe (2020) said in his paper, "This is in line with the view that economic states last for at least

a few consecutive periods". This means that when estimating the $\beta_t$ at a certain time t, its neighbours $\beta_{t-2}$, $\beta_{t-1}$, $\beta_{t+1}$ and $\beta_{t+2}$ should be taken into account in the estimation process. It comes down to shrinking $\beta_t$ to be close to its neighbours. Coulombe (2020) calls this shrinkage the random walk regularization. This is the Macroeconomic Random Forest (MRF) model of Coulombe (2020).

Coulombe (2020) implements the random walk regularization by using a small rolling window view in the estimation in the leaves and the splitting rule. Instead of solving a least squares problem in the splitting rule and using a normal OLS model in the leaves, he proposes to solve and use a small weighted least squared (WLS) problem which will take into account the neighbours of the observation but with smaller weights. For the weights of the WLS, he uses a symmetric 5-step Olympic podium. Which puts a weight 1 on observation $t$, weight $\zeta < 1$ on observations $t+1$ and $t-1$ and weight $\zeta^2$ on observations $t+2$ and $t-2$. Since in the sum of the splitting rule some $t$'s will occur multiple times, he takes the maximum weight allocated to that observation. For example, when $t$ and $t+1$ both occur in a leaf, they get weight $\zeta$ from each other and weight 1 from themselves. Then the maximum weight is taken, so they both get assigned a weight of 1.

To define the Olympic podium, we first need to define the lags of the leaves. Define $l_{-1}$ as the lagged version of the leaf and $l_{-2}$ as the second lag. $l_{+1}$ and $l_{+2}$ are the one-step and two-step forwarded versions of the leaf. The lagged version of a leaf contains all the time observations in a certain leaf lagged with one step. So if, for example, $l_8$ contains $t = 4, 7, 9$, then the lagged version of this leaf contains $t = 3, 6, 8$. The symmetric 5-step Olympic podium is then defined as in Coulombe (2020):

$$w(t;\zeta) = \begin{cases} 1 & \text{if } t \in l, \\ \zeta & \text{if } t \in (l_{+1} \cup l_{-1}) \setminus l, \\ \zeta^2 & \text{if } t \in (l_{+2} \cup l_{-2}) \setminus (l \cup (l_{+1} \cup l_{-1})), \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

To define the splitting rule let us first redefine the definition of the leaves in equation 2. We define our random walk regularized leaf $l_i^{RW}$ for $i = 1, 2$ as:

$$l_i^{RW}(j,c) \equiv l_i(j,c) \cup l_i(j,c)_{-1} \cup l_i(j,c)_{+1} \cup l_i(j,c)_{-2} \cup l_i(j,c)_{+2}. \tag{9}$$

Now we are able to define the new splitting rule. The splitting rule for a MRF is defined as:

$$\begin{aligned} \min_{j \in \mathcal{J}^-, c \in \mathbb{R}} [ & \min_{\beta_1} \Big( \sum_{t \in l_1^{RW}(j,c)} w(t;\zeta)(y_t - X_t\beta_1)^2 + \lambda \|\beta_1\|_2 \Big) \\ & + \min_{\beta_2} \Big( \sum_{t \in l_2^{RW}(j,c)} w(t;\zeta)(y_t - X_t\beta_2)^2 + \lambda \|\beta_2\|_2 \Big) ]. \end{aligned} \tag{10}$$

The splitting rule now minimizes a weighted least squares problem. The weight assigned to each observation t is defined in equation 8. The splitting rule is almost the same as in equation 7 ex-

cept the addition of WLS. The definitions of $j$, $\mathcal{J}^-$, $c$, $y_t$, $X_t$ and $\lambda$ is the same as in equation 7. When an observation gets assigned multiple weights, the maximum weight is taken. The Ridge shrinkage is still in the function. This means this formula now has 2 forms of regularization, Ridge and random walk (Coulombe, 2020). When we set $\zeta = 0$ we return to the local linear forest model. In every leaf there is now the same type of WLS model as used in the splitting rule. For all $t$ in a certain leaf, all 4 observations around $t$ are now also used for estimating the GTVPs ($\beta_t$s) but with smaller weights.

The MRF uses a more sophisticated bootstrap procedure than the regular RF. RF takes single observations and reproduces a new bootstrap replication of the original dataset. Coulombe (2020) proposes Block Bayesian Bootstrap,based on the Bayesian bootstrap method of Rubin (1981) and the Block bootstrap (Mackinnon, 2006). The Bayesian bootstrap method uses a Dirichelet distribution to generate weights for each observation in the original dataset. We can then generate multiple random sets of weights and use these to compute a new type of observation. With enough repetitions of this, the new type of observation will simulate a population which has a distribution close to the original one. We can take random draws from this distribution to generate a new replication of the dataset, which we can use for the fitting of the tree. With Block bootstrapping, instead of taking single observations with bootstrapping, you take blocks of observations. These blocks will be of size $s = \frac{T}{\#Blocks}$. So instead of $Z_t = (Y_t, X_t, S_t)$, we now have $Z_b = (Y_b, X_b, S_b)$, where b represents a block of size $s$, which includes multiple $t$'s which come after each other. Combining these two creates Block Bayesian Bootstrap. This means we draw blocks from a distribution determined by Block Bayesian Bootstrap to create a new replication of the dataset. Coulombe (2020) states that it is better the use the Block Bayesian Bootstrap for forecasting. However, when estimating and modelling, it is better to use regular Block Bootstrap as it is faster and computationally easier.

Setting $X_t = \imath$, $\lambda = 0$ and $\zeta = 0$ we return to the standard RF model described before. A standard RF is not good at capturing linear relationships. A standard RF will model a linear trend as some sort of step function. The RF will likely waste many splits trying to model the linear relationship and at the end will not have many left to focus on the non-linear relationship (Coulombe, 2020). The MRF is able to model long- and short-term relationships and therefore will have enough splits left to also focus on the non-linear relationships.

The MRF has similarities with OLS. Instead of estimating a scalar value like a RF, the MRF tries to estimate a linear model for each data point. OLS estimates a linear model for all data points. The difference between the OLS and MRF estimate is that the $\beta$ is able to vary over time in a MRF. This gives the MRF an advantage over OLS when the $\beta$ varies over time. Furthermore if we compare the MRF to classical time-varying model, where the variation of $\beta$ is determined by time itself, it is found that the MRF is able to investigate all time-varying variables and determine the time variation of $\beta$ based on that and not just on time. This gives MRF the opportunity to better identify when the time variation occurs. In a case of a recession for example, the MRF is able to identify the change of the $\beta$ by the change of other variables.

