Bay Area MSA Analysis

Filipp Krasovsky

5/20/2019

Overview: This analysis serves to identify the temporal signal in housing price growth in the San Francisco/Santa Cruz metropolitan areas, as well as to provide an exercise in vector autoregressive models between the two time series and the shock/effect they have on one another. Data is provided by Freddie Mac, and is fed through both a VAR regression as well as a univariate ARMA analysis for each series; these models are then compared on mean squared errors to determine accuracy. We use a prediction space of the last 20 values for each series. All data is taken in monthly aggregation, and all unit measurements of temporal analysis (ie lags) refer to x-month differences.

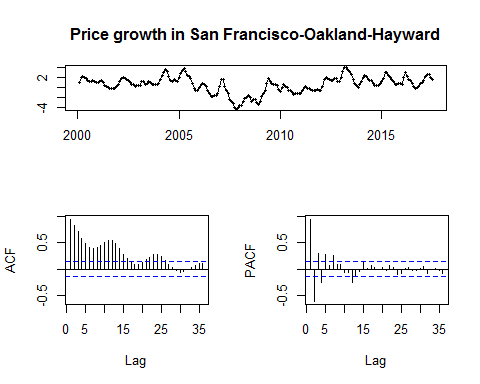
1. Sample Description: non-adjusted MSA prices for San Francisco/Oakland/Hayward and Santa-Cruz/Watsonville from Jan 2001 to March 2019 transformed into price growth time series.

For this analysis, we use the forecast, vars, combMsc, and VAR libraries:

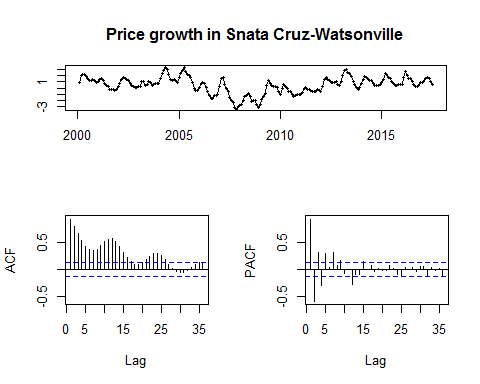
library('forecast')  
require('vars')  
require('CombMSC')  
library('VAR.etp')  
require('MLmetrics')

Data Source: Freddie Mac, Accessed 5-20-2019: <http://www.freddiemac.com/research/indices/house-price-index.page>

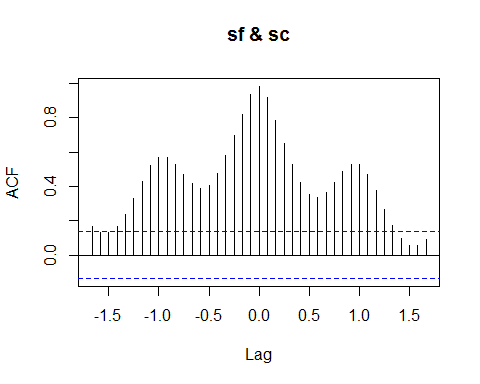
#initialize data   
z = read.csv('prices.csv',header=TRUE)  
sf = ts(z$SF,start=c(2000,1),freq=12)  
sc = ts(z$SC,start=c(2000,1),freq=12)  
  
#growth transformation   
sf = diff(sf)  
sc = diff(sc)  
#create a prediction sample and estimation sample   
sfp=window(sf,start=c(2017,8),freq=12)  
scp=window(sc,start=c(2017,8),freq=12)  
sf=window(sf,start=c(2000,1),end=c(2017,7),freq=12)  
sc=window(sc,start=c(2000,1),end=c(2017,7),freq=12)  
  
tsdisplay(sf,main="Price growth in San Francisco-Oakland-Hayward")



tsdisplay(sc,main="Price growth in Snata Cruz-Watsonville")

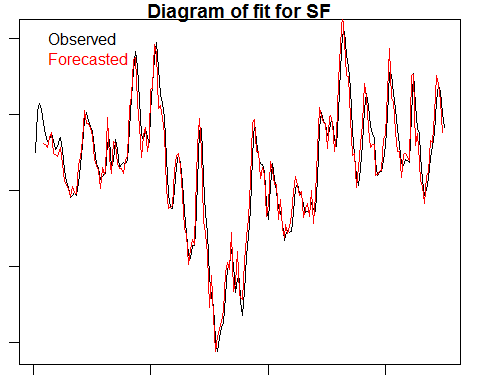


ccf(sf,sc)

