Applying GLM Model and ARIMA Model to the Analysis Of Monthly Temperature of Stockholm

Author: Xier Li

Supervisor: Mikael Möller

June 10, 2009

D-level Essay in Statistics in Spring 2009

Department of Economics and Society, Dalarna University

Abstract

Over the past centuries, climate change has had a great influence on natural ecosystems and social economic, so studies on temperature have become increasingly important in recent years. Stockholm temperature has been recorded for a long time from 1756 to 2007. The main aim of this paper is to check whether the Stockholm monthly temperature series can be analyzed by statistical method, and try to build general linear model (GLM) and ARIMA models to fit the data. Data used in this paper has been adjusted by Anders Moberg and his colleagues. Based on the features of the data, we divide the time period between the year of 1756 and that of 2007 into three periods: 1756-1925, 1926-1985 and 1986-2007, and try to build general linear models (GLM) and ARIMA models to fit the data in the three periods. Then we forecast the monthly temperature of 2008 and compare them with the true values. Compare the results and we find the Seasonal ARIMA (SARIMA) model for the series fits the data better than the general linear model.

Key words: General linear model, Box-Jenkins methodology, ARIMA model, SARIMA model

Contents

1.Introduction

1.1 Background

The temperature increase effects the managed and human systems, such as agricultural and forestry management, human health and human activities in Arctic.[1] So it is meaningful to study the temperature. In this paper, we study the long range of the Stockholm temperature data, which ranges from 1756 to 2007. Moberg[1] and his colleagues corrected the data, due to different placement of thermometers, the quality of thermometers, missing observation and so on.

1.2 Aim

The aim of this paper is to apply general linear model(GLM) and ARIMA (autoregressive integrated moving average) model to analyze the series and check whether we can use statistical method to analyze the Stockholm monthly temperature series. We calculate the average temperature of every 10 years between 1756-2005 and 2006-2007. According to the average temperature, we divide the time period between the year of 1756 and that of 2007 into three periods: 1756-1925, 1926-1985, 1986-2007. Then we check whether the structures in the three periods are stable. At last, we the forecast the monthly temperature of 2008 and compare the results with the true values.

[1] Department of Physical Geography, Stockholm University

2. Data Description

As early as in 1778, the temperature series from Stockholm and Uppsala were used in a study of climatic changes. We used the daily average data given by Moberg from 1756 to 2007. The data is after homogenization and with gaps filled in using data from Uppsala and transformed. We transform it into monthly average temperature. The temperature records from Uppsala and Stockholm are homogeneity [2] tested by Alexandersson[3] and and Moberg. Using the sum of daily data to divide the number of the days in the month, then we get monthly data. There are 3024 months measured from the first month of 1756 to the last month of 2007. The table of the transformed temperature series sample is given in the Appendix I: Table 1. The plot of the sample of the monthly temperature series is given in Figure 1 below.

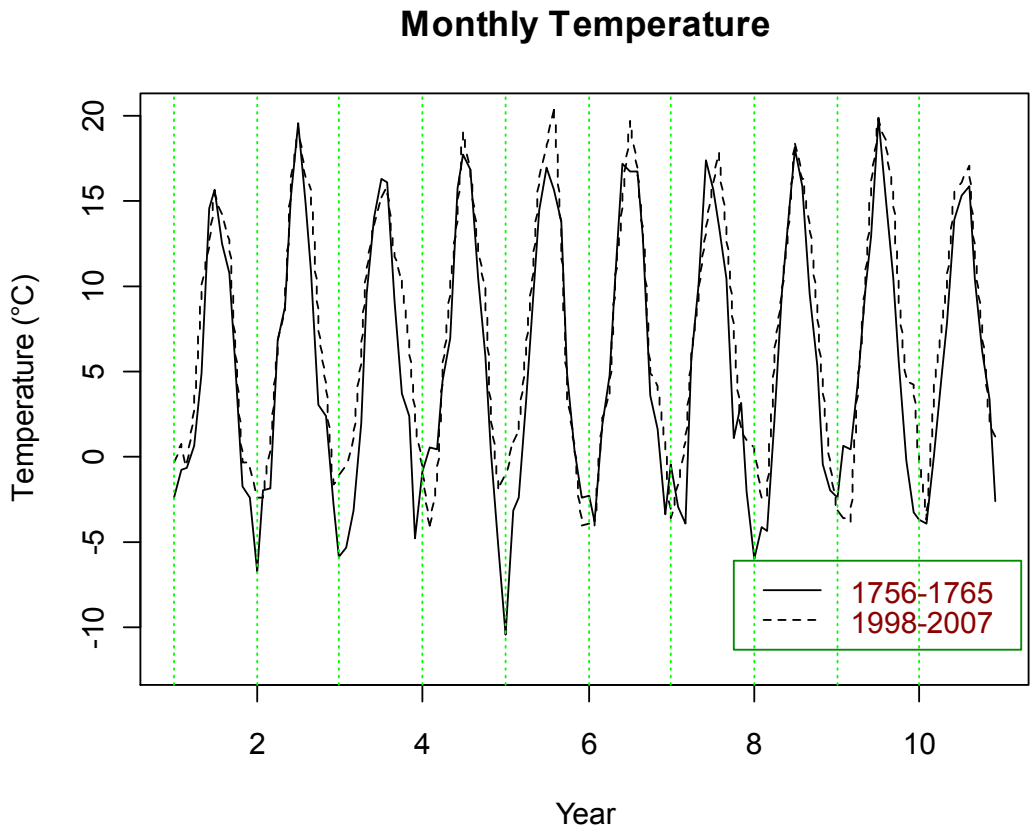

Figure 1 The monthly average temperature of 1756-1765 and 1998-2007

[2] A time series of a climatological variable where the variations are caused by variations of weather and climate only is said to be homogenous.[2]
[3] Swedish Meteorogicaland Hydrological Institute

3. Statistical Methodology

3.1 General Linear Models (GLM)

In a general linear model (GLM), the observed value of the dependent variable y for observation number i (i =1,2,…,n) is modeled as a linear function of (p-1), so called independent variables $x_1, x_2, ..., x_{p-1}$ as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i(p-1)} + e_i$$

or in matrix terms

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{e}$$

Y is a vector of observations on the dependent variable. **X** is a known matrix of dimension $n \times p$, called a design matrix that contains the values of the independent variables and one column corresponding to the intercept. $\boldsymbol{\beta}$ is a vector containing p parameters to be estimated (including the intercept) and **e** is a vector of residuals. It is common to assume that the residuals e are independent, normally distributed and that the variances are the same for all e_i.

