8-2010

# ARIMA model for forecasting Poisson data: Application to long-term earthquake predictions

Wangdong Fu
*University of Nevada, Las Vegas*

# ARIMA MODEL FOR FORECASTING POISSON DATA: APPLICATION TO LONG-TERM EARTHQUAKE PREDICTIONS

by

Wandong Fu

Bachelor of Science
Nanjing University of Technology, Nanjing
2004

A thesis submitted in partial fulfillment
of the requirements for the

**Master of Science in Mathematical Sciences**
**Department of Mathematical Sciences**
**College of Sciences**

**Graduate College**
**University of Nevada, Las Vegas**
**August 2010**

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

THE GRADUATE COLLEGE

We recommend that the thesis prepared under our supervision by

## Wandong Fu

entitled

### ARIMA Model for Forecasting Poisson Data: Application to Long-Term Earthquake Predictions

be accepted in partial fulfillment of the requirements for the degree of

### Master of Science in Mathematical Sciences

Chih-Hsiang Ho, Committee Chair

Amei Amei, Committee Member

Kaushik Ghosh, Committee Member

LeinLein Chen, Graduate Faculty Representative

Ronald Smith, Ph. D., Vice President for Research and Graduate Studies
and Dean of the Graduate College

**August 2010**

ABSTRACT

**ARIMA Models for Forecasting Poisson Data: Application to Long-Term Earthquake Predictions**

by

Wandong Fu

Dr. Chih-Hsiang Ho, Examination Committee Chair
Professor of Mathematical Sciences
University of Nevada, Las Vegas

Earthquakes that occurred worldwide during the period of 1896 to 2009 with magnitude greater than or equal to 8.0 on the Richter scale are assumed to follow a Poisson process. Autoregressive Integrated Moving Average models are presented to fit the empirical recurrence rates, and to predict future large earthquakes. We show valuable modeling and computational techniques for the point processes and time series data. Specifically, for the proposed methodology, we address the following areas: data management and graphic presentation, model fitting and selection, model validation, model and data sensitivity analysis, and forecasting.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to say thank you to everyone who helped me throughout this journey. Especially, I would like to show appreciation to my advisor, Dr. Ho, for everything. His encouragement, guidance, and support kept me on the right track. His excellence in both research and teaching will always be a great example for me.

In addition, I am thoroughly thankful to these respectable committee members, Dr. Amei, Dr. Ghosh and Dr. Chen, for their positive inputs and mentoring during my graduate studies.

Last but not least, I would also like to thank my family for their love and support. In particular, I am grateful for my husband, Ting, who always cares about me even though we are far apart from each other.

# CHAPTER 1

# INTRODUCTION

On January 12, 2010, a 7.0 magnitude earthquake hit Port-au-Prince, Haiti. The earthquake lasted one minute, just enough time to kill thousands of people and destroy numerous buildings. The earthquake caused major damage to Port-au-Prince and the surrounding area. According to the government's estimate, 200,000 people were killed, 250,000 were injured, and consequently, 1.5 million people became homeless. Many notable landmark buildings were significantly damaged or destroyed, including the Palace (President René Préval survived), the National Assembly building, the Port-au-Prince Cathedral, and the main jail. The whole country was in ruins. On February 27, 2010, a magnitude of 8.8 earthquake occurred off the coast of the Maule Region of Chile, which lasted 90 seconds. Six Chilean cities experienced intense vibrations. Tremors were also felt in many Argentine cities, including Buenos Aires, Córdoba, Mendoza and La Rioja. The earthquake triggered a tsunami which devastated several coastal towns in south-central Chile and damaged the port at Talcahuano. Tsunami warnings were issued in 53 countries, causing minor damage in the San Diego area of California. The earthquake also generated a blackout that affected 93% of the country's population and which went on for several days in some locations.

Earthquakes always strike suddenly without warning, and lead to disaster -- lack of basic necessities, loss of lives, general property damage, road and bridge damage, and collapse of buildings. Therefore, forecasting earthquake has been the focus of numerous studies (Bakun et al., 2005; Felzer et al., 2003; Helmstetter et al., 2006; Hong and Guo, 1995; Jackson and Kagan, 2006; Kagan, 1993; Savage and Cockerham, 1987; and

references therein).

In this thesis, we use the earthquake data worldwide from 1896 to 2009 with magnitude greater than or equal to 8.0 on the Richter scale. We assume that they follow a Poisson process. We then constructed a discrete time series based on the empirical recurrent rates (ERRs) of the assumed Poisson process, computed sequentially at equidistant time intervals during the observation period. The time-plot of the ERRs, referred to as the “fingerprint” or the ERR plot, offers the possibility of further insight into the data and provides a technical basis for model developments for the earthquake data. In short, we present three main ideas: (1) convert point process to ERR time series, (2) study the time series using the ARIMA modeling techniques (to be defined later), and (3) the develop methods to retrieve the counterparts of the predicted ERRs.

In summary, we define ERR, introduce ARIMA models and some related theories and methods in Chapter 2. Chapter 3 applies the modeling techniques in Chapter 2 to the earthquake data. The sensitivity analysis based on the process size and the time-steps are presented in Chapter 4. We then conclude our studies in Chapter 5.

# CHAPTER 2

# THEORIES AND METHODS

## 2.1 Empirical Recurrence Rates

Let $t_1,\ldots,t_n$ be the time of the $n$-ordered earthquakes during an observation period $(t_0,0)$, where $t_0$ is the time-origin and 0 is the present time. If $h$ is the time-step, then a discrete time series $\{z_\ell\}$ is generated sequentially at equidistant time intervals $t_o+h,\ t_0+2h,\ \ldots,\ t_0+\ell h,\ \ldots,\ t_0+Nh$ (= 0 = present time). $z_\ell$ is regarded as the observation at time $t\ (=t_0+\ell h)$, for the earthquakes to be modeled. A key parameter desired by the modelers is the recurrence rate of the targeted earthquake data. Therefore, a time series of the empirical recurrence rates (Ho, 2008) is generated as follows:

$$z_\ell = \frac{n_\ell}{\ell h} = \frac{\text{total number of earthquakes in}(t_0,t_0+\ell h)}{\ell h},$$

where $\ell$=1, 2, …, *N*. Note that $z_\ell$ evolves over time and is simply the maximum likelihood estimator (MLE) of the mean, if the underlying process observed in $(t_0,\ t_0+\ell h)$ is a homogeneous Poisson process. The time-plot of the empirical recurrence rate (ERR-plot), offers the possibility of further insights into the data. Also, if we start at time $T$, the value $z_{T+k},\ k\geq 1$ needs to be predicted based on the sample observation $(z_1,\ldots,z_T)$ of an ERR time series. In a regression modeling, let $X$ denote the time index, $z$ be the response values, and then use the fitted regression model to obtain $z_{T+k}$. However, a regression model assumes that the observations are independent and this is not a reasonable assumption for a process that evolves over time. Thus the ARIMA models are introduced.

## 2.2 ARIMA Models

Autoregressive integrated moving average (ARIMA) models are mathematical models of persistence, or autocorrelation, in a time series. It was introduced by Box and Jenkins (1976). ARIMA models allow us not only to uncover the hidden patterns in the data but also to generate forecasts and predict a variable's future values from its past values.

ARIMA models can be expressed by a series of equations. One subset of ARIMA models is called autoregressive, or AR models. The name autoregressive refers to the regression on self. An AR model describes a time series as a linear function of its past values plus a noise term $\varepsilon_t$. The order of the AR model shows the number of past values included. The simplest AR model is the first-order autoregressive, or AR (1) model. The equation for this model is given by:

$$X_t = \phi X_{t-1} + Z_t,$$

where $t = 1, 2, ..., N$, $X_t$ is a stationary zero-mean time series and $\phi$ is the first-order autoregressive coefficient. We can see that the AR (1) model has the form of a regression model in which $z_t$ is regressed on its previous value, and the error term $Z_t$ is analogous to the regression residuals and represents a "white noise" (uncorrelated with mean 0 and variance $\sigma^2$) process.

The moving average (MA) model is another form of ARIMA model in which the time series is described as a linear function of its prior errors plus a noise term $\varepsilon_t$. The first-order moving average, or MA (1), model is given by:

$$X_t = Z_t - \theta Z_{t-1},$$

where $t = 1, 2, ..., N$, $z_t$ is a stationary zero-mean time series, $Z_t$, $Z_{t-1}$ are the error terms at time t and t-1, and $\theta$ is the first-order moving average coefficient.

A general autoregressive moving average (ARMA) model, ARMA (p, q), is given by:

$$X_t - \phi_1 X_1 - ... - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q}$$

The integrated ARMA (ARIMA) is a broadening of the class of ARMA that includes differencing. We will explain the differencing in Section 2.4. Moreover, ARIMA modeling involves three stages. The first stage is to identify the model. Identification consists of specifying the appropriate model (AR, MA, ARMA, or ARIMA) and order of model. Sometimes identification is done by looking at plots of the sample autocorrelation function (ACF) and sample partial autocorrelation function (PACF). Sometimes it is done by an auto fit procedure – fitting many different possible model structures and orders and using a goodness-of-fit statistic to select the best model. The second stage is to estimate the order of the model. At this stage, the coefficients are estimated, so that the sum of squared residuals is minimized. The final stage is model diagnostics. One of the important elements in this stage is to make sure that the residuals of the candidate model are random and normally distributed. And the other one is to ensure that the estimated parameters are statistically significant. The fitting process is usually guided by the principle of parsimony, by which the best model is the one which has fewest parameters among all models that fit the data.

## 2.3 Data Splitting and ERR Plotting

If the data set is large enough, it can be split into two sets: training sample and prediction set. Training sample is used to develop a model for prediction. Prediction set is

used to evaluate the reasonableness and predictive ability of the selected model. This validation procedure named cross-validation is the statistical practice of splitting a sample of data into subsets so that the analysis is initially performed on a single subset, while the other subset is retained for subsequent use in confirming and validating the initial analysis. The application in this regard will be detailed in Section 3.1 and 3.2.

## 2.4 Data Transformation

Our main goal is to model and predict the occurrences of the large earthquakes in the future. In proving a fitted ARMA model meaningful, it must be at least plausible that the data are in fact a realization of an ARMA process and in particular a realization of a stationary process. A stationary time series is the one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. By using of some mathematical transformations, we can render our time series approximately stationary. We will introduce three common transformations that are called Box-Cox, differencing and subtracting the mean as follows.