By making the $\beta$ dependent on other variables, the MRF is also able to adjust to the market as the linear model within is flexible.

### 4.5 Dense and Sparse Dimensionality reduction

The construction of $S_t$ is also of importance. With so many variables, lags and transformations at one's disposal, the size of $S_t$ can quickly become quite large. Often the number of predictors becomes larger than the number of observations. This leads to statistical dimensionality problems with RFs (Coulombe, 2020).

There are two different dimensionality reduction methods, sparse and dense. Sparse dimensionality reduction methods select a smaller number of variables out of the total pool of all available variables. An example of this is Ridge and LASSO, where the values of the $\beta$s are penalized and as such the model is forced to select fewer variables for estimation. Dense dimensionality reduction methods summarize the data in a set of factors that span most of the regressors. Often it is required to include one of them. The MRF, however, is a regularized model and in that case both can be included and the algorithm will select the optimal combination between the two.

The MRF model already includes the Ridge penalty which is the sparse dimensionality reduction technique. The model does still need dense dimensionality reduction. The goal is to summarize the available information from our variables in fewer factors. For each variable, we can take multiple lags. If there is residual autocorrelation left, we might want to include more lags. This, however, increases the size of $S_t$ quickly. So, a solution is to summarize multiple lags in one variable. Coulombe (2020) proposes using Moving Average Factors of the lag polynomial of a specific variable. Let us consider a panel of $P$ lags of variable $j$:

$$X_{t,j}^{1:P} \equiv [X_{t-1,j}, ..., X_{t-P,j}]. \tag{11}$$

Then we want weighted averages of the lags such that we can extract the most information of the $P$ lags of variable $j$. These weighted averages can be extracted with Principal Component Analysis and taking the first few factors. Through this, the recent lags of $X_{t,j}$ are summarized in a few variables and use these instead of many lags of the regressor.

### 4.6 Forecasting

The MRF will be used to forecast the dependent variable. The forecasting horizons used are 1,2 and 4 quarters in the extension part. For the forecasts an expanding window estimation will be used in which the model will be re-estimated every two years (or every eight observations). I will use direct forecasts instead of repeatedly iterating the one-step-ahead forecast. This means the model is directly fit to $h$ periods ahead. The formula for direct forecasts is:

$$y_{t+h} = \beta_t X_t + \upsilon_{t+h} \tag{12}$$

$X_t$ is the matrix of the independent variables and stays the same for every $h$. $\beta_t$ is the vector containing the parameters and $\upsilon$ the error. So, with direct forecast we fit the same model but

just to different observations of the dependent variable. In the case of a AR(1), for $h = 1$, our model is $y_t = y_{t-1} + \upsilon_t$ or $y_{t+1} = y_t + \upsilon_{t+1}$. In the case of $h = 4$, the model is $y_t = y_{t-4} + \upsilon_t$ or $y_{t+4} = y_t + \upsilon_{t+4}$.

The forecasts will be evaluated with the Root Mean Squared Prediction Error (RMSE). For the out-of-sample (OOS) forecasts at time $t$ for model $m$ and forecast horizon $h$, the OOS RMSE is computed as:

$$RMSE_{h,m} = \sqrt{\frac{1}{\#OOS} \sum_{t \in OOS} (y_t - \hat{y}_t^{h,m})} \tag{13}$$

Where $\hat{y}_t^{h,m}$ is the $h$-step ahead prediction of model $m$ and $\#OOS$ is the number of out of sample observations. The Diebold & Mariano (2002) (DM) test statistic is used to compare the predictive accuracy of the model against the benchmark which in this paper is the OLS estimation of the model used in the MRF. So in case the independent variables are an intercept and the first 2 lags of the dependent variables, the benchmark is an AR(2). This means that the only difference between the models is how they are estimated. This will give a clear indication of whether a MRF can estimate a model and predict with it significantly more accurately than OLS. The null hypothesis of the DM test is that the two models have equal predictive accuracy. The alternative hypothesis is that the MRF has a lower RMSE than its OLS counterpart. If the RMSE of a MRF model is lower than the MSPE of its OLS counterpart and the Diebold Mariano test has a low p-value it means the MRF model had better forecasts.

### 4.7 Variable Importance Measures (VI)

The variable importance measures were originally proposed by Breiman (2001). As a RF is kind of a black-box model, it is hard to examine the driving variables behind the prediction. While the MRF does give some more insight with its GTVPs, it is still not possible to see which variables drive the prediction. VIs are able to give insights in the driving variables behind the prediction.

Originally Breiman (2001) proposed the out-of-bag VI. Coulombe (2020) used 3 different VIs, the out-of-bag VI ($VI_{OOB}$), the out-of-sample VI ($VI_{OOS}$) and VIs for the GTVPs ($VI_{\beta_k}$). VIs are calculated by randomly permuting all observations of a certain variable $j$ and then examining what influence this has on the estimation error. The more the error worsens, the larger the influence of this variable was for estimation. The VI is measured in relative RMSE gains from including the predictor versus not including the predictor.

The ($VI_{OOB}$) calculates the error for the out-of-bag samples for every tree. The out-of-bag samples are the set of observations not used for the creation of a certain tree. Earlier it was explained that a tree is created from a bootstrapped sample. This bootstrapped sample does not contain certain observations from the original dataset, these observations are called out-of-bag. So, the observations of a certain variable $j$ are permuted and then we use the out-of-bag observations to calculate the gain in error. This can be used to evaluate the importance of a

certain variable on the accuracy in the estimation of a certain model. The ($VI_{OOS}$) is similar to the ($VI_{OOB}$). The difference is that it uses the observations out-of-sample to calculate the RMSE. With the $VI_{OOS}$ we can evaluate the importance of a certain variable on the prediction of a model. The ($VI_{\beta_k}$) is also similar to the ($VI_{OOB}$). The difference is that it uses a different loss-function. The loss function used in the ($VI_{\beta_k}$) calculates how much the path of a GTVP ($\beta$) is changed. This measure is used to evaluate the importance of a variable in the estimation of a GTVP.

