Michael Tran
Stats 153 - Fall 2015
Instructor: David Brillinger

Modeling the Strength of Earth's Magnetic Fields

1. Introduction

The scientific question motivating my work is whether there exists a time series model that can model the strength of Earth's magnetic fields with a mean absolute percentage error (MAPE) less than 25%. The magnetic fields have been very important in terms of helping human beings navigate around the globe, protecting Earth from solar winds that would otherwise result in the depletion of Earth's atmosphere, and guiding animals during migration seasons. These fields originate from a geo-dynamo process takes place between Earth's outer and inner core [1]. This process creates electric currents, which in turn produce the magnetic fields. In addition, there are records that indicate that the North Magnetic Pole and South Magnetic Pole have switched places in the past with the weakening of the field's strength as a precursor. Thus, finding a time series model that can model the strength of the magnetic fields may also help in predicting when the next reversal will happen.

2. Data

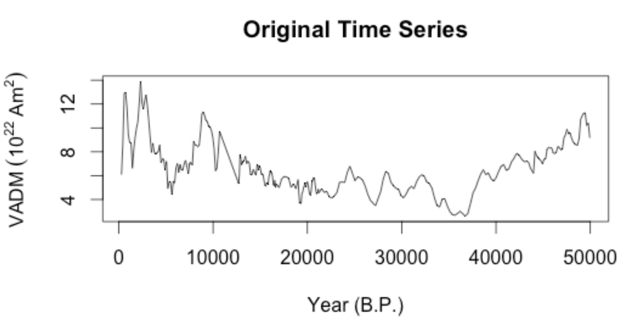
The original data comes from GEOMAGIA, which can be found at http://geomagia.gfz-potsdam.de/index.php [2]. The website contains databases, which store records of magnetic intensity measurements along with the location of the measurement site, descriptions of the samples taken, and methods used to determine the sample's age and magnetic intensity. When rocks form, cool down, and solidify, there are iron-bearing minerals that store the strength and direction of the local magnetic field and remain inside the rock [3]. Virtual axial dipole moment

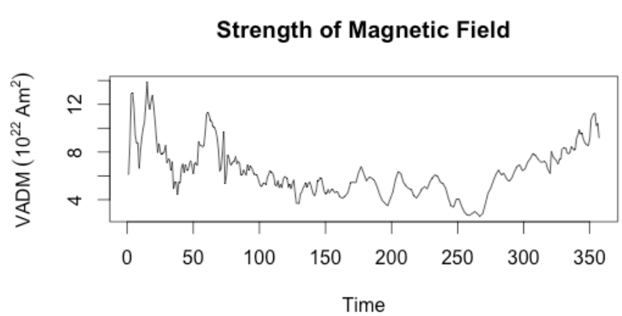
(VADM) is a measurement of that strength with units in Ampere meter squared (Am^2). There are two main databases on the site; one contains records of sediment data and the other contains volcanic samples. For now, I will be looking at the sediment samples. The data was narrowed down and the samples will come from a fixed site, which is located at 52.4616˚ Latitude and 106.1397˚ Longitude, which point to Lake Baikal, a freshwater lake in Eastern Russia.

Before any data analysis can be done, one problem with the current data is that it is not equidistant since there are only have 276 data points starting from 288 B.P. and ending at 49,946 B.P. There were data points that were spaced 139 years apart from each other and others spaced 277 years apart from each other. I decided to create NA values right in the middle of data points that were spaced 277 years apart to create an equispaced time series. In order to fill the gaps, I used R’s na.approx function, which uses linear interpolation in order to replace the NA values. After the interpolation process, the new data now has 357 data points, with each point being 139 years apart from each other, starting from 288 B.P. and ending at 49,946 B.P. Finally, I looked at other geophysical databases to see if there were any new sediment data from Lake Baikal. Currently, there are no online sediment records that date from the past 500 years.

3. Graphing the Data & Detrending

Below are the plots of the original time series and the new, interpolated time series.