Estimation of parameters in general linear model is often done using the method of least squares. For normal theory models this is equivalent to Maximum Likelihood estimation. The parameters are estimated with those values for which the sum of the squared residuals, $\sum e_i^2$ is minimal. In matrix terms, the sum of squares is

$$\mathbf{e'e = (Y - X\beta)'(Y - X\beta)}$$

Minimizing the above equation with respect to the parameters in $\boldsymbol{\beta}$ gives the normal equations

$$\mathbf{X'X\beta = X'Y}$$

If the **X'X** is nonsingular, this yields, as estimators of the parameters of the model, [3]

$$\hat{\boldsymbol{\beta}} = \mathbf{(X'X)^{-1}X'Y}$$

3.2 Box-Jenkins methodology

In econometrics, the Box-Jenkins methodology, named after the statisticians George Box and Gwilym Jenkins, applies autoregressive moving average ARMA or ARIMA models to find the best fit of a time series to past values of this time series, in order to make forecasting. The Box-Jenkins methodology consists of a four-step iterative procedure: tentative identification, estimation, diagnostic checking and forecasting.[4]

The first step in developing a Box-Jenkins model is to determine if the time series is stationary and if there is any significant seasonality that needs to be modeled. Stationarity can be assessed from an autocorrelation plot. Specifically, nonstationarity is often indicated by an autocorrelation plot with very slow decay. In time series models, a linear stochastic process has a unit root if 1 is a root of the process's characteristic equation. The process will be non-stationary. If the other roots of the characteristic equation lie inside the unit circle, then the first difference of the process will be stationary.[5] An augmented Dickey–Fuller test (ADF) is a test for a unit root in a time series sample. The augmented Dickey–Fuller (ADF) statistic, used in the test, is a negative number. The more negative it is, the stronger the rejection of the hypothesis that there is a unit root at some level of confidence.[6] At the model identification stage, the goal is to detect seasonality, if it exists, and to identify the order for the seasonal autoregressive and seasonal moving average terms. For many series, the period is known and a single seasonality term is sufficient. However, it may be helpful to apply a seasonal difference to the data and regenerate the autocorrelation and partial autocorrelation plots. This may help in the model identification of the non-seasonal component of the model. In some cases, the seasonal differencing may remove most or all of the seasonality effect.

3.2.1 Autoregressive integrated moving average (ARIMA) models

In statistics, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average or (ARMA) model. These models are fitted to time series data either to better understand the data or to predict future points in the series. The ARIMA model is applied in some cases where data show evidence of nonstationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied to remove the nonstationarity.
The model is generally referred to as an ARIMA(p,d,q) model where p, d, and q are integers greater than or equal to zero and refer to the order of the autoregressive, integrated, and moving average parts of the model respectively. The first parameter p refers to the number of autoregressive lags (not counting the unit roots), the second parameter d refers to the order of integration, and the third parameter q gives the

number of moving average lags. ARIMA models form an important part of the Box-Jenkins approach to time-series modelling.

A process, $\{x_t\}$ is said to be ARIMA (p,d,q) if $\nabla^d x_t = (1-B)^d x_t$ is ARMA(p,q).In general, we will write the model as

$$\phi(B)(1-B)^d x_t = \theta(B)\omega_t, \quad \{\omega_t\} \sim WN(0,\sigma^2)$$

WN stands for white noise. Here, we define the backshift operator by $B^k x_t = x_{t-k}$ and the autoregressive operator and moving average operator are defined as follows:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$$

$\phi(B) \neq 0$ for $|B| \leq 1$, the process $\{x_t\}$ is stationary if and only if d=0, in which case it reduces to an ARMA(p,q) process.

Table 1 Behavior of the ACF and the PACF for ARMA Models [7]

	AR(p)	*MA(q)*	ARMA(p,q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

3.2.2 Multiplicative Seasonal ARIMA Models

Often, the dependence on the past tend to occur most strongly at multiples of some underlying seasonal lag s. Natural phenomena such as temperature also have strong components corresponding to seasons. Because of this, it is appropriate to introduce autoregressive and moving average polynomials that identify with the seasonal lags. The resulting pure seasonal autoregressive moving average model, say, ARMA$(p,q)_s$, then takes the form

$$\Phi_P(B^S)x_t = \Theta_Q(B^S)\omega_t$$

with the following definition.
The operators

$$\Phi_P(B^S)x_t = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdots - \Phi_P B^{Ps}$$

and

$$\Theta_Q(B^S) = 1 + \Theta_1 B^S + \Theta_2 B^{2S} + \cdots + \Theta_Q B^{Qs}$$

are the seasonal autoregressive operator and the seasonal moving average operator of orders P and Q, respectively, with seasonal period s.

The multiplicative seasonal autoregressive integrated moving average model, or SARIMA model, of Box and Jenkins is given by

$$\Phi_P(B^S)\phi(B)\nabla_S^D \nabla^d x_t = \alpha + \Theta_Q(B^S)\theta(B)\omega_t$$

where ω_t is the usual Gaussian white noise process. The general model is denoted as ARIMA(p,d,q) $\times (P,D,Q)_S$. The non-seasonal difference components are $\nabla^d = (1-B)^d$ and the seasonal are $\nabla_S^D = (1-B^S)^D$.

