8-2016

Spatio-Temporal Analysis of Point Patterns

Abdul-Nasah Soale
East Tennessee State University

Spatio-Temporal Analysis of Point Patterns

A thesis

presented to

the faculty of the Department of Mathematics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

---

by

Abdul-Nasah Soale

June 2016

Edith Seier, Ph.D., Chair

Michele L. Joyner, Ph.D.

Robert M. Price Jr., Ph.D.

Keywords: earthquakes, K-function.

ABSTRACT

Spatio-Temporal Analysis of Point Patterns

by

Abdul-Nasah Soale

In this thesis, the basic tools of spatial statistics and time series analysis are applied to the case study of the earthquakes in a certain geographical region and time frame. Then some of the existing methods for joint analysis of time and space are described and applied. Finally, additional research questions about the spatial-temporal distribution of the earthquakes are posed and explored using statistical plots and models. The focus in the last section is in the relationship between number of events per year and maximum magnitude and its effect on how clustered the spatial distribution is and the relationship between distances in time and space in between consecutive events as well as the distribution of the distances.

ACKNOWLEDGMENTS

First and foremost, I would like to thank the Almighty Allah for giving me the strength and ability to be able to come this far. I would also like to acknowledge my gratitude to my adviser and the other members of my thesis committee, from whom I learned so much.

Finally, I would like to acknowledge the faculty and colleagues who have, directly and indirectly, contributed to the completion of the thesis.

TABLE OF CONTENTS

ABSTRACT 2

ACKNOWLEDGMENTS 4

LIST OF TABLES 7

LIST OF FIGURES 10

1 INTRODUCTION 11

2 BACKGROUND THEORY IN SPATIAL STATISTICS 13

2.1 Edge-Effects 13

2.2 First-Order Properties 15

2.3 Second-Order Properties 16

2.4 Complete Spatial Randomness (CSR) 16

2.5 The $\mathcal{G} - function$ 17

2.6 The $\mathcal{F} - function$ 18

2.7 The $\mathcal{K} - function$ 18

2.8 The $\mathcal{J} - function$ 19

3 CASE STUDY 20

3.1 Data Set 20

3.2 Spatial Analysis 23

3.2.1 First-Order Properties 23

3.2.2 Second-Order Properties 25

3.3 Analysis With Respect to Time 30

3.3.1 Time Series of the Frequency of Events Per Year and Per Year by Months 30

3.3.2 Time Series Analysis of the Frequency of Events in a Given Month for All Years 36
4 EXPLORING THE SPATIAL POINT PATTERNS SIMULTANEOUSLY WITH RESPECT TO SPACE AND TIME 43
4.1 First-Order Property 47
4.2 Second-Order Property 49
5 FURTHER EXPLORATION OF EARTHQUAKES WITH RESPECT TO SPACE AND TIME 54
5.1 Spatial Analysis of Events in Years and Months With Extreme Number of Events 54
5.2 Data Simulation With an Inhomogeneous Poisson Process . . 56
5.3 Analysis of Consecutive Events 59
5.4 Time Between Consecutive Events 62
5.5 Relationship Between Magnitude and Frequency of Events . . 63
6 CONCLUSIONS 67
BIBLIOGRAPHY 69
VITA 71

LIST OF TABLES

1 Quadrat Count 23

2 $\mathcal{F}$ Estimates 26

3 $\mathcal{G}$ Estimates 26

4 $\mathcal{K}$ Estimates 27

5 $\mathcal{K}$ and $\mathcal{L}$ Estimates 28

LIST OF FIGURES

1 Buffer zone method of edge correction 14

2 Study region 21

3 UTM and Lat-Long compared 22

4 Quadrat Count 24

5 Spatial Intensity 24

6 F ,G, J and K functions plots 25

7 Envelopes plot of K and L functions 28

8 Histogram of distance to nearest neighbor and $G - function$ for earthquakes 29

9 Time series plots for frequency events from 1974 - 2014 per year by months (top) and frequency of events from 1974 - 2014 per year(bottom). 30

10 Box plots for frequency events from 1974 - 2014 per year (left) and frequency of events from 1974 - 2014 per year by months (right). . . . 31

11 Autocorrelation and Partial-autocorrelation for earthquakes from 1974 - 2014 33

12 Periodogram and Cumulative Periodogram 34

13 Histograms of frequency of earthquakes (A) and log of frequency of earthquakes (B) 35

14 Monthly time series for the frequency of earthquakes from 1974 - 2014 37

15 Autocorrelation function for monthly time series for the frequency of earthquakes from 1974 - 2014 38

16 Monthly time series for the frequency of earthquakes from 1974 - 2014 39

17 Maximum magnitude of earthquakes per year 40
18 ACF and PACF of maximum magnitude of earthquakes per year . . . 41
19 Histogram and normal qqplot of maximum magnitude of earthquakes per year 42
20 Scatter plot of earthquakes 46
21 Plot of earthquakes through time 46
22 Estimated temporal intensity of the time of occurrence of events in weeks 48
23 Kernel Estimate of Spatial Intensity 49
24 Inhomogeneous K-function estimate (Khat) 51
25 Khat 52
26 PCFhat 52
27 PCF Perspective Plot 53
28 Years with the least frequency of events 55
29 Years with the most frequency of events 56
30 Months with the most frequency of events 57
31 Random Inhomogeneous Poisson Process with estimated kernel. The red circles indicate different concentration within the same region from the two simulations. 58
32 Random Homogeneous Poisson Process with estimated kernel 58
33 Distance and days between consecutive earthquakes 60
34 Empirical density and cumulative density distribution of time between consecutive events 62
35 Cullen Frey plot 63

36 Comparison of Weibull, exponential, and gamma distribution 64

37 Quadratic model fit for the frequency of events per year on maximum magnitude of events 65

38 Non-parametric quadratic model fit for the frequency of events per year on maximum magnitude of events 66

1 INTRODUCTION

Earthquakes remain a constant nightmare for most South American countries especially areas along the coast. Earthquakes occur when there is a sudden slip of faults within the earth's crust resulting in ground shaking and release of energy called seismic energy. Earthquakes occur randomly in time and with regard to space, they tend to occur more in certain regions near the boundaries of tectonic plates. Thus, a plot of earthquakes in a region looks like a spatial point pattern.

According to Adrian [2], a **spatial point pattern** gives the location of events occurring in a region under study. The study region can be 2-dimensional, 3-dimensional or multidimensional. Numata [13] shows 65 Japanese black pine saplings in a square with side 5.7 meters. This is an example of a two dimensional study region. In some cases, the location and an additional factor is considered. In such cases, we refer to the additional factor as mark and the point pattern is referred to as a *marked point pattern*. Marks could be categorical, multivariate or take other complicated forms. Diggle and Rowlingson [6] studied the location of the residence of asthmatic and non-asthmatic cases in North Derbyshire in 1992. Asthmatic and non-asthmatic cases were assigned different labels. This is an example of a marked process.

Spatial point patterns may also include not only the location of the events but also the times at which these events occur. In such a case, the point pattern is called *Spatio-temporal point pattern*. Thus, **Spatio-temporal Point Pattern** is a set of events that happen in a given study region during a certain interval of time. An example of spatio-temporal point pattern is the location and time of 10,572 cases of non-specific gastro-intestinal disease in the county of Hampshire, UK reported to the

National Health Service between January 2001 and December 2003 [5].

This thesis focuses on spatio-temporal analysis of earthquakes in South America between latitude $-18°$ and $0°$ and longitude $-83°$ and $-68°$. Our main objective is to study the second-order properties of the earthquakes using the nonparametric inhomogeneous $K - function$ proposed by Gabriel and Diggle [7].

In Section 2, we reviewed basic concepts and tools of spatial statistics that are used later such as edge effects, completely spatial randomness, first and second order properties. We apply separate analysis of space and time in Section 3. Next, in Section 4, we apply space-time analysis using the STIKHAT and PCF functions in stpp [8] package in R developed by Gabriel and Diggle [7]. Finally, in Section 5 we propose some additional analysis of space and time together. There we address the following questions:

- Is the spatial distribution of events per year related to the number of events in the year?
- Is the earthquake temporal space process a homogeneous or inhomogeneous point process?
- How is time and space related for consecutive events?
- Is the distribution of the number of events in a year related to the intensity of the strongest earthquake?

2 BACKGROUND THEORY IN SPATIAL STATISTICS

2.1 Edge-Effects

In analyzing spatial point patterns, there are instances where events occurring outside the study region interact with some of the events being observed. However, because these events are not observed, it is difficult to keep track of them. This phenomenon is known as edge-effects. Edge-effects may or may not be ignored in the exploratory analysis depending on the type of study.

More generally, we can distinguish between three broad approaches to handling edge-effects: the use of buffer zones, explicit adjustments to take account of unobserved events, and when the region is rectangular, wrapping the region onto a torus by identifying opposite edges [5].

First, with the buffer zone method, we choose a region, say B, within a specified distance, say d_0, from the edge of the study region S, as the buffer zone. Then we perform the statistical analysis after conditioning on all events that fall in the buffer zone. For any event x in the remaining region, say R, within the region S outside the buffer zone, if $d \leq d_0$, then the observed number of events within a distance d from x must equal the actual number of events within a distance d from x within the underlying process. However, if $d > d_0$ the observed number of events within a distance d from x may be less than the actual number of events within a distance d from x within the underlying process. Therefore, estimates based on these observed values may be biased. There is no specific choice of d_0. Hence, depending on the statistical analysis, d_0 may be varied to avoid residual edge-effects or leaving out

of data unnecessarily. Figure 1 below, illustrates the buffer zone method of edge correction.

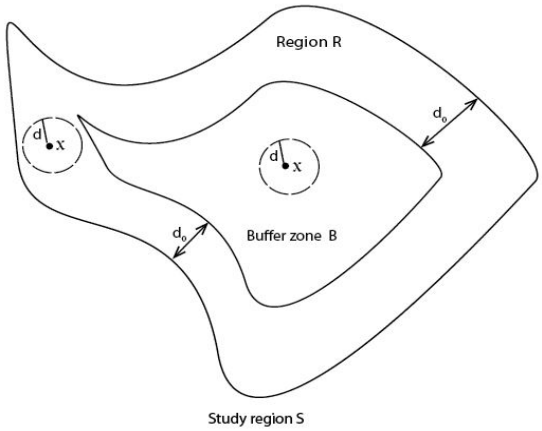

Figure 1: Buffer zone method of edge correction

Secondly, with the adjustment method, we adjust for the unobserved events outside region S. Usually, this adjustment is based on an average estimate of the number of observed events within a distance d of any point x. We do this by letting a denote the area of the circle with a radius d centered at x, then we estimate the actual number of events, n, within a distance d from x as $\frac{n\pi d^2}{a}$. This method is very good, because it makes use of all the observed data but has the tendency to increase the sample variance.

Lastly, we can reduce edge-effects by wrapping a rectangular study region on a torus. This method is mostly used for simulating various point process realizations.

2.2 First-Order Properties

First-order properties measure the distribution of events in the study region: intensity and spatial intensity [14]. Intensity is the expected number of points per unit area. In other words, the average density of points. A point process with constant intensity is called a homogeneous or uniform process, while that with varying intensity is termed inhomogeneous process. The intensity function is defined by;

$$\lambda(x) = \lim_{|dx|\to 0} \left\{ \frac{E[N(dx)]}{|dx|} \right\} \tag{1}$$

where $N(dx)$ is the number of events in a small region dx and $|dx|$ is the area of the region dx. [5]. For a homogeneous process, the intensity estimate $\widehat{\lambda}(x) = \frac{N}{|S|}$ where N is the total number of events and $|S|$ is the area of the study region.

For an inhomogeneous process as in our case (earthquake epicenters are usually concentrated along fault lines), quadrat counting or kernel smoothing maybe used for determining the intensity. The unbiased estimator of the intensity is the kernel density estimator,

$$\widehat{\lambda}(x) = \frac{1}{h^2} \sum_{i=1}^{n} \frac{\kappa(\frac{\|x-x_i\|}{h})}{q(\|x\|)} \tag{2}$$

where $x_i \in \{x_1, x_2, ...x_n\}$ is an observed point, h is the bandwidth and $q(\|x\|)$ is the border correction [14]. The level of smoothing depends on the bandwidth h. There is no general rule for selecting the bandwidth. However, Berman and Diggle (1989) proposed a criterion for choosing a bandwidth that minimizes the mean square error (MSE).

2.3 Second-Order Properties

Here we are interested in investigating the level and type of interaction between events. That is, whether the events are independent, regular or clustered. In an informal sense, for any two points x and y, the second-order intensity is the probability of any pair of events occurring around location x and y, respectively. Second-order properties can be measured using the $K-function$ or other type of functions we will see later.

2.4 Complete Spatial Randomness (CSR)

A completely spatially random process is a homogeneous Poisson Point Process (HPP) [3]. For Complete Spatial Randomness(CSR), the following must hold:

1. Events in a given region are independent and uniformly distributed.

2. The number of events in a given region, say S, follows a Poisson distribution with mean $\lambda|S|$, where λ is the intensity or mean number of events per unit area and $|S|$ is the area of the region.

The first property implies that the occurrence of an event say, x does not affect the probability of occurrence of another event, say y nearby. Thus, there is no interaction between events in the same neighborhood. Similarly, the second property implies that the intensity or the mean number of events is the same everywhere in the region. Hence, we say the process is a **Homogeneous**. When the intensities vary within the region, the process is termed an **Inhomogeneous**. CSR is an ideal process and usually not achievable in reality. However, CSR is useful in exploratory analysis of a data set and also for pattern distinction as regular, clustered or random.

In this thesis, we test for CSR using a distance method. We are considering nearest distances, pairwise distances and empty space distances.

Pairwise distances $d_{ij} = \|x_i - x_j\|, \forall i \neq j \in S$.

Nearest distances $d_i = min\{d_{ij}, \forall i \neq j \in S\}$ for each point i in S.

Empty space distances $d(u) = min_i\|u - x_i\|, \forall i \in S$, the distance between from a fixed reference location u in S to the nearest data point [2]. From these distances we estimate the $\mathcal{G} - function$, $\mathcal{F} - function$, $\mathcal{K} - function$ and $\mathcal{J} - function$.

2.5 The $\mathcal{G} - function$

The $\mathcal{G} - function$ measures the distribution of distances from an arbitrary event to its nearest neighbors [14]. Let d_i denote the distance from event i to other nearest events in the region. For n events in the region S,

$$\widehat{\mathcal{G}}(r) = \frac{1}{n}\sum_{i=1}^{n}(d_i \leq r) \tag{3}$$

$$d_i = \begin{cases} 1 & \text{if } d_i \leq r, \forall i \\ \\ 0 & otherwise \end{cases}$$

where, $d_i = min_j\{d_{ij}, \forall j \neq i \in S\}, i = 1, 2, ..., n$. The expected value of the $\mathcal{G} - function$ under CSR with intensity λ is $\mathcal{G}(r) = 1 - e^{-\lambda\pi r^2}$. When $\mathcal{G}(r) > 1 - e^{-\lambda\pi r^2}$, a clustering pattern is suggested, while $\mathcal{G}(r) < 1 - e^{-\lambda\pi r^2}$ suggests a regular pattern. We compare point processes with the CSR by plotting the empirical function $\widehat{\mathcal{G}}(r)$ against the theoretical expectation $\mathcal{G}(r)$.

2.6 The $\mathcal{F} - function$

The $\mathcal{F} - function$ measures the distribution of all distances from an arbitrary point k in the plane to the nearest observed event j [14].

$$\widehat{\mathcal{F}}(r) = \frac{1}{m}\sum_{k=1}^{m}(d_k \leq r) \tag{4}$$

$$d_k = \begin{cases} 1 & \text{if } d_k \leq r, \forall k \\ 0 & otherwise \end{cases}$$

where, $d_k = min_j\{d_{kj}, \forall j \in S\}, k = 1, 2, ..., m, j = 1, 2, ..., n$. The expected value of the $\mathcal{F} - function$ under CSR with intensity λ is $\mathcal{F}(r) = 1 - e^{-\lambda\pi r^2}$.

