UNIVERSITY OF CALGARY

Chemical and Petroleum Engineering department

A statistical analysis on global warming: is there evidence it exists?

Course: Error Analysis and Experimental Design (Winter 2020)

Name: Lucas Santos Queiroz

Prof: Dr. Josephine Hill

Executive Summary

As the spread of fake news and theories are becoming quite popular, certain well-proven facts have been labeled as lies, such as global warming. The objective of this study is to answer whether the global annual temperature is increasing or it is due to chance. The dataset is a compilation of several global temperature reports from 1900 to 2015 consolidated by a third-party organization. Considering a level of confidence equals 0.05, at first, a simple linear regression is performed. The linear regression showed an increasing tendency, however, after residual analysis, it presented aucorroleation on the residuals, which violates the regression assumptions. Then, time series analysis techniques are required. Based on data 2D-plot and the Augmented Dickey-Fuller test results, there is evidence of the increasing tendency of global annual temperature. Therefore, considering the association between global warming and temperature, there is enough evidence that global warming exists.

Additionally, ARIMA, a time series analysis model, is used to model the data. The best predictor model selected is ARIMA(0,1,2) as it presented the least sum of squares, no correlation within the residuals and all coefficients statistically significant (highest p-value = 0.008 < 0.05). Further steps on the temperature forecast and ways to enrich the statistical analysis are suggested.

Table of contents

List of figures

1) Introduction

Considering that fake news or internet information reliability is one of the major challenges of the current century[1], well-established concepts and discoveries have been facing discussions that try to invalidate them, e.g. if the Earth is flat or not, if global warming exists, among others. Thus, projects that aim to analyze and clarify basic topics are somehow needed nowadays.

All the life on Earth is similar to an experiment performed in the lab - they require special conditions (temperature, pressure, specific surroundings, among others) to happen and be stable. Consequently, specifically to temperature, even small changes in the global profile can have an enormous impact in nature, e.g. the reproduction process of green sea turtles[2]. Therefore, a clarification on the global warming aspect and statistical analysis on global temperature changes are essential. The present report chooses to understand the global temperature trending throughout the years (1900 – 2015).

Moreover, the project aims to develop a predictor model for the present data. The goal is to build critical thinking about the temperature tendency. As next steps, the model may be used on proactive planning to reduce potential casualties on the environment due to temperature changes.

2) Experimental Methods

First, a linear regression is performed on the dataset. The goal is to evaluate the regression coefficients and check if the global temperature shows an increasing tendency, statistically. The linear regression follows three assumptions that must be attained by performing a residual analysis and they are: normality of the error, homoscedasticity, and independence of the errors [5]. If all assumptions are met and the statistical test for the slope rejects the null hypothesis, then we obtain a model to predict the data.

Time series data is well known by its autocorrelation along the time axis (x-axis), which violates certain assumptions of previous analysis. Then, alternative methods to study this scenario are presented.

In this report, as the relative difference between the values is more important than the value itself, a multiplicative regression model of time series was performed – ARIMA model on the natural log of annual temperature. ARIMA is the autoregressive integrated moving average model that is based upon the autocorrelation of the values and their random error. Basically, it considers the same assumptions as of the previous linear regression plus the stationarity of the data, which is going to be explained later [4].

As ARIMA takes the advantage of autocorrelation, two graphical methods are used to assist finding the best set of parameters: ACF (Auto Correlation Function) and PACF (Partial Auto Correlation Function) [4]. Analogous to the linear regression that calculates the correlation coefficient between y and x, these functions evaluate the correlation between the data and the data itself but shifted k periods – periods are called lags. Different values for k are used to check the level of autocorrelation within the data. It is worth mentioning that the difference between ACF and PACF is that the first one considers the impact of all lags whereas the second one focuses only on the two lags that are being analyzed.

$$autocorrelation\ (k) = \frac{\sum_{i=1}^{N-k}(Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{\sum_{i=1}^{N}(Y_i - \bar{Y})}$$

Finally, stationarity of the data means that the average and variance do not change over time, and it is commonly tested by plotting the data on a 2-D plot [4]. Additionally, the Augmented Dickey-Fuller (ADF) test is performed to reassure the conclusions about stationarity. This test assumes the null hypothesis as there is a unit root and the alternate hypothesis as the data is stationary. If there is enough evidence that data is not stationary, a transformation of the data is needed prior to the application of the regression model, e.g. differentiate the dataset.

