STAT3301/4601/6002 Time Series Analysis Project
Topic: Fitting time series model in Hong Kong monthly average temperature
Name: Leung Ming Tak
UID: 3035053381
Due Date: 5 December 2015

Project Specification

Find a time series with sample size n >100, and delete the last 5 values for the sake of comparison at the end of this project.

Introduction

Background: Weather data is one of the most common time series data. In order to interpret and forecast weather in Hong Kong, the Hong Kong Observatory (HKO) has been recording various meteorological data since 1883. Raw data, however, is not fully accessible by the public. Since only the latest data will be provided from official website, in this project, data originally disturbed by HKO will be collected from The Journalism and Media Studies Centre (JMSC), the University of Hong Kong, under project of Hong Kong Open Government Data. Source data are available from
http://data.jmsc.hku.hk/hongkong/observatory/.

Objective: I will fit a most suitable model from raw data and predict the trend of average temperature data (measured in Celsius) monthly from January 1997 to March 2015 (size=219). To verify the performance of selected model, raw data will be separated into training data of 17 years (size=207) for model building, and testing data of one year (size=12) for validation.

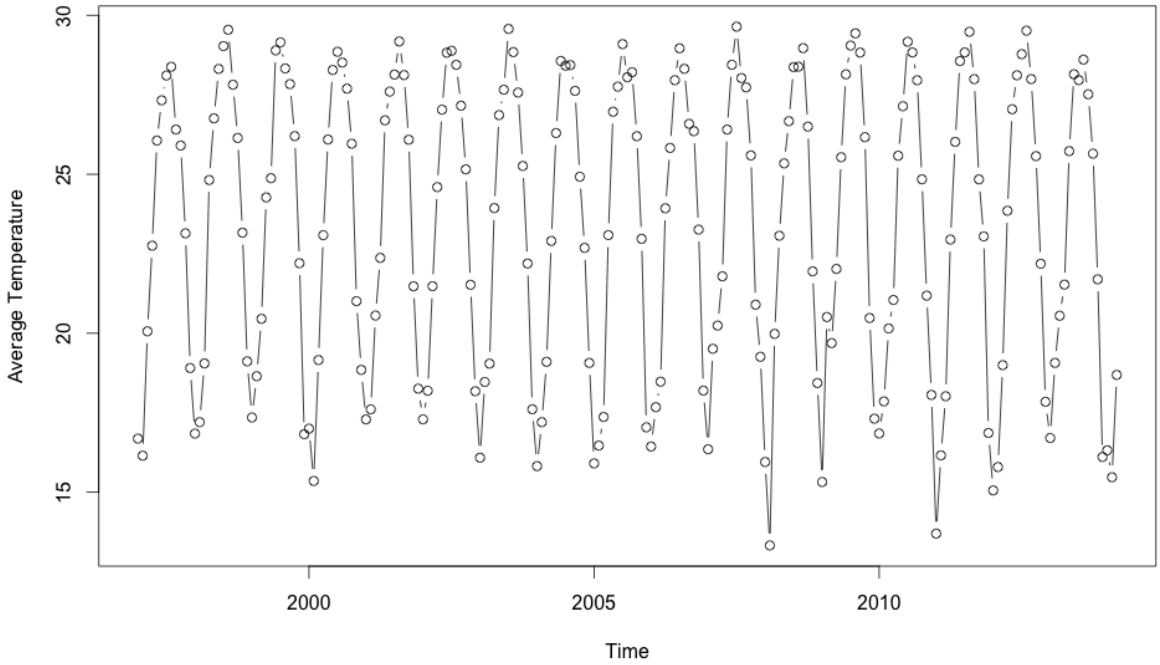
Methodology: In this project, I will perform the analysis using programming language R. Instead of core packages, I will also include two external packages "astsa" and "forecast". Since both packages are available on CRAN, those packages are relatively easy to install and use. The two packages are essential for fitting seasonal time series data and visualize the data in with one line of code. Source code will be provided in appendix section.

Data Exploration

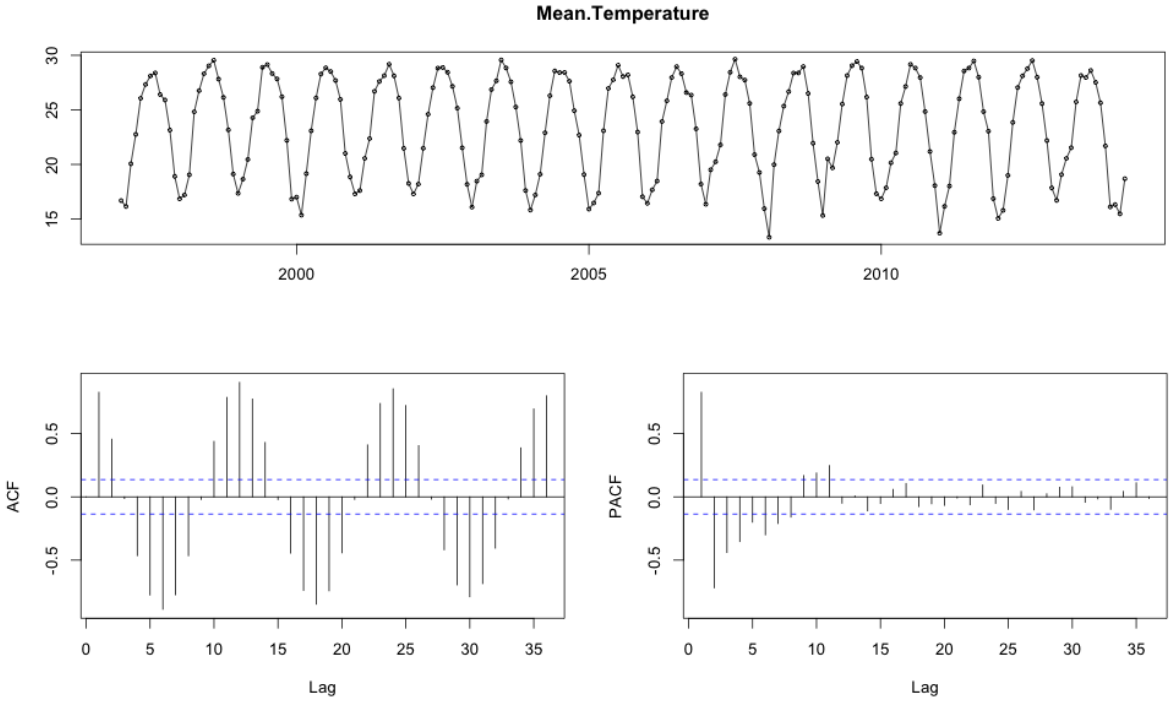
Raw Data: The raw data is composed of various meteorological data, including temperature, pressure, rainfall, wind speed, sea level etc., measured by day from January 1997 to March 2015. Since model fitted from daily sample through 18 years would be over-fitted, I will convert the data into monthly mean temperature by taking average of daily mean temperature.
Before proposing suitable models, we would conduct some preliminary data exploration for the raw data in the sections. Mean temperature versus time plot, sample autocorrelation function (ACF) and partial autocorrelation function (PACF) will be explored for preliminary analysis.

