Time Series Analysis of U.S. Housing Price Index (1991 – 2013)

Suryansh Mishra

Dataset Description & Objective

- **Dataset:**
  - Monthly U.S. Housing Price Index (1/1/1991 – 4/1/2013)
- **Variables:**
  - HPI
  - Number of houses sold (in thousands)
  - Researched federal funds rate and added it to our dataset
  - Create events variable to account for drop in 2007 (before is 0, after is 1)
- **Goals:**
  - Decompose HPI
  - Model and accurately forecast its behavior
  - Explore relationships with external variables

Original and Modified Datasets

date	hpi	numsold (k)
1/1/1991	100	30
2/1/1991	100.48	40
3/1/1991	100.74	51
4/1/1991	100.75	50
5/1/1991	100.92	47
6/1/1991	101.4	47
7/1/1991	101.36	43
8/1/1991	101.31	46
9/1/1991	101.41	37
10/1/1991	101.62	41
11/1/1991	102.16	39
12/1/1991	102.21	36

Original dataset had date, HPI (target variable, and numsold (exogenous variable)

Any missing values for date were imputed using monthly pattern

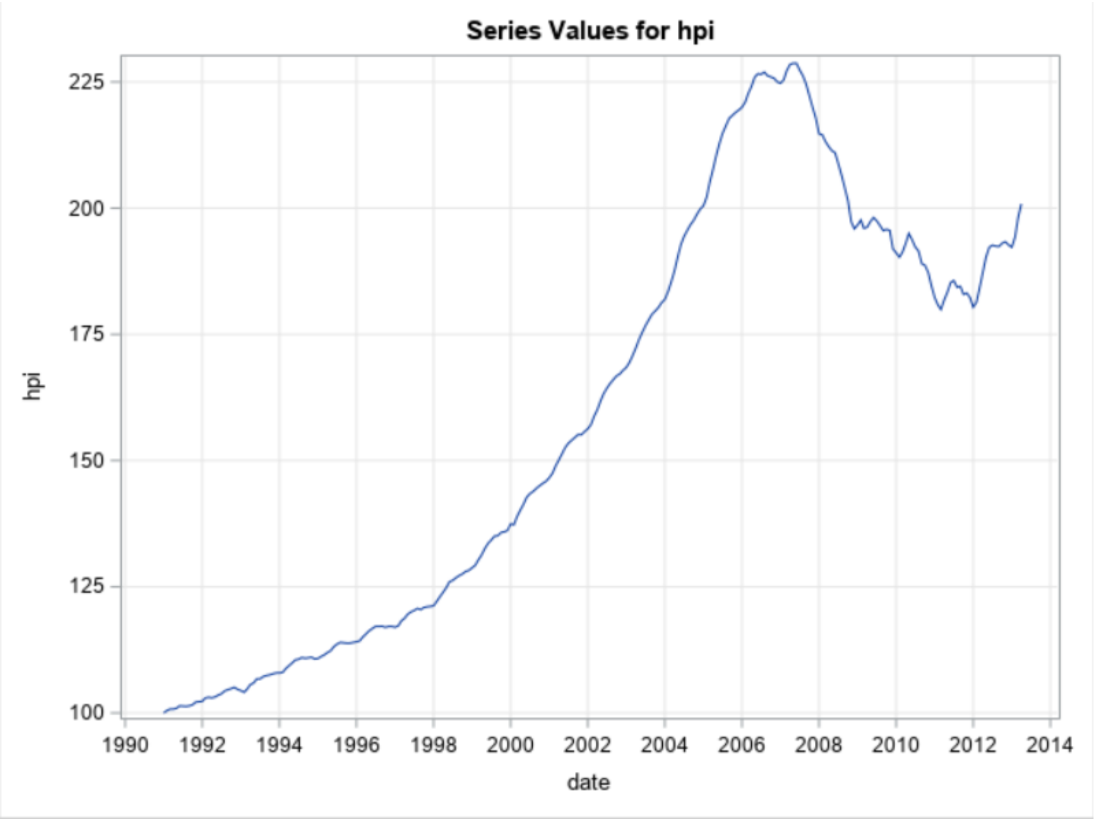
date	hpi	numsold (k)	fed funds rate	events
1/1/1991	100	30	6.91	0
2/1/1991	100.48	40	6.25	0
3/1/1991	100.74	51	6.12	0
4/1/1991	100.75	50	5.91	0
5/1/1991	100.92	47	5.78	0
6/1/1991	101.4	47	5.9	0
7/1/1991	101.36	43	5.82	0
8/1/1991	101.31	46	5.66	0
9/1/1991	101.41	37	5.45	0
10/1/1991	101.62	41	5.21	0
11/1/1991	102.16	39	4.81	0
12/1/1991	102.21	36	4.43	0

Federal funds rate was researched for each month in our dataset and added as another exogenous variable