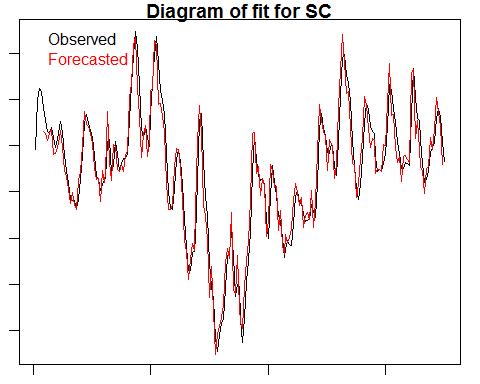


VAR Estimation using BIC suggests an optimal lag-5 VAR fit for the two series.

y = cbind(sf,sc)  
y\_tot = data.frame(y)  
  
bestP = VAR.select(y\_tot,ic="sc",pmax=20)  
bestP = as.numeric(bestP$p)#==>5   
m1 = VAR(y\_tot,p=5)  
  
par(mar=c(1,1,1,1))  
plot(sf,main="Diagram of fit for SF")  
lines(ts(m1$varresult$sf$fitted.values,start=c(2000,6),freq=12),col='red')  
legend("topleft",legend=c("Observed","Forecasted"),text.col=c("black","red"),bty="n")



plot(sc,main="Diagram of fit for SC")  
lines(ts(m1$varresult$sc$fitted.values,start=c(2000,6),freq=12),col='red')  
legend("topleft",legend=c("Observed","Forecasted"),text.col=c("black","red"),bty="n")



summary(m1)

##   
## VAR Estimation Results:  
## =========================   
## Endogenous variables: sf, sc   
## Deterministic variables: const   
## Sample size: 205   
## Log Likelihood: 286.216   
## Roots of the characteristic polynomial:  
## 0.8943 0.8654 0.8654 0.777 0.777 0.7699 0.7699 0.5575 0.5575 0.4208  
## Call:  
## VAR(y = y\_tot, p = 5)  
##   
##   
## Estimation results for equation sf:   
## ===================================   
## sf = sf.l1 + sc.l1 + sf.l2 + sc.l2 + sf.l3 + sc.l3 + sf.l4 + sc.l4 + sf.l5 + sc.l5 + const   
##   
## Estimate Std. Error t value Pr(>|t|)   
## sf.l1 3.295484 0.514163 6.409 1.08e-09 \*\*\*  
## sc.l1 -1.616468 0.555286 -2.911 0.00402 \*\*   
## sf.l2 -2.742338 1.127553 -2.432 0.01592 \*   
## sc.l2 1.304534 1.203964 1.084 0.27992   
## sf.l3 1.152324 1.306506 0.882 0.37887   
## sc.l3 0.020775 1.390438 0.015 0.98809   
## sf.l4 -1.680743 1.157509 -1.452 0.14811   
## sc.l4 0.964190 1.238418 0.779 0.43718   
## sf.l5 1.267764 0.547019 2.318 0.02151 \*   
## sc.l5 -1.085416 0.594695 -1.825 0.06951 .   
## const 0.009315 0.026541 0.351 0.72598   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
##   
## Residual standard error: 0.345 on 194 degrees of freedom  
## Multiple R-Squared: 0.9573, Adjusted R-squared: 0.9551   
## F-statistic: 435.1 on 10 and 194 DF, p-value: < 2.2e-16   
##   
##   
## Estimation results for equation sc:   
## ===================================   
## sc = sf.l1 + sc.l1 + sf.l2 + sc.l2 + sf.l3 + sc.l3 + sf.l4 + sc.l4 + sf.l5 + sc.l5 + const   
##   
## Estimate Std. Error t value Pr(>|t|)   
## sf.l1 1.225436 0.475911 2.575 0.0108 \*  
## sc.l1 0.478680 0.513975 0.931 0.3528   
## sf.l2 -1.207320 1.043668 -1.157 0.2488   
## sc.l2 -0.241456 1.114394 -0.217 0.8287   
## sf.l3 0.415513 1.209308 0.344 0.7315   
## sc.l3 0.746328 1.286995 0.580 0.5627   
## sf.l4 -1.215647 1.071395 -1.135 0.2579   
## sc.l4 0.502481 1.146285 0.438 0.6616   
## sf.l5 1.088626 0.506323 2.150 0.0328 \*  
## sc.l5 -0.907324 0.550452 -1.648 0.1009   
## const 0.003667 0.024567 0.149 0.8815   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
##   
## Residual standard error: 0.3194 on 194 degrees of freedom  
## Multiple R-Squared: 0.9473, Adjusted R-squared: 0.9446   
## F-statistic: 348.7 on 10 and 194 DF, p-value: < 2.2e-16   
##   
##   
##   
## Covariance matrix of residuals:  
## sf sc  
## sf 0.1191 0.1091  
## sc 0.1091 0.1020  
##   
## Correlation matrix of residuals:  
## sf sc  
## sf 1.0000 0.9903  
## sc 0.9903 1.0000

