A3

Zoe Zhou

29/04/2020

Question 1. Build a Seasonal Factor model of the data (2000 to 2018.3).

```
summary(sf.CO2.fit)
```

```
##
## Call:
## lm(formula = reduced.CO2.ts[-1] ~ Time[-1] + Time.break[-1] +
##     Quarter[-1] + reduced.CO2.ts[-75])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48990 -0.12209 -0.00581  0.11306  0.53995
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         108.82449   29.39775   3.702 0.000435 ***
## Time[-1]              0.14548    0.03860   3.769 0.000349 ***
## Time.break[-1]        0.04182    0.01196   3.496 0.000843 ***
## Quarter[-1]2          0.43876    0.07593   5.778 2.14e-07 ***
## Quarter[-1]3          1.14763    0.07477  15.348  < 2e-16 ***
## Quarter[-1]4          0.41510    0.06543   6.344 2.22e-08 ***
## reduced.CO2.ts[-75]   0.70187    0.08043   8.727 1.18e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1889 on 67 degrees of freedom
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9997
## F-statistic: 4.282e+04 on 6 and 67 DF,  p-value: < 2.2e-16
```

Prediction

```
# Predict for T76 - 2018 Q4
t76.sf.pred = sf.CO2.fit$coef[1] + sf.CO2.fit$coef[2] * 76 +
          sf.CO2.fit$coef[3] * (76 - 50) + sf.CO2.fit$coef[6] +
  sf.CO2.fit$coef[7] * reduced.CO2.ts[75]
t76.sf.pred
```

```
## (Intercept)
##    406.0347
```

```
# Predict for T77 - 2019 Q1, baseline
t77.sf.pred = sf.CO2.fit$coef[1] + sf.CO2.fit$coef[2] * 77 +
          sf.CO2.fit$coef[3] * (77 - 50) + sf.CO2.fit$coef[7] * t76.sf.pred
t77.sf.pred
```

```
## (Intercept)
##    406.1401

# Predict for T78 - 2019 Q2
t78.sf.pred = sf.CO2.fit$coef[1] + sf.CO2.fit$coef[2] * 78 +
  sf.CO2.fit$coef[3] * (78 - 50) + sf.CO2.fit$coef[4] +
  sf.CO2.fit$coef[7] * t77.sf.pred
t78.sf.pred

## (Intercept)
##    406.8401

# Predict for T79 - 2019 Q3
t79.sf.pred = sf.CO2.fit$coef[1] + sf.CO2.fit$coef[2] * 79 +
  sf.CO2.fit$coef[3] * (79 - 50) + sf.CO2.fit$coef[5] +
  sf.CO2.fit$coef[7] * t78.sf.pred
t79.sf.pred

## (Intercept)
##    408.2276

sf.pred = c(t76.sf.pred, t77.sf.pred, t78.sf.pred, t79.sf.pred)
names(sf.pred) = c("2018.4", "2019.1", "2019.2", "2019.3")
sf.pred

##   2018.4   2019.1   2019.2   2019.3
## 406.0347 406.1401 406.8401 408.2276
```

Calculate RMSEP to compare actual values and predicted values

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

## [1] 0.2384888
```

Comment:

We have fitted a Time variable, a Time Break variable, a seasonal factor and a lagged response variable in the Seasonal Factor 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 plot of the autocorrelation function of the residuals shows lags 1, 11 and 16 are weakly significant, but of no real concern. The residuals appear to be normally distributed (Shapiro-Wilk P-value = 0.852) with isolated values at each end of the reasonably symmetric distribution due to the large residuals discussed above. The assumptions appear to be satisfied.

From the summary output of the model, we can see all terms are significant. Compared to the baseline level Quarter 1, the coefficients for Quarters 2 – 4 CO2 concentrations are all positive. This means they are all larger than Q1. Quarter 3 is the largest with a difference of 1.15 ppm.

The RMSEP was 0.238 ppm.

Question 2. Find the best predicting Harmonic model of the data (2000 to 2018.3).

```
fh.CO2.fit

##
## Call:
## lm(formula = reduced.CO2.ts[-1] ~ Time[-1] + Time.break[-1] +
##     c1[-1] + s1[-1] + c2[-1] + reduced.CO2.ts[-75])
##
## Coefficients:
##         (Intercept)             Time[-1]       Time.break[-1]
##           109.32486              0.14548              0.04182
##              c1[-1]               s1[-1]               c2[-1]
##            -0.01183             -0.57381             -0.07344
## reduced.CO2.ts[-75]
##             0.70187

