Forecasting the U.S. real house price index

Vasilios Plakandaras [a], Rangan Gupta [b], Periklis Gogas [a,*], Theophilos Papadimitriou [a]

[a] Department of Economics, Democritus University of Thrace, Greece
[b] Department of Economics, Pretoria University, South Africa

ARTICLE INFO

ABSTRACT

The 2006 sudden and immense downturn in U.S. house prices sparked the 2007 global financial crisis and revived the interest about forecasting such imminent threats for economic stability. In this paper we propose a novel hybrid forecasting methodology that combines the Ensemble Empirical Mode Decomposition (EEMD) from the field of signal processing with the Support Vector Regression (SVR) methodology that originates from machine learning. We test the forecasting ability of the proposed model against a Random Walk (RW), a Bayesian Autoregressive and a Bayesian Vector Autoregressive model. The proposed methodology outperforms all the competing models with half the error of the RW model with and without drift in out-of-sample forecasting. Finally, we argue that this new methodology can be used as an early warning system for forecasting sudden house price drops with direct policy implications.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction and literature review

The last global financial crisis restored a keen interest in both academic and policy circles on the role of asset prices and in particular housing prices in economic activity. As Leamer (2007) notes the housing market predicted eight of the ten post-World War II recessions, acting as a leading indicator for the real sector of the economy. In fact he goes as far as to state that "Housing is the business cycle". Vargas and Silva (2008) argue that housing price adjustments play an important role in the determination of the phase of the business cycle. When the economy booms, construction and employment in the housing sector expand rapidly to respond to excess demand, rapidly pushing nominal house prices upwards. During the contraction phase, the drop in private income reduces aggregate demand and nominal house prices. Nevertheless, nominal house prices typically fall sluggishly since householders are unwilling to lower their prices. Most of the adjustment is achieved through decreases in sales volume resulting in a drop in the construction sector and the housing based employment. Moreover, during contraction and recession real house prices fall rapidly as general inflationary trends reduce real house prices even with sticky nominal prices.

Recently, several authors reach to empirical findings that house prices can be instrumental in forecasting output (Das et al., 2009, 2010, 2011; Forni et al., 2003; Gupta and Das, 2010; Gupta and Hartley, 2013; Stock and Watson, 2003). The housing construction sector represents a large part of total economic activity expressed in the

GDP. Consequently, as it reflects a large portion of the overall wealth of the economy, house price fluctuations can be an indicator of the evolution of GDP (Case et al., 2005). As it is the case with other assets, the movement of house prices can be also an indicator of the future direction of inflation (Gupta and Kabundi, 2010). Overall, accurate forecasting of the evolution path of house prices can be a useful tool both to house market participants and monetary policy authorities.

There is a vast literature regarding U.S. house price forecasting. Rapach and Strauss (2007) use an autoregressive distributed lag (ARDL) model framework, containing 25 determinants to forecast real housing price growth for the individual states of the Federal Reserve's Eighth District. They find that ARDL models tend to outperform a benchmark AR model. Rapach and Strauss (2009) extend the same analysis on the 20 largest U.S. states based on ARDL models examining state, regional and national level variables. Once again, the authors reach similar conclusions on the importance of combining forecasts of models with different lag structures. Gogas and Pragidis (2011) use the risk premium calculated as the difference between various long-term interest rates and the agents' expectations about future short-term rates as an input variable in predicting the future direction of house prices. They conclude that investors and analysts can use effectively the information provided by the interest rate risk premium today in order to estimate the probability of obtaining a below-trend S&P CS-10 index three months ahead.

Gupta and Das (2010) also forecast the recent downturn in real house price growth rates for the twenty largest U.S. states. The authors use Spatial Bayesian VARs (BVARs), based only on monthly real house price growth rates, to forecast their downturn over the period 2007:01 to 2008:01. They find that BVAR models are well-equipped in forecasting the future direction of real house prices, though they

* Corresponding author at: Department of Economics, Democritus University of Thrace, University Campus, Komotini, Greece. Tel.: + 30 2531039555.
E-mail address: pgkogkas@ierd.duth.gr (P. Gogas).

significantly underestimate the decline. They attribute this under-prediction of the BVAR models to the lack of any information on fundamentals in the estimation process.

Das et al. (2010) use small-scale BVARs, Bayesian Factor Augmented VARs (BFAVARs) and large-scale BVARs in forecasting house prices of nine census regions. The authors use the standard Minnesota Bayesian prior to estimate the Bayesian models. They indicate that the BFAVARs are best-suited for forecasting the house price growth rates of the nine census divisions. Gupta et al. (2011) examine the explanatory power of small and large sets of economic variables, using atheoretical models such as VAR, BVAR, FAVAR, BVAR, BFAVAR, and forward-looking structural Dynamic Stochastic General Equilibrium (DSGE) models. Based on the average root mean-squared errors for the one-, two-, three-, and four-quarters-ahead forecasts, they find that the small-scale Bayesian-shrinkage model fed with 10 variables outperforms all the other input sets. Gupta (2013) uses dynamic factor and Bayesian shrinkage models in a large number of predictors (145 variables) and forecasts of house prices for four U.S. census regions and for the aggregate economy. The results show that the BFAVAR models exhibit the best forecasting ability. Similar results were also obtained by Gupta and Kabundi (2010) and Aye and Gupta (forthcoming) using Bayesian predictive regressions in forecasting the overall U.S. house price index. More recently, Balcilar et al. (forthcoming) compared the ability of non-linear AR models in forecasting nominal house price growth rates of four U.S. census regions and the aggregate economy, relative to an AR model. Interestingly, even though they could detect nonlinearity in the in-sample for all the 5 growth rates of house prices, when it came to out-of-sample point, interval and density forecasting, the evidence in favor of the nonlinear model was virtually non-existent.[1] In light of the above discussion, it is clear that, in general, Bayesian models are best-suited for forecasting house prices — regional or for the aggregate U.S.

In this paper we build on previous empirical studies by comparing the econometric Bayesian Vector Autoregressive (BVAR) and Bayesian Autoregressive (BAR) models, instead of Bayesian predictive regressions to avoid issues of endogeneity, with a novel forecasting methodology on one-year-ahead forecasting. We propose a methodology that combines Ensemble Empirical Mode Decomposition from the field of signal processing with the machine learning Support Vector Regression methodology for constructing forecasting models. In comparison to previous studies we use a much longer sample spanning annual observations from 1890 to 2012 and evaluate the use of 11 macroeconomic variables. Finally, we expand our research framework into multi-period ahead forecasting in an effort to evaluate our forecasting methodology as a house price early-warning system, focusing on the 2006–2007 house price downturn.

The rest of the paper is organized as follows: the methodology is discussed in Section 2, the dataset is presented in Section 3 while the empirical results are reported in Section 4. Finally, Section 5 concludes.

2. Methodology

2.1. Overview

A common issue among econometricians when developing forecasting models is that almost all time series are aggregations of more volatile elements. For instance, the annual dataset of real annual housing prices in the U.S. consists of quarterly or even monthly price surveys which are aggregated. As the sampling frequency decreases short-term dynamics diminish and long-term characteristics such as trend or seasonality are to be observed. The use of such input variables in a forecasting system incorporates error, making it difficult to create accurate models. Smoothing techniques can reduce the influence of noise

and errors in observed data, offering a less volatile representation of the underlying phenomenon. Nonetheless, essential information may be lost during the smoothing process.

