STA457 : Time Series Analysis of Earth's Geomagnetic Pole Intensities

Prajwol Poudel

April 12, 2023

Abstract

The Earth's magnetic field plays a crucial role in various aspects of our planet. Over the course of Earth's history the north and south poles vary in power and orientation referred to as "geomagnetic secular variation". In this study, the time series data was analyzed and modelled using ARIMA and ARMA time series models in an attempt to identify any anomalies or interesting conclusions about the north geomagnetic field intensity(NMI) and south geomagnetic field intensity(SMI). The results of the study showcase a seasonal or periodic pattern in the time series but the forecasts from the models fail to capture this phenomenon.

1 Introduction

1.1 Geomagnetic Intensity Data

The data provided to us has records of geomagnetic intensity of both the North and the South poles ranging from 65,000-70,000 nanoTeslas. To put the data into perspective,a single household fridge magnet has a strength of $1 * 10^8$ nanoteslas, so we are dealing with very weak intensities. The data recorded is measured from a singular location not specified and has records ranging from 1590-2023.Due to the phenomena known as "geomagnetic secular variation", the power and orientation of these intensities change over time and the poles have been moving away from their original location. From 1900 to 2010, the Nnorth Magnetic Pole shifted, left the territory of Canada in the 1990s, and is continuing moving in Canada's Arctic waters(Zverva, 2012). The movement of these poles have a huge impact on Earth and its inhabitants. One real life impact of this phenomenon is evident the migratory patterns of birds and how they use the Earth's magnetic fields to orient themselves. In a study conducted by Ritz et.al, migratory birds were exposed to oscillating geomagnetic fields parallel to the Earth's magnetic fields but immediately became disoriented when the fields were shifted away from the Earth's magnetic fields (Ritz, Thalan, Philips, Wiltschko Wiltschko, 2004). Even if these magnetic intensities are weak they have profound effects with small changes.

1.2 Objective

The motivation behind this study is to analyze the geomagnetic data and find the step ahead forecasts of the geomagnetic intensities of both the north and south poles. These analyses might provide us with a hint as to where the poles are moving next,if they are moving along a pattern or the rate at which these changes are occuring. There are several important questions that can be considered when analyzing this dataset but this is a complex problem, so no concrete result can be found as well.For this study, we are going to follow the box-Jenkins methodology to create two robust ARMA and ARIMA models that can forecast the geomagnetic intensities to a high accuracy.

2 Model Specification

The data we have is time dependent, so the realization of the data is dependent on a specified unit of time. Traditional statistical approaches fail to capture the underlying probabilty from the observed stochastic time series data. Therefore, we are going to use classical time series linear methods to model our time series data, more specifically ARMA and ARIMA models.

The ARMA(p,q) model is a mixture of two commonly used time series models (AR and MA models) containing p AR(auto-regressive) terms and q MA(moving-average):

$$X_t = \phi X_{t-1} + ... + \phi X_{t-p} - W_t - \theta_1 W_{t-1} - .... - \theta_q W_{t-q} \tag{1}$$

Here, W_t is a purely random process with mean zero and variance σ_w^2 and X_{t-1} represents the one time step backward stochastic point.

For a time series data, it is important to analyze the degree of affect one variable has on another variable as the stochastic processes are dependent on past processes. This process is formally defined as taking the covariance of the variables. For time sreis data with a stochastic process, we define the autocovariance function as the covariance between X_t and $X_{t+\tau}$, τ represents the time lag parameter:

$$\gamma_{t,t+\tau} = Cov(X_t, X_{t+\tau}) = E[(X_t - \mu_t)(X_{t+\tau} - \mu_{t+\tau})]t, \tau = 0, \pm 1, \pm 2... \tag{2}$$

Then, the autocorrelation function standardizes the autocovariance results as X_t and $X_{t+\tau}$ can be of different units:

$$\rho_{t,t+\tau} = \frac{Cov(X_t, X_{t+\tau})}{\sqrt{Var(X_t)Var(X_{t+\tau})}} = \frac{\gamma_{t,t+\tau}}{\sqrt{\gamma_{t,t}\gamma_{t+\tau,t+\tau}}} \tag{3}$$

The acf coefficients range from 1 to -1 and the closer the coefficients are to 1 or -1 the stronger the evidence of a linear relationship between the two stochastic points.

$$\rho_{t,t+\tau} = \frac{Cov(X_t, X_{t+\tau})}{\sqrt{Var(X_t)Var(X_{t+\tau})}} = \frac{\gamma_{t,t+\tau}}{\sqrt{\gamma_{t,t}\gamma_{t+\tau,t+\tau}}} \tag{4}$$

While the ACF coefficeints are a good indicator of a MA process is zero for lags beyond q, however for an AR process the the coefficients beyond lag p do not reduce to zero. So, we also use another function for identifying the AR and MA signatures known as the PACF (partial auto-correlation function):

$$\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1}, \rho k - j}{1 - \sum_{j=1}^{k-1} \phi_{k-1}, \rho j} \tag{5}$$

where $\phi_{k,j} = \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j}$ for j = 1,2,...,k-1

In order to identify the required number of AR and MA terms for the model we need to analyze the ACF(autocorrelation) and PACF(partial-autocorrelation) plots and identify the time signature of each MA(q) and AR(p) process.

Before we fit our data to this model, one important assumption about the data must be satisfied. The time series we have must be 'stationary' which satisfies the assumptions:

- E[X_t] is constant over time
- Var[X_t] is constant over time
- The autocovariance: $\gamma_\tau = \gamma_{t,t+\tau}$ and $\rho_\tau = \rho_{t,t+\tau}$ can be expressed in the form of the time lag parameter τ

By analyzing the times series, ACF and PACF plots of the North geomagnetic intensity data we can see that the data is non-stationary.

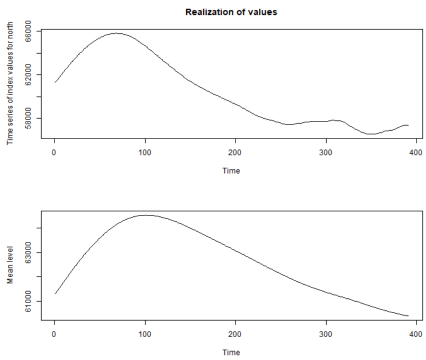

Figure 1: Time series plot and mean plot of the realizations of the North geomagnetic time series

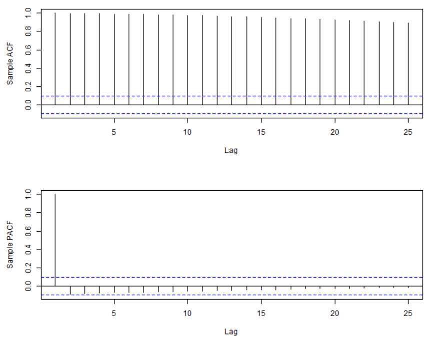

Figure 2: ACF and PACF plots of the time series

By analyzing the times series, ACF and PACF plots of the South geomagnetic intensity data we can see that the data is non-stationary. From the time series plot we can see that the data has a linear trend.

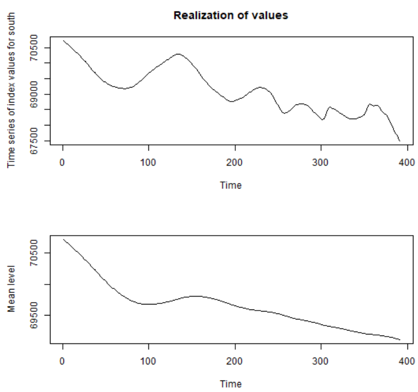

Figure 3: Time series plot and mean plot of the realizations of the North geomagnetic time series

To formally test if a time series is stationary we conduct three statistical hypothesis tests for a time series: ADF(Augmented-Dickey Fuller), KPSS for level stationary and KPSS for trend stationary.

