University of Nairobi

School of Mathematics

ISSN: 2410-1397

Project in Social Statistics

Poisson Process Modelling of the Temporal Behaviour of Volcanic Eruptions in the East African Rift System

Research Report Number 41, 2020

James Kinyanjui

Nov 2020

Submitted to the School of Mathematics in partial fulfilment of the requirements for the award of the degree of Master of Science in Social Statistics

Poisson Process Modelling of the Temporal Behaviour of Volcanic Eruptions in the East African Rift System

Research Report Number 41, 2020

James Kinyanjui

School of Mathematics
College of Biological and Physical Sciences
Chiromo, off Riverside Drive
30197 - 00100 Nairobi, Kenya

Project

Submitted to the School of Mathematics in partial fulfilment of the requirements for the award of the degree of Master of Science in Social Statistics

Submitted to: The Graduate School, University of Nairobi, Kenya

Abstract

The social costs of volcanic eruptions are severe for eruptions of considerable magnitude. Understanding the temporal behaviour of volcanic eruptions is thus key to hazard assessments and prevention of future loss of life and damage to property. This project aimed to demonstrate that the temporal behaviour of volcanic eruptions in the East African Rift System can be effectively modelled using Poisson processes. The data used for the analysis was from the Smithsonian Institution's Global Volcanism Program, which is freely available on line. Three models were chosen for analysis: homogeneous Poisson model, log-linear non-homogeneous Poisson model and Weibull power non-homogeneous Poisson model. The assumptions and theory underpinning Poisson processes were presented and the eruption data considered was shown to fit into a Poissonian framework. The method of maximum likelihood was used to estimate the parameters of each model. The Akaike Information Criterion was used to select the optimal model. The log-linear non-homogeneous Poisson model with intensity function $\lambda(t) = exp(-0.911036781 + 0.009976769t)$ was found to best fit the empirical data with 95% confidence and based on it basic forecasts were issued. It was recommended that governments and disaster management authorities incorporate these findings in their preparedness frameworks as the log-linear model predicts an increasing trend in volcanic activity.

Master Thesis in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics

DISTRIBUTOR: School of Mathematics, University of Nairobi, Kenya

Declaration and Approval

I, the undersigned, hereby declare that work presented herein is an authentic record of my own endeavour and, furthermore, has not been submitted for the award of any degree in any other institution of learning.

Signature Date

James Kinyanjui
Reg. No. I56/8520/2017

I hereby certify that this project has been submitted to the University with my approval as the supervisor

Signature Date

Dr Goerge Muhua
School of Mathematics,
University of Nairobi,
Box 30197 - 00100 Nairobi, Kenya.
E. mail: muhuageorge@gmail.com

Dedication

This project is dedicated to *my late Dad* and *my late Uncle John*.

Acknowledgments

First and foremost I want to express my appreciation to the University of Nairobi's School of Mathematics for all its support throughout my period of study. Special mention to Director Dr Stephen W. Luketero for his pro-active approach and concern. I am also extremely grateful to my supervisor, Dr George Muhua. His guidance, insights and constructive critique were invaluable and had a direct impact on the form and quality this project took.

Deepest gratitude to my mother, Rosemary Nyambura. It is an understatement to say that I wouldn't have made it this far without her love, material and moral support and encouragement.

Heartfelt thanks to Irene Nduva: mentor, sponsor, neighbour, and family friend. This journey started in undergraduate and we continue to march forward! I also acknowledge the role that the late John Kamanu and Jane Kariuki played in gently steering me towards statistics at a time I was conflicted about my future.

A nod and wink to my dear friend, Kareen Durand. We have been through a lot together and all I can say is Thank You. We are almost there! *C'est possible!* Special shout out to Joseph Muindu for his friendship and fierce loyalty and for helping me with my presentation skills and to Simon Mwaniki for his unique eye and for generally being supportive. Lastly, a big *Ahsante* to Leonard Salasya for always being willing to check my output.

James Kinyanjui

Contents

Figures and Tables

Figures

Tables

Chapter One
Introduction

1.1 Introduction

This chapter provides the background of the topic, gives a statement of the problem, states the objectives, furnishes a brief justification for the study as well as the scope and gives an the overview of the entire body of this work.

1.2 Background

Volcanic eruptions represent the visible and sometimes violent manifestation of the dynamic funnelling of the Earth's internal energy to the surface (Wilson, 2009). The most active volcanic zones are located at the boundaries between tectonic plates, either at rift zones (zones of crustal separation) or subduction zones (zones of crustal collision). The distribution is such that 75% of the world's active and dormant volcanoes are situated in the Pacific Ring of Fire, a horseshoe-shaped region of intense geological activity hemming in the Pacific Ocean along the Western coast of the Americas and Eastern Asia stretching from Russian through Japan, the Malay Archipelago, Micronesia to New Zealand (Cottrell, 2014). The next greatest concentration of volcanoes is located in Eastern Africa, roughly following the course of the Great Rift Valley. Of the 148 volcanoes found in Africa, 120 are found in the East African Rift System. Ethiopia contains 59 volcanoes followed in second place by Kenya with 22 volcanoes (*Global distribution of volcanism: Regional and country profiles*, 2015). Other volcanic systems are found in Anatolia, the Italian Peninsular, the Middle East, Antarctica, Central Asia, The Caucasus, Iceland, South Caribbean, and scattered across the Atlantic Ocean and Indian Ocean. (Cottrell, 2014). The figure below shows this distribution.

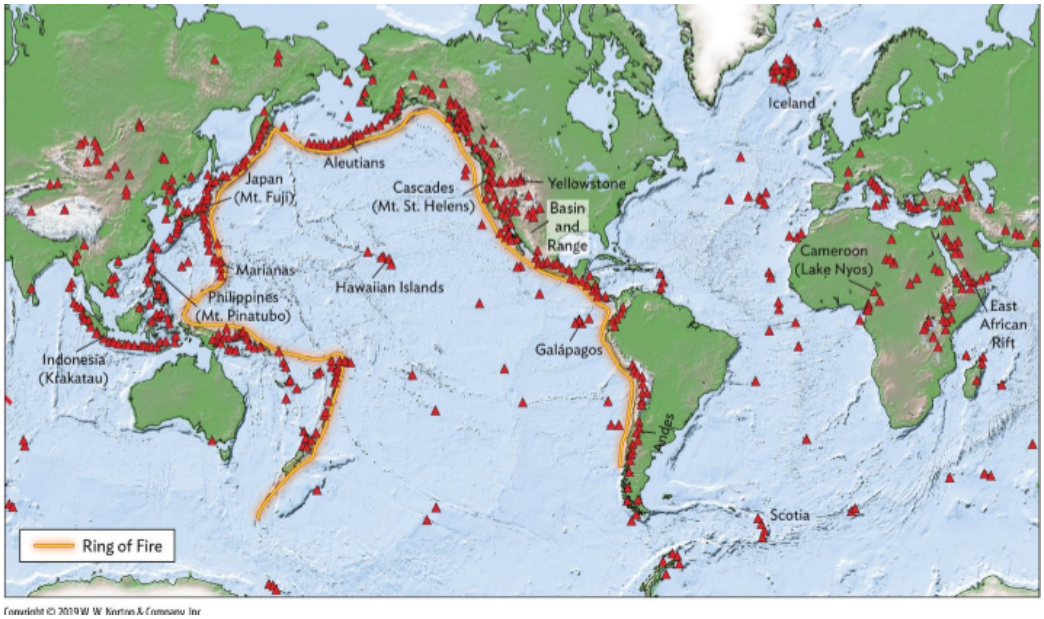

Figure 1. Worldwide distribution of Active Subaerial Volcanoes

Volcanoes possess enormous destructive power and volcanic eruptions constitute a major natural hazard. Eruptions of considerable magnitude can have localized effects, causing destruction through ballistic projections, pyroclastic flows, tephra (fragmented material), lava flows, lahar flows (slurry of volcanic material), landslides and avalanches, as well as global environmental impact through the dispersion of ashes, aerosols and other fine material from eruption plumes to the atmosphere where they can remain suspended for long periods of time (*Global Volcanic Hazards and Risk*, 2015). Eruptions can change weather patterns and lead to disruption of established climatic patterns, and, in the past, caused large-scale destruction that led to mass extinctions (Gilbert & Sparks, 1998).

Over 800 million people currently reside within 100 km of a volcano, of which more than 120 million of these are in Africa (*Global distribution of volcanism: Regional and country profiles*, 2015). These populations continues to grow, magnifying the potential effects of an eruption on life and property. Ethiopia and Kenya are ranked 5th and 10th respectively in terms of volcanic threat as measured by hazard index, number of volcanoes and population exposed within 30 km of volcanoes (*Global Volcanic Hazards and Risk*, 2015). It is further estimated that in the 20th century alone volcanic eruptions claimed the lives of about 80,000 people (Sigurdsson et al., 2015). The social costs are potentially high in the case of eruptions of significant magnitude. Statistical analysis and modelling of the uncertain behaviour of volcanoes, then, becomes an important task.

The field of statistical volcanology, specifically the temporal behaviour of volcanic activity, is driven by the increasing awareness of the hazards that volcanoes pose to life and property as the population swells. Many authors have contributed to the field, from the pioneering Poissonian modelling works of Wickman (1966) and Reyment (1969) to the successor studies of Ho (1991, 1992) and Bebbington and Lai (1996a, 1996b) who considered renewal models through to the more recent use of Gumbelian extreme value models as well as cluster models considered by authors such as Marzocchi and Zaccarelli (2006) and Gusev (2008) and novel models such as cellular automata models used by Sanchez (2014). Poisson models, however, seem to the most popular among many authors despite the merits of other models because of the ease of their use and their ability to capture effectively the count and temporal aspect of volcanic activity. The African perspective is largely absent from these efforts and this author has not come across any modelling enterprise with respect to the temporal aspect of volcanism. It is hoped that this project will go some way in remedying this.

1.3 Statement of the Problem

East Africa sits astride an area of seismic and geological importance with a significant number of active and dormant volcanoes. Eruptions of considerable size are rare events yet the when they do occur they have the potential for devastating consequences. Indeed, most volcanoes pose the greatest hazard over considerably long time scales in the order of decades and centuries; at longer time scales they have the potential for global impact and catastrophe. Even volcanoes thought to be dormant or extinct can suddenly erupt with little warning. The problem, then, becomes building probabilistic forecasts that account for this long-scale uncertainty using potential eruption scenarios and relevant data. An important consideration is that the historical record is short, biased and incomplete. The instrumented record is even more problematic, being shorter and, for most volcanoes, spanning only the last few decades of uninterrupted surveillance — a infinitesimal fraction of their long lifetime.

1.4 Objectives

The main objective of this project was fit eruption data to Poisson process models and obtain a model with an intensity function that would best approximate the observed trend in the data. This was accomplished by achieving the following specific objectives:

1. Test eruption data for the assumptions of Poisson processes;

2. Estimate the parameters of the intensity functions and compare model fit;

3. Determine goodness of fit for the model selected;

4. Forecast the number and probability of future eruptions.

1.5 Justification

The use of Poisson processes has been successful in varying degrees in modelling volcanism in the last five decades. This is because of its attractive quality of being able to combine the discrete property of count with the continuous element of time. The three specific models used were selected based on the literature review, as well as their parsimony: the homogeneous model based on the need to incorporate a stationary model that could capture any underlying constant trend in the data especially in light of the potential that the data considered was incomplete and the log-linear and Weibull non-homogeneous models based on the the assumption the data set was complete and inherently non-stationary coupled with the need to capture any trend observed in the data.

1.6 Significance

Most of the research done in statistical volcanology has concentrated on volcanoes in other tectonic-geological settings, particularly in the Pacific Ring of Fire, with scant attention paid to African volcanoes in general and East African ones in particular. It is hoped that this project will enrich the literature and add to our understanding of the stochastic processes that govern the temporal behaviour of volcanic eruptions and assist in hazard assessments that might be useful in mitigating risk to life and property.

1.7 Scope

The sampling frame for the sample used for the project was a global database of documented eruptions known to have occurred over the last 12,500 years compiled by the Smithsonian Institution's Global Volcanism Program. This catalogue contains all documented geological and historical-observation eruptions known to have occurred and consists of a set of just under 10,000 eruptions from approximately 10,500 B.C.E. till present day. The project was limited in scope to eruptions from 1 January 1919 to 1 January 2019, encompassing 13 volcanoes that have erupted in this period out of a total of the 120 volcanoes found in the East African Rift. The volcanoes are spread out across eight countries located in the East African Rift System: Ethiopia, Eritrea, Djibouti, Kenya, Uganda, Tanzania, Rwanda, and the Democratic Republic of Congo. The database contains a variety of information related to eruptions; however, the variable of interest was date of onset, which was either given in exact or approximated form.

1.8 General Outline of the Project

Chapter 1 introduced the topic and laid the groundwork for later chapters. Chapter 2 will be a review of important papers that present what other researchers have done and inform the basic motivation for this project, with some critique offered and gaps identified, as well the theory underpinning Poisson processes. Chapter 3 will present tests of data validation, the method of maximum likelihood estimation for the intensity functions selected, model selection criteria and a goodness-of-fit statistical test. Chapter 4 will involve analysis of the data in accordance with the framework set out in Chapter 3. Goodness of fit will then be performed and forecasts will then be given based on the model found to best suit the data. Finally, Chapter 5 will summarize the results, give recommendations and provide direction for further work.