Table 2 Behavior of the ACF and the PACF for Pure Seasonal ARMA Models [7]

	AR(P)s	*MA(Q) s*	ARMA(P,Q) s
ACF	Tails off at lags ks, k=1,2,…,	Cuts off after lag Qs	Tails off at lag ks
PACF	Cuts off after lag Ps	Tails off at lags ks k=1,2,…,	Tails off at lag ks

4. Model Building

4.1 General Linear Model

Figure 1(in page 2), suggests that the monthly temperature series behave like a trigonometric function curve. So we try to build a model y_t=f(sin$t$,cos$t$).

Let y_t represents the monthly temperature series and z_t denote the time effect. Because we want the annual cycle is of length 1 and try to find a variable relate it to time , we set:

z_t= i / (the number of monthly measurements of y_t/the amount of the years in the period)

i=1,2,3…., the number of monthly measurements.

$$y_t = 5.71 - 6.53 \times \sin(2\pi \times z_t) - 7.95 \times \cos(2\pi \times z_t) \qquad \text{(Model 1)}$$

T-statistic: (142) (-115.4) （-140.6）

R-squared: 0.9163, Adjusted R-squared:0.9613, F-statistic: 1.654e+04

All of the coefficients are significant. When we observe the plot (Figure 1 in page 2) of the monthly temperature series, we can find that the trend of the temperature is on a slight increase. Thus, we try to divide the temperature series into sub series in order to study the trend. The monthly data consist of 252 years, we divide the monthly data from 252 years into 26 series with 10 years each and the last series from 2006 to 2007. The mean temperature of the sub series is shown in Appendix I: Table 2. We find the mean temperature of the sub series from 1756 to 1925 are less than 6.0°C except 1816-1825, from 1926 it becomes higher than 6.0°C and from 1986 it increases to 6.4 °C .

The mean temperatures of the three periods are:5.5°C, 5.9°C, 6.7°C. Study the mean temperatures of the three periods, we can find Stockholm temperature remains relatively constant in the 1756-1925 period. In the 1926-1985 period, annual temperature rises at a rate of 0.007°C. In the 1986-2007 period, the rate 5 times to 0.036°C per year. Therefore, we try to build three models according to the three periods to see the trend of the series, and the results are as follows:

1756-1925: $y_t = 5.50 - 6.58 \times \sin(2\pi \times z_t) - 8.04 \times \cos(2\pi \times z_t)$ (Model 2)

T-statistic: (112.31) （-94.96） (-116.16)

R-squared: 0.917, Adjusted R-squared:0.9169, F-statistic: 1.126e+04

All the coefficients are significant.

1926-1985: $y_t = 5.93 - 6.59 \times \sin(2\pi \times z_t) - 7.80 \times \cos(2\pi \times z_t)$ (Model 3)

T-statistic: (77.58) (-61.00) (-72.16)

R-squared: 0.9257, Adjusted R-squared:0.9255, F-statistic: 4464

All the coefficients are significant.

1986-2007: $y_t = 6.73 - 5.99 \times \sin(2\pi \times z_t) - 7.66 \times \cos(2\pi \times z_t)$ (Model 4)

T-statistic: (51.56) (-32.46) (-41.51)

R-squared: 0.9141, Adjusted R-squared:0.9134, F-statistic: 1388

All the coefficients are significant.

Compare intercepts and coefficients of the above 3 models: Model 2, Model 3 and Model 4, the intercepts are equal to the mean temperature of the periods and increase. The reason may be different trends in different periods and the trends can be reflected in the models for different periods.

Using Model 1 and Model 4 to forecast the monthly temperature of 2008, we can see the results in Table 3 (next page). Comparing the prediction values of the two models, we can find all of the prediction values in Model 1 are lower than the true value of 2008 and the prediction value of Model 4 is higher than the prediction value of Model 1. The prediction value of Model 4 is closer to the true value than the prediction value

of model 2. The reason why these prediction values are lower than the true value may be that the temperature increases year by year, but it may not be reflected in one general linear model for the series. Observing the prediction temperature of Model 4, we can find the prediction values in December, January, February differ greatly with the true value, but the value of other months are close to the true value.

Table 3 Prediction values of the general linear model

2008	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
prediction temperature of model 1	-4.4	-3.9	-0.8	4.0	9.3	13.7	15.9	15.3	12.2	7.4	2.1	-2.2
Prediction temperature of model 4	-2.8	-2.2	0.7	5.3	10.4	14.4	16.4	15.7	12.7	8.0	3.0	-0.9
true temperature	1.4	2.1	0.8	6.5	10.7	15.4	17.8	15.3	11.1	8.1	2.9	0.9

4.2 Seasonal ARIMA (SARIMA) model

4.2.1 Identification

Noting the Figure 2 (next page) ACF for the series, the peak at 12,24,36 and 48 with relatively slow decay suggested nonstationary and a seasonal difference. In order to obtain a stationary series, we decide to take 12 months differences of data to remove the seasonal influence. After differencing the series, we check the ACF and PACF plot and use ADF test to see whether the seasonal differenced series is stationary. At the nonseasonal levels, the ACF has significant spikes at lag 1 and tails off after lag 1. The PACF has a spike at lag 1 and cuts off at lag 1. At the seasonal level, the ACF has spike at lag 12 and cuts off after lag 12 and the PACF tails off after lag 12. There are only a few lags slightly outside the confidence limits. The p value of the ADF test[4]is smaller than 0.01, this indicates the seasonal differenced series is stationary.

[4]The null hypothesis of the ADF test is that a unit root exists in the time series.