### 2.4.1 Box-Cox Transformation

If the original observations are $Y_1, Y_2, Y_3, ...Y_n$, the Box-Cox transformation $f_\lambda$ converts them to $f_\lambda(Y_1), f_\lambda(Y_2), ... f_\lambda(Y_n)$, where:

$$f_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

This transformation is useful when the variability of the data increases or decreases with the level. By suitable choice of $\lambda$, the variability can be made nearly

constant. For instance, positive data whose standard deviation increases linearly with level, the variability can be stabilized by choosing $\lambda = 0$ (Brockwell et al., 2002).

### 2.4.2 Differencing

Differencing is an important technique in transforming data, which attempts to de-trend to control autocorrelation and achieve stationary time series. The first difference is denoted as:

$$\nabla X_t = X_t - X_{t-1} = (1-B)X_t$$

where B is the backshift operator. We may extend the notion further and define the differences of order d as:

$$\nabla^d X_t = (1-B)^d X_t$$

Usually, single differencing is used to remove linear trends and double differencing is used to remove quadratic trend. We can eliminate seasonality and trend of period $d$ by introducing the lag d difference operator $\nabla_d$:

$$\nabla_d X_t = X_t - X_{t-d} = (1-B^d)X_t$$

This operator should not be confused with the operator $(1-B)^d$ defined earlier (Ho, 2010a).

Normally, the correct amount of differencing is the lowest order of differencing that yields a time series which fluctuates around a well-defined mean value and whose autocorrelation function (ACF) plot decays rapidly to zero, either from above or below. Thus, at every stage of differencing, we check the plots of sample autocorrelation function (ACF) and the sample partial autocorrelation function (PACF) to see where the ACF/PACF "cuts off" the bounds $\pm 1.96/\sqrt{n}$. It is desirable to find a sample ACF that decays fairly rapidly. We say that a series is stationary if the sample ACF has very few

significant spikes at very small lags and then cuts off drastically or dies down very quickly. If the sample ACF dies slowly, the series still has some trend. If ACF has periodicity, the series has seasonality. We should do some more differencing of the data before continuing.

### 2.4.3 Subtracting the Mean

The term, ARMA model, is used in the program ITSM2000 (Brockwell et al., 2002) to denote a zero-mean ARMA process. Therefore, the sample mean of the data should be small before modeling. Once the apparent deviations from stationarity of the data have been removed, we subtract the sample mean of the transformed data from each observation. The search for a fitted ARMA model for a mean-corrected data set then follows.

## 2.5 Model Diagnostics

We will check the residual ACF/PACF of the models that we develop. Also the models need to pass the test for randomness of the residuals. After the model diagnostics process, we can do further predictions and comparisons.

### 2.5.1 The Sample ACF of the Residuals

For large n, the sample autocorrelations of an independent and identically distributed (iid) sequence $Y_1,\ldots, Y_n$ with finite variance are approximately iid with distribution N (0, 1/n). We can therefore test whether or not the observed residuals are consistent with iid noise by examining the sample correlations of the residuals and rejecting the iid noise hypothesis if more than two or three out of 40 fall outside the bounds $\pm 1.96/\sqrt{n}$ or if one falls far outside the bounds (Brockwell et al., 2002).

### 2.5.2 Ljung-Box Test for Lack of Fit in Time Series Models

Ljung-Box Test was proposed by Ljung and Box (1978). It is commonly used to check whether the residuals of a fitted model are iid in ARIMA modeling. It is based on the autocorrelation plot, and it tests the overall independence based on a few of lags. Because of this, it is often referred to as a portmanteau test. Formally, the definition of Ljung-Box test is as follows.

$H_0$ : The sequence data are iid

$H_a$ : The sequence data are not iid

The test statistic is $\hat{Q}(\hat{r}) = n(n+2)\sum_{k=1}^{m}(n-k)^{-1}\hat{r}_k^2$,

where $\hat{r}_k = \sum_{l=k+1}^{n}\hat{a}_l\hat{a}_{l-k} \Big/ \sum_{l=1}^{n}\hat{a}_l^2$, the estimated autocorrelation at lag $k$,

$n$ = sample size,

$m$ = number of lags being tested

$\hat{a}_1,...,\hat{a}_n$ are the residuals after a model has been fitted to a series $z_1,...,z_n$. If no model is being fitted, then $\hat{a}_1,...,\hat{a}_n$ are the "mean corrected" series of $z_1,...,z_n$.

For large $n$, the distribution of $\hat{Q}(\hat{r})$ is approximately $\chi^2_{m-p-q}$ under the null hypothesis, where $p+q$ is the number of parameters of the fitted model. The hypothesis of iid is rejected if $\hat{Q} > \chi^2_{1-\alpha;m-p-q}$ at level $\alpha$. Therefore, there is dependence among the sequence data. Or we can say the sequence data do have autocorrelations significantly different from zero.

## 2.6 Model Comparison

### 2.6.1 AIC, BIC and AICC Statistics

In this thesis, we will use the AICC statistic as an information criterion to select candidate models using the ITSM2000 package. The AICC statistic, the bias-corrected version of the AIC statistic, was introduced by Akaike in 1974. Small value of AICC is indication of a good model, but it should be used only as rough guide. Final decisions between models are based on maximum likelihood estimation. Some other Model-selection statistics, such as the BIC statistic, are also available in ITSM2000. The BIC statistic (Schwarz, 1978) is a Bayesian modification of the AIC statistic. It is evaluated at the same time as the AICC, and it is used in the same way as the AICC. Each information statistic is defined as following,

$$AIC_{p,q} = N \log \hat{\sigma}_{\varepsilon}^2 + 2r$$

$$AICC_{p,q} = N \log \hat{\sigma}_{\varepsilon}^2 + 2rN/(N-r-1)$$

$$BIC_{p,q} = N \log \hat{\sigma}_{\varepsilon}^2 + r \log N$$

where $\hat{\sigma}_{\varepsilon}^2$ is the maximum likelihood estimator of $\sigma_{\varepsilon}^2$, and $r = p + q + 1$ is the number of parameters estimated in the model, including a constant term. The second term in all three equations is a penalty for increasing *r*. Hence, if we want to minimize the values of these criteria, we should minimize the number of parameters. Therefore, the best model is the model adequately describes data and has fewest parameters.

### 2.6.2 Forecasting

The candidate ARIMA models will be used to predict future values of the time series from the past values. The forecasting function $z_t = f(z_{t-1}, \ldots, z_1) + a_t$ has the

minimum mean square error. The first part of the above equation $f(z_{t-1,}...,z_1)$ is a function of the past values of the series and it should be determined by the data. The second part $a_t$, called noise part, is a sequence of independent and identically distributed (iid) variables. Predictions will be achieved by forecasting the residuals and then inverting the transformations adopted to arrive at forecasts of the original series. Also, we will see which model is the best fitting model by comparing the prediction from training set with the prediction set. Then, we will combine the training sample and the prediction set as a full data set to forecast earthquakes for the future based on the same techniques as before. Note that the cumulated mean numbers inverted from the forecasted ERRs should be nondecreasing, and should sometimes be adjusted accordingly (e.g., Ho, 2010a.)

### 2.6.3 The Subset Model Checking

In the ITSM2000 package, the coefficients of models are given with the ratio of each estimate to 1.96 times its standard error, if it is a causal model (p85, Brockwell et al., 2002). The denominator (1.96×standard error) is the critical value (at level 0.05) for the coefficient. Thus, if the ratio is greater than 1 in absolute value, we may conclude (at level 0.05) that the corresponding coefficient in the model may be zero (Brockwell et al., 2002). After dropping the non-significant coefficients, the subset model comes up. We will do more comparisons between the full model and the subset model.

# CHAPTER 3

# APPLICATION

## 3.1 Data

Earthquakes that occurred worldwide during the period of 1896 to 2009, with Magnitude (M) $\geq 8.0$ on the Richter scale, are obtained from the *U.S. Geological Survey* (*http://www.usgs.gov/*). In the data set, the year of 1896 is the time origin $t_0$, and 2009 is the present time 0. There were 55 earthquakes that occurred during the 114 years (Appendix Table 1).

By using the raw data, we constructed a dot plot to observe any possible trends (Figure 1). It is clear that the dot plot has limited value in delivering the temporal trend presented by the data.

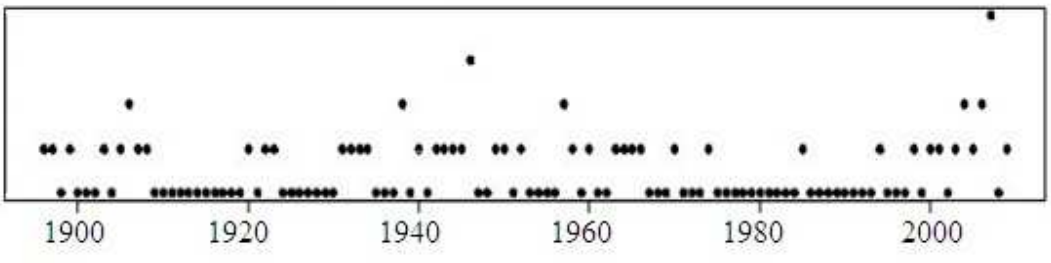


**Figure 1**. Dot plot of large earthquakes worldwide between 1896 and 2009

Then, we count the number of earthquakes with each time-step (1, 2 and 3 years) and calculate the $z_l$ values to do further analysis (Appendix Table 2 - 4). Three ERR plots with three different time-steps are shown in Figure 2.

