Forecasting Home Sales in the Four Census Regions and the Aggregate US Economy Using Singular Spectrum Analysis

Hossein Hassani*, Zara Ghodsi*, Rangan Gupta ** and Mawuli Segnon ***

* *The Statistical Research Centre, Bournemouth University, 89 Holdenhurst Road, Bournemouth BH8 8EB, UK*
** *Department of Economics, University of Pretoria Pretoria, 0002, South Africa*
*** *Christian-Albrechts-University of Kiel Leibnizstraße 3, Room 110, 24118 Kiel, Germany*

December 26, 2014

Abstract

Accurate forecasts of home sales can provide valuable information for not only, policy makers, but also financial institutions and real estate professionals. Given this, our analysis compares the ability of two different versions of Singular Spectrum Analysis (SSA) methods, namely Recurrent SSA (RSSA) and Vector SSA (VSSA), in univariate and multivariate frameworks, in forecasting seasonally unadjusted home sales for the aggregate US economy and its four census regions (Northeast, Midwest, South and West). We compare the performance of the SSA-based models with classical and Bayesian variants of the autoregressive and vector autoregressive models. Using an out-of-sample period of 1979:8-2014:6, given an in-sample period of 1973:1-1979:7, we find that the univariate VSSA is the best performing model for the aggregate US home sales, while the multivariate versions of the RSSA is the outright favorite in forecasting home sales for all the four census regions. Our results highlight the superiority of the nonparametric approach of the SSA, which in turn, allows us to handle any statistical process: linear or nonlinear, stationary or non-stationary, Gaussian or non-Gaussian.

JEL Codes: C32, R31.
Keywords: Home Sales, Forecasting, Singular Spectrum Analysis, Classical and Bayesian (Vector) Autoregressive Models.

1 Introduction

That there is a strong link between the housing market and the economic activity in the US, has been reported in a large number of recent papers (see for example Aye *et al.*, (2014) and Nyakabawo *et al.*, (forthcoming) for a detailed literature review). Leamer (2007) goes as far as suggesting that housing *is* the business cycle, with housing market activity affecting the economy at both macroeconomic and microeconomic levels. On one hand, at the macroeconomic level, with housing representing a large share of the total economy, movements in the housing sector spills over to the entire macro economy through new construction, renovations of existing

property and the volume of home sales (Dua and Smyth, 1995; Dua and Miller, 1996; and Dua *et al.*, 1999). On the other hand, at the microeconomic level, performances of financial institutions and real estate firms depend crucially on housing market activity, as the recent financial crisis would vouch for. Hence, timely and accurate forecasts of home sales can provide valuable information not only for policy makers, but also for financial institutions and real estate professionals, i.e., housing market participants.

The literature on forecasting home sales in the US, both at aggregate and regional levels, is however, only limited to four studies, namely Dua and Smyth (1995), Dua and Miller (1996), Dua *et al.* (1999), and Gupta *et al.*, (2010). While Dua and Smyth (1995) used Bayesian VAR (BVAR) models to predict home sales for the aggregate US economy, Dua and Miller (1996) used its extended versions to forecast home sales for the state of Connecticut. In their original model, Dua and Smyth (1995) considered home sales, price of homes, mortgage interest rate, real disposable income and unemployment rate. Dua and Miller (1996) extended this model by including a leading index for the Connecticut economy and showed that by doing so, one can improve the forecast performance of the benchmark model of Dua and Smyth (1995) substantially. Inspired by this result, Dua *et al.*, (1999) extended the model described in Dua and Smyth (1995) by adding six different leading indicators, namely housing permits authorized, housing starts, the US Department of Commerce's composite index of eleven leading indicators, the short- and long-leading indices developed by the Center for International Business Cycle Research (CIBCR) at Columbia University and the leading index constructed by CIBCR that focussed solely on employment related variables. They found that the benchmark BVAR model, which included home sales, price of homes, mortgage rate, real personal disposable income, unemployment rate, supplemented by the building permits authorized as the leading indicator consistently produced the most accurate forecasts. Gupta *et al.*, (2010) analyzed the ability of a model portfolio that comprises the BVAR models, the random walk (RW) model, the autoregressive (AR) model, and the classical and Bayesian vector-error correction (VEC) models in forecasting home sales for the four US census regions (Northeast, Midwest, South, West). In their analysis Gupta *et al.*, (2010) used home sales, price of homes, mortgage rate, real personal disposable income, unemployment rate and building permits authorized. They found that except for the South, the Bayesian type models outperformed all the other models in forecasting home sales at all forecasting time horizons and are also capable of predicting the peaks and declines in home sales with tremendous accuracy. In sum, it seems clear that the Bayesian type models are better equipped in forecasting home sales than their classical counterparts.

