5-2015

Exploring Ways of Identifying Outliers in Spatial Point Patterns

Jie Liu
East Tennessee State University

Exploring Ways of Identifying Outliers in Spatial Point Patterns

A thesis

presented to

the faculty of the Department of Mathematics and Statistics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

by

Jie Liu

May 2015

Edith Seier, Ph.D., Chair

Michele Joyner, Ph.D.

Yali Liu, Ph.D.

Keywords: spatial data, outlier, distance method, statistics.

ABSTRACT

Exploring Ways of Identifying Outliers in Spatial Point Patterns

by

Jie Liu

This work discusses alternative methods to detect outliers in spatial point patterns. Outliers are defined based on location only and also with respect to associated variables. Throughout the thesis we discuss five case studies, three of them come from experiments with spiders and bees, and the other two are data from earthquakes in a certain region. One of the main conclusions is that when detecting outliers from the point of view of location we need to take into consideration both the degree of clustering of the events and the context of the study. When detecting outliers from the point of view of an associated variable, outliers can be identified from a global or local perspective. For global outliers, one of the main questions addressed is whether the outliers tend to be clustered or randomly distributed in the region. All the work was done using the R programming language.

DEDICATION

I would like to dedicate this work to my beloved family for always standing behind me, and my sweet boyfriend Kai Pei for the love and support.

ACKNOWLEDGMENTS

First, I would like to express my sincerest appreciation to my thesis advisor, Dr. Seier Edith, for all her support, guidance, advice, encouragement and patience for this work. I could not have expected to have a better advisor during my Masters studies. Also, I would like to thank the other committee members, Dr. Michele Joyner and Dr. Yali Liu, for the assistance and trust they offered. Furthermore, I must acknowledge the Department of Mathematics and Statistics at East Tennessee State University for providing me with financial support. In the end, a very special thanks goes to my parents and my boyfriend, Bingshan Liu, Yuzhen Li and Kai Pei for being by my side all the way.

TABLE OF CONTENTS

LIST OF FIGURES

1 INTRODUCTION

Spatial point pattern is a set of points distributed irregularly within a specific space [1]. In Section 2 of the thesis we review the literature on the analysis of spatial point patterns. There are three types of distributions for spatial point patterns, random, regular and clustered [2]. There are three estimated functions which can be used to determine which distribution a given spatial point pattern exhibits, the G function, K function and F function. We use as motivation 5 case studies of spatial point patterns that have not been studied in the literature. Three come from data produced at ETSU by the NSF funded CRAWL (Collaborative Research in the Arthopod Way of Life) project. Two of those three cases refer to spiders' location in a web at a certain time and the other one is the location of waggle dances in an experimental bee hive during one day. The other two cases have been created retrieving data from a large data base on earthquakes. We do basic analysis of the five cases based on the methods from the literature review and then we propose some methods for outliers' identification and apply them to those case studies. For the first three cases, we apply the G function, K function and F function using Euclidean distance. The earthquake data are given in latitude and longitude. The R package *sp* [7] is used to handle these coordinates.

Regarding the outliers in a spatial point pattern, I focus on two topics. The first one is when only the location of the events is considered and the second one is when there is one or more variables associated to the event besides the coordinates. In Section 3, we discuss alternative methods to detect outliers in spatial point patterns when we only consider the location of the events. Most of the methods explored are based on the distances to the nearest neighbor (dnn). To determine whether a point is an outlier or not, we need to discuss the distribution of *dnn* because a single method might not be appropriate for all shapes of distributions. The method to detect outliers might depend on the shape of the distribution. The methods considered are (1) the usual application of the boxplot, (2) the first gap method, (3) applying Method 1 to the transformed data, (4) three standard deviations from the mean of the transformed data, (5) using the adjusted boxplot based on a linear model and (6) using the adjusted boxplot based on an exponential model [9].

Section 4 addresses the problem of detecting outliers with respect to one or more associated variables. In some cases, like the earthquakes examples, there might be other variables such as magnitude and depth of epicenter associated to each event. We discuss methods to identify outliers with respect to these variables, i.e., when the value of the variable is unusually high or low considering the values associated to all of the surrounding events. We consider the concepts of global and local outliers. We study the spatial distribution of the global outliers. To identify local outliers, we define a circle around the event, with a defined value of radius, and check whether the value for the event would be an outlier compared to the values for the points inside the circle. We first work with each associated variable separately and then with both variables together to detect outliers in the spatial point pattern. To work with two associated variables at the same time, we standardize the values of each variable and plot their absolute standardized values in one plot.

It is necessary to write computer code to perform the proposed analysis. All the work is done using the R programming language.

2 ANALYSIS OF SPATIAL POINT PATTERNS

2.1 Spatial Point Patterns

A set of points distributed irregularly within a specific space, is called spatial point pattern [1]. Classical examples from the literature include the location of trees in a forest, nests of birds, cell nuclei in tissue, people with a certain illness in a region. The trees, nest, nuclei, sick people in the examples are called ‘events’, and we care for the location of the events. We will work with the coordinates of where events happen. Here we will work mainly with five case studies. The first two come from observations of the location of spiders in a web; the third case is about the location of waggle dances done by bees in an experimental hive. These three data sets were produced in the undergraduate research project CRAWL (Collaborative Research in the Arthropod Way of Life) at ETSU. The last two case studies are the location of earthquakes within certain values of longitude and latitude; these two cases are examples where the information about the value of an associated variable of interest in each location is available, such as the intensity or the depth at which the earthquake happened. Next we will describe these examples that will be analyzed in this work.

2.1.1 Case Studies

Cases 1 and 2 refer to experiments done with spiders of the species *Anelosimus studiosus*. Brooding spiders were located each one in an enclosure of 28cm x 28cm. When the offspring was visible, the location of each juvenile spider was recorded in four different stages of development (observation day 0, 23, 35 and 48) 6 times a day. Colony 35 in Case 1 is formed by a mother and her offspring; mother is no longer alive or visible in Colony 32 for Case 2.

Case 1. Spiders in Colony 35, first day of observation at midnight, the mother is absent

The location of the spiders in Case 1 is displayed in Figure 1A.

Case 2. Spiders in Colony 32, first day of observation at 8am in the morning, mother is present

The location of the spiders in Case 2 is displayed in Figure 1B. The mother is the point distant from the others in Figure 1B.

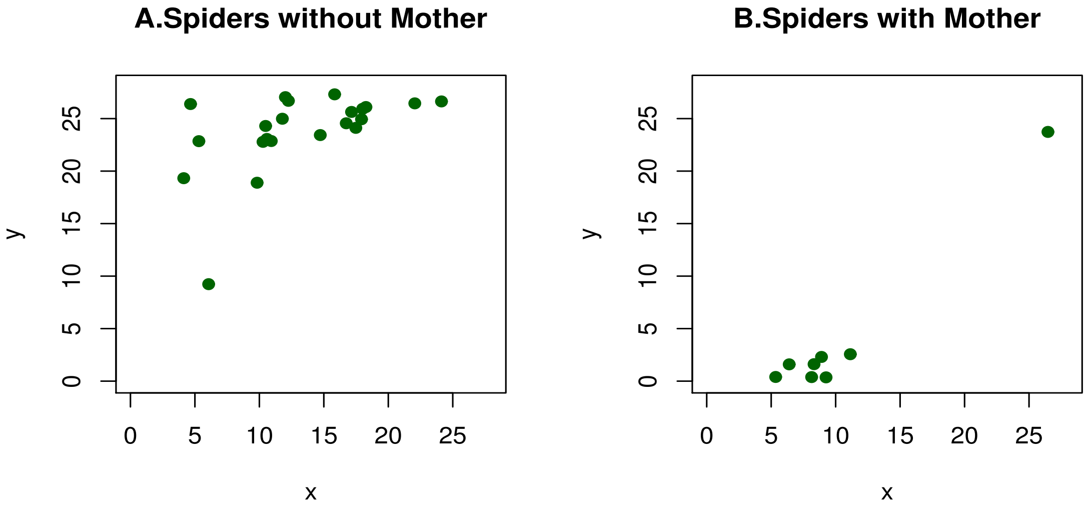

Figure 1: Locations of spiders in two different colonies

Case 3. The Location of waggle dances in a hive

As part of the data collection in the CRAWL project, the location of the waggle dances that happened during the day in a square (2 connected panels, each 50x25cm), experimental hive were recorded. Figure 2 displays the location of the waggle dances during observation day 3 in a given hive.

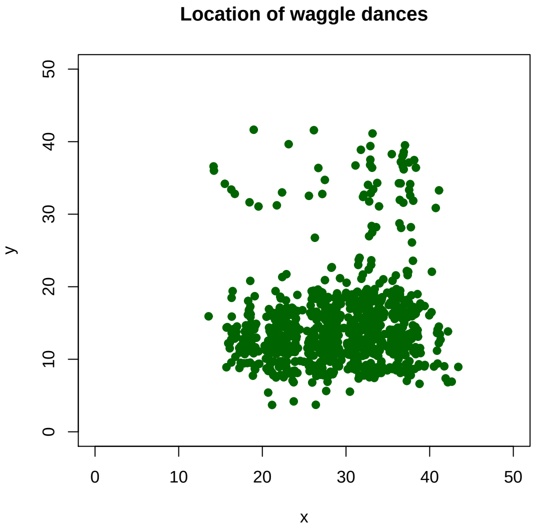

Figure 2: The location of waggle dances of bees in a hive

Case 4 and 5 refer to earthquakes that happened within certain longitude (-68°, -83°) and latitude (0°, -18°). That region includes the country Peru. The location of the earthquakes (since 1973) can be found at the website of Earthquake Hazards Program [3]. As for the earthquakes, I prepared two data sets, one for all the earthquakes with magnitude 5 or more and the other is for all the earthquakes with magnitude 6 or more.

Case 4. Earthquakes of magnitude 5 or more within certain longitude (-68°, -83°) and latitude (0°, -18°)

The location of the earthquakes of magnitude 5 or more within certain longitude (-68°, -83°) and latitude (0°, -18°) are in Figure 3A.

Case 5. Earthquakes of magnitude 6 or more within certain longitude (-68°, -83°) and latitude (0°, -18°)

The location of the earthquakes with magnitude 6 or more within certain longitude (-68°, -83°) and latitude (0°, -18°) are displayed in Figure 3B.

Case studies 1-5 are examples of spatial point patterns. In the next section, the tools used to describe the distribution of spatial point patterns will be summarized.

2.2 Types of Distributions

Basically, there are three types of distributions for the spatial point patterns, which are 'random', 'regular' and 'aggregated' distributions [2].

Random distribution indicates a completely random pattern, which shows no obvious structure, i.e., there is equal chance for any point to occur at any location within the space and the events will not influence each other at all. Regular distribution means the events are distributed regularly to some degree, like every event within the space is at nearly the same distance from all of its neighbors. Aggregated distribution means clustered points on the space, where many points are distributed close together.

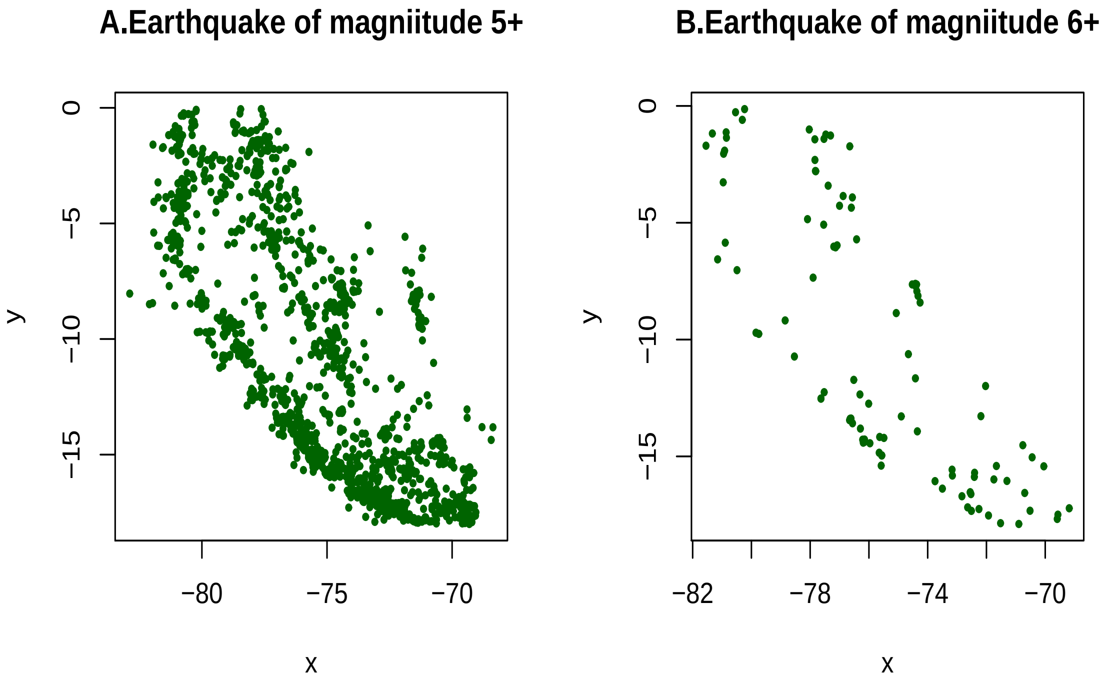

Figure 3: Locations of earthquakes within certain longitude (-68°, -83°) and latitude (0°, -18°)

2.2.1 Testing for Complete Spatial Randomness

Before doing further analysis of the spatial point patterns, we should test for complete spatial randomness (CSR). CSR assumes the points within the region follow a homogeneous Poisson point process. There are several methods to check for CSR, the most important ones are the distance methods [6]. They are nearest neighbour distance method, inter-event distance method and point to nearest event distance method. The following statistical tests can be conducted to test for significant patterns in our data.

Ho: events exhibit complete spatial randomness (CSR)

Ha: events are spatially clustered or dispersed

If the hypothesis of the complete spatial randomness for a spatial point patten is not rejected, then it is assumed that the number of events in the region follows a Poisson distribution and the events in the region are distributed randomly and independently. In other words, all the events have equal chance to occur anywhere over the space and will not influence each other [4].

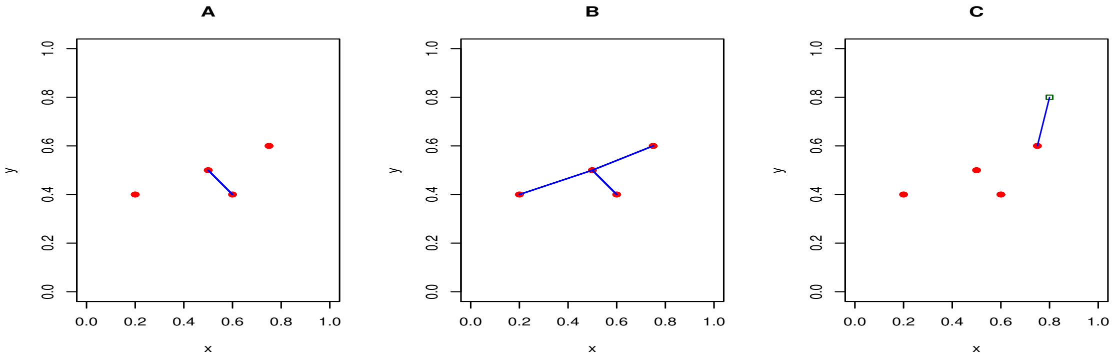

Figure 4: Different types of distances

There are three distance methods for testing CSR. The decision about CSR is made based on the cumulative distribution function of a variable that is a distance; the difference between different methods is the type of distance they consider. Figure 4 depicts these different types of distributions. Consider a spatial point pattern with only 4 events, Figure 4A shows the distance to the nearest neighbor for the particular event in the center of the region. Figure 4B shows all the inter-event distances for the events in the space. Figure 4C shows a random point in the upper corner of the region and the distance from that random point to the closest event.

