Fitting ARIMA Models For Sugar

Rolande Mbatchou

June 8, 2016

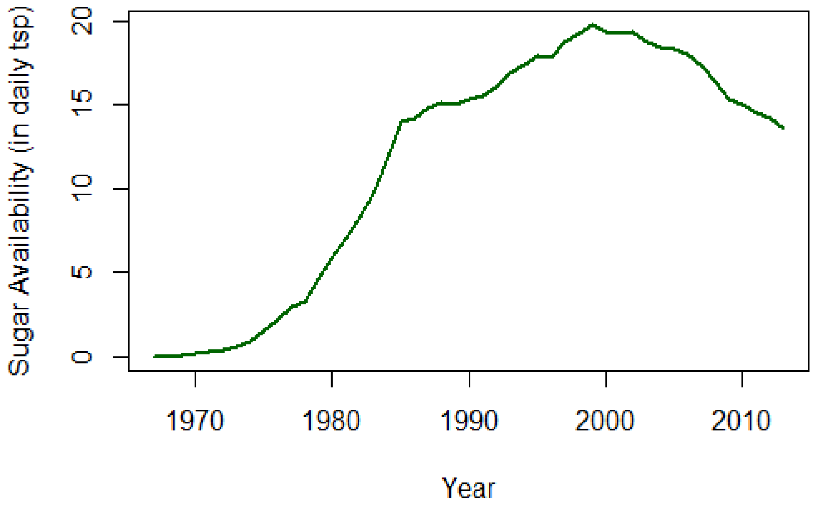
TABLE OF CONTENTS

INTRODUCTION

Excessive sugar consumption does not promote healthy living. Nonetheless, very few Americans consume sugar in recommended moderate amounts. In fact, worldwide, we consume about 500 extra calories a day from sugar. With all these shocking statistics and facts concerning sugar, I wanted to analyze whether the supply of sugar was now decreasing over time. The ERS Food Availability Data System (FADS) includes three distinct but related data series on food and nutrient availability for consumption. High Fructose Corn Syrup is one of the most concentrated source of sugar, thus, we focused our analysis on this commodity, as a substitute for sugar. Our dataset provide US yearly high fructose corn syrup availability for the years 1960 to 2013.

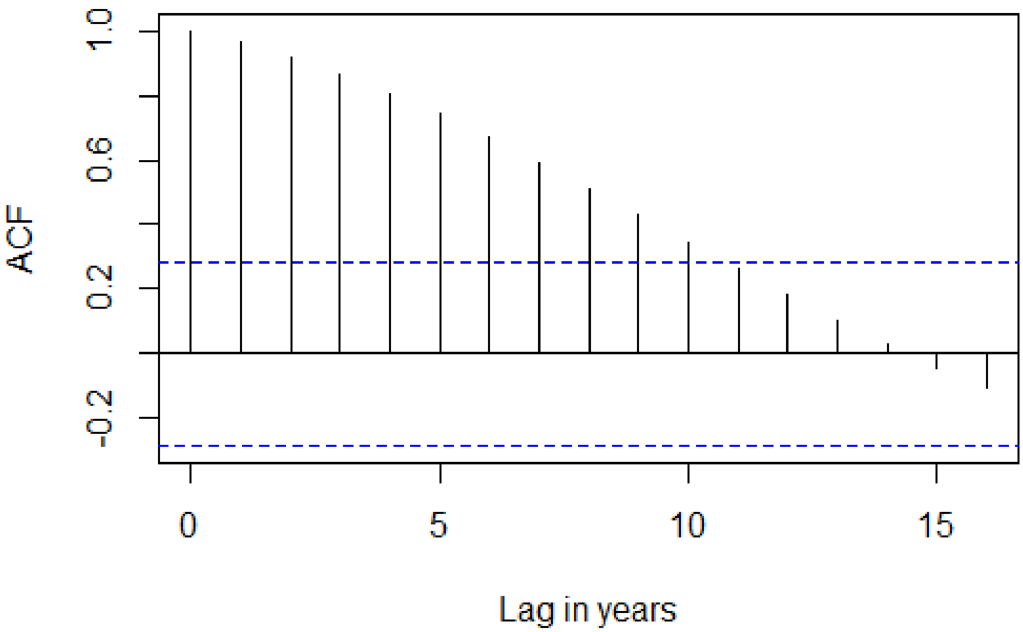
As we can see from Figure 1, the availability of high fructose corn syrup increased significantly from 1980 to 2000. Thereafter, the availability of the commodity decreased. Could we assume that it is due to increase awareness of the adverse effects of sugar? How could we best model this dataset to forecast future trends? We decided to use ARIMA models, as described in the "Time Series Analysis with Applications in R" text by Cryer and Chan, in order to determine a model fitted for our data.

