1-1-2006

# An Arima-model-based approach with hazard area for the probability of volcanic disruption of the proposed high-level radioactive waste repositiory at Yucca Mountain, Nevada, Usa

XiaoJuan Liu
*University of Nevada, Las Vegas*

# AN ARIMA-MODEL-BASED APPROACH WITH HAZARD AREA FOR THE PROBABILITY OF VOLCANIC DISRUPTION OF THE PROPOSED HIGH-LEVEL RADIOACTIVE WASTE REPOSITORY AT YUCCA MOUNTAIN, NEVADA, USA

by

XiaoJuan Liu

Bachelor of Science
Nanjing Normal University, China
2001

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**
**December 2006**

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

# Thesis Approval

The Graduate College
University of Nevada, Las Vegas

November 17, 2006

The Thesis prepared by

XiaoJuan Liu

**Entitled**

AN ARIMA-MODEL-BASED APPROACH WITH HAZARD AREA FOR THE PROBABILITY OF VOLCANIC DISRUPTION OF THE PROPOSED HIGH-LEVEL RADIOACTIVE WASTE REPOSITORY AT YUCCA MOUNTAIN, NEVADA, USA

is approved in partial fulfillment of the requirements for the degree of

Master of Science in Mathematical Sciences

*Examination Committee Chair*

*Dean of the Graduate College*

*Examination Committee Member*

*Examination Committee Member*

*Graduate College Faculty Representative*

ABSTRACT

**An ARIMA-Model-Based Approach with Hazard Area for the Probability of Volcanic Disruption of the Proposed High-level Radioactive Waste Repository at Yucca Mountain, Nevada, USA**

by

XiaoJuan Liu

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

An interesting extension of advanced time-series analysis techniques is introduced into the domain of volcanological data exploration. A new and innovative use of the well-known ARIMA method for modeling the recurrence rate of volcanism ranging from simple Poissonian volcanoes to those showing cyclic trends is presented. Specifically, we propose a new tool to fingerprint the eruptive behavior of a volcano, which also links some modeling tools of two of the most developed areas in the literature of statistics: stochastic processes and time series. Valuable modeling and computing insights are discussed using a data set from the volcanic database at Yucca Mountain, Nevada, a potential site for an underground geologic repository of high-level radioactive waste in the USA.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to express my sincerely gratitude and appreciation to my advisor, Dr. Ho, for his warm encouragement, patience, time and dedication to keep me on the right track throughout this undertaking. His excellence in both research and teaching will always be a great example to me.

My deeply indebtedness also gives to respectable committee members, Dr. Ananda, Dr. Catlin and Dr. Qian, for their positive inputs and mentoring during my graduate studies.

I would also like to thank my husband, Hui Li, and my daughter, Julianna Li, for their care and mental support that make me to achieve my goal. Last but not least, I would like to thank my families for their love and support throughout my education.

# CHAPTER 1

## INTRODUCTION

The application of statistical methods to volcanic eruptions is put onto a sound analytical footing by Wickman (1966, 1976) in a series of papers that discuss the applicability of the methods and the evaluation of recurrence rates for a number of volcanoes. Wickman observes that for some volcanoes, the recurrence rates are independent of time. Volcanoes of this type are called " Simple Poissonian Volcanoes." A simple Poisson process had been state-of-the-art (e.g., Crowe et al. 1982; Scandone et al. 1993) until a Power-law process coupled with Bayesian analysis were proposed in a number of studies related to the volcanic hazard assessment of the Yucca Mountain high-level nuclear waste repository site (Ho, 1990, 1991a, 1991b, 1992). Volcanic risk models have advanced along related paths over the last decade. A key parameter for volcanic hazard and risk assessments is the recurrence rate. This becomes a motivation of developing a discrete time series based on the empirical recurrence rates (ERR), which is computed sequentially at equidistant time intervals during an observation period (Ho et al., 2006). It is been demonstrated that the time-plot of the empirical recurrence rates, to be referred as the "fingerprint" or the "ERR-plot" offers the possibility of further insights into the data and it can provide a valuable technical basis for model developments in volcanic hazard and risk assessment studies.

This thesis, firstly, demonstrates how to build a discrete time series based on the

empirical recurrence rates. Basic modeling theory for the ERR time series and the background information of application to the volcanism at Yucca Mountain (YM) regions, Nevada then follow. Secondly, the three stages of identification, estimation, and diagnostics along with several practical modeling techniques are presented with the YM volcanic data. Thirdly, hazard area (Ho et at., 2006) and probability of volcanic disruption of the proposed high-level radioactive waste repository at Yucca Mountain are calculated. General pattern-classification, the potential impacts of this work, and other areas of application are noted.

# CHAPTER 2

## METHOD AND BASIC THEORIES

### 2.1 Time Series Based on the Empirical Recurrence Rates

Let $t_1,\ldots,t_n$ be the time of the $n$ ordered eruptions during an observation period $(t_0,0)$ from oldest to youngest. Then a discrete time series $\{z_\ell\}$ is generated sequentially at equidistant time intervals $t_o+h$, $t_0+2h$, ..., $t_0+\ell h$, ..., $t_0+Nh$ (= 0 = present time). If $t_0$ is adopted as the time-origin and $h$ as the time-step, then $z_\ell$ can be regarded as the observation at time, $t=t_0+\ell h$, for the volcanism to be modeled. A key parameter, most sought after by the modelers of volcanic hazard and risk assessments, is the recurrence rate of targeted volcanism worldwide. Therefore, a time series of the empirical recurrence rates is proposed and is defined as follows:

$$z_\ell = n_\ell / \ell h = \text{total number of eruptions in} (t_0, t_0+lh)/lh,$$

where $\ell$=1, 2, ..., $N$. Note that $z_\ell$ evolves over time and it is simply the MLE of the mean, if the underlying process observed in $(t_0,\ t_0+\ell h)$ is a simple Poisson process. The time-plot of the empirical recurrence rate (ERR-plot) offers the possibility of further insights into the data. Also, suppose, starting at time $T$, that a value $z_{T+k}$, $k \geq 1$ is needed to be predicted based on the sample observation $(z_1,\ldots,z_T)$ of an ERR time series. This forecast is said to be made at (forecast) origin T for lead time (or forecast horizon) $k$. In

a regression situation, let $X$ denote the time index, $z$ 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 class of models is introduced.

## 2.2 ARIMA Model

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

ARIMA modeling involves three stages. The first stage is to identify the model. Identification consists of specifying the appropriate model (AR, MA, or ARMA) and order of model. Identification is sometimes done by looking at plots of the sample autocorrelation function (ACF) and sample partial autocorrelation function (PACF). Sometimes identification 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 third stage is to check the model. This step is also called diagnostic checking. One of the two important elements of checking is to ensure that the residuals of the model are random and normally distributed; the other 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 one who has fewest parameters among all models that fit the data.

Definition Stationarity and white noise ( Peña et al., 2001)

The assumption of stationarity has various forms and we state first the weak form, that

1. $E(z_t) = \mu_t$ is constant for all t
2. $Var(z_t) = \sigma_z^2$ is constant for all t
3. $Cov(z_{t-k}, z_t) = \gamma_{z,k}$ depends only the separation lag k and not on t

The sequence $\gamma_{z,k}$ is the autocovariance function of the series and, dropping the suffix $z$ for simplicity, $\rho_k = \gamma_k / \gamma_0$ is the autocorrelation function. Strict stationarity of a time series means that the probability density functions of $(z_{t_1}, ..., z_{t_1+k})$ and $(z_{t_2}, ..., z_{t_2+k})$ are of identical forms for any arbitrary choice of the integers $(t_1, t_2, k)$. In practice, this is saying that the overall behavior of the series remains the same over time. Also, a stationary time series (mean = 0) for which there is no autocorrelation is known as white noise.

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 (auto). 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

$$z_t = \phi z_{t-1} + \varepsilon_t$$

where $t = 1, 2, ..., N$, $z_t$ is a stationary zero-mean time series. 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 $\varepsilon_t$ is analogous to the regression residuals and represents a "white noise" (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

$$z_t = \varepsilon_t - \theta\, \varepsilon_{t-1},$$

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

The basic AR (1) and MA (1) models are insufficient to describe the autocorrelation structure of time series in most cases. For the more complex situations, there is a general Box-Jenkins ARIMA model, built on the simpler AR (1) and MA (1), may be more appropriate for time series data. They are contained in many books and are summarized in the Appendix.

# CHAPTER 3

# APPLICATION

## 3.1 Data

In the Nuclear Waste Policy Act of 1982, the US Congress directed the Department of Energy (DOE) to investigate potential sites for the location of an underground geologic repository to contain the growing volume of high-level radioactive waste. In 1987, Congress amended the Act, directing DOE to study only Yucca Mountain (YM), Nevada, USA. As the first US DOE nuclear program subject to external regulation, the YM Site Characterization Project is one of the most closely reviewed programs ever undertaken by the federal government.