2.2.2 Distance Methods Testing the Complete Spatial Randomness

Nearest Neighbor Distance Method

Illian [1] indicates that in a specific space with n events, r_{min} denotes the distance from the ith event to its nearest neighbour (Figure 4A). The G function is the CDF of the variables 'distance to the nearest neighbor'. The $\hat{G}$ function is the cumulative frequency distribution of the nearest neighbor distances calculated from the data.

$$\hat{G}(r) = \frac{N[r_{min}(s_i) \leq r]}{n} = \frac{\text{Number of point pairs where } \mathrm{r}_{min} \leq r}{\text{Number of points in study area}} \tag{1}$$

In the analysis of case studies, the simulated confidence envelope is constructed. The confidence envelope includes the central 95% of the values obtained by simulation assuming complete spatial randomness. If the $\hat{G}$ for a given spatial point pattern is within the envelope, it indicates the spatial point pattern is randomly distributed; if the $\hat{G}$ is above the envelope, it indicates the spatial point pattern is clustered. An example of $\hat{G}$ can be seen in Figure 5b.

Inter-event Distance Analysis

The second type of distance is the inter-event distance. Compared to the nearest neighbour distance, which only considers the shortest distance, the inter-event distance method is based on all the distances between events in the study area [1]. Figure 4B shows all the inter-event distances for an sample with just four events.

Note that if there are n events, there should be $0.5n(n-1)$ inter-event distances [5]. In the inter-event analysis, we calculate the empirical distribution function of the inter-event distance called the $\hat{K}$ function.

$$\hat{K}(h) = \frac{R}{n^2} \sum \sum_{i \neq j} \frac{I_h(d_{ij})}{w_{ij}} \tag{2}$$

where R is the area of the region, n is the number of points, I_h is a dummy variables that takes value 1 if $d_{ij} \leq h$ and 0 otherwise, where d_{ij} is the inter-event distance between the i event and the j event. The symbol w_{ij} in the equation is the edge correction, i.e., the proportion of circumference of a circle centered on point i, containing point j that is in the study area (proportion is 1 if the whole circle is in the study area).

The simulated confidence envelope helps to check the distribution of a given spatial point pattern. If the $\hat{K}$ function we estimated from the spatial point pattern is within the envelope, it indicates that it is random distributed; if the $\hat{K}$ function is above the envelope, it indicates a clustered spatial point pattern. An example of $\hat{K}$ can be seen in Figure 5d.

Point to the Nearest Event Distance Analysis

In this process, we randomly select m points on the space, then we calculate the distance from each of the m points to the closest event located on the space. Figure 4C shows the distance from a random point to its nearest event. The F function is the CDF of those 'points to the nearest event distance'. The $\hat{F}$ is defined as

$$\hat{F}(r) = \frac{N[d_{min}(p_i) \leq r]}{m} = \frac{\text{Number of points pairs where } \mathrm{r}_{min} \leq r}{\text{Number of sample points}} \tag{3}$$

The simulated confidence envelope is constructed in the analysis of case studies. If the $\hat{F}$ function is within the envelope, it means the spatial point pattern is completely randomly distributed; if the $\hat{F}$ function is below the envelope, it indicates a clustered spatial point pattern. An example of $\hat{F}$ can be seen in Figure 5c.

2.3 Analysis of Case Studies

Now we will use the methods described in Section 2.2 to analyze the case studies mentioned in Section 2.1.

For Case 1 (Spiders in Colony 35), Case 2 (Spiders in Colony 32) as well as Case 3 (Location of waggle dances in a hive), we are dealing with Euclidean distance, the $\hat{G}$ function, the $\hat{K}$ function as well as the $\hat{F}$ function can be calculated as mentioned in Section 2.2. However, with regard to the earthquake locations in Case 4 and Case 5, Euclidean distance can not be applied since the earth is round and we are dealing with longitude and latitude. Therefore, in Case 4 (Earthquakes of magnitude 5 or more within certain longitude (-68°, -83°) and latitude (0°, -18°)) and Case 5 (Earthquakes of magnitude 6 or more within certain longitude (-68°, -83°) and latitude (0°, -18°)), the R package *sp* [7] is used to calculate the distances between points where coordinates are given in longitude and latitude. To calculate such distances, the great circle method is used.

Case 1. Spiders in Colony 35, first day of observation at midnight, the mother is absent

Based on Figure 5, we can see that the $\hat{G}$ and the $\hat{K}$ are above the envelope, while the $\hat{F}$ is below the envelope, indicating that the spatial point pattern in Case 1 is clustered.

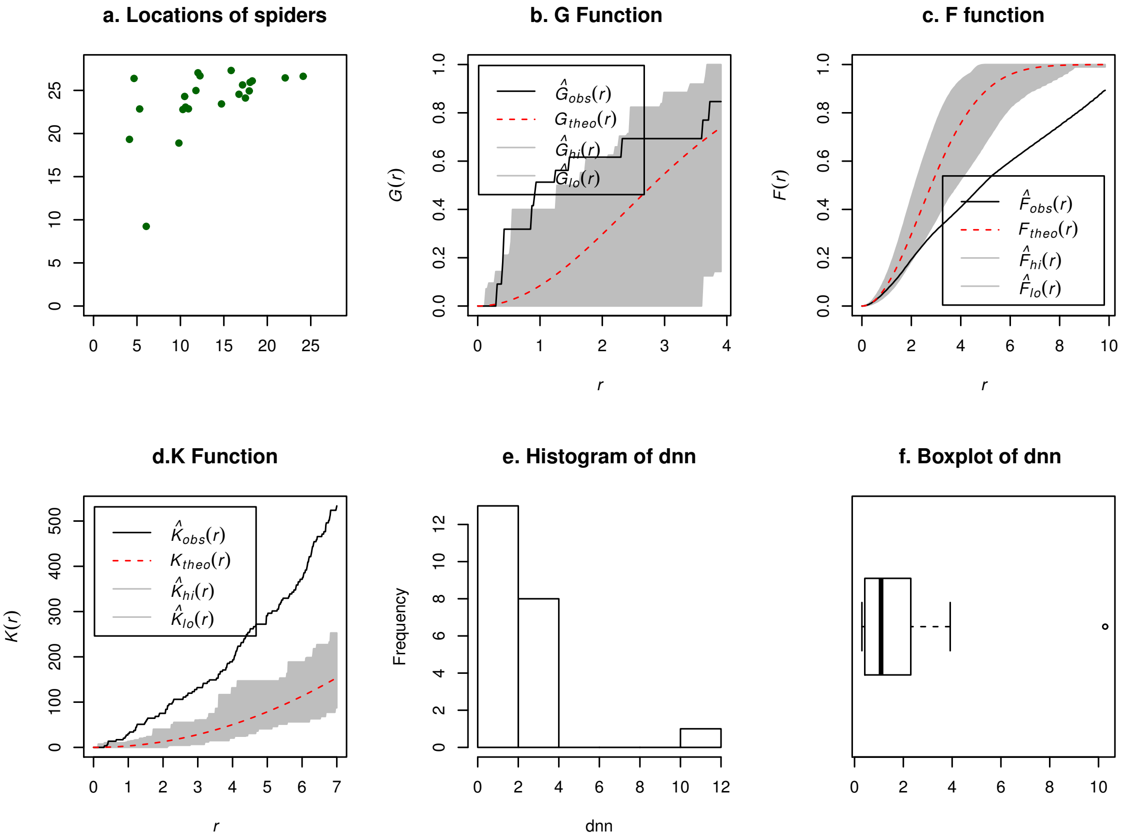

Figure 5: Analysis of Case Study 1

Case 2. Spiders in Colony 32, first day of observation at midnight, the mother is present

Figure 6 clearly indicates that the $\hat{G}$ and the $\hat{K}$ are at least partially above the envelope, while the $\hat{F}$ is below the envelope. Thus, the spatial point pattern in Case 2 is considered to be clustered.

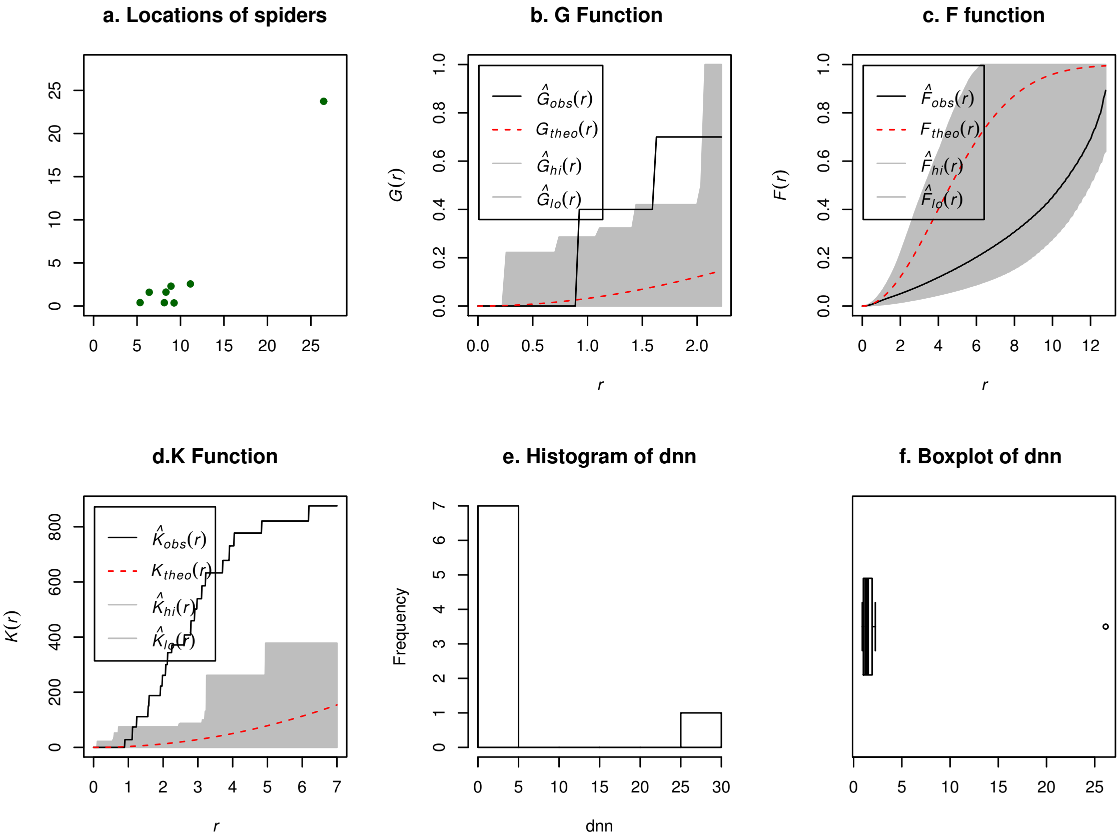

Figure 6: Analysis of Case Study 2

Case 3. The location of waggle dances in a hive

Based on Figure 7, the $\hat{G}$ and the $\hat{K}$ are above the envelope, while the $\hat{F}$ is below the envelope. Thus, the spatial point pattern in Case 3 is clustered.

Case 4. Earthquakes of magnitude 5 or more within certain longitude (-68°, -83°) and latitude (0°, -18°)

Based on Figure 8, by comparing the $\hat{G}$ estimated from Case 4 with the $\hat{G}$ under CSR calculated with a simulated large data set, we conclude the spatial pattern of the location of the earthquakes with magnitude 5 is clustered. Because the $\hat{G}$ estimated from the given spatial point pattern is above the $\hat{G}$ estimated from the simulated data set assuming spatial randomness.

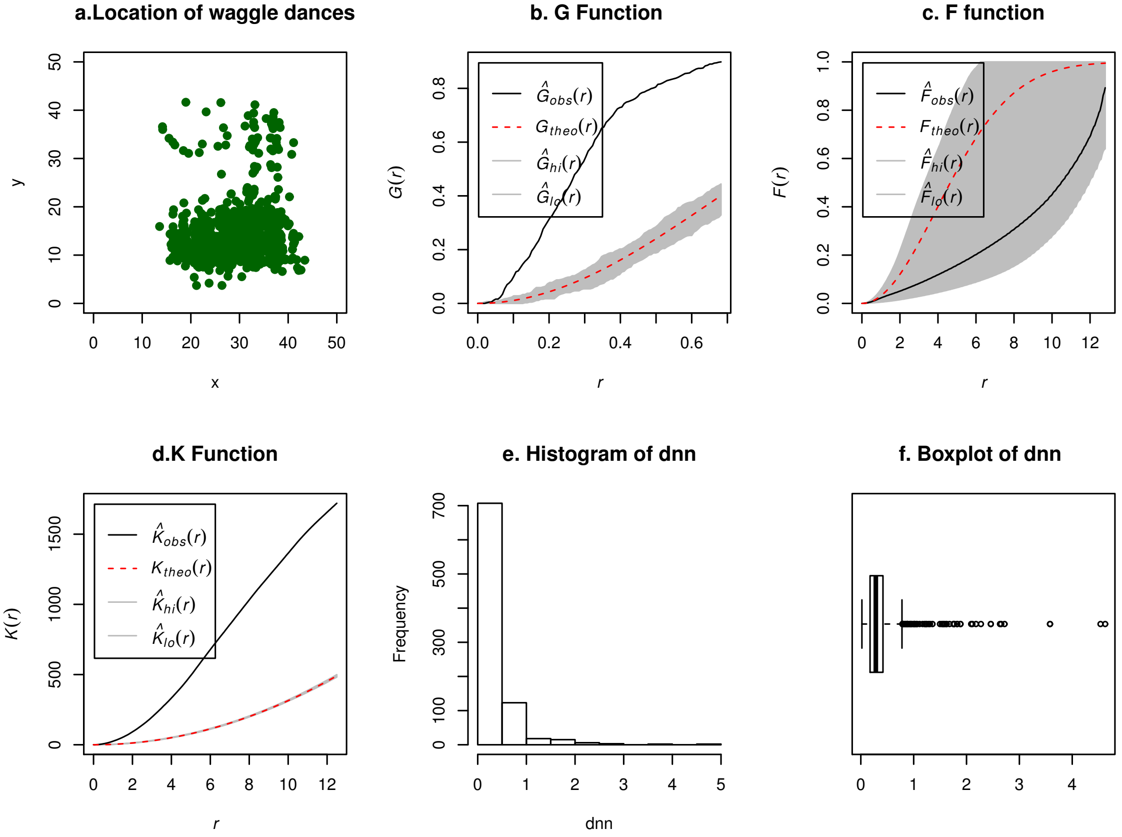

Figure 7: Analysis of Case Study 3

Earthquakes do not randomly happen anywhere, they are determined by geology. That explains why Case 4 shows a clustered pattern as well. Earths outer shell is formed of approximately ten large and about twenty small rigid tectonic plates that move slowly but continuously. According to the plate tectonics theory, earthquakes happen when tectonic plates touch each other or separate. The region under study is located between longitudes -68° and -83° and from the equatorial line to latitude -18°. Part of that region is situated on the encounter of two tectonic plates: the Nazca Plate and the South American plate. Peru is located along the boundary of two tectonic plates. These two plates are located closely putting huge strain on the Earth's crust. The pressures are periodically released through earthquakes. Earthquakes do not randomly happen anywhere, they are determined by geology. That explains why Case 4 shows a clustered pattern as well.