Time plot

The following shows mean temperature versus time plot from 1997 to 2014:



From the time plot, we can observed that current samples are non-stationary with some seasonality. Following sample ACF and PACF plot of raw time series data further suggest the seasonal effect:



Hence, we might need to perform transformation before specify suitable models. Details

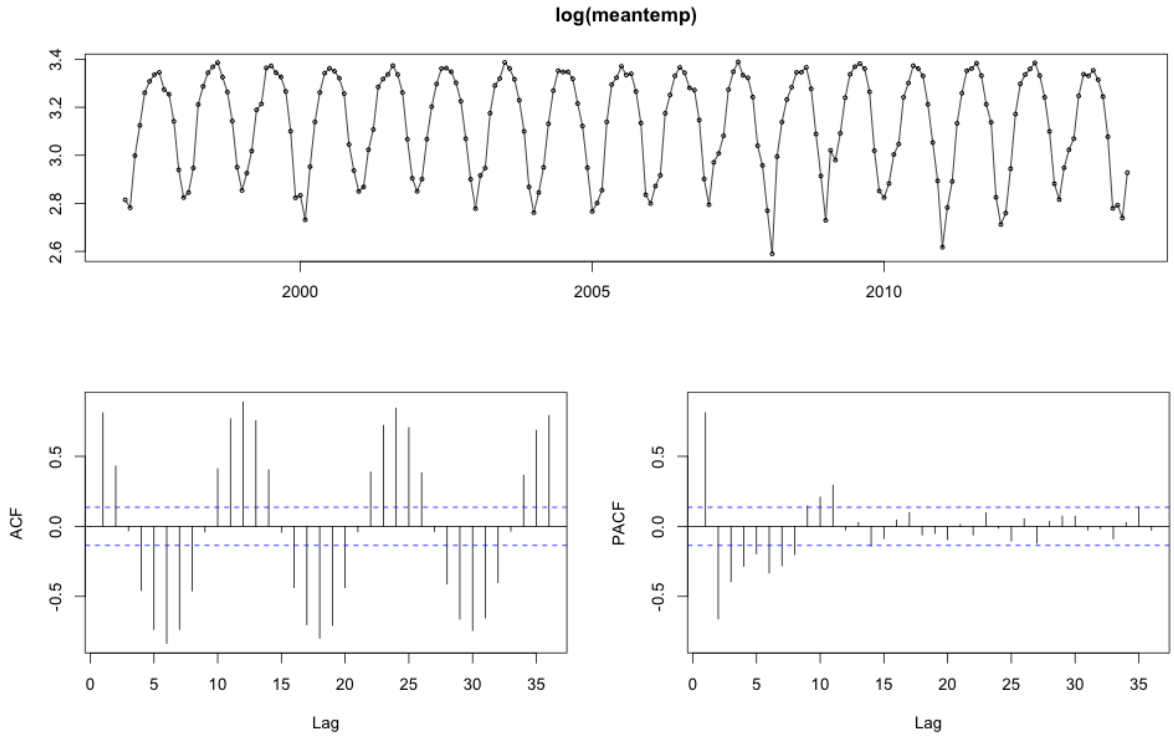
of transformation will be discussed in the following section.

Transformation

In this section, we will perform log, 1st and seasonal difference to the raw data. Time plot of transformed data, sample ACF and sample PACFs will be used to select a suitable transformation.

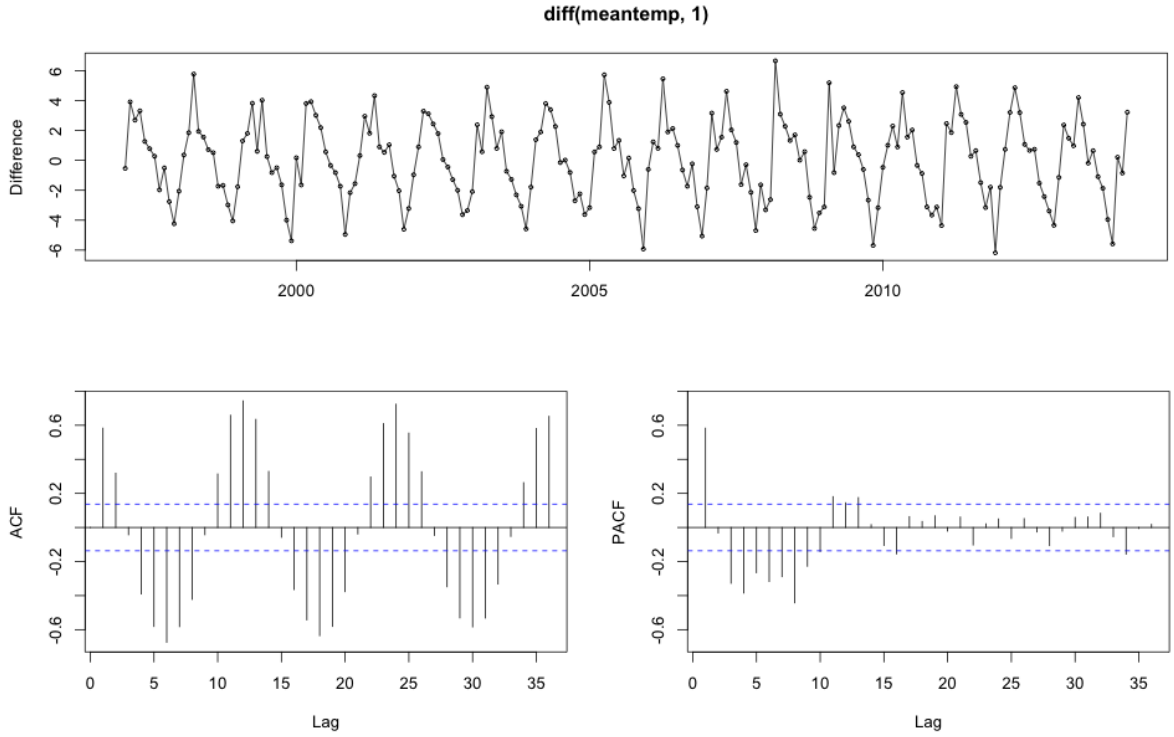
Log Transformation

Log Transformation is used to remove the non-stationary effect in variance. It is clear that seasonal effect remains after log transformation. Besides, ACF and PACF plot suggested that most of the lags are significantly correlated. In this case, log transformation does not needed in the time series transformation.



