VEE Time Series Student Project Write-Up

By Seah Chun Leong, Leon (AXA Healthcare Management)

Note: This write-up should be read with its accompanying excel spreadsheet.

1) Introduction

In this project, we fit an ARIMA model to the Global land-ocean temperature index, 1880 to present, with base period 1951 to 1980. The data used in this project is publicly available on the NASA giss website (http://data.giss.nasa.gov/gistemp/graphs_v3/Fig.A2.txt). The goal of this project is to investigate the relationship between the global average temperatures of adjacent years, that is, the relationship between temperature of year x and the temperature of year x-1, x-2, ... , x-n. This relationship will be based on ARIMA. We shall use this model to visualise the dynamics of global warming.

2) Data

The time interval between each data point is one year. There are a total of 134 data points (year 1880 to year 2013). We partition this set of data into "in-sample" (year 1880 to year 1999 -> 120 points) and "out-of-sample" (year 2000 to year 2013 –> 14 points) sets. "In-sample" data shall be used for parameter fitting; "Out-of-sample" data shall be used for residual analysis.

Let t be year and y(t) be the annual mean surface temperature in degrees Celsius. A preliminary assessment on the "in-sample" data suggests that MA(1) and ARIMA(0,1,0) might be good fits.

Figure 1: A plot of y(t) against t and y(t+1) against y(t). y(t+1) and t shows a somewhat positive relationship with y(t).

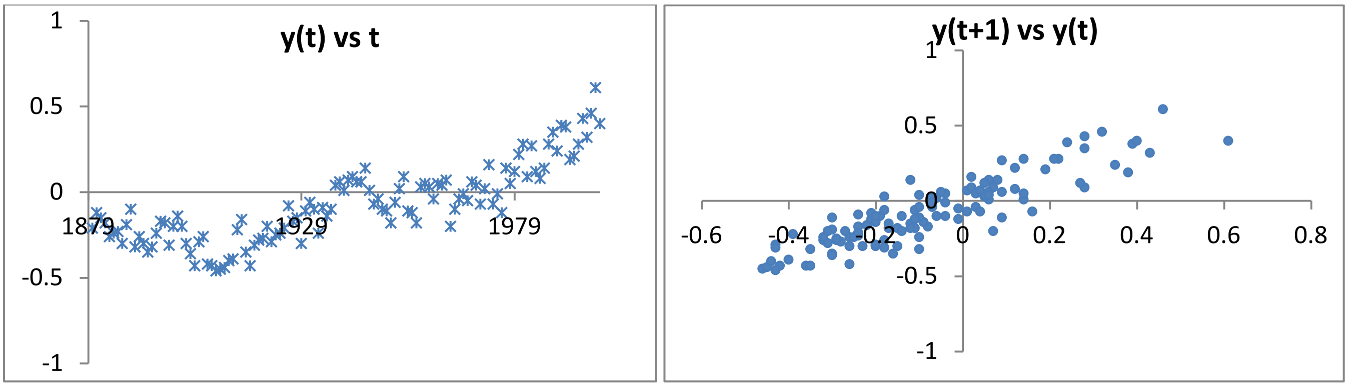
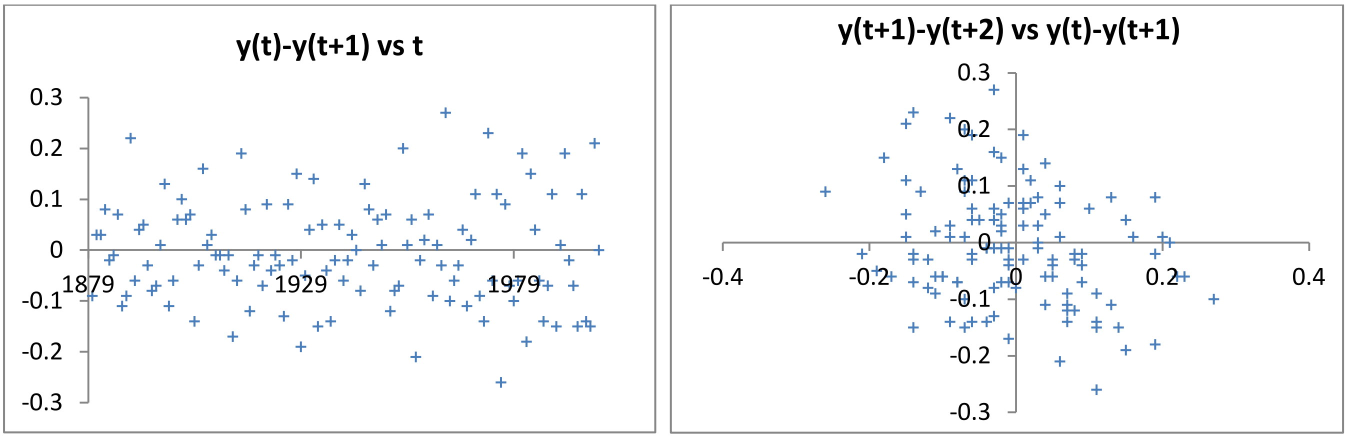