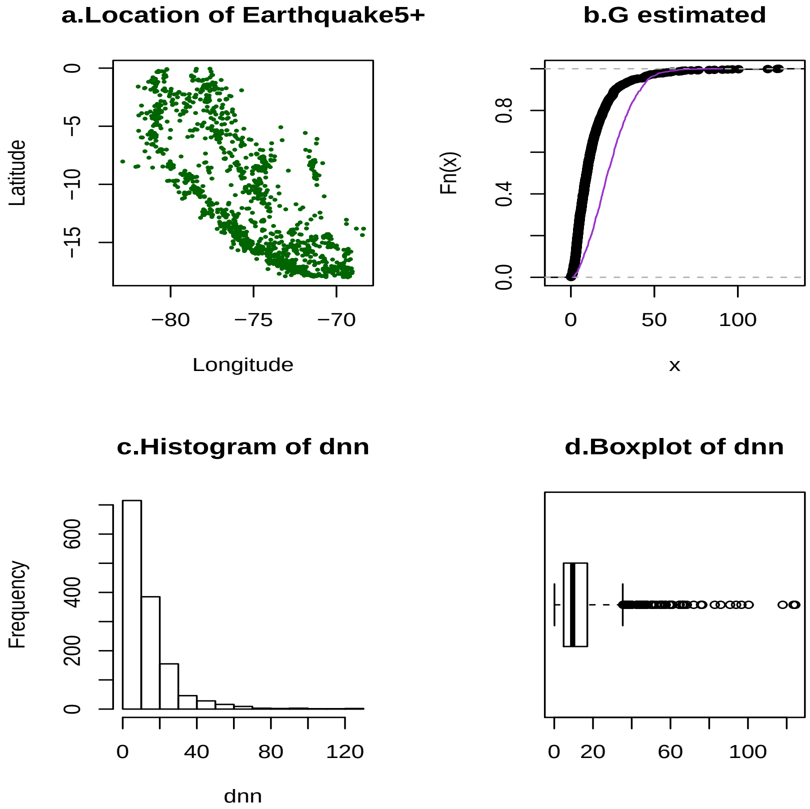

Figure 8: Analysis of Case Study 4

Case 5. Earthquakes of magnitude 6 or more within certain longitude (-68°, -83°) and latitude (0°, -18°)

Based on Figure 9, a conclusion can be reached that the spatial pattern concerning the location of earthquakes with magnitude 6 or greater is clustered.

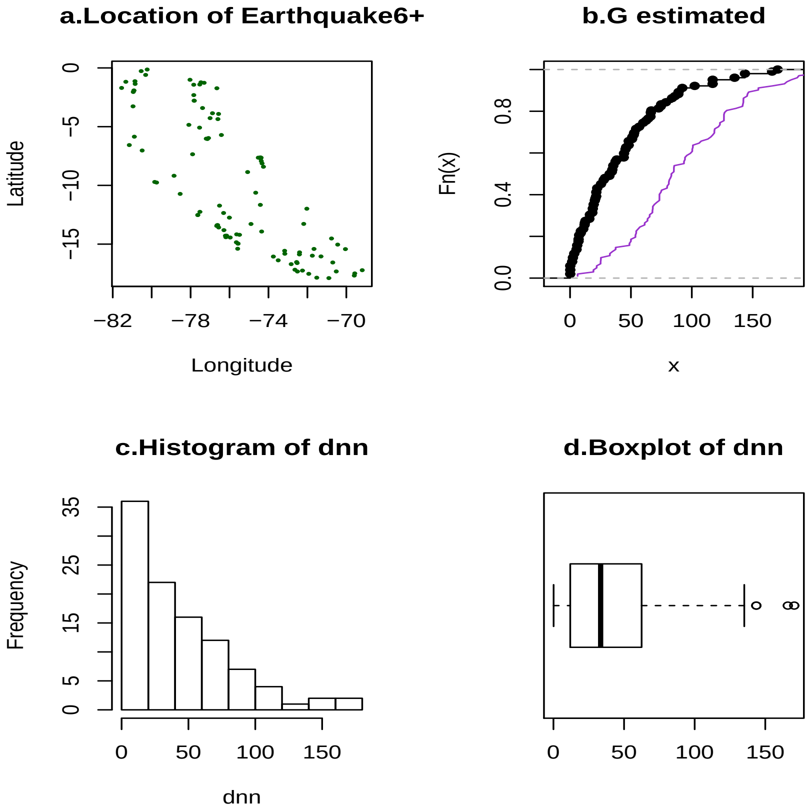

Figure 9: Analysis of Case Study 5

Cases 1-5 show a clustered or aggregated spatial distribution. However, looking at the distributions of *dnn* in Figures 5-9, we can see that the degree of aggregation is not the same. The locations of spiders in Case 2 are more clustered except for an outlier than the spiders in Case 1. The spatial point pattern in which the clustering or aggregation is stronger than the other cases is Case 3, the location of the waggle dances in an experimental hive during one day. The location of earthquakes in Case 4 are much more clustered than that in Case 5. The boxplot in Figure 8d indicates a more skewed distribution than the boxplot in Figure 9d.

3 OUTLIERS IN SPATIAL POINT PATTERNS WITH RESPECT TO LOCATION ONLY

3.1 Outliers Regarding Location Only

We will consider the term ‘outlier’ with two different meanings in the context of spatial point patterns depending on whether we are considering only the location of the events or we are considering the values of a variable associated with the events. For example, in the earthquake example we can consider only the location of the epicenter or the location and the magnitude or the depth of the earthquake. The outliers with respect to an associated variable will be considered in Section 4. Within the study area, the points which are not expected to occur on the space according to the general structure of the pattern, but they appear on the space are considered as outliers [1].

3.2 Outlier Detection Methods

In Illian et al. [1] the following comment is found ‘the basic statistical idea for outlier detection is quite simple: assign numerical or functional marks to all points in the pattern, analyze these marks statistically and regard points with extreme marks as outliers’. We will work with the distances to the nearest neighbor, which is the value assigned to each point but several criteria to determine above which distances are to be considered outliers are to be explored and compared. For our case studies, we consider outliers as the ones far away from their neighbors. Thus, we will only care for the large *dnn* values, but not the small ones.

As indicated in Section 2, the distances to the nearest neighbor can be calculated using packages in R. For points in a plane, the function *nndist* in the package *spatstat* [7] calculates the distances to the nearest neighbor using Euclidean distance. For points expressed in longitudes and latitudes, we will use the package *sp* [7] to handle with latitude and longitude coordinates [8]. The function *spDists* in the package *sp* is able to calculate distances using the great circle method. We will discuss alternative ways of identifying outliers based on the distribution of the distances to the nearest neighbor. The different ways to identify outliers that we will use are:

Method 1. The usual application of the boxplot

One simple method of identifying outliers would be to prepare a boxplot with those distances. Any value greater than $Q3 + 1.5IQR$ would be considered as an outlier.

Method 2. First gap method

We define this by looking at the histogram or the stem and leaf display and consider outliers the values beyond the first large gap that is highly visible at the right side of the distribution.

Method 3. Applying Method 1 to the transformed data

The usual boxplot is applied to the transformed data after applying the logarithm transformation or any other $Box - Cox$ transformation.

Method 4. Three standard deviations from the mean of the transformed data

Values beyond $\bar{y} + 3sd$ where y is the transformed data are considered to be outliers.

Method 5. Using the adjusted boxplot based on a linear model

In the original boxplot, we use the cutoff value $Q3 + 1.5IQR$. However, when the distribution is highly skewed, we use a different value instead of 1.5. For the linear model approach, we use $1.5 + aMC$, where MC is the medcouple value [9].

Method 6. Using the adjusted boxplot based on an exponential model

There is another model for the adjusted boxplot, the R package *Robustbase* is used to make the adjusted boxplot based on the exponential model [9].

In the next section, those methods will be presented, applied and discussed in the context of analyzing these case studies.

3.3 Case Studies

Case 1. Spiders in Colony 35, first day of observation at midnight, the mother is absent

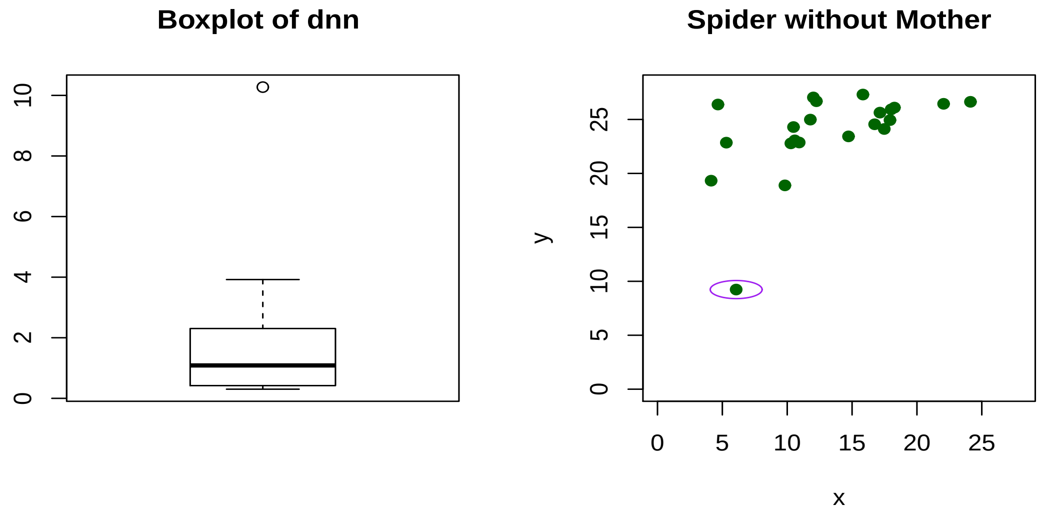

Figure 10: Outliers detection based on boxplot for Case 1

Based on Figure 10, it can be seen that an obvious outlier is identified which is shown in the purple circle by the general method that events with values greater than $Q3 + 1.5IQR$ can be considered as outliers. The distribution of dnn is not highly skewed except for the outlier and Method 1 works satisfactory.

Case 2. Spiders in Colony 32, first day of observation at 8am in the morning, mother is present

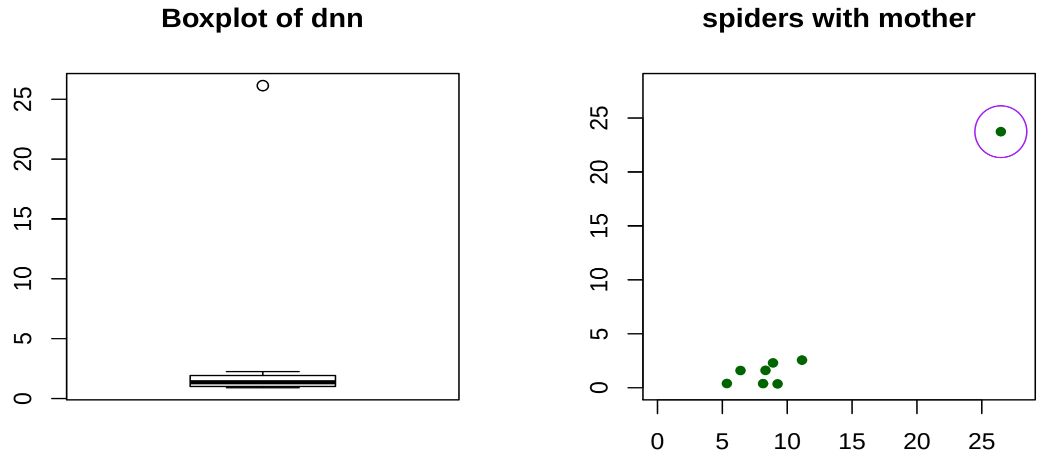

Figure 11: Outliers detection based on boxplot for Case 2

Based on Figure 11, it is clear that an obvious outlier is identified shown in the purple circle by the general method that events can be considered as outliers if they have values greater than $Q3 + 1.5IQR$. In this case, the outlier is the mother; a guess is that she went on a hunting trip to catch a prey. The boxplot without the distant outlier indicates a fairly symmetric distribution and Method 1 works well.

Case 3. Location of waggle dances in a hive

Method 1. The usual application of the boxplot

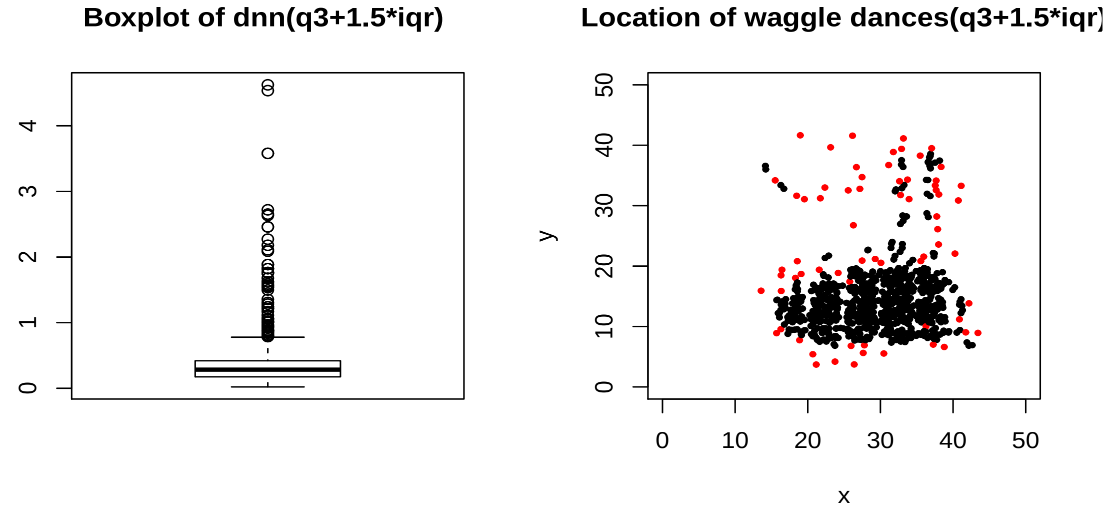

Figure 12: Outliers detection based on boxplot for Case 3

Based on Figure 12, it can be seen that a large number of outliers are detected which is shown in red by the usual method that events can be considered as outliers with values greater than $Q3 + 1.5IQR$. Figure 12 is an application of Method 1. There is a considerable number of events that are located at a moderate distance from the nearest neighbor are identified as outliers with this method. When most of the points are highly clustered, the ones that are even at a moderate distance from the nearest neighbor are identified as outliers. In this example, most distances (in cm) were very small and any point with distance to the nearest neighbor larger than 0.7844 is considered an outlier by applying the usual rule of the boxplot. Below are the basic statistics for the *dnn* of Case 3.

```
Min.     1st Qu.  Median    Mean     3rd Qu.    Max.
0.02054  0.17450  0.28390   0.38540  0.41850   4.62400
```

In cases like this we might want to rethink the way outliers are identified. Due to the overcrowding of the events on the space, some otherwise acceptable nearest neighbor distances for the events turned out to be outliers based on the Method 1.

Method 2. First gap method

In order to get a better view of how our data are clustered, a histogram and a stemplot of the *dnn* for Case 3 are displayed.

