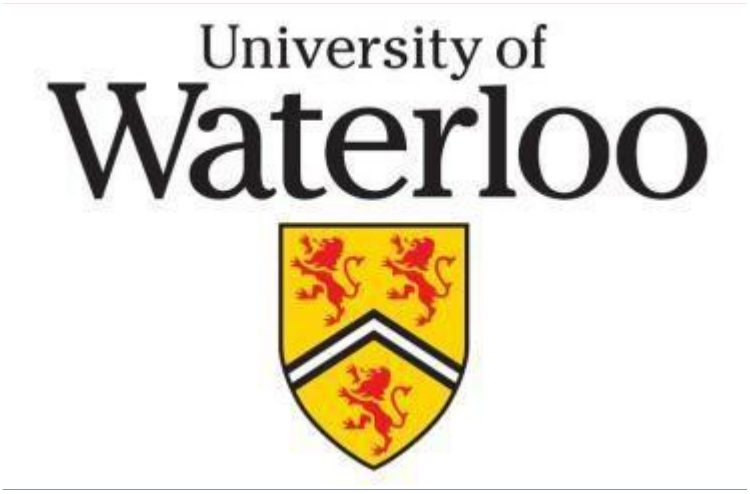


SYDE 631
PROJECT REPORT

ANNUAL SNOWFALL IN BUFFALO: 1910 - 1972

Name: Dibya Darshini Mohanty
20566881
Department of Systems Design Engineering

Abstract

Buffalo, USA has been experiencing heavy snowfall for a very long time. It has recorded one of the highest records of snow in USA. The snowfall is also governed by the Lake-Effect which sometimes makes the situation worse. I have analyzed the annual snowfall records for Buffalo using statistical tools and forecasting techniques and models have been used to forecast the annual snowfall values in the next few years. Data is for 63 years of snowfall where each data point represents the Annual snowfall for a single year.

ACKNOWLEDGEMENT

My sincere appreciation and gratitude goes to **Prof. Keith W Hipel**, my course professor, for his guidance, valuable suggestions, and constant encouragement throughout the course of this semester.

Special thanks go to **Prof. A. I. McLeod** and **Prof. Hipel** for developing one of the best software package which I have used for my project.

TABLE OF CONTENT

1. INTRODUCTION

Fig 1: Map of Buffalo, USA

Buffalo's first snowfall of winter usually arrives in November. And the last snowfall typically happens in April. Although sometimes early snow has been known to appear in October and late snow in May. It is normally free of snow every year from June to September.[1]

Most days of snowfall in Buffalo result in less than an inch left on the ground. Snowstorms of over five inches a day normally occur four times a year. But major blizzards that dump ten inches or more only happen on average once a year.

Thus, for about two-thirds of winter days, Buffalo has at least an inch of snow on the ground. Snow mostly accumulates during January. Typically, on five or six days in January and another one or two days in each of December, February and March, the snow covering Buffalo gets to ten or more inches deep.

As of this writing, Buffalo, has one of the highest measured snowfall total from the Lake-Effect. Now, what is the Lafe-Effect?[2]

When cold air moves over a warmer lake, the warm, moist air close to the surface of the lake rapidly rises, cools and condenses, forming clouds that churn out snow. Big lake-effect snow events are much more likely if and when cold air enters the equation earlier in the season, because the temperature of the lake, and the air above it, is warmer, increasing atmospheric instability, and in turn, the chance for heavy snow.

Cold air and a warm lake aren't enough to produce accumulating lake-effect snow, though. The shape and size of the lake matter, too. The larger the lake, the more time and space the encroaching cold air has to suck in moisture-rich warm air and eventually generate precipitation.

From Fig 1, one can start to understand why Lake Erie -- the lake responsible for generating the Buffalo-area's lake-effect snow -- is such an efficient lake-effect snow producer: Erie's near horizontal orientation acts like a funnel, giving cold air moving across the lake more time to suck up warmer, moist air, and in turn, produce more snow.

So, the single factor most responsible for all of these huge snowfall totals might just be the unseasonably cold air.

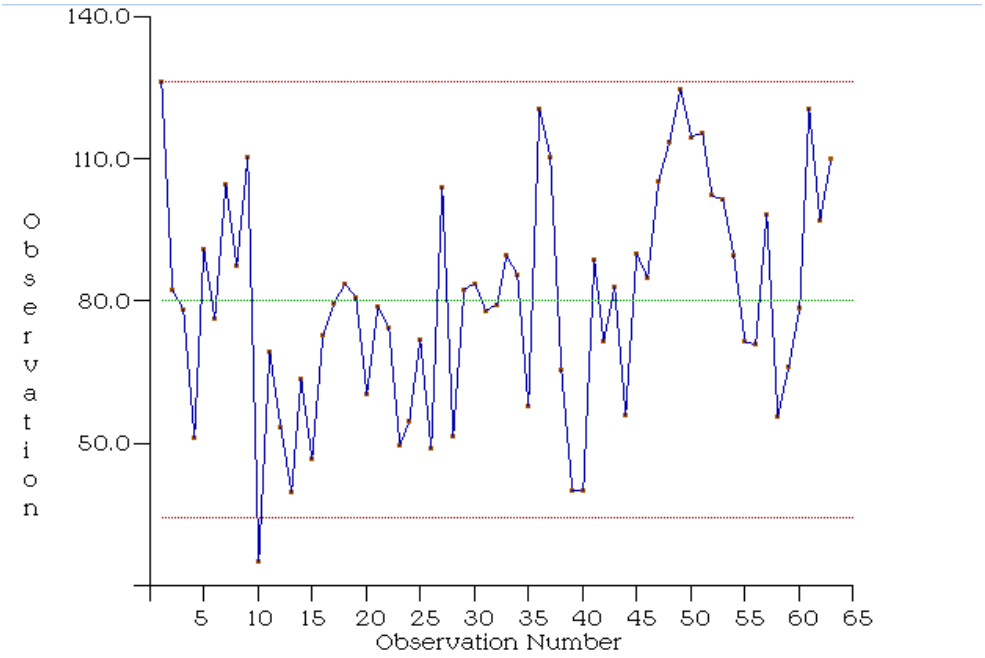
We can analyze the dataset acquired from Data Market which has the snowfall values from 1910 to 1972 in Buffalo, averaged for every year and recorded as the annual snowfall value. We can also accurately predict the annual snowfall in the next few years using forecasting.[9]

There are basically few steps involved in forecasting (McLeod, Hipel, 1977):

1. Exploratory analysis.
2. Trend analysis.
3. Determination of correlation functions
4. Estimating model parameters
5. Diagnostic Checks
6. Forecasting and Simulation using various methods.

2. Exploratory Analysis

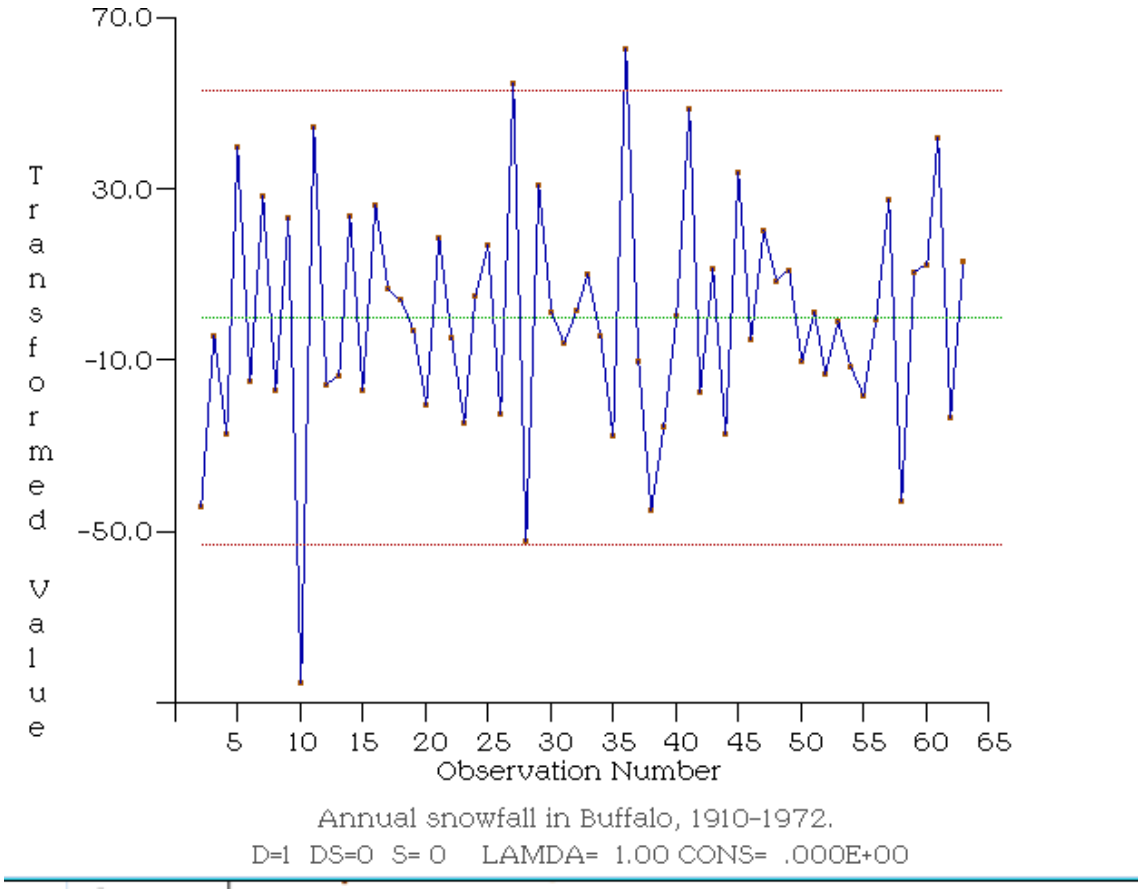
The trace plot for z_t; t = 1; 2; : : : is simply a plot of z_t vs. t. Although it is very simple, it is very useful in practice. Many interesting features of the series are often visible such as outliers, variance changes or shifts, trends (deterministic or stochastic), interventions, seasonal effects, non-Gaussian or non-linear features. Z_t is on y-axis and time on x-axis. Three graphs with differencing d=0, 1 and 2 have been plotted. All have different behaviour as shown below:



