STAT33600 Final Project : Time Series Analysis on Central England Temperature

Bryan Chang
Department of Statistics, University of Chicago

December 9, 2019

Abstract

This report aims to provide an extensive analysis on the central England temperature dataset, which includes the seasonal temperature records from year 1659 to 2019. An exploratory data analysis of the warming trend is performed, and a statistical model (SARIMA) is fitted.

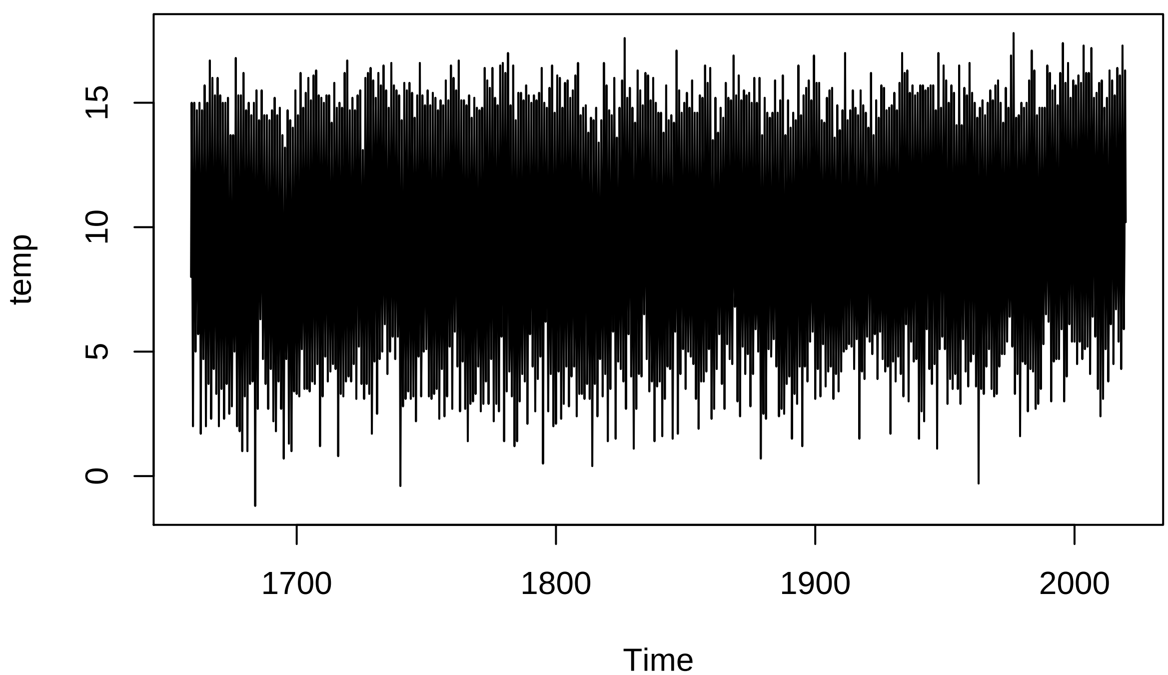
Keywords: CET dataset, exploratory data analysis, warming trend, SARIMA

1 Introduction

The central England temperature (CET) dataset, collected and maintained by the Hadley Centre, is the longest instrumental record of temperature in the world. Mean monthly temperature starting from 1659 to 2019 is provided, which we then divided into four seasons (DJF, MAM, JJA, SON) for this analysis. Although the data is only representative of the central England (a roughly triangular area enclosed by Lancashire, London, and Bristol), it should provide some insights into the long term trend and variability of Earth's temperature.

Illustration of the CET dataset:

	DJF	MAM	JJA	SON
2014	6.1	10.0	15.9	12.1
2015	4.5	8.7	15.3	11.0
2016	6.7	8.6	16.4	10.8
2017	5.4	10.3	16.1	10.9
2018	4.3	9.3	17.3	10.9
2019	5.9	9.3	16.3	10.2



2 Exploratory Data Analysis

2.1 Trend Analysis and Smoothing

An important aspect of the CET dataset we would like to investigate is the overall trend of the temperature. To achieve that, a common technique is smoothing. By smoothing the data, we remove the noise in the observations and could get a better picture of the data. In this section, two smoothing techniques are presented: the kernel smoother and Lowess.