Against this backdrop, the objective of this paper is to contribute to the literature on forecasting non-seasonally adjusted home sales for the aggregate US economy, as well as, the four census regions. We propose a new modeling approach that is based on univariate and multivariate models of Singular Spectrum Analysis (SSA). We compare the forecasting performance of the SSA models with the univariate and multivariate versions of B(V)ARs, the standard benchmarks such as RW, AR and classical VAR. Our data set contains monthly home sales over the period of 1973:1 till 2014:6. The start and end points of our data set are purely driven by the home sales data available at the time of writing this paper. To determine the starting date for the out-of-sample analysis, we apply the Bai and Perron (2003) tests of multiple structural breaks to our five home sales series. The earliest possible break date happened on 1979:8 for the Midwest, so that our out-of-sample data covers the period of 1979:8-2014:6. Since, we estimate our models recursively over the out-of-sample period, we are able guard against the change in the parameter estimates of the models when producing our forecasts. Note that since home sales can provide leading information for the economy at the macro and micro levels, unlike the existing literature, we do not incorporate the information of any macroeconomic predictors in

our multivariate models. This helps us to ensure that our search for the best-suited forecasting model for home sales can be provided independent of economic fundamentals. In our case, for the multivariate models involving the four census regions, we only consider the lagged home sales of all the four regions in a specific equation relating to particular region to accommodate for possible leading information content in the home sales of the other regions over and above the information carried by the lags of the specific region under consideration. Since the aggregate home sales for the US is a sum of all the homes sold in the four census regions, we cannot include the total US home sales in our multivariate models, understandably to avoid issues of multicollinearity. Hence, for the aggregate US economy, we only consider univariate models. Further, since we do not transform the data to remove seasonality, which can be handled by the SSA approach, we decided to rule out the use of economic fundamentals as possible predictors of home sales.

At this stage, it is important to emphasize the decision to use the SSA technique in forecasting home sales, which has recently evolved as a powerful technique in the field of time series analysis (Hassani, 2007; Hassani *et al.*, 2009). We are motivated to use SSA because it is a non-parametric technique that works with arbitrary statistical processes, whether linear or non-linear, stationary or non-stationary, Gaussian or non-Gaussian. Given that the dynamics of real time series, in our case home sales, has usually gone through structural changes during the time period under consideration, one needs to make certain that the method of prediction is not sensitive to the dynamical variations. Moreover, contrary to the standard methods of time series forecasting that assume normality and stationarity of the series (though the latter is not an issue for BVAR models), SSA method is non-parametric and makes no prior assumptions about the data, with forecasts being obtained through bootstrapping. Additionally, SSA method decomposes a series into its component parts, and reconstructs the series by leaving the random (noise) component behind. Clearly then, the SSA is a much more general approach that allows us to handle issues of non-stationarity, seasonality, non-normality and non-linearity. To the best of our knowledge, this is the first paper to use SSA in forecasting home sales for the aggregate and the regional US economies. Rest of the paper is organized as follows: Section 2 discusses the econometric models, while Section 3 presents the data and the empirical results. Finally, Section 4 concludes.

2 Forecasting Models

2.1 Random Walk

We use the random walk as a benchmark, as it is a widely accepted practice that a forecasting technique which is recommended for a particular forecast should at least be more accurate than a random walk[1]. In brief, the current month's number of home sales is forecasted to be the next month's number of home sales.

2.2 Classical and Bayesian Autoregression and Vector Autoregression

The Vector Autoregressive (VAR) model, though "atheoretical" is particularly useful for forecasting purposes. Note that an unrestricted VAR model, as suggested by Sims (1980), can be written as follows:

$$y_t = C + A(L)y_t + \epsilon_t, \tag{1}$$

[1] http://robjhyndman.com/hyndsight/benchmarks/.

where: $y : (p \times 1)$ vector of variables (in our case, the home sales of the four census regions, namely the Northeast, the Midwest, the South and the West) being forecasted; $A(L) : (p \times p)$ polynomial matrix in the backshift operator L with lag length m, i.e., $A(L) = A_1L + A_2L^2 + ... + A_mL^m$; $C : (p \times 1)$ vector of constant terms, and $\varepsilon : (p \times 1)$ vector of white-noise error terms.

The VAR model uses equal lag length for all the variables of the model. One drawback of VAR models is that many parameters are needed to be estimated, some of which may be insignificant. This problem of overparameterization, resulting in multicollinearity and loss of degrees of freedom leads to inefficient estimates and large out-of-sample forecasting errors. One solution, often adapted, is simply to exclude the insignificant lags based on statistical tests. Another approach is to use near VAR, which specifies unequal number of lags for the different equations.

However, an alternative approach to overcome this overparameterization, as described in Littermann (1981), Doan *et al* (1984), Todd (1984), Littermann (1986), and Spencer (1993), is to use a Bayesian VAR (BVAR) model. Instead of eliminating longer lags, the Bayesian method imposes restrictions on these coefficients by assuming that these are more likely to be near zero than the coefficient on shorter lags. However, if there are strong effects from less important variables, the data can override this assumption. The restrictions are imposed by specifying normal prior distributions with zero means and small standard deviations for all coefficients with the standard deviation decreasing as the lags increase. The exception to this is, however, the coefficient on the first own lag of a variable, which has a mean of unity. Note that Littermann (1981) used a diffuse prior for the constant. This is popularly referred to as the "Minnesota prior" due to its development at the University of Minnesota and the Federal Reserve Bank at Minneapolis.