### 4.8 Models, $S_t$ and Tuning Parameters

Table 2: *All the models that are used in this paper.*

| Macroeconomic Random Forest models | Acronym | OLS counterpart model | Acronym |
|---|---|---|---|
| Autoregressive Random Forest | ARRF | Autoregressive Model with first 2 lags | AR(2) |
| Factor-Autoregressive Random Forest | FA-ARRF | Factor Autoregressive Model | FA-AR |
| Vector Autoregressive Random Forest 1 | VARRF1 | Autoregressive model with lags and three variables | VAR1 |
| Vector Autoregressive Random Forest 2 | VARRF2 | Autoregressive model with just three variables | VAR2 |

In the table above, the different models which are used in the thesis are visible. On the left side we have the MRF estimated models we use and on the right side we have the OLS counterparts of the models. As described earlier the MRF models will be compared with its OLS estimated counterpart to investigate whether an MRF can yield significant forecasting gains over OLS estimation. A note on the VARRF/VAR model. Those models are not really vector autoregressive but this is how Coulombe (2020) named this type of model, so I stick to the same name. The difference between VARRF1/VAR1 and VARRF2/VAR2 is that the first two include lags of the dependent variable while the other two do not. The equations for the models are presented in the table below.

Table 3: *Equations of all the models.*

| Acronym | estimated with | Equation |
|---|---|---|
| ARRF | MRF | $y_t = \mu_t + \phi_{1,t} y_{t-1} + \phi_{2,t} y_{t-2} + \epsilon_t$ |
| FA-ARRF | MRF | $y_t = \mu_t + \phi_{1,t} y_{t-1} + \phi_{2,t} y_{t-2} + \gamma_{1,t} F_{1,t} + \gamma_{2,t} F_{2,t} + \epsilon_t$ |
| VARRF1 | MRF | $y_t = \mu_t + \phi_{1,t} y_{t-1} + \phi_{2,t} y_{t-2} + \beta_{1,t} IR_{t-1} + \beta_{2,t} IF_{t-1} + \beta_{3,t} EA_{t-1} + \epsilon_t$ |
| VARRF2 | MRF | $y_t = \mu_t + \beta_{1,t} IR_{t-1} + \beta_{2,t} IF_{t-1} + \beta_{3,t} EA_{t-1} + \epsilon_t$ |
| AR(2) | OLS | $y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$ |
| FA-AR | OLS | $y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \gamma_1 F_{1,t} + \gamma_2 F_{2,t} + \epsilon_t$ |
| VAR1 | OLS | $y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \beta_1 IR_{t-1} + \beta_2 IF_{t-1} + \beta_3 EA_{t-1} + \epsilon_t$ |
| VAR2 | OLS | $y_t = \mu + \beta_1 IR_{t-1} + \beta_2 IF_{t-1} + \beta_3 EA_{t-1} + \epsilon_t$ |

The first thing to note is that the coefficients $(\phi, \gamma, \beta, \mu)$ of all the OLS estimated models are not time-varying. In the table this is easily visible as the coefficients in the OLS estimated models do not depend on $t$. This is because the MRF does estimate time-varying parameters but OLS does not. $y_t$ is the dependent variable which in this thesis is the house price index. $y_{t-1}$ and $y_{t-2}$ are its first and second lag respectively. $\mu_t$ is the time-varying intercept of the MRF models and $\mu$ is the fixed intercept from the OLS models. $\phi_{1,t}$ and $\phi_{2,t}$ are the time varying coefficients estimated by the MRF models for $y_{t-1}$ and $y_{t-2}$ respectively. $\phi_1$ and $\phi_2$ are the fixed OLS estimated coefficients for $y_{t-1}$ and $y_{t-2}$ respectively. $F_{1,t-1}$ and $F_{2,t-1}$ are

the first lags of the first and second traditional factors of the FRED-QD dataset computed by PCA. These factors are also described in the data section. $\gamma_{1,t}$ and $\gamma_{2,t}$ are their time-varying coefficients respectively estimated using MRF and $\gamma_1$ and $\gamma_2$ are their fixed coefficients estimated using OLS. $IR_{t-1}, IF_{t-1}, EA_{t-1}$ are the first lags of interest rate, inflation and economic activity respectively. More detailed information on these variables is available in the data section. These variables are chosen based on Adams & Füss (2010). $\beta_{1,t}$, $\beta_{2,t}$ and $\beta_{3,t}$ are their time-varying coefficients respectively estimated using MRF, while $\beta_t$, $\beta_2$ and $\beta_3$ are their fixed coefficients estimated using OLS. In all the equations $\epsilon_t$ represents the error term.

Table 4: *Composition of $S_t$.*

| What? | How? |
|---|---|
| 8 lags of $y_t$ | - |
| 2 lags of all variables in FRED-QD | - |
| Trend $t$ | - |
| 8 lags of first five traditional factors of FRED-QD | PCA on whole FRED-QD dataset |
| 2 MAFs of each variable in FRED-QD | PCA on 8 lags of each variable |

Next, the composition of $S_t$ is discussed. Coulombe (2020) shows that the MRF is able to handle a lot of data and that using a lot of data improves the accuracy. Therefore $S_t$ is going to be large. The composition of $S_t$ will be similar as Coulombe (2020) described it. In table 4 the exact composition of $S_t$ can be seen. The composition contains the first five traditional factors described in the data section, which contain a lot of information on the whole FRED-QD dataset. It also contains MAFs of each variable to help with dimensionality problems. The economic activity variable will be considered as a variable, therefore it gets its own MAFs and will be included in the computation of the traditional factors. $S_t$ can not contain variables with empty observations. These therefore need to be deleted. As there are many missing observations at the beginning of the dataset, the size of $S_t$ depends on the dependent variable. The later the dependent variable begins, the more variables you can keep. For the deleted variables the MAFs will also not be included. In the end, the size of $S_t$ for the house price index is 987 variables and the size for the unemployment rate is 907.

Then the tuning parameters. For the estimation of the MRFs the package macrorf created by Coulombe (2020) will be used in R. This package has many different options and parameters for estimation. Coulombe (2020) states in his paper that none of the parameters were tuned as this yields miniscule performance gains. He states that the importance is in the linear part. Furthermore, runtimes with an expanding window estimation for several models are more than a day, so tuning is hard with those times. So, we keep most of the parameters at their standard value assigned by the model. So the minimal node size is 10 and the fraction of variables to consider for the split ($m$) is 0.33. The subsampling rate is the fraction to determine how big the bootstrapped dataset should be compared to the original one, this tuning parameter is set at 0.75. The random walk regularization ($\zeta$), which is used for the Olympic podium, is also set at 0.75. The Ridge lambda ($\lambda$) in the splitting rules is set to 0.1. Then there is a parameter which determines the least amount of observations a node may contain, less is not allowed. If there is a situation in which the split creates a leaf with less, a different split is considered.

This is called the minimum leaf fraction and it is multiplied by the number of regressors in the regression to determine the minimum amount of observations in the leaf. This is set to 1. For example for the ARRF model it means that each leaf cannot contain less than 3 observations.