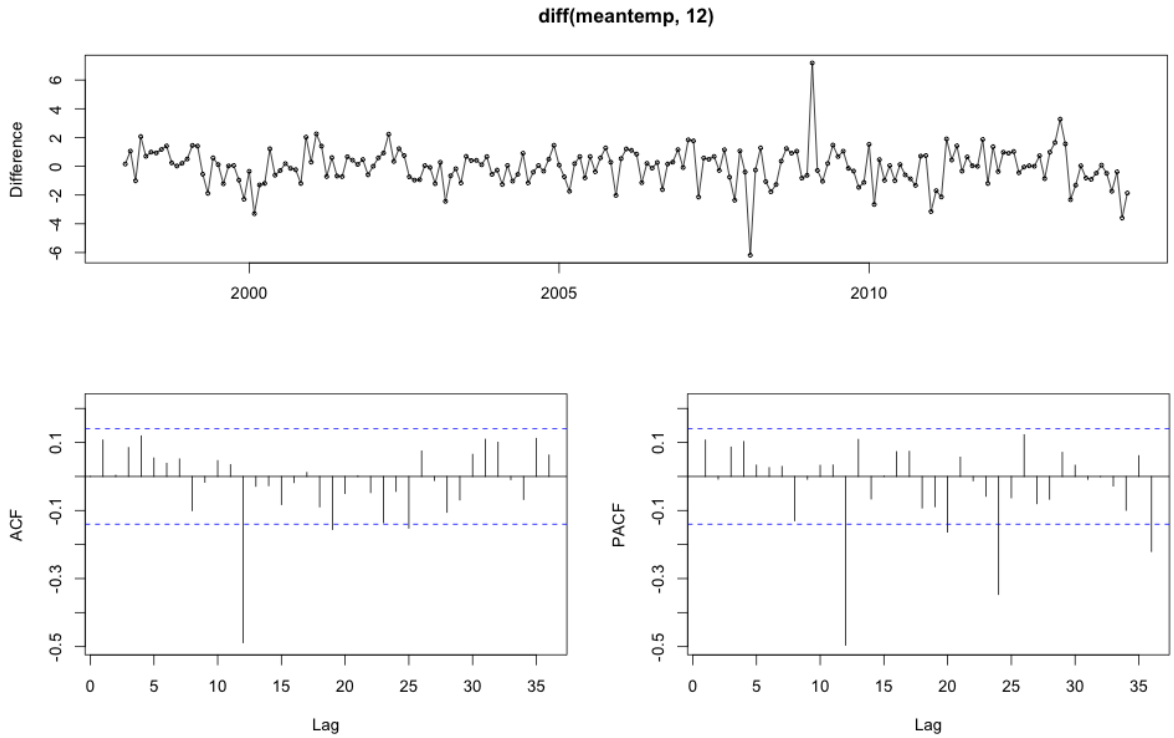
1st Difference

1st differencing is commonly used to eliminate the effect of local linear trend. After taking difference operations, similarly, seasonal effect remains in the transformed sequence, indicated by sample ACF, PACF plot. Therefore, 1st differencing is not necessary in the model.



Seasonal Difference

Since the plot shows a seasonality effect and the data is measured by month, we will examine sample ACF and PACF of 12^{th} differences.

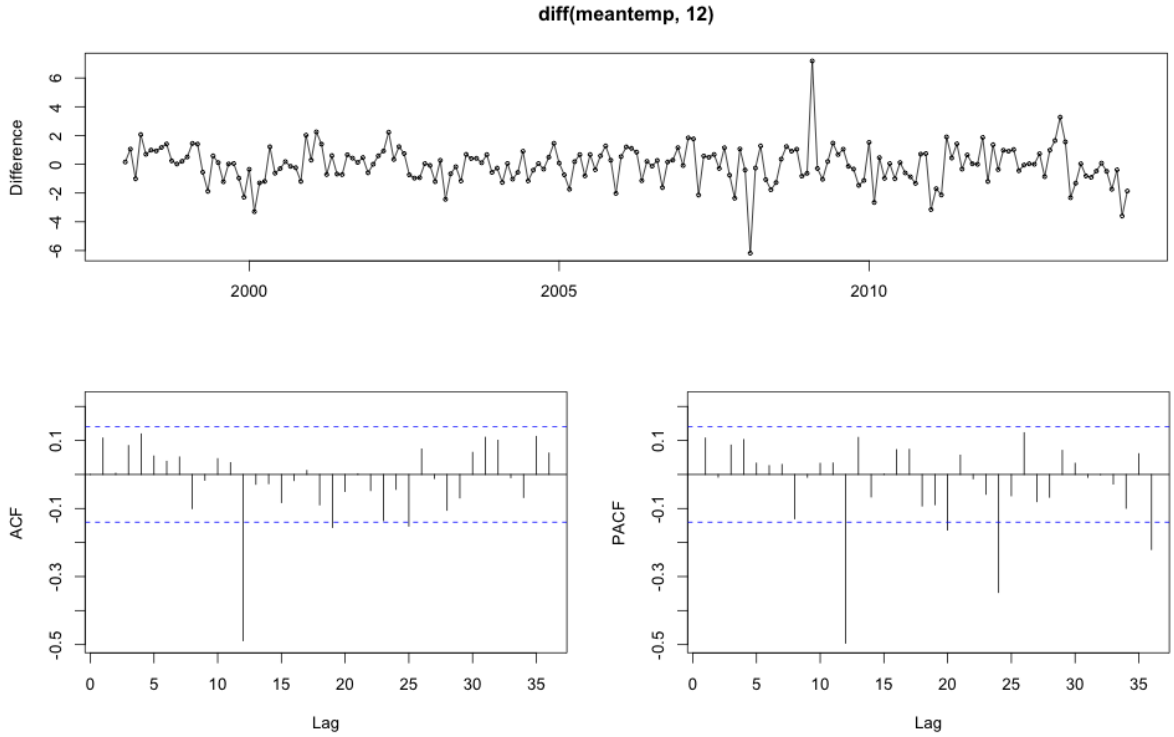


From the difference plot, it is observed that mean and variance are stationary over time.

Seasonal effect does not appear from the graph. Besides, most of the lags are not significantly correlated in ACF and PACF, excepted lag=12, 24 and 36. Since we can identify lags that have significant correlations from ACF and PACF, seasonal ARIMA models will be suggested from the time series. Details will be covered in next section.

Model Specification

Followed by the previous section, we concluded that taking seasonal transformation with $D = 12$. In this section, we will discuss the some possible models for fitting the data based on sample ACFs and the sample PACFs.



