STAT 429 Final Project: Analyzing Bacon Prices

Max Candocia

December 9, 2013

Introduction

The dataset I have chosen is a time series of bacon prices in Chicago in the early-mid twentieth century, found on the data marketplace. The reason I chose this is because I really like bacon, and it frustrates me when it's too expensive. However, the causes that relate to it more recently have to do with a bad harvest (from bad weather) in 2012 and a recent outbreak in a virus that affects piglets. Since it's impractical to use the time series we've applied so far to estimate current prices, the dataset I have will suffice for a general analysis of prices. The original units of the values of the dataset at http://data.is/188XaQt are in cents per pound of bacon, and each time point represents one month.

Detrending and Deseasoning

There are a few things that I did to this dataset to make it easier to analyze. Firstly, I only took a subset of the data. As you can see below in Figure 1, there appears to be a change in the pattern of prices post-World War I. Looking at the ACF also in Figure 1, it is apparent that this dataset is nonstationary and needs to be differenced. Therefore, from this point on, the data should be interpreted as the *change* in the price of bacon, $\frac{cents}{lb * month}$. I also removed some constant data padded on at the end of the time series.

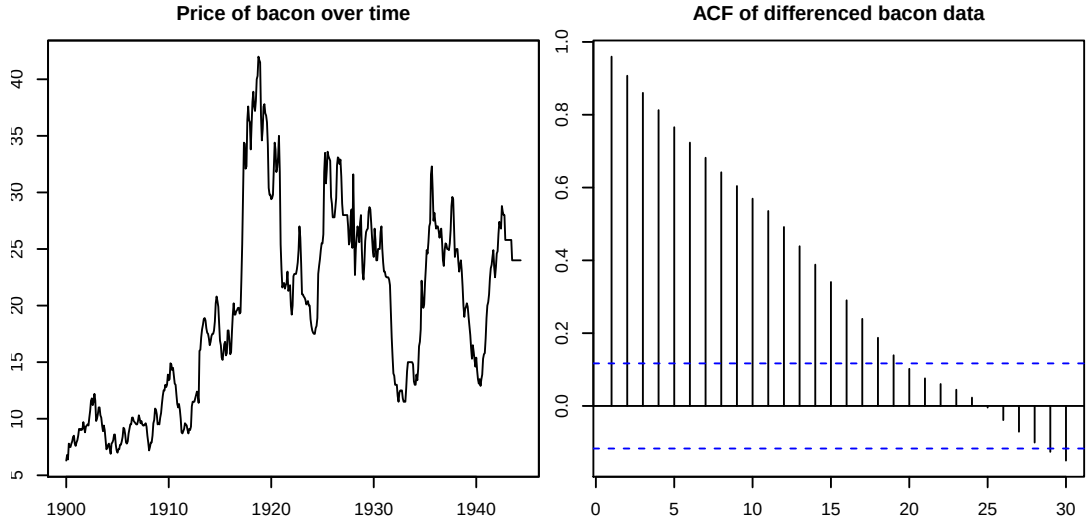

Figure 1: Left: original time series. Right side: ACF of truncated time series.

I also made a cut so that the data set, which is originally 532 points (after differencing), is now 281 points, starting in June of 1921. After applying the difference, the dataset is at 280 points, which is a relatively composite number. As for seasonality, I noticed a persistent trend in the ACF over large lags with values of roughly 20.