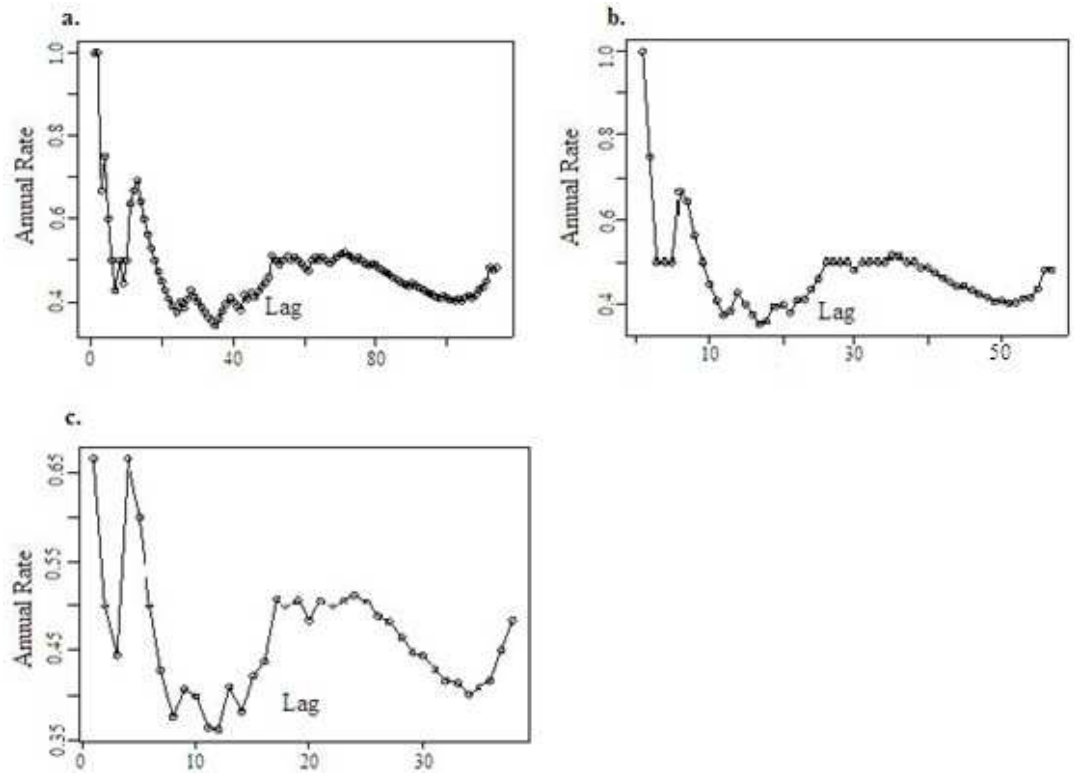


**Figure 2**. ERR plots with different time-steps (h): **a** h=1 year, **b** h=2 years, **c** h=3 years

Since there are 55 large earthquakes in 114 years, which indicates there is approximately one large earthquake in every two years. We consider to choose h = 2 years as the time-step. Therefore, we will try to predict earthquakes with h = 2 years.

### 3.2 ARIMA Modeling with h = 2

We will use the ITSM2000 software to model the ERR data with h = 2 years. The data set with time-step h = 2 years has 57 lags in total. At First, we use the technique described in Section 2.3 to split the data into two sets: training sample and prediction set. In this case, our training sample is the original data set excluding the last 5 ERRs, which is the prediction set (Figure 3).These five ERR values in the prediction set, representing a

decade of earthquake data, will be used to compare with those of the one to five-step predictions, produced by a candidate model. Of course, the size of a predict set is quite flexible as long as it fits a common goal of model selection. Then we focus on the training sample set and plot the sample ACF and PACF to observe the data set (Figure 4). From the plot of sample ACF, we found that the spikes die slowly and have periodicity. This indicates nonstationary behavior. As mentioned in Section 2.4, it has some trend and seasonality. Thus differencing is considered.

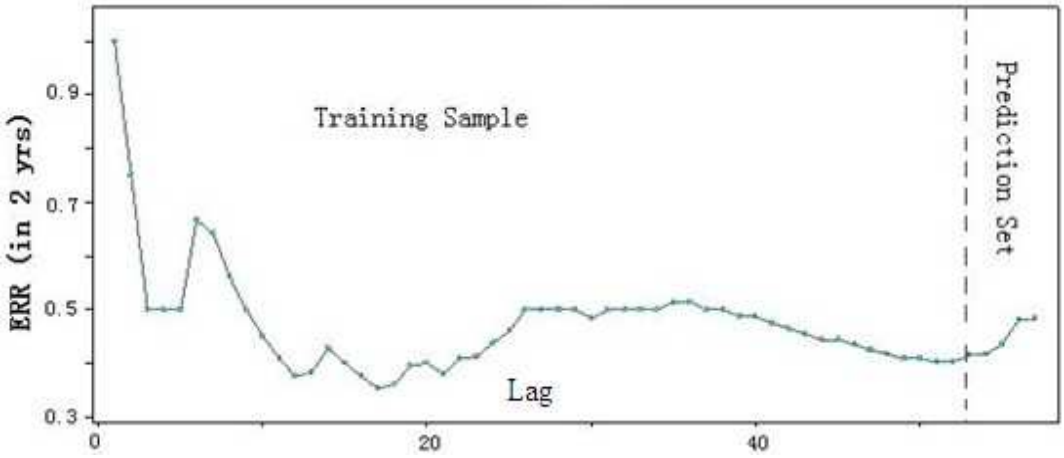


**Figure 3**. Training sample and prediction set of data set with h = 2 years. Each lag corresponds to 2 years

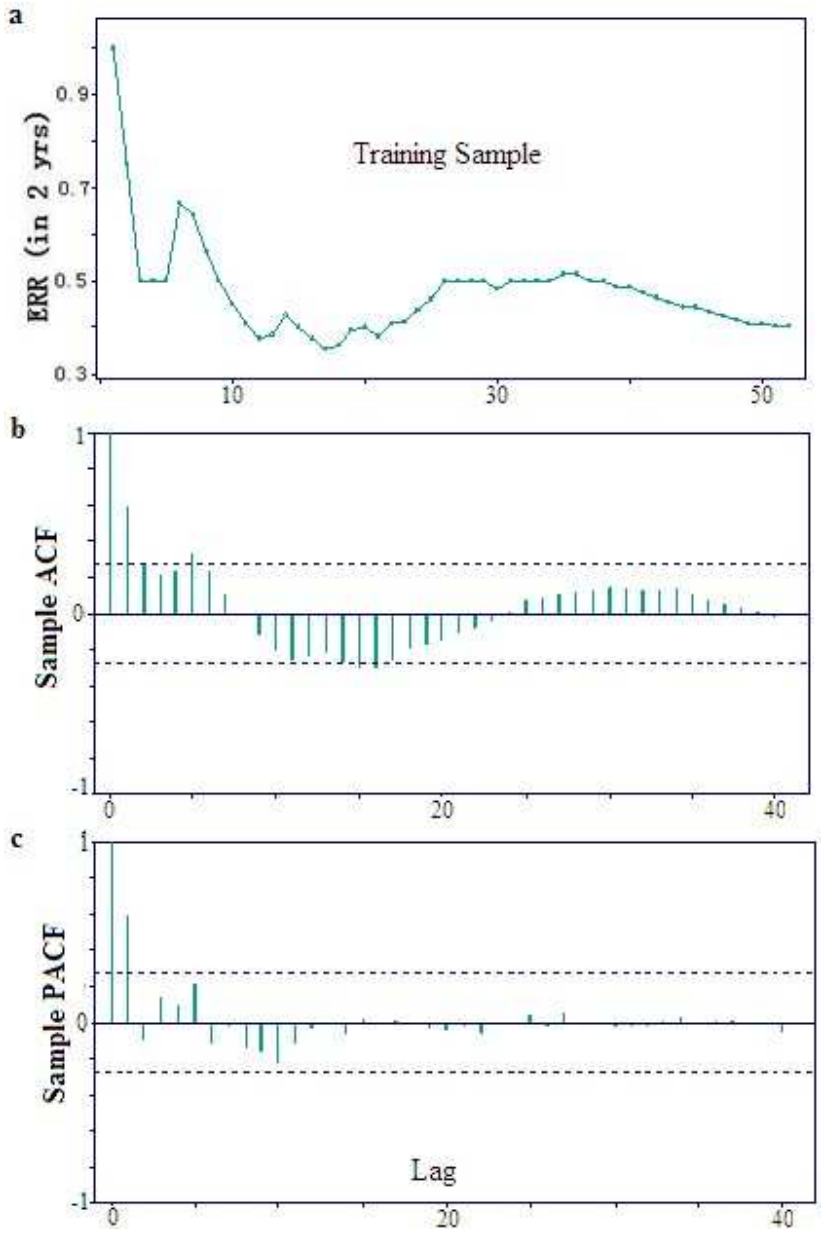


**Figure 4**. **a**, time-plot; **b**, sample ACF; **c**, sample PACF of the training sample with h = 2 years. Each lag corresponds to 2 years

Applying the differencing operator $\nabla$ on the training sample, we take a difference at lag 1. Figure 5 tells us that the stationarity has not been achieved. So we do further

difference at lag 1. Then we subtract the sample mean from each observation of the differenced series to generate a stationary zero-mean time series (Figure 6). We feel that the ACF is cutting off at lag 3 and the PACF is tailing off. This would suggest that an MA (3) should be considered. Indeed, our initial model selection process concludes that the estimated (MLE) model is:

$X_t = Z_t - 0.2475Z_{t-1} + 0.1471Z_{t-2} - 0.4985Z_{t-3}$

Estimated WN Variance = 0.002240

Standard Error of MA Coefficients

0.141517 0.168863 0.114696

Note that $X_t$ represents a twice-differenced stationary zero-mean time series and the error term $Z_t$ represents a white noise process.

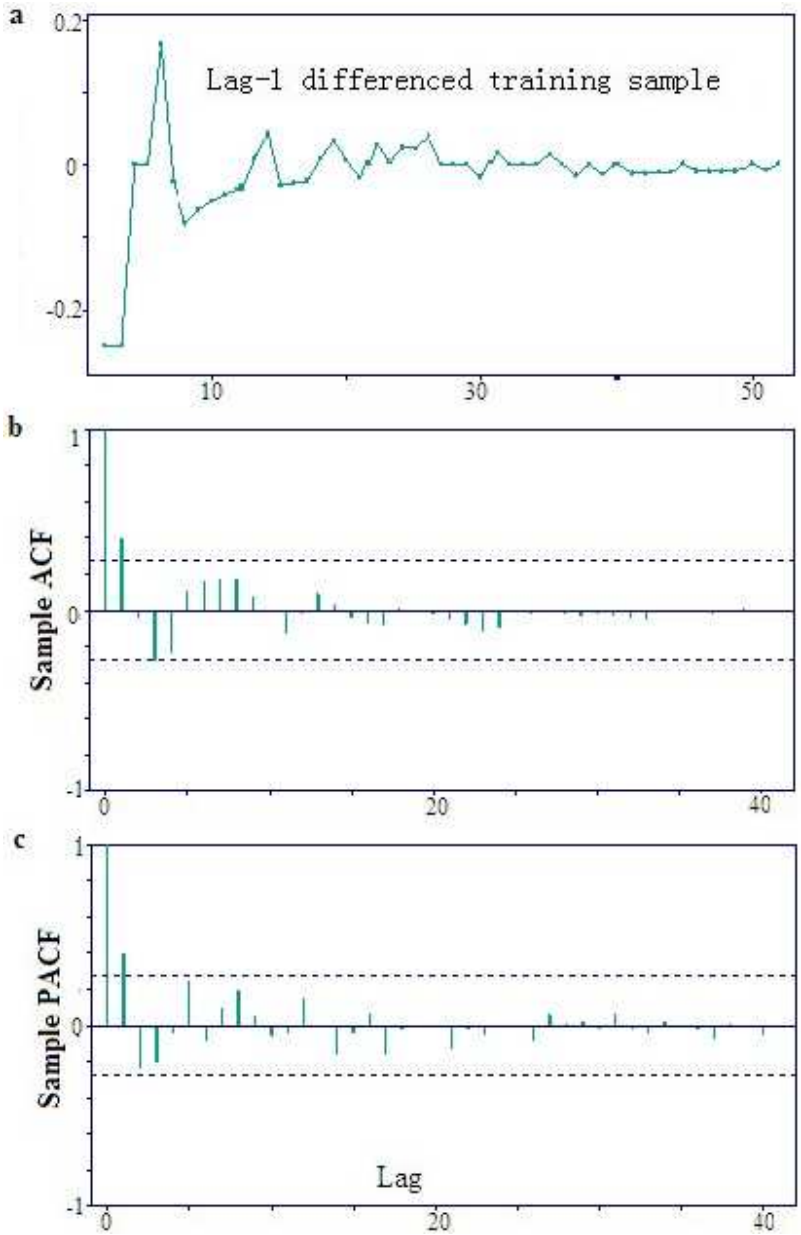


**Figure 5 a,** time-plot; **b,** sample ACF; **c**, sample PACF of a lag-1 differenced training sample with h = 2 years. Each lag corresponds to 2 years

