STA2202 Final Project: Time Series Analysis on Global Temperature

Abstract

Global warming is more and more important global issue especially for this summer, there are many countries and areas have reached high temperature records that occur only once in decades or even hundreds of years. To investigate global temperature and improve public's awareness of environmental protection, we will build a reliable model to predict future temperature changes. In this report, we will use R tools and statistical theory to analyze the world temperature from 1880 to 2009. Firstly, we introduce `gtemp2` dataset from library(`astsa`) and tried to establish two ARIMA model with the help of ACF and PACF. After testing the significance of the model perform diagnostics, we also select the model based on performance to predict the future values and the conclusion is that there is increasing trend on global temperature over years which remind us of the global warming issues.

To make the fitted model better, we can consider introduce more dataset (including temperatures from 2010 to 2021) and try more models for further research.

Keywords: global temperature, ARIMA, global warming, prediction, spectrum analysis

Introduction

Studying global temperature is a very important sub-task for solving global temperature warming. If there is a reliable model to predict future changes in global temperature, it will be very helpful for researchers to investigate global temperature warming. We used the dataset of `gtemp2` in the library of `astsa`. This dataset includes the world temperature records from 1880 to 2009. There are totally 130 data points, which enough to help us build a reliable statistical model for prediction.

In short, this report will use these data to build a reliable prediction model, detection model and prediction for global temperature, and perform spectrum analysis.

Statistical Methods

In this part, we will do (1) explore the data and check whether it's stationary, (2) apply transformation if it required, (3) identify the dependence orders of ARMA models.

Firstly, we should explore the yearly time series plot of global temperature. As Figure 1 shown, there is a generally upward trending on global temperature over these years instead of floating around the mean.

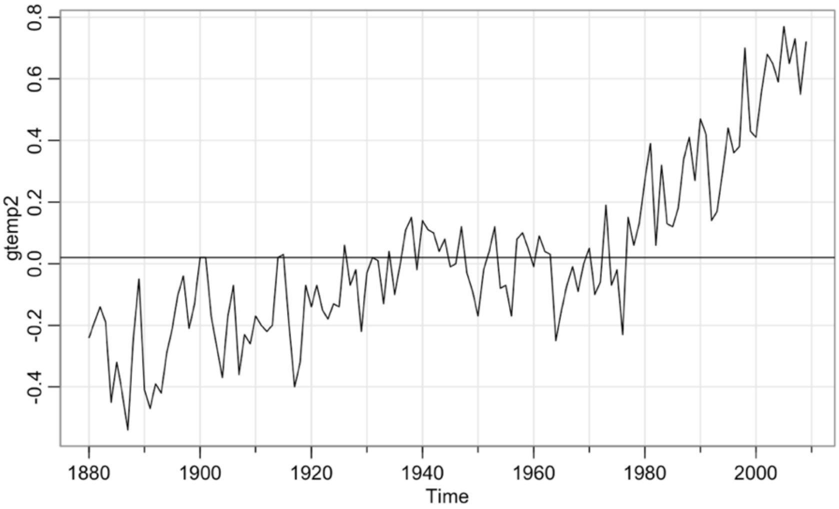

Figure 1: The yearly time series plot of global temperature from 1880 to 2009

From Figure 2, it's really easy to conclude that it's not stationary process cause the ACF shows a very slow decay to zero as lag increases. And by the conclusion from Figure 1, we may need to do transformation on global temperature to get a stationary process.

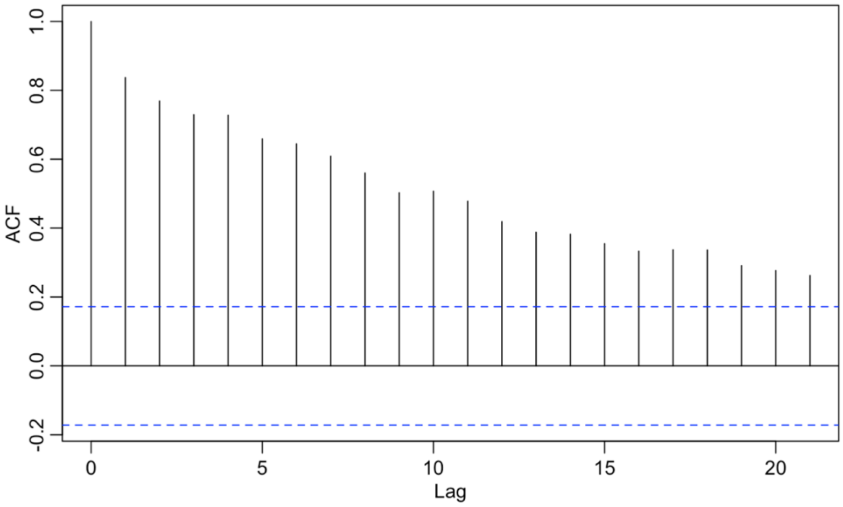

Figure 2: Sample ACF for global temperature (Xt)

By differencing the global temperature, we get a new time series dataset and Figure 3 shows the yearly time series plot of differences of global temperature (∇Xt).