Figure 1: US Yearly High Fructose Corn Syrup Availability



MODEL SPECIFICATION

Figure 2: Autocorrelation of US High Fructose Corn Syrup Availability



To specify a model for our data, we first begin examining the behavior of the autocorrelation factors at various yearly lags. We can see in Figure 2 that at later lags of (13-15), the autocorrelation factor decreased down to values close to zero, which is our desired outcome since we would like to work with a stationary time series. Nevertheless, due to the large amount of parameters it would require to reach a stationary process and since most of the autocorrelation factors were beyond the confidence interval of standard errors, we decided to take differences in order to reduce the amount of parameters used in the model. Concerning the type of model that could best fit our data, we recognized that the tailing off effect of the autocorrelation function most likely mean that our data followed an autoregressive process: **$Yt= \phi 1Y(t-1)+ \phi 2Y(t-2)+\ldots+ \phi pY(t\text{-}p) + \epsilon t$.**

Figure 3: Autocorrelation of Differenced Data

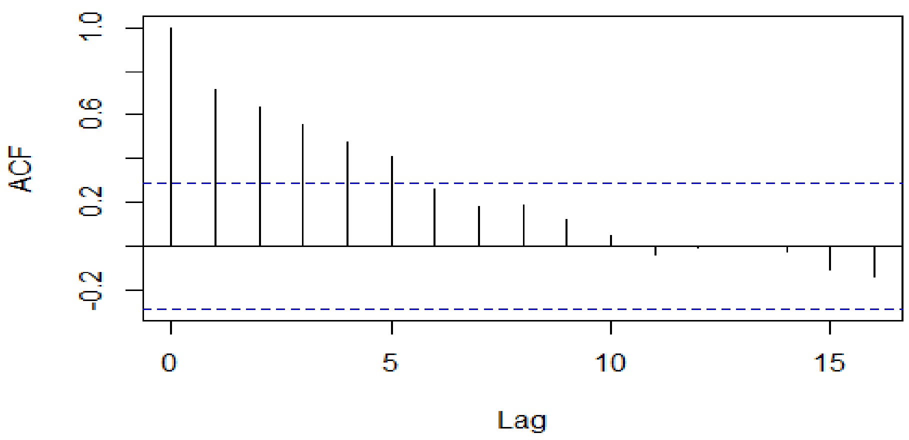
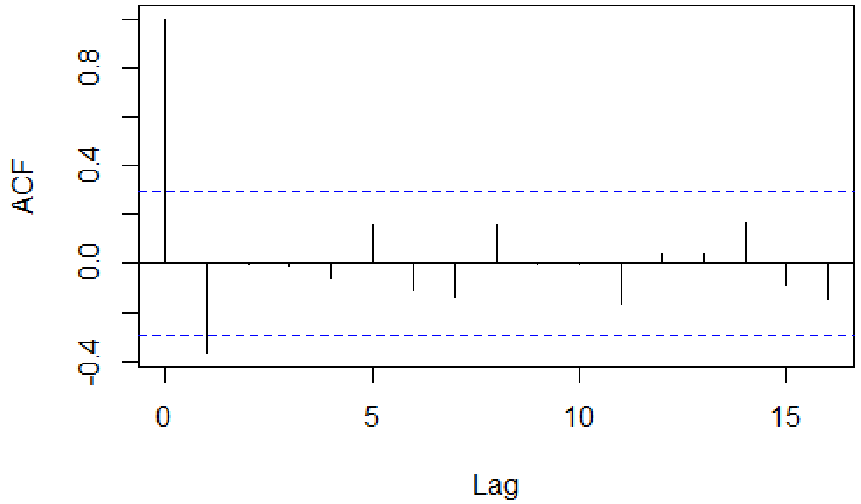