Box.test(m1$varresult$sf$residuals)

##   
## Box-Pierce test  
##   
## data: m1$varresult$sf$residuals  
## X-squared = 0.02279, df = 1, p-value = 0.88

Box.test(m1$varresult$sc$residuals)

##   
## Box-Pierce test  
##   
## data: m1$varresult$sc$residuals  
## X-squared = 0.041431, df = 1, p-value = 0.8387

Remarks: The model fit is, on face value, sufficient enough to reduce the residuals to what appears to be white noise, and qualitatively fits the data provided. The Box-test concludes that we do not have enough evidence to reject the null hypothesis that there is no serial correlation in the residuals for either model. San Francisco’s price growth is largely dependent on lags of itself at l=1,2,5 and only significantly depends on price growth in Santa Clara at lag 1. Santa Clara price growth is significantly correlated to prices in San Francisco at lag = 1,5.

1. Granger Causality Analysis: Time series seem to granger-cause one another; with limited subject matter knowledge, this outcome is intuitive given the level of economic activity in both areas, although there is a higher change of a type-1 error for the rejection of the null hypothesis that San Francisco housing prices do not granger-cause santa cruz prices.

grangertest(sf ~ sc, order = bestP)

## Granger causality test  
##   
## Model 1: sf ~ Lags(sf, 1:5) + Lags(sc, 1:5)  
## Model 2: sf ~ Lags(sf, 1:5)  
## Res.Df Df F Pr(>F)   
## 1 194   
## 2 199 -5 5.1526 0.0001839 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

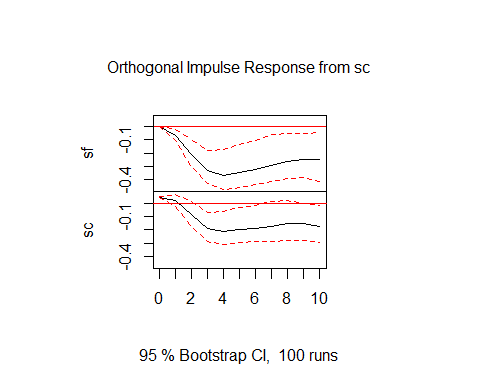
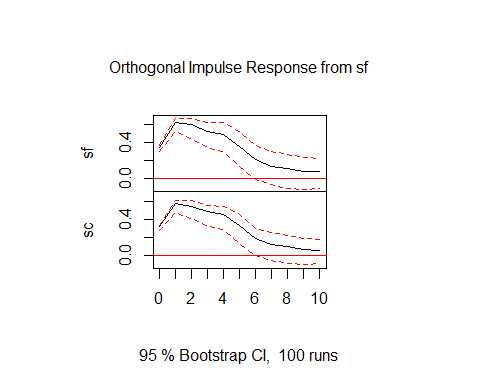
grangertest(sc ~ sf, order = bestP)

## Granger causality test  
##   
## Model 1: sc ~ Lags(sc, 1:5) + Lags(sf, 1:5)  
## Model 2: sc ~ Lags(sc, 1:5)  
## Res.Df Df F Pr(>F)   
## 1 194   
## 2 199 -5 4.2015 0.001198 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