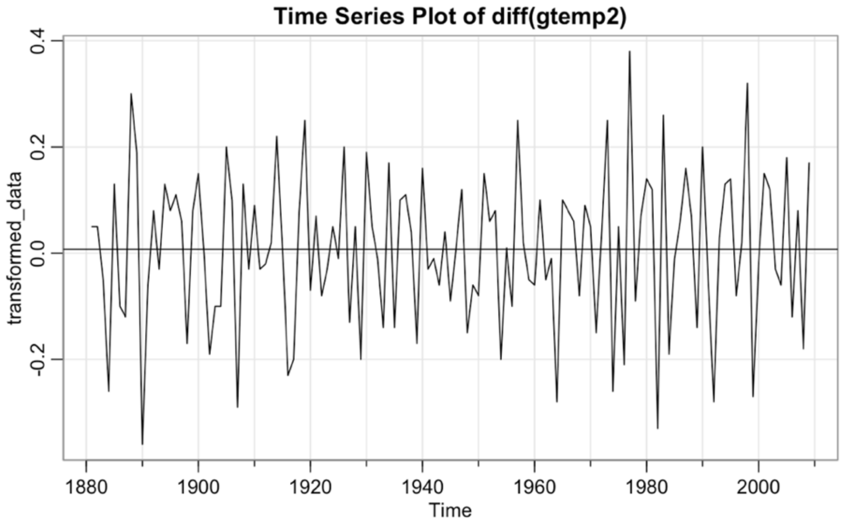

Figure 3: The yearly time series plot of ∇Xt (differences of global temperature)

Based on Figure 3, it shows the differences of data keep floating around mean and have same variance over these years and it looks much better than our Figure1. But we may still need to check ACF and PACF to check its property and fit model.

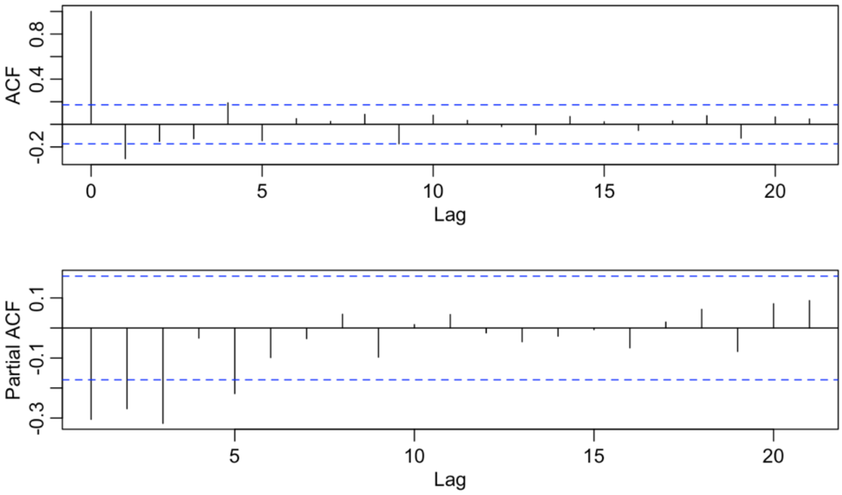

Figure 4: Sample ACF and PACF of ∇Xt

The ACF and PACF of ∇Xt are shown at Figure 4, which can help us to determine the order of ARMA model. Firstly, PACF appears to cuts off at lag = 3 and ACF cuts off at lag = 2 thus we can get ARMA(2,3) but for ACF, it also can say that ACF cuts off at lag =1 cause it's not significant on lag=2, thus we also can propose ARMA(1,3). Since it's for ∇Xt, thus we will try (1) ARIMA(2,1,3), and (2) ARIMA(1,1,3) for global temperature.

Result

For this part, we will have presentation on the results of our findings from data analytic strategies and will include (1) estimated parameters for two proposed models with precise interpretations of these parameters and significance testing, (2) all necessary diagnostics for two proposed models, (3) model selections with explanation, (4) forecasts future 10 years of global temperature with 95% confidence interval, and (5) spectral analysis to identify the first three predominant periods and obtain the confidence interval of the identified periods.