Results of the stationary tests also suggest that the time series data is non-stationary as ADF p-values > 0.05 and KPSS p-values are < 0.05:

Method	North data	South data
ADF test	0.06945(lag= 17)	0.1872(lag=17)
KPSS for level	0.01(lag=4)	0.01(lag=4)
KPSS for trend	0.01(lag=4)	0.01(lag=4)

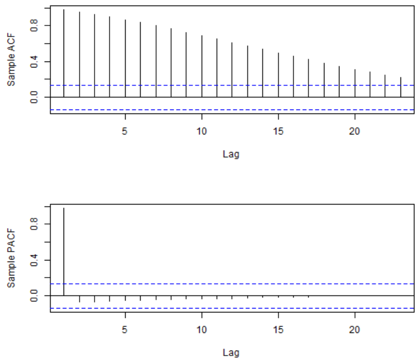

Figure 4: ACF and PACF plots of the time series

Now, in order to make the time series data into stationary we will first try two methods: for the NMI data we will use simple differencing applied twice and for the SMI data we will detrend the data.

The ADF and KPSS test results of this first attempt at making the data stationary can be seen in Appendix 7.1. These attempts failed to produce a stationary model, so we will explore different techniques to make the series stationary.

Since, detrending the data did not work for the SMI data, we differenced the data twice which produced a stationary series:

Method	North data	Test stat
ADF test	0.01(lag= 17)	-4.9406
KPSS for level	0.1(lag=4)	0.058337
KPSS for trend	0.1(lag=4)	0.024827

The differenced SMI time series, ACF and PACF plots are displayed below:

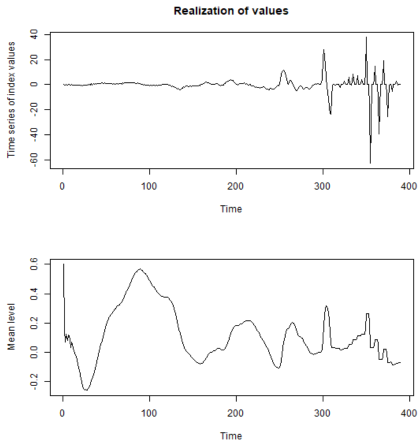

Figure 5: Time series plot and mean plot of the realizations of the differenced SMI time series

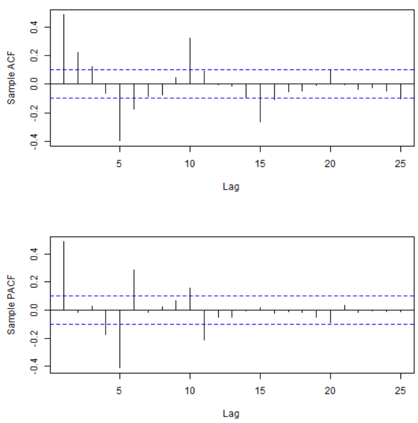

Figure 6: ACF and PACF plots of the SMI time series

The ACF plot has a damped wave pattern which might suggest a AR(2) process as it does not damp down fast enough for a AR(1) process. The PACF plot does not have a significant lag point at 2, so it does not necessarily match the PACF signature of an AR(2) model. The lag points 5,6,10 and 11 are points of interest. So, from the ACF and PACF plots we cannot model the data as a pure MA(q) or AR(p) processes. Furthermore, we can look at the EACF table to identify the AR and MA terms (Appendix 7.2). The EACF table suggests a ARMA(4,1) model

For the NMI time series, various methods were applied: log transformation, BoxCox transformation, seasonal differncing with different lag points, differencing with log transformation but none of these methods produced a stationary series that can pass our tests. Therefore, as a last resort attempt the number of realizations in the process were reduced. the data points from 1590-1799 were removed as we are interested in forecasting future data points. Therefore, our new training set consists of 202 observations compared to 391 from before.

After the reduction in the training set, the data was differenced twice which resulted in a stationary time series that passes all of the stationary tests:

Method	North data	Test stat
ADF test	0.01(lag= 17)	-3.4818
KPSS for level	0.1(lag=4)	0.084884
KPSS for trend	0.1(lag=4)	0.085492

The differenced NMI time series, ACF and PACF plots are displayed below:

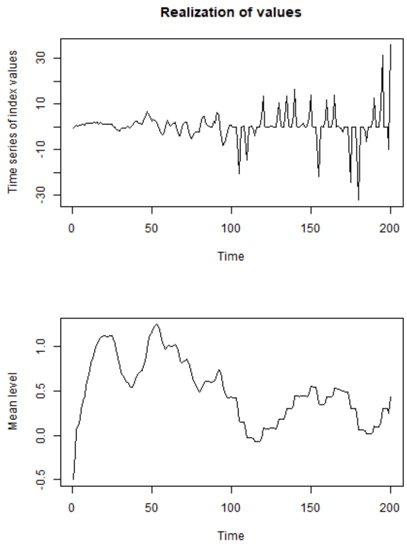

Figure 7: Time series plot and mean plot of the realizations of the differenced NMI time series

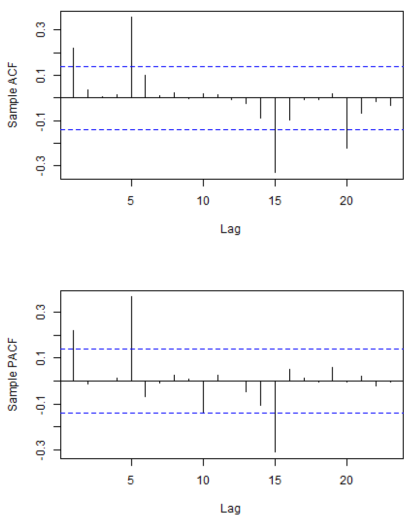

Figure 8: ACF and PACF plots of the NMI time series

The ACF plot and PACF plots have a very significant ACF at lag 1 which is a time signature for a MA(1) model. Other ACF values that are significant are at lag points 5 and 15. Furthermore, we can look at the EACF table to identify the AR and MA terms (Appendix 7.3). The EACF table suggests a ARMA(1,0), ARMA(0,1) and ARMA(4,0) model

The EACF models and the signatures are suggestions for an ARMA or ARIMA model. For more sophisticated model selection criterias we will look at the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) scores of various ARMA and ARIMA models:

AIC chooses the model with the best fit measured by the likelihood function which is subjected to a penalty term in order to punish models that have more parameters which are more likely to overfit. For our ARIMA model we desire a parsimonious model with as fewer parameters as possible. Since, our dataset is small we will use the bias corrected AIC and is formally defined as:

$$AIC_c = -2log(max.likelihood) + \frac{2rN}{N-r-1} \tag{6}$$

Here, the penalty term is $\frac{2rN}{N-r-1}$ and r represents the number of parameters.

Conversely, BIC is another model selection criteria that punishes models that overfit more severely than AIC and is formally defined as:

$$BIC = -2log(max.likelihood)(r + logN) \tag{7}$$

The AIC and BIC scores of ARMA models for the differenced SMI and NMI time series with parameters ranging from p,q = 1 to 5 (Appendix 7.3)

Candidate models:

SMI(candidate models)	NMI(candidate models)
ARMA(5,5)	ARMA(3,5)
ARMA(5,4)	ARMA(5,4)
ARIMA(5,2,5)	ARMA(3,1,5)
ARMA(4,1)	ARMA(4,0)
ARMA(3,5)	ARMA(3,2,5)
ARMA(5,4)	ARMA(2,1,2)
ARMA(1,5)	ARMA(5,2,4)
ARIMA(5,2,4)	ARMA(4,5)
ARMA(5,1,5)	ARMA(4,1,5)

Now, that we have a pool of models to choose from we will move onto choosing the best model and estimating its parameters in the next section.