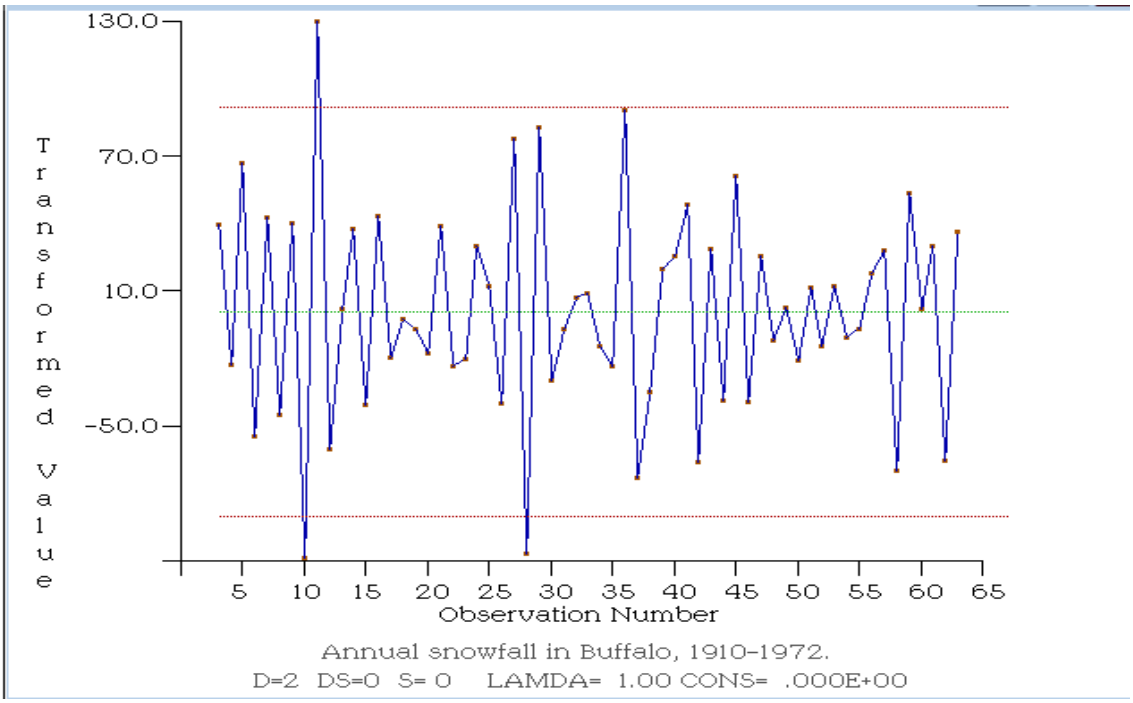
Annual snowfall in Buffalo, 1910-1972.

Above graph shows there is no seasonality with d=0.

Next, d=1 means present value is subtracted from the previous value one time, so total number of values decreased by 1, this process can be used to remove non-stationarity, graph below is stationary graph with almost constant mean and standard deviation:

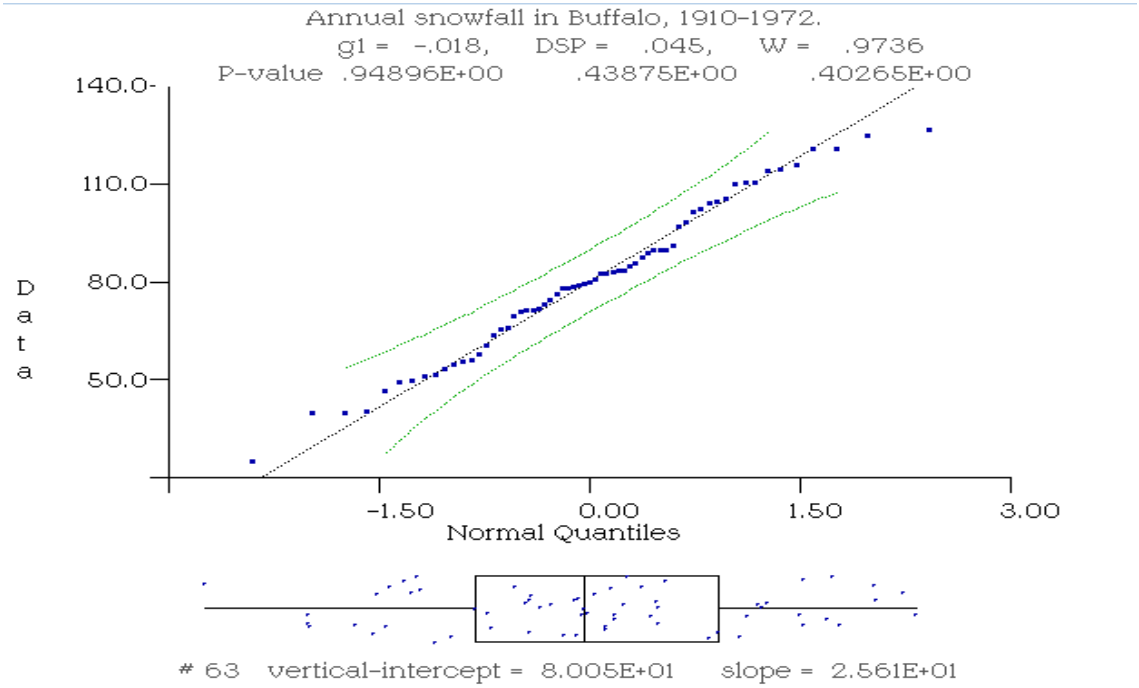


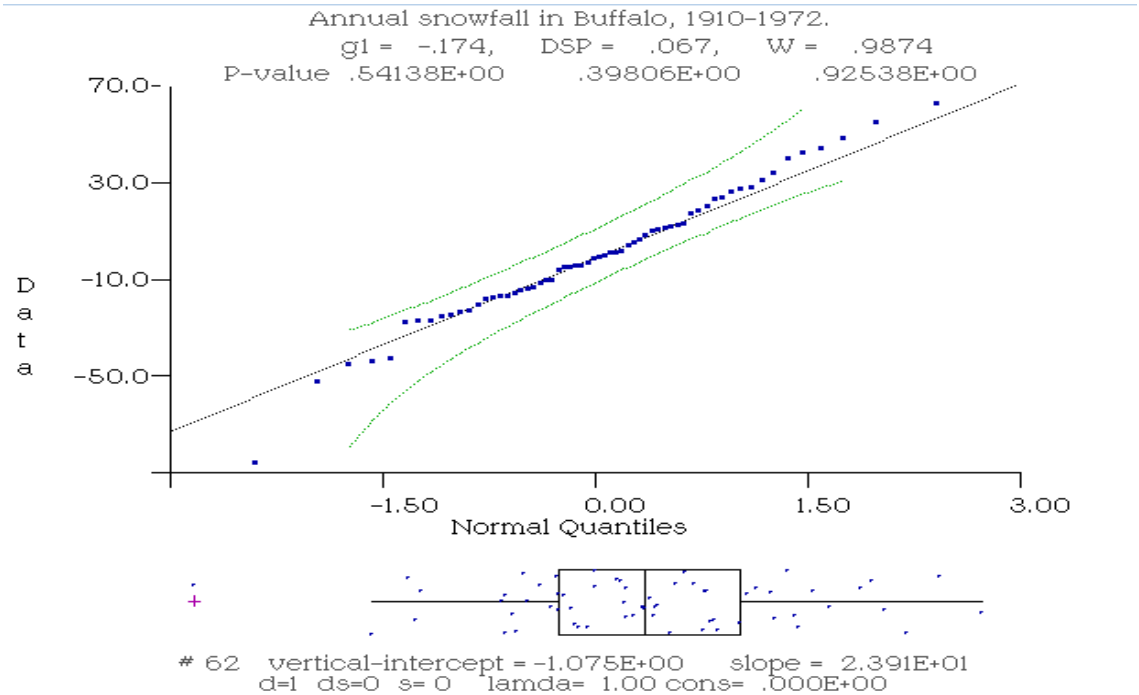
Next differencing = 2, means values subtracted twice from previous value.



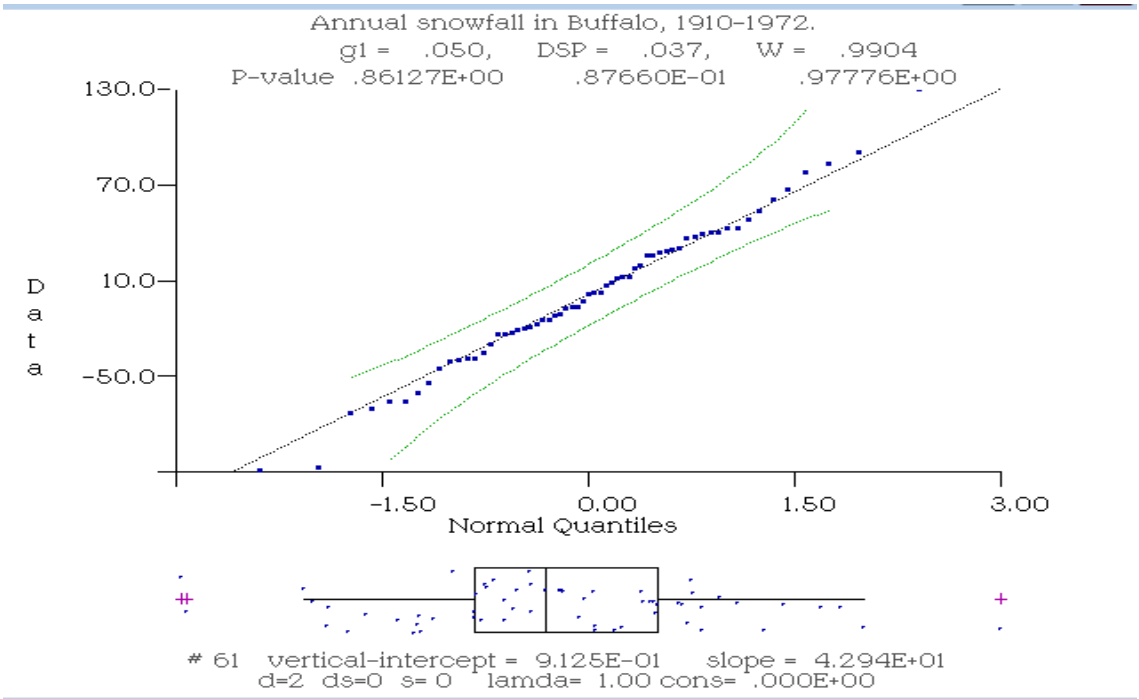
2.1 Normal Probability plot:

Graph below shows data points are normally distributed without any differencing.





After differencing =1, we see there is a better normal distribution along the tails.