3) Data

The global temperature dataset was obtained from the website Kaggle as referenced in the appendix. However, the original data were collected by the Berkeley Earth – affiliation with Lawerence Berkeley National Laboratory. The dataset is a historical compilation from different temperatures reports – the main report datasets are NOAA's MLOST, NASA's GISTEMP and the UK's HadCrut [6]. Here, it is considered the data from 1900-2015 due to the inexistence of missing data – it is assumed that it would reduce "bad data" errors. Moreover, it is considered that "data management" errors may appear, but it will not impact the results of the present work [5].

The global temperature presented in the Appendix is an annual temperature in degrees Celsius calculated by the average of different temperatures measured on land and sea in several locations around the globe. However, only the absolute values are provided for each year. Unfortunately, it may bring errors to the statistical analysis as the confidence interval of each annual temperature cannot be estimated and the assumption that this average annual temperature represents the globe, which may not be statistically right.

Historically, different types of temperature measurement equipment were used as well. However, the source of data does not provide ways to bring the equipment error to the statistical analyzes, therefore it is going to be treated as random error.

Table 1: Annual Global Temperature subset of the original dataset from 1900 to 1962

year	Annual Temp	year	Annual Temp	year	Annual Temp
1900	15.144	**1921**	15.102	**1942**	15.325
1901	15.073	**1922**	15.012	**1943**	15.345
1902	14.958	**1923**	15.028	**1944**	15.449
1903	14.837	**1924**	15.039	**1945**	15.331
1904	14.810	**1925**	15.090	**1946**	15.263
1905	14.955	**1926**	15.241	**1947**	15.310
1906	15.032	**1927**	15.122	**1948**	15.261
1907	14.875	**1928**	15.132	**1949**	15.223
1908	14.838	**1929**	14.967	**1950**	15.140
1909	14.790	**1930**	15.173	**1951**	15.311
1910	14.819	**1931**	15.228	**1952**	15.363
1911	14.776	**1932**	15.188	**1953**	15.421
1912	14.877	**1933**	15.013	**1954**	15.231
1913	14.909	**1934**	15.145	**1955**	15.174
1914	15.076	**1935**	15.119	**1956**	15.106
1915	15.144	**1936**	15.166	**1957**	15.374
1916	14.901	**1937**	15.301	**1958**	15.382
1917	14.810	**1938**	15.293	**1959**	15.341
1918	14.960	**1939**	15.302	**1960**	15.292
1919	15.007	**1940**	15.372	**1961**	15.380
1920	15.026	**1941**	15.385	**1962**	15.326

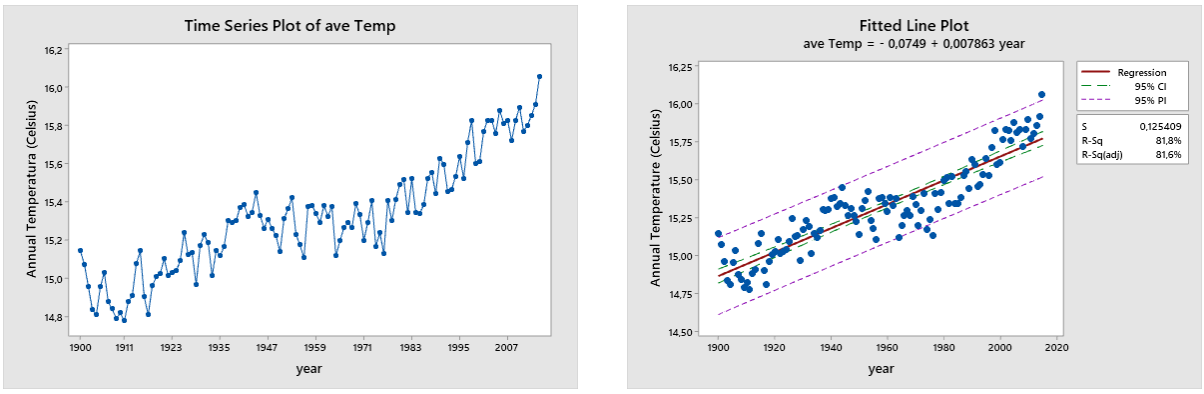
4) Results and Discussion

For all the results presented below, a level of confidence equals to 0.05 was fixed.

i. Linear Regression

Per linear regression, the final equation is presented as follows: $T(celsius) = 0.008\,t - 0.075$, where t in years. Graphically, the regression seems to be a great method to adjust the data (R^2 = 81.8%). However, residual analysis shows a different reality. Per Fig.2, even though the residual shows normality following the red line in the Normal Probability Plot and presenting a bell-shape in the Histogram, and homoscedasticity is not an issue (Versus Order plot), the residuals are dependent as Versus fits plot shows. The Versus fits plot presents a sinusoidal tendency of the residuals which violates the linear regression assumptions.