3 Fitting and Diagnostics

For this section, we will look at the pool of our candidate models and choose the model that best fits our training data by analysing the residuals of the fitted models.

Residuals are defined as the difference between the predicted values by the model versus the training data we provided to the model. It is formally defined as: residual = actual- predicted and can be represented for ARMA models as:

$$\hat{e} = Y_t - \hat{\pi} Y_{t-1} - \hat{\pi} Y_{t-2} - .... \tag{8}$$

If a fitted model is correctly specified and the parameter estimates are close to the true values then the residuals should behave like a random stochastic process with normal distribution, constant mean, variance and no correlation between the residuals. Models that are not the best for our data tend to deviate from these properties of the residuals.
By analyzing the residual plot we can determine if the residuals have a constant mean and variance. From the ACF plot we can check to see if there exists correlations at individuals lag points. Furthermore, the Ljung-Box test is another formal test that checks the magnitudes of the residuals as a group. The final formal test is the Shapiro-Wilks test for normality which checks for the normality of the residuals.

From the pool of ARMA and ARIMA models for the SMI data, the models with residuals that best satisfied the desirable properties are: ARMA(5,0,5), ARMA(5,0,4) and ARIMA(5,2,5):

From the residual plot we see that the residuals seem uncorrelated except the ACF coefficient at lag 10 is significant and lag 15 is close to being significant. From the Ljung-box test we can see that there is a significant p-value present at lag 10 which supports the residual plot that lag 10 is a point of interest. For comparison the plots of the other models are in Appendix 7.4.

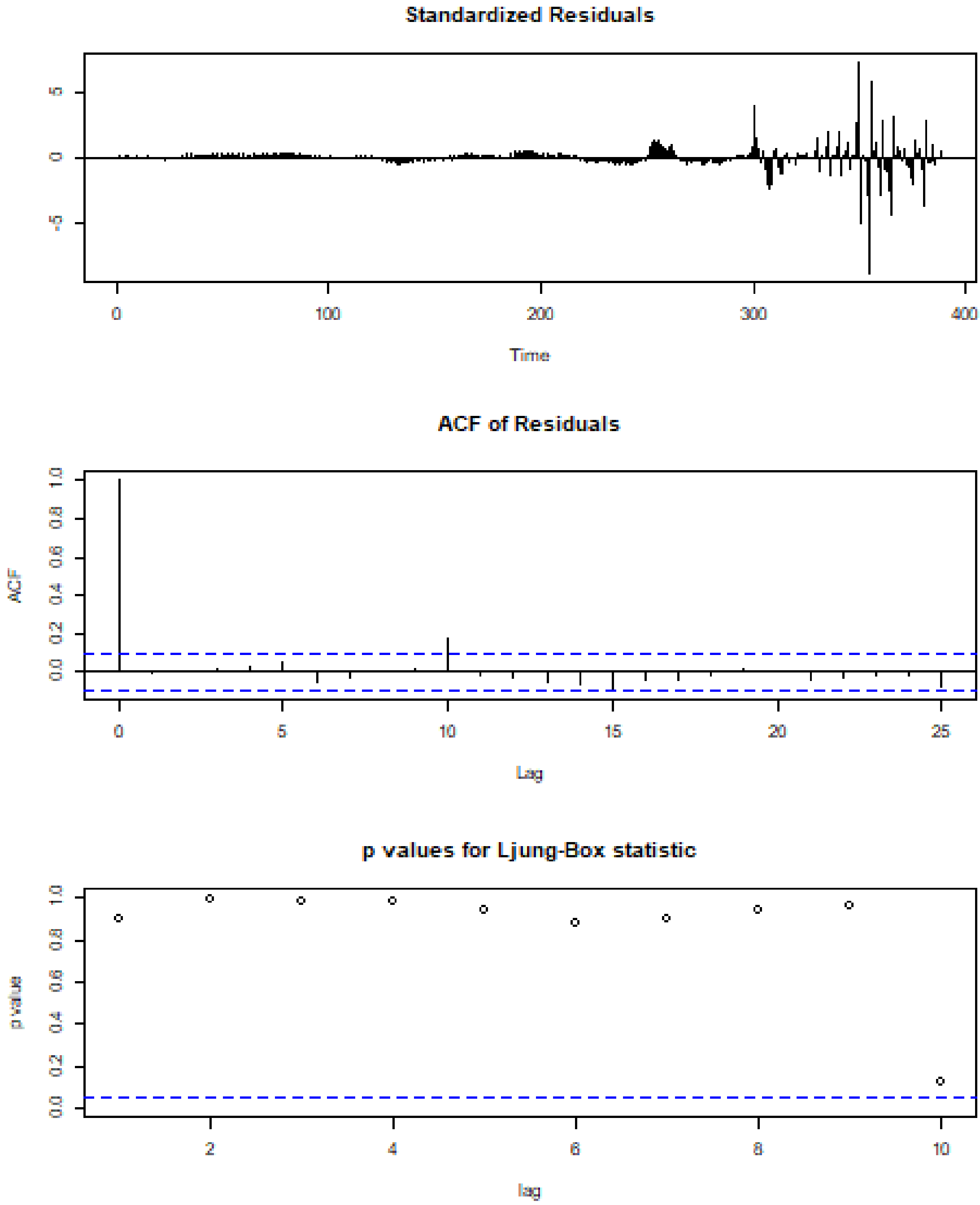

Figure 9: Residual, ACF and Ljung Box test plots for ARMA(5,0,4)

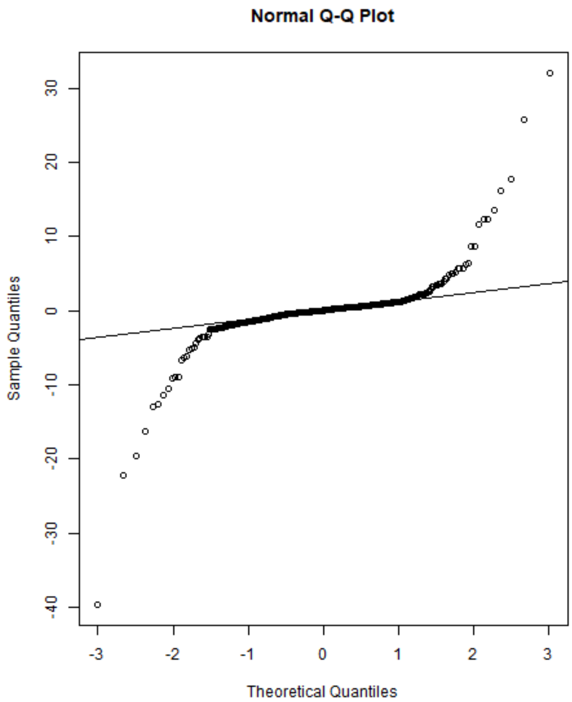

Figure 10: Q-Q plot for ARMA(5,0,4)

From the Q-Q plot we can see that the residuals are not perfectly normal as it does not exhibit a linear trend and also has a lot of outliers. The Shapiro-Wilks test provided a test-statistic of W = 0.60108 and a p-value = 2.2e-16. So, we do not reject the null-hypothesis and conclude that the data is not generated from a normal distribution. the other two models exhibit similar results for the normality test which displays that the failure of the normality test is unavoidable. In conclusion, even though the model does not pass the normality test we still choose this model as it produces the best results from following the box-jenkins methodology.

There are also other statistics that can determine the best model such as the AIC and BIC scores, standard errors and Residual mean squared error(RMSE) or Harmonic mean error(HME).