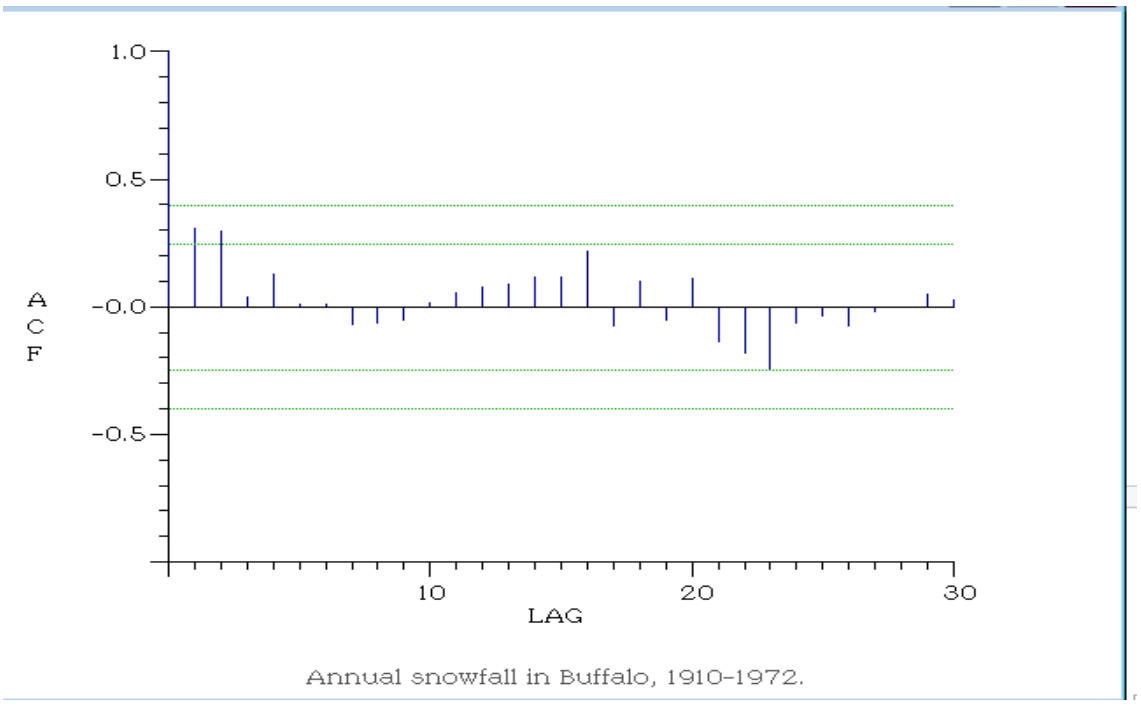
The above graph shows the normal distribution at differencing = 2. Thus, from the above graphs we can deduce that no kind of Box-Cox transformation is required.

3. Correlation Functions

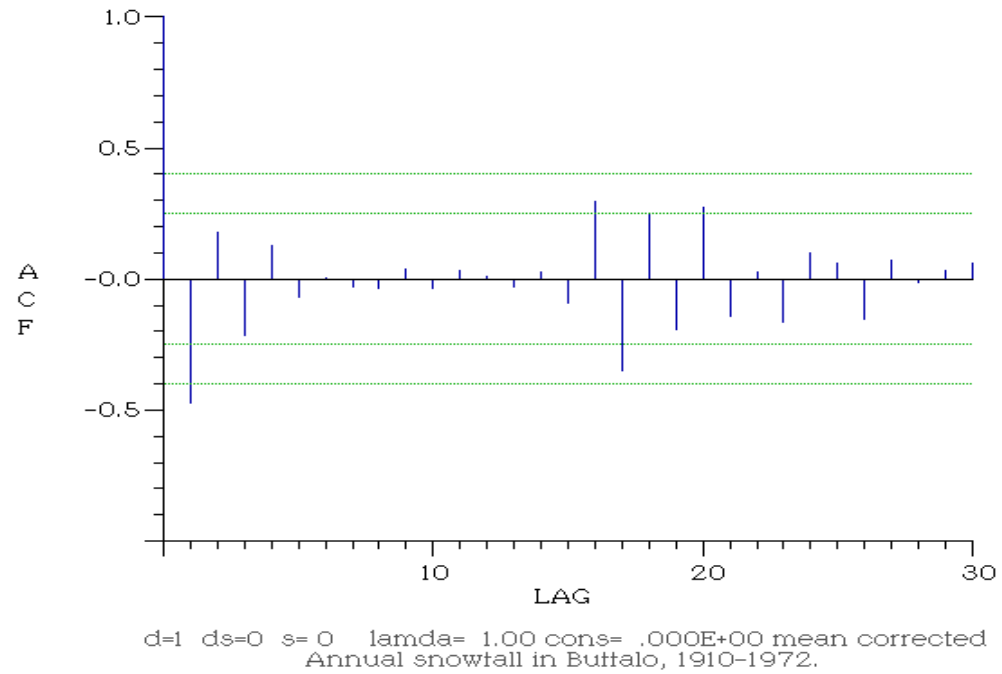
3.1 Auto Correlation Function

Autocorrelation function (ACF) at lag k is the correlation between z_t and z_{t-k}, when the intervening $z_{t-1}, \ldots, z_{t-k+1}$ are held constant. It is cross correlation of a signal with itself. It is the similarity between observations as a function of the time separation between them. It is mathematical tool for finding repeating patterns, such as the presence of a periodic signal which has been buried under noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies [3]. This also aids in moving – average identification. If the sample values are outside the limits, the model cannot be represented as an MA (q). Various graphs have been constructed with different differencing, shown below:

Graph 1: ACF with d=0. It cuts off at lag 2. So MA(2) can be taken.

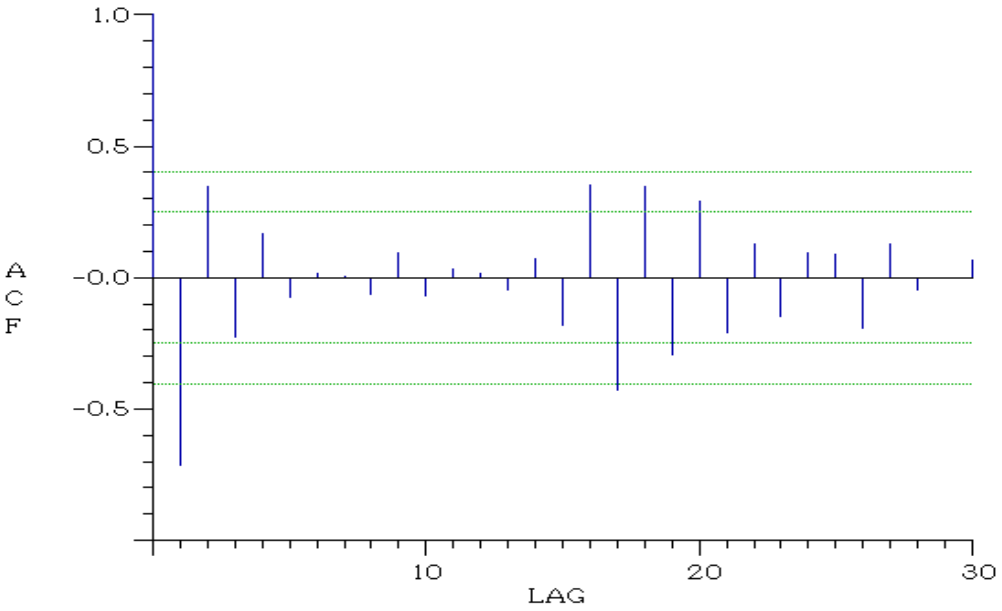


Graph 2: ACF with differencing d=1, here it cut off at lag 1, so MA(1) can be taken



d=1 ds=0 s= 0 lamda= 1.00 cons= .000E+00 mean corrected
Annual snowfall in Buffalo, 1910-1972.

Graph 3: ACF with differencing d=2, here it cuts off at lag 2, so again MA (2) can be taken.



d=2 ds=0 s= 0 lamda= 1.00 cons= .000E+00 mean corrected
Annual snowfall in Buffalo, 1910-1972.

3.2 Partial Correlation Function

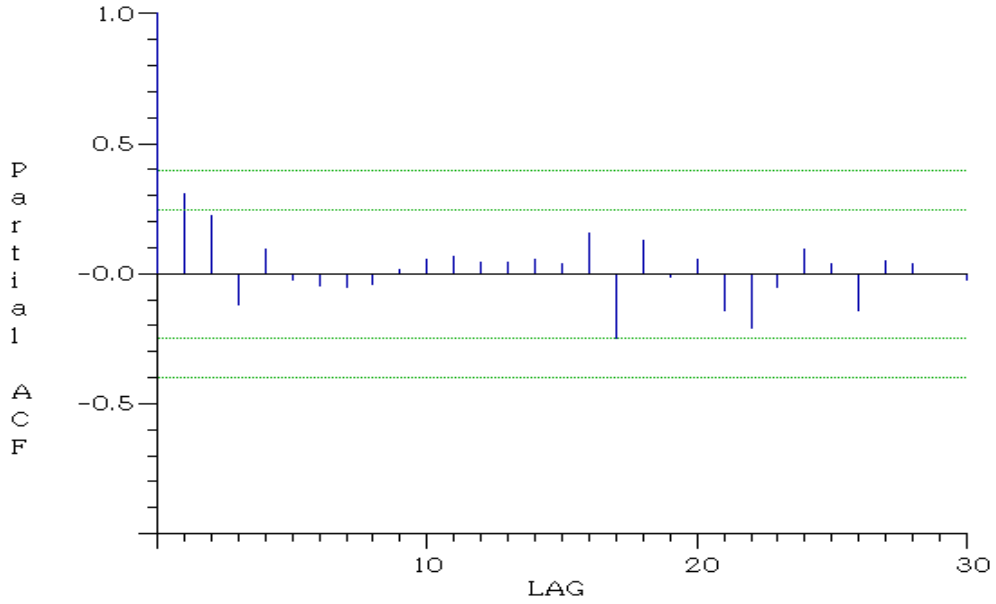
The partial correlation function (PACF) at lag k is the correlation between z_t and z_{t-k},

when the intervening $z_{t-1}, \ldots, z_{t-k+1}$ are held constant. The PACF at lag k is denoted by $\Phi_{k,k}$. It can be shown that $\Phi_{k,k}$ can be estimated by fitting an AR(k) [3],

$$Z_t = \mu + \Phi_{1,1}(z_{t-1} - \mu) + \ldots + \Phi_{k,k}(z_{t-k} - \mu) + a_t$$