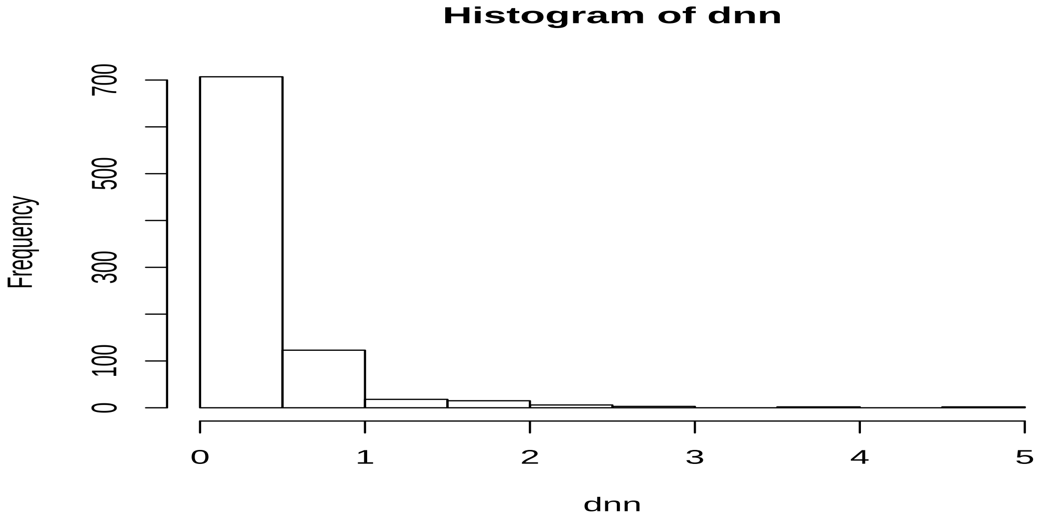

Figure 13: Histogram of the *dnn* for Case 3

Based on Figure 13, the histogram of *dnn* for Case 3, it is clear that a gap exists. The gap can indicate what the cutoff value to detect the outliers should be. However the stemplot give us a better idea of where the gaps in the *dnn* exist.

```
The decimal point is 1 digit(s) to the left of the |

   0 | 2233444444555566666666667777777777777777777777888888888888888*
   2 | 0000000000000000011111111111111111112222222222222222222222333*
   4 | 0000111111111111111122233334444444455555555555566666777777888*
   6 | 000000111111223333344444555566667788888990122222223334668999
   8 | 011446688889233557
  10 | 1112256617
  12 | 1144816
  14 | 03579
  16 | 2285666
  18 | 2299
  20 | 918
  22 | 7
  24 | 66
  26 | 352
  28 |
  30 |
  32 |
  34 | 88
  36 |
  38 |
  40 |

  42 |

  44 | 4

  46 | 2

  Note: *indicates that the line has been truncated for formatting purposes.
```

All the points after the first gap in the sorted distances could be considered as outliers. For Case 3, the dances with distance to the nearest neighbor equal or greater than 3.48 would be the natural candidates. After applying the first gap method, four outliers are detected which are shown in the darkgreen circles in Figure 14.

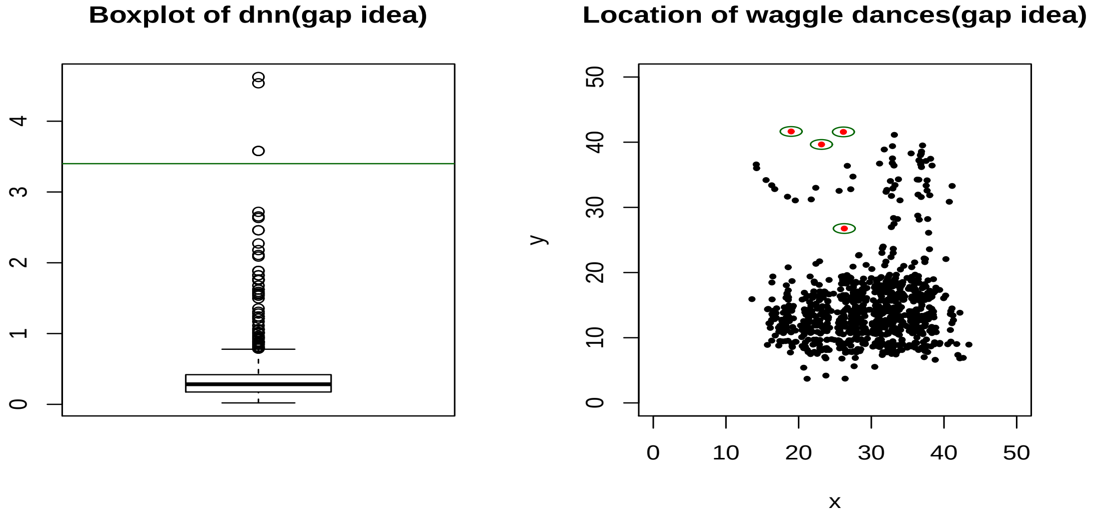

Figure 14: Outliers detection based on first gap for Case 3

Method 3. Applying Method 1 to the transformed data

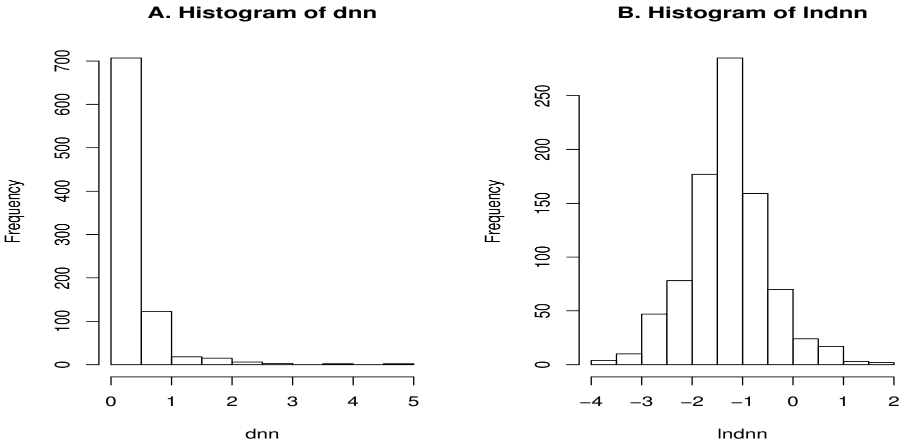

Figure 15: Histogram of *dnn* and *lndnn* for Case 3

Based on the histogram in Figure 15A, the distribution of *dnn* is highly skewed, and it looks like the distribution of *dnn* for the waggle dances is log normal. We calculated $lndnn = log(dnn)$; the distribution of the *lndnn* shown in Figure 15B is fairly symmetric.

Figure 16 is based on the method that identifies outliers with values greater than $Q3 + 1.5IQR$ using the transformed distances *lndnn*. There is a considerable number of events that are located at moderate distances from the nearest neighbor but are identified as outliers.

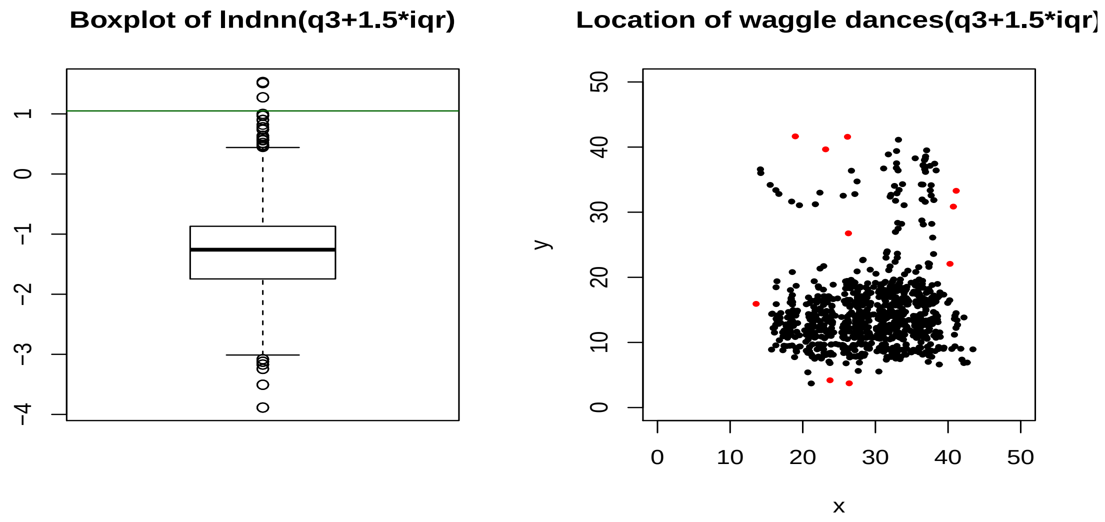

Figure 16: Outliers detection based on boxplot for *lndnn* for Case 3

Method 4. Three standard deviations from the mean of the transformed data

Another idea for detecting outliers is based on the 68-95-99.7% rule for the normal distribution. Strictly speaking, we should check first for the normality of the *dnn* or their non-linear transformation. Figure 15B indicates a fairly symmetric distribution apparently with higher skewness than the normal distribution. The *Shapiro-Wilk* test rejected the assumption of normality in this case. We need to be aware that with such a large number of observations (876), the test is very sensitive to any small departure from normality. We went ahead and applied this method that has the normal model in mind.

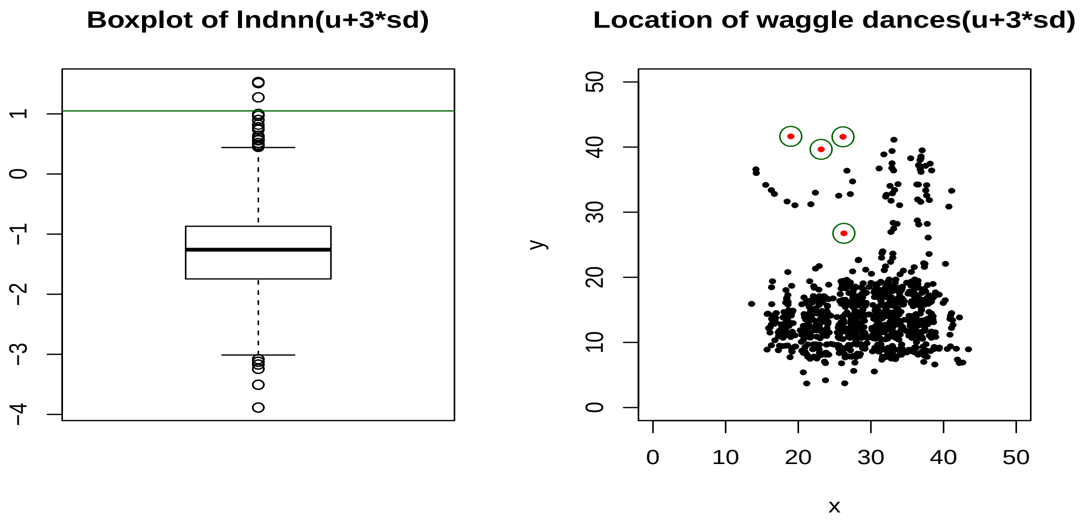

Figure 17: Outliers detection based on three standard deviations from mean of *lndnn* for Case 3

In Figure 17, the darkgreen horizontal line on the boxplot is the cutoff value of $\bar{y} + 3sd$. By applying this cutoff value, four outliers are identified shown in the darkgreen circle. Notice that these are the same four outliers identified in Figure 17 with applying the the first gap method shown in Figure 14.

Method 5. Using the adjusted Boxplot based on a linear model

Figure 18 is an application of adjusted boxplot using a linear model [9], In this case, events with *dnn* values greater than 3.5259 or lower than 0.0306 are identified as outliers. We only care about the large values, so only the ones with values of *dnn* greater than 3.5259 will be considered as outliers. They are shown in the darkgreen cicles in Figure 18.

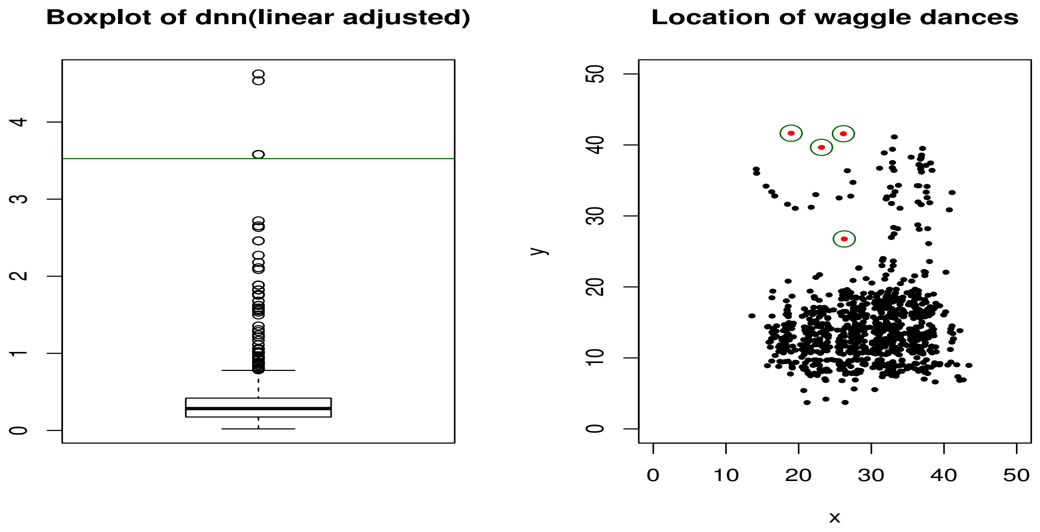

Figure 18: Outliers detection based on linear adjusted boxplot for Case 3

Method 6. Using the adjusted boxplot based on an exponential model

There is an exponential model for the adjusted boxplot applied to skewed distribution. Values greater than $Q3 + 1.5exp(3MC)IQR$ can be considered as outliers, in which MC is the medcouple value [9]. In R, the function *adjust* in the package *Robustbase* [7] can be used to make adjusted boxplot based on the exponential model. The adjusted boxplot is applied in Figure 19. We loaded the *robustbase* package and use the function *adjust*, a new method to identify outliers for a skewed distribution proposed in the reference [9]. However, a considerable number of outliers of which nearest neighbor distances are not too high can be detected. Adjusted boxplot works well for the moderate skewness, but the distribution of *dnn* in this case is highly skewed.

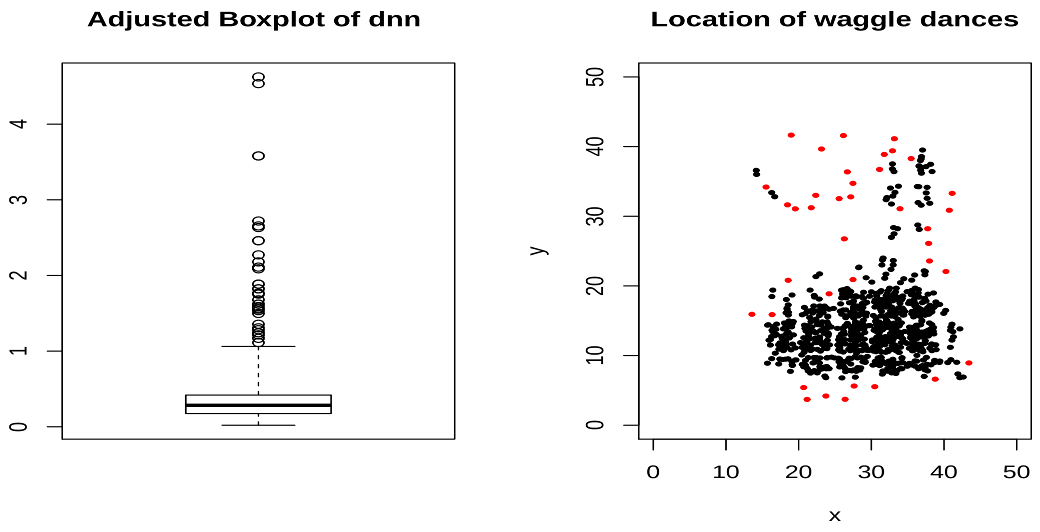

Figure 19: Outliers detection based on exponential adjusted boxplot for Case 3

In summary, for Case 3 where the events are highly clustered, some methods (Method 1, Method 3, Method 6) identify a high number of outliers. Some of those outliers are actually pretty close to other events. However, the application of the first gap method (Figure 14), three standard deviations from the mean of the transformed data (Figure 17) and using the adjusted boxplot based on a linear model (Figure 18) identify the same four outliers that stand out in the spatial point pattern.