Chapter Two

Literature Review

2.1 Introduction

This chapter presents a review of the theory underpinning Poisson processes as well a review of some existing papers that motivated and informed this project. The papers selected for review are by no means exhaustive but serve to provide a general snapshot of the work done in the area of Poissonian volcanic modelling over the past fifty years.

2.2 Theoretical Review

2.2.1 Homogeneous Poisson Process

A simple Poisson process is a mathematical model that describes a temporal-spatial series of events that occur randomly and independent of each other. The broad characteristics of a homogeneous Poisson process are: events occur singly with probability of near zero that two events occurring simultaneously; the rate of occurrence of events is constant; the probability of future events is independent of the past; and lack of time trend, i.e., stationarity. One useful characteristic to investigate is the distribution of inter-event times, which for a homogeneous Poisson process have an exponential distribution which is completely defined by a single parameter commonly denoted by λ, which represents the rate of occurrence of events or arrivals (Cox & Lewis, 1966).

Definition 2.2.1. *A collection of random variables {N(t): $t \in [0,\infty)$} indexed by time t is called a continuous-time stochastic process. Further, such a stochastic process is a (homogeneous) Poisson process if:*

(a) starting from N(0) = 0 the process N(t) takes non-negative integers 0,1,2.... for all

$t \geq 0$;

(b) the increment N(t + s) - N(t) is surely non-negative for any $s > 0$;

(c) the increments $N(t_1), N(t_2) - N(t_1)$.....$N(t_n) - N(t_{n-1})$

are independent for any $0 < t_1 < t_2$.....$t_{n-1} < t_n$;

(d) the increment N(t + s) - N(t) has a distribution which is dependent on the values

$s > 0$ *but independent of* $t > 0$; *and*

(e) the increment N(t + s) - N(t) has a Poisson distribution with mean λt, *i.e., for any*

s, $t \geq 0$.

$$Pr(N(t+s) - N(s) = n) = \frac{(\lambda t)^n exp(-\lambda t)}{n!}; \quad n = 0, 1, 2..... \tag{1}$$

The definition given by Ross (2010) lays the foundation for looking at the Poisson process as an integrated collection of three random variables: a counting process, a sequence of arrival or onset times and a sequence of inter-arrival or inter-event times.

A stochastic process satisfying (a) and (b) is called a *counting process* in which *N(t)* represents the total number of 'events' (from here onwards 'events' will refer to volcanic eruptions). Properties (c) and (d) are respectively called the *independent* and *stationary* increments.

Events counted by a Poisson process $\{N(t), t \geq 0\}$ are called *Poisson events*. Now, let T_n denote the time when the n-th Poisson event occurs. T_n is called the *arrival*, *event*, *onset* or *occurrence* time (in this case the onset of an eruption) and we can then define the $inter-arrival$, $inter-event$ or (in a volcanological context) *repose* times W_n as

$$W_n = T_n - T_{n-1}; \quad n = 1, 2, ... \tag{2}$$

where T_0 = 0 by convention and for convenience.

If we apply (a) and (d) together (see Definition 2.2.1) we obtain

$$Pr(N(t+s) - N(t) = n) = Pr(N(s) = n); \quad n = 0, 1, 2, ... \tag{3}$$

We observe that the event $\{W_n > s\}$ for an inter-arrival time can equivalently be expressed by the event $\{N(T_{n-1} + s) - N(T_{n-1}) = 0\}$, i.e., no event has occurred in the waiting period less than s, and that $N(s)$ = 0. This will justify the properties that inter-arrival time random variable W_n has a distribution independent of n and inter-arrival times W_1, W_2, W_3..... are independent.

Consider the survival function $S(s) = Pr(W_1 > s)$. It then follows that

$$\begin{aligned} S(t+s) &= Pr(W_1 > t+s) \\ &= Pr(N(t) = 0, N(t+s) - N(t) = 0) \\ &= Pr(N(t) = 0) Pr(N(t+s) - N(t) = 0) \\ &= Pr(N(t) = 0) Pr(N(s) = 0) \\ &= Pr(W_1 > t) Pr(W_1 > s) \\ &= S(t) S(s) \end{aligned} \tag{4}$$

The property above is called the *memoryless* property. The memeoryless property has the following formal definition. Let $G(a)$ = *Pr(X > a)*. If X is memoryless, then G has the following properties: (i) $G(a+b) = G(a)G(b)$; and (ii) G is monotonically decreasing, i.e., if $a \leq b$ then *G(a)* $\geq$ *G(b)* (Sigman, 2009). This means that waiting time until an event occurs does not depend on how much time has already elapsed. This implies that W_1 *must* have an exponential distribution since it is the only continuous distribution with this unique property. This can be easily proved.

The survival function of an exponential distribution is

$$S(t) = e^{-\lambda t} \tag{5}$$

such that

$$\begin{aligned} S(t+s) &= e^{-\lambda(t+s)} \\ &= e^{-\lambda t} e^{-\lambda s} \\ &= S(t)S(s) \end{aligned} \tag{6}$$

Let us define T_n, the time to the n-th arrival. It can be seen that the arrival time random variable can be expressed as $T_n = \sum_{k=1}^{n} W_k$. Since $W_1, W_2, W_3 \ldots\ldots\ldots\ldots W_n$ are iid exponential random variables with the common parameter λ, T_n has an $Erlang$ density with parameters (n, λ). The Erlang density can be derived by performing an n-fold convolution of the exponential distribution. Alternatively, it can be derived using the duality equation $W_n > t \iff N(t) < n$, which implies that the n-th event by a certain time t has not occurred if the waiting time to the n-th event is beyond t (Ross, 2010).

The Erlang distribution takes the following form:

$$f_{T_n}(t) = \frac{(\lambda^n) t^{n-1} exp(-\lambda t)}{(n-1)!}; \quad t, \lambda \geq 0 \tag{7}$$

The joint density of arrival times $T_1, T_2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots T_n$ conditional on $N(t) = n$ is identical to the order statistics $U_{(1)}, U_{(2)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots U_{(n)}$ of iid uniform random variables on (0,t]. This follows from the fact that if a homogeneous point process is defined on the real line as a mathematical model for occurrences of some observable event, then the positions of these occurrences on the real line (the real line in this case representing time) will be uniformly distributed. More concretely, if an event occurs according to this process in an interval $(a, b]$ where $a \leq b$, then its location in the interval will be a uniform random variable. The homogeneous point process is sometimes called the uniform Poisson point process for this reason (Ross, 2010).

An interesting property of the Poisson process is that it can be partitioned. If $\{N(t), t \geq 0\}$ is a Possion process with expectation λt each arrival independently either of type 1 or type 2 with probability p and $q = 1 - p$ respectively, then, the process can be partitioned into two independent sub-Poisson processes $\{N_1(t), t \geq 0\}$ and $\{N_2(t), t \geq 0\}$ each with its unique rate function such that $N_1(t) \sim Pois(\lambda t p)$ and $N_2(t) \sim Pois(\lambda t q)$ and $E[N(t)] = E[N_1(t)] + E[N_2(t)]$. The opposite (where separate Poisson process can be combined) is also true and is called *supposition* (Sigman, 2009).

2.2.2 Renewal Process

A simple Poisson process imposes the constraint of exponentially distributed inter-arrival times, which might be untenable in certain situations. A renewal process relaxes this assumption and permits other distributions to describe inter-event times. A renewal process is a generalization of a Poisson process consisting of inter-arrival times that are identically and independently distributed with a common distribution D (Ross, 1996). Some of the most common distributions for D are the Weibull distribution, gamma distribution, log-normal distributions and log-logistic distribution.

Definition 2.2.2. *A point process $\{t_n, n \geq 1\}$ is a strictly increasing sequence $0 < t_1 < t_2$ ..... If $\{N(t), t \geq 0\}$ is defined as a counting process with t_n being random variables, then $\{N(t)\}$ has iid inter-arrival times and is called a renewal process and it has a rate λ defined as $1/E(D)$, where W is the inter-event time random variable.*

From this definition adopted from by Sigman (2009) it becomes evident that if D is exponentially distributed then a renewal process dissolves to a HPP.

The distribution of $N(t)$ can be obtained by using the equality $N(t) \geq n \iff T_n \leq t$, i.e., the number of events or renewals by time time t is greater than or equal to n if and only if the n-th event or renewal occurs before or at time t (Ross, 1996). Therefore

$$\begin{aligned} Pr(N(t) = n) &= Pr(N(t) \geq n) - Pr(N(t) \geq n+1) \\ &= Pr(T_n \leq t) - Pr(T_{n+1} \leq t) \end{aligned} \tag{8}$$

Since the inter-event times are iid and have a common distribution D and $T_n = \sum_{k=1}^{n} D_k$ it follows that T_n is distributed as D_n, the n-fold convolution of D with itself. The distribution of $N(t)$ therefore becomes as shown below.

$$Pr(N(t) = n) = D_n(t) - D_{n+1}(t) \tag{9}$$

Let us define the expectation function $m(t)$. This function is called the *renewal function* in the context of a renewal process and much analysis is devoted to its properties (Ross, 1996). It is defined by the relationship given below.

$$m(t) = E[N(t)] = \sum_{n=1}^{\infty} D_n(t) \tag{10}$$

2.2.3 Non-Homogeneous Poisson Process

The restrictions in the properties of the homogeneous Poisson process make it inadequate for many other real world systems and natural phenomena which are prone to wild unpredictability but nonetheless possess Poisson-like characteristics with parameters dependent on time. Such systems can be described by non-homogeneous (or non-stationary) Poisson processes. The empirical review will show that while Poisson models have been found to be adequate in certain volcanic systems, universal application for all volcanic systems is found to be untenable as certain data sets show considerable deviation. The introduction of the non-homogeneous Poisson process as a generalization of the homogeneous Poisson model allows for some of this randomness to be captured. A non-homogeneous Poison process satisfies the same assumptions as a homogeneous Poisson process but with λ dependent on time, i.e., $\lambda(t)$. Utilizing a non-homogeneous time-dependent Poisson process in the context of volcanic activity implies that a number of underlying processes conflate together and the balance of these processes is a function of time (Sanchez, 2014).

Definition 2.2.3. *The counting process $\{N(t), t \geq 0\}$ is said to be a NHPP with intensity function $\lambda(t), t \geq 0$ if it satisfies the following conditions:*

(a) N(0) = 0 almost certainly;

(b) N(t) has independent increments;

(c) Pr(N(t + h) - N(t) = 0) = 1 - $\lambda(t)h + o(h)$;

(d) Pr(N(t + h) - N(t) = 1) = $\lambda(t)h + o(h)$;

(e) Pr(N(t + h) - N(t) $\geq 2)) = o(h)$.

This definition from Ross (2010) introduces another way of defining a Poisson process. Parts (c) and (d) of the definition may look awkward at first sight but are, in fact, insightful and intuitive. They state that having two or more events in a small time interval is extremely unlikely while the probability of a single event is approximately proportional to the length of that small interval. The notation $o(h)$ refers to some function g for which $\lim_{h\to 0} \frac{g(h)}{h} = 0$. The intensity function $\lambda(t)$ is a function of time and is often called the *instantaneous arrival rate.*

The distribution of the number of events in an interval is as follows:

$$Pr(N(t+s) - N(t) = n) = \frac{[\Lambda(t+s) - \Lambda(t)]^n e^{-[\Lambda(t+s)-\Lambda(t)]}}{n!} \tag{11}$$

This logically follows from the HPP since both processes have independent increments. This means that that the distribution of $N(t+s) - N(t)$ is, in fact, Poisson with parameter $\Lambda(t + s)$ - $\Lambda(t)$.

The relationship between the average number of events which occur in the interval (0,t] and the intensity function can be expressed as

$$\begin{aligned} E[N(t)] &= \int_0^t \lambda(u)du \\ &= \Lambda(t) - \Lambda(0) \\ &= m(t) \end{aligned} \tag{12}$$

This expectation function $m(t)$ completely defines the NHPP and is a monotonic non-decreasing right-continuous function such that

$$0 \leq \int_R \lambda(u)du < \infty$$

for all bounded subsets $\mathbb{R}$ of the state space S of the process.

This concept can be extended to the number of events between times t and $t + s$ to yield

$$E[N(t+s)-N(t)] = \int_{t}^{t+s} \lambda(u)du = \Lambda(t+s)-\Lambda(t) \tag{13}$$

The above results were adopted from Çinlar (2013).

For a NHPP $\{N(t), t \geq 0\}$ with intensity function $\lambda(t)$, the integrated function between two successive event times (which is, in fact, the mean function) T_n and T_{n+1} follows an exponential distribution with unit mean, i.e.,

$$m(T_n, T_{n+1}) = \int_{T_n}^{T_{n+1}} \lambda(t)dt \tag{14}$$

In addition, as a result of the independent increments property in non-overlapping intervals, $m(T_0, T_1), m(T_1, T_2)$.....$m(T_{n-1}, T_n)$ are iid exponential random variables (Smethurst, 2009). This serves as a useful link between the HPP and NHPP, which can be exploited, for example, if NHPP event times are to be converted to HPP event times.