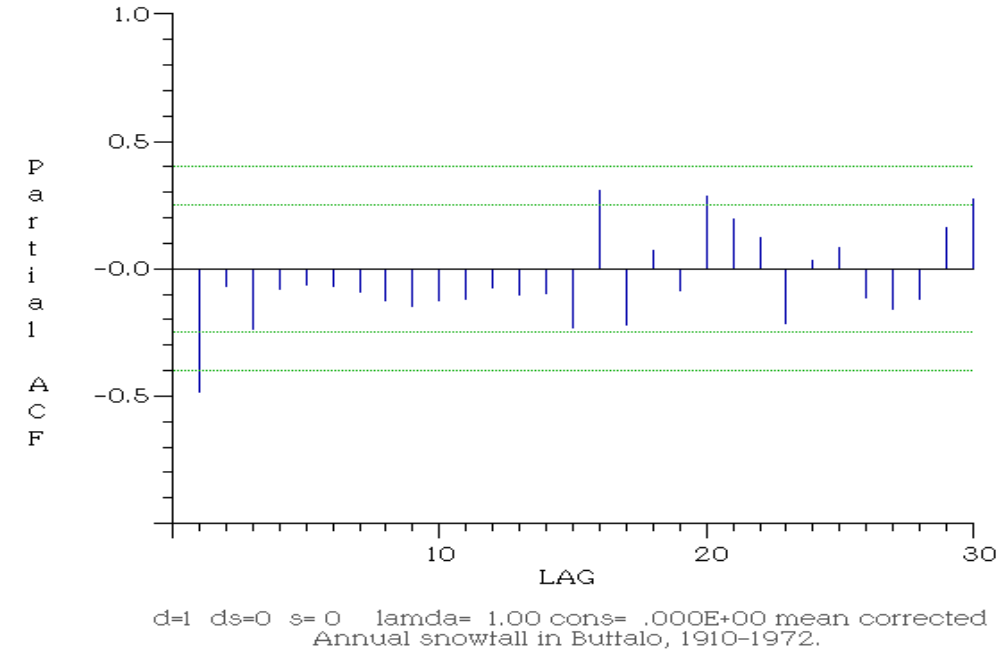
Thus by fitting successive AR(k) models for k=1,2, . . . the sequence of sample partial autocorrelations $\Phi_{k,k}$, k=1,2, . . .is generated. It is also used for identification of AR(p).

Graph 1: PACF with d=0, here it cuts off at lag 1, so AR (1) can be taken.

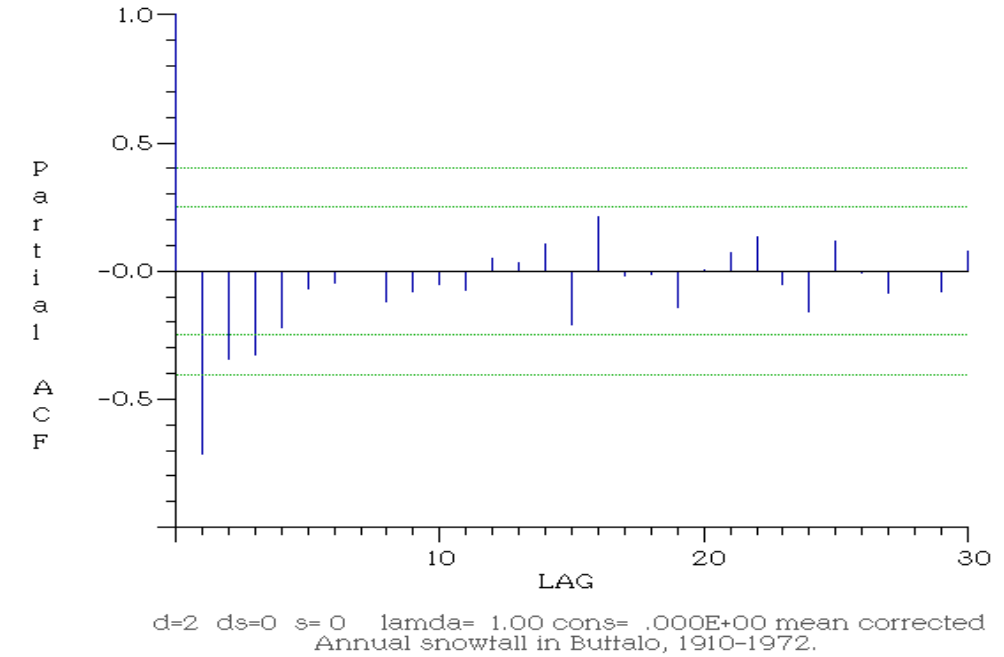


Annual snowfall in Buffalo, 1910-1972.

Graph 2: PACF with d=1, here it cuts off at lag 1 again, so AR (1) can be taken.



Graph 3: PACF with d=2, cuts off at lag 3, so AR (3) can be chosen.



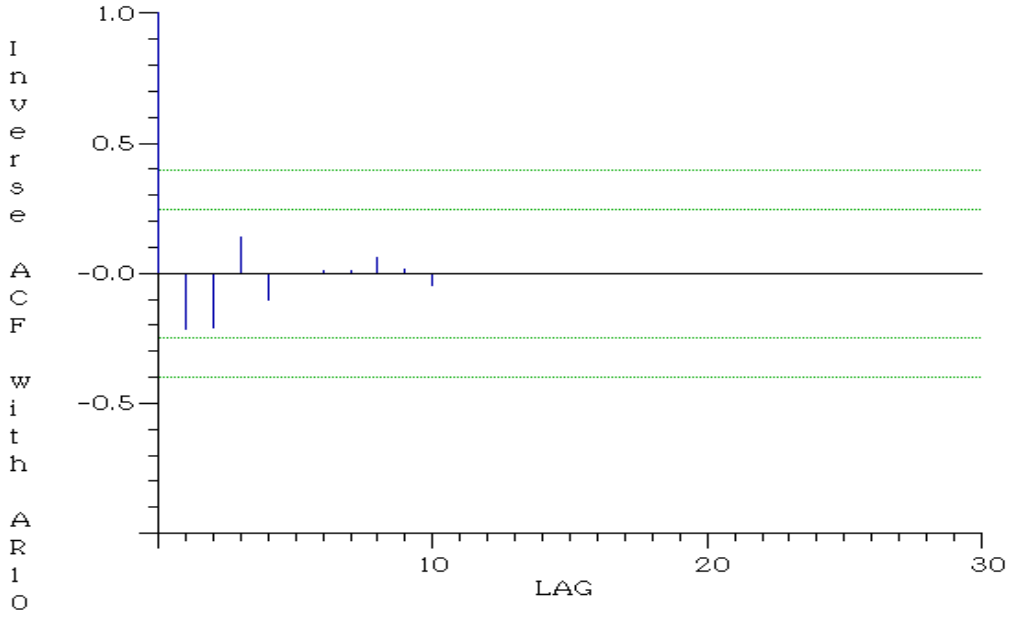
3.3 Inverse Autocorrelation Function

Inverse correlation function could be used to identify ARMA (p,1) model. Suppose we have model,

$$\Theta(B)\,(z_t - \mu) = \Phi(B)\,a_t$$

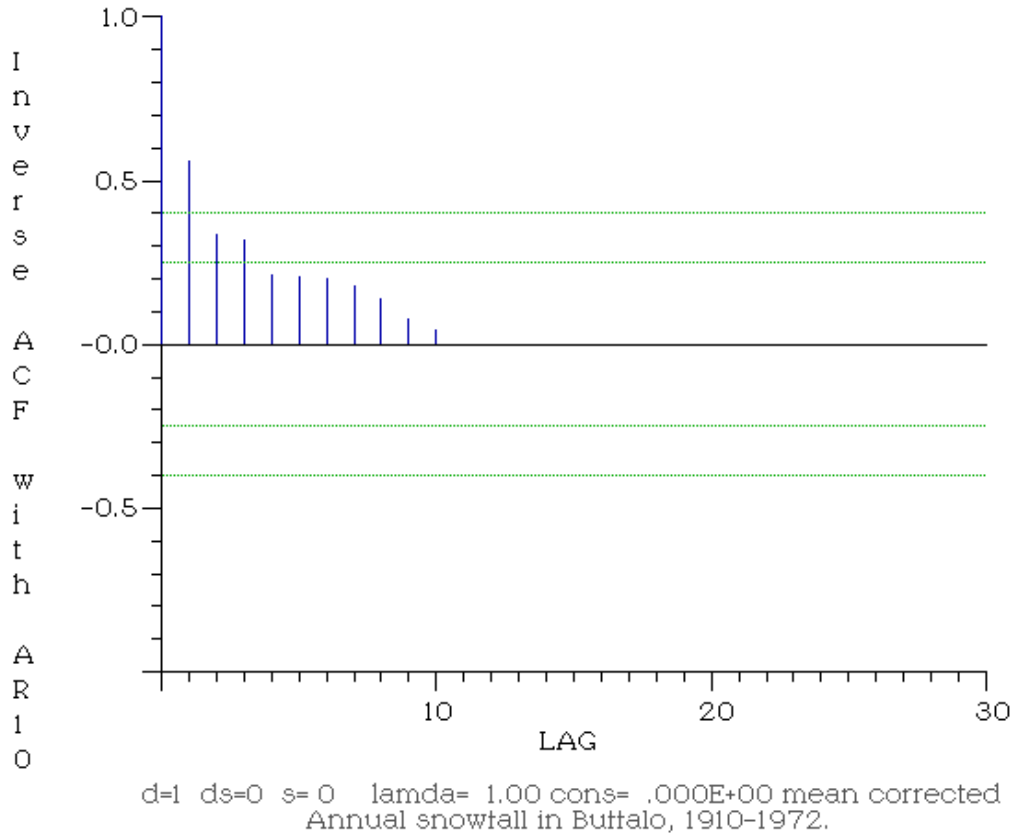
The first model can be referred to as the primal model and the second model is the corresponding dual. Notice, that the dual of dual gives back primal.

Graph 1: IACF with d=0, is not cutting off at any point.