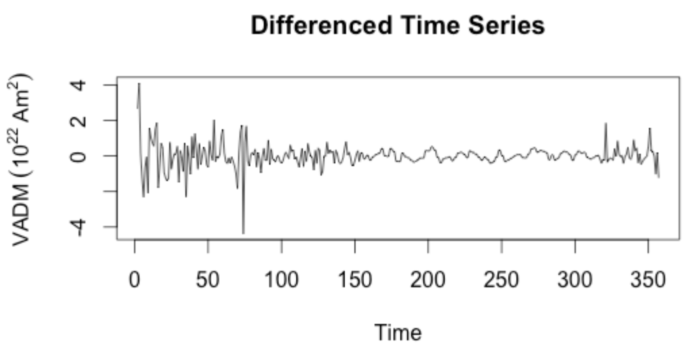


It appears that the time series isn't stationary because there does not seem to be a constant mean. In order to determine whether or not it is necessary to difference the time series, we can look at the Augmented Dickey-Fuller test. The test has a null hypothesis that the given time series has a unit root and an alternative hypothesis that the time series is stationary. The results came out as followed.

Augmented Dickey-Fuller Test

data: CPTS
Dickey-Fuller = -1.9241, Lag order = 7, p-value = 0.6089
alternative hypothesis: stationary

With a p-value greater than 0.05, we can not assume the time series is stationary. Thus, I differenced the time series and below is its plot and Dickey-Fuller test results.



Augmented Dickey-Fuller Test

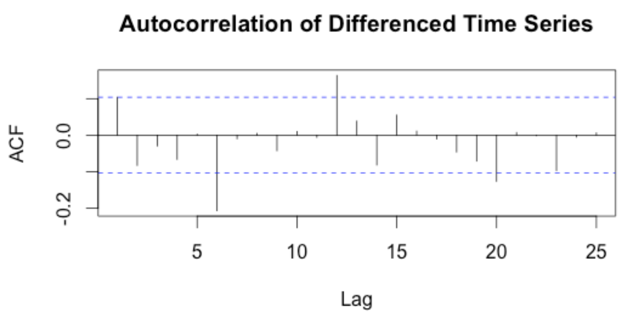
data: dCPTS
Dickey-Fuller = -7.8908, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary

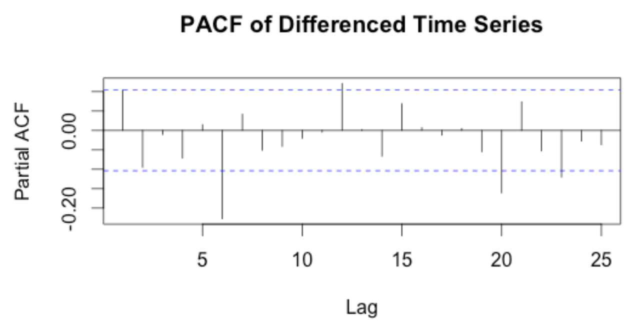
Warning message:
In adf.test(x = dCPTS, alternative = "stationary") :
  p-value smaller than printed p-value

With a p-value less than 0.01, the once differenced time series is now stationary.

4. ARIMA Modeling

Since the differenced time series is stationary, I will work with the differenced time series henceforth. To determine p and q, lets look at the autocorrelation and partial autocorrelation plot.