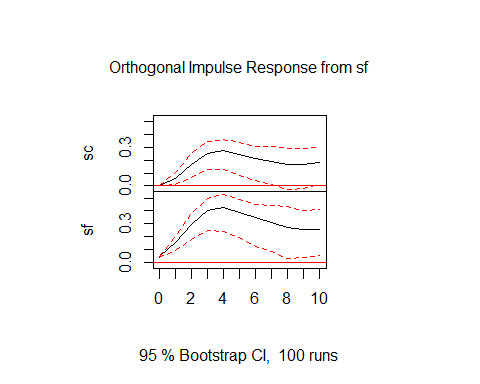
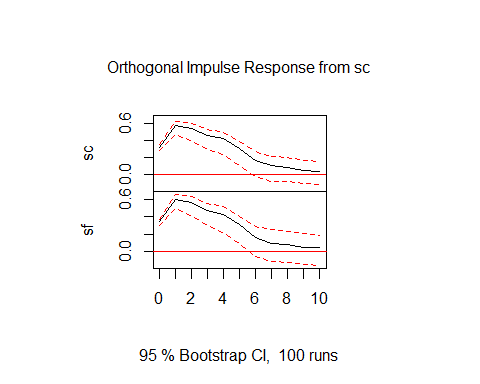
1. Impulse Response:

Order does affect the impulse response function visually (sign change), but does not affect the absolute value of the impulse. Analysis will be done for the original order, preceded by proof of effect from the order change.

#order: 1. sf 2.sc   
ir1 <- irf(m1)   
  
#order: 1.sc 2.sf   
y=cbind(sc,sf)  
y\_tot=data.frame(y)  
m2 = VAR(y\_tot,p=5)  
  
ir2 <-irf(m2)  
plot(ir1)



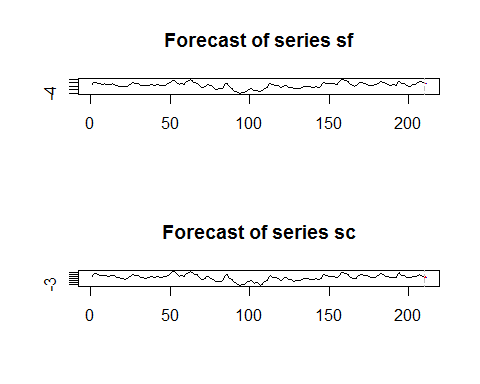
plot(ir2)



The impulse effect of San Francisco’s housing price growth remains relatively constant for 5 months, become insignificant at month 6 in affect future housing prices in San Francisco. For Santa Cruz, the impulse retains the same behavior, becoming insignificant after 7 months and slowly declines in effect beforehand with a sharp spike in month 1. The effect of Santa Cruz’s price increase seems to have a long term effect on San Francisco housing prices that eventally terminalizes at a price drop of -.4 into month 10, with similar behavior in affecting future prices of itself. In short, San Francisco shocks have a large effect that decays slowly, while Santa Cruz shocks have no effect at first but increase over time in absolute magnitude.

1. Prediction/Forecast: We begin with a 1-step ahead forecast for the VAR model, and compare it with 1 step ahead forecast for a univariate model of each series. The ACF+PACF of San Francisco prices suggest an AR-12 model as a possible univariate fit, while Santa Cruz possibly follows an AR series of order 15. A summary conclusion of the difference between the VAR forecast and AR forecasts suggests that the univariate model performs slightly better for the 1 step ahead forecast.

m1.predict = predict(m1,n.ahead=1)  
plot(m1.predict)



#univariate models:  
u1= arima(sf,order=c(12,0,0))  
u2= arima(sc,order=c(15,0,0))  
  
u1.p = predict(u1,n.ahead=1)  
u2.p = predict(u2,n.ahead=1)  
  
#residual computation for 1 step ahead   
sf.h1 <- sfp[1]  
sc.h1 <- scp[1]  
  
"MSE for VAR predict: San Francisco Prices"

## [1] "MSE for VAR predict: San Francisco Prices"

(m1.predict$fcst$sf[1] - sf.h1)^2

## [1] 0.01706678

"MSE for VAR predict: Santa Cruz Prices"

## [1] "MSE for VAR predict: Santa Cruz Prices"

(m1.predict$fcst$sc[1] - sc.h1)^2

## [1] 0.01426316

"MSE for ARMA forecast: San Francisco Prices"

## [1] "MSE for ARMA forecast: San Francisco Prices"

(u1.p$pred[1] - sf.h1)^2

## [1] 0.003015401

"MSE for ARMA forecast: Santa Cruz Prices"