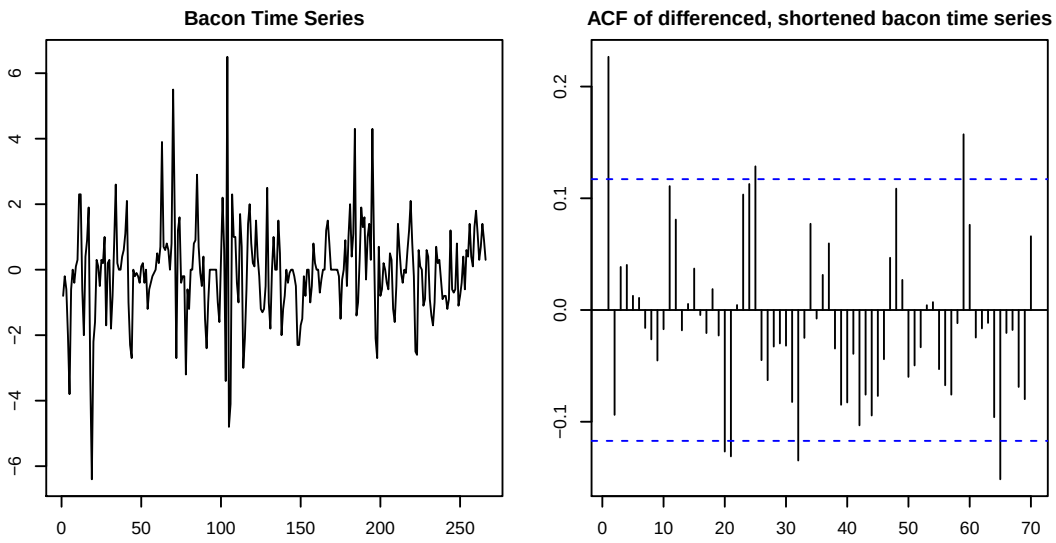

Figure 2: Left: Differenced time series. Right: ACF of differenced time series

I played around with a few different values for season lengths, and I noticed that a biennial seasonality (lag=24) had some effects. I tested different pure seasonal models, with P and Q ranging from 0 to 2, each, and I noticed that the only improvement in BIC is for the $P = 1, Q = 1$ model, which only has an improvement in BIC over the original model by a value less than 1. I decided to not control for seasonal effects.

I also looked for a linear trend to the series, and after performing a simple linear regression, none of the coefficients were close to being significant, so I decided not to do any detrending, which would needlessly complicate the model.

Choosing the best ARMA model

In addition to looking at the ACF seen before, it helps to look at the PACF, as seen in Figure 3. Looking at both the ACF and PACF, the one significant lag for the ACF and the two significant lags for the PACF imply that the model is probably an MA(1) process.

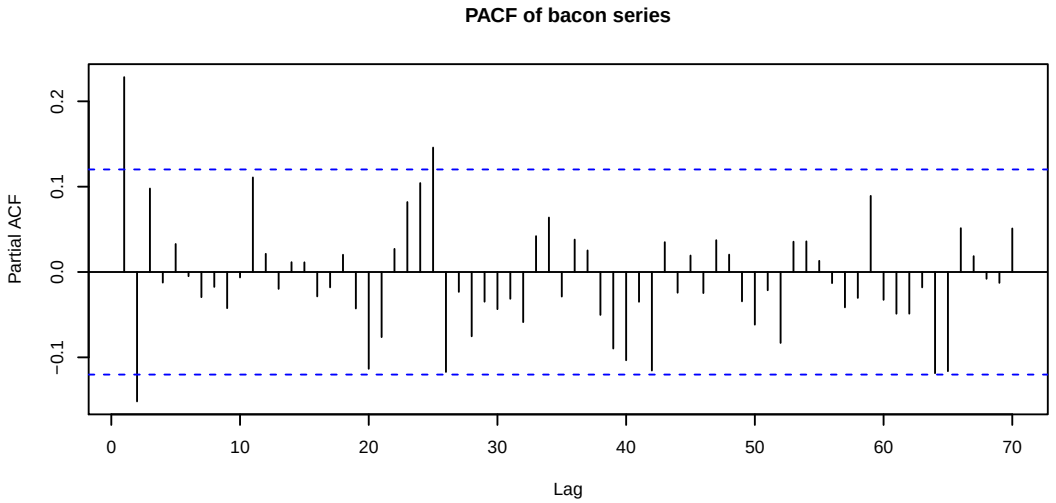

Figure 3: PACF of the differenced, shortened time series

I model 16 different ARMA(p,q) models, such that $\{(p, q) \in \mathbb{Z} : 0 \leq p, q \leq 3\}$, and then I compare the different models based on their AIC, BIC, and AICc. The results of these tests can be seen below in Figure 4.