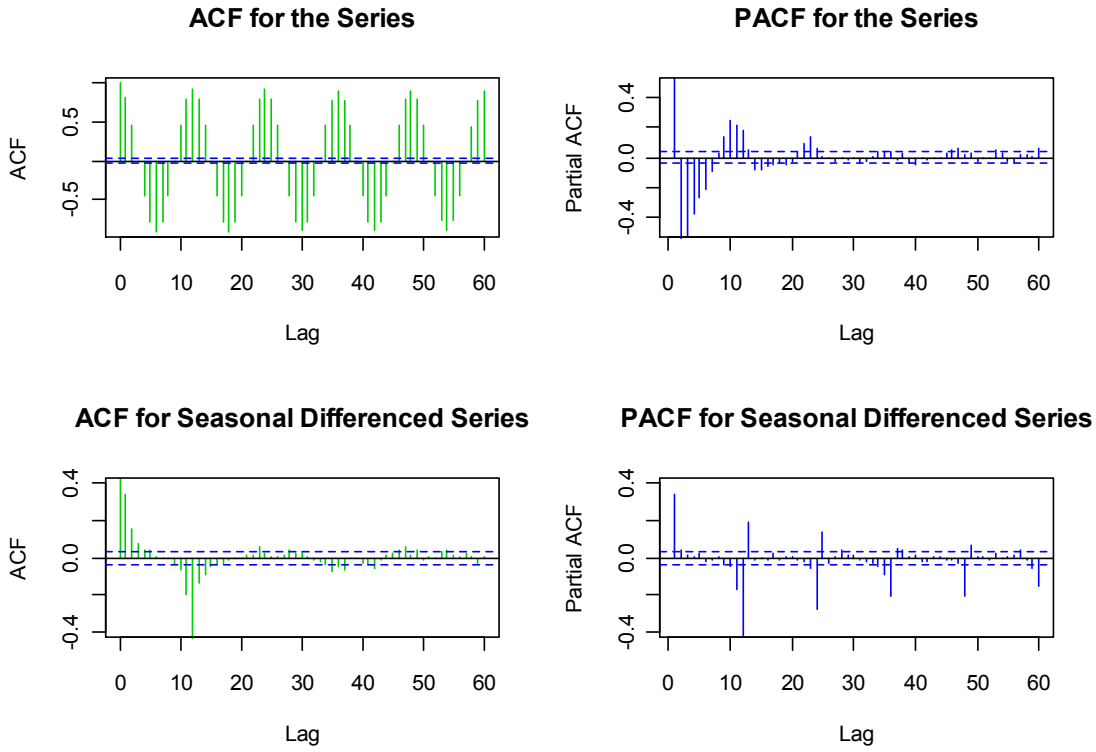

Figure 2 the ACF and PACF for monthly average temperature series and the seasonal differenced series

According to the Figure 2, the PACF has significant spikes at lag 1 and cuts off after lag 1 at the nonseasonal level and the ACF are tailing off. Using Table 1（page 5）, it suggests a nonseasonal autoregressive of order p=1.Using x_t to denote the differenced monthly temperature series and y_t present the original monthly temperature series, we can identify the following nonseasonal autoregressive model.

$$x_t = \phi_1 x_{t-1} + \omega_t$$

From Figure 2,characteristics of the ACF and PACF of this series tells us that the ACF has a spike at the seasonal lag 12 and cuts off after lag 12 and the PACF tails off at the seasonal level, then we might tentatively conclude that the time series values are described by the seasonal moving average model of order Q=1 depending on the Table, this is to say,

$$x_t = \omega_t + \Theta_1 \omega_{t-12}$$

Combing these two models, we obtain the overall model

$$x_t = \phi_1 x_{t-1} + \omega_t + \Theta_1 \omega_{t-12}$$

4.2.2 Estimation

The fitted ARIMA$(1,0,0)\times(0,1,1)_{12}$ into

$$\phi(B)\nabla^d x_t = \Theta_Q(B^S)\omega_t$$

then

$$(1-\phi_1 B)(1-B)^0 x_t = (1+\Theta_1 B^{12})\omega_t$$

Because taking 12^{th} differences, the original series can be written as $x_t = \nabla_{12}^1 y_t = (1-B^{12})y_t$,

$$(1-\phi_1 B)(y_t - y_{t-12}) = \omega_t + \Theta_1\omega_{t-12}$$

or in difference equation form:

$$y_t = \phi_1 y_{t-1} + y_{t-12} - \phi_1 y_{t-13} + \omega_t + \Theta_1\omega_{t-12}$$

Therefore, we obtain the final model for the series from1756 to 2007 as follows:

$$y_t = 0.3663y_{t-1} + y_{t-12} - 0.3663y_{t-13} + \omega_t - 0.9754\omega_{t-12} \qquad \text{(Model 5)}$$

s.e. (0.0170) (0.0170) (0.0050)

with $\hat{\sigma}_t^2 = 3.858$.

Similarly, we can build the seasonal ARIMA models for the sub series of period 1756-1925, 1926-1985,1986-2007. Checking the behavior of the ACF and PACF plots, they also fit well by the SARIMA model above. The models are as follows:

1756-1925: $y_t = 0.3406y_{t-1} + y_{t-12} - 0.3406y_{t-13} + \omega_t - 0.9795\omega_{t-12}$ (Model 6)

s.e. (0.0209) (0.0209) (0.0063)

with $\hat{\sigma}_t^2 = 4.019$.

1926-1985: $y_t = 0.4177y_{t-1} + y_{t-12} - 0.4177y_{t-13} + \omega_t - 0.9752\omega_{t-12}$ (Model 7)

s.e. (0.0342) (0.0342) (0.0274)

with $\hat{\sigma}_t^2 = 3.32$.

1986-2007: $y_t = 0.3417y_{t-1} + y_{t-12} - 0.3417y_{t-13} + \omega_t - 0.9546\omega_{t-12}$ (Model 8)

s.e. (0.0601) (0.0601) (0.0909)

with $\hat{\sigma}_t^2 = 3.891$.

Observing Model 5, Model 6 and Model 8, we can also find they are similar because the coefficients are very close, One of the coefficients 0.4177 in Model 7 is slightly

different from others.

4.2.3 Diagnostic checking

We can use the white noise test[5] to check whether the residuals of the Seasonal ARIMA model is white noise .We find the p value of the white noise test is 0.727 ,we can't reject the null hypothesis, so the residuals of the SARIMA model is white noise . From the Figure 3, it displays the standardized residuals, the ACF of the residuals and the value of the Q-statistic[6] .