Figure 3 displays the autocorrelation of the first difference and we can see that many autocorrelation factors exceed the confidence interval bound, which is 2 times standard error.

Figure 4: Autocorrelation of 2nd Differenced Data



However, in Figure 4, we can see that most autocorrelation factors fall in between the confidence interval bounds. The only two values that are outside these bounds are lag 1 and lag2, which caused us to assume that an ARIMA(2,2,0) model would probably be the best model fit. To confirm our assumptions, we fit 2 models to our 2nd difference dataset: ARIMA(1,2,0) and ARIMA(2,2,0).

MODEL FITTING

From our model fitting, we were able to determine the parameters, along with the intercept values for ARIMA(1,2,0) and ARIMA(2,2,0) models, respectively (Table 1 and 2).

$W(t)= -.0123 -.3607W(t-1) +\epsilon t$, under ARIMA(1,2,0):

Table 1

```
## arima(x = sec_dif, order = c(1, 0, 0))
##
## Coefficients:
##          ar1      intercept
##       -0.3607     -0.0123
## s.e.  0.1373      0.0531
##
## sigma^2 estimated as 0.2321:  log likelihood = -31.06,  aic = 68.12

## [1] -0.62976009 -0.09154645
```

$W(t)= -.0124 -.4165W(t-1) -.1502W(t-2) +\epsilon t$, under ARIMA(2,2,0):

Table 2

```
## arima(x = sec_dif, order = c(2, 0, 0))
##
## Coefficients:
##          ar1     ar2  intercept
##      -0.4165  -0.1502    -0.0124
## s.e.   0.1462   0.1446     0.0457
##
## sigma^2 estimated as 0.2264:  log likelihood = -30.53,  aic = 69.05

## [1] -0.7030861 -0.1298659

## [1] -0.4336917  0.1332109
```

MODEL DIAGNOSTIC

To determine which of our two ARIMA models was the best fit, we analyzed the residuals and came up with two main hypothesis:

Ho: corr(es,et) =0 (data are also independently distributed)

H1: corr(es,et) ≠ 0 (residual exhibit serial corroletion)

We analyzed whether any of the autocorrelations were different from zero by using Box-Pierce Q-statistic and correlograms. We also determined whether our residuals were normally distributed by using q-q plots.

We can see from the Box-Ljung's p-values below that both models do NOT reject the null hypothesis (Ho). Thus, we can assume that the residuals are independently distributed and follow a white noise process.

```
##  Box-Ljung test for ARIMA(1,2,0)
## data:  AR1$residuals
## X-squared = 6.7032, df = 10, p-value = 0.7531

##  Box-Ljung test for ARIMA(2,2,0)
## data:  AR2$residuals
## X-squared = 4.2369, df = 10, p-value = 0.936
```

The correlograms in Figure 6 and 8, for both ARIMA models, also confirm that we should NOT reject the null hypothesis as residual values are all closed to zero.