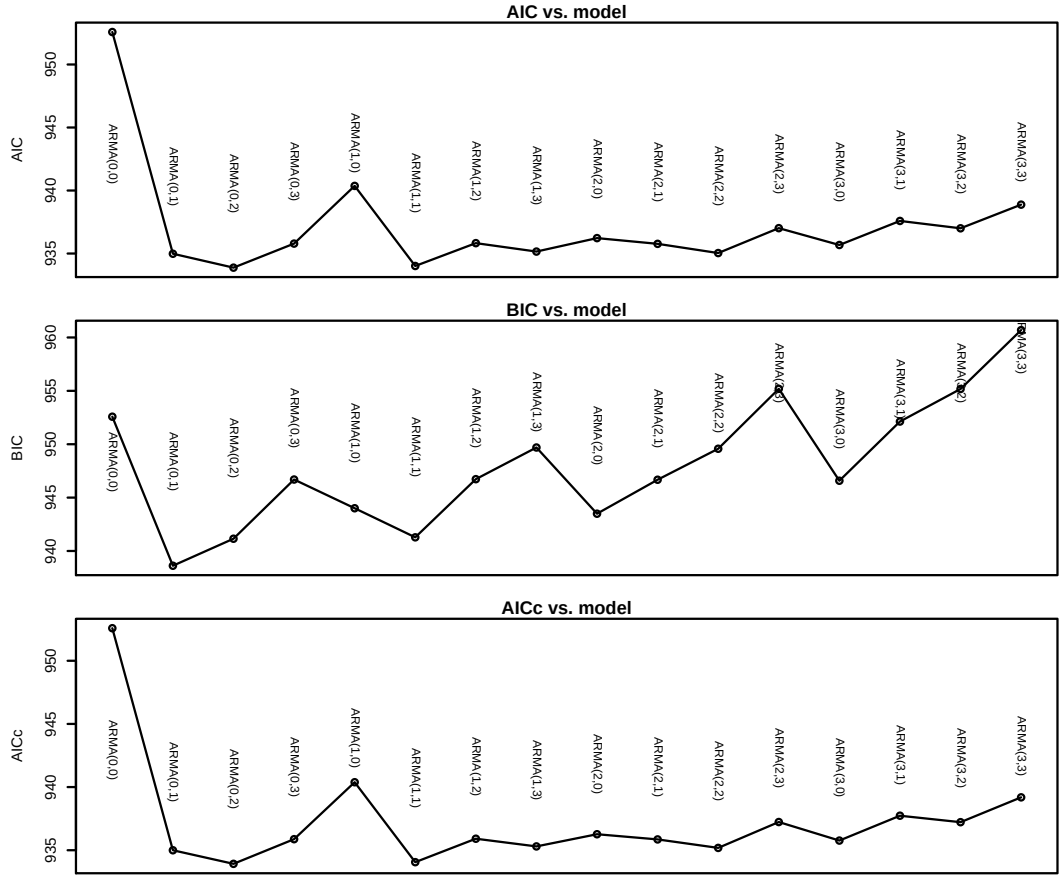

Figure 4: Stacked plots of the AIC, BIC, and AICc of the different ARMA fits for the bacon time series.