The following application is motivated by the recent developments in connection with the studies of volcanic risk to the proposed high-level radioactive waste repository at YM. We commence the investigation with an YM database containing 33 dates (Smith et al., 2002, and references therein). Quaternary events [1.6 Ma, 0) in the YM region include:

(1) 0.08 Ma Center: Lathrop Wells

(2) 0.4 Ma Centers (2 events): Sleeping Butte Cones

(3) 0.9 Ma Centers (2 events): Little Cone

(4) 1.0 Ma Center: Black Cone

(5) 1.0 Ma Center: Red Cone

(6) 1.2 Ma Center: Northern Cone

Pliocene volcanic events [5.3 Ma, 1.6 Ma) in the YM region include:

(1) 2.7 Ma Center: Buckboard Mesa

(2) 3.7 Ma Centers (2 events): Pliocene Crater Flat

(3) 3.7 Ma Centers (5 events): Aeromagnetic buried centers

(4) 4.8 Ma Center: Thirsty Mesa

Post-12-Ma events [12 Ma, 5.3 Ma) in the YM region include:

(1) 6.8 Ma Centers (2 events): Basalt of Nye Canyon

(2) 7.2 Ma Centers (2 events): Basalt of Nye Canyon

(3) 8.0 Ma Center: Basalt of Rocket Wash

(4) 8.5 Ma Centers (2 events): Basalt of Paiute Ridge

(5) 8.7 Ma Center: Basalt of Scarp Canyon

(6) 8.8 Ma Center: Basalt of Pahute Mesa

(7) 9.0 Ma Center: Basalt of Pahute Mesa

(8) 9.1 Ma Center: Basalt of Pahute Mesa

(9) 10.0 Ma Center: Solitario Canyon Dike

(10) 11.0 Ma Center: Jackass Flat basalt

(11) 11.0 Ma Center: SE Crater Flat basalt

(12) 11.2 Ma Center: Jackass Flat basalt

(13) 11.2 Ma Center: SE Crater Flat basalt

A very important issue in the sensitivity analysis is to specify the observation period, $(t_0, 0)$, in modeling the volcanic history at YM. All the dates were recorded later

than 12 Ma, which is adopted as time-origin for the following tests analysis. The aggregated volcanic eruptive episodes are presented by a dot plot (Figure 1A). It is clear that the dot plot has limited value in delivering the information behavior presented by the data.

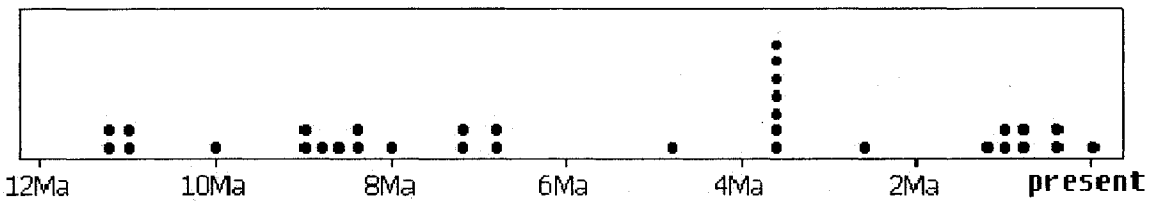


Figure 1A Dot plot of raw data

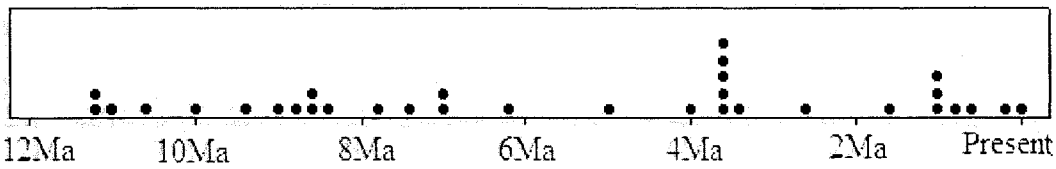


Figure 1B Dot plot of the smoothed raw data

For further development, data smoothing techniques are considered. The most common technique is "the moving average smoothing" (Kutner et al., 2004), which uses the mean of the adjacent z values to obtain the smoothed values. This smoothing technique, using 3 adjacent z values, was first applied to the raw data and the result is displayed in Figure 1B. The ERR-plot based on the smoothed raw data is shown in Figure 1C. Note that: (1) the ERR-plots presented in this thesis are using 12.0 Ma as the time-origin and 0.1 m.y. for the time-step (a total of 120 time-steps); (2) we keep the first and the last values of the original data after smoothing. So, the total number of the time steps

remains the same; and (3) for the sake of simplicity, the unit of the time series is consistently presented as annual rate (number of eruptions per year). In contrast, the process was reversed to smooth the time series produced directly from the raw data (Figure 1A), and the resulting ERR-plot is displayed in Figure 1D. Clearly, there is a similarity in their patterns. However, the smoothing technique appears to be more effective in Figure 1D than Figure 1C. Therefore, the data based on the smoothed ERR-plot (Figure 1D) are used for further model development.

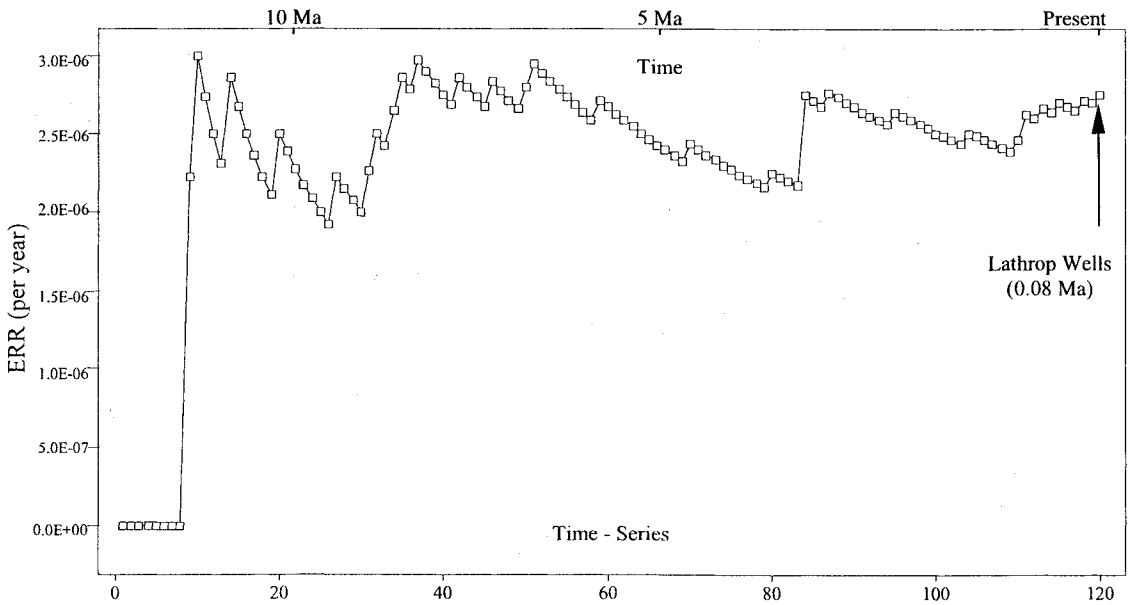


Figure 1C ERR-plot for the smoothed raw data

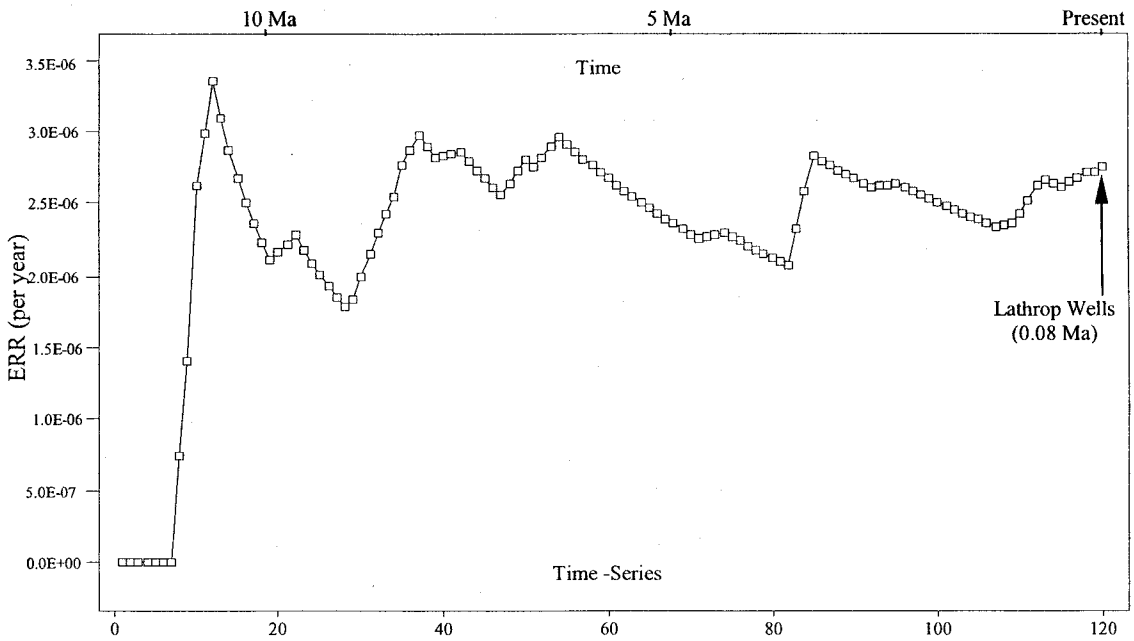

Figure 1D Smoothed ERR-plot using the raw data

## 3.2 Pattern Classification via ARIMA

### 3.2.1 Plotting Data

The ERR-plot, exhibited in Figure 1D starts with 7 zeros due to the selected time-origin, which causes a spike at lag 8. Therefore, a revised time series excluding the first seven data points (Fig. 2A, with 113 time-steps) is used for further analyses.