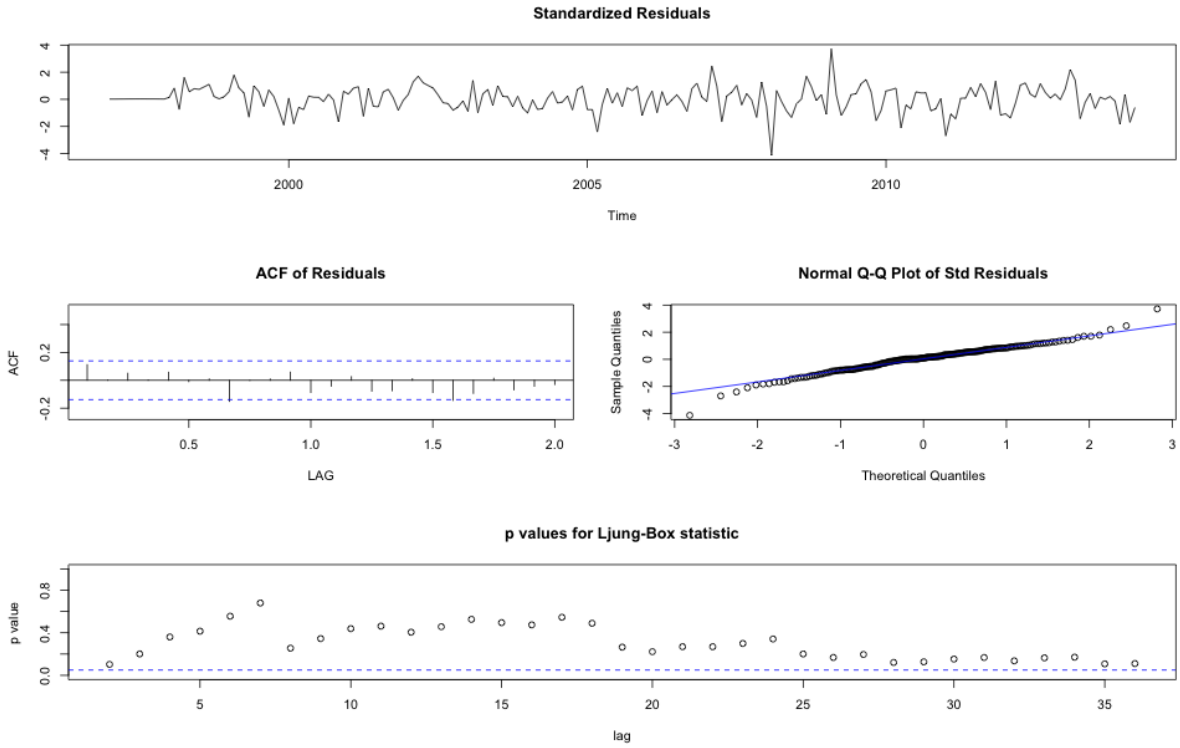
For the seasonal part, since a spike at ρ_{12} and other insignificant spikes ρ as $k \neq 12$ in ACF plot, the feature matches one of the key property of seasonal MA(1) model. Hence, we will select some several SMA(1) model for model diagnostics.
Similarly, based on the observation of exponentially decayed trend at ρ_{12},ρ_{24} and ρ_{36}, we would also suggest seasonal AR(1) model for model diagnostics.
For the non-seasonal part of the model, none of the plot shows significant correlations. In that case, we would not suggest any models for the non-seasonal part.
Apart from the above specified models, general seasonal model ARIMA(0,1,1)x(0,1,1), i.e. an MA(1)xSMA(1) model with both seasonal and non-seasonal difference, are the most commonly used in most seasonal prediction models. Although the correlation at lag=1 does not significantly shown on sample ACF plot, this model is also worth examining and will be included in next section.
To conclude, we will examine the following three models:

- ARIMA(0,0,0)x$(0,1,1)_{12}$
- ARIMA(0,0,0)x$(1,1,0)_{12}$
- ARIMA(0,1,1)x$(0,1,1)_{12}$

Model Diagnostics

1. ARIMA(0,0,0)x(0,1,1)$_{12}$

Overview: After fitting to ARIMA(0,0,0)x(0,1,1) model, the sample ACF, PACF and residual plots are shown in the following graphs:



From the QQ-Plot, most of the residuals point fits to to plot, which suggest that it normality assumption holds in this model.
Since there are no specific pattern shown in the residual plot, and correlations shown in sample ACF and PACF are not significant, it is most likely an adequate model.

Maximum Likelihood Estimation: The following table shows the value, standard error and p-value of each estimated valuables:

Parameter	Estimate	Standard Error	p-value
Constant	-0.0027	-0.0011	0.0124
SMA1	-1.0000	0.2127	0.0000
σ^2	0.8502		
log likelihood	-255.03		
AIC	516.06		

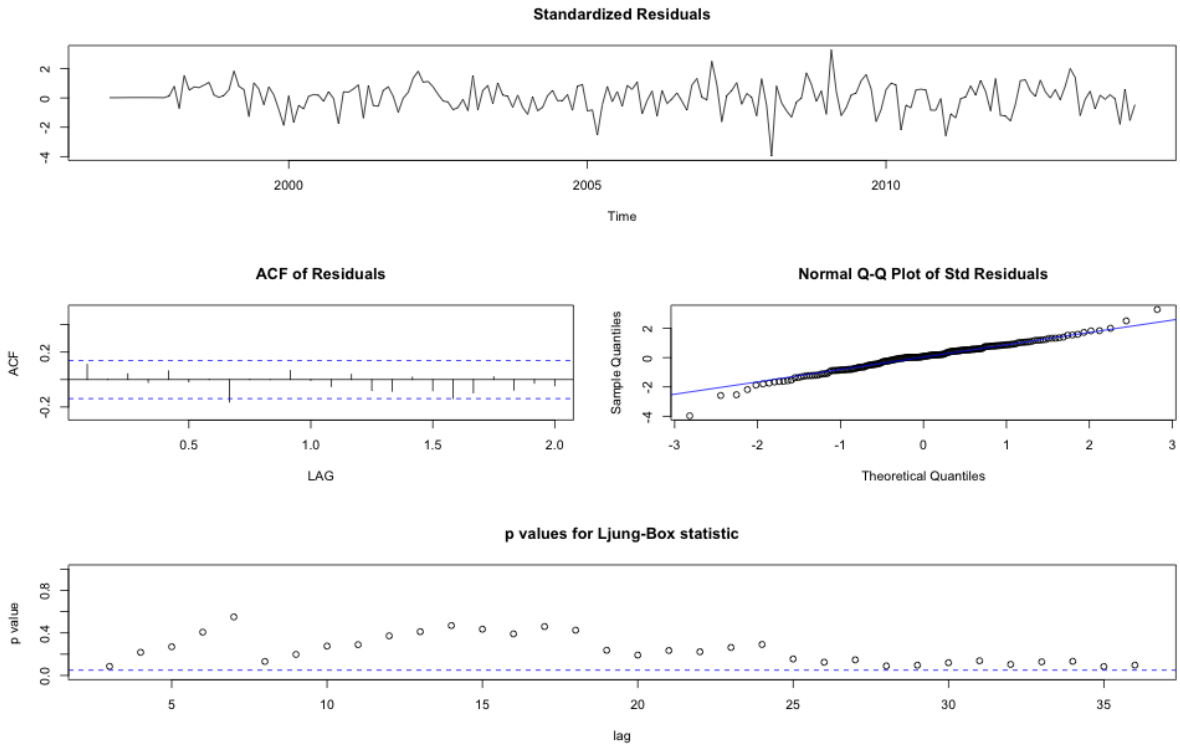
In this case ,estimated SMA1 coefficient is significant in 5% significance level. Hence the parameter has effect on the model.

Ljung-Box Test: The following tables shows the Ljung-Box statistics and its p-value:

Lag	Chi-Square	DF	p-value
6	4.042	5	0.5434
12	10.9834	11	0.4447
18	14.6346	17	0.6218
24	21.4827	23	0.5516
30	32.579	29	0.295

As shown in the table, residual ACFs are jointly insignificant in 5% significance level. Therefore, we consider ARIMA(0,0,0)x$(0,1,1)_{12}$ an adequate model.

Overparameterization: For checking the adequacy of fitted model, it is suggested to check the significance of an over fitted ARIMA(0,0,0)x$(0,1,2)_{12}$ model.