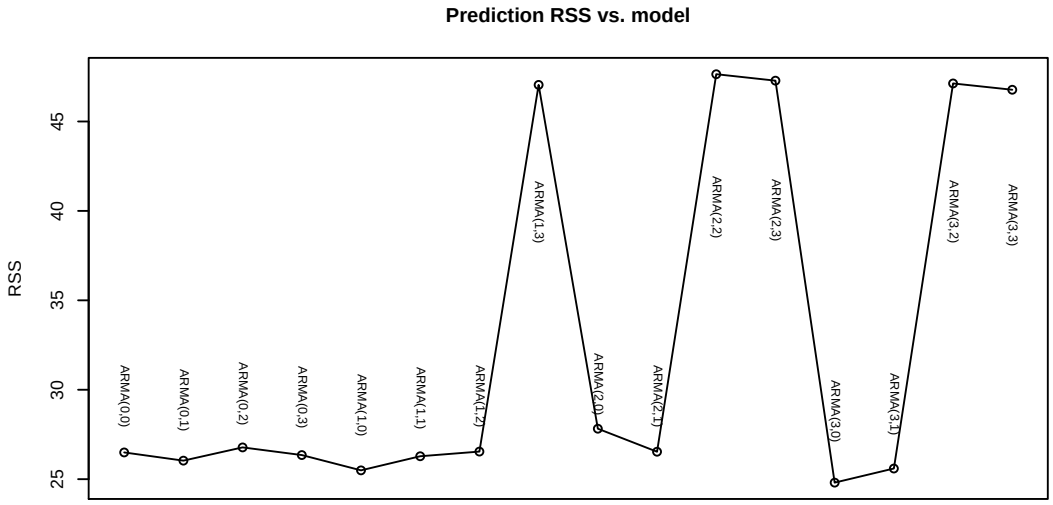

Figure 5: Residual sum of squares for predctions originating from different ARMA models. The first 95% of the data is used to train the model, which is tested with the remaining 5% of the data.

Above in Figure 5 you can see the different residual sums of squares from different models used to predict the last 5%, or 14, data points in the series. It appears as if there seems to be an overfitting issue when there are too many

AR *and* MA terms. The three best models appear to be MA(1) ($\theta = 0.3129$), MA(2) ($\theta_1 = 0.2819, \theta_2 = -0.1046$), and ARMA(1,1) ($\phi = -0.2923, \theta = 0.5764$). I find that the improvement of the MA(1) BIC being greater than the MA(2) model's improvement of AIC is enough grounds to choose it, since the RSS from the forecast is based on very few points because MA models quickly decay, and the coefficient of the MA(1) model, $\theta = 0.3129$, is much more significant than any of the other coefficients from the other models. Using this model, the predictions with 95% confidence bands are plotted below alongside the actual values.