Firstly, we will estimate parameters for two proposed models.

	AR1	AR2	MA1	MA2	MA3	Xmean
Estimate	-0.5955	-0.8104	0.0376	0.3084	-0.6431	0.0072
P-value	0.0000	0.0000	0.7357	0.0226	0.0000	0.0265

Table1: Estimated parameters and p-values of fitted ARIMA(2,1,3)

Table 1 shows the estimated parameters and p-values of fitted ARIMA(2,1,3) model and as we can see the p-value of MA1 is 0.7357 which is much larger than 0.05 and thus not statistical significant but other parameters are statistical significant (since their p-values are all smaller than 0.05). The fitting model would be: $(1 + 0.5955B + 0.8104B^2)\nabla X_t = 0.0072 + (1 + 0.0376B + 0.3085B^2 - 0.6431B^3)w_t$

To interpretate each parameter, it also can be represented as: $\nabla X_t = 0.0072 - 0.5955\,\nabla X_{t-1} - 0.8104\,\nabla X_{t-2} + w_t + 0.0376w_{t-1} + 0.3085w_{t-2} - 0.6431w_{t-3}$. That is: for every single unit more on differences of global temperature in last year, there would be less 0.5955 (AR1) unit on this year's differencing value. For the differencing value in the year before, if it changes one unit, then the differencing value in this year will change -0.8104 (AR2) unit. And the parameters for white noise terms can be explained as same way.

	AR1	MA1	MA2	MA3	Xmean
Estimate	-0.8775	0.3947	-0.6239	-0.2959	0.0074
P-value	0.0000	0.0046	0.0000	0.0011	0.0104

Table2: Estimated parameters and p-values of fitted ARIMA(1,1,3)

Table 2 shows the estimated parameters and p-values of fitted ARIMA(1,1,3) model and all p-values are smaller than 0.05 thus all parameters for this model are statistical significant, and the fitting model would be: $(1 + 0.8775B)\nabla X_t = 0.0074 + (1 + 0.3947B \; - \; 0.6239B^2 - 0.2959B^3)w_t$

The interpterion of this formula is similar with ARIMA(2,1,3) model but will not count the year before last.

Secondly, we will do all necessary diagnostics for two proposed models. To do that, we should show residual analysis for the two models separately.