A key issue in all smoothing methodologies is the definition of the optimal point, beyond which smoothing results to distortion rather than noise reduction. In real applications, noise-free time series are unobservable and we can only approximate them through their smoothed counterparts. Many smoothing implementations select *ad hoc* the smoothness degree imposed; see for example the Hodrick and Prescott (1997) filter and the discussion regarding the optimum value of the smoothness parameter λ with respect to data frequency. To mitigate proper model selection dilemmas, our implementation exploits a relatively novel signal decomposition method called Ensemble Empirical Mode Decomposition (EEMD) used as a smoothing function. The main advantages of this smoothing framework are a) the absence of *a priori* assumptions and that b) the tuning of the model's parameters is solely based on data characteristics.

In this paper we propose a three step forecasting methodology as follows:

Step 1 A smoothing of the initial series employing the EEMD
Step 2 A variable selection process using the Elastic Net approach
Step 3 Fitting a Support Vector Regression model for forecasting.

In the first step we decompose all input series with EEMD into a low frequency (smoothed) and a remaining highly volatile (fluctuating) component that approximate the long-term trend and short-term dynamics,[2] respectively. By de-composing the initial time series into two distinct components we are able to use a distinct forecasting model for each of the individual components. In this way each model is better suited to the individual data characteristics. The model regarding the smoothed component of the output time series is fed with the smoothed component of the input time series. The same mechanism is used accordingly for the fluctuating component. The variable selection in every case is conducted using the Elastic Net (Zou and Hastie, 2005) and the selected variable set is forwarded to an SVR model. The proposed system yields one forecasted time series for the smoothed and one for the fluctuating component of the initial time series. Overfitting is avoided by adopting a 4-fold cross-validation scheme in the training step. The sum of the two forecasted components is the final output of our methodology and it is evaluated for its out-of-sample forecasting accuracy. An overview of the method is depicted in Fig. 1.

2.2. VAR and BVAR specifications

Following the work of Sims (1980), we create an unrestricted VAR model as follows:

$$\boldsymbol{y}_t = A_0 + A(L)\boldsymbol{y}_t + e_t, \tag{1}$$

assuming that $\varepsilon_t \sim N(0, \sigma^2 I_n)$ and I_n is an $n \times n$ identity matrix where $\boldsymbol{y}_t$ is an $n \times 1$ vector of variables, $A(L)$ is an $n \times n$ polynomial matrix in the backshift operator L with lag length p $[A(L) = A_0 + A_1 L + A_2 L^2 + \cdots + A_p L^p]$, A_0 is an $n \times n$ vector of constant terms, and e_t is an $n \times n$ vector of error terms.

VAR models use equal lag lengths for all variables in the model, which implies that the researcher must estimate many parameters, including some that may be statistically insignificant. This over-parameterization problem can create multicollinearity, leading to possibly large out-of-sample forecasting errors. Litterman (1981, 1986), Doan et al. (1984) and Spencer (1993) use a Bayesian VAR (BVAR) model to overcome the over-parameterization problem. Rather
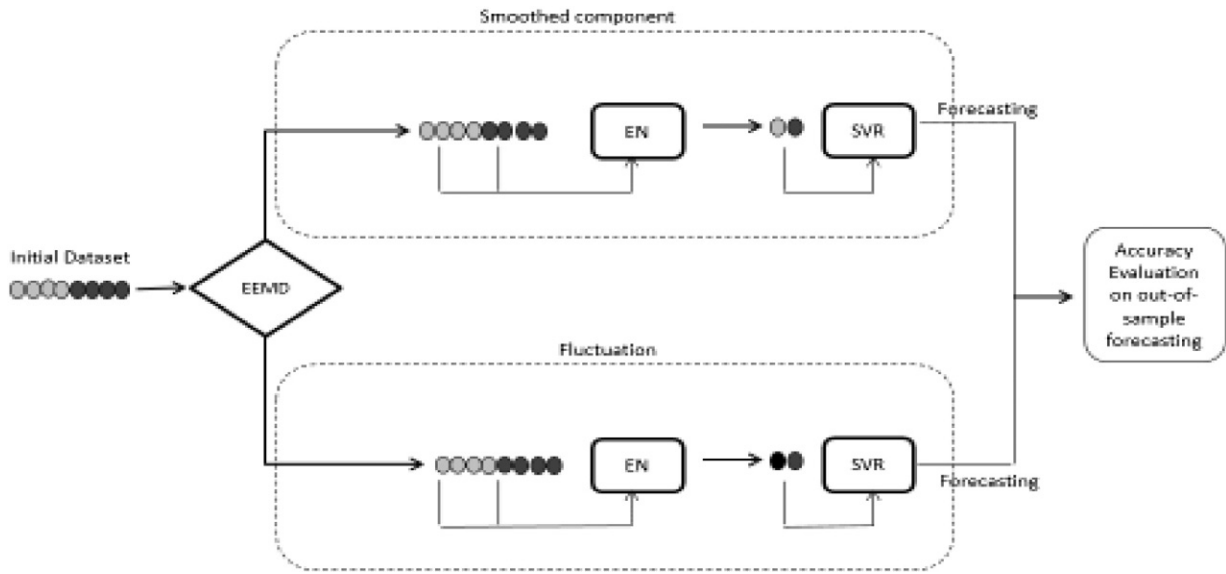

Fig. 1. The hybrid EEMD–EN–SVR structural model.

than eliminating statistically insignificant coefficients, they impose different weights on the coefficients across different lag lengths, assuming that longer lag coefficients are less significant than shorter lag ones. In our Bayesian variants of the classical VAR and VEC models we follow the propositions of Litterman (1981) that proposes a diffuse prior to the constant, often named in literature as the "Minnesota prior". Formally, the means and variances of the Minnesota prior take the form:

$$\beta_i \sim N\left(1, \sigma_{\beta_i}^2\right) \text{ and } \beta_j \sim N\left(0, \sigma_{\beta_j}^2\right) \qquad (2)$$

where β_i are the coefficients associated with the lagged dependent variables in each equation of the VAR model and β_j are all other coefficients. Setting all variances to zero, the prior specification for every variable reduces to a Random Walk model with a drift. The prior variances, $\sigma_{\beta_i}^2$ and $\sigma_{\beta_j}^2$, specify uncertainty about the prior with means $\overline{\beta}_i = 1$, and $\overline{\beta}_i = 0$, respectively.

Doan et al. (1984) propose that standard deviations could be calculated based on a small number of hyper-parameters: q, g, and a weighting matrix $f(i, j)$. The standard deviation of the prior imposed on variable j in equation i at lag m, for all i, j and m is as follows:

$$S(i, j, m) = [q \times g(m) \times f(i, j)]\frac{\hat{\sigma}_i}{\hat{\sigma}_j}, \qquad (3)$$

$$\text{Subject to} \begin{cases} f(i, j) = 1, \text{ if } i = j \\ f(i, j) = k_{ij} \text{ otherwise, with } 0 \leq k_{ij} \leq 1 \\ g(m) = m^{-d}, \text{ with } d > 0 \end{cases}$$

where $\hat{\sigma}_i$ is the estimated standard error of the univariate autoregression for variable i. The ratio $\frac{\hat{\sigma}_i}{\hat{\sigma}_j}$ scales the variables to account for differences in the units of measurement, while term q indicates the overall tightness of the model, with the prior getting tighter as the value falls. The parameter $g(m)$ measures the tightness on lag m with respect to the first lag and tightens the prior as the lag order increases. The weighting matrix $f(i, j)$ equals the tightness of variable j in equation i and by increasing the interaction (i.e., the value of k_{ij}), we loosen the prior.[3] Sims et al. (1990) argue that with the Bayesian approach entirely based on the likelihood function, the associated inference does not require stationarity, since the likelihood function exhibits the same Gaussian shape regardless of the presence of unit roots. Thus, in this paper we specify BVAR and BAR models in levels.

