A2

Zoe Zhou

29/03/2020

### Question 1. Use the Holt-Winters technique to build a model.

(HW.CO2 = HoltWinters(reduced.CO2.ts))

## Holt-Winters exponential smoothing with trend and additive seasonal component.  
##   
## Call:  
## HoltWinters(x = reduced.CO2.ts)  
##   
## Smoothing parameters:  
## alpha: 0.9267355  
## beta : 0.0813906  
## gamma: 1  
##   
## Coefficients:  
## [,1]  
## a 405.1011655  
## b 0.5889777  
## s1 0.2311236  
## s2 -0.2791367  
## s3 -0.2150001  
## s4 0.4588345

#### Prediction

# Predict for the next 4 quarters  
(HW.CO2.pred = predict(HW.CO2, n.ahead = 4))

## Qtr1 Qtr2 Qtr3 Qtr4  
## 2018 405.9213  
## 2019 406.0000 406.6531 407.9159

# Compare to actual values  
(actual = CO2.ts[(length(CO2.ts) - 3):length(CO2.ts)])

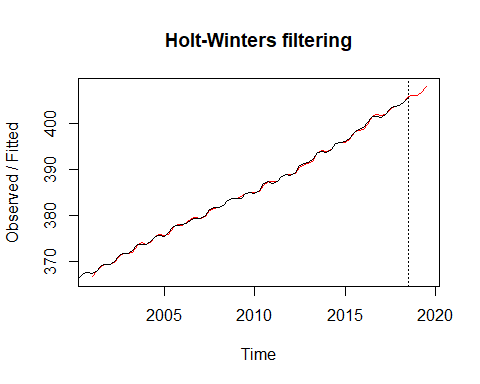
## [1] 405.83 405.73 406.71 408.25

#### RMSEP

(RMSEP.HW.pred = sqrt(1/4 \* sum((actual - HW.CO2.pred) ^ 2)))

## [1] 0.2214015

plot(HW.CO2, HW.CO2.pred)



#### Comment:

The model did a pretty good job at prediction as we can see very small differences between the predicted and actual line. The predicted values are very close to the actual values. The prediction error is pretty small (0.2214 ppm).

Model answer:

The plot of the Holt-Winters model shows an excellent fit with the model (red line) being very close to the actual observations (black line). There is a slightly higher peak in 2016.

The predictions are above the actual values for 2018.4 and 2019.1 and below for 2019.2 and 2019.3.

The RMSEP shows that, on average, the prediction error is 0.22 ppm.

### Question 2. Using the de-seasonalising techniques (Moving Averages and Seasonal Trend Lowess), build Seasonally Adjusted models of the data (2000 to 2018.3).

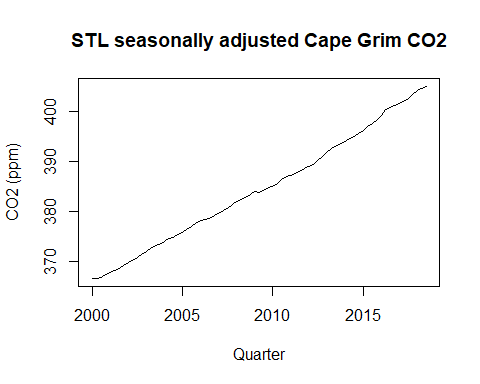
#### Seasonal Tread Lowess  
# Seasonal estimates  
decomp.stl.CO2$time.series[1:4, 1]

## [1] -0.3906880 -0.3028320 0.4787971 0.2147230

#### Comment:

The seasonal estimates from stl decomposition are showed above. We can see that, on average Quarter 1 has the smallest value of CO2 concentration while Quarter 3 has the largest.

# Subtract seasonal component  
stl.sa.CO2.ts = reduced.CO2.ts - decomp.stl.CO2$time.series[, 1]  
plot(stl.sa.CO2.ts, main = "STL seasonally adjusted Cape Grim CO2",  
 xlab = "Quarter", ylab = "CO2 (ppm)")



#### Comment:

After removing the seasonal components, we can see a clear positive linear trend. Interestingly there are some small bumps in the data. We possibly have a “break” - change in slope somehwere around 2010.

The seasonally adjusted plot is reasonably linear but the slope increases around time point 50 (2012.2).