# Predict for T76 - 2018 Q4
t76.fh.pred = fh.CO2.fit$coef[1] + fh.CO2.fit$coef[2] * 76 +
  fh.CO2.fit$coef[3] * (76 - 50) +
  fh.CO2.fit$coef[4] * cos(2*pi*76*(1/4)) + #c1
  fh.CO2.fit$coef[5] * sin(2*pi*76*(1/4)) + #s1
  fh.CO2.fit$coef[6] * cos(2*pi*76*(2/4)) + #c2
  fh.CO2.fit$coef[7] * reduced.CO2.ts[75]
t76.fh.pred

## (Intercept)
##    406.0347

# Predict for T77 - 2019 Q1, baseline
t77.fh.pred = fh.CO2.fit$coef[1] + fh.CO2.fit$coef[2] * 77 +
  fh.CO2.fit$coef[3] * (77 - 50) +
  fh.CO2.fit$coef[4] * cos(2*pi*77*(1/4)) + #c1
  fh.CO2.fit$coef[5] * sin(2*pi*77*(1/4)) + #s1
  fh.CO2.fit$coef[6] * cos(2*pi*77*(2/4)) + #c2
  fh.CO2.fit$coef[7] * t76.fh.pred
t77.fh.pred

## (Intercept)
##    406.1401

# Predict for T78 - 2019 Q2
t78.fh.pred = fh.CO2.fit$coef[1] + fh.CO2.fit$coef[2] * 78 +
  fh.CO2.fit$coef[3] * (78 - 50) +
  fh.CO2.fit$coef[4] * cos(2*pi*78*(1/4)) + #c1
  fh.CO2.fit$coef[5] * sin(2*pi*78*(1/4)) + #s1
  fh.CO2.fit$coef[6] * cos(2*pi*78*(2/4)) + #c2
  fh.CO2.fit$coef[7] * t77.fh.pred
t78.fh.pred
```

```
## (Intercept)
##    406.8401

# Predict for T79 - 2019 Q3
t79.fh.pred = fh.CO2.fit$coef[1] + fh.CO2.fit$coef[2] * 79 +
  fh.CO2.fit$coef[3] * (79 - 50) +
  fh.CO2.fit$coef[4] * cos(2*pi*79*(1/4)) + #c1
  fh.CO2.fit$coef[5] * sin(2*pi*79*(1/4)) + #s1
  fh.CO2.fit$coef[6] * cos(2*pi*79*(2/4)) + #c2
  fh.CO2.fit$coef[7] * t78.fh.pred
t79.fh.pred

## (Intercept)
##    408.2276

fh.pred = c(t76.fh.pred, t77.fh.pred, t78.fh.pred, t79.fh.pred)
names(fh.pred) = c("2018.4", "2019.1", "2019.2", "2019.3")
fh.pred

##   2018.4   2019.1   2019.2   2019.3
## 406.0347 406.1401 406.8401 408.2276

#### Calculate RMSEP to compare actual values and predicted values
(RMSEP.fh.pred = sqrt(1/4 * sum((actual - fh.pred) ^ 2)))

## [1] 0.2384888
```

Comment:

We find that, the Full Harmonic model was the best predicting Harmonic model. It had the smallest RMSEP (0.2384 ppm) of all Harmonic models. It is the same as the Seasonal Factor Model.

The Full Harmonic model included a Time variable, a Time Break variable, 3 harmonics while c1 with a P-value = 0.79 being non-significant and a lagged response variable. 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 plot of the autocorrelation function of the residuals shows lags 1, 11 and 16 are weakly significant, but of no real concern. The residuals appear to be normally distributed (Shapiro-Wilk P-value = 0.74) with isolated values at each end of the reasonably symmetric distribution due to the large residuals discussed above. The assumptions appear to be satisfied.

Comment about other models

We had fitted a single Cosine Model but it has the largest RMSEP = 0.2648 ppm.

In the summary of the full harmonic model, we find that the cosine harmonic with frequency 1/4 was not significant. So we dropped that term and fitted a Reduced Harmonic model. The summary were very similar to the Full Harmonic model. But the RMSEP of the Reduced Harmonic model was higher than the Full Harmonic Model at 0.244 ppm.

We did not fit a model removing pairs of harmonics of the same frequency when both are not significant. Because it will be the same as the Full Harmonic model.

Question 3. Technical Notes.

The Seasonal Factor model included a Time variable, a Time Break variable, a seasonal factor and a lagged response variable to take care of autocorrelation.

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 plot of the autocorrelation function of the residuals shows lags 1, 11 and 16 are weakly significant, but of no real concern. The residuals appear to be normally distributed (Shapiro-Wilk P-value = 0.852) 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 that the Time variable is not 0 (P-value = 0.000349) and strong evidence that the Time.break variable is not 0 (P-value = 0.000843).

We have strong evidence that Quarter 2 is larger than the omitted baseline (Quarter 1) level (P-value = 2.14e-07), extremely strong evidence that Quarter 3 is larger the omitted baseline level (P-value ≈ 0) and strong evidence that Quarter 4 is is larger than Quarter 1 (P-value = 2.22e-08).

We have very strong evidence against the hypothesis of no autocorrelation (P-value = 1.18e-12).

The F-statistic provides extremely strong evidence against the hypothesis that none of the variables are related to the CO2 concentration (P-value ≈ 0). The Multiple R2 is 0.9997, almost equal to 1 indicating that nearly all the variation in the CO2 concentration is explained by the model.

The Residual Standard Error is 0.1889 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 2019 predictions was 0.2384 which was smaller than the Reduced Harmonic model (0.2439) and a single Cosine Model (0.2648). It was the same as that of the Full Harmonic model, as expected.

Our predictions for 2019 were (in ppm):

2018 Quarter 4: 406.03

2019 Quarter 1: 406.14

2019 Quarter 2: 406.84

2019 Quarter 3: 408.23

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

