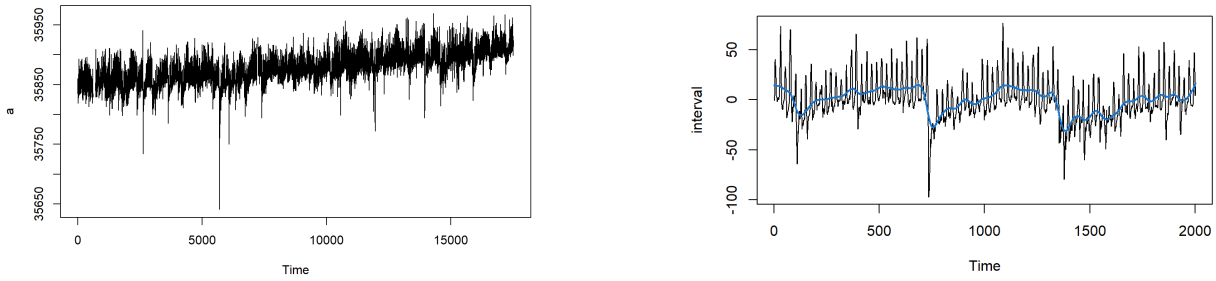
Time Series Analysis of Geomagnetic Data

20307130030 Shen Jianzhi

2023.6.24

1 Introduction

The data is the geomagnetic data from Jan. 2018 to Dec. 2019, with hourly temporal solution. It is collected from Archived data of WDC for geomag., Kyoto., and the observation was done by Guam Observatory, since low-latitude geomagnetic field has more variation. We take the horizontal component X as our uni-dimensional time series. And a plot of the data is shown in Figure 1 (a), where upward trend of the time series is probably caused by the west-ward drift of geomagnetic field.



(a) Original

(b) Detrended Local Plot with Smoothing Splines

Figure 1: Time Series Plot for the Geomagnetic Data

The problem we are addressing here is: 1. The possible underlying data generation mechanism and coarse forecast for geological research and navigation; 2. The forecast of sudden change, which prewarns geomagnetic storms, solar eruptions and even earthquakes. Since ARIMA may fail to model abrupt disturbance, we further try LSTM on fine-grained modeling.

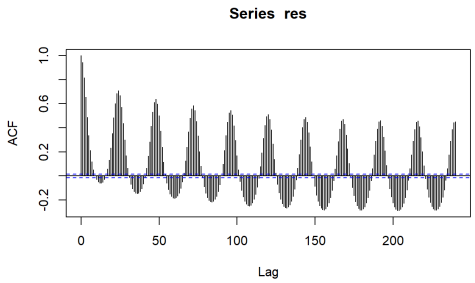
The data is split into 15,000 train items and 2,295 test items.

2 EDA

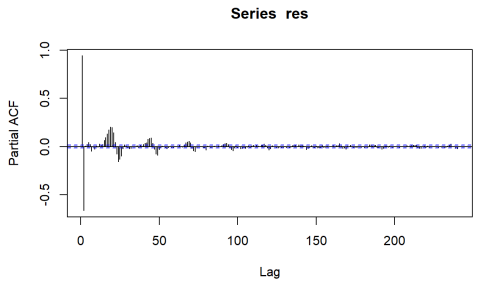
First, we linearly detrend the data, and smooth the time series with $\lambda = 0.5$ Smoothing Splines, as Figure 1 (b) shows.

As can be seen in the smoothing splines, there is a seasonal pattern about every 700 lapses, which is approximately one month. It is plausible because the moon may affect the geomagnetic field through change of gravity.

The ACF and PACF plot are shown in Figure 2. The ACF suggests that the time series has seasonal pattern every 24 hours, which is exactly a day.