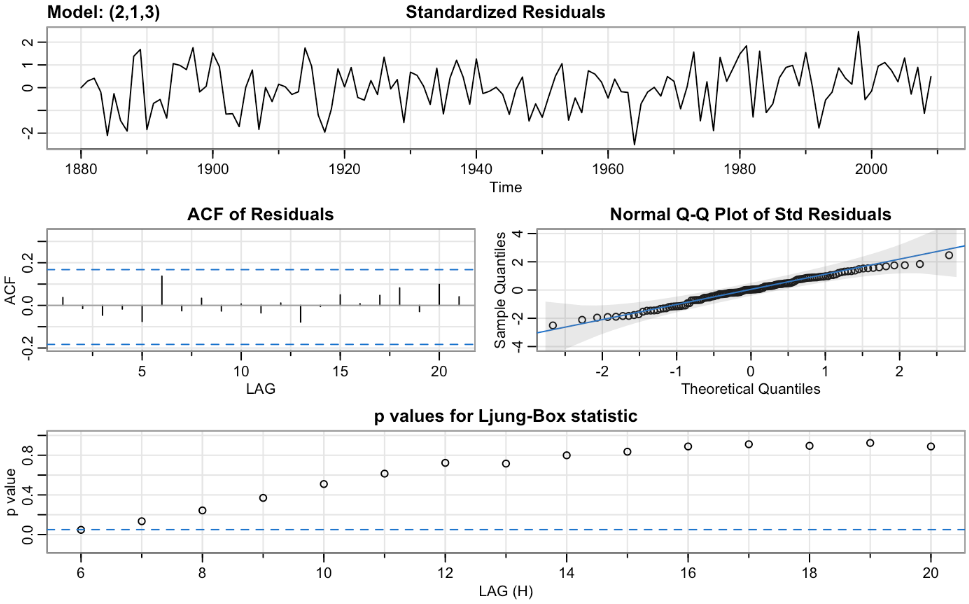

Figure 5: Residual Analysis for ARIMA(2,1,3) on global temperature

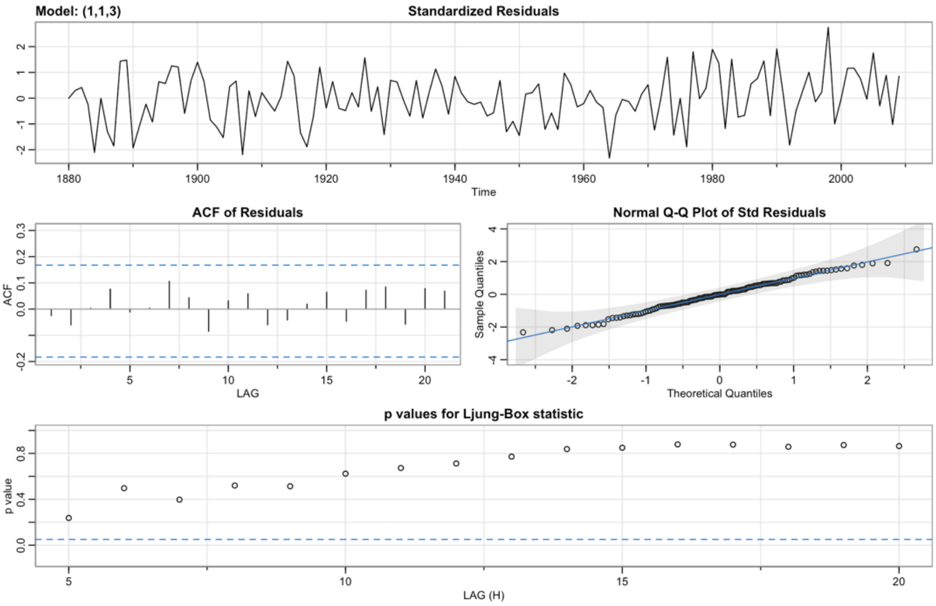

Figure 6: Residual Analysis for ARIMA(1,1,3) on global temperature

Figure 5 and Figure 6 shows residual analysis for our two proposed model. From the residuals plot, there are few outliers here for both plots but neither of them don't have particular patterns. The all ACF of both series don't over the line and no such patterns when both of two normal Q-Q plots shows their residuals are normally distributed. The only difference between two model is Ljung-Box statistic: for ARIMA(1,1,3) model(Figure 6), their p-values are all above 0.05 thus fail to reject null hypothesis (the data are independently distributed) but for ARMA(2, 1, 3) model(Figure 5), there is one p-value = ~0.05 but all other p-value > 0.05.

After we finish diagnostic on residuals, we can evaluate performance of two models and select which one is better.

Table 2 provides AIC, AICc, BIC for two fitted models. As we can see, all values of ARIMA(1,1,3) are smaller than the values of ARIMA(2, 1, 3) thus we will select ARIMA(1, 1, 3) as our preferred model to make prediction.

Model	AIC	AICc	BIC
ARIMA(2,1,3)	-1.2459	-1.2406	-1.0907
ARIMA(1,1,3)	-1.2554	-1.2516	-1.1224

Table 1: AIC, AICc, BIC for two fitted models

Figure 8 shows the predictions of next ten years' global temperature based on ARIMA(1,1,3) model and red line denote the predicted value when gray areas are prediction intervals.

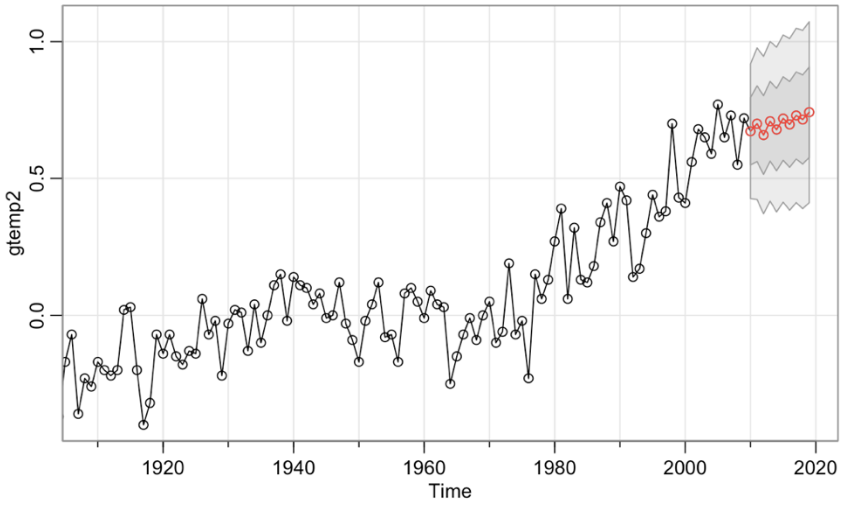

Figure 5: 10-years prediction on global temperature via ARMA(1,1,3)

To show details value of predictions and confidence intervals, I list them as table and Table 3 below just provides 10 years' forecasting values, lower bound and upper bound by 95% confidence intervals in detail.