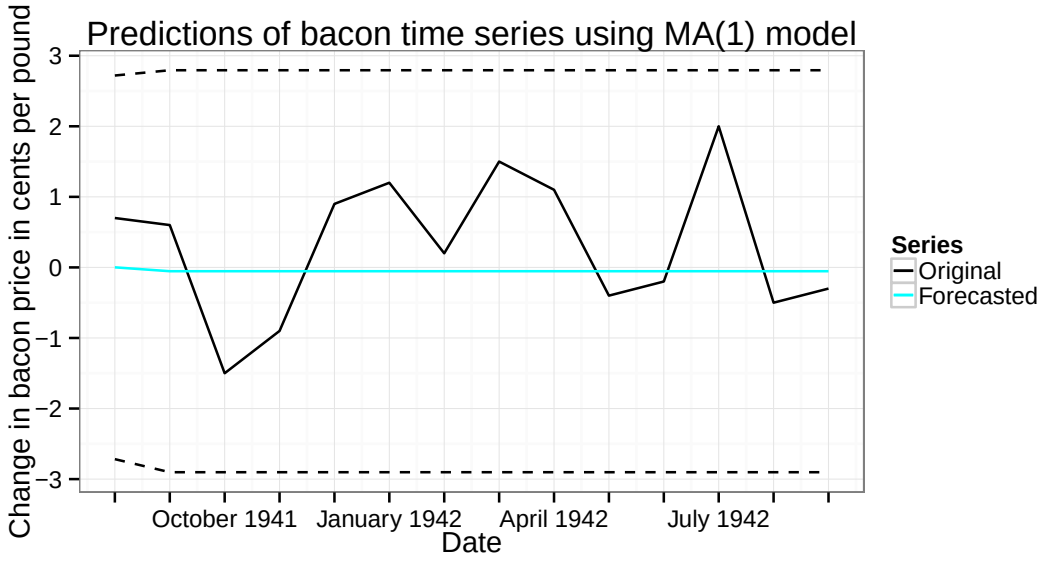

Figure 6: End of model with 95% confidence bands alongside actual data.

Spectral Analysis

The last part of the analysis regards analyzing the frequency composition. Before I compute the estimate, I consider the different theoretical spectral densities based off the MA(1), MA(2), and ARMA(1,1) models. Plugging in the coefficients from the sample models, one would expect the best model to most closely match its corresponding figure below.

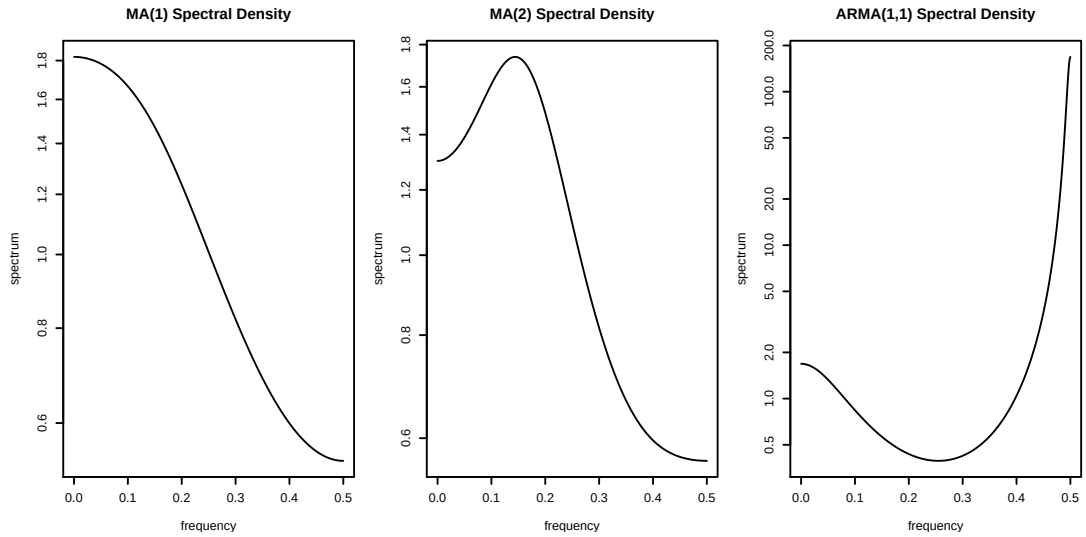

Figure 7: Theoretical spectral densities of the best models using the various information criteria

Using four different bandwidths with Modified Daniell kernels, I will estimate the spectral density.

