

HW3

Part 3



a. Consider the task of modeling this data using a SARIMA model. Based on your knowledge of monthly variation in temperature, what value would be most appropriate for the seasonal lag term, S ?

We could consider that the temperature changes each season of the year, i.e., Spring, Summer, Autumn and winter. So, each three months the temperature will change. Hence, given that we have the averaged monthly maximum daily temperature, we have that

$$S = 3 \text{ (months)}$$

b. Using the seasonal lag selection in the previous subquestion, fit the $\text{SARIMA}(p,d,q) \times (P,D,Q)_s$ model to the full aMDT time series for all combinations of p, d, q, P, D , and Q in $\{0, 1\}$ except the four cases where $P = 1, D = 0, Q = 1$, and $d = 0$ (Hint: this means you should be checking 60 different combinations). In answering this question, you should fit the various models to the full data set (do not split it into a training/test split) and assume that $\delta = 0$ (where δ is the constant term). Identify which of these models best fits the data and report the AICc value for this model and the estimated values of the unknown parameters.

The best fitted model is

$\text{SARIMA}(1, 1, 0) \times (1, 0, 1)_3$

with

$$\text{AIC}_c = 6.663289197$$

$$\phi_1 = 0.5152$$

Values
of the
unknown
parameters

$$H = -0.6361$$

$$\sigma^2 = 44.57$$

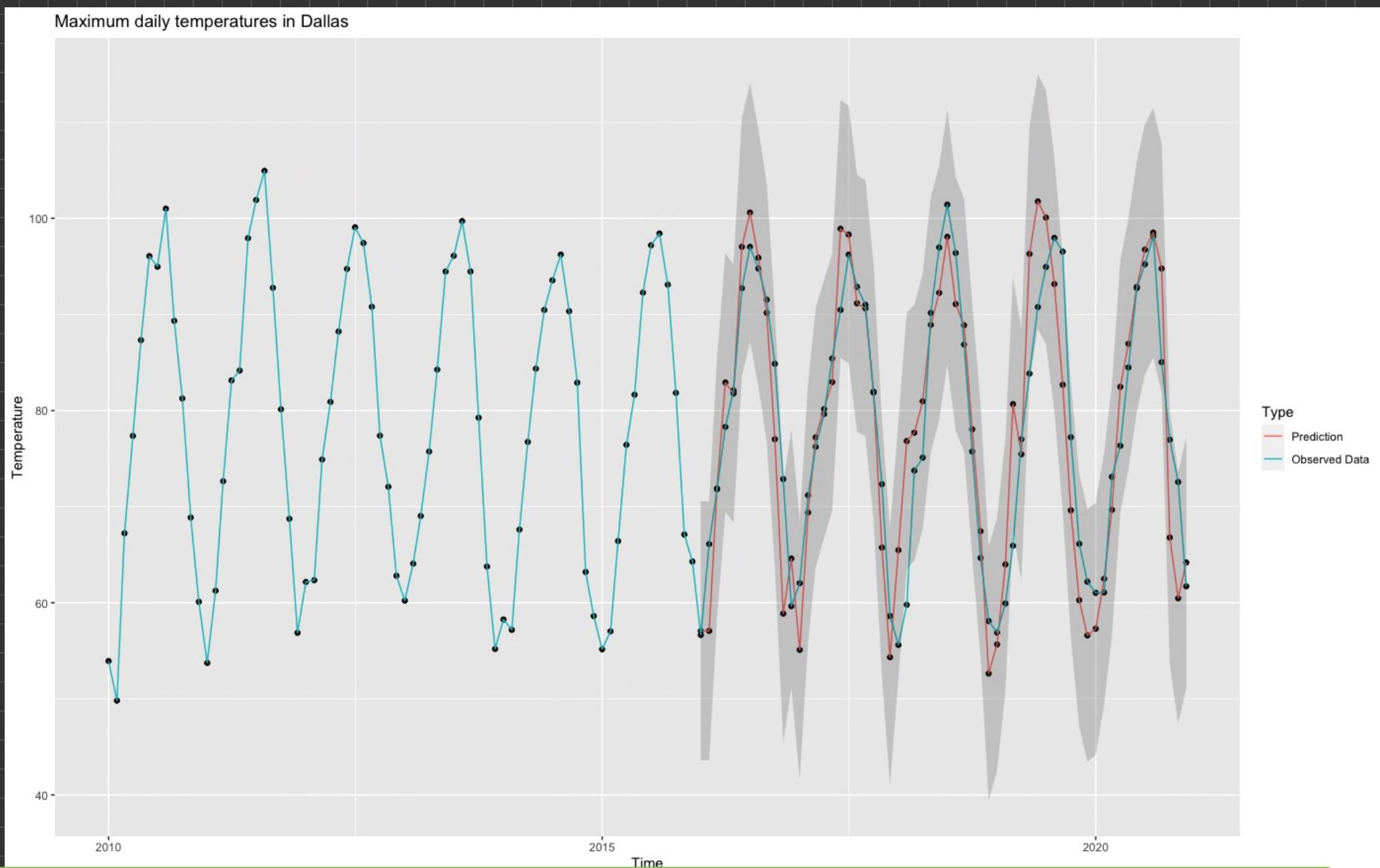
CODE FOR FITTING ALL MODELS,
AND OBTAIN THE BEST ONE

```

11 data1 <- read.csv("/Users/rafa/Documents/Master Austin/MAESTRÍA_AUSTIN/Advanced Predictive Models/HW2/file3.xls", header=TRUE, stringsAsFactors=FALSE)
12 avg_temps<-data1$AvgTemp
13 #Fitting to data all combinations except for P=1,D=0,Q=1 and d=0
14 count<-0
15 lowest<-100
16 for (p_ar in 0:1)
17 {
18   for (q_ar in 0:1)
19   {
20     for (d_ar in 0:1)
21     {
22       for (P_ar in 0:1 )
23     {