The standard deviation of the distribution of the prior for lag m of variable j in equation i for all i, j and m, defined as $S(i, j, m)$, can be specified as follows:

$$S(i, j, m) = [w \times g(m) \times f(i, j)]\frac{\sigma_i}{\sigma_j}, \tag{2}$$

where $f(i, j) = 1$, if $i = j$ and k_{ij} otherwise, with $(0 \leq k_{ij} < 1)$, and $g(m) = m^{-d}$ with $d > 0$. Note that σ_i is the standard error of the univariate autoregression for variable i. The ratio $\frac{\sigma_i}{\sigma_j}$ scales the variables so as to account for differences in the units of measurement and, hence, causes specification of the prior without consideration of the magnitudes of the variables. The term w indicates the overall tightness and is also the standard deviation on the first own lag, with the prior getting tighter as we reduce the value of w. The parameter $g(m)$ measures the tightness on lag m with respect to lag 1, and is assumed to have a harmonic shape with a decay factor of d, increasing which tightens the prior on increasing lags. The parameter $f(i, j)$ represents the tightness of variable j in equation i relative to variable i, and by increasing the interaction, i.e., the value of k_{ij}, we can loosen the prior. Following the extant literature on BVAR models, we look at the following combinations of w and d: (0.3, 0.5), (0.2, 1.0), (0.1, 1.0), (0.2, 2.0) and (0.1, 2.0), with k_{ij} set at 0.5. These models are, respectively named, BVAR1, BVAR2, BVAR3, BVAR4 and BVAR5. Univariate versions of the BVAR models, which we call Bayesian autoregressive (BAR) models, are estimated for the same values of w and d as above, but with k_{ij} set at 0.001, since a small interaction value basically reduces the multivariate model to its corresponding univariate version. These models, in turn, are named as BAR1, BAR2, BAR3, BAR4 and BAR5, respectively. Note that for the overall US economy, we only consider the univariate models in order to avoid the issues of perfect multicollinearity that will

arise in multivariate models if one puts together the aggregate US home sales and home sales for one of the four census regions.

The BVAR model is estimated using Theil's (1971) mixed estimation technique, which involves supplementing the data with prior information on the distribution of the coefficients. In an artificial way, the number of observations and degrees of freedom are increased by one, for each restriction imposed on the parameter estimates. The loss of degrees of freedom due to overparameterization associated with a VAR model is, therefore, not a concern in the BVAR model. Further, note that one major advantage of the BVAR and BAR models is that we can use non-stationary data for their estimation. Sims *et al.* (1990) indicate that with the Bayesian approach entirely based on the likelihood function, the associated inferences do not require special treatment for non-stationarity, since the likelihood function exhibits the same Gaussian shape regardless of the presence of non-stationarity.

2.3 Singular Spectrum Analysis (SSA)

The principal objective of the SSA consists in decomposing original series into a sum of small number of noise and interpretable components (trend, periodic or quasi-periodic components) which in turn should be reconstructed for forecasting purposes (see Hassani *et al.*, 2009). In the following we briefly describe the main steps of the basic SSA algorithm developed by Golyandina *et al.* (2001) and its two versions we use in our forecasting exercises (see Hassani, 2007).

2.3.1 Basic SSA

As proposed in Golyandina *et al.* (2001) the two complementary stages of the SSA, i.e., the decomposition and reconstruction stages can be carried out in four steps.

Step 1: Embedding
Embedding can be considered as a mapping that transfers a one-dimensional time series $Y_N = (y_1, \ldots, y_N)$ into the multi-dimensional series $X_1, \ldots, X_K$ with vectors $X_i = (y_i, \ldots, y_{i+L-1})^T \in \mathbf{R}^L$, where L, $(2 \leq L \leq N-1)$ is the window length and $K = N - L + 1$. This first step provides the trajectory matrix $\mathbf{X} = [X_1, \ldots, X_K] = (x_{ij})_{i,j=1}^{L,K}$.

Step 2: Singular Value Decomposition (SVD)
In this step we perform the SVD of $\mathbf{X}$ into a sum of rank-one bi-orthogonal elementary matrices, $\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_L$, with $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$ and $i = 1, \ldots, L$. $\lambda_1, \ldots, \lambda_L$ are the eigenvalues of $\mathbf{X}\mathbf{X}^T$ $(\lambda_1 \geq \ldots \geq \lambda_L \geq 0)$, $U_1 \ldots U_L$ the corresponding eigenvectors and $V_1 \ldots V_L$ are the principal components defined as $V_i = X^T U_i / \sqrt{\lambda_i}$.

Step 3: Grouping
The grouping step consists in splitting the elementary matrices in *Step 2* into m disjoint groups and summing the matrices within each group to obtain new matrices, $\mathbf{W}_1, \ldots, \mathbf{W}_m$, so that the trajectory matrix $\mathbf{X}$ can be rewritten as: $\mathbf{X} = \mathbf{W}_1 + \ldots + \mathbf{W}_m$ (see Hassani, 2007 for technical details.).

Step 4: Diagonal averaging
The purpose of diagonal averaging is to transform each new matrix $(\mathbf{W}_{i(i=1,\ldots,m)})$ to the form of a Hankel matrix, which can be subsequently converted to a time series.