Some parameters are changed. The first one is the block size for the block bootstrap methods. Coulombe (2020) states that for quarterly data 8 is a good size, this is two years. Furthermore, he states that Bayesian Block bootstrap yields better forecasts, so for this extension that type of bootstrap is used instead of the standard block bootstrap the model normally uses. The standard block bootstrap is, however, used for the replication as Coulombe (2020) does this too. I will also set the number of trees ($B$) to 100 to get more accurate results. For the replication, it is kept at its standard value 50, which Coulombe (2020) uses too. Furthermore, he has a parameter which increases the probability that the trend is included as a potential splitting variable, he states that a reasonable value for this is 4. So it is changed to 4. For the replication it is kept at 1. Lastly, an important thing to note is that Coulombe (2020) uses a fast random walk regularization. This means the algorithm only considers the random walk regularization in the estimation of the leaves and not in the splitting rule. It is possible to disable this, however run times then increase a lot. So I use this both for the extension and replication.

# 5 Analysis and Results

In this section, a small replication is done on the Unemployment rate results of Coulombe (2020). After that, a thorough analysis is done on the MRF estimations of the House Price Index. First, I do a small replication and discuss the similarities and differences. After that, I discuss the model of Adams & Füss (2010). Then I forecast the house price index using different MRF models. At last I discuss the model that performs best in forecasting, the ARRF. I investigate the GTVPs and the VIs of that model.

## 5.1 Replication of Unemployment rate

In this section, a small replication of the results in Coulombe (2020) will be done. After that, the results will be discussed. Because the runtimes of a MRF expanding window forecast are quite large the replication of forecasts is only done for the variable unemployment rate and the two best MRF models, ARRF and FA-ARRF. For the forecasts Coulombe (2020) uses an AR(4) model as benchmark, as such I will do the same. The pseudo-out-of-sample period starts in 2003Q1 and ends in 2014Q4.

Table 5: *Forecast MSPE for ARRF and FA-ARRF relative to AR(4) model MSPE.*

| | ARRF | FA-ARRF |
|---|---|---|
| h = 1 | 0.9033*** | 0.9092 |
| h = 2 | 0.8848*** | 1.0902 |
| h = 4 | 0.8434*** | 0.8721** |
| h = 8 | 0.9034*** | 0.9820 |

*Note: ***, **, * represent significance at 1%, 5% and 10% respectively for the Diebold Mariano test*

Looking at the MSPEs of the ARRF relative to the AR(4), it is found that they are quite

similar but not exactly the same as in Coulombe (2020). The performance is, as in Coulombe (2020) very good, with very significant smaller MSPEs for every horizon. The performance I achieve actually seems significantly better than his performance for the ARRF. The results for the FA-ARRF differ a lot from his results. This difference will be discussed later in this section. The small difference for the ARRF can be caused by different things, which will be discussed below.

First of all, my methods for selecting which data to keep is built upon keeping as much variables as possible. It is not possible to have empty observations in certain variables, those variables need to be eliminated from the dataset. A lot of variables have empty observations in the beginning. I take advantage of the fact that the lags of the dependent variable will also have empty observations in the beginning. Using this, I know that certain $t$'s in the beginning are going to be eliminated anyway because of the lags of $y_t$. I also take into consideration the number of lags I need from the variables used to construct $S_t$. Combining these two things, I construct $S_t$ as large as possible without getting rid of too many observations. I suspect Coulombe (2020) uses a cheap approach by just deleting all variables with empty observations at $t = 1$, which is the time at which the first observation of $y_t$ occurs.

Secondly, it is unclear how Coulombe (2020) exactly transforms his data. It is not clear whether he transforms all his available variables to be stationary or just the dependent variables. Stationarized data and non stationarized data will contain different information. This could lead to differences in the estimation. It could be an improvement to include both in $S_t$.

Thirdly, I have more data available. I use the full dataset to compute my MAFs and traditional factors of the FRED-QD. It seems that Coulombe (2020) computes the factors once and then incorporates them into the data set. The same goes for the MAFs. This is the same as my approach. When this is the case, I use the full data set available and it seems he does the same. However, I have more data available, so this could generate small differences. However, it could be the case that he re-computes the factors every time the expanding window expands. It is unclear what Coulombe (2020) exactly does, which makes exact replication hard.

Lastly, it is machine learning, which is often seen as a black-box model. It is hard to see exactly what happens on the inside, which makes exact replication hard. Furthermore, the randomness used by the algorithm to decorrelate the trees will likely also give small deviations in the results.