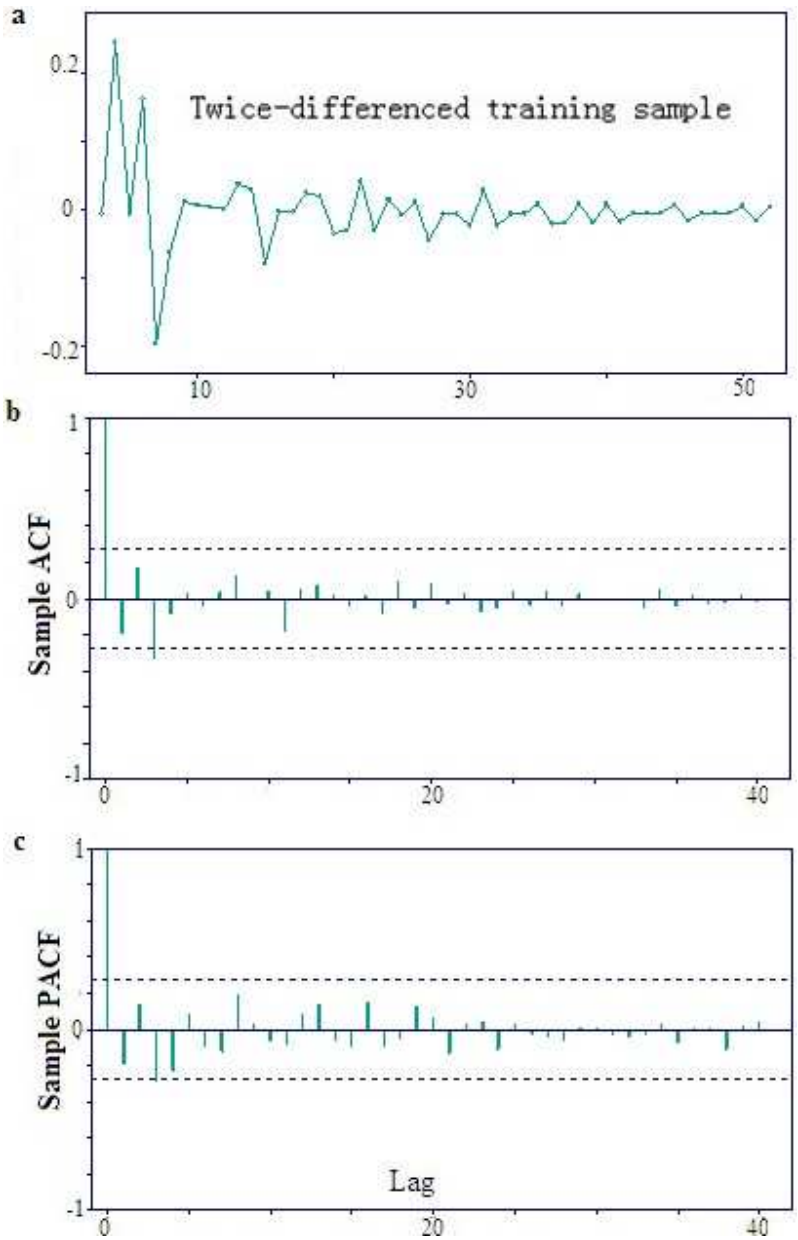


**Figure 6. a**, time-plot; **b**, sample ACF; **c**, sample PACF of the twice-differenced training sample with h = 2 years. Each lag corresponds to 2 years

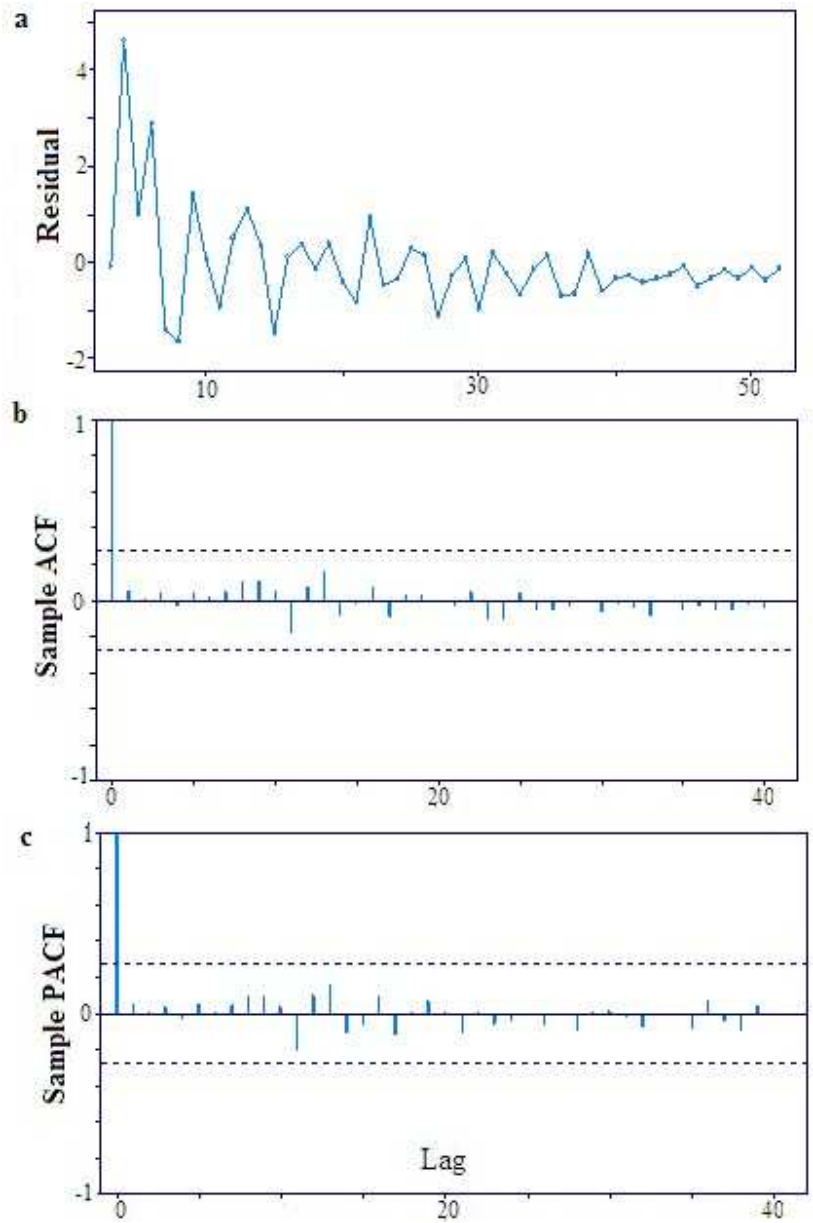


**Figure 7.** Diagnostics for the MA (3) fitted to the mean-corrected and twice-differenced training sample. Residual **a**, time-plot; **b**, sample ACF; **c**, sample PACF. Each lag corresponds to 2 years

A set of diagnostic plots (Figure 7) is produced by the ITSM2000 package, consisting of the plot of the residuals, its ACF and PACF for the MA (3) model. The

AICC statistic is -153.367. And the Ljung - Box test is not significant (p-value = 0.96067), indicating that the residuals are approximately white noise. The numerical values of the actual ERRs and mean numbers in the prediction set, and the predicted ERRs by the model MA (3) with their counterparts are shown in Table 5.

**Table 5**. The numerical values of the actual ERRs and mean numbers in the prediction set, and the predicted ERRs using the MA (3) with their counterparts (the corresponding mean values derived from the predicted ERRs)

| Year | Annual ERR | | Mean number | |
|---|---|---|---|---|
| | Actual | Prediction | Actual | Prediction |
| 2000-2001 | 0.415094 | 0.41238 | 2 | 1.71228 |
| 2002-2003 | 0.416667 | 0.43368 | 1 | 2.83744 |
| 2004-2005 | 0.436364 | 0.46318 | 3 | 5.9498 |
| 2006-2007 | 0.482143 | 0.49771 | 6 | 7.74352 |
| 2008-2009 | 0.482456 | 0.53728 | 1 | 7.24992 |

We list the ratios of (estimated coefficients)/(1.96×standard error) for each coefficient, calculated from the output of an MA (3) model, shown in Section 3.2. The ratios are:

-0.892166 0.444319 -2.217303

Note that the ratio at lag 3 in absolute value is greater than 1, which indicates the corresponding coefficient is nonzero. We will keep the corresponding coefficient. Table 6 shows the AICC statistics and the p-values of the Ljung-Box test for a variety of subset MA (3) models. All of these models pass the residual diagnostic tests.

**Table 6**. The AICC statistics and the p-values of the Ljung-Box test for a variety of subset MA (3) models

| Lags | | | MLE Model | AICC | p-value of the Ljung-Box Test |
|---|---|---|---|---|---|
| 1 | 2 | 3 | **1**. $X_t = Z_t - 0.2475Z_{t-1} + 0.1471Z_{t-2} - 0.4985Z_{t-3}$ | -153 | 0.961 |
| | 2 | 3 | **2**. $X_t = Z_t - 0.1995Z_{t-2} - 0.4659Z_{t-3}$ | -153 | 0.898 |
| 1 | | 3 | **3**. $X_t = Z_t - 0.1995Z_{t-1} - 0.4659Z_{t-3}$ | -155 | 0.978 |
| | | 3 | **4**. $X_t = Z_t \quad - 0.5136Z_{t-3}$ | -155 | 0.828 |

We then use these models to make predictions. Figure 8 shows the comparisons of the results with the prediction set. Model 1 – 4 are defined in Table 6. The predicted values are very similar, indicating that these models are all acceptable.