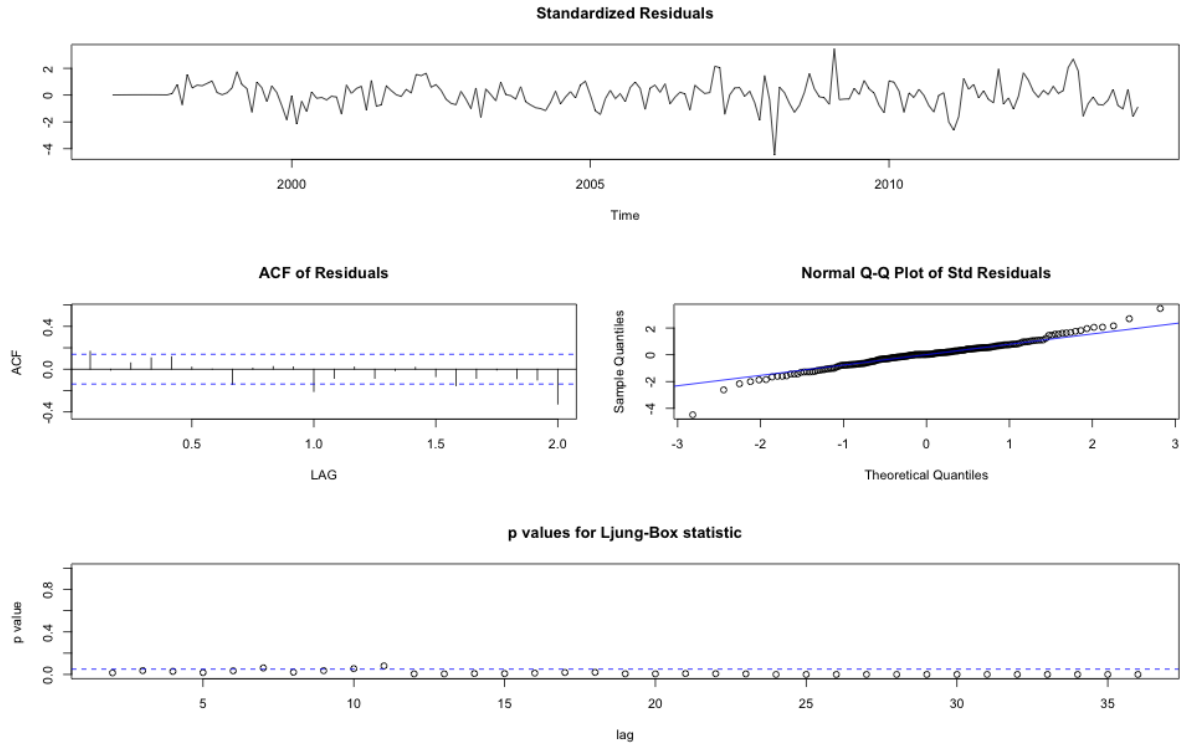


Parameter	Estimate	Standard Error	p-value
Constant	-0.0027	0.0011	0.03259
SMA1	-1.0278	0.0906	0.0000
SMA2	0.1127	0.0842	0.1462
σ^2	0.8924		
log likelihood	-254.2		
AIC	516.41		

In this case, the estimate of SMA1 does not differ too much from original model. However, coefficient SMA2 is not significant in 5% confidence level, parameter SMA2 is over fitted. It is indicated that Seasonal SMA1 model is an adequate model from the above diagnostics.

2. ARIMA(0,0,0)x(1,1,0)$_{12}$

Overview: After fitting time series to ARIMA(0,0,0)x(1,1,0)$_{12}$ model, the sample ACF, PACF and residual plots are shown in the following graphs:



Although normality assumption holds and we cannot indicate any patterns from residual plot, correlations are significant in some specific lags like lag 1, lag 12 and lag 24. Based on observation, it is most likely to propose an alternative model.

Maximum Likelihood Estimation: The following table shows the maximum likelihood estimation value, standard error and p-value of each parameters:

Parameter	Estimate	Standard Error	p-value
Constant	-0.0020	0.0047	0.7489693
SAR1	-0.5192	0.0637	0.0000000
σ^2	1.345		
log likelihood	-282.76		
AIC	571.52		

In this case, estimated SAR1 coefficient is significant in 5% significant level. Hence the parameter has effect on the model.

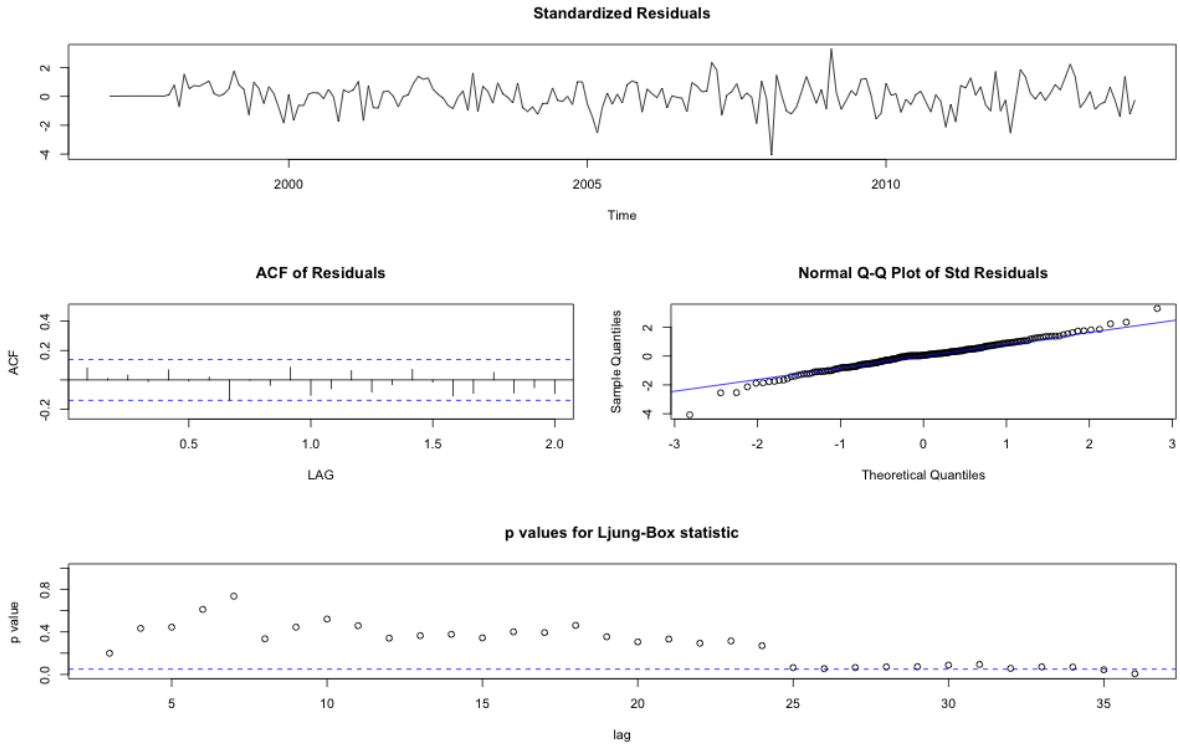
Ljung-Box Test: The following tables shows the Ljung-Box statistics and its p-value:

Lag	Chi-Square	DF	p-value
6	10.8705	5	0.05401
12	24.7356	11	0.010
18	29.0921	17	0.03369
24	61.0204	23	2.717e-05
30	74.9262	29	6.24e-06

As shown in the table, most of the p-values are less than 0.05, therefore, ARIMA(0,0,0)x$(1,1,0)_{12}$ is not an adequate model.