Unlike the $\mathcal{G} - function$, $\mathcal{F}(r) < 1 - e^{-\lambda\pi r^2}$ suggests a clustering pattern, while $\mathcal{F}(r) > 1 - e^{-\lambda\pi r^2}$ suggests a regular pattern. We compare point processes with the CSR by plotting the empirical function $\widehat{\mathcal{F}}(r)$ against the theoretical expectation $\mathcal{F}(r)$.

2.7 The $\mathcal{K} - function$

The $\mathcal{K} - function$ also known as the Ripley's $\mathcal{K} - function$ is denoted by $\mathcal{K}(r)$. For a *stationary process*, it is the expected number of points within a distance r of an arbitrary event in the point process. That is,

$$\widehat{\mathcal{K}}(r) = \frac{E(\#\text{ of events within r distance of an arbitrary event})}{\lambda} \tag{5}$$

where λ is the intensity of the point process.

Considering a region S, the expected number of events is $\lambda|S|$. Thus, under CSR, $\mathcal{K}(r) = \pi r^2$. Again, $\mathcal{K}(r) > \pi r^2$ suggests clustering, while $\mathcal{K}(r) < \pi r^2$ suggests a

regular pattern.

Various estimators for $\mathcal{K} - function$ have been proposed. We are considering the estimator based on pairwise distances that takes edge-effects into account.

$$\widehat{\mathcal{K}}(r) = \frac{1}{\widehat{\lambda} N} \sum_{i}^{N} \sum_{j \neq i} w_{ij}^{-1} (d_{ij} \leq r) \tag{6}$$

$$d_{ij} = \begin{cases} 1 & \text{if } d_{ij} \leq r, \forall i \neq j \\ \\ 0 & otherwise \end{cases}$$

where $w_{ij} = \frac{a_{ij}}{2\pi d_{ij}}$ is the edge correction, a_{ij} is the length of the arc of the circle defined by radius r within the region S and $d_{ij} = \|x_i - x_j\|, \forall i \neq j \in S$ [9].

2.8 The $\mathcal{J} - function$

The $\mathcal{J} - function$ is a combination of the $\mathcal{G} - function$ and the $\mathcal{F} - function$ defined as

$$\mathcal{J}(r) = \frac{1 - \mathcal{G}(r)}{1 - \mathcal{F}(r)} \tag{7}$$

[2]. For CSR, $\mathcal{G}(r) = \mathcal{F}(r)$; therefore, $\mathcal{J}(r) \equiv 1$. Values of $\mathcal{J}(r) < 1$ suggest clustering, while $\mathcal{J}(r) > 1$ suggest regularity.

3 CASE STUDY

To illustrate the spatial-temporal analysis of spatial point patterns through time, the earthquakes in a given region of the world from 1974 to September 2015 will be considered. In this chapter the data set will be described first; separate analyses from the point of view of space and time will be done. Later, the joint spatial-temporal analysis will be applied.

3.1 Data Set

Our data is obtained from the US Geological Survey website [12]. The region of interest is from latitude $-18°$ and $0°$ and longitude $-83°$ and $-68°$. This region includes Peru, Ecuador, some small parts of Brazil, Bolivia and Chile. We are considering only earthquakes magnitude 5 and above. Earthquakes with magnitudes less than 5 rarely cause significant damage and can be difficult to locate especially when it occurs outside the United States of America [12]. The location of the earthquakes is in terms of longitude and latitude. For each earthquake, the magnitude, depth and date and time is considered as well. There are a total of 1359 earthquakes of magnitude 5 or more since January 1974 to September 2015 in this region within those lines of longitude and latitude where the earthquakes actually happen. There are some sections in the Pacific Ocean within those lines of longitude and latitude where no earthquake of magnitude 5 or more happen. Thus, we defined our region as shown in the Figure 2. Figure 2 (left) shows the region in red lines and the maps of the countries within the region in black lines. On the right of Figure 2 is the location of the earthquakes within the region.

Our data is originally in latitude-longitude format but to avoid negatives or east-west designation we converted the data to the Universal Transverse Mercator (UTM) coordinate system for some of our estimations. UTM system makes calculations easier since distance between points is in metric just like the Cartesian coordinates unlike the latitude-longitude system that is in angular format. We want to stress that this change however, does not change the location of the earthquakes as we can see in Figure 3 below. Just a note, although you need to make no changes, when plotting these points in 2D, the software is making a conversion already from 3D to 2D to plot.

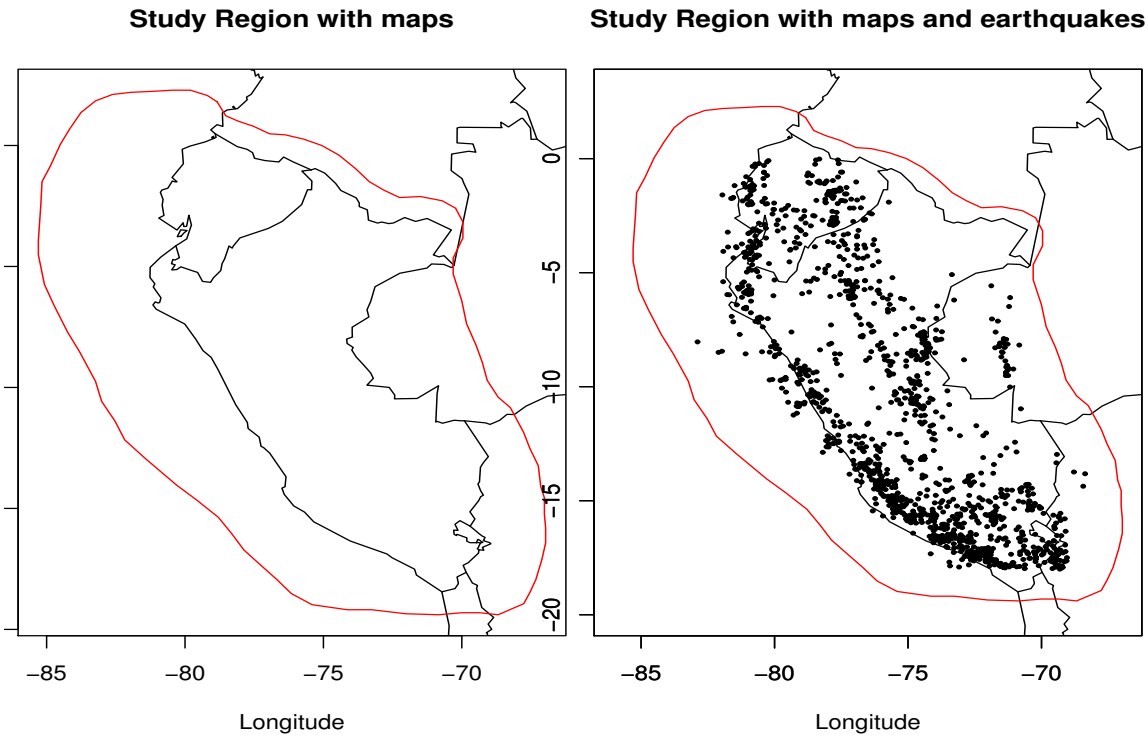

Figure 2: Study region

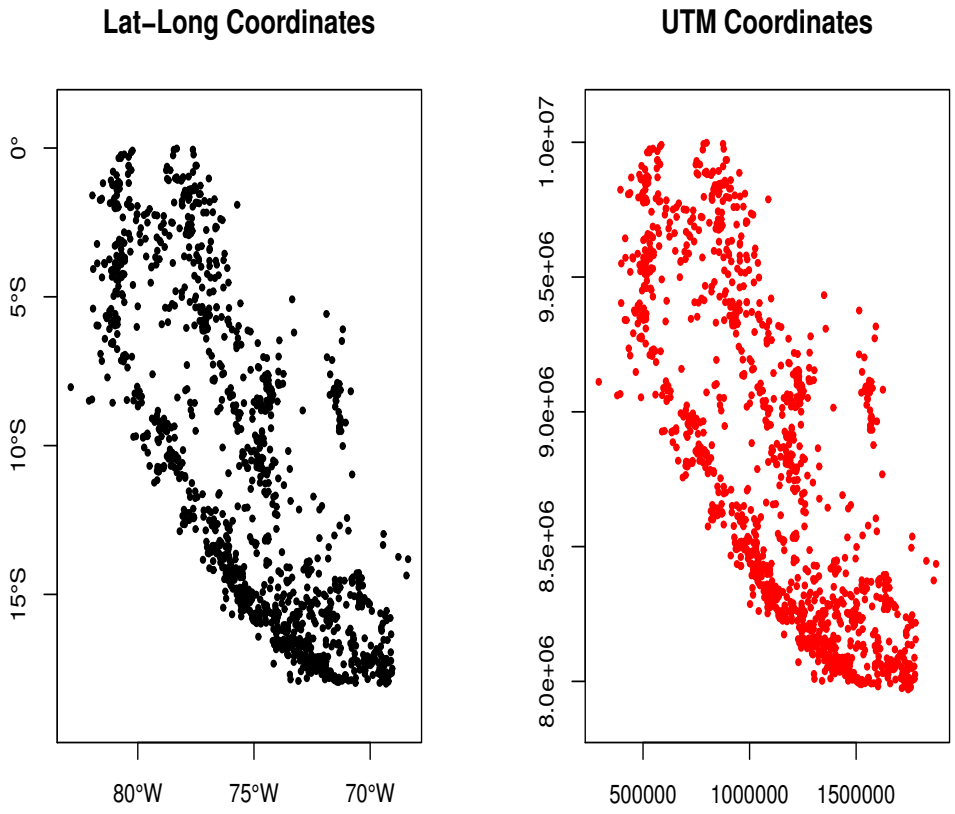

Figure 3: UTM and Lat-Long compared

3.2 Spatial Analysis

In terms of space, we want to analyze the distribution of events in the region (first order properties) and the level of interaction between events (second order properties).

3.2.1 First-Order Properties

We begin our spatial analysis by considering the distribution of the events in the region using quadrat count. The quadrat count in Table 1 shows that the intensity of events is not constant. We see from the table that most of the events (326 events) happen within the area bounded by longitude ($-75°$, $-71°$) and latitude ($-20°$, $-14.75°$), and the least is zero in three different regions. All other regions experience varying intensities between these two extremes . The distribution of points and their

Table 1: Quadrat Count

Lat \ Long	$[-83°,-79°]$	$(-79°, -75°]$	$(-75°, -71°]$	$(-71°, -67°]$
$(-4.25°, 1°]$	114	118	0	0
$(-9.5°, -4.25°]$	88	127	92	1
$(-14.75°, -9.5°]$	20	197	109	19
$[-20°, -14.75°]$	0	64	302	108

count is shown in Figure 4.

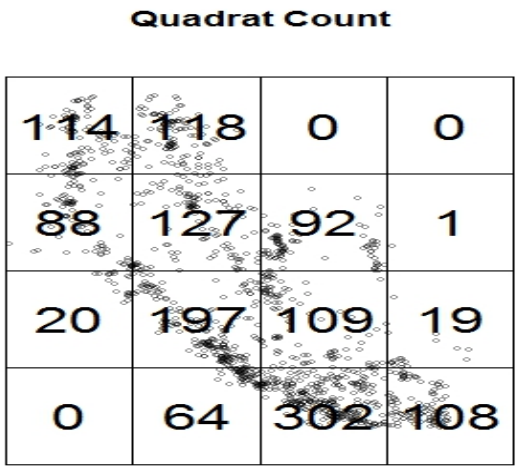

Figure 4: Quadrat Count

Figure 5 shows the plot of the kernel intensity estimates. The blue areas indicate lower intensity while the yellow regions indicate highest intensity. Intensity increases from blue to yellow as seen in the color scale on the right. We observe that there are more events at the lower diagonal part of the region than the other parts. In addition, the intensities decreases as we move up diagonally. At the top right and bottom left, there is little or no intensity at all. Unequal intensities show that the events are not homogeneous.

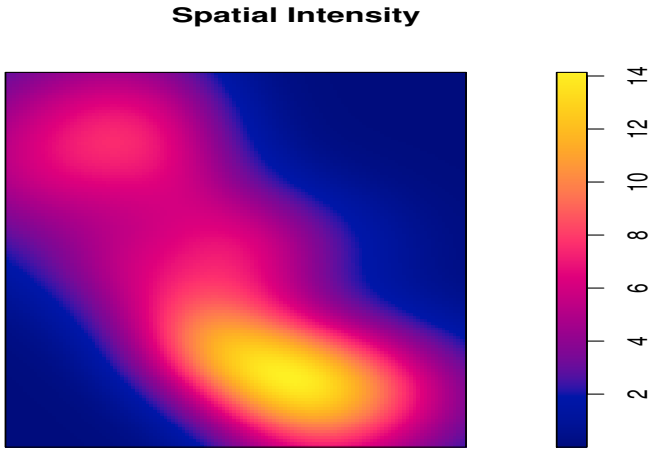

Figure 5: Spatial Intensity

3.2.2 Second-Order Properties

The estimated $\mathcal{F}, \mathcal{G}, \mathcal{J}$ and $\mathcal{K}$ functions is shown inin Figure 6. Four different plots are produced for each function. For each of the plots, we are only interested in the theoretical Poisson and the border corrected estimated.

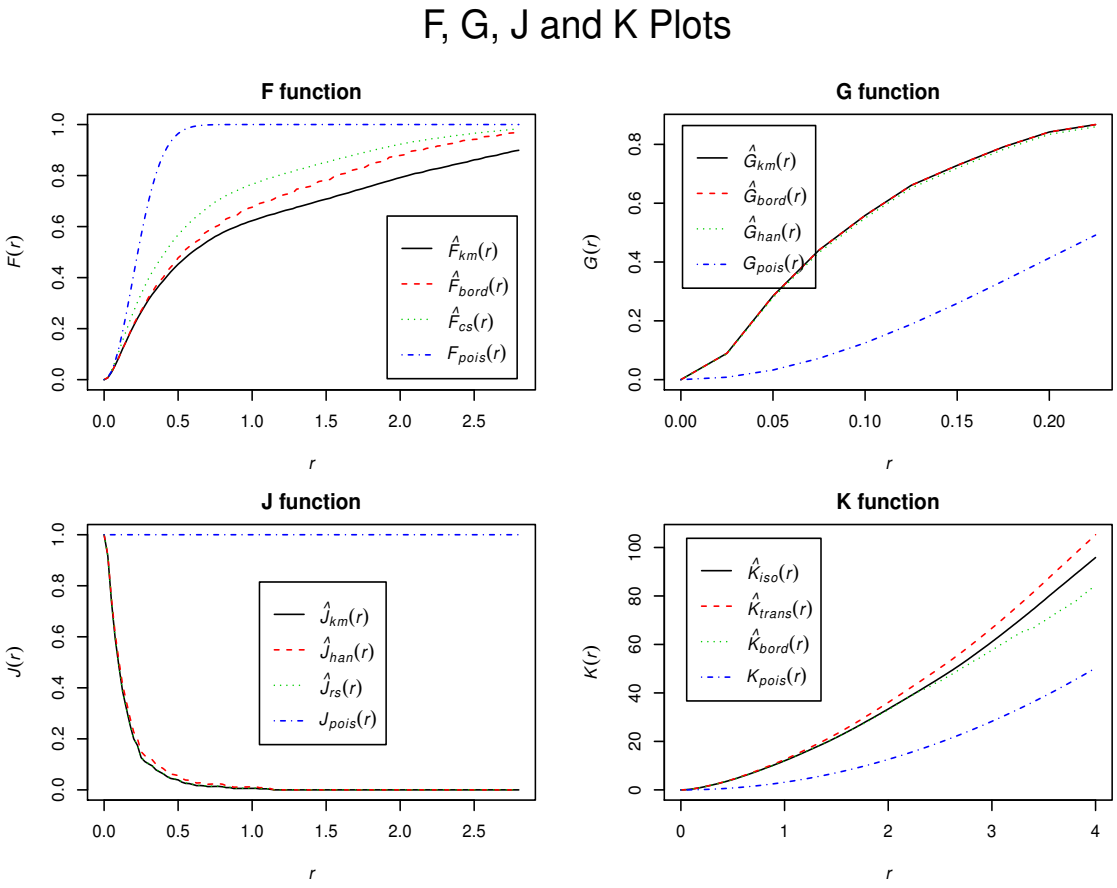

Figure 6: F ,G, J and K functions plots

For the $\mathcal{F} - function$ we have the estimates in Table 2 below:

Table 2: $\mathcal{F}$ Estimates

key	meaning
km	Kaplan-Meier estimate
rs	border corrected estimate
cs	Chui-Stoyan estimate
$theo$	theoretical Poission estimate

As seen in Figure 6, the border corrected estimated is below the theoretical Poisson estimate, indicating that the earthquakes are clustered. This is because the observed points are further away from an arbitrary point x_i for the clustered process than in CSR.

