An Analysis of Munic-Reim Summer Temperature Data

Michael Najarro

06/06/2020

```
library(pacman)
p_load(astsa, knitr)
```

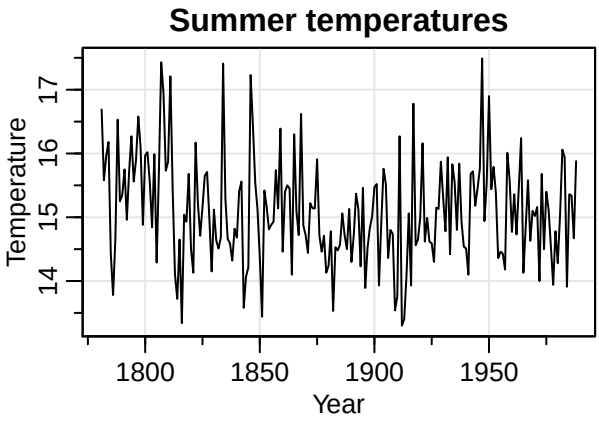
Introduction

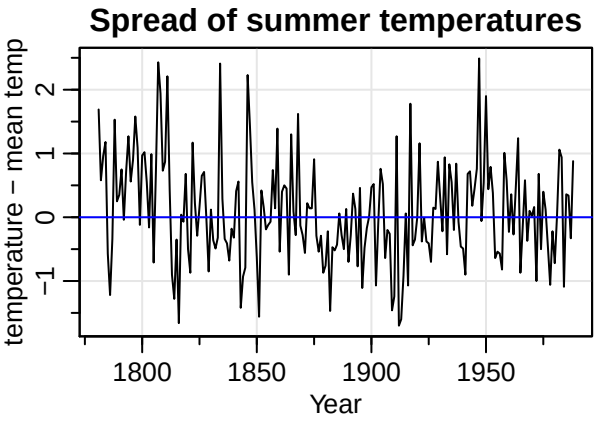
The summer data set contains mean summer temperatures (Celcius) across 153 days between the years 1781 to 1988 (208 years), measured in the Munich-Riem region of Germany. SInce each point represents one average seasonal temperature I consider no seasonal pattern in this analysis.

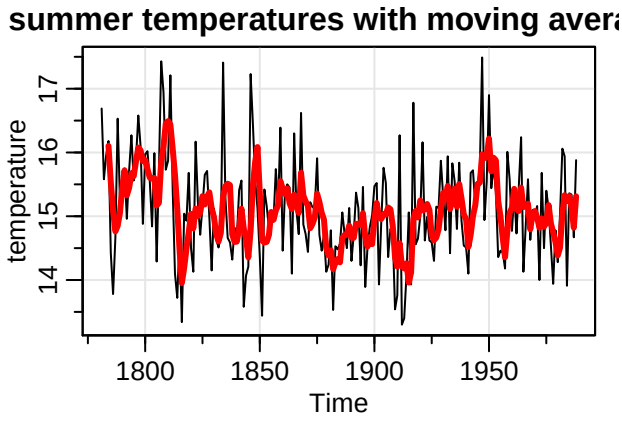
Initial Data Investigation

```
summer <- read.csv(file = "./data/summer.txt", header = FALSE)
summer <- ts(summer, start = c(1781,1), end = c(1988,1), frequency =1)
str(summer)
```

```
##  Time-Series [1:208, 1] from 1781 to 1988: 16.7 15.6 16 16.2 14.4 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr "V1"
```







An initial time series plot of the data yeilds noticable findings. The trend of the data appears to be complex in that multiple, different linear functions could describe different portions of the data; a decrease from 1781 to 1875, an increase between 1875 to 1950, and then a decrease from 1950 to 1988. The data does not appear stationary.

The average temperatures fluctuate between 13 to 18 degrees Celcius. The spread in temperature varies between a -1.5 to 2.5 degree change around 0 Celcius, however the volatility in temperature does not show a growth or reduction in its behavior over time.

Starting from the 25th year (about 1806) and roughly every 40 to 50 years after, unusual warm summers appear.

Applying a moving average to the data to further investigate cyclical patterns enahnaced slightly the 40-50 year cyclical pattern, but bore no new information on cyclical patterns. A noticable cooling period occured between 1850 to 1950. The equation used to smoothen the data is:

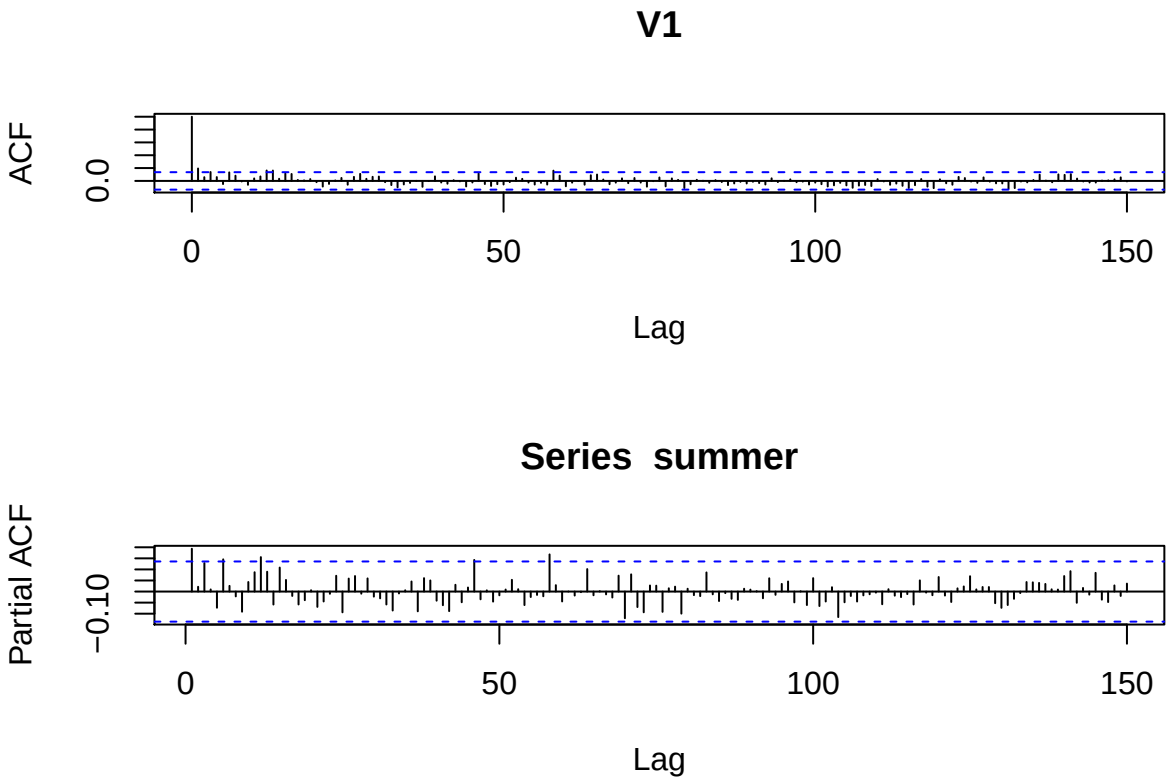
$$v_t = \frac{1}{4}(x_t + x_{t-1} + x_{t-2} + x_{t-3})$$