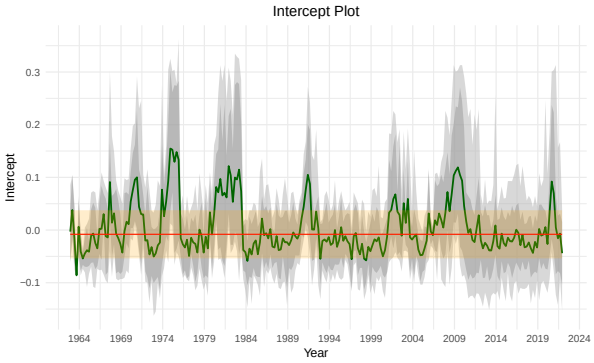


Figure 4: Intercept FA-ARRF

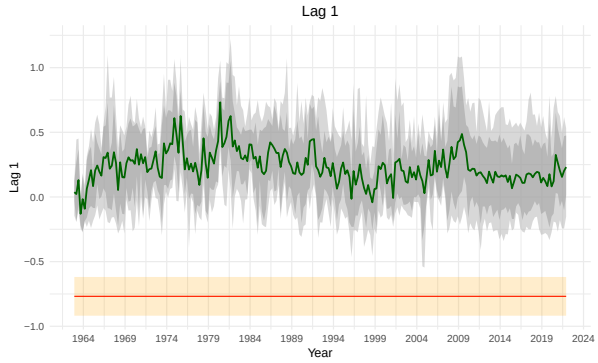


Figure 5: First lag FA-ARRF

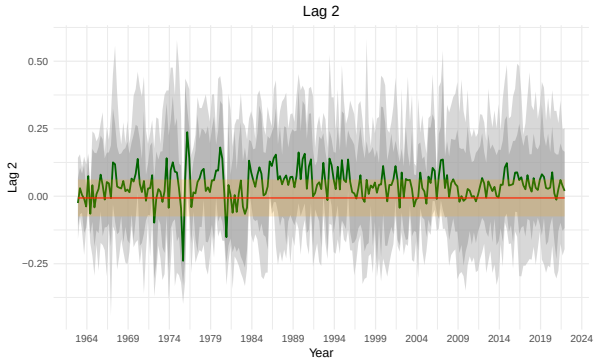


Figure 6: Second lag FA-ARRF

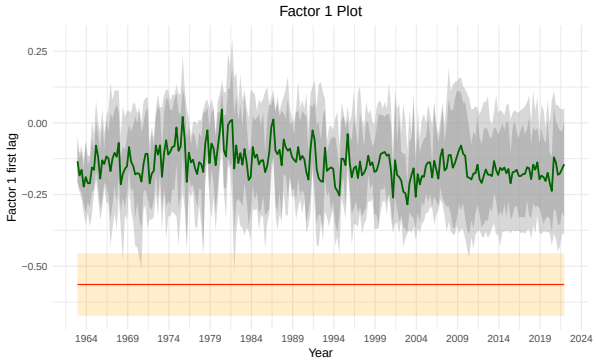


Figure 7: Second Factor FA-ARRF

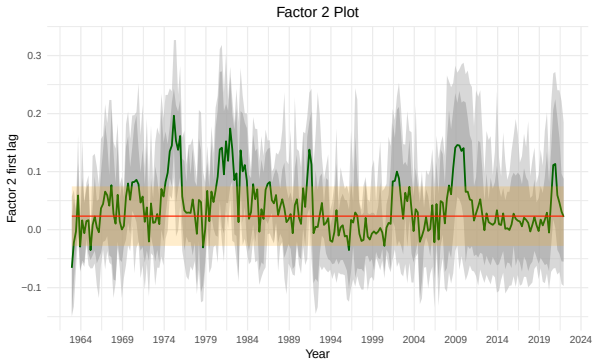


Figure 8: Second Factor FA-ARRF

*Note: The dark and light grey bands are the 68% and 90% confidence regions. The red line is the OLS coefficient and the orange band is the OLS coefficient ± one standard error.*

In the figures above, the graphs for the GTVPs of the FA-ARRF model are visible. These graphs are obtained by using a one-step-ahead forecast with $h = 1$ and estimating the model once at 2007Q2. The forecast and these graphs differ from those in Coulombe (2020). The intercept looks kind of similar but the GTVPs differ. This likely has to do with the fact that I, first of all, compute the traditional factors differently and secondly, I have more data available and therefore my factors changed. Coulombe (2020) says he uses usual PCA to compute the factors, however it is not said how he deletes the variables with empty data, or if he deletes outliers or not. This is also the reason why the forecasts MSPEs are different. However, I did compare my factors to those of McCracken & Ng (2020) and they are somewhat similar. So, this confirms my factors are not necessarily incorrect. Thus it is very likely the difference is due to difference

in computation.

## 5.2 Model of Adams & Füss (2010)

In this section, the model of Adams & Füss (2010) is discussed. In their paper they investigate the long-term impact and short-term dynamics of several macroeconomic variables. They choose their variables based on the statistical equilibrium model of DiPasquale & Wheaton (1996). The chosen variables are economic activity, long-term interest rates and construction costs. These variables influence the demand and supply of housing stock from which they derive a function for the house price.

The first variable, economic activity, is often represented by disposable income. However, Adams & Füss (2010) argue that disposable income does not link well to house prices as disposable income is a measure of average income while home buyers often have an income above that. Thus they construct economic activity by taking the first factor of the matrix of real money supply, real consumption, real industrial production, real GDP, and employment. They describe that an increase in economic activity positively shifts the demand curve for housing space. Since the supply of houses cannot increase in a short time, rents increase which in turn leads to higher house prices. So, economic activity influences the demand positively and therefore the house price positively.

For the following variable, long-term interest rate, Adams & Füss (2010) describe that it rather influences the demand to own a house than the demand for housing space. A higher long-term interest rate increases the demand for other assets and decreases the demand for real estate. It also increases mortgage rates which in turn decreases demand for owning a house even more. These two effects should decrease the house price. They describe this as an increase in the capitalization rate, which is the rents to house price ratio. This results in lower real estate prices and therefore in less construction which will result in even higher rents. The increase in rents will normally lead to higher house prices however because the capitalization rate changed, this is not the case. So, they expect the interest rate to influence the demand negatively and therefore the price negatively.

The last variable, construction cost, influences the supply schedule of new construction. An increase in for example construction materials is likely to influence the supply of new construction negatively. This in turn influences the supply of housing negatively. Less supply of housing will lead to less housing space which will increase rents and eventually increase house prices. (Adams & Füss, 2010)

Next Adams & Füss (2010) create a demand and supply function for the housing stock using the variables just described. This models the long-run housing market. The functions looks like:

$$\begin{aligned} D_t &= \alpha - \beta_1 hp_t + \beta_2 EA_t - \beta_3 long_t + \epsilon_t, \\ S_t &= \eta + \gamma_1 hp_t - \gamma_2 constr_t + \upsilon_t. \end{aligned} \tag{14}$$

As supply equals demand, we can equate the two functions and have the house price on one side and the other variables on the other side. The model becomes:

$$hp_t = \alpha^* + \beta_1^* EA_t + \beta_2^* constr_t + \beta_3^* long_t + \epsilon_t^*. \tag{15}$$

It is expected that $\beta_1^*$ and $\beta_2^*$ are positive while $\beta_3^*$ is negative.

As described in the data section, the variable construction costs will be exchanged for the regular consumer price index variable. The series from FRED-QD used for the other variables are also described in the data section. The variables I use are transformed to be stationary. This makes it hard to compare the MRF with the model of Adams & Füss (2010). Furthermore, the model also represents a long-run model, these long-run effects will likely not come to light with the first lags I use. However, it is still interesting to see if these variables could provide gains in forecasting, especially in longer horizons. As Case & Shiller (1990) state for example that the effects of construction costs in one year affect the house price the next year.