2.3. Ensemble Empirical Mode Decomposition and time series smoothing

The EEMD is a data driven algorithm that decomposes a time series into finite additive oscillatory components called Intrinsic Mode Functions (IMFs). Proposed by Wu and Huang (2009), the decomposition into IMFs is achieved through an iterative scheme until the residual has no local extrema. The amplitude of the added Gaussian noise during the decomposition procedure is chosen according to the maximized relative RMSE criterion proposed by Guo and Tse (2013).[4]

An example of a decomposition applied on the real annual U.S. house prices is depicted in Fig. 2.

During decomposition, the frequency of every IMF drops as its index number increases. Wu et al. (2007) propose that each IMF represents different dynamics of the time series, with the more volatile IMFs depicting short-run dynamics and the less volatile ones long-run trends of the phenomenon. Moghtaderi et al. (2013) build on this framework proposing a trend extraction technique based on decomposition. They argue that long run trends should be examined as summations rather than independent IMFs, since individual characteristics are dispersed between IMFs and are not exclusively isolated in only one IMF. In other words they suggest that the long-run trend is the summation of 1) the lower frequency IMFs and 2) the final residual of the EEMD decomposition.

Expanding the proposition of Moghtaderi et al. (2013) we argue that all IMFs can be evaluated for constructing a representation of short-run dynamics of the initial time series and a long-run trend function. The former is the result of the summation of more volatile IMFs while the latter can be obtained following the procedure described by Moghtaderi et al. (2013). Thus, the segregation problem breaks down to selecting the most appropriate IMF index which defines the limit between short- and long-term dynamics. With mathematical notation, the above smoothing function is expressed with:

$$\boldsymbol{h}_{i*} = \sum_{i=i_*}^{R} \text{IMF}^i + r, \quad 1 \leq i \leq L \qquad (4)$$

where $\boldsymbol{h}_{i*}$ is the smoothed variant of the initial series for index i_*, R the total number of IMFs, r the final EEMD residual and i_* the index of the IMF beyond which we start the summation. For instance, in Fig. 3 the decomposition of the real house price time series results into 5 IMFs

---

[3] For an illustration, see Dua and Ray (1995).

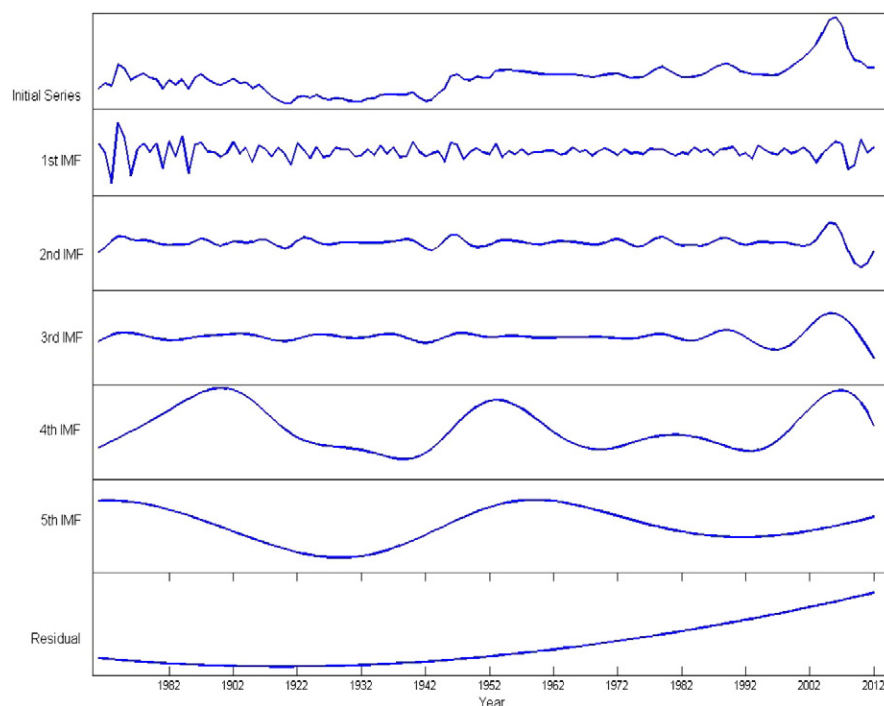[4] For more details on the method see Wu and Huang (2009).

Fig. 2. Decomposition of the original (1st row) real annual U.S. house prices into 6 series. The last row (straight line) is the residue of the EEMD process.

and a final residual. By summing the 3rd, 4th and 5th IMFs with the final residual, we obtain the smoothed function of the red curve of Fig. 3.

2.4. Support Vector Regression (SVR)

The Support Vector Regression is a direct extension of the classic Support Vector Machine algorithm proposed by Vapnik et al. (1992) and originates from the field of statistical learning. Despite the ability of the methodology to provide accurate forecasts with high generalization ability, it has not attracted significant interest in forecasting economics and financial time series. Among the few empirical applications, Rubio et al. (2011) forecast electric loads based on SVR models, while Papadimitriou et al. (2014) develop SVM models for directional price forecasting in electric energy markets. Härdle et al. (2009) evaluate the default risk of companies with SVM and Öğüt et al. (2012) extend default risk forecasting in the banking sector. Khandani et al. (2010) use SVR models for private credit risk evaluation and Gogas et al.

(forthcoming) use SVR models for recession forecasting. Finally, Gogas et al. (2013) compare Simple Sum and Divisia monetary aggregates under a machine learning framework in order to forecast the U.S. GDP.

The basic idea is to find a linear function that has at most a predetermined deviation from the actual values of the dataset. In other words, we do not care about the error of each forecast as long as it doesn't violate the predefined threshold, but we penalize any deviations higher than the threshold. The set that bounds this "error-tolerance band" is the Support Vector (SV). This, is located through a minimization procedure.

One of the main advantages of the SVR in comparison to other machine learning techniques is that, it typically yields a convex minimization problem, avoiding local minima (Vapnik et al., 1992). The model is built in two steps: the training and the testing step. In the training step, the largest part of the dataset is used for the estimation of the function (i.e. the detection of the Support Vectors that define the "error-tolerance band"); in the testing step, the generalization ability of the model is
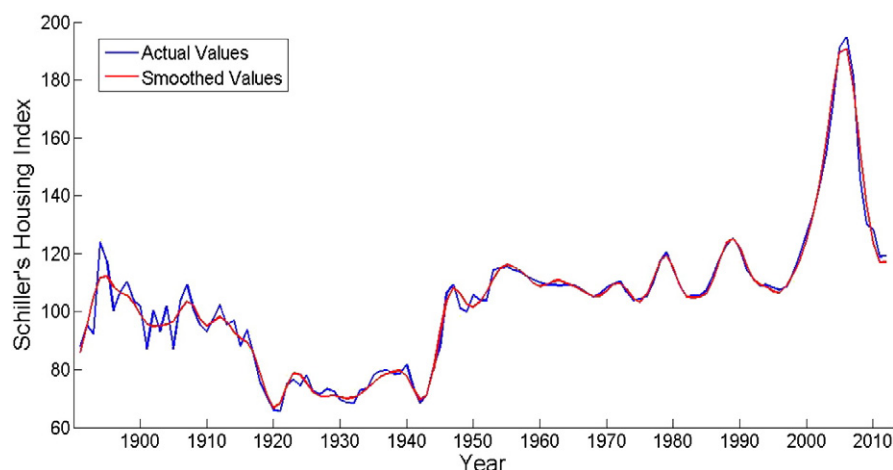

Fig. 3. Real U.S. house prices and its EEMD smoothed series.

evaluated by checking the model's performance in the small subset that was left aside during training.