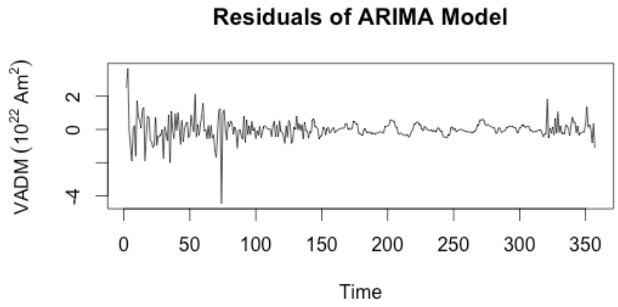


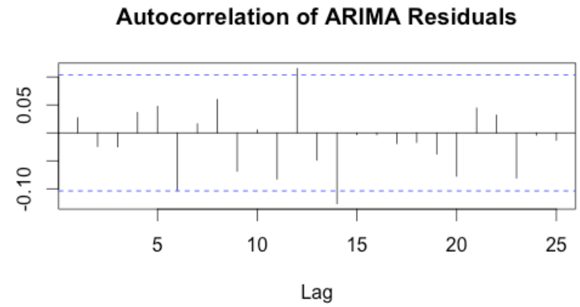


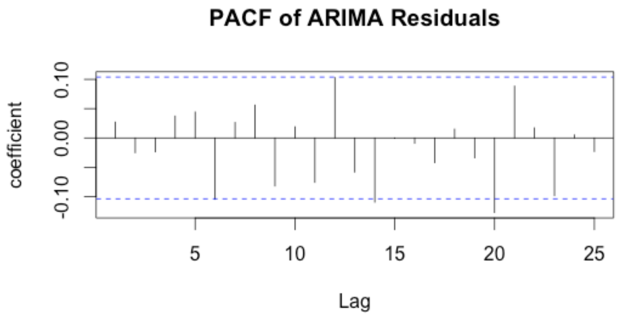
It looks like at the very beginning, the lags are below the significance line, with a few significant lags occurring afterwards. In both plots, a cut-off happens after lag two. To ensure those are the right values for both p and q, lets look at another method. I will search through every ARIMA(p, 0, q) model, p and q both ranging from 0 to 5, and select the model with the lowest AIC. Below is a table of each ARIMA model and its AIC.

Order	AR(0)	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
MA(0)	698.4254	695.4948	694.4360	696.4303	691.4391	692.6632
MA(1)	696.3231	688.6863	689.3531	691.0593	691.2258	692.0255
MA(2)	695.0670	689.9966	685.8938	684.4329	686.2957	689.9593
MA(3)	696.9537	691.4956	684.3154	**674.6875**	676.0000	677.8935
MA(4)	696.6796	693.0023	686.3065	676.0320	675.6130	679.0858
MA(5)	698.6146	695.8215	688.3018	677.8786	683.1022	683.9055

Based on the AIC of each model, ARIMA(3, 0, 3) came out as the best model. With ARIMA(3, 0, 3) as the time-side model, lets look at its residuals plot, its autocorrelation and partial autocorrelation plot, the Ljung-Box test on its residuals, and its statistical estimates.







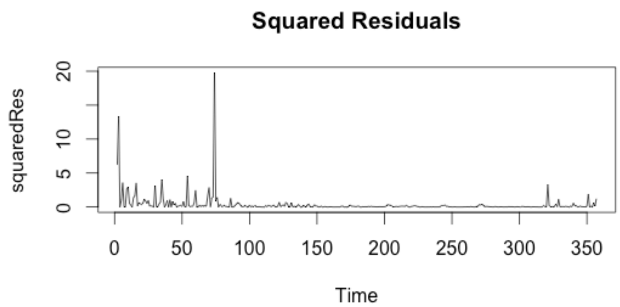

```
Coefficients:
         ar1     ar2      ar3      ma1      ma2     ma3  intercept
      0.9178  0.6190  -0.8564  -0.8641  -0.8104  0.9573     0.0089
s.e.  0.0470  0.0808   0.0473   0.0325   0.0435  0.0297     0.0285
```

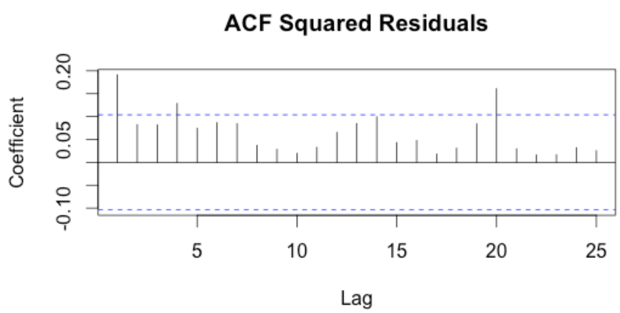
Box-Ljung test
data: resid(ArcheoFit)
X-squared = 0.26435, df = 1, p-value = 0.6071

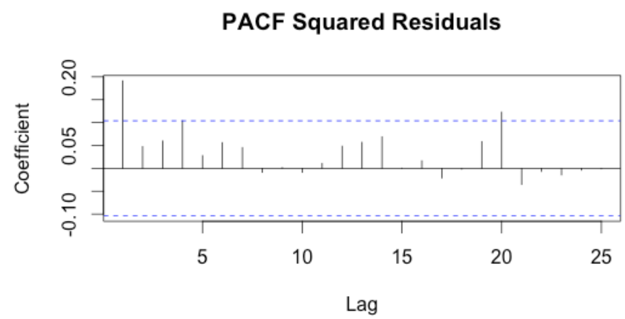
Looking at the ARIMA residuals plot, it appears the residuals seem to have shocks occurring at different times. Furthermore, there are very few significant lags in both the autocorrelation and partial autocorrelation plot. In addition, the Ljung-Box test has a null hypothesis that the residuals are independently distributed and uncorrelated, with the alternative hypothesis saying the residuals are correlated. With a p-value of 0.6071, we accept the null hypothesis. Finally, the bottom right plot contains the estimates and standard errors of each term.