### 3.2.2 Ljung-Box Test for lack of fit in time series models

Ljung-Box Test, proposed by Ljung and Box (1978), is commonly used in ARIMA modeling for checking whether the residuals or noise sequence of a fitted model are independent and identically distributed random variables (iid). It is based on the autocorrelation plot, and it tests the overall independence based on a number of lags. Because of which, it is often referred to as a portmanteau test. More formally, the Ljung-Box test can be defined 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 (As a rule of thumb, the sample ACF and PACF are good estimates of the ACF and PACF of a stationary process for lags up to about a third of the sample size.)

$\hat{a}_1,...,\hat{a}_n$ 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 $\alpha$ , and therefore, there is dependence among the sequence data, or the sequence data do have sample autocorrelations significantly different from zero.

The sample value of the Ljung-Box statistic $\hat{Q}$ with $m = 20$ is 282.6 for the series data $z_1,...,z_n$ based on Figure 1D. The corresponding $p-value$ displayed by ITSM (Brockwell and Davis, 2002) is 0.000 < 0.05. Therefore, the hypothesis of iid is rejected at level 5%, which implies that the series are not stationary and there is significant evidence that there is autocorrelation among the $z_i$'s .

### 3.2.3 Differencing

Differencing is a data processing step, which attempts to de-trend to control autocorrelation and achieve stationary by subtracting each datum in a series from its predecessor. For example, single differencing is used to remove linear trends; double differencing is used to remove quadratic trend. Furthermore, the volcanism displayed in Figure 2A exhibits seasonal component (or seasonality, a statistical term) with peaks occurring at the following time steps: 11, 36, 54, 85, and 113. This distinctive signature, marked by systematic peaks and troughs, can be described as cyclical volcanism with a gradually stabilizing period of approximately 25 time-steps or 2.5 m.y. In order to remove this seasonal component with a period approximately equal to 25 from the series of Figure 2A, $\{z_t\}$, we generate the transformed series ( differencing at lag 25),

$$\nabla_{25} z_t = z_t - z_{t-25}.$$

Note that with each degree of differencing, the time series is shortened by one. Figure 2B shows the transformed series by differencing at lag 25. Inspection of the graph (Figure 2B) suggests a further differencing at lag 1 to eliminate the remaining trend. Once the apparent deviations from stationarity of the data have been removed, the sample mean is then subtracted from each observation of the twice-differenced series to generate a "mean-corrected" series. The resulting series is now stationary with zero mean and is displayed in Figure 2C. Note that a full analysis that allows for changing periodicity is beyond the scope of this thesis.

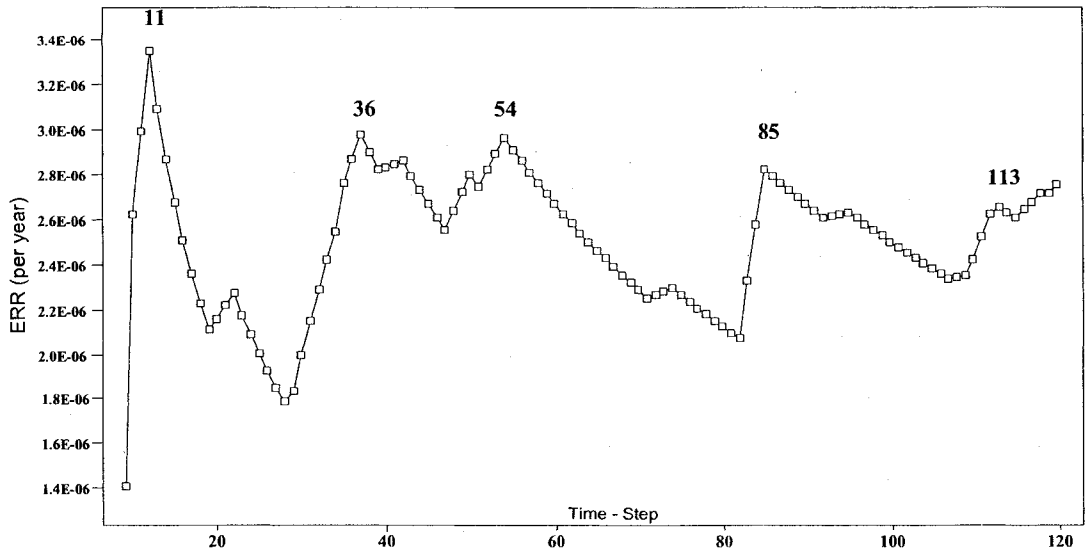


Figure 2A ERR-plot after dropping zeros

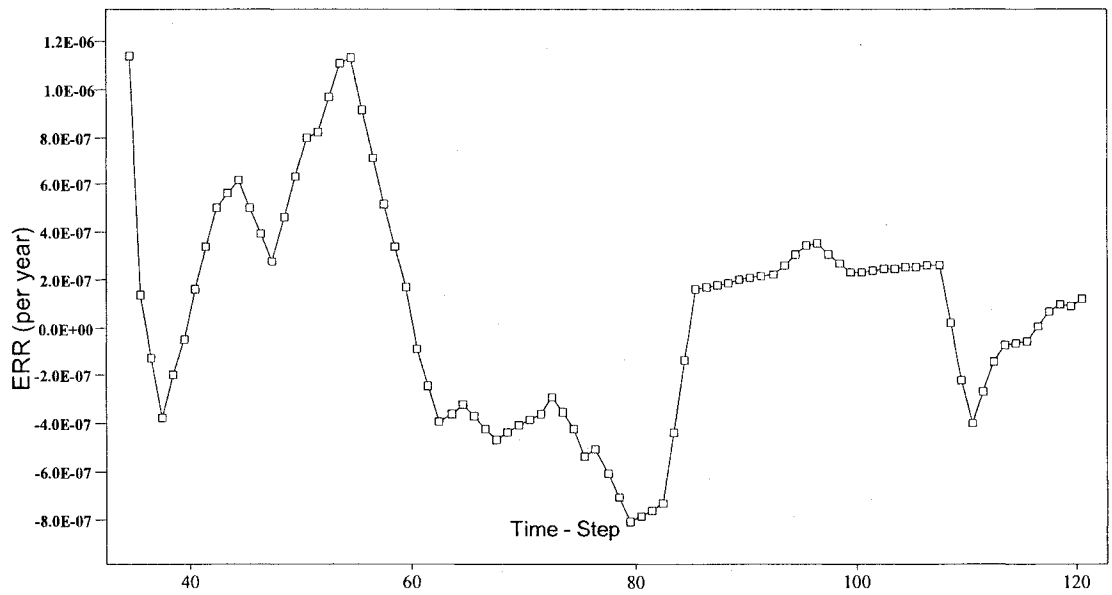


Figure 2B ERR-plot after differencing at lag 25 ( $\nabla_{25} z$ )

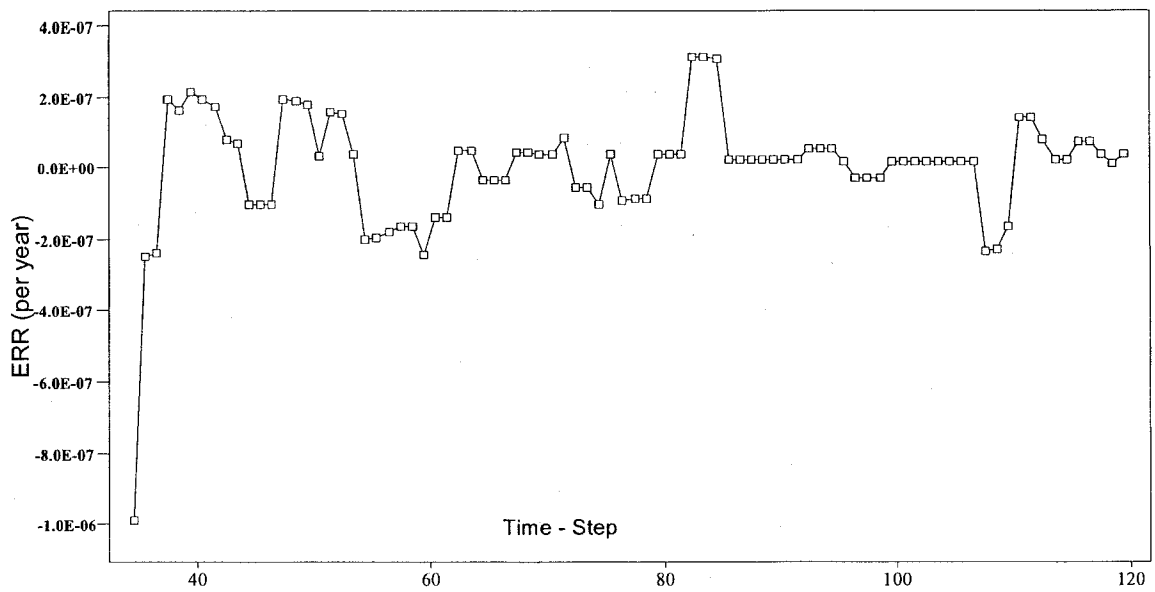


Figure 2C ERR-plot for the "mean corrected" and twice-differenced data ( $\nabla\nabla_{25}z$ )

### 3.2.4 Sample ACF and PACF

After a time series has been stationarized by differencing, the next step in fitting an ARIMA model is to determine AR or MA terms, needed to correct any autocorrelation that remains in the differenced series. This can be tentatively done by looking at the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the differenced series. The sample ACF plot is merely a bar chart of the coefficients of correlation between a time series and lags of itself. The PACF plot is a plot of the partial correlation coefficients between the series and lags of itself. The sample ACF of the data are shown, respectively, in Figures 3A, 3B (after differencing at lag 25), and 3C (after differencing twice). A persistently high sample ACF signals the need for differencing. Figure 3A supports the above argument and suggests that seasonal differencing with period 25 might work. The sample PACF of the data shown in Figures 4A, 4B

(differencing at lag 25), and 4C (after differencing twice) is another convenient tool for tentative model specification. A low order moving-average model is suggested by sample ACF exhibiting a small number of large values at low lags, and a low order autoregressive model is suggested by sample PACF marking a similar "cutting off" pattern.