Using mathematical notation and starting from a training dataset $D = [(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_n, y_n)]$, $\boldsymbol{x}_i \in \mathbb{R}^m$, $y_i \in \mathbb{R}$, and $i = 1, 2, ..., n$ where for each observation pair, $\boldsymbol{x}_i$ values are the observation samples and y_i is the dependent variable (the target of the regression system) and the linear regression function takes the form of $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$. The SVR methodology tries to reach two contradictory goals: a) to find a solution that best approximates the given dataset (*i.e.* a large part of the data points should be inside the tolerance "belt", while a few points will lie out of bounds) and b) to find a solution that generalizes to the underlying population. This is achieved by solving:

$$\min\left(\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*)\right)$$
$$\text{subject to} \begin{cases} y_i - (\boldsymbol{w}\boldsymbol{x}_i + b) \leq \varepsilon + \zeta_i \\ (\boldsymbol{w}\boldsymbol{x}_i + b) - y_i \leq \varepsilon + \zeta_i^* \\ \zeta_i,\ \zeta_i^* \geq 0 \end{cases} \quad (5)$$

where ε defines the tolerance belt around the regression, and ζ_i and ζ_i^* are slack variables controlled through a penalty parameter C (see Fig. 4). All the points inside the tolerance belt have $\zeta_i, \zeta_i^* = 0$. Problem (5) is a convex quadratic optimization problem with linear constraints and has a unique solution. The first part of the objective function controls the generalization ability of the regression, by imposing a smaller possible $\|\boldsymbol{w}\|$. This is not an obvious statement and a detailed analysis of the SVR minimization process is not in the scope of this paper; however we can hint that the smaller the $\|\boldsymbol{w}\|$, the closer to parallel to the x-axes is the regression function. Geometrically we can see that a parallel line to the x-axes, maximizes the covered area by the tolerance belt, which means maximum generalization ability. The second part of the objective function controls the regression approximation to the training data points (by increasing C we penalize with a bigger weight any point outside the tolerance belt *i.e.* with $\zeta_i \geq 0$ or $\zeta_i^* \geq 0$). The key element in the SVR concept is to find the balance between the two parts in the objective function, controlled by the ε and C parameters, as presented in Fig. 4, and the solution is given by:

$$\boldsymbol{w} = \sum_{i=1}^{n}(a_i - a_i^*)\boldsymbol{x}_i \quad (6)$$

and

$$y = \sum_{i=1}^{n}(a_i - a_i^*)\boldsymbol{x}_i^T\boldsymbol{x}. \quad (7)$$

Real life phenomena are rarely described correctly by a linear regression; they are too complex for such a simplistic approximation. An alternative method to treat real phenomenon datasets would be to project them into a higher dimensional space where the transformed dataset may be described by a linear function. The "kernel trick" follows the projection idea while ensuring minimum computational cost: the dataset is mapped in an inner product space, where the projection is performed using only dot products within the original space through special "kernel" functions, instead of explicitly computing the mapping of each data point. Non-linear kernel functions have evolved the SVR mechanism to a non-linear regression model.

In our simulations we employed four kernels: the linear, the radial basis function (RBF), the sigmoid and the polynomial. The mathematical representation of each kernel is:

Linear $\quad K_1(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1^T\boldsymbol{x}_2 \quad\quad (8)$

RBF $\quad K_2(\boldsymbol{x}_1, \boldsymbol{x}_2) = e^{-\gamma\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2} \quad\quad (9)$

Polynomial $\quad K_3(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left(\gamma\boldsymbol{x}_1^T\boldsymbol{x}_2 + r\right)^d \quad\quad (10)$

Sigmoid $\quad K_4(\boldsymbol{x}_1, \boldsymbol{x}_2) = \tanh\left(\gamma\boldsymbol{x}_1^T\boldsymbol{x}_2 + r\right) \quad\quad (11)$

with factors d, r, and γ representing kernel parameters.

2.5. Elastic Net

The Elastic Net is a variable selection method proposed by Zou and Hastie (2005) that linearly combines the LASSO (Tibshirani, 1996) and ridge regression techniques. According to the LASSO, when solving a regression problem we try to minimize the squared difference of actual to forecasted values, while imposing the constraint that only the coefficients with an absolute value greater than a threshold are acceptable. Coefficients with a value lower than the threshold are set to zero and thus discarded. In this way the LASSO methodology tries to minimize the forecasting error of the regression model, while pruning regressors. The difference of ridge regression to the LASSO methodology is that instead of imposing a threshold on the absolute value of each coefficient, it evaluates its squared value and discards all coefficients with a squared value below a threshold.

The Elastic Net combines the aforementioned methods in a more flexible framework. Instead of using a fixed form (absolute or squared value) for each coefficient, Zou and Hastie (2005) propose a regularization parameter that fluctuates from zero to a unit. When zero is selected, the Elastic Net reverts to the LASSO model, while when the regularization parameter is set to one, we get a ridge regression representation. In other words, the LASSO and the ridge regression are only special cases of an Elastic Net model. Overall, the aforementioned methodology is flexible, adapts to data characteristics and can select regressors that lead to accurate regression models.

3. The dataset

In comparison to previous studies we use a much longer sample spanning annual observations from 1890 to 2012, with an in-sample size of 1890–1988 and an out-of-sample size of 1989–2012. A very short sample may describe only episodic trends of the entire phenomenon, while longer periods allow observing the forecasting performances of these models over prolonged periods and changes in the trend of house prices. The decision to use an out-of-sample starting in 1989 is to cover a period of sharp increase in activity in the housing market (Rapach and Strauss, 2007; 2009). Our dataset consists of ten annual U.S. macroeconomic variables spanning the period 1890 to 2012 which are selected based on literature:

- The real GDP per capita (RGDPPC) (Agnello and Schuknecht, 2011; Case and Shiller, 1990)
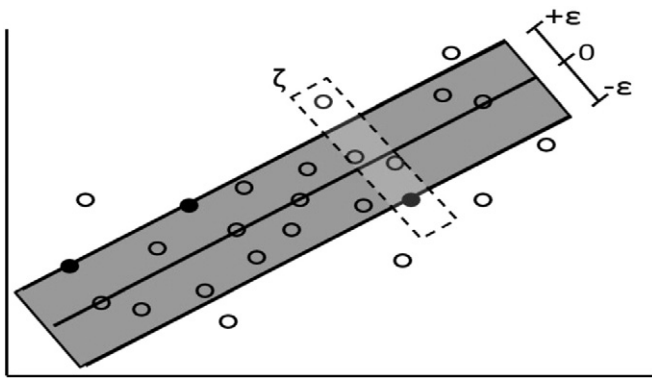- The long-term interest rate (LTR) and the short-term interest rate