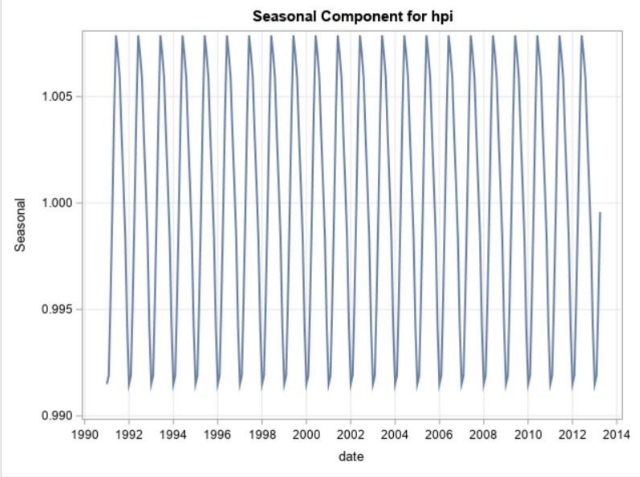
Events variable had a 0 before 2007 and 1 after 2007 to account for the sudden drop in HPI after the recession

Raw Series Visualization

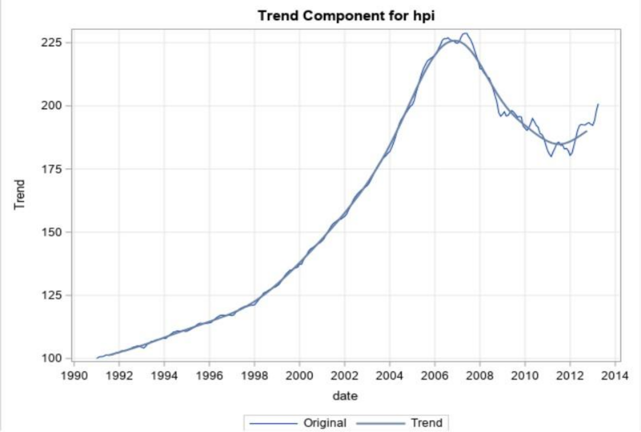
- Steady rise until 2007
- Sharp decline around the 2008 financial crisis
- Partial recovery post-2012
- Event appears to have been a step post-2007



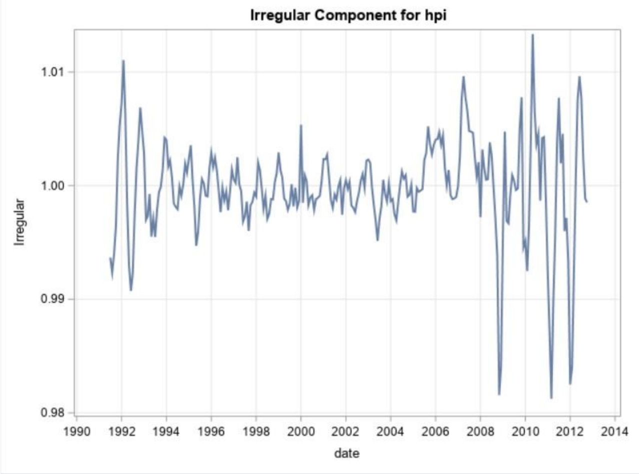
Time Series Decomposition



Annual cyclical patterns (peaks/troughs)