5. GARCH Modeling

Although there were very few significant lags in the autocorrelation and partial autocorrelation plot of the residuals, there were clusters of volatility in the residuals plot. Thus, this might present an opportunity to model the volatility of the series with an ARCH or GARCH model. To check if there is volatility, let's look at the squared residuals plot and its corresponding autocorrelation and partial autocorrelation plot.





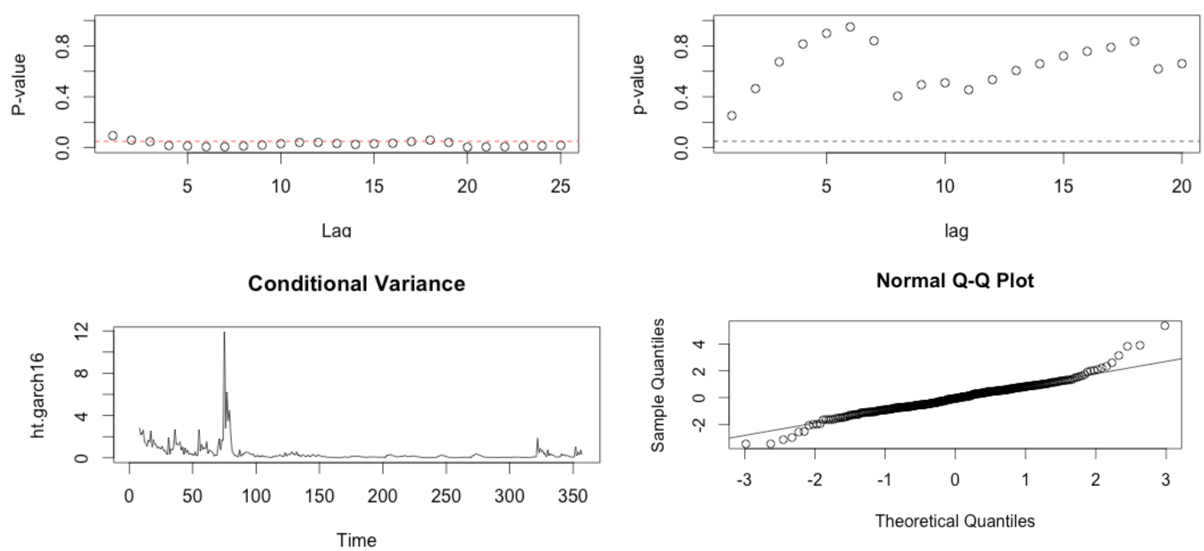


The squared residuals plot shows patterns of volatility that can be modeled. In addition, the autocorrelation and partial autocorrelation plot die off around lag seven. So, I'll start with an ARCH model and check every ARCH(q) model, q ranging from 1 to 10, and see which one has the lowest AIC. Below to the left is the corresponding table.

Model	AIC
ARCH(1)	511.0652
ARCH(2)	491.3146
ARCH(3)	430.6950
ARCH(4)	415.1476
ARCH(5)	389.8900
ARCH(6)	**386.7644**
ARCH(7)	405.2139
ARCH(8)	399.6839
ARCH(9)	409.7515
ARCH(10)	406.4659

Model	AIC
GARCH(1, 6)	**390.3949**
GARCH(2, 6)	396.3631
GARCH(3, 6)	391.2402
GARCH(4, 6)	396.6634
GARCH(5, 6)	397.3488
GARCH(6, 6)	398.8567
GARCH(7, 6)	403.8308
GARCH(8, 6)	401.6318
GARCH(9, 6)	402.4390
GARCH(10, 6)	402.3920

The ARCH(6) model came out with the lowest AIC. Now, I will look at every GARCH model (p, 6), p ranging from 1 to 10, and select the GARCH model with the lowest AIC. GARCH(1, 6) came out as the best model. To ensure this is a good model, lets look at the McLeod-Li test, the Generalized Portmanteau test, a plot of the predicted conditional variance, and a Q-Q plot of residuals.