Similarly, for the $\mathcal{G} - function$, the estimates are given in Table 3.

Table 3: $\mathcal{G}$ Estimates

key	meaning
km	Kaplan-Meier estimate
rs	border corrected estimate
han	Hanisch estimate
$theo$	theoretical Poission estimate

Again, as seen in Figure 6, the border corrected estimated is above the theoretical Poisson estimate, indicating that the earthquakes are clustered. This is because the observed points are closer to each other for the clustered process than the CSR. The $\mathcal{J} - function$ produce similar estimates. From Figure 6, the estimated border corrected estimate is below 1, which indicates that the earthquakes are clustered.

Finally, the $\mathcal{K} - function$ estimates are given in Table 4.

Table 4: $\mathcal{K}$ Estimates

key	meaning
trans	translation-corrected estimate
border	border corrected estimate
iso	Ripley isotropic correction estimate
theo	theoretical Poission estimate

The plot of the $\mathcal{K} - function$ in Figure 6 shows that the border corrected estimated is above the theoretical Poisson estimate for CSR, indicating that the earthquakes are clustered.

Therefore, from all the four estimates above, we see that the events do not follow complete spatial randomness. However, we cannot conclude at this stage that the events are clustered, because even with completely random pattern it is hard to obtain the theoretical Poisson estimate, say K_{pois}, due to random variability. Thus, we need to construct envelopes that will give the bounds of the estimated functions under CSR. If the estimated functions falls outside the bounds, we can conclude the pattern is not CSR. Here, we construct the envelop for only the $K - function$ and a transformation of it which is the $L - function$, given by

$$L(r) = \sqrt{\frac{K(r)}{\pi}} \tag{8}$$

[2]. Transforming the estimator with the square root approximately stabilises the variance of the estimator. Thus, making it easier to access any deviation. We see for both functions in Figure 7 that the estimates lie very far away from the confidence bounds and look similar as the initial estimates. Therefore, we conclude that the events are indeed clustered. The estimates for both functions are the same and are

given in Table 5.

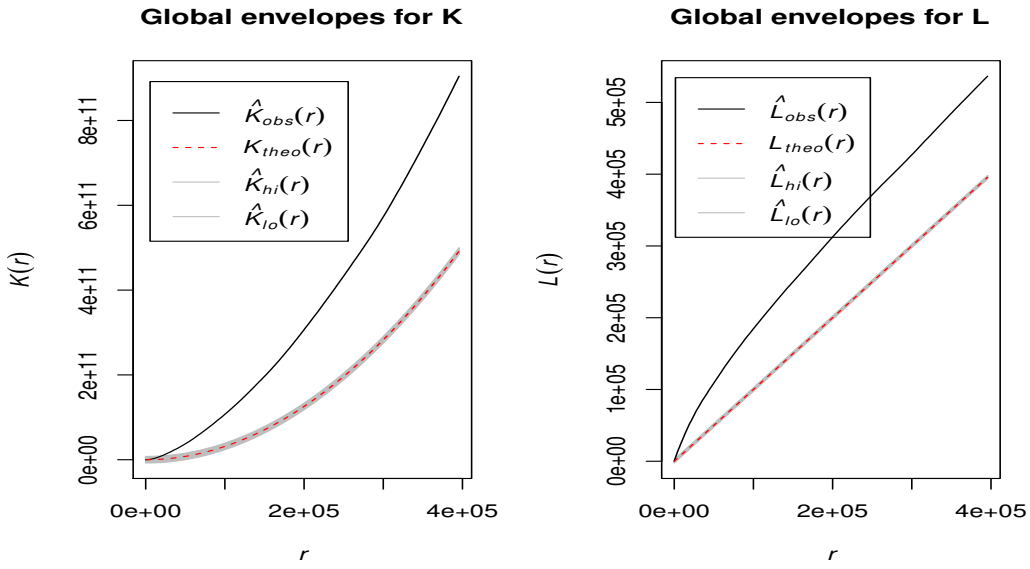

Figure 7: Envelopes plot of K and L functions

Table 5: $\mathcal{K}$ and $\mathcal{L}$ Estimates

key	meaning
obs	observed values estimate
theo	theoretical values for CSR estimate
hi	upper pointwise envelope from simulations estimate
lo	lower pointwise envelope from simulations estimate

Clearly the estimated K and L functions for the observed values lie outside the range of highest and lowest values of the K and L under complete spatial randomness. Therefore, we conclude that the events are clustered. This is not surprising because earthquakes happen along fault planes and high magnitude earthquakes tend to have a lot of aftershocks. Hence, these events happen only around the faults in the region

and very close to each other which brings about clustering.

We are especially interested in how clustered the points are. So we plot the histogram of the nearest neighbor distances. Figure 8 (left) below is the histogram of the distribution of the distance to the nearest neighbors. The histogram shows that majority of the earthquakes occur within a distance of 20,000m. In all the nearest neighbors are not more than 70km away which explains why the $G - function$ in Figure 8 (right) increases high at shorter distances above the theoretical Poisson.

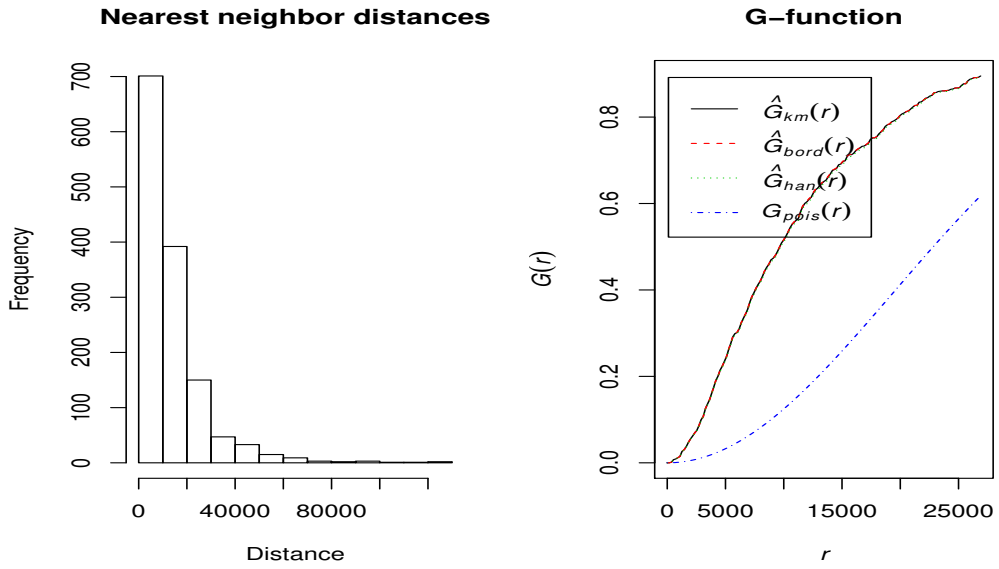

Figure 8: Histogram of distance to nearest neighbor and $G-function$ for earthquakes

A possible explanation of this could be due to aftershocks. Aftershocks are earthquakes that occur after the main earthquakes usually within a day or two after the main earthquake. They are usually smaller than the main earthquake and occur in the same area. If an aftershock is larger than a main earthquake, it is recorded as a main earthquake and the previous one recorded as the foreshock.

3.3 Analysis With Respect to Time

With respect to time, we are interested in the behaviour of the frequency of events in terms of:

- The frequency of events per year.
- The frequency of events per year by months.
- The frequency of events in a given month for all years.

3.3.1 Time Series of the Frequency of Events Per Year and Per Year by Months

The time series of the frequency of earthquakes per year by months is shown in Figure 9 (top) and the time series of the frequency of earthquakes per year year is shown in Figure 9 (bottom).

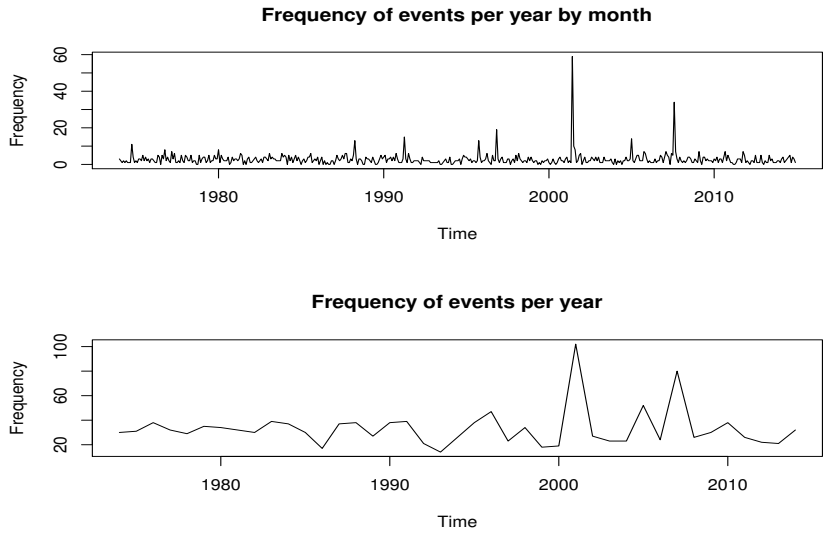

Figure 9: Time series plots for frequency events from 1974 - 2014 per year by months (top) and frequency of events from 1974 - 2014 per year(bottom).

The time series plots of the frequency of events of magnitude 5 or more per year by months shows that the frequency of events appear to be stationary with two potential outliers between 2000 and 2010. The same behavior is observed for the time series plot of the frequency of earthquakes per year. However, the variability of the frequency of events per year seems to slightly increase as time passes. A summary of the behavior of the two time series is shown in the box plots in Figure 10.

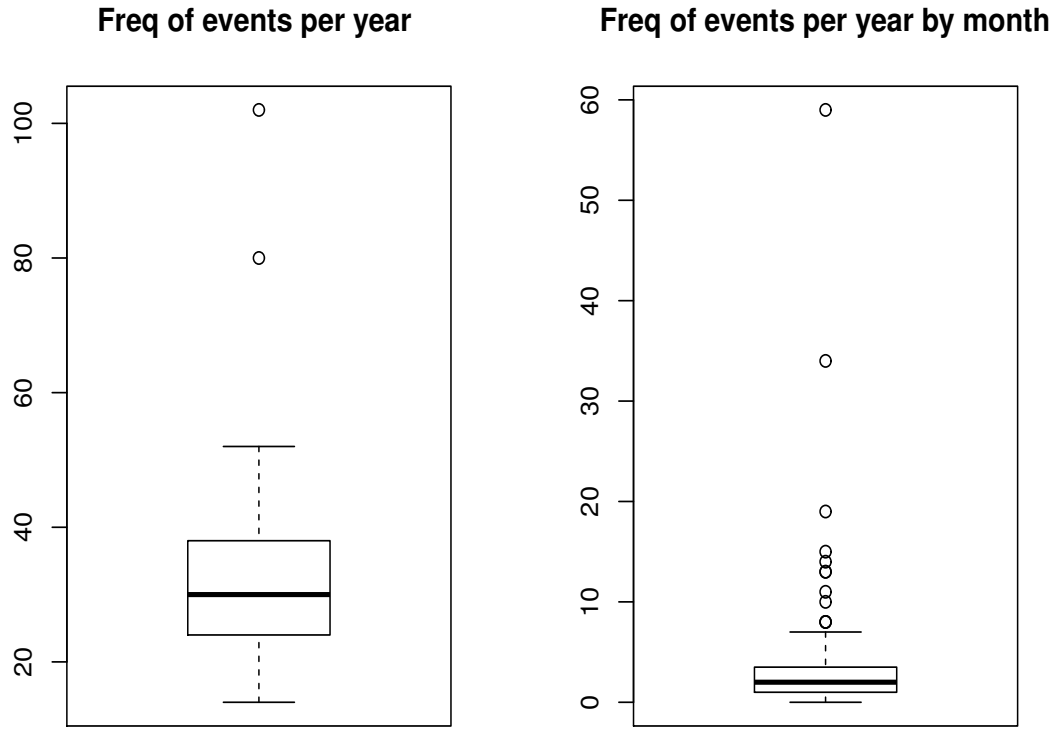

Figure 10: Box plots for frequency events from 1974 - 2014 per year (left) and frequency of events from 1974 - 2014 per year by months (right).

Figure 10 (left) indicates that the frequency of events per year is right skewed

with two outliers: 2001 with 102 events, and 2007 with 80 events. For the frequency of events per year by months (right), the distribution is also skewed to the right with nine outliers and a maximum of 59 events in June, 2001. This is an indication that some months are more prone to earthquakes in a year than others. This led us to consider the frequency of events in a given month for all years. Over all, the lower quartile, median, and upper quartile of the frequency of events per year is higher than that of the frequency of events per year by months.

Next, we want to identify the model for each of the time series in Figure 9. The plotted autocorrelation (ACF) and partial autocorrelation (PACF) functions for each of the time series in Figure 11 reveals that the frequency of events per year by month looks like a first order moving average, MA(1). All the ACF and PACF decays after lag 1 which is typical of moving average of order 1. For the frequency of events per year, the ACF and PACF decays with a spike at lag 6 indicating that the time series was generated by white noise.

The follow up is to determine whether there is periodicity. The cumulative periodogram for the frequency of events per year by months in Figure 12 (top right) indicates that the time series for the frequency of earthquakes per year by months in Figure 9 appeared to be an MA(1) while the cumulative periodogram for the frequency of events per year (bottom right) in Figure 12 indicates that the time series for the frequency of earthquakes per year in Figure 9 seems to be generated by a white noise process.