As discussed in Section 3, the reason for the violation may be the autocorrelation of the data itself. The annual global temperature is specified as a Time Series, which is defined as a series that intrinsically possesses autocorrelation through time. Then, alternative statistical models must be used.



(a) (b)

Figure 1: (a) Annual Temperature plot (b) Linear Regression on Annual Temperature

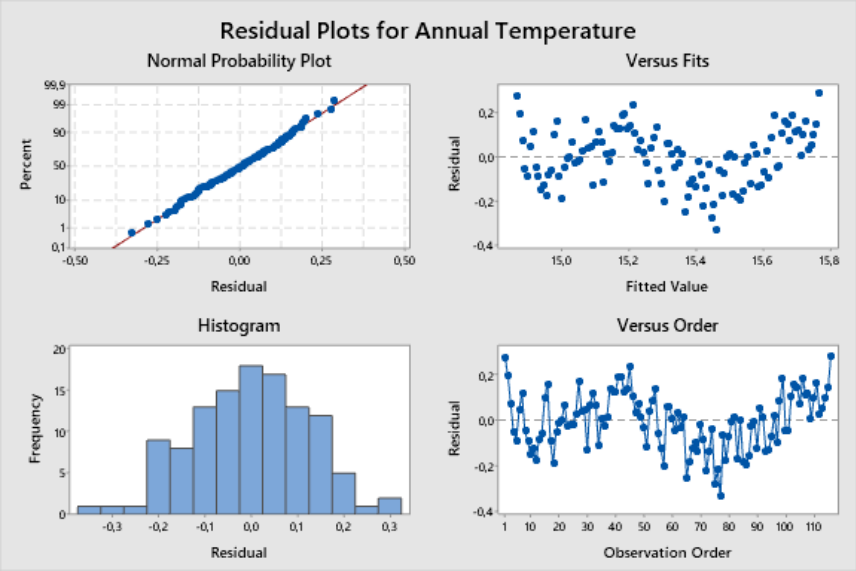


Figure 2: Residual analysis of the Linear Regression

ii. ARIMA

As previously discussed, the ARIMA model is applied to model the time series. However, the stationarity of the dataset must be tested.

In Fig.1 (a), the dataset presents a clear increasing tendency that includes changes in the average through time (non-stationarity). Therefore, there is evidence of the non-stationarity of the dataset. To reassure the findings, the Augmented Dickey-Fuller (ADF) test is performed and the results are in Table 2 (a). Having a p-value equals to 0.404, we fail to reject the null hypothesis and there is no evidence of stationarity of the data.

Now, a data transformation is needed and the differentiation method is chosen - the first derivative of the data. Again, the ADF test is used after transformation and the p-value is 0.003. Thus, there is enough evidence to reject the null hypothesis and conclude the time series is stationary. The graphical result is presented in the Appendix.

Table 2: ADF test: (a) Original data (b) Data after transformation

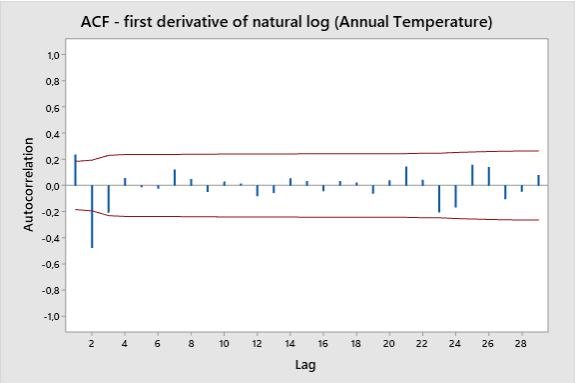
```
Original data

Augmented Dickey-Fuller Test Results:
ADF Test Statistic       -2.354037
P-Value                   0.404322
# Lags Used               3.000000
# Observations Used     112.000000
Critical Value (1%)      -4.041963
Critical Value (5%)      -3.450443
Critical Value (10%)     -3.150465
dtype: float64
Is the time series stationary? False
```

(a)