RMSE is more sensitive to outliers, so we will use HME as it is more robust statistic not as sensitive to outliers for comparison between the residuals of the models. It is formally defined as:

$$HME = \frac{N}{\frac{1}{x1} + \frac{1}{x1} + \frac{1}{x1} + ... + \frac{1}{xn}} \tag{9}$$

Here, the x_i are the residuals at the lag point i and N is the total number of observations.
The models with the best results for the differenced SMI data:

Method	(5,0,5)	(5,0,4)	(5,2,5)
Harmonic mean error	1.79	0.61	0.184
AIC	2284.07	2286.78	2302.18
BIC	2346.403	2371.157	2347.297
σ^2	19.52	19.77	20.34

Here, from the results, ARMA(5,0,4) model has the best results as it has a much lower harmonic mean than the ARMA(5,0,5) model, lower AIC and BIC scores than the ARIMA(5,2,5) model, as well as having less parameters which is a desirable property. The estimates of this model and its standard errors are:

	estimates	s.e.
AR(1)	-0.1167	0.0848
AR(2)	0.0302	0.0807
AR(3)	0.1636	0.0669
AR(4)	0.2538	0.0540
AR(5)	-0.5496	0.0522
MA(1)	0.7000	0.0982
MA(2)	0.3553	0.1244
MA(3)	0.0691	0.1150
MA(4)	-0.2089	0.0764

Similarly, from the pool of ARMA and ARIMA models for the NMI data, the models with residuals that best satisfied the desirable properties are: ARIMA(3,1,5), ARMA(5,2,4) and ARIMA(4,1,5).

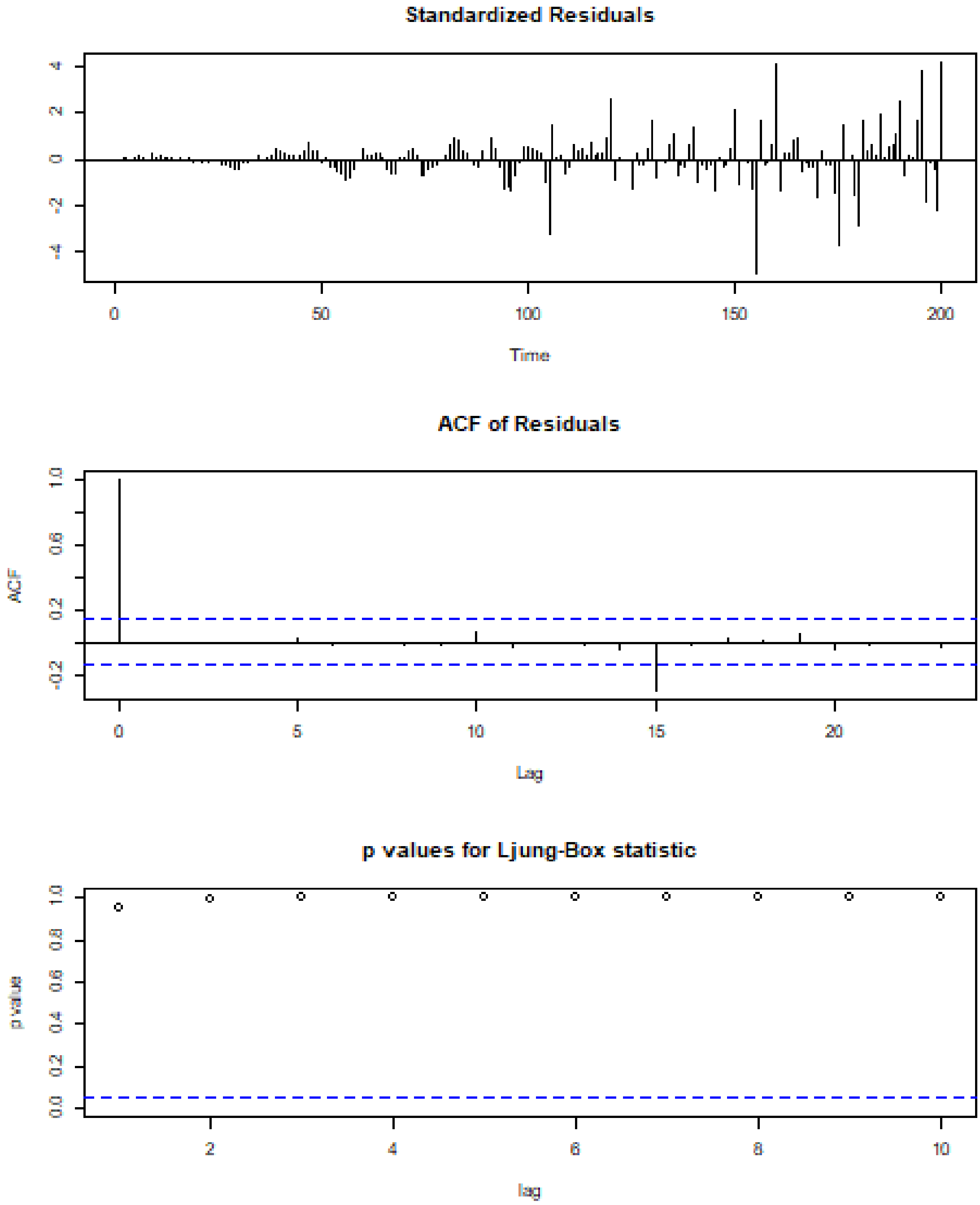

Figure 11: Residual, ACF and Ljung Box test plots for ARIMA(4,1,5)

From the residual plot we see that the residuals seem uncorrelated except the ACF coefficient at lag 15 is significant. From the Ljung-box test we can see that there is no significant p-value present at lag 10, however the range is only till lag 10, so we still consider lag 15 as a point of interest. For comparison the plots of the other models are in Appendix 7.5

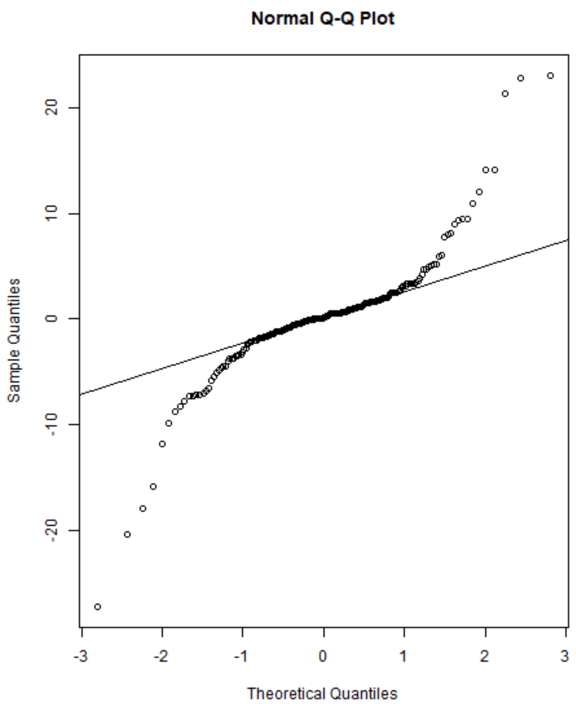

Figure 12: Q-Q plot for ARIMA(4,1,5)

From the Q-Q plot we can see that the residuals are not perfectly normal as it does not exhibit a linear trend and also has a lot of outliers. The Shapiro-Wilks test provided a test-statistic of W = 0.8399, p-value = 1.427e-13. So, we do not reject the null-hypothesis and conclude that the data is not generated from a normal distribution. the other two models exhibit similar results for the normality test which displays that the failure of the normality test is unavoidable. In conclusion, even though the model does not pass the normality test we still choose this model as it produces the best results from following the box-jenkins methodology.