Alternative: ARIMA(0,0,0)x(2,1,0) Since seasonal AR(1) is not an adequate model, we suggest ARIMA(0,0,0)x$(2,1,0)_{12}$ as a alternative model.

Checking of alternative method: Following table shows the estimated parameters by maximum likelihood estimation method:



Parameter	Estimate	Standard Error	p-value
Constant	-0.0021	0.0030	0.5530
SAR1	-0.7346	0.0685	0.0000
SAR2	-0.4198	0.0686	5.7131e-10
σ^2	1.101		
log likelihood	-266.63		
AIC	541.25		

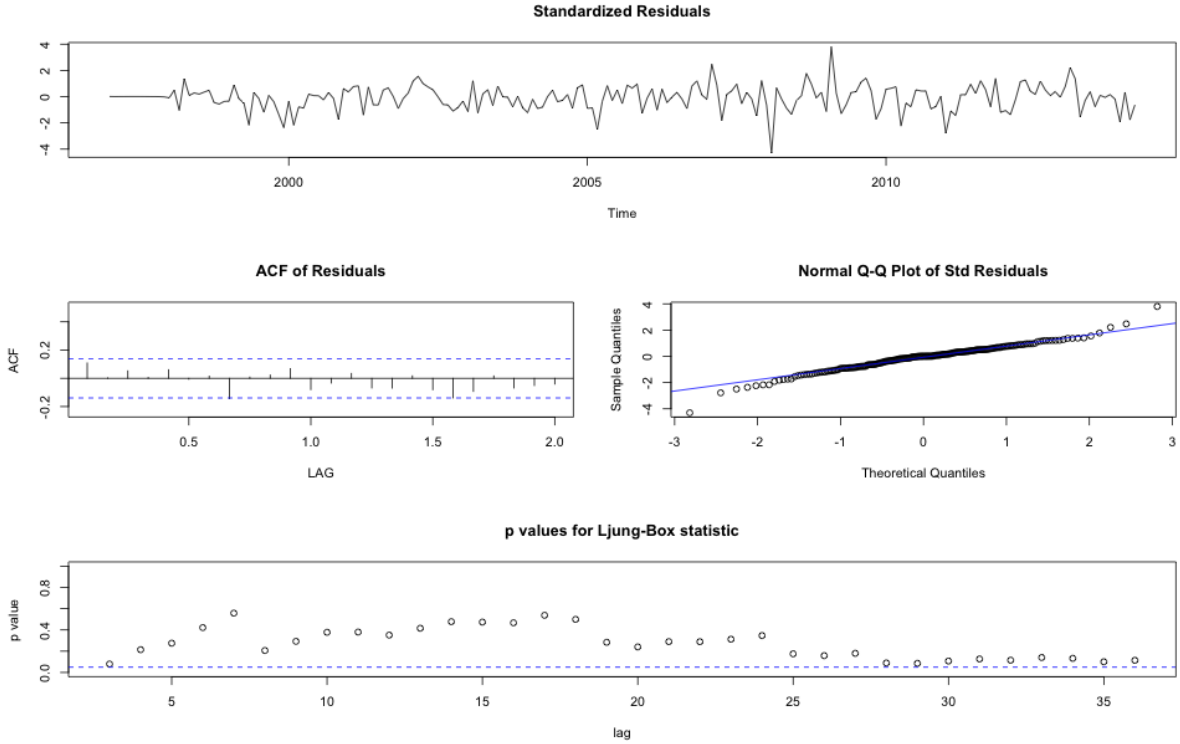
From the table, estimated SAR1 and SAR2 are significant in 5% significant levels, we will perform Ljung-Box test to ARIMA(0,0,0)x$(2,1,0)_{12}$ for further diagnostics.

Lag	Chi-Square	DF	p-value
6	2.6946	4	0.6102
12	10.5191	10	0.3962
18	14.9772	16	0.5263
24	23.1597	22	0.3928
30	35.2534	28	0.1626

As shown in the table, residual ACFs are jointly insignificant in 5% significance level. Therefore, we choose an alternative ARIMA(0,0,0)x$(2,1,0)_{12}$ model as a replacement of Seasonal AR(1) model.

3. ARIMA(0,1,1)x$(0,1,1)_{12}$ model

Overview: After fitting time series to ARIMA(0,1,1)x(0,1,1) model, the sample ACF, PACF and residual plots are shown in the following graphs:



From the QQ-Plot, most of the residuals point fits to to plot, which suggest that it normality assumption holds in this model. Since no patterns or signals can be indicated from the residual plot, and correlations shown in sample ACF and PACF are not significant, it is most likely an adequate model.

Maximum Likelihood Estimation: The following table shows the value, standard error and p-value of each estimated valuables:

Parameter	Estimate	Standard Error	p-value
MA1	-1.0000	-0.0273	0.0000
SMA1	0.9830	0.3718	6.7502e-14
σ^2	0.815		
log likelihood	-275.21		
AIC	556.43		

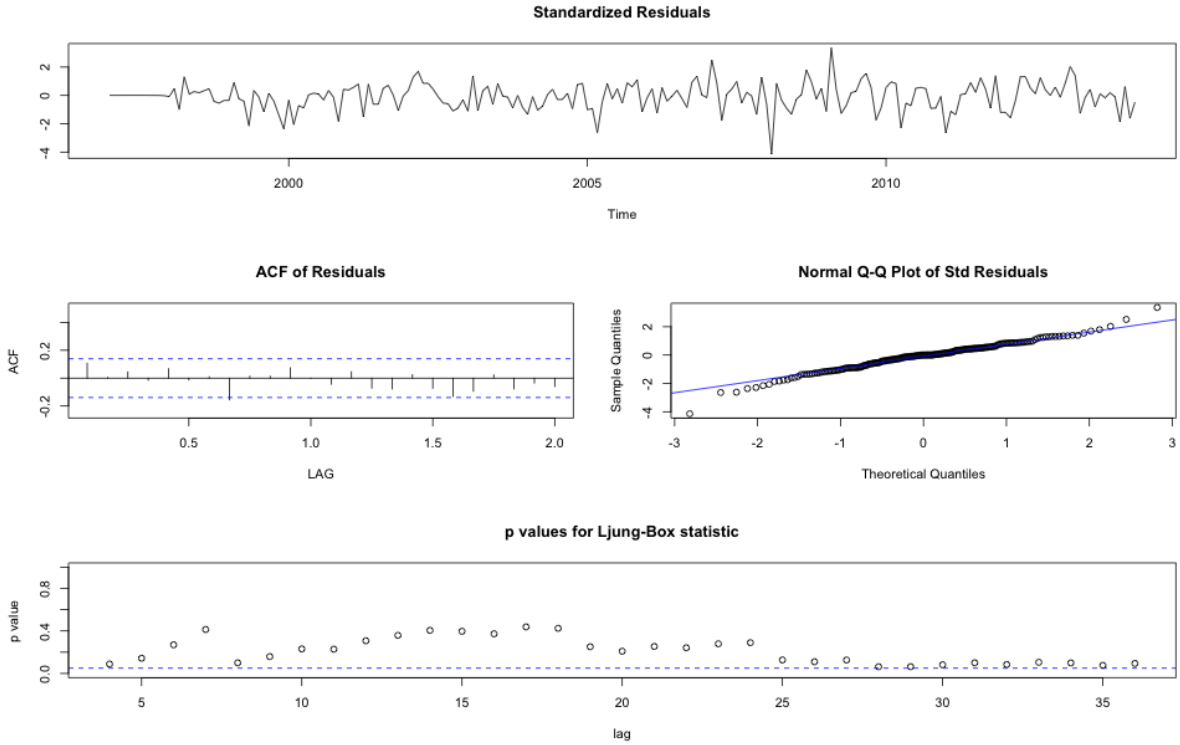
In this case, estimated seasonal MA1 and non-seasonal MA1 coefficient are both significant in 5% significant level. Hence the parameter has effect on the model and proceed to Ljung-Box Test.

