Merced/Modesto MSA Analysis

Filipp Krasovsky

6/12/2019

#function for installing libraries  
doCheck<-function(x){  
 if (!require(x, character.only = TRUE)) {  
 install.packages(x, dependencies = TRUE)  
 library(x, character.only = TRUE)  
 }  
}

Introduction

Context: Merced is a city in the San Jouaqin Valley and the county seat of Merced County in California, located relatively close to Modesto. The city has a population of roughly ~86,000 and since 2017, a GDP of 8.65 Billion dollars (unadjusted), according to the federal reserve bank of St. Louis. Modesto is the county seat of Stanislaus Country, with a popular of roughly ~201,000 as of 2010, and has a GDP of 20.64 Billion dollars (unadjusted) as of 2017.

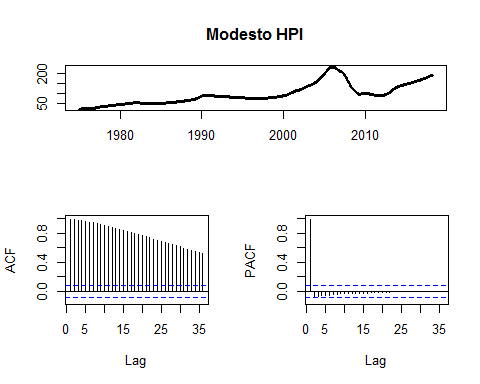
The motivation for this statistical inquiry is to observe the relationship between housing prices in both MSAs based on their proximity to one another; Merced is relatively close to Modesto, with the two city centers only 40 miles away from one another. The data provided measures the Merced and Modesto MSAs, and therefore captures housing data for cities inbetween the two locations included Turlock, Livingston, Patterson, etc.

This project will cover a series of modeling approaches for bivariate regression; namely, we will observe the correlograms of the two time series and attempt to several models which will subsequently be evaluated for best fit using AIC and residual analysis, and will then be tested with a forecast against a holdout set procured from the last year of housing prices.

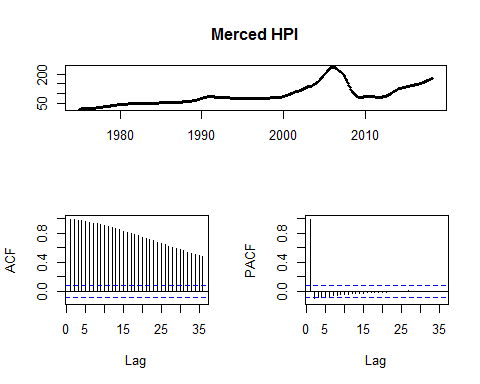
Interpretation of HPI: Taken from the Federal Housing Finance Agency: The FHFA House Price Index (HPI) is a broad measure of the movement of single-family house prices. The HPI is a weighted, repeat-sales index, meaning that it measures average price changes in repeat sales or refinancings on the same properties. This information is obtained by reviewing repeat mortgage transactions on single-family properties whose mortgages have been purchased or securitized by Fannie Mae or Freddie Mac since January 1975.

First, we begin by constructing time series for both data sets, as well as a dummy variable for recession values (Dec 2007 - June 2009):

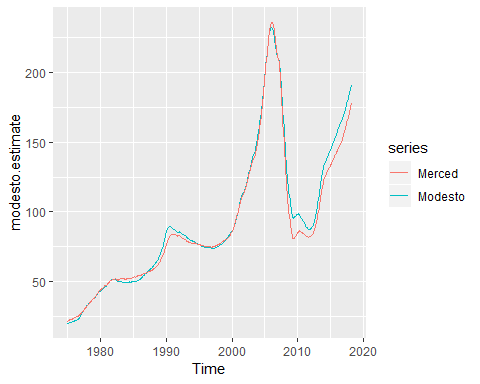
require('forecast')  
require('tseries')  
require('VAR.etp')  
require('vars')  
require('fpp2')  
require('tis')  
  
hpiList<-read.csv("MSA.csv",header=T)  
merced.gdp<-read.csv("GDPMerced.csv",header=T)  
modesto.gdp<-read.csv("GDPModesto.csv",header=T)  
  
#hpi and gdp declaration  
merced.hpi <- ts(hpiList$Merced,start=c(1975,1),freq=12)  
modesto.hpi<- ts(hpiList$Modesto,start=c(1975,1),freq=12)  
recession <- ts(hpiList$Dummy,start=c(1975,1),freq=12)  
  