The standardized residuals plot in Figure 3 top shows us the residuals are with the same mean 0 and variance σ^2. From the Figure 3 middle, it shows none of the autocorrelations is individually statistically significant, it means the residuals of the series are independent. The p values for the Ljung-Box statistic are all greater than 0.05, so we can't reject the null hypothesis, it also means the data is independent.

After checking the residuals, we can say that the identified model fits the data well, leading to the conclusion that the model is adequate.

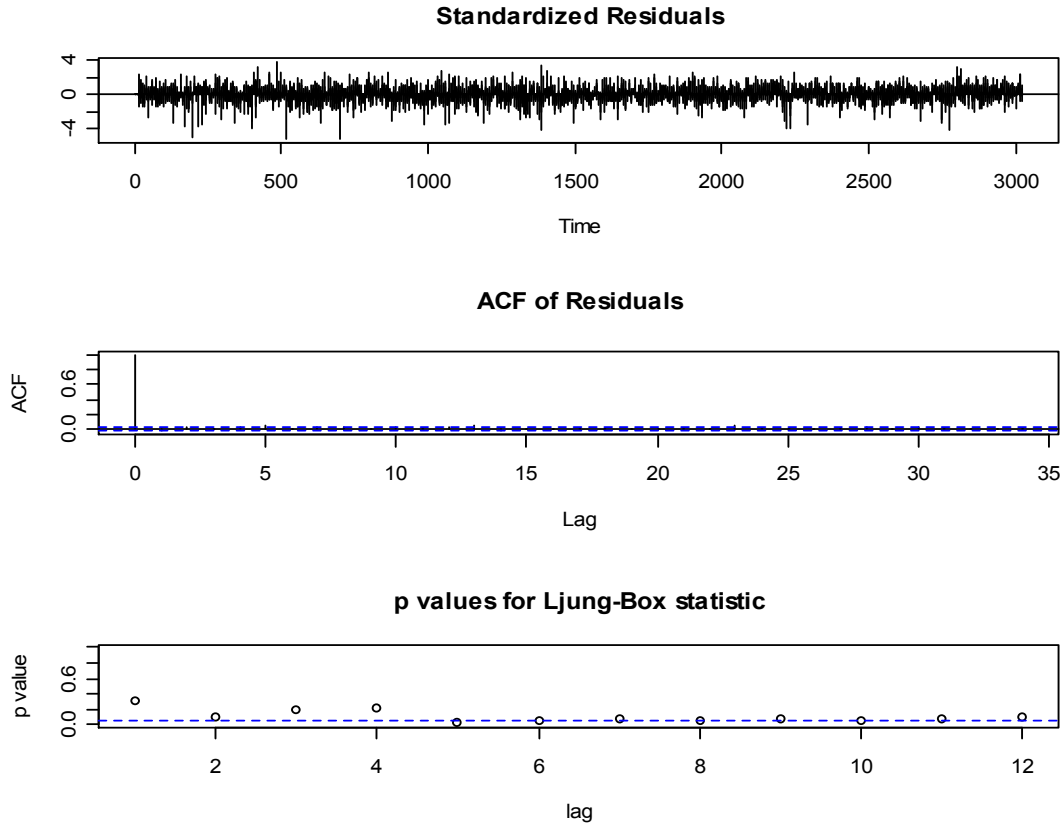

Figure 3 Diagnostics for the SARIMA model

[5] From the normwn.test package in R, perform a univariate test for white noise. The null hypothesis is the residuals are white noise.

[6] The Q-statistic is a test statistic output by either the Box-Pierce test or, by the Ljung-Box test. It follows the chi-squared distribution.

4.2.4 Forecasting

We use the SARIMA model for the whole series to forecast the monthly temperature of the 2008 and compare it with the true monthly temperature of 2008. We also compare the prediction results of the whole series and the sub series from 1986-2007, they are very close. The result is in Appendix I Table 3. The result for the whole series is in Table 3 and the plot of the result is in Figure 4. Compared the prediction temperature with the true temperature, we can find the prediction value is very close to the true value except January and February, and all of the true values fall into the confidence interval except February. So, we can say, the SARIMA model can be used to analyze the Stockholm temperature series.

Table 4 Forecasts based on the SARIMA model for the next 12 month:

2008	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
predict temperature	-2	-2.9	-0.4	4.2	9.4	14.3	16.9	16.1	11.9	7.1	2.3	-0.9
95%lowerCI	-5.9	-7.1	-4.6	0	5.2	10	12.7	11.9	7.7	2.8	-2	-5.1
95%upperCI	1.9	1.3	3.8	8.4	13.7	18.5	21.1	20.3	16.2	11.3	6.5	3.3
true temperature	1.4	2.1	0.8	6.5	10.7	15.4	17.8	15.3	11.1	8.1	2.9	0.9