Model development

Stationarity Dvelopment

In order to develop a model around the temperature data, A look at plots of the auto-covariance functions and partial auto-covariance functions on the raw data show distinct patterns(ACF and PACF graphs not depicted in report).

```
par(mfrow = c(2,1))
acf(summer, lag.max = 150)
pacf(summer, lag.max = 150)
```



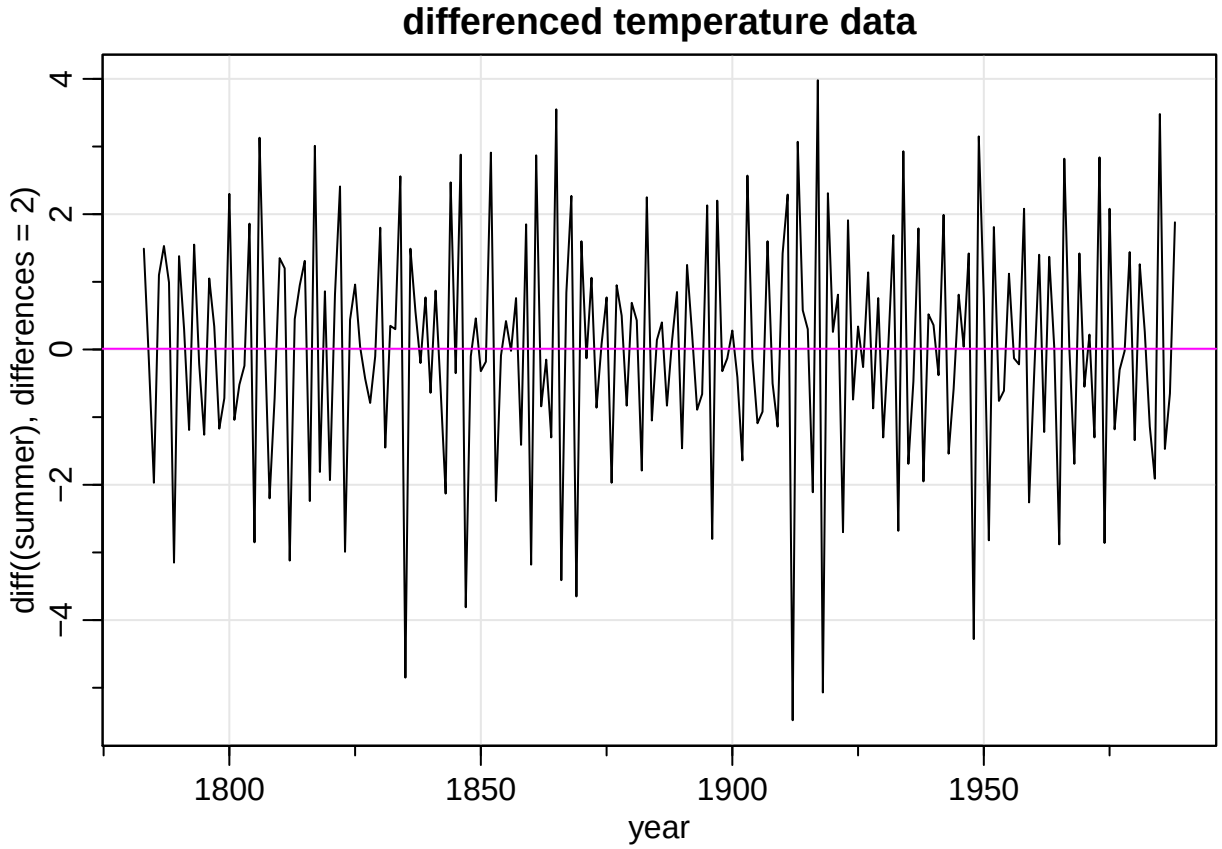
The ACF shows a drop off after the first year. however, there are several significant autocavariances when the lag is at 3, 12, and 13. There may be cyclical patterns. In the ACF, It appears that every 7 years from lag 1 to lag 30 there is some sort of oscilation occuring. This may by analagous to el nino events which occur every 5 to 7 years after a proceeding event, and their duration is roughly a few months up to two years. This pattern appears more noticable in the PACF between lags 10 and 50.

The PACF shows nearly a white-noise pattern except at lags 1, 6, 12, 46, and 58 being significantly larger than 0. Their appears to be no distinct pattern. Seasonality does not appear to exist in the data.

Because of the sharp drop in the ACF and the gradual decline in the PACF, a non-seasonal Moving Average model can explain the time series data.

Note that a transformation won't be necessary because the data does not show any trumpetting effects, or growth in volitility. To make the model stationary, differencing the data by 2 past measurements appears to de-trend the data towards a constant mean and a white-noise pattern.