2.1.1 Kernel Smoother

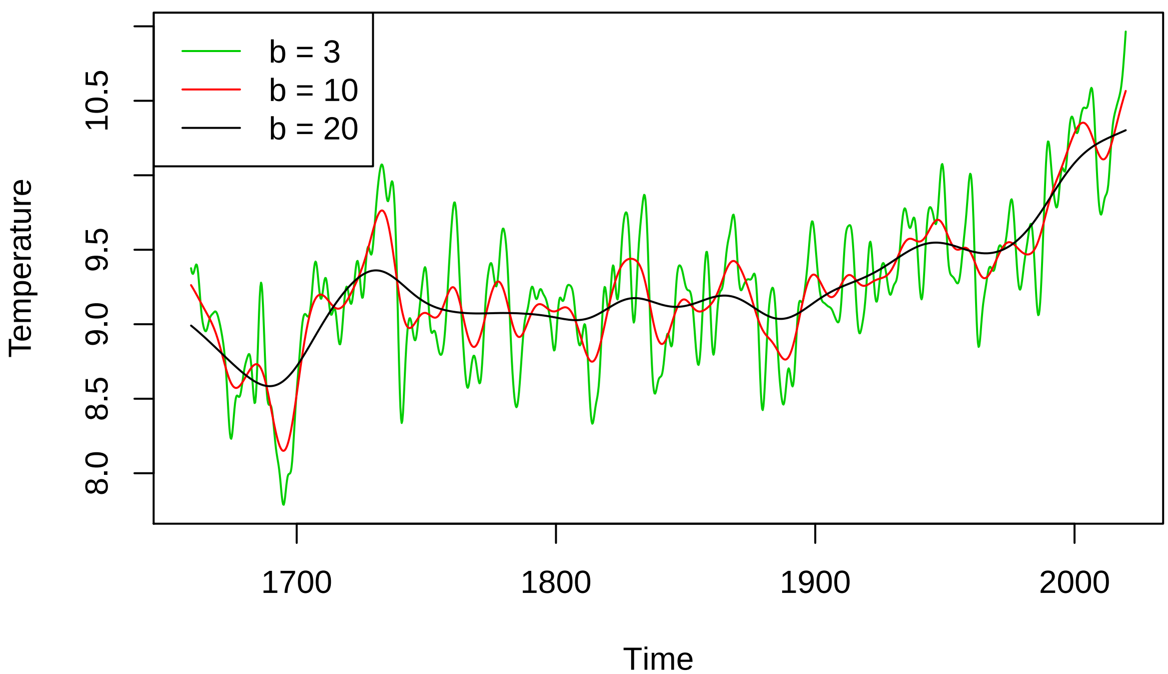
Let x_t be the observations, then

$$m_t = \sum_{i=1}^{n} w_i(t) x_i \tag{1}$$

is the kernel smoother, where

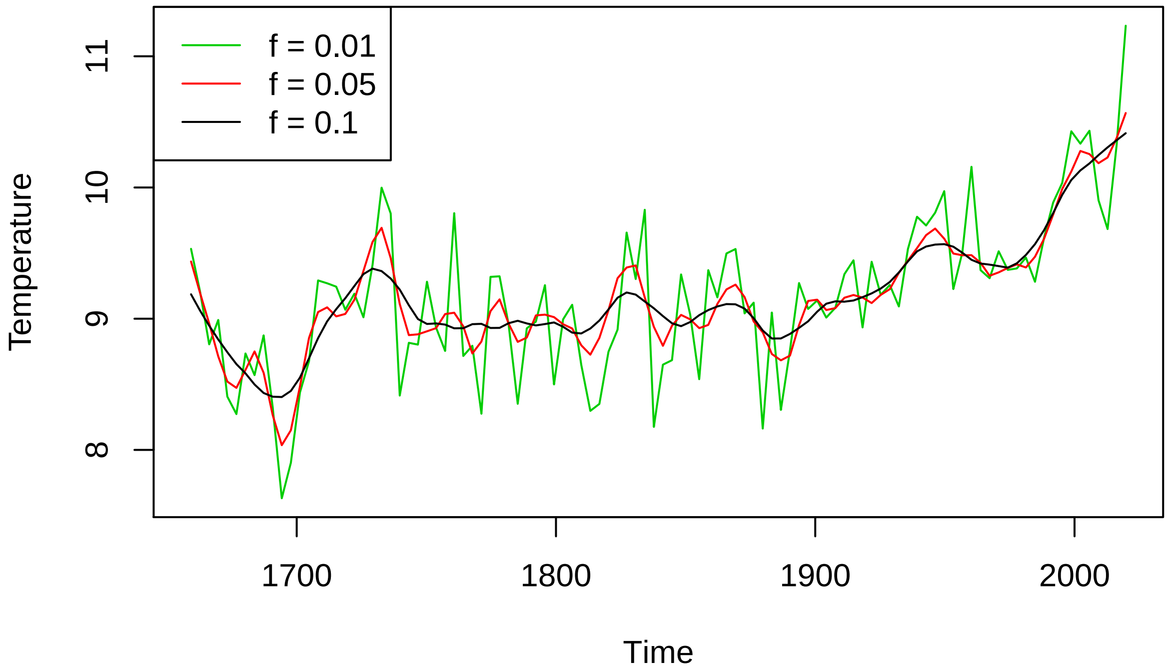
$$w_i(t) = K(\frac{t-i}{b}) / \sum_{j=1}^{n} K(\frac{t-i}{b}) \tag{2}$$

are the weights and K is a kernel function. Here I choose to use the Gaussian Kernel ($K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$) with different choices of bandwidth b. The wider the bandwidth, the smoother the line would be.



2.1.2 Lowess

In Lowess, each smoothed value m_t is given by a weighted linear regression over the span specified by the user. The wider the span, the smoother the line would be.



2.1.3 Comments on the smoothed results

The minimum temperature is reached around 1700, and is quite variable between 1659 and 1750. The trend of the temperature stays rather flat and stable from 1750 to 1800, and we can observe a steady increase in temperature from 1900. The claim of global warming reaching hiatus (Vaidyanathan, 2016) doesn't seem conclusive based on simply smoothing the data. Of course, if one looks at the smoothest curve between 1980 to 2000 and 2000 to 2019, it may seem that the slope of the line has decreased, but the choice of bandwidth or span is quite arbitrary. Only if we are able to justify the choice of bandwidth/span can we verify the claim of global warming's hiatus.

3 Statistical Model

In this section we fit a few different statistical models to the central England temperature data. All models fitted here would fall in the category of the autoregressive integrated moving average (ARIMA) model. In particular, since there is a seasonal pattern in our data, the seasonal autoregressive integrated moving average (SARIMA) will be used.