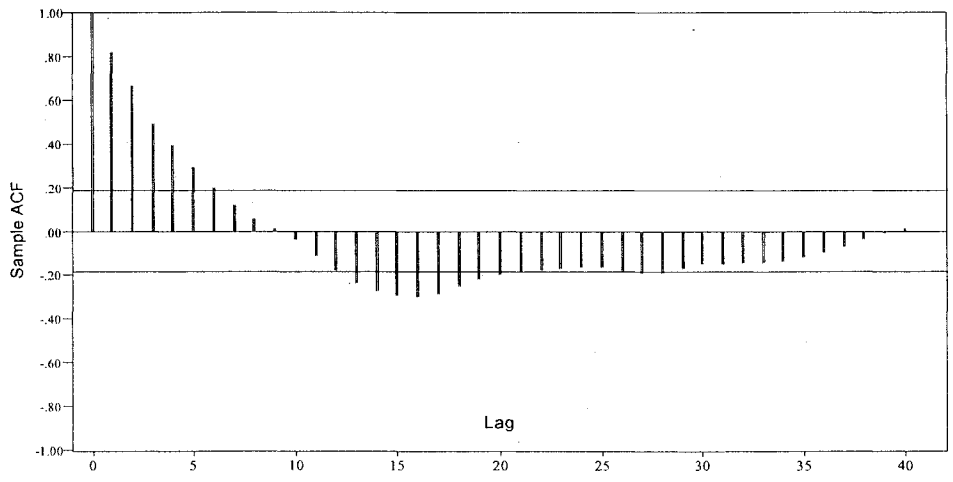


Figure 3A Sample ACF of the series data $z$

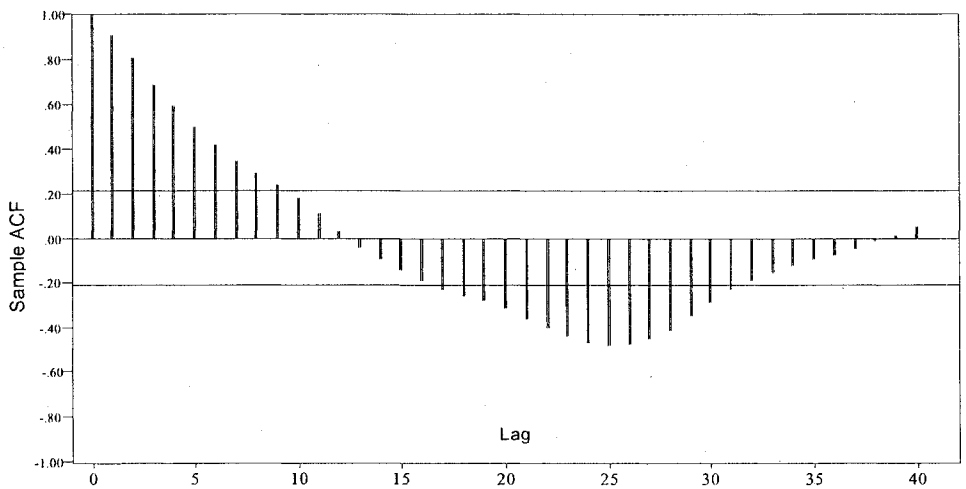


Figure 3B Sample ACF of the series data after differencing at lag 25 ($\nabla_{25}z$)

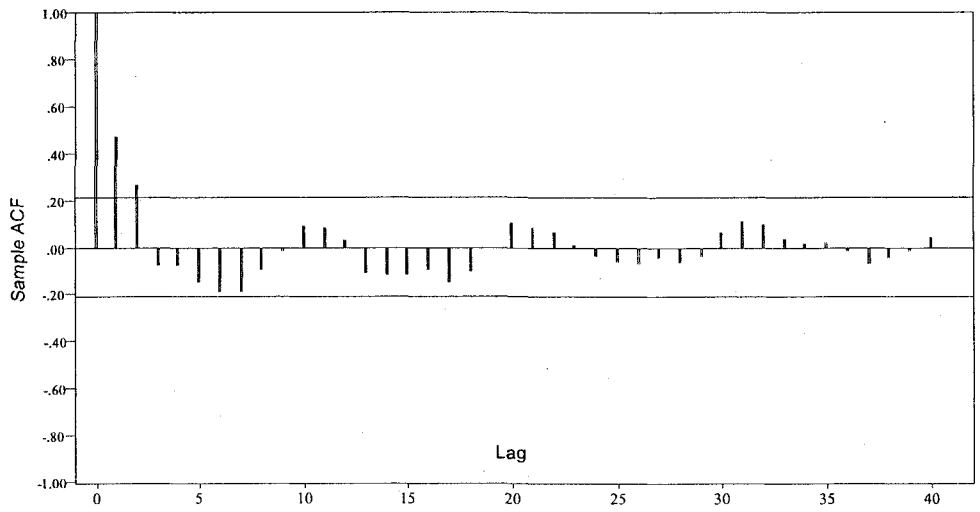


Figure 3C Sample ACF of the series data after differencing twice ($\nabla\nabla_{25}z$)

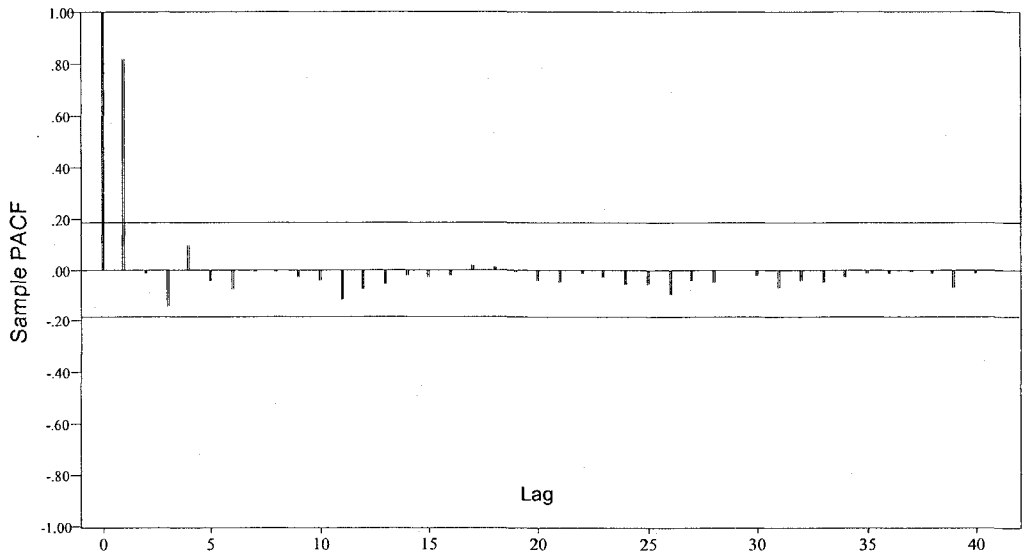


Figure 4A Sample PACF of the series data z

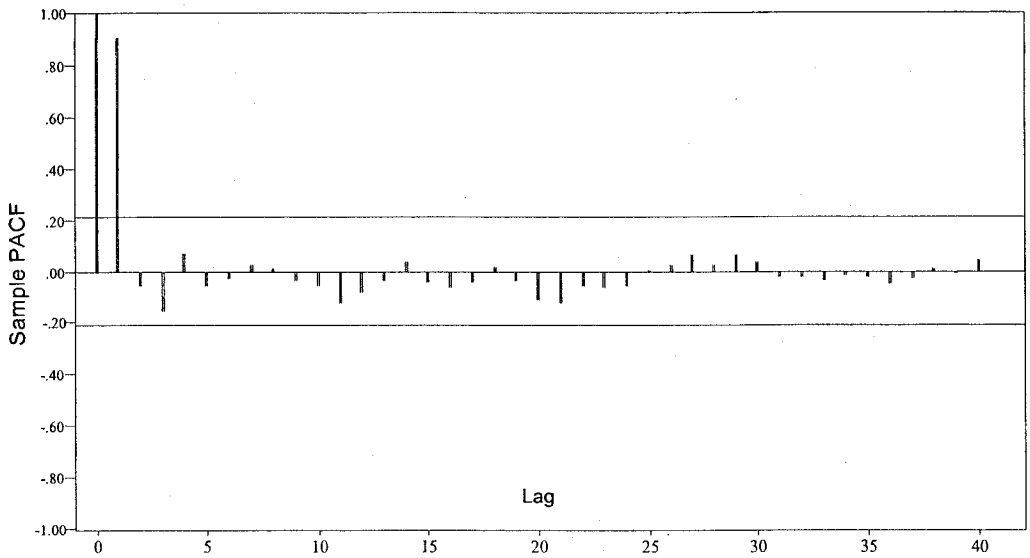


Figure 4B Sample PACF of the series data after differencing at lag 25 ($\nabla_{25} z$)