```
Time.new = 1:79
Time.break.new = c(rep(0, 49), Time.new[50:79] - Time.new[50])
Quarter.new = factor(c(rep(1:4, 19), (1:3)))

full.sf.CO2.fit = lm(CO2.ts[-1] ~ Time.new[-1] + Time.break.new[-1] + Quarter
.new[-1] + CO2.ts[-79])
summary(full.sf.CO2.fit)

##
## Call:
## lm(formula = CO2.ts[-1] ~ Time.new[-1] + Time.break.new[-1] +
##     Quarter.new[-1] + CO2.ts[-79])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50285 -0.12867 -0.00066  0.12102  0.53787
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        108.36491   28.78665   3.764 0.000341 ***
## Time.new[-1]         0.14500    0.03785   3.831 0.000273 ***
## Time.break.new[-1]   0.04082    0.01114   3.665 0.000474 ***
## Quarter.new[-1]2     0.46150    0.07438   6.205 3.25e-08 ***
## Quarter.new[-1]3     1.16834    0.07242  16.132  < 2e-16 ***
## Quarter.new[-1]4     0.41786    0.06380   6.549 7.79e-09 ***
## CO2.ts[-79]          0.70309    0.07876   8.927 3.20e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1888 on 71 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9997
## F-statistic: 5.127e+04 on 6 and 71 DF,  p-value: < 2.2e-16

# Predict for T80 - 2019 Q4
t80.sf.pred = full.sf.CO2.fit$coef[1] + full.sf.CO2.fit$coef[2] * 80 +
  full.sf.CO2.fit$coef[3] * (80 - 50) +
  full.sf.CO2.fit$coef[6] +
  full.sf.CO2.fit$coef[7] * CO2.ts[79]
t80.sf.pred

## (Intercept)
##    408.6447

# Predict for T81 - 2020 Q1, baseline
t81.sf.pred = full.sf.CO2.fit$coef[1] + full.sf.CO2.fit$coef[2] * 81 +
  full.sf.CO2.fit$coef[3] * (81 - 50) +
  full.sf.CO2.fit$coef[7] * t80.sf.pred
t81.sf.pred
```

```
## (Intercept)
##    408.6901

# Predict for T82 - 2020 Q2
t82.sf.pred = full.sf.CO2.fit$coef[1] + full.sf.CO2.fit$coef[2] * 82 +
  full.sf.CO2.fit$coef[3] * (82 - 50) +
  full.sf.CO2.fit$coef[4] +
  full.sf.CO2.fit$coef[7] * t81.sf.pred
t82.sf.pred

## (Intercept)
##    409.3694

# Predict for T83 - 2020 Q3
t83.sf.pred = full.sf.CO2.fit$coef[1] + full.sf.CO2.fit$coef[2] * 83 +
  full.sf.CO2.fit$coef[3] * (83 - 50) +
  full.sf.CO2.fit$coef[5] +
  full.sf.CO2.fit$coef[7] * t82.sf.pred
t83.sf.pred

## (Intercept)
##    410.7396

sf.pred.full = c(t79.sf.pred, t81.sf.pred, t82.sf.pred, t83.sf.pred)
names(sf.pred.full) = c("2019.4", "2020.1", "2020.2", "2020.3")
sf.pred.full

##   2019.4   2020.1   2020.2   2020.3
## 408.2276 408.6901 409.3694 410.7396
```

Comment:

The model including the full data has similar estimates to our previous model. The intercept is slightly smaller while the estimate for Quarter 2 and Quarter 3 is slightly larger.The autocorrelation estimate and Q4 estimates are very similar. The Residual Standard Error is (0.1888 ppm) so the prediction intervals should be reasonably narrow. Our predictions should be reliable.

Question 5.

The best predicting model is the STL Seasonally Adjusted model as it has the lowest RMSEP (0.195 ppm).