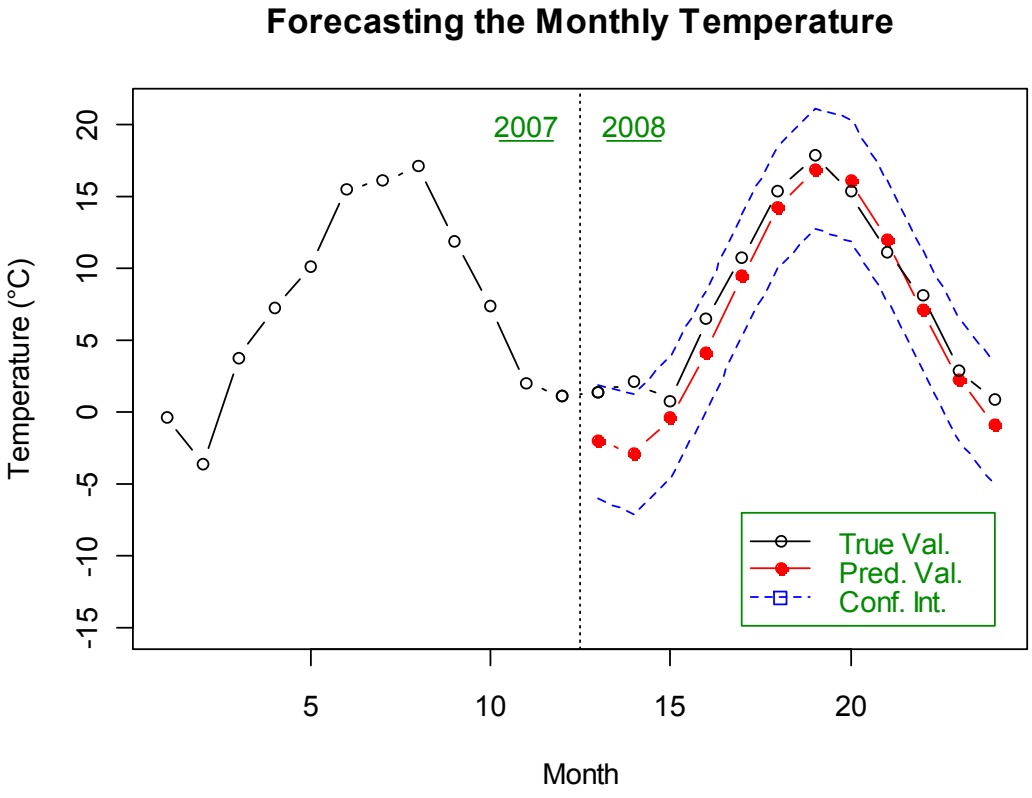

Figure 4 Forecasting the monthly temperature base on the SARIMA model

5. Discussion

5.1 Residuals of General Linear Model

From the ACF plot of general linear model in Figure 5, there are some spikes, it tells us there are autocorrelations in the general linear models. The GLM models need to be adjusted to make it better. If we need further study the we should take the pseudoreplication into account in the general linear model, but we don't emphasis on this papar.

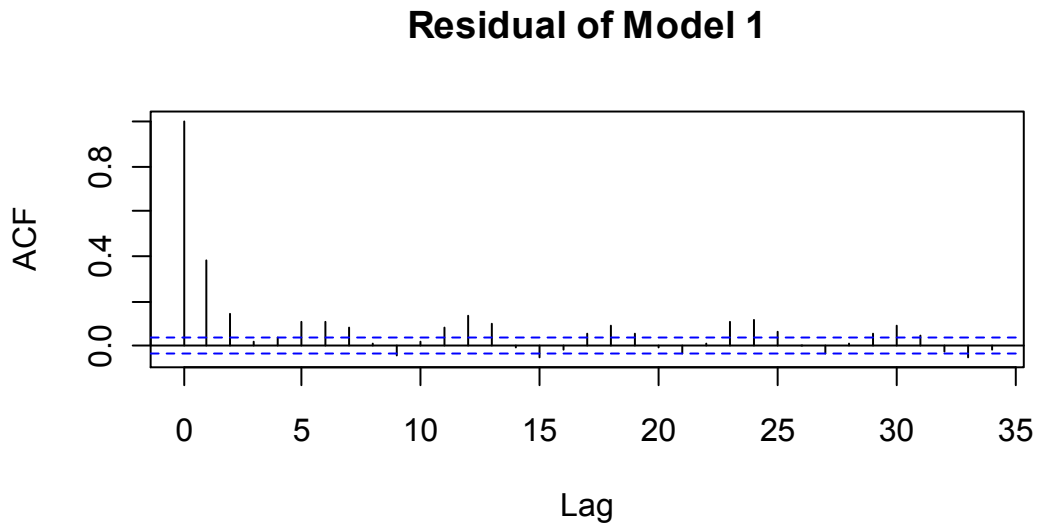

Figure 5: Residual of General Linear Model

5.2 Forecasting values

From Table 5 (next page), the forecasting values of ARIMA model (Model 5) are closer to the true value than general linear model (Model 1). Both forecasting values of January and February predicted by the general linear model and SARIMA model are inaccuracy. We still need further studies the reason why these two values are inaccuracy.

Table 5 Forecasting Values of Model 1 and Model 5

2008	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
prediction temperature of Model 1	-4.4	-3.9	-0.8	4.0	9.3	13.7	15.9	15.3	12.2	7.4	2.1	-2.2
Prediction temperature of Model 5	-2.0	-2.9	-0.4	4.1	9.4	14.3	16.9	16.1	11.9	7.0	2.2	-0.9
true temperature	1.4	2.1	0.8	6.5	10.7	15.4	17.8	15.3	11.1	8.1	2.9	0.9

6. Conclusion

The sub series in the three periods: 1756-1925, 1926-1985, 1986-2007 have the similar stable structures in general linear models and Seasonal ARIMA models. However, the monthly temperature in the three periods has different trends. Thus, the temperature series from 1756-2007 may not reflect precisely the trend of the series in general linear model. The forecasting results also tell us the GLM model of the sub series 1986-2007 fits better than the series 1756-2007. When we use the SARIMA model, we differentiate the series, so the seasonal differenced series removes the trend effect and the series become stationary. The SARIMA models for sub series also show the monthly average temperature have a stable structure. The forecasting results show the SARIMA models fits the data well. So the Stockholm monthly temperature can be analyzed by statistical method.