Long-term growth and crash

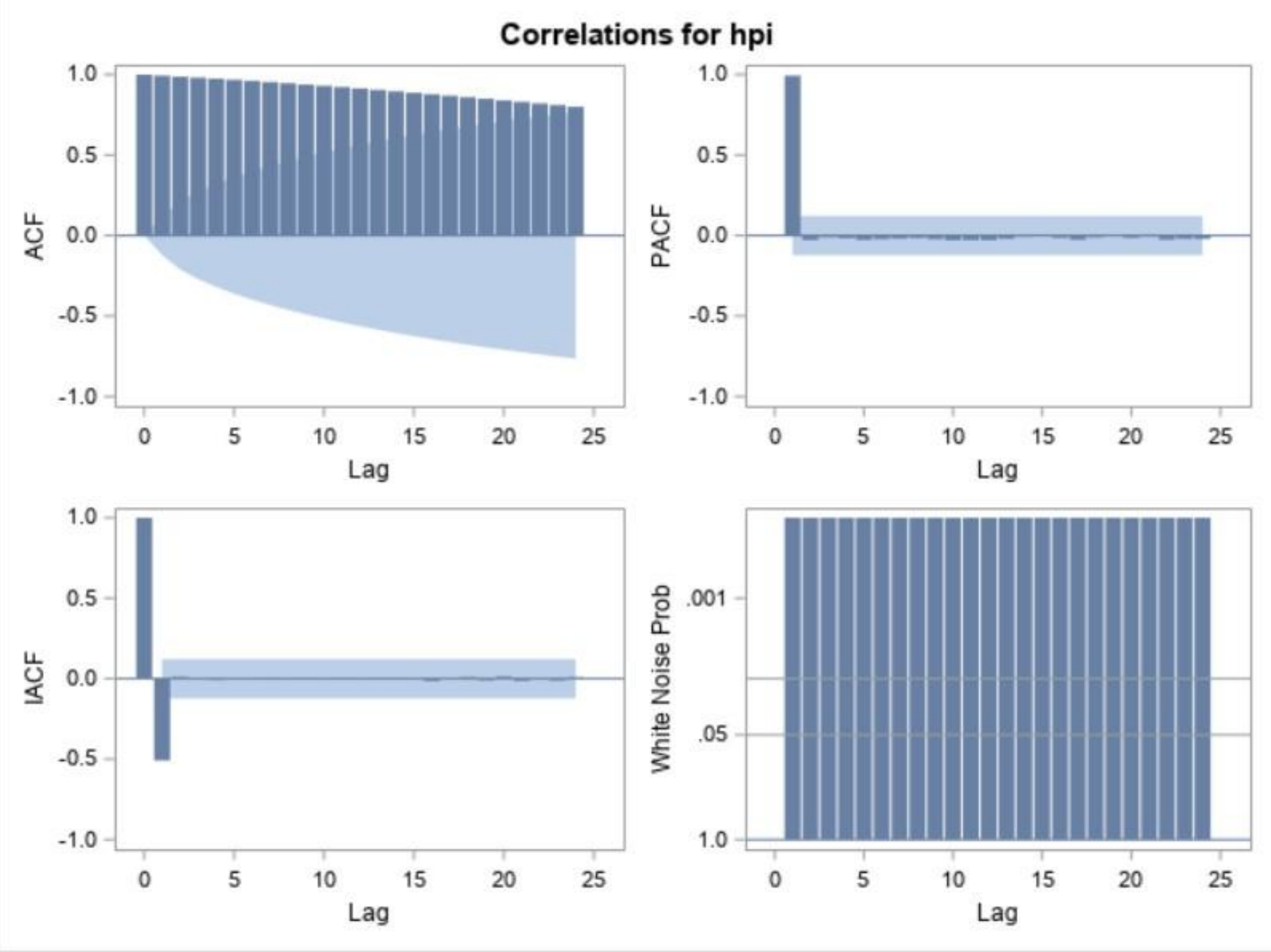


Volatility spikes post-2007

Autocorrelation Diagnostics

Gradual decay – strong persistence and trend

Significant spike at lag 1 – possibility of AR(1)



Significant spike at lag 1 – possibility of AR(1)

