Travis Klipp

Pstat 174 W 10AM

Earthquake Time Series Analysis Project

Introduction

The U.S. Geological Survey estimates millions of earthquakes occur every year, but only a fraction (about 20,000 out of millions) of them are located due to the magnitude of the earthquake or the remote location (smaller and isolated earthquakes are much harder to locate). There are several different scales to classify earthquakes, including using the magnitude scale (which is the measure of the amplitude of seismic waves using the Richter Scale). Major earthquakes are usually classified as having a magnitude of 7.0-7.9 and great earthquakes are classified as having a magnitude of 8.0 or greater. It's been estimated that on average there are 18 major earthquakes and 1 great earthquake throughout the world every year.

The data set used in this project displays the frequency of major earthquakes annually with a 7.0 magnitude or greater during the period from 1900-1998. The data was collected from the National Earthquake Information Center (in which they note that different lists may give different values of magnitudes depending on the formulas used). For this project, the data was obtained from Rob Hyndman's Time Series Data Library (*Hyndman, R.J.* Time Series Data Library*,* https://datamarket.com/data/list/?q=provider:tsdl (citing: National Earthquake Information Center), *Accessed on 12/2/1015*.)

Aim of Modeling

The purpose of modeling this data set will be to predict the possible frequency of earthquakes greater than 7.0 10 years in the future (relative to 1998). We will see a comparison of the real data from years 1999-2008, and conclude whether the model was an accurate fit.

To analyze if the forecast (prediction) of the frequencies of future years was accurate, we will also:

1) Determine if the original time series is stationary using ACFs, PACFs, and general observation
2) Determine whether there's any seasonality, trend, random walk processes
3) Determine whether transformations are necessary
4) Transform and possibly difference the data if applicable
5) Examine several different potential models that may fit the data
6) Find a final fit, estimate coefficients, confidence intervals, diagnostic checking
7) Forecast/prediction intervals of major earthquakes 10 years into the future
8) Compare to real-life observations

Original Time Series Stationarity

A time series is stationary if the mean value function is constant and does not depend on time. We may also determine if a time series is stationary if the variances are also constant over time. If a time series is stationary, then the Autocovariance (ACVF) function is $\gamma X(h) = Cov(Xt+h, Xt)$ at lag h, meaning h being the lag point, and the autocovariance simply being the variance at lag h=0. The Autocorrelation (ACF) function for a stationary time series is $\rho X(h) = Corr(Xt+h, Xt)$. The ACF and PACF (partial autocorrelation) functions are used to determine if a time series is stationary. For ARMA (auto-regressive moving average of order) time series models, we may potentially see exponential decay (tailing off of values) or abrupt cut off to near 0 in the PACF for stationary models. Obviously, autocorrelation values must be between -1 and 1.

Seasonality is when there is evidence of some kind of pattern for multiple periods. The time series may decrease and increase, but it follows a pattern throughout the series. Seasonality is evidence of a non-stationary model, and usually transformations and possibly differencing of the data is necessary in order to make the time series stationary.

Trend is when the time series has either a linear upward or downward tendency. Trend is also evidence of a non-stationary model, and differencing is usually needed to produce a stationary model.

Observations of Original Time Series Data

Figure 1 shows the plot of the original time series of earthquakes. Upon first glance, stationarity seems to be in the grey area. It possibly looks like the mean may not be constant, and there is possibly some evidence of a downward trend. We can take a look at the ACF and PACF of the time series to determine if there is stationarity. We also see that the data isn't exactly normally distributed, so a transformation may be needed (**Figure 3**).

Judging from the ACF and PACF in **Figure 2AB**, there is evidence that leads us to believe that this time series may already be stationary, however it is curious that there are a handful of lags after lag 40 that aren't zero. For a stationary ACF, we would like the ACF to have values at 0 or exponentially decreasing, so maybe we will apply a first order difference and compare results.