Figure 6: ACF of residuals from the ARIMA(1,2,0) Model

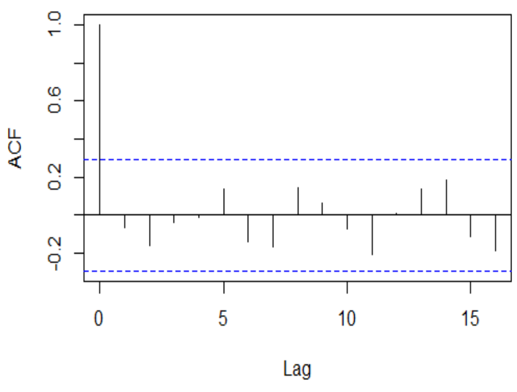
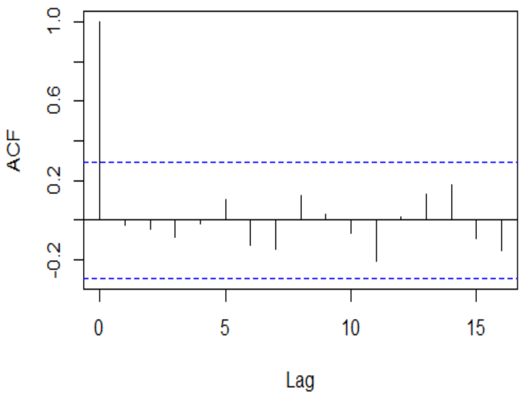


Figure 8: ACF of residuals from the ARIMA(2,2,0) Model



However, when examining the q-q plots for normality in Figure 5 and 7, for ARIMA(1,2,0) and ARIMA(2,2,0) respectively, we realized that the latter was a better fit. In fact, the q-q plot for ARIMA(2,2,0) showed that the residuals followed a normal distribution as they were closer to the regression line than under ARIMA(1,2,0), which we rejected.