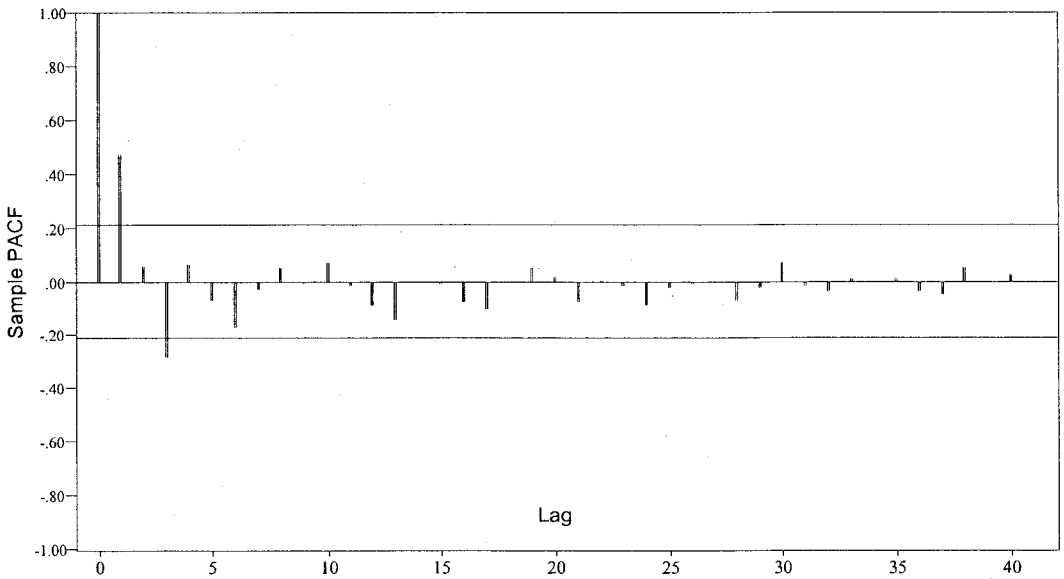


Figure 4C Sample PACF of the series data after differencing twice ($\nabla\nabla_{25} z$)

### 3.2.5 Entering a Model

The horizontal lines on the graphs of the sample ACF and sample PACF are the bounds $\pm 1.96/\sqrt{N}$ (N = the sample size). If the data constitute a large sample from an independent white noise sequence, approximately 95% of the sample autocorrelations should lie between these bounds. As a rough guide, if the sample ACF falls between the plotted bounds $\pm 1.96/\sqrt{N}$ for lags $h > q$, then an MA ($q$) model is suggested, while if the sample PACF falls between the plotted bounds $\pm 1.96/\sqrt{N}$ for lags $h > p$, then an AR ($p$) model is suggested. If neither the sample ACF nor sample PACF "cuts off" as previously described, a more refined model selection technique is required. Even if the sample ACF or sample PACF does cut off at some lag, it is still advisable to explore models other than those suggested by the sample ACF and sample PACF.

Figures 3C and 4C show the sample ACF and sample PACF of the time series $\nabla\nabla_{25} z_t$. These graphs suggest considering an MA model of order 2 since sample ACF seems to cut off at 2, or alternatively an AR model of order 3 since sample PACF seems to cut off at 3. In other words, these characteristics of the sample ACF and sample PACF suggest models without a seasonal component; the ARIMA $(p,1,q)\times(0,1,0)_{25}$ could be fitted to the time series $z_t$.

### 3.2.6 AIC, BIC and AICC Statistics

The AICC statistic, the bias-corrected version of the AIC statistic (Akaike, 1974), is the information criterion used in this thesis to help search for an appropriate model in the ITSM package (Brockwell and Davis, 2002). Smallness of AICC value is indication of a good model, but it should be used only as rough guide. Final decisions between models should be based on maximum likelihood estimation. Model-selection statistics other than

AICC are also available in ITSM. A Bayesian modification of the AIC statistic known as the BIC statistic (Schwarz, 1978) 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 $r$; so to minimize the values of these criteria is to minimize the number of parameters. Therefore, the best model is the model adequately describes data and has fewest parameters.

### 3.2.7 Model Diagnostics

Models MA (2), AR (3), and several ARMA$(p,q)$ with $0 \le p,q \le 6$ are considered here to fit the time series $\nabla\nabla_{25}z_t$. For each model, AICC value was evaluated and a set of diagnostic plots (not displayed here) including the residual sample ACF and sample PACF were produced by the ITSM package (Brockwell and Davis, 2002). After testing these models, we narrow down to two models MA(2) and ARMA(1, 1). For the model MA(2), ITSM gives the value AICC = -2512 while model ARMA (1, 1) has AICC = -2478. Due to the lower AICC value criterion, the final choice of the model is MA(2). Its residual ACF and PACF plots (Figures 5B and 5C) exhibiting no significant spike. The portmanteau goodness-of-fit test (Ljung and Box, 1978) is not significant ($p$-value =

0.97), indicating that the residuals (Fig. 5A) are approximately white noise but it is also heteroscedastic (it has changing variance). Therefore the ARIMA $(0,1,2)\times(0,1,0)_{25}$ model, seems to be an appropriate model for $z_t$, and the estimated (MLE) model is

$\nabla\nabla_{25}z_t = 0.8318\hat{\varepsilon}_{t-1} + 0.9911\hat{\varepsilon}_{t-2} + \hat{\varepsilon}_t$, and $\hat{\sigma}_\varepsilon^2 = 1.034\times10^{-14}$