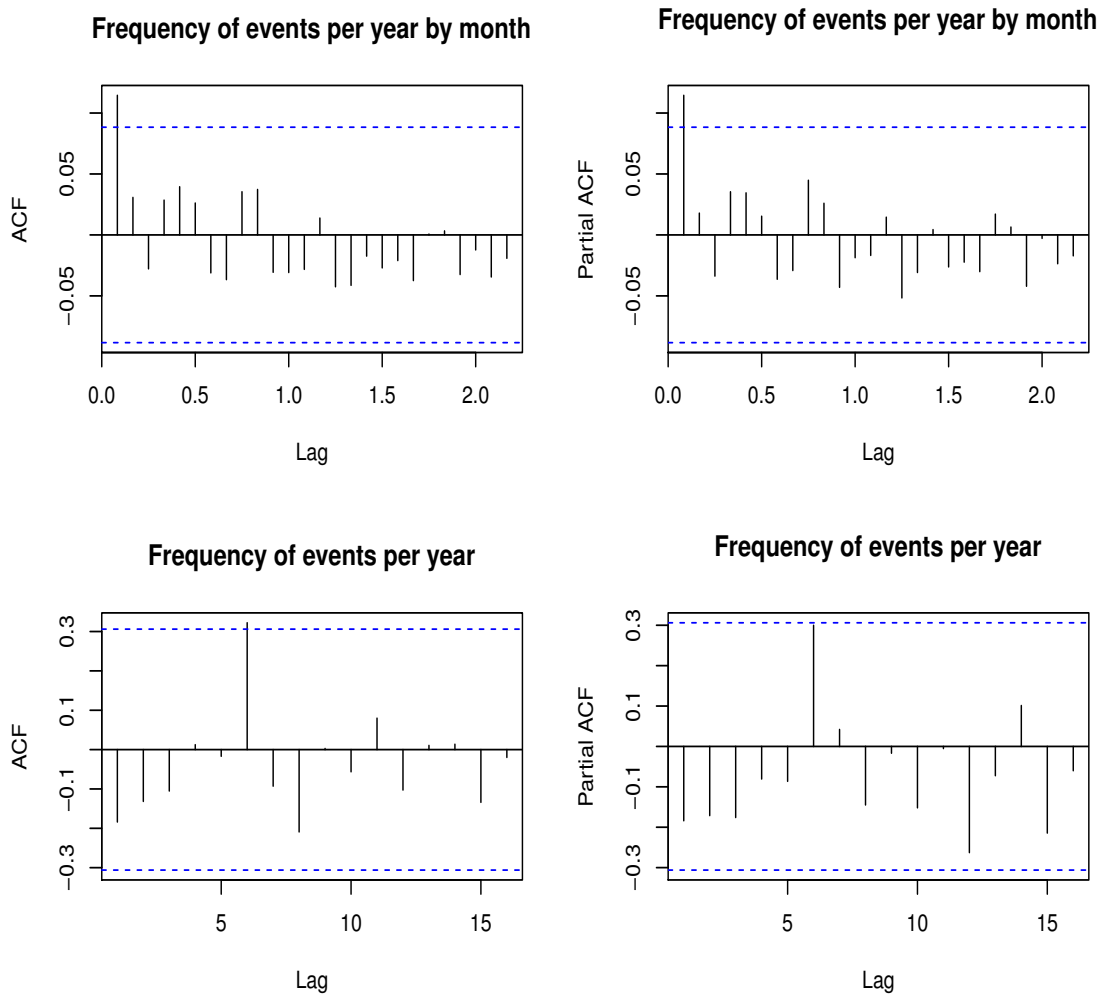

Figure 11: Autocorrelation and Partial-autocorrelation for earthquakes from 1974 - 2014

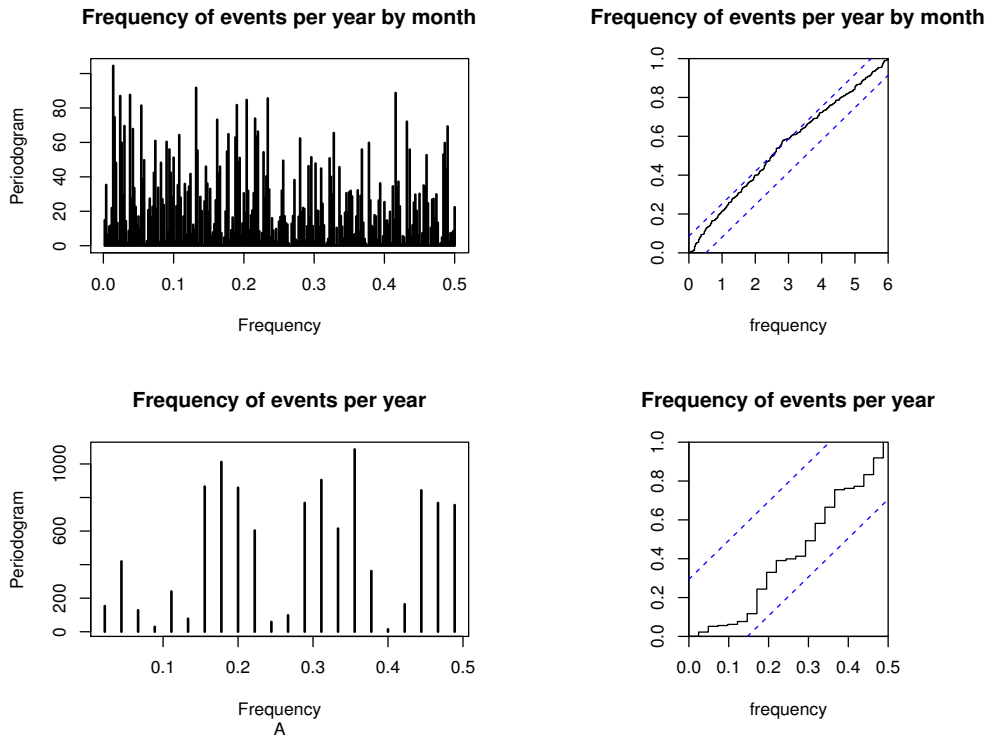

Figure 12: Periodogram and Cumulative Periodogram

There appears to be non-constant variability in the time series of the frequency of events per year, but this has no effect on the behavior of the time series in Figure 9 as seen in Figure 13. Comparing the histogram of the frequency of events and the natural logarithm of the frequency of events in Figure 13 we see that the two distributions look similar. Thus, the logs transformation has no significant effect on the shape of the distribution.

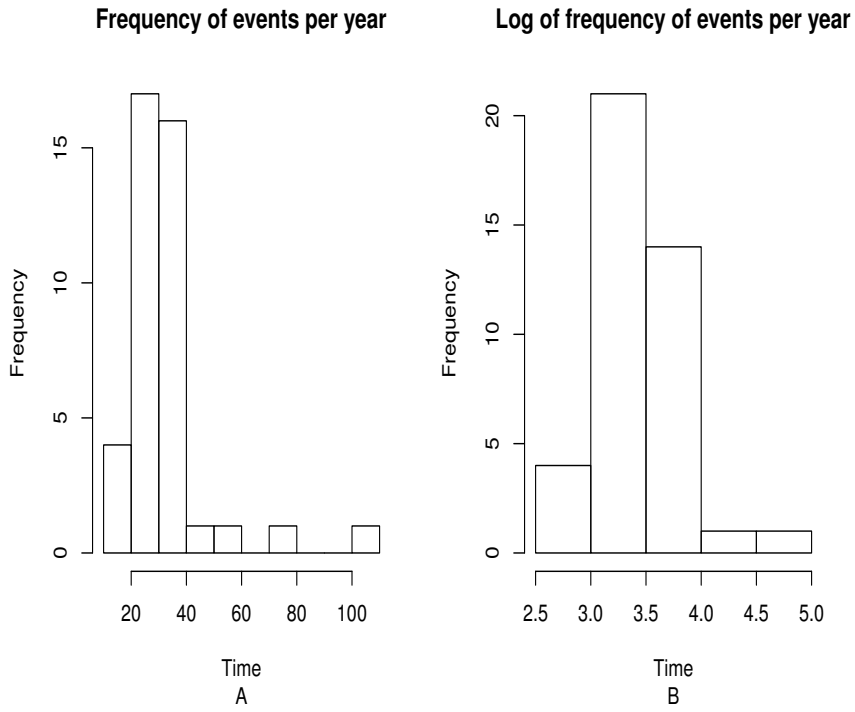

Figure 13: Histograms of frequency of earthquakes (A) and log of frequency of earthquakes (B)

The augmented Dickey-Fuller test applied to test for random walk against a stationary alternative for the frequency of events per year has $p - value = 0.02802$ favoring the alternative hypothesis of stationary process. Similarly, for the frequency of events per year by month, the $p - value = 0.01$ also favoring the alternative hypothesis of stationary process. Due to the fact that the time series of the frequency of events per year and the frequency of events per year by month is close to be considered a white noise, no attempt was made of fitting ARIMA models.

3.3.2 Time Series Analysis of the Frequency of Events in a Given Month for All Years

As seen earlier, some months appear to be more prone to incidents of earthquakes than others. Here, we want to look at the behavior of the number of earthquakes in each month for every year from 1974 - 2014. The frequency of events for all months except for June and August ranges between 0 and 15 as seen in Figure 14. February, March, May, July, September and December are very similar with a maximum of eight events. January, April, October and November have one or two peaks at about 14 or 15 for some years, but all others are between 0 and 8.

With the case of June and August, we observe that the extreme numbers just happened once while in all other years they behave similar to the other months. The ACFs in Figure 15 and the PACFs in Figure 16 shows no trend or seasonality, thus, there is no indication that June and August are actually prone to having more earthquakes. The extreme cases are probably due to some unusual occurrence in the faults. In all, the frequency of events for all the months appear stationary with few outliers for some months.

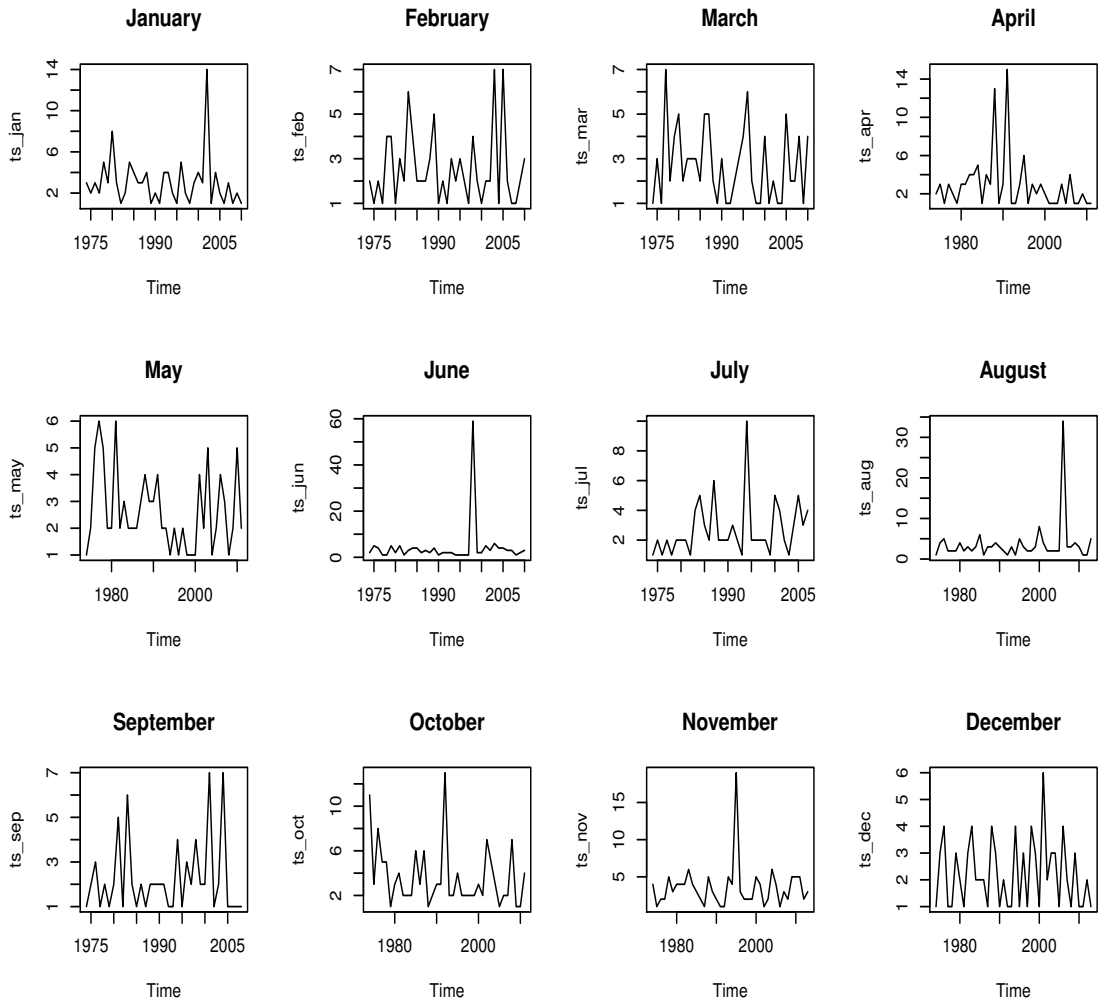

Figure 14: Monthly time series for the frequency of earthquakes from 1974 - 2014

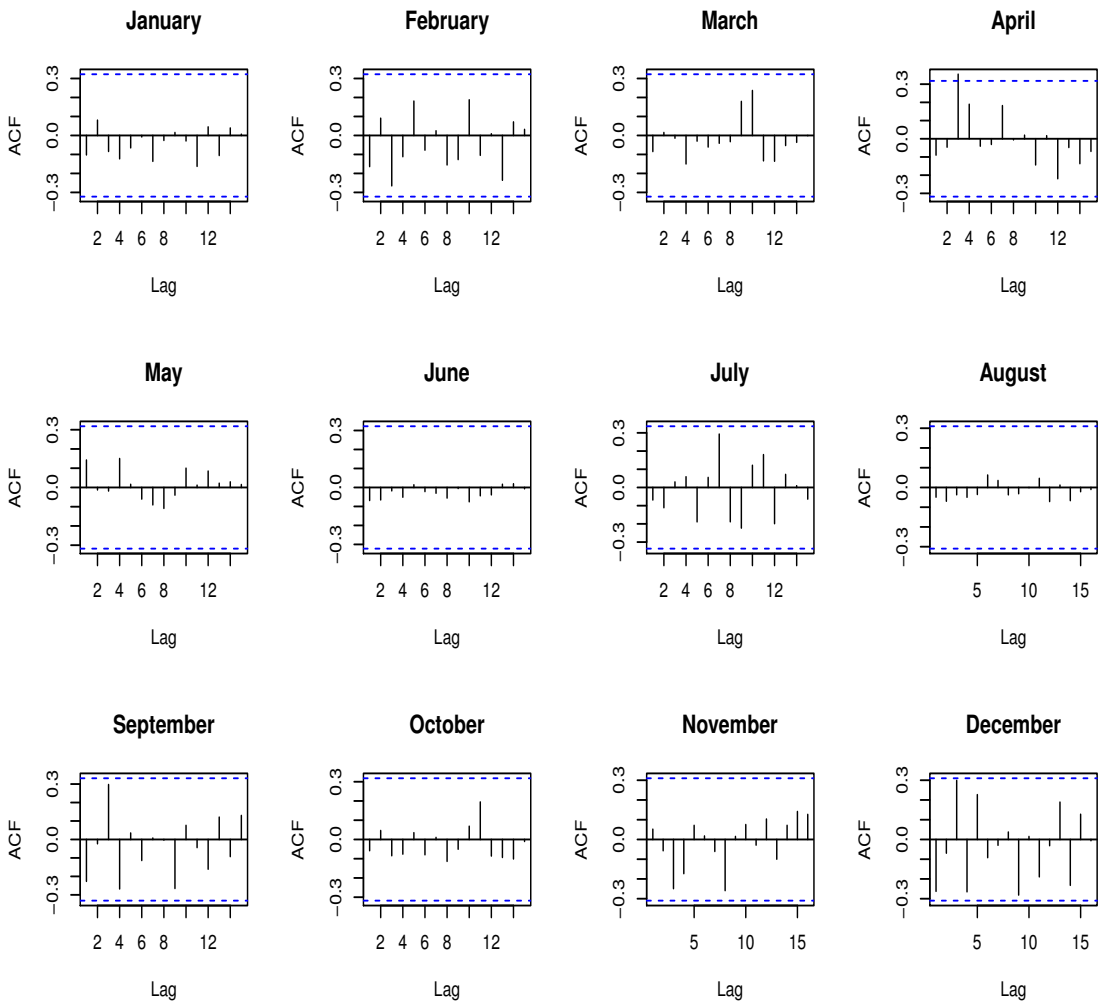

Figure 15: Autocorrelation function for monthly time series for the frequency of earthquakes from 1974 - 2014