The final set of results are adopted from Cox & Lewis (1966).

For a NHPP $\{N(t), t \geq 0\}$ with mean function $m(t)$ with an intensity function λ(t) which is absolutely continuous, the arrival times $t_1, t_2 \ldots\ldots\ldots\ldots t_n$ for n observed events in the interval $(0,T]$ are distributed as order statistics from a sample with probability density function

$$f_T(t) = \frac{\lambda(t)}{\Lambda(T) - \Lambda(0)} \tag{15}$$

We use the equality $W_n > t \Longleftrightarrow N(t) < n$ and observe that $Pr(T > t) = Pr(N(t+s) - N(t) = 0)$, i.e., time of the next arrival from start of observation is greater than t only if there is no event in the interval $(t, t+s]$. Using Eq. 11 it can be seen that for any $t, s \geq 0$

$$Pr(N(t + s) - N(t) = 0) = e^{-[\Lambda(t+s) - \Lambda(t)]} \tag{16}$$

and the cdf of arrival time becomes

$$F_T(t) = 1 - e^{-[\Lambda(t+s) - \Lambda(t)]} \tag{17}$$

To obtain the density function of conditional arrival time we get the derivative of the cdf wrt to s.

$$\begin{aligned} \frac{d}{ds} F_T(t) &= \frac{d}{ds} \{1 - e^{-[\Lambda(t+s) - \Lambda(t)]}\} \\ &= \lambda(t+s) e^{-[\Lambda(t+s) - \Lambda(t)]} \end{aligned} \tag{18}$$

For n observed events in the interval $(0,T]$ at times $t_1, t_2 \ldots\ldots\ldots\ldots t_n$ the joint density becomes

$$\lambda(t_1)e^{-[\Lambda(t_1)-\Lambda(t_0)]}\lambda(t_2)e^{-[\Lambda(t_2)-\Lambda(t_1)]}\ldots\ldots\ldots\ldots\ldots\lambda(t_n)e^{-[\Lambda(t_n)-\Lambda(t_{n-1})]}e^{-[\Lambda(T)-\Lambda(t_n)]}$$

$$\begin{aligned} &= e^{-[\Lambda(T)-\Lambda(0)]}\prod_{i=1}^{n}\lambda(t_i) \\ &= e^{-\int_0^T \lambda(u)du}\prod_{i=1}^{n}\lambda(t_i) \end{aligned} \tag{19}$$

where the term $e^{-[\Lambda(T)-\Lambda(t_n)]}$ denotes the probability that no event occurs in the interval $(t_n, T]$.

Eq. 19 can also be expressed as

$$\frac{[\Lambda(T)-\Lambda(0)]^n e^{-[\Lambda(T)-\Lambda(0)]}}{n!}\prod_{i=1}^{n}\frac{\lambda(t_i)}{[\Lambda(T)-\Lambda(0)]^n} \tag{20}$$

if we consider *unordered* event times.

This, then, yields the conditional density function of T_i as shown below.

$$f_T(t_i|N(t)=n) = \frac{\lambda(t)}{\Lambda(T)-\Lambda(0)}; \quad i = 1,2\ldots\ldots\ldots\ldots n \tag{21}$$

2.3 Empirical Review

Wickman (1966) carried out one of the pioneering works of statistical analysis on volcanic data where he described the applicability of Poisson modelling. The Poisson process was defined as a model for describing random temporal-spatial events. He noted that certain volcanoes showed eruptive rates that were independent of time and were thus memoryless: past events had no bearing on future events. Such volcanoes were described as Simple Poissonian Volcanoes. This lack of memory naturally implied the use of an exponential distribution to model inter-event times. The exponential distribution is completely defined by its single parameter λ, which represents the rate of occurrence, which is constant in the case of stationarity. Wickman tested the homogeneous Poisson process hypothesis on two Hawaiian volcanoes: Mauna Loa and Kilauea. He found that the eruptive activity of Mauna Loa was well described by a stationary Poisson process but that the activity of Kilauea deviated from the model and showed non-stationarity and hence a non-constant event rate. Wickman's study left open the question of the nature of this inhomogeneity but suggested modelling the intensity using a step function.

Reyment (1969) expanded on Wickman's work by studying the activity of three Japanese volcanoes (Asama, Aso and Kirisima), Mount Etna, three Indonesian volcanoes (Bromo, Semeru and Peak of Ternate) and Mauna Loa. Mauna Lao was found to approximately follow a simple Poisson process, a finding that coincided with Wickman's earlier work. Bromo, Semeru and Peak of Ternate showed Poissonian behaviour consistent with either some form of renewal process. The three Japanese volcanoes and Mt. Etna showed some trend in the eruption rates, suggesting inhomogeneity and a log-linear intensity function was proposed. The conclusion was that the Poisson model was not universal in application and individual volcanoes exhibited unique eruption patterns. As with Wickman (1966), no direction was given on what sort of renewal model was likely to fit the data set considered.

Settle and Mcgetchin (1980) took a different approach. Instead of examining a number of volcanoes, they chose to concentrate on one volcano with multiple sites of volcanic activity. They examined the three-day 1971 activity of Stromboli (a volcano off the Sicilian coast in Italy) through the eruption sequence of three different vents. They found that when considered separately the repose times of the three vents showed a Gaussian distribution. However, a critical analysis of the entire record showed correlated behaviour between two of the vents with remaining one. The repose time distribution for two of the three vents could be fitted by an exponential distribution, and therefore their activity could be modelled by a Poisson process. This vent dependency suggested the direct connection of the two vents to the magma reservoir of the remaining vent.

In another study, Klein (1982) returned to the Hawaiian twins of Mauna Loa and Kilauea and investigated their holistic activity using an updated catalogue. He used the customary definition of inter-event times of volcanic eruptions as the time passed between one eruption onset to the next. The absence of major periodicity with future eruptions relatively independent of past eruptions suggested that their random behaviour was typical of a Poisson process though it was noted that long repose times observed in both volcanoes were associated with large eruptions.

Noting that previous studies had been restricted in scope, De la Cruz-Reyna (1991) opted for a different approach by considering worldwide volcanic eruptions. Exception was taken to the lumping together of all eruptions without regard to their strength. The study found that when sorted based on size, higher magnitude eruptions followed a Poisson process with a constant rate. This was explained in terms of a load-and discharge heat transfer mechanism, where the vast energy pent up in volcanic systems was released to the surface at a constant rate in small amounts.

In the aforementioned investigations the unifying thread was the hypothesis of volcanic events being a series of independent random events with some sort of constant underlying rate implied. However, considerable deviations for a number of volcanic systems studied suggested the inadequacy of the homogeneous Poisson model. A number of investigators turned their attention to the non-homogeneous Poisson model as means of addressing these weaknesses.

Ho (1991) tested the appropriateness of a homogeneous Poisson model on five individual volcanoes in four different tectonic settings: Kilauea, St. Helens, Etna, Aso and Yake. He considered a Weibull process with an eruptive rate λ(t) such as $\lambda(t) = (\frac{\beta}{\theta})(\frac{t}{\theta})^{\beta-1}$ where β and θ were parameters to be estimated and t the time from a pre-defined origin. The parameter β was of special interest because it was used to characterize volcanoes into three types: volcanoes with increasing eruption rate ($\beta > 1$); volcanoes with decreasing eruptive rate ($\beta < 1$); and volcanoes with constant eruptive rate ($\beta = 1$). Essentially Ho used the Weibull model as a goodness of fit test to investigate if a volcano was Simple Poissonian under the null hypothesis that $\beta = 1$. The results showed none of the volcanoes he considered had a constant eruptive rate.

Bebbington and Lai (1996a) were critical of Ho's method and argued for a more general approach. They especially questioned the use of a non-stationary monotonic-trend model in the face of the available and potentially incomplete data arguing that a monotonic trend parameter would likely overestimate the underlying increase in activity with data sets that become more complete with time. They subjected the same data set considered by Ho (1991) to a renewal process. They proposed a generalized Weibull distribution and log-normal distribution for the common distribution D describing inter-arrival times, noting that while the exponential distribution represented pure random behaviour as espoused by the memoryless trait unique to the distribution, the log-normal distribution represented periodicity and the Weibull distribution represented both periodicity and clustering. They found the Weibull renewal process produced more plausible results than Ho's model but found log-normal inter-arrival times untenable.

Bebbington and Lai (1996b) followed up on their earlier efforts by using a general Poisson process and Weibull renewal process to study the activity of Mt. Ruapehu and Mt. Ngauruhoe in New Zealand. They found that a homogeneous Poisson process (where D is an exponential distribution) described the behaviour of Ngauruhoe well but Ruapehu showed a more complex eruption pattern even though fairly reasonable forecasts were obtained. The study showed no correlation between the eruptive behaviour between these two closely neighbouring volcanoes. Ho (1992) had in fact used the type of renewal models suggested by Bebbington and Lai (1996a, 1996b), using one such to give forecasts on the activity of Vesuvius where he defined D as a gamma distribution and the frequency distribution of eruptions in a given interval of equal size governed by a negative binomial distribution.

Salvi et al (2006) analysed the lateral (flank) activity of Mt. Etna over the last five centuries with the aim of defining a space-time distribution and obtaining estimates of lava flow hazard. The conclusion reached was that a non-homogeneous Poisson process was a likely model considering that there was strong statistical evidence for increasing intensity, notice being taken of the increase in the number of eruptions over the period of 20 years prior to the study. This result was corroborated by Smethurst et al (2009), who found a nearly linear increase in intensity from the 1950s and little evidence of periodicity. The findings of the two studies contradicted those of Mulargia et al (1985), who had found a general Poisson model suitable. This was attributed to use of an updated time series and indicated the need for re-examination of past models based on improved catalogues.

Marzocchi and Zaccarelli (2006) were skeptical of past studies that relied on data from one or a few volcanoes. They argued that it was difficult to gauge whether the behaviour noticed was generic or represented activity specific to the system studied. They opted to use a more diverse catalogue of volcanoes from all over the world in order to obtain results that reconciled the universal and peculiar of different eruptive styles within the framework of Poissonian modelling. Their study came to the conclusion that there were two competing eruptive regimes: time clustering of short inter-event times following a time-dependent model characteristic of open conduit systems and randomness of long repose times associated with closed conduit systems behaving according to a Poisson process. Gusev (2008) also observed clustering in time and size of eruptions of similar order with an event rate non-uniform in time. He posited the existence of an underlying global mechanism whose distribution was not well understood.

Mendoza-Rosas and De la Cruz-Reyna (2009) averred that the Bebbington and Lai (1996b) approach was useful but found it complex in its calculation and proposed a simpler model for repose times. They opted for a mixed exponentials approach to study Colima and Popocatépetl volcanoes in Mexico and compared the results to those of a Weibull distribution. This approach involved modelling non-stationary eruptive series with different occurrence rates as a sum of exponential random variables. They found that their approach fared better and was easier in application than the Weibull renewal process and recommended their method when eruption rates had well-defined patterns evident from a cumulative series of arrivals.

Dzierma and Wehrmann (2010) also chose a dual approach and used both the homogeneous and non-homogeneous process on ten volcanoes in the South Chile Volcanic Zone with the aim of forecasting the likelihood of future eruption. They modelled repose times using the exponential distribution, Weibull distribution and log-logistic (Pareto III) distribution. They were able to show that each volcano was better fitted by a particular distribution, that is, the nature of D appeared to be unique depending on the volcano considered. This further highlighted the issue of volcanoes having their own individual eruptive regimes even for volcanoes in general proximity to each other. Their study also raised the issue of the delineation of different eruption regimes of a single volcano, especially in light of historically incomplete data.

The use of a log-logistic distribution to model interval data already had precedence. Connor et al (2003) had approached the issue of modelling of repose times from the point of view of the balance of competing geological processes operating in a volcanic conduit system at different times and argued that any model should take into account. To that effect they used a log-logistic model to study a curtate eruption sequence of the then highly active Soufriere Hills on the island of Montserrat. Their model produced an excellent fit and they attributed this to the fact the parameters of the model were linked to the underlying geological processes in a meaningful way. They stated that the elegance of the log-logistic distribution was that some parameters worked to increase probability of an event in time while others working on a different time scale operated in the opposite direction, operating in a similar fashion to the shape parameter of a Weibull model. They were, however, wary of modelling repose times using Weibull failure models, because of their unreliability in explaining eruptive patterns with significant variation.

Chapter Three
Methodology

3.1 Introduction

This chapter presents the tests used to validate the data in order to ascertain that the assumptions of Poisson processes are met. The method of maximum likelihood estimation is elaborated upon, followed by the inversion simulation method to be used in the R algorithm. Finally, the goodness-of-fit statistic to be used is briefly outlined.