Series NOT distributed as white noise – definitely some modeling to be done

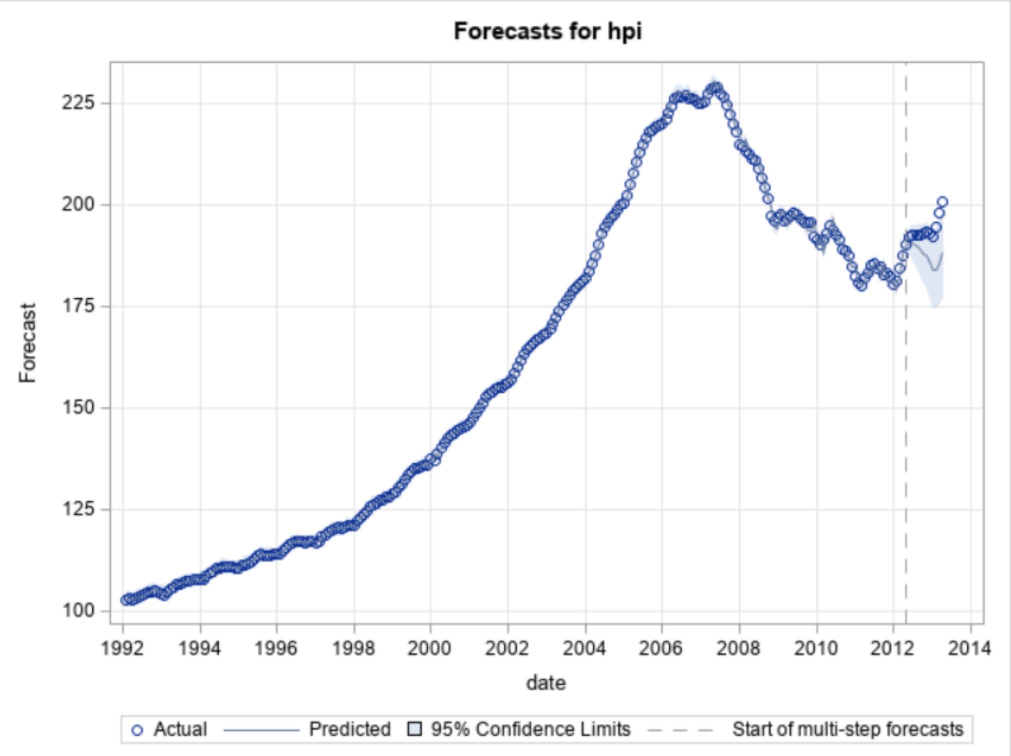
SARIMA(1,1,0)(1,1,0) Model

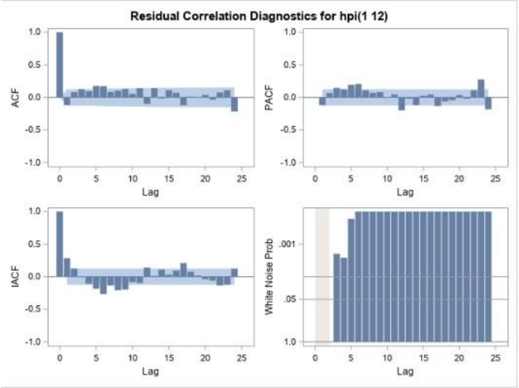
- Model successfully captures both seasonal and non-seasonal first-order autoregression.
- However, residuals still resemble a pattern, indicating that our model needs to be refined.
- The MAPE for this model is 3.06% and the RMSE is 7.2
  - This makes it the least accurate of all models we considered by far

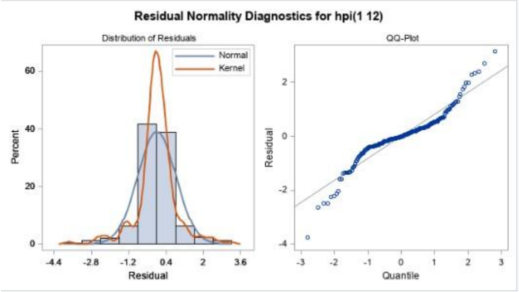
Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	0.03597	0.08056	0.45	0.6552	0
AR1,1	0.53909	0.05345	10.09	<.0001	1
AR2,1	-0.39415	0.06023	-6.54	<.0001	12

Constant Estimate	0.023113
Variance Estimate	0.67789
Std Error Estimate	0.823341
AIC	629.8732
SBC	640.497
Number of Residuals	255