The models with the best results for the differenced NMI data:

Method	(3,1,5)	(5,2,4)	(4,1,5)
Harmonic mean error	-0.104	-0.052	-0.0503
AIC	1269.63	1270.41	1266.52
BIC	1301.268	1305.29	1301.451
σ^2	31.57	31.03	30.73

Here, we can see that the ARIMA(4,1,5) model has the best results as it has the least HME, AIC,BIC and σ^2 values. We could consider the ARIMA(3,1,5) model as it has less terms but the results of ARIMA(4,1,5) is better in every statistic.The estimates of this model and its standard errors are:

	estimates	s.e.
AR(1)	-0.8142	0.2217
AR(2)	-0.6739	0.3010
AR(3)	-0.4337	0.3069
AR(4)	-0.4440	0.1243
MA(1)	0.1901	0.2197
MA(2)	-0.0938	0.2082
MA(3)	-0.2065	0.1703
MA(4)	-0.1746	0.2204
MA(5)	0.0401	0.1723

In conclusion, the model chosen for the SMI data is ARMA(5,4) and for the NMI data is ARIMA(4,1,5) following model diagnostics tests and comparison of important statistics while considering overfitting and the principle of parsimony.

4 Forecasting

For this section, we will use the final models for SMI and NMI data to forecast data into the future. The data used in this section will come from the test dataset. Since, we transformed our data from the original non-stationary series will have to apply inverse transformations in order to obtain the forecast in original terms:

For the ARMA(5,4) model the forecasts are given by:

1-step ahead:

$$\begin{aligned}\hat{Y}_t(1) = \mathit{Intercept} + \phi_1 Y_t + \phi_2 Y_{t-1} + \phi_3 Y_{t-2} + \phi_4 Y_{t-3} + \phi_5 Y_{t-4} \\ -\theta_1 \hat{W}_t - \theta_2 \hat{W}_{t-1} - \theta_3 \hat{W}_{t-2} - \theta_4 \hat{W}_{t-3}\end{aligned} \tag{10}$$

2-step ahead:

$$\hat{Y}_t(2) = Intercept + \phi_1 \hat{Y}_t(1) + \phi_2 Y_t + \phi_3 Y_{t-1} + \phi_4 Y_{t-2} + \phi_5 Y_{t-3} \\ -\theta_2 \hat{W}_t - \theta_3 \hat{W}_{t-1} - \theta_4 \hat{W}_{t-2} \tag{11}$$

This process can continue until however many n-step ahead forecast we want. Then, inverse transformations are applied to these forecasts to get the original forecasts.

The forecasted time series versus the actual observations for SMI data:

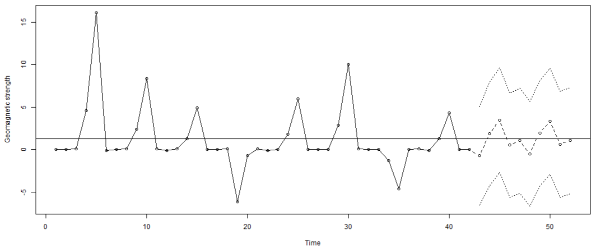

Figure 13: Forecast for the SMI time series

Forecasted 5-step ahead data:

Lag	Forecasted value
1	66755.07
2	66756.92
3	66760.41
4	66760.93
5	66761.99

Original SMI series with forecasted tail:
For the ARIMA(4,1,5) model the forecasts are given by:

1-step ahead:

$$\hat{Y}_t(1) = Intercept + \phi_1 Y_t + \phi_2 Y_{t-1} + \phi_3 Y_{t-2} + \phi_4 Y_{t-3} \\ -\theta_1 \hat{W}_t - \theta_2 \hat{W}_{t-1} - \theta_3 \hat{W}_{t-2} - \theta_4 \hat{W}_{t-3} - \theta_5 \hat{W}_{t-4} \tag{12}$$

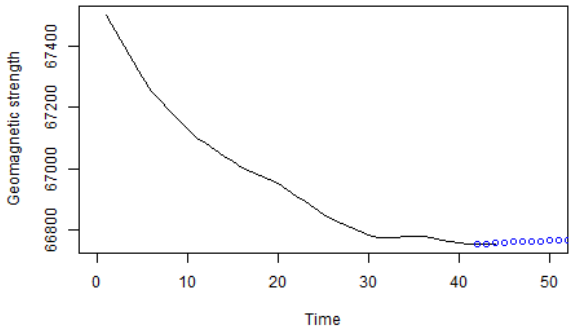

Figure 14: Forecast for the SMI time series

2-step ahead:

$$\begin{aligned} \hat{Y}_t(2) = Intercept + \phi_1 \hat{Y}_t(1) + \phi_2 Y_t + \phi_3 Y_{t-1} \\ -\theta_2 \hat{W}_t - \theta_3 \hat{W}_{t-1} - \theta_4 \hat{W}_{t-2} - \theta_5 \hat{W}_{t-3} \end{aligned} \tag{13}$$

Then,inverse transformations are applied to these forecasts to get the original forecasts.

The forecasted time series versus the actual observations for NMI data:

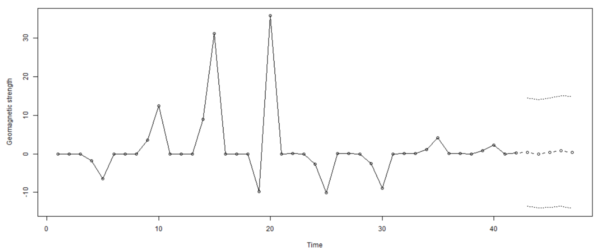

Figure 15: Forecast for NMI time series

Forecasted 5-step ahead data:

Lag	Forecasted value
1	57418.89
2	57418.87
3	57419.22
4	57419.98
5	57420.36

Original NMI series with forecasted tail:

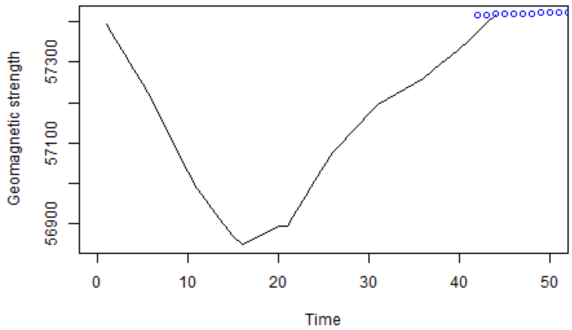

Figure 16: Forecast for the NMI time series

In this section, we used the models we developed to forecast the test data into future lag points. The forecasts generated from an ARMA or ARIMA model attempt to capture the auto correlation of the data points which would be missed in deterministic models. The RMSE for the models fitted with test data:

Data	RMSE
SMI	2.834352
NMI	6.904572

5 Discussion

The purpose of this project was to identify any significant information we can extrapolate from the SMI and NMI data. From the introduction section we know that the movement and power of these intensities can affect the daily lives of every being on Earth. The analysis of the time series data as well as the ACF coefficients of the transformed data set showcases evidence of seasonality. For the SMI and the NMI data, the ACF and PACF points at every 5 time lag units showcase a flipping trend or a significant increase in the coefficients. Therefore, some change might be occuring every five years whether it be in the intensity or the movement of the direction of these intensities. Although, these conclusions cannot be stated with confidence, the historical evidence suggest that major changes do occur and there is seasonality in the original dataset. The final forecasts do not necessarily support this finding as the tail prediction for the SMI data closely follow the trend and start to damp out, whereas the forecasts of the NMI data schowcases very minor changes in the intensities.