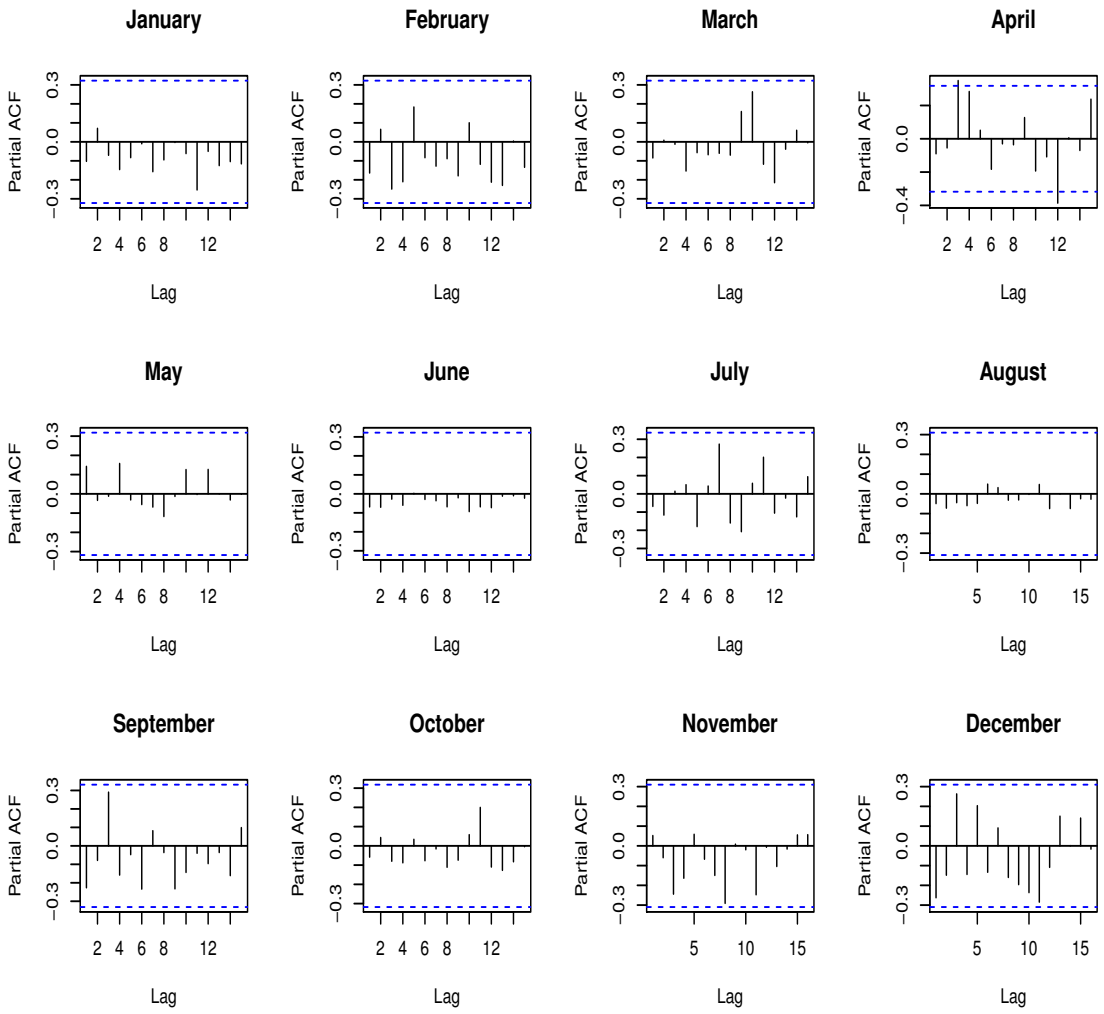

Figure 16: Monthly time series for the frequency of earthquakes from 1974 - 2014

Even though this work is focused on time and space, it is interesting also to look at the magnitude of the earthquakes. Figure 17 displays the maximum magnitude per year. We also look at the autocorrelation and partial autocorrelation functions of the maximum magnitudes of the earthquakes per year in Figure 18. The ACF and PACF reveal that the maximum magnitude per year also appears to be a white noise.

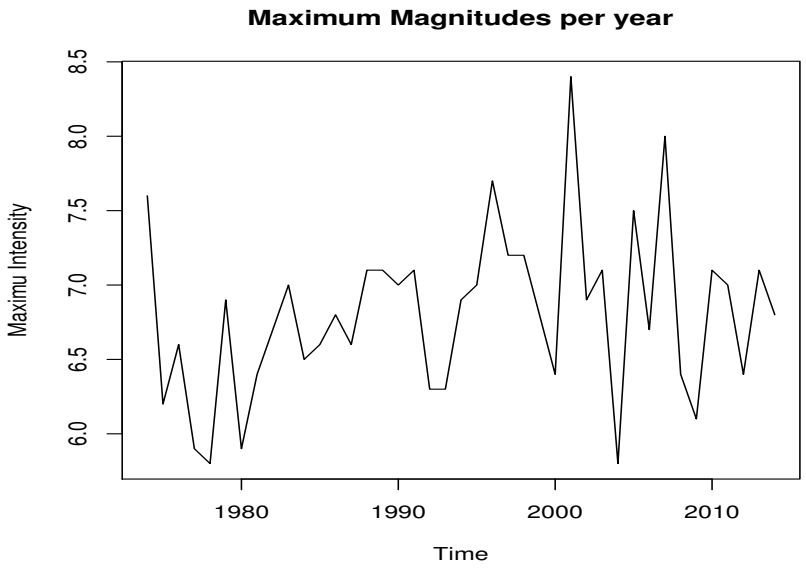

Figure 17: Maximum magnitude of earthquakes per year

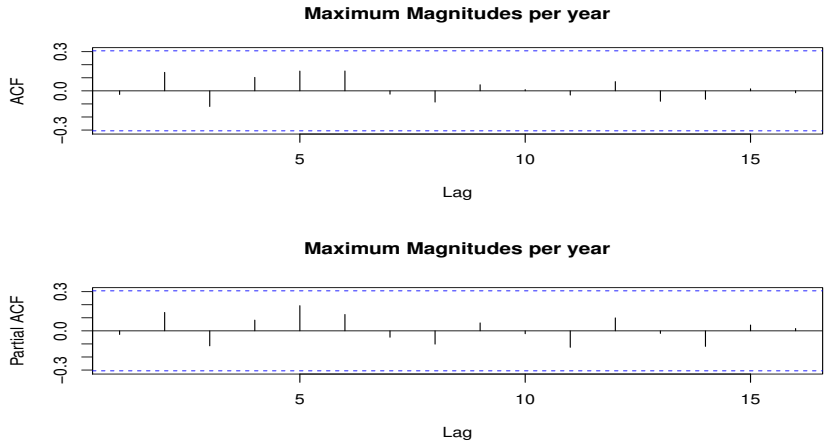

Figure 18: ACF and PACF of maximum magnitude of earthquakes per year

The histogram of the distribution of the maximum intensity of earthquakes in Figure 19A looks symmetric. To find out if the distribution is normal, we plot the normal quantile-quantile plot (qqplot) and also perform Shapiro-Wilk's normality test. From the qqplot in Figure 19B we see that the initial and final points deviate from the line but most of the points lie on the line or close to the line. Also, the $p - value = 0.2546$ for the Shapiro-Wilk's test at 5% significance level, hence we fail to reject the null hypothesis and conclude that the distribution of the intensity of earthquake magnitude 5 or more is normal.

```
        Shapiro-Wilk normality test

data:  ts_max

W = 0.96605, p-value = 0.2546
```

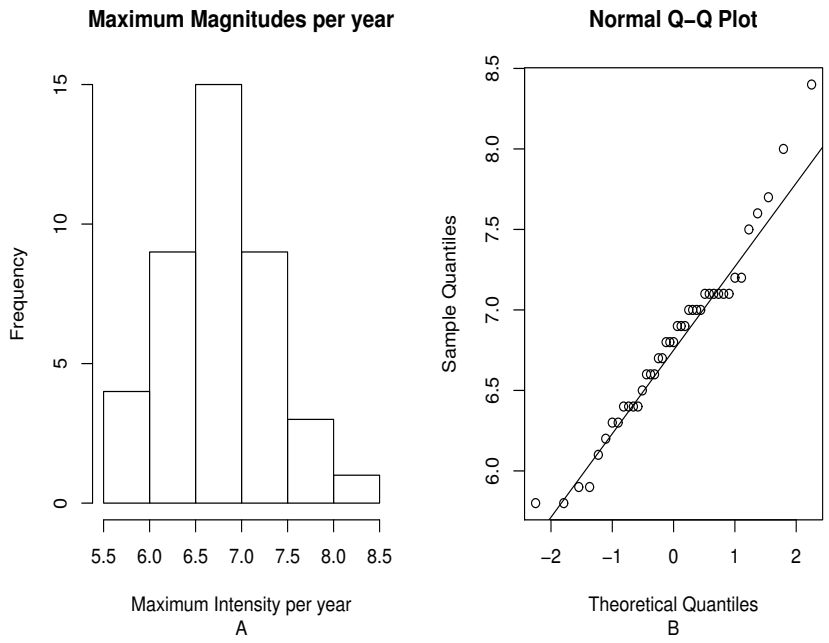

Figure 19: Histogram and normal qqplot of maximum magnitude of earthquakes per year

4 EXPLORING THE SPATIAL POINT PATTERNS SIMULTANEOUSLY WITH RESPECT TO SPACE AND TIME

Spatio-temporal point processes can be considered as a two-plus-one space-time distributions ($R^2 \times R^+$) that is, two dimensional for the space and one for time which is fundamentally different from either of the two space dimensions. In this sense, two-plus-one does not equal three, thus $R^2 \times R^+ \neq R^3$. Thus, for any event we consider the location say $\vec{x}_i$ and time of occurence say t_i, hence $\{(\vec{x}_i, t_i) : i = 1, 2..., n\}$ where $(\vec{x}_i, t_i) \in S \times T$ for some predefined spatial region S and temporal region T. Diggle [5] classified spatio-temporal point processes as either *continuous, spatially discrete or temporally discrete* [5].

- **Continuous** : The process is classified as continuous if an event can occur at any place and time. Here both location and time are continuous variables.

- **Spatially Discrete** : This is a process that can occur only at specific locations at any time. In this case, the location is a discrete variable but time is a continuous variable.

- **Temporally Discrete**: A temporally discrete process can happen anywhere but within specific times. Therefore, the location is a continuous variable while the time is a discrete variable.

- **First Order Separable** : A spatio-temporal point process is first-order separable if its intensity $\lambda(s, t)$ can be factorized as

$$\lambda(s, t) = m(s)\mu(t) \tag{9}$$

, for all $(s,t) \in S \times T$

- **Second Order Separable** : A stationary spatio-temporal point process is second-order separable if the covariance density,

$$\gamma(u,v) = \lambda_2(u,v) - \lambda^2 \tag{10}$$

can be factorized as

$$\gamma(u,v) = \gamma_s(u)\gamma_t(v) \tag{11}$$

. This does not mean independence of space and time; it is implied.

For our application, we are assuming that the earthquakes occur anywhere in an ordered sequence through time. The date and time of each event is recorded, thus, even with aftershocks, there is a time difference in hours, minutes or seconds. To distinctly represent each event with a unique time, we assign the event with an integer value of the date and time in seconds, calculated using the function *unclass in base package* in R [11]. The first event happened on 1974-01-05 at 08:33:50 GMT and the integer value for this time is 126606830. The next two events both happened on 1974-01-14 at 15:52:47 GMT and 17:35:17 GMT respectively. And their integer values are 127410767 and 127416917 respectively, which implies that even if we have events that are separated by a second, they will have different integer time values.

Initially, we were considering January 1, 1974 as the start date for our observations, so that the time of occurrence of an event is the number of days from the start date to the date of occurrence. The first event in 1974 was on January 5, 1974, thus we assigned a time of 4 to the event. However, problems arose with *aftershocks* happening within the same day. Thus, we had two or more events with the same

time. The maximum number of aftershocks in our data is 25, which occurred on June 24, 2001 between 11:13AM to 23:27PM, within latitude [-17.463, -16.888] and longitude [-72.409,-17.29]. Because these aftershocks happen very close to the main events, and on the same day, they look as one big event on the map. Hence, to distinguish these events is very difficult. But, we solved this problem easily when we assigned an integer value to each event. In a future study, it will be quite interesting to study dispersion and frequency of aftershocks within a day, but this study is not on aftershocks.

Figure 20 shows the scatter plot of the spatio-temporal data without marking points by time. The plot on the left is based on the locations, X is the longitude and Y is the latitude. The size of the points are all same, the dark areas are simply due to overlaps of points. This suggests that we have more that one earthquake occurring at particular places. Another reason is because of the aftershocks. The right plot is the cumulative plot of the time part of the spatio-temporal data.

In comparison, in Figure 21, the earthquakes are plotted through time using the plot function in *stpp* [8]. Here, points have different sizes and shades of color. The size and shade of the points depend on the time the earthquake occurred. Points representing recent events are larger and darker. The size shrinks and the color fades as time passes. Therefore, it is easier to distinguish between events in terms of time here as compared to the plot in Figure 20. Another way to visualize the data points is through animation plots using *stan and animation functions in stpp package* [8]

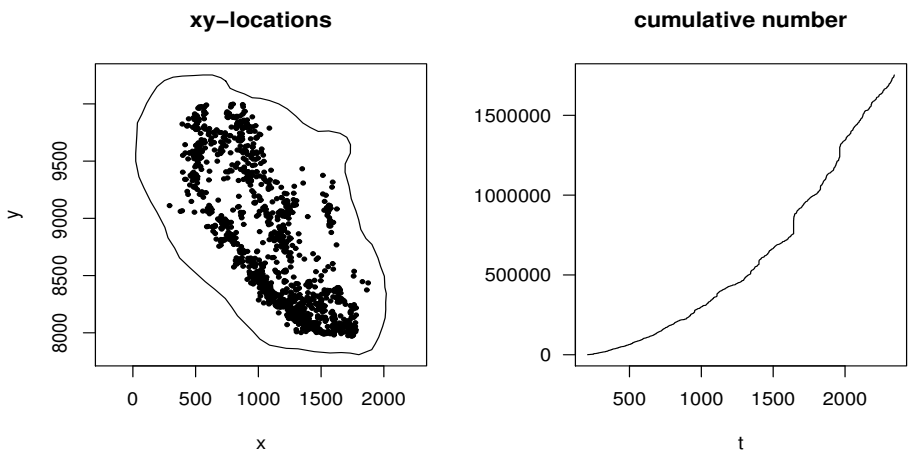

Figure 20: Scatter plot of earthquakes

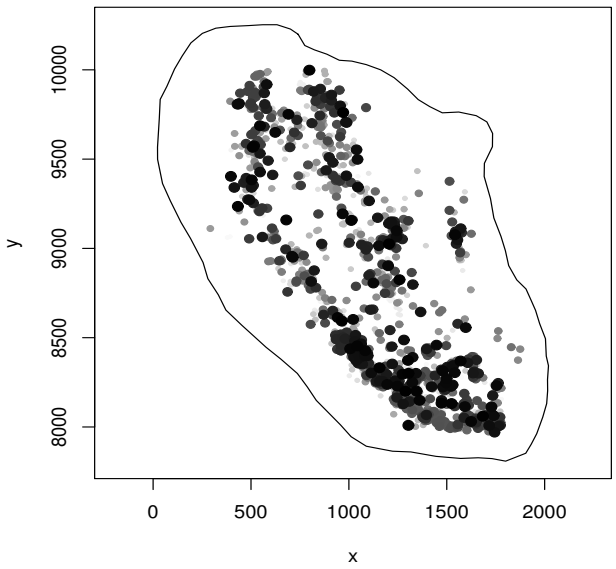

Figure 21: Plot of earthquakes through time

4.1 First-Order Property

For any spatio-temporal process, the expected number of events per unit area per unit time known as the *intensity*, characterize the first-order property. In other words, the density of the point patterns describe the first-order effects. In what follows from now, we denote the number of events and expected number of events in a given region S and within time T as $N(S \times T), E[N(S \times T)]$ respectively, area of S as $|S|$, time interval as $|T|$ and the first-order intensity as $\lambda_1(S, T)$. Thus, going by this, $\widehat{\lambda}_1 = \frac{n}{|S \times T|} = \frac{1359}{270 \times 14942.} = 0.000337$ earthquakes per kilometer squared per day and 0.1213 per kilometer squared per year. This is only true for a homogeneous spatio-temporal point process. In general, Diggle [5] defined the first order property as

$$\lambda(x, t) = \lim_{|dx|,|dt| \to 0} \left\{ \frac{E[N(dx, dt)]}{|dx||dt|} \right\} \tag{12}$$

[5].

Practically, the distinction between first-order and second order intensity is difficult without making some assumptions. We need the assumption of separability to distinguish between first-order and second-order effects. Thus, we assume that the first-order effects are separable. That is $\lambda(s, t) = m(s)\mu(t)$. We find a non-parametric Gaussian kernel estimate for the spatial intensity $m(s)$ with an appropriate bandwidth. There are several methods for choosing the bandwidth. We want to choose a bandwidth that minimizes the mean square error of the estimated spatial intensity. For the temporal intensity $\mu(t)$, we use a parametric log-linear model. Figure 22 shows the density plot of the temporal intensity. The temporal intensity estimate is based on the density of the time (in weeks) of occurrence of the earthquakes. The

plot shows that the temporal intensity has been slightly decreasing from late 1974 to mid 1975. After that, there has been a rise in intensity till mid 1976, and it became stationary afterwards.