### 5.3 Forecasting results

In this section the forecast results for the all-transaction house price index in the United States are going to be discussed. The period forecasted is 2005Q1 to 2020Q4. For the forecasting process an expanding window approach is used, where the models are re-estimated every 2 years (8 observations). The forecasts for the MRFs are evaluated with their MSPEs relative to the MSPEs of its OLS counterpart. The significance is tested using the DM-statistic. (Diebold & Mariano, 2002)

Table 6: *MSPEs of different MRF models for the All-Transaction House Price Index for different forecast horizons using expanding window forecasting.*

| | ARRF | FA-ARRF | VARRF1 | VARRF2 |
|---|---|---|---|---|
| h = 1 | 0.0110 | 0.0115 | 0.0121 | 0.0131 |
| h = 2 | 0.0117 | 0.0122 | 0.0121 | 0.0132 |
| h = 4 | 0.0107 | 0.0113 | 0.0115 | 0.0136 |

Table 7: *MSPEs of different MRF models relative to the MSPE of its OLS counterpart model for the All-Transaction House Price Index for different forecast horizons using expanding window forecasting.*

| | ARRF | FA-ARRF | VARRF1 | VARRF2 |
|---|---|---|---|---|
| h = 1 | 0.9771 | 1.0324 | 1.0027 | 0.9297*** |
| h = 2 | 0.9678 | 1.0150 | 0.9910 | 0.9447** |
| h = 4 | 0.9536* | 0.9604* | 0.9855 | 0.9311* |

*Note: ***, **, * represent significance at 1%, 5% and 10% respectively for the Diebold Mariano test*

In table 6 and 7 the MSPEs of different MRF models and their MSPEs relative to the MSPE of their OLS counterpart are shown. All values in table 7 were tested with the Diebold Mariano test (Diebold & Mariano, 2002) to examine whether they are significantly better than their OLS counterpart. Examining table 7, we can see that all models are better or slightly

worse than their OLS counterparts. The ARRF is for every forecasting horizon better than its OLS counterpart. The VARRF2 is significantly better in all horizons. This indicates that the MRF could yield significant forecasting gains in predicting the housing market if we compare it to OLS. The FA-ARRF seems to be significantly better once and worse in the other cases.

Looking at table 6 we can see that the MSPEs lie very close to each other. But the ARRF model is clearly the best model, with the lowest MSPE in all forecast horizons. It looks like the house price index is primarily dependent on its past lags. The VARRF 2 is actually the worst in all cases. Doing a DM-test on the ARRF and VARRF2 with a two-sided alternative hypothesis, I do find that they are significantly different for $h = 1$ and $h = 4$ but not for $h = 2$. It is not bad to see that there actually is some predictive accuracy in the VARRF2 model. As the VARRF2 contains more useful information on the effects of macroeconomic variables on the house price index, it could still be helpful in real-life situations. Including the first two lags gives the VARRF1 model. This is also worse than the ARRF in all cases, indicating that adding those variables does not really contribute to the accuracy in forecasting. The factors of the FA-ARRF model also seem to not really contribute to the forecast accuracy.

Coulombe (2020) has results for forecasting housing starts in his appendix. He does not discuss them in his paper. The results for the housing starts in Coulombe (2020) are quite bad. He uses an AR(4) model as benchmark. Almost all MRF models in his paper are not able to predict housing starts better than the AR(4) in the forecasting horizons 1,2 and 4.

### 5.4 Analysis of ARRF

While the ARRF is only once significantly better than the AR(2), it does always give slightly better predictions for every horizon. It may therefore be interesting to dive deeper into this model. Furthermore, we could gain some insights into the main drivers of the house price index by investigating the VIs. So, in this section we are going to investigate the GTVPs and the VIs of the ARRF model.

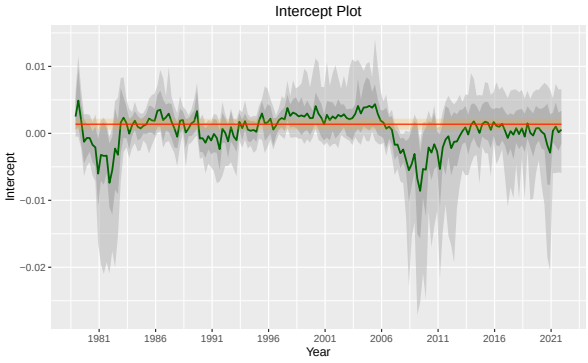


Figure 9: Intercept ARRF

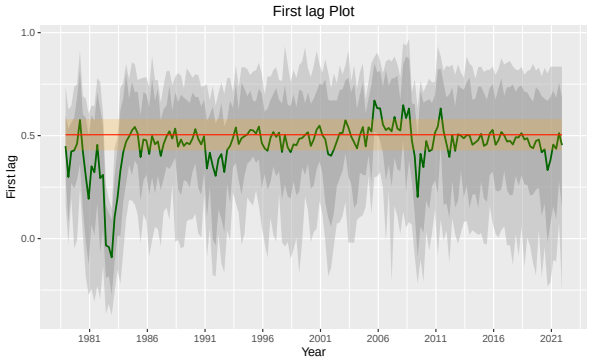


Figure 10: First lag ARRF

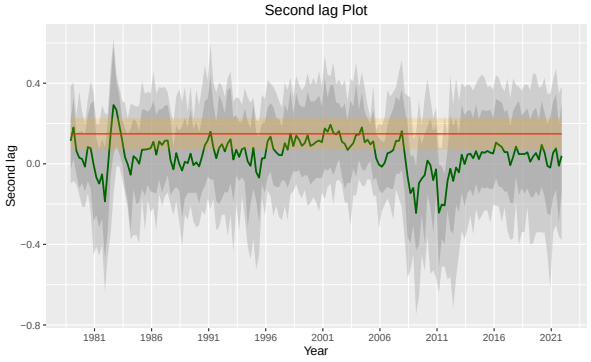


Figure 11: Second lag ARRF

*Note: The dark and light grey bands are the 68% and 90% confidence regions. The red line is the OLS coefficient and the orange band is the OLS coefficient ± one standard error.*

In the figures above, we can see the GTVPs for the ARRF model on the house price index. The graphs are obtained by the MRF function in R of Coulombe (2020) and using an out-of-sample period from 2012Q1 to 2021Q4. The forecasts are done with a one-step-ahead forecast for horizon 1 and estimating the model once at 2011Q4. Examining the graphs we see that the GTVPs are often within the one standard error of the OLS coefficient but during some periods leave this band. This directly indicates where the differences in forecasting accuracy come from. In all the graphs we also see a downward spike around the 2008 recession in the US, which is the period between 2006 and 2011. It looks like the MRF is able to adjust during the recessions where the AR(2) is not. We also see a small downward spike around 2021, which means the model was able to recognize something in the corona crisis even though it was not used in the estimation.

