Journal of Real Estate Finance & Economics (2022) 64:523–545
https://doi.org/10.1007/s11146-020-09813-1

# Machine Learning Predictions of Housing Market Synchronization across US States: The Role of Uncertainty

**Rangan Gupta[1] · Hardik A. Marfatia[2] · Christian Pierdzioch[3] · Afees A. Salisu[4]**

Accepted: 10 December 2020 / Published online: 12 January 2021


## Abstract

We analyze the role of macroeconomic uncertainty in predicting synchronization in housing price movements across all the United States (US) states plus District of Columbia (DC). We first use a Bayesian dynamic factor model to decompose the house price movements into a national, four regional (Northeast, South, Midwest, and West), and state-specific factors. We then study the ability of macroeconomic uncertainty in forecasting the comovements in housing prices, by controlling for a wide-array of predictors, such as factors derived from a large macroeconomic dataset, oil shocks, and financial market-related uncertainties. To accommodate for multiple predictors and nonlinearities, we take a machine learning approach of random forests. Our results provide strong evidence of forecastability of the national house price factor based on the information content of macroeconomic uncertainties over and above the other predictors. This result also carries over, albeit by a varying degree, to the factors associated with the four census regions, and the overall house price growth of the US economy. Moreover, macroeconomic uncertainty is found to have predictive content for (stochastic) volatility of the national factor and aggregate US house price. Our results have important implications for policymakers and investors.

**Keywords** Machine learning · Random forests · Bayesian dynamic factor model · Forecasting · Housing markets synchronization · United States

**JEL Classification** C22 · C32 · E32 · Q02 · R30

We would like to thank an anonymous referee for many helpful comments. However, any remaining errors are solely ours.

The research of C. Pierdzioch was supported by the German Science Foundation (Project: Exploring the experience-expectation nexus in macroeconomic forecasting using computational text analysis and machine learning; Project number: 275693836). The usual disclaimer applies.

✉ Hardik A. Marfatia
h-marfatia@neiu.edu

Extended author information available on the last page of the article.

## Introduction

In the wake of the Global Financial Crisis (GFC), many researchers have studied the impact of uncertainty on the overall macroeconomy of the United States (US), as well as its financial markets (see for example, Chuliá et al. 2017; Gupta et al. 2018, 2020a, for detailed reviews). Moreover, given that the root of the GFC was the subprime mortgage market of the US, a growing number of studies (see, for example, Antonakakis et al. 2015, 2016; El Montasser et al. 2016; André et al. 2017; Christou et al. 2017, 2019; Christidou and Fountas 2018; Nguyen Thanh et al. 2018; Aye et al. 2019; Bouri et al. 2020) have highlighted the role of uncertainty in predicting (primarily) aggregate and regional housing returns and (to some extent) aggregate volatility of the US.[1]

We aim to add to this burgeoning literature by analyzing the comovement of housing markets across all the US states plus District of Columbia (DC), and investigate the ability of macroeconomic uncertainty to forecast synchronous movements in states' housing markets, after controlling for a wide-array of macroeconomic variables. Following the GFC, the comovement of the housing markets and its connections with the macroeconomy has been at the center of discussions among researchers, policymakers, and market participants (Ghent and Owyang 2010; Christiansen et al. 2019; Sun and Tsang 2019). This is particularly highlighted by the recent synchronized booms and busts in the housing cycles with the business cycle (Leamer 2007, 2015; Balcilar et al. 2014; Nyakabawo et al. 2015; Emirmahmutoglu et al. 2016).

To study the nature of synchronization in housing prices across the different states, we use a (Bayesian) dynamic factor model (DFM), as discussed by Stock and Watson (1989) , and decompose the movement in the growth rate of real house prices for all the states in the US plus DC into a national factor, which captures the fluctuations that are common across all the states, besides four regional factors, which document the common movements in a particular region, and state-specific factors, which are unique to each state. This modeling strategy allows us to study the nature of synchronization over time, and the relative importance of each latent factor in influencing the housing price dynamics in each state.

We highlight that, distinguishing the national factor from local factors in the housing market, and determining what fraction of the variation in house prices across the states is explained by the common component, allow us to deduce whether the US economy is facing a "national bubble" or "local bubbles". While "local bubbles" are attributable to circumstances that are specific to each geographic market, by forecasting the national factor with a large set of predictors including uncertainty, we gauge the predictable part of common regional housing market movement attributable to changes in fundamentals, and the (unpredictable) portion that could be due to speculation or pricing errors. Naturally, our analysis has tremendous significance from a policy perspective, as accurate forecast of the path of the "national bubble" would

[1] Similar observations for the US Real Estate Investment Trusts (REITs) have also been drawn by Ajmi et al. (2015), and Akinsomi et al. (2016).

allow policy authorities to undertake appropriate policies ahead of time and prevent the negative influence on the overall economy, in case the "national bubble" bursts.

Our paper builds on the works of Del Negro and Otrok (2007), Marfatia (2018), Gupta et al. (2020b), and Sheng et al. (2020), who too use a Bayesian DFM to obtain the national and regional factors of house prices associated with the US states, and highlight the in-sample role of macroeconomic variables and oil in driving the same, especially focusing on nonlinearities in the relationships. These studies are indeed insightful, but as pointed out by Campbell (2008), the ultimate test of any predictive model (in terms of econometric methodologies and the predictors used) is in its out-of-sample performance. Becuase existence of in-sample predictability does not necessarily ensure out-of-sample forecasting gains (Rapach and Zhou 2013). Consequently, our paper is a robust extension of these papers as it involves a full-fledged out-of-sample forecasting exercise of the "national bubble" component of the overall US housing market.

To the best of our knowledge, this is the first paper to analyze the role of macroeconomic uncertainty in forecasting the national and regional housing price factors over the monthly period from 1975:02 to 2019:12 using univariate and multivariate machine learning (ML) methods, known as random forests (Breiman 2001), which in turn has multiple advantages. First, random forests can accurately analyze the links between the housing price factors and a large number of predictors in a fully data-driven way. The need for many predictors is motivated from the results of studies dealing with aggregate and regional house price forecasts (see for example, Rapach and Strauss 2009, Gupta et al. 2011a, b; Gupta 2013; Bork and Møller 2015; Plakandaras et al. 2015; Bork et al. 2020). Our predictors of comovement include overall macroeconomic and financial uncertainties, 8 factors extracted from a large data set aggregate and regional macroeconomic and financial variables. In addition, we include four structural (oil-specific supply and demand, inventory accumulation, and global demand) shocks, rather than just movements of the oil price per se. This choice is supported by the finding that oil-price movements can have a different impact on US asset (stock and housing) markets and the overall economy depending on the cause of the oil-price change (Kilian 2009; Kilian and Park 2009; Killins et al. 2017). Second, random forests automatically capture potential nonlinear links between the housing price factors and its predictors, the importance of which has been discussed by the in-sample-based studies discussed above, as well as any interaction effects between the predictors.

We organize the remainder of this research as follows. "Methodologies" section outlines the basics of the Bayesian DFM and random forests. Third section discusses the data. Section "Empirical Results" presents the results from the Bayesian DFM and the forecasting experiments. "Conclusion" section concludes.

## Methodologies

In this segment, we briefly discuss the basics of the methodologies of Bayesian DFM and the ML methods.

### The Dynamic Factor Model

House prices across the states largely comove with each other and understanding the reasons for such synchronous pattern is important to fully understand the real estate landscape. However, prima facie, these reasons are unobserved, even while we could contemplate that the comovement is related at different levels. We model the unobserved common movement in house prices across states by decomposing the real house price growth rate ($h_{i,t}$) for each $i$ state ($i = 1, ..., N$) into three latent factors. First, the common national level shocks that have a varied impact on the state housing market. Second, the regional level shocks, named as a regional factor, shared by all the states within a census region (namely, Northwest, South, Midwest, and West). Third, the state-specific factor unique to each state. This decomposition can be represented as follows:

$$h_{i,t} = \beta_i^n f_t^n + \beta_i^r f_t^r + \epsilon_{i,t} \tag{1}$$

where subscript $i$ represents each of the $N$ states. Coefficients $\beta_i^n$ and $\beta_i^r$ are the loadings of the national factor ($f_t^n$) and the regional factor ($f_t^r$), respectively. The factor loadings show the extent to which house prices in each state respond to the national and regional forces. The idiosyncratic component unique to each state's housing market dynamics is captured by $\epsilon_{i,t}$. We are particularly interested in the shared national factor that captures the common movement of house prices across all the states, as well as the four regional factors shared by the housing markets of a particular region.

