MAS6011 Time Series Project - Analysis of Daily Temperatures in Melbourne

179746491-6011-P2

28 April 2018

The time series data set consisting of daily maximum temperatures ($°C$) in Melbourne, from 1st January 1981 to 31 December 1990 were analysed using **R Studio** package . This report investigated the structure of the data, modelled the data using a state space model and provided forecasts. All code is included in the Appendix.

Data structure

The text file containing the data were converted into a time series format, consisting of temperatures for each of the 3650 consecutive days in the analysed time period. This represented 365 daily temperatures per year.

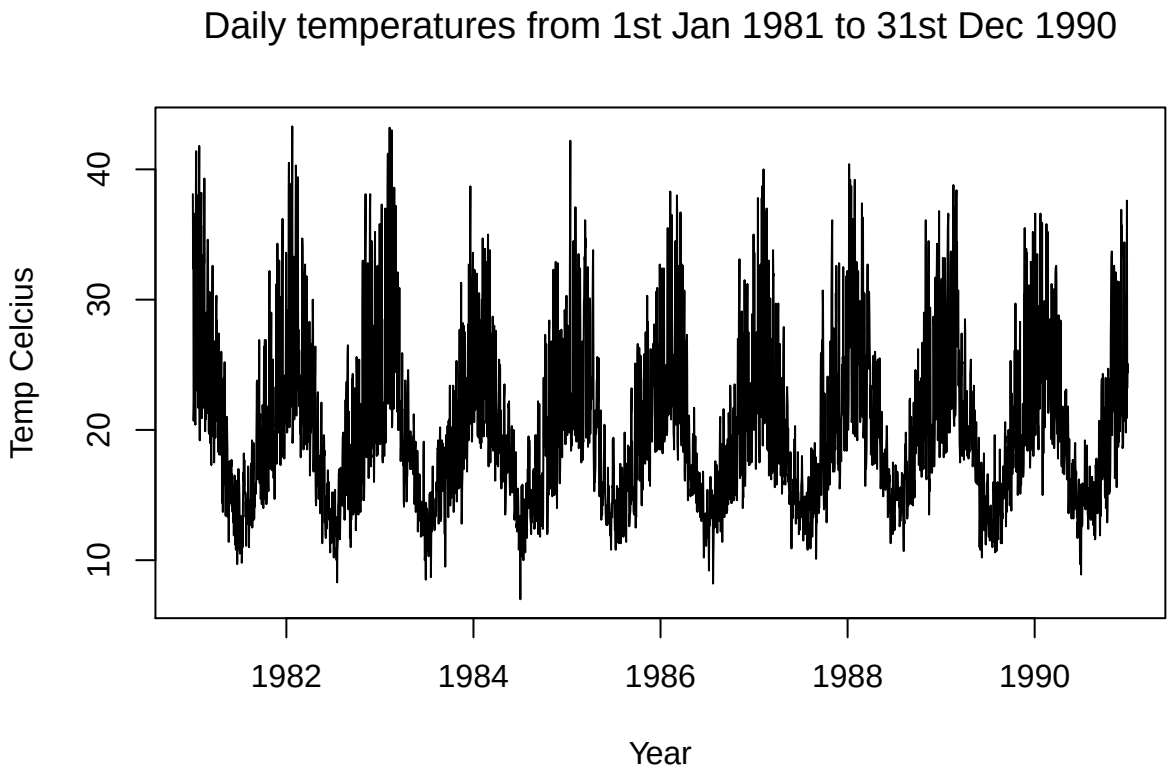

Figure 1: Daily plot of temperatures

The time series data is displayed in Figure 1.

The daily temperature plot displays much daily variation. However, it is clear that seasonal variation exists. A seasonal state space model was chosen to analyse the data. A full effects seasonal model that would use

all 3650 data points would require a too large transition matrix F_t. In addition, there would be little gain in using such a large transition matrix compared to a reduced Fourier form model. In practice, many time series data are averaged over a quarter year period and this technique was also employed in this report to produce a reduced Fourier form model.

The time points for each quarter year period (91 days) were averaged to obtain a single averaged temperature value. As there were a non-multiple of 4 time points per year (365), every sequential 365th time point from the data was removed.