Figure 5: Normal Quantile plot of ARIMA(1,2,0) Residuals

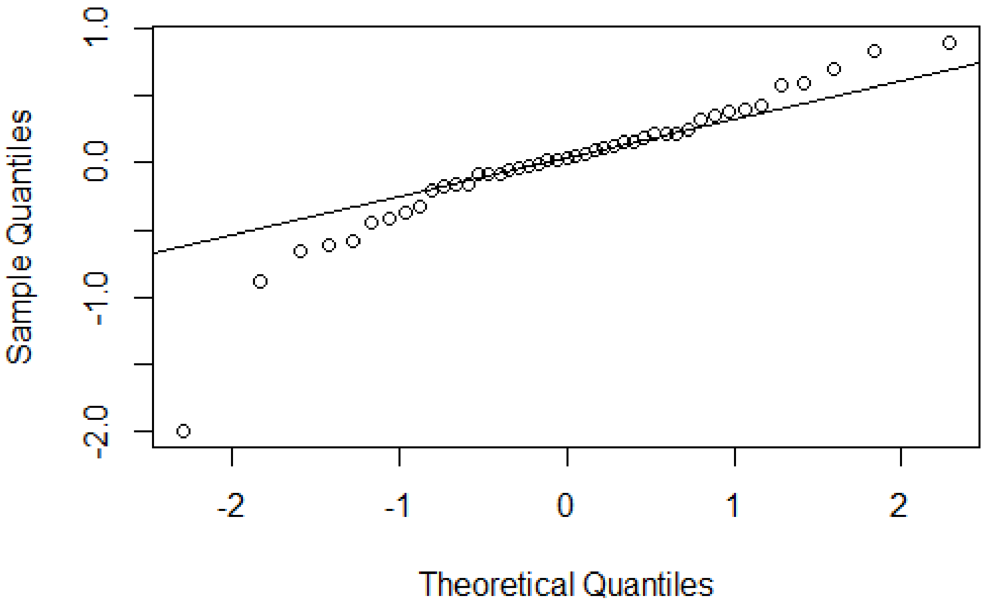
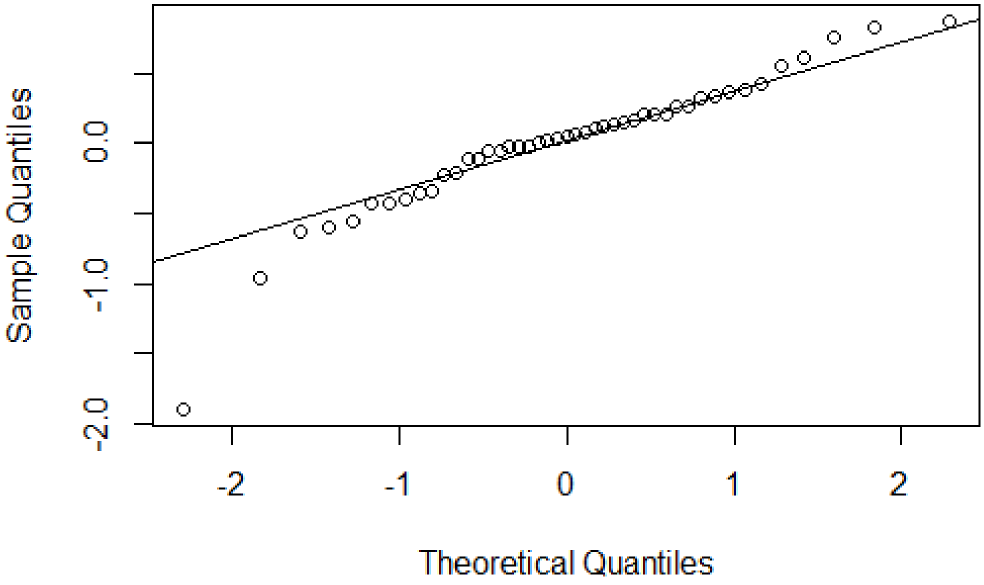
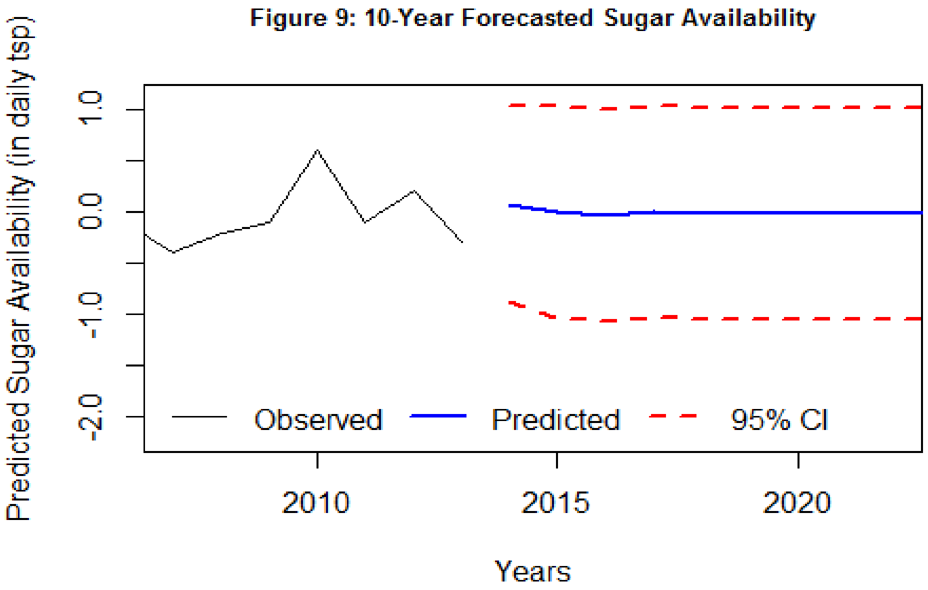


Figure 7: Normal Quantile plot of ARIMA(2,2,0) Residuals



FORECASTING & CONCLUSION



Using our selected ARIMA(2,2,0) model, we can see that sugar availability is forecasted to decrease slightly before stabilizing in the next couple of years. The confidence intervals show that the accuracy of the forecasts remains constant over time. Overall, we can conclude that ARIMA(2,2,0) is an appropriate model for our data.

BIBLIOGRAPHY

Cryer, Jonathan D and Kung-Sik Chan. Time Series Analysis with Applications in R. 2nd Edition. New York: Springer Science & Business Media, LLC, 2010. Textbook.

Data Source: http://www.ers.usda.gov/data-products/food-availability-(per-capita)-data-system/.aspx, USDA Website