2.3.2 Recurrent SSA

With the reconstructed series, $\widetilde{y}_i$ $(i = 1, \ldots, N)$, the recurrent SSA forecasts can be easily obtained as follows:

Let denote π_i the last component of the eigenvector U_i $(i = 1, \ldots, r)$ and $v^2 = \pi_1^2 + \ldots + \pi_r^2 < 1$. Moreover suppose for any vector $U \in \mathbf{R}^L$ and denote by $U^\nabla \in \mathbf{R}^{L-1}$ the vector consisting of the first $L-1$ components of the vector U. Then, the h-step ahead forecasts are given by

$$y_i = \begin{cases} \widetilde{y}_i & \text{for } i = 1, \ldots, N \\ \sum_{j=1}^{L-1} \alpha_j y_{i-j} & \text{for } i = N+1, \ldots, N+h \end{cases} \tag{3}$$

where α_j's are wrapped in a vector $A = (\alpha_1, \ldots, \alpha_{L-1})$ that can be computed by $A = (1 - v^2)^{-1} \sum_{i=1}^{r} \pi_i U_i^\nabla$.

2.3.3 Vector SSA

Consider a matrix, Π, that is given by

$$\Pi = V^\nabla (V^\nabla)^T + (1 - v^2) A A^T,$$

where $V^\nabla = [U_1^\nabla, ..., U_r^\nabla]$, v^2 and A are as defined above. Defining a linear operator $\theta^{(v)} : \mathfrak{L}_r \mapsto \mathbf{R}^L$ by the following formula

$$\theta^{(v)} U = \begin{pmatrix} \Pi U^\nabla \\ A^T U^\nabla \end{pmatrix}.$$

The h-step ahead forecasts, $y_{N+1}, ..., y_{N+h}$, can be computed in two steps:

In the first step we define a vector Z_i as follows:

$$Z_i = \begin{cases} \widetilde{X}_i & \text{for } i = 1, \ldots, K \\ \theta^{(v)} Z_{i-1} & \text{for } i = K+1, \ldots, K+h+L-1 \end{cases} \tag{4}$$

where, $\widetilde{X}_i$'s are the reconstructed columns of trajectory matrix.

In the second step we construct the matrix Z, as $\mathbf{Z} = [Z_1, ..., Z_{K+h+L-1}]$, and make its diagonal averaging to obtain a series $y_1, ..., y_{N+h+L-1}$ that contain our VSSA forecasts.

3 Data Description and Empirical Results

3.1 Data

Following the literature on home sales forecasting, we use data on new, single-family houses for the US census regions (Northeast, Midwest, South and West) and the overall US economy (which is the sum of home sales in the four census regions), with the data derived from the US Census Bureau. The data sample covers the monthly period of 1973:1-2014:6, and is purely driven by availability of data at the time of writing this paper.[2] We use the seasonally unadjusted data to retain strong seasonal characteristics in home sales behavior, and hence model the data in its

[2]It must be pointed out though, that the data on aggregate US home sales is available from 1963:1, but we start from 1973:1 for the sake of keeping our results comparable across models over the same in-sample and out-of-sample periods.

original form, rather than the seasonally adjusted data. The data is transformed to its natural logarithmic levels. Figure 1 plots the data for the overall US economy and the four census regions, and clearly depicts the seasonal patterns.