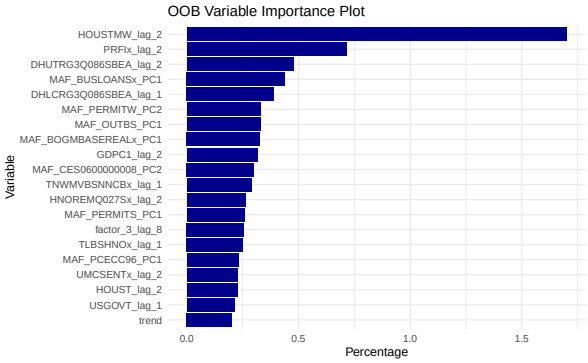


Figure 12: Variable Importance OOB

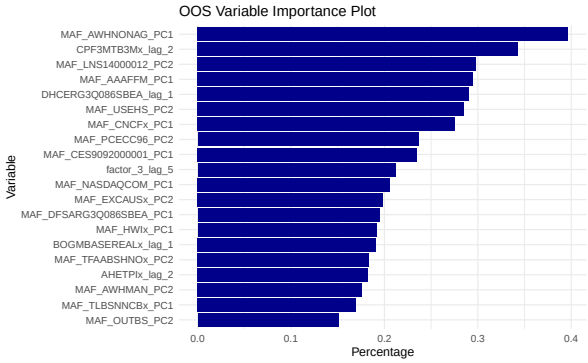


Figure 13: Variable Importance OOS

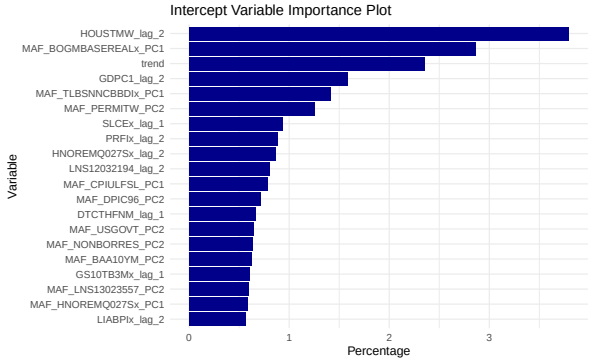


Figure 14: Variable Importance Intercept

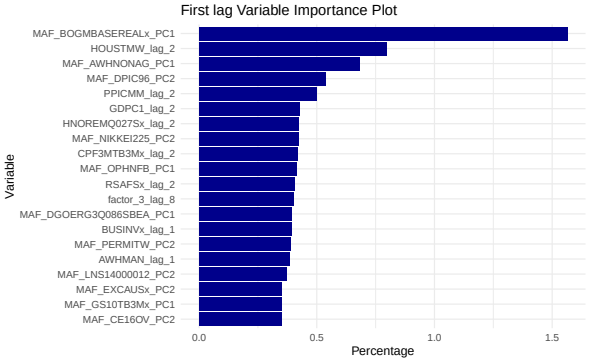


Figure 15: Variable Importance First lag

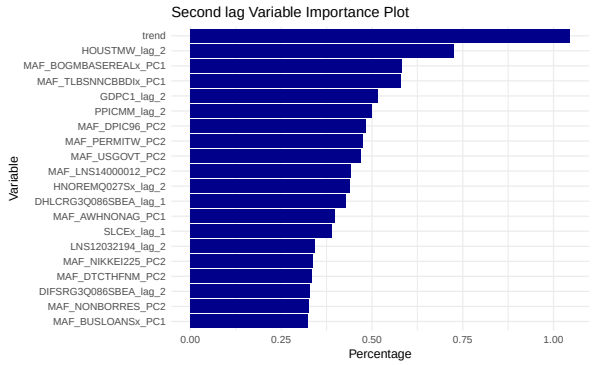


Figure 16: Variable Importance Second lag

In the figures above, the best 20 VIs for the out-of-bag sample, out-of-sample sample and the betas are displayed. The numbers displayed are relative gains in MSPE in percentage points when excluding the variable from the dataset. First for the oob-sample, we find that the second lag of HOUSTMW, which is the housing starts in the midwest, is the most important variable by far compared to the other variables, with a change of more than 1.5 percent. That means that for the accuracy in estimation of the model, the housing starts in the midwest two periods ago is the most important variable. Housing starts influence the supply of houses a few periods later, it is therefore quite logical that it influences the housing price two quarters later. Then observing the out-of-sample VIs we find that the most important variable is the first MAF of Average Weekly Hours Of Production And Nonsupervisory Employees: Total private. The increase in MSPE is quite low, with only about 0.4 percent. The other variables do also not seem that important.

For the VI of the intercept, we find that again the second lag of HOUSTMW is the most important variable. It alters the MSPE by more than 3.5 percent, which means it alters the path of the intercept quite a lot. Likely this relates to the importance of the HOUSTMW in the out-of-bag sample. For the first lag, the first MAF of the BOGMBASEREALx, which is the real monetary base, is the most important with 1.5 percentage points. We also find this variable in the OOB VI and the first lag instead of MAF in the OOS VI. While the real monetary base is often not used as the money supply variable, it strongly relates to it. This result is in line with Tripathi (2019), Garriga et al. (2019) and Adams & Füss (2010), whom all found a relation between money supply and the house price. For the second lag, the trend is the most important variable followed by the second lag of HOUSTMW and first MAF of real monetary base. The importance is not very big however. The importance of the trend, which represents time, could indicate some form of time variation. Investigating the GTVP of the second lag, it looks like there could be a structural change after 2008 which could be the cause of the importance of the trend.

# 6 Conclusion and Recommendations

In this thesis I use the Macroeconomic Random Forest of Coulombe (2020) to estimate and predict the US all transaction house price index. First, decision trees and random forest are explained to give insights into the workings of a random forest. Then I describe the local linear forest of Friedberg et al. (2021) to build to the macroeconomic random forest. The MRF is used to forecast the house price index. The MRF models used are ARRF, FA-ARRF, VARRF1 and VARRF2. The forecasts of these models are compared with their OLS counterpart and using the Diebold Mariano test (Diebold & Mariano, 2002) it is evaluated whether the MRF performs significantly better than its OLS counterpart. After that, the GTVPs and VIs of the ARRF model are investigated to gain further on the model.

From the results, I can conclude that the MRF is able to realize significant forecasting gains over OLS. In most of the forecast situations the MRF has a slightly lower MSPE than its OLS counterpart. For almost half of the situations it is significantly better, which is also what I hypothesized. The model of Adams & Füss (2010) or the VARRF2 performs quite well as an MRF in comparison to its OLS counterpart. For all forecast horizons it is significantly better than its OLS counterpart. However, it is the worst model in terms of MSPE. The VARRF1 is still not better than the ARRF in terms of MSPE, indicating that adding the variables of Adams & Füss (2010) does not really contribute to forecast accuracy. The VARRF1 and VARRF2 models are however not that bad and do have some ability to forecast. As I hypotesized I did not expect the VARRF1 or VARRF2 to realize gains over other MRFs as the house price index largely depends on past trends. For the GTVPs of the ARRF we find that they change over time and adjust in periods of recessions as I hypothesized. For the VIs I do find that the number of housing starts has a big influence, which is a variable which is highly correlated with house prices. However also other variables have come to light which I did not hypothesize. So to answer the main question, yes, it is possible to estimate and predict US house prices with a macroeconomic random forest.