#prediction and estimation samples for HPI and recession  
modesto.estimate = window(modesto.hpi,start=c(1975,1),end=c(2018,4),freq=12)  
modesto.predict = window(modesto.hpi,start=c(2018,5),freq=12)  
merced.estimate = window(merced.hpi,start=c(1975,1),end=c(2018,4),freq=12)  
merced.predict = window(merced.hpi,start=c(2018,5),freq=12)  
  
recession.estimate = window(recession,start=c(1975,1),end=c(2018,4),freq=12)  
recession.predict = window(recession,start=c(2018,5),freq=12)  
  
#individual and joint graphs of HPI prices.  
tsdisplay(modesto.estimate,main="Modesto HPI")



tsdisplay(merced.estimate,main="Merced HPI")

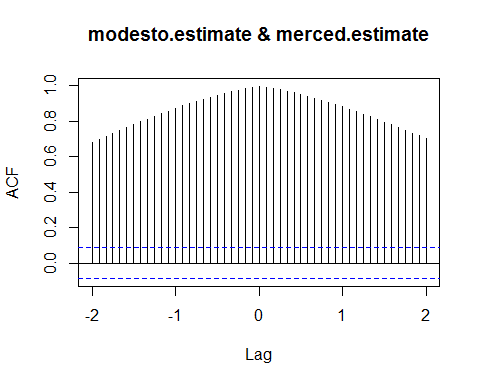


autoplot(modesto.estimate,series="Modesto")+autolayer(merced.estimate,series="Merced")



An initial review of both time series demonstrates gradual decay in the ACF values with a significant spike in the PACF at lag-1, suggesting a textbook case of an AR(1) model. More importantly, each time series has a prominent spike around the 2008 recession which deflates closed to the beginning of 2010. Insofar as this is the case, it seems that an AR(1) model alone will not be able to account for the fluctuation in HPI alone, and suggests the requirement of an indicator variable to explain the noise. The joint graph shows us an obvious similarity in movement between the two HPIs. An STL decomposition suggests minimal presence of seasonality in the series.

ccf(modesto.estimate,merced.estimate) #best correlation at lag zero!



The cross-correlation function also suggests that the highest level of correlation is at zero-lags.This suggests that we can bypass an AR-1 model in exchange for a regular linear model as a possibility for a parsimonious way of forecasting future housing prices. Regardless, we will now attempt several layers of complexity in modeling to properly forecast the signal, check for best-fit, and conclude with a summary and interpretation of the model for both data sets.

Modeling & Stationarity

We begin modeling with an examination of the stationarity of both series. For this, we use the Augmented Dickey Fuller test as a guideline:

adf.test(modesto.estimate)

##   
## Augmented Dickey-Fuller Test  
##   
## data: modesto.estimate  
## Dickey-Fuller = -3.429, Lag order = 8, p-value = 0.04913  
## alternative hypothesis: stationary

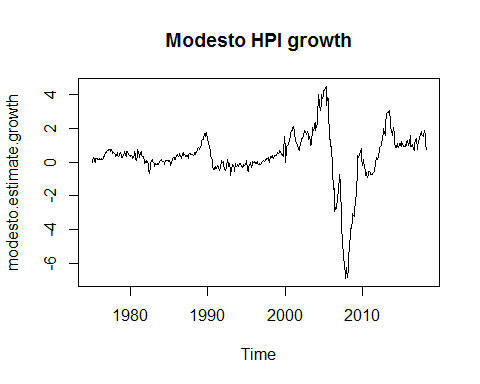
adf.test(merced.estimate)

##   
## Augmented Dickey-Fuller Test  
##   
## data: merced.estimate  
## Dickey-Fuller = -3.4852, Lag order = 8, p-value = 0.04373  
## alternative hypothesis: stationary

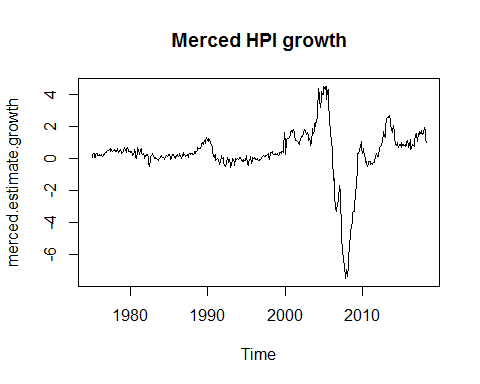
While both series technically are classified as stationary by the ADF test, the closeness of both p-values raises questions that conflict with an intuitive evaluation of both series.