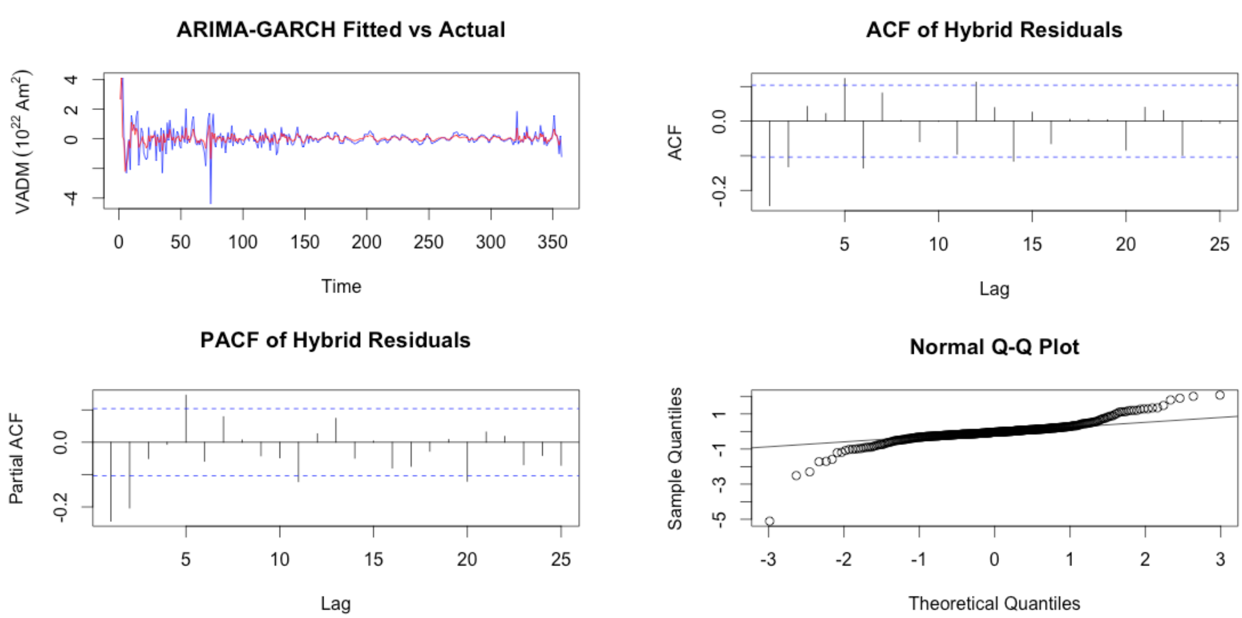
On the upper left is a plot of the McLeod-Li test. The tests are all significant when the number of lags of the autocorrelations of the squared residuals is greater than one, which points to strong evidence of conditional heteroscedasticity. The upper right is a plot of the Generalized Portmanteau Test on the standardized squared residuals, which shows that all p-values are higher than 0.05, suggesting that the squared residuals are uncorrelated over time, and hence the standardized residuals are independent. On the lower left is a plot of the predicted conditional variances from the GARCH(1, 6) model. The plot captures for the most part the same volatility as shown in the squared residuals plot. The lower right is a Q-Q plot of the model's residuals.

The plot shows a largely straight-line pattern with long tails on both ends. Thus, GARCH(1, 6) looks to be a good fit and the hybrid model will be an ARIMA(3, 0, 3) + GARCH(1, 6). Right below is a table of the GARCH model's estimates and uncertainties for its coefficients.

	a0	a1	a2	a3	a4	a5	a6	b1
Estimate	1.082E-02	5.445E-01	9.200E-16	2.359E-01	1.064E-01	1.665E-01	5.939E-02	1.361E-01
Std. Error	6.490E-03	1.042E-01	2.426E-01	4.377E-02	1.102E-01	9.978E-02	9.229E-02	3.376E-01

6. Hybrid Analysis

With the hybrid model being ARIMA(3, 0, 3) + GARCH(1, 6), I used fGarch's garchFit() function to create the hybrid model. Below are plots of the fitted vs actual values, the autocorrelation and partial autocorrelation plot, and a Q-Q plot of the hybrid's residuals.



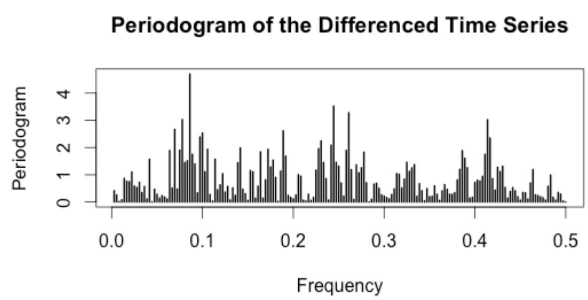
With red as the fitted values and blue as the actual values, the hybrid model looks like it was able to capture some volatile periods to a certain degree. The autocorrelation and partial autocorrelation plot both cut off after lag two, with few significant lags afterwards; thus, exhibiting no serial dependence. Finally, the Q-Q plot of the hybrid residuals shows much longer tails than the GARCH Q-Q plot, so normality can't be assumed for all intents and purposes. Overall, the hybrid model fits the magnetic field dataset well and below are the estimates.