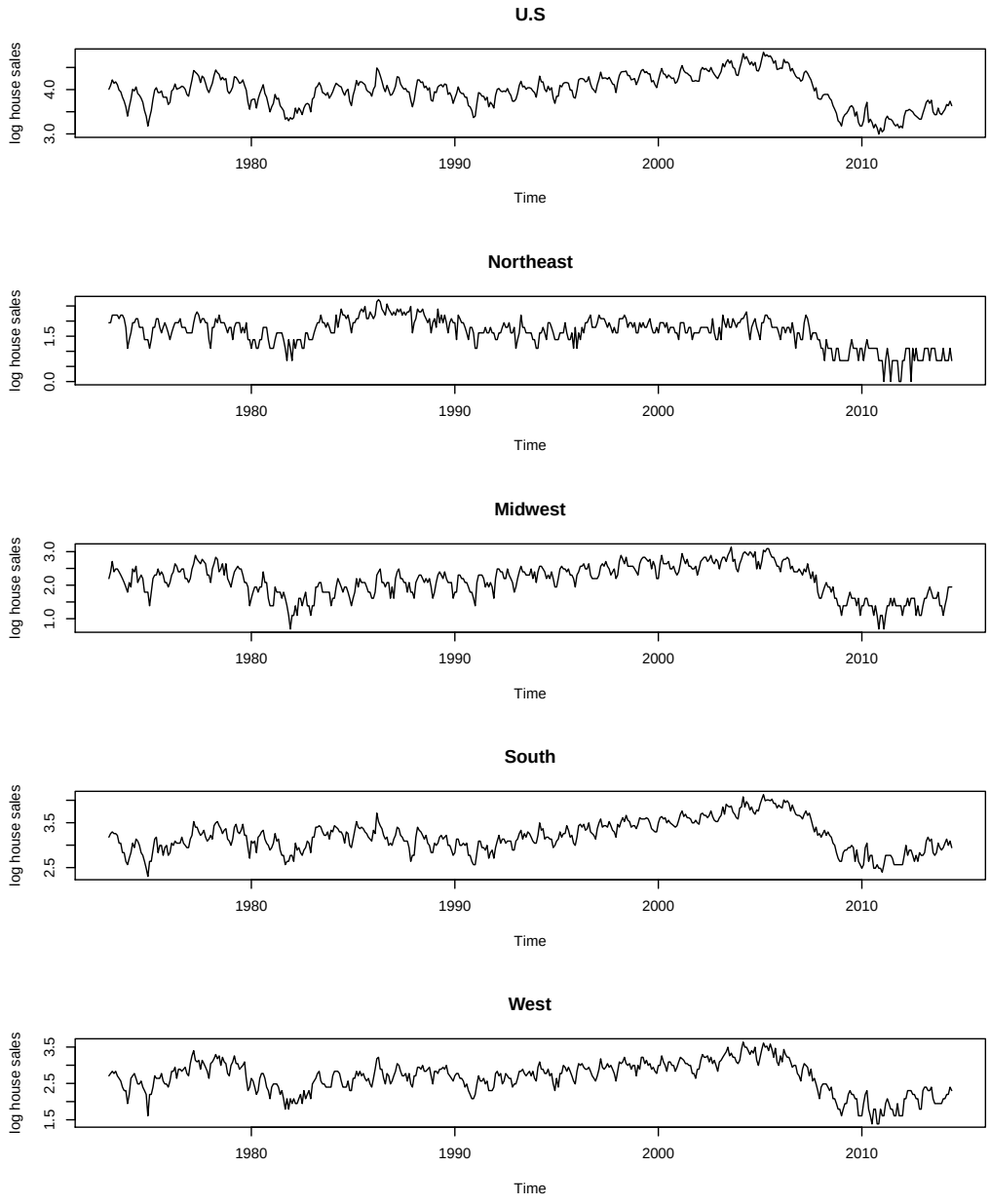

Figure 1: Natural logarithms of home sales for the US and the four census regions.

3.2 Empirical Results

In this paper, Root Mean Squared Error (RMSE) is used to evaluate the forecasting performance of our model portfolio. Note that here, we prefer the Ratio Root Mean Squared Error that displays the RRMSE for our model portfolio relative to the RRMSE of a random walk (RW) model. Formally, the Ratio Root Mean Squared Error (RRMSE) criterion is defined as

$$\text{RRMSE} = \frac{\text{Alternate Model}}{\text{RW Model}} = \frac{\left(\sum_{i=1}^{n}(\widehat{y}_{N+h,i} - y_{N+h,i})^2\right)^{1/2}}{\left(\sum_{i=1}^{n}(\widetilde{y}_{N+h,i} - y_{N+h,i})^2\right)^{1/2}}, \tag{5}$$

where $\widehat{y}_{N+h}$ is the h-step ahead forecast obtained via an alternate model, $\widetilde{y}_{N+h}$ is the h-step ahead forecast from the RW model, y_t is the actual values and n is the number of the forecasts. If RRMSE < 1, then the alternate model outperforms the RW method by $\left(1 - \frac{\text{Alternate Model}}{\text{RW Model}}\right)$ percent.

Besides the benchmark random-walk (RW) model, the models evaluated in this study include classical and Bayesian autoregressive and vector autoregressive models, as well as, two different version of Singular Spectrum Analysis (SSA) forecasting methods, namely, Recurrent SSA (RSSA) and Vector SSA, both in univariate (UV) and multivariate (MV) set-ups. We select the period of 1973:1-1979:7 as in-sample for the model training and the rest of the data (1979:8-2014:6) as out-of-sample for evaluating the forecasting performance, given that the earliest structural break amongst the five series occurred in 1979:8 for the AR model, as well as the VAR model, of the Midwest region based on the Bai and Perron (2003) test of multiple structural breaks. All models are estimated based on 2 lags, suggested by the Schwarz Information Criterion (SIC) for the in-sample exercise. We forecast the home sales for one-month-ahead horizon till twelve-months-ahead horizon, and thus covering short, medium and long-term forecasting horizons. Following the extant literature on forecasting with BVARs (see Gupta *et al.*, 2014 for a further details in this regard), we estimate five variants of each the BAR and BVAR models. We only report the so-called "optimal" BVAR models, i.e., the models which produce the lowest average RRMSEs over the 12 forecasting horizons considered.

We start off the discussion with the case of the overall US home sales. Recall, due to issues of multicollinearity, we only consider univariate models for the aggregate US home sales. As can be seen from Table 1, on average, only the UVRSSA and UVVSSA models outperform the RW model, with the UVVSSA being the best model, based on the average RRMSE. Though the tightly preferred BAR5 outperforms the AR model, but both have average RRMSEs greater than unity, even though the BAR5 model beats the RW model at lower horizons ($h \leq 6$). However, for longer horizons, $h > 6$, the results show that the BAR5 cannot improve upon the RW model. The UVRSSA and UVVSSA consistently outperform the RW model at all horizons, except at the one-year-ahead horizon. For both models the RRMSEs keep declining until the five-months-ahead horizon, and, then start increasing until the twelve-months-ahead horizon.