For cautionary purposes, we will begin by taking the first difference of both series and evaluating data signatures from there:

modesto.estimate.growth = diff(modesto.estimate)  
modesto.predict.growth = diff(modesto.predict)  
merced.estimate.growth = diff(merced.estimate)  
merced.predict.growth = diff(merced.predict)  
  
plot(modesto.estimate.growth,main="Modesto HPI growth")



plot(merced.estimate.growth,main="Merced HPI growth")



The Augmented Dickey Fuller test confirms that these series are significantly likelier to be stationary with p-values reasonably further away from alpha:

adf.test(merced.estimate.growth)

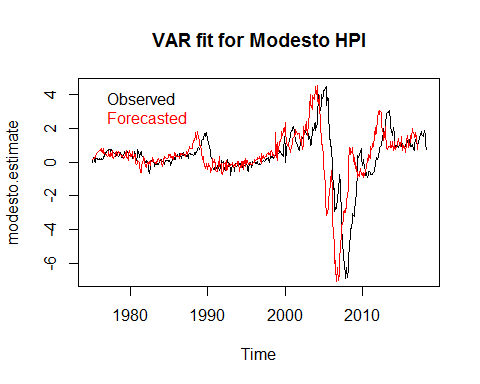
##   
## Augmented Dickey-Fuller Test  
##   
## data: merced.estimate.growth  
## Dickey-Fuller = -3.7591, Lag order = 8, p-value = 0.02105  
## alternative hypothesis: stationary

adf.test(modesto.estimate.growth)

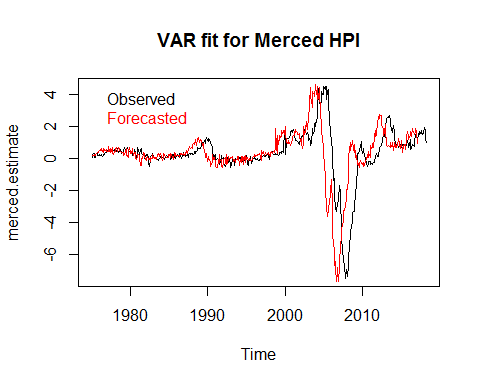
##   
## Augmented Dickey-Fuller Test  
##   
## data: modesto.estimate.growth  
## Dickey-Fuller = -3.3288, Lag order = 8, p-value = 0.06572  
## alternative hypothesis: stationary

From here, we select the best possible Vector Autoregressive model using the Var.select model, and graph the model-fit and residuals: (note: due to R’s graphing setup, model and observed values may have a noticable gap)

merced.estimate = merced.estimate.growth  
modesto.estimate = modesto.estimate.growth  
merced.predict = merced.predict.growth  
modesto.predict = modesto.predict.growth   
  
y\_tot = data.frame(cbind(merced.estimate,modesto.estimate))  
y\_model = VAR(y\_tot,p=14)  
merced.residuals <- y\_model$varresult$merced.estimate$residuals  
modesto.residuals <- y\_model$varresult$modesto.estimate$residuals  
  
plot(modesto.estimate,main="VAR fit for Modesto HPI")  
lines(ts(y\_model$varresult$modesto.estimate$fitted.values,start=c(1975,2),freq=12),col='red')  
legend("topleft",legend=c("Observed","Forecasted"),text.col=c("black","red"),bty="n")

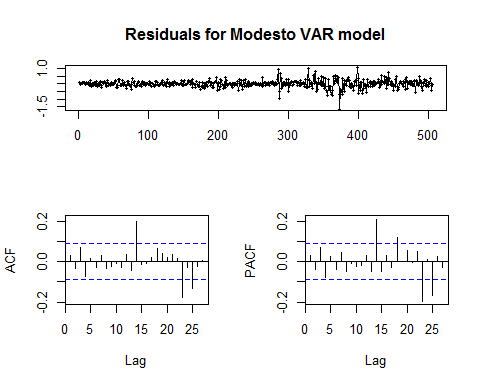


plot(merced.estimate,main="VAR fit for Merced HPI")  
lines(ts(y\_model$varresult$merced.estimate$fitted.values,start=c(1975,2),freq=12),col='red')  
legend("topleft",legend=c("Observed","Forecasted"),text.col=c("black","red"),bty="n")