Case 4. Earthquakes of magnitude 5 or more within certain longitude (-68°, -83°) and latitude (0°, -18°)

Method 1. The usual application of the boxplot

Figure 20 shows that a large number of outliers (in red) are detected by applying Method 1. There is a considerable number of events located at moderate distances from their nearest neighbors are identified as outliers. When most of the points are highly clustered, the ones that are even at a moderate distance from the nearest neighbor are identified as outliers. In this example, most distances (in km) were very small and any point with dnn larger than 0.2892 is considered as an outlier by applying the usual rule that values greater than $Q3+1.5IQR$ can be considered as outliers.

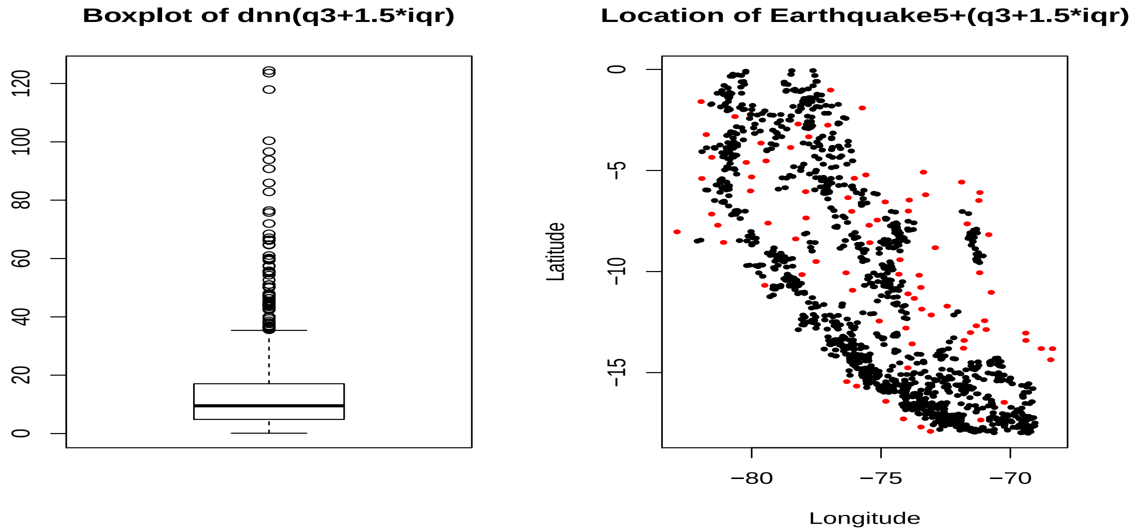

Figure 20: Outliers Detection based on boxplot for Case 4

Method 2. First gap method

It is clear that Method 1 does not work well for Case 4. To apply the first gap method, we would look at the histogram and stemplot of dnn for Case 4.

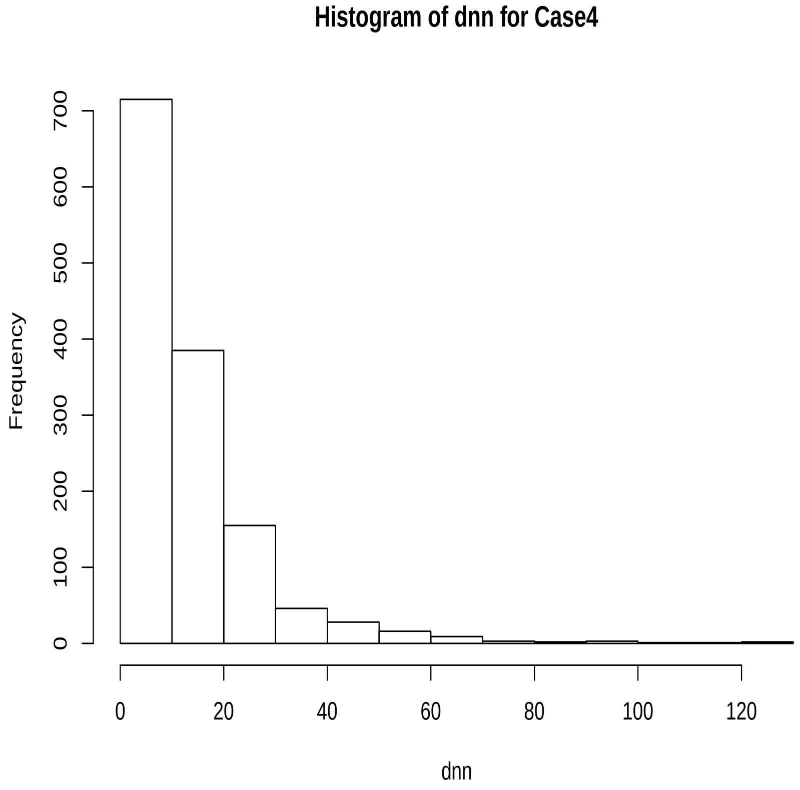

Figure 21: Histogram of the *dnn* for Case 4

```
The decimal point is 2 digit(s) to the left of the |

   0 | 111111444444445566889999

   1 | 0000111111111111111112222333344444455555555666666667777788888*

   2 | 000000111112222222222233333333444444444445555566666666667777*

   3 | 0000000000000000111111111111111222222333333333333333444444444*

   4 | 0000000000011111122222222222222333333333344444444455555555555*

   5 | 0000000000011111222222222233344444444555556666666666777777778*

   6 | 0001111112222222223333333334444444444445555555666666666777788*
```

```
 7 | 0000000000111122222222222333333333334444444555555566666666667*
 8 | 000000000001122233333344444444445555566666666666677777888888 8*
 9 | 00000000000111222222222334444444445555555555556666667777777 88*
10 | 0000001111122222333333333333444444455555566666677777777788888*
11 | 00000112222223333333334445555555566666778888888899999999
12 | 000111122223333444445556666667777778889999
13 | 0001122222222233334556666667778899999999
14 | 01111113333444555555666666777888999
15 | 01111222233334555566666677778888999
16 | 001123334555556777888888888899
17 | 0233333334455555567778899
18 | 002333333344556667777788999
19 | 000001111444555666688899
20 | 0000011333333455566799
21 | 0022235555568899
22 | 01223666788899999
23 | 111111122233344445577788
24 | 12333456678
25 | 0056667999
26 | 22344489
27 | 00444447
28 | 036668
29 | 2379
```

```
30 | 00134458
31 | 136
32 | 00125
33 | 1223369
34 | 136
35 | 0
36 | 2224
37 |
38 | 488
39 | 35568
40 | 12334457
41 | 11
42 | 145
43 | 003366
44 |
45 | 456
46 | 66
47 | 5
48 |
49 | 68
50 | 235
51 | 8
52 | 5
```

```
53 | 88
54 | 55
55 | 2
56 |
57 |
58 | 79
59 | 7
60 | 228
61 |
62 | 2
63 |
64 |
65 | 2
66 |
67 |
68 | 5
69 | 2
70 |
71 |
72 |
73 |
74 |
75 | 0
```

```
76 |
77 | 9
78 |
79 |
80 |
81 | 8
82 |
83 |
84 |
85 | 3
86 |
87 | 4
88 |
89 |
90 |
91 | 1
92 |
93 |
94 |
95 |
96 |
97 |
98 |
 99 |
100 |
101 |
102 |
103 |
104 |
105 |
106 | 9
107 |
108 |
109 |
110 |
111 | 9
112 |
113 | 0
Note: *indicates that the line has been truncated for formatting purposes.
```

Case 4 regarding the earthquakes of magnitude 5 experience the same situation as Case 3 regarding the waggle dances. Based on the histogram and the stemplot, the large gap exists between 92 and 105. We consider the *dnn* values greater than 100 can be identified as outliers. Applying the first gap method, four outliers are identified and shown in red in Figure 22.

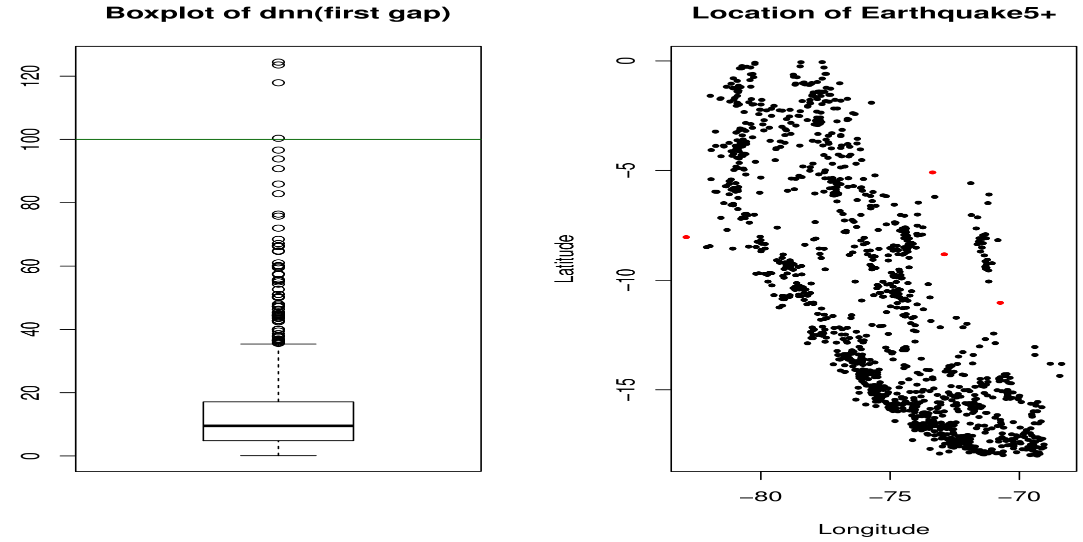

Figure 22: Outliers detection based on first gap for Case 4

Method 3. Applying Method 1 to the transformed data

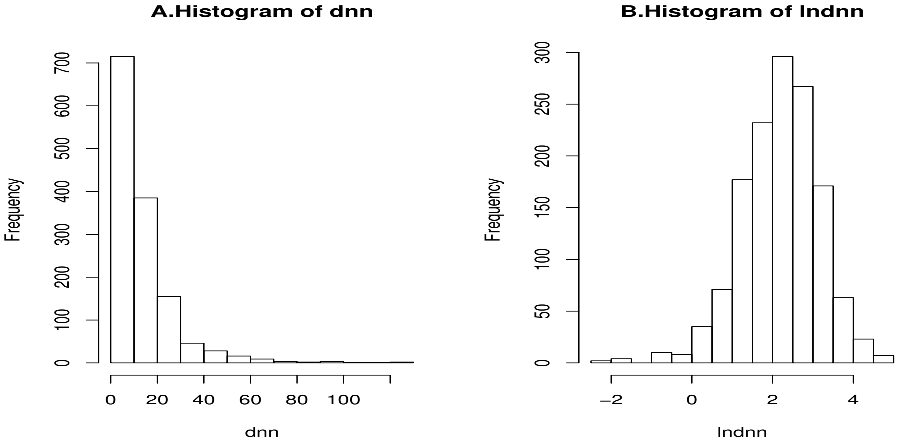

Figure 23: Histogram of *dnn* and *lndnn* for Case 4

In Figure 23A, we can see that the distribution of *dnn* is highly skewed. The distribution of the transformed distances $lndnn = log(dnn)$ is fairly symmetric except for a few extreme small outlier shown in Figure 23B. We are only concerned with large outliers but not small ones, so these small values should not matter.

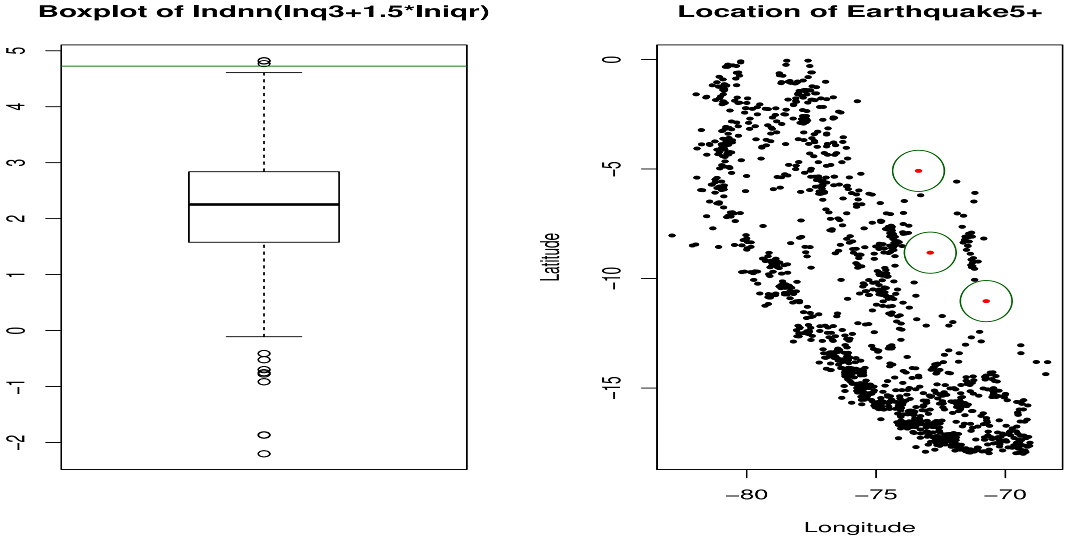

Figure 24: Outliers detection based on boxplot for *lndnn* for Case 4

Figure 24 is prepared by applying Method 1 to the transformed data *lndnn*. Three outliers were identified in this method shown in the darkgreen circles.

Method 4. Three standard deviations from the mean of the transformed data

The boxplot in Figure 25 is the boxplot of *lndnn* with a horizontal line $y = \bar{y} + 3sd$ where $\bar{y}$ is the mean of *lndnn* and sd is the standard deviation of *lndnn*. For a normal distribution, 99.7% of the population should fall into the interval $(\bar{y} - 3sd, \bar{y} + 3sd)$. Based on this idea, we want to detect the outliers greater than $\bar{y} + 3sd$ since we only care about the large outliers. No event had values of *lndnn* greater than

$\bar{y} + 3sd = 5.072$. Thus, no outlier is identified using this method.

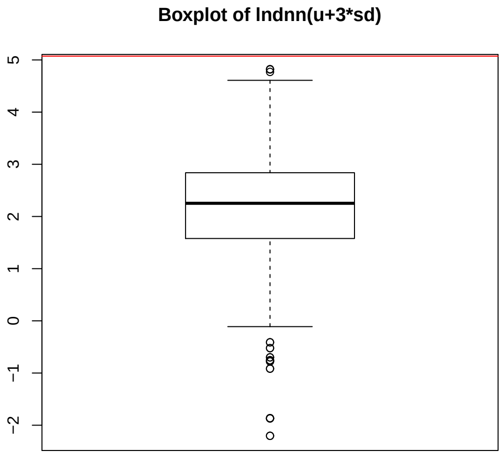

Figure 25: Outliers detection based on three standard deviations from mean of *lndnn* for Case 4

Method 5. Using the adjusted boxplot based on a linear model

As we indicated in Case 3, there is a way of identifying outliers with the adjusted boxplot using a linear model [9]. Figure 25 is an application of this method. Events with *dnn* value greater than 94.4810 are identified as outliers. Five outliers shown in red are identified by this method in Figure 26.