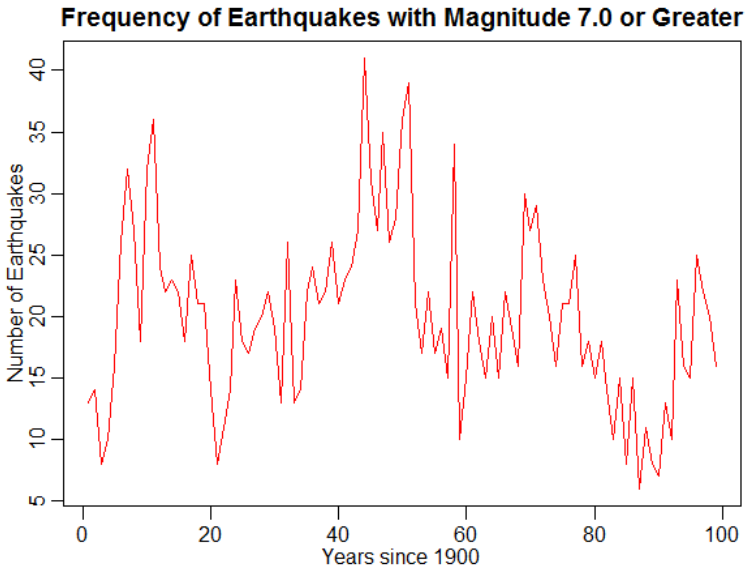

Figure 1: Plot of major earthquake time series from 1900-1998

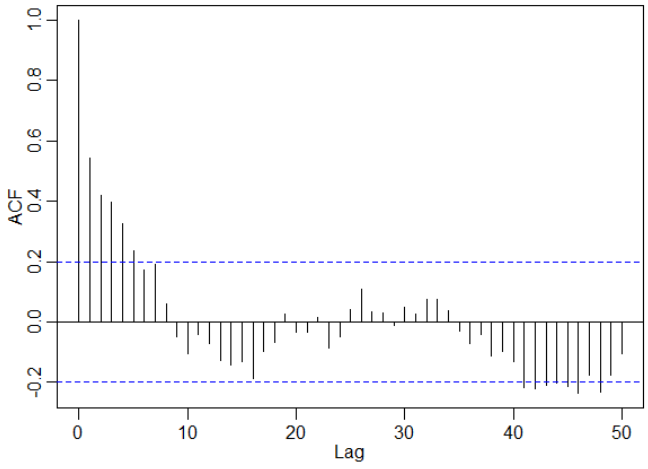

Figure 2A: ACF has lags increasing after lag 40.

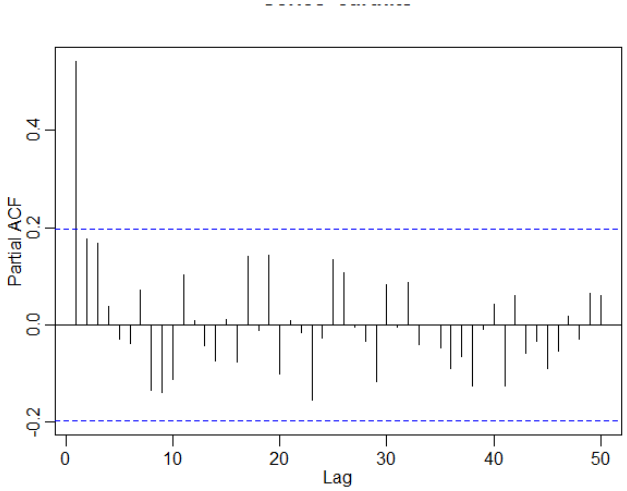

Figure 2B: PACF of original time series demonstrating MA model activity.

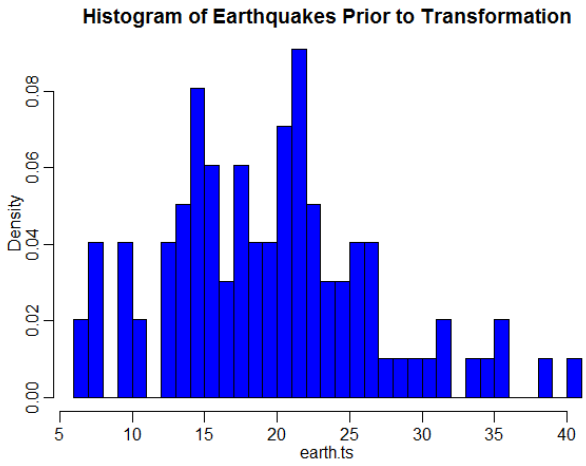

Figure 3: Distribution of data doesn't follow a normal distribution.