Therefore, there is some evidence that there exists seasonal change in the data but the forecasts predicted by the models developed in this project fail to capture these changes. This can mean that the chosen models were not the best choice to model this dataset and other models can lead to better forecasts with more evidence.More data might have been more useful as the movement of these poles might not have been a random process but might have been an influence of global factors and changes. So, analysing this dataset with more data might be more useful but might make the problem even more complex. Therefore, further analysis on this dataset is needed in order to have more concrete results.

6 Bibliography

Ritz, T., Thalau, P., Phillips, J. B., Wiltschko, R., Wiltschko, W. (2004). Resonance effects indicate a radical-pair mechanism for avian magnetic compass. Nature, 429(6988), 177–180. https://doi.org/10.1038/nature02534

Zvereva, T. I. (2012). Motion of the Earth's magnetic poles in the last decade. Geomagnetism and Aeronomy, 52(2), 261–269. https://doi.org/10.1134/S0016793212020168

7 Appendix

7.1 Detrending SMI data

```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 17
  STATISTIC:
    Dickey-Fuller: -3.1824
  P VALUE:
    0.01

Description:
 Tue Apr 11 23:09:06 2023 by user: poude

        KPSS Test for Level Stationarity

data:  y.diff2
KPSS Level = 0.20864, Truncation lag parameter = 5, p-value = 0.1


        KPSS Test for Trend Stationarity

data:  y.diff2
KPSS Trend = 0.19278, Truncation lag parameter = 5, p-value = 0.01871
```

Figure 17: R output of detrending SMI data

```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 17
  STATISTIC:
    Dickey-Fuller: -3.3641
  P VALUE:
    0.01

Description:
 Tue Apr 11 23:12:43 2023 by user: poude

        KPSS Test for Level Stationarity

data:  x.diff
KPSS Level = 0.69819, Truncation lag parameter = 5, p-value = 0.01371


        KPSS Test for Trend Stationarity

data:  x.diff
KPSS Trend = 0.17089, Truncation lag parameter = 5, p-value = 0.02926
```

Figure 18: First attempt at differencing NMI data

7.2 EACF tables

AR/MA	0	1	2	3	4	5
0	x	x	x	x	x	x
1	x	x	x	x	x	x
2	x	x	x	o	x	x
3	o	x	x	o	x	o
4	x	o	o	o	x	x
5	x	o	o	x	x	x

Figure 19: EACF table for SMI

AR/MA	0	1	2	3	4	5
0	x	o	o	o	x	o
1	o	o	o	o	x	o
2	x	o	o	o	x	o
3	x	o	o	o	x	o
4	o	o	o	x	x	o
5	x	o	o	o	x	o

Figure 20: EACF table for NMI

7.3 Full AIC and BIC scores

r5:c5	r3:c5	r5:c4	r4:c5	r2:c5	r5:c3	r1:c5	r4:c4	r4:c3	r3:c4	r5:c2	r5:c1	r2:c4
2304.767	2319.594	2333.485	2337.723	2346.214	2348.318	2348.827	2350.886	2351.432	2355.928	2376.635	2391.971	2394.115
r5:c5	r3:c5	r5:c4	r1:c5	r4:c5	r2:c5	r4:c3	r5:c3	r4:c4	r3:c4	r5:c2	r2:c1	r2:c2
2346.403	2353.303	2371.157	2374.609	2375.395	2375.959	2381.177	2382.027	2384.594	2385.673	2406.380	2411.852	2417.723

Figure 21: AIC scores first row and BIC scores second row for SMI

```
   r3:c5    r5:c4    r4:c5    r5:c3    r4:c4    r2:c5    r5:c5    r1:c5    r5:c2    r3:c3    r2:c2    r3:c2    r2:c4
1269.692 1271.142 1272.009 1273.444 1275.492 1278.068 1278.075 1280.970 1283.990 1284.913 1285.165 1285.748 1285.970
   r3:c5    r2:c2    r1:c2    r2:c1    r5:c3    r1:c5    r5:c4    r2:c5    r4:c5    r4:c4    r3:c2    r1:c3    r3:c1
1298.078 1300.358 1300.457 1300.906 1301.831 1302.760 1302.827 1303.156 1303.694 1303.878 1304.239 1305.723 1305.870
```

Figure 22: AIC scores first row and BIC scores second row for NMI

7.4 Residual, ACF and QQ-plots of SMI data

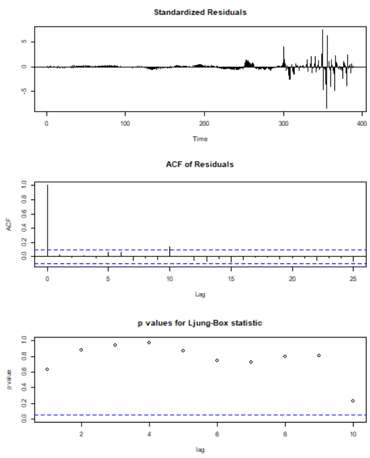

Figure 23: ARMA(5,5)

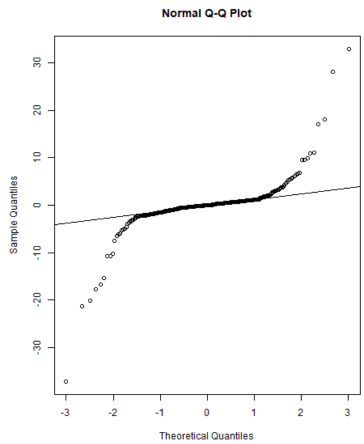

Figure 24: Q-Q plot ARMA(5,5)

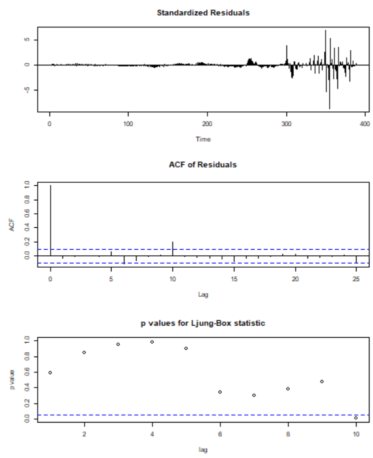

Figure 25: ARIMA(5,2,5)

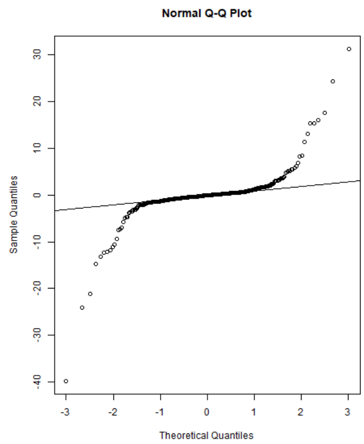

Figure 26: Q-Q plot ARMA(5,5)

7.5 Residual, ACF and QQ-plots of NMI data

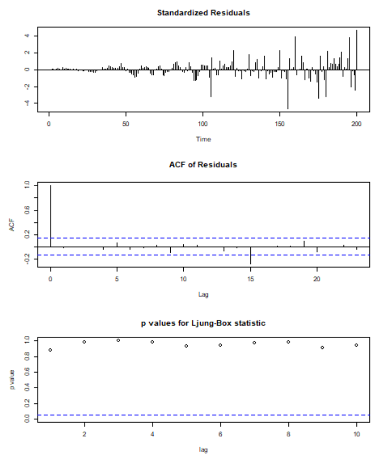

Figure 27: ARIMA(3,1,5)