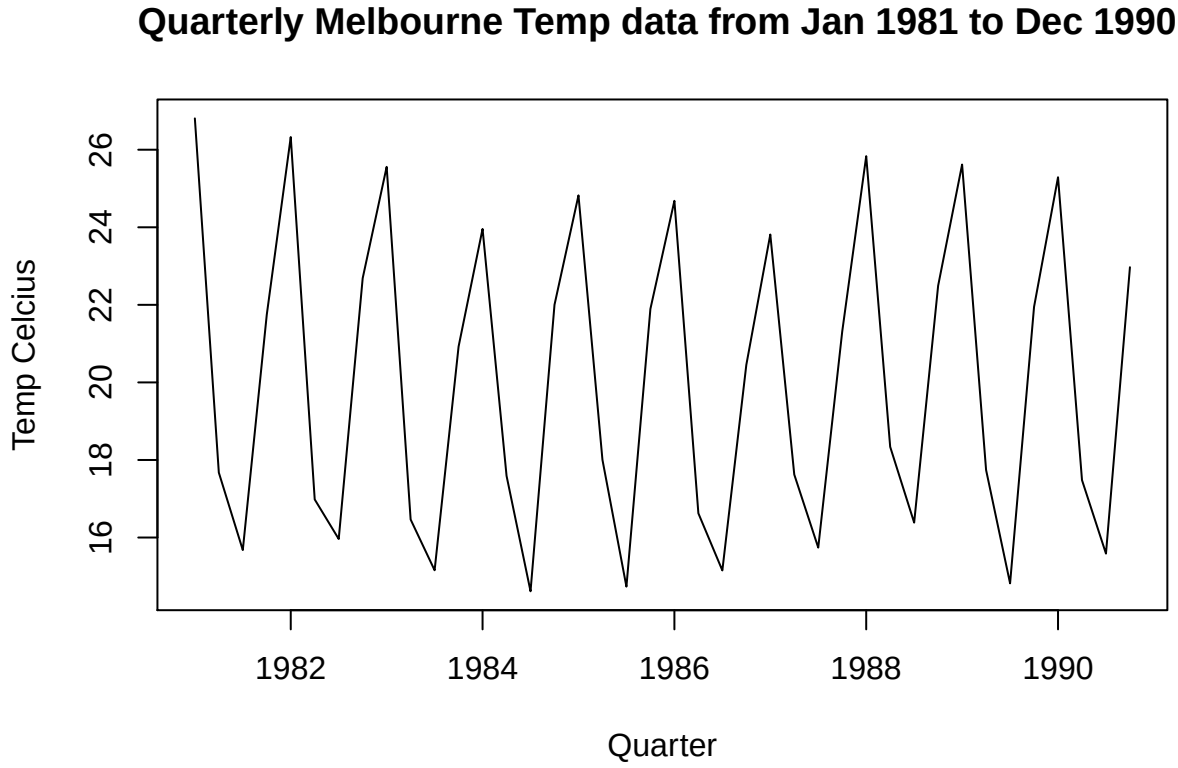

Figure 2: Quarter yearly averaged of temperatures

The relationship between averaged temperature and time can be seen in Figure 2. A seasonal component to the data can be clearly seen. Also, ignoring the effects of seasonality, there appears a decreasing trend in average quarter yearly temperature from the beginning of 1981 until around mid 1984. Afterwhich, the trend appears to disappear. Therefore, a state space model for trend and seasonal effects was chosen to investigate the quarter yearly data.

Data modelling

The superposition of state space models, where N time series $\{y_{it}\}$ $(i = 1, \ldots, N)$ each of which follow a state space model, follow the formula

$$\begin{aligned} y_t &= x_{it}^T \beta_{it} + \epsilon_{it} \quad \epsilon_{it} \tilde{} N(0, \sigma_i^2) \quad \text{(observation model)} \\ \beta_{it} &= F\beta_{i,t-1} + \zeta_{it} \quad \zeta_{it} \tilde{} N(0, Z_{it}) \quad \text{(transition model)} \end{aligned}$$

In our model the components can be written as

$$\mathbf{X}^T = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \end{bmatrix}^T$$

$$\mathbf{F} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix}$$

$$\beta_{\mathbf{t}} = \begin{bmatrix} \beta_{1,1,t} \\ \beta_{1,2,t} \\ \beta_{2,1,t} \\ \beta_{2,2,t} \\ \beta_{2,3,t} \end{bmatrix}$$

$$\beta_{\mathbf{t-1}} = \begin{bmatrix} \beta_{1,1,t-1} \\ \beta_{1,2,t-1} \\ \beta_{2,1,t-1} \\ \beta_{2,2,t-1} \\ \beta_{2,3,t-1} \end{bmatrix}$$

The value for the beta prior $\beta_{0|0}$ was 20, with a variance $P_{0|0} = 1000$.

The **dlm** package was used to fit the model to the data. Estimation of the values for σ^2 and Z_{it} were obtained using the **dlmMLE** command. A build function was created and the maximum likelihood estimates obtained using initial values for $\sigma^2 = 0.5$ and a 5*5 diagonal Z matrix with values of 0.01.

The maximum likelihood estimate for $\sigma^2 = 0.251$ and for the diagonal matrix
$Z = [0.1000014, 0.000001, 0.000001, 0.000001, 0.010]$. These values were then used in conjunction with those above to define the model. The resulting log likelihood value for the model was -36.46. This was low when compared to other adhoc values of σ^2 and Z_t. Therefore the estimated values for σ^2 and Z_t were utilised in the model.

A fit of the data was performed using the the **dlmFilter** command. A smoothed average and 95% prediction intervals were calculated using **dlmSmooth** and **dlmdlmSvd2var** packages respectively. The resulting plot is displayed in Figure 3.

Figure 3 demonstrates that initially the one-step forecasts and prediction intervals are extremely inaccurate. This is expected as seasonality is not taken into account for the first four data points. Another plot is shown below (with same legend), whereby the y-axis range has been adjusted to the data points.

Figure 4 is a scaled up version of Figure 4. When ignoring the first five data points, the actual observations (red crosses) appear within the 95% prediction interval. This suggests that the model fits the data well.

Error Analysis

Error analysis was performed using residual plots and ACF plots. Because the variance of the residual at time t may vary over time, error analysis was performed using standardised residuals, defined as

$$e_t^* = \frac{e_t}{\sqrt{(}q_{t|t-1})},$$

which follows a standard normal distribution, N(0,1).

Figure 5 demonstrates a plot of standardised residuals. The standardised residuals e_t^* appear to fluctuate randomly around zero. All but one e_t^* are within 95% credible bounds of the standard normal distribution.