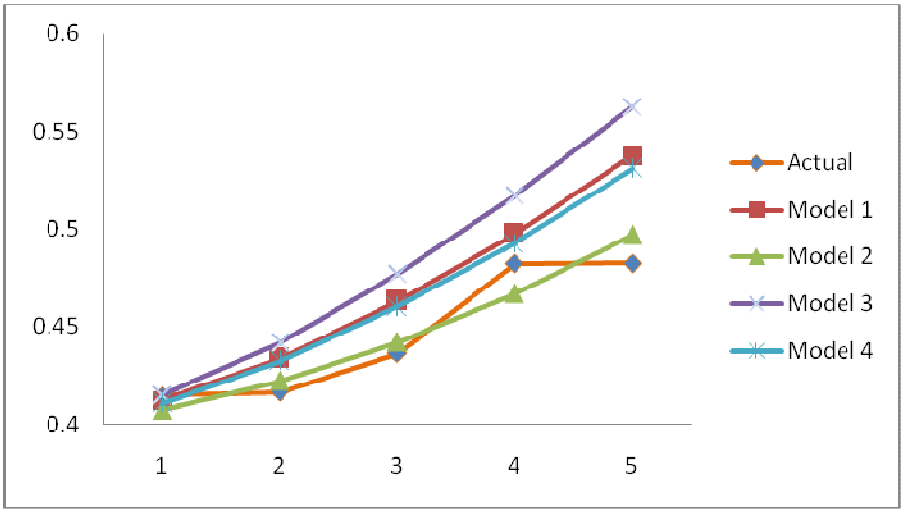


**Figure 8**. Comparison of five forecasted ERRs with the prediction set. Each lag corresponds to 2 years

### 3.3 Full-Data Forecasting

Finally, we will use the full ERR time series to forecast the number of earthquakes in the future. This yields the best-fitted MA (3) model for the mean-corrected and twice-differenced at lag 1 data (same as before). The estimated (MLE) model is:

$X_t = Z_t - 0.2708Z_{t-1} + 0.1450Z_{t-2} - 0.5025Z_{t-3}$

Estimated WN Variance = 0.002100

Standard Error of MA Coefficients

0.134042 0.165004 0.114643

The AICC statistic is -173.294, and the Ljung - Box test is not significant (p-value = 0.96568). Then, we check the ratios as following:

-1.030746 0.448350 -2.236312

This leads to a subset MA (3), which has the AICC statistic -174.738, and the p-value of the Ljung-Box test is 0.97676. The estimated (MLE) model is:

$X_t = Z_t - 0.2213\ Z_{t-1} - 0.4597\ Z_{t-3}$

Estimated WN Variance = .002137

Standard Error of MA Coefficients

0.114263 0.000000 0.108700

The AICC statistics and the p-values of the Ljung-Box test of the subset MA (3) are a little better than MA (3). But there is no big difference, and we will keep both of them.

The predictions of the next ten years, from 2010 to 2019, are shown in the Table 7.

**Table 7**. The predicted ERRs using the MA (3) and the subset MA (3) with their counterparts (the corresponding mean values derived from the predicted ERRs)

| Year | Full model ERR | | Mean number | |
|---|---|---|---|---|
| | MA (3) | Subset MA (3) | MA (3) | Subset MA (3) |
| 2010-2011 | 0.50365 | 0.49733 | 3.4234 | 2.69028 |
| 2012-2013 | 0.50785 | 0.50620 | 1.5029 | 2.04132 |
| 2014-2015 | 0.54365 | 0.54526 | 5.3117 | 5.69960 |
| 2016-2017 | 0.58401 | 0.58886 | 6.01122 | 6.40972 |
| 2018-2019 | 0.62891 | 0.63702 | 6.73562 | 7.14956 |

# CHAPTER 4

# SENSITIVITY ANALYSIS

In Chapter 3, we have shown that, for h = 2 years, there are at least two adequate ARIMA models for the chosen earthquake data. We are now ready to address the following issues: (1) Will the technique be applicable for a small point process or data set? And (2) how will the choices of time-step affect the results? The investigation of the first part could be done by increasing the time-step of our ERRs to a desired level. In this thesis, however, we simply cite the work of Ho (2010b) to demonstrate that the proposed technique works for the Parkfield earthquake prediction experiment, which represents a small point process. We then discuss the sensitivity analysis based on our own data with three different time-steps.

## 4.1 Sensitivity on Process Size - Parkfield Earthquake Prediction

Since the large earthquake of Jan. 2, 1857, earthquake sequences with main shocks of magnitude (M) 6 have occurred near Parkfield, on Feb. 2, 1881, Mar. 3, 1901, Mar. 10, 1922, June 8, 1934, June 28, 1966, and Sep. 28, 2004 (Bakun et. al, 2005). This is a small point process with a somewhat periodic recurrence rate. A focused earthquake prediction experiment has been in progress along this area since then. The Parkfield Experiment (*http://earthquake.usgs.gov/research/parkfield/index.php*), which is led by the USGS and the State of California, is a comprehensive, long-term earthquake research project on the San Andreas Fault. Scientists hope to better understand the physics of earthquakes -- what actually happens on the fault and in the surrounding region before, during, and after an earthquake, and to provide a scientific basis for earthquake prediction. The experiment

has involved more than 100 researchers at the USGS, collaborating universities and government laboratories. Their coordinated efforts have led to a dense network of instruments poised to "capture" the anticipated earthquake and reveal the earthquake process in unprecedented detail.

In 1985, the National Earthquake Prediction Evaluation Council (NEPEC) issued a statement that an earthquake of about M 6 would probably occur before 1993 on the San Andreas Fault near Parkfield (Shearer, 1985). However, no such event occurred until Sep. 28, 2004. The ARIMA model of Ho (2010b) predicted a new earthquake to occur between Dec. 5, 2002 and Dec. 4, 2004.

## 4.2 Sensitivity on Time-Step, h

### 4.2.1 ARIMA Modeling with h = 1

When we choose the time-step h = 1 year, the data set has 114 lags in total. The training sample with 104 lags and the prediction set with 10 lags are shown in Figure 9. The plots of sample ACF and PACF on the training sample (Figure 10) indicate nonstationary behavior. Therefore, we need differencing. We used the same method described in the last chapter. After taking twice difference at lag 4 and 1, we find the best fitting model ARMA (5, 5). The AICC statistic is -362.268, and the Ljung and Box test is not significant (p-value = 0.86584), which gives us evidence to believe that the residuals are approximately white noise. The estimated (MLE) model is:

$X_t = -0.1717X_{t-1} + 0.1597X_{t-2} + 0.5383X_{t-3} – 0.08758X_{t-4} - 0.4073X_{t-5} + Z_t +$

$0.2547Z_{t-1} + 0.4622Z_{t-2} - 0.2237Z_{t-3} - 0.2444Z_{t-4} - 0.6749Z_{t-5}$

Estimated WN Variance = .001080 Standard Error of AR Coefficients

| | | | | |
|---|---|---|---|---|
| 0.120296 | 0.151424 | 0.122791 | 0.138724 | 0.130128 |

Standard Error of MA Coefficients

| | | | | |
|---|---|---|---|---|
| 0.104708 | 0.131067 | 0.120401 | 0.119647 | 0.105785 |