The results in this thesis could be used for controlling the house market. The forecasting gains achieved could be of use to prevent future housing crises. The VIs and GTVPs could give insights on how to achieve this. However, there could still be a lot of improvement. Further research could be done on which kind of model would perform really well as a MRF. The models I used were quite basic, however with more research on the main drivers of the housing market a new type of equation could be created which would likely perform better than the ARRF. Furthermore, more research could be done on the construction of $S_t$. I build $S_t$ the same way Coulombe does. However, it could be argued that for example including the not transformed variables could give extra information as it gives information on whether we are in a high or low for example. Lastly, better methods to investigate the effects of certain variables could be used to gain more insights into the working of their effects. These insights can then can be used for policies concerning the housing market.

## References


Adams, Z., & Füss, R. (2010). Macroeconomic determinants of international housing markets. *Journal of Housing Economics*, *19*(1), 38-50. Retrieved from https://www.sciencedirect.com/science/article/pii/S1051137709000552 doi: https://doi.org/10.1016/j.jhe.2009.10.005

Alexander, W. P., & Grimshaw, S. D. (1996). Treed regression. *Journal of Computational and Graphical Statistics*, *5*(2), 156–175. Retrieved 2022-06-15, from http://www.jstor.org/stable/1390778

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Comput.*, *9*(7), 1545-1588. Retrieved from http://dblp.uni-trier.de/db/journals/neco/neco9.htmlAmitG97

Breiman, L. (1996, aug). Bagging predictors. *Mach. Learn.*, *24*(2), 123–140. Retrieved from https://doi.org/10.1023/A:1018054314350 doi: 10.1023/A:1018054314350

Breiman, L. (2001, 10). Random forests. *Machine Learning*, *45*, 5-32. doi: 10.1023/A:1010950718922

Case, K. E., & Shiller, R. J. (1990). Forecasting prices and excess returns in the housing market. *Real Estate Economics*, *18*(3), 253-273. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6229.00521 doi: https://doi.org/10.1111/1540-6229.00521

Cirillo, P., & Muliere, P. (2013, 12). An urn-based bayesian block bootstrap. *Metrika*, *76*. doi: 10.1007/s00184-011-0377-1

Cohen, V., & Karpavičiūtė, L. (2017, 03). The analysis of the determinants of housing prices. *Independent Journal of Management Production*, *8*, 49-63. doi: 10.14807/ijmp.v8i1.521

Coulombe, P. G. (2020, June). The Macroeconomy as a Random Forest [Papers]. (2006.12724). Retrieved from https://ideas.repec.org/p/arx/papers/2006.12724.html

Coulombe, P. G., Leroux, M., Stevanovic, D., & Surprenant, S. (2021). Macroeconomic data transformations matter.

Diebold, F., & Mariano, R. (2002, 02). Comparing predictive accuracy. *Journal of Business Economic Statistics*, *20*, 134-44. doi: 10.1080/07350015.1995.10524599

DiPasquale, D., & Wheaton, W. C. (1996). *Urban economics and real estate markets* (Vol. 23) (No. 7). Prentice Hall Englewood Cliffs, NJ.

Égert, B., & Mihaljek, D. (2007, 10). Determinants of house prices in central and eastern europe. *Comparative Economic Studies*, *49*. doi: 10.1057/palgrave.ces.8100221

Friedberg, R., Tibshirani, J., Athey, S., & Wager, S. (2021). Local linear forests. *Journal of Computational and Graphical Statistics*, *30*(2), 503-517. Retrieved from https://doi.org/10.1080/10618600.2020.1831930 doi: 10.1080/10618600.2020.1831930

Garriga, C., Manuelli, R., & Peralta-Alva, A. (2019, June). A macroeconomic model of price swings in the housing market. *American Economic Review*, *109*(6), 2036-72. Retrieved from https://www.aeaweb.org/articles?id=10.1257/aer.20140193 doi: 10.1257/aer.20140193

Goodhart, C., & Hofmann, B. (2008, 03). House prices, money, credit, and the macroeconomy. *Oxford Review of Economic Policy*, *24*(1), 180-205. Retrieved from https://doi.org/10.1093/oxrep/grn009 doi: 10.1093/oxrep/grn009

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York, NY, USA: Springer New York Inc.

Ho, T. K. (1995). Random decision forests. , *1*, 278-282 vol.1. doi: 10.1109/ICDAR.1995.598994

Hossain, B., & Latif, E. (2009). Determinants of housing price volatility in Canada: a dynamic analysis. *Applied Economics*, *41*(27), 3521-3531. Retrieved from https://ideas.repec.org/a/taf/applec/v41y2009i27p3521-3531.html doi: 10.1080/00036840701522861

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in r*. Springer. Retrieved from https://faculty.marshall.usc.edu/gareth-james/ISL/

Mackinnon, J. G. (2006, September). Bootstrap Methods in Econometrics. *The Economic Record*, *82*(s1), 2-18. Retrieved from https://ideas.repec.org/a/bla/ecorec/v82y2006is1ps2-s18.html doi: 10.1111/j.1475-4932.2006.

McCracken, M. W., & Ng, S. (2020, March). *FRED-QD: A Quarterly Database for Macroeconomic Research* (Working Papers No. 2020-005). Federal Reserve Bank of St. Louis. Retrieved from https://ideas.repec.org/p/fip/fedlwp/87608.html doi: 10.20955/wp.2020.005

Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, *9*(1), 130 – 134. Retrieved from https://doi.org/10.1214/aos/1176345338 doi: 10.1214/aos/1176345338

Sutton, G. D. (2002, September). Explaining changes in house prices. *BIS Quarterly Review*. Retrieved from https://ideas.repec.org/a/bis/bisqtr/0209f.html

Tripathi, S. (2019, November). Macroeconomic Determinants of Housing Prices: A Cross Country Level Analysis [MPRA Paper]. (98089). Retrieved from https://ideas.repec.org/p/pra/mprapa/98089.html

Tsatsaronis, K., & Zhu, H. (2004, 02). What drives housing price dynamics: Cross-country evidence. *BIS Quarterly Review*, *March*.

Wang, Y., & Witten, I. (1997, 01). Induction of model trees for predicting continuous classes. *Induction of Model Trees for Predicting Continuous Classes*.