References

[1] PICC, *Climate Change 2007 Synthesis Report*, 17 November 2007, pp.11 http://www.ipcc.ch/pdf/assessment-report/ar4/syr/ar4_syr.pdf

[2] Anders Moberg, Hans Bergström, *"Homogenization of Swedish temperature data. Part III: The long temperature records from Uppsala and Stockholm"*

[3] Ulf Olsson, *Generalized linear models : an applied approach*, pp.2-10

[4] BoIrman, O'Connell, Koehler, *Forecasting, Time Series, And Regression*, Fourth Edition

[5] James D. Hamilton, *Time Series Analysis*, Princeton University Press, pp.12-18

[6] Elliott, G., Rothenberg, T. J. & J.H. Stock (1996) "Efficient Tests for an Autoregressive Unit Root", *Econometrica*, Vol. 64, No. 4., pp. 813–836

[7] Robert H. Shumway，David S. Stoffer，*Time Series Analysis and Its Applications With R Examples*，Second Edition, pp.84-154

[8] J.Xiang, " *Applying ARIMA model to the analysis of Monthly Temperature of Stockholm*" http://www.statistics.du.se/essays/D08D_XiangJunquan.pdf

[9] Torbjorn Lorentzen, *Global warming an univariate estimation of sea temperature data*, SNF Project No. 5015

[10] G. M. Ljung, G. E. P. Box (1978), "On a Measure of a Lack of Fit in Time Series Models". *Biometrika* 65: pp. 297-303.

Appendix I: Tables

Table 1 The Monthly Temperature（°C）of 1756-1765 and 1998-2007

Year	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1756	-2.4	-0.8	-0.6	0.7	4.9	14.5	15.7	12.5	10.7	5.6	-1.7	-2.4
1757	-6.7	-1.9	-1.8	6.9	8.8	15.3	19.6	14.9	10.9	3.1	2.4	-2.7
1758	-5.8	-5.3	-3.2	2.0	9.6	14.1	16.3	16.1	9.1	3.7	2.5	-4.7
1759	-0.9	0.5	0.5	4.4	6.9	15.8	17.8	16.9	10.6	5.9	0.3	-5.2
1760	-10.4	-3.1	-2.4	3.3	9.5	14.5	16.9	15.7	13.9	5.0	0.5	-2.3
1761	-2.3	-3.7	1.8	4.8	10.4	17.2	16.8	16.7	12.8	3.6	1.7	-3.3
1762	-0.5	-2.9	-3.8	5.7	10.2	17.3	15.7	13.5	10.4	1.2	3.1	-2.1
1763	-5.9	-4.1	-4.3	2.2	9.6	14.7	18.1	15.9	9.6	5.3	-0.4	-2.0
1764	-2.4	0.6	0.5	4.5	9.8	12.9	19.9	15.1	10.6	5.7	-0.2	-3.2
1765	-3.6	-3.8	0.0	4.7	7.9	13.9	15.4	15.9	10.4	6.2	3.4	-2.5
1998	-0.3	0.8	-0.5	3.1	9.8	12.4	15.5	14.1	12.7	6.6	-0.3	-0.4
1999	-2.3	-2.4	0.7	6.6	8.9	16.0	19.2	16.4	15.7	7.8	4.3	-1.6
2000	-1.0	-0.4	1.2	6.0	11.1	13.6	15.3	15.7	11.7	10.3	6.6	2.3
2001	-0.7	-4.2	-1.3	5.0	9.9	14.3	19.0	16.6	12.3	9.7	2.6	-2.0
2002	-1.1	0.9	1.6	6.5	11.4	16.4	18.2	20.6	13.2	3.7	0.5	-4.0
2003	-3.9	-4.0	2.3	3.7	10.3	15.0	19.7	16.9	13.3	4.9	4.1	0.9
2004	-3.5	-1.6	0.9	6.0	9.6	13.4	16.1	17.8	13.2	7.5	1.5	1.0
2005	0.6	-2.5	-2.3	6.0	9.7	14.0	18.5	16.3	13.5	8.7	4.0	-0.5
2006	-3.1	-3.5	-3.8	4.5	10.1	15.9	19.8	18.4	15.2	9.6	4.4	4.3
2007	-0.3	-3.6	3.8	7.3	10.2	15.5	16.1	17.1	11.9	7.4	2.0	1.2

Table 2 The Mean temperature for the sub series

Year	Mean temperature (°C)
1756-1765	5.5
1766-1775	5.7
1776-1785	5.7
1786-1795	5.8
1796-1805	5.3
1806-1815	5.2
1816-1825	6.0
1826-1835	5.5
1836-1845	5.0
1846-1855	5.6
1856-1865	5.6
1866-1875	5.2
1876-1885	5.2
1886-1895	5.3
1896-1905	5.6
1906-1915	5.9
1916-1925	5.6
1926-1935	6.0
1936-1945	6.0
1946-1955	6.2
1956-1965	5.6
1966-1975	6.1
1976-1985	5.5
1986-1995	6.4
1996-2005	6.9
2006-2007	7.5

Table 3 Forecasting based on the SARIMA Model 8 for the next 12 month:

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
temperature	-1.3	-2.0	0.5	4.9	9.9	14.4	17.4	16.6	12.3	7.3	2.5	-0.4
95%lowerCI	-5.2	-6.2	-3.8	0.7	5.7	10.2	13.1	12.4	8.1	3.1	-1.7	-4.7
95%upperCI	2.7	2.2	4.7	9.1	14.2	18.6	21.6	20.9	16.6	11.5	6.8	3.8
true temperature	1.4	2.1	0.8	6.5	10.7	15.4	17.8	15.3	11.1	8.1	2.9	0.9