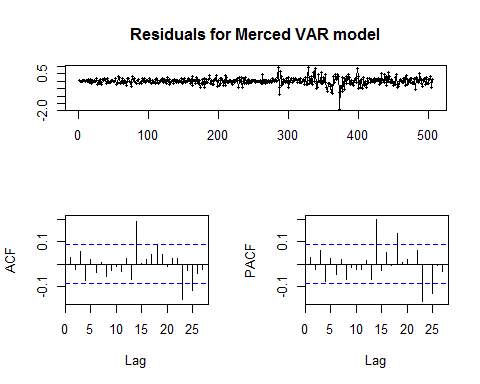


Both data sets are modeled reasonably close from a qualitiatve perspective by a VAR-14 model. We can confirm that the model is adequate by analyzing the residuals for both series:

tsdisplay(modesto.residuals,main="Residuals for Modesto VAR model")

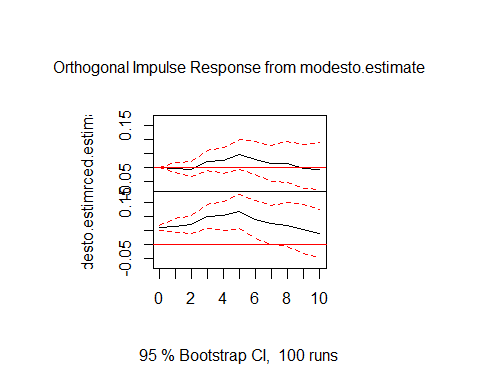
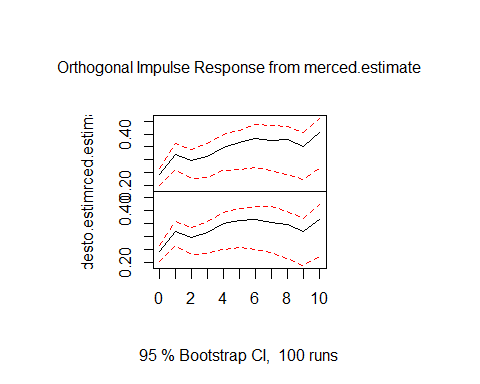


tsdisplay(merced.residuals,main ="Residuals for Merced VAR model")



Both residual sets exhibit a white noise pattern, allowing us to conclude that a VAR model fit is appropriate for the situation. We can then move on to examine the impulse response function for both time series and determine the effect housing prices have on one another.

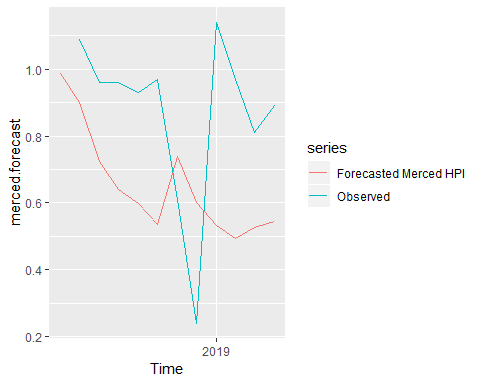
ir1<-irf(y\_model)  
plot(ir1)



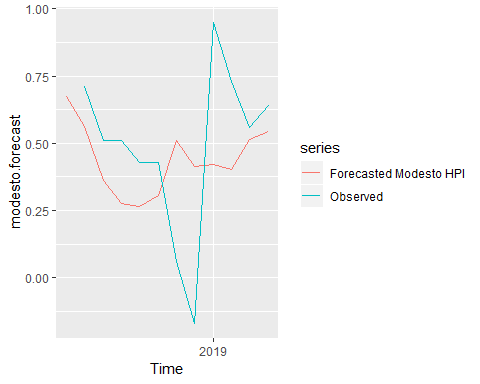
Summarily, the interpretation of the Impulse Response Function leads us to conclude that a 1-unit change in the HPI of Merced is significantly associated with an increase in the HPI of both housing markets. Conversely, the signal from Modesto’s housing market seems to be relatively insignificant in determining Merced’s HPI, while a unit increase in Modesto’s HPI does significantly affect future HPI values of itself. The 95% confidence interval for the impulse on Merced crosses zero very early, suggesting a possibility that the orthogonal response is altogether insignificant.

