**Gabriel Tellez**

**Take Home Exam 2 – Time Series Analysis**

**DA 6823**

**Exam 2 – Time Series Analysis**

1. First data set without seasonality:

Number of earthquakes per year with magnitude 7.0 or greater from 1900 – 1998

<https://datamarket.com/data/set/22p8/number-of-earthquakes-per-year-magnitude-70-or-greater-1900-1998#!ds=22p8&display=line>

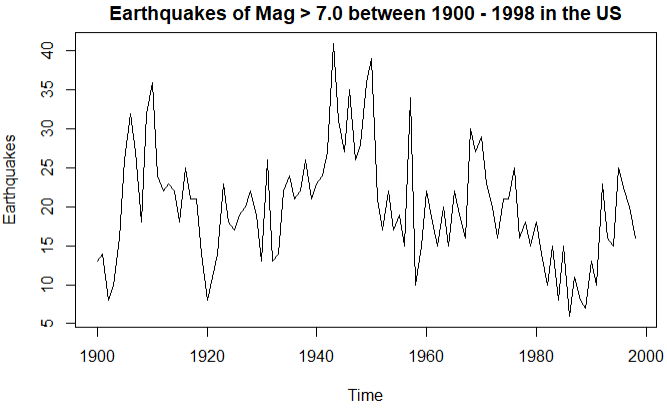
Second data set with seasonality:

Temperature (in Kelvin) in San Antonio from 01/01/13 until 11/30/17 taken every hour of every day. I made data start in 2013 instead of original 10/01/12

<https://www.kaggle.com/selfishgene/historical-hourly-weather-data>

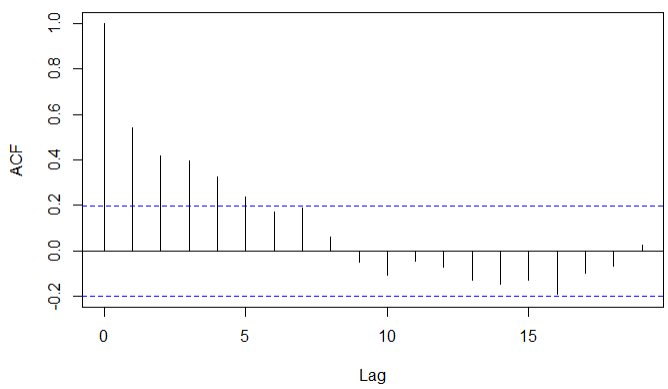
**Without Seasonality**

1. Plot of time series variable.



The time series data is NOT stationary by just using the eyeball test. The mean and variance vary throughout the time series. For example, there is an increase in the number of earthquakes during the 1940s and 1950s and then a decrease at the end of the 1970s and going throughout the 1980s. The mean is closer to being constant if small time periods are looked at but not for the entire data set. The variance is not constant.

1. ACF – Earthquakes Data Set



The ACF plot of the earthquakes data set indicates a non-stationary series because the decay is gradual and stays above the significance range up until lag 5.

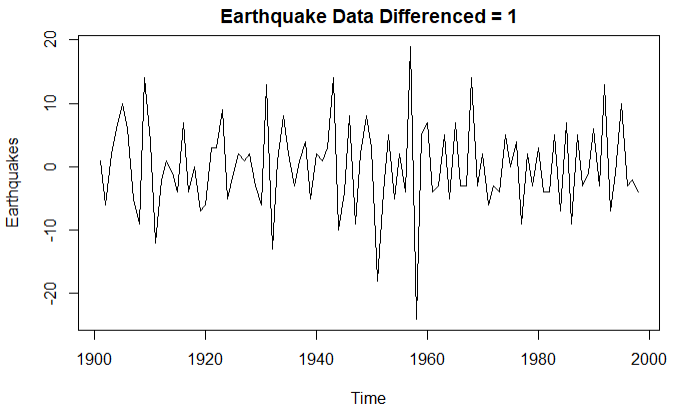
1. KPSS and ADF Tests

The KPSS test for trend stationarity presented a p-value smaller than 0.01 (the limit for the KPSS test in R). The null hypothesis for the KPSS test is that the data is stationary. We can accept the null hypothesis if the p-value is greater than the chosen alpha level, in this case 0.05. For the earthquake data, the KPSS test suggests the data is non-stationary (since p-value 0.01 < alpha = 0.05).