Appendix II: R codes

```
####################################
# Get the monthly data from daily data    #
####################################
d<-read.table("D:/stockholm_td_adj.dat",header=T)
month<-tapply(d$X.8.7.2,list(d$X1756,d$X1),mean)
month    #monthly temperature
###############
#    data frame    #
###############
c<-b1<-e<-numeric(0)
b1<-e<-NULL
for (i in 1756:2007){
c<-rep(i,12)
b1<-c(b1,c)
}
b2<-rep(1:12,252)
for (i in 1:252){
        for (j in 1:12){e<-c(e,month[i,j])}
}
b<-data.frame(cbind(b1,b2,e))
colnames(b)<-c("Year","Month","Temperature")
b      #b is data frame: year, month, temperature

############
# Regression #
############

# model1 1756-2007#
y<-b
attach(y)
temperature<-b$Temperature
tp<-temperature
index<-1:length(tp)                  # vector 1,2,...,3024, last 252 years
time<-index/(length(tp)/252)    # a complete annual cycle is of length 1
model1<-lm(tp~sin(time * 2 * pi) + cos(time * 2 * pi))
summary(model1)

## then mean temperature every 10 years##
tpf<-factor(1+(index>120)+(index>240)+(index>360)+(index>480)+(index>600)
+(index>720)+(index>840)+(index>960)+(index>1080)+(index>1200)+(index>1320)
+(index>1440)+(index>1560)+(index>1680)+(index>1800)+(index>1920)+(index>2
```

```
040)
+(index>2160)+(index>2280)+(index>2400)+(index>2520)+(index>2640)+(index>2
760)+(index>2880))
tpf
tapply(tp,tpf,mean)

##model2 1756-1925##
y1<-b[1:2040,]
attach(y1)
temperature<-b$Temperature
tp1<-temperature[1:2040]
index1<-1:length(tp1)
time1<-index1/(length(tp1)/170)
model2<-lm(tp1~sin(time1 * 2 * pi) + cos(time1* 2 * pi))
summary(model2)

## model3 1926-1985##
y2<-b[2041:2760,]
attach(y2)
temperature<-b$Temperature
tp2<-temperature[2041:2760]
index2<-1:length(tp2)
time2<-index2/(length(tp2)/60)
model3<-lm(tp2~sin(time2 * 2 * pi) + cos(time2 * 2 * pi))
summary(model3)


##model4 1986-2007##
y3<-b[2761:3024,]
attach(y3)
temperature<-b$Temperature
tp3<-temperature[2761:3024]
index3<-1:length(tp3)
time3<-index3/(length(tp3)/22)
model4<-lm(tp3~sin(time3 * 2 * pi) + cos(time3 * 2 * pi))
summary(model4)



################
# ARIMA models #
################
d1<-b$Temperature
d2<-diff(d1,12) ###differenced the whole series with lag 12##
```

```
####################
# plot the whole series #
####################
temperature<-ts(d1,start=c(1756,1),end=c(2007,12),frequency=12)
plot.ts(temperature,main="monthly
temperature(1756-2007)",col="red",ylab="Temperature (°C)",xlab="Year")

################################################
#plot the subset series the year 1756-1766 and 1997-2007#
################################################
temperature1<-ts(d1[1:120],frequency=12)
temperature2<-ts(d1[2905:3024],frequency=12)
ts.plot(temperature1,temperature2,ylim=c(-12,20),lty=c(1:2),main="Monthly
Temperature",ylab="Temperature (°C)",xlab="Year")
abline(v=1:10,lty="dotted",col="green")
legend(7.75,-6,c('1756-1765,'1998-2007'),lty=c(1,2),text.col='red4',box.col='green4')

###############
#ACF and PACF#
###############
par(mfrow=c(2,2))
acf(d1,60,col=3,main="ACF for the Series")
pacf(d1,60,col=4,ylim=c(-0.5,0.5),main="PACF for the Series")
acf(d2,60,col=3,ylim=c(-0.4,0.4),main="ACF for the Differenced Series")
pacf(d2,60,col=4,ylim=c(-0.4,0.4),main="PACF for the Differenced Series")


#############
# Estimation #
#############
library(tseries)
adf.test(d2)        ##ADF test for the differenced series, null hypothesis is the series has
                        a unit root
m<- arima (d1, order = c (1,0,0),seasonal = list(order = c(0,1,1),period=12) );
# ARIMA model for the whole series#
m
m1<- arima (d1[1:2040], order = c (1,0,0),seasonal = list(order =
c(0,1,1),period=12) );
m1
m2<- arima (d1[2041:2760], order = c (1,0,0),seasonal = list(order =
c(0,1,1),period=12) );
m2
m3<- arima (d1[2761:3024], order = c (1,0,0),seasonal = list(order =
```

```
c(0,1,1),period=12) );
m3

## residuals checking ##
library(stats)
r <- resid(m)
Whitenoise.test(r) ####null hypthsis :The residuals are white noise
tsdiag(m, gof.lag=12)
qqnorm(r)
qqline(r)

########################################
#forecasting the monthly temperature of 2008#
########################################
temp.pr=predict(m,n.ahead=12)
U=temp.pr$pred+2*temp.pr$se
L=temp.pr$pred-2*temp.pr$se
prediction <- matrix(c(temp.pr$pred,L,U), nrow = 12, ncol=3,
                     dimnames                                                    =
list(c("JAN","FEB","MAR","APR","MAY","JUN","JUL","AUG","SEP","OCT","NO
V","DEC"),
                                                  c("temperature",          "95%lowerCI",
"95%upperCI")))
month=c(1:24)
da<-as.numeric(d1[3013:3024])
plot(1:12,da,type="b",xlim=c(1,24),main="Forecasting          the          Monthly
Temperature",ylim=c(-15,21),ylab="Temperature (°C)",xlab="Month")
mypred<-as.numeric(temp.pr$pred)
lines(13:24,mypred,col="red",type="b",pch=19)
lines(13:24,as.numeric(U),col="blue",lty="dashed")
lines(13:24,as.numeric(L),col="blue",lty="dashed")
abline(v=12.5,lty="dotted")
lines(13:24,c(1.4,2.1,0.8,6.5,10.7,15.4,17.8,15.3,11.1,8.1,2.9,0.9),type='b')
lines(12:13,c(da[12],1.4),type='b')
legend(17,-7,c('True Val.','Pred. Val.','Conf. Int.'),
lty=c(1,1,2),pch=c(1,19,0),col=c(1,2,4),text.col='green4',box.col='green4')
text(11,20,'2007',col='green4');text(14,20,'2008',col='green4')
lines(c(10+.26,12-.26),rep(18.8,2),col='green4');lines(c(13+.26,15-.26),rep(18.8,2),col
='green4')
```