From here, we can begin moving to a 12-step ahead forecast and examining the residuals from the model:

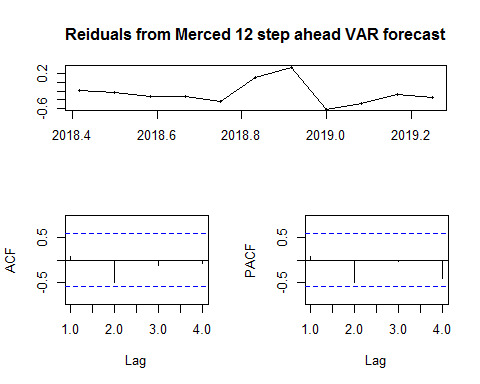
y\_forecast <- predict(object=y\_model,n.ahead=12)  
  
merced.forecast <- ts(y\_forecast$fcst$merced.estimate[,1],start=c(2018,5),freq=12)  
modesto.forecast<- ts(y\_forecast$fcst$modesto.estimate[,1],start=c(2018,5),freq=12)  
autoplot(merced.forecast,series="Forecasted Merced HPI")+autolayer(merced.predict,series="Observed")



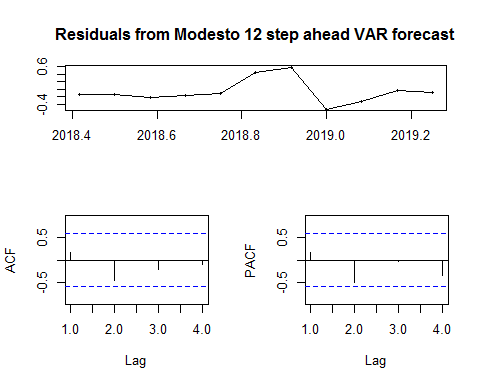
autoplot(modesto.forecast,series="Forecasted Modesto HPI")+autolayer(modesto.predict,series="Observed")



merced.forecast.residuals <- merced.forecast - merced.predict  
modesto.forecast.residuals<-modesto.forecast -modesto.predict  
tsdisplay(merced.forecast.residuals,main="Reiduals from Merced 12 step ahead VAR forecast")



tsdisplay(modesto.forecast.residuals,main="Residuals from Modesto 12 step ahead VAR forecast")



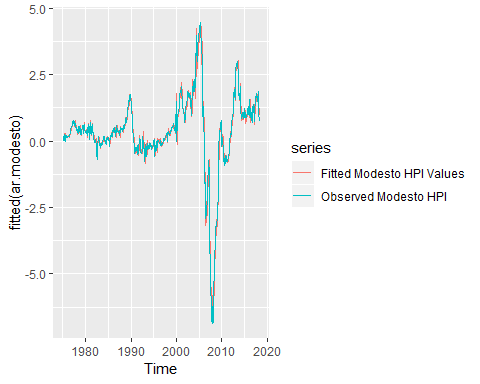
Interpretation: While the model fit is overall satisfactory, the forecasting strength of the VAR model is somewhat challenge by a large gap in values right before the start of 2019. Nevertheless, the observed values outperform the mode significantly, and other methods of forecasting are worth inspection.

From here, we can compare and contrast the performance of a VAR model with a univariate ARIMA process to get a better model fit for both series.

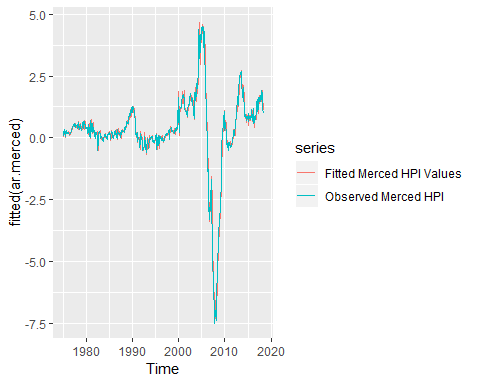
Univariate Forecast for the Modesto series

A previous correlogram as displayed above suggests to us that the modesto HPI growth has a strong AR and MA signal, as well as the possibility of a seasonal AR signature. Although we know inituitive that a major spike in volatility happens around the time of the recession, we will ignore this for now and proceed with a regular ARIMA model, and compare performance to an ARIMA+Dummy variable model afterwards for both series. The Merced HPI growth exhibits a dominant AR signature, but the includsion of an MA component to forego using 15 lags may still be optimal. Regardless, we will proceed ahead with a high-order AR process. (Note: AIC comparrison shows that an AR-15 outperforms auto.arima()).