Fig. 4. Upper and lower thresholds on error tolerance indicated with letter ε. The boundaries of the error tolerance band are defined by the Support Vectors (SVs) denoted by the black filled points. Forecasted values greater than ε get a penalty ζ according to their distance from the tolerance accepted band.
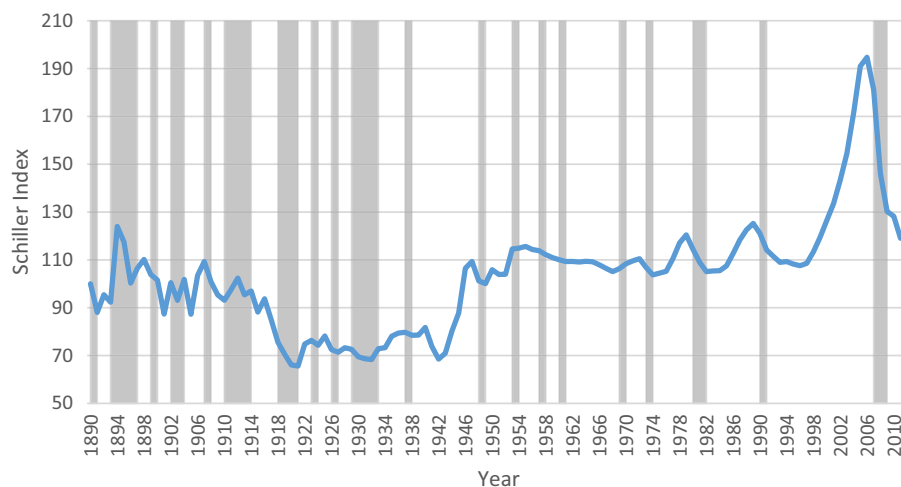
Fig. 5. Yearly real housing prices in the U.S. as measured by the Shiller housing price index (1890 = 100). Gray areas indicate recession periods of the U.S. economy as reported by the National Bureau of Economic Research (NBER).

(STR) (Agnello and Schuknecht, 2011; McGibany and Nourzad, 2004; Mikhed and Zemčík, 2009)
- The population (POP) (Agnello and Schuknecht, 2011; Case and Mayer, 1996; Case and Shiller, 1990; Mikhed and Zemčík, 2009)
- The real stock price (RSP) (Abelson et al., 2005; Mikhed and Zemčík, 2009; Rapach and Strauss, 2009)
- The real construction cost (RCONSTR) (Case and Shiller, 1990; Jud and Winkler, 2002; Mikhed and Zemčík, 2009; Zeno and Füss, 2010)
- The unemployment rate (UNEMPL) (Abelson et al., 2005; Case and Mayer, 1996; Rapach and Strauss, 2007)
- The inflation rate (INFL) (Abelson et al., 2005; Rapach and Strauss, 2007; Stevenson, 2000)
- The real oil price (ROILP) (Beltratti and Morana, 2010; Padilla, 2005)
- The ratio of budget deficit/surplus as a fiscal policy indicator (FISPOL) (Afonso and Sousa, 2011, 2012; Agnello and Schuknecht, 2011; Agnello and Sousa, 2011).

The dataset consists of 11 annual U.S. macroeconomic variables spanning the period 1890 to 2012. These include the real house price (RHP), fiscal policy variable (FISPOL), real GDP per capita (RGDPPC), unemployment (UNEMPL), long-term interest rate (LTR), short-term interest rate (STR), inflation rate (INFL), population (POP), real construction cost (RCONSTR), real stock price (RSP) and real oil price (ROILP). All variables are from the Robert J. Shiller web page, barring real GDP, population, unemployment and part of the budget surplus/deficit data used for computing the fiscal policy variable which are from the Global Financial data (GFD) base. For the budget surplus/deficit, we obtain the 1890 to 2006 data from GFD while 2007 to 2012 data are obtained from the Federal Reserve Bank, St. Louis. We use the ratio of budget surplus/deficit to GDP as our measure of fiscal policy. Inflation rate is computed as the annual rate of growth in consumer price index. Real oil price is obtained by deflating the nominal West Texas Intermediate with CPI. Fig. 5 depicts real house prices and the gray areas represent recession periods reported by the National Bureau of Economic Research. During 2006 to 2009, the historically highest price levels are followed by a sharp drop in real house prices of unprecedented magnitude.

4. Empirical results

In order to test the generalization ability of our selected model, our dataset is split into two parts for in- and out-of-sample forecasting. The ratio chosen is 80/20. The forecasting accuracy is measured by the Mean Absolute Percentage Error (MAPE) and Directional Symmetry

(DS):

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \qquad (12)$$

$$DS = \frac{100}{n} \sum_{i=1}^{n} d_i, \text{ where } d_i = \begin{cases} 1 \text{ if } (y_i - y_{i-1})(\hat{y}_i - \hat{y}_{i-1}) > 0 \\ 0 \text{ otherwise} \end{cases} \qquad (13)$$

where $\hat{y}_i$ and y_i stand for the forecasted and actual values respectively while n is the total number of out-of-sample observations. The MAPE measures the absolute percentage error in the forecast, while the DS is a measure of directional forecasting accuracy. Directional forecasting is of key interest to house market participants and policy authorities, since as discussed, the future direction of house prices can be an indicator of the current and future phases of the business cycle.

In order to test the forecasting ability of the proposed EEMD–EN–SVR methodology in forecasting the real house price index, we compare it with several alternative forecasting models. Apart from the EEMD–EN–SVR we develop a Random Walk (RW) and a simple autoregressive version without any extra-explanatory variables labeled EEMD–AR–SVR; only the past values of the smoothed and the fluctuating part of the house price index are used as input variables. The optimum lag structure for both the EEMD–AR–SVR and EEMD–EN–SVR models is selected according to the minimum in-sample MAPE. For the EEMD–AR–SVR model we select two lags for the fluctuating component and six lags for the smoothed one. Following the shrinkage scheme of the BVAR model described in Section 2, we introduce lagged values of all predictors and leave EN to choose among them the most informative ones. In other words as in the BVAR framework the EN may select the

Table 1
In- and out-of-sample forecasting results.

Model	In-sample accuracy		Out-of-sample accuracy	
	MAPE (%)	DS (%)	MAPE (%)	DS (%)
RW	4.732	56.701	5.352	70.833
RW with a drift	6.588	52.577	5.387	70.833
BAR	4.952	55.670	5.422	70.833
BVAR	4.809	56.701	11.931	75.000
EEMD–AR–SVR (linear)	1.985*	84.615*	2.151*	87.500*
EEMD–EN–SVR (linear)	2.337	78.495	5.990	87.500*