Ljung-Box Test: The following tables shows the Ljung-Box statistics and its p-value:

Lag	Chi-Square	DF	p-value
6	4.0674	4	0.397
12	10.5829	10	0.3909
18	13.359	16	0.6463
24	19.5158	22	0.6133
30	32.3269	28	0.2614

As shown in the table, residual ACFs are jointly insignificant in 5% significance level. Therefore, we consider ARIMA(0,1,1)x$(0,1,1)_{12}$ an adequate model.

Overparameterization: For checking the adequacy of fitted model, we suggest to check the significance of an over fitted model ARIMA(0,1,1)x$(0,1,2)_{12}$ model.



Parameter	Estimate	Standard Error	p-value
MA1	-1.0000	0.0257	0.0000
SMA1	-1.0159	0.0857	0.0000
SMA2	0.1133	0.0812	0.1299
σ^2	0.848		
log likelihood	-274.32		
AIC	556.65		

In this case, the estimate of SMA1 model does not variate too much from original model, however, coefficient SMA2 is not significant in 5% confidence interval, which suggest that parameter SMA2 is over fitted.

Conclusion:

Based on model diagnostics, we conclude that the following model are candidate models that can be used for model selection:

Model
ARIMA(0,0,0)x$(0,1,1)_{12}$
ARIMA(0,0,0)x$(2,1,0)_{12}$
ARIMA(0,1,1)x$(0,1,1)_{12}$

Model Selection

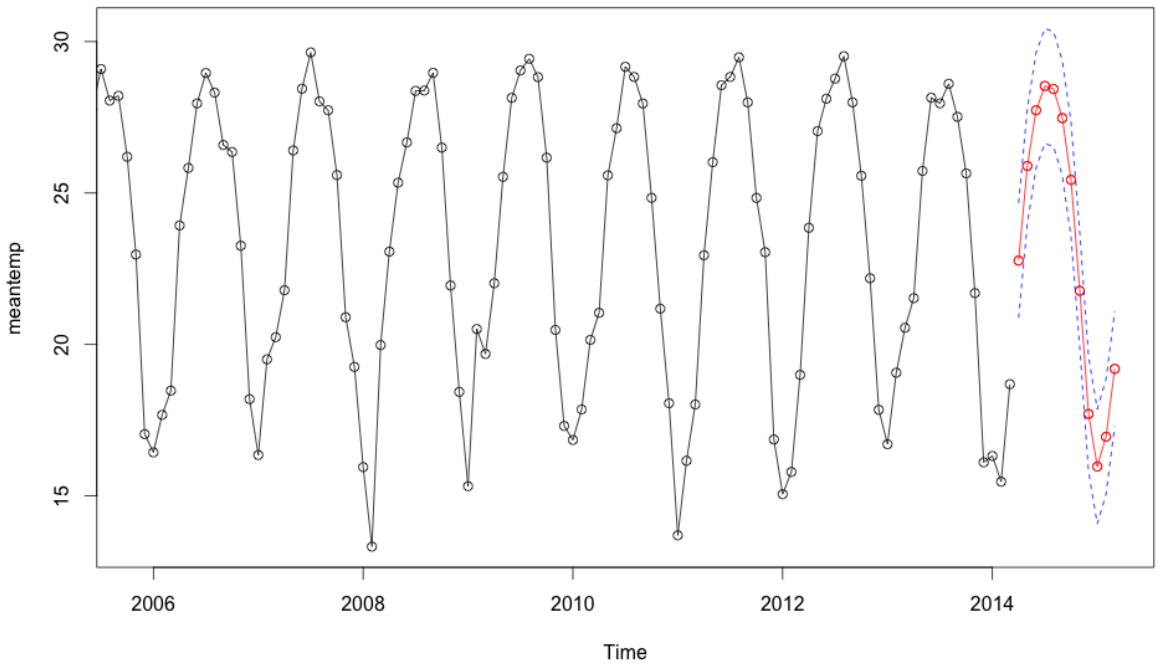
From the above section, we select three candidates model for model selection. In this section, we will identify the most suitable model by Akaike Information Criterion (AIC). The following table shows AIC of each candidate model:

Model	AIC
ARIMA(0,0,0)x$(0,1,1)_{12}$	516.06
ARIMA(0,0,0)x$(2,1,0)_{12}$	541.25
ARIMA(0,1,1)x$(0,1,1)_{12}$	556.43

Since the AIC of ARIMA(0,0,0)x$(0,1,1)_{12}$ is the smallest among other candidate models, we conclude that ARIMA(0,0,0)x$(0,1,1)_{12}$ would be the most suitable model.

Forecasting

After selecting a most suitable model, in this section, we are going to evaluate its forecasting power, by predicting the average monthly data in 2014-2015 (size=12). By fitting training data to ARIMA(0,0,0)x$(0,1,1)_{12}$ from 1997 to 2014, the following graph illustrates the forecast trend from 2014 to 2015.



Besides, the following table shows the forecast data with standard error and 95% confidence interval, as well as the comparison of true value and percentage error.

Obs	Forecast	Std Error	95% Confidence Limits	True Value	%Error
Apr 2014	22.76933	0.9487361	(20.9098, 24.6289)	22.58	0.84%
May 2014	25.89436	0.9487361	(24.0435, 27.7539)	26.39	-1.88%
Jun 2014	27.73325	0.9487361	(25.8737, 29.5928)	29.01	-4.40%
Jul 2014	28.53345	0.9487361	(26.6739, 30.3930)	29.78	-4.19%
Aug 2014	28.43383	0.9487361	(26.5743, 30.2934)	29.02	-2.02%
Sep 2014	27.47502	0.9487361	(25.6155, 29.3345)	28.99	-5.23%
Oct 2014	25.43610	0.9487361	(23.5760, 27.2950)	26.25	-3.10%
Nov 2014	21.77286	0.9487361	(19.9133, 23.6324)	22.64	-3.83%
Dec 2014	17.70631	0.9487361	(15.8468, 19.5658)	16.34	8.36%
Jan 2015	15.97050	0.9472709	(14.1138, 17.8272)	16.45	-2.91%
Feb 2015	16.95337	0.9472709	(15.0967, 18.8100)	17.48	-3.01%
Mar 2015	19.19828	0.9472709	(17.3416, 21.0549)	19.90	-3.53%

All true values are within 95% confidence limits and below 10% absolute percentage error. From the above results, the model is reasonably good for predicting future temperatures.