3.1 Fitting SARIMA

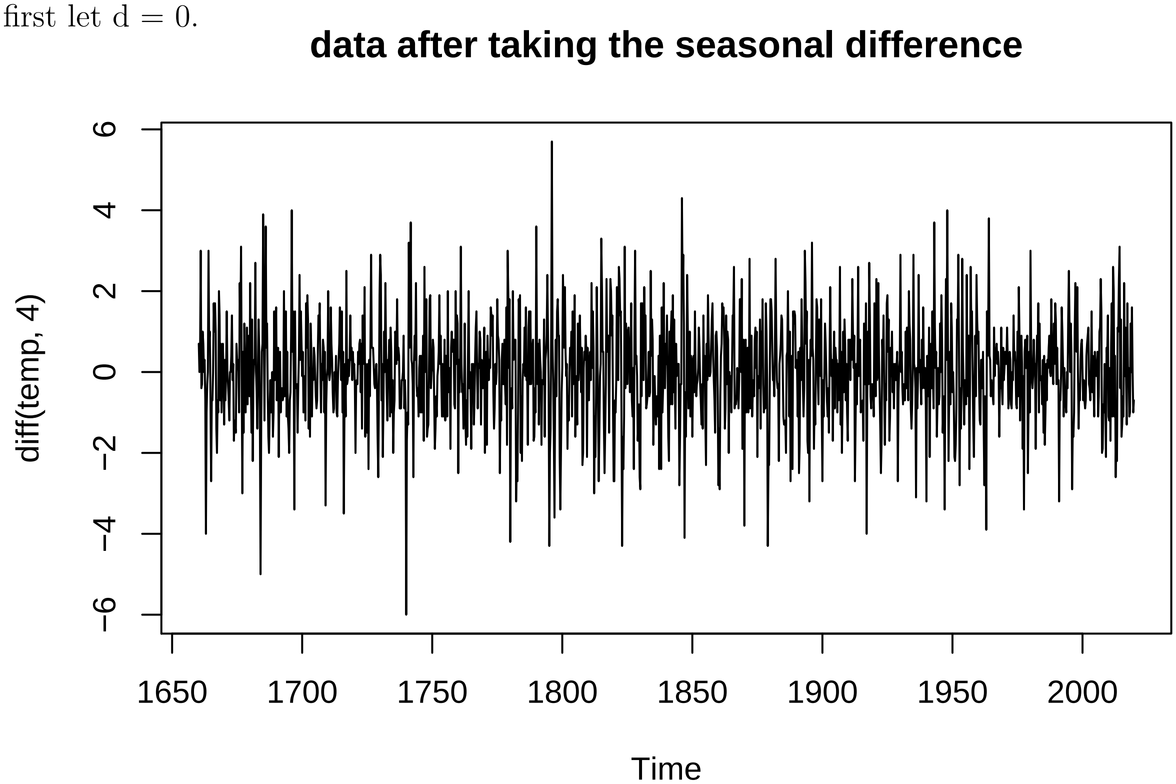
By definition, the SARIMA model is given by

$$\Phi_p(B^s)\phi(B)\nabla_s^D \nabla^d x_t = \delta + \Theta_Q(B^s)\theta(B)w_t \tag{3}$$

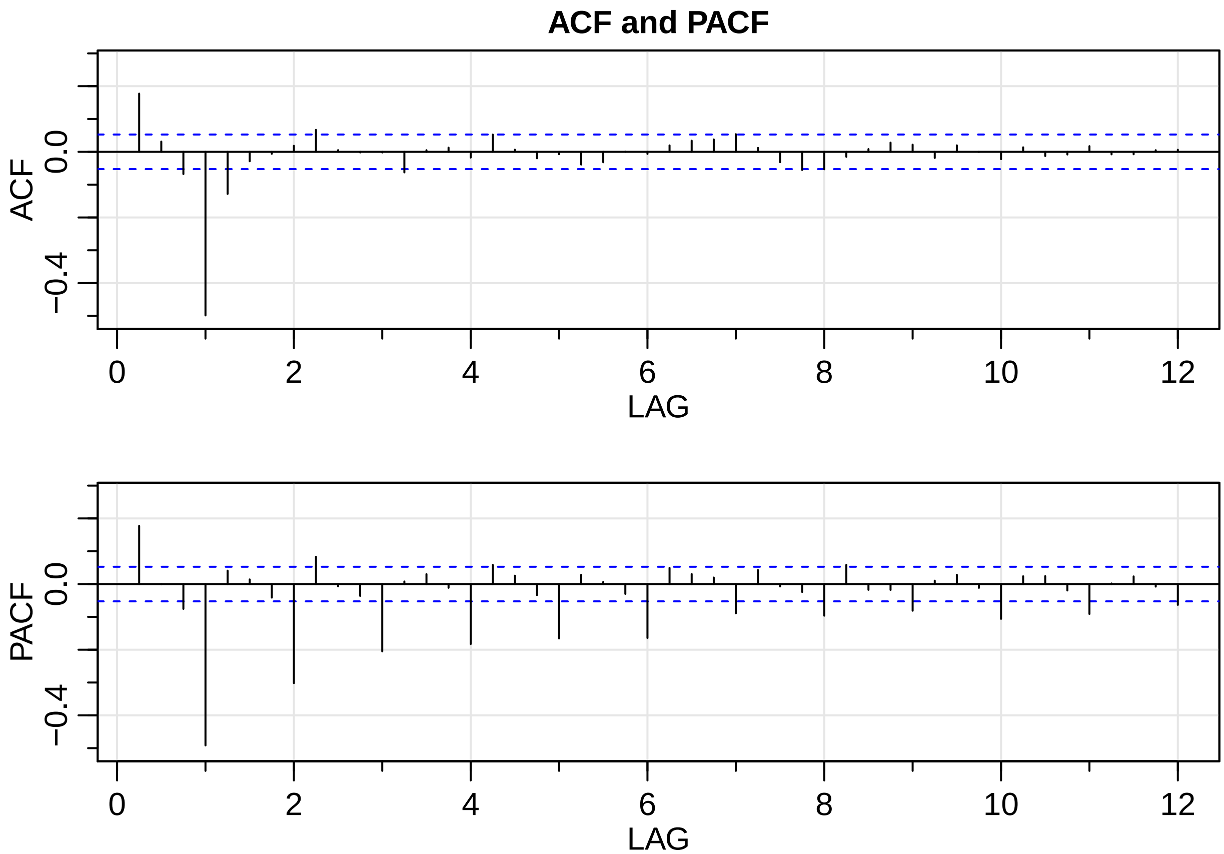
, denoted as $ARIMA(p, d, q) \times (P, D, Q)_s$