3.2 Data Validation

Temporal Poissonian processes are built on a number of assumptions. Events are assumed to occur singly, i.e., over an infinitesimally small period of time only one event can occur. A point of reference is needed so it is also assumed that at the start of observation there is almost certainly no chance of observing an arrival. These arrivals occur randomly in time and of each other. The type of Poisson process determines whether arrival is constant in time or time-dependent and hence whether the process is stationary or non-stationary. The first two assumptions are axiomatic; the other assumptions were tested to see if the eruption data considered fit into a Poissonian framework. This took the form of plots and formal statistical tests.

3.2.1 Testing for Randomness of Arrivals

To test if arrivals depend on date Brown et al (2004) suggested choosing a short interval of time over which $\lambda(t)$ is plausibly constant. If $\lambda(t) = \lambda_{date}$ is constant on this short interval, then the counts over an extended period of time are approximately uniform in distribution as a function of date. The implication is that an eruption can occur on any given calendar day, thus making the arrival process random in time. A plot of the eruption data arranged as above should show a rectangular shape for uniform data. The Kolmogorov-Smirnov (K-S) one-sample test (see Section 3.5 for more details) was used to test this assumption. Below are the hypotheses.

H_0: Eruption data arranged by week is uniform in distribution

vs

H_1: Eruption data arranged by week is not uniform in distribution

3.2.2 Testing for Independence of Arrivals

For events to be said to be occurring independently of each other Brown et al (2004) stated that the inter-arrival times need to be independent of ordering and serially uncorrelated. They suggested not only checking for independence of the raw inter-event times but also of their various transformations. The hypotheses are shown below.

H_0: Inter-event times (and their transformations) are serially uncorrelated

vs

H_1: Inter-event times (and their transformations) are serially correlated

Two statistics were used to test this assumption: the co-efficient of determination and the Ljung-Box statistic.

The co-efficient of determination R^2 measures the proportion of total variation explained by the regressor(s) in the regression model, in this case a regression of inter-event times against their index of ordering.

$$R^2 = 1 - \frac{SSE}{SST} \tag{22}$$

where SSE is sum of squared errors and SST is total sum of squares.

Arrivals are judged as independent of ordering and serially uncorrelated if regression produces a poor fit as evidenced by a low R^2.

The Ljung-Box statistic is shown below.

$$Q_k = n(n+2)\sum_{j=1}^{k}\frac{r_j 2}{n-j} \tag{23}$$

where r_j is the sample autocorrelation co-efficient at lag j, k is the lag order and n is the sample size. The statistic has a χ^2 distribution with k degrees of freedom.

3.2.3 Testing for Stationarity

Dzierma and Wehrmann (2010) suggested testing stationarity by computing moving averages for repose times and plotting them as a function of time. For a formal test the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test is utilized to see if a time series is stationary around a mean or linear trend. The KPSS test decomposes a time series into a deterministic trend, a random walk and an error term. The hypotheses to be tested were as follows.

H_0: Eruption data is trend stationary

vs

H_1: Eruption data is not trend stationary

3.2.4 Testing for Exponentiality of Transformed Inter-event Times

The exponentiality of inter-arrival times of a HPP was presented in Chapter 2. To test whether a time series is generated by a Poisson process Brown et al (2005) suggested transforming the empirical event times into those of a HPP by assuming a piecewise-constant arrival rate. The method is discussed below.

The duration of time of interest is broken down into relatively short blocks of time, preferably of equal length. These blocks should be short enough to assume a constant arrival rate but long enough to include between five and seven observations. Arrivals are then considered within these block in either a vertical manner (arrivals in the same time block) or a horizontal manner (arrivals staggered across blocks). The first approach tests for homogeneity within a block while the second approach tests for homogeneity across blocks. The second approach was the approach of interest.

Divide the time period of length N into blocks of equal length L such that $J(i)$ represents the number of events in the i-th block. Define the starting point for each block as $T_{i0} = 0$ and let T_{ij} denote the j-th ordered arrival time in the i-th block. Finally define R_{ij} such that

$$R_{ij} = (J(i)+1-j)(-ln(\frac{L-T_{ij}}{L-T_{i(j-1)}})); \quad j = 1,2 \ldots\ldots\ldots\ J(i) \tag{24}$$

R_{ij} represents iid exponential random variables with unit mean. This is how that comes about. Let U_{ij} represent the j-th unordered arrival time in the i-th block. Assuming a constant arrival rate in each block then conditional on $J(i)$, $U_{ij} \sim iid\ Unif(0,L]$ and$T_{ij} = U_{i(j)}$ (see p. 10). It then follows that (see Lehmann (1986), Problem 6.14.33 [345-346])

$$\frac{L-T_{ij}}{L-T_{i(j-1)}} \sim iid\ Beta(j, J(i)+1-j) \tag{25}$$

A change of variable yields the exponential distribution of R_{ij} conditional on $J(i)$. The null hypothesis then becomes that R_{ij} are iid unit-mean exponential random variables.

Once the transformed event times are obtained a diagnostic plot can be produced and augmented with a formal test, in this case the K-S one-sample test.

3.3 Parameter Estimation and Model Selection

In this section we examine the estimation of the intensity function parameters using MLE. This is followed by a brief description of the Akaike Information Criteria (AIC) as a method of model selection.

3.3.1 Maximum Likelihood Estimation

MLE involves optimizing the likelihood function with the goal of estimating parameters which make it more probable to observe the given data. The advantage of MLE it takes into account the real distribution of the data and is robust in case of deviation from normality, a key assumption of OLS estimation (Myung, 2003).

Let us consider a random sample from an unknown population. MLE attempts to make inference about the population that generated that sample. Assume we have a set of iid random variables $(t_1, t_2 \ldots\ldots\ldots\ldots t_n)$, each indexed by a unique parameter vector $\underline{\theta}$ = $\{\theta_1, \theta_2 \ldots\ldots\ldots\ldots \theta_k\}$ whose values can lie anywhere in the parameter space Θ. To obtain the ML estimates of $\underline{\theta}$ we need to get the *likelihood function*, which is the joint distribution of the observed sample.

$$\mathscr{L}(\underline{\theta}; \mathbf{t}) = \prod_{i=1}^{N} f_i(t_i; \underline{\theta}) \tag{26}$$

The goal of MLE is to find the specific values that maximize the likelihood function over the parameter space, i.e,

$$\underline{\hat{\theta}} = \underset{\theta \in \Theta}{\operatorname{argmax}} \mathscr{L}\left(\underline{\theta}; \mathbf{t}\right) \tag{27}$$

This maximum point is called the *maximum likelihood estimate*.

This entails the selection of parameter values that make the observed data most probable. It is customary and convenient to deal with the *log-likelihood*, the natural logarithm of the likelihood function.

$$\ell(\underline{\theta}; \mathbf{t}) = ln\mathscr{L}(\underline{\theta}; \mathbf{t}) \tag{28}$$

If $\ell(\underline{\theta}; \mathbf{t})$ is a differentiable function then the maxima are the solutions to the likelihood equations obtained by getting the derivative with respect to $\underline{\theta}$ and setting the results to zero, i.e.,

$$\frac{\partial \ell}{\partial \theta_1} = 0, \frac{\partial \ell}{\partial \theta_2} = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots \frac{\partial \ell}{\partial \theta_k} = 0 \tag{29}$$

Two problems might arise. For some models explicit solutions for $\underline{\hat{\theta}}$ can be derived. For other models the likelihood equations are intractable; no closed-form solutions exist and they can only be solved using numerical methods such as Newton-Raphson estimation. The other problem is the possibility of the existence of multiple solutions to the likelihood equations. A particular solution is defined as a maximum if the matrix of second-order partial derivatives of the log-likelihood, known as the $Hessian$, as shown below,

$$\mathbf{H} = \begin{bmatrix} \left.\frac{\partial^2 \ell}{\partial \theta_1{}^2}\right|_{\theta=\hat{\theta}} & \left.\frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2}\right|_{\theta=\hat{\theta}} & \cdots & \left.\frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_k}\right|_{\theta=\hat{\theta}} \\ \left.\frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1}\right|_{\theta=\hat{\theta}} & \left.\frac{\partial^2 \ell}{\partial \theta_2{}^2}\right|_{\theta=\hat{\theta}} & \cdots & \left.\frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_k}\right|_{\theta=\hat{\theta}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial^2 \ell}{\partial \theta_k \partial \theta_1}\right|_{\theta=\hat{\theta}} & \left.\frac{\partial^2 \ell}{\partial \theta_k \partial \theta_2}\right|_{\theta=\hat{\theta}} & \cdots & \left.\frac{\partial^2 \ell}{\partial \theta_k{}^2}\right|_{\theta=\hat{\theta}} \end{bmatrix} \tag{30}$$

is *negative semi-definite* at $\hat{\theta}$, indicating local concavity.

3.3.2 Maximum Likelihood Estimators of Models Selected

This section presents the method of MLE with respect to the three Poisson models considered, namely: the HPP; the log-linear NHPP; and the Weibull power NHPP.

The joint distribution of arrival times conditional on number of arrivals was derived in Section 2.2.3 (see Eq. 16) and was found to be as follows:

$$f_T(t_i|N(t)=n) = e^{-\int_0^T \lambda(u)du} \prod_{i=1}^{n} \lambda(t_i) \tag{31}$$

Homogeneous Poisson Process

The intensity function for a HPP is as shown below.

$$\lambda(t) = \lambda \tag{32}$$

The likelihood function is

$$\begin{aligned} \mathscr{L}(\lambda;t) &= e^{-\int_0^T \lambda dt} \prod_{i=1}^{n} \lambda \\ &= e^{-\lambda T} \lambda^n \end{aligned} \tag{33}$$

and the log-likelihood becomes

$$\ell(\lambda;t) = nln\lambda - \lambda T \tag{34}$$

Getting the derivative wrt λ and setting the result to zero yields the ML estimator of a HPP, which is

$$\hat{\lambda} = \frac{n}{T} \tag{35}$$

Log-Linear Non-Homogeneous Poisson Process

The intensity function for a log-linear NHPP is as shown below.

$$\lambda(t) = e^{\gamma_0+\gamma_1 t} \tag{36}$$

The likelihood function is

$$\begin{aligned}\mathscr{L}(\gamma_0,\gamma_1;t) &= e^{-\int_0^T e^{\gamma_0+\gamma_1 t}dt}\prod_{i=1}^{n} e^{\gamma_0+\gamma_1 t_i} \\ &= e^{-\frac{e^{\gamma_0}}{\gamma_1}(e^{\gamma_T}-1)} e^{n\gamma_0+\gamma_1\sum_{i=1}^{n} t_i}\end{aligned} \tag{37}$$

and the log-likelihood becomes

$$\ell(\gamma_0,\gamma_1;t) = n\gamma_0 + \gamma_1\sum_{i=1}^{n} t_i - \frac{e^{\gamma_0}}{\gamma_1}(e^{\gamma_1 T}-1) \tag{38}$$

Getting the derivative wrt γ_0 and γ_1 and setting the results to zero yields

$$\begin{aligned}\hat{\gamma_0} &= ln(\frac{n\hat{\gamma_1}}{e^{\hat{\gamma_1}T}-1}) \\ \sum_{i=1}^{n} t_i + \frac{n}{\hat{\gamma_1}} &= \frac{nTe^{\hat{\gamma_1}T}}{e^{\hat{\gamma_1}T}-1}\end{aligned} \tag{39}$$

which have no closed-form solution.

Weibull Power Non-Homogeneous Poisson Process

The intensity function for a Weibull power NHPP is as shown below.

$$\lambda(t) = \frac{\beta}{\eta})(\frac{t}{\eta})^{\beta-1} \tag{40}$$

The likelihood function is

$$\begin{aligned} \mathscr{L}(\beta,\eta;t) &= e^{-\int_0^T (\frac{\beta}{\eta})(\frac{t}{\eta})^{\beta-1}dt} \prod_{i=1}^{n} (\frac{\beta}{\eta})(\frac{t_i}{\eta})^{\beta-1} \\ &= e^{-(\frac{T}{\eta})^{\beta}} \frac{\beta^n}{\eta^{\beta n}} \prod_{i=1}^{n} t_i{}^{\beta-1} \end{aligned} \tag{41}$$

and the log-likelihood becomes

$$\ell(\beta,\eta;t) = nln(\beta) - n\beta ln(\eta) + (\beta-1)\sum_{i=1}^{n} ln(t_i) - (\frac{T}{\eta})^{\beta} \tag{42}$$

Getting the derivative wrt β and η and setting the results to zero yields the ML estimators of a Weibull power NHPP, which are

$$\begin{aligned} \hat{\beta} &= \frac{n}{\sum_{i=1}^{n} ln(\frac{T}{t_i})} \\ \hat{\eta} &= \frac{T}{n^{1/\hat{\beta}}} \end{aligned} \tag{43}$$

3.3.3 Akaike Information Criterion

The AIC is an estimator of resampling prediction error and therefore a measure of the relative quality of a statistical model for a given set of data. A statistical models can never perfectly represent the process that generated a sample; inevitably some information gets lost in the process. AIC estimates the relative amount of information lost by a given model and gives a score expressing this loss: the less information a model loses, the better the quality of that model. In determining the amount of information lost AIC performs a balancing act between model fit (as determined by maximized likelihood) and parsimony (as determined by k, the dimension of the parameter vector), penalizing both overfitting and underfitting. If a number of models are considered, then, in the most simplistic sense, the model with the lowest score is the one selected (Aho et al, 2014).