24       for (Q_ar in 0:1 )
25     {
26       for (D_ar in 0:1 )
27     {
28       if (d_ar!=0 | P_ar !=1 | D_ar!=0 | Q_ar != 1)
29     {
30       # Fit model
31       fit <- sarima(avg_temps,
32                     p = p_ar,
33                     d = d_ar,
34                     q = q_ar,
35                     P=P_ar,
36                     D=D_ar,
37                     Q=Q_ar,
38                     S=3,
39                     no.constant = TRUE,
40                     details = FALSE)
41
42       if (fit$AICc < lowest)
43     {
44         p_low<-p_ar
45         d_low<-d_ar
46         q_low<-q_ar
47         P_low<-P_ar
48         D_low<-D_ar
49         Q_low<-Q_ar
50         lowest<-fit$AICc
51     }
52     # Examine estimated model parameters
53     print(paste("SARIMA(p=",p_ar,"d=",d_ar, "q=", q_ar, " ) x (P=",P_ar,",D=",D_ar,"Q=",Q_ar,") with AICc =", fit$AICc))
54   }
55 }
56 }
57 }
58 }
59 }
60 }
61 }
62 }
```

c. Consider the task of forecasting the aMDT twelve months in advance. For the last five years of data (2016-2020), predict the value of aMDT using all of the data up until one year prior to the prediction (i.e. predict the aMDT for January 2016 using all of the data up to and including January 2015, then add in the observed aMDT for February 2015 and predict aMDT for February 2016, etc.). Use the values of p, d, q, P, D , and Q as determined to be best in part b, but update your coefficients at every time step using the new data. Create a plot of the one-year-in-advance predictions and 95% confidence bands superimposed on a time series plot of the observed aMDT values from January 2010 to December 2020. Report the one-year-in-advance prediction of aMDT for January 2018, along with the upper and lower bounds of the prediction interval. (Hint: Making one-year-in-advance predictions with newly added data at each time step may require a for loop)