Note: Best values are noted with an *. Best kernel for the SVR models is noted in parentheses.
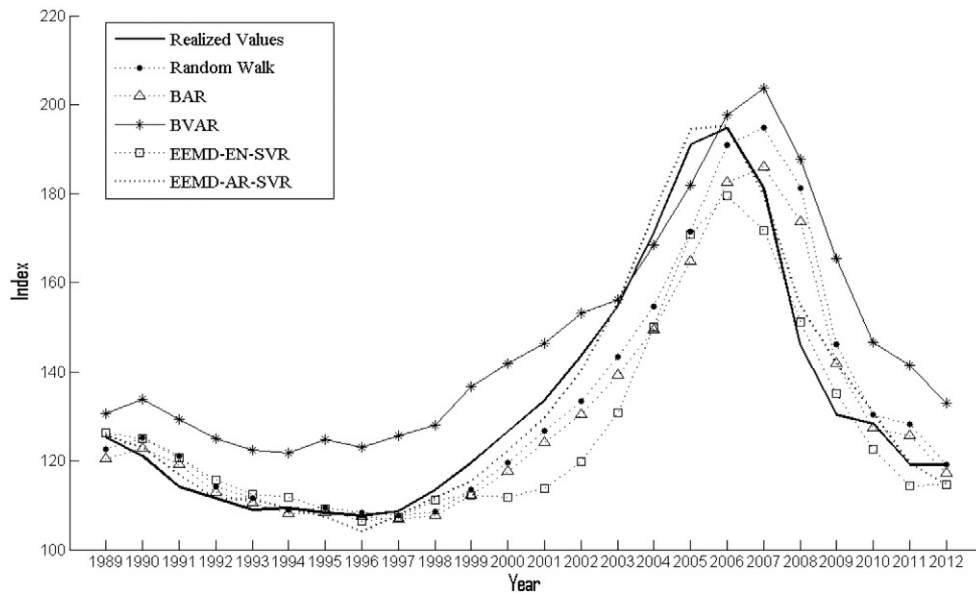
Fig. 6. Out-of-sample forecasted values.

second lag of a variable but exclude the first one, on the contrary to classic VAR models. The EN selects one input variable for the fluctuating part and 36 for the smoothed one.[5]

Moreover, we also develop an autoregressive BAR and a BVAR model that have been used extensively in house price forecasting literature. A key advantage of the shrinkage procedure of the BVAR models is that they adapt to data characteristics, excluding irrelevant variables and focusing on the more informative ones with respect to forecasting. The lag structure selection for the Bayesian models is based on the minimum Schwartz Information Criterion (SIC) (Schwarz, 1978) value. For both the BAR and the BVAR model the selected lag is one.[6] We consider various values for the tightness term q, the decay factor g and the standard deviation of the distribution of the prior k_{ij}. The values $q = 0.3$, $g = 0.5$ and $k_{ij} = 0.5$ yield the highest in-sample accuracy and they are used to train the BVAR to attain the out-of-sample forecasts. The BAR model is obtained by setting the interaction term $k_{ij} = 0.001$, while q and g remain the same as in the BVAR model.[7] As proposed by the literature, we fix the parameter a of the EN penalty term to 0.5, since simulations with $a = 0.1, 0.3, 0.7$ and 0.9 yield quantitative similar results.

In Table 1 we report the in-sample and out-of-sample forecasting accuracy in one period ahead static forecasting.

We observe that the highest in-sample accuracy according to the MAPE is achieved with the EEMD–AR–SVR model closely followed by the EEMD–EN–SVR model. The latter is also the most accurate model according to the Directional Symmetry criterion. When we evaluate the out-of-sample forecasting ability of our models in the last two columns of Table 1, we observe that the EEMD–AR–SVR model outperforms all other models for both the MAPE and DS criteria. The performance of the EEMD–EN–SVR model drops significantly in terms of the MAPE (the forecasting error doubles) and it is now outperformed by all other models with the exception of the BVAR. In terms of the DS criterion the EEMD–EN–SVR model provides the second most accurate directional forecasting after the EEMD–AR–SVR model. Finally, both the BAR

and BVAR models perform poorly and exhibit a lower forecasting accuracy in both in- and out-of-sample forecasting as compared to the SVR models with the exception of the BAR that provides a better MAPE than the EEMD–EN–SVR model. Moreover, both the BAR and BVAR models provide less accurate forecasts than the simple naïve Random Walk model in-sample and out-of-sample. On the other hand, since the autoregressive EEMD–AR–SVR model outperforms the RW model, we find evidence against the Efficient Market Hypothesis (EMH) of Eugene Fama (1965). Since with the EEMD–AR–SVR model the real house price index can be adequately forecasted with only historical (lagged) values of itself we can conclude that even weak form efficiency is not supported by the data in the U.S. housing market.

In order to have a visual representation of the forecasting ability of the models presented in Table 1, we depict in Fig. 6 the out-of-sample forecasts.

As we observe from Fig. 6, the EEMD–AR–SVR model is the only one of the five models tested that timely and efficiently forecasts the 2006–2009 sudden downturn in real U.S. housing prices that sparked the latest financial crisis. The EEMD–EN–SVR model performs very poorly with the out-of-sample data. This may be an indication of overfitting for the EEMD–EN–SVR model. The other three models, capture the downturn in real house prices with one lag, only one year later than it actually occurred.

In Fig. 7 we depict only the actual values and the EEMD–AR–SVR model forecasts. It is clear that the proposed methodology traces very closely the actual prices on the U.S. housing market.

Next, we produce forecasts for longer forecasting windows. From one period (year) ahead that we used thus far, we now extend our forecasting up to ten periods (years) ahead. In this analysis we only use the EEMD–AR–SVR model that outperformed all other tested models in the one period ahead out-of-sample forecasts. The RW model is also included as a benchmark.

As expected, the forecasting accuracy of the model deteriorates as the forecasting horizon increases. Both in in- and out-of-sample forecasting and for both the MAPE and DS criteria the EEMD–AR–SVR model outperforms the RW model for forecasting windows from 1 to 4 years ahead. This superiority of the EEMD–AR–SVR model is maintained in out-of-sample forecasting for both criteria up to 6 years ahead. Moreover, according to the DS the EEMD–AR–SVR outperforms the RW model up to 9 years ahead. Only for the last year (10) of out-

---

[5] The selected variables can be accessed at http://utopia.duth.gr/~vplakand/Selected_Variables_by_the_Elastic_Net.pdf.

[6] The application of the Akaike Information Criterion (Akaike, 1974) and Hannan–Quinn Information Criterion (Hannan and Quinn, 1979) as expected selected a higher lag order. We follow the SIC that results in more parsimonious models.

[7] As suggested by the extant Bayesian literature, we also experimented with the following combinations of q and g respectively: (0.2, 1.0); (0.1, 1.0); (0.2, 2.0); and (0.1, 2.0).
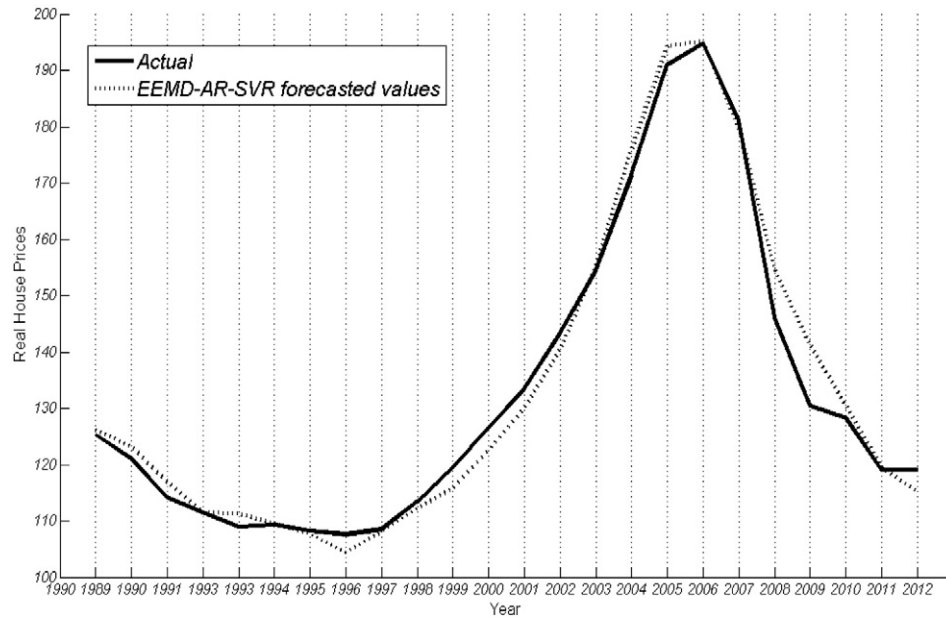
Fig. 7. EEMD–AR–SVR out-of-sample forecasted and actual values.