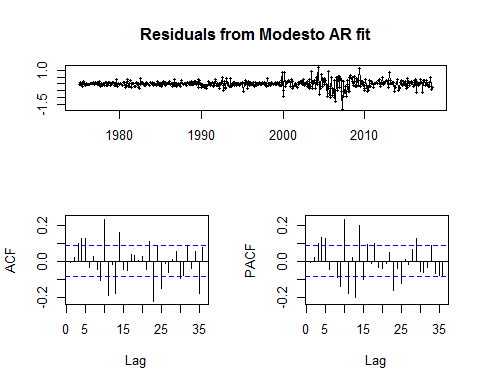
ar.modesto<-arima(modesto.estimate.growth,order=c(3,0,1),seasonal=list(order=c(1,0,0)))   
ar.merced <-arima(merced.estimate.growth,order=c(3,0,1),seasonal=list(order=c(1,0,0)))  
  
autoplot(fitted(ar.modesto),series="Fitted Modesto HPI Values")+autolayer(modesto.estimate.growth,series="Observed Modesto HPI")



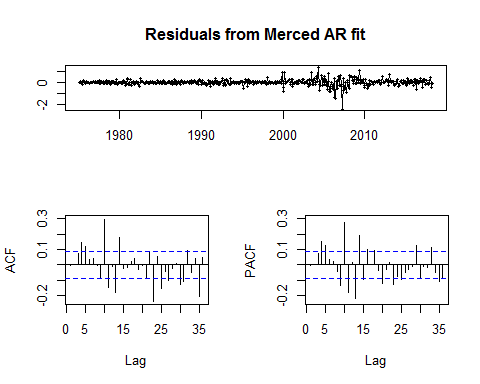
autoplot(fitted(ar.merced ),series="Fitted Merced HPI Values ")+autolayer(merced.estimate.growth,series="Observed Merced HPI")



tsdisplay(ar.modesto$residuals,main="Residuals from Modesto AR fit")

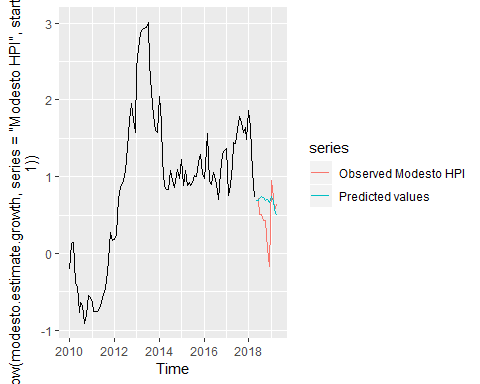


tsdisplay(ar.merced$residuals,main="Residuals from Merced AR fit")

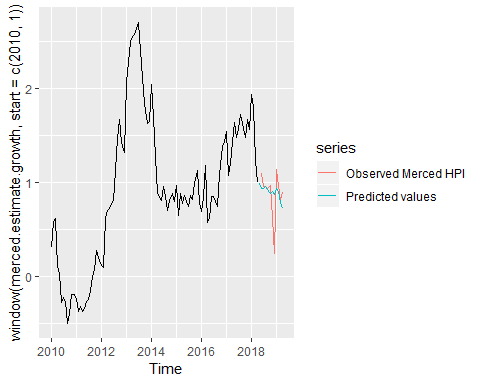


Overall, an ARMA model for both series seems to fit the model relatively well. The graph of the residuals for both models reveals significant clustering around the 2008-2009 region, suggesting that the model does not account for the effect of the 2008/2009 recession. Regardless, the residuals follow a white noise pattern, and we can therefore use the model to forecast.

ar.modesto.predict <- predict(ar.modesto,n.ahead=12)  
ar.merced.predict <- predict(ar.merced, n.ahead=12)  
autoplot(window(modesto.estimate.growth,series="Modesto HPI",start=c(2010,1)))+autolayer(modesto.predict.growth,series="Observed Modesto HPI")+autolayer(ar.modesto.predict$pred,series="Predicted values")

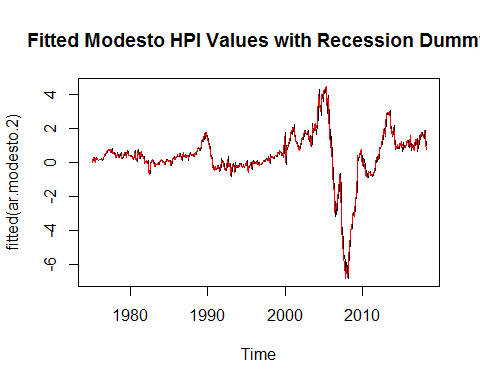


autoplot(window(merced.estimate.growth,start=c(2010,1)))+autolayer(merced.predict.growth,series="Observed Merced HPI")+autolayer(ar.merced.predict$pred,series="Predicted values")