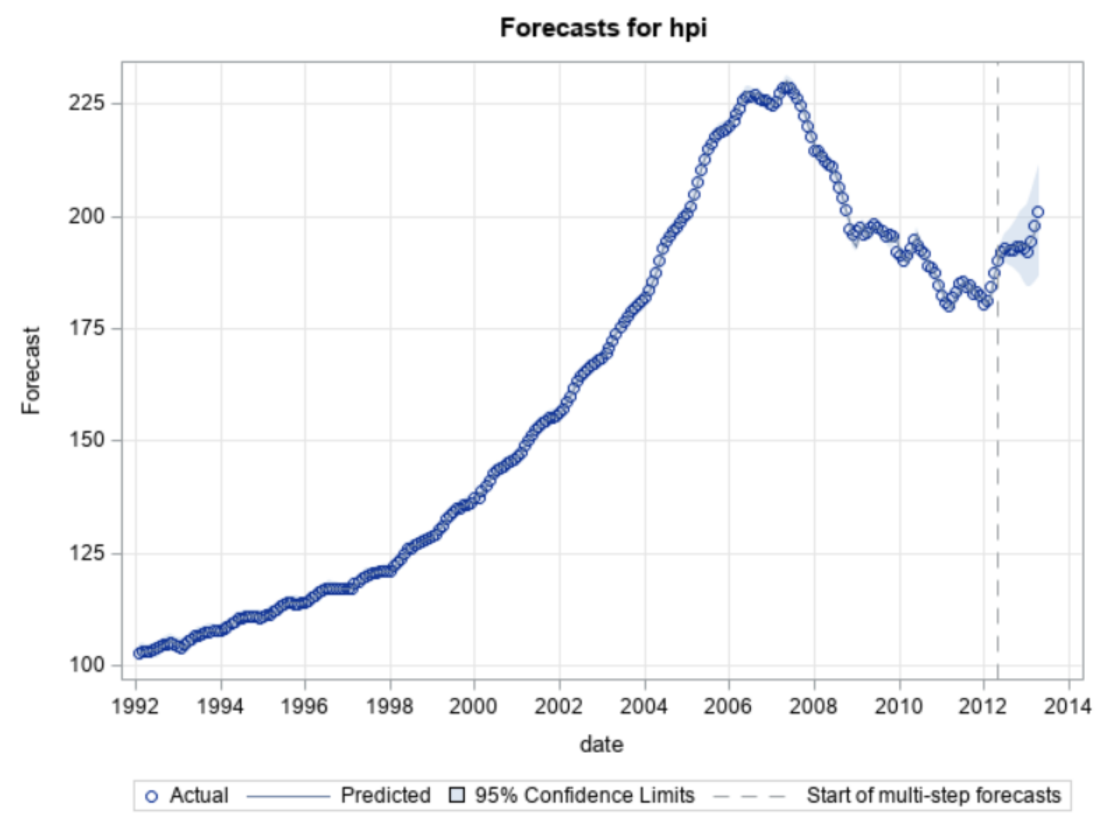
SARIMA(1,1,1)(1,1,1) Model

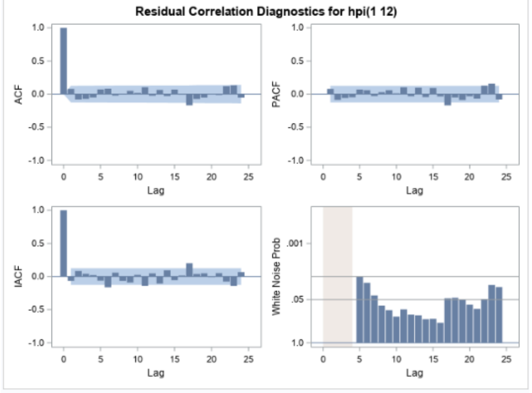
- This model appears to better fit the data, given that it has a lower AIC and SBC.
- Autocorrelation of residuals is also minimal.
- However, seasonal autoregressive parameter is not statistically significant.
- The MAPE of this model was .4078% and the RMSE was .9354
  - This shows a good degree of predictive accuracy, but later models attempted yielded better results

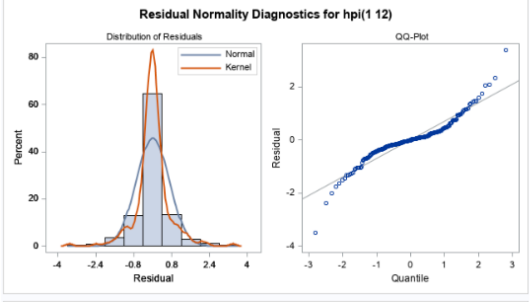
Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	0.04182	0.11158	0.37	0.7078	0
MA1,1	0.66929	0.05742	11.66	<.0001	1
MA2,1	0.74733	0.08037	9.30	<.0001	12
AR1,1	0.96321	0.02334	41.27	<.0001	1
AR2,1	0.05008	0.09856	0.51	0.6114	12

Constant Estimate	0.001462
Variance Estimate	0.496053
Std Error Estimate	0.70431
AIC	558.6947
SBC	576.401
Number of Residuals	255