(a) ACF

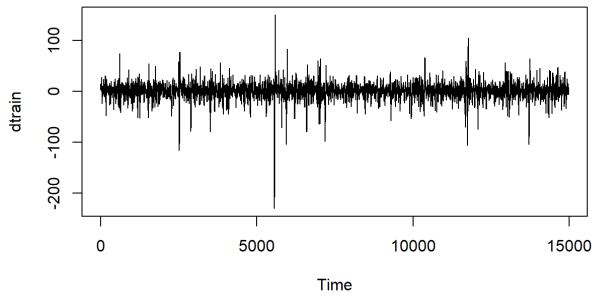


(b) PACF

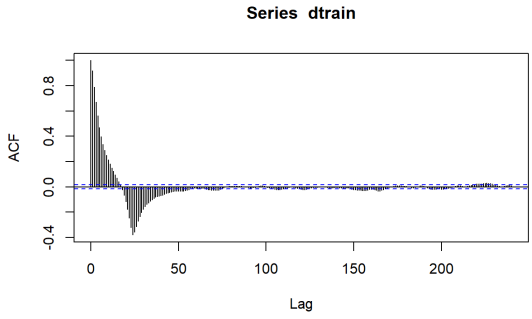
Figure 2: ACF and PACF Plot

3 Build ARIMA Models

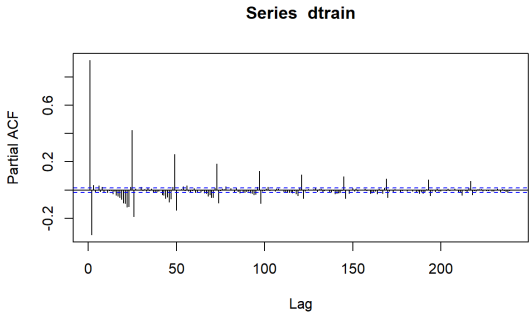
3.1 Model Selection



(a) Differenced Time Series



(b) Differenced ACF



(c) Differenced PACF

Figure 3: Seasonal Differencing

According to the pre-analysis in Section 2, we first adopt a 24-lag seasonal differencing on the time series. As can be seen in the ACF in Figure 3, there exists seasonal cut-off, which suggests seasonal $MA(1)$; and the non-seasonal cut-off in the PACF suggests non-seasonal $AR(1)$. Based on those characteristics, our candidate models revolve around $ARIMA(1,0,q)(P,1,1)_{24}$

We choose our model based on AICc criterion, since we want to prioritize fitting, rather than control the complexity of the model.

(p,d,q) \ (P,D,Q)	(2,1,1)	(3,1,1)	(3,1,2)
(1,0,2)	5.8601	5.8597	5.8596
(1,0,3)	5.8584	5.8580	5.8579
(2,0,3)	5.8580	NA	NA
(2,1,3)	5.8553	5.8552	5.8554

Table 1: AICc of Different Models

In Table 1, "NA" means the convergence fails even if outliers are dropped, so we try a higher differencing order. As the number of parameter increases, the optimization process gets harder to converge. Based on the AICc, we finally choose $ARIMA(2,1,3)(3,1,1)_{24}$.

3.2 Diagnostics

The fitted model is $(1-1.36B+.47B^2)(1-.01B^{24}-.03B^{48}-.02B^{72})\nabla\nabla_{24}X_t = (1-1.04B+.03B^2+.04B^3)(1+.89B^{24})W_t$, and $\sigma_W^2 = 20.36$. The fitting RMSE is 4.51. As can be seen in Figure 4, none of the residual tests is sufficed. There are still some significant auto-correlation at lag 23, 25, 47, 49. The distribution of residuals is more heavy-tailed than normal. The Ljung-Box test also suggests the auto-correlation remains between residuals. Therefore, the model fail to capture all the information in the time series.