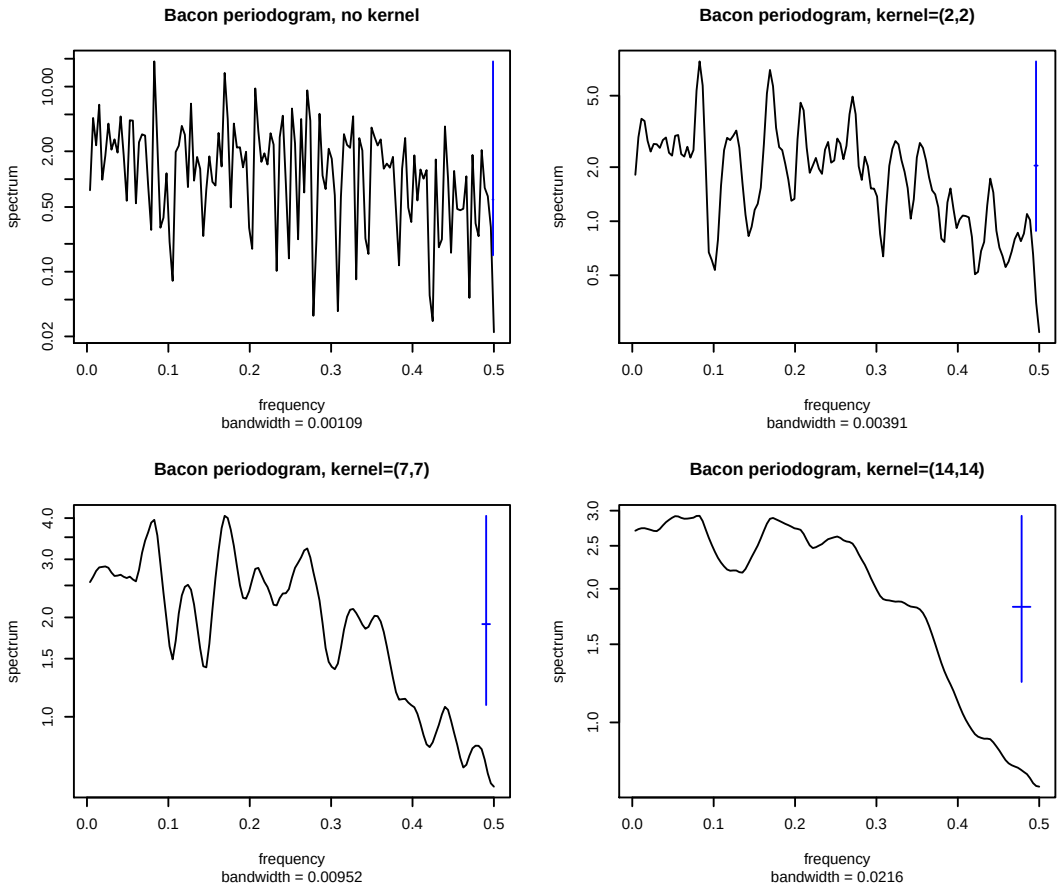

Figure 8: Various periodograms for the bacon time series using Modified Daniell kernels

Looking at the periodograms, it appears that after much smoothing, the periodogram with the (14,14) kernel looks similar to the one predicted by the theoretical model for the MA(1) process. It also shares some similarity to the MA(2) process, but it looks nothing like the ARMA(1,1) process.

Conclusions

The bacon data set is somewhat interesting. Apart from it needing to be differenced, it is described by a relatively simple process, even if there appear to be imperfections. While the AIC gave preference to the larger MA(2) and ARMA(1,1) models, the ACF, PACF, and BIC gave preference to the MA(1) model. Additionally, the heavily smoothed spectral density estimate indicated that the ARMA(1,1) model was not likely accurate.

Unfortunately this data is from the 1920s to 1940s (although thankfully from Chicago), so the model probably doesn't hold anymore. The fact that the behavior changed wildly after World War I supports this, and chaotic current events do a much better job of explaining the prices than the MA(1) model does.