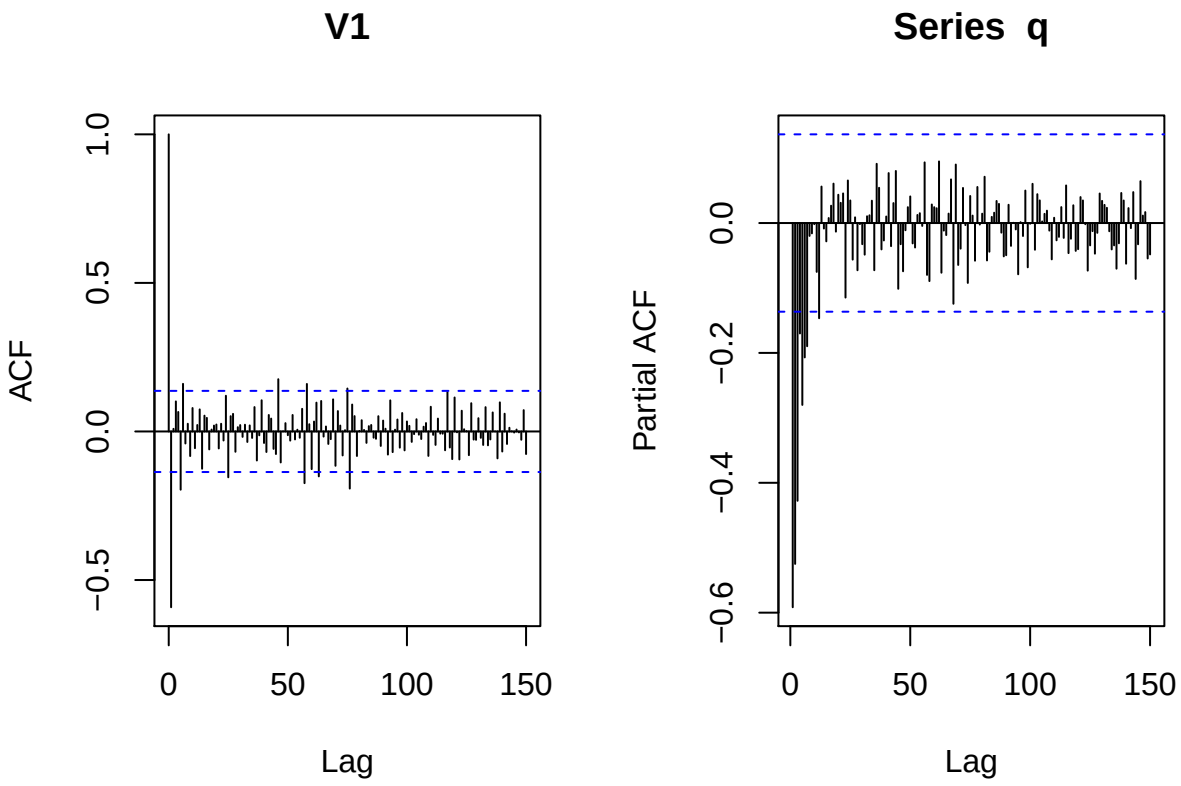
```
tsplot(diff((summer), differences = 2), main = "differenced temperature data", xlab = "year")
abline(h=mean(diff((summer), differences = 2)), col=6)
```



Model Development

A re-analysis of the ACF and PACF plots of the differenced data identify patterns typical of a Moving Average process.

```
q<- (diff((summer), differences = 2))
par(mfrow = c(1,2))
acf(q, lag.max  = 150)
pacf(q, lag.max = 150)
```



I conclude that an ARMA(0,2,3) Model applied to the differenced data most accruately describes the data. The Munich summer temperature can be modelled by an MA(3) process with a diffrence of 2 applied to the data in the following model:

$$x_t = (w_t - 1.7579w_{t-1} + 0.5597w_{t-2} + 0.1991w_{t-3})$$

Since this model does not include any Auto regression terms, this is the most reduced form of the model that can exist.

The model will be stationary, as the expecation of X_t will be zero due to the incorporation of white-nose terms within the model and the autocovariance will be the product of:

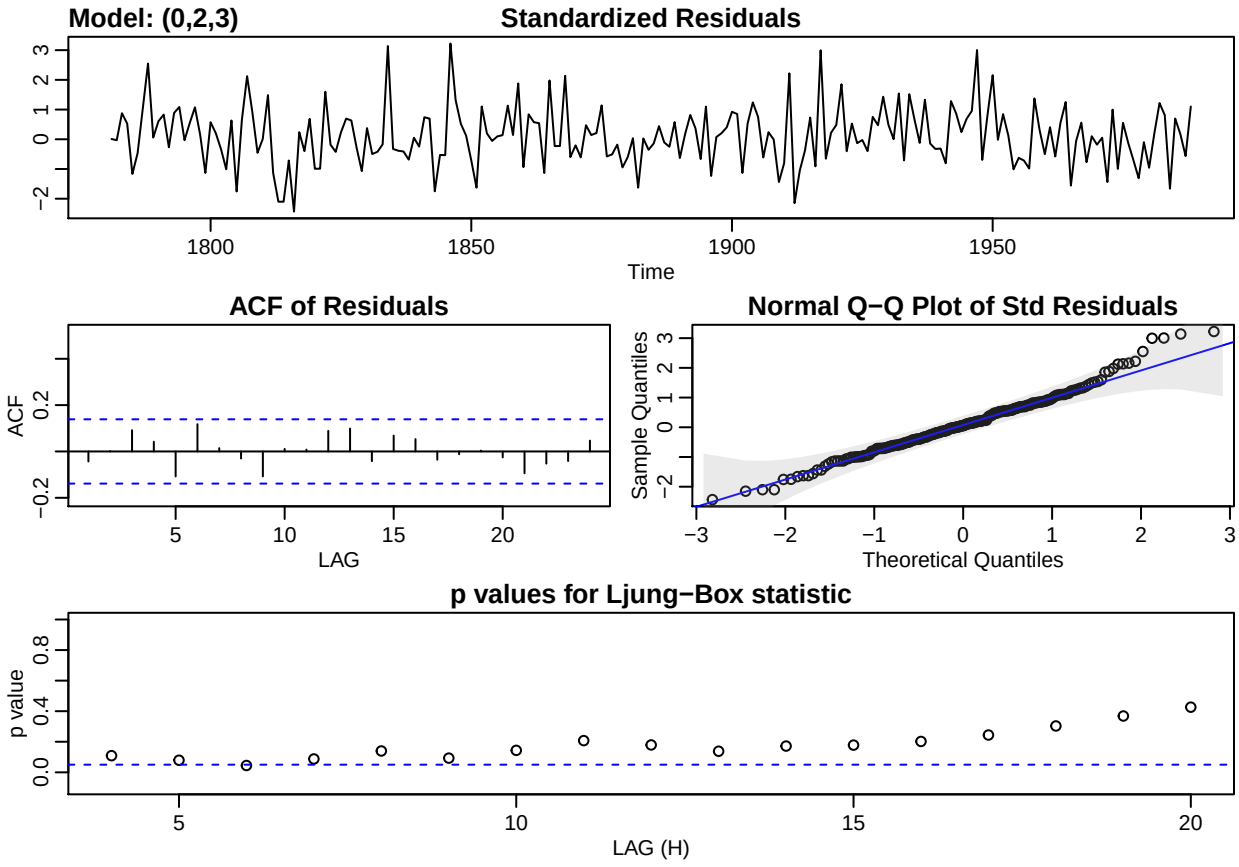
$$(w_s - 1.7579w_{s-1} + 0.5597w_{s-2} + 0.1991w_{s-3})((w_t - 1.7579w_{t-1} + 0.5597w_{t-2} + 0.1991w_{t-3})$$