Potential Models

Without any transformation or differencing, we see evidence of AR (auto-regressive) and MA (moving average order) models, which are types of ARMA (mixed auto-regressive moving average order) models. An MA(1) PACF model would cut off to 0 after lag 1, which is the case here (**figure 2b**). An MA (1) process has the equation Xt = Zt + θ1Zt−1 where Zt ∼ W N(0, σ2 Z) (WN is white noise which is defined as when Xt=Zt, E(Xt)=0,Ext^2= σ2z, γZ(k) = ρZ(k) = 0, k > 0), |θ1| < 1, with ACF:

$$\rho_X(k) = \frac{\theta_k + \theta_1\theta_{k+1} + \ldots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \ldots + \theta_q^2},\ k = 1, 2, \ldots, q$$

$$\rho_X(k) = 0,\ k > q.$$

An AR(1) process exponentially decreases to zero in the ACF (although it increases again) as is the case here (**figure 2a**). An AR(1) process has the equation Xt = ϕ1Xt−1 + Zt , |ϕ1| < 1 with ACF:

$$\rho_X(k) = \phi_1^{|k|}, \forall k : |k| \geq 1 \text{ and } \sigma_X^2 = \gamma_X(0) = \frac{\sigma_Z^2}{1-\phi_1^2}.$$

Since the ACF and PACF display evidence of MA and AR models, we can guess an ARIMA (1,1) model might be a good fit for this data. An ARMA model has the equation:

Xt − ϕ1Xt−1 = Zt + θ1Zt−1, |ϕ1| < 1, |θ1| < 1 with ACF:

$$\rho_X(k) = \frac{(\phi_1+\theta_1)(1+\phi_1\theta_1)}{(1+2\phi_1\theta_1+\theta_1^2)}\phi_1^{k-1},\quad k \geq 1$$

Transformation

Transformations of data are usually needed to stabilize variance or seasonal effect. There isn't significant evidence of seasonality or periodicity, so if a transformation was needed, it would be to stabilize the variance or make the data more normally distributed in order to forecast. Figure 3 shows a histogram of the data before the transformation. Looking at the figure, you can see that the data doesn't exactly fit a normal distribution, however we can try transforming the data to make it better distributed using a Box-Cox Transformation.

A Box-Cox Transformation uses the formula:

$$f_\lambda(U_t) = \begin{cases} \ln U_t, & \text{if } U_t > 0,\ \lambda = 0; \\ \lambda^{-1}(U_t^\lambda - 1), & \text{if } U_t \geq 0,\ \lambda > 0 \end{cases}$$

for specific gamma values (usually 0, .5, or 1). Using R programming, we determine gamma to be .5 since .5 lies in the confidence interval; so we use the second equation to transform the data. The transformed time series plot doesn't significantly change our data (stabilized variance) (**figure 4**) and the histogram of the transformed data (**figure 5**) doesn't make it more normally distributed, so we continue on without a transformation.

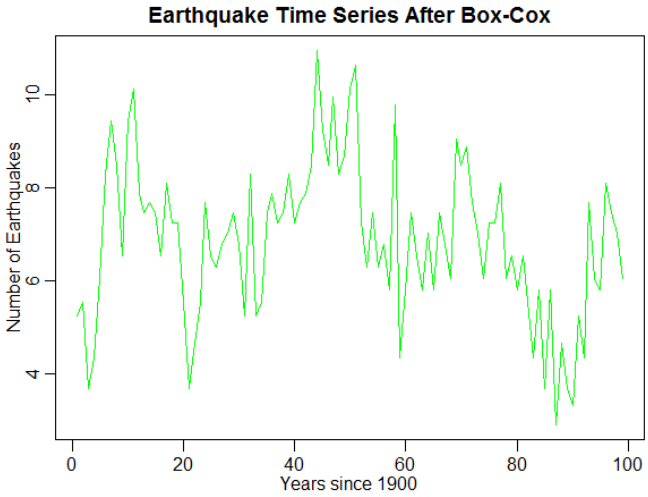

Figure 4

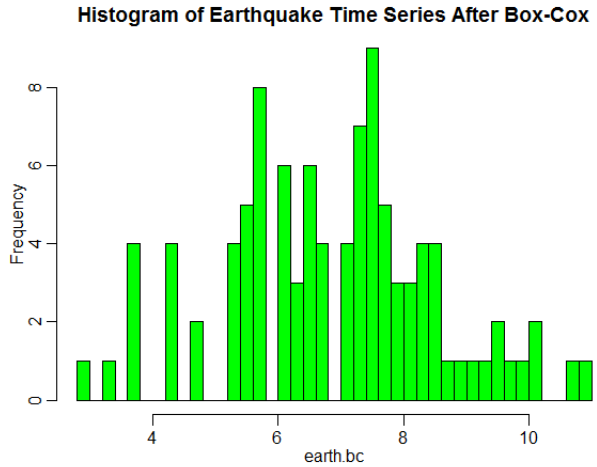

Figure 5

Differencing

Differencing is used to remove any trend or seasonality in the data. I noticed that there may be a slight downward trend in the data, and the autocorrelations of the original ACF increasing, so a first order difference was used. The time series plot of the differenced data is shown below in **figure 6**. The plot clearly shows stationarity as there's a constant mean, variance, and no trend or seasonality. **Figure 7** shows a clearly stationary ACF/PACF, and possibly suggests an MA(1) model with negative autocorrelation at lag 1, with one order of differencing ARIMA(0,1,1) or an