#### Fit the best predicting model.

stl.CO2.fit3 = lm(stl.sa.CO2.ts[-1] ~ Time[-1] + Time.break[-1] + stl.sa.CO2.ts[-75])  
summary(stl.CO2.fit3)

##   
## Call:  
## lm(formula = stl.sa.CO2.ts[-1] ~ Time[-1] + Time.break[-1] +   
## stl.sa.CO2.ts[-75])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.51687 -0.13691 0.01665 0.11813 0.51237   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 113.68138 29.12337 3.903 0.000216 \*\*\*  
## Time[-1] 0.15131 0.03826 3.955 0.000181 \*\*\*  
## Time.break[-1] 0.04304 0.01189 3.620 0.000554 \*\*\*  
## stl.sa.CO2.ts[-75] 0.68994 0.07972 8.654 1.14e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.189 on 70 degrees of freedom  
## Multiple R-squared: 0.9997, Adjusted R-squared: 0.9997   
## F-statistic: 8.543e+04 on 3 and 70 DF, p-value: < 2.2e-16

#### Prediction

t76.sa.pred = stl.CO2.fit3$coef[1] + stl.CO2.fit3$coef[2] \* 76 +  
 stl.CO2.fit3$coef[3] \* (76 - 50) + stl.CO2.fit3$coef[4] \* stl.sa.CO2.ts[76 - 1]  
t76.pred = t76.sa.pred + decomp.stl.CO2$time.series[4, 1]  
  
t77.sa.pred = stl.CO2.fit3$coef[1] + stl.CO2.fit3$coef[2] \* 77 +  
 stl.CO2.fit3$coef[3] \* (77 - 50) + stl.CO2.fit3$coef[4] \* t76.sa.pred  
t77.pred = t77.sa.pred + decomp.stl.CO2$time.series[1, 1]  
  
  
t78.sa.pred = stl.CO2.fit3$coef[1] + stl.CO2.fit3$coef[2] \* 78 +  
 stl.CO2.fit3$coef[3] \* (78 - 50) + stl.CO2.fit3$coef[4] \* t77.sa.pred  
t78.pred = t78.sa.pred + decomp.stl.CO2$time.series[2, 1]  
  
  
t79.sa.pred = stl.CO2.fit3$coef[1] + stl.CO2.fit3$coef[2] \* 79 +  
 stl.CO2.fit3$coef[3] \* (79 - 50) + stl.CO2.fit3$coef[4] \* t78.sa.pred  
t79.pred = t79.sa.pred + decomp.stl.CO2$time.series[3, 1]  
  
stl.pred = c(t76.pred, t77.pred, t78.pred, t79.pred)

#### Calculate RMSEP to compare actual values and predicted values

(RMSEP.stl.pred = sqrt(1/4 \* sum((actual - stl.pred) ^ 2)))

## [1] 0.1951761

### Question 3. Technical Notes.

The seasonal estimates show that for the last 2 quarters, the CO2 concentration is above the overall trend with Quarter 3 being the largest(0.48 ppm). And for the first 2 quarters, the CO2 concentration is below the overall trend with Quarter 1 being the lowest (-0.39 ppm).

The seasonal estimates show that the CO2 concentration is below the overall trend for the first 2 quarters with Quarter 1 being the lowest (-0.39) and above the overall trend in the last 2 quarters with Quarter 3 being the highest (0.48).

The plot of the seasonally adjusted data shows a clear positive linear trend. Interestingly there is a “break” - change in slope somehwere around 2010.

The plot of the seasonally adjusted series shows an increasing reasonably linear trend. There is a break in the trend at time point 50 (2012.2)

The final model included a Time variable for the linear trend, a Time break variable for the change in slope and a lagged response variable to deal with the autocorrelation found in the residual series.

The final model included a trend term, a trend break term and a lagged response to take care of autocorerelation detected in the Residual Series.

For the final model, the plot of the autocorrelation function of the residuals still shows lags 1, 11 and 16 are slightly significant. It is ok to ignore them as they are small.

The plot of the autocorrelation function of the residuals shows lags 1, 11 and 16 are weakly significant, but of no real concern.

The residual series shows no evidence of non-linearity. There is no evidence against the underlying errors having come from a noraml distribution (P-value = 0.834). So our residual series is normally distributed.

For the final model, the Residual Series appears to be reasonably random scatter about 0 with a slight positive trend for the first 2 – 3 years. There is a large negative residual for time period 38 (2009.2) and a large positive residual for time period 66 (2016.2).

The residuals appear to be normally distributed (Shapiro-Wilk P-value = 0.834) with isolated values at each end of the reasonably symmetric distribution due to the large residuals discussed above. The assumptions appear to be satisfied.

We have strong evidence against the hypothesis that the coefficient associated with the Time variable is 0 (P-value = 0.000181) and strong evidence that the coefficient associated with the Time.break variable is 0 (P-value = 0.000554). We have very strong evidence against the hypothesis of no autocorrelation (P-value ≈ 0).

All variables (time, time break and lagged response) in the model are highly significant with P-values almost close to 0. The F-statistic (P-value ≈ 0) also showed at least one of the variables is significantly important.