Result ↗



**Prediction of
AMDT For
January 2018
with Lower Bound = 52.12599
Upper Bound = 78.82339**

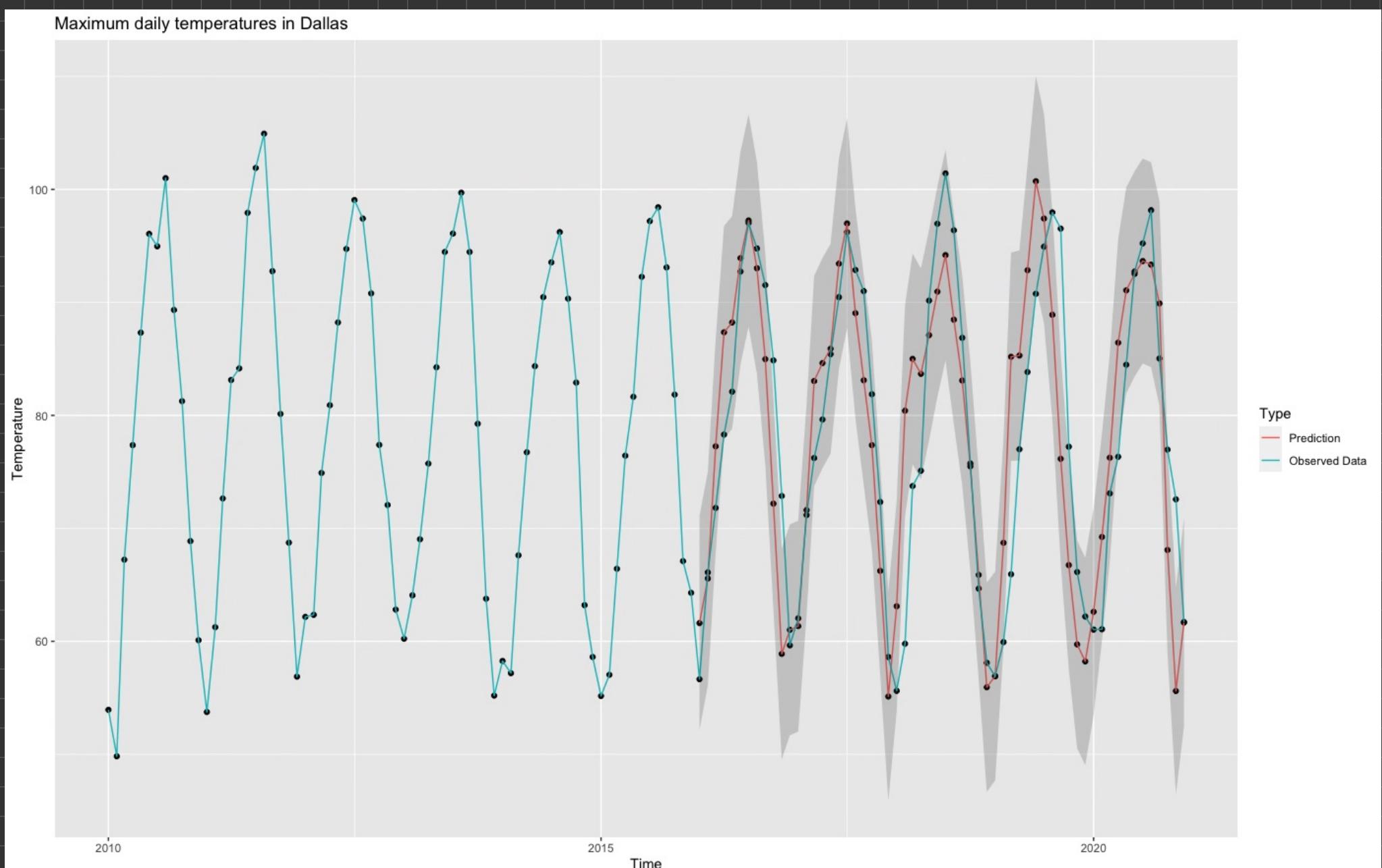
List of predictions

Time	Temperature
2016-01-01	57.07270
2016-02-01	57.07241
2016-03-01	71.88720
2016-04-01	82.91683
2016-05-01	81.77165
2016-06-01	97.03684
2016-07-01	100.59824
2016-08-01	95.90098
2016-09-01	90.16670
2016-10-01	77.01494
2016-11-01	58.87265
2016-12-01	64.59362
2017-01-01	55.09468
2017-02-01	69.38830
2017-03-01	77.20814
2017-04-01	80.14104
2017-05-01	82.95736
2017-06-01	98.90755
2017-07-01	98.31506
2017-08-01	91.15451
2017-09-01	90.63220
2017-10-01	81.96791
2017-11-01	65.74363
2017-12-01	54.33418
2018-01-01	65.47467
2018-02-01	76.82222
2018-03-01	77.69348
2018-04-01	80.95802
2018-05-01	88.91214
2018-06-01	92.24464
2018-07-01	98.07511
2018-08-01	91.07659
2018-09-01	88.86568
2018-10-01	78.04116
2018-11-01	67.44103
2018-12-01	52.63905
2019-01-01	55.66022
2019-02-01	63.98522
2019-03-01	80.66796
2019-04-01	75.45207
2019-05-01	96.29569
2019-06-01	101.75588
2019-07-01	100.07681
2019-08-01	93.15700
2019-09-01	82.67885
2019-10-01	69.61756
2019-11-01	60.26915

2019-12-01	56.60320
2020-01-01	57.29675
2020-02-01	62.48962
2020-03-01	69.67311
2020-04-01	82.46451
2020-05-01	86.94302
2020-06-01	92.85946
2020-07-01	96.74324
2020-08-01	98.51787
2020-09-01	94.77013
2020-10-01	66.78724
2020-11-01	60.46400
2020-12-01	64.18933

d. Now consider an alternative model for the aMDT data that does not have a seasonal component. Report the AICc value for an ARIMA(3,1,1) model fit to the full aMDT data set. Refit the model to make one-year-in-advance predictions of aMDT for the last five years of the observation window (2016-2020) as you did in the previous subquestion. Plot your predictions and 95% confidence bounds, along with the true observed values shown. Set your x-axis to span January 2010 to December 2020. Additionally, report your one-year-in-advance prediction for the aMDT for January 2018, along with your upper and lower bounds of your prediction interval. Does the fitted model produce predictions that capture seasonal behavior? How do the predictions from the ARIMA(3,1,1) model that does not include a specific seasonal component compare to the predictions from the model fitted in part c?

The AICc for ARIMA(3,1,1)
is 5.99606 Result plot ↗



Prediction of
aMDT For
January 2018
with Lower Bound = 53.81456
Upper bound = 72.38862

* The model indeed produce predictions that capture seasonal behaviour. The model captures cycles

* Comparison to model in C:

The two models are extremely similar in accuracy of prediction. However, the variance of this model is smaller, and the prediction band always cover the observed data. Hence, taking into account this point, this last model is slightly better than the last model.