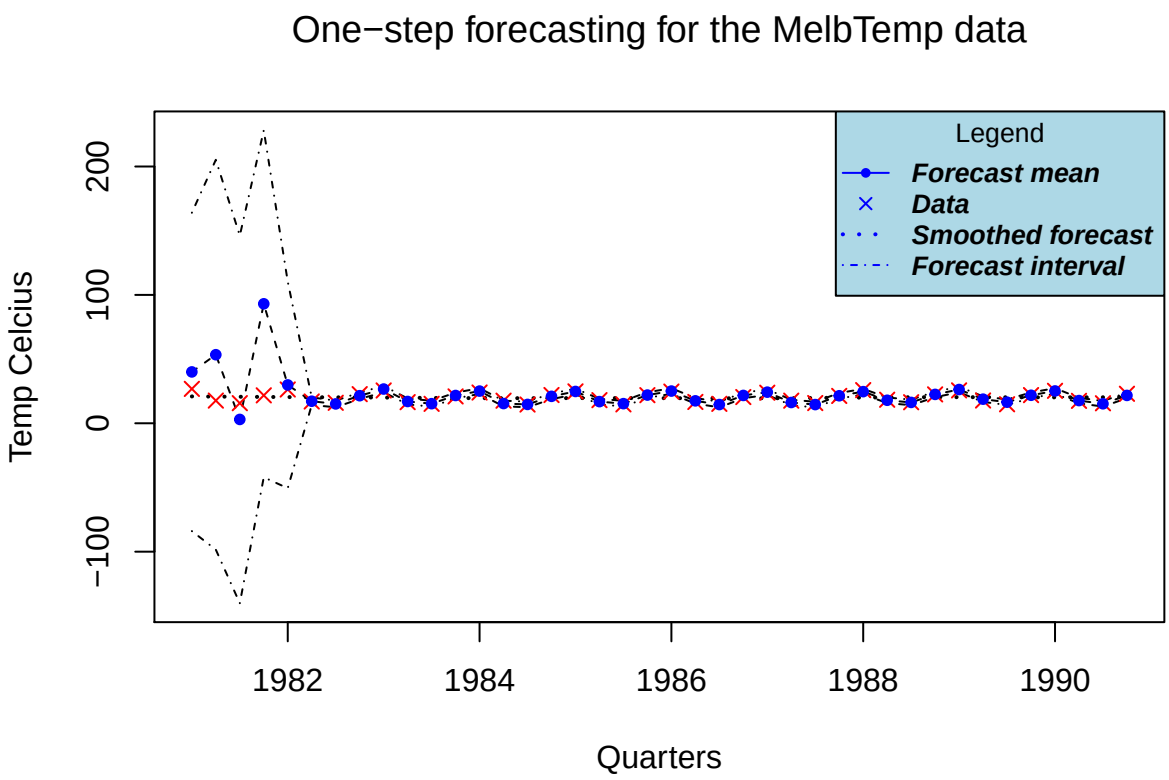

Figure 3: One-step forecasting and smoothing

The mean of the standardised residuals is close to zero $E(e_t^*) = 0.01$ and variance ($q_{t|t-1} = 0.71$). As 2.5% of the residuals e_t^* lie outside the 95% credible bounds, and that $E(e_t^*)$ is close to zero, it appears the model fits the data well.

The mean of squared residuals (MSE), the mean of the standardised residuals (MSSE) and the mean absolute deviation (MAD), defined by

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e_t^2, \quad MSSE = \frac{1}{n}\sum_{i=1}^{n}(e_t^*)^2, \quad MSE = \frac{1}{n}\sum_{i=1}^{n}|e_t|,$$

respectively, were calculated. For a good fit, MSE and MAD should be close to 0, while MSSE should be close to 1. It was found that

$$MSE = 168.4, \quad MSSE = 0.56, \quad MAD = 4.02.$$

The MSSE is higher than 1, which indicates that either the average residual value is high or that the variance $q_{t|t-1}$ is low. The MSE is high but this value is dependent on the scale of the observations. The MAD is much lower but interpretation is difficult because MAD is also dependent on scale of observations.

The plot of the ACF for standardised residuals (figure 6) demonstrates that all but lag k = 0 and lag k = 5 are within the 95% credible bounds. As lag k=0 is always 1 we can ignore this value. Lag k=5 is more concerning, however we expect the value of one lag out of 20 to be outside the credible interval. Lag 5 does not correspond with seaonality (lag k=4), therefore I am willing to accept the large value at lag k=5 was caused by chance. Hence we conclude that the standardised residuals appear uncorrelated.