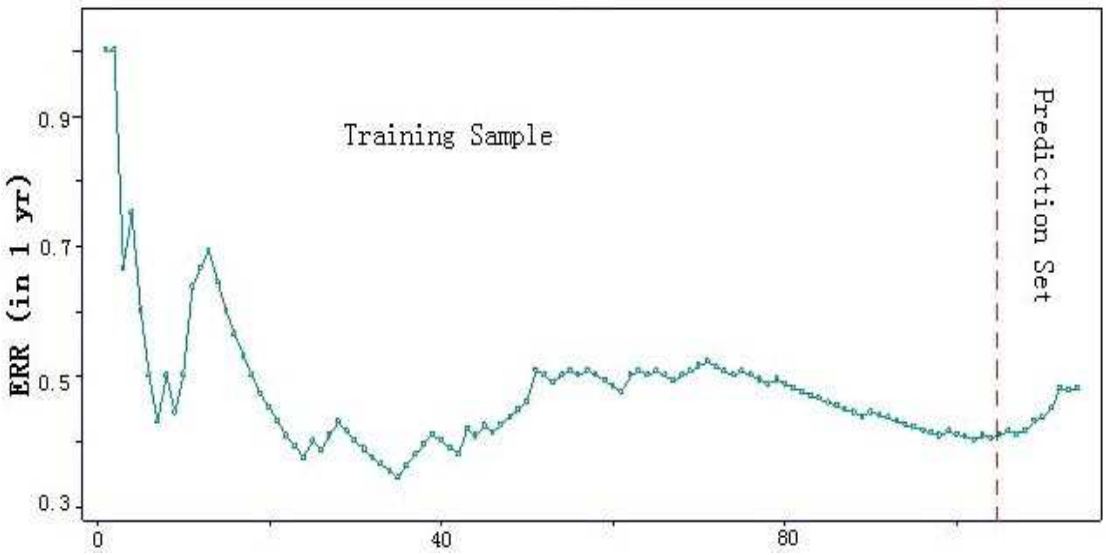


**Figure 9**. Training sample and prediction set of data set with h = 1 year. Each lag corresponds to 1 year

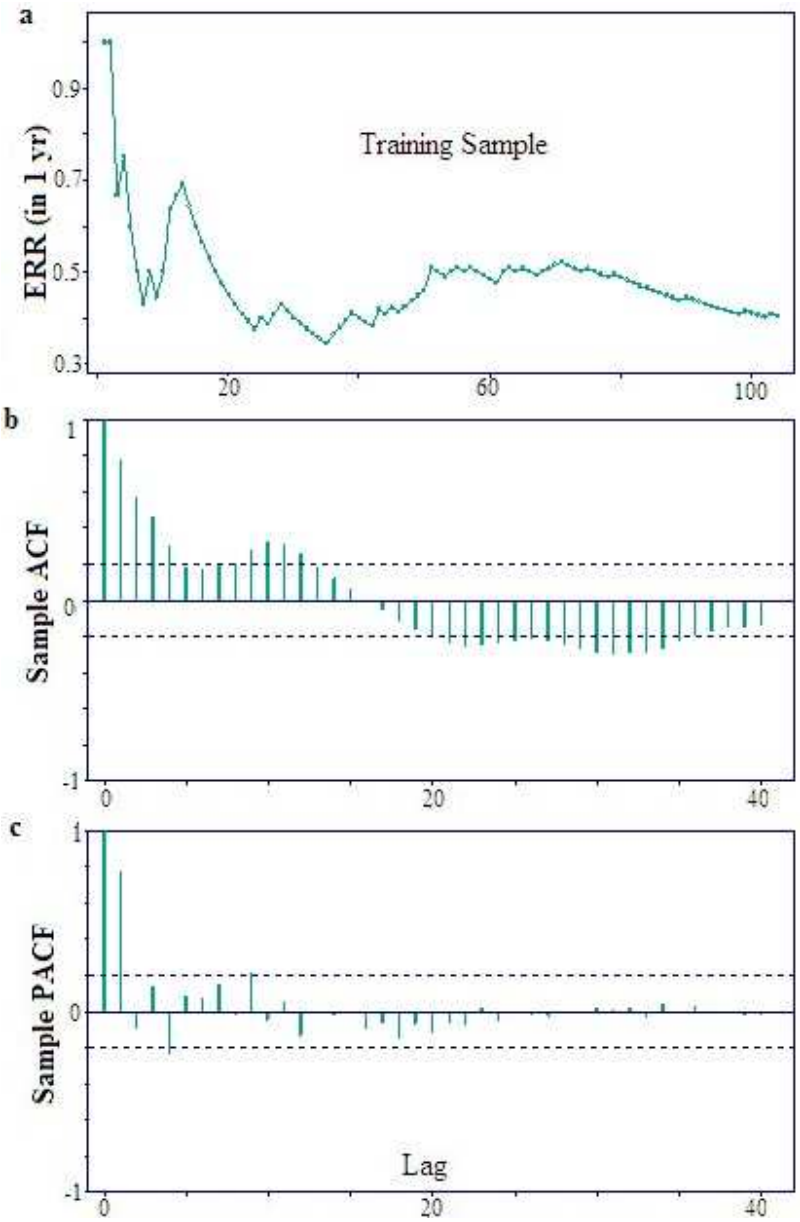


**Figure 10**. **a**, time-plot; **b**, sample ACF; **c**, sample PACF of the training sample with h = 1 year. Each lag corresponds to 1 year

Unfortunately, all the subset ARMA (5, 5) models neither pass the model diagnostic tests nor outperform the ARMA (5, 5). The forecasting results for this model are summarized in Table 8.

**Table 8**. The numerical values of the actual ERRs and mean numbers in the prediction set, and the predicted ERRs using the ARMA (5, 5) with their counterparts (the corresponding mean values derived from the predicted ERRs)

| Year | Annual ERR | | Mean number | |
|---|---|---|---|---|
| | Actual | Prediction | Actual | Prediction |
| 2000 | 0.40952381 | 0.41519 | 1 | 1.59495 |
| 2001 | 0.41509434 | 0.4218 | 1 | 1.11585 |
| 2002 | 0.41121495 | 0.43678 | 0 | 2.02466 |
| 2003 | 0.41666667 | 0.44336 | 1 | 1.14742 |
| 2004 | 0.43119266 | 0.46441 | 2 | 2.73781 |
| 2005 | 0.43636364 | 0.47262 | 1 | 1.36751 |
| 2006 | 0.45045045 | 0.49327 | 2 | 2.76477 |
| 2007 | 0.48214286 | 0.50355 | 4 | 1.64463 |
| 2008 | 0.47787611 | 0.52445 | 0 | 2.86525 |
| 2009 | 0.48245614 | 0.53609 | 1 | 1.85141 |

$X_t = - 0.1766X_{t-1} + 0.1507X_{t-2} + 0.5407X_{t-3} - 0.08186X_{t-4} - 0.4143X_{t-5} + Z_t + 0.2635Z_{t-1} + 0.4700Z_{t-2} - 0.2179Z_{t-3} - 0.2439Z_{t-4} - 0.6681Z_{t-5}$

Estimated WN Variance = 0.000990

Standard Error of AR Coefficients

0.112016 0.112036 0.100380 0.102246 0.102535

Standard Error of MA Coefficients

0.096521 0.098182 0.121989 0.097102 0.092292

The above output is the best fitted mode using a complete ERR time series for the mean-corrected and twice-differenced data (same as before). It's also an ARMA (5, 5). The predictions are shown in Table 9.

**Table 9**. Predictions of large earthquakes with h = 1 year

| Year | Full model ERR | Mean number |
|---|---|---|
| 2010 | 0.4976 | 2.224 |
| 2011 | 0.50603 | 1.47548 |
| 2012 | 0.50149 | 0 (adjusted) |
| 2013 | 0.53084 | 3.96479 |
| 2014 | 0.54151 | 1.80057 |
| 2015 | 0.5598 | 2.73631 |
| 2016 | 0.57991 | 2.99311 |
| 2017 | 0.60608 | 3.77265 |
| 2018 | 0.62047 | 2.37605 |
| 2019 | 0.65601 | 5.02743 |

### 4.2.2 ARIMA Modeling with h = 3

The data set with the time-step h = 3 years has 38 lags. The training sample with 35 lags and the prediction set with 3 lags are shown in Figure 11. The plots of sample ACF and PACF on the training sample (Figure 12) indicate nonstationary behavior. Thus differencing is considered. We took the first difference at lag 2 and the second difference at lag 1. This is also a suggestion of the ARMA (2, 2) model. The estimated (MLE) model is:

$X_t = 0.1439\ X_{t-1} - 0.7329\ X_{t-2} + Z_t - 0.08275\ Z_{t-1} - 0.3587\ Z_{t-2}$

Estimated WN Variance = 0.003556

Standard Error of AR Coefficients

0.309275 0.244192

Standard Error of MA Coefficients

0.325282 0.285052

The AICC statistic is -74.570857. The Ljung - Box statistic is 14.846 and the p-value is 0.78513, which indicates that the residuals are approximately white noise. Additionally, all the subset ARMA (2, 2) models neither pass the model diagnostic tests nor outperform the ARMA (2, 2). Table 10 gives us the comparison of actual ERRs and prediction values.

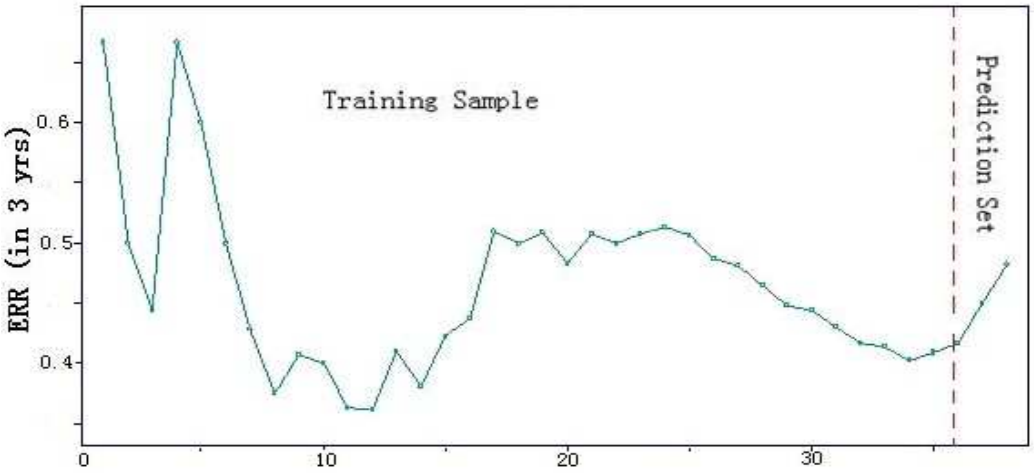


**Figure 11**. Training sample and prediction set of data set with h = 3 years. Each lag corresponds to 3 years

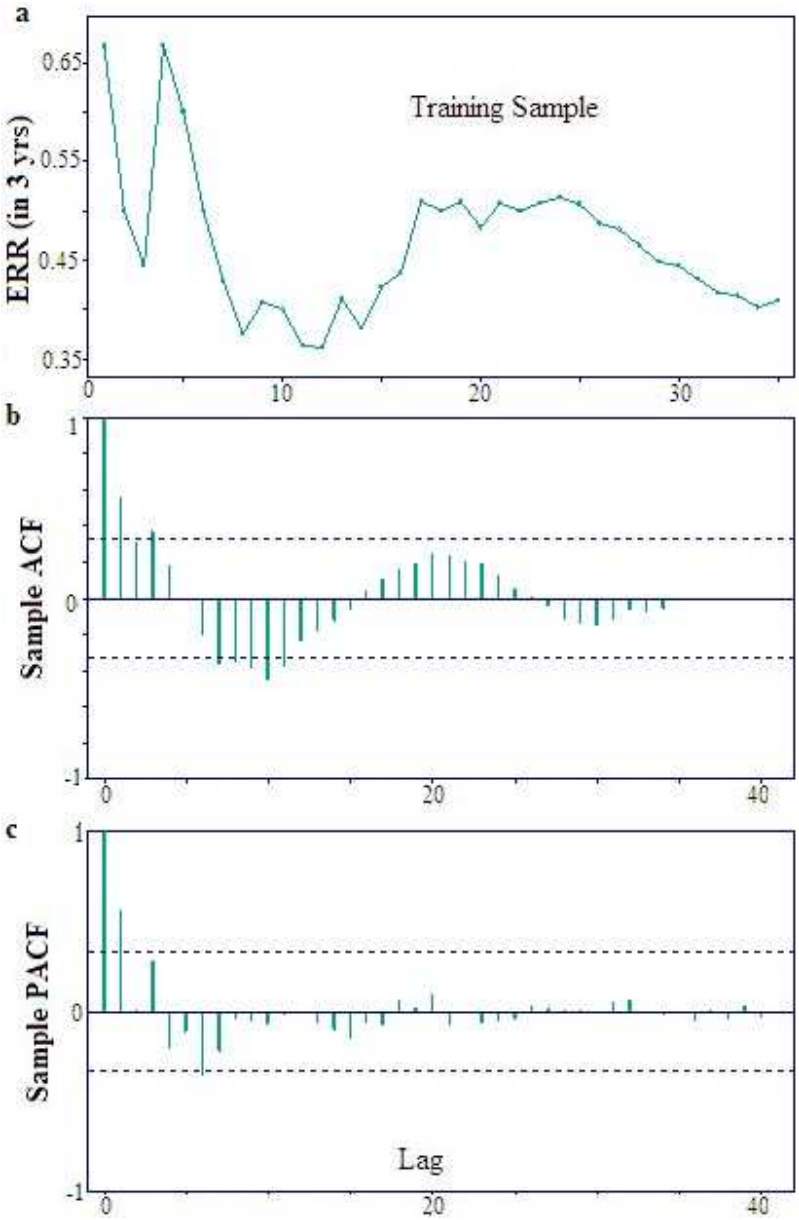


**Figure 12**. **a**, time-plot; **b**, sample ACF; **c**, sample PACF of the training sample with h = 3 years. Each lag corresponds to 3 years

**Table 10**. The numerical values of the actual ERRs and mean numbers in the prediction set, and the predicted ERRs using the ARMA (2, 2) with their counterparts (the corresponding mean values derived from the predicted ERRs)

| Year | Annual ERR | | Mean number | |
|---|---|---|---|---|
| | Actual | Prediction | Actual | Prediction |
| 2001-2003 | 0.41666667 | 0.41764 | 2 | 2.10512 |
| 2004-2006 | 0.45045045 | 0.43075 | 5 | 2.70813 |
| 2007-2009 | 0.48245614 | 0.4356 | 5 | 1.84515 |

A complete ERR time series yields the following best fitted model for the mean-corrected and twice-differenced data (same as before).