ARIMA(1,1,1) model (ARMA(1,1) with one order of differencing). ARMA(p,q) models are classes of ARIMA(p,d,q) models with d being the order of difference.

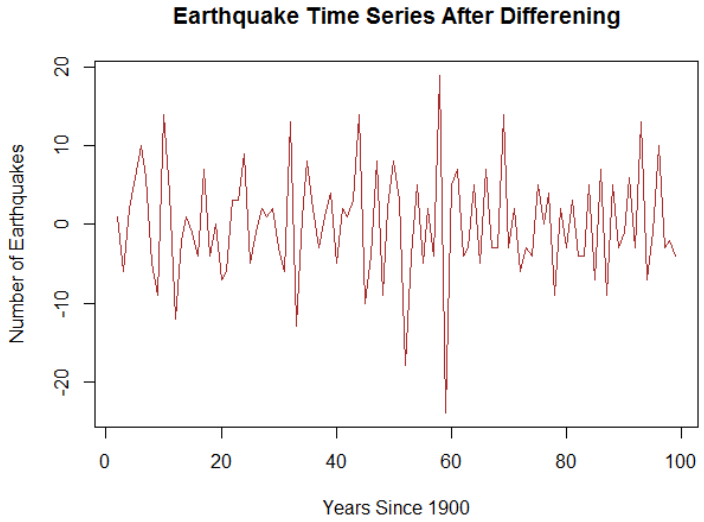

Figure 6:Time series shows clear signs of stationarity.

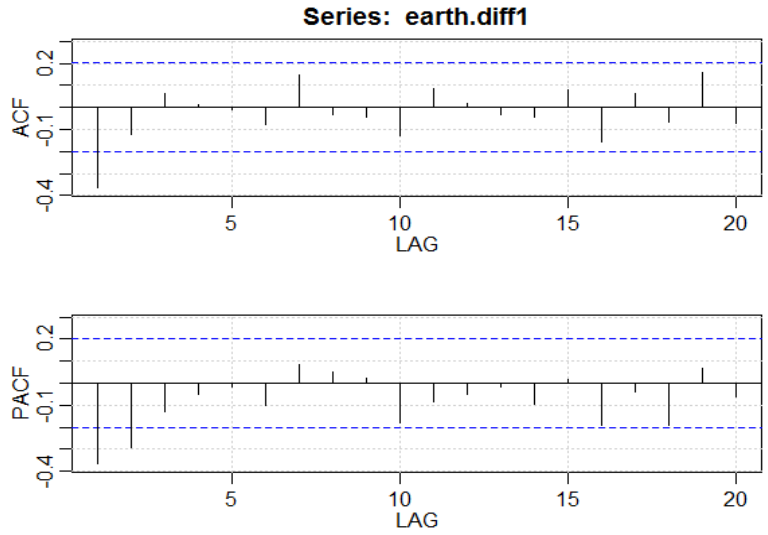

Figure 7: ACF/PACF plots show MA activity, possibly AR activity. Notice the osciliations.

Fitting a Model

We've narrowed down several possible models, including ARIMA(1,1,1)(ARMA(1,1) with one order of difference), and ARIMA(0,1,1) (MA(1) model with one order of difference).The best model fit has the lowest AIC (Akaike Information Criterion)/AICc (adjusted for bias)/BIC

(Bayesian Information Criterion). In most cases, we will choose the model that minimizes the AIC/AICc. The AICc has the following equation for following p,q values in ARMA(p,q):

$$AICC = -2\ln L(\underline{\theta_q}, \underline{\phi_p}, S(\underline{\theta_q}, \underline{\phi_p})/n) + 2(p+q+1)n/(n-p-q-2)$$

Using R Programming, we determine the AIC of the ARIMA(1,1,1) model to be 640.69 and the AIC of ARIMA(0,1,1) to be 639.06. This suggests the best model fit is the ARIMA(0,1,1) model.