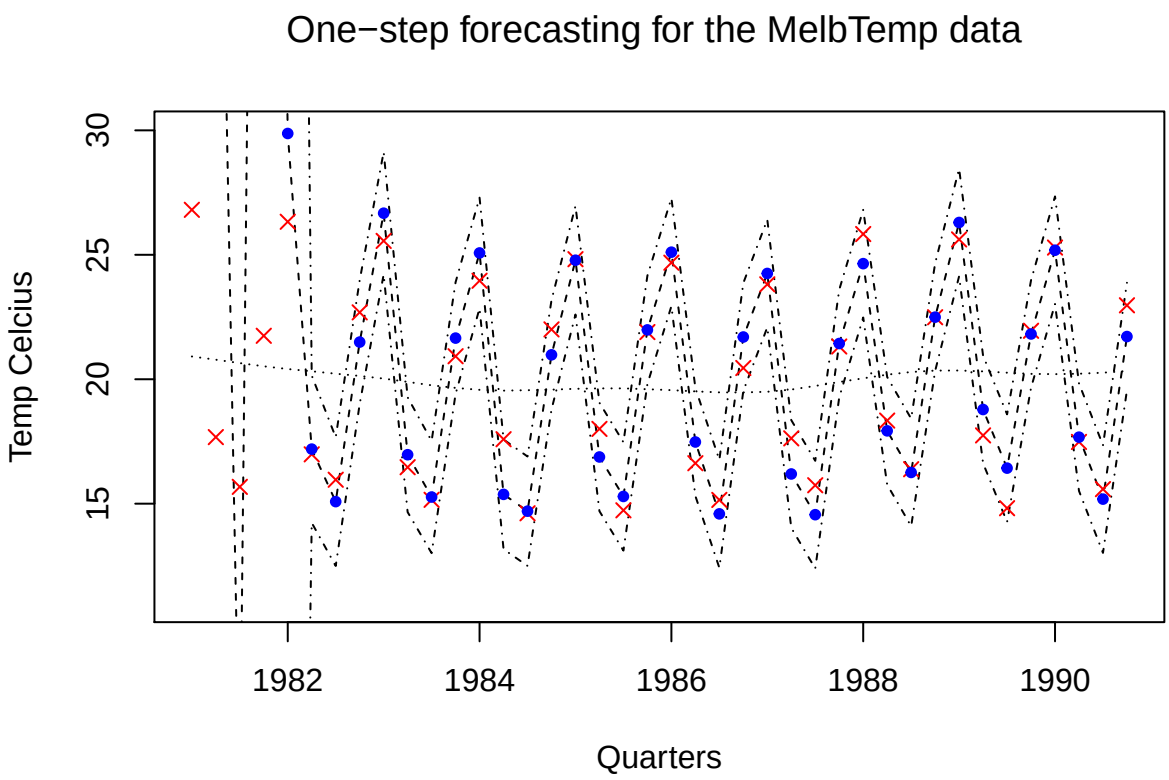

Figure 4: Reduced scale one-step forecasting and smoothing

Finally, the sum of the residuals $\sum_{i=1}^{n}(e_t^*)^2 = 27.2$. This is lower than the one-sided 95% quantile of the chi-squared distribution with 40 degrees of freedom (55.8). This implies the overall fit of the model is good.

In summary, the fit of the model is good, with little evidence of auto-correlation.

Forecasting

Forecasting k-steps ahead was performed using the **dlmForecast** command. In order to assess the accuracy of the forecasting model, the first 20 values of the dataset were used as training data for the model that the forecasting function would require. The σ^2 and Z values estimated previously were used to create the training data model. The latter 20 k-step ahead quarter values were then predicted with the **dlmForecast** function. These k-step ahead prediction values were subtracted from the 21st-40th observed values in the dataset.

Figure 7 displays the observed temperatures subtracted from the predicted temperatures from quarter 21-40. There appears a bias to over-predict the temperature. The mean value of prediction temperature above observed temperature is 0.42 and the standard deviation is 0.66. A paired t-test was performed to test whether the bias was real. The 95% confidence interval (0.11, 0.73) demonstrates moderate evidence that the prediction estimate over-estimates the observed temperature by a mean value of 0.42.

The 5th and 6th k-step predictions are outside the 95% credible bounds. Due to the prediction bias, and With a small sample size, this is unsurprising.

There appears possible autocorrelation in Figure 7. However, the ACF plot of the data (Figure8) fails to show significant values of lags above zero, although lag 5 value is noted as almost being significant. Hence