The appropriateness of using census regions in the present context as compared to alternative classification schemed is supported on several grounds. First, it avoids the risk of weak identification and/or misspecification that may arise from fewer cross-sections in the alternative classification scheme. Second, the latent national factor is orthogonal to the latent regional factors by construction. Thus, the choice of a regional or divisional classification scheme would not affect the estimated synchronization in housing price movements at the national level. Hence, the result that, using machine learning, macroeconomic uncertainties can accurately forecast synchronization in housing price movements at the national level remains unaffected by the choice of the classification scheme. Third, the estimation process does not achieve convergence in the case of alternative classification schemes. We ran estimations considering the 9 census divisions, the 8 BEA regions, and other related schemes. Convergence was not achieved in any of these schemes. Fourth, pragmatic cost-benefit considerations (in terms of computing time) and the need to be consistent with earlier literature (Del Negro and Otrok 2007) led us to adopt the classification scheme that we use in our research.

We follow the literature and model the three latent factors - $f_t^n$, $f_t^r$, and $\epsilon_{i,t}$ - as an autoregressive (AR) process. We have,

$$f_t^n = \phi_1^n f_{t-1}^n + ... + \phi_p^n f_{t-p}^n + \upsilon_t^n, \qquad \upsilon_t^n \sim i.i.d.N(0, \sigma_n^2), \tag{2}$$

$$f_t^r = \phi_1^r f_{t-1}^r + ... + \phi_p^r f_{t-p}^r + \upsilon_t^r, \qquad \upsilon_t^r \sim i.i.d.N(0, \sigma_r^2), \tag{3}$$

$$\epsilon_{i,t} = \phi_{i,1}\epsilon_{i,t-1} + ... + \phi_{i,q}\epsilon_{i,t-q} + \upsilon_{i,t}, \qquad \upsilon_{i,t} \sim i.i.d.N(0, \sigma_i^2). \tag{4}$$

For a meaningful identification of latent components and in order to estimate the coefficients, we assume that the shocks are orthogonal contemporaneously as well as at all leads and lags. Thus, $E(\upsilon_t^n, \upsilon_{t-s}^n) = E(\upsilon_t^r, \upsilon_{t-s}^r) = E(\upsilon_{i,t}, \upsilon_{i,t-s}) = 0$. We normalize the sign and scale following the strategy well established in the literature (Kose et al. 2003, 2008; Neely and Rapach 2011).[2]

We emphasize that the factors are latent. Consequently, the usual regression apparatus is not applicable for estimating the model. In consequence, we use the Bayesian technique developed by Otrok and Whiteman (1998). We use a Markov chain Monte Carlo (MCMC) procedure to successively draw from a series of conditional distributions the complete posterior distribution of all the parameters together with the latent factors. We use the standard specification for priors, similar to Kose et al. (2003) and Del Negro and Otrok (2007).[3]

In addition to estimating the latent factors, we also explore the role of the three latent factors in house price movements. The fraction of variance due to the national ($\theta_i^n$), regional ($\theta_i^r$), and state-specific ($\theta_i^s$) factors in the overall variation in house prices can be derived as

$$\theta_i^n = \frac{(\beta_i^n)^2 var(f_t^n)}{var(h_{i,t})}, \qquad \theta_i^r = \frac{(\beta_i^r)^2 var(f_t^r)}{var(h_{i,t})}, \qquad \theta_i^s = \frac{var(\epsilon_{i,t})}{var(h_{i,t})}. \tag{5}$$

The estimates of $\theta_i^n$, $\theta_i^r$, and $\theta_i^s$ show the proportion of variance in the national, regional, and state-specific factors, respectively, relative to the overall variance in house price movements of each state.

## Random Forests

We use random forests for forecasting national and regional housing price factors. Random forests render it possible to forecast with many predictors, they account in a natural way for interaction effects between the predictors, and they capture in a data-driven way potential nonlinearities in the data. In addition, their output is easy to interpret and an intuitive algorithm governs their computation. It is, therefore, not surprising that random forests are a very popular ensemble machine-learning technique.

A random forest is formed by growing an ensemble of individual regression trees. The basic idea motivating the construction of a regression tree, $T$, is that branches partition the space of predictors, $\mathbf{x} = (x_1, x_2, ...)$, into $l$ non-overlapping regions, $R_l$ (for a comprehensive introduction, see Hastie et al. 2009; we largely use their notation). These regions are formed in a top-down way (that is, from root to the leaves of a tree) by means of a recursive search-and-split algorithm.

[2] In particular, for sign identification, national factor for Alaska is restricted to be positive, whereas the sign restriction on regional factor loadings is chosen arbitrarily. We achieve scale normalizations by following Sargent and Sims (1977), Stock and Watson (1989, 1993), and Del Negro and Otrok (2007) and restrict $\sigma_n^2$ and $\sigma_r^2$ to unity. The signs and scale normalization do not have any economic content and do not affect any economic inference (Neely and Rapach 2011).

[3] The prior for idiosyncratic state-specific shocks follows an inverse-gamma distribution with parameters 6 and 0.001. The prior for the AR polynomial follows a normal distribution with tighter centering on zero (at the geometric rate of 0.5). The priors for factor loadings are standard normal.

We sketch how this algorithm works by starting at the root of a regression tree. At the root of a tree, search-and-split algorithm loops over the predictors, $s$, and their corresponding realizations can be used to construct a splitting point, $p$. For every combination of a predictor and a splitting point, the search-and-split algorithm then forms two half-planes, $R_1(s, p) = \{x_s|x_s \leq p\}$ and $R_2(s, p) = \{x_s|x_s > p\}$. The optimal half-planes and, thus, the optimal pair of a predictor and the splitting point is chosen such as to minimize the following squared-error loss criterion:

$$\min_{s,p} \left\{ \min_{\bar{f}_1} \sum_{x_s \in R_1(s,p)} (f_i - \bar{f}_1)^2 + \min_{\bar{f}_2} \sum_{x_s \in R_2(s,p)} (f_i - \bar{f}_2)^2 \right\}, \tag{6}$$

where the index $i$ identifies those data of the housing factor, $f$, that is being studied that belong to a half-plane, and $\bar{f}_k = \text{mean}\{f_i \; |x_s \in R_k(s, p)\}, k = 1, 2$ denotes the half-plane-specific mean of a housing factor (for ease of notation, we have dropped the time index). The intuition behind Eq. 6 is simple: The outer minimization loops over all pairs of $s$ and $p$, while the inner minimization minimizes, for a given pair of $s$ and $p$, identifies the half-plane-specific means of a housing factor by minimizing the half-plane-specific squared error loss.

The solution of the minimization problem given in Eq. 6 gives the first optimal splitting predictor, the first optimal splitting point, and the two region-specific means of a housing factor. The minimization problem, thus, already results in a regression tree, albeit a very simple one. This simple regression tree, however, can be grown further by applying the same search-and-split algorithm. Hence, one solves again the minimization problem given in Eq. 6, but now separately for the two optimal top-level half-planes, $R_1(s, p)$ and $R_2(s, p)$, resulting in up to two second-level optimal splitting predictors and their optimal splitting points, and four second-level region-specific means of the housing factor under study.

Further application of the search-and-split algorithm then renders it possible to form a regression tree whose hierarchical structure becomes increasingly complex. The search-and-split algorithm stops when a the number of terminal nodes reaches a maximum or the number of observations per terminal node reaches minimum, both criteria are specified in advance by a researcher. Naturally, a larger number of maximum possible terminal nodes results in a larger tree, so does a smaller number of observations per terminal node.

Once the search-partition algorithm stops, the data on the predictors are sent from the root to the terminal nodes down a regression along its nodes and branches, and the region-specific means at the terminal nodes of a regression tree can then be used to compute a forecast of a housing factor. When there are $L$ regions, a forecast is computed as follows (**1** denotes the indicator function):

$$T\left(\mathbf{x}_i, \{R_l\}_1^L\right) = \sum_{l=1}^{L} \bar{f}_l \mathbf{1}(\mathbf{x}_i \in R_l). \tag{7}$$

Extensive application of the search-and-split algorithm results in fine granular forecasts of a housing factor. This does not mean, however, that growing a more complex regression tree always improves forecast accuracy because a growing complexity of the hierarchical structure of a regression tree naturally results in an overfitting and data-sensitivity problem. A random forest is a vehicle to overcome this data-sensitivity problem. A random forest is grown in three steps:

1. A certain number of bootstrap samples is generated from the data. Sampling is done with replacement.
2. A random regression tree is estimated on every bootstrap sample. A random regression tree selects for every splitting decision a randomly chosen subset of the predictors. This random sampling of predictors curbs the influence of influential predictors on tree growing.