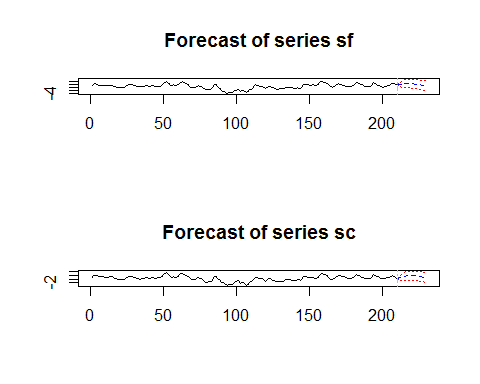
## [1] "MSE for ARMA forecast: Santa Cruz Prices"

(u2.p$pred[1] - sc.h1)^2

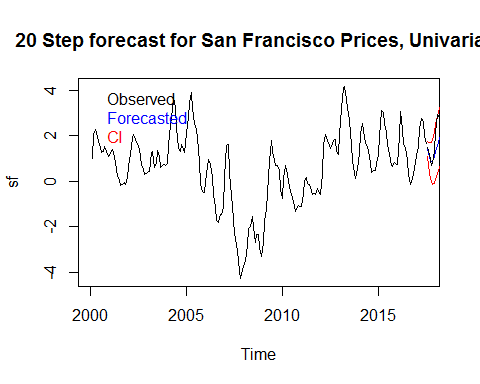
## [1] 0.003526601

1. After a 1-step ahead forecast, we move to a 20-step ahead forecast. After this, we include data from a third MSA outside the Bay Area for VAR analysis and Granger Causality.

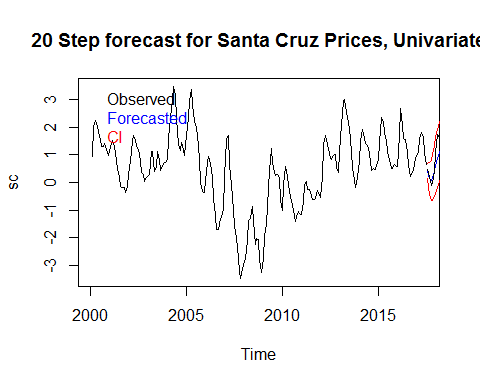
m2 = predict(m1,n.ahead=20)  
u1.p = predict(u1,n.ahead=20) #sf  
u2.p = predict(u2,n.ahead=20) #sc  
  
plot(m2)



plot(sf,main="20 Step forecast for San Francisco Prices, Univariate")  
lines(u1.p$pred,col='blue')  
lines(u1.p$pred+u1.p$se,col='red')  
lines(u1.p$pred-u1.p$se,col='red')  
lines(sfp)  
legend("topleft",legend=c("Observed","Forecasted","CI"),text.col=c("black","blue","red"),bty="n")



plot(sc,main="20 Step forecast for Santa Cruz Prices, Univariate")  
lines(u2.p$pred,col='blue')  
lines(u2.p$pred+u2.p$se,col='red')  
lines(u2.p$pred-u2.p$se,col='red')  
lines(scp)  
legend("topleft",legend=c("Observed","Forecasted","CI"),text.col=c("black","blue","red"),bty="n")



#MSE analysis  
m2.sfe <- MSE(m2$fcst$sf[,1],sfp)  
m2.sce <- MSE(m2$fcst$sc[,1],scp)  
u1.e <- MSE(u1.p$pred,sfp)  
u2.e <- MSE(u2.p$pred,scp)  
  
"VAR Model: San Francisco Prices, Santa Cruz Prices Mean Squared Error"

## [1] "VAR Model: San Francisco Prices, Santa Cruz Prices Mean Squared Error"

m2.sfe

## [1] 2.936384

m2.sce

## [1] 2.896519

"ARMA Models: San Francisco Prices, Santa Cruz Prices Mean Squared Error"