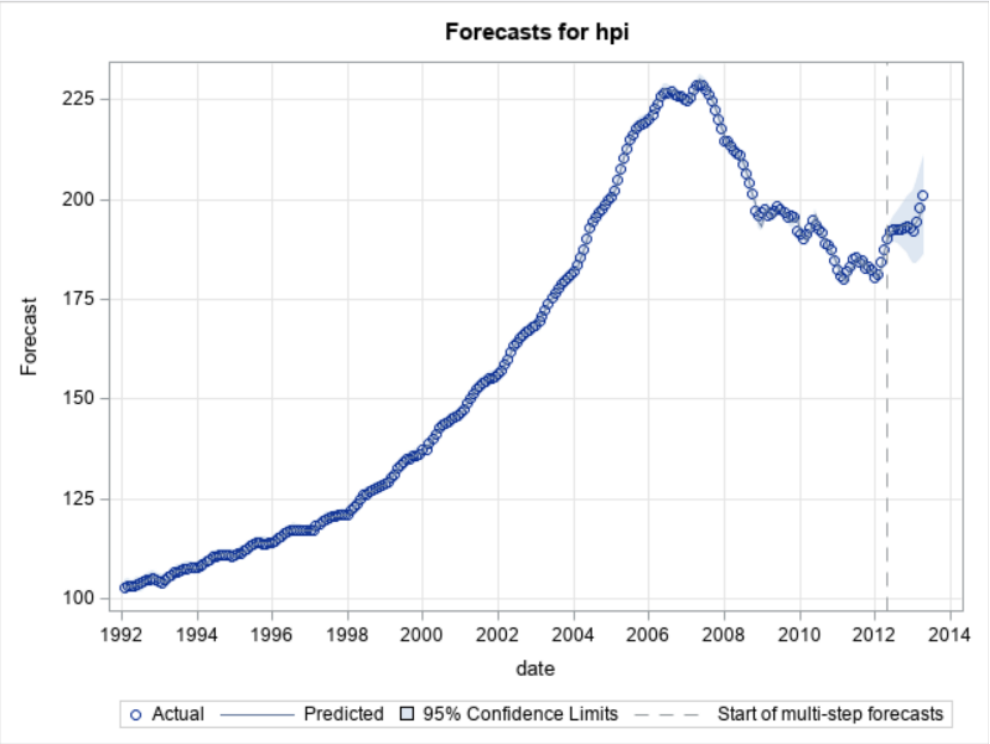


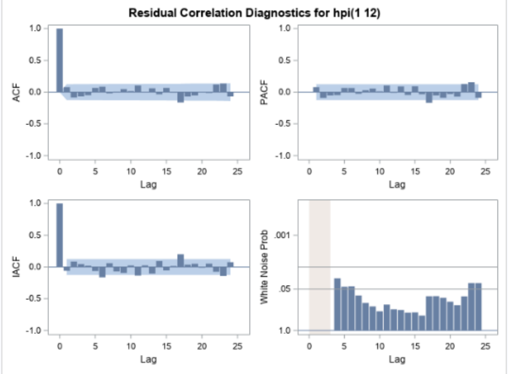
SARIMA(1,1,1)(0,1,1) Model

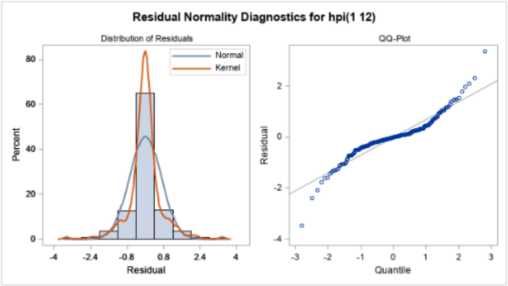
- This model fits the data better than the previous two, given that it has the lowest AIC and SBC values. Nearly all parameters are statistically significant, and autocorrelation among residuals has been further reduced.
- MAPE for this model was 0.38% and RMSE was .9912
  - These values indicate a high degree of near-term forecasting accuracy from the model
- Let us now see what effect exogenous variables may have in improving our model

Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	0.04714	0.11999	0.39	0.6944	0
MA1,1	0.67163	0.05661	11.86	<.0001	1
MA2,1	0.72385	0.05484	13.20	<.0001	12
AR1,1	0.96554	0.02249	42.93	<.0001	1

Constant Estimate	0.001624
Variance Estimate	0.494809
Std Error Estimate	0.703426
AIC	557.0433
SBC	571.2084
Number of Residuals	255