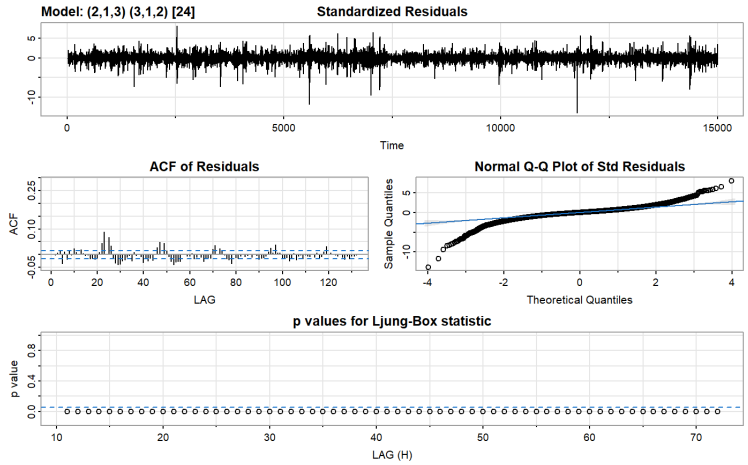


Figure 4: Diagnostics of the Residual

There are some possible explanation for the failure of ARIMA:

1. Due to frequent geomagnetic storms, the geomagnetic data may present strong local disturbance, which affects the fitting of ARIMA model. Introducing external regressors like solar activity index may help.

2. The seasonal pattern does not only exists in a 24-lag. A 700-lag seasonal pattern exists as well, and traditional seasonal ARIMA model may fail to utilize that information. Further more, the seasonal lag may be nonuniform.

3.3 Forecast

The forecasting result seems plausible. However, as the lag increases, the bias in the constant term is accumulated, and the long-term forecasting will thus get astray. The overall RMSE on test set is 73.85.

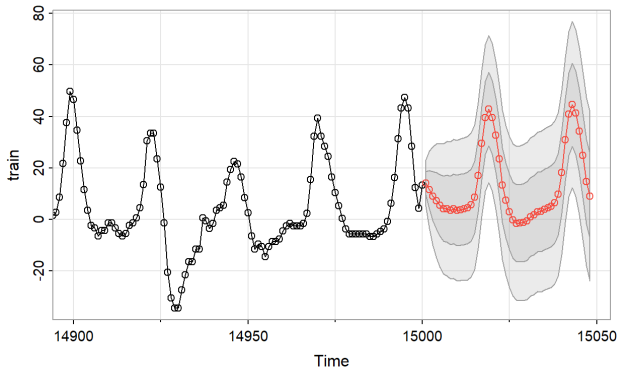
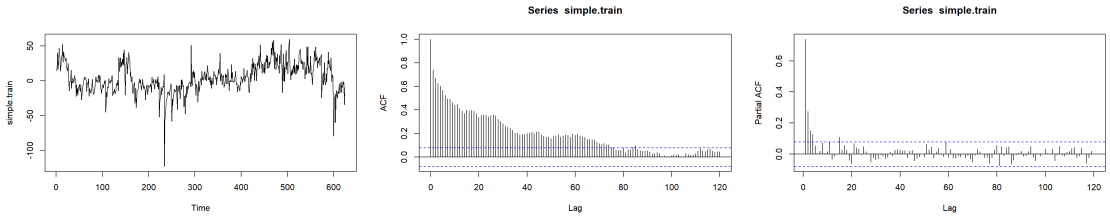


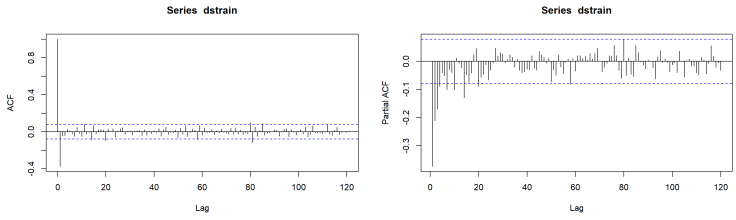
Figure 5: Local Forecasting Result

3.4 A Simplified Task

We choose the geomagnetic at 1 AM each day, and try to fit an $ARIMA$ model on a daily basis. The simplified train set has 624 observations and test set has 97. The plots of the time series are shown in Figure 6. With similar process we choose $ARIMA(3,1,4)$ with minimum AICc 8.01. The fitted model is $(1-.34B+.53B^2-.74B^3)\nabla X_t = (1-.87B+.65B^2-1.10B^3+.41B^4)W_t - .05$. This time, the residual ACF is only slightly significant in two points, the distribution is closer to normal and Ljung-Box test cannot reject the null hypothesis. Although the forecasted lacks permutation, the overall trend is plausible. The RMSE on the test set is 16.91.