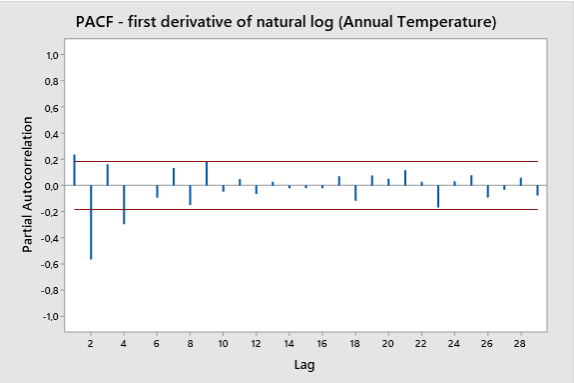
```
Data differentiated

Augmented Dickey-Fuller Test Results:
ADF Test Statistic       -4.269355
P-Value                   0.003509
# Lags Used               8.000000
# Observations Used     107.000000
Critical Value (1%)      -4.045971
Critical Value (5%)      -3.452348
Critical Value (10%)     -3.151576
dtype: float64
Is the time series stationary? True
```

(b)

ARIMA (p,d,q) is a regression method based on "p" parameters of autoregressive methods (AR), d parameters of integrative methods (I) and "q" parameters of moving average models (MA). The value for d is equal to 1 since the stationarity assumption was achieved after one differentiation of the data. To find the best pairs of "p" and "q", two correlograms are used: ACF and PACF.



(a)



(b)

Figure 3: (a) ACF and (b) PACF plot for first derivative of natural log (Annual Temperature)

In both graphs, the red lines are the significant limits with $\alpha = 0.05$. In Fig. 3 (b), the autocorrelation exponentially decreases after lag 2 having significant peaks at lag 1, 2 and

4. It suggests that the ARIMA model should focus on lag 1 and/or 2 – here, lag 4 will be discarded as we believe the model considering autocorrelation up to lag 1 and/or 2 is sufficient. Then, it suggests the two ARIMA models: ARIMA(0,1,1) and ARIMA(0,1,2). On the other hand, in Fig.3 (a), only two correlations are significant (lag 1 and 2) and, after lag 2, it decays exponentially. Therefore, it reassures the usage of the moving average [4].

To test the potential models, they were simulated using MINITAB and the results are in the Appendix.

Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
MA 1	0,6262	0,0742	8,44	0,000
Constant	0,000521	0,000236	2,21	0,029

Residual Sums of Squares

DF	SS	MS
113	0,0051080	0,0000452

(a)

Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
MA 1	0,4729	0,0915	5,17	0,000
MA 2	0,2471	0,0916	2,70	0,008
Constant	0,000541	0,000172	3,16	0,002

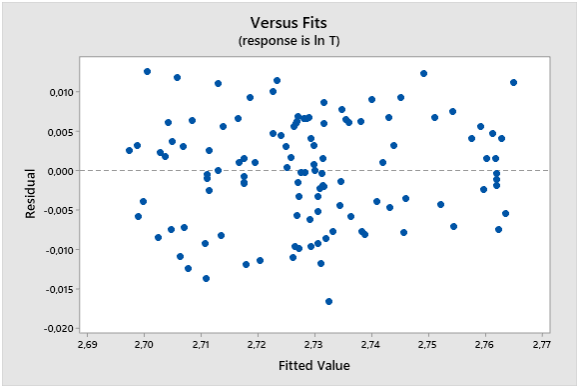
Residual Sums of Squares

DF	SS	MS
112	0,0046922	0,0000419

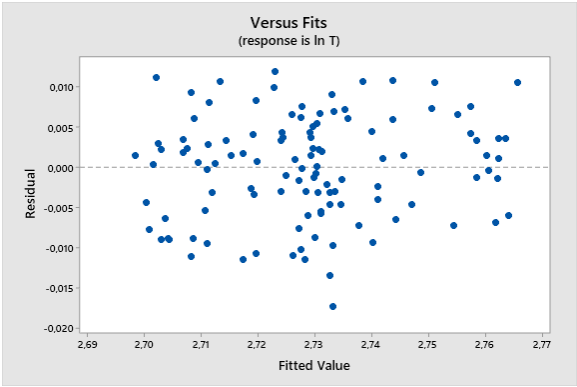
(b)

Figure 4: Partial results for ARIMA (0,1,1) (a) and ARIMA (0,1,2) (b)

Both presented statistically significant regression parameters having a p-value less than the level of confidence. However, the ARIMA(0,1,2) shows better accuracy in regards to residual sums of squares. Also, when the residuals plots are compared, the ARIMA(0,1,1) residuals show a small dependency on the residuals, which violates the regression assumptions. Therefore, we have enough evidence to use ARIMA (0,1,2) as a model to predict the temperatures through time.