Method 6. Using the adjusted boxplot based on an exponential model

There is an adjusted boxplot using the exponential model. Values greater than $Q3 + 1.5exp(3MC)IQR$ can be considered as outliers, in which MC is the medcouple

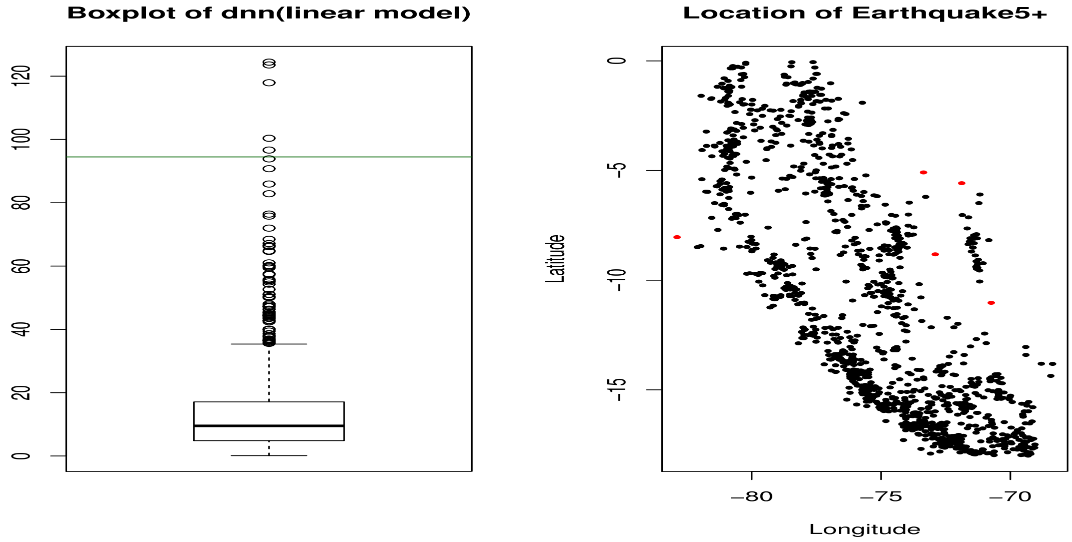

Figure 26: Outliers detection based on linear adjusted boxplot for Case 4

value [9]. The function *adjust* in the package *Robustbase* is used to make the adjusted boxplot based on an exponential model [7].

In Figure 26, the adjusted boundary value is 63.1390, which indicates that any value greater than 63.1390 can be considered as outlier. However, a considerable number of outliers shown in red are identified; some of them have moderate *dnn* values.

In summary, the events in Case 4 are highly clustered. Some methods identify a large number of outliers, but some of these outliers are pretty close to other events. However, the first gap method (Figure 22), applying Method 1 to the transformed data (Figure 23) and using the adjusted boxplot based on a linear model (Figure 26) identify the outliers that stand out in the spatial point pattern.

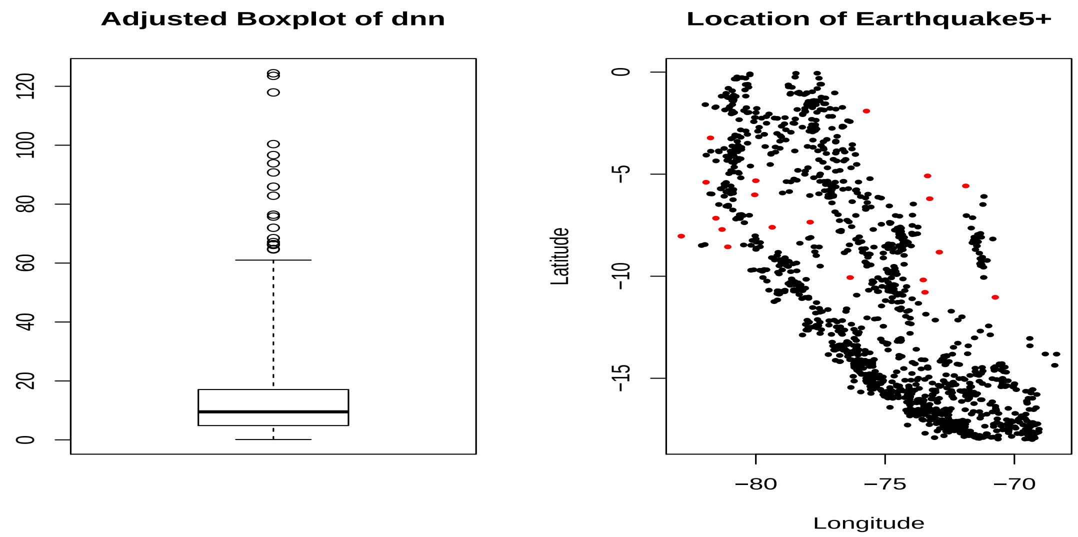

Figure 27: Outliers detection based on exponential adjusted boxplot for Case 4

Case 5. Earthquakes of magnitude 6 or more within certain longitude (-68°, 83°) and latitude (0°, -18°)

Figure 28 shows that there are four events that can be considered as outliers, because that have *dnn* values greater than $Q3 + 1.5IQR$. The distribution of *dnn* is not extremely skewed, Method 1 is good enough to identify outliers.

As to detecting outliers with respect to the location of the event, most of the current methods are based on the distances to the nearest neighbor. However, to determine whether a point is an outlier we need to discuss the distribution of *dnn* and the method to define outliers might depend on the shape of that distribution. The usual definition of outlier associated to the boxplot might not be enough. When most of the events are moderately clustered we need to consider the adjusted boxplot proposed by Huber [9] for skewed distributions. When the data are highly skewed

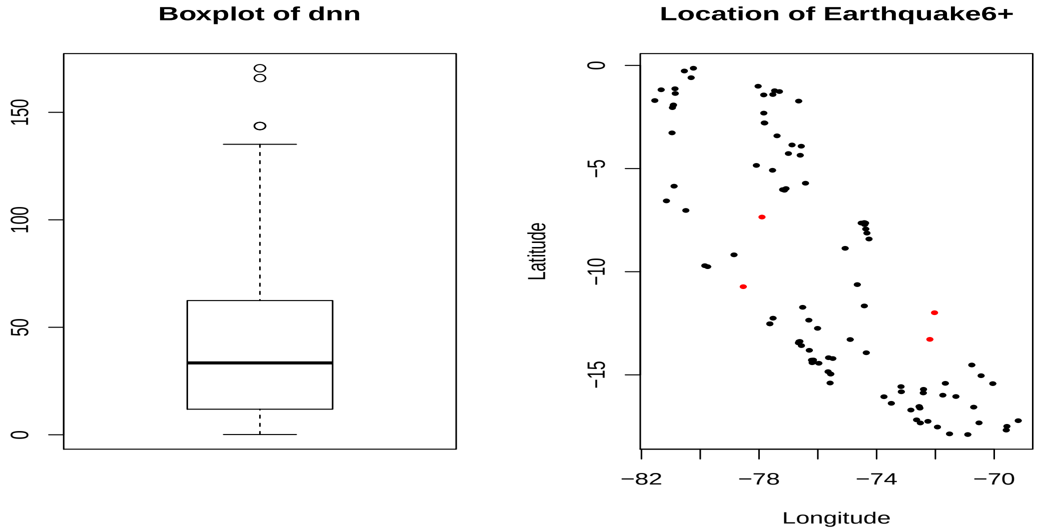

Figure 28: Outliers detection based on boxplot for Case 5

in Case 3 and Case 4, we might consider alternatives. The first gap method is an easy and intuitive method to deal with highly skewed distributions. Working with transformed data is also a good way to detect outliers for a highly skewed distribution.

4 OUTLIERS IN SPATIAL POINT PATTERNS WITH RESPECT TO AN ASSOCIATED VARIABLE

In this section we will address the problem of identifying outliers with respect to the values of one or more associated variables. In some cases, like the earthquake examples, there might be other variables such as magnitude and depth of the epicenter associated to each event. We will discuss methods to identify outliers with respect to these variables. We will consider two types of outliers, global and local. Global outliers are the events with unusually high or low values of their associated variables. Local outliers are the outliers with respect to the values associated to the surrounding events.

There is very little in the statistical literature about outliers with regard to an associated variable in spatial point patterns. However, recently a book written from the point of view of computer science has been published [11]. The following rule of spatial data is found in *Outlier Analysis* [11]: 'Everything is related to everything else, but nearby objects are more related than distant objects'. It indicates that we need to detect the events with behavioral attribute values varying much from the neighboring spatial data. The same reference [11] also indicates that when dealing with multiple behavioral attributes, we need to work with each variable separately, then make a combination of these two to get the final outlier score to do a further analysis of spatial outliers.

4.1 Working with One Associated Variable

When it comes to the earthquake data, there are two associated variables, which are depth and magnitude. In this section, we will discuss methods to identify outliers when we consider each associated variable separately.

Global Outliers

Case 4. Earthquakes of magnitude 5 or more within certain longitude (-68°, -83°) and latitude (0°, -18°)

We can determine outliers for each variable based on the usual concept that values greater than $Q3 + 1.5IQR$ can be identified as outliers. An event with an extremely high value for the associated variable can be identified as an outlier [11]. We first prepare histograms and boxplots for each associated variable.

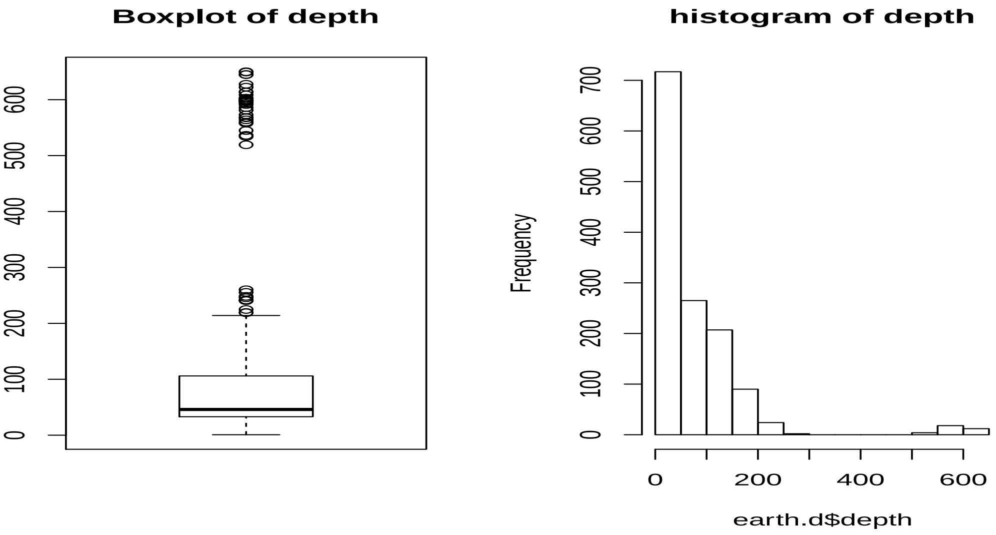

Figure 29: Boxplot and histogram of depth for Case 4

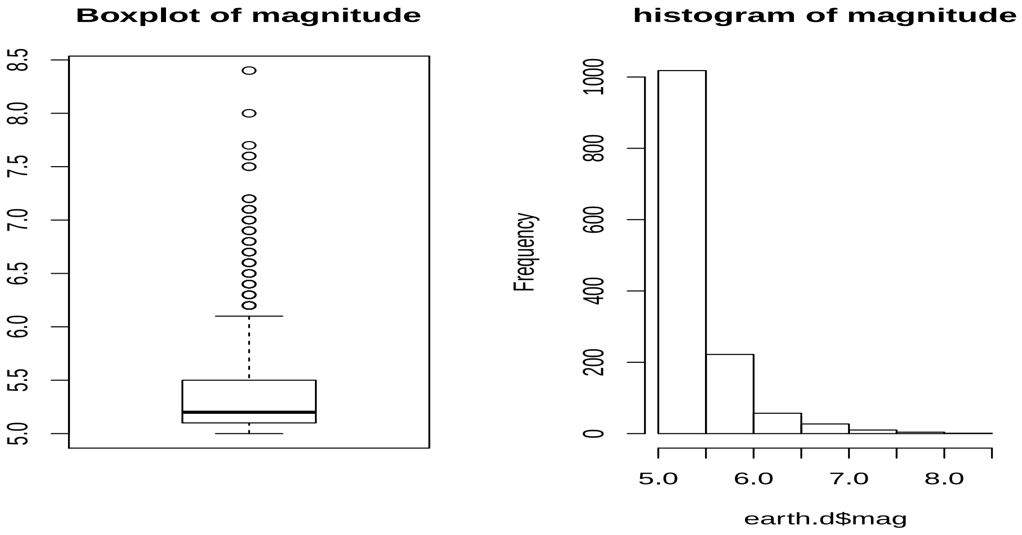

Figure 30: Boxplot and histogram of magnitude for Case 4

We notice that there seems to be a lot of outliers with respect to both associated variables according to the boxplots in Figure 29 and Figure 30. However, since these data have a spatial location, we want to see how those outliers are located. Two 3D plots are created, one for depth and one for magnitude. The points in red are those considered to be outliers according to the boxplots.

After examining the 3D scatter plots in Figure 31 and Figure 32, something that calls our attention is that the outliers from the depth point of view seem to be clustered in a certain region and the outliers with respect to magnitude are more scattered all over the region [10]. It would be nice to have a tool to visualize if the outliers with respect to an associated variable in a spatial point pattern tend to be more clustered than the events in general. We propose to compare the $\hat{G}$ for all the data with the $\hat{G}$ for just the outliers.