Evaluation

Regarding to the raising concern of extreme weather and global warming, what I originally believed is that temperature in Hong Kong will more or less depends on non-seasonal factors. Eventually, only seasonal factor is reflected from the model, which give us an insight that the non-seasonal trend might not be significant within 18 years. If we can observe for a longer time interval, it might be possible to find out non-seasonal factors

from trend.

References

[1] Robert H. Shumway & David S.Stoffer *Time Series Analysis and Its Applications With R Examples*.

[2] *OTexts*[online]. 8.7 ARIMA modelling in R Available from World Wide Web: (https://www.otexts.org/fpp/8/7).

[3] [onlinecourses]. 4.2 Identifying Seasonal Models and R Code Available from World Wide Web: (https://onlinecourses.science.psu.edu/stat510/node/68).

Appendix

Source Code

```
# Preliminary Packages astsa and forecast
# astsa: Applied Statistical Time Series Analysis
# forecast: Forecasting Functions for Time Series and Linear Models
# Install astsa and forecast by
# install.packages("astsa")
# install.packages("forecast")
library("astsa")
library("forecast")
#
# Self-defined function for batch
# Ljung-box test in lag 6,12,18,24 and 30
#
batchBoxTest <-function(model){
  df=model$fit$arma
  K=df[1]+df[3]+df[4]+df[6]
  for(h in c(6,12,18,24,30)){
    print(Box.test(model$fit$residuals,lag=h,fitdf=K))
  }
}
#
# Data Exploration and Transformation
#
# Defining time series data and
meantemp=ts(meantemp,freq=12,start=c(1997,1))
# display time plot, sample ACF
# and PACF of raw data
plot(meantemp,type="b")
tsdisplay(meantemp)
# Display time plot, sample ACF
# and PACF of log transformation
tsdisplay(log(meantemp))
```

```
# Display time plot, sample ACF
# and PACF of first common difference
tsdisplay(diff(meantemp,1))
# Display time plot, sample ACF
# and PACF of seasonal difference
tsdisplay(diff(meantemp,12))
#
# Model Diagnostics and Selection
#
# Fitting ARIMA(0,0,0)x(0,1,1) model and parameter estimation
sma1 <-sarima(meantemp,0,0,0,0,1,1,12,details = FALSE)
(1-pnorm(abs(sma1$fit$coef)/sqrt(diag(sma1$fit$var.coef))))*2
sma1$fit
batchBoxTest(sma1)
# Fitting ARIMA(0,0,0)x(0,1,2) model and parameter estimation
sma2 <-sarima(meantemp,0,0,0,0,1,2,12,details = FALSE)
sma2$fit
(1-pnorm(abs(sma2$fit$coef)/sqrt(diag(sma2$fit$var.coef))))*2
batchBoxTest(sma2)
# Fitting ARIMA(0,0,0)x(1,1,0) model and parameter estimation
sar1 <-sarima(meantemp,0,0,0,1,1,0,12,details = FALSE)
(1-pnorm(abs(sar1$fit$coef)/sqrt(diag(sar1$fit$var.coef))))*2
sar1$fit
batchBoxTest(sar1)
# Fitting ARIMA(0,0,0)x(2,1,0) model and parameter estimation
sar2 <-sarima(meantemp,0,0,0,2,1,0,12,details = FALSE)
sar2$fit
(1-pnorm(abs(sar2$fit$coef)/sqrt(diag(sar2$fit$var.coef))))*2
batchBoxTest(sar2)
# Fitting ARIMA(0,1,1)x(0,1,2) model and parameter estimation
ma1sma1 <-sarima(meantemp,0,1,1,0,1,1,12,details = FALSE)
ma1sma1$fit
(1-pnorm(abs(ma1sma1$fit$coef)/sqrt(diag(ma1sma1$fit$var.coef))))*2
batchBoxTest(ma1sma1)
# Fitting ARIMA(0,1,1)x(0,1,1) model and parameter estimation
ma1sma2 <-sarima(meantemp,0,1,1,0,1,2,12,details = FALSE)
ma1sma2$fit
(1-pnorm(abs(ma1sma2$fit$coef)/sqrt(diag(ma1sma2$fit$var.coef))))*2
batchBoxTest(ma1sma2)
#
# Forecasting
#
# Forecasting future 12 values by
# ARIMA(0,1,1)x(0,1,1) Model
sarima.for(meantemp,12,0,0,0,0,1,1,12)
```

Raw Data

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1997	16.69	16.15	20.06	22.75	26.06	27.33	28.11	28.38	26.41	25.90	23.14	18.91
1998	16.85	17.20	19.05	24.82	26.76	28.31	29.03	29.55	27.82	26.14	23.16	19.12
1999	17.35	18.65	20.45	24.27	24.88	28.90	29.15	28.33	27.84	26.20	22.20	16.83
2000	17.00	15.36	19.16	23.08	26.09	28.28	28.85	28.51	27.69	25.96	21.01	18.85
2001	17.29	17.61	20.56	22.37	26.70	27.60	28.14	29.18	28.12	26.09	21.48	18.26
2002	17.29	18.19	21.48	24.60	27.03	28.82	28.88	28.44	27.16	25.15	21.52	18.18
2003	16.09	18.47	19.05	23.94	26.86	27.66	29.57	28.84	27.57	25.26	22.19	17.61
2004	15.82	17.20	19.10	22.90	26.30	28.56	28.42	28.43	27.62	24.92	22.69	19.07
2005	15.91	16.47	17.37	23.09	26.97	27.76	29.09	28.05	28.21	26.20	22.97	17.04
2006	16.44	17.68	18.48	23.93	25.83	27.96	28.96	28.32	26.59	26.35	23.26	18.20
2007	16.35	19.51	20.24	21.79	26.41	28.44	29.64	28.03	27.73	25.59	20.90	19.26
2008	15.95	13.33	19.98	23.07	25.34	26.67	28.37	28.38	28.97	26.50	21.95	18.44
2009	15.32	20.51	19.69	22.02	25.54	28.14	29.05	29.43	28.83	26.17	20.48	17.31
2010	16.85	17.86	20.15	21.05	25.58	27.14	29.17	28.83	27.96	24.84	21.18	18.06
2011	13.70	16.16	18.02	22.95	26.02	28.56	28.83	29.48	27.99	24.84	23.05	16.86
2012	15.06	15.79	19.00	23.85	27.05	28.11	28.78	29.52	27.99	25.57	22.19	17.85
2013	16.71	19.07	20.55	21.53	25.73	28.15	27.96	28.61	27.52	25.65	21.70	16.11
2014	16.32	15.47	18.69	22.58	26.39	29.01	29.78	29.02	28.99	26.25	22.64	16.34
2015	16.45	17.48	19.90									