Year	Predicted value	Lower Bound (95% CI)	Upper Bound (95% CI)
2010	0.6721935	0.4313023	0.9130848
2011	0.6998159	0.4286127	0.9710192
2012	0.6584304	0.3766830	0.9401778
2013	0.7085839	0.4228759	0.9942919
2014	0.6784118	0.3835721	0.9732514
2015	0.7187254	0.4195889	1.0178619
2016	0.6971877	0.3899519	1.0044236
2017	0.7299247	0.4181654	1.0416841
2018	0.7150356	0.3959513	1.0341199
2019	0.7419384	0.4181892	1.0656876

Table 2: 10-year predictions on differences of global of change(∇X_t) with 95% Confidence interval via ARIMA(1,1,3)

From Table 3, we can find the temperature will increase for every other year and decrease back somewhat for other year but there is a general increasing trend on global temperatures. And there is other interesting thing is that the 95% confidence interval become wider and wider when we predict later and later year’s global temperature.

Lastly but not least, we will do spectral analysis to identify the first three predominant periods and obtain the confidence interval of the identified periods.

The periodogram of global temperature is shown in below Figure 9. We can tell from Figure 9 that there is one peak at around 0.01 and it’s much larger than others. And second peak is at around 0.008 (just before first peak) when third one is around 0.06-0.07.

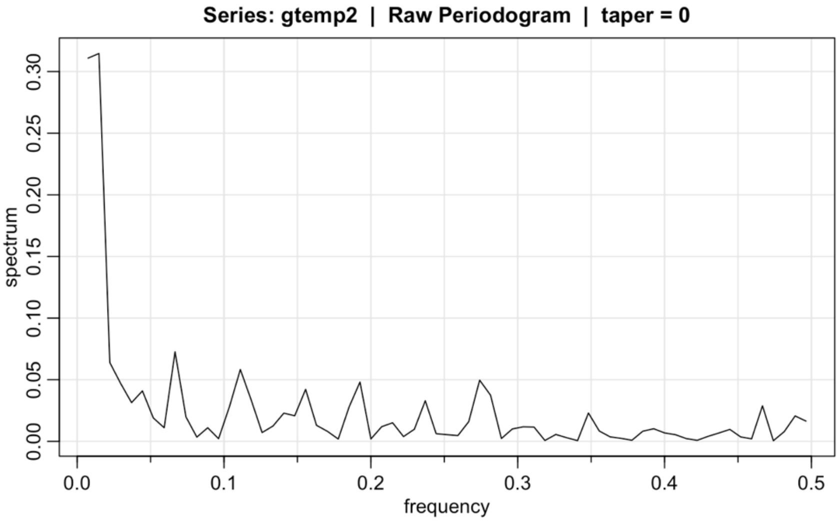

Figure 6: Periodogram of global temperature

But it didn’t provide the details of three dominant periodogram thus I made Table 4 to show first three predominant spectrums with the 95% confidence intervals.

	Frequency	Period	Spectrum	Lower Bound	Upper Bound
Period1	0.0148	67.56757	0.3147	0.08531046	12.429986
Period2	0.0074	135.13514	0.3109	0.08428034	124.204485
Period3	0.0667	14.99250	0.0727	0.01970788	2.871497

Table 3: First three dominant periodogram periods of global temperature

From Table 4, we can find that the confidence intervals for all three periods are wide thus it's hard to test significance. And we can find that all three intervals are overlapping to each other. In this case, we cannot test significance of first two peaks, but we can test third one's significance. The first peak (0.3147) and second peak (0.3109) are lying all three intervals when the third peak (0.0727) only lie third interval and not lie either first or second interval, so we can make difference on third peak. However, we still need to be careful with the overlapping and large intervals.

Discussion

To conclude for our project, we use some statistical tools to fit model which can make forecasting for global temperature and the conclusion for prediction is that there is general increasing trend on global temperature and will increase every other year in general. That's consistent with the global warming issues which we are worried. Even if we get the conclusion from fitted model, there are still some limitations like outliers in standardized residuals, and relatively small sample size (it's over 100 but may still not enough) but they're not trivial. To make our prediction more accurate, we can try more models to test performances or add more dataset if it's available (and currently, the original dataset didn't cover 2010-2021).