which will produce coefficients that are not functions of time at any lag. Support for this model are noted in residuals that approximately follow a normal distribution (see qqplot of residuals), acf and pacf lag values not being significantly different from zero, and lag measurements from the Ljun-Box statistic not being signficantly different from zero for all but one lag.

```
h<- sarima((summer),0,2,3)
```

```
## initial  value 0.553352
## iter   2 value 0.232854
## iter   3 value 0.178409
## iter   4 value 0.084976
```

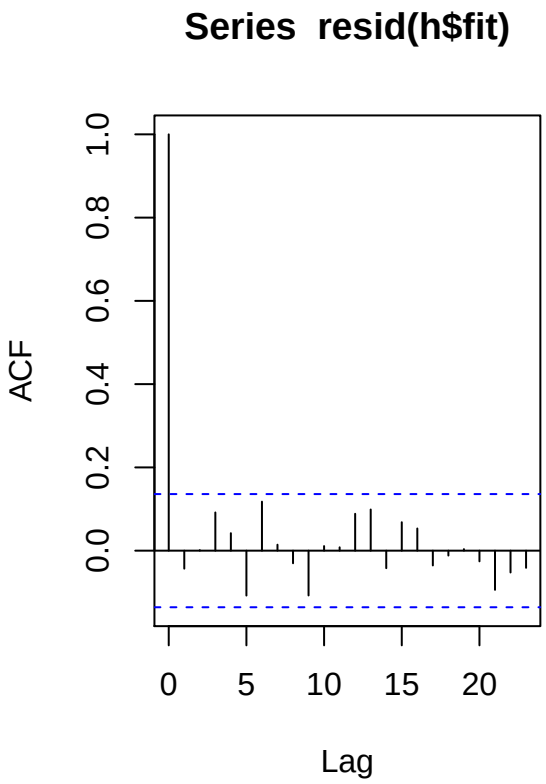
```
## iter   5 value 0.057966
## iter   6 value 0.009481
## iter   7 value -0.018108
## iter   8 value -0.025249
## iter   9 value -0.025879
## iter  10 value -0.033798
## iter  11 value -0.041194
## iter  12 value -0.041439
## iter  13 value -0.041445
## iter  14 value -0.041446
## iter  15 value -0.041446
## iter  15 value -0.041446
## final  value -0.041446
## converged
## initial  value -0.078564
## iter   2 value -0.113972
## iter   3 value -0.120004
## iter   4 value -0.166004
## iter   5 value -0.175369
## iter   6 value -0.179070
## iter   7 value -0.184068
## iter   8 value -0.185773
## iter   8 value -0.185773
## final  value -0.185773
## converged
```

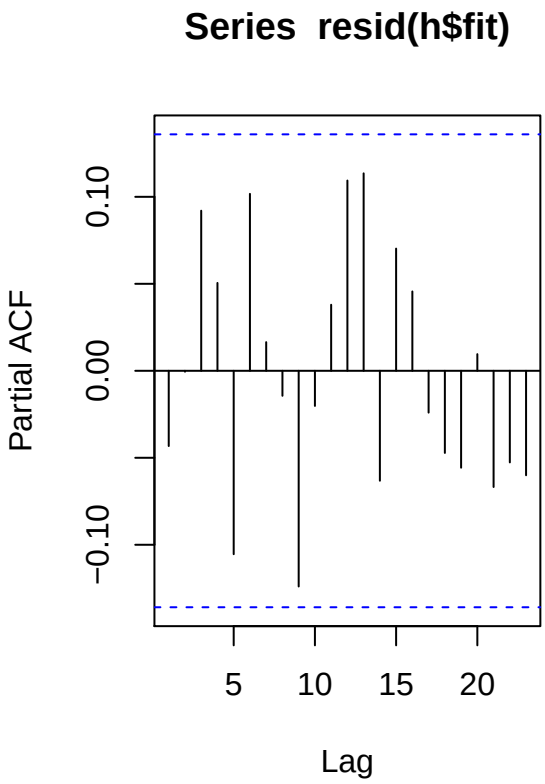


```
h
```

```
## $fit
##
## Call:
## stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
##     Q), period = S), include.mean = !no.constant, transform.pars = trans, fixed = fixed,
##     optim.control = list(trace = trc, REPORT = 1, reltol = tol))
##
## Coefficients:
##           ma1     ma2     ma3
##       -1.7579  0.5597  0.1991
## s.e.   0.1109  0.1777  0.0788
##
## sigma^2 estimated as 0.651:  log likelihood = -254.03,  aic = 516.06
##
## $degrees_of_freedom
## [1] 203
##
## $ttable
##     Estimate     SE  t.value p.value
## ma1  -1.7579 0.1109 -15.8480  0.0000
## ma2   0.5597 0.1777   3.1500  0.0019
## ma3   0.1991 0.0788   2.5256  0.0123
##
## $AIC
## [1] 2.505166
##
## $AICc
## [1] 2.505743
##
## $BIC
## [1] 2.569785
```

```
par(mfrow = c(1,2))
acf(resid(h$fit))
pacf(resid(h$fit))
```