Growing a large number of random regression trees in this way decorrelates their forecasts, while averaging the forecasts across the trees stabilizes the forecasts computed by means of a random forest.

## Data

For the house price data, we use seasonally adjusted nominal house prices of 50 states plus DC derived from Freddie Mac,[4] with the indices based on an ever-expanding database of loans purchased by either Freddie Mac or Fannie Mae. In order to obtain the real house prices, the nominal values are deflated by the seasonally adjusted personal consumption expenditures (PCE) deflator obtained from the Fred database of the Federal Reserve Bank of St. Louis. To ensure stationarity, required for the estimation of the DFM, we work with month-on-month growth rates of real housing prices.

We include a plethora of predictors in the study. We use 8 factors derived from the 134 macroeconomic variables of Ludvigson and Ng (2009, 2011).[5] Including these series gives us the advantage of capturing broad categories of overall and regional macroeconomic time series namely, real output and income, employment and hours, real retail, manufacturing and sales data, international trade, consumer spending, housing starts, housing building permits, inventories and inventory sales ratios, orders and unfilled orders, compensation and labor costs, capacity utilization measures, price indexes, interest rates and interest rate spreads, stock market indicators, and foreign exchange measures.

Next set of predictors include the four structural oil-shocks obtained from the structural vector autoregressive (SVAR) model of Baumeister and Hamilton (2019).[6] The advantage is that Baumeister and Hamilton (2019) formulate a less restrictive framework than what has been traditionally used in the literature following Kilian

[4]The data is available for download from: http://www.freddiemac.com/research/indices/house-price-index.page.

[5]The factors are available for download from: https://www.sydneyludvigson.com/data-and-appendixes.

[6]The data is downloadable from: https://sites.google.com/site/cjsbaumeister/research.

(2009), by incorporating uncertainty about the identifying assumptions of the SVAR. In other words, the obtained oil shocks do not contain overlapping information and can be considered to be relatively more accurately estimated, with each of them capturing distinct aspects regarding the demand and supply sides of the oil market.

Finally, for our main predictor of interest - uncertainty - we again take a comprehensive approach because of the inherit latent nature of uncertainty. In this regard, besides the various alternative metrics of uncertainty associated with financial markets (such as the Chicago Board Options Exchange (CBOE) implied-volatility index, popularly called the VIX), there are primarily three broad approaches to quantify uncertainty (Gupta et al. 2019, 2020c): (1) A newspapers-based approach, where searches of major newspapers are conducted for terms related to economic and policy uncertainty (EPU), and then the results are used to construct indexes of uncertainty. (2) Measures of uncertainty derived from stochastic-volatility estimates of various types of small and large-scale structural models related to macroeconomics and finance. (3) Uncertainty as obtained from dispersion of professional forecaster disagreements. In this research, we use the macroeconomic uncertainty (MU) and financial uncertainty (FU) measures of Jurado et al. (2015) and Ludvigson et al. (Forthcoming), which, in turn, is the average time-varying variance in the unpredictable component of 134 macroeconomic and 148 financial time-series respectively, i.e., it attempts to capture the average volatility in the shocks to the factors that summarize real and financial conditions.[7] In other words, it is derived based on the second approach outlined above. Note that, the same 134 variables are used in computing the factors used as predictors and the metric of macroeconomic uncertainty. Unlike existing alternative measures of uncertainty, based on newspaper (text-based) approaches, forecast disagreement associated with certain variables, or volatility of a specific financial market, the metrics that we use are the broadest measures of macroeconomic and financial uncertainties available for the US. The uncertainty indices are available for three forecasting horizons of 1-, 3-, and 12-month-ahead.

Based on data availability and transformation of the data, our monthly sample period covers 1975:02 to 2019:12. Note that all our predictors are stationary by design.

## Empirical Results

### The National Factor, Regional Factors, and Variance Decomposition

Figure 1 shows the behavior of the national factor over time. We find that, in the 1979–1981 period, the national factor dived down sharply with a quick recovery in the next four years. This is followed by another housing cycle in the 1982–1990 period. From 1990–2005, the national factor rose steadily until the onset of the

[7] The MU and FU indices are available for download from: www.sydneyludvigson.com/data-and-appendixes.

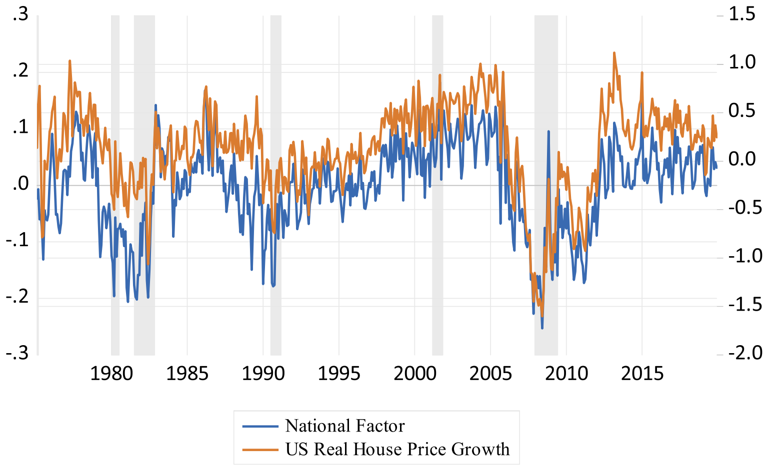


**Fig. 1** National factor and aggregate real housing growth rate of the US. The left scale corresponds to the national factor, while the right scale is associated with the US real house price Growth; the shaded regions correspond to the NBER recession dates

financial crisis, which had its roots in the housing markets. This pattern of national factor mimics the well-known housing boom of 2005, the following crash in 2005–2008 period, and slow recovery thereafter (2011–2017). When we compare the unobserved national factor of our model with that of the aggregate real housing price growth rate of the US (derived from the national Freddie Mac house price index deflated by the PCE deflator), also plotted in Fig. 1, we find that the common component is able to map accurately the aggregate house price movements, which we also forecast as a matter of comparability. In fact, the correlation between these two series is 0.88, which is statistically significant at the highest possible level. Our results suggest that the otherwise unobserved national factor of our model aptly captures the aggregate movements in the US house prices.

The estimates of variance decomposition of the national factor are shown in Table 1. Results show that the national factor plays a big role in New Jersey and Pennsylvania in the Northeast, District of Columbia and Delaware in the South, Nebraska, Illinois, Michigan, and Missouri in the Midwest, and Arizona, California, and Nevada in the West. The national factor explains nearly 40–50 % of the variation in house prices in these states. These results imply that national-level forces significantly impact coastal America and parts of Midwest which has some of the biggest financial hubs in the country. In contrast, mountain and central states are less linked with the national factor, and more associated with the regional and state-specific dynamics. Overall, the national factor explains 34 % of the real house price growth rates of the Northeast and West census regions states, while for the South this number stands at 38 % and the lowest explanatory power is found for the Midwest at 32 %. Finally, for the overall US involving the 50 states plus DC, the national factor explains 35 % of the variation.