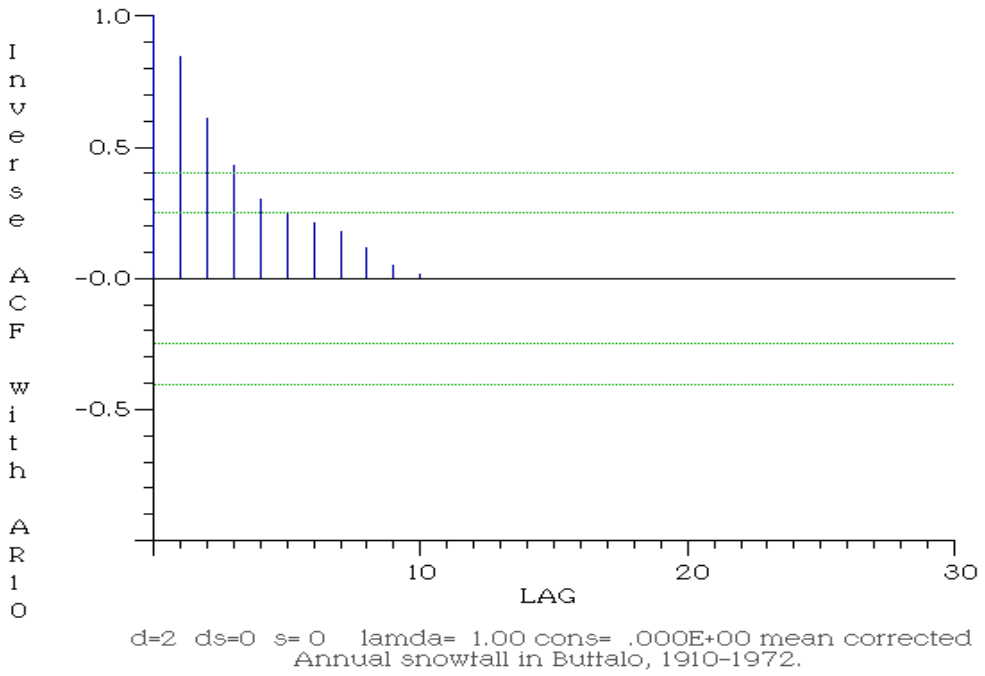


Annual snowfall in Buffalo, 1910-1972.

Graph 2: IACF with d=1, cuts off at lag 3.



Graph 3: IACF with d=2, cuts off at lag 4.

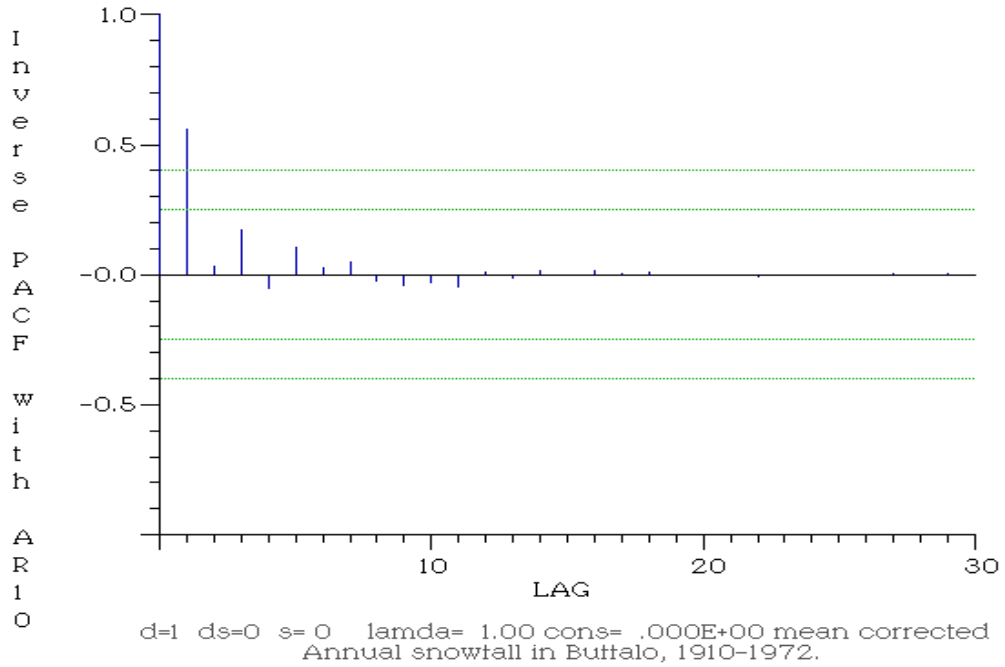


3.4 Partial Inverse Autocorrelation Function

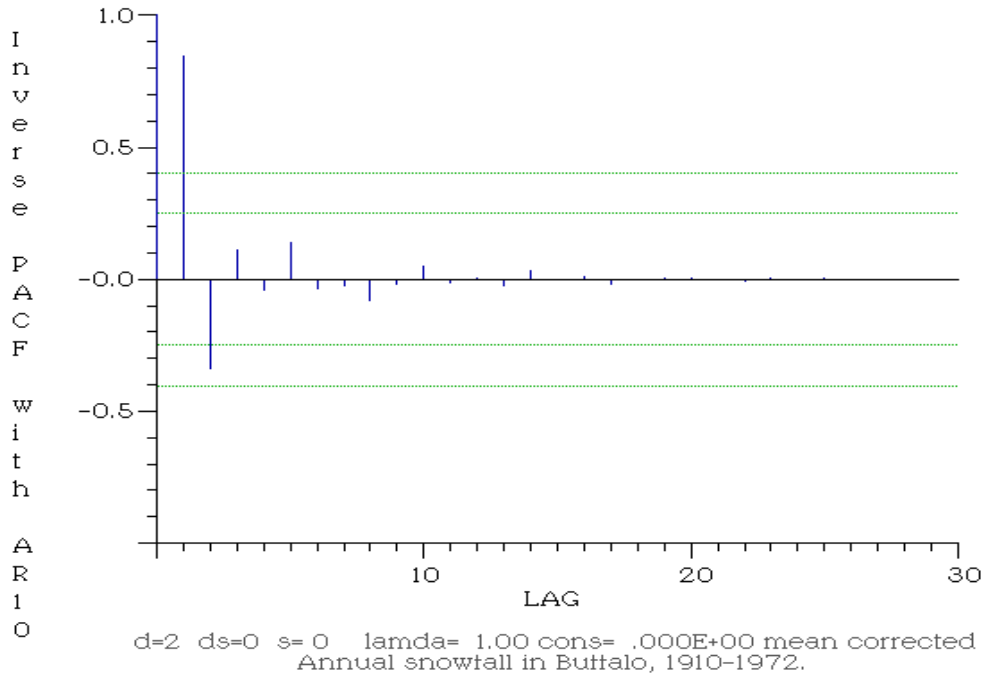
The IPACF is sometimes helpful in identifying MA(q) models. If the true model is MA(q), the IPACF should cut-off after lag q and this can readily be checked from the confidence limits shown on the plot. The ACF also cuts of after lag q, but to check the significance of higher-order lags it is necessary to look at the revised confidence limits using Bartlett's formula.

Graph 1: IPACF with d=0, cuts off at lag 2.

Graph 2: IPACF with d=1, cuts off at lag 1



Graph 3: IPACF with d=2, cuts off at lag 2.



3.5 Summarizing Table

	d	MA(q)	AR(p)
ACF	**0**	**2**	
	1	**1**	
	2	**2**	
PACF	**0**		**1**
	1		**1**
	2		**3**
IACF	**0**		**NA**
	1		**3**
	2		**4**
IPACF	**0**	**2**	
	1	**1**	
	2	**2**	

From the above table we have five models to fit. Now we will proceed with these models and choose the best fit. Models are ARIMA (1,0,2) , ARIMA(1,1,1) , ARIMA (3,1,1) and ARIMA (3,2,2), ARIMA (4,2,2)

4. Model Estimation

The tables are estimated using MHTS package [13]. Diagnostic checks for each model are being done and showed later in report. Model estimation is used to estimates parameters using McLeod approximate maximum likelihood algorithm. In models, we used ARIMA. Below are estimations of each model with different tests (Hillmer, Tiao, 1982).

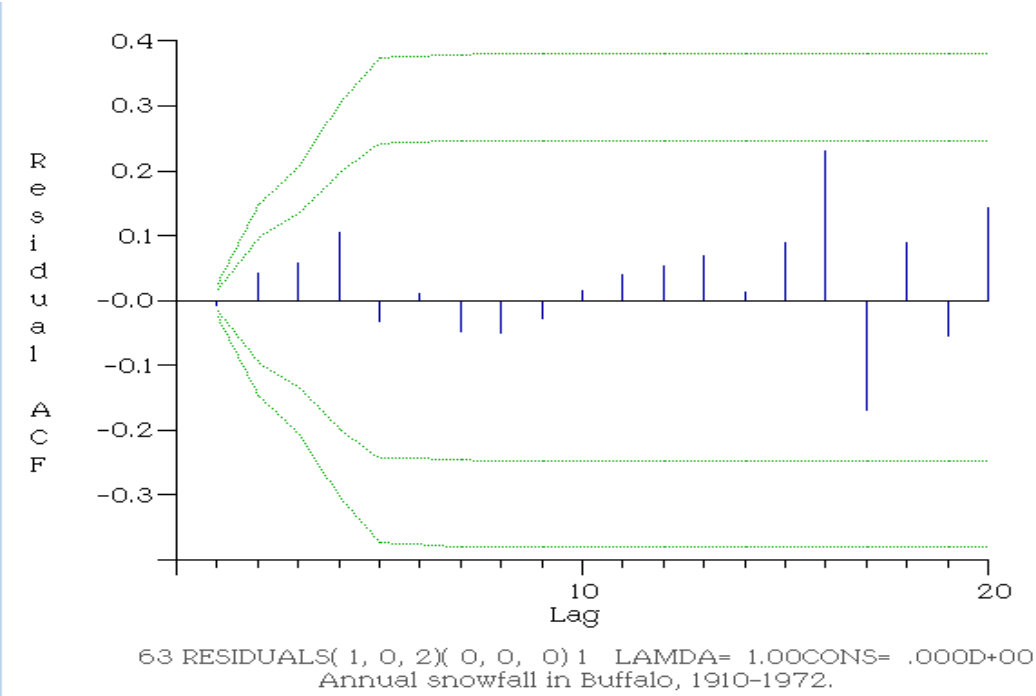
4.1 Estimation 1.