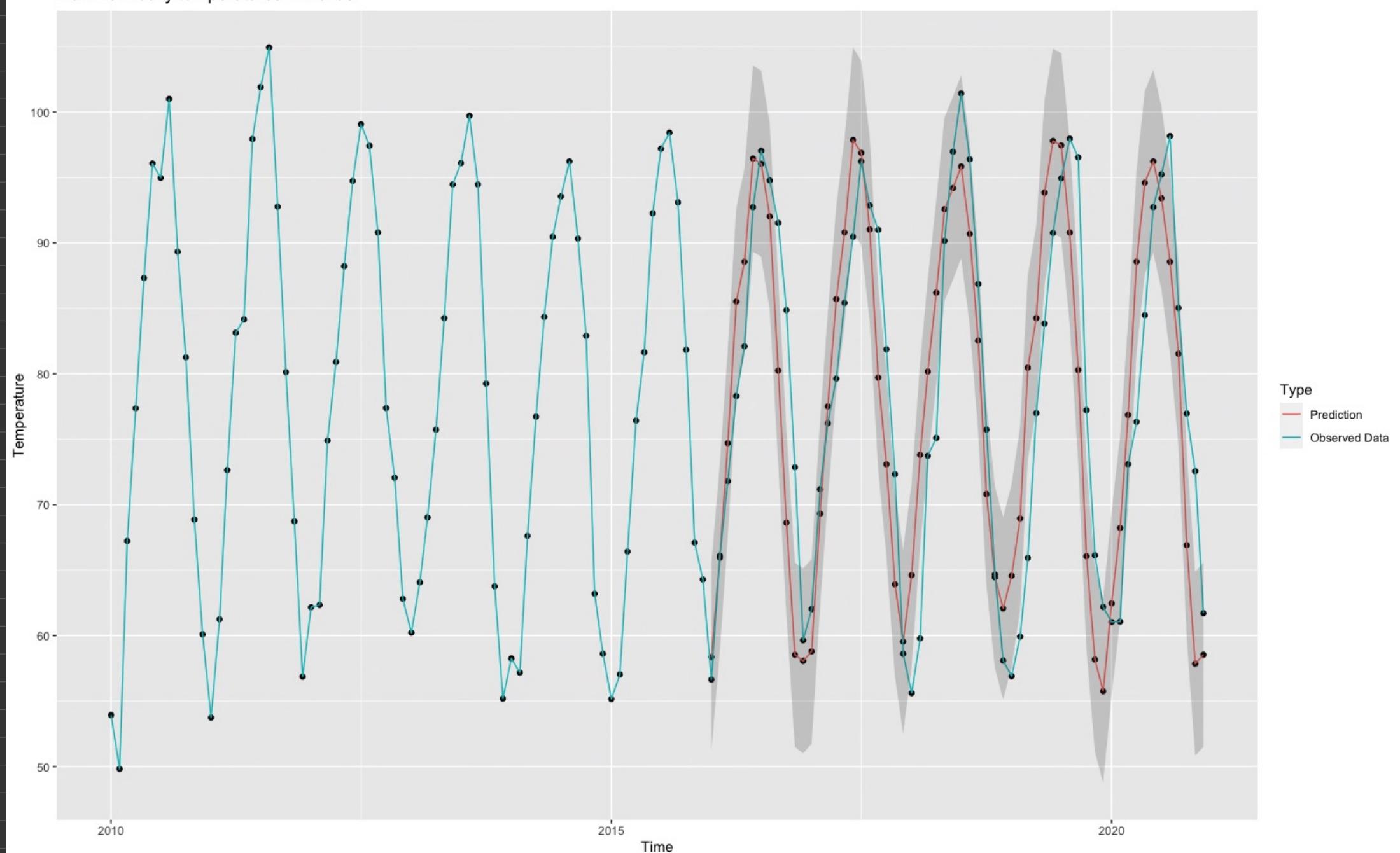
List of Predictions

Time	Temperature
2016-01-01	61.61228
2016-02-01	65.56088
2016-03-01	77.25958
2016-04-01	87.35701
2016-05-01	88.22386
2016-06-01	93.91831
2016-07-01	97.25352
2016-08-01	93.02446
2016-09-01	84.97709
2016-10-01	72.18685
2016-11-01	58.88669
2016-12-01	61.01334
2017-01-01	61.34631
2017-02-01	71.62038
2017-03-01	83.04204
2017-04-01	84.63006
2017-05-01	85.89751
2017-06-01	93.43148
2017-07-01	96.99275
2017-08-01	89.04905
2017-09-01	83.10895
2017-10-01	77.36570
2017-11-01	66.23405
2017-12-01	55.11625
2018-01-01	63.10129
2018-02-01	80.41388
2018-03-01	85.00038
2018-04-01	83.67929
2018-05-01	87.09940
2018-06-01	90.95488
2018-07-01	94.18931
2018-08-01	88.47249
2018-09-01	83.08779
2018-10-01	75.48621
2018-11-01	65.87853
2018-12-01	55.93385
2019-01-01	56.92328
2019-02-01	68.72013
2019-03-01	85.18663
2019-04-01	85.29939
2019-05-01	92.85184
2019-06-01	100.72957
2019-07-01	97.42489
2019-08-01	88.90629
2019-09-01	76.15304
2019-10-01	66.74547
2019-11-01	59.72610
2019-12-01	58.22010
2020-01-01	62.61100
2020-02-01	69.23801
2020-03-01	76.25243
2020-04-01	86.43425
2020-05-01	91.06753
2020-06-01	92.52847
2020-07-01	93.65904
2020-08-01	93.34246
2020-09-01	89.90958
2020-10-01	68.08289
2020-11-01	55.60077
2020-12-01	61.65736

e. Report the AICc value for an ARIMA(12,1,0) model fit to the full aMDT data set. Refit the model to make one-year-in-advance predictions of aMDT for the last five years of the observation window (2016-2020) as you did in the previous subquestion. Plot your predictions and 95% confidence bounds, along with the true observed values shown in. Set your x-axis to span January 2010 to December 2020. Additionally, report your one-year-in-advance prediction for the aMDT for January 2018, along with your upper and lower bounds of your prediction interval. Does the fitted model produce predictions that capture seasonal behavior? How do the predictions from the ARIMA(12,1,0) model compare to the predictions from the models fitted in parts c and d?