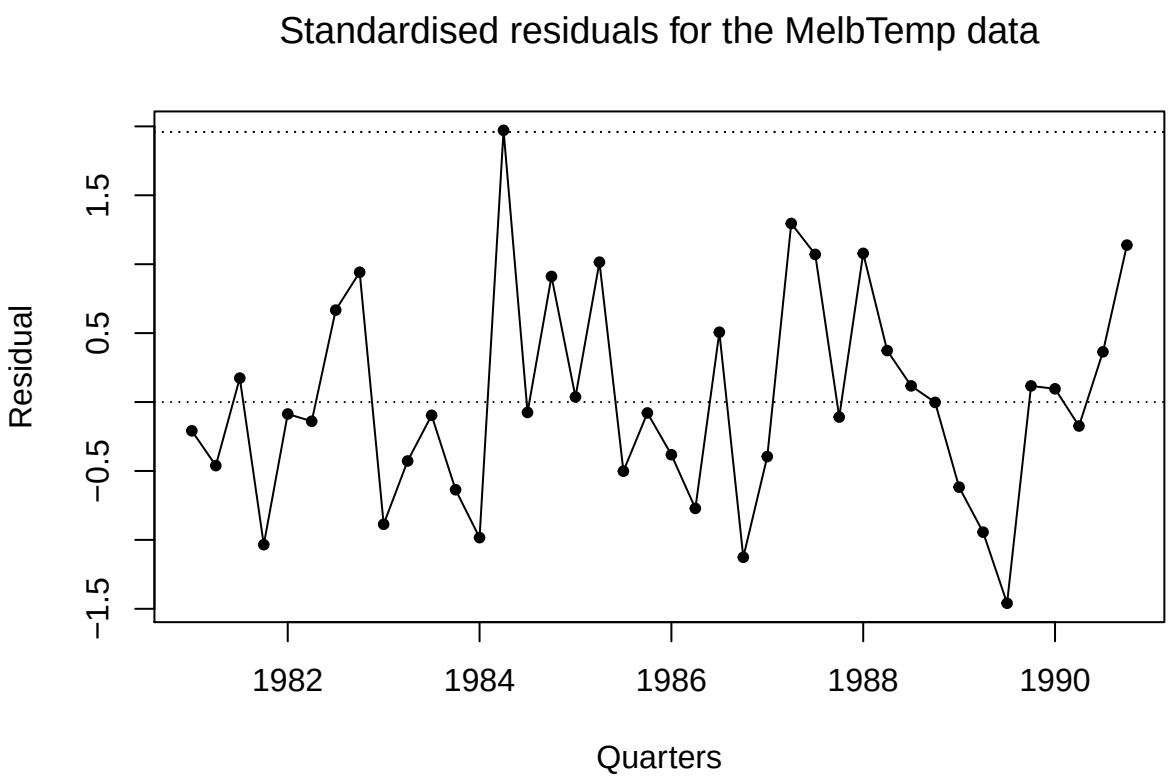

Figure 5: Standardised residual plot

there is little evidence of autocorrelation in the prediction results.

In conclusion, the k-step predictions for the Quarters 21-40 are biased by a mean temperature value of 0.42. Informally, the variance of the difference between the k-th prediction estimate and observed value doesn't appear to increase with k. Accounting for the bias, it appears the prediction model is reasonable.

Utilising the model that was constructed from the 40 quarter yearly, a plot was constructed displaying the predicted temperatures from 1991 and beyond, seasonally adjusted (Figure 9). By removing the seasonal effect, there appears a small upward trend in the k-step forecast. This trend is however masked by the large 95% prediction bounds.

The k-step non-seasonally adjusted forecasts are shown below

```
##          Qtr1     Qtr2     Qtr3     Qtr4
## 1991 25.59756 18.09618 15.90073 22.72797
## 1992 25.87164 18.37026 16.17481 23.00205
## 1993 26.14572 18.64434 16.44889 23.27612
```

and the k-step variance from 1 to 12 respectively are

```
##  [1]  1.215479  1.409139  1.736439  2.074024  3.174367  3.880086  4.825081
##  [8]  5.787482  7.887025  9.424802 11.307493 13.214710
```

The increase in average quarterly temperature between for each year is $0.27°C$, while the variance $q_{t+4|t}$ is 2.07. Such a large variance compared to mean temperature limits the usefulness of the k-step prediction.