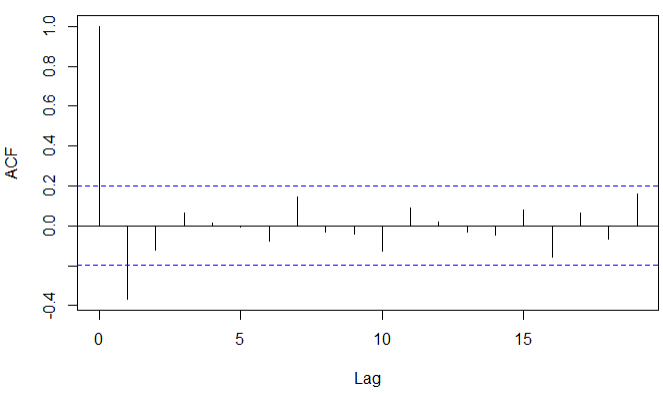
The ADF test null hypothesis states the data is non-stationary. The null hypothesis can be rejected if p-value < alpha (chosen to be 0.05). For the earthquake data, the ADF test presented a p-value of 0.08583. Thus, the null hypothesis is accepted, and we can state that the earthquake data is non-stationary.

Both tests suggest that the earthquake time series data is non-stationary, which means the data does not have a constant mean.

1. Differencing
2. After differencing the data once, the data seems to have a constant mean. The estimated number of differences suggested by the NDIFF function in R returned 1 difference.



1. The ACF plot for the difference time series shows that there is no trend.



1. KPSS and ADF tests to differenced time series

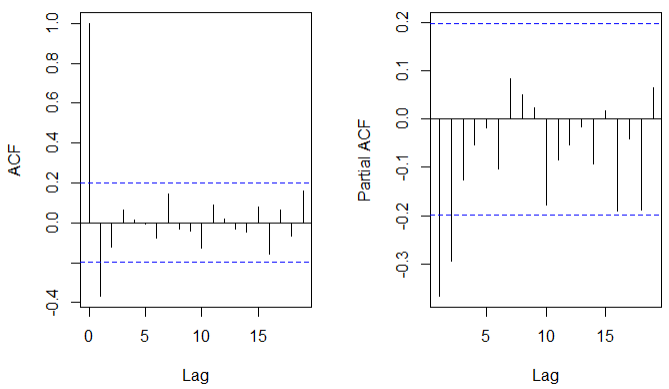
After differencing the time series, the KPSS test returns a p-value greater than 0.1 which means the null hypothesis is accepted. The KPSS test now shows that the differenced data is stationary. The trend has disappeared.

The ADF test returns a p-value smaller than 0.01 which means the null hypothesis is rejected. The ADF test suggests that the data is now stationary.

1. ARCH Test

The differenced earthquake time series data returned a p-value of 0.79 with the ARCH test which shows that the data has constant variance and it is stationary.

1. PACF on differenced time series



From the ACF plot, the sharp cutoff after lag 1 points to a MA(1) process since the sharp decrease is after the first lag. From the PACF plot, there is no sharp cutoff which means there is no evidence of an AR process. The PACF does increase but it is not an exponential change from one lag to the next. This would make the following model: ARIMA(0,1,1) because of the 0 AR processes, the differencing of 1, and 1 MA process.

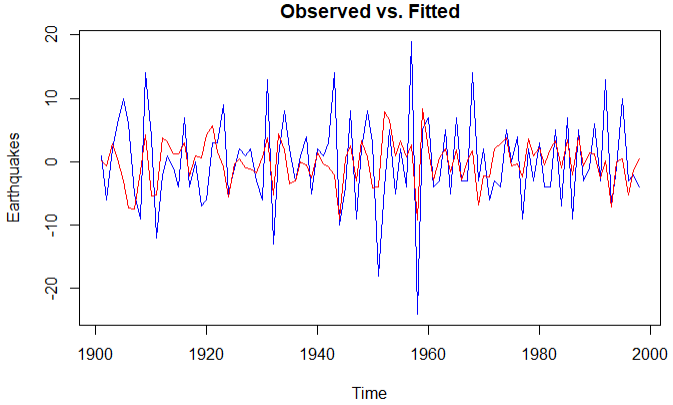
To check the interpretations of the ACF and PACF, I used the auto.arima function in R which finds the optimal time series model for the data. The auto.arima function returned ARIMA(0,0,1) with the differenced earthquake time series. So there is more evidence of a MA(1) process and of no AR processes in this time series.

1. Experimenting with different ARIMA models
2. Earthquake Differenced = 1

|  |  |  |  |
| --- | --- | --- | --- |
| MODEL | AIC | BIC | Ljung-Box |
| (0,0,1) | 639.04 | 646.79 | 0.7692 |
| (1,0,1) | 640.69 | 651.03 | 0.9992 |
| (1,1,1) | 648.45 | 656.17 | 0.2631 |
| (1,1,2) | 641.5 | 651.8 | 0.9139 |
| (0,0,2) | 640.58 | 650.92 | 0.9218 |
| (0,1,2) | 639.74 | 647.46 | 0.8838 |

For the AIC and BIC metrics, the smaller the better. For the Ljung-Box a p-value greater than 0.05 means the model fits the data well. All the model tested above had at least 1 MA process. Without a MA process, the goodness of fit and parsimony of the model would decrease meaning that the AIC and BIC increased. The p-values of the Ljung-Box test were still non-significant (>0.05) meaning that there were still no signs of dependence of values on each other. So the main takeaway from this time series is that there is definitely one MA process one the data has been differenced once.