Taking ARIMA (1, 0, 2)

ARIMA (1,0,2)			
Series Mean	80.295	Standard Error	4.1578
Estimated Parameters	Φ_1 = -.1823		
	Θ_1 = -.4709		
	Θ_2 = -.3344		
Skewness	-.2694	Significance Level	.345541
Kurtosis	-.0002	Significance Level	.999759
AIC	395.4857		
BIC	404.0582		

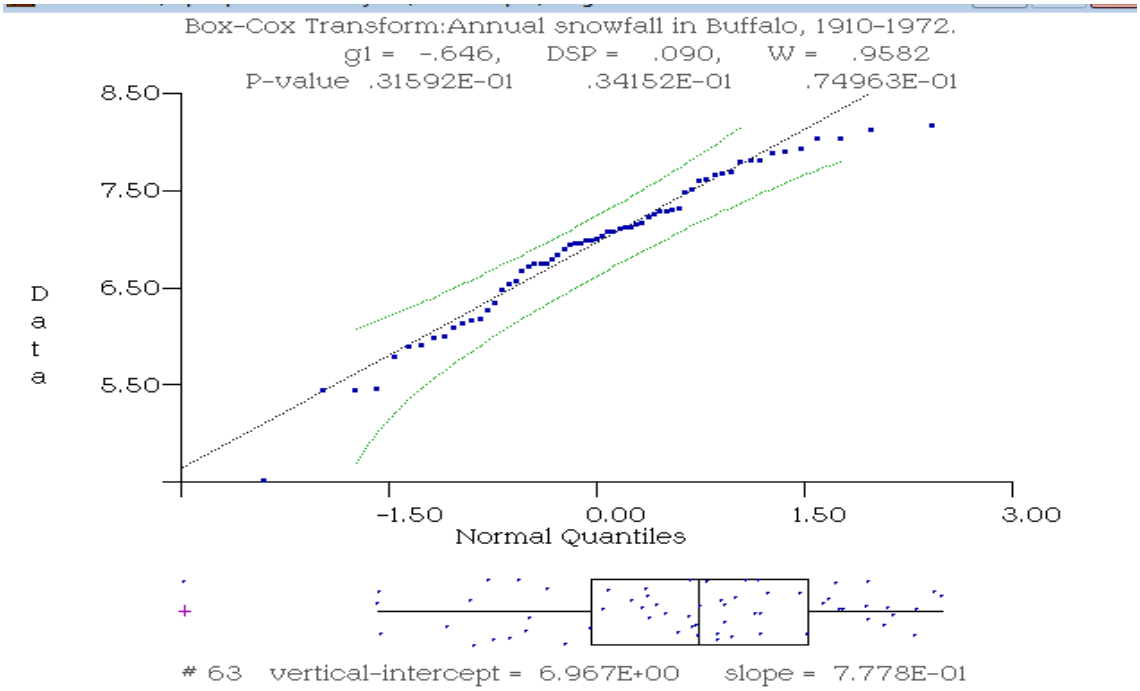
4.1.1 Diagnostic Checks

1. Plotting Residual Autocorrelation Function (RACF)



From the above graph out of 30 lags (>63/4=16), none of the values are above the confidence limit.

2. Normal Probability Plot



Normal probability curve (Jarque, Bera, 1987) is within limits but it is not a perfect straight line.

3. Homoscedastic Test

If the value of Standard Error is greater than Variance (SE) and Chi- statistic is greater than SE (chi-statistic) then it can be assumed that residual are Homoscedastic (Bai, Ng, 2005)

Variance over time:

$\chi = 0.027584$ SE = 0.020364 (χ >SE)

So residual are Heteroscedastic.

4.2 Estimation 2.

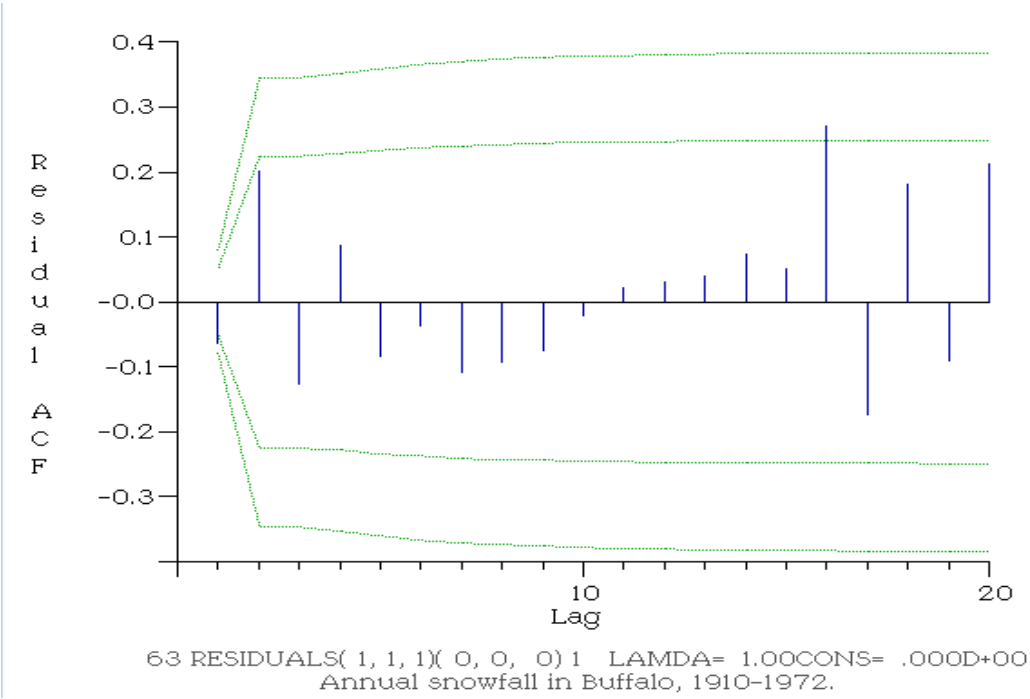
Taking ARIMA (1, 1, 1)

ARIMA (1,1,1)			
Series Mean	-2.645	Standard Error	5.0395

Estimated Parameters	Φ_1 = .2335		
	Θ_1 = .8670,		
Skewness	-.3363	Significance Level	.2422
Kurtosis	.2704	Significance Level	.6608
AIC	402.357		
BIC	410.9298		

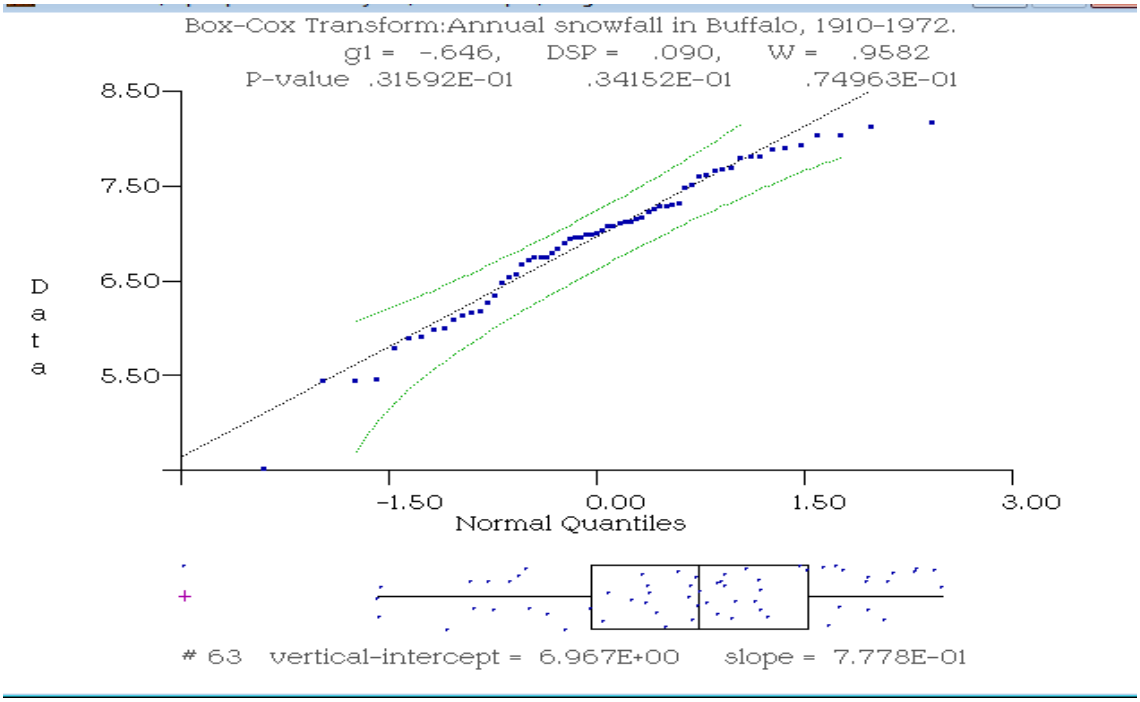
4.2.1 Diagnostic Checks

1. Plotting Residual Autocorrelation Function (RACF)



From the above graph, we can infer that most of its values are below confidence limit, with 20 lags. Only 1 lag crosses the confidence limit.

2. Normal Probability Plot



Normal probability curve (Jarque, Bera, 1987) is within limits but it is not a perfect straight line. Al though values around the tails cross the confidence limits.

3. Homoscedasticity Test

Variance over time:

$\chi = .002553$ SE = .006252 (SE > χ)

So residual are Homoscedastic.

4.3 Estimation 3.

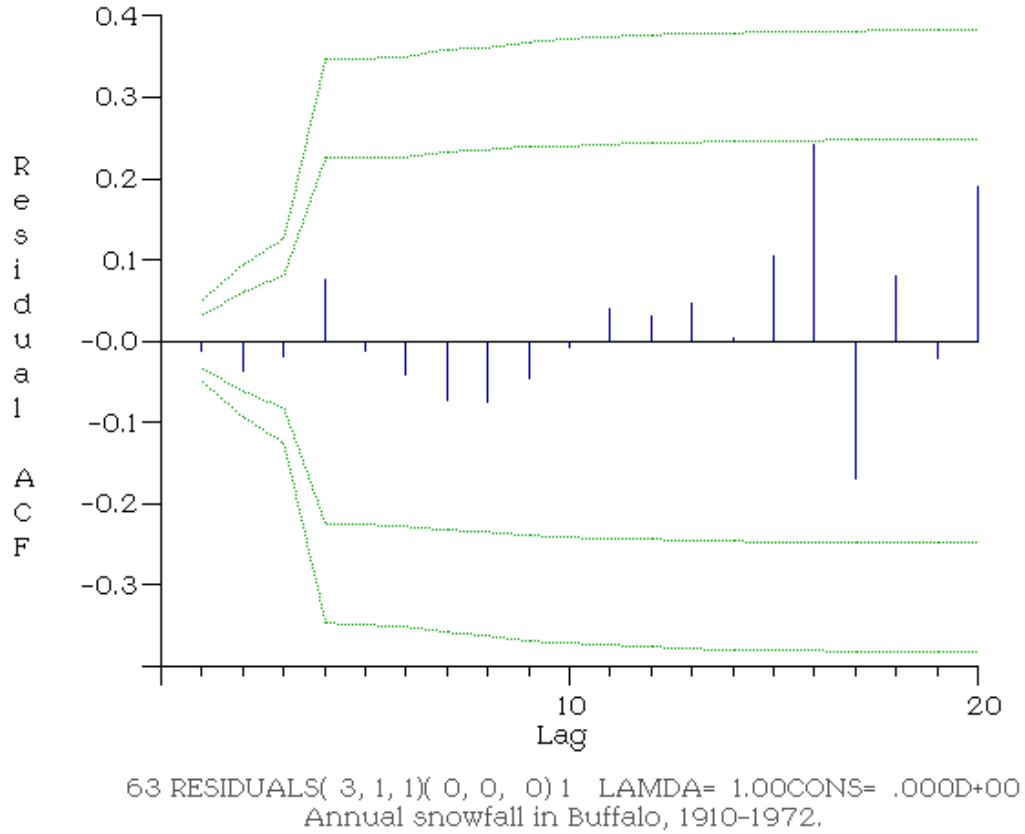
Taking ARIMA (3,1,1)