Table 1: Out-of-sample RRMSE results for total U.S. House sales.

h	RW	AR	BAR5	UVRSSA	UVVSSA
1	0.115	1.000	0.993	0.922	0.922***
2	0.170	0.993	0.982	0.820	0.814***
3	0.212	0.998	0.970	0.728	0.737***
4	0.239	1.011	0.966	0.675	0.696***
5	0.249	1.036	0.976	0.666	0.666***
6	0.258	1.065	0.989	0.722	0.668***
7	0.258	1.111	1.017	0.729	0.710***
8	0.256	1.161	1.051	0.773	0.738***
9	0.245	1.248	1.114	0.815	0.791***
10	0.226	1.386	1.220	0.894	0.872***
11	0.210	1.530	1.340	0.972	0.967***
12	0.204	1.613	1.419	1.053	1.053
Average	0.220	1.182	1.085	0.803	0.790

Notes: *** indicates significance of the $MSE - F$ statistic at 1% level.

Next, we turn our attention to the four census regions and discuss them individually. Note that for the four census regions, we consider both univariate and multivariate versions of our model portfolio. For the Northeast region and based on average RRMSEs, the best model seems to be the MVRSSA followed by the MVVSSA and the UVVSSA. For all the other models, namely the AR, BAR1, VAR, BVAR3 and the UVRSSA the average RRMSEs is greater than one. The MVRSSA outperforms the RW model for all horizons. Similar results can be obtained for the MVVSSA, except for the one-month-ahead horizon. One important point to note here is that, the multivariate models outperform their corresponding univariate ones consistently - a result which tends to suggest the importance of lagged home sales information of the other regions in the multivariate models. This result is not surprising given the fact that agents can move freely across the regions within the country. For the other three regions (Midwest, South and West) the MVRSSA outperforms the RW model at all forecasting horizons, barring the twelve-months-ahead horizon. As is the case for the Northeast region, the forecasting performance of the MVRSSA is followed by that of the MVVSSA, the UVVSSA and the UVRSSA. However, we note that UVRSSA produces relatively low RMSEs compared to the RW model at all horizons and for Midwest, South and West regions. The UVVSSA performs poorly relative to the RW model for lower horizons, $h \leq 5$ for the Midwest, South and West and $h \leq 3$ for the Northeast. Though the classical and Bayesian variants of the AR and VAR models perform poorly relative to the RW model, we observe that the Bayesian variants outperform their classical counterparts based on the average RRMSEs. Also the multivariate versions of the AR and VAR, classical or Bayesian do better than their corresponding univariate ones. However, the forecasting performance of the optimal BAR and BVAR models differs widely in the four regions. While the BAR1 and BVAR3 seem to be the best models to properly fit home sales data for the Northeast, BAR3 and BVAR5 were found optimal for the Midwest and West and BAR5 and BVAR5 for the South census region.

Overall, the superiority of the SSA models, especially the multivariate ones, and in particular the MVRSSA is clearly evident for the four census regions. For the aggregate US economy, as is the case for univariate models for the four census regions, the UVVSSA is the best performing model. One thing to note is that the poor performance of the B(V)AR models relative to the RW in this paper is possibly due to the fact that we do not use macroeconomic and financial variables as predictors in our models, and concentrate solely on information carried by the autoregressive components of the home sales.

Table 2: Out-of-sample Uni-/Multivariate RRMSE results for House sales in Northeast Region.

h	RW	AR	BAR1	UVRSSA	UVVSSA	VAR	BVAR3	MVRSSA	MVVSSA
1	0.280	0.945	0.946	1.008	1.051	0.883	0.915	0.892***	1.120
2	0.322	0.956	0.957	1.005	1.038	0.908	0.926	0.829***	0.841
3	0.350	0.967	0.969	1.008	1.020	0.938	0.950	0.797***	0.822
4	0.355	1.008	1.010	1.008	0.974	0.997	1.006	0.816***	0.827
5	0.366	1.034	1.036	1.010	0.946	1.027	1.038	0.823***	0.818
6	0.360	1.092	1.094	1.010	0.894	1.097	1.103	0.842***	0.842
7	0.376	1.099	1.102	1.008	0.888	1.101	1.098	0.820***	0.820
8	0.387	1.108	1.111	1.008	0.881	1.118	1.111	0.806***	0.806
9	0.375	1.170	1.174	1.008	0.834	1.181	1.169	0.841***	0.838
10	0.370	1.209	1.213	1.007	0.807	1.221	1.205	0.861***	0.861
11	0.363	1.252	1.257	1.008	0.778	1.265	1.245	0.895***	0.895
12	0.355	1.305	1.310	1.007	0.748	1.316	1.291	0.955***	0.955
Average	0.355	1.100	1.103	1.008	0.890	1.094	1.094	0.848	0.871

Notes:*** indicates significance of the $MSE - F$ statistic at 1% level.

Table 3: Out-of-sample Uni-/Multivariate RRMSE results for House sales in Midwest Region.

h	RW	AR	BAR3	UVRSSA	UVVSSA	VAR	BVAR5	MVRSSA	MVVSSA
1	0.210	0.986	0.985	0.999	1.013	0.854	0.912	0.804***	1.193
2	0.273	0.986	0.981	0.996	1.015	0.897	0.903	0.692***	0.696
3	0.330	0.972	0.969	0.996	1.028	0.909	0.902	0.628***	0.640
4	0.359	0.975	0.971	0.995	1.024	0.937	0.917	0.608***	0.636
5	0.375	0.989	0.985	0.994	1.008	0.960	0.936	0.598***	0.798
6	0.376	1.017	1.012	0.993	0.980	0.998	0.975	0.604***	0.612
7	0.377	1.046	1.041	0.992	0.952	1.025	1.006	0.615***	0.634
8	0.369	1.094	1.089	0.992	0.909	1.083	1.062	0.654***	0.662
9	0.356	1.157	1.152	0.991	0.858	1.155	1.130	0.699***	0.705
10	0.317	1.311	1.303	0.988	0.757	1.324	1.293	0.788***	0.788
11	0.279	1.516	1.506	0.987	0.654	1.533	1.490	0.899***	0.896
12	0.263	1.652	1.641	0.987	0.600	1.663	1.612	1.009	1.009
Average	0.324	1.129	1.123	0.992	0.882	1.101	1.083	0.716	0.772

Notes: *** indicates significance of the $MSE - F$ statistic at 1% level.