(a)



(b)

Figure 5: Residuals Versus Fits: (a) ARIMA (0,1,1) and (b) ARIMA (0,1,2)

It is worth to mention that all assumptions were tested to ARIMA(0,1,2), as can be seen in the Appendix, and all of them are met.

5) Conclusions and Recommendations

It is a fact that the development of the internet is a milestone in human progress and it has changed several aspects in our society – from connecting families around the globe to improve management systems using the cloud. However, it has also opened a space for several threats, such as hackers, leaking of personal information, the spread of fake theories, among others. This project aims to clarify uncertainties related to global warming and the increasing tendency of global temperature – topics that have been labeled as potential lies as well as the shape of the Earth.

The data is collected by third-parties organizations that consolidated all values into the present dataset. However, neither information about measurement equipment nor data management tracking was provided. Then, this report may include errors such as "bad data" and "data management" [5],but we are going to consider these factors have no impact on the main findings.

A simple linear regression method was performed to understand the tendency/slope of the data. With a p-value less than the level of confidence, the regression model was significant statistically. However, after residual analysis, the correlation of residuals was noticed and alternative methods were needed. For time series analysis, the chosen model was ARIMA due to its vast application and rich statistical meaning.

Two statistical approaches were conducted to check the stationarity of the data – a critical assumption for ARIMA – and they were: data 2D-plot and the Augmented Dickey-Fuller (ADF) test. The plot showed an increasing tendency and, after the ADF test, the null hypothesis was not rejected with p-value equals to 0.404 (>0.05). Therefore, there are no enough evidence of data stationarity. In other words, there is evidence to confirm the increasing tendency of global annual temperature within the periods considered in this work. Considering global warming is associated with the increase in global temperature, we have evidence that global warming exists.

Further statistical tools and data transformation discussed in section 5 were used to find the best set of ARIMA parameters, and the candidates were: ARIMA (0,1,1) and ARIMA(0,1,2). Both regression parameters are statistically significant with p-values less than 0.05, but the sum of squares for the second model is smaller than the first (SS_{second} = 0.004692 < SS_{first} = 0.005180). Additionally, after residual analysis, the ARIMA (0,1,1) model presented small autocorrelation where the second model did not. Therefore, the ARIMA(0,1,2) is the best model to predict the temperature through the years (1900-2015).

This project suggests as next steps a deeper study on the ARIMA(0,1,2) model to forecast future global temperature. Predicting values outside the range in which the model was adjusted is dangerous, but it may assist in best & worst case-scenarios studies and help organizations in proactive planning to reduce potential casualties in the envinronment. Moreover, additional information is required to estimate the uncertainties of the statistical tests and models. Knowing that the global annual temperature is an average of temperatures around the globe (on land and sea), having the confidence interval of this mean and the precision of the measurement equipment may give different conclusions about global warming.

6) References

[1] BBC Future, *"Lies, propaganda and fake news: A challenge for our age"*. Available: https://www.bbc.com/future/article/20170301-lies-propaganda-and-fake-news-a-grand-challenge-of-our-age [Accessed 15th March 2020]

[2] NASA Global Climate Change, *"A Degree of Concern: Why global temperatures matter"*. Available: https://climate.nasa.gov/news/2878/a-degree-of-concern-why-global-temperatures-matter/ [Accessed 15th March 2020]

[3] Wooldridge, Jeffrey M. "*Introductory Econometrics: A Modern Approach".* Michigan: South Western, 2013. 5 ed.

[4] David M. Levine, Patricia P. Ramsey, Robert K. Smidt. "*Applied Statistics for Engineers and Scientists".* Prentice-Hall, Inc, 2001.

[5] Brown, A. W. ; Kaiser, K. A.; Allison, D. B., "Issues with data and analyses: Errors, underlying themes, and potential solutions"; PNAS, vol 115-11

[6] Kaggle, "Climate Change: Earth Surface Temperature Data". Available: https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data [Accessed 26th February,2020]