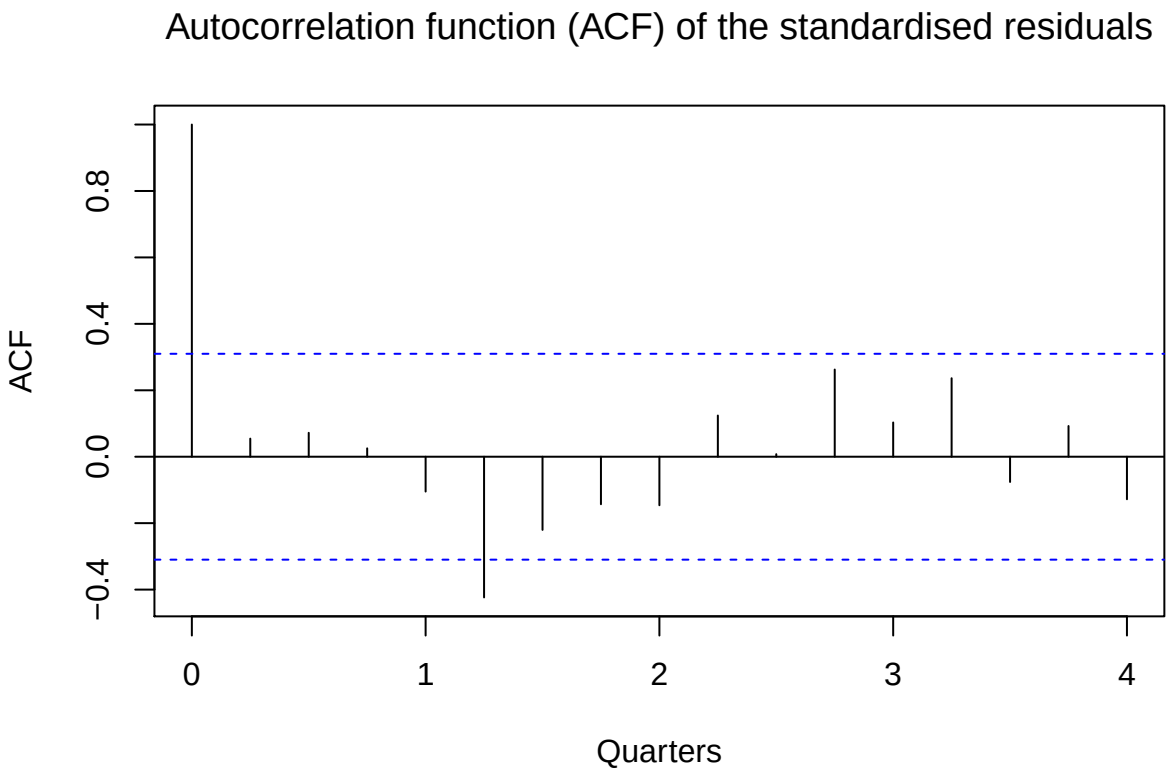

Figure 6: ACF of the standarised residuals

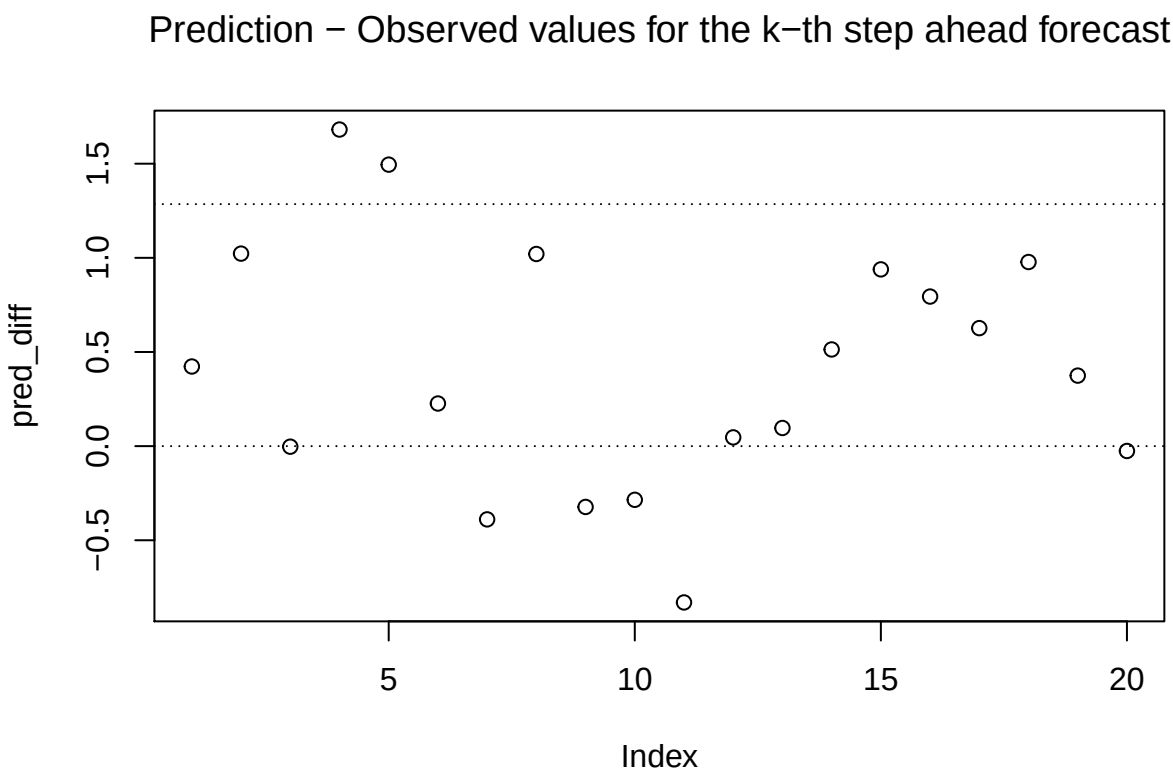

Figure 7: Prediction - Observed temperatues from Quarter 21-40

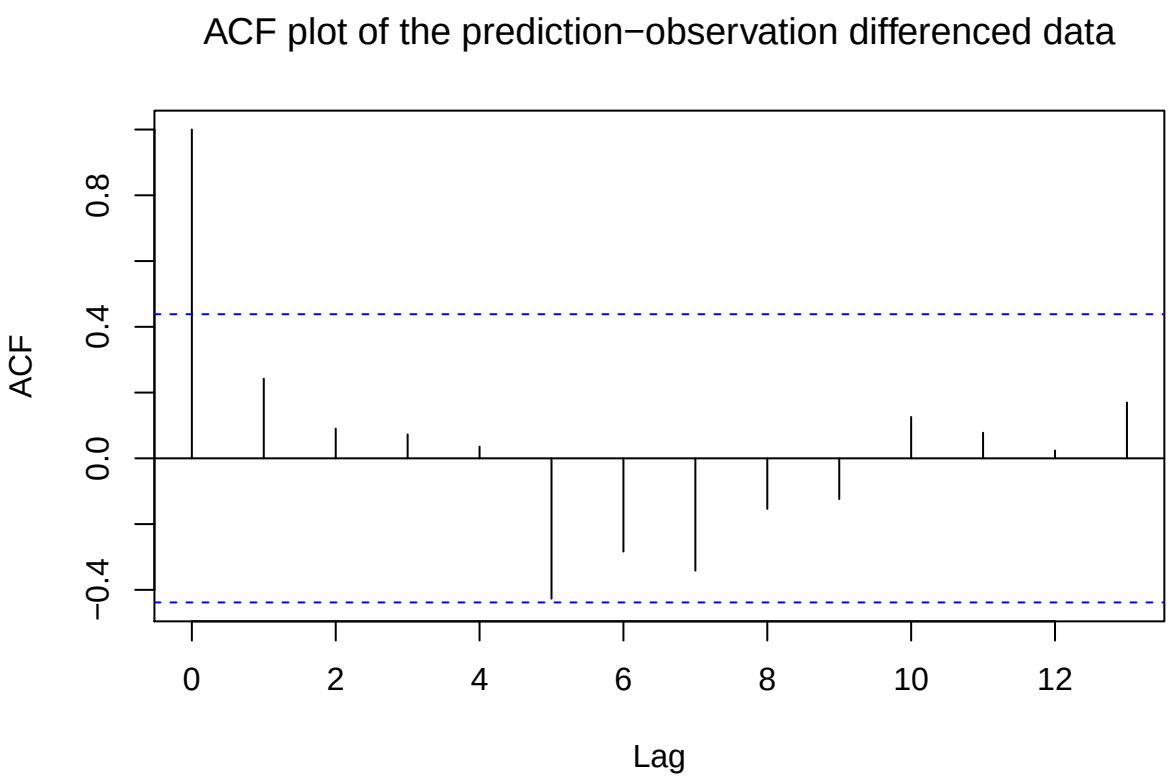

Figure 8: ACF of prediction-observation differences from Quarter 21-40

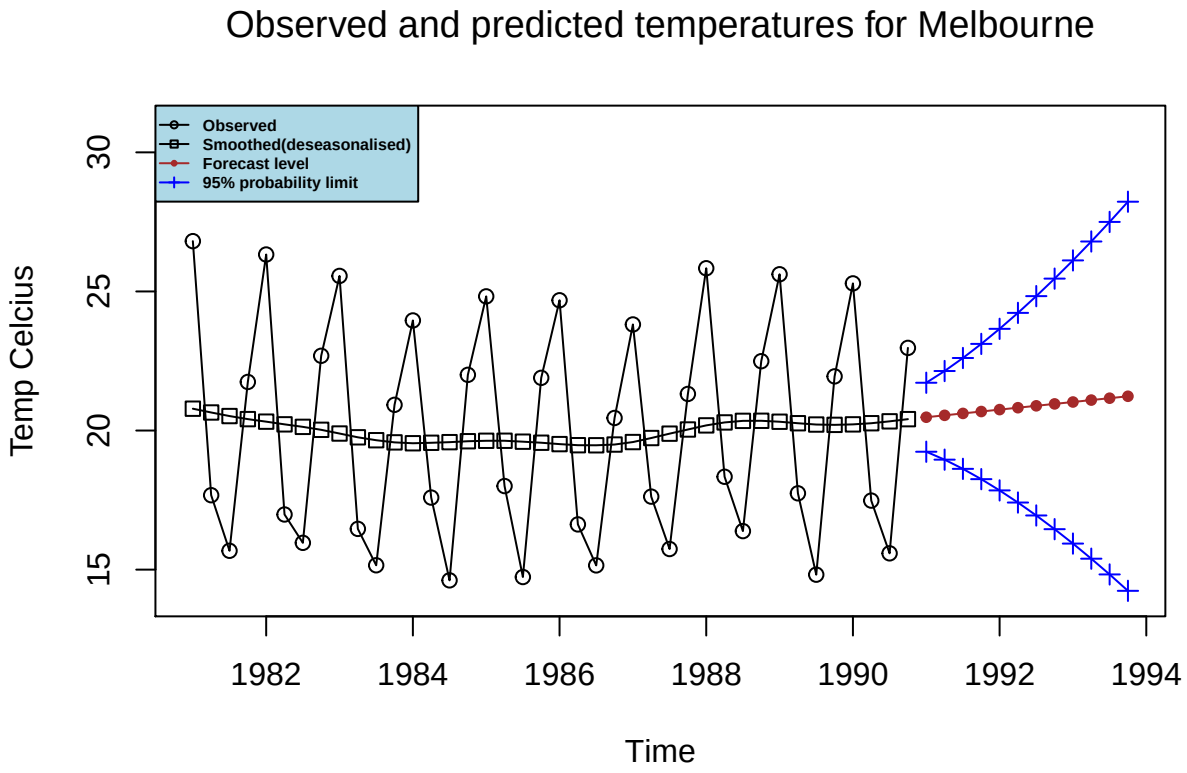

Figure 9: Forecast of Melbourne temperatures without seasonal effect

Appendix

```
library(dlm)
melbtemp <- read.table("G:\\Masters Sheffield\\MAS6061\\Sem2
                       \\TempMelb.csv", skip=11, header = T, sep = "," )
#create time series for data
melbtemp <- (melbtemp[,2])
daily <- ts(melbtemp, start=c(1981,1), frequency = 365)
ts.plot(daily, main=expression("Daily temperatures from 1st Jan 1981 to 31st Dec 1990"),
        ylab="Temp Celcius", xlab="Year")


#removing 31st december from each year
rmdays <- seq(365,3650, by=365)
revrmdays <- rev(rmdays)
dailynew <- daily[-revrmdays]

#create vector ranges that each mean will be calculated
ind2 <- seq(91,length(dailynew),91) # last day for each quarter
ind1 <- ind2-90                     #first day of each quarter

#create function that will calculate mean of each quarter
day2quart <- function(tseries, list1, list2, ...){
  n <- length(list1)
  d <- rep(0, n)
  for (i in 1:length(list1)){
  d[i] <- (mean(tseries[list1[i]:list2[i]]))
  }
  return(d)
}


qtrmeans <- day2quart(dailynew, ind1, ind2)
qtrmts <- ts(qtrmeans, start=c(1981,1), frequency = 4)
ts.plot(qtrmts, ylab="Temp Celcius", xlab="Quarter",
        main="Quarterly Melbourne Temp data from Jan 1981 to Dec 1990 ")

#trend and seasonal model with
x <- matrix(c(1,0,1,0,1),1)
F1 <- matrix(c(1,0,1,1),2,2)
F2 <- matrix(c(0,-1,1,0),2,2)
F3 <- matrix(c(-1),1,1)
F <- bdiag(F1,F2,F3)
beta0 <- rep(20,5)
P0 <- 1000*diag(5)

#Create build function to determine MLEs of sigma and Z
build <- function(parm){
  sigma2=parm[1]
  diag(Z)[1:4] <- parm[2:5]
  return(list(FF=x, GG=F, V=sigma2, W=Z, m0=beta0, C0=P0))
}
```

```r