Estimated Coefficients and Confidence Intervals

The estimated coefficients for the ARIMA(0,1,1) model are -.5601 with standard error .0869. We can determine if the MA coefficients are significant by calculating a confidence interval and determining if 0 lies in it. If 0 is contained in the confidence interval, we conclude the coefficients are not significant and we should possibly find a better model. Using a 95 percent confidence interval, we calculated [-.7305, -.398]. 0 does not lie in the confidence interval, so we conclude that the MA(1) coefficient (theta) is significant (-.5601).

Diagnostic Checking

We determine if a model is an accurate fit by analyzing the residuals. The residuals must not be significant for all lags (near 0), be stationary, and not be particularly skewed (has a sense of normality). **Figure 8** below shows that the residuals are stationary, near 0 for all lags, and has a sense of normality. There are a few residuals outside the line of normality so it's not perfectly normal, but there isn't a real sense of skewedness (bottom right plot). The ACF of the residuals (bottom left) show that the residuals are not significant for all lags. The time series plot (top plot) shows that the residuals are indeed stationary.

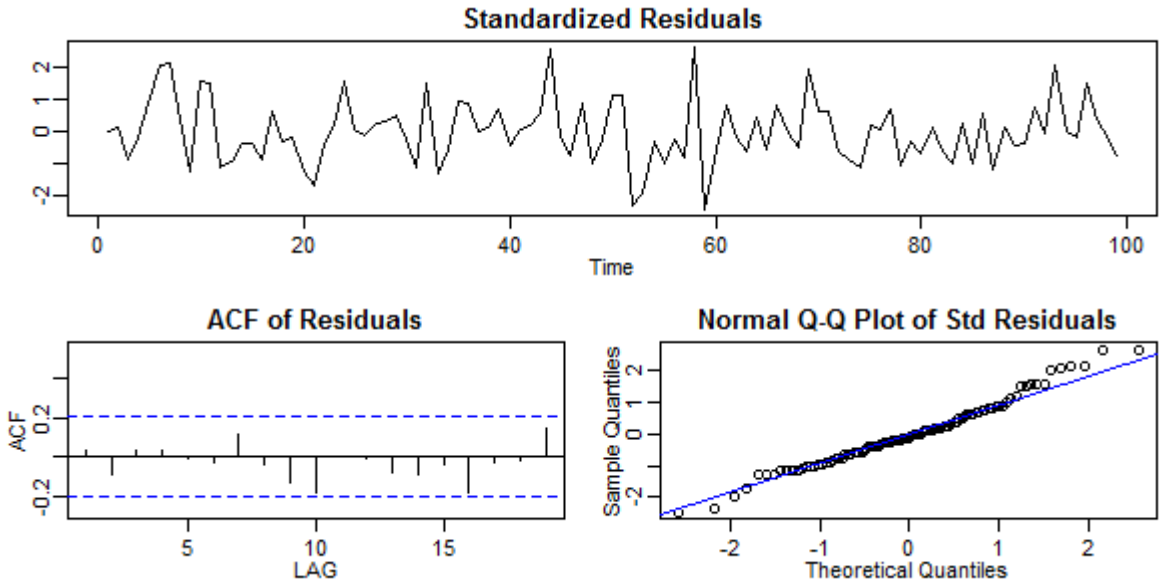

Figure 8

Ljung-Box Test

We may also use the Ljung-Box Test Statistic to see if this model is an accurate fit. We have a null hypothesis that the residual errors are uncorrelated/white noise. The alternative hypothesis is that the residual errors are correlated/significant. The Ljung-Box Test is calculated using this formula: $\tilde{Q}_W = n(n+2)\sum_{i=1}^{h}\hat{\rho}_W^2(j)/(n-j) \sim \chi^2(h-p-q)$. If the P values are greater than .05, we fail to reject the null hypothesis and thus the errors are uncorrelated. After using R programming for the Ljung-Box Test (**figure 9**), we conclude that the errors are uncorrelated since all p values are greater than .05 (lie above the dotted blue line).