Error Analysis:

	Estimate	Std. Error		Estimate	Std. Error
mu	1.477e-02	1.035e-02	alpha1	3.537e-01	1.334e-01
ar1	7.987e-01	1.792e-01	alpha2	1.000e-08	NA
ar2	2.124e-01	1.516e-01	alpha3	4.496e-01	1.406e-01
ar3	-4.198e-01	7.763e-02	alpha4	1.077e-01	1.519e-01
ma1	-4.240e-01	1.814e-01	alpha5	8.827e-02	6.333e-02
ma2	-3.502e-01	1.041e-01	alpha6	1.108e-01	8.147e-02
ma3	3.922e-01	7.778e-02	beta1	1.573e-01	NA
omega	1.087e-02	2.770e-03			

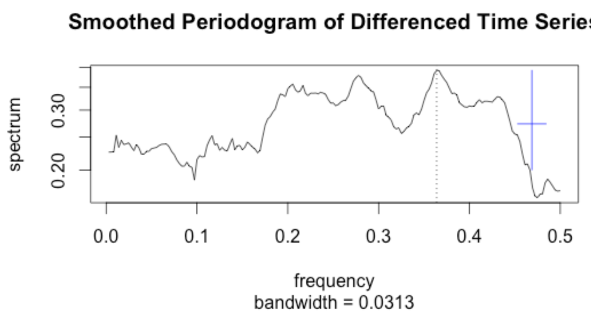
7. Frequency-Side Analysis

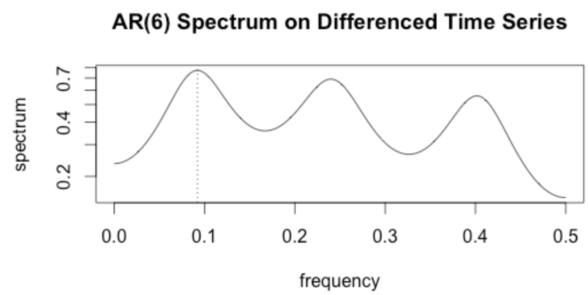
To start off the frequency-side analysis, lets take a look at the raw periodogram.



There are 3 dominant peaks around the frequencies 0.90, 0.25, and 0.41. The maximum peak has a spectrum value of 4.7, which corresponds to a frequency of 0.0861. This frequency corresponds to a period of 11.61 time units, which is approximately 1614 years. Thus, there appears to be a dominant periodicity of about 1614 years in magnetic field activity. Looking at the history of Earth's magnetic field activities, this actually doesn't correspond to any magnetic field reversals, as the last one happened about 800,000 years ago [4]. This periodicity most likely refers to the movement of the magnetic field and may have passed through Lake Baikal. Now, let's look at a smoothed periodogram using a Daniell Window where m will be 19, which is an approximation of the square root of the number of observations in the time series.

In the smoothed version, the dominant peak occurs when the frequency is 0.364. This frequency corresponds to a period of 2.75 time units, which is equivalent to 381 years. Alternatively, we can find the best fitting AR model for the series and plot the spectral density of that function. Using R's spec.ar() function with AIC as the criteria in model selection, AR(6) came out to be the best fit. Below to the right is the plot of its spectral density.





In this spectral density, the appearance looks similar to the raw periodogram, with about three dominant peaks occurring at relatively the same frequencies. The max spectrum value occurs at the frequency value of 0.092, which is equivalent to approximately 1507 years.