## [1] "ARMA Models: San Francisco Prices, Santa Cruz Prices Mean Squared Error"

u1.e

## [1] 1.222686

u2.e

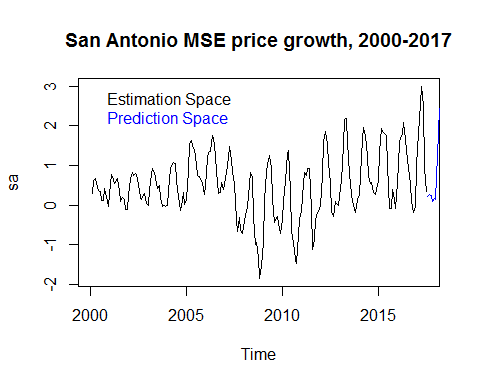
## [1] 0.8013183

Again, the univariate forecast outperforms the VAR model considerably, judging by the MSE figures provided above. From this point, we include analysis on MSA price growth from the San Antonio-New Braunfels, TX MSA for the same period of time and calculate growth rates as well as divide it into a prediction and estimate space.

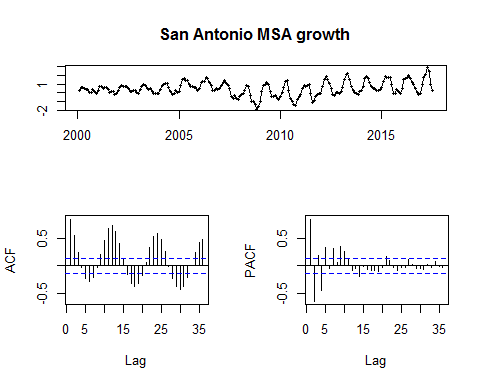
sa = ts(z$SA,start=c(2000,1),freq=12)  
sa = diff(sa)  
sap=window(sa,start=c(2017,8),freq=12)  
sa =window(sa,start=c(2000,1),end=c(2017,7),freq=12)

## Warning in window.default(x, ...): 'start' value not changed

plot(sa,main="San Antonio MSE price growth, 2000-2017")  
lines(sap,col='blue')  
legend("topleft",legend=c("Estimation Space","Prediction Space"),text.col=c("black","blue"),bty="n")



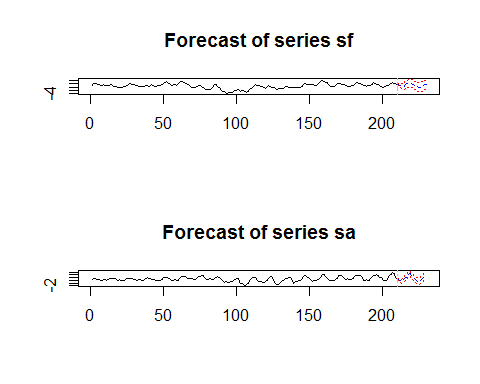
tsdisplay(sa,main="San Antonio MSA growth")



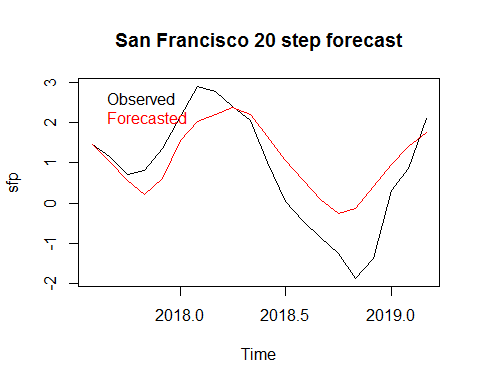
A qualitative asessment of the series shows a persistent seasonal component with a relatively linear trend and possibly an AR process. THe strength of the seasonal component seems to increase with time, as evidence by the increase in variance between 2000-2005 and 2008-2010, for instance. The drop in price growth right before 2010 can be interpreted in the context of subject matter as being affected by the 2008 recession.

1. From here, we use this MSA and compare it against the San Francisco MSA in a VAR model with an optimally structured lag, which we will use to construct a 20 step ahead forecast. Afterwards, we inspect for Granger causality and conclude with some remarks on the forecast.