#without initially setting sigma2 and Z values the dlmMLE doesn't converge
sigma2 <-.5
Z <- 0.01*diag(5)
mod <- dlm(qtrmts, FF=x, GG=F, V=sigma2, W=Z, m0=beta0, C0=P0)
dlmLL(y=qtrmts, mod=mod) #negative Log likelihood

#find MLE of parameters sigma2 and Z
mlefit <- dlmMLE(y=qtrmts, parm=c(0.5, rep(0.01,5)), build=build, lower=c(1e-6,1e-6))

#set values for sigma2 and Z according to MLE
sigma2 <-  mlefit$par[1]
Z <- bdiag(mlefit$par[2], mlefit$par[3],mlefit$par[4],mlefit$par[5],mlefit$par[6])

#fit model with Kalman filter
fit <- dlmFilter(qtrmts, mod)
#fit model by smoothing
fit.s <- dlmSmooth(qtrmts, mod)

#calculate prediction intervals
R <- dlmSvd2var(fit$U.R, fit$D.R)
vl <- length(qtrmts)
q <- rep(0,vl)
for(i in 1:vl){
  q[i] <- x%*%R[[i]]%*%t(x)+sigma2
}
a <- 0.05
L <- fit$f-qnorm(1-a/2)*sqrt(q)
U <- fit$f+qnorm(1-a/2)*sqrt(q)
#plot prediction limits and mean of seasonally adjusted data
ts.plot(fit$f, fit.s$s[1:vl], L, U,
        ylim=c(min(fit$f, fit$s[1:vl], qtrmts, L),
               max(fit$f, fit$s[1:vl], qtrmts, U) ),
        lty=c(2,3,4,4), main=expression("One-step forecasting for the MelbTemp data"),
        xlab="Quarters", ylab="Temperature Celcius", lwd=c(1,2,1,1))
points(qtrmts, pch=4, col="red")
points(fit$f, pch=20, col="blue")
legend("topright", legend=c("Forecast mean","Data", "Smoothed forecast", "Forecast interval"),
       col=c("blue"), pch=c(20,4,NA,NA), lty=c(1,NA,3,4), lwd=c(1,NA,2,1), cex=0.8,
       title="Legend", text.font=4, bg='lightblue')

#error analysis
#residuals
e <- qtrmts - fit$f
#standardised residuals
estar <- e/sqrt(q)
#plot of standardised residuals
ts.plot(estar, main=expression("Standardised residuals
                               for the MelbTemp data"), xlab="Quarters", ylab="Residual")
points(estar, pch=20)
abline(h=1.96, lty=3)
abline(h=1.96, lty=3)
#other error measures
mean(estar)
```

```r
var(estar)
#MSE -affected by scale of obs
mean(e^2)
#MSSE - should be close to 1
mean(estar^2)
#MAD - affected by scale of obs
mean(abs(e))
#acf