R Code

```
> #code used to create this file (outside of the .Rnw file)
> filename="STAT429Final"
> Sweatex<-function(filename,extension='Rnw',command='pdflatex',silent=FALSE,preview=FALSE)
+ {
+ if (command=='latex') command='simpdftex latex --maxpfb'
+ extension<-paste('.',extension,sep='')
+ path=options('latexcmd')[[1]]
+ path=substr(path,start=1,stop=nchar(path)-5)
+ Sweave(paste(filename,extension,sep=''))
+ system(paste(path,command,' ',filename,sep=''),intern=silent)
+ if (preview)
+ {
+ system(paste(options('pdfviewer')[[1]],' ',filename,'.pdf',sep=''))
+ }
+ }
> setwd("C:/Users/maxcan2/Desktop/Dropbox/STAT 429")
> pdf.options(pointsize=7)
> #Sweatex(filename) #I have to comment out this line otherwise
> #it will try making this document an infinite number of times

> #code used in the header of the file in order to get results
>
> require(forecast)
> require(astsa)
> require(TSA)
> require(ggplot2)
> require(rdatamarket)
> require(MASS)
> #STAT 429 Final Project
>
> #bacon data
> bacon_url='http://data.is/188XaQt'
> bacon_base=dmseries(bacon_url)
> bacon=bacon_base[233:(length(bacon_base)-20),]#composite: 281; changes to 280 after diff
> bacon_prediff=bacon
> bacon=diff(bacon)#needs to be differenced because of acf
> bacon=as.ts(bacon)
> #check for linear trend
> t=1:(length(bacon))
> bacon_lm=lm(bacon~t)
> summary(bacon_lm)#both coefficients found to be insignificant (p-values around 0.1-0.2)
> ###SEASONAL AREA
> N=length(bacon)
> blist=list()
> saics=numeric(9)
> sbics=numeric(9)
> for (P in 0:2)
+         for (Q in 0:2){
+                 n=P*3+Q+1
+                 blist[[n]]=arima(bacon,order=c(0,0,0),
+                 seasonal=list(order=c(P,0,Q),period=24))
+                 saics[n]=blist[[n]]$aic
+                 sbics[n]=saics[n]-2*(P+Q)+log(N)*(P+Q)
+ }
> bacons=bacon
> baconstr=bacons[1:266]
```

```
> modlist=list()
> aics=numeric(16)
> bics=numeric(16)
> aiccs=numeric(16)
> for (p in 0:3)
+ for (q in 0:3){
+         k=p+q
+         n=4*p+q+1
+         modlist[[n]]=Arima(baconstr,order=c(p,0,q))
+         aics[n]=modlist[[n]]$aic
+         bics[n]=aics[n]-2*(p+q)+log(N)*(p+q)
+         aiccs[n]=aics[n]+2*k*(k+1)/(N-k-1)
+ }
> orders=paste0("ARMA(",rep(0:3,rep(4,4)),",",rep(0:3,4),")")
> bacon_eacf=eacf(baconstr)#results fairly inconsistent with rest of model
> rss2=numeric(16)
> first=1:266
> modlist2=list()
> for (p in 0:3)
+         for (q in 0:3){
+                 k=p+q
+                 n=4*p+q+1
+                 modlist2[[n]]=Arima(baconstr,order=c(p,0,q))
+
+
+                 fc=forecast.Arima(modlist2[[n]],14)
+                 rss2[n]=sum((bacons[last]-as.numeric(fc$mean)))^2
+ }

> #HERE IS THE CODE INSIDE OF THE SWEAVE FILE WHICH IS USED FOR OUTPUT
> #Fig. 1
> par(mfrow=c(1,2),mai=c(0.3,0.2,0.3,0.1))
> plot(bacon_base,main="Price of bacon over time",xlab="Time",ylab="Price in cents per pound")
> acf(as.numeric(bacon_prediff),lag.max=30,main="ACF of differenced bacon data")
> par(mfrow=c(1,1))