7) Appendix

A. **Data -** Annual temperature in degrees Celsius.

Table 3: Annual global temperature dataset

year	Annual Temp	year	Annual Temp	year	Annual Temp
1900	15.144	**1941**	15.385	**1982**	15.342
1901	15.073	**1942**	15.325	**1983**	15.52
1902	14.958	**1943**	15.345	**1984**	15.344
1903	14.837	**1944**	15.449	**1985**	15.341
1904	14.81	**1945**	15.331	**1986**	15.384
1905	14.955	**1946**	15.263	**1987**	15.525
1906	15.032	**1947**	15.31	**1988**	15.556
1907	14.875	**1948**	15.261	**1989**	15.442
1908	14.838	**1949**	15.223	**1990**	15.629
1909	14.79	**1950**	15.14	**1991**	15.598
1910	14.819	**1951**	15.311	**1992**	15.453
1911	14.776	**1952**	15.363	**1993**	15.466
1912	14.877	**1953**	15.421	**1994**	15.535
1913	14.909	**1954**	15.231	**1995**	15.638
1914	15.076	**1955**	15.174	**1996**	15.525
1915	15.144	**1956**	15.106	**1997**	15.714
1916	14.901	**1957**	15.374	**1998**	15.826
1917	14.81	**1958**	15.382	**1999**	15.6
1918	14.96	**1959**	15.341	**2000**	15.611
1919	15.007	**1960**	15.292	**2001**	15.768
1920	15.026	**1961**	15.38	**2002**	15.829
1921	15.102	**1962**	15.326	**2003**	15.827
1922	15.012	**1963**	15.377	**2004**	15.757
1923	15.028	**1964**	15.117	**2005**	15.879
1924	15.039	**1965**	15.196	**2006**	15.814
1925	15.09	**1966**	15.265	**2007**	15.827
1926	15.241	**1967**	15.294	**2008**	15.721
1927	15.122	**1968**	15.264	**2009**	15.827
1928	15.132	**1969**	15.391	**2010**	15.896
1929	14.967	**1970**	15.333	**2011**	15.77
1930	15.173	**1971**	15.2	**2012**	15.802
1931	15.228	**1972**	15.293	**2013**	15.854
1932	15.188	**1973**	15.405	**2014**	15.913
1933	15.013	**1974**	15.168	**2015**	16.059
1934	15.145	**1975**	15.239		
1935	15.119	**1976**	15.131		
1936	15.166	**1977**	15.408		
1937	15.301	**1978**	15.301		
1938	15.293	**1979**	15.414		
1939	15.302	**1980**	15.492		
1940	15.372	**1981**	15.516		

B. **Descriptive analysis – Global temperature**

Statistics

Variable	Mean	StDev	Variance	Minimum	Q1	Median	Q3	Maximum	Range
ave Temp	15,317	0,292	0,0855	14,776	15,120	15,302	15,485	16,059	1,283