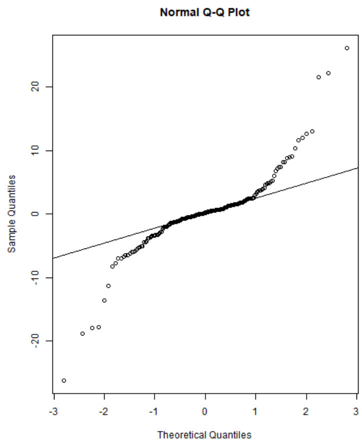

Figure 28: Q-Q plot ARIMA(3,1,5)

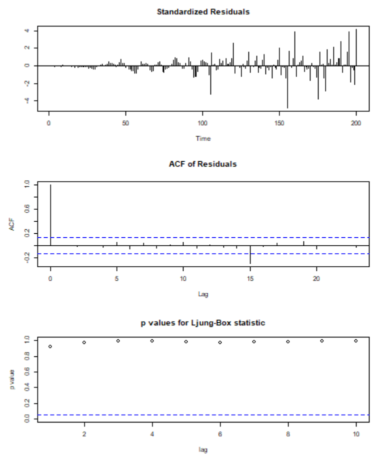

Figure 29: ARIMA(5,2,4)

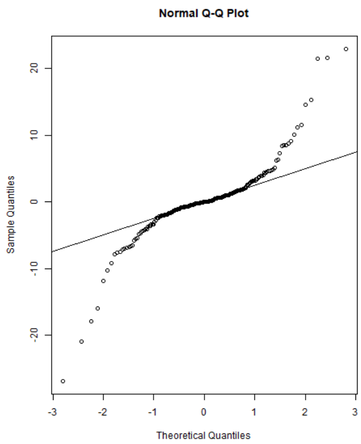

Figure 30: Q-Q plot ARIMA(5,2,4)

7.6 Code

Included in the following page

STA457_project

prajwol poudel

08/04/2023

```
library(fUnitRoots)
library(caTools)
library(tseries)
library(TSA)
library(psych)
setwd("C:\\Users\\poude\\Desktop\\time_series project")
df = read.csv("Geomagnetic_Intensity_Data.csv")
```

Including Plots

You can also embed plots, for example:

```
x = new_train$North_Geomagnetic_Pole
cummeanx <- cumsum(x)/ seq_along(x)
par(mfrow=c(2,1), mar = c(4,4,4,4))
plot(x, type="l", xlab = "Time",
     ylab="Time series of index values for north", main = "Realization of values")
plot(cummeanx, type = "l", xlab="Time", ylab= "Mean level")
acf(x,xlab= "Lag",ylab = "Sample ACF", main ="")
acf(x,type="partial",xlab = "Lag", ylab= "Sample PACF", main="")
x.adf <- adfTest(x, lags = 17)
x.adf2 <- adfTest(x, lags = 8)
x.adf3 <- adfTest(x, lags = 3)
x.kpss_level <- kpss.test(x, null = "Level")
x.kpss_trend <- kpss.test(x, null = "Trend")
x.adf
x.adf2
x.adf3
x.kpss_level
x.kpss_trend
```

```
x.diff2 = diff(diff(x))
cummeanx.diff2 <- cumsum(x.diff2)/ seq_along(x.diff2)
par(mfrow=c(2,1), mar = c(4,4,4,4))
plot(x.diff2, type="l", xlab = "Time",
     ylab="Time series of index values", main = "Realization of values")
```

```
plot(cummeanx.diff2, type = "l", xlab="Time", ylab= "Mean level")
acf(x.diff2,xlab= "Lag",ylab = "Sample ACF", main ="")
acf(x.diff2,type="partial",xlab = "Lag", ylab= "Sample PACF", main="")
x.diff2.adf <- adfTest(x.diff2, lags = 17)
x.diff2.adf2 <- adfTest(x.diff2, lags = 4)
x.diff2.adf3 <- adfTest(x.diff2, lags = 8)
x.diff2.kpss_level <- kpss.test(x.diff2, null = "Level")
x.diff2.kpss_trend <- kpss.test(x.diff2, null = "Trend")
x.diff2.adf
x.diff2.adf2
x.diff2.adf3
x.diff2.kpss_level
x.diff2.kpss_trend
```

North From the ACF diagram we can see a decay signature similar to an AR(1)

```
x.diff2.eacf <-eacf(x.diff2,ar.max=5,ma.max=5)
```

```
x2 <- x.diff2
x2.aic <- matrix(0,5,5)
x2.bic <- matrix(0,5,5)

for (i in 0:4) for (j in 0:4){
  library(gdata)
  x2.fit <- arima(x2, order=c(i,0,j), method = "ML",  include.mean = TRUE)
  x2.aic[i+1,j+1] <-x2.fit$aic
  x2.bic[i+1,j+1] <- BIC(x2.fit)
}

x2.aic_vec <- sort(unmatrix(x2.aic, byrow=FALSE))[1:13]
x2.bic_vec <- sort(unmatrix(x2.bic, byrow = FALSE))[1:13]

x2.aic_vec
x2.bic_vec
```

Checking the model with the lowest AIC and BIC: ARMA(3,5) model

```
x2.fit_1 <- arima(x2, order=c(3,0,5), method="ML",include.mean=TRUE)
tsdiag(x2.fit_1)
qqnorm(residuals(x2.fit_1))
qqline(residuals(x2.fit_1))
shapiro.test(residuals(x2.fit_1))
harmonic.mean(x2.fit_1$residuals)
```

From the residual plot the errors look to be uncorrelated but There is some significant acf values in the residual acf plot at lag 16.