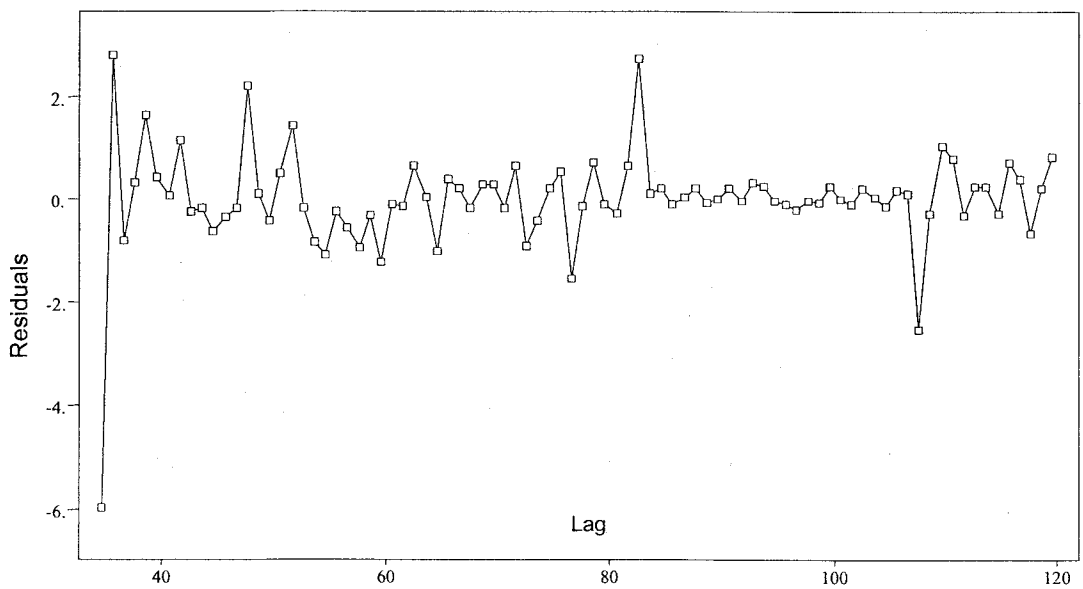


Figure 5A Time plot of residuals after fitting MA (2) model

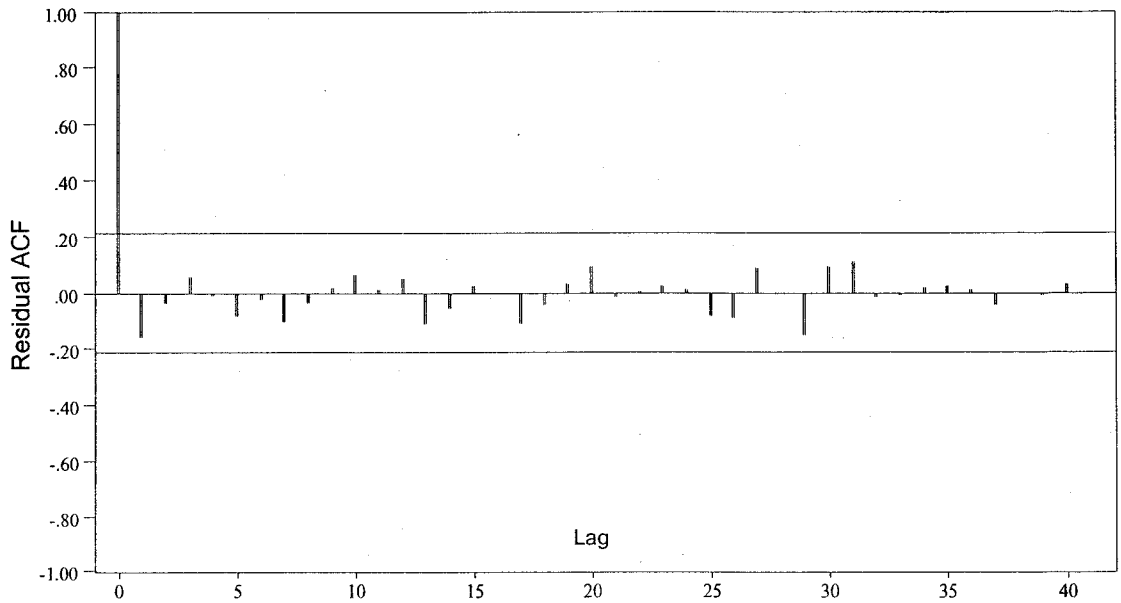


Figure 5B ACF of residuals after fitting MA (2) model

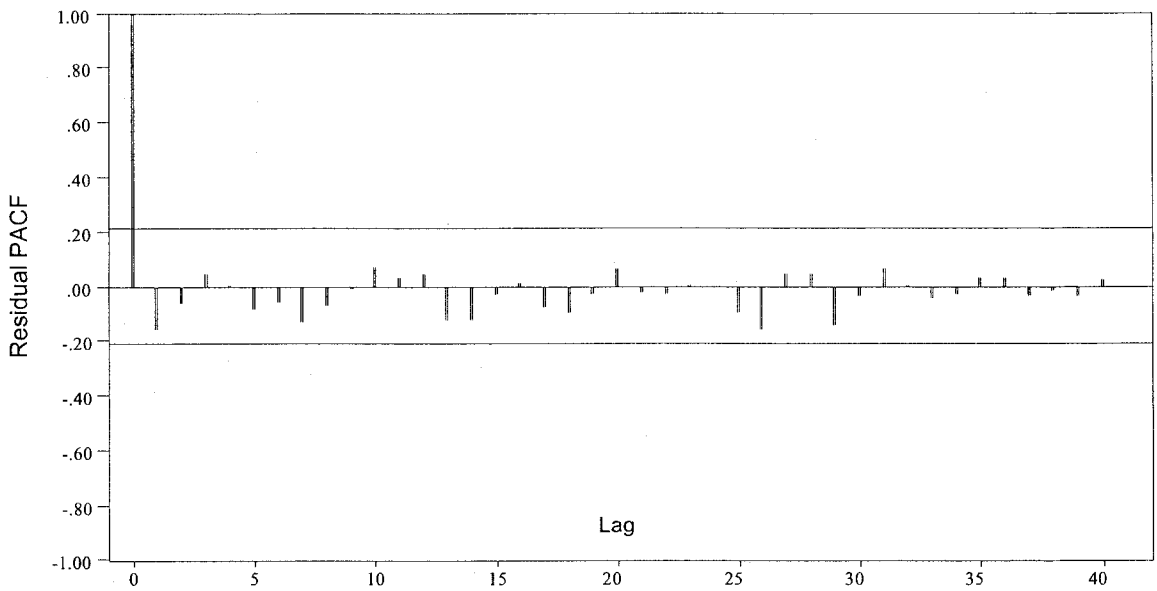


Figure 5C PACF of residuals after fitting MA (2) model

### 3.2.8 Forecasting

An ARIMA model (or other time series model) predicts future values of the time series from past values. The forecast function $z_t = f(z_{t-1},\ldots,z_1)+a_t$ is minimum mean square error forecast. The first part of the above equation $f(z_{t-1},\ldots,z_1)$ is a function of the past values of the series and it should be determined by the data while the second part $a_t$ is a sequence of independent and identically distributed (iid) variables. This part is also called noise part, which is independent from previous values and hence it is unpredictable from its past values. In some cases, obtaining the structure of the function $f$ is the main objective of the analysis while in other cases our interest is mostly in getting forecasts.

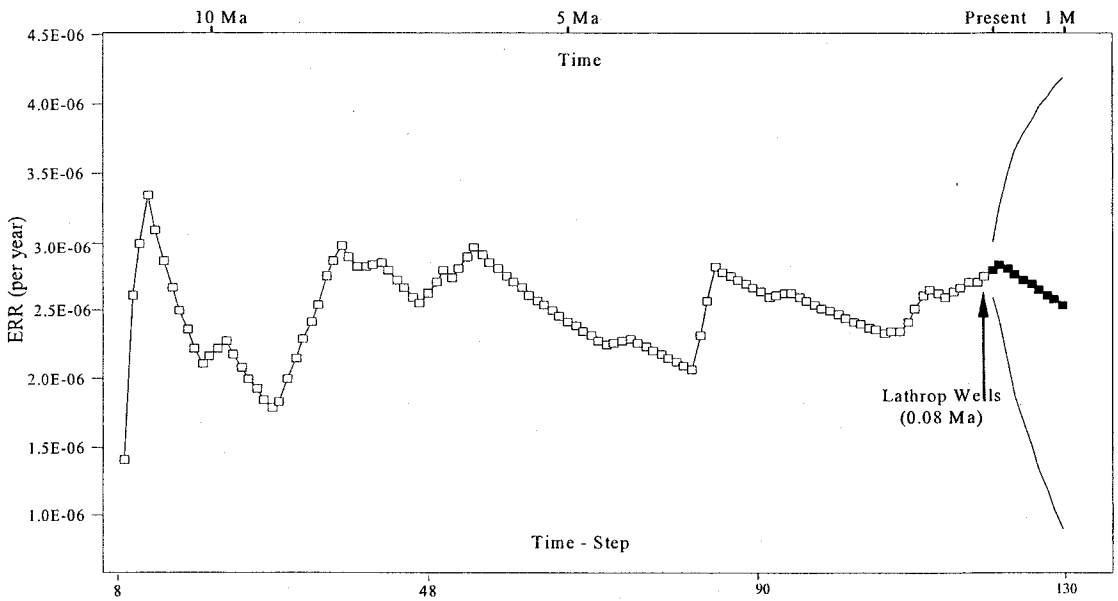


Figure 6 ERR-plot with 10 forecasts appended and 95% confidence bounds

For the data in this thesis, the one-step-ahead forecast (0.1 m.y. from now) for future recurrence rate $z_{N+1}$, is $2.8030\times10^{-6}$ per year, which should not be linearly used to predict longer horizons because the recurrence rate is not constant in this case. For the purpose of pattern recognition, we produce Figure 6 to depict the YM data with 10 forecasts and 95% confidence bounds appended. The confidence bounds are necessarily wider for predictions with longer horizons. They predict a short-term waning trend, concluding the present cycle until a new one (trough to trough) commences at about 1 m.y. later, while maintaining a similar momentum for the long-term forecasting.

Furthermore, a 95% confidence interval, (LB, UB), can be calculated for $z_{N+1}$, and it is $2.60\times10^{-6}$ per year to $3.00\times10^{-6}$ per year. That is, at the 95% confidence level, the model predicts scenarios of 0 to 3.33 $(=3.00\times120\times0.1-33)$ new events that may occur in the next 0.1 m.y., which lays a solid groundwork for the probabilistic estimation of the repository site disruption, to be discussed in the next chapter. Apparently, the predicted lower bound (LB = $2.60\times10^{-6}$ per year) is not valid in this case and needs to be adjusted because the way the ERR is defined depends, effectively, on the cumulative sum of past events. Thus, a meaningful lower bound for every future recurrence rate should be adjusted to reflect the maximum of the following two values: the predicted LB and the rate calculated by incorporating zero future events. Table 1 shows 10 forecasts with the adjusted 95% confidence prediction bounds generated from ARIMA $(0,1,2)\times(0,1,0)_{25}$. The estimate future recurrence rates peak at the second time-step and decrease all the way to the end from there (2.5536 to 2.8450 eruptions per m.y.). Also, the adjusted 95% prediction bounds for the next 1 m.y., ranging from 2.5384 to 4.1968 (eruptions per m.y.) will be used to bound the probability of site disruption in Chapter 4.

Table 1 Ten ERR predictions (first to tenth-step-ahead forecasts) from ARIMA $(0,1,2)\times(0,1,0)_{25}$: the length of time-step is 0.1m.y; the numbers are annual rate$\times 10^6$

| Lead time | Prediction | 95% prediction Lower bound (adjusted) | 95% prediction Upper bound |
|---|---|---|---|
| 1 | 2.8030 | 2.7273 | 3.0023 |
| 2 | 2.8450 | 2.7049 | 3.2610 |
| 3 | 2.8068 | 2.6829 | 3.5066 |
| 4 | 2.7692 | 2.6613 | 3.6671 |
| 5 | 2.7321 | 2.6400 | 3.7918 |
| 6 | 2.6954 | 2.6190 | 3.8952 |
| 7 | 2.6592 | 2.5984 | 3.9844 |
| 8 | 2.6235 | 2.5781 | 4.0633 |
| 9 | 2.5883 | 2.5581 | 4.1341 |
| 10 | 2.5536 | 2.5384 | 4.1968 |
| Maximum | 2.8450 | | |
| Minimum | 2.5536 | | |
| Mean | 2.7076 | | |
| Median | 2.7138 | | |

# CHPATER 4

# HAZARD AREA AND PROBABILITY OF VOLCANIC DISRUPTION

## 4.1 Hazard Area

Models that calculate the probability that a new volcano or a dike from a nearby eruption will intersect the footprint of the proposed high-level nuclear waste repository are generalized by Ho et al. (2006) based on a conceptual model developed for the space transportation industry. The proposed hazard area, defined such that every new eruption that occurs there will disrupt the repository, plays a fundamental role in developing probability models. This hazard area is used not only to hedge the uncertainties in predicting patterns of future volcanic activity, but also to account for the characteristics of a new eruption during the post-closure performance period of an underground geologic repository.

In space transportation industry, the licensing for the execution of a commercial space launch and reentry is directed by the US Federal Aviation Administration (FAA) Office of the Associate Administrator for Commercial Space Transportation. This licensing process is established to limit risks to public health, public safety, and the safety of property, as well as to ensure national security and foreign policy interests of the United States.