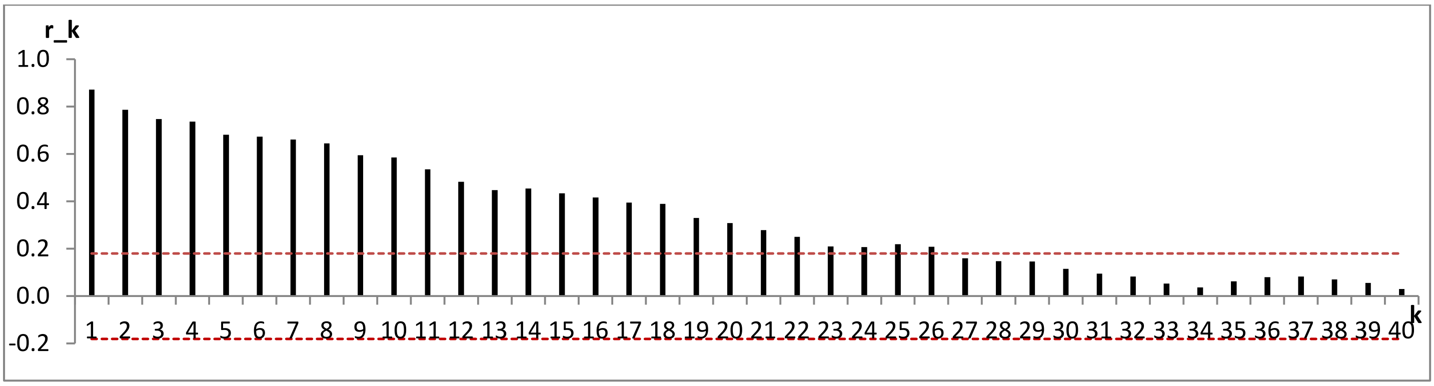
Figure 2: A plot of y(t)-y(t+1) against t and y(t)-y(t+1) against y(t+1)-y(t+2). These suggest that the first difference might be white noise.



3a) Correlogram

We start by plotting the sample autocorrelation function (r_k) against lag (k). Clearly, the time series is not stationary; r_k dies off very slowly, only to become statistically insignificant at lag 27. The statistical significance (red dotted line in the correlogram below) is calculated based on the t-value at 95% level and the standard deviation of white noise for a sample size of 120.