The F-statistic provides extremely strong evidence against the hypothesis that none of the variables are related to the seasonally adjusted CO2 concentration (P-value ≈ 0). The Multiple R 2 is almost 1 (0.9997) indicating that nearly all the variation in the seasonally adjusted CO2 concentration is explained by the model.

The Multiple R2 is 0.9997 indicating that nearly 99.97% the variation in the seasonally adjusted CO2 concentration is explained by the model. The model is a good fit. The residual standard error is 0.189 ppm so prediction intervals will be narrow. The model predictions can be relied on as the assumptions appear to be satisfied.

The RMSEP for the 2018.4-2019.3 predictions was 0.195 which was smaller than that for the moving average model (0.233).

The Residual Standard Error is 0.19 ppm so prediction intervals should be reasonably narrow. The model predictions can be relied on as the assumptions appear to be satisfied. The RMSEP for the predictions was 0.20 ppm which was smaller than that for the Moving Average model (0.23 ppm). Our predictions for 2018.4 and 2019.1-3 were: …

Predictions are 2018.4: 406.00 ppm, 2019.1: 406.07 ppm, 2019.2 406.82 ppm, 2019.3: 408.25 ppm.

### Question 4. Use full data and predict for the 4 quarters of 2019.4 to 2020.3.

CO2.fit4 = lm(sa.CO2.ts[-1] ~ Time.new[-1] + Time.break.new[-1] + sa.CO2.ts[-79])  
  
# Predict  
t80.sa.pred = CO2.fit4$coef[1] + CO2.fit4$coef[2] \* 80 +  
 CO2.fit4$coef[3] \* (80 - 50) + CO2.fit4$coef[4] \* sa.CO2.ts[79]  
  
t80.pred = t80.sa.pred + decomp.CO2$time.series[4, 1]  
  
t81.sa.pred = CO2.fit4$coef[1] + CO2.fit4$coef[2] \* 81 +  
 CO2.fit4$coef[3] \* (81 - 50) + CO2.fit4$coef[4] \* t80.sa.pred  
t81.pred = t81.sa.pred + decomp.CO2$time.series[1, 1]  
  
  
t82.sa.pred = CO2.fit4$coef[1] + CO2.fit4$coef[2] \* 82 +  
 CO2.fit4$coef[3] \* (82 - 50) + CO2.fit4$coef[4] \* t81.sa.pred  
t82.pred = t82.sa.pred + decomp.CO2$time.series[2, 1]  
  
  
t83.sa.pred = CO2.fit4$coef[1] + CO2.fit4$coef[2] \* 83 +  
 CO2.fit4$coef[3] \* (83 - 50) + CO2.fit4$coef[4] \* t82.sa.pred  
t83.pred = t83.sa.pred + decomp.CO2$time.series[3, 1]  
  
pred = c(t80.pred, t81.pred, t82.pred, t83.pred)  
  
names(pred) = c("2019.4","2020.1","2020.2","2020.3")  
  
pred

## 2019.4 2020.1 2020.2 2020.3   
## 408.5988 408.6075 409.3381 410.7559

summary(CO2.fit4)

##   
## Call:  
## lm(formula = sa.CO2.ts[-1] ~ Time.new[-1] + Time.break.new[-1] +   
## sa.CO2.ts[-79])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5294 -0.1417 0.0186 0.1271 0.5095   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 114.64012 28.54054 4.017 0.000140 \*\*\*  
## Time.new[-1] 0.15269 0.03754 4.067 0.000118 \*\*\*  
## Time.break.new[-1] 0.04254 0.01109 3.837 0.000260 \*\*\*  
## sa.CO2.ts[-79] 0.68731 0.07813 8.797 4.02e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1896 on 74 degrees of freedom  
## Multiple R-squared: 0.9998, Adjusted R-squared: 0.9997   
## F-statistic: 1.016e+05 on 3 and 74 DF, p-value: < 2.2e-16

#### Comment:

The model including the full data’s Multiple R-squared is nearly 1 (0.9998). The residual standard error is small (0.1896 ppm) so the prediction intervals will be very narrow. The predictions should be reliable.

The STL Seasonally Adjusted model using all the data to 2019.3 has almost exactly the same estimates to the original model.

The Multiple R-squared is 99.9 showing that almost all the variability is explained by the model.

The residual standard error is 0.19 ppm so prediction intervals should be narrow.

Our predictions for 2019.4 and 2020.1 – 3 are:

Quarter 4: 408.60 ppm

Quarter 1: 408.61 ppm

Quarter 2: 409.34 ppm

Quarter 3: 410.76 ppm