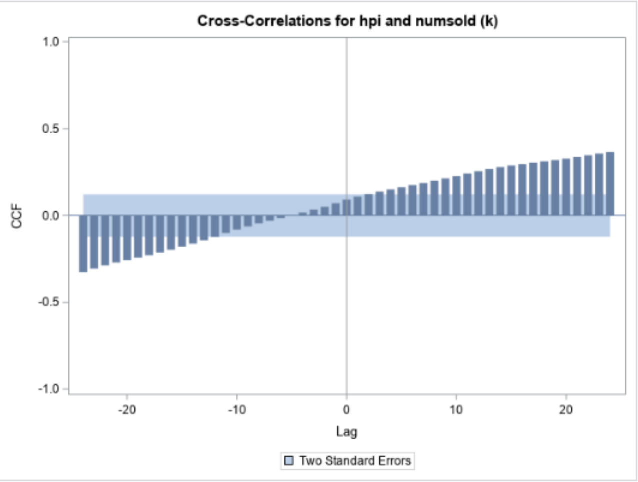


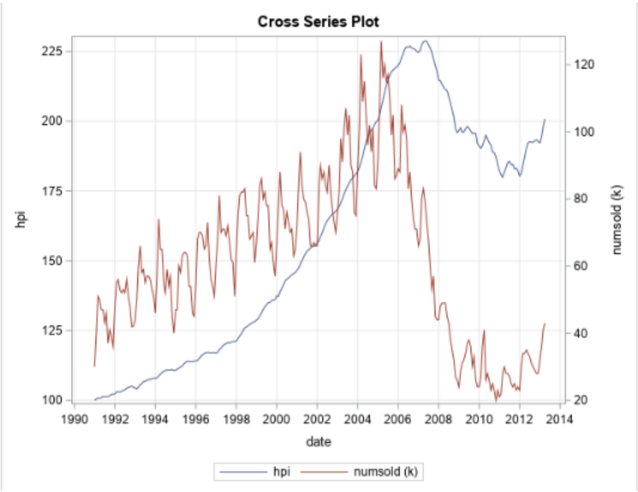


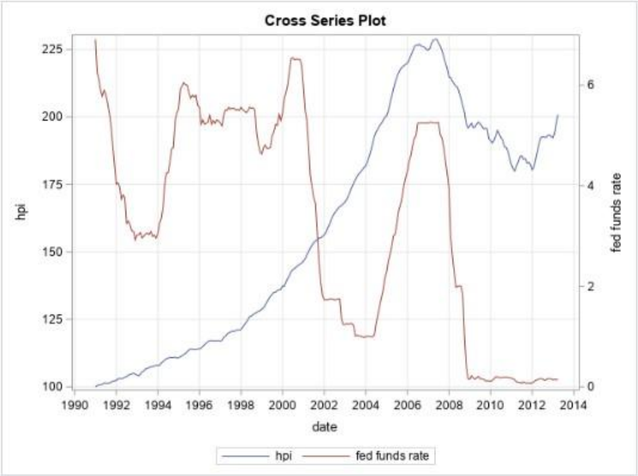


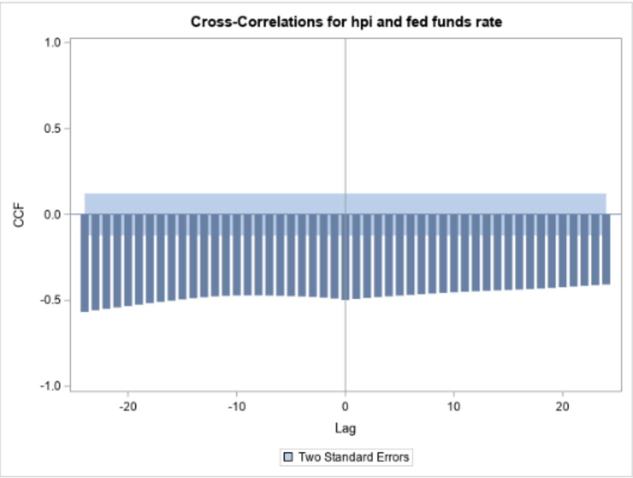
Cross-Correlation with Exogenous Variables

Statistically significant positive correlations at positive lags, indicating that changes in HPI tend to precede shifts in housing sales volume









Statistically significant negative correlations at positive lags, indicating that federal funds rate leads HPI. Since our project's goal is to predict HPI, we decided to not run a model with fed funds rate.

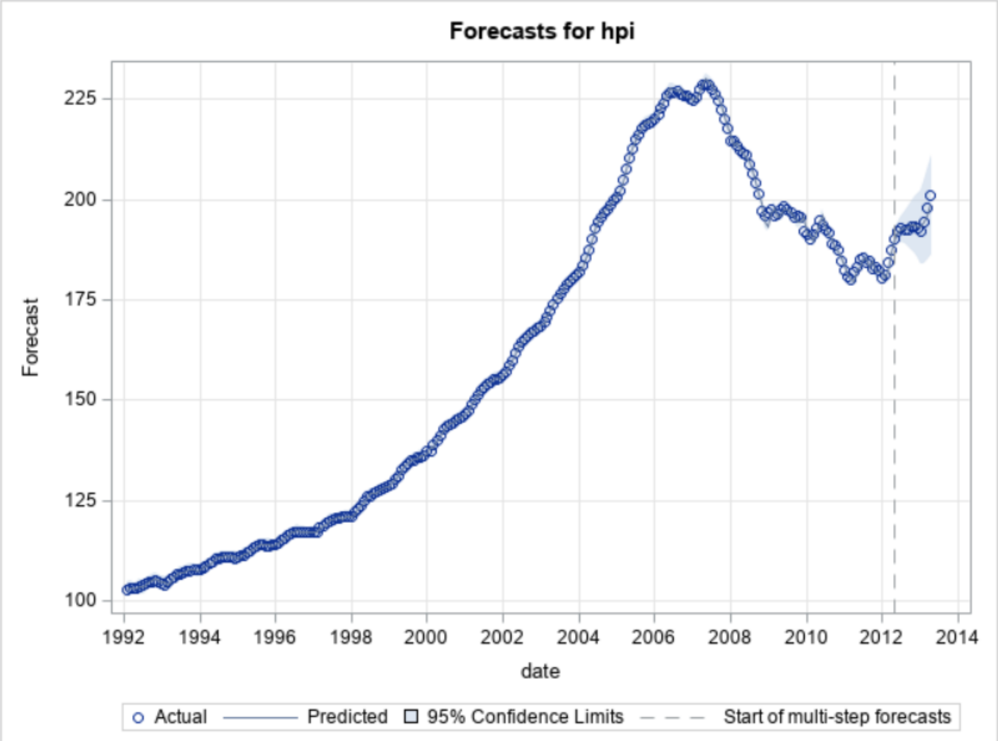
SARIMAX(1,1,1)(0,1,1) Model

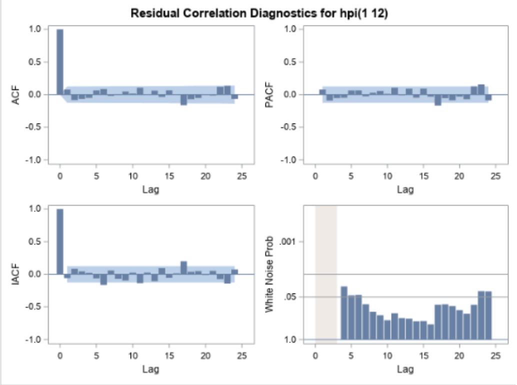
- While the autoregressive and moving average parameters remain significant, the numsold coefficient is not.
- The model seems to have been a worse fit than without exogenous variables.
- Autocorrelation was not reduced by adding NumSold.
- MAPE for this model was 0.3816% and RMSE was .9883
  - This show good predictive performance with values nearly identical to the previous model
  - However, this model is less parsimonious due to the exogenous variable
- Hence, our best candidate model appears to have been SARIMA(1,1,1)(0,1,1)