of-sample forecasts the RW is superior to the EEMD–AR–SVR model according to both criteria used. For the MAPE and in-sample forecasting the EEMD–AR–SVR model is superior to the RW model for all 10 forecasting windows. In order to better observe the forecasting ability of the proposed methodology, we depict the forecasted values of the EEMD–AR–SVR model for the first four horizons in the out-of-sample period in Fig. 7.

The model forecasts the significant fall in 2006–2009 of U.S. housing market prices up to 2 periods (years) ahead. Overall, the proposed methodology forecasted correctly the U.S. housing price collapse nearly two years before the actual event occurred an observation that according to Leamer (2007) can be evaluated as an indicator of the business cycle phase of the economy. As observed in Fig. 7, the drop in house prices starting in 2006 signals a contraction period of the U.S. economy followed by the 2007–2008 recession period. The same phenomenon is repeated just before the brief 2 quarters of the 1991 recession where dropping prices during 1989–1990 mark a contraction period. In other

words predicting house price deterioration can be an indicator that the economy is entering a contraction period and thus monetary policy should be adjusted accordingly.

5. Conclusion

We examine eleven predictors in forecasting the real U.S. house price index comparing the econometric BAR and BVAR techniques and a novel EEMD–SVR forecasting methodology. The empirical findings suggest that the EEMD–AR–SVR model outperforms all competing models in both in-sample and out-of-sample forecasting. The reported research framework is extended to single and multi-period ahead forecasting (Fig. 8, Table 2) reporting early accurate detection of the 2006–2009 price downturn. Overall, we argue that the proposed forecasting methodology can be used effectively as a policy instrument for determining the business cycle of the economy.
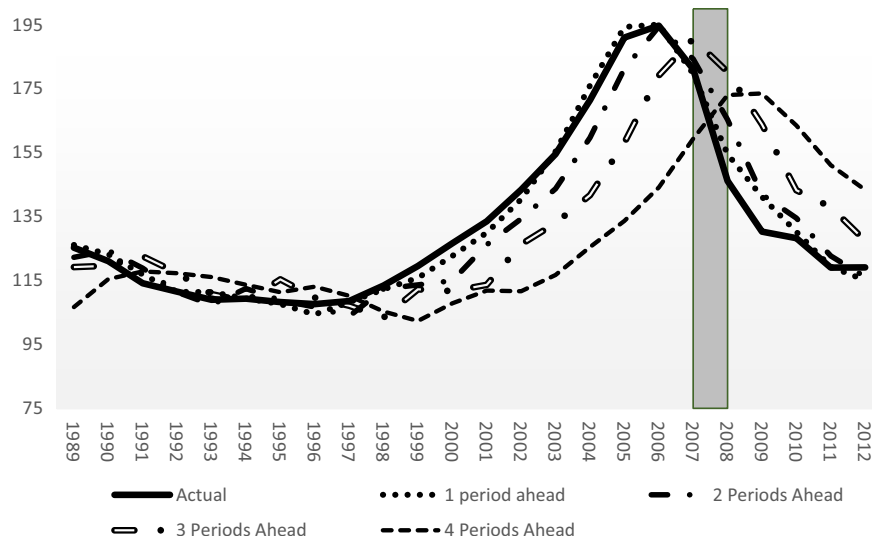

Fig. 8. Multi-period ahead forecasts. The gray area denoted the recession period of the U.S. economy as reported by the National Bureau of Economic Research.

Table 2
Multi-period ahead forecasting results.

Period	In-sample forecasting				Out-of-sample forecasting			
	RW		EEMD–AR–SVR		RW		EEMD–AR–SVR	
	MAPE (%)	DS (%)	MAPE (%)	DS (%)	MAPE (%)	DS (%)	MAPE (%)	DS (%)
1	4.389	55.670	1.985*	84.615*	5.351	70.833	2.016*	87.500*
2	6.688	52.577	3.481*	71.111*	10.462	58.333	4.164*	83.333*
3	8.353	45.833	5.669*	50.000*	14.903	54.167	9.895*	62.500*
4	8.664	40.000	6.766*	42.046*	18.672	41.667	15.344*	45.833*
5	9.667	46.809*	7.987*	42.529	20.735	37.500	16.528*	41.667*
6	10.232	48.387*	8.180*	48.148	20.760	33.333	20.566*	37.500*
7	10.640	54.348*	8.655*	50.633	19.874*	25.000	21.630	29.167*
8	10.594	56.044*	9.419*	55.844	18.133*	29.167	21.085	37.500*
9	11.201	55.556	9.757*	56.000*	16.364*	33.333	20.300	41.667*
10	11.426	57.303*	10.210*	55.405	14.713*	37.500*	18.049	29.167

Note: Best forecasts for each forecasting horizon are noted with an *.

Acknowledgments

References

Abelson, P., Joyeux, R., Milunovich, G., Chung, D., 2005. Explaining house prices in Australia: 1970–2003. Econ. Rec. 81 (S1), S96–S103.

Afonso, A., Sousa, R.M., 2011. What are the effects of fiscal policy on asset markets? Econ. Model. 28, 1871–1890.

Afonso, A., Sousa, R.M., 2012. The macroeconomic effects of fiscal policy. Appl. Econ. 44, 4439–4454.

Agnello, L., Schuknecht, L., 2011. Booms and busts in housing markets: determinants and implications. J. Hous. Econ. 20, 171–190.

Agnello, L., Sousa, R.M., 2011. Fiscal policy and asset prices. Bull. Econ. Res. 65 (20), 154–177.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19 (6), 716–723.

Aye, G.C., Gupta, R., 2014. Forecasting Real House Price of the U.S.: An Analysis Covering 1890 to 2012. J. Econ. Comput. Econ. Cybern. Stud. Res (forthcoming).

Balcilar, M., Gupta, R., Miller, M.S., 2014. The out-of-sample forecasting performance of non-linear models of regional housing prices in the US. Appl. Financ. Econ (forthcoming).

Beltratti, A., Morana, C., 2010. International house prices and macroeconomic fluctuations. J. Bank. Financ. 34 (3), 533–545.

Case, K.E., Quigley, J.M., Shiller, R.J., 2005. "Comparing Wealth Effects: The Stock Market versus the Housing Market," The B.E. Journal of Macroeconomics, De Gruyter 5 (1), 1–34.

Case, K.E., Mayer, C.J., 1996. Regional housing price dynamics within a metropolitan area. Sci. Urban Econ. 26, 387–407.