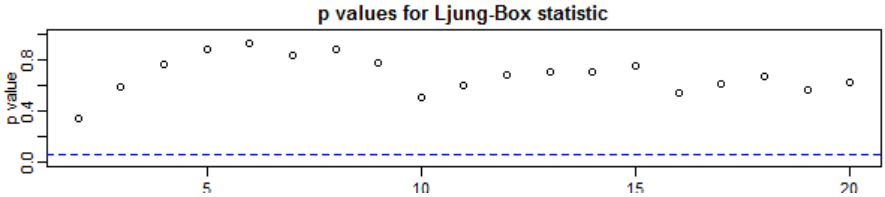

Figure 9

Forecasting and Prediction Intervals

We can now predict/forecast possible future values. The goal was to determine how many major earthquakes (over 7.0 magnitude) would occur in the next 10 years (up to 2008) and compare to real data. We may use the innovation algorithm in R which determines the minimized error of prediction to provide the best forecast:

$$\hat{X}_{n+1} = \theta_1(X_n - \hat{X}_n) + \ldots + \theta_q(X_{n+1-q} - \hat{X}_{n+1-q})$$

And prediction error:

$$v_n = E\left((X_{n+1} - \hat{X}_{n+1})^2 | X_1 = x_1, \ldots, X_n = x_n\right).$$

The resulting forecast displays an approximate of 18-20 earthquakes for years 1999-2008 (**figure 10**):

```
Time Series:
Start = 100
End = 109
Frequency = 1
 [1] 18.58501 18.63962 18.69422 18.74883 18.80344 18.85804 18.91265 18.96726
 [9] 19.02186 19.07647
>
```

A prediction interval (**figure 11)** is provided using the prediction interval formula used in the mathematical statistics/regression analysis class. Each number corresponds to the year past 1998. Notice for the last couple years, the lower prediction is slightly below 0, and since we can't have negative earthquakes, we must interpret this as possibly 0 major earthquakes for that year.

```
                [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
lower.pred  6.621486  5.569751  4.604622  3.708473  2.868951  2.077003  1.325757
upper.pred 30.548534 31.709483 32.783825 33.789188 34.737924 35.639085 36.499545
                [,8]        [,9]      [,10]
lower.pred  0.609853 -0.07498844 -0.7322459
upper.pred 37.324663 38.11871814 38.8851893
>
```

Figure 11

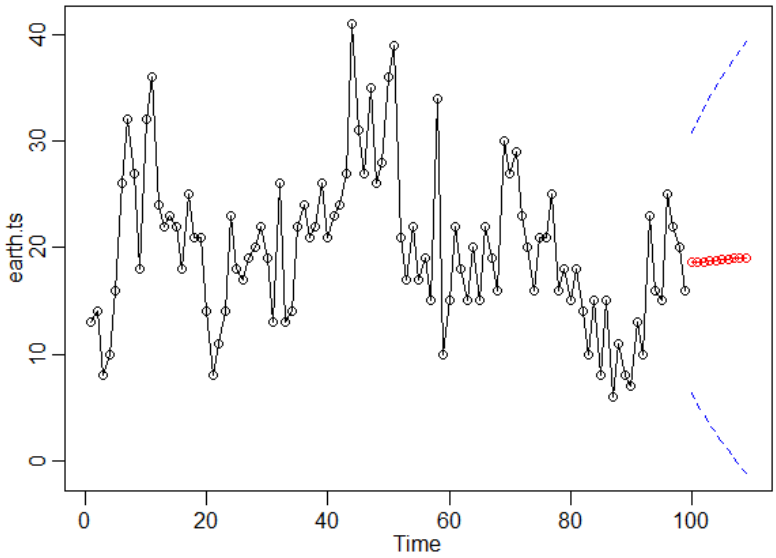

Figure 12

Figure 12 (above) shows a time series plot include the prediction interval minimum and maximum. Notice the forecasts suggests the number of earthquakes will hover slightly below 20 throughout the next 10 years, which is approximately the average number of earthquakes previously calculated.

Was Forecasting Accurate?

The United States National Geographical Survey listed all of the major earthquakes with magnitude 7.0 and above from years 1999-2008. The following table illustrates how many major earthquakes there were between those years:

1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
18	15	16	13	16	17	12	12	20	12

All of these values lie in the prediction intervals we created. Although the forecast estimates aren't 100 percent exact, we provided a solid forecast.

Sources

(*Hyndman, R.J.* Time Series Data Library*,* https://datamarket.com/data/set/22p8/number-of-earthquakes-per-year-magnitude-70-or-greater-1900-1998#!ds=22p8&display=line (citing: National Earthquake Information Center), *Accessed on 12/2/2015.*)

Earthquake information obtained from earthquake.usgs.gov (National Earthquake Information Center)(no particular author noted other than US Geological Survey), information gathered 12/2/2015

Dr. Wendy Meiring lecture notes from Fall 2015 PSTAT 174, Katherine Shatskikh notes from lab.