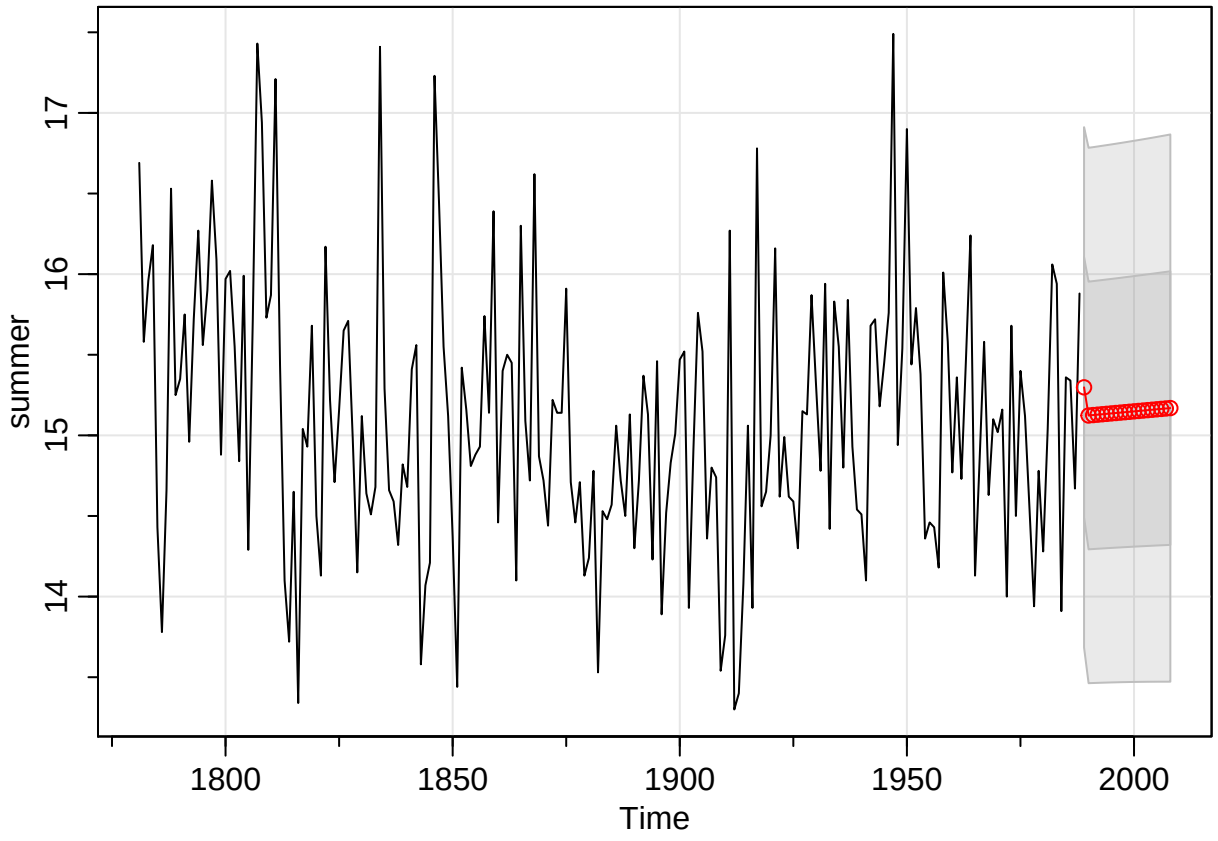

Series resid(h$fit)
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Lag


Series resid(h$fit)
Partial ACF
−0.10
0.00
0.10
5
10
15
20
Lag

Model Forecasting: Testing the Model

Forecasting the data 20 years into the future, the model simply predicts the average summer temperature (general trend of the data), and no spread of the variation in mean temperatures.The model cannot fully explain the complexity and spread of this data.

```
j<- sarima.for(summer, plot.all = TRUE, n.ahead=20,0,2,3)
```




```
j
```