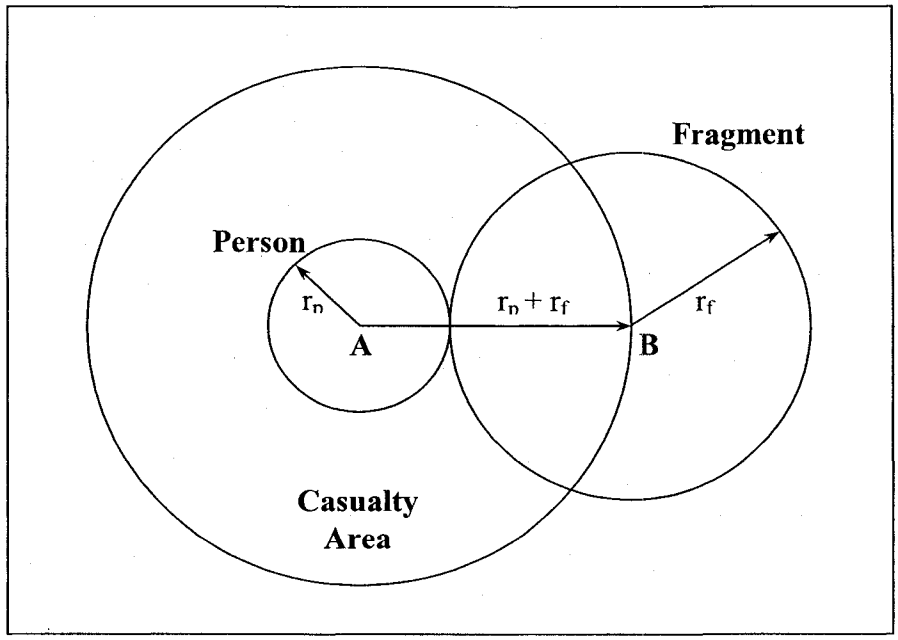


Figure 7 Casualty area for fragment falling vertically (FAA 2000, Figure 1) $r_p$ = radius of person (1 ft); $r_f$ = radius of the fragment

The concept of "casualty area" is involved in one of the factors that will be considered by the US government before approving the licensing of a commercial launch. This "casualty area" for each piece of vehicle debris is determined by finding the area where 100% of the exposed population on the ground is a casualty, specifically defined as any human contact with vehicle debris that can cause injury or any exposure to explosive pressure 0.25 kg/cm2 or greater. A sample case for determining the casualty area for the simplest scenario is demonstrated in Figure7 (FAA 2000, Figure 1). For this example, the desired casualty area for a vertically falling inert piece of debris is a circle whose radius is the sum of the radius of a circle enclosing the largest cross sectional area of the piece and the radius of a human being (1.0 ft).

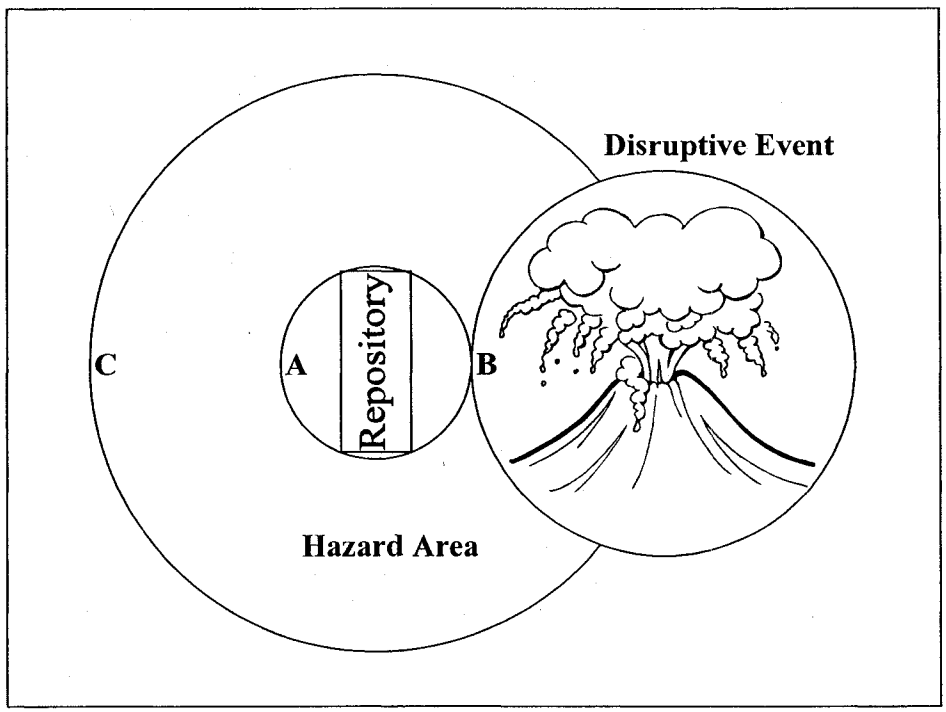


Figure 8 Hazard area for a disruptive event (Ho et al., 2006, Figure 2). *Circle A* represents a minimal circle enclosing the repository; *circle B* quantifies the effective size (including the associated dike and lava) of a disruptive eruption; *circle C*, with radius the sum of those of *A* an *B*, is the desired area and is referred as "hazard area" in the text

Great similarities are found between volcanic hazard area and those of licensing commercial space missions. Thus, the comprehensiveness of FAA's approach provides an acceptable alternative to worldwide modelers of volcanic hazard and risk studies. Therefore, the following two-dimensional transformation from Figure 7 to Figure 8 is straightforward:

1. The circle representing a person is replaced with a minimal circle (A in Figure 8) enclosing the repository. This circle may be generalized to an ellipse or another irregular shape depending on geologic structures of the target sites or other controlling factors.

2. The circle depicting a vertically falling inert piece of debris becomes a circle (B in Figure 8) quantifying the effective size (including the associated dike and lava) of a disruptive eruption. This area is quite flexible in providing likely bounds for uncertainties associated with the magnitude of future eruptions.

3. The largest circle (C in Figure 8), with radius the sum of those of circles A and B, is the desired area to be referred as "hazard area" in the following development.

Knowing that the casualty area for each piece of debris is the area within which 100% of the unprotected population on the ground is assumed to be a casualty. Analogously, the hazard area, in a defined volcanic field, is the area where every new eruption will disrupt the repository. Hence, the probability of a volcanic site disruption is equal to the chance that a new eruption occurs within the hazard area. Furthermore, repository failure modes, justified by geologically meaningful scenarios of a volcanic disruption (or consequence models), will facilitate the definition of the hazard area.

## 4.2 Probability of Volcanic Disruption

Assuming that the compliance period is $(0, t)$, a simple way to represent the probability of site disruption is:

$p_{sd}$ = P [site disruption event occurs during $(0, t)$]

= P [at least one volcanic event occurs in $(0, t)$, which disrupts the repository]

= P [at least one event occurs in $(0, t)$] × P [events occur within the hazard area]

$$= P_e \times P_h \qquad (1)$$

In general, evaluations of $P_e$ and $P_h$ in equation (1) depend on the probability models fitted to the targeted volcanism. For the following parameter estimates, a homogeneous Poisson process (HPP) is assumed to model future eruptions. Therefore, For future YM volcanism, the model assumption of an HPP leads equation (1) to (Ho et al., 1991a, and Smith, 1998):

$P_e$ = P [at least one event occurs in $(0, t)$ ]

$$= 1 - \exp(-\lambda t) \tag{2}$$

$P_h$ = P [events occur within the hazard area]

$$= \pi(r_s + r_d)^2 / A \tag{3}$$

$$P_{sd} = [1 - \exp(-\lambda t)] \times [\pi(r_s + r_d)^2 / A] \tag{4}$$

where,

$\lambda$ = recurrence rate of the volcanism

$t$ = observation period

$r_s$ = radius of a circle enclosing the repository

$r_d$ = radius of a circle quantifying the size of the eruption

$A$ = area of the defined volcanic field

The k-step-ahead forecast (Table 1) for future recurrence rate, $z_{N+k}$, ($k = 1,2, \ldots, 10$), based on ARIMA $(0,1,2) \times (0,1,0)_{25}$ , will be used to evaluate $P_e$, and consequently, the probability of site disruption $p_{sd}$ .

### 4.2.1 Estimates of Future Recurrence Rates and $P_e$

For the following development, we assume that the compliance period is 1 m.y. into the future. Therefore, the value of t in equation (2) for $P_e$ is $10^6$. The confidence bounds concluded from Table 1 will be used to estimate the other parameter, $\lambda$, for $P_e$. The values are 2.5384 to 4.1968 (eruptions per m.y.). Therefore, assuming that the future eruption follows a simple Poisson process, the estimated probability that at least one eruption occurs at the YM region during the next 1 m.y. ($=P_e$) ranges from 0.9210 to 0.9850.

### 4.2.2 Estimates of $P_h$

The area of the actual repository is currently undetermined but is estimated to be 6-8 $km^2$, which prescribes a circle with a radius, $r_s \approx 1.5\,km$ for the hazard area. The area of the defined volcanic field, $A = 3,532\,km^2$, was obtained (Ho et al., 2006) by setting the probability of Crowe et al. (1982) to match the base value, $r_d = 0$. Although the soundness of $A = 3,532\ km^2$ remains to be challenged, for the sake of consistency, we shall use the same value for the following calculations. In addition, the values of "$r_d$ equivalence" are calculated by Ho et al. (2006) as 1.85 and 6.0 km, respectively, for $P_h = 0.01$ (Sheridan, 1992) and 0.05 (Ho, 1992), using the same set of known parameter values. Therefore, we shall use 0, 1.85, and 6.0 km for $r_d$ to evaluate $P_h$.

### 4.2.3 Probability of Site Disruption: $p_{sd}$

We now are ready to link equation (4) to the two components, $P_e$ and $P_h$, defined in equation (1). And the calculated results, incorporating all the parameters previously estimated, of the probability of site disruption, $p_{sd}$, are summarized in Table 2.