acf(estar,
    main=expression("Autocorrelation function
                    (ACF) of the standardised residuals"), xlab="Quarters")
#sum estar^2 and compare with chi-sq dist with 40 df
sum(estar^2)
qchisq(0.95,vl)

#forecasting from 1991
fcast <- dlmForecast(mod=fit, nAhead=12)
fcast$f # beta k step ahead
fcast$Q # q k step ahead

sqrtR <- sapply(fcast$R, function(x) sqrt(x[1,1]))
pl <- fcast$a[,1] + qnorm(0.025, sd= sqrtR)
pu <- fcast$a[,1] + qnorm(0.975, sd= sqrtR)
w <- ts.union(window(qtrmts, start=c(1981,1)),
              window(fit.s$s[,1], start=c(1981,1)),
              fcast$a[,1], pl, pu)
plot(w, plot.type = "single", type="o", pch = c(1,0,20,3,3),
     col=c("black", "black", "brown", "blue", "blue"),
     ylab="Temp Celcius", main=expression("Observed and
      predicted temperatures for Melbourne"),ylim=c(14,31))
legend("topright", legend=c("Observed", "Smoothed(deseasonalised)",
                            "Forecast level", "95% probability limit"),
       col=c("black", "black", "brown", "blue", "blue"),
       pch = c(1,0,20,3,3), lty=1,
        text.font=4, bg='lightblue')

#create training data (1:20) for model
#then predict next 20 data values and
#then compare with actual data values to determine
#accuracy of prediction model

train <- qtrmts[1:20]
trainmod <- dlm(train, FF=x, GG=F, V=sigma2, W=Z, m0=beta0, C0=P0)
dlmLL(y=train, mod=mod)
trainfit <- dlmFilter(train, mod)
fcast21_40 <- dlmForecast(mod=trainfit, nAhead=20)
real21_40 <- qtrmts[21:40]
pred_diff <- as.vector(fcast21_40$f) - real21_40

avgp <- mean(pred_diff)
stdevp <- sd(pred_diff)
plot(pred_diff, main=expression("Prediction - Observed values for the k-th step ahead forecast"))
abline(h=1.96*stdev, lty=3)
```

```r
abline(h=0, lty=3)
abline(h=-1.96*stdev, lty=3)

#acf of differenced data
acf(pred_diff, main=expression("ACF plot of the prediction-observation differenced data"))
#paired t.test to determine bias between pred and obs values
t.test(as.vector(fcast21_40$f), real21_40, paired = TRUE )
```