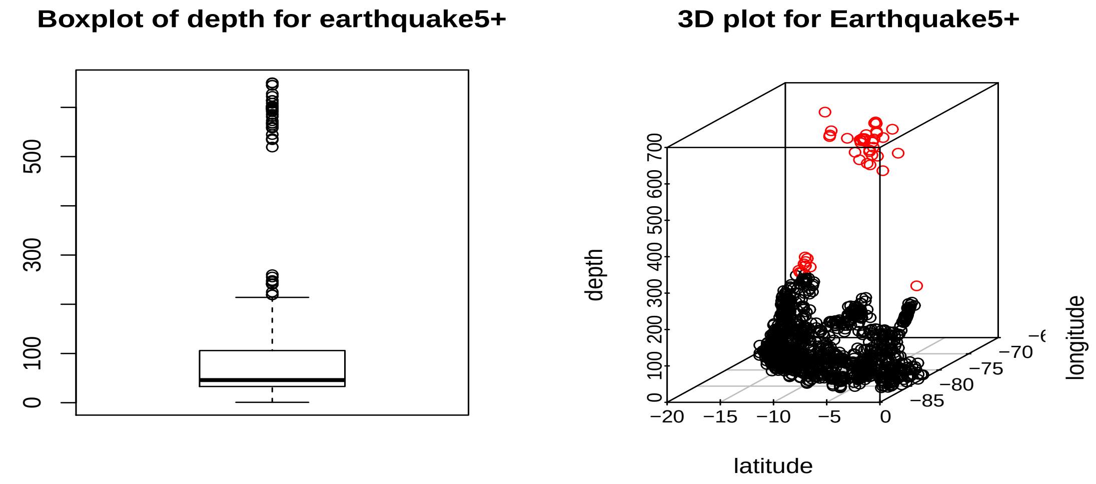

Figure 31: 3D-plot with outliers identified based on depth for Case 4

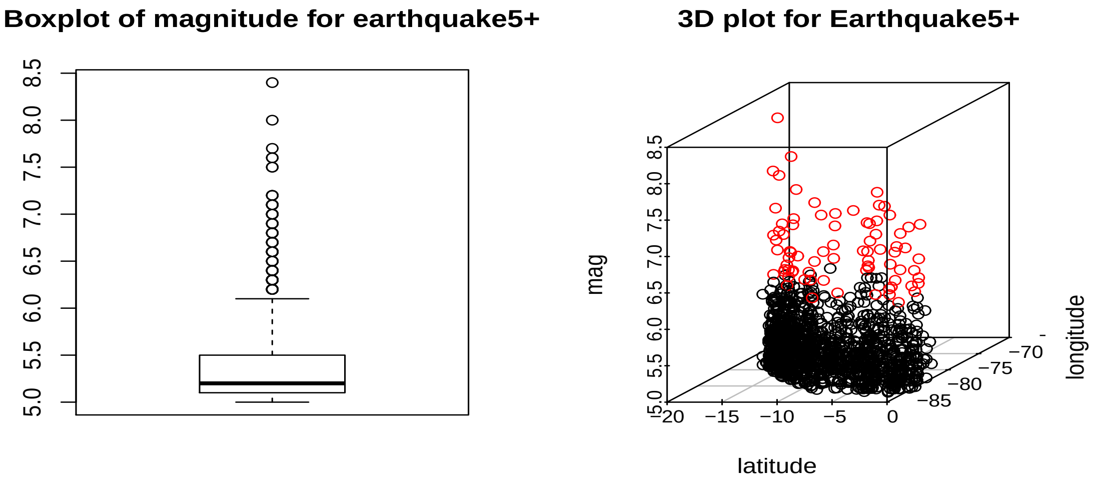

Figure 32: 3D-plot with outliers identified based on magnitude for Case 4

There are three $\hat{G}$ in Figure 33 and in Figure 34; the black one is for the whole data set, the green one is for the outliers only and the purple one is for the simulated data under the condition of complete spatial randomness.

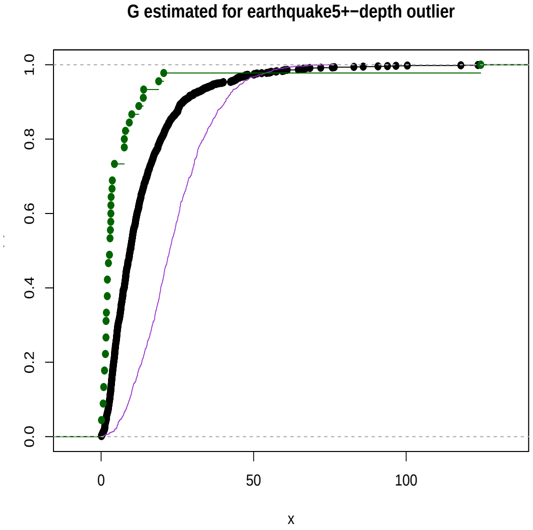

Figure 33: $\hat{G}$ with depth outliers for Case 4

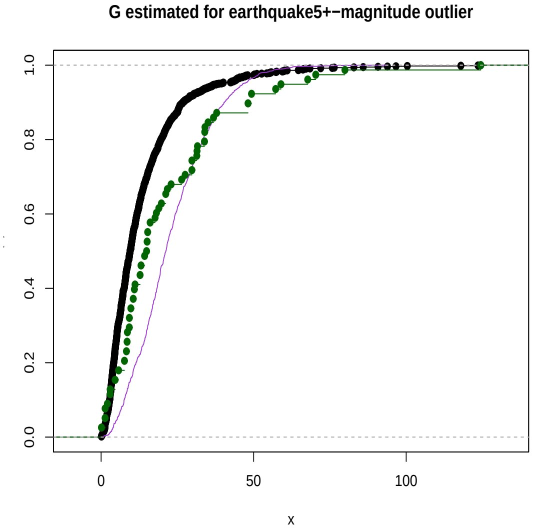

Figure 34: $\hat{G}$ with magnitude outliers for Case 4

Based on Figure 33, it is clear that the outliers with regard to depth are super aggregated, which is much more clustered than all the events in Case 4, even though the whole data set is very aggregated as we discussed in Section 2. The $\hat{G}$ for the outliers with regard to magnitude in Figure 34 indicates that they are less clustered than all the events of earthquakes of magnitude 5 or more.

Figure 35 and Figure 36 display the 2D plots for the location with the outliers shown in red for depth and magnitude respectively. We make 2D plots of the location to have a better perception of where they are located.

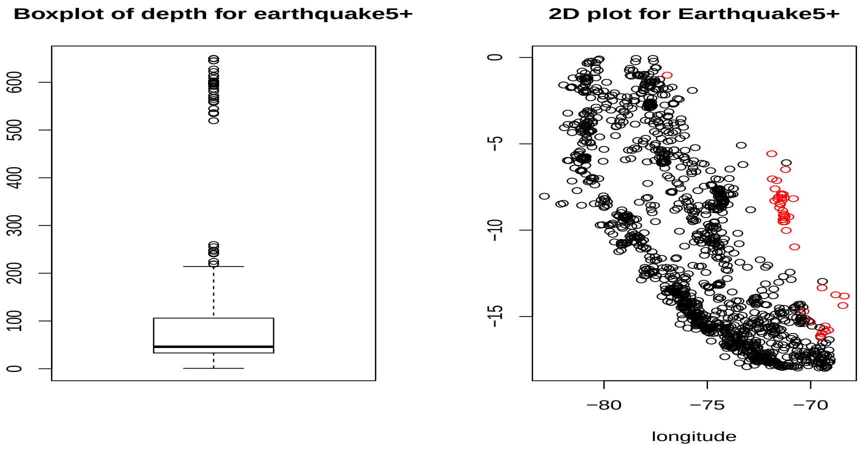

Figure 35: 2D-plot with outliers identified based on depth for Case 4

Figure 35 shows that the outliers with regard to depth in Case 4 are clustered and practically all in one region. However, the outliers with respect to magnitude (Figure 36) are scattered all over the whole pattern.

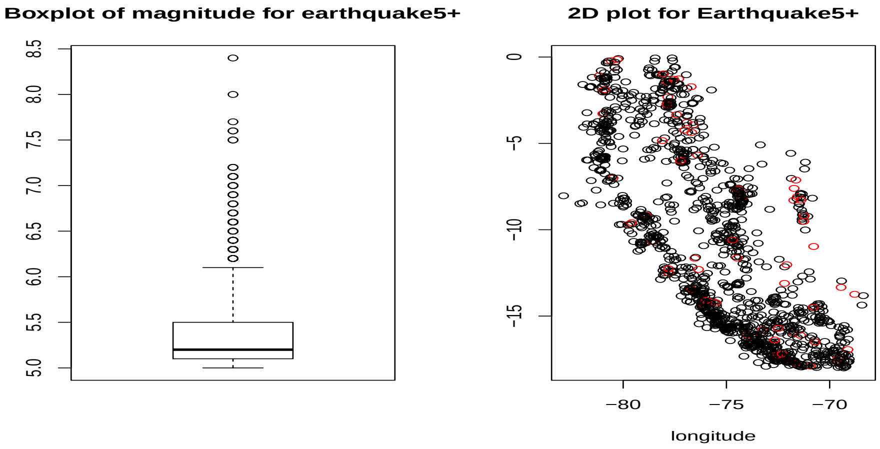

Figure 36: 2D-plot with outliers identified based on magnitude for Case 4

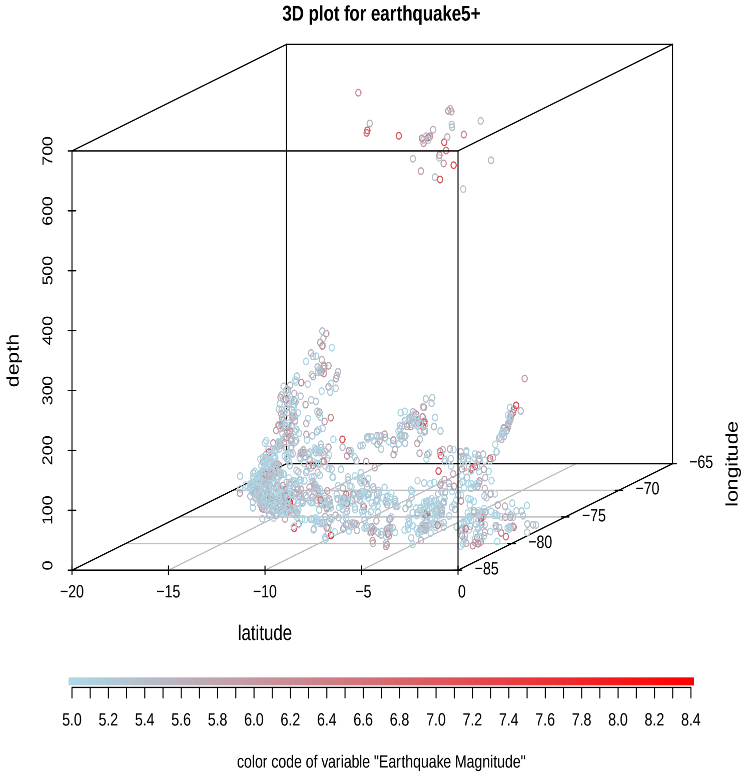

Figure 37: 3D-plot for earthquakes of magnitude 5 or more

In Figure 37, we make a 3D-plot with varying color to represent the fourth variable which is magnitude in our case to get a better perception of how the outliers are located in terms of two different associated variables.

Case 5. Earthquakes of magnitude 6 or more within certain longitude (-68°, -83°) and latitude(0°, -18°)

We can determine outliers with respect to an associated variable as those values greater than $Q3 + 1.5IQR$, Figure 38 and Figure 39 show the outliers identified with boxplot in terms of depth and magnitude respectively.

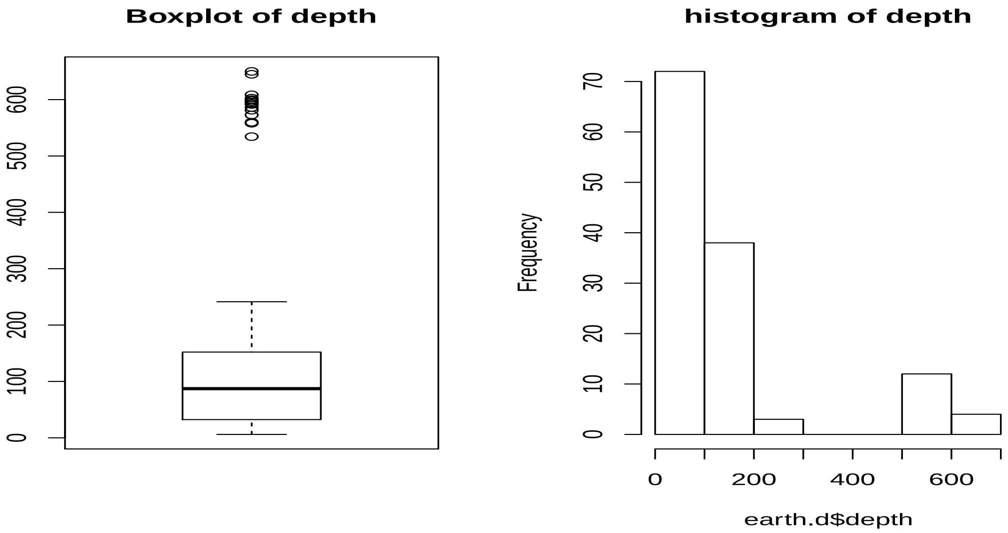

Figure 38: Boxplot and histogram of depth for Case 5

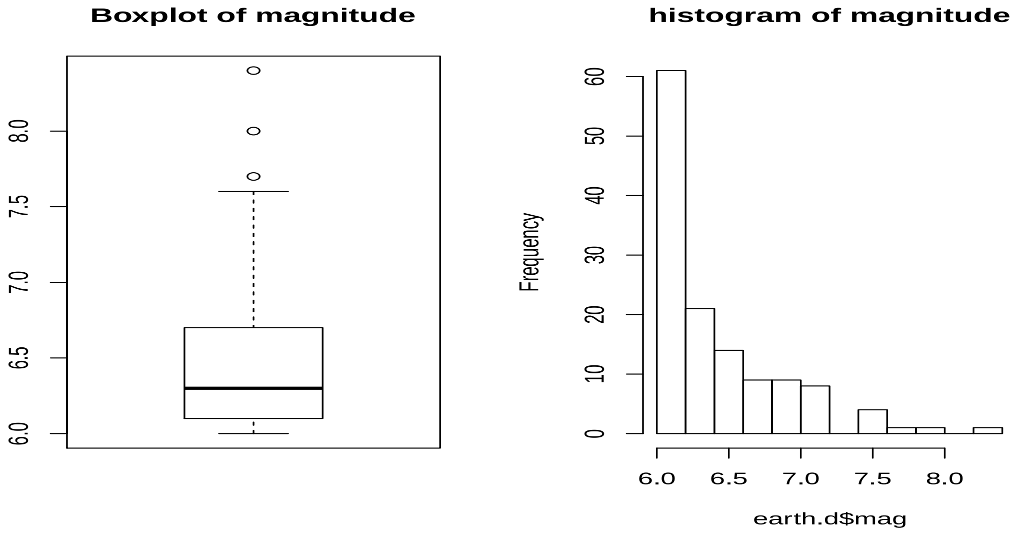

Figure 39: Boxplot and histogram of magnitude for Case 5

Notice that both variables seem to have outliers according to the usual rule applied by the boxplots. However, since these data have a spatial location, we want to see where those outliers are located.

Figure 40 and Figure 41 are 3D plots, one for depth and one for magnitude marking the points in red if they were considered to be outliers by the boxplots. In this case, both with respect to depth and magnitude, it seems that the outliers identified by the usual boxplot are clustered in one region.

Figure 42 and Figure 43 display the $\hat{G}$ for outliers with respect to depth and magnitude respectively for Case 5. Each one has three $\hat{G}$, the black one is for all the events, the green one is for the outliers only and the purple one is for complete spatial randomness obtained by simulation.