$X_t = 0.1512\ X_{t-1} - 0.7353\ X_{t-2} + Z_t - 0.05566\ Z_{t-1} - 0.3374\ Z_{t-2}$

Estimated WN Variance = 0.003382

Standard Error of AR Coefficients

0.307757 0.242601

Standard Error of MA Coefficients

0.322960 0.273897

Again, it's ARMA (2, 2). The predictions are shown in Table 11.

**Table 11.** Predictions of large earthquakes with h = 3 years

| Year | Full model ERR | Mean number |
|---|---|---|
| 2010-2012 | 0.50662 | 4.27454 |
| 2013-2015 | 0.52456 | 3.67266 |
| 2016-2018 | 0.56671 | 6.75813 |

In conclusion, Table 12 shows the prediction values of all the comparable models with different time-steps.

**Table 12**. Predictions of earthquakes with different time-steps, h

| h | Fitted model | Full data forecasting |
|---|---|---|
| | | 2010-2015 |
| 1 | ARMA (5, 5) | 12.176 |
| 2 | Subset MA (3) | 10.431 |
| 3 | ARMA (2, 2) | 7.9472 |

# CHAPTER 5

# CONCLUSIONS

Earthquakes that occurred during 1896 to 2009 with magnitude greater than or equal to 8.0 on the Richter scale are assumed to follow a Poisson process. Time series (ARIMA) models are well developed, and are applied in many fields. The integrated ARMA, or ARIMA model, is an extension of the class of ARMA models that include differencing. In this thesis, we build a linking bridge between a Poisson process and the classical time series via a sequence of the empirical recurrent rates (ERR), calculated sequentially at equidistant time intervals.

We split the earthquake data set into a training sample and a prediction set. The training sample is used to develop the candidate models. For time-step h = 2 years, we used the last five ERRs as a prediction set to make model comparisons by checking the predictive ability of the candidate models developed from the training sample. Before modeling, we must make sure the ARMA process is stationary. After taking twice difference at lag 1, an MA (3) model with the lowest AICC statistic passes the randomness test for residuals and has all the residual ACF lags falling within the bounds $\pm 1.96/\sqrt{n}$ (Figure 7). The ratio (estimated coefficient)/(1.96×standard error) is a critical value (at level 0.05) for the coefficient. If the ratio is greater than 1 in absolute value, we conclude that the corresponding coefficient is not significant. We then obtained some subset MA (3) models by dropping the non-significant coefficients. Based on the full data forecasting, there will be 12 large earthquakes in the next 6 years from the prediction of h = 1 year, 10 earthquakes from the prediction of h = 2 years, and 8 from h = 3 years, which are pretty similar (Table 12).

The application of ARIMA models for long-term earthquake prediction is a natural extension of the methodologies developed for the volcanic risk assessment studies (Ho, 2008, 2010a). Likewise, this work will further facilitate the research in the areas of monitoring the occurrence rates of cancer death, car accident, teen pregnancy, suicide, dust storm, hurricane, bank failure, foreclosure, genetic mutation, etc.

APPENDIX

DATA

**Table 1**. Large earthquakes worldwide since 1896 ($M \geq 8.0$)

| Date | Location | Magnitude |
|---|---|---|
| 06/15/1896 | Sanriku, Japan | 8.5 |
| 06/12/1897 | Assam, India | 8.3 |
| 09/10/1899 | Yakutat Bay, Alaska | 8 |
| 08/11/1903 | Southern Greece | 8.3 |
| 07/09/1905 | Mongolia | 8.4 |
| 01/31/1906 | Off the Coast of Esmeraldas, Ecuador | 8.8 |
| 08/17/1906 | Valparaiso, Chile | 8.2 |
| 10/21/1907 | Qaratog, Tajikistan | 8 |
| 12/12/1908 | Off the Coast of Central Peru | 8.2 |
| 06/05/1920 | Taiwan region | 8 |
| 11/11/1922 | Chile-Argentina Border | 8.5 |
| 02/03/1923 | Kamchatka | 8.5 |
| 08/10/1931 | Xinjiang, China | 8 |
| 06/03/1932 | Jalisco, Mexico | 8.1 |
| 03/02/1933 | Sanriku, Japan | 8.4 |
| 01/15/1934 | Bihar, India - Nepal | 8.1 |
| 02/01/1938 | Banda Sea, Indonesia | 8.5 |
| 11/10/1938 | Shumagin Islands, Alaska | 8.2 |
| 05/24/1940 | Callao, Peru | 8.2 |
| 08/24/1942 | Off the coast of central Peru | 8.2 |
| 04/06/1943 | Illapel - Salamanca, Chile | 8.2 |
| 12/07/1944 | Tonankai, Japan | 8.1 |
| 11/27/1945 | Makran Coast, Pakistan | 8 |
| 04/01/1946 | Unimak Island, Alaska | 8.1 |
| 08/04/1946 | Samana, Dominican Republic | 8 |
| 12/20/1946 | Nankaido, Japan | 8.1 |
| 08/22/1949 | Queen Charlotte Islands, British Columbia, Canada | 8.1 |
| 08/15/1950 | Assam - Tibet | 8.6 |
| 11/04/1952 | Kamchatka | 9 |
| 03/09/1957 | Andreanof Islands, Alaska | 8.6 |
| 12/04/1957 | Gobi-Altay, Mongolia | 8.1 |
| 11/06/1958 | Kuril Islands | 8.3 |
| 05/22/1960 | Chile | 9.5 |
| 10/13/1963 | Kuril Islands | 8.5 |
| 03/28/1964 | Prince William Sound, Alaska | 9.2 |
| 02/04/1965 | Rat Islands, Alaska | 8.7 |

| Date | Location | Magnitude |
|---|---|---|
| 10/17/1966 | Near the Coast of Peru | 8.1 |
| 07/31/1970 | Colombia | 8 |
| 10/03/1974 | Near the Coast of Central Peru | 8.1 |
| 09/19/1985 | Michoacan, Mexico | 8 |
| 06/09/1994 | Bolivia | 8.2 |
| 03/25/1998 | Balleny Islands Region | 8.1 |
| 11/16/2000 | New Ireland Region, Papua New Guinea | 8 |
| 06/23/2001 | Near the Coast of Peru | 8.4 |
| 09/25/2003 | Hokkaido, Japan Region | 8.3 |
| 12/23/2004 | North of Macquarie Island | 8.1 |
| 12/26/2004 | Sumatra-Andaman Islands | 9.1 |
| 03/28/2005 | Northern Sumatra, Indonesia | 8.6 |
| 05/03/2006 | Tonga | 8 |
| 11/15/2006 | Kuril Islands | 8.3 |
| 01/13/2007 | East of the Kuril Islands | 8.1 |
| 04/01/2007 | Solomon Islands | 8.1 |
| 08/15/2007 | Near the Coast of Central Peru | 8 |
| 09/12/2007 | Southern Sumatra, Indonesia | 8.5 |
| 09/29/2009 | Samoa Islands region | 8.1 |

**Table 2**. ERR with time-step h = 1 year

| Time-Step | Count | ERR | Time-Step | Count | ERR |
|---|---|---|---|---|---|
| 1896* | 1 | 1 | 1935 | 0 | 0.4 |
| 1897 | 1 | 1 | 1936 | 0 | 0.390244 |
| 1898 | 0 | 0.666667 | 1937 | 0 | 0.380952 |
| 1899 | 1 | 0.75 | 1939 | 0 | 0.409091 |
| 1900 | 0 | 0.6 | 1940 | 1 | 0.422222 |
| 1901 | 0 | 0.5 | 1941 | 0 | 0.413043 |
| 1902 | 0 | 0.428571 | 1942 | 1 | 0.425532 |
| 1903 | 1 | 0.5 | 1943 | 1 | 0.4375 |
| 1904 | 0 | 0.444444 | 1944 | 1 | 0.44898 |
| 1905 | 1 | 0.5 | 1945 | 1 | 0.46 |
| 1906 | 2 | 0.636364 | 1946 | 3 | 0.509804 |
| 1907 | 1 | 0.666667 | 1947 | 0 | 0.5 |
| 1908 | 1 | 0.692308 | 1948 | 0 | 0.490566 |
| 1909 | 0 | 0.642857 | 1949 | 1 | 0.5 |
| 1910 | 0 | 0.6 | 1950 | 1 | 0.509091 |
| 1911 | 0 | 0.5625 | 1951 | 0 | 0.5 |
| 1912 | 0 | 0.529412 | 1952 | 1 | 0.508772 |
| 1913 | 0 | 0.5 | 1953 | 0 | 0.5 |
| 1914 | 0 | 0.473684 | 1954 | 0 | 0.491525 |
| 1915 | 0 | 0.45 | 1955 | 0 | 0.483333 |
| 1916 | 0 | 0.428571 | 1956 | 0 | 0.47541 |
| 1917 | 0 | 0.409091 | 1957 | 2 | 0.5 |
| 1918 | 0 | 0.391304 | 1958 | 1 | 0.507937 |
| 1919 | 0 | 0.375 | 1959 | 0 | 0.5 |
| 1920 | 1 | 0.4 | 1960 | 1 | 0.507692 |
| 1921 | 0 | 0.384615 | 1961 | 0 | 0.5 |
| 1922 | 1 | 0.407407 | 1962 | 0 | 0.492537 |
| 1923 | 1 | 0.428571 | 1963 | 1 | 0.5 |
| 1924 | 0 | 0.413793 | 1964 | 1 | 0.507246 |
| 1925 | 0 | 0.4 | 1965 | 1 | 0.514286 |
| 1926 | 0 | 0.387097 | 1966 | 1 | 0.521127 |
| 1927 | 0 | 0.375 | 1967 | 0 | 0.513889 |
| 1928 | 0 | 0.363636 | 1968 | 0 | 0.506849 |
| 1929 | 0 | 0.352941 | 1969 | 0 | 0.5 |
| 1930 | 0 | 0.342857 | 1970 | 1 | 0.506667 |
| 1931 | 1 | 0.361111 | 1971 | 0 | 0.5 |
| 1932 | 1 | 0.378378 | 1972 | 0 | 0.493506 |
| 1933 | 1 | 0.394737 | 1973 | 0 | 0.487179 |
| 1934 | 1 | 0.410256 | 1974 | 1 | 0.493671 |