checking if the residuals are normal by using a Q-Q plot and a shapiro-wilk test

```
x2.fit_1 <- arima(x2, order=c(3,2,5), method="ML",include.mean=TRUE)
tsdiag(x2.fit_1)
qqnorm(residuals(x2.fit_1))
```

```
qqline(residuals(x2.fit_1))
shapiro.test(residuals(x2.fit_1))
harmonic.mean(x2.fit_1$residuals)
```

```
x2.fit_1 <- arima(x2, order=c(3,1,5), method="ML",include.mean=TRUE)
tsdiag(x2.fit_1)
qqnorm(residuals(x2.fit_1))
qqline(residuals(x2.fit_1))
shapiro.test(residuals(x2.fit_1))
harmonic.mean(x2.fit_1$residuals)
BIC(x2.fit_1)
```

```
x2.fit_1 <- arima(x2, order=c(2,1,2), method="ML",include.mean=TRUE)
tsdiag(x2.fit_1)
qqnorm(residuals(x2.fit_1))
qqline(residuals(x2.fit_1))
shapiro.test(residuals(x2.fit_1))
harmonic.mean(x2.fit_1$residuals)
```

```
x2.fit_1 <- arima(x2, order=c(5,2,4), method="ML",include.mean=TRUE)
tsdiag(x2.fit_1)
qqnorm(residuals(x2.fit_1))
qqline(residuals(x2.fit_1))
shapiro.test(residuals(x2.fit_1))
harmonic.mean(x2.fit_1$residuals)
BIC(x2.fit_1)
```

```
x2.fit_1 <- arima(x2, order=c(4,1,5), method="ML",include.mean=TRUE)
tsdiag(x2.fit_1)
qqnorm(residuals(x2.fit_1))
qqline(residuals(x2.fit_1))
shapiro.test(residuals(x2.fit_1))
harmonic.mean(x2.fit_1$residuals)
BIC(x2.fit_1)
```

```
#candidate model
arima(x2, order=c(3,1,5), method="ML",include.mean=TRUE)
arima(x2, order=c(5,2,4), method="ML",include.mean=TRUE)
arima(x2, order=c(4,1,5), method="ML",include.mean=TRUE)
```

```
x = test$North_Geomagnetic_Pole
test.diff = diff(diff(x))
test.fit = arima(test.diff,order=c(4,1,5), method="ML",include.mean=TRUE)
plot(test.fit,n.ahead=5,type='b',xlab='Time',
ylab='Geomagnetic strength')
abline(h=coef(test.fit)[names(coef(test.fit))=='intercept'])
prediction<- predict(test.fit,n.ahead = 10)
pred <- prediction$pred
test
library(stats)
forecast<- diffinv(pred, xi = x[44])
forecast
```

```r
RMSE <- sqrt(sum((test.fit$residuals-mean(test.fit$residuals))**2)/nrow(test))
RMSE
plot(x,type = "l",xlab='Time',ylab='Geomagnetic strength', xlim = c(0,50)) +
  lines(forecast,col="blue", type = "b")
```

```r
y = train$South_Geomagnetic_Pole
cummeany <- cumsum(y)/ seq_along(y)
par(mfrow=c(2,1), mar = c(4,4,4,4))
plot(y, type="l", xlab = "Time",
     ylab="Time series of index values for south", main = "Realization of values")
plot(cummeany, type = "l", xlab="Time", ylab= "Mean level")
acf(y,xlab= "Lag",ylab = "Sample ACF", main ="")
acf(y,type="partial",xlab = "Lag", ylab= "Sample PACF", main="")
y.adf <- adfTest(y, lags = 17)
y.adf2 <- adfTest(y, lags = 4)
y.adf3 <- adfTest(y, lags = 8)
y.kpss_level <- kpss.test(y, null = "Level")
y.kpss_trend <- kpss.test(y, null = "Trend")
y.adf
y.adf2
y.adf3
y.kpss_level
y.kpss_trend
```

```r
library(astsa)
y.diff2 = diff(diff(y))
#y.diff2 = detrend(y, lowess = TRUE)
cummeany.diff2 <- cumsum(y.diff2)/ seq_along(y.diff2)
par(mfrow=c(2,1), mar = c(4,4,4,4))
plot(y.diff2, type="l", xlab = "Time",
     ylab="Time series of index values", main = "Realization of values")
plot(cummeany.diff2, type = "l", xlab="Time", ylab= "Mean level")
acf(y.diff2,xlab= "Lag",ylab = "Sample ACF", main ="")
acf(y.diff2,type="partial",xlab = "Lag", ylab= "Sample PACF", main="")
y.diff2.adf <- adfTest(y.diff2, lags = 17)
y.diff2.adf2 <- adfTest(y.diff2, lags = 4)
y.diff2.adf3 <- adfTest(y.diff2, lags = 8)
y.diff2.kpss_level <- kpss.test(y.diff2, null = "Level")
y.diff2.kpss_trend <- kpss.test(y.diff2, null = "Trend")
y.diff2.adf
y.diff2.kpss_level
y.diff2.kpss_trend
```

```r
y.diff2.eacf <-eacf(y.diff2,ar.max=5,ma.max=5)
```

```r
y2 <- y.diff2
y2.aic <- matrix(0,5,5)
y2.bic <- matrix(0,5,5)

for (i in 0:4) for (j in 0:4){
  library(gdata)
  y2.fit <- arima(y2, order=c(i,0,j), method = "ML",  include.mean = TRUE)
  y2.aic[i+1,j+1] <-y2.fit$aic
  y2.bic[i+1,j+1] <- BIC(y2.fit)
}

y2.aic_vec <- sort(unmatrix(y2.aic, byrow=FALSE))[1:13]
y2.bic_vec <- sort(unmatrix(y2.bic, byrow = FALSE))[1:13]

y2.aic_vec
y2.bic_vec
```

```r
y.fit_1 <- arima(y2, order=c(5,0,5), method="ML",include.mean=TRUE)
tsdiag(y.fit_1)
qqnorm(residuals(y.fit_1))
qqline(residuals(y.fit_1))
shapiro.test(residuals(y.fit_1))
harmonic.mean(y.fit_1$residuals)
```

```r
y.fit_1 <- arima(y2, order=c(5,1,5), method="ML",include.mean=TRUE)
tsdiag(y.fit_1)
qqnorm(residuals(y.fit_1))
qqline(residuals(y.fit_1))
shapiro.test(residuals(y.fit_1))
harmonic.mean(y.fit_1$residuals)
```

```r
y.fit_1 <- arima(y2, order=c(5,2,5), method="ML",include.mean=TRUE)
tsdiag(y.fit_1)
qqnorm(residuals(y.fit_1))
qqline(residuals(y.fit_1))
shapiro.test(residuals(y.fit_1))
harmonic.mean(y.fit_1$residuals)
BIC(y.fit_1)
```

```r
y.fit_1 <- arima(y2, order=c(3,0,5), method="ML",include.mean=TRUE)
tsdiag(y.fit_1)
qqnorm(residuals(y.fit_1))
qqline(residuals(y.fit_1))
shapiro.test(residuals(y.fit_1))
harmonic.mean(y.fit_1$residuals)
```

```r
y.fit_1 <- arima(y2, order=c(3,1,5), method="ML",include.mean=TRUE)
tsdiag(y.fit_1)
qqnorm(residuals(y.fit_1))
qqline(residuals(y.fit_1))
shapiro.test(residuals(y.fit_1))
harmonic.mean(y.fit_1$residuals)
```

```r
y.fit_1 <- arima(y2, order=c(5,0,4), method="ML",include.mean=TRUE)
tsdiag(y.fit_1)
qqnorm(residuals(y.fit_1))
qqline(residuals(y.fit_1))
shapiro.test(residuals(y.fit_1))
harmonic.mean(y.fit_1$residuals)
```

```r
y.fit_1 <- arima(y2, order=c(5,1,4), method="ML",include.mean=TRUE)
tsdiag(y.fit_1)
qqnorm(residuals(y.fit_1))
qqline(residuals(y.fit_1))
shapiro.test(residuals(y.fit_1))
harmonic.mean(y.fit_1$residuals)
```

```r
y.fit_1 <- arima(y2, order=c(4,0,5), method="ML",include.mean=TRUE)
tsdiag(y.fit_1)
qqnorm(residuals(y.fit_1))
qqline(residuals(y.fit_1))
shapiro.test(residuals(y.fit_1))
harmonic.mean(y.fit_1$residuals)
```

```r
y.fit_1 <- arima(y2, order=c(4,1,5), method="ML",include.mean=TRUE)
tsdiag(y.fit_1)
qqnorm(residuals(y.fit_1))
qqline(residuals(y.fit_1))
shapiro.test(residuals(y.fit_1))
harmonic.mean(y.fit_1$residuals)
```

```r
library(psych)
arima(y2, order=c(5,0,5), method="ML",include.mean=TRUE) #candidate model
arima(y2, order=c(5,0,4), method="ML",include.mean=TRUE)
arima(y2, order=c(5,2,5), method="ML",include.mean=TRUE)
arima(y2, order=c(5,1,5), method="ML",include.mean=TRUE)
```

```r
y = test$South_Geomagnetic_Pole
test.diff2 = diff(diff(y))
test.fit2 = arima(test.diff2,order=c(5,0,4), method="ML",include.mean=TRUE)
plot(test.fit2,n.ahead=10,type='b',xlab='Time',
ylab='Geomagnetic strength')
abline(h=coef(test.fit2)[names(coef(test.fit2))=='intercept'])
prediction<- predict(test.fit2,n.ahead = 10)
pred <- prediction$pred
test
library(stats)
forecast <- diffinv(pred, xi = y[44])
forecast
```

```r
RMSE <- sqrt(sum((test.fit2$residuals-mean(test.fit2$residuals))**2)/nrow(test))
RMSE
plot(y,type = "l",xlab='Time',ylab='Geomagnetic strength', xlim= c(0,50))
+ lines(forecast,col="blue", type = "b")
```

Southern data