The AICc for ARIMA(12,1,0)
is 5.532295

Maximum daily temperatures in Dallas



Prediction of

64.6168

AMDT For

January 2018

with Lower Bound = 57.60232

Upper Bound = 71.63128

*

The model indeed produce predictions that capture seasonal behaviour. The model captures cycles

The model is better than models in b and c. The variance is smaller and the prediction bandwidth covers the observed data. The prediction line is smoother, so much less noise than the other two models.

List of predictions

Time	Temperature
2016-01-01	58.36861
2016-02-01	65.95009
2016-03-01	74.70879
2016-04-01	85.52100
2016-05-01	88.56363
2016-06-01	96.44432
2016-07-01	96.04925
2016-08-01	92.01881
2016-09-01	80.24911
2016-10-01	68.63265
2016-11-01	58.53560
2016-12-01	58.06971
2017-01-01	58.79656
2017-02-01	69.32594
2017-03-01	77.51095
2017-04-01	85.70879
2017-05-01	90.80865
2017-06-01	97.86221
2017-07-01	96.88261
2017-08-01	91.03218
2017-09-01	79.71185
2017-10-01	73.08946
2017-11-01	63.91679
2017-12-01	59.54528
2018-01-01	64.61680
2018-02-01	73.81076
2018-03-01	80.17303
2018-04-01	86.20109
2018-05-01	92.56782
2018-06-01	94.18889
2018-07-01	95.84012
2018-08-01	90.69666
2018-09-01	82.53590
2018-10-01	70.80967
2018-11-01	64.44962
2018-12-01	62.07836
2019-01-01	64.57401
2019-02-01	68.96034
2019-03-01	80.47175
2019-04-01	84.25959
2019-05-01	93.84566
2019-06-01	97.78971
2019-07-01	97.45003
2019-08-01	90.79812
2019-09-01	80.28995

Time	Temperature
2019-01-01	64.57401
2019-02-01	68.96034
2019-03-01	80.47175
2019-04-01	84.25959
2019-05-01	93.84566
2019-06-01	97.78971
2019-07-01	97.45003
2019-08-01	90.79812
2019-09-01	80.28995
2019-10-01	66.06609
2019-11-01	58.17680
2019-12-01	55.75895
2020-01-01	62.46463
2020-02-01	68.23116
2020-03-01	76.86224
2020-04-01	88.56803
2020-05-01	94.59211
2020-06-01	96.22852
2020-07-01	93.41267
2020-08-01	88.56267
2020-09-01	81.53426
2020-10-01	66.90264
2020-11-01	57.85802
2020-12-01	58.54054