ARIMA (3,1,1)			
Series Mean	-2.64516	Standard Error	4.5397
Estimated Parameters	$\Phi_1 = .2401$		
	$\Phi_2 = .2360$		
	$\Phi_3 = -.1393$		
	$\Theta_1 = -.8931$		

Skewness	-.3601	Significance Level	.211715
Kurtosis	.3386	Significance Level	.582798
AIC	402.617365		
BIC	415.476174		

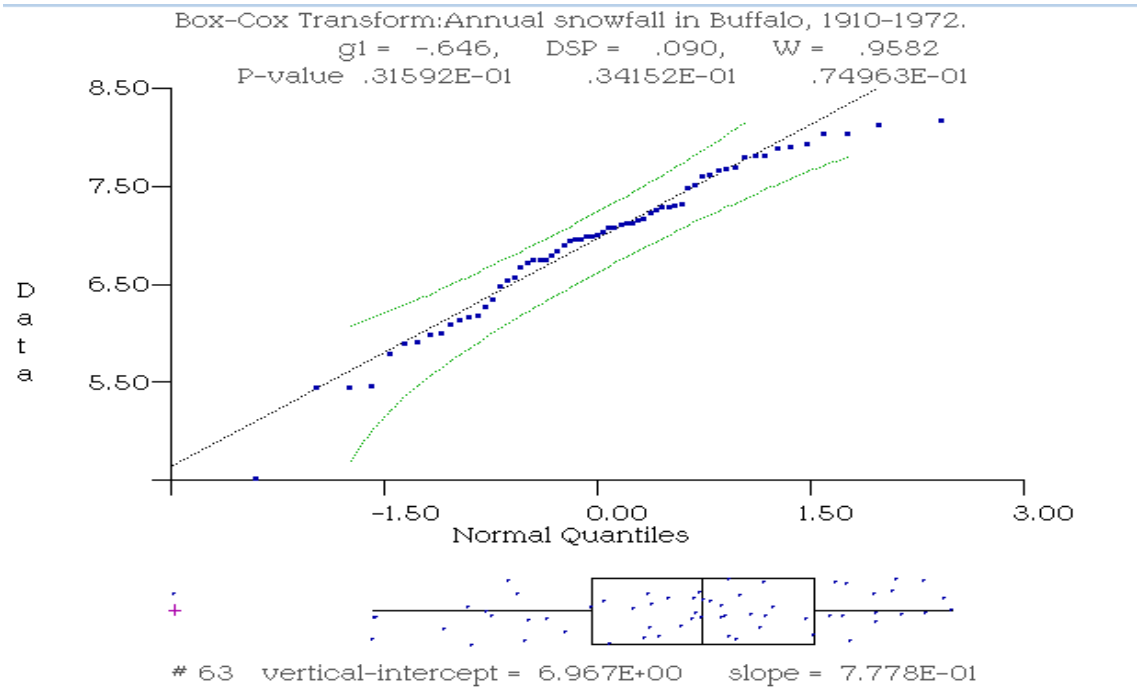
4.3.1 Diagnostic Checks

1. Plotting Residual Autocorrelation Function (RACF)



From the graph above we can say that all lags lie within the confidence limit.

2. Normal Probability Test



Normal probability curve (Jarque, Bera, 1987) is within limits but it is not a perfect straight line.

3. Homoscedasticity Test Variance over time:

$\chi = 0.002105$ SE = 0.006456 (SE > χ)

So residual are Homoscedastic.

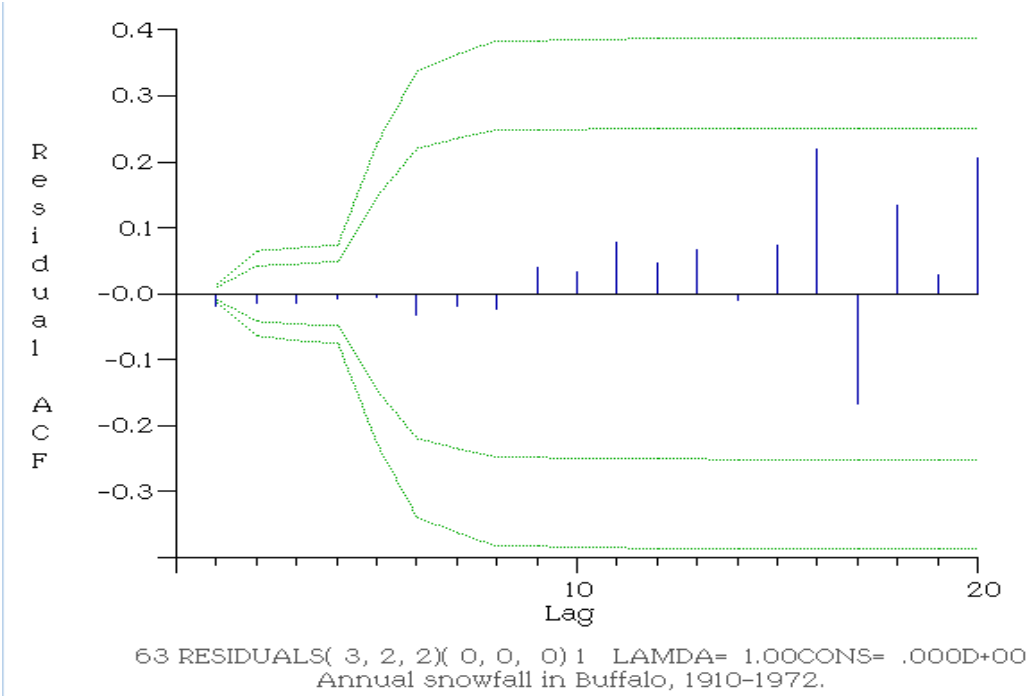
4.4 Estimation 4.

Taking ARIMA (3,2,2)

ARIMA (3,2,2)			
Series Mean	9.344262	Standard Error	4.157327
Estimated Parameters	Φ_1 = -.3828		
	Φ_2 = -.1171		
	Φ_3 = -.2106		
	Θ_1 = .9413		
	Θ_2= -.1577		
Skewness	-.2674	Significance Level	.348994
Kurtosis	.3727	Significance Level	.545586
AIC	425.21878		
BIC	442.36385		

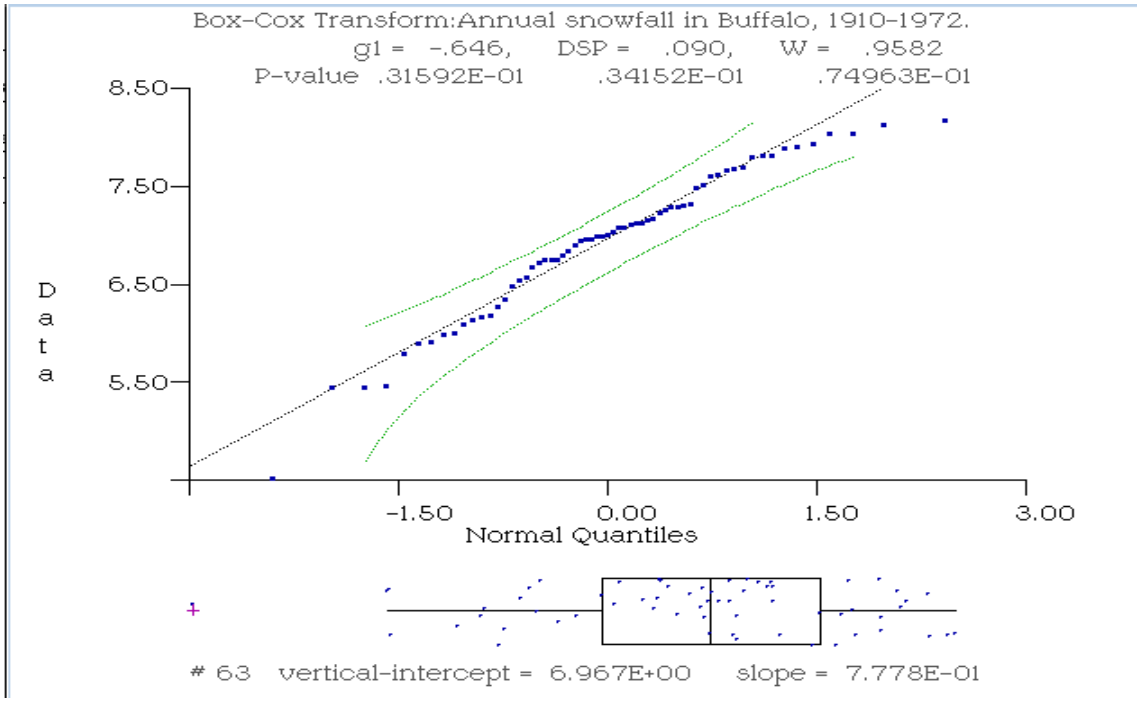
4.4.1 Diagnostic Checks

1. Plotting Residual Autocorrelation Function (RACF)



Above graph shows that there are no lags which lie above the confidence limit.

2. Normal Probability Test



Normal probability curve (Jarque, Bera, 1987) is within limits but it is not a perfect straight line.

4. Homoscedasticity Test

Variance over time:

$\chi = 0.008125$ SE = .005961 (SE < χ)

So residual are Heteroscedastic.