Let us a consider a model with k parameters and let L be the value of the maximized likelihood of the model. The AIC score of the model is given below.

$$AIC = -2lnL(\hat{\theta}) + 2k \tag{44}$$

3.4 Simulation

Once the parameters have been estimated the next step is to to use them to simulate a data set that will be compared with the sample data. The inversion algorithm used in generating NHPP event times stems directly from the theory presented in Chapter 2.2.3.

Consider arrival times $T_1 = t_1, T_2 = t_2 \ldots\ldots\ldots\ldots\ldots\ldots T_i = t_i$. The distribution function of *(i + 1)*-th inter-arrival time conditional on i arrivals takes the following form (see Eq. 17):

$$F_{W\,i+1}(x) = 1 - e^{-(\Lambda(t_i+x)-\Lambda(t_i))} \tag{45}$$

Given the *i*-th event time, the *(i + 1)*-th event time is generated as the sum of the *i*-th event time and the *(i + 1)*-th inter-arrival time distributed according to $F_{W_{i+1}}$. Klein & Roberts (1984) suggested the following steps when developing an algorithm for simulation:

1. Initialise *t = 0*

2. Generate *x* from F_W

3. Set $t \leftarrow t + x$

4. Deliver *t*

5. Return to Step 2

If the NHPP has a rate function $\lambda(t)$ this means finding x satisfying

$$\Lambda(t_i + x) - \Lambda(t_i) = \int_{t_i}^{t_i+x} \lambda(y)dy = -ln(1-u) \tag{46}$$

where $u \sim$ *Unif (0,1)*.

The explanation of the algorithm is as follows. If we consider a finite interval $(0,t]$ then a single event can occur in this interval at any point, its random positioning following a uniform distribution. The length of time to the event occurring follows the distribution function given in Eq. 45. If a particular random number from *Unif (0,1)* is generated representing where the event falls in the interval, then all that is required is to find the particular time value that causes the distribution function to integrate to the uniform distribution value generated, i.e., $x = F^{-1}(u)$. This inter-arrival time is then be added to the previous event time, the process being repeated until the requisite number of arrival times is obtained. This can be done manually or using software.

3.5 Goodness of Fit

The Kolmogorov-Smirnov test (K-S test) is a formal statistical test used to augment the customary plots used to check goodness of fit. The K-S test is a non-parametric and agnostic test used to detect differences between distributions. It examines a single maximum difference between distributions. If a statistical difference exits, the test does not provide insight into the cause of the difference nor does it indicate the nature of the common distribution if there is no statistical difference between the two distributions. The differences could be as a result of difference in: location; variation; skewness; kurtosis; and modality, or presence of outliers, among other things (Daniel, 1990).

Consider two random variables X and Y with specific distributions (not necessarily known) from which are drawn samples of equal size *n*. Under the hypotheses

$$H_0 : F(x) = F(y), \text{ i.e., the two distributions are the same}$$

vs

$$H_1 : F(x) \neq F(y), \text{ i.e., the two distributions are not equal}$$

the K-S test statistic is as follows:

$$d = max|F(x) - F(y)| \tag{47}$$

where $F(x)$ is the empirical cdf of *n* iid ordered observations, $F(y)$ is the comparison cdf of size *m* and *max* is the maximum of the absolute differences of the set of ordered distances.

If the two samples come from the same distribution then the statistic *d* converges to zero almost surely as *n, m* $\rightarrow \infty$. For large samples H_0 is rejected at α level if

$$d > c(\alpha)\sqrt{\frac{n+m}{nm}} \tag{48}$$

where in general $c(\alpha) = \sqrt{-ln(\frac{\alpha}{2}) * 0.5}$

The one-sample version of the test is used to check whether data follows a hypothesized distribution. Critical values for comparison are obtained from statistical tables.

It should be noted that the test only applies to continuous distributions and tends to be more sensitive near the centre of the distributions than at the tails. However, the test becomes more robust if a large sample size is used, in which case the minimum bound becomes more sensitive.

Chapter Four

Data Analysis, Interpretation and Results

4.1 Introduction

This chapter presents the testing of assumptions of a Poisson process and fits the observed data to three selected models in order to chose an appropriate one followed by simulation of an idealized data set with which to compare the empirical data. Tests are then carried out to check how well the data fits the model.

4.2 Source and Brief Description of the Data

The data used for the analysis was from the Smithsonian Institution's Global Volcanism Program. It contains information on name and index number of the volcano, eruption index number, eruption category (in this case all confirmed eruptions), dates for onset and ending of eruptions, each eruption's corresponding Volcanic Explosivity Index (a relative metric developed by Newhall and Self (19282) taking the form of a gradated logarithmic scale which measures the strength of eruptions), dating method for each eruption (in this case all historical observations), and longitudinal and latitudinal location of each volcano. The data contains (in chronological order) all the known and/or documented volcanic eruptions of the Holocene Epoch (the current geological epoch, which began approximately 12,000 years ago), a set of just under 10,000 volcanic eruptions. However, a choice was made to limit the project to confirmed eruptions from 1919 to 2019, representing 69 eruptions. The justification for this judgement call is fourfold: the historical record is better documented in the last century as a result of increased surveillance of volcanoes, as well as better scientific methods; the majority of eruption dates are attested (the seven missing data points were obtained by interpolation); the sample size is sufficiently large; and the time span encompasses the recent historic component of activity.

4.3 Data Validation

The first task was to produce a plot to have a sense of how the eruptions were arranged in time.

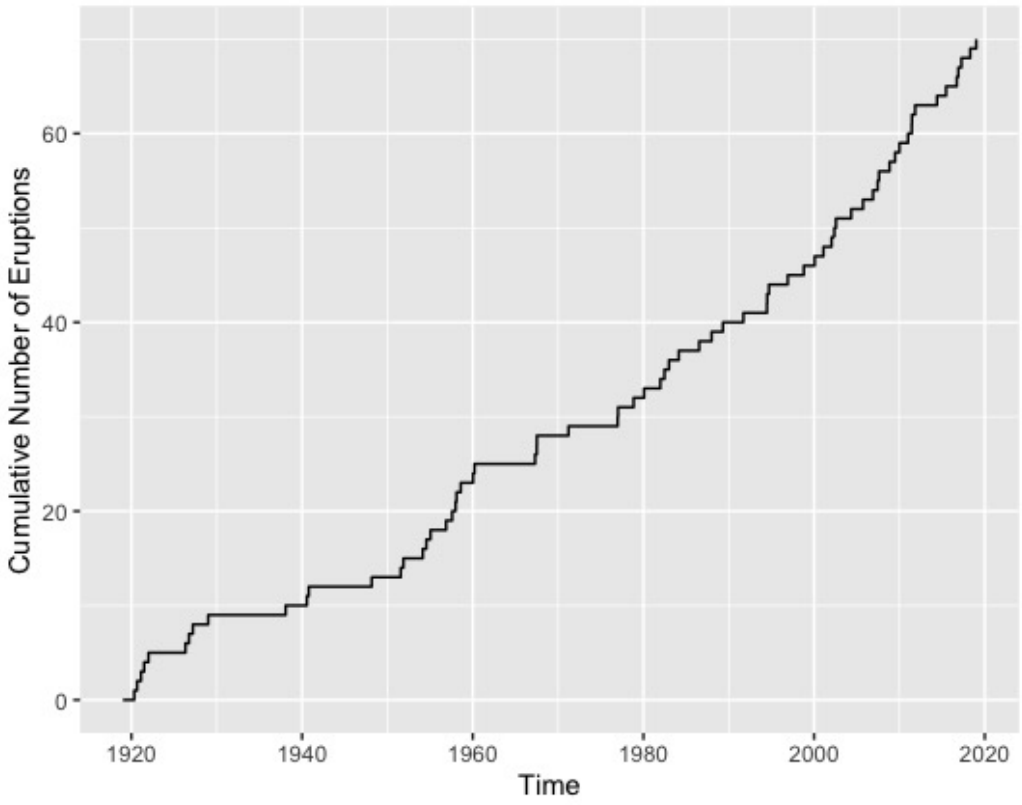

Figure 2. Step Plot of Cumulative Number of Eruptions against Time

The data shows two noticeable regimes: a regime featuring a number of long reposes and another (starting from around 1980) dominated by shorter reposes. The slight curvature was an indication that the eruption rate was non-constant. To check if this was the case the intensity was plotted as a function of time using a non-parametric method with points smoothed out using a Gaussian kernel function (see Gelissen (2016a) for the R code used in obtaining the intensity).

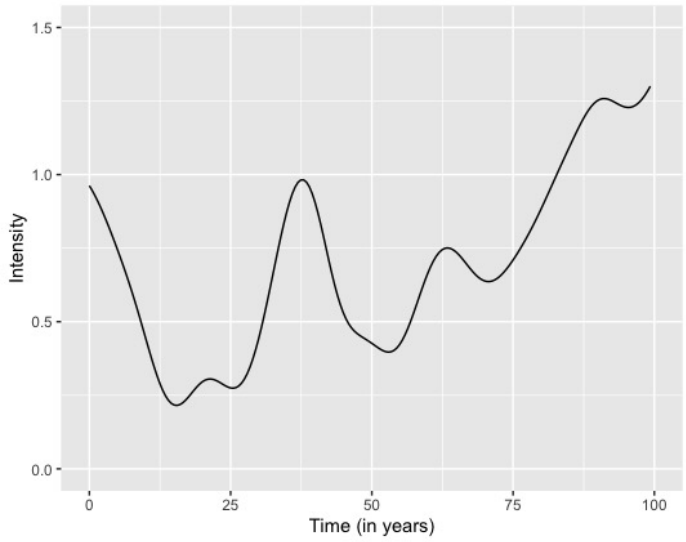

Figure 3. Line plot of Empirical Intensity

4.3.1 Testing for Randomness of Arrivals

A short interval of approximately a week (taken to be 8 days on average since some months contain 31 days) was chosen. The bar plot below shows Week 1 eruption counts over the 100-years data set, representing 18 eruptions.

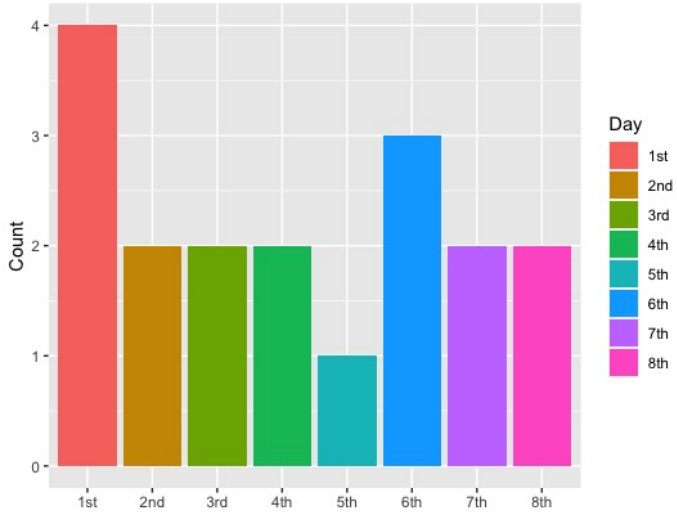

Figure 4. Bar Plot of Eruption Counts by First Week of Occurrence

The bar plot shows some uniformity in the counts but not conclusively. The K-S test was used to check the uniformity assumption by comparing the empirical cdf with the uniform cdf under the null hypothesis that there was no difference between the two distributions. Since $d = 0.139 < d_{18,0.05} = 0.453$, the null hypothesis was not rejected and the conclusion taken that Week 1 counts were uniform in distribution at 5% level of significance. An investigation of the other three intervals also led to the same conclusion though Week 2 showed a result on the boundary of statistical significance. This was because the calendar date 16th recorded 6 counts, quite a significant deviation from the mean of Week 2. There was no *a priori* reason to believe that this date was particularly special and it was concluded that its frequency was a matter of chance. The results are shown in the table below.

	n	d	$d_{n,0.05}$
Week 1	18	0.139	0.309
Week 2	12	0.375*	0.375
Week 3	20	0.200	0.294
Week 4	19	0.075	0.301

Table 1. Results for K-S Test for Uniformity on the Four Short Intervals

4.3.2 Testing for Independence of Arrivals

Three transformations of the inter-arrival times were chosen: natural logarithm; power; and first order differencing. Independence was visually demonstrated by plotting inter-event times against their position in the order of events. The four plots generated all showed 'white noise' behaviour typical of random data.