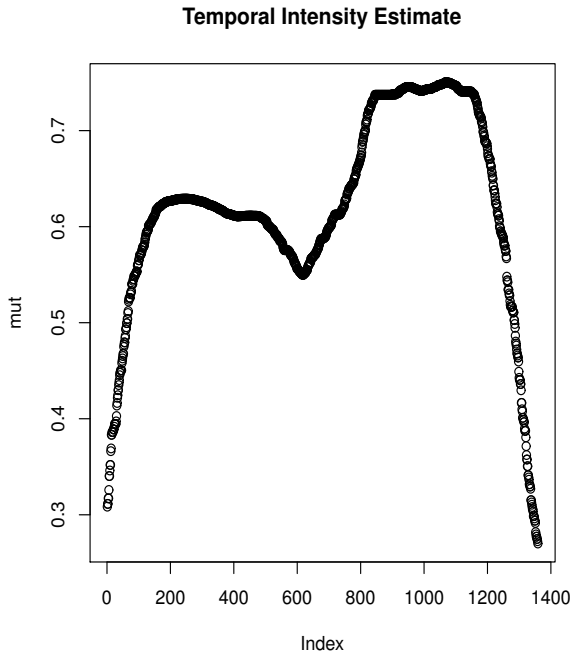

Figure 22: Estimated temporal intensity of the time of occurrence of events in weeks

In Figure 23 we have the 2-dimensional kernel estimate of the spatial intensity. The yellow regions in the kernel estimate show the intensity of the earthquakes. We see from the yellow strip running from the bottom to the top on the left of the region that, a lot of earthquakes happen near the coast of Peru with more clustering around the coast of Central and Southern Peru. We also see some clustering around the Peru-Brazil border.

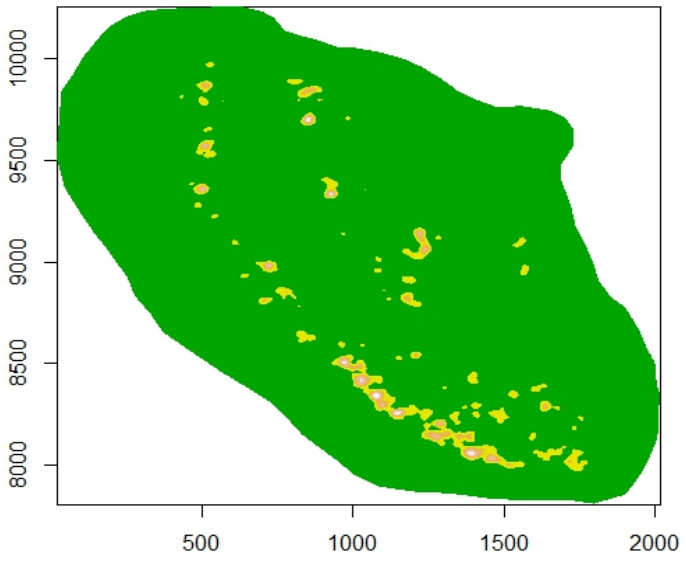

Figure 23: Kernel Estimate of Spatial Intensity

4.2 Second-Order Property

Here, we are interested in the pairwise correlation between two pairs of events in a sub-region. We now consider the joint spatio-temporal intensity function for any two locations. That is,

$$\lambda_2(\vec{x}, \vec{y}, s, t) = \lim_{|d\vec{x}|,|d\vec{y}|,|ds|,|dt|\to 0} \left\{ \frac{E[N(d\vec{x} \times d\vec{y})N(ds \times dt)]}{|d\vec{x}||d\vec{y}||ds||dt|} \right\} \tag{13}$$

where, $(\vec{x}, s)$ and $(\vec{y}, t)$ are any two locations within the region [5].

Now, we are considering only the $\mathcal{K} - function$ for the spatio-temporal point process. We define the homogeneous Poisson Process $\mathcal{K} - function$ for the joint

space and time for any radius u and time v by

$$\mathcal{K}_{ST}(u,v) = \pi u^2 v \tag{14}$$

. For any inhomogeneous Poisson process (STIK-function), $\mathcal{K}_{ST}(u,v) > \pi u^2 v$ indicates clustering while $\mathcal{K}_{ST}(u,v) < \pi u^2 v$ indicates regularity. Diggle [5] defined the second-order intensity reweighted stationary $\mathcal{K} - function$ as

$$\mathcal{K}_{ST}(u,v) = 2\int_0^u \int_0^v g(u',v')u'du'dv' \tag{15}$$

where $g(u',v') = \frac{\lambda(u,v)}{\lambda(s,t)\lambda(s',t')}, u = \|s-s'\|,$ and $v = \|t-t'\|$ [8]. The STIK-function is estimated by using a non-parametric estimator function called STIKhat, defined by:

$$\widehat{\mathcal{K}}_{ST}(u,v) = \frac{1}{S\times T}\frac{n}{n_v}\sum_{i=1}^{n_v}\sum_{j=1;j>i}^{n_v}\frac{1}{w_{ij}}\frac{1}{\lambda(s_i,t_i)\lambda(s_j,t_j)}\left\{\|s_i-s_j\| \le u; t_i - t_j \le v\right\} \tag{16}$$

where n_v is the number of events for which $t_i \le T_1 - v, T = [T_0, T_1]$ [8]. This function is estimated using an approximated unbiased estimator, based on the event location $\vec{x}_i, i = 1,2,...,n$ in the region $S\times T$ as:

$$\widehat{\mathcal{K}}_{ST}(u,v) = \frac{1}{S\times T}\frac{n}{n_v}\sum_{i=1}^{n_v}\sum_{j=1;j>i}^{n_v}\frac{1}{w_{ij}}\frac{1}{\lambda(\vec{x}_i)\lambda(\vec{x}_j)}1\left\{u_{ij}\le u\right\}1\{t_i - t_j \le v\} \tag{17}$$

where $\lambda(x_i)$ is the intensity of ordered events x_i such that $t_i < t_{1+1}$ at $x_i = (s_i, t_i)$. Temporal and spatial edge effects are accounted for by n_v and w_{ij}.

Figure 24 shows the plot of the nonparametric estimated $K - function$ in space and time up to a radius of 100,000m = 100km and time period of 35 weeks. Not much can be said about the plot until we compare the estimated values of $\widehat{\mathcal{K}}_{ST}(u,v)$ with the $\pi u^2 v$, that is the K-estimate under CSR. Hence, we plot $\widehat{\mathcal{K}}_{ST}(u,v) - \pi u^2 v$ as shown in Figure 25.

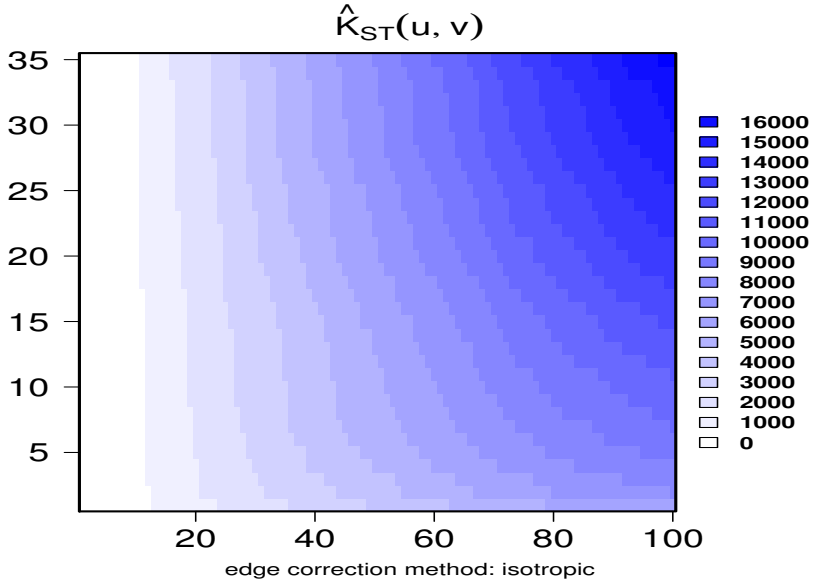

Figure 24: Inhomogeneous K-function estimate (Khat)

Figure 25 shows the contour plot of the estimated $\widehat{\mathcal{K}}_{ST}(u,v) - \pi u^2 v$ using different bandwidths. To tell the point in space and time at which we have spatio-temporal clustering, we compare the contour plot of $\widehat{\mathcal{K}}_{ST}(u,v) - \pi u^2 v$ using different bandwidth, h. On the extreme left, $h = 4$, the middle has a bandwidth that gives the minimum Mean Squared Error (mse) and on the extreme right $h = 20$. There is no rule of thumb for choosing h. Though the h based on the mse is usually preferred, for our case $h = 20$ gives a better estimate. The positive regions indicate spatio-temporal clustering while the negative regions indicate regularity.

Diggle [5] defined the Pair Correlation Function (PCF) as

$$g((s,t),(s',t')) = \frac{\lambda_2((s,t),(s',t,))}{\lambda(s,t)\lambda(s',t')} \tag{18}$$

Informally, this is interpreted as the standardized density probability that an event occurs in $ds \times dt$ and $ds' \times dt'$. For a Poisson process, $g((s,t),(s',t')) = 1$. Values of $g((s,t),(s',t')) < 1$ suggest inhibition between points while $g((s,t),(s',t')) > 1$

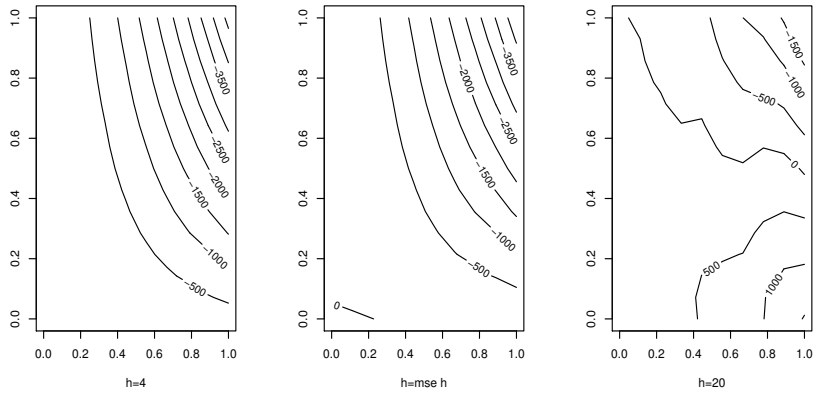

Figure 25: Khat

suggest clustering. We have contour plot of the PCF in Figure 26 and the perspective plot in Figure 27 for the 3-dimensional view. Figure 26 suggests clustering up to a distance of about 4km and a time of 15 weeks.

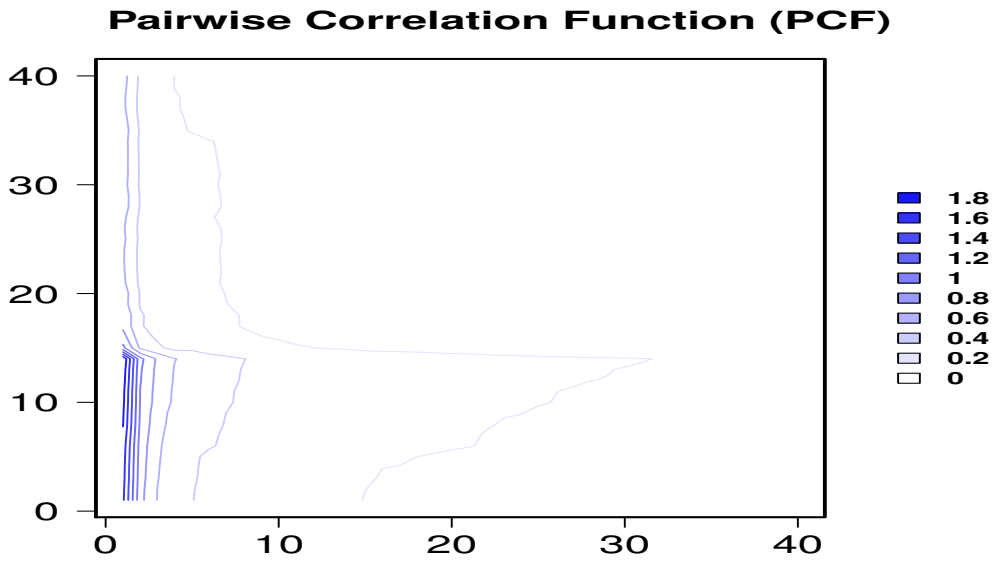

Figure 26: PCFhat

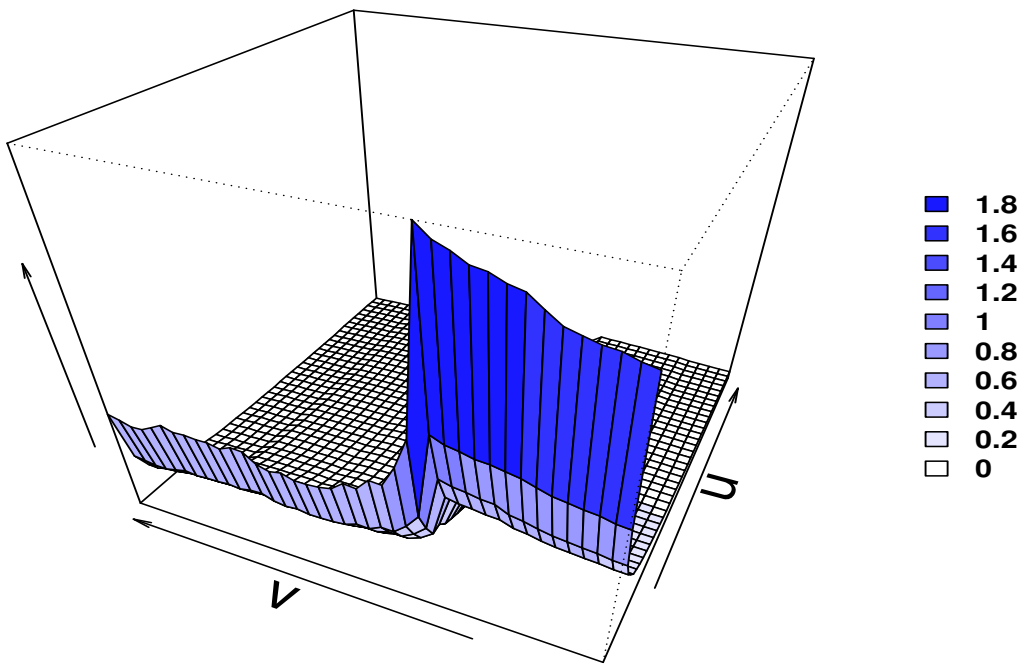

Figure 27: PCF Perspective Plot

5 FURTHER EXPLORATION OF EARTHQUAKES WITH RESPECT TO SPACE AND TIME

In the previous section we applied the methods found in the literature to the earthquakes example. In this section, we are proposing the use of some simple plots that would help us to further explore the joint analysis of time and space together in the case of the earthquakes.

5.1 Spatial Analysis of Events in Years and Months With Extreme Number of Events

We pose the question: Is the spatial distribution in a given year associated to the number of events happening that year? To answer that question, we plotted the location of the events for each year and sorted the years with respect to the number of events. We considered, two years with the least number of events, and two years with the highest number of events.

Figure 28 shows that for the two years with least events, the distribution of events appear random. In 1986 there were a total of 17 earthquakes which occurred all over the region. In 1993, the lowest for the 41 years from 1974 to 2014 was observed with a total of 14 events. These appeared even more sparse than was observed for 1986 with three more events. We observed that in both years the earthquakes tend to occur more around the central part of the region than the other areas and the events appear to follow a random process.