**Table 1** Variance decomposition

| State name | National | Regional | State | State name | National | Regional | State |
|---|---|---|---|---|---|---|---|
| Northeast | | | | Midwest | | | |
| Connecticut | 0.426 | 0.057 | 0.517 | Iowa | 0.353 | 0.029 | 0.618 |
| Massachusetts | 0.266 | 0.004 | 0.730 | Illinois | 0.391 | 0.060 | 0.548 |
| Maine | 0.433 | 0.027 | 0.540 | Indiana | 0.198 | 0.146 | 0.655 |
| New Hampshire | 0.251 | 0.024 | 0.725 | Kansas | 0.211 | 0.002 | 0.787 |
| New Jersey | 0.454 | 0.018 | 0.528 | Michigan | 0.405 | 0.022 | 0.572 |
| New York | 0.402 | 0.020 | 0.578 | Minnesota | 0.378 | 0.004 | 0.618 |
| Pennsylvania | 0.023 | 0.046 | 0.931 | Missouri | 0.462 | 0.002 | 0.536 |
| Rhode Island | 0.440 | 0.017 | 0.543 | N. Dakota | 0.219 | 0.059 | 0.721 |
| Vermont | 0.386 | 0.014 | 0.599 | Nebraska | 0.506 | 0.008 | 0.487 |
| South | | | | Ohio | 0.134 | 0.261 | 0.605 |
| Alabama | 0.402 | 0.181 | 0.417 | S. Dakota | 0.305 | 0.057 | 0.639 |
| Arkansas | 0.466 | 0.096 | 0.438 | Wisconsin | 0.287 | 0.068 | 0.645 |
| D.C. | 0.506 | 0.195 | 0.300 | West | | | |
| Delaware | 0.460 | 0.072 | 0.468 | Alaska | 0.444 | 0.256 | 0.300 |
| Florida | 0.354 | 0.005 | 0.641 | Arizona | 0.516 | 0.188 | 0.296 |
| Georgia | 0.430 | 0.067 | 0.503 | California | 0.467 | 0.234 | 0.300 |
| Kentucky | 0.283 | 0.004 | 0.713 | Colorado | 0.417 | 0.211 | 0.372 |
| Louisiana | 0.433 | 0.059 | 0.507 | Hawaii | 0.386 | 0.007 | 0.607 |
| Maryland | 0.344 | 0.004 | 0.652 | Idaho | 0.310 | 0.005 | 0.685 |
| Mississippi | 0.264 | 0.132 | 0.604 | Montana | 0.445 | 0.027 | 0.528 |
| N. Carolina | 0.435 | 0.019 | 0.547 | N. Mexico | 0.135 | 0.090 | 0.774 |
| Oklahoma | 0.559 | 0.002 | 0.439 | Nevada | 0.371 | 0.047 | 0.582 |
| S. Carolina | 0.416 | 0.006 | 0.577 | Oregon | 0.204 | 0.004 | 0.792 |
| Tennessee | 0.087 | 0.002 | 0.911 | Utah | 0.379 | 0.033 | 0.588 |
| Texas | 0.353 | 0.056 | 0.592 | Washington | 0.257 | 0.044 | 0.698 |
| Virginia | 0.375 | 0.036 | 0.589 | Wyoming | 0.107 | 0.115 | 0.778 |
| West Virginia | 0.345 | 0.046 | 0.609 | | | | |

This table reports the variance decomposition of real house price into the national, regional, and state-specific factors

## Forecasting Results

We now turn our attention to the out-of-sample forecasting analysis of the national factor and the four regional factors using random forests.[8] Our random forests consist of 50, 75, or 100 random trees, where we use a minimum number of observations

[8]We use the R language for statistical computing (R Core Team 2019) to carry out our forecasting experiments, and the add-on package “MultivariateRandomForest” (Rahman 2017) to estimate random forests.

of five per terminal node and a number of round (number of predictors/3) of randomly chosen predictors for splitting. Further, we study a ten-year rolling-estimation window and three forecast horizons: one, three, and twelve months. We consider the following forecasting models:

1. The forecasting model ("Aggregate") for aggregate real housing price growth rate includes the (lagged) aggregate real housing in the array of predictors, but excludes the national factor and the regional factors.
2. The forecasting model ("National") for the national factor includes the (lagged) national factor in the array of predictors, but excludes the aggregate real housing price growth rate and the regional factors.
3. The forecasting models for the four regional factors include all four (lagged) regional factors, but exclude the aggregate real housing price growth rate and the national factor.

In order to set the stage for our analysis, we compare in Table 2 the various forecasting models with the rolling-historical-mean forecasts of the dependent variables. To this end, we compute the ratio of the root-mean-squared forecast errors (RMSFE) by dividing the RMSFE from rolling-historical mean forecasts by the RMSFE from the random-forest-forecasting models that exclude macroeconomic uncertainty. A ratio larger than unity signals superior performance of the random-forest models. Results show that, with the Midwest factor being the sole exception, the random-forest models produce a smaller RMSFE than the rolling historical mean. This means our predictors help to improve forecast accuracy relative to a naive forecast.

Table 3 reports the ratio of RMSFE from the forecasting models that exclude macroeconomic uncertainty from the set of predictors over the RMSFE from the model that includes macroeconomic uncertainty in the set of predictors. Again, a ratio larger than unity indicates that macroeconomic uncertainty improves forecast accuracy. Results show that RMSFE ratios exceed unity for most configurations of random forests and forecast horizons, indicating that macroeconomic uncertainty provides incremental information in predicting common movements in the housing markets. We observe RMSFE ratios smaller than unity in the case of the Midwest factor (West factor), when we use 50 (75) random trees, and the South factor. In case of the Midwest factor, further inspection shows that adding macroeconomic uncertainty to the array of predictors does not suffice to outperform a naive forecast (results are available upon request).

The results we summarize in Table 4 for the mean-absolute-forecasting-error (MAFE) ratios corroborate the results for the RMSFE ratios. Broadly, result show that macroeconomic uncertainty helps to improve forecast accuracy, where the results are relatively weak for the South factor.[9]

[9]We also analyzed the relative importance of the predictor variables using the full sample of data. The importance of a predictor measures how often a predictor is used for splitting. Results show that, across the different model configurations and forecast horizons, the three measures of macroeconomic uncertainty have a relative importance of about 10 % percent, which corresponds to a top average rank of 1.3 among all predictors (detailed results are available upon request from the authors).

**Table 2** Root-mean-squared-forecasting-error ratios (historical mean)

| Trees | $h = 1$ | $h = 3$ | $h = 12$ |
|---|---|---|---|
| Panel A: Aggregate | | | |
| 50 | 1.9512 | 1.6338 | 1.6572 |
| 75 | 1.9262 | 1.6893 | 1.6836 |
| 100 | 1.9755 | 1.6996 | 1.6477 |
| Panel B: National factor | | | |
| 50 | 1.6393 | 1.4120 | 1.4234 |
| 75 | 1.6323 | 1.4053 | 1.4088 |
| 100 | 1.6327 | 1.4022 | 1.4161 |
| Panel C: Midwest factor | | | |
| 50 | 0.9519 | 0.9484 | 0.9005 |
| 75 | 0.9694 | 0.9427 | 0.8911 |
| 100 | 0.9579 | 0.9434 | 0.8943 |
| Panel D: Northeast factor | | | |
| 50 | 1.7361 | 1.4493 | 1.5656 |
| 75 | 1.7655 | 1.4421 | 1.5769 |
| 100 | 1.7576 | 1.4569 | 1.5837 |
| Panel E: South factor | | | |
| 50 | 1.2054 | 1.1455 | 1.1770 |
| 75 | 1.2414 | 1.1387 | 1.1865 |
| 100 | 1.2134 | 1.1318 | 1.1947 |
| Panel F: West factor | | | |
| 50 | 1.0731 | 1.0689 | 1.0340 |
| 75 | 1.0895 | 1.0766 | 1.0441 |
| 100 | 1.0827 | 1.0713 | 1.0368 |

This table reports results of RMSFE ratio obtained by dividing the RMSFE of a model that uses the rolling historical mean (that is, the mean computed using a rolling-estimation window) of the dependent variable to compute forecasts by the RMSFE restricted model that excludes macroeconomic uncertainty from the array of predictors. For the latter model, random forests are estimated using a rolling-estimation window of length 120 months, a minimum nodesize of 5, and a selection of random splitting predictors of round(number of predictors/3). The parameter $h$ denotes the forecast horizon (in months). The random forests estimated on aggregate data only include, in addition to the other predictors, aggregate data as a predictor, but no regional factors. The same is true for the random forests estimated on data for the national factor. The random forests estimated for the regional factors include the own regional factor along with all other regional factors in the array of predictors

Table 5 summarizes the results of the Clark and West (2007) test of the null hypothesis of equal mean-squared prediction errors of the models that include/exclude the realizations of macroeconomic uncertainty as predictors. The alternative hypothesis is that the performance of the model that includes macroeconomic uncertainty in the array of predictors is better than the performance of the model that excludes macroeconomic uncertainty from the array of predictors. The results of the Clark-West test are in line with the results of the analyses of the RMSFE and MAFE ratios. Most test results are significant at the 5 % level of significance. The results for the Midwest factor and West factor depend on the number of random trees that we use to grow random forest, and only two of the tests are significant in the case of the South factor.