To determine the orders of the SARIMA, we have to first examine the trend and the seasonality of the data to decide (d, D). After the differencing is done, we then look at the ACF and PACF to determine (p,q) and (P,Q).

It seems that there is no linear trend after taking the seasonal difference (D = 4), so we first let d = 0.



Now we can determine the orders (p,q) and (P,Q). The ACF indicates that the seasonal MA order (Q) should be 1, and the PACF indicates that the seasonal AR order (P) could be either 0 or 1. Similarly, the ACF indicates that the non-seasonal MA order (q) should be 1, and the PACF indicates that the non-seasonal AR order (p) should be 1.



For $ARIMA(1,0,1) \times (0,1,1)_4$ model, the mathematical equation is:

$$(1 - \phi_1 B)(1 - B^4)x_t = \delta + (1 + \theta_1 B)(1 + \Theta_1 B^4)w_t \tag{4}$$

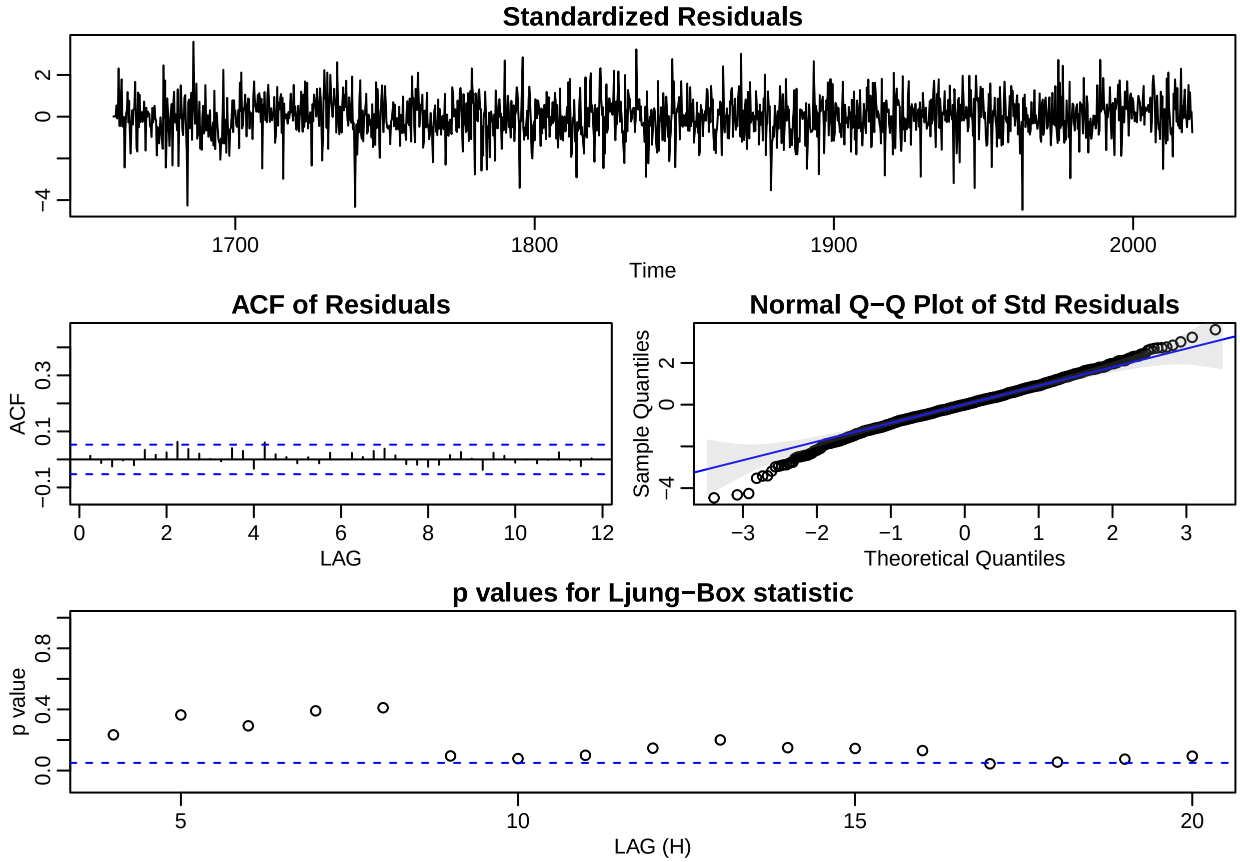
For $ARIMA(1,0,1) \times (1,1,1)_4$, the mathematical equation is:

$$(1 - \phi_1 B)(1 - \Phi_1 B^4)(1 - B^4)x_t = \delta + (1 + \theta_1 B)(1 + \Theta_1 B^4)w_t \tag{5}$$

3.2 Model Diagnostics

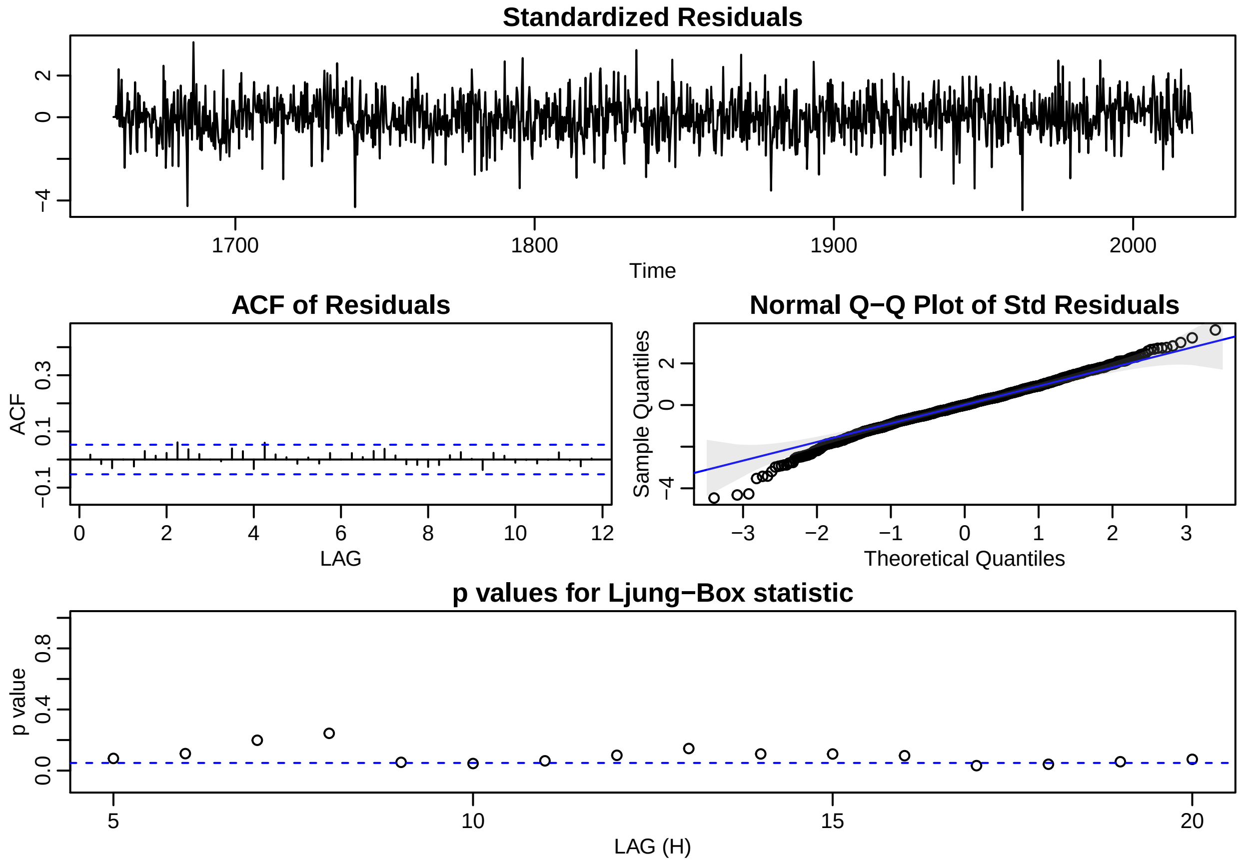
For $ARIMA(1,0,1) \times (0,1,1)_4$:

The time plot of the standardized residuals show no obvious pattern, although we can observe that there are a few outliers. The ACF of the standardized residuals indicates that there is little correlation structure in the residuals. The Q-statistic plot looks okay (most of them are not significant at the lags shown). The normal Q-Q plot shows that there are some outliers, but in general the residuals follow normal distribution.



For $ARIMA(1,0,1) \times (1,1,1)_4$:

Like the previous model, most of the diagnostic plots look okay. However, many of the p-values of the Q-statistics are close to being significant.



3.3 Model Estimates

For $ARIMA(1,0,1) \times (0,1,1)_4$:

	Estimate	SE	t.value	p.value
ar1	0.6819	0.1060	6.4345	0.0000
ma1	-0.4943	0.1273	-3.8821	0.0001
sma1	-0.9697	0.0086	-112.6525	0.0000
constant	0.0009	0.0003	2.6645	0.0078

For $ARIMA(1,0,1) \times (1,1,1)_4$:

	Estimate	SE	t.value	p.value
ar1	0.7112	0.1834	3.8785	0.0001
ma1	-0.5287	0.2198	-2.4055	0.0163
sar1	-0.0098	0.0473	-0.2070	0.8360
sma1	-0.9698	0.0088	-110.5906	0.0000
constant	0.0009	0.0003	2.6277	0.0087

The seasonal AR(1) term is not significant. So the $ARIMA(1,0,1) \times (0,1,1)_4$ should be a better choice.