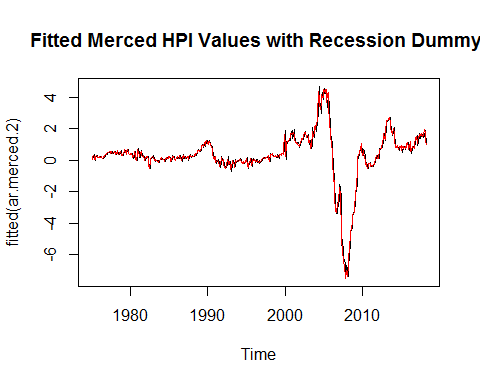


An evaluation of the forecast compared to the observed Modesto & Merced HPI shows that the forecast performs rather poorly for the Modesto time series, while the Merced forecast has considerable success with the exception of the large dip in HPI process closed to the end of 2019. From here, we can attempt another model fit where we control for the effects of the recession but retain the same ARMA models for each series.

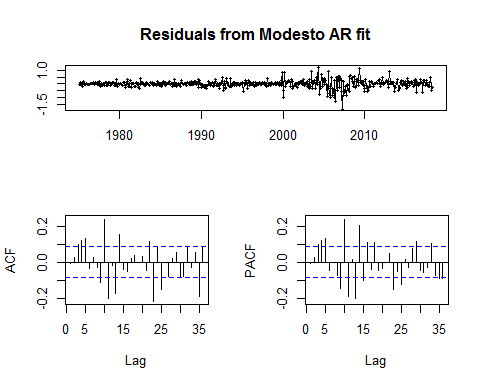
recession.estimate=window(recession.estimate,start=c(1975,2))  
  
ar.modesto.2<-Arima(modesto.estimate.growth,order=c(3,0,1),seasonal=c(1,0,0),xreg=recession.estimate)  
ar.merced.2 <-Arima(merced.estimate.growth,order=c(3,0,1),seasonal=c(1,0,0),xreg=recession.estimate)  
  
plot(fitted(ar.modesto.2),main="Fitted Modesto HPI Values with Recession Dummy")  
lines(modesto.estimate.growth,col='red')



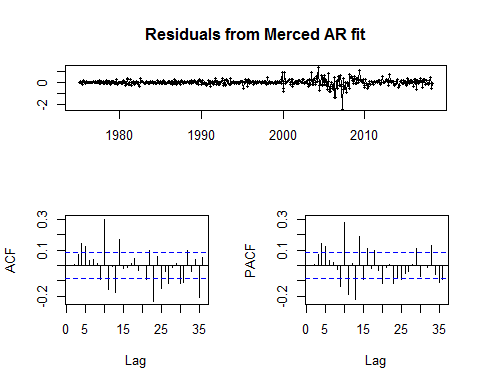
plot(fitted(ar.merced.2 ),main="Fitted Merced HPI Values with Recession Dummy ")  
lines(merced.estimate.growth,col='red')



tsdisplay(ar.modesto.2$residuals,main="Residuals from Modesto AR fit")

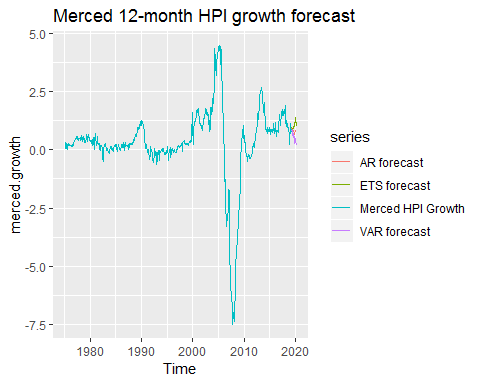


tsdisplay(ar.merced.2$residuals,main="Residuals from Merced AR fit")

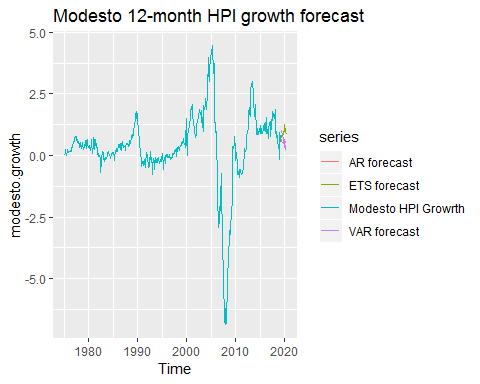


Overall, accounting for the recession has almost no effect on the model fit, and thus the inclusion of a dummy variable is not useful in meaningfully determining future values of the HPI for either county. At this point, we can aggregate our two types of forecasts and combine them with a standard ETS, and make an assessment of the future values of the HPI prices.