1. Observed vs. Fitted

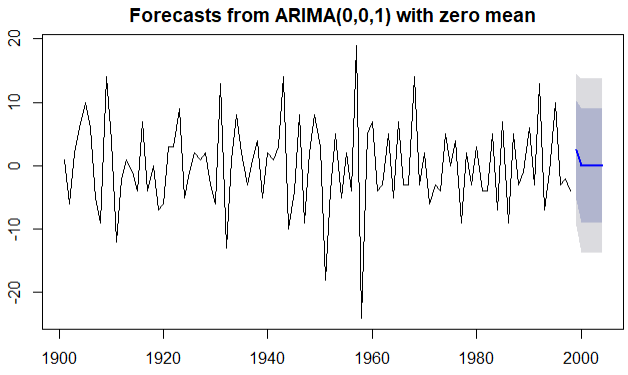


The observed data is in blue and the fitted data is shown in red. I think the fitted data for the time series is following the observed data very well. There are no drastic changes. If anything, there seems to be a small negative lag. The fitted data has its peaks just a little behind the observed data.

1. Favorite Model

As my favorite model, I definitely pick the ARIMA(0,0,1) model because it has the best fit and the most parsimonious. Like all other models, there are no signs of dependence of lag values on each other.

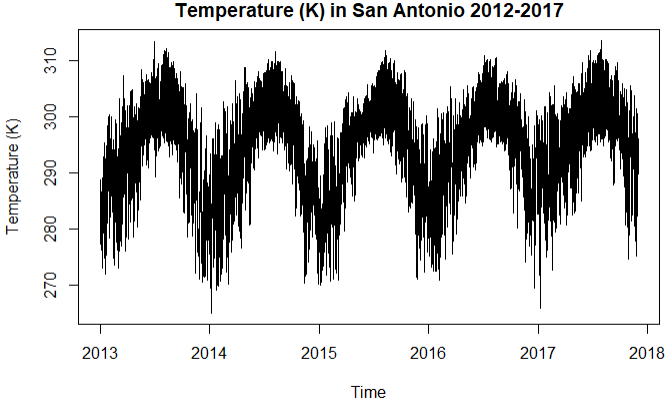
1. Forecasting Favorite Model



After forecasting the ARIMA(0,0,1) model for the next 6 time periods it seems like the forecast is good for the first one or two years because there is some change (decreasing). However, after the initial one or two years the forecast literally flat lines. That definitely does not look good. I am no earthquakes expert but I think having the same number of earthquakes for several years is very unlikely.

**With Seasonality**

1. For this section, I am using the data set with seasonality. The original data set includes temperatures from 30 different cities starting in 10/01/12 until 11/30/17. However, I focused on temperatures from San Antonio and started from 01/01/13. Temperatures are in Kelvin. For reference, 270 K ~ 26 °F and 310 K ~ 98 °F.

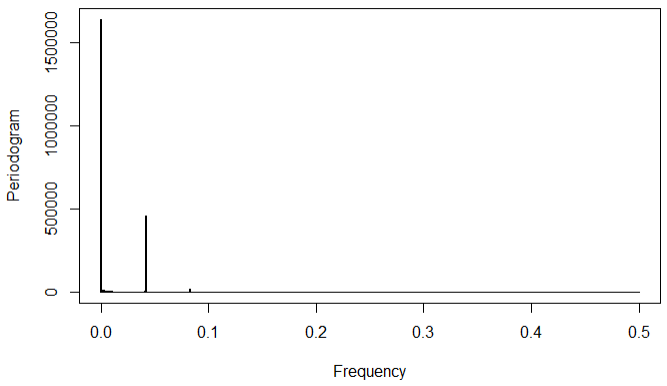


From the plot above, I suggest a type 2 Holt Winter model should be applied: additive Holt Winter model. There is no apparent trend and the seasonality is very apparent (which is expected from the seasonal changes).

b) Eyeball Period Size

The period is 1 year for this data set. I would be worried if it was different. The reason why it is a 1-year period is because of the yearly seasonal changes due to the Earth’s trip around the Sun. We can clearly see the lower temperatures at the end and beginning of every year due to Earth’s tilt away from the Sun in the northern hemisphere. The hot Texas summer days are also on full display in the middle of every year. An interesting pattern is that the transition between the coldest season and the hottest season is longer and more gradual than the transition from hot to cold. In other words, spring seems to be a gentle transition between winter and summer while autumn is more drastic in terms of temperature changes.