8. Conclusion and Inferences

The final model with respect to the differenced time series is ARIMA(3, 0, 3): $Y_t = 0.01477 + 0.7987Y_{t-1} + 0.2124Y_{t-2} - 0.4198Y_{t-3} + \varepsilon_t + 0.424\varepsilon_{t-1} + 0.3502\varepsilon_{t-2} - 0.3922\varepsilon_{t-3}$ and

GARCH(1, 6): $\sigma_t^2 = 0.01087 + 0.3537\varepsilon_{t-1}^2 + 1e\text{-}08\varepsilon_{t-2}^2 + 0.4496\varepsilon_{t-3}^2 + 0.1077\varepsilon_{t-4}^2 + 0.08827\varepsilon_{t-5}^2 + 0.1108\varepsilon_{t-6}^2 + 0.1573\sigma_{t-1}^2$

Using the accuracy function, MAPE came out to be 173.17%. Thus, the answer to my question is there does not exist a time series model with a MAPE less than 25%. If I were to redo this research process, the next iteration would account for the movement of the magnetic fields, include data that shows a reversal of the magnetic fields, and incorporate other variables from the data, such as declination, inclination, and dating methods used on the sediment rock samples.

9. Citations & References

[1] "What causes the magnetic field to reverse its polarity?" USGS FAQs. United States Geological Survey. Web. 11 Dec. 2015. < http://www.usgs.gov/faq/categories/9782/2741>

[2] Brown, M.C., F. Donadini, A. Nilsson, S. Panovska, U. Frank, K. Korhonen, M. Schuberth, M. Korte and C.G. Constable, GEOMAGIA50.v3 : 2. A new paleomagnetic database for lake and marine sediments, Earth Planets Space 67:70, doi:10.1186/s40623-015-0233-z.

[3] "Magnetism in Rocks." University of Texas, Dallas. Web. 11 Dec. 2015. < http://www.utdallas.edu/~pujana/oceans/mag.html>

[4] "2012: Magnetic Pole Reversal Happens All The (Geologic) Time" Topics of Earth. NASA. Web. 11 Dec. 2015. < http://www.nasa.gov/topics/earth/features/2012-poleReversal.html>