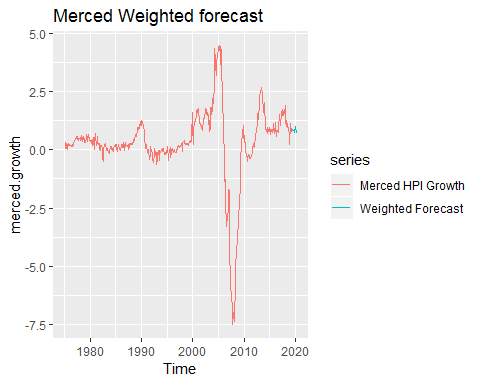
merced.growth <- diff(merced.hpi)  
modesto.growth <- diff(modesto.hpi)  
  
#out of sample forecasts  
  
#AR  
ar.modesto<-arima(modesto.growth,order=c(3,0,1),seasonal=list(order=c(1,0,0)))   
ar.merced <-arima(merced.growth,order=c(3,0,1),seasonal=list(order=c(1,0,0)))  
f.ar.modesto<-predict(ar.modesto,n.ahead=12)  
f.ar.merced <-predict(ar.merced, n.ahead=12)  
  
#VAR  
y\_tot = data.frame(cbind(merced.growth,modesto.growth))  
y\_model = VAR(y\_tot,p=14)  
y\_fcst = predict(y\_model,n.ahead=12)  
var.modesto = ts(y\_fcst$fcst$merced.growth[,1],start=c(2019,5),freq=12)  
var.merced = ts(y\_fcst$fcst$modesto.growth[,1],start=c(2019,5),freq=12)  
  
#ETS  
ets.modesto<-predict(ets(modesto.growth,model="ZAA"),h=12)  
ets.merced <-predict(ets(merced.growth ,model="ZAA"),h=12)  
  
#Weighted  
modesto.weighted <- (1/4) \* f.ar.modesto$pred + (1/4) \* var.modesto + (1/2) \* ets.modesto$mean  
merced.weighted <- (1/4) \* f.ar.merced$pred + (1/4)\*var.merced + (1/2)\*ets.merced$mean   
  
autoplot(merced.growth,series = "Merced HPI Growth",main="Merced 12-month HPI growth forecast")+autolayer(f.ar.merced$pred,series="AR forecast")+autolayer(var.merced,series="VAR forecast")+autolayer(ets.merced$mean,series="ETS forecast")



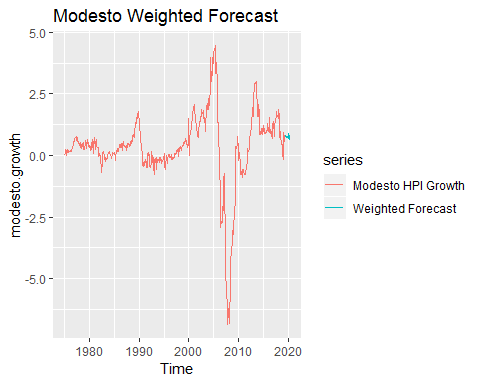
autoplot(modesto.growth,series="Modesto HPI Growrth",main="Modesto 12-month HPI growth forecast")+autolayer(f.ar.modesto$pred,series="AR forecast")+  
 autolayer(var.modesto,series="VAR forecast")+autolayer(ets.modesto$mean,series="ETS forecast")



autoplot(merced.growth,series="Merced HPI Growth",main="Merced Weighted forecast")+autolayer(merced.weighted,series="Weighted Forecast")



autoplot(modesto.growth,series="Modesto HPI Growth",main="Modesto Weighted Forecast")+autolayer(modesto.weighted,series="Weighted Forecast")



Conclusion: Overall , we can conclude from current modeling attempts that HPI growth will stay relatively constant, and thus we can interpret this to mean that the housing market in both areas will experience steady growth at a rate of approximately 1%-1.2%.