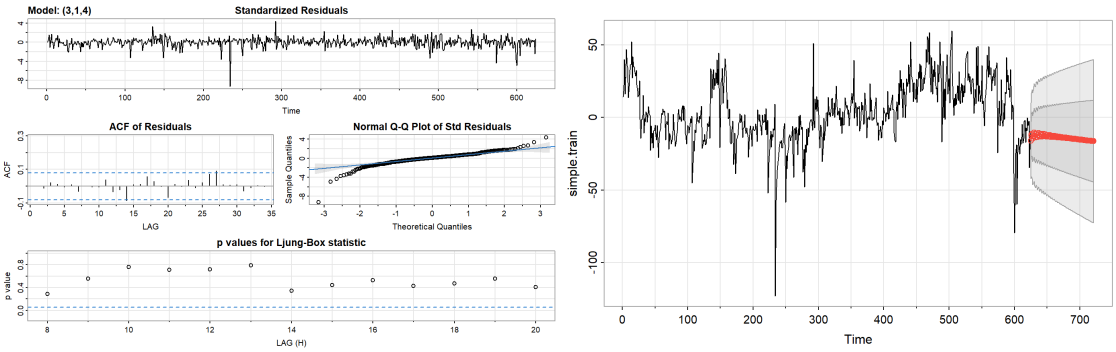


(a) Simplified (b) ACF (c) PACF



(d) Differenced ACF (e) Differenced PACF

Figure 6: Plots of the Simplified Time Series

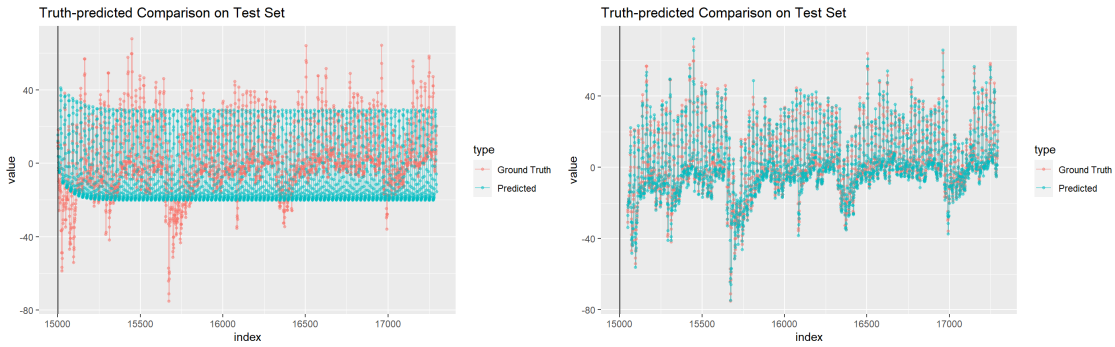


(a) Diagnostics (b) Forecast

Figure 7: Diagnostics and Forecast

4 Machine Learning Approach

We try LSTM for forecasting, using the previous 48 observations (2 days) as uni-feature input sequence to predict the next value. The number of hidden units is 24, the optimizer is Adam, and the batch size is 36. After 40 epochs of training, the fitting RMSE is 4.37. As Figure 8 shows, if we let the model forecast only based on the train data, then the forecasting result will be periodic and the test RMSE is 26.26. If we give the ground truth of the previous 48 lapses, the one-step forecasting performs well and the test RMSE is 3.66.



(a) Multi-step Prediction (b) Single-step Prediction

Figure 8: Predicted-Truth Comparison

5 Conclusion and Discussion

During this project, we try to build a seasonal ARIMA model to fit the geomagnetic data. Although the complexity and disturbance of the original data fails ARIMA, we simplify the data and ARIMA fits the new data well. During the analysis we suggest an idea of a multi-seasonal adaptation of ARIMA, which may handle this kind of data which have mixed daily and monthly periods. Generally speaking, LSTM can fit better, but the result is incompatible for further statistical analysis.