Table 2 Probability of site disruption ( $p_{sd}$, during the next 1 m.y.) summary for 3 sizes of eruption, $r_d$

| $r_d$ | $p_e = 0.921$ | $p_e = 0.985$ |
|---|---|---|
| 0 | $1.842\times10^{-3}$ | $1.97\times10^{-3}$ |
| 1.85 | $9.189\times10^{-3}$ | $9.827\times10^{-3}$ |
| 6 | $4.606\times10^{-2}$ | $4.926\times10^{-2}$ |

In conclusion, the probability of volcanic disruption of the proposed high-level radioactive waste repository at YM for the next 1 m.y. is bounded by $1.842\times10^{-3}$ and $4.926\times10^{-2}$ for $r_d$ ranging from 0 to 6 km.

# CHAPTER 5

## CONCLUSIONS

In this thesis we showed tremendous merits in building a linking bridge between a point process and the classical time series via a sequence of the empirical recurrence rates, calculated sequentially at equidistant time intervals. The distinctive technique, generating the unique eruptive pattern of a volcano or a volcanic field, is demonstrated with an empirical recurrence rate plot (ERR-plot), designed to fingerprint the temporal pattern of the targeted volcanism.

We also presented a strategy for the evaluation and use of "hazard area" based on a model developed for licensing commercial space launch and reentry operations in the space transportation industry. We assumed that every new eruption that occurs within the hazard area would disrupt the proposed high-level radioactive waste repository. Then the probability of site disruption by volcanic activity is equal to the chance that a new eruption will occur in the same area.

Autoregressive Integrated Moving Average models (ARIMA) were presented to find the best fitting model to predict the future recurrence rates, which were applied to calculate the probability of site disruption. The chosen model is MA(2), which has the lowest AICC value (= - 2512), and the residuals of this model are approximately white noise. The one-step-ahead forecast is $2.8030 \times 10^{-6}$ per year, and the adjusted 95%

prediction bounds for the annual recurrence rate are ($2.7273\times10^{-6}$, $3.0023\times10^{-6}$). Along with the other parameters' ($r_s$ and $A$) estimates, we conclude that the probability of volcanic disruption of the proposed high-level radioactive waste repository at YM for the next 1 m.y. is bounded by $1.842\times10^{-3}$ and $4.926\times10^{-2}$ for $r_d$ ranging from 0 to 6 km.

In summary, time series modeling are well developed and are largely applied in many other fields, which will greatly facilitate the needs of volcanologists using the proposed methods.

# APPENDIX

## ARIMA MODELS

Notation is first presented for a nonseasonal model, and then extended to include seasonal components in the model (Heiberger and Teles, 2002).

*Nonseasonal Models*

Assume $z_t$ follows the autoregressive integrated moving average ARIMA $(p, d, q)$ model $\phi(B)\nabla^d z_t = \theta(B)\varepsilon_t$,

where $B$ is the backshift operator; $B^j z_t = z_{t-j}$ is used to indicate *lagged* observations, that is, earlier observations of the same time series.

$\nabla^d = (1-B)^d$; $\nabla$ is the differencing operator and $d$ is the order of differencing, for example, $\nabla^2 z_t = (1-B)^2 z_t = z_t - 2z_{t-1} + z_{t-2}$;

$\phi(B) = (1-\phi_1 B - ... - \phi_p B^p)$, is the autoregressive operator;

$\theta(B) = (1-\theta_1 B - ... - \theta_q B^q)$, is the moving average operator;

$\varepsilon_t$ is a white noise process with zero mean and var$(\varepsilon_t) = \sigma_\varepsilon^2$.

*Seasonal Models*

When there is a seasonal component in the time series, $z_t$ is assumed to follow the more general multiplicative seasonal ARIMA $(p, d, q) \times (P, D, Q)_s$ model.

$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D z_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t$, where $s$ is the seasonal period of the time series;

$\nabla_s^D = (1-B^s)^D$; $\nabla_s$ is the seasonal differencing operator and $D$ is the order of seasonal differencing;

$\Phi_P(B^s) = (1-\Phi_1 B^s - ... - \Phi_P B^{Ps})$, is the seasonal autoregressive operator;

$\Theta_Q(B^s) = (1-\Theta_1 B^s - ... - \Theta_Q B^{Qs})$, is the seasonal moving average operator.

For various technical reasons, there are certain restrictions on the values that the roots of these polynomials may assume. The roots of the four polynomials $(\phi(B), \theta(B), \Phi_P(B), \text{and } \Theta_Q(B))$ must be outside the unit circle (if not, the model is not stationary and/or not invertible). The polynomials $\phi(B)$ and $\theta(B)$ must have no roots in common. Likewise, the polynomials $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ must have no roots in common. If the polynomials have common roots, these roots can be factored out. The reader interested in a deeper analysis of the basic concepts in time series should consult the books by Box and Jenkins (1976), and Box et al. (1994). The identification steps of ARIMA$(p,d,q)\times(P,D,Q)_s$ modeling can be difficult and will be demonstrated in the applications.

## REFERENCES


Akaike H (1974) A new look at the statistical model identification. IEEE Trans: Aut. Contr. AC 19: 203-217

Box GEP, Jenkins GM (1976) Time series analysis: forecasting and control: Holden-Day San Francisco (592pp.)

Box GEP, Jenkins GM, Reinsel GC (1994) Time series analysis: forecasting and control: 3rd ed. Prentice Hall, Englewood Cliffs, New Jersey (598pp.)

Brockwell PJ, Davis RA (2002) Introduction to time series and forecasting: 2nd ed. Springer-Verlag, New York (434 pp.)

Crowe BM, Johnson ME, Beckman RJ (1982) Calculation of the probability of volcanic disruption of a high-level radioactive waste repository within southern Nevada, USA: *Radioactive Waste Management* 3 (2): 167-190

Heiberger RM, Teles P (2002) Displays for direct comparison of ARIMA models: *The American Statistician* 56(2): 131-138

Ho CH (1990) Bayesian analysis of volcanic eruptions: *Jour. Volcanol. Geotherm. Res.* 43 (2): 91-98

Ho CH (1991a) Time trend analysis of basaltic volcanism for the Yucca Mountain site: *Jour. Volcanol. Geotherm. Res.* 46 (2), 61-72

Ho CH (1991b) Nonhomogenous Poisson model for volcanic eruptions: *Math. Geology* 23 (2): 167–173

Ho CH (1992) risk assessment for the Yucca Mountain high-level nuclear waste repository site: estimation of volcanic disruption. *Math Geol* 24: 347-364

Ho CH, Smith EI (1998) A spatial-temporal/3-D model for volcanic hazard assessment: application to the Yucca Mountain Region, Nevada: *Math. Geology* 30 (5): 497-510

Ho CH, Smith EI, Keenan DL (2006) Hazard area and probability of volcanic disruption of the proposed high-level radioactive waste repository at Yucca Mountain, Nevada, USA: *Bull. Vocanol* 69: 117-123

Ljung GM, Box GEP (1978) On a measure of lack of fit in time series models: Biometrika 65: 297-303

Kutner MH, Nachtsheim CJ, Neter J (2004) Applied Linear Regression Models: 4nd ed. McGraw-Hill (701 pp.)

Peña D, Tiao GC, Tsay RS (2002) A course in time series analysis: Wiley & Sons, Inc. (460pp.)

Scandone R., Arganese G., Galdi F (1993) The evaluation of volcanic risk in the Vesuvian area: *Jour. Volcanol. Geotherm. Res.* 58: 263-271

Schwarz G (1978) Estimating the dimensions of a model: *Ann. Stat.* 6 (2): 461-464

Smith EI, Feuerbach DL, Naumann TR, Faulds JE (1990) The area of most recent volcanism near Yucca Mountain, Nevada: implications for volcanic risk assessment, in Proc: Topical Meeting, High-level Radioactive Waste Management. Am. Nuclear Soc. /Am. Soc. Civil Engineers 1, (La Grange, Illinois), pp. 81-90

Smith EI, Keenan DL, Terry P (2002) Episodic volcanism and hot mantle: implications for volcanic hazard studies at the proposed nuclear waste repository at Yucca Mountain, Nevada: *GSA Today* 12 (4): 4-10

Sheridan MF (1992) A Monte Carlo technique to estimate the probability of volcanic dikes: high-level radioactive waste management. In: *Proceedings of the Third int. conf.*, 12-16 April 1992, Las Vegas, Nevada, pp 2033-2038

Wickman FE (1966) Repose-period patterns of volcanoes: *Ark. Mineral. Geol.* 4 (8): 291-367

Wickman FE (1976) Markov models of repose-period patterns of volcano: in D. F. Merriam, ed., Random processes in geology: Springer, New York: 135-161

# VITA

Graduate College
University of Nevada, Las Vegas

XiaoJuan Liu

Home Address:
9169 Whatley Street
Las Vegas, NV 89148
USA

Degrees:
Bachelor of Science, Computer Science, 2001
Nanjing Normal University, China

Thesis Title: An ARIMA-Model-Based Approach with Hazard Area for the Probability of Volcanic Disruption of the Proposed high-level Radioactive Waste Repository at Yucca Mountain, Nevada, USA

Thesis Examination Committee:
Chairperson, Dr. Chih-Hsiang Ho, Ph. D.
Committee Member, Dr. Malwane Ananda, Ph. D.
Committee Member, Dr. Sadra Catlin, Ph. D.
Graduate Faculty Representative, Dr. ShiZhi Qian, Ph. D.