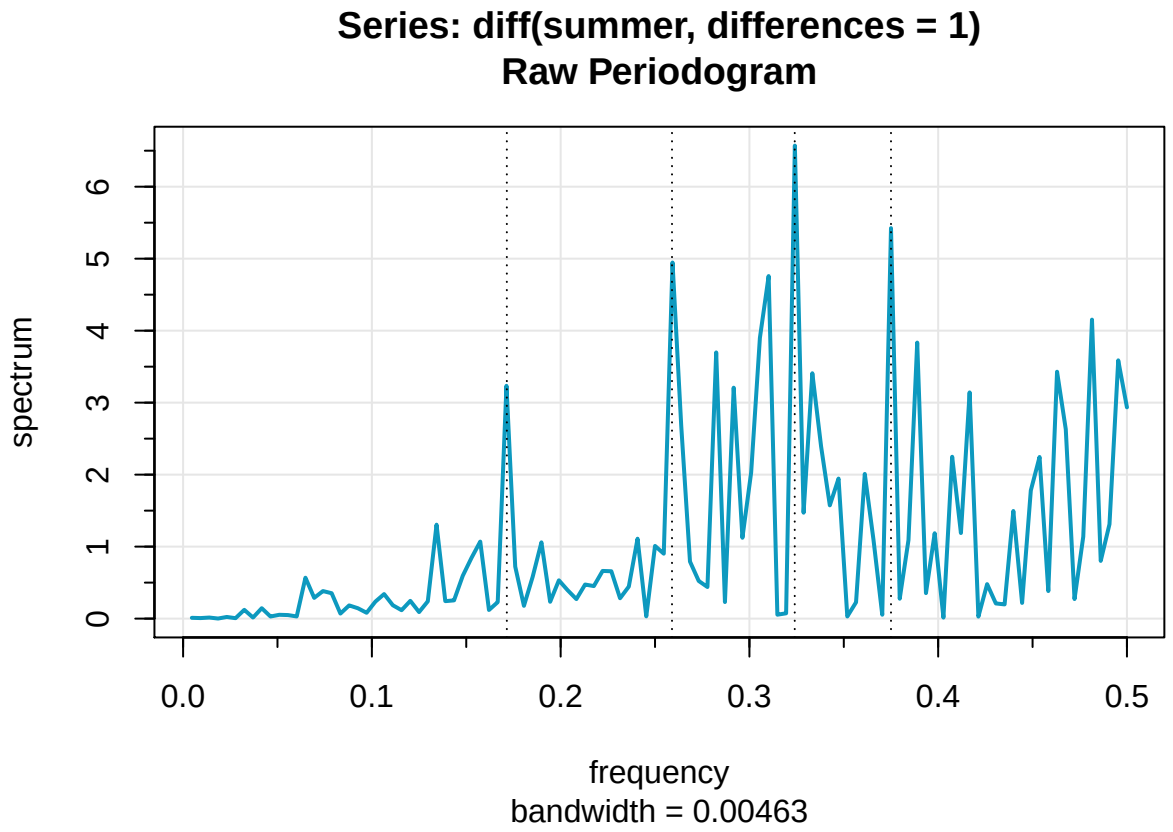
```
## $pred
## Time Series:
## Start = 1989
## End = 2008
## Frequency = 1
##  [1] 15.29799 15.12332 15.12587 15.12842 15.13097 15.13352 15.13608 15.13863
##  [9] 15.14118 15.14373 15.14628 15.14883 15.15138 15.15393 15.15648 15.15904
## [17] 15.16159 15.16414 15.16669 15.16924
##
## $se
## Time Series:
## Start = 1989
## End = 2008
## Frequency = 1
##  [1] 0.8068498 0.8301625 0.8309158 0.8316979 0.8325094 0.8333507 0.8342223
```

```
##  [8] 0.8351246 0.8360582 0.8370235 0.8380209 0.8390508 0.8401139 0.8412103
## [15] 0.8423407 0.8435054 0.8447048 0.8459394 0.8472095 0.8485155
```

Spectral Analysis

A spectral analysis can help understand which measurements dominate and explain the patterns within the series.

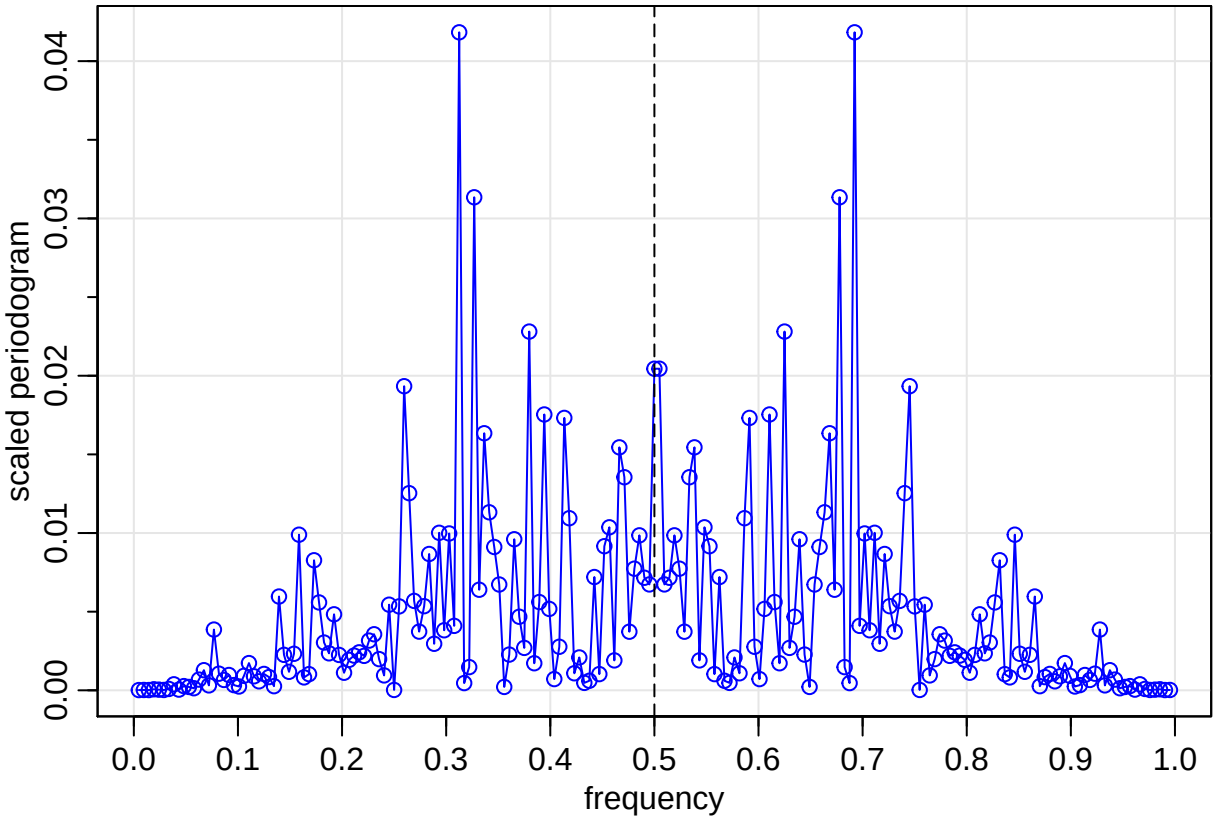
```
verano <- mvspec(diff(summer, differences = 1), col=rgb(.05,.6,.75), lwd=2)
abline(v=.1715, lty='dotted')
abline(v=.259, lty='dotted')
abline(v=.324, lty='dotted')
abline(v=.375, lty='dotted')
```




```
# code below gives values of frequency, period and spectrum.
#verano$details[37:100, ]
```