As an extension, we use the Clark-West test to study whether macroeconomic uncertainty contributes to improve forecast accuracy in a multivariate setting. Specifically, we study the four regional factors by means of multivariate random forests

**Table 3** Root-mean-squared-forecasting-error ratios

| Trees | $h = 1$ | $h = 3$ | $h = 12$ |
|---|---|---|---|
| Panel A: Aggregate | | | |
| 50 | 1.0141 | 1.0561 | 1.0257 |
| 75 | 1.0304 | 1.0298 | 1.0131 |
| 100 | 1.0175 | 1.0071 | 1.0423 |
| Panel B: National factor | | | |
| 50 | 0.9955 | 1.0019 | 1.0509 |
| 75 | 1.0160 | 1.0262 | 1.0677 |
| 100 | 1.0308 | 1.0344 | 1.0565 |
| Panel C: Midwest factor | | | |
| 50 | 0.9814 | 1.0139 | 0.9803 |
| 75 | 1.0035 | 1.0340 | 1.0047 |
| 100 | 0.9859 | 1.0113 | 1.0083 |
| Panel D: Northeast factor | | | |
| 50 | 1.0246 | 1.0349 | 1.0344 |
| 75 | 1.0072 | 1.0428 | 1.0269 |
| 100 | 1.0182 | 1.0266 | 1.0196 |
| Panel E: South factor | | | |
| 50 | 0.999 | 0.9792 | 1.0012 |
| 75 | 0.968 | 0.9910 | 0.9883 |
| 100 | 0.995 | 1.0083 | 0.9835 |
| Panel F: West factor | | | |
| 50 | 1.0105 | 1.0095 | 1.0060 |
| 75 | 0.9950 | 0.9804 | 0.9918 |
| 100 | 1.0065 | 1.0028 | 1.0134 |

This table reports results of RMSFE ratio obtained by dividing the RMSFE of the restricted model that excludes macroeconomic uncertainty from the array of predictors by the RMSFE of the model that includes macroeconomic uncertainty in the array of predictors. Random forests are estimated using a rolling-estimation window of length 120 months, a minimum nodesize of 5, and a selection of random splitting predictors of round(number of predictors/3). The parameter $h$ denotes the forecast horizon (in months). The random forests estimated on aggregate data only include, in addition to the other predictors, aggregate data as a predictor, but no regional factors. The same is true for the random forests estimated on data for the national factor. The random forests estimated for the regional factors include the own regional factor along with all other regional factors in the array of predictors

(see Segal and Xiao 2011). In the case of multivariate random forests, we forecast a four-component vector with our array of predictors.[10] Because the full-sample correlations of the regional factors are not particularly strong (and often negative), we do not expect noticeable forecasting gains from a multivariate approach. It is interesting, however, whether the predictive value of macroeconomic uncertainty that we observe in the univariate case carries over to a multivariate setting.

Table 6 summarizes the results of the Clark-West test. The results are mixed. The test yields significant results across all model configurations and forecast horizons for the Northeast factor. For the South factor, only the test results for the models that feature fewer trees are significant, and for the Midwest and West factors a few

[10]The splitting function is the same as in the case of univariate random forests, but the region-impurity measure is the Mahalanobis distance (for a detailed description and a recent application in the forecasting literature, see Behrens et al. 2018).

**Table 4** Mean-absolute-forecasting-error ratios

| Trees | $h = 1$ | $h = 3$ | $h = 12$ |
|---|---|---|---|
| Panel A: Aggregate | | | |
| 50 | 1.0290 | 1.0513 | 1.0114 |
| 75 | 1.0569 | 1.0288 | 0.9982 |
| 100 | 1.0317 | 1.0065 | 1.0160 |
| Panel B: National factor | | | |
| 50 | 1.0176 | 1.0195 | 1.0603 |
| 75 | 1.0305 | 1.0318 | 1.0657 |
| 100 | 1.0367 | 1.0424 | 1.0542 |
| Panel C: Midwest factor | | | |
| 50 | 0.9862 | 1.0074 | 0.9624 |
| 75 | 1.0129 | 1.0205 | 1.0034 |
| 100 | 0.9912 | 1.0186 | 0.9909 |
| Panel D: Northeast factor | | | |
| 50 | 1.0423 | 1.0415 | 1.0183 |
| 75 | 1.0109 | 1.0504 | 1.0285 |
| 100 | 1.0122 | 1.0312 | 1.0163 |
| Panel E: South factor | | | |
| 50 | 1.0076 | 0.9752 | 0.9869 |
| 75 | 0.9700 | 0.9950 | 0.9943 |
| 100 | 1.0034 | 1.0206 | 0.9985 |
| Panel F: West factor | | | |
| 50 | 1.0160 | 1.0018 | 1.0170 |
| 75 | 0.9979 | 0.9714 | 0.9724 |
| 100 | 1.0015 | 0.9895 | 1.0020 |

This table reports results of MAFE ratio obtained by dividing the MAFE of the restricted model that excludes macroeconomic uncertainty from the array of predictors by the MAFE of the model that includes macroeconomic uncertainty in the array of predictors. Random forests are estimated using a rolling-estimation window of length 120 months, a minimum nodesize of 5, and a selection of random splitting predictors of round(number of predictors/3). The parameter $h$ denotes the forecast horizon (in months). The random forests estimated on aggregate data only include, in addition to the other predictors, aggregate data as a predictor, but no regional factors. The same is true for the random forests estimated on data for the national factor. The random forests estimated for the regional factors include the own regional factor along with all other regional factors in the array of predictors

tests are significant but there is no systematic pattern in the results. We conclude that spotting the predictive value of macroeconomic uncertainty is more difficult with multivariate than univariate random forests, perhaps reflecting that the latter tend to have a better overall forecasting performance. But also possibly highlighting the accuracy of our Bayesian DFM in isolating the unique region-specific factors with non-overlapping information in explaining the associated price movements of the overall housing market.

We summarize the results of some robustness checks in Table 7. To this end, we vary the minimum node size, the number of randomly sampled predictors used for splitting, the length of the rolling-estimation window, and the number of trees. The key insight from Table 7 is that macroeconomic uncertainty significantly improves forecast accuracy in the vast majority of model configurations.

As yet another robustness check, it is interesting to study cumulated returns (the sum of continuously compounded returns over the forecast horizon). While the

**Table 5** Out-of-sample tests

| Trees | $h = 1$ | $h = 3$ | $h = 12$ |
|---|---|---|---|
| Panel A: Aggregate | | | |
| 50 | 2.3770 | 3.9832 | 2.0291 |
| 75 | 2.6559 | 3.0497 | 1.7417 |
| 100 | 2.5095 | 1.6388 | 1.8658 |
| Panel B: National factor | | | |
| 50 | 2.3676 | 2.5973 | 4.2274 |
| 75 | 3.4982 | 2.7519 | 3.8172 |
| 100 | 4.0320 | 3.1922 | 3.1200 |
| Panel C: Midwest factor | | | |
| 50 | 0.9753 | 2.0274 | 0.7015 |
| 75 | 2.1130 | 2.2708 | 1.6866 |
| 100 | 0.5179 | 2.3302 | 1.4362 |
| Panel D: Northeast factor | | | |
| 50 | 3.2500 | 2.3378 | 2.5762 |
| 75 | 1.8668 | 2.7278 | 2.0960 |
| 100 | 2.2966 | 2.2760 | 1.7604 |
| Panel E: South factor | | | |
| 50 | 1.9416 | 0.9871 | 1.6394 |
| 75 | −0.9671 | 1.1684 | 1.1170 |
| 100 | 1.2551 | 2.8951 | 0.6264 |
| Panel F: West factor | | | |
| 50 | 3.2789 | 2.5434 | 2.4249 |
| 75 | 1.4007 | −0.4405 | 1.0893 |
| 100 | 2.0447 | 2.0689 | 2.3808 |

This table reports results of the Clark-West test. The null hypothesis is that the full model that includes macroeconomic uncertainty has the same forecasting performance out-of-sample as the restricted model that excludes macroeconomic uncertainty from the array of predictors. The alternative hypothesis is that the full model performs better than the restricted model. Results are based on Newey-West robust standard errors. Critical values are 1.28 and 1.65 at the 10 % and 5 % level of significance. Random forests are estimated using a rolling-estimation window of length 120 months, a minimum nodesize of 5, and a selection of random splitting predictors of round(number of predictors/3). The parameter $h$ denotes the forecast horizon (in months). The random forests estimated on aggregate data only include, in addition to the other predictors, aggregate data as a predictor, but no regional factors. The same is true for the random forests estimated on data for the national factor. The random forests estimated for the regional factors include the own regional factor along with all other regional factors in the array of predictors

results in Table 5 show whether macroeconomic uncertainty predicts how returns change $h$ months in the future, cumulated returns show whether macroeconomic uncertainty predicts returns during the next $h$ months. Results for cumulated returns are broadly in line with our other results. Aggregate ($h = 3$, $h = 12$; 100 trees): 3.21,