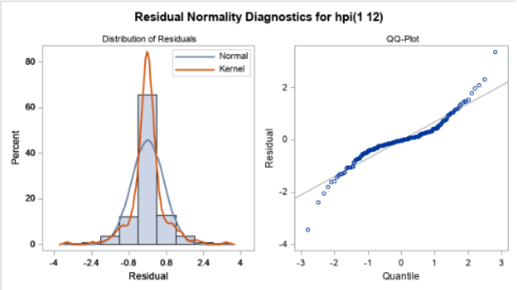
Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	0.04717	0.12001	0.39	0.6943	0	hpi	0
MA1,1	0.67184	0.05672	11.84	<.0001	1	hpi	0
MA2,1	0.72403	0.05510	13.14	<.0001	12	hpi	0
AR1,1	0.96552	0.02254	42.83	<.0001	1	hpi	0
NUM1	0.0018404	0.0076767	0.24	0.8105	0	numsold (k)	0

Constant Estimate	0.001627
Variance Estimate	0.496666
Std Error Estimate	0.704745
AIC	558.9863
SBC	576.6926
Number of Residuals	255



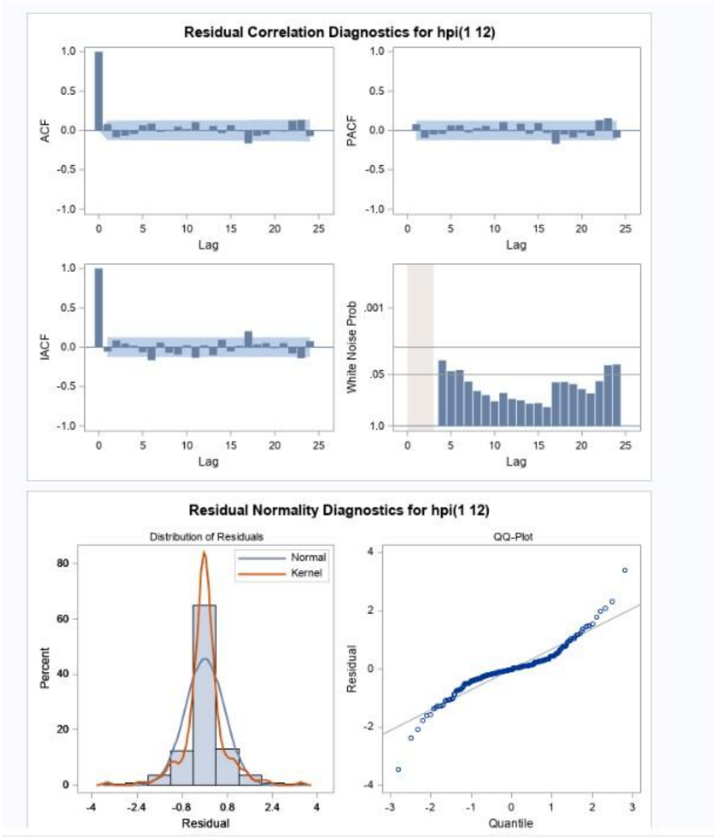




Effect of Event Variable

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	0.04767	0.12093	0.39	0.6934	0	hpi	0
MA1,1	0.67340	0.05656	11.91	<.0001	1	hpi	0
MA2,1	0.72363	0.05496	13.17	<.0001	12	hpi	0
AR1,1	0.96594	0.02242	43.09	<.0001	1	hpi	0
NUM1	0.18371	0.59895	0.31	0.7591	0	events	0

Constant Estimate	0.001624
Variance Estimate	0.496631
Std Error Estimate	0.70472
AIC	558.9513
SBC	576.6576
Number of Residuals	255



Not much of an impact that the event variable had on predicting HPI given our best model so far

Conclusion & Key Takeaways

- **Summary of Findings:**
  - HPI exhibits strong trend and seasonal components
  - Cross-correlation analysis showed that numsold follows HPI, while fed funds rate showed a lagged inverse relationship
  - SARIMA(1,1,1)(0,1,1) provided the best univariate fit, with minimal autocorrelation among residuals and lowest AIC/SBC, as well as the best accuracy metrics with the lowest MAPE and RMSE
  - Our best model showed good short-term accuracy, making it useful for short-term investment and development decision-making
  - The accuracy of the model decays over time, so it may be less useful for long term planning
- **Future Directions:**
  - Test multiple other additional macroeconomic indicators (e.g., unemployment rate, inflation)
  - Explore transfer function models for richer multivariate analysis