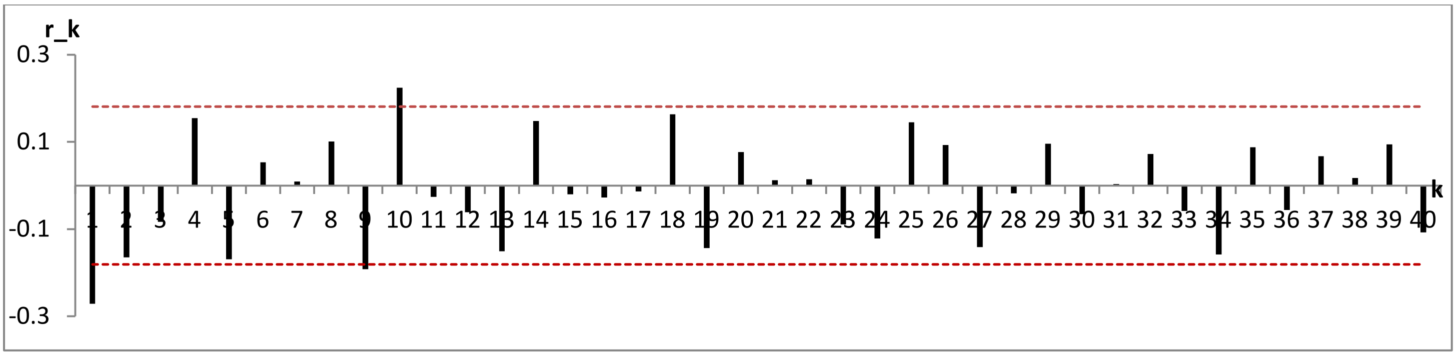
Figure 3: Sample ACF of Earth's annual mean surface temperature. The time series is likely to be nonstationary.



3b) Correlogram

By the Principle of Parsimony, we reject the notion of a complicated ARMA(p,q) for some large p and q. This leads us to consider the first difference, y(t)–y(t+1). True enough, we arrive at a stationary process, no obvious seasonality though.

Figure 4: Sample ACF of the first differences of Earth's annual mean surface temperature. The first difference is a stationary process.



4) Model fitting & Checking

We fit an ARIMA(0,1,1) process to the Global land-ocean temperature index. The model is:

$$y(t) - y(t-1) = \mu + e(t) - \theta\, e(t-1)$$

The drift μ is estimated by taking the mean of the first differences of the "in-sample" data. The IMA parameter θ is estimated using method of moments, that is, equating the sample autocorrelation function to the associated MA(1)

Yule-Walker equation. Our estimates for μ = -0.005126 and θ = 0.2945099. This yields the following ARIMA(0,1,1) process:

$$y(t) – y(t-1) = -0.005126 + e(t) – 0.2945099\ e(t-1)$$

Last but not least, we embark on the residual analysis. We first forecast the annual mean surface temperature into the “out-of-sample” period (year 2000 to year 2013).

Table 5: A comparison between the forecast made using the ARIMA(0,1,1) model and the actual values.

t	Actual value (℃)	One period ahead forecast (℃)	Residue
2000	0.4	0.394874	0.005126
2001	0.52	0.393364	0.126636
2002	0.61	0.477578	0.132422
2003	0.6	0.565875	0.034125
2004	0.51	0.584824	-0.07482
2005	0.66	0.52691	0.13309
2006	0.59	0.615678	-0.02568
2007	0.62	0.592436	0.027564
2008	0.49	0.606756	-0.11676
2009	0.59	0.51926	0.07074
2010	0.66	0.56404	0.09596
2011	0.54	0.626613	-0.08661
2012	0.57	0.560382	0.009618
2013	0.59	0.562041	0.027959

Figure 6: QQ plot of the residuals. Notice how the data points oscillate around the y = x line. Stochasticity dominates in our small sample size of 14.