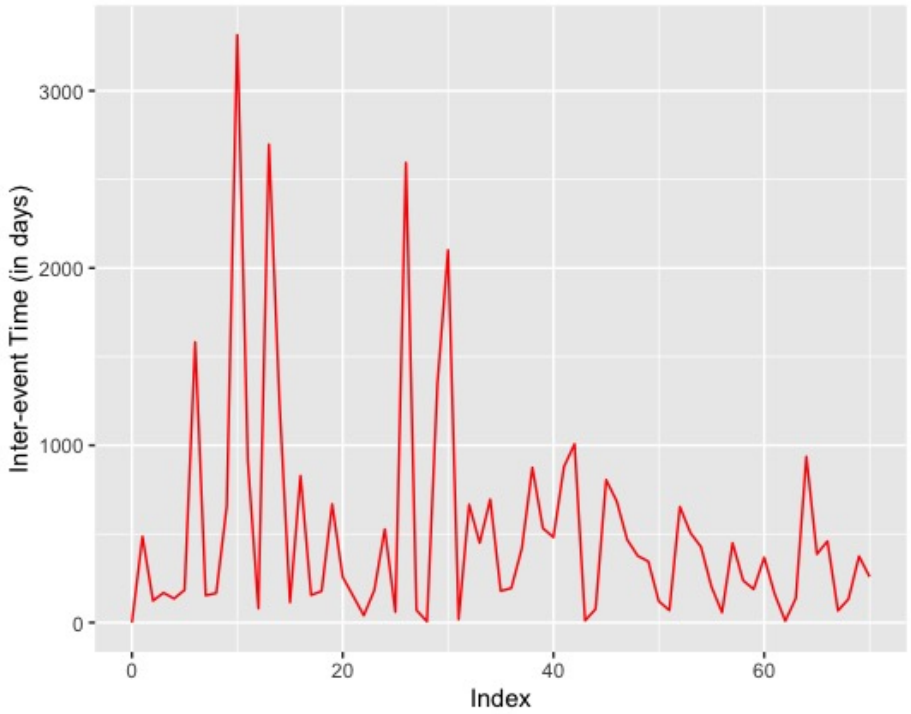

Figure 5. Plot of Raw Inter-Event Times

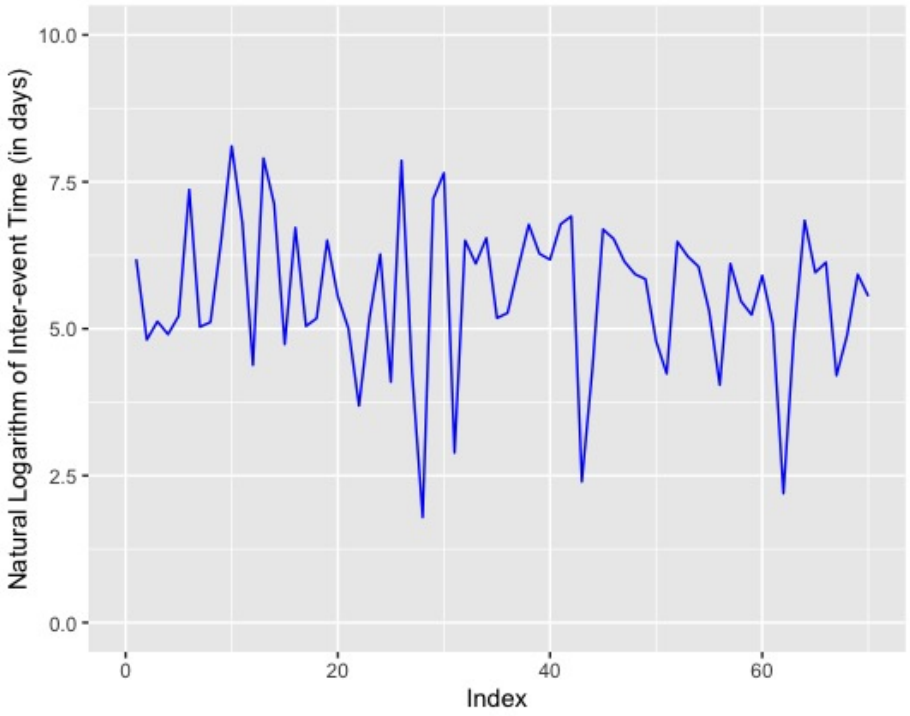

Figure 6. Plot of Natural Logarithm of Inter-Event Times

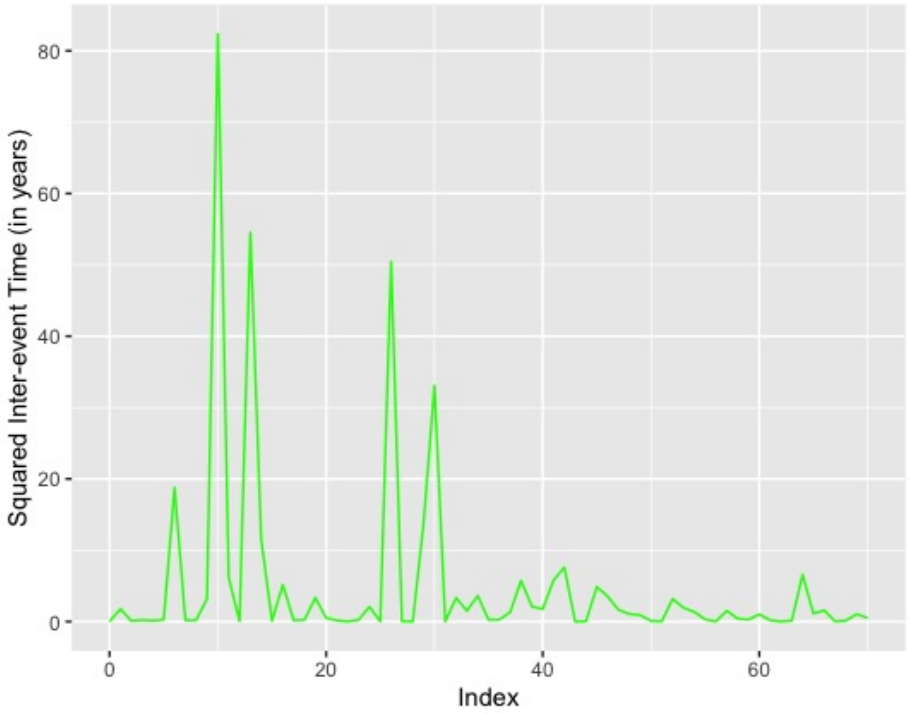

Figure 7. Plot of Squared Annualized Inter-Event Times

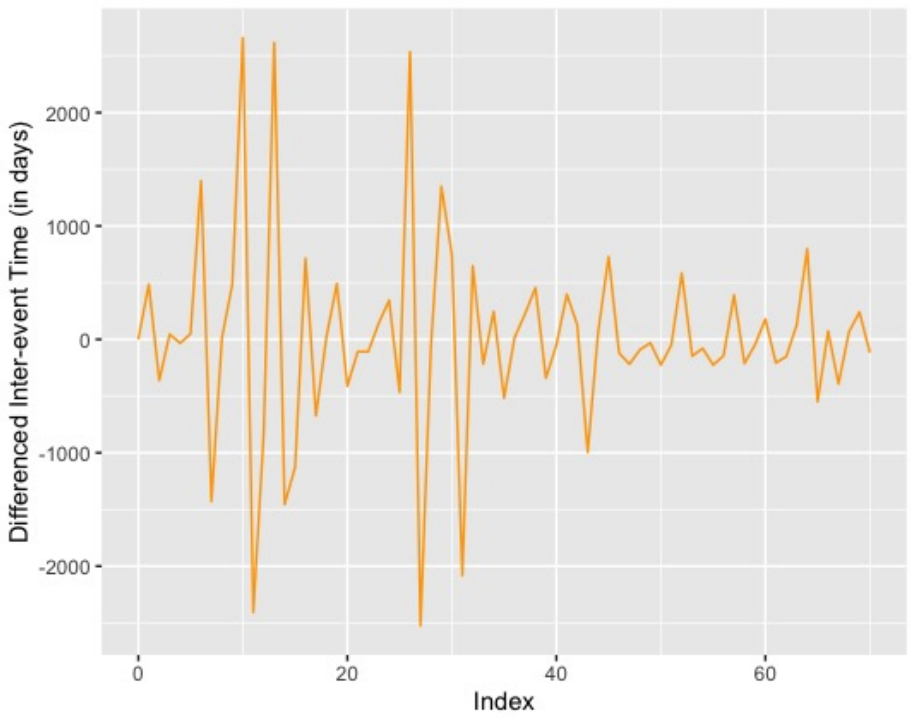

Figure 8. Plot of Differenced Inter-event Times

The plot of inter-event times against their first-order lagged values was used to augment the previous plots.

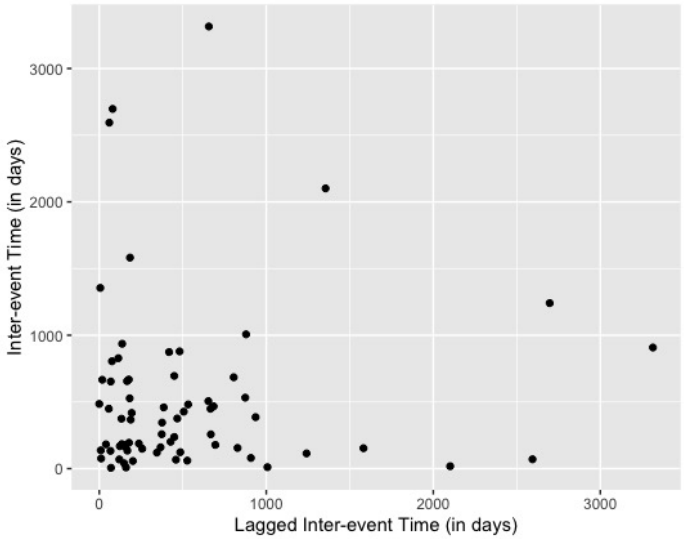

Figure 9. Plot of Inter-Event Times against First-Order Lagged Inter-Event Times

The argument for independence of arrivals was reinforced as the scatter plot showed no distinct pattern, which suggested that successive inter-arrival times were not correlated with each other.

Two procedures were carried out to confirm what was observed visually: regression of the inter-event times against their ordering and the Ljung-Box test. The null hypothesis for the Ljung-Box test was that arrivals were independent with no serial correlation for the first-order lag (any significant non-zero correlations are a result of chance). The results of the two tests are shown below.

	n	Multiple R^2	Ljung-Box Test p value
Raw Inter-event Times	69	0.0500	0.7423
Natural Logarithm of Inter-event Times	69	0.0142	0.8069
Squared Annualized Inter-event Times	69	0.0589	0.9061
First Order-Differenced Inter-event Times	69	0.0004	0.0007*

Table 2. Results of Tests for Serial Correlation of Various Transformations of the Inter-event Times

From the results it was concluded that inter-event times and their transformations are independent and arrivals are not serially correlated from one observation to the next. (though differenced inter-events times showed a deviant result for the Ljung-Box test).

4.3.3 Testing for Stationarity

The plot below shows five-year moving averages for repose times.

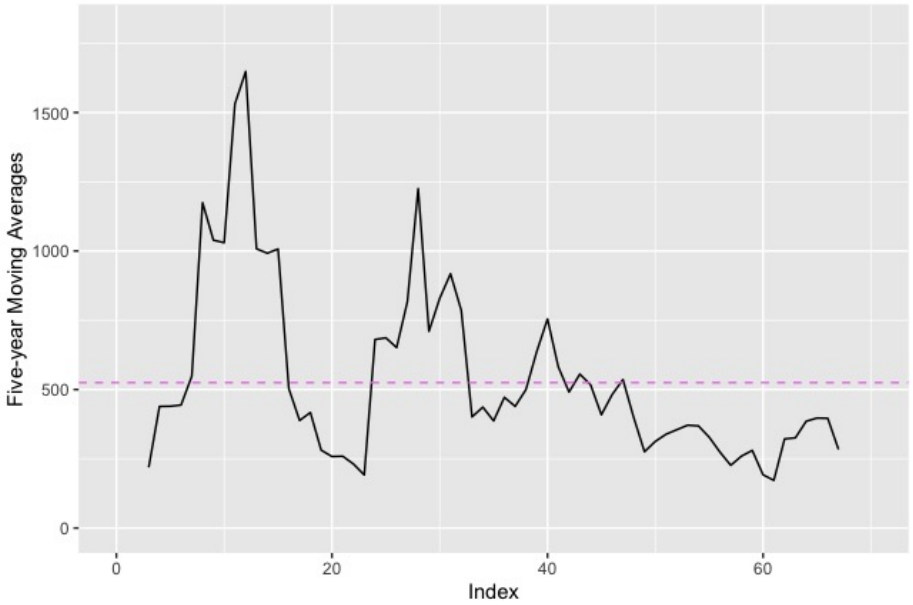

Figure 10. Line plot of Five-Year Moving Averages of Inter-Event Times

The shape of the line indicated stationarity; noticeably, the moving averages drift away from the mean (shown by a violet dashed line) towards the end, a sign of time trend. The KPSS test for trend stationarity on cumulative event times, however, returned a negative result (p value < 0.01) and it was concluded that the data was non-stationary at 5% level of significance.

4.3.4 Testing for Exponentiality of Transformed Inter-Event Times

For the analysis exponentiality was tested on the eruption time series for the period 1994 - 2019 with five blocks of equal length five years. A probability-to-probability (p-p) exponential plot was then produced to check the validity of the null hypothesis. The p-p exponential plot is shown below.