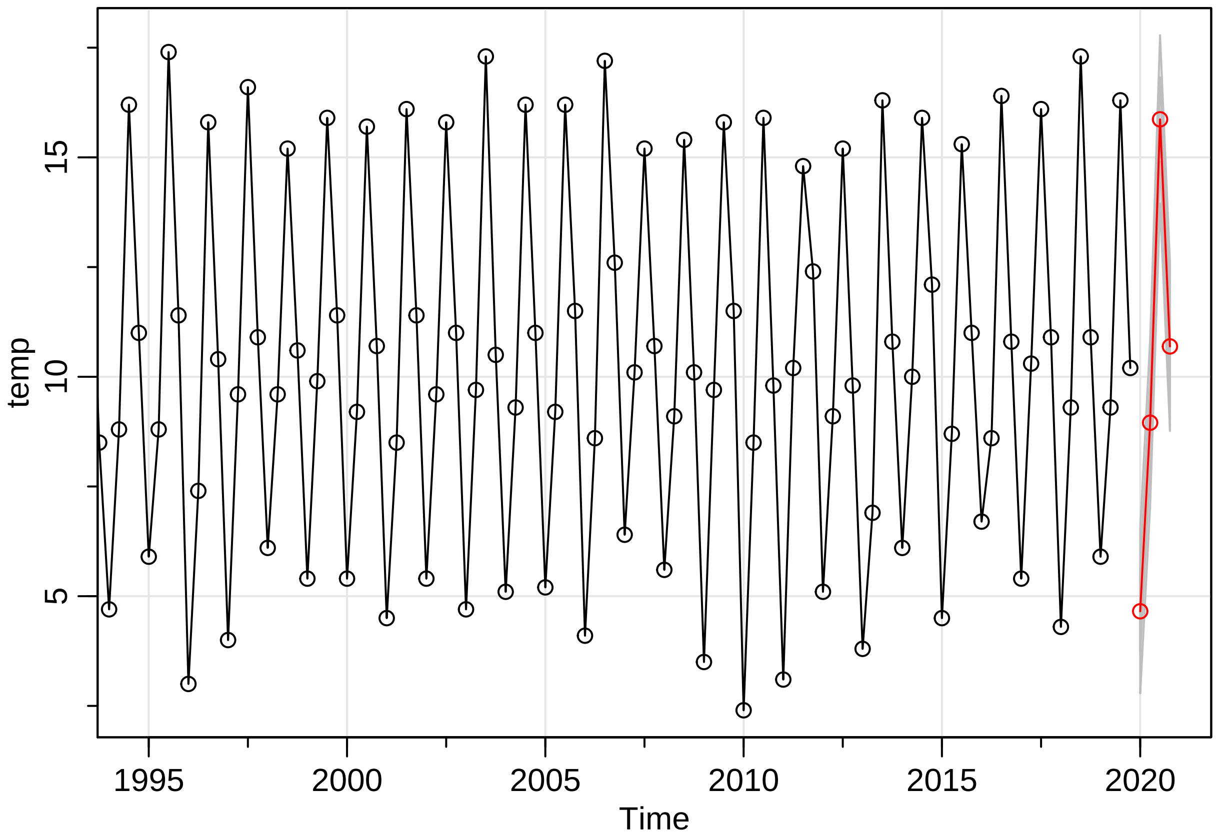
3.4 Model Selection by Information Criteria

Other than $ARIMA(1,0,1) \times (0,1,1)_4$ and $ARIMA(1,0,1) \times (1,1,1)_4$, I also tried fitting $ARIMA(1,1,1) \times (0,1,1)_4$ and $ARIMA(2,1,1) \times (0,1,1)_4$ (so taking an additional difference). Here I compare the AICc and BIC of those models, and the results show that all information criteria prefer the $ARIMA(1,0,1) \times (0,1,1)_4$ model.

	AIC	AICc	BIC
ARIMA(1,0,1)(0,1,1)_4	2.713838	2.713857	2.732117
ARIMA(1,0,1)(1,1,1)_4	2.715190	2.715219	2.737125
ARIMA(1,1,1)(0,1,1)_4	2.722098	2.722109	2.736729
ARIMA(2,1,1)(0,1,1)_4	2.720340	2.720359	2.738630

3.5 Forecasting

Finally, we can make predictions using the fitted $ARIMA(1,0,1) \times (0,1,1)_4$ model. Here I display the predictions for the next four seasons.



Predictions for 2020 and data from 2014:

	DJF	MAM	JJA	SON
2014	6.100000	10.000000	15.90000	12.10000
2015	4.500000	8.700000	15.30000	11.00000
2016	6.700000	8.600000	16.40000	10.80000
2017	5.400000	10.300000	16.10000	10.90000
2018	4.300000	9.300000	17.30000	10.90000
2019	5.900000	9.300000	16.30000	10.20000
2020	4.656055	8.953694	15.86693	10.69189

4 Conclusion

The trend analysis in this report gives no conclusive answer to the hiatus or slowdown of global warming in the 2000s, but it is clear that the temperature has been rising fastly since 1900 compared to the period before. This report also provides justifications for the choice of the seasonal autoregressive integrated moving average (SARIMA) model by plots of ACF/PACF and different information criteria.

5 References

Harvey, D. I. and Mills, T. C. (2003). Modelling trends in central England temperatures. J. Forecasting 22, 35ā47.

Shumway, Robert H. and David S. Stoffer. (2017). Time Series Analysis and Its Applications: with R Examples. Springer.

Vaidyanathan (2016). Did Global Warming Slow Down in the 2000s, or Not? Scientists clarify the recent confusion.