A scaled periodogram produces a similar result as the raw periodogram. Surprisingly most peaks are having some, albeit a tiny influence, in the series and are not scaling to 0.

```
n<- diff(summer)
P= Mod(fft(n)/sqrt(208))^2 # periodogram
sP= (1/208)*P #scaled periodogram
Fr = 1:207/208 # fundmanetal frequencies
tsplot(Fr, sP, type = "o", xlab = "frequency", ylab = "scaled periodogram", col = 4)
abline(v=0.5, lty= 5)
axis(side=1, at=seq(0.1,0.9,by=0.2))
```



Recall that frequency is the number of events occuring in a given time interval. Thus for each temperature measurement, its frequency would be the number of times a particular summer temperature occured across 208 years. However, since the same temperature measurement can repeat with the data, the frequency is multiplied by the reciprocal of the number of occurances. The periods represent the number of years to pass for a particular measure to be recorded at a later point in time.

Based on the raw periodogram, we can see that spectral densities rise rapidly after frequencies of 0.2. However the densities are highly volatile. This can be interpretted as the first and second MA process coefficients as eliminating any correlation between temperature measurements, while the last term may be producing a sufficient level of correlation between later terms to produce eratic and volatile noise between measurements.

The frequencies with the greatest spectral densities, along with their temperatures are:

```
# find the largest periods first
#library(dplyr)
y <- verano$details
y[,3]
```

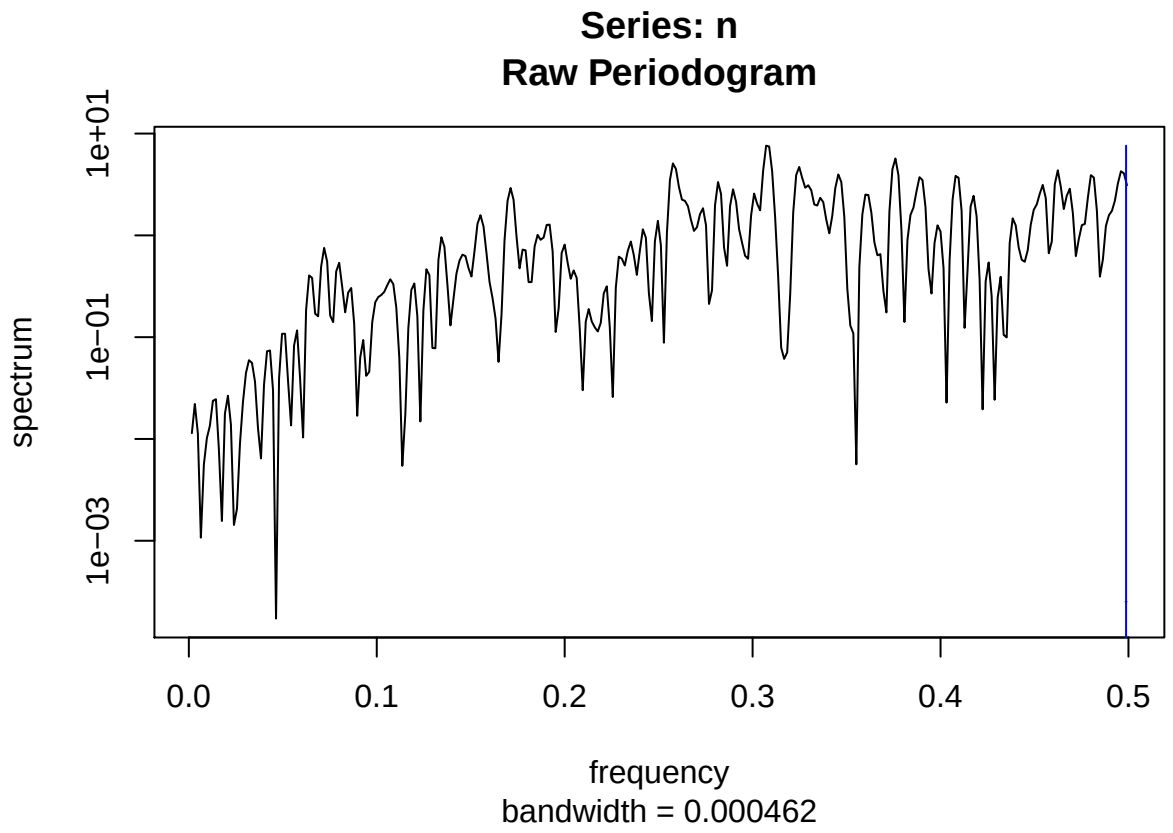
```
##   [1] 0.0110 0.0074 0.0149 0.0012 0.0231 0.0059 0.1219 0.0148 0.1444 0.0293
##  [11] 0.0531 0.0493 0.0299 0.5682 0.2882 0.3826 0.3527 0.0699 0.1833 0.1441
##  [21] 0.0810 0.2360 0.3407 0.1856 0.1168 0.2467 0.0908 0.2399 1.3052 0.2428
##  [31] 0.2535 0.5944 0.8439 1.0699 0.1198 0.2289 3.2338 0.7234 0.1775 0.5811
##  [41] 1.0594 0.2339 0.5320 0.3944 0.2713 0.4733 0.4520 0.6622 0.6579 0.2825
##  [51] 0.4452 1.1103 0.0309 1.0099 0.9020 4.9469 2.6772 0.7926 0.5231 0.4388
##  [61] 3.6977 0.2299 3.2079 1.1228 2.0310 3.8999 4.7584 0.0543 0.0736 6.5703
##  [71] 1.4712 3.4084 2.3929 1.5730 1.9457 0.0296 0.2266 2.0099 1.0931 0.0539
##  [81] 5.4244 0.2764 1.0905 3.8345 0.3541 1.1860 0.0142 2.2487 1.1888 3.1425
```

```
##  [91] 0.0278 0.4789 0.2106 0.1973 1.4950 0.2177 1.7800 2.2453 0.3826 3.4297
## [101] 2.6306 0.2715 1.1386 4.1533 0.8015 1.3112 3.5878 2.9344
```

```
# find the frequencies with greatest spectral desnities.
#years obtained by adding 37,67,70,81 to the starting year 1781.
k<-summer[c(37,67,70,81),]
yrs<- c(1817,1847,1850,1860)
kable(cbind(y[c(37,67,70,81), ],k, yrs))
```