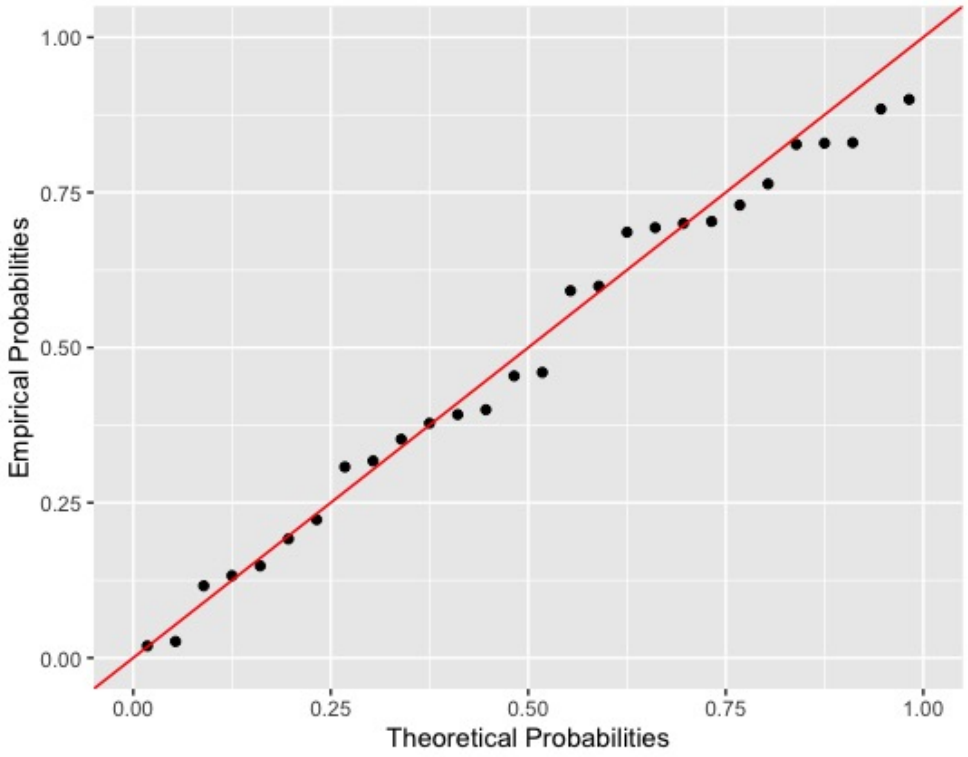

Figure 11. P-P Plot of Transformed Event Times from 1994 - 2019

The points fall mostly on or near the 45° line indicating that the null hypothesis should not be rejected and it was concluded that the transformed time points were iid exponential. To compliment the visual findings, the K-S test was carried out to determine if the transformed event times were the random realizations from an exponential distribution. The K-S test statistic obtained was $d = 0.00823$, which was compared with the table value $d_{28,0.05} \approx 0.257$, again leading to non-rejection of the null hypothesis. Similar investigation of other time periods came to the same conclusion.

4.4 Parameter Estimation and Simulation

MLE was used to obtain parameter estimates for the three models (R codes used in fitting the data adopted from Gelissen (2016a)). The results are shown in the table below.

	Parameter Estimates	ℓ	AIC
HPP	$\hat{\lambda}$ = 0.69	-94.60339	191.2068
Log-Linear NHPP	$\hat{\gamma}_0$ = -0.911036781	-91.81112	187.6222
	$\hat{\gamma}_1$ = 0.009976769		
Weibull Power NHPP	$\hat{\beta}$ = 1.167440	-93.81759	191.6352
	$\hat{\eta}$ = 2.660038		

Table 3. Results of Model Fitting

Based on the AIC the best model is the log-linear NHPP. The fitted models and empirical data were plotted on the same axes to see if the choice of the log-linear NHPP was justified.

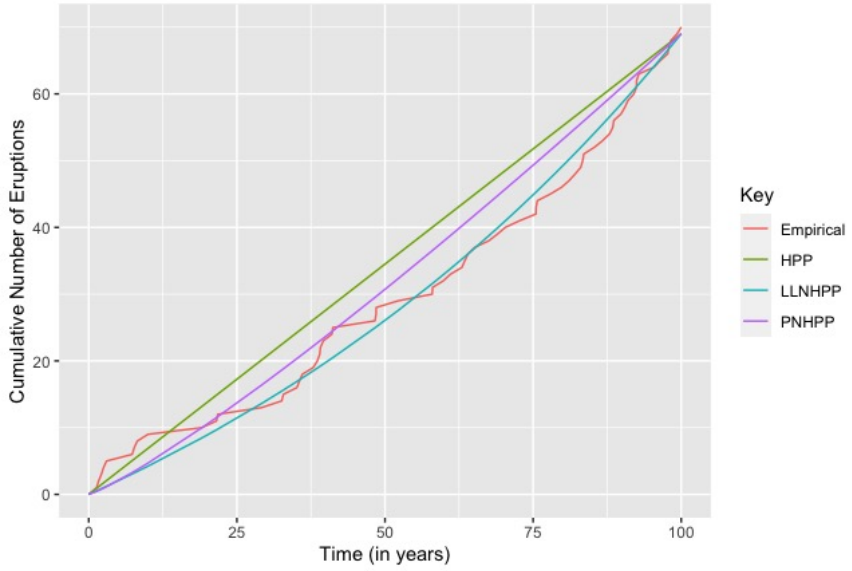

Figure 12. Plot of Cumulative Number of Eruptions Against Observed and Fitted Event Times

The plot shows that the log-linear NHPP was the best model of the three considered: it gave the best fit to the data.

4.5 Goodness of Fit

The log-linear intensity function was used to simulate a set of NHPP event times (R code used in the simulation adopted from Gelissen (2016b)). The eruption data was compared with the simulated data to check how well the model performed. This was done visually and through the K-S test.

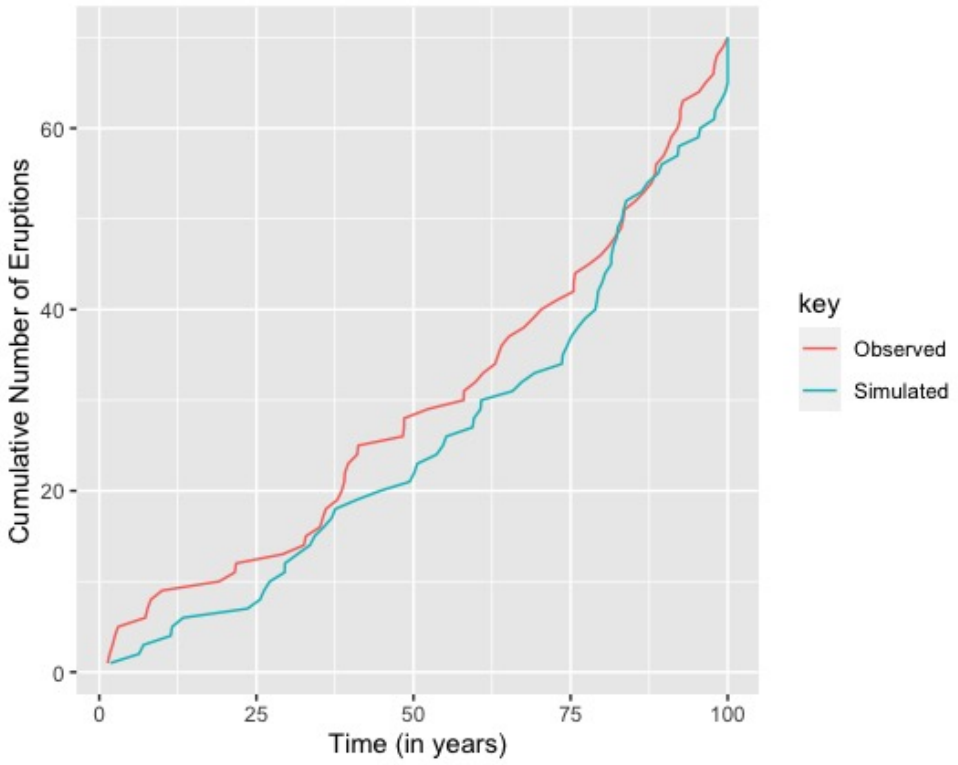

Figure 13. Plot of Cumulative Number of Eruptions Against Observed and Simulated Event Times

The plot showed a fairly good fit. The K-S test was performed to confirm the visual conclusion. The null hypothesis was that the two event times were similar in distribution. The K-S test statistic was d = 0.0875 compared against a critical value of d = 0.236 (p value = 0.9315). The null hypothesis was therefore not rejected and the conclusion drawn was that the two event times have similar distributions at 5% level of significance.

4.6 Prediction

The log-linear model was then used to forecast the cumulative number of eruptions in the next hundred years, together with confidence intervals for the estimates. The plot shown below illustrates this.

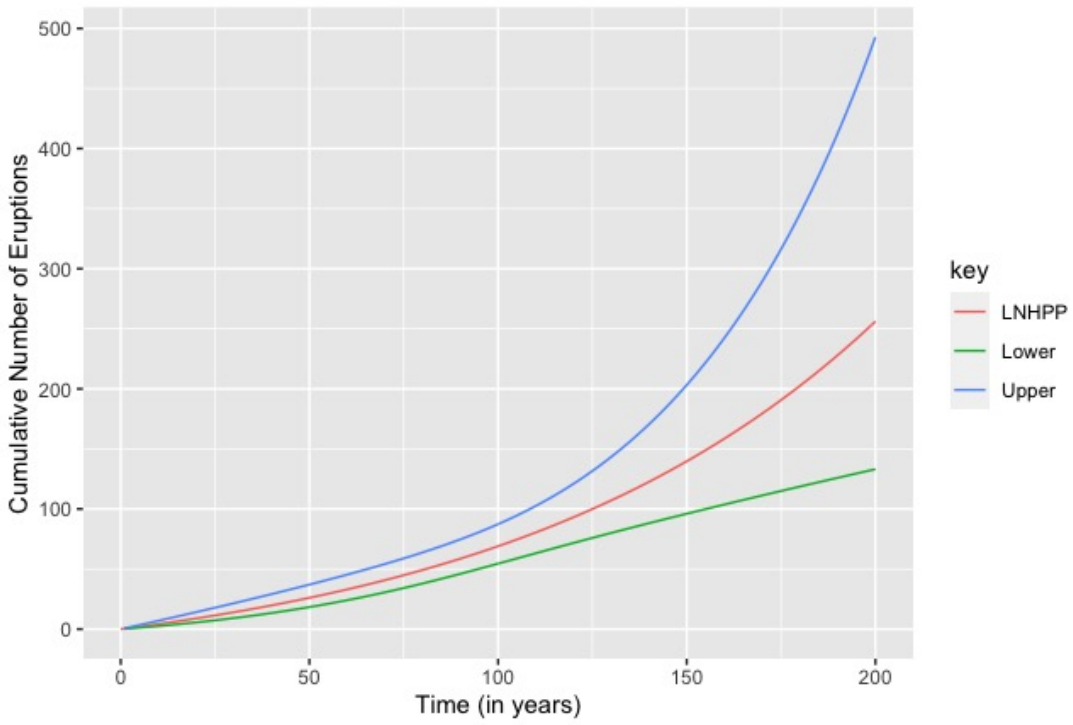

Figure 14. Plot of Cumulative Number of Eruptions Against Times

We might want, for example, to predict the number of eruptions between January 2019 and December 2034, i.e., $E[N(125) - N(100)]$. The model predicts that there will be about 31 (whole-number approximate of 30.96372) eruptions. This result was obtained by using Eq. 13. The confidence interval of this estimate is (18.49463, 51.83949). The standard errors were calculated by R using the delta method formula, i.e.,

$$Var[\hat{\Lambda}(t)] = \left(\frac{\partial \hat{\Lambda}(t)}{\partial \underline{\theta}}\Big|_{\underline{\theta}=\underline{\hat{\theta}}}\right)^2 Var[\hat{\lambda}(t)]$$

$Var[\hat{\lambda}(t)]$ is, in fact, the inverse of the Hessian and the square root of the diagonal gives the standard errors of the parameter estimates.

By the partitioning of a Poisson process a forecast on the number of eruptions of a particular volcano and a of particular VEI can be issued. The number of observed eruptions for Nyamuragira, for example, are 33 and so the model predicts it will have approximately 15 eruptions [(33/69)*31] from January 2019 to December 2034. The number of eruptions of VEI ≥ 3 are 13 and so the model predicts approximately 6 eruptions [(13/69)*31] in the next 25 years.

Probabilities can also be computed for a particular number of eruptions over an interval of choice. For instance, the probability of two or more eruptions from January 2019 to December 2021, i.e., $Pr(N(103) - N(100) \geq 2) \approx 0.8439$. This result was calculated using Eqs. 13 and 11. The plot below illustrates predictions for three different cumulative eruptions.