c) Periodogram with R



From the periodogram using R, there are two spikes. The first is close to 0 Hz and the second is around 0.05 Hz. The following table shows the exact numbers:

|  |  |
| --- | --- |
| Freq | Spec |
| 0.0001157407 | 1638551 |
| 0.0416666667 | 458730.8 |

To get the period, the following equation is used:

This translates to the following periods:

|  |  |
| --- | --- |
| Periods (Hours) | |
| 8640 | 24 |

So the main seasonality is detected at 8640 hours which is close to the 8760 hours in one year. The second seasonality is every 24 hours. Both are expected periods. The 1-year period is obvious by looking at the time series plot. The 24-hour period is expected because the data is collected every hour for every day of the year.

d) KPSS or ADF test for trend

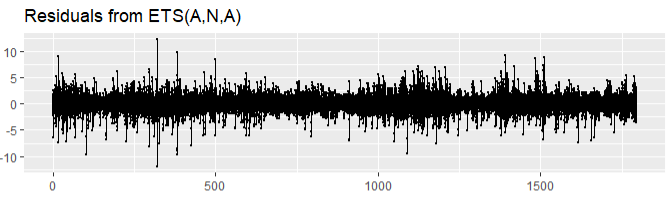
I applied the ADF test to check for trend. The test returned a p-value < 0.05. For the ADF test this means that the data is stationary so there is no trend.

e) Weights for the three components of the Holt Winters smoothing

I used the ETS function in R to find the optimal weights for the components of the Holt Winters smoothing process. This function uses alpha = error, beta = trend, and gamma = seasonality. The following table shows the optimal weights:

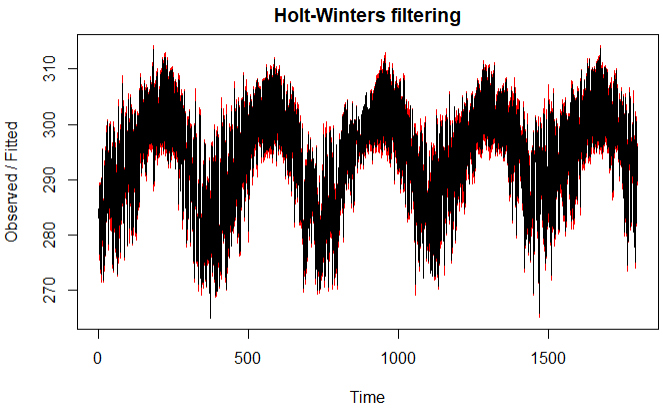
|  |  |  |
| --- | --- | --- |
| α | β | γ |
| 0.9558 | 0.000 | 0.0095 |

Both the error and seasonality components are additive. The residuals stay constant over time. This indicates that the additive model is appropriate. The model has additive errors and additive seasonality with no trend.



f) Holt Winters model

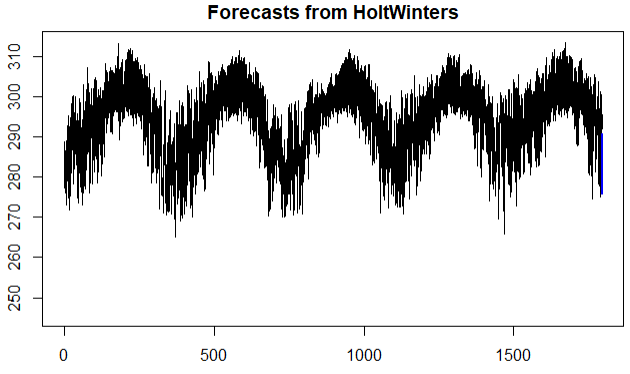
The following Holt Winters model includes the alpha and gamma components found in the previous section, the beta components was set to NULL so the HoltWinters function in R does exponential smoothing, and seasonality was specified to “additive”.



The time series had to be adjusted to find the optimal components for the model. Initially, the time series was plotted with a frequency of 8760 which is the number of hours in one year. However, the ETS function in R cannot process data with such a large number of frequencies. The function has a limit of 24 frequencies. So, the frequency was changed to 24 hours. As seen in the x-axis of the Holt Winters model above, that modification in frequency changed the axis to days instead of years. That is the only difference.

The plot above shows the actual data in black and the fitted data in red. Throughout the plot, the Holt Winters model did very well. The discrepancies are so small that the red line for the fitted data is barely visible.

Forecasting 12 periods



The plot above shows the original time series with a forecast of 12 periods. That is 12 days and it is barely visible at the end.