APPENDIX (CODE FOR C, D and E)

```
78 #-----question c-----  
79  
80 #Index of Month of January 2015  
81 limit<-12*15+1  
82 # Number of Predictions from 2015 to 2020 (5 years)  
83 limit2<-limit+12*5-1  
84  
85 preds<-numeric(limit2-limit)  
86 SE<-numeric(limit2-limit)  
87  
88 #predicting for each day in one year advance  
89 for (i in limit:limit2){  
90 {  
91 x_train<-data1[-(i+1:length(data1$Month)),]  
92 fit_for <- sarima.for(x_train$AvgTemp,  
93 n.ahead = 12,  
94 p = 1,  
95 d = 1,  
96 q = 0,  
97 P=1,  
98 D=0,  
99 Q=1,  
100 S=3,  
101 plot = F)  
102 preds[i-limit+1]<-fit_for$pred[1]  
103 SE[i-limit+1]<-fit_for$se[1]  
104  
105 print(x_train$Month[length(x_train$Month)])  
106 print(preds[i-limit+1])  
107 print(preds[i-limit+1]- 1.96*SE[i-limit+1])  
108 print(preds[i-limit+1]+1.96*SE[i-limit+1])  
109 }  
110 #Creating the data frames  
111 pred_lim1<-limit+12  
112 pred_lim2<-limit2+12  
113  
114 lim10<-12*10+1  
115 lim20<-pred_lim2  
116 pred_time<-data.frame(Type=factor(rep("Prediction", pred_lim2-pred_lim1+1)),Time=as.Date(data1$Month[pred_lim1:pred_lim2]),Temperature=preds)  
117 f2010_2020<-data.frame(Type=factor(rep("Observed Data", lim20-lim10+1)),Time=as.Date(data1$Month[lim10:lim20]),Temperature=data1$AvgTemp[lim10:lim20])  
118  
119 fit_data <- bind_rows(pred_time,f2010_2020)  
120 fit_data  
121 # Plot data and forecasts  
122 gg_fit <- ggplot(fit_data,aes(x = Time,y=Temperature,group=Type),colour=Type) +geom_point() +geom_line(aes(color=Type)) +  
123 geom_ribbon(data = pred_time,  
124 aes(x = Time,  
125 ymin = Temperature - 1.96*SE,  
126 ymax = Temperature + 1.96*SE),  
127 alpha = .2)+  
128 labs(title = "Maximum daily temperatures in Dallas")  
129  
130 gg_fit  
131
```

```

133 #-----question d-----
134 data1 <- read.csv("/Users/rafa/Documents/Master Austin/MAESTRÍA_AUSTIN/Advanced Predictive Models/HW2/file3.xls", header=TRUE, stringsAsFactors=FALSE)
135 #fitting to obtain AIC_c
136 fit <- sarima(data1$AvgTemp,
137   p = 3,
138   d = 1,
139   q = 1,
140   no.constant = TRUE,
141   details = FALSE)
142 fit
143 #Index of the month January
144 limit<-12*15+1
145 #Number of predictions in 5 years
146 limit2<-limit+12*5-1
147 preds<-numeric(limit2-limit)
148 SE<-numeric(limit2-limit)
149 for (i in limit:limit2) {
150   {
151     x_train<-data1[-(i+1:length(data1$Month)),]
152     fit_for <- sarima.for(x_train$AvgTemp,
153       n.ahead = 12,
154       p = 3,
155       d = 1,
156       q = 1,
157       plot = F)
158     preds[i-limit+1]<-fit_for$pred[1]
159     #Standard error
160     SE[i-limit+1]<-fit_for$se[1]
161     #Getting the prediction and upper and lowe bound
162     print(x_train$Month[length(x_train$Month)])
163     print(preds[i-limit+1])
164     print(preds[i-limit+1]- 1.96*SE[i-limit+1])
165     print(preds[i-limit+1]+1.96*SE[i-limit+1])
166   }
167   #Making dataframes
168   pred_lim1<-limit+12
169   pred_lim2<-limit2+12
170   lim10<-12*10+1
171   lim20<-pred_lim2
172
173   pred_time<-data.frame(Type=factor(rep("Prediction", pred_lim2-pred_lim1+1)),Time=as.Date(data1$Month[pred_lim1:pred_lim2]),Temperature=preds)
174   f2010_2020<-data.frame(Type=factor(rep("Observed Data", lim20-lim10+1)),Time=as.Date(data1$Month[lim10:lim20]),Temperature=data1$AvgTemp[lim10:lim20])
175   fit_data <- bind_rows(pred_time,f2010_2020)
176   fit_data
177   # Plot data and forecasts
178   gg_fit <- ggplot(fit_data,aes(x = Time,y=Temperature,group=Type),colour=Type) +geom_point() +geom_line(aes(color=Type)) +
179     geom_ribbon(data = pred_time,
180       aes(x = Time,
181           ymin = Temperature - 1.96*SE,
182           ymax = Temperature + 1.96*SE),
183       alpha = .2) +
184     labs(title = "Maximum daily temperatures in Dallas")
185
186 gg_fit