Case, K.E., Shiller, R.J., 1990. Forecasting prices and excess returns in the housing market. AREUEA J. 8, 253–273.

Das, S., Gupta, R., Kabundi, A., 2009. Could we have predicted the recent downturn in the South African housing market? J. Hous. Econ. 18 (4), 325–335.

Das, S., Gupta, R., Kabundi, A., 2010. The blessing of dimensionality in forecasting real house price growth in the nine census divisions of the US. J. Hous. Res. 19 (1), 89–109.

Das, S., Gupta, R., Kabundi, A., 2011. Forecasting regional house price inflation: a comparison between dynamic factor models and vector autoregressive models. J. Forecast. 30 (2), 288–302.

Doan, T.A., Litterman, R.B., Sims, C.A., 1984. Forecasting and conditional projections using realistic prior distributions. Econ. Rev. 3, 1–100.

Dua, P., Ray, S.C., 1995. A BVAR Model for the Connecticut Economy. J. Forecast. 14, 167–180.

Fama, E., 1965. The behavior of stock market prices. J. Bus. 38, 34–105.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2003. Do financial variables help forecasting inflation and real activity in the Euro area? J. Monet. Econ. 50 (6), 1243–1255.

Ghysels, E., Plazzi, A., Torous, W., Valkanov, R., 2013. Forecasting real estate prices. In: Alberto, W., E., Valkanov, R. (Eds.), In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting vol. II. Elsevier.

Gogas, P., Pragidis, I., 2011. Does the interest risk premium predict housing prices? Int. J. Econ. Res. 8 (1), 1–10.

Gogas, P., Papadimitriou, T., Matthaiou, M., Chrysanthidou, E., 2014. Yield curve and recession forecasting in a machine learning framework. Comput. Econ (forthcoming).

Gogas, P., Papadimitriou, T., Takli, E., 2013. Comparison of simple sum and Divisia monetary aggregates in GDP forecasting: a support vector machines approach. Econ. Bull. 33 (2), 1101–1115.

Guo, W., Tse, P., 2013. A novel signal compression method based on optimal ensemble empirical mode decomposition for bearing vibration signals. J. Sound Vib. 332, 423–441.

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

Gupta, R., Das, S., 2010. Predicting downturns in the US housing market: a Bayesian approach. J. Real Estate Financ. Econ. 41 (3), 294–319.

Gupta, R., Hartley, F., 2013. The role of asset prices in forecasting inflation and output in South Africa. J. Emerg. Mark. Financ. 12 (3), 239–291.

Gupta, R., Kabundi, A., 2010. Forecasting real US house price: principal components versus Bayesian regressions. Int. Bus. Econ. Res. J. 9 (7), 141–152.

Gupta, R., Kabundi, A., Miller, S.M., 2011. Forecasting the US real house price index: structural and non-structural models with and without fundamentals. Econ. Model. 28, 2013–2021.

Hannan, E.J., Quinn, B.G., 1979. The Determination of the Order of an Autoregression. J. R. Stat. Soc. B. 41, 190–195.

Härdle, Wolfgang, Lee, Yuh-Jye, Schäfer, Dorothea, Yeh, Yi-Ren, 2009. Variable selection and oversampling in the use of smooth support vector machines for predicting the default risk of companies. J. Forecast. 28 (6), 512–534.

Hodrick, R.J., Prescott, E.C., 1997. Postwar US business cycles: an empirical investigation. J. Money Credit Bank. 29, 1–16.

Jud, G.D., Winkler, D.T., 2002. The dynamics of metropolitan housing prices. J. Real Estate Res. 23 (1/2), 29–45.

Khandani, Amir E., Kim, Adlar J., Lo, Andrew W., 2010. Consumer credit-risk models via machine-learning algorithms. J. Bank. Financ. 34 (11), 2767–2787.

Leamer, E.E., 2007. Housing is the business cycle. NBER Working Paper No. 13428 (Retrieved from http://www.nber.org/papers/w13428).

Litterman, R.B., 1981. A Bayesian procedure for forecasting with vector autoregressions. Working Paper. Federal Reserve Bank of Minneapolis.

Litterman, R.B., 1986. Forecasting with Bayesian vector autoregressions — five years of experience. J. Bus. Econ. Stat. 4 (1), 25–38.

McGibany, J.M., Nourzad, F., 2004. Do lower mortgage rates mean higher housing prices? Appl. Econ. 36 (4), 305–313.

Mikhed, V., Zemčík, P., 2009. Do house prices reflect fundamentals? Aggregate and panel data evidence. J. Hous. Econ. 18, 140–149.

Moghtaderi, A., Flandrin, P., Borgnat, P., 2013. Trend filtering via empirical mode decomposition. Comput. Stat. Data Anal. 58, 114–126.

Öğüt, Hulisi, Mete Doğanay, M., Ceylan, Nildağ Başak, Aktaş, Ramazan, 2012. Prediction of bank financial strength ratings: the case of Turkey. Econ. Model. 29 (3), 632–640.

Padilla, M.A., 2005. The Effects of Oil Prices and Other Economic Indicators on Housing Prices in Calgary, Canada. Massachusetts Institute of Technology. Dept. of Architecture.

Papadimitriou, T., Gogas, P., Stathakis, E., 2014. Forecasting energy markets using support vector machines. Energy Econ. 44, 135–142.

Rapach, D.E., Strauss, J.K., 2007. Forecasting real housing price growth in the eighth district states. Federal Reserve Bank of St. Louis. Reg. Econ. Dev. 3 (2), 33–42.

Rapach, D.E., Strauss, J.K., 2009. Differences in housing price forecastability across US states. Int. J. Forecast. 25, 351–372.

Rubio, Ginés, Pomares, Héctor, Rojas, Ignacio, Herrera, Luis Javier, 2011. A heuristic method for parameter selection in LS-SVM: application to time series prediction. Int. J. Forecast. 27 (3), 725–739.

Schwarz, Gideon E., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.

Sims, C.A., 1980. Macroeconomics and reality. Econometrica 48 (1), 1–48.

Sims, C.A., Stock, J.H., Watson, M.W., 1990. Inference in linear time series models with some unit roots. Econometrica 58, 113–144.

Spencer, D.E., 1993. Developing a Bayesian vector autoregression model. Int. J. Forecast. 9 (3), 407–421.

Stevenson, S., 2000. A long term analysis of housing and inflation. J. Hous. Econ. 9 (1–2), 24–39.

Stock, J.H., Watson, M.W., 2003. Forecasting output and inflation: the role of asset prices. J. Econ. Lit. 41 (3), 788–829.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B. 58 (1), 267–288.

Vapnik, V., Boser, B., Guyon, I., 1992. A training algorithm for optimal margin classifiers. Fifth Annual Workshop on Computational Learning Theory. ACM, Pittsburgh, pp. 144–152.

Vargas, Silva, C., 2008. The effect of monetary policy on housing: a factor augmented approach. Appl. Econ. Lett. 15 (10), 749–752.

Wu, Z., Huang, N.E., Long, S., Peng, C.-K., 2007. On the trend, detrending and variability of nonlinear and nonstationary time series. Proc. Natl Acad. Sci. 4 (38), 14889–14894.

Wu, Z., Huang, N.E., 2009. Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv. Adapt. Data Anal. 1 (1), 1–41.

Zeno, Füss, 2010. Macroeconomic determinants of international housing markets. J. Hous. Econ. 19, 38–50.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320.