| Time-Step | Count | ERR | Time-step | Count | ERR |
|---|---|---|---|---|---|
| 1975 | 0 | 0.4875 | 1992 | 0 | 0.412371 |
| 1976 | 0 | 0.481481 | 1993 | 0 | 0.408163 |
| 1977 | 0 | 0.47561 | 1994 | 1 | 0.414141 |
| 1978 | 0 | 0.46988 | 1995 | 0 | 0.41 |
| 1979 | 0 | 0.464286 | 1996 | 0 | 0.405941 |
| 1980 | 0 | 0.458824 | 1997 | 0 | 0.401961 |
| 1981 | 0 | 0.453488 | 1998 | 1 | 0.407767 |
| 1982 | 0 | 0.448276 | 1999 | 0 | 0.403846 |
| 1983 | 0 | 0.443182 | 2000 | 1 | 0.409524 |
| 1984 | 0 | 0.438202 | 2001 | 1 | 0.415094 |
| 1985 | 1 | 0.444444 | 2002 | 0 | 0.411215 |
| 1986 | 0 | 0.43956 | 2003 | 1 | 0.416667 |
| 1987 | 0 | 0.434783 | 2004 | 2 | 0.431193 |
| 1988 | 0 | 0.430108 | 2005 | 1 | 0.436364 |
| 1989 | 0 | 0.425532 | 2006 | 2 | 0.45045 |
| 1990 | 0 | 0.421053 | 2007 | 4 | 0.482143 |
| 1991 | 0 | 0.416667 | 2008 | 0 | 0.477876 |
| | | | 2009 | 1 | 0.482456 |

* January 1, 1896 – December 31, 1896

**Table 3**. ERR with time-step h = 2 years

| Time-Step | Count | ERR |
|---|---|---|
| 1896* | 2 | 1 |
| 1898 | 1 | 0.75 |
| 1900 | 0 | 0.5 |
| 1902 | 1 | 0.5 |
| 1904 | 1 | 0.5 |
| 1906 | 3 | 0.666666667 |
| 1908 | 1 | 0.642857143 |
| 1910 | 0 | 0.5625 |
| 1912 | 0 | 0.5 |
| 1914 | 0 | 0.45 |
| 1916 | 0 | 0.409090909 |
| 1918 | 0 | 0.375 |
| 1920 | 1 | 0.384615385 |
| 1922 | 2 | 0.428571429 |
| 1924 | 0 | 0.4 |
| 1926 | 0 | 0.375 |
| 1928 | 0 | 0.352941176 |
| 1930 | 1 | 0.361111111 |
| 1932 | 2 | 0.394736842 |
| 1934 | 1 | 0.4 |
| 1936 | 0 | 0.380952381 |
| 1938 | 2 | 0.409090909 |
| 1940 | 1 | 0.413043478 |
| 1942 | 2 | 0.4375 |
| 1944 | 2 | 0.46 |
| 1946 | 3 | 0.5 |
| 1948 | 1 | 0.5 |
| 1950 | 1 | 0.5 |
| 1952 | 1 | 0.5 |
| 1954 | 0 | 0.483333333 |
| 1956 | 2 | 0.5 |
| 1958 | 1 | 0.5 |
| 1960 | 1 | 0.5 |
| 1962 | 1 | 0.5 |
| 1964 | 2 | 0.514285714 |
| 1966 | 1 | 0.513888889 |
| 1968 | 0 | 0.5 |
| 1970 | 1 | 0.5 |
| 1972 | 0 | 0.487179487 |
| 1974 | 1 | 0.4875 |
| 1976 | 0 | 0.475609756 |
| 1978 | 0 | 0.464285714 |
| 1980 | 0 | 0.453488372 |

| Time-Step | Count | ERR |
|---|---|---|
| 1982 | 0 | 0.443181818 |
| 1984 | 1 | 0.444444444 |
| 1986 | 0 | 0.434782609 |
| 1988 | 0 | 0.425531915 |
| 1990 | 0 | 0.416666667 |
| 1992 | 0 | 0.408163265 |
| 1994 | 1 | 0.41 |
| 1996 | 0 | 0.401960784 |
| 1998 | 1 | 0.403846154 |
| 2000 | 2 | 0.41509434 |
| 2002 | 1 | 0.416666667 |
| 2004 | 3 | 0.436363636 |
| 2006 | 6 | 0.482142857 |
| 2008 | 1 | 0.48245614 |

* January 1, 1896 – December 31, 1897

**Table 4**. ERR with time-step h = 3 years

| Time-Step | Count | ERR |
|---|---|---|
| 1896* | 2 | 0.666667 |
| 1899 | 1 | 0.5 |
| 1902 | 1 | 0.444444 |
| 1905 | 4 | 0.666667 |
| 1908 | 1 | 0.6 |
| 1911 | 0 | 0.5 |
| 1914 | 0 | 0.428571 |
| 1917 | 0 | 0.375 |
| 1920 | 2 | 0.407407 |
| 1923 | 1 | 0.4 |
| 1926 | 0 | 0.363636 |
| 1929 | 1 | 0.361111 |
| 1932 | 3 | 0.410256 |
| 1935 | 0 | 0.380952 |
| 1938 | 3 | 0.422222 |
| 1941 | 2 | 0.4375 |
| 1944 | 5 | 0.509804 |
| 1947 | 1 | 0.5 |
| 1950 | 2 | 0.508772 |
| 1953 | 0 | 0.483333 |
| 1956 | 3 | 0.507937 |
| 1959 | 1 | 0.5 |
| 1962 | 2 | 0.507246 |
| 1965 | 2 | 0.513889 |
| 1968 | 1 | 0.506667 |
| 1971 | 0 | 0.487179 |
| 1974 | 1 | 0.481481 |
| 1977 | 0 | 0.464286 |
| 1980 | 0 | 0.448276 |
| 1983 | 1 | 0.444444 |
| 1986 | 0 | 0.430108 |
| 1989 | 0 | 0.416667 |
| 1992 | 1 | 0.414141 |
| 1995 | 0 | 0.401961 |
| 1998 | 2 | 0.409524 |
| 2001 | 2 | 0.416667 |
| 2004 | 5 | 0.45045 |
| 2007 | 5 | 0.482456 |

* January 1, 1896 – December 31, 1898

REFERENCES

[1] Bakun, W.H. and Aagaard, B. and Dost, B. (2005). Implication for Prediction and Hazard Assessment From the 2004 Parkfield Earthquake. *Nature*, **437**, 969-974.

[2] Box, G.E.P. and Jenkins G.M. (1976) *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

[3] Brockwell, P.J. and Davis, R.A. (2002). *Introduction to Time Series and Forecasting*. 2nd Edition. Springer-Verlag, New York.

[4] Felzer, K.R. and Abercrombie, R.E. and Ekstrom, G. (2003). Secondary Aftershocks and Their Importance for Aftershock Forecasting. *Bulletin of the Seismological Society of America*, **93**, 1433-1448.

[5] Helmstetter, A. and Kagan, Y.Y. and Jackson D.D. (2006). Comparison of short-term and long-term earthquake forecast models for southern California. *Bulletin of the Seismological Society of America*, **96**, 90-106.

[6] Ho, C.-H. (2008). Empirical recurrent rate time series for volcanism: application to Avachinsky volcano, Russia. *Volcanol Geotherm Res*, **173**, 15-25.

[7] Ho, C.-H. (2010a). Hazard area and recurrence rate time series for determining the probability of volcanic disruption of the proposed high-level radioactive waste repository at Yucca Mountain, Nevada, USA. *Bulletin of Volcanology*, **72**,205-219.

[8] Ho, C.-H. (2010b). *Empirical Recurrence Rate for Point Processes, Data Smoothing and Statistical Inference.* Manuscript in progress.

[9] Hong, L.-L. and Guo, S.-W. (1995). Nonstationary Poisson Model for Earthquake Occurrences. *Bulletin of the Seismological of America,* **85**, 814-824.

[10] Jackson, D.D. and Kagan, Y.Y. (2006). The 2004 Parkfield Earthquake, the 1985 Prediction, and Characteristic Earthquakes: Lessons for the Future. *Bulletin of the Seismological Society of America*, **96**, 397-409.

[11] Kagan, Y.Y. (1993) Statistics of Characteristic Earthquakes. *Bulletin of the Seismological Society of America*, **83**, 7-24.

[12] Ljung, G.M. and Box, G.E.P. (1978) On A Measure of Lack of Fit in Time Series Models. *Biometrika*, **65**, 297-303.

[13] Savage, J.C. and Cockerham, R.S. (1987). Quasi-Periodic Occurrence of Earthquakes in the 1978-1986 Bishop-Mammoth Lakes Sequence, Eastern California. *Bulletin of the Seismological Society of America*, **77**, 1347-1358.

[14] Shearer, R. (1985). Minutes of the National Earthquake Prediction Evaluation Council. *U.S. Geological Survey*, Open-File, 85-507.

# VITA

Graduate College
University of Nevada, Las Vegas

Wandong Fu

Degree:
Bachelor of Science in Applied Mathematics, 2004
Nanjing University of Technology, Nanjing

Thesis Title: ARIMA Models for Forecasting Poisson Data: Application to Long-Term Earthquake Predictions

Thesis Examination Committee:
Chairperson, Chih-Hsiang Ho, Ph. D.
Committee Member, Amei Amei, Ph. D.
Committee Member, Kaushik Ghosh, Ph. D.
Graduate Faculty Representative, LeinLein Chen, Ph. D.