4.5 Estimation 5.

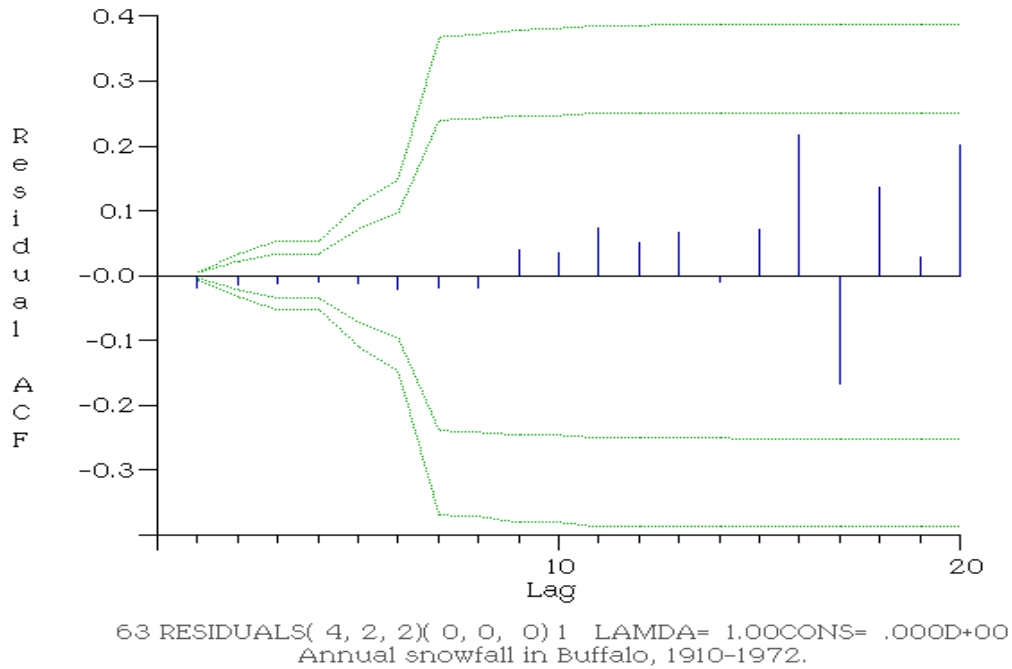
Taking ARIMA (4,2,2)

ARIMA (4,2,2)			
Series Mean	9.344262	Standard Error	4.14922
Estimated Parameters	Φ_1 = -.7681		

	Φ_2 = -.3365		
	Φ_3 = -.2916		
	Φ_4 = -.0859		
	Θ_1 = .5558		
	Θ_2 = .1308		
Skewness	-.2799	Significance Level	.327604
Kurtosis	.4088	Significance Level	.507481
AIC	427.209029		
BIC	446.497241		

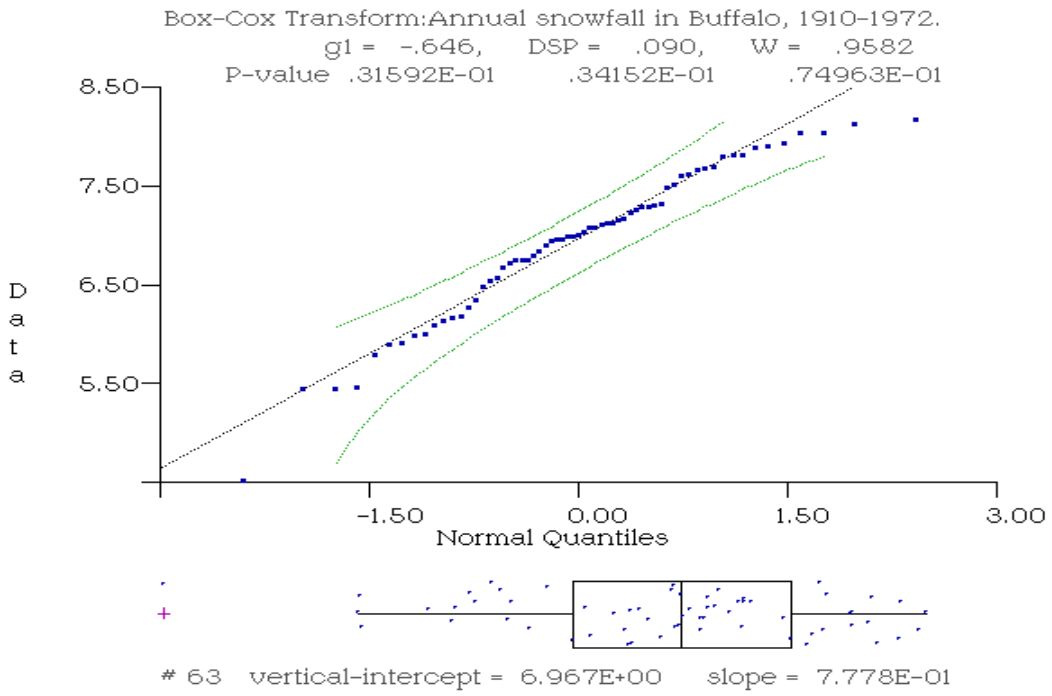
4.5.1 Diagnostic checks

1. Plotting Residual Autocorrelation Function (RACF)



Above plot shows no lags lie above the confidence limit.

2. Normal Probability Plot



Normal probability curve (Jarque, Bera, 1987) is within limits but it is not a perfect straight line. Al though values around the tails cross the confidence limits.

4. Homoscedasticity Test

Variance over time:

$\chi = .008216$ SE = .005955 (SE > χ)

So residual are Homoscedastic.

5. COMPARING AIC AND BIC

Estimations	AIC	BIC
ARIMA (1,0,2)	395.4857	404.0582
ARIMA (1,1,1)	402.357	410.9298
ARIMA (3,1,1)	402.617365	415.476174
ARIMA (3,2,2)	425.21878	442.36385
ARIMA (4,2,2)	427.209029	446.497241

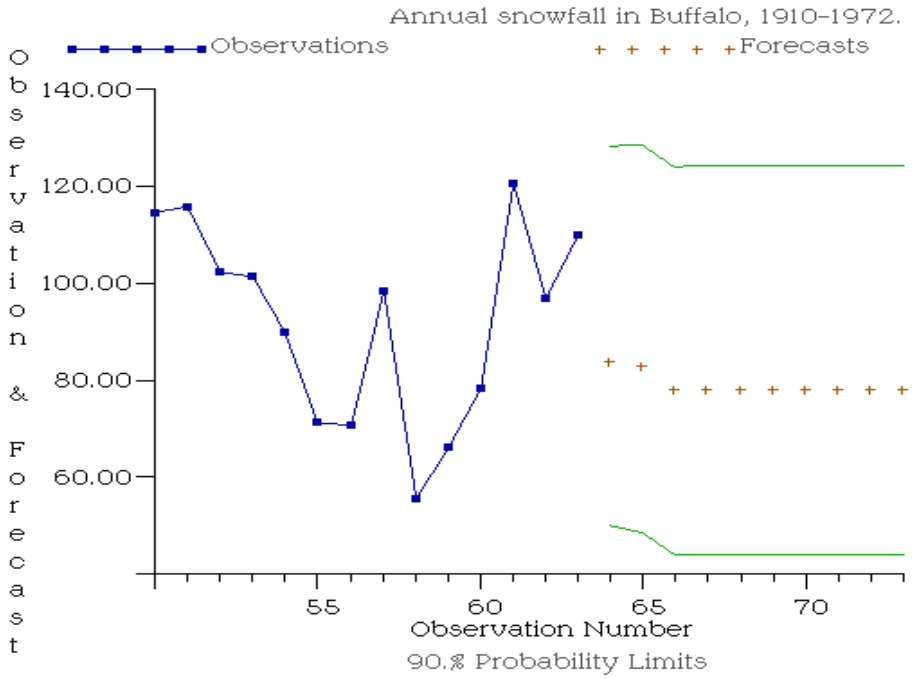
On the basis of above calculation, we found AIC and BIC for ARIMA (1,0,2)is minimum with respect to other values. If we look at the graphs of ARIMA (1,0,2) we noticed that it is normal distributed. Also RACF plot of ARIMA (1,0,2) can be taken into account in comparison to others. But the model does not pass the Homoscedasticity test since the χ > SE.

Thus we test different values of Box-Cox Transformation Parameter and find that when lamda=0.4, χ=.292808 and SE=.295005. Thus it is Homoscedastic. Moreover, AIC value = 402.0907, which is still the minimum for all models.

Thus, the best fit model for the time series is ARIMA (1,0,2) with Box-Cox Transformation parameter lamda=0.4.

6. FORECASTING

ARIMA (1,0,2)



Using MMSE we can find error ARIMA (1,0,2):

Original Values	Forecasted from model	Errors ARIMA (7,1,5)
4.518270D+00	12.54131D+00	-8.023041D+00
6.671231D+00	1.663924D+00	5.007307D+00
6.082577D+00	10.42947D+00	-4.346889D+00
5.445918D+00	1.694674D+00	3.751244D+00
6.472589D+00	9.399784D+00	-2.927195D+00
5.785329D+00	1.719253D+00	4.066076D+00

7. Conclusion

For the Annual Snowfall in Buffalo, ARIMA (1,0,2) with Box-Cox value of 0.4 is said to be the best model fit. The forecasted values when compared with the original observed values give an average MMSE of 2.174. Hence, the forecasted values quite well imitate the original values of the series.

8. REFERENCES

1) "Buffalo Lake-Snow Effect", http://www.weather.com/safety/winter/news/buffalo-newyork-lake-effect-snow

2) "Buffalo Snowfall Totals & Accumulation Averages", Current Results weather and science facts, http://www.currentresults.com/Weather/New-York/Places/buffalo-snowfall-totals-snow-accumulation-averages.php

3) Pfeifer, P.E. and Deutsch, S.J. (1981). Variance of the Sample Space- Time Autocorrelation Function. Royal Statistical Society. Series B (Methodological), Vol. 43, No. 1 (1981), pp.28-33.

4) Jarque, C.M. and Bera, A.K (1987). *A Test for Normality of Observations and Regression Residuals.* International Statistical Review / Revue Internationale de Statistique, Vol. 55, No. 2(Aug., 1987), pp. 163-172.

5) Hillmer, S.C. and Tiao, G.C. (1982). *An ARIMA-Model-Based Approach to Seasonal Adjustment*. Journal of the American Statistical Association, Vol. 77, No. 377 (Mar., 1982), pp. 63-70.

6) Kim, J.S., Shin, K.Y. and Ahn S.E. (2003). *A Multiple Replenishment Contract with ARIMA Demand Processes.* The Journal of the Operational Research Society, Vol. 54, No. 11 (Nov., 2003), pp.1189-1197.

7) Hawkins, D.M. (1981). *A New Test for Multivariate Normality and Homoscedasticity. Americal Statistical Association*. Technometrics, Vol. 23, No. 1 (Feb., 1981), pp. 105-110.

8) "Annual Snowfall in Buffalo, 1910-1972", DataMarket, https://datamarket.com/data/set/22yn/annual-snowfall-in-buffalo-1910-1972#!ds=22yn&display=line

9) Thomas A. R. Jr. and Brian L. J. (1976). *Normal probability plot and normality test.* Statistics Department, The Pennsylvania State University.

10) Wikipedia Contributor, *Partial Autocorrelation* Function Retrieved, November, 2012, http://en.wikipedia.org/wiki/Partial_autocorrelation_function

11) Wikipedia contributor, *Autocorrelation Function*, Retrieved, November 2012, http://en.wikipedia.org/wiki/Partial_autocorrelation_function

12) McLeod, A.I. & Hipel, K.W. (2002) Time Series Package MHTS Reference Manual.

13) Yang, Y. (2005). Can the strength of AIC and BIC be shared? A conflict between Model Identification and Regression Estimation. Biometrica, vol.92, No. 4 (Dec, 2005), pp. 937-950