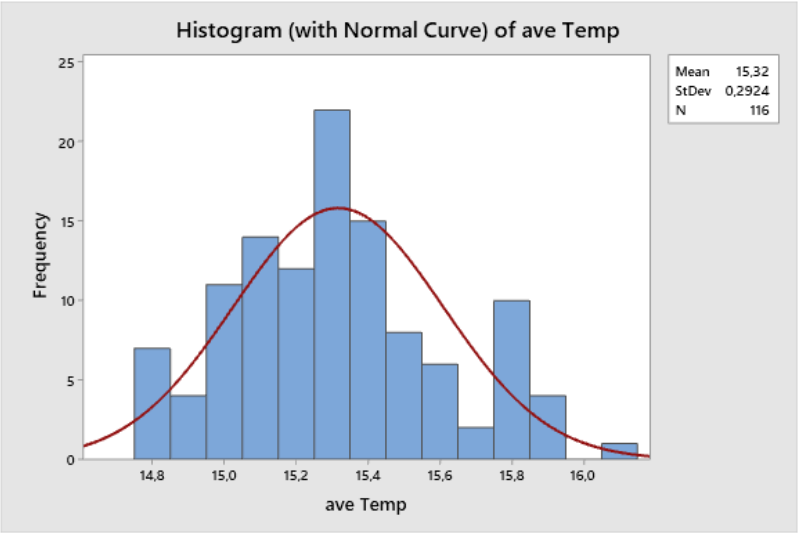


Figure 6: Descriptive analysis on global temperature

C. Time series plot: First derivative of natural log (Annual Temperature)

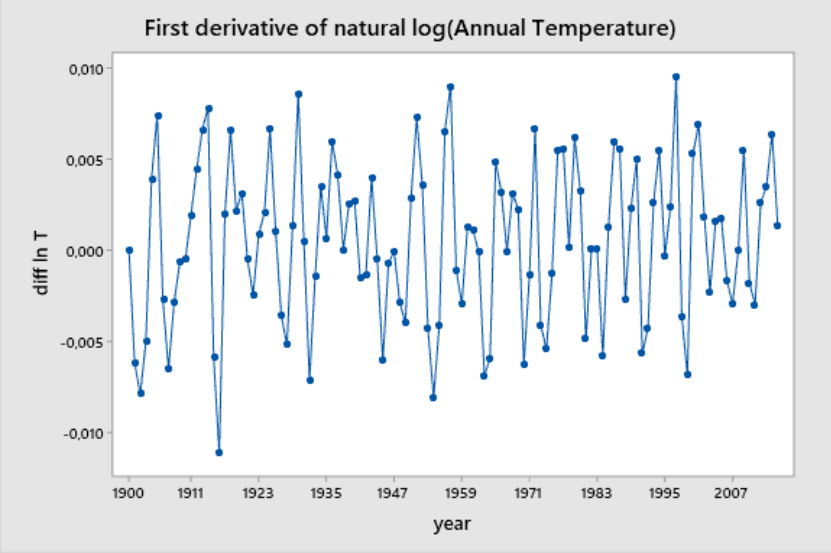


Figure 7: Time series plot after data transformation

D. ARIMA (0,1,2)

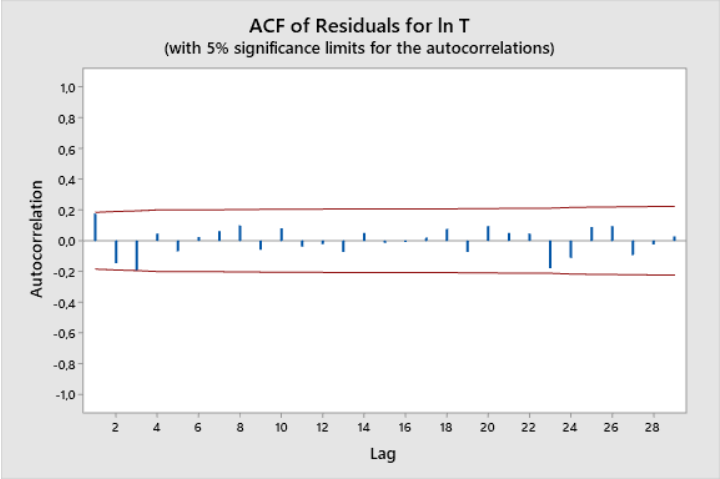


Figure 8: ARIMA (0,1,2) - ACF plot for residuals

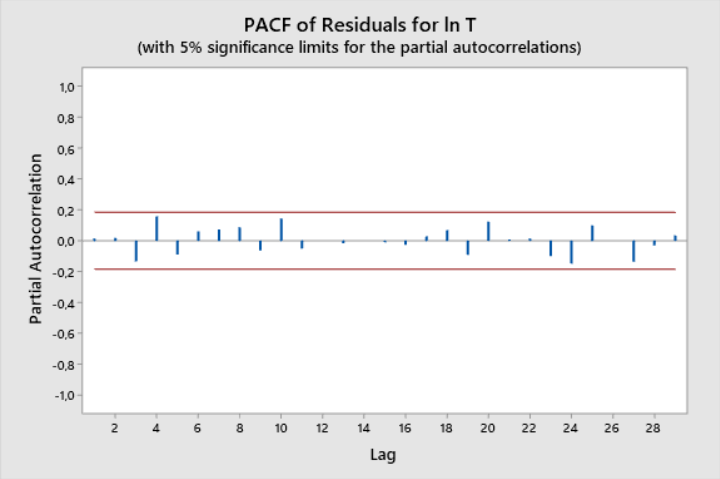