```

```

192
193 #Same procedure as d) ...just changing the model to ARIMA(12,1,0)
194 data1 <- read.csv("/Users/rafa/Documents/Master Austin/MAESTRÍA_AUSTIN/Advanced Predictive Models/HW2/file3.xls", header=TRUE, stringsAsFactors=FALSE)
195 fit <- sarima(data1$AvgTemp,
196   p = 12,
197   d = 1,
198   q = 0,
199   no.constant = TRUE,
200   details = FALSE)
201 fit
202 limit<-12*15+1
203 limit2<-limit+12*5-1
204 preds<-numeric(limit2-limit)
205 SE<-numeric(limit2-limit)
206 for (i in limit:limit2)
207 {
208   x_train<-data1[-(i+1:length(data1$Month)),]
209   fit_for <- sarima.for(x_train$AvgTemp,
210     n.ahead = 12,
211     p = 12,
212     d = 1,
213     q = 0,
214     plot = F)
215   preds[i-limit+1]<-fit_for$pred[1]
216   SE[i-limit+1]<-fit_for$se[1]
217   print(x_train$Month[i])
218   print(preds[i-limit+1])
219   print(preds[i-limit+1]- 1.96*SE[i-limit+1])
220   print(preds[i-limit+1]+1.96*SE[i-limit+1])
221 }
222 pred_lim1<-limit+12
223 pred_lim2<-limit2+12
224 lim10<-12*10+1
225 lim20<-pred_lim2
226 pred_time<-data.frame(Type=factor(rep("Prediction", pred_lim2-pred_lim1+1)),Time=as.Date(data1$Month[pred_lim1:pred_lim2]),Temperature=preds)
227 f2010_2020<-data.frame(Type=factor(rep("Observed Data", lim20-lim10+1)),Time=as.Date(data1$Month[lim10:lim20]),Temperature=data1$AvgTemp[lim10:lim20])
228 fit_data <- bind_rows(pred_time,f2010_2020)
229 fit_data
230 # Plot data and forecasts
231 gg_fit <- ggplot(fit_data,aes(x = Time,y=Temperature,group=Type),colour=Type) +geom_point() +geom_line(aes(color=Type)) +
232   geom_ribbon(data = pred_time,
233     aes(x = Time,
234       ymin = Temperature - 1.96*SE,
235       ymax = Temperature + 1.96*SE),
236     alpha = .2) +
237   labs(title = "Maximum daily temperatures in Dallas")
238
239 gg_fit
240
241

```