**Table 6** Out-of-sample tests (multivariate model)

| Trees | $h = 1$ | $h = 3$ | $h = 12$ |
|---|---|---|---|
| Panel A: Midwest factor | | | |
| 50 | 1.8981 | 0.8400 | 2.3781 |
| 75 | 0.4728 | 1.3852 | 1.1952 |
| 100 | 2.0584 | 0.1186 | 0.1131 |
| Panel B: Northeast factor | | | |
| 50 | 1.4548 | 3.1005 | 1.3579 |
| 75 | 1.8219 | 2.9023 | 1.9933 |
| 100 | 1.6163 | 3.0877 | 1.5289 |
| Panel C: South factor | | | |
| 50 | 1.5547 | 1.6099 | 2.0961 |
| 75 | 0.7210 | 1.8492 | 1.5176 |
| 100 | 0.8045 | 1.0303 | 1.4482 |
| Panel D: West factor | | | |
| 50 | 0.0572 | −0.5005 | 1.9156 |
| 75 | 1.5066 | 1.8225 | 0.1056 |
| 100 | 1.1375 | −0.0872 | 1.6087 |

This table reports results of the Clark-West test. The null hypothesis is that the full model that includes macroeconomic uncertainty has the same forecasting performance out-of-sample as the restricted model that excludes macroeconomic uncertainty from the array of predictors. The alternative hypothesis is that the full model performs better than the restricted model. Results are based on Newey-West robust standard errors. Critical values are 1.28 and 1.65 at the 10 % and 5 % level of significance. Multivariate random forests are estimated using a rolling-estimation window of length 120 months, a minimum nodesize of 5, and a selection of random splitting predictors of round(number of predictors/3). The parameter $h$ denotes the forecast horizon (in months). The multivariate random forests estimated on aggregate data only include, in addition to the other predictors, aggregate data as a predictor, but no regional factors. The same is true for the multivariate random forests estimated on data for the national factor. The multivariate random forests estimated for the regional factors include the own regional factor along with all other regional factors in the array of predictors

1.81. National factor: 3.21, 3.31. Midwest factor: 1.68, −0.46. Northeast factor: 2.91, 1.75. South factor: 2.41, 2.17. West factor: 1.40, 2.17.

As a final exercise, we return to the univariate random forests, but we inspect whether macroeconomic uncertainty helps to improve the accuracy of forecasts of volatility of both the aggregate real house price growth rate and the national factor, which in turn is motivated by the works of Antonakakis et al. (2016), André et al. (2017), and Bouri et al. (2020) in particular. Residential real estate represents about 83.98 % of total household non-financial assets, 30.64 % of total household net worth, and 26.64 % of household total assets (Financial Accounts of the US, First Quarter, 2020).[11] Therefore, as pointed out by Shiller (1998), the risk or volatility of

[11] The reader is referred to https://www.federalreserve.gov/releases/z1/20200611/z1.pdf for further details.

**Table 7** Out-of-sample tests (robustness checks)

| Factor | $h = 1$ | $h = 3$ | $h = 12$ |
|---|---|---|---|
| Panel A: Minimum nodesize = 10 | | | |
| Aggregate | 3.4488 | 2.7678 | 1.5461 |
| National factor | 2.8583 | 2.298 | 3.6223 |
| Midwest factor | 0.7499 | 2.2697 | 1.5723 |
| Northeast factor | 2.8784 | 2.5124 | 1.6025 |
| South factor | 0.9770 | 2.1587 | 2.9959 |
| West factor | 1.4283 | 1.5740 | 2.6518 |
| Panel B: Sampling = round[sqrt(number of predictors)] | | | |
| Aggregate | 2.7621 | 3.0374 | 1.5620 |
| National factor | 2.9739 | 2.7814 | 3.2260 |
| Midwest factor | 2.6724 | 3.8081 | 1.7013 |
| Northeast factor | 2.8336 | 2.8350 | 1.6373 |
| South factor | 1.1953 | 1.6563 | 3.6160 |
| West factor | 0.3093 | 2.4214 | 4.0461 |
| Panel C: Rolling-estimation window of 240 months | | | |
| Aggregate | 1.9455 | 2.8338 | 2.9659 |
| National factor | 2.6416 | 2.8026 | 3.4091 |
| Midwest factor | 3.0722 | 2.5858 | -0.1336 |
| Northeast factor | 2.8336 | 2.8350 | 1.6373 |
| South factor | 1.2296 | 3.1343 | 1.6046 |
| West factor | 2.4390 | 0.1056 | 1.9095 |
| Panel D: Number of random regression trees = 500 | | | |
| Aggregate | 2.7222 | 3.1384 | 1.4338 |
| National factor | 2.4219 | 2.5345 | 3.7432 |
| Midwest factor | 2.1073 | 3.0473 | -0.5216 |
| Northeast factor | 2.9672 | 2.6112 | 1.3757 |
| South factor | 0.1613 | 0.5232 | 3.5996 |
| West factor | 1.8468 | 0.7643 | 1.6992 |

This table reports results of the Clark-West test. The null hypothesis is that the full model that includes macroeconomic uncertainty has the same forecasting performance out-of-sample as the restricted model that excludes macroeconomic uncertainty from the array of predictors. The alternative hypothesis is that the full model performs better than the restricted model. Results are based on Newey-West robust standard errors. Critical values are 1.28 and 1.65 at the 10 % and 5 % level of significance. Unless otherwise stated, the models are estimated using 100 trees. The parameters that are not changed are the same as in the baseline calibration of the model. The parameter $h$ denotes the forecast horizon (in months)

house prices is among the largest personal economic risks faced by the individual. Note that housing assets differ from financial assets, such as stocks, since they serve the dual role of investment and consumption (Henderson and Ioannides 1987). Naturally, the effects of housing on savings and portfolio choices are extremely important questions. Hence, understanding the source of housing market price volatility is important because it has individual portfolio implications, as it affects households' investment decisions regarding tenure choice and housing quantity (Miles 2008a). Moreover, knowledge about house price volatility is also an important input into housing policy (Zhou and Haurin 2010). In light of these points, the variations in the overall housing market are important with regard to the development of the overall economy and the welfare of society.

**Table 8** Out-of-sample tests (volatility)

| Trees | $h = 1$ | $h = 3$ | $h = 12$ |
|---|---|---|---|
| Panel A: Aggregate | | | |
| 50 | 2.0530 | 2.5862 | 3.1903 |
| 75 | 2.4794 | 2.7435 | 3.5491 |
| 100 | 1.8948 | 2.4380 | 2.6915 |
| Panel B: National factor | | | |
| 50 | 2.3135 | 2.8904 | 2.6696 |
| 75 | 2.6061 | 2.5157 | 2.2382 |
| 100 | 2.4053 | 2.1143 | 2.2025 |

This table reports results of the Clark-West test. The null hypothesis is that the full model that includes macroeconomic uncertainty has the same forecasting performance out-of-sample as the restricted model that excludes macroeconomic uncertainty from the array of predictors. The alternative hypothesis is that the full model performs better than the restricted model. Results are based on Newey-West robust standard errors. Critical values are 1.28 and 1.65 at the 10 % and 5 % level of significance. Multivariate random forests are estimated using a rolling-estimation window of length 120 months, a minimum nodesize of 5, and a selection of random splitting predictors of round(number of predictors/3). The parameter $h$ denotes the forecast horizon (in months). The multivariate random forests estimated on aggregate data only include, in addition to the other predictors, aggregate data as a predictor, but no regional factors. The same is true for the multivariate random forests estimated on data for the national factor. The estimates of volatility for the aggregate data and the national factor are based on a stochastic volatility model with moving-average innovations, i.e., SV-MA, and a stochastic volatility-model-in-mean model, i.e., SV-M, respectively

But volatility is unobservable, and the literature on house price volatility has primarily relied on various types of models of conditional volatility from the generalized autoregressive conditional heteroskedasticity (GARCH) family (see, Miles 2008b and Segnon et al. 2020 for detailed reviews). Given this, to generate our measure of volatility, we consider seven commonly-used GARCH models. These include the standard GARCH(1,1) model, and the more flexible models of GARCH(2,1), GARCH with jumps, GARCH in mean (M), GARCH with moving average (MA) innovations, GARCH with $t$-distributed innovations, and GARCH with an asymmetric leverage effect. In addition, we also use seven stochastic volatility (SV) models that are close counterparts of these GARCH models, and are very popular following the work of Kim et al. (1998), whereby the latent volatility follows a stochastic process. The reader is referred to (Table 1 of Chan and Grant 2016 for complete technical details of these 14 models). All the models are estimated using Bayesian techniques, and the marginal likelihoods are computed using the improved cross-entropy method of Chan and Eisenstat (2015). Based on the in-sample fits derived from the marginal likelihoods, the SV-M and SV-MA are the best suited models for the aggregate real house price growth rate and the national real house price growth factor, respectively.[12] The Clark-West test results associated with the forecasting of

[12]Complete details of the estimation results for the volatility models are available upon request from the authors.

volatilities from the full model and the model excluding the measures of macroeconomic uncertainties for the aggregate real house price growth and national factor are reported in Table 8. The test statistics depict that statistically significant forecasting gains for both the aggregate housing price growth and the national factor can be obtained by incorporating the information content of the metrics of macroeconomic uncertainty for all the three forecast horizons considered.