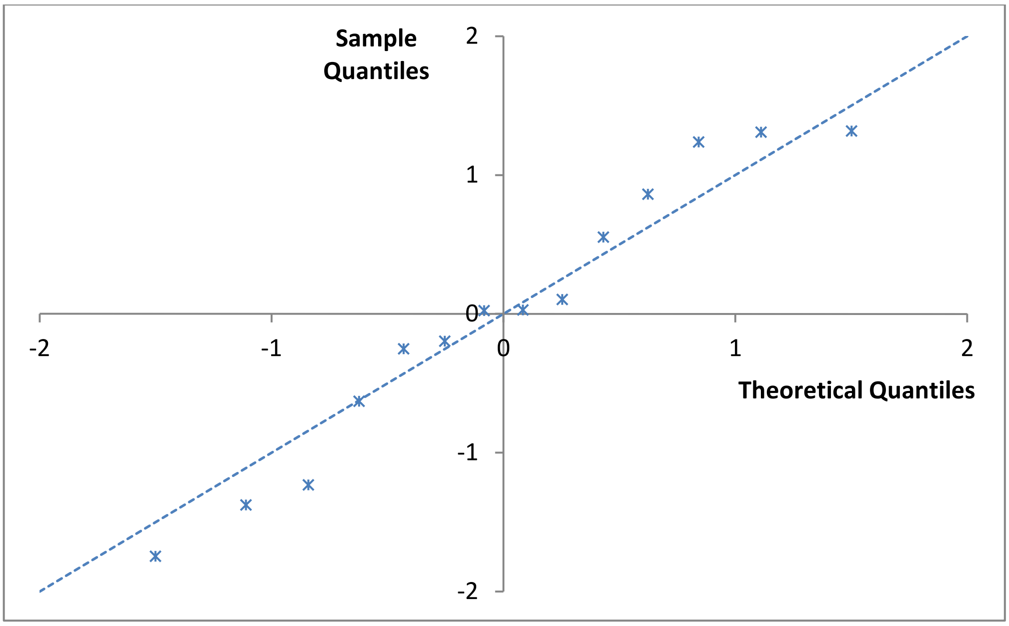
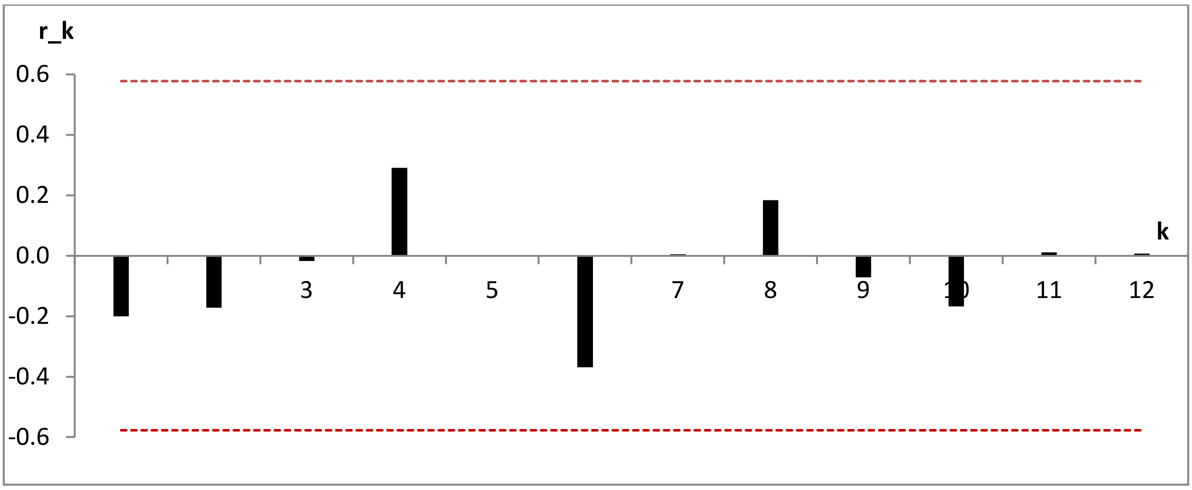


Figure 7: Correlogram of the sample autocorrelation function of the residuals. Notice that the values are all statistically insignificant even though they are large in absolute value terms. This is another example of how stochasticity obscure statistical interpretation when sample size is small.



Finally, we test at 95% significant level H_0 : The error terms are uncorrelated; against H_1: error terms are correlated.

Table 8: We calculate the Ljung-Box test statistic using a lag threshold of K=7.

K	7
Sample size	14
Degrees of freedom	6
Ljung-Box test statistic (Q)	6.94634783
Critical value	12.591587
Result	Do not reject H_0

5) Conclusion

The dynamics of Earth's annual mean surface temperature can be modelled by:

$$y(t) - y(t-1) = -0.005126 + e(t) - 0.2945099\ e(t-1)$$

The above is subject to a multitude of interpretations. One way of reading the equation is that the mean temperature in succeeding year y(t) is the sum of the previous year's temperature y(t-1) and the random terms (caused by both man-made and natural processes) from the current year and the previous year. This is in line with known scientific theories on the retention of greenhouse gasses in the atmosphere.