In contrast, Figure 29 shows two of the years with the most events with 2001

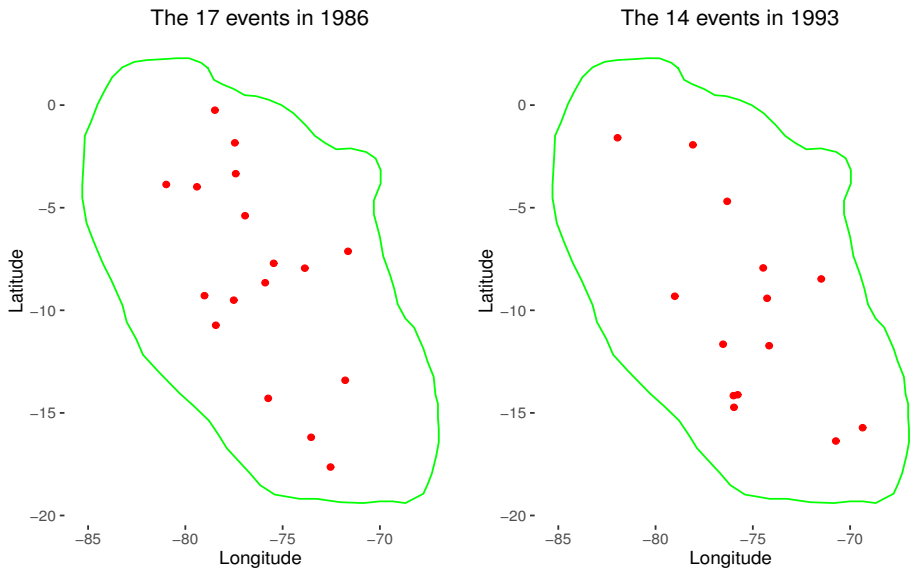

Figure 28: Years with the least frequency of events

having the most earthquakes for the 41 years, that is 102, and 2007 following with 80 events. In 2001, we observed that most of the events are clustered in the lower part of the region while the rest are sparse randomly within the region. In 2007, there appears to be more clustering in the lower part of the region and slight clustering in the upper part, but the events in the middle portion appears to be more random.

From the plot of those two years, the natural explanation is that for years with more earthquakes, there is a higher chance that most of the events will be clustered around a particular area probably due to the presence of aftershocks. A natural question that follows is whether the clustering occurs in specific months or not?

Figure 30 gives us more insight about what happened in the years with very high number of earthquakes. We see from the plots that the clustered events all happened within the same month. From this analysis, we conclude that at least in this region,

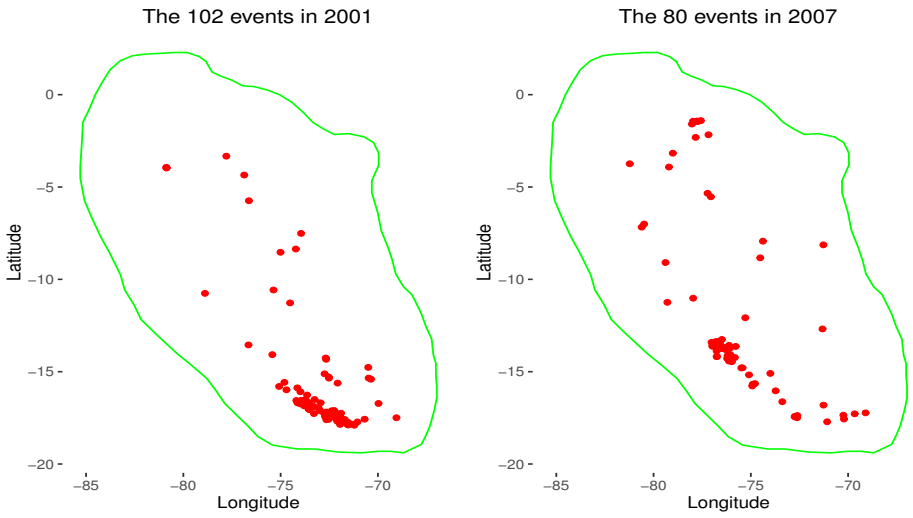

Figure 29: Years with the most frequency of events

when there is a low number of earthquakes, they tend to be randomly distributed. However, when the number of earthquakes in a year is high, they tend to be clustered both with respect to space and time because several of them tend to happen in a certain subregion and month.

5.2 Data Simulation With an Inhomogeneous Poisson Process

From all the previous analysis, it can be concluded that the earthquakes in this region follow an Inhomogeneous Poisson Process. The stpp package [8] offers the possibility of generating random processes. Therefore, we wanted to explore how the simulated data assuming an Inhomogeneous Poisson Process for the study region would look if data were simulated using the same intensity as that of the earthquake data.

The random inhomogeneous Poisson process generated using the inhomogeneous

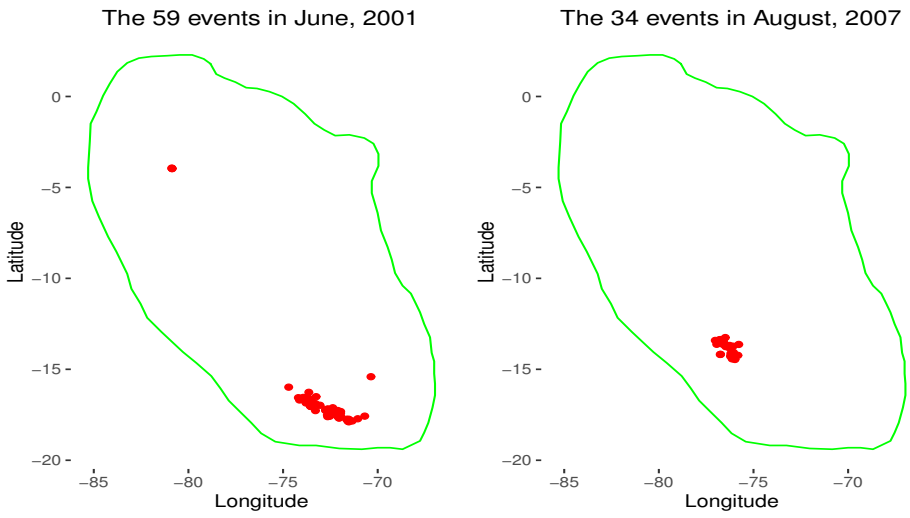

Figure 30: Months with the most frequency of events

intensity estimate shown in Figure 31 looks very similar to the original data. The generated data follows the density of the nonparametric intensity, thus areas with higher concentration turn to have more generated points than areas with low density. The two simulations below look very similar, but there are still some differences. The red circles show some of the different concentration of events happening within the same region in the two different simulations. By contrast if we assumed a homogeneous process as in Figure 32, the data would be scattered all over the region which is entirely different from the original data.

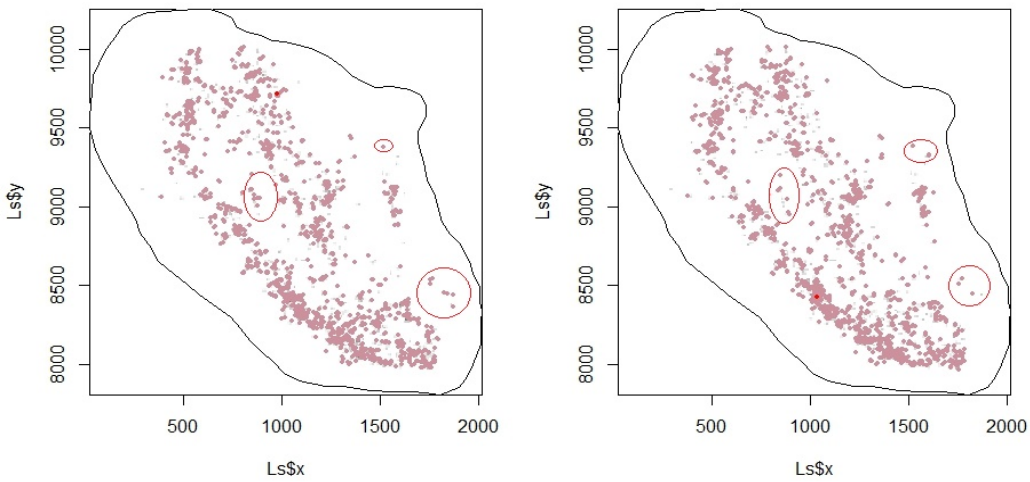

Figure 31: Random Inhomogeneous Poisson Process with estimated kernel. The red circles indicate different concentration within the same region from the two simulations.

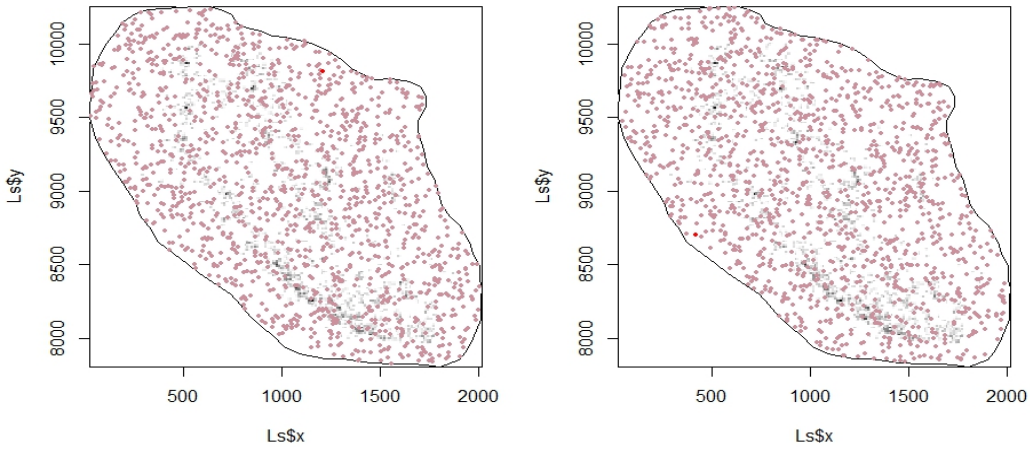

Figure 32: Random Homogeneous Poisson Process with estimated kernel

5.3 Analysis of Consecutive Events

We consider the case of discrete time and space so that we can calculate the probability of an event happening within a certain number of days and distance away from the previous event. The maximum distance and number of days between any two earthquakes is 1239km and 101 days respectively. The distances are calculated using *sp package* [1] in R, because the coordinates are in longitude and latitude and not in the Cartesian plane. Thus, we cannot use the Euclidean distances method. We divide the distance into equal intervals of 5 from 0 to 1240. Similarly, for the days we divide them into intervals of 10.1 from 0 to 101. We find the quadrat count in Table 5.3 and hence the joint and marginal probabilities in Table 5.3. The x and y in the tables represent the days and distances respectively.

The plot in Figure 33 below shows that most of the earthquakes happen within the first 20 days, the majority of which are within the first 10 days. This is not surprising because of the aftershocks.

For instance, from Table 5.3, we see that the likelihood of an earthquake occurring within 428km of the previous within the first 10 days is 27.3%. Again for all consecutive earthquakes, the likelihood of occurrence within the first 10 days is 62.5%, and the likelihood of occurrence within the first 428km is 36.38%.

Overall, the chance of a consecutive earthquake occurring decreases as the distance from the events increases and the number of days increases. Another probability of interest for us is to find out how far the consecutive earthquakes can occur given a fixed number of days. In Table 7, we estimate the conditional probabilities of consecutive earthquakes occurring within the five distance intervals for given time

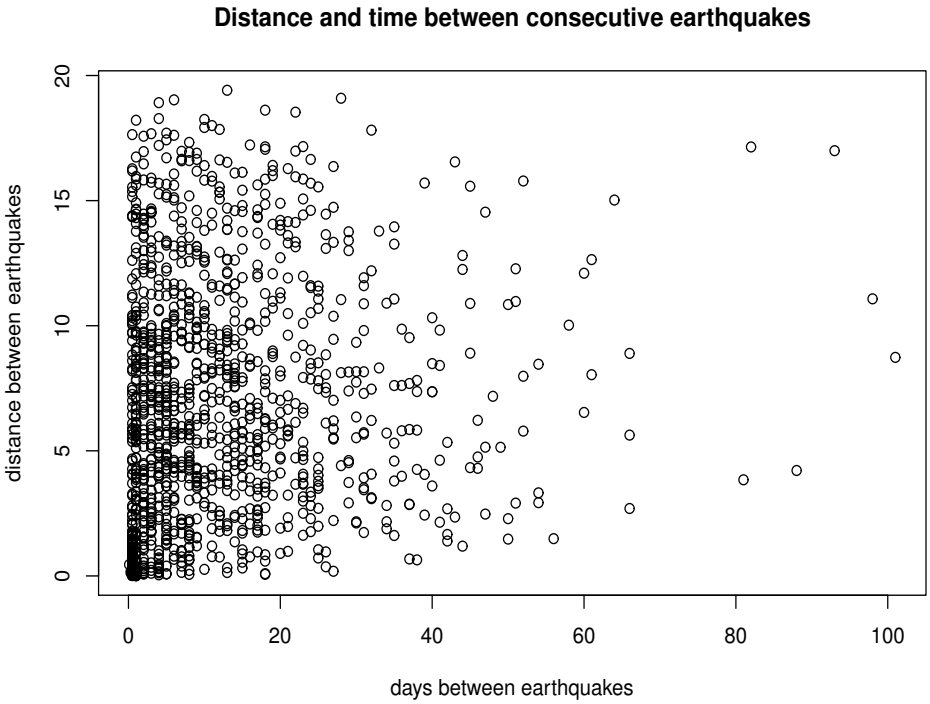

Figure 33: Distance and days between consecutive earthquakes

periods. For instance, within the first 10 days, there is a probability of 43.7% that another earthquake will occur from the previous within 428Km.

Distance and days between consecutive events

y \x	0–10.1	–20.2	–30.3	–40.4	–50.5	–60.6	–70.7	–80.8	–90.9	–101	Total
–1240	40	19	11	2	2	1	0	0	1	1	77
–1710	82	40	19	5	3	2	2	0	0	0	153
–1280	146	68	28	12	5	4	2	0	0	2	267
–856	210	86	36	22	9	2	1	0	1	0	367
0–428	371	61	31	16	9	4	1	0	1	0	494
Total	849	274	125	57	28	13	6	0	3	3	1358

Joint probabilities of distance and days between consecutive events

y \x	0–10.1	–20.2	–30.3	–40.4	–50.5	–60.6	–70.7	–80.8	–90.9	–101	Total
–124	0.030	0.014	0.008	0.002	0.002	0.001	0	0	0.001	0.001	0.057
–1710	0.06	0.030	0.014	0.004	0.002	0.002	0.002	0	0	0	0.113
–1280	0.108	0.050	0.021	0.009	0.004	0.003	0.002	0	0	0.002	0.20
–856	0.155	0.063	0.027	0.016	0.007	0.002	0.001	0	0.001	0	0.270
0–428	0.273	0.045	0.023	0.012	0.007	0.003	0.001	0	0.001	0	0.364
Total	0.625	0.202	0.092	0.042	0.0206	0.010	0.004	0	0.002	0.002	1

Conditional probabilities of distance given days between consecutive events

y \x	0–10.1	–20.2	–30.3	–40.4	–50.5	–60.6	–70.7	–80.8	–90.9	–101	Total
–1240	0.047	0.069	0.088	0.035	0.071	0.077	0	0	0.333	0.333	0.058
–1710	0.097	0.146	0.152	0.088	0.107	0.154	0.333	0	0	0	0.113
–1280	0.172	0.248	0.224	0.211	0.179	0.308	0.333	0	0	0.667	0.197
–856	0.247	0.314	0.288	0.386	0.321	0.154	0.167	0	0.333	0	0.270
0–428	0.437	0.223	0.248	0.281	0.321	0.308	0.167	0	0.333	0	0.364
Total	1.0000	1	1	1	1	1	1	0	1	1	1

5.4 Time Between Consecutive Events

We want to consider the distribution of the time until the occurrence of the next event. Intuitively, we expect this to follow an exponential distribution. However, to be sure for certainty, we would fit the distribution using *fitdistrplus* [4] package in R. The empirical density and cumulative distribution in Figure 34 suggests that the time between consecutive earthquakes follows an exponential or a gamma distribution as we can see in Cullen Frey plot in Figure 35.