In sum, our results suggest that macroeconomic uncertainty cannot only improve the forecast performance of the growth rate of real prices of the overall US housing market, but also the associated volatility.

## Conclusion

This paper examines synchronization in house price movements across the US states and the role of macroeconomic uncertainty in forecasting the comovement, over and above a wide-array of predictors, namely, factors derived from a large macroeconomic dataset, oil shocks, and financial market-related uncertainties. In this regard, we use a Bayesian dynamic factor model (DFM) to decompose the house price movements for each state into a national factor that affects all the states, regional factors capturing linkages of the housing markets at the level of the four census regions, and state-specific factors which capture the idiosyncratic dynamics of each of 50 US states plus DC. We find that the national factor explains 35 % of the variation of real housing price growth rates of the 50 states and DC.

We then take machine-learning approach of random forests to analyze the ability of macroeconomic uncertainty to forecast the common movements captured by the national factor. In the forecasting exercise, we also control for eight macroeconomic factors (derived from a wide number of macroeconomic variables associated with the aggregate and regional US economy), four structural oil shocks (oil-specific supply and demand, inventory accumulation, and global demand shocks), and financial uncertainty. Random forests can not only optimally accommodate many predictors that we consider, but also account for nonlinearity in the relationship between the national factor and the predictors. We find statistically significant evidence of forecastability of the national housing factor based on the incremental information content of macroeconomic uncertainties over and above the other predictors. This result carries over with a varying degree to the four regional factors associated with the census regions, and also the overall (non-decomposed) real house price growth of the aggregate US economy. In addition, using our random-forests-modeling approach, we show that macroeconomic uncertainty can also forecast best-fitting models of volatility of the national factor and aggregate US house price, chosen from a battery of GARCH- and stochastic volatility-based models.

While accurate forecasting of volatility has important implications for investors operating in the real estate market in terms of designing optimal portfolios, the role of uncertainty in future predictions of the housing market comovement carries important information for policymakers. With the national and regional real house price growth rate factors capturing the “national bubble” and “regional bubbles”, the forecast of these important variables can guide appropriate monetary policy response of the Federal Reserve, and also possibly fiscal policy, in case there is not enough space to maneuver interest rates, to ensure that the bubbles do not burst and negatively impact the overall macroeconomy, as witnessed during the GFC.

As part of future research, it would be interesting to go beyond the state-level analysis performed in this paper, and, computations-permitting, delve into the housing markets of the 384 Metropolitan Statistical Areas (MSAs) to obtain the national and regional housing factors, and then re-conduct our predictive analysis. This is an important issue, since even though looking at the US states does allow us to incorporate the heterogeneity in the housing market, but, as rightly pointed out to us by an anonymous referee, we are still dealing with highly aggregated "markets", even at the state-level. Further, contingent on the availability of regional house price data, our analysis can also be applied to other developed and emerging countries.

As part of future research, it is interesting to go beyond the state-level analysis that we have performed in this research and, if computations are feasible, study in detail the housing markets of the 384 Metropolitan Statistical Areas (MSAs) to obtain the national and regional housing factors, and then re-conduct our predictive analysis. This is an important issue because even though looking at the US states does allow us to account for the heterogeneity in the housing market, we are still dealing with highly aggregated "markets", even at the state-level. Further, contingent on the availability of regional house price data, our analysis can also be applied to other developed and emerging countries.

## References

Ajmi, A.N., Babalos, V., Economou, F., Gupta, R. (2015). Real estate markets and uncertainty shocks: a variance causality approach. *Frontiers in Finance and Economics*, *12*(2), 56–85.

Akinsomi, O., Aye, G.C., Babalos, V., Economou, F., Gupta, R. (2016). Real estate returns predictability revisited: novel evidence from the US REITs market. *Empirical Economics*, *51*(3), 1165–1190.

André, C., Bonga-Bonga, L., Gupta R., Mwamba, J.W.M. (2017). Economic policy uncertainty, US real housing returns and their volatility: a nonparametric approach. *Journal of Real Estate Research*, *39*(4), 493–513.

Antonakakis, N., Gupta, R., André, C. (2015). Dynamic co-movements between economic policy uncertainty and housing market returns. *Journal of Real Estate Portfolio Management*, *21*(1), 53–60.

Antonakakis, N., André, C., Gupta, R. (2016). Dynamic spillovers in the United States: stock market, housing, uncertainty and the macroeconomy. *Southern Economics Journal*, *83*(2), 609–624.

Aye, G.C., Clance, M.W., Gupta, R. (2019). The effect of economic uncertainty on the housing market cycle. *Journal of Real Estate Portfolio Management*, *25*(1), 67–75.

Balcilar, M., Gupta, R., Miller, S.M. (2014). Housing and the great depression. *Applied Economics*, *46*(24), 2966–2981.

Baumeister, C., & Hamilton, J.D. (2019). Structural interpretation of vector autoregressions with incomplete identification: revisiting the role of oil supply and demand shocks. *American Economic Review*, *109*(5), 1873–1910.

Behrens, C., Pierdzioch, C., Risse, M. (2018). A test of the joint efficiency of macroeconomic forecasts using multivariate random forests. *Journal of Forecasting*, *37*(5), 560–572.

Bork, L., & Møller, S.V. (2015). Forecasting house prices in the 50 states using dynamic model averaging and dynamic model selection. *International Journal of Forecasting*, *31*(1), 63–78.

Bork, L., Møller, S.V., Pedersen, T.Q. (2020). A new index of housing sentiment. *Management Science*, *66*(4), 1563–1583.

Bouri, E., Gupta, R., Kyei, C.K., Shivambu, R. (2020). Uncertainty and daily predictability of housing returns and volatility of the United States: evidence from a higher-order nonparametric causality-in-quantiles test. University of Pretoria, Department of Economics, Working Paper No. 2020–71.

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

Campbell, J.Y. (2008). Viewpoint: estimating the equity premium. *Canadian Journal of Economics*, *41*(1), 1–21.

Chan, J.C.C., & Eisenstat, E. (2015). Marginal likelihood estimation with the cross-entropy method. *Econometric Reviews*, *34*(3), 256–285.

Chan, J.C., & Grant, A. (2016). Modeling energy price dynamics: GARCH versus stochastic volatility. *Energy Economics*, *54*, 182–189.

Christiansen, C., Eriksen, J.N., Møller, S.V. (2019). Negative house price co-movements and US recessions. *Regional Science and Urban Economics*, *77*(C), 382–394.

Christidou, M., & Fountas, S. (2018). Uncertainty in the housing market: evidence from US states. *Studies in Nonlinear Dynamics & Econometrics*, *22*(2), 20160064.

Christou, C., Gupta, R., Hassapis, C. (2017). Does economic policy uncertainty forecast real housing returns in a panel of OECD countries? A Bayesian approach. *Quarterly Review of Economics and Finance*, *65*, 50–60.

Christou, C., Gupta, R., Nyakabawo, W. (2019). Time-varying impact of uncertainty shocks on the US housing market. *Economics Letters*, *180*, 15–20.

Chuliá, H., Gupta, R., Uribe, J.M., Wohar, M.E. (2017). Impact of US uncertainties on emerging and mature markets: evidence from a quantile-vector autoregressive approach. *Journal of International Financial Markets Institutions & Money*, *48*(C), 178–191.

Clark, T.D., & West, K.D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, *138*, 291–311.

Del Negro, M., & Otrok, C. (2007). 99 Luftballons: monetary policy and the house price boom across US states. *Journal of Monetary Economics*, *54*(7), 1962–1985.

El Montasser, G., Ajmi, A.N., Chang, T., Simo-Kengne, B.D., André, C., Gupta, R. (2016). Cross-country evidence on the causal relationship between policy uncertainty and house prices. *Journal of Housing Research*, *25*(2), 195–211.

Emirmahmutoglu, F., Balcilar, M., Apergis, N., Simo-Kengne, B.D., Chang, T., Gupta, R. (2016). Causal relationship between asset prices and output in the US: evidence from state-level panel Granger causality test. *Regional Studies*, *50*(10), 1728–1741.