Table 4: Out-of-sample Uni-/Multivariate RRMSE results for House sales in South Region.

h	RW	AR	BAR5	UVRSSA	UVVSSA	VAR	BVAR5	MVRSSA	MVVSSA
1	0.131	0.993	0.991	0.999	1.009	0.985	0.978	0.888***	1.455
2	0.173	0.987	0.986	0.998	1.012	1.005	0.973	0.791***	0.779
3	0.209	0.976	0.976	0.999	1.023	1.041	0.976	0.723***	0.713
4	0.233	0.974	0.974	0.998	1.025	1.051	0.988	0.687***	0.678
5	0.243	0.992	0.991	0.997	1.005	1.077	1.012	0.679***	0.666
6	0.251	1.011	1.010	0.996	0.986	1.104	1.041	0.690***	0.670
7	0.251	1.048	1.045	0.994	0.951	1.151	1.086	0.709***	0.709
8	0.249	1.088	1.085	0.993	0.915	1.198	1.137	0.747***	0.747
9	0.239	1.162	1.156	0.991	0.857	1.288	1.222	0.803***	0.795
10	0.221	1.283	1.274	0.990	0.777	1.427	1.355	0.890***	0.858
11	0.210	1.387	1.378	0.989	0.718	1.540	1.470	0.956***	0.952
12	0.204	1.470	1.460	0.990	0.678	1.620	1.552	1.037	1.037
Average	0.218	1.114	1.111	0.994	0.895	1.212	1.152	0.800	0.838

Notes:*** indicates significance of the $MSE - F$ statistic at 1% level.

Table 5: Out-of-sample Uni-/Multivariate RRMSE results for House sales in West Region.

h	RW	AR	BAR3	UVRSSA	UVVSSA	VAR	BVAR5	MVRSSA	MVVSSA
1	0.172	0.991	0.989	0.999	1.011	1.002	0.988	0.994***	1.506
2	0.232	0.981	0.980	0.999	1.019	1.015	0.982	0.868***	0.855
3	0.274	0.971	0.971	0.999	1.028	1.021	0.983	0.788***	0.791
4	0.302	0.970	0.970	0.999	1.028	1.029	0.991	0.752***	0.769
5	0.309	0.993	0.991	0.996	1.003	1.061	1.020	0.758***	0.761
6	0.318	1.010	1.008	0.995	0.985	1.086	1.043	0.770***	0.764
7	0.323	1.034	1.032	0.995	0.961	1.115	1.072	0.789***	0.783
8	0.322	1.067	1.066	0.994	0.930	1.153	1.112	0.804***	0.804
9	0.315	1.116	1.116	0.993	0.887	1.210	1.168	0.815***	0.822
10	0.293	1.213	1.212	0.992	0.815	1.323	1.275	0.879***	0.876
11	0.275	1.320	1.318	0.991	0.749	1.441	1.389	0.956***	0.952
12	0.275	1.363	1.361	0.991	0.725	1.482	1.428	1.004	1.004
Average	0.284	1.087	1.086	0.995	0.914	1.166	1.124	0.848	0.890

Notes:*** indicates significance of the MSE $- F$ statistic at 1% level.

To formally test whether forecasts from a specific model (UVVSSA in case of the US and MVRSSA for the four census regions) are significantly more accurate than the RW model forecasts, we use the McCracken (2007) MSE $- F$ statistic – designed for nested models, given that the UVVSSA or MVRSSA models nests the RW model. The MSE $- F$ statistic tests the null

hypothesis that the UVVSSA or the MVRSSA forecast mean squared error (MSE) equals the RW model forecast MSE against the one-sided (upper-tail) alternative that the MSE of the UVVSSA or the MVRSSA falls below the MSE of the RW model. The MSE $-$ F statistic is based on the loss differential, and is given as:

$$\text{MSE} - F = (N - R - h + 1)\frac{\bar{d}}{\text{MSE}_1} \tag{6}$$

where N is the number of observations in the total sample, R denotes number of observations used to estimate the model from which we calculate the first forecast (i.e., the in-sample portion of N), h is the forecast horizon, $\hat{\text{MSE}}_i = (N - R - h + 1)^{-1} \sum_{t=R}^{N-h} (u_{i,t+1})^2$ with $i = 1, 0$, $\bar{d}$ $= \hat{\text{MSE}}_0 - \hat{\text{MSE}}_1$, $\hat{\text{MSE}}_1$ corresponds to the MSE of the unrestricted model (i.e., the UVVSSA or MVRSSA), and ($\hat{\text{MSE}}_0$) corresponds to the MSE of the restricted model (i.e., the RW-benchmark model). A significant $MSE - F$ statistic indicates that the unrestricted model forecasts are statistically more accurate than those of the restricted model. The results show that the MSE $-$ F test statistics are significant at the 1% level for all regions and for all time horizons. The strong rejection of the null hypothesis of the MSE $-$ F statistic confirms the superiority of the UVVSSA and MVRSSA over the RW model in terms of forecasting seasonally unadjusted home sales data.