> #Fig. 2
> par(mar=c(2,2,2,2)+0.1,mfrow=c(1,2))#YEAHHHHH
> plot(baconstr,main="Bacon Time Series",cex=0.4,type='l')
> acf(as.numeric(bacon),lag.max=70,main="ACF of differenced, shortened bacon time series")
> par(mar=c(5, 4, 4, 2) + 0.1,mfrow=c(1,1))#resets margin values
> #Fig. 3
> pacf(as.numeric(baconstr),lag.max=70,main="PACF of bacon series")
> #Fig. 4
> par(mfrow=c(3,1), mai = c(0.1, 0.4, 0.1, 0.1))
> sh=c(7,rep(0,15))#shift the first point down
> plot(aics,type='o',ylab="AIC",axes=FALSE,main="AIC vs. model",frame.plot=TRUE,xlab="")
> Axis(side=2, labels=TRUE)
> text(aics/2+mean(aics)/2+5-sh,labels=orders,cex=0.9,srt=270)
> plot(bics,type='o',ylab="BIC",axes=FALSE,main="BIC vs. model",frame.plot=TRUE,xlab="")
> Axis(side=2, labels=TRUE)
> text(bics/2+mean(bics)/2+5-sh,labels=orders,cex=0.9,srt=270)
> plot(aiccs,type='o',ylab="AICc",axes=FALSE,main="AICc vs. model",frame.plot=TRUE,xlab="")
> Axis(side=2, labels=TRUE)
> text(aiccs/2+mean(aiccs)/2+5-sh,labels=orders,cex=0.9,srt=270)
> par(mfrow=c(1,1))
> #Fig. 5
> plot(rss2,type='o',ylab="RSS",axes=FALSE,main="Prediction RSS vs. model"
+ ,frame.plot=TRUE,xlab="")
```

```
> Axis(side=2, labels=TRUE)
> text(rss2/2+mean(rss2)/2,labels=orders,cex=0.7,srt=270)
> #Fig. 6
> fc=forecast.Arima(modlist[[2]],14)#MA(1) forecast
> times=time(bacons)
> latter_times=times[last]
> fcd=as.data.frame(fc)
> cust_times=c("","","October 1941","","","January 1942","","","April 1942","","","July 1942","","")
> colnames(fcd)[c(1,4,5)]<-c("Est","Upper95","Lower95")
> print(
+ ggplot()+geom_line(aes(x=latter_times,y=bacons[last],color='black'))+
+ geom_line(aes(x=latter_times,y=fcd$Est,color='cyan'))+
+ geom_line(aes(x=latter_times,y=fcd$Lower95,linetype='dashed'))+
+ geom_line(aes(x=latter_times,y=fcd$Upper95,linetype='dashed'))+
+ guides(alpha="none",fill="none",color=guide_legend(title="Series"),linetype='none')+
+ scale_color_identity("Series",guide="legend",labels=c("Original","Forecasted"))+
+ xlab("Date")+ylab("Change in bacon price in cents per pound")+
+ ggtitle("Predictions of bacon time series using MA(1) model")+
+ scale_x_continuous(breaks=latter_times,labels=cust_times)+theme_bw()+
+ scale_linetype_identity()
+ )
> #Fig. 7
> par(mfrow=c(1,3),mai=c(0.3, 0.3, 0.3, 0.1))
> co1=coef(modlist[[2]])
> co2=coef(modlist[[3]])
> co3=coef(modlist[[6]])
> arma.spec(co1,main="MA(1) Spectral Density")
> arma.spec(co2,main="MA(2) Spectral Density")
> arma.spec(co3,main="ARMA(1,1) Spectral Density")
> par(mfrow=c(1,1))
> #Fig. 8
> bts=as.numeric(baconstr)#used to fix the frequency axis
> par(mfrow=c(2,2))
> spec.pgram(bts,taper=0.0,fast=FALSE,main="Bacon periodogram, no kernel")#it's already composite enough
> spec.pgram(bts,taper=0.0,spans=c(2,2),fast=FALSE,main="Bacon periodogram, kernel=(2,2)")
> spec.pgram(bts,taper=0.0,spans=c(7,7),fast=FALSE,main="Bacon periodogram, kernel=(7,7)")
> spec.pgram(bts,taper=0.0,spans=c(14,14),fast=FALSE,main="Bacon periodogram, kernel=(14,14)")
> par(mfrow=c(1,1))
```