Ghent, A.C., & Owyang, M.T. (2010). Is housing the business cycle? Evidence from US cities. *Journal of Urban Economics*, *67*(3), 336–351.

Gupta, R. (2013). Forecasting house prices for the four census regions and the aggregate US economy in a data-rich environment. *Applied Economics*, *45*(33), 4677–4697.

Gupta, R., Kabundi, A., Miller, S.M. (2011a). Using large data sets to forecast housing prices: a case study of twenty US states. *Journal of Housing Research*, *20*(2), 161–190.

Gupta, R., Kabundi, A., Miller, S.M. (2011b). Forecasting the US real house price index: structural and non-structural models with and without fundamentals. *Economic Modelling*, *28*(4), 2013–2021.

Gupta, R., Ma, J., Risse, M., Wohar, M.E. (2018). Common business cycles and volatilities in US states and MSAs: The role of economic uncertainty. *Journal of Macroeconomics*, *57*, 317–337.

Gupta, R., Lau, C.-K.-M., Wohar, M.E. (2019). The impact of US uncertainty on the Euro area in good and bad times: evidence from a quantile structural vector autoregressive model. *Empirica*, *46*, 353–368.

Gupta, R., Marfatia, H.A., Olson, E. (2020a). Effect of uncertainty on u.s. stock returns and volatility: evidence from over eighty years of high-frequency data. https://doi.org/10.1080/13504851.2019.1677846.

Gupta, R., Ma, J., Theodoridis, K., Wohar, M.E. (2020b). Is there a national housing market bubble brewing in the United States? University of Pretoria, Department of economics, Working Paper No. 2020–23.

Gupta, R., Olasehinde-Williams, G., Wohar, M.E. (2020c). The impact of US uncertainty shocks on a panel of advanced and emerging market economies. Journal of International Trade & Economic Development. https://doi.org/10.1080/09638199.2020.1720785.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*, 2nd edn. New York: Springer.

Henderson, J.V., & Ioannides, Y. (1987). Owner occupancy: consumption vs. investment demand. *Journal of Urban Economics*, *21*, 228–241.

Jurado, K., Ludvigson, S.C., Ng, S. (2015). Measuring uncertainty. *American Economic Review*, *105*(3), 1177–1216.
Kilian, L. (2009). Not all oil price shocks are alike: disentangling demand and supply shocks in the crude oil market. *American Economic Review*, *99*, 1053–1069.
Kilian, L., & Park, C. (2009). The impact of oil price shocks on the US stock market. *International Economic Review*, *50*(4), 1267–1287.
Killins, R.N., Egly, P.V., Escobari, D. (2017). The impact of oil shocks on the housing market: evidence from Canada and US. *Journal of Economics and Business*, *93*, 15–28.
Kim, S., Shephard, N., Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies*, *65*, 361–93.
Kose, M.A., Otrok, C., Whiteman, C.H. (2003). International business cycles: world, region, and country-specific factors. *American Economic Review*, *93*(4), 1216–39.
Kose, M.A., Otrok, C., Whiteman, C.H. (2008). Understanding the evolution of world business cycles. *Journal of International Economics*, *75*(1), 110–130.
Leamer, E.E. (2007). Housing is the business cycle. In *Proceedings, economic policy symposium, Jackson Hole, Federal Reserve Bank of Kansas City* (pp. 149–233).
Leamer, E.E. (2015). Housing really is the business cycle: what survives the lessons of 2008–09? *Journal of Money, Credit and Banking*, *47*(S1), 43–50.
Ludvigson, S., Ma, S., Ng, S. (Forthcoming). Uncertainty and business cycles: exogenous impulse or endogenous response?
Ludvigson, S.C., & Ng, S. (2009). Macro factors in bond risk premia. *The Review of Financial Studies*, *22*(12), 5027–5067.
Ludvisgon, S.C., & Ng, S. (2011). A factor analysis of bond risk premia. In Ulah, A., & Giles, D. (Eds.) *Handbook of empirical economics and finance* (pp. 313–372). London: Chapman and Hall.
Marfatia, H. (2018). Modeling house price synchronization across the u.s. states and their time-varying macroeconomic linkages. Downloadable from SSRN at: https://ssrn.com/abstract=3549114.
Miles, W. (2008a). Boom-bust cycles and the forecasting performance of linear and non-linear models of house prices. *Journal of Real Estate Finance and Economics*, *36*, 249–264.
Miles, W. (2008b). Volatility clustering in U.S. home prices. *Journal of Real Estate Research*, *30*, 73–90.
Neely, C.J., & Rapach, D.E. (2011). International comovements in inflation rates and country characteristics. *Journal of International Money and Finance*, *30*(7), 1471–1490.
Nguyen Thanh, B., Strobel, J., Lee, G. (2018). A new measure of real estate uncertainty shocks. Real Estate Economics. https://doi.org/10.1111/1540-6229.12270.
Nyakabawo, W.V., Miller, S.M., Balcilar, M., Das, S., Gupta, R. (2015). Temporal causality between house prices and output in the U.S.: a bootstrap rolling-window approach. *North American Journal of Economics and Finance*, *33*(1), 55–73.
Otrok, C., & Whiteman, C.H. (1998). Bayesian leading indicators: measuring and predicting economic conditions in Iowa. *International Economic Review*, *39*(4), 997–1014.
Plakandaras, V., Gupta, R., Gogas, P., Papadimitriou, T. (2015). Forecasting the U.S. real house price index. *Economic Modelling*, *45*(C), 259–267.
Rahman, R. (2017). Multivariaterandomforest: models multivariate cases using random forests. R Package Version 1.1.5. https://CRAN.R-project.org/package=MultivariateRandomForest.
Rapach, D.E., & Strauss, J.K. (2009). Differences in housing price forecastability across US States. *International Journal of Forecasting*, *25*(2), 351–372.
Rapach, D.E., & Zhou, G. (2013). Forecasting stock returns. In Elliott, G., & Timmermann, A. (Eds.) *Handbook of economic forecasting*, (Vol. 2A pp. 328–383). Amsterdam: Elsevier.
R Core Team (2019). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Sargent, T., & Sims, C.A. (1977). Business cycle modeling without pretending to have too much a priori economic theory. In *New methods in business cycle research: proceedings from a conference, 45–109, Federal Reserve Bank of Minneapolis*.
Segal, M., & Xiao, Y. (2011). Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1.1*, 80–87.
Segnon, M., Gupta, R., Lesame, K., Wohar, M.E. (2020). High-frequency volatility forecasting of US housing markets. Journal of Real Estate Finance and Economics. https://doi.org/10.1007/s11146-020-09745-w.

Sheng, X., Marfatia, H.A., Gupta, R., Ji, Q. (2020). House price synchronization across the US states: the role of structural oil shocks. University of Pretoria, Department of Economics, Working Paper No. 2020–76.
Shiller, R.J. (1998). *Macro markets: creating institutions for managing society's largest economic risks*. New York: Oxford University Press.
Stock, J.H., & Watson, M.W. (1989). New indexes of coincident and leading economic indicators. *NBER Macroeconomics Annual*, *4*, 351–409.
Stock, J.H., & Watson, M.W. (1993). A procedure for predicting recessions with leading indicators: econometric issues and recent experience. In Stock, J.H., & Watson, M.W. (Eds.) *Business cycles, indicators and forecasting. NBER studies in business cycles*, Vol. 28. Chicago: University of Chicago Press for the NBER.
Sun, X., & Tsang, K.P. (2019). Large price movements in housing markets. *Journal of Economic Behavior & Organization*, *163*, 1–23.
Zhou, Y., & Haurin, D.R. (2010). On the determinants of house value volatility. *The Journal of Real Estate Research*, *32*, 377–396.



## Affiliations

**Rangan Gupta[1] · Hardik A. Marfatia[2] · Christian Pierdzioch[3] · Afees A. Salisu[4]**

Rangan Gupta
rangan.gupta@up.ac.za

Christian Pierdzioch
macroeconomics@hsu-hh.de

Afees A. Salisu
adebare1@yahoo.com

1 Department of Economics, University of Pretoria, Pretoria 0002, South Africa

2 Department of Economics, Northeastern Illinois University, 5500 N St Louis Ave, BBH 344G, Chicago, IL 60625, USA

3 Department of Economics, Helmut Schmidt University, Holstenhofweg 85, P.O.B. 700822, 22008 Hamburg, Germany

4 Centre for Econometric & Allied Research, University of Ibadan, Ibadan, Nigeria