frequency	period	spectrum	k	yrs
0.1713	5.8378	3.2338	15.04	1817
0.3102	3.2239	4.7584	16.44	1847
0.3241	3.0857	6.5703	14.37	1850
0.3750	2.6667	5.4244	15.40	1860

```
spec.pgram(n, taper = 0.1, pad = 2)
```



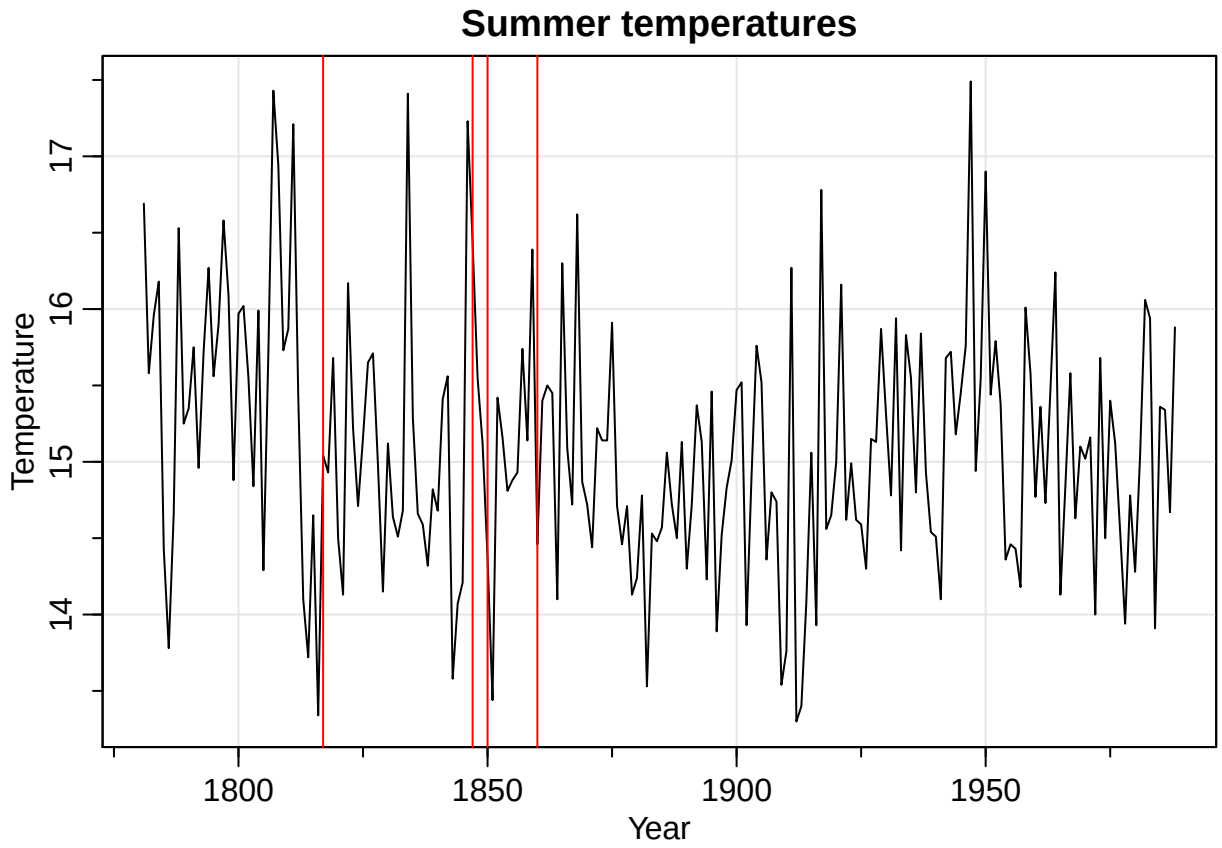
Discussion and Conclusion

The temperature data measured over the past 208 years is highly complex with few to no discenrable patterns. Based on the spectral analyses, we can conclude that the sumemrs with the greatest influence on the time

series had frequencies or rates of occurance at 0.3 events per year, with periods somewhere between every 2 to 5 years; lag measures were every 15 years.

I Initially suspected that the warmest temperatured years in the data would have had the largest effect in the data. Surprisingly, the summers with cooler temperatures (troughs within the time series, excluding 1847) had the greatest effect on the behavior of the series, as identified in the periodogram:

```
tsplot(summer, main = "Summer temperatures", xlab= "Year", ylab = "Temperature")
abline(v=1817, col = "red")
abline(v=1847, col='red')
abline(v=1850, col='red')
abline(v=1860, col='red')
```



A variety of sources could have contributed to model complexity. Between the years of 1300 to 1800, Europe was locked in a "little Winter," where global temperatures had decreased for small periods of time. Such intervals could explain the decreasing trend in temperature between 1781 to 1850.

The data lacks multiple measures of centrality per year. Increasing the number of average samples per summer could help to reduce the volatility of the data, reduce the amplitudes in spectral analysis to produce more accurate measures, and contribute to increasing frequency measures. I recommend investigating the raw data to average multiple temperature sub-samples per year and to re-analyze the data.