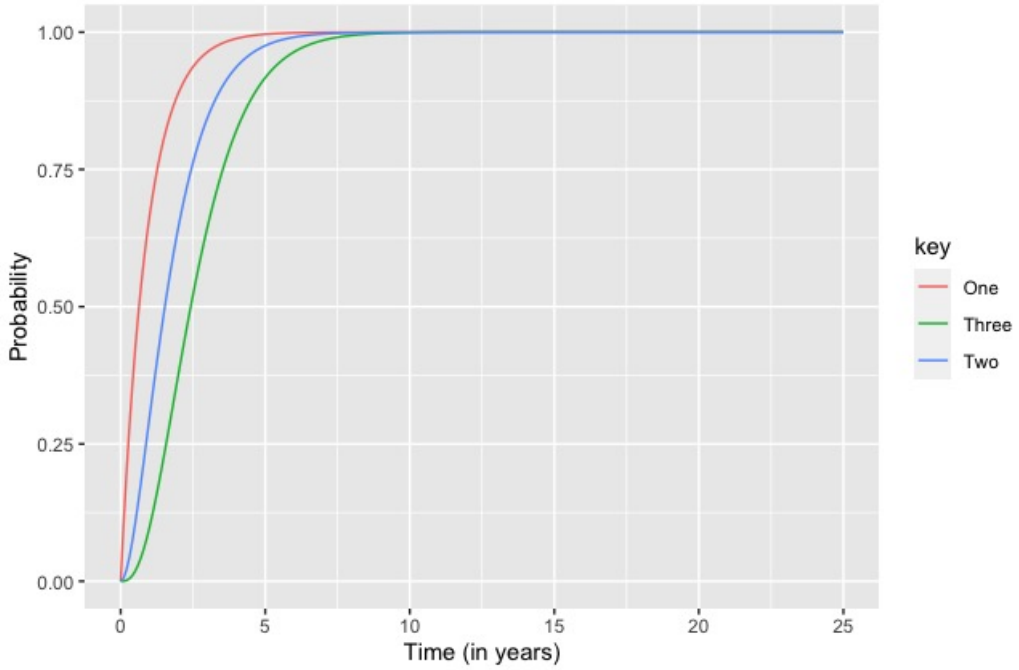

Figure 15. Plot of Predicted Probabilities for a 25-Year Period for One or More, Two or More and Three or More Eruptions

Chapter Five

Summary, Conclusion and Recommendations

5.1 Introduction

This chapter concludes the project by presenting a summary of the results, gives the challenges encountered and offers recommendations for further work in the area.

5.2 Summary

This project sought to find a Poisson model most appropriate in describing a curtate catalogue of eruptions. Before model fitting was carried out the major assumptions of Poisson models were elaborated upon and the eruption data was found to meet those assumptions. Of the three models considered a NHPP with a log-linear intensity function was found to best explain the data. The most parsimonious models are usually preferred and in this case the most parsimonious model considered was the single-parameter HPP. However, it was not justified with respect to the model results. In fact, the data validation process established that the data was non-stationary, making the HPP the least obvious candidate for selection as a model. While the quality of the eruption data was not called into question because it was beyond the scope of the project, the issue of incomplete eruption records (which usually favour rejection of stationary models) is a long standing concern in statistical volcanology. Indeed, the data considered showed a number of long reposes in the early record. The Weibull power function performed the poorest. Of note, however, is the fact that shape parameter $\beta = 1.167440$. Ho (1991) identified this parameter as the indicator of waxing or waning of eruptive activity. Because $\beta > 1$, it was concluded that there was increase in volcanic eruptions with time. This coincides with the increasing trend that the log-linear model predicts. The forecasting possibilities of the log-linear model were demonstrated with the model predicting of an increase in eruptive activity with time though the monotonic-increasing nature of the model (in general contravention of the real-life observable behaviour of volcanoes) means that only short-term forecasting of, say, one or two decades, will give plausible results.

5.3 Conclusion

The project was able to meet all four objectives and the log-linear NHPP model with intensity function $\lambda(t) = exp(0.911036781 + 0.009976769t)$ was shown to be statistically tenable and a good fit was found between the observed data and the simulated data with 95% confidence.

5.4 Recommendations

The model results forecast an increase in activity and it is only a matter of time before a serious eruption occurs again. The last major eruption occurred in the East African region was in June 2011. Nabro in Eritrea animated and erupted violently despite having had no historic eruptions and being thought to be extinct by the scientific community. The resulting ash cloud was dispersed northwesterly, disrupting air travel throughout the Horn of Africa and the Middle East. The eruption caused thousands to be evacuated and led to some fatalities (*Global distribution of volcanism: Regional and country profiles*, 2015). A volcanic eruption in a more densely populated area would have wreaked more havoc. It is recommended that East African governments and their disaster management authorities incorporate these findings in their disaster preparedness frameworks and hazard assessments and institute robust monitoring of the volcanoes in their jurisdictions to avoid being caught unawares in event of another major, possibly more violent eruption.

5.5 Challenges Encountered

The main challenge encountered was the completeness of the data set considered. A few of the dates in the sample extracted were approximate and the dates for a few data points had to be interpolated, which no doubt reduced the accuracy of the model results. The near-total lack of documentation of historical eruptions meant that the project had to be constrained to a time span where the historical record is more reliable.

5.6 Future Work

Three volcanoes (Nyamuragira, Ol Doinyo Lengai and Nyiragongo) represent about 80% of all eruptions in the East African region in the last century. A narrow focus on these three volcanoes might yield useful results.

References

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, 95(3), 631–636.

Bebbington, M. S., & Lai, C. D. (1996a). On non-homogeneous models for volcanic eruptions. *Math. Geol.*, 28(5), 585–600.

Bebbington, M. S., & Lai, C. D. (1996b). Statistical analysis of New Zealand volcanic occurrence data. *J. Volcanol. Geotherm. Res.*, 74(1-2), 101–110.

Brown, L. D., Zhao, L. H., Shen, H., & Mandelaum, A. (2004). *Multifactor poisson and gamma-poisson models for call center arrival times.* *https://repository.upenn.edu/statistics_papers/148*

Brown, L. D., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., & Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469), 36-50.

Brown, S. K., Sparks, R. S. J., Mee, K., Vye-Brown, C., Ilyinskaya, E., Jenkins, S., Loughlin, S. C., et al. (2015). *Global distribution of volcanism: Regional and country profiles. Report IV of the GVM/IAVCEI contribution to the UN-ISDR global assessment report on disaster risk reduction 2015.* Global Volcano Model and International Association of Volcanology and Chemistry of the Earth's Interior.

Çinlar, E (2013). *Introduction to stochastic processes.* Dover Publications.

Connor, C. B., Sparks, R. S. J., Mason, R. M., & Bonadonna, C., (2003). Exploring links between physical and probabilistic models of volcanic eruptions: The Soufrière Hills Volcano, Montserrat. *Geophys. Res. Lett.*, 30(13), 1701.

Cottrell, E., (2014). Global distribution of active volcanoes. In J. F. Shroder & P. Papale (Eds.). *Volcanic Hazards, Risks, and Disasters* (pp. 1-16). Elsevier.

Cox, D. R., & Lewis, P. A. W. (1966). *The statistical analysis of series of events.* John Wiley & Sons.

Daniel, W. W. (1990). *Applied nonparametric statistics* (2nd ed.). PWS-Kent.

De la Cruz-Reyna, S. (1991). Poisson-distributed patterns of explosive eruptive activity. *Bulletin of Volcanol.*, 54(1), 57–67.

Dzierma, Y., & Wehrmann, H. (2010). Statistical eruption forecast for the Chilean southern volcanic zone: typical frequencies of volcanic eruptions as baseline for possibly enhanced activity following the large 2010 Concepction earthquake. *Nat. Hazards. Earth Sys. Sci.*, 10, 2093-2108.

Gelissen, S. (2016a). *R code for fitting a nonhomogeneous temporal Poisson process model* [R Source code]. *https://www.blogs2.datall-analyse.nl*

Gelissen, S. (2016b). *R code for fitting a nonhomogeneous temporal Poisson process model using the spatstat package* [R Source code]. *https://www.blogs2.datall-analyse.nl*

Gilbert, J. S., & Sparks, R. S. J. (1998). Future research directions on the physics of explosive volcanic eruptions. *Geol. Soc. London Spec. Publ.*, 145(1), 1–7.

Gusev, A. A. (2008). Temporal structure of the global sequence of volcanic eruptions: Order clustering and intermittent discharge rate. *Phys. Earth Planet. Inter.*, 166(3-4), 203–218.

Ho, C. H. (1991). Non-homogeneous Poisson model for volcanic eruptions. *Math. Geol.*, 23(2), 167–173.

Ho, C. H. (1992). Predictions of volcanic eruptions at Mt. Vesuvius, Italy. *J. Geodyn.*, 15(1), 13-18.

Klein, F. W. (1982). Patterns of historical eruptions at Hawaiian volcanos. *J. Volcanol. Geotherm. Res.*, 12(1-2), 1–35.

Klein, R. W., & Roberts, S. D. (1984). A time-varying Poisson arrival process generator. *Simulation*, 43(4), 193-195.

Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). Chapman & Hall.

Loughlin, S. C., Sparks, S., Brown, S. K., Jenkins, S. F., & Charlotte Vye-Brown, C. (Eds.). (2015). *Global volcanic hazards and risk*. Cambridge University Press.

Marzocchi, W., & Zaccarelli, L. (2006). Quantitative model for the time-size distribution of eruptions. *J. Geophys. Res.*, 3, B04204.

Mendoza-Rosas, A. T., & De la Cruz-Reyna, S. (2009). A mixture of exponentials distribution for a simple and precise assessment of the volcanic hazard. *Nat. Hazards Earth Syst. Sci.*, 9(2), 425–431.

Mulargia, F., Tinti, S., & Boschi, E. (1985). A statistical analysis of flank eruptions on Etna volcano. *J. Volc. Geotherm. Res.*, 23, 263–272.

Myung, I. J. (2003). Tutorial on Maximum Likelihood Estimation. *Journal of Mathematical Psychology*. 47(1), 90–100.

Newhall, C., & Self, S. (1982). The Volcanic Explosivity Index (VEI): An estimate of explosive magnitude for historical volcanism. *Journal of Geophysical Research*, 87(C2), 1231–1238.

Reyment, R. A. (1969). Statistical analysis of some volcanologic data regarded as series of point events. *Pure Appl. Geophys.*, 74(1), 57–77.

Ross, S. M. (1996). *Stochastic processes* (2nd ed.). John Wiley & Sons.

Ross, S. M. (2010). *Introduction to probability models* (10th ed.). Academic Press.

Salvi, F., Scandone, R., & Palma, C. (2006). Statistical analysis of the historical activity of Mount Etna, aimed at the evaluation of volcanic hazard. *J. Volcanol. Geotherm. Res.*, 154, 159–168.

Sanchez, L. A. (2014). *Statistical analysis and computer modelling of volcanic eruptions* (Doctoral dissertation). *https://ir.lib.uwo.ca/etd/1912*

Settle, M., & Mcgetchin, T. R. (1980). Statistical analysis of persistent explosive activity at Stromboli, 1971: Implications for eruption prediction. *J. Volcanol. Geotherm. Res.*, 8(1), 45–58.

Sigman, K. (2009). *Poisson processes, elementary renewal theorem with proof* [Course notes]. IEOR 6711. Stochastic Modeling I.
http://www.ieor.columbia.edu/∼ sigman/stochastic-I.html

Sigurdsson, H., Houghton, B., McNutt, S. R., Rymer, H., & Stix, J. (2015). *Encyclopedia of volcanoes* (2nd ed.). Academic Press.

Smethurst, L., James, M. R., Pinkerton, H., & Tawn, J. A. (2009). A statistical analysis of eruptive activity on Mount Etna, Sicily. *Geophys. J. Int.*, 179, 655–666.

Wickman, F. E. (1966). Repose period patterns of volcanoes. I. Volcanic eruptions regarded as random phenomena. *Ark. Kem. Mineral. Geol.*, 4(4), 291-367.

Wilson, L. (2009). Volcanism in the solar system. *Nat. Geosci.*, 2(6), 388-396.

[Worldwide distribution of active subaerial volcanoes]. (n.d). *https://www.chegg.com*. W.W.Norton & Company, Inc. (Copyright 2019).

Appendix: List of Volcanoes

Name	Location	Number of Eruptions*	Date of Last Eruption*
Alu-Dalafilla	Ethiopia	1	3-11-2008
Ardoukoba	Djibouti	1	7-11-1978
The Barrier	Kenya	1	31-12-1921
Dabbahu	Ethiopia	1	26-9-2005
Dallol	Ethiopia	2	4-1-2011
Erta Ale	Ethiopia	3	2-7-1967
Ol Doinyo Lengai	Tanzania	15	9-4-2017
Manda Hararo	Ethiopia	2	28-6-2009
Manda-Inakir	Djibouti/Ethiopia	1	31-12-1928
Nabro	Eritrea	1	13-6-2011
Nyamuragira	DRC	33	18-4-2018
Nyiragongo	DRC	7	17-5-2002
Visoke	DRC/Rwanda	1	1-8-1957

*for time period considered