Figure 9: ARIMA (0,1,2) - PACF plot for residuals

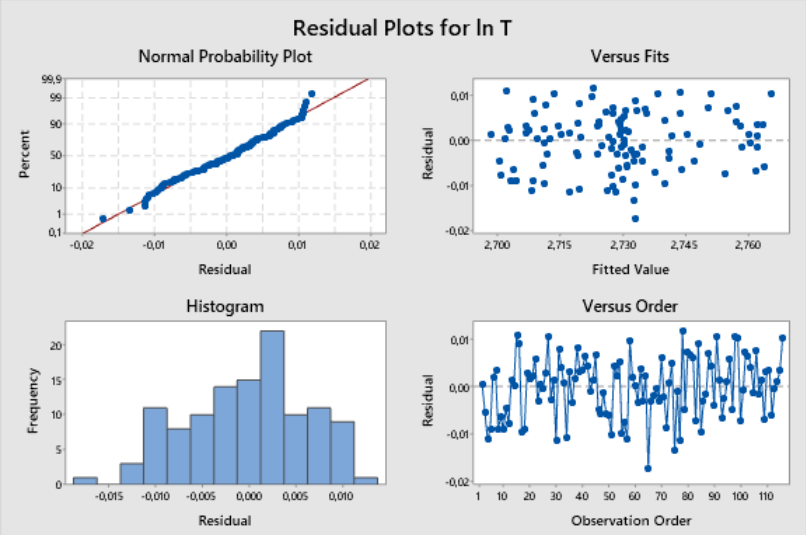


Figure 10: ARIMA (0,1,2) - Residual analysis

E. ARIMA (0,1,1)

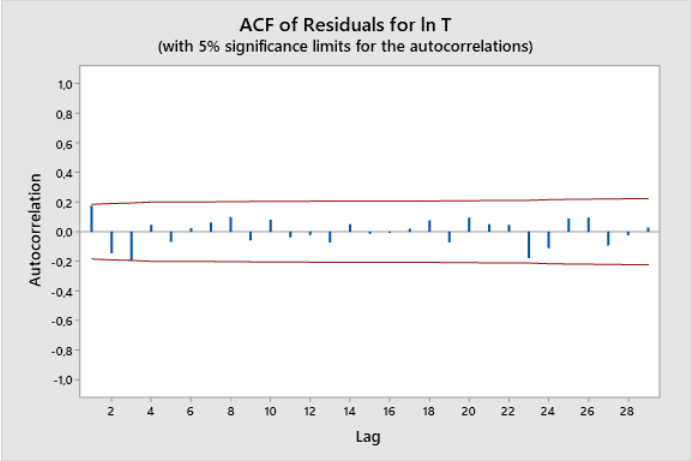


Figure 11: ARIMA (0,1,1) - ACF plot for residuals

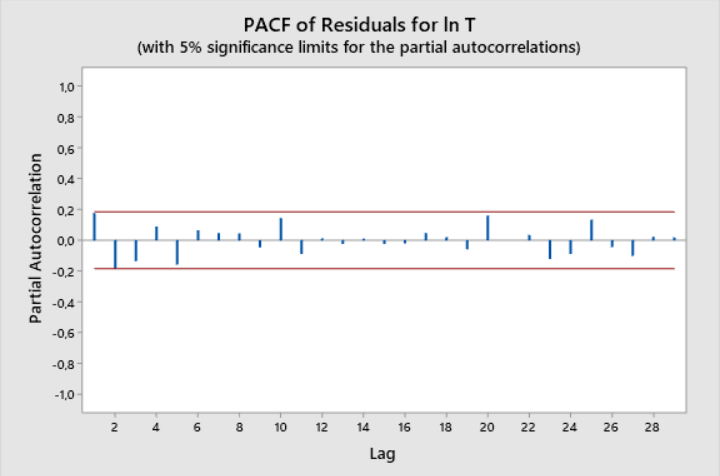


Figure 12: ARIMA (0,1,1) - PACF plot for residuals

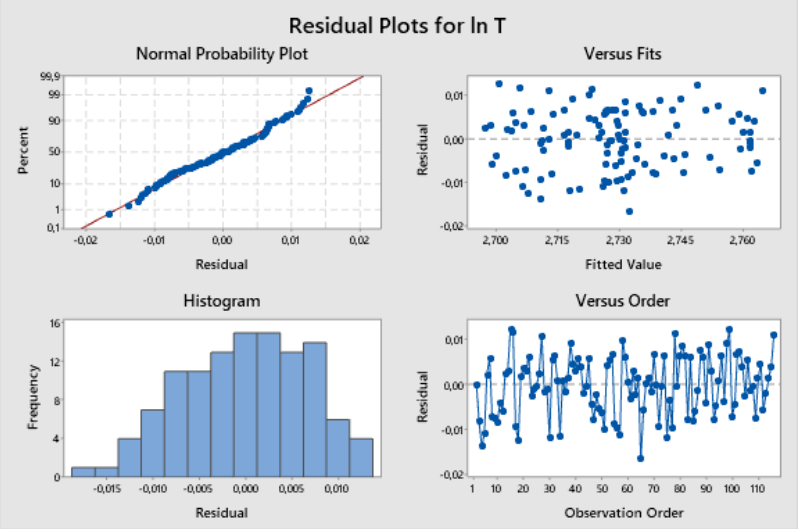


Figure 13: ARIMA (0,1,1) - Residual analysis