4 Conclusions

Housing market activity in the US has been shown to affect the economy at both macroeconomic and microeconomic levels. Hence, timely and accurate forecasts of home sales can provide valuable information not only, for policy makers, but also for housing market participants (financial institutions and real estate professionals). Against this backdrop, we made use of the Recurrent and Vector Singular Spectrum Analysis methods, both in uni- and multivariate frameworks, to forecasting seasonally unadjusted home sales for the aggregate US and for the four census regions (Northeast, Midwest, South and West). We have compared the forecasting performance of the SSA-based models with Classical and Bayesian variants of the autoregressive and vector autoregressive models. Since the aggregate home sales for the US is a sum of all the homes sold in the four census regions, we cannot include the total US economy in our multivariate models, understandably to avoid issues of multicollinearity, we only look at univariate models when forecasting home sales for the overall US economy. Using an out-of-sample period of 1979:8-2014:6, given an in-sample period of 1973:1-1979:7, we found that the univariate VSSA is the best performing model for the overall US home sales, while the multivariate version of the RSSA is the outright favorite in forecasting home sales for all the four census regions. Our results highlight the superiority of the nonparametric approach of the SSA, which in turn, allows us to handle any statistical process: linear or nonlinear, stationary or non-stationary, Gaussian or non-Gaussian.

References

Aye, G. C., Balcilar, M., Bosh, A, and Gupta, R. (2014). Housing and the Business Cycle in South Africa. *Journal of Policy Modeling*, **36 (3)**, pp. 471–491.

Bai J., and Perron P. (2003). Computation and Analysis of Multiple Structural Change Models. *Journal of Applied Econometrics*, **18 (1)**, pp. 1–22.

Doan, T. A., Litterman, R. B. and Sims, C. A. (1984). Forecasting and Conditional Projections Using Realistic Prior Distributions. *Econometric Reviews*, **3 (1)**, pp. 1-100.

Dua, P. and Smyth, D.J. (1995). Forecasting U.S. Home Sales Using BVAR Models and Survey Data on Households' Buying Attitudes for Homes. *Journal of Forecasting*, **14 (3)**, pp. 167–180.

Dua, P. and Miller, S. M. (1996). Forecasting Connecticut Home Sales in a BVAR Framework Using Coincident and Leading Indexes. *Journal of Real Estate Finance and Economics*, **13 (3)**, pp. 219–235.

Dua, P., Miller, S. M. and Smyth, D. J. (1999). Using Leading Indicators to Forecast US Home Sales in Bayesian VAR Framework. *Journal of Real Estate Finance and Economics*, **18 (2)**, pp. 191–205.

Golyandina, N., Nekrutkin, V. and Zhigljavsky, S. (2001). Analysis of Time Series Structure: SSA and Related Techniques. Chapman and Hall/CRC.

Gupta, R., Tipoy, C. K. and Das, S. (2010). Could We Have Predicted the Recent Downturn in Home Sales of the Four US Census Regions? *Journal of Housing Research*, **19 (2)**, pp. 111–128.

Gupta, R., Kabundi, A., Miller, S. M. and Uwilingiye, J. (2014). Using Large Data Sets to Forecast Sectoral Employment. *Statistical Methods and Applications*, **23 (2)**, pp. 229–264.

Hassani, H. (2007). Singular Spectrum Analysis: Methodology and Comparison. *Journal of Data Science*, **5**(2), pp. 239-257.

Hassani, H., Heravi, H., and Zhigljavsky, A. (2009). Forecasting European Industrial Production with Singular Spectrum Analysis. *International Journal of Forecasting*, **25 (1)**, pp. 103–118.

Leamer, E. E. (2007). Housing is the business cycle. *Working Paper 13428, National Bureau of Economic Research.*

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. *Journal of Business and Economic Statistics*, **4 (1)**, pp. 25–38.

McCracken, M. W. (2007). Asymptotics for Out-of-Sample Tests of Granger Causality. *Journal of Econometrics*, **140 (2)**, pp. 719–752.

Nyakabawo, W. V., Miller, S. M., Balcilar, M., Das, S. and Gupta, R. (forthcoming). Temporal Causality between House Prices and Output in the U.S.: A Bootstrap Rolling-window Approach. *North American Journal of Economics and Finance.*

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, **48:1**, pp. 1–48.

Spencer, D. E. (1993). Developing a Bayesian Vector Autoregression Model. *International Journal of Forecasting*, **9 (3)**, pp. 407–421.

Theil, H. (1971). Principles of Econometrics. New York: John Wiley.

Todd, R. M. (1984). Improving Economic Forecasting with Bayesian Vector Autoregression. *Quarterly Review*, Federal Reserve Bank of Minneapolis, **Fall**, pp. 18–29.