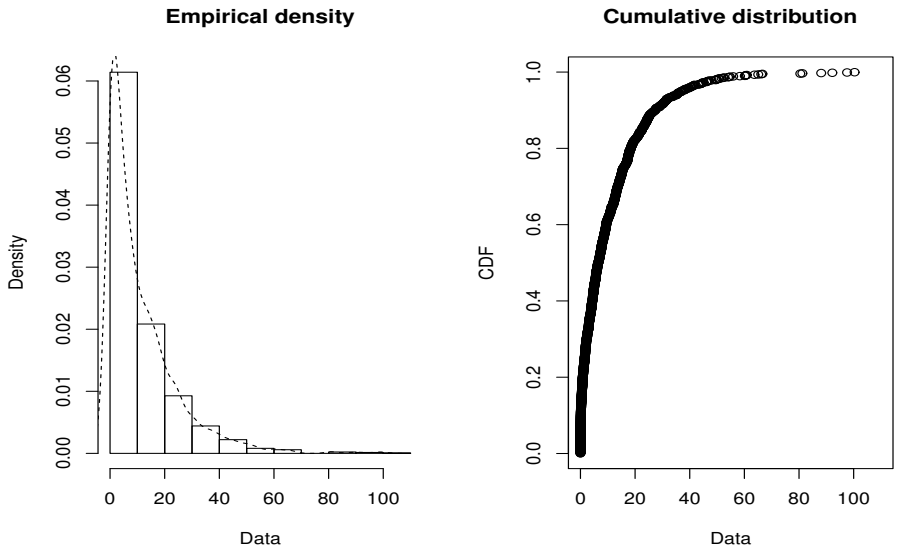

Figure 34: Empirical density and cumulative density distribution of time between consecutive events

The cumulative distributions(CDF), probability plot (P-P plot) and quantile plots (Q-Q plot) in Figure 36 are calculated based on the default Hazens rule. The CDF plot shows that the distribution follows an exponential or gamma distribution. Thus, for more insight we look at the Q-Q plot. The Q-Q plot shows lack of fit at the right

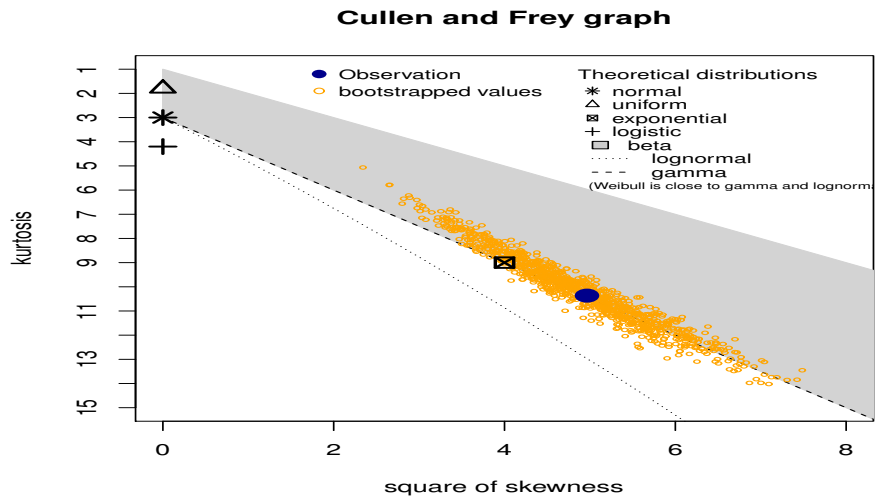

Figure 35: Cullen Frey plot

tail of the distribution while the P-P plot shows lack of fit at the center. In any case, the gamma is a better fit than the exponential. The fitted distribution is:

```
    Fitting of the distribution ' gamma ' by maximum likelihood
Parameters:
        estimate  Std. Error
shape 0.53480414 0.017045816
rate  0.04859697 0.002375889
```

5.5 Relationship Between Magnitude and Frequency of Events

Now we pose the question: Is the number of earthquakes in a year associated to the intensity of the strongest earthquake in the year?

The scatter plot of the frequency of events per year and the maximum magnitude per year reveals a positive relationship. We see from the plot that the higher the magnitude of the strongest earthquake, the higher the frequency and vice versa. However, the relationship does appear to be rather quadratic than linear. Below is

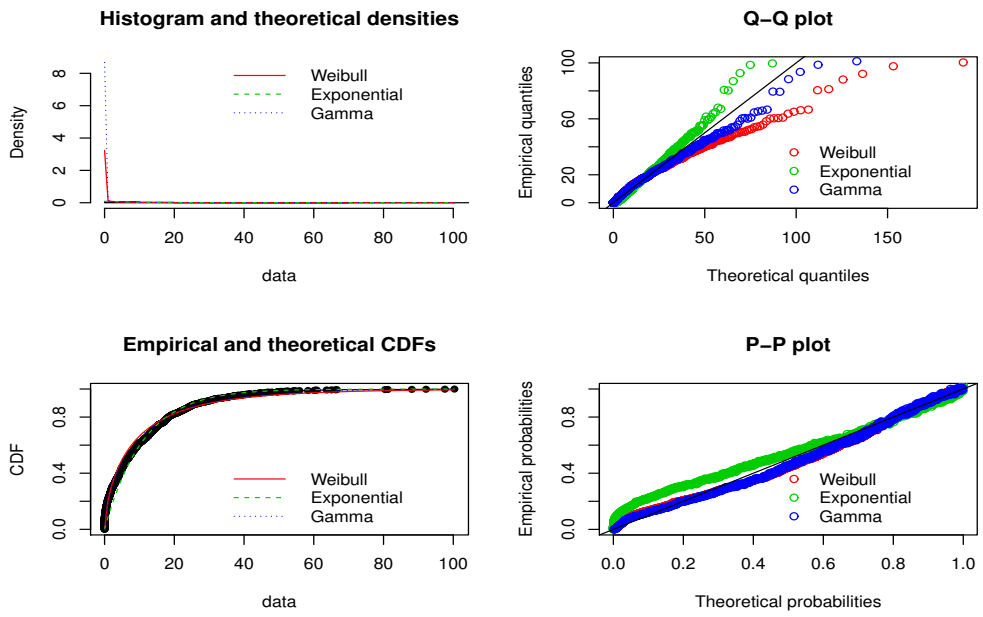

Figure 36: Comparison of Weibull, exponential, and gamma distribution

the summary of the quadratic model.

```
    Call:
lm(formula = num1 ~ mag1 + mag2)
Residuals:
    Min      1Q  Median      3Q     Max
-19.878  -5.837   1.315   6.163  12.708
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  815.777    127.780   6.384 1.69e-07 ***
mag1        -245.132     36.973  -6.630 7.82e-08 ***
mag2          18.994      2.666   7.125 1.67e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 8.079 on 38 degrees of freedom
Multiple R-squared:  0.7486,    Adjusted R-squared:  0.7354
F-statistic: 56.58 on 2 and 38 DF,  p-value: 4.042e-12
```

The p-values of the coefficients are all significant at a significance level $\alpha = 0.05$. In addition, the R-squared is 74.86% indicating that about 75% of the variability is explained by the model, which is good. Figure 37 shows the scatter plot and the fitted regression curve. A more flexible representation of the relationship between

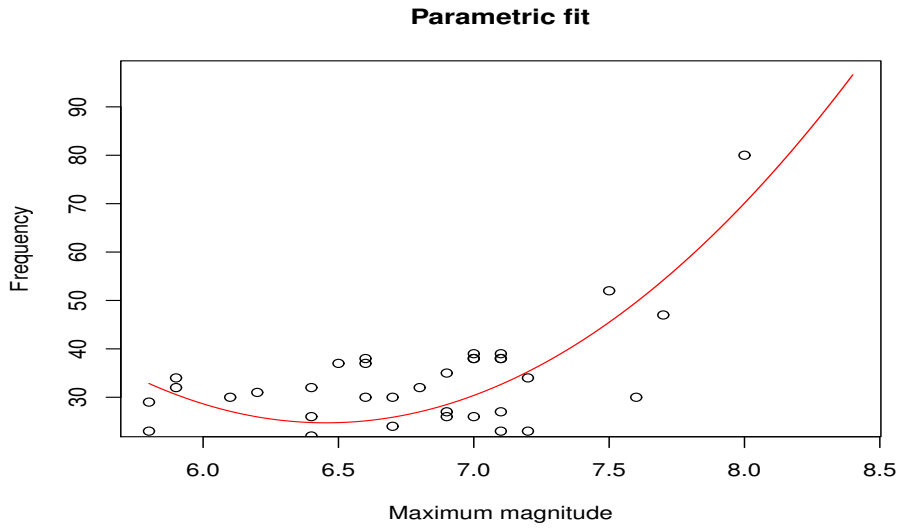

Figure 37: Quadratic model fit for the frequency of events per year on maximum magnitude of events

maximum intensity and number of events might be described. With that purpose, a nonparametric regression using $np - package$ [10] was applied. Figure 38 shows the scatter plot and the nonparametric curve dictated by the data. The nonparametric regression approach seems to be more appropriate in this case. The second order polynomial is a more rigid model and suggests that as magnitude increases at first, the

number of events goes down first and then up. On the other hand, the nonparametric regression suggests that up to magnitude 7.3, the number of earthquakes in the year is partly stable. But, when the strongest earthquake has intensity 7.5 or more, the number of earthquakes in the year increases dramatically.

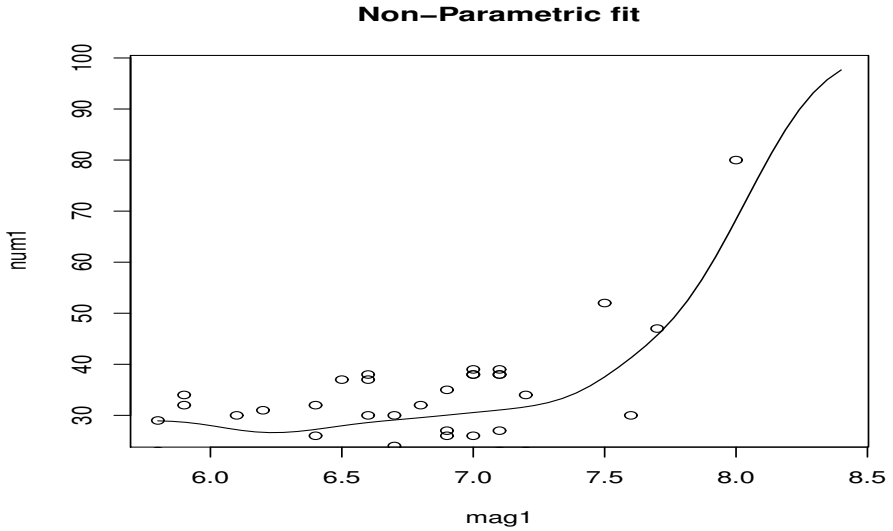

Figure 38: Non-parametric quadratic model fit for the frequency of events per year on maximum magnitude of events

6 CONCLUSIONS

In this thesis, we have applied both spatial and spatio-temporal tools to analyze the behavior of earthquakes between longitude -83° and -68°, and latitude -18° and 0° which includes Peru, Ecuador, and some parts of Brazil, Bolivia and Chile.

From the analysis of space only, we see that the intensity of the earthquakes is not uniform. Most of the earthquakes happen in southern Peru and along the coast of Northern Peru and Ecuador. The analysis in terms of time only shows that the rate of occurrence of the earthquakes appears to be white noise. In addition, the number of earthquakes in a year is more or less the same for each year. There were years with exceptionally high number of earthquakes, but there is no pattern to this effect. The same explanations goes for the rate of earthquakes in a given month for all the years.

Further, the simultaneous space-time analysis behaved similar to the space only analysis. The intensity is inhomogeneous. There is a high interaction between earthquakes at very short distances and short time periods as can be seen from the Pair Correlation Function (PCF). Also, when there are fewer earthquakes in a year, these tend to be more randomly located than when there are many earthquakes. The probable explanation of this, coming from the regression analysis of magnitude on frequency is that the years that have one or more earthquakes of great magnitude (greater than or equal to 7.5), there are several aftershocks that increases the number of earthquakes for that year. The aftershocks are naturally located near the location of the original strong earthquake, and thus the events are more clustered..

Random simulations of the earthquakes also strongly supports that the intensity of the earthquakes is inhomogeneous. The random simulation using the inhomogeneous

intensity estimate is almost the same as the original data which is not the case for the simulation with homogeneous intensity.

The application of tools for spatial temporal analysis provides an additional perspective beyond what separate analysis of time alone or space alone provide. The existing tools of spatial-temporal analysis such as the $K - function$ and the pairwise correlation were applied to the case study. However the additional analysis proposed in this thesis regarding the distances in time and space between consecutive events and the joint analysis of time, space and a third variable (magnitude) proved also to be useful to understand earthquakes.

Although this thesis only revealed the behavior of earthquakes, the same tools could be applied to other cases of spatial-temporal point patterns. In a similar way, more tools, such as spatial models, could be applied to the analysis of earthquakes.

BIBLIOGRAPHY

[1] Spatial point patterns: Methodology and applications with r. 2015.

[2] Adrian Baddeley. *Analysing spatial point patterns in R.* CSIRO, CSIRO and University of Western Australia, third edition edition, 2010.

[3] Noel Cressie and Christopher K. Wikle. *Statistics for Spatio-Temporal Data.* John Wiley & Sons, Noel Cressie, Department of Statistics, Ohio State University and Christopher K. Wikle, Department of Statistics, Missouri University, 2011.

[4] Marie Laure Delignette-Muller and Christophe Dutang. fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34, 2015.

[5] Peter J. Diggle. *Statistical Analysis of Spatial and SpatioTemporal Point Patterns*, volume Third Edition. CRC Press Taylor & Francis Group, Lancaster University England, UK, third edition edition, 2014.

[6] P.J. Diggle and B.S. Rowlingson. *A conditional approach to point process modelling of elevated risk.*, volume 157. Journal of the Royal Statistical Society, A, 1994.

[7] Edith Gabriel and Peter J. Diggle. Second-order analysis of inhomogeneous spatio-temporal point process dat. 63, 2009.

[8] Edith Gabriel, Peter J Diggle, and stan function by Barry Rowlingson. *stpp: Space-Time Point Pattern simulation, visualisation and analysis*, 2012. R package version 1.0-2.

[9] Jon Graham. Which earthquakes are included on the map and list, January 9 2016 8:05PM.

[10] Tristen Hayfield and Jeffrey S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 2008.

[11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[12] USGS. Which earthquakes are included on the map and list, September 28,2015 3:45PM.

[13] Forest vegetation in the vicinity of Choshi. Coastal flora and Chiba Prefecture IV. vegetation at Choshi. volume 3. Bulletin of Choshi Marine Laboratory, Chiba University, 1961.

[14] M. Zhukov Yuri. *Applied Spatial Statistics in R, Section 4*. 01 2010. Lecture.

VITA

ABDUL-NASAH SOALE

Education	MS Mathematical Sciences, East Tennessee State University, 2016.
	B.Sc Actuarial Science, Kwame Nkrumah Univ. of Sci and Technology, 2012.
Professional Experience	Mathematics Tutor, Center for Academic Achievement, ETSU, Jan. 2015 - May. 2016.
	Research and IT Officer, Choice Ghana, Feb. 2014 - Jan. 2015
	Mathematics Teacher, Tamale Sr. High School, Oct. 2012 - Aug. 2013.
Professional Development (Software)	Statistical and Mathematical: SAS, R, SPSS, Minitab, STATA, QM, GLPK
	Programming Languages: C++, C#, Visual Basic, Java
	Scripting Languages: PHP, HTML, JavaScript, CSS, ASP.NET, Python, Latex
	Database Server: MySQL, SQL Server
	Microsoft Office Suite : MS Access VBA, Word, Excel, PowerPoint, Publisher, Outlook

	Tools/Servers: Matlab, Apache server
	Adobe Design: Photoshop, Fireworks, Illustrator, Dreamweaver
	Web Development: Wordpress, Drupal
	Operating Systems: Windows and Linux
Honors and Awards	Vice Chancellor's Excellence Awards Best Student , Faculty of Physical Sciences, KNUST, 2016.