10. Code

```
library("TSA")
library("forecast")
library("tseries")
library("fGarch")
set.seed(8639)
# Data Processing for CPData
CPData = read.csv(file.choose(), header = TRUE)
baiIndex = which(CPData$LocationCode == "BAI")
non_outliers = c()
for (i in 1:289) {
  if (CPData$VDM_VADM[baiIndex[i]] < 200) {
    non_outliers = c(non_outliers, baiIndex[i])
  }
}
timeData = CPData$Age.yr.BP.[non_outliers]
teslaData = CPData$VDM_VADM[non_outliers]
op <- par(mar = c(5,4.5,4,2) + 0.1)
```

```
plot(x = timeData, y = teslaData, type = "l", main = "Original Time Series",
     xlab = "Year (B.P.)", ylab = "VADM" ~ (10^{22} ~ Am^{2}))
par(op)
# 154 is start, 233 is end
# Understand the missing values
timeT = rep(NA, length(timeData) + 81)
timeT[1:153] = teslaData[1:153]
p = 155
for (j in 154:233) {
  timeT[p] = teslaData[j]
  p = p + 2
}
timeT[315:357] = teslaData[234:276]
interpolatedTesla = na.approx(timeT, na.rm = F, rule = 2)

# TS Creation
CPTS = ts(data = interpolatedTesla)
op <- par(mar = c(5,4.5,4,2) + 0.1)
plot(CPTS,  ylab = "VADM" ~ (10^{22} ~ Am^{2}), main = "Strength of Magnetic Field")
par(op)
# Augmented Dickey Fuller
adf.test(x = CPTS, alternative = "stationary")
# Since p-value is 0.6089, we take a difference to make it stationary.
dCPTS = diff(CPTS)
op <- par(mar = c(5,4.5,4,2) + 0.1)
plot(dCPTS,  ylab = "VADM" ~ (10^{22} ~ Am^{2}), main = "Differenced Time Series")
par(op)
adf.test(x = dCPTS, alternative = "stationary")
# P-value is now less than 0.01, thus stationary
# Thus, d is 1

# p, d, and q analysis
acf(dCPTS, main = "Autocorrelation of Differenced Time Series")
pacf(dCPTS, main = "PACF of Differenced Time Series")

AICMatrix = matrix(data = 0, nrow = 6, ncol = 6)
colnames(AICMatrix) = c("AR(0)","AR(1)","AR(2)","AR(3)","AR(4)","AR(5)")
rownames(AICMatrix) = c("MA(0)","MA(1)","MA(2)","MA(3)","MA(4)","MA(5)")
for (k in 0:5) {
  for (p in 0:5) {
    CPFit = Arima(dCPTS, order = c(k, 0, p), method = "CSS-ML")
      AICMatrix[k + 1, p + 1] = AIC(CPFit)
  }
```

```
}
# Based on AIC criterion, final model should be ARIMA(3, 1, 3) with respect to CPTS

ArcheoFit = Arima(dCPTS, order = c(3, 0, 3), method = "CSS-ML")
op <- par(mar = c(5,4.5,4,2) + 0.1)
plot(resid(ArcheoFit), main = "Residuals of ARIMA Model",  ylab = "VADM" ~ (10^{22} ~
Am^{2}))
par(op)
acf(resid(ArcheoFit), main = "Autocorrelation of ARIMA Residuals", ylab = "coefficient")
pacf(resid(ArcheoFit), main = "PACF of ARIMA Residuals", ylab = "coefficient")
Box.test(x = resid(ArcheoFit), type = "Ljung-Box")
# The p-value came out to be 0.6071, so accept the null that the residuals are
# independently distributed and uncorrelated.

# ARCH / GARCH
res.ArcheoFit = ArcheoFit$residuals
squaredRes = res.ArcheoFit^2
plot(squaredRes, main = "Squared Residuals")
acf.SquaredRes = acf(squaredRes, main = "ACF Squared Residuals", ylab = "Coefficient")
pacf.SquaredRes = pacf(squaredRes, main = "PACF Squared Residuals", ylab = "Coefficient")

# AIC and Log Liklihood
listOfAIC = c()
listofLL = c()
for (i in 1:10) {
  arch = garch(res.ArcheoFit, order = c(0, i), trace = F)
  listOfAIC = c(listOfAIC, AIC(arch))
  listofLL = c(listofLL, logLik(arch))
}
# ARCH(6) has lowest AIC
listOfGAIC = c()
listofGLL = c()
for (k in 1:10) {
  garchk6 = garch(res.ArcheoFit, order = c(k, 6), trace = F)
  listOfGAIC = c(listOfGAIC, AIC(garchk6))
  listofGLL = c(listofGLL, logLik(garchk6))
}
# GARCH(1, 6) has lowest AIC
garchModel = garch(res.ArcheoFit, order = c(1, 6), trace = F)
#arch = garch(res.ArcheoFit, order = c(0, 6), trace = F)
gBox(garchModel, lags = 1:20, x = res.ArcheoFit, method = "squared")
# Test if residuals are independent, uncorrelated
summary(garchModel)
```

```
McLeod.Li.test(object = ArcheoFit, y = dCPTS) # Test for conditional heteroskedasity
# Evidence of strong ARCH effects after lag 1. The ACF and PACF confirm it

ht.garch16 = garchModel$fitted.values[,1]^2
plot(ht.garch16, main = "Conditional Variance")
plot(garchModel$residuals)
qqnorm(garchModel$residuals)
qqline(garchModel$residuals)

# Hybrid Analysis
hybridModel = garchFit(formula = ~arma(3, 3) + garch(6, 1), data = dCPTS, trace = F)

fit31316 = hybridModel@fitted
op <- par(mar = c(5,4.5,4,2) + 0.1)
plot(dCPTS, col = "blue", main = "ARIMA-GARCH Fitted vs Actual",  ylab = "VADM" ~
(10^{22} ~ Am^{2}))
par(op)
lines(fit31316, col = "red")

hybridRes = hybridModel@residuals
acf(hybridRes, main = "ACF of Hybrid Residuals")
pacf(hybridRes, main = "PACF of Hybrid Residuals")
qqnorm(hybridRes)
qqline(hybridRes)

# Spectrogram
# Cross Correlation for series relation
x = periodogram(dCPTS, main = "Periodogram of the Differenced Time Series")
maxIndex = which(x$spec == max(x$spec))
freq = x$freq[maxIndex]

k = kernel("daniell", 19) # Square root of number of obs
alpha = spec.pgram(x = dCPTS, k, detrend = F, plot = F)
plot(alpha, main = "Smoothed Periodogram of Differenced Time Series")
maxSmooth = which(alpha$spec == max(alpha$spec))
smoothFreq = alpha$freq[maxSmooth]
abline(v = smoothFreq, lty = "dotted")

specValues = spec.ar(x = dCPTS, log = "no", plot = F)
plot(specValues, main = "AR(6) Spectrum on Differenced Time Series")
maxSpec = which(specValues$spec == max(specValues$spec))
topFreq = specValues$freq[maxSpec]
abline(v = topFreq, lty = "dotted")
```

```
# Conclusion
accuracy(fit31316, dCPTS)[5]
```