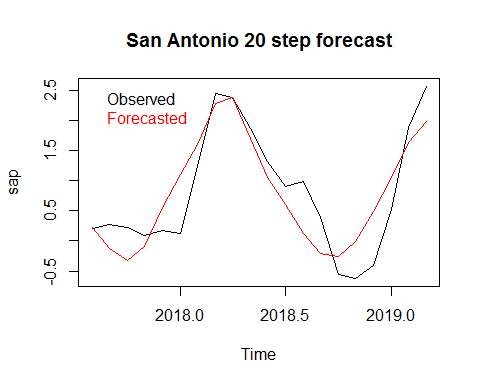
y2 = cbind(sf,sa)  
y\_tot2 = data.frame(y2)  
bestP <- VAR.select(y\_tot2,ic="sc",pmax=20) #optimal lag = 10  
  
m3 <- VAR(y\_tot2,p=10)  
m3.predict <- predict(m3,n.ahead=20)  
plot(m3.predict)



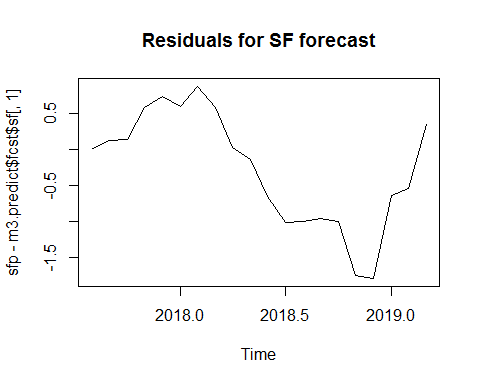
plot(sfp,main="San Francisco 20 step forecast")  
lines(ts(m3.predict$fcst$sf[,1],start=c(2017,8),freq=12),col='red')  
legend("topleft",legend=c("Observed","Forecasted"),text.col=c("black","red"),bty="n")



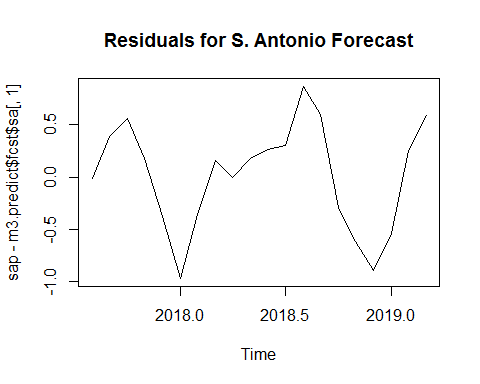
plot(sap,main="San Antonio 20 step forecast")  
lines(ts(m3.predict$fcst$sa[,1],start=c(2017,8),freq=12),col='red')  
legend("topleft",legend=c("Observed","Forecasted"),text.col=c("black","red"),bty="n")



plot(sfp-m3.predict$fcst$sf[,1],main="Residuals for SF forecast")



plot(sap-m3.predict$fcst$sa[,1],main="Residuals for S. Antonio Forecast")



#residual construction (MSE)  
sf\_res<- MSE(m3.predict$fcst$sf[,1],sfp)  
sa\_res<- MSE(m3.predict$fcst$sa[,1],sap)  
"MSE: San Francisco, San Antonio"

## [1] "MSE: San Francisco, San Antonio"

sf\_res

## [1] 0.692869

sa\_res

## [1] 0.250003

Overall, the forecasted data is roughly consistent with the observed prediction space from a qualitatitve level, but analysis of the residuals for both forecasts reveals that the VAR forecast did not capture all dynamics within the data. This is apparent in the residual plot of the forecast for San Antonio home price growth, wherein residuals seem to fluctuate seasonally year-yo-year. Overall, while sufficient at capturing the inter-data dynamic, the VAR model alone does not account for the signal.

Finally, we move the granger analysis:

grangertest(sf,sa)

## Granger causality test  
##   
## Model 1: sa ~ Lags(sa, 1:1) + Lags(sf, 1:1)  
## Model 2: sa ~ Lags(sa, 1:1)  
## Res.Df Df F Pr(>F)   
## 1 206   
## 2 207 -1 4.4876 0.03534 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

grangertest(sa,sf)

## Granger causality test  
##   
## Model 1: sf ~ Lags(sf, 1:1) + Lags(sa, 1:1)  
## Model 2: sf ~ Lags(sf, 1:1)  
## Res.Df Df F Pr(>F)   
## 1 206   
## 2 207 -1 2.9489 0.08744 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The granger causality test leads us to reject the null hypothesis that San Francisco home prices do not explain the variance in San Antonio price growth, and therefore have evidence to conclude that San Francisco MSA growth granger-causes San Antonio’s MSA growth. The converse, however, is not true, and there is no evidence to accept the alternative hypothesis that San Antonio granger-causes San Francisco home price growth.