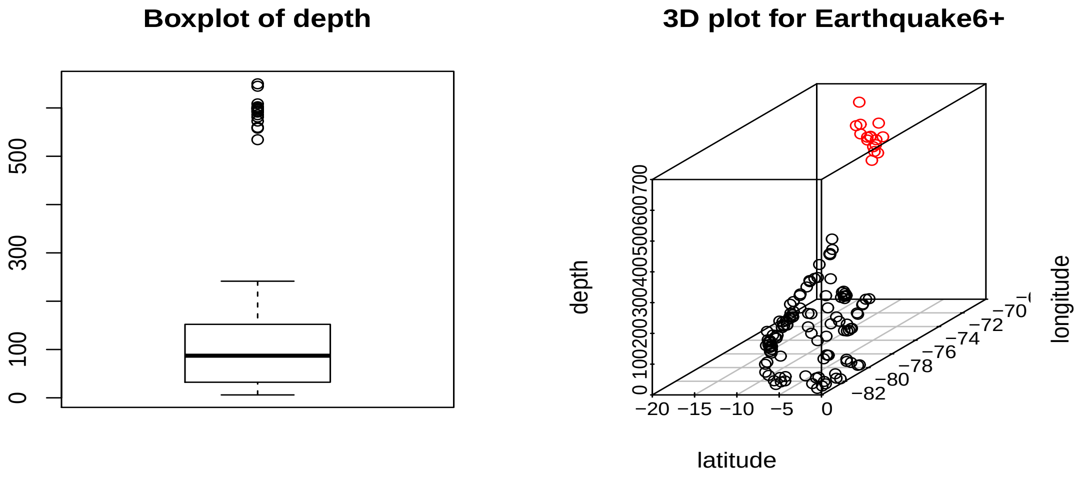

Figure 40: 3D-plot with outliers identified based on depth for Case 5

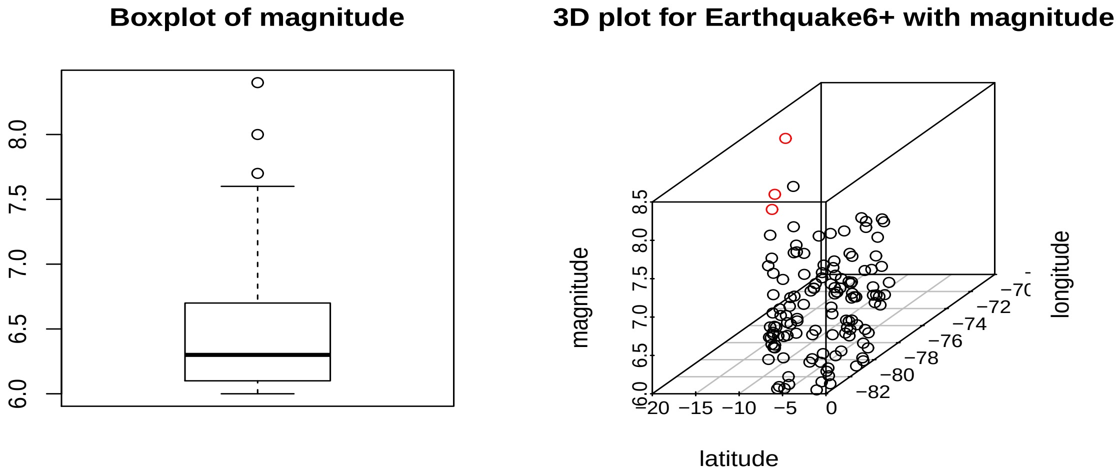

Figure 41: 3D-plot with outliers identified based on magnitude for Case 5

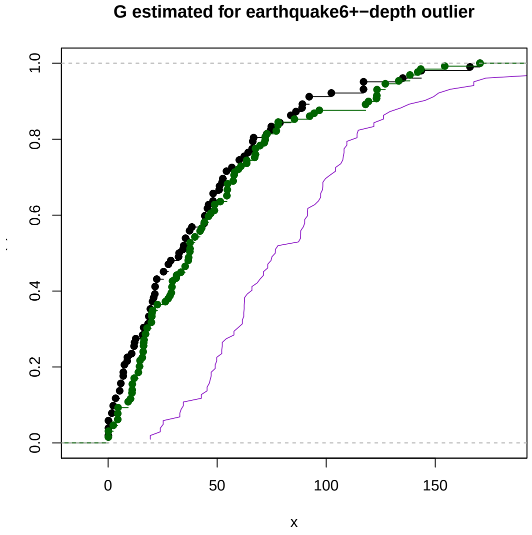

Figure 42: $\hat{G}$ with depth outliers for Case 5

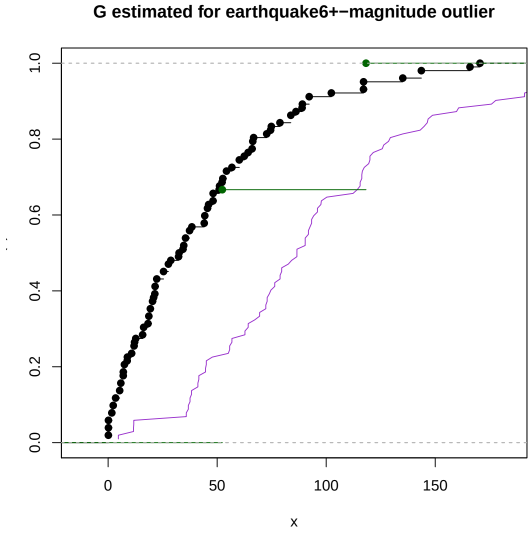

Figure 43: $\hat{G}$ with magnitude outliers for Case 5

Also, we want to make 2D plots, one for depth and one for magnitude, with location of the points and marking the points in red if they are considered as outliers to get a better view of how the outliers are located. They are shown in Figure 44 and Figure 45.

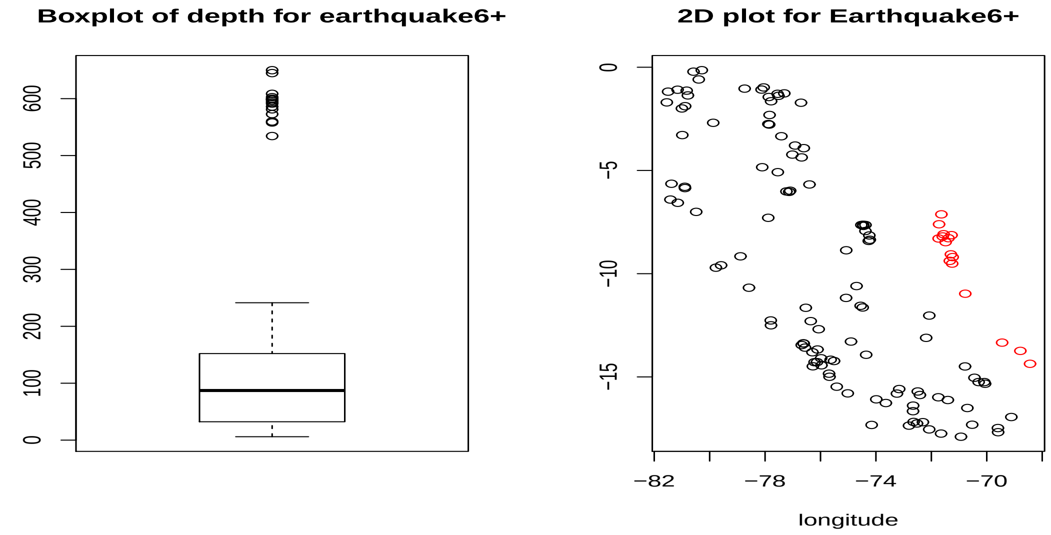

Figure 44: 2D-plot with outliers identified based on depth for Case 5

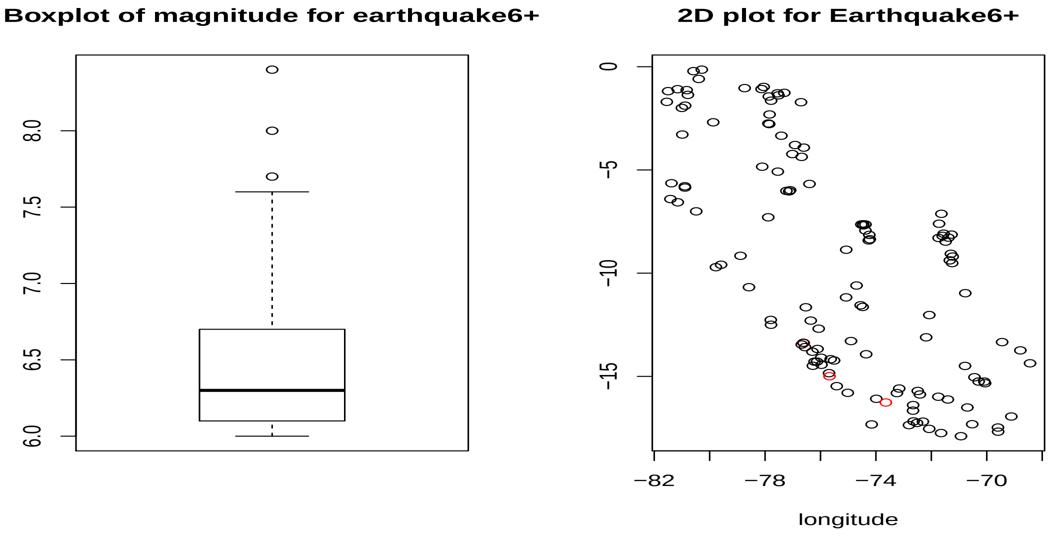

Figure 45: 2D-plot with outliers identified based on magnitude for Case 5

In Figure 46, we make a 3D plot with varying color to represent the fourth variable which is magnitude in our case to get a better perception of how the outliers are located in terms of two different associated variables.

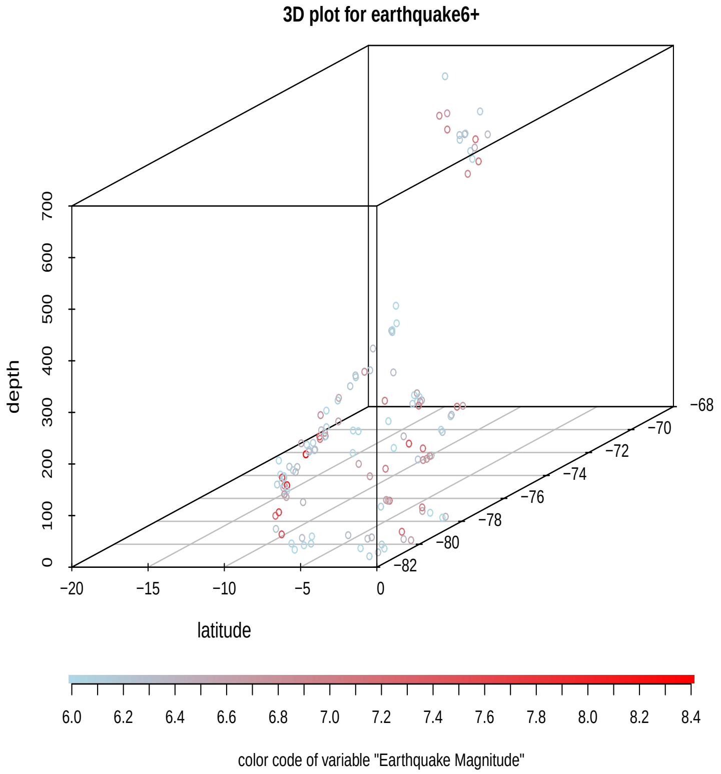

Figure 46: 3D-plot for earthquakes of magnitude 6 or more

Local Outliers

In the previous section, global outliers for one variable were identified using the usual method of the boxplot considering the values of the associated variable for all the events. Now we will address the question: Would they be outliers if we consider the values of the variable only for the events around them? In reference [11], we find the phrase ‘spatial outliers are objects which have behavioral attribute values that are distinct from those of their surrounding spatial neighbors’. We will call this ‘local outliers'. We explore the idea of defining a circle around the point (event) and see if it would be an outlier compared to the other points or events inside the circle. The definition depends on the value of the radius, and we arbitrarily choose radius (r) equals 50km for the case of the earthquakes.

For each event in the spatial point pattern, we define a circle with the event as the center and r equalling to 50km, for each event we form a subset of points that are of distance r or less from it. Because the location for earthquakes is given in longitude and latitude, distances are calculated accordingly using the *sp* [7] package that implements the great circle method in R. We calculate the average depth (without the event in the center) and compare it with the depth for all the events inside the circle, those differences are stored in an object. Also, we compare the maximum depth with the depth for all the events inside the circle and store those differences in an object so that we can later examine them. The usual criterion to identify outliers in a boxplot is applied to those differences.

Case 4. Earthquakes of magnitude 5 or more within certain longitude (-68°, -83°) and latitude (0°, -18°)

We apply two criteria, the difference between the value for the event and the mean of the surrounding events and the difference between the value for the event and the maximum value of the surrounding events. In Figure 47 and Figure 48, we make boxplots and histograms in terms of those two differences for depth and magnitude respectively.

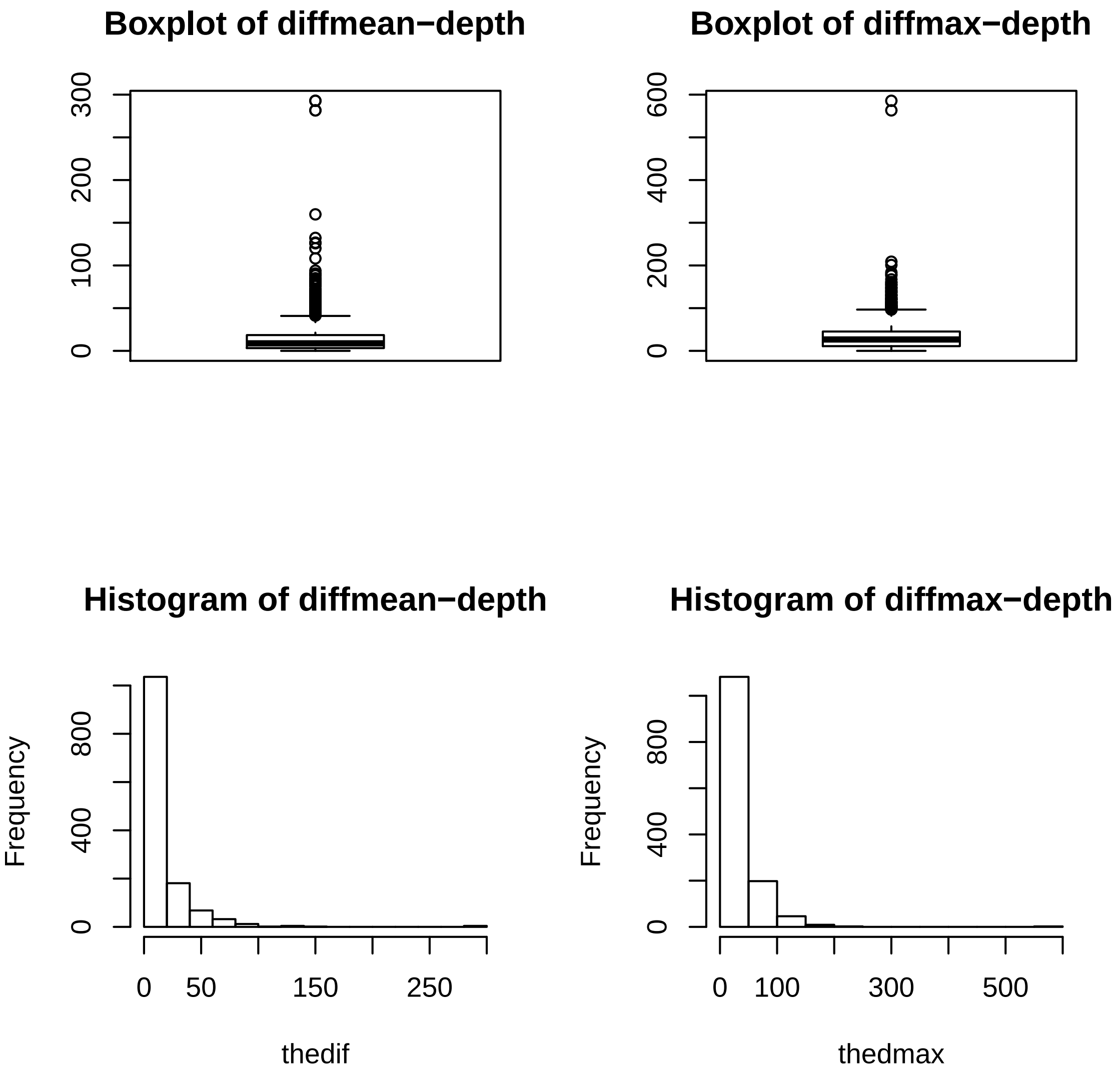

Figure 47: Boxplots and histograms of differences from mean and max for depth for Case 4

Based on Figure 47, we can see there is a number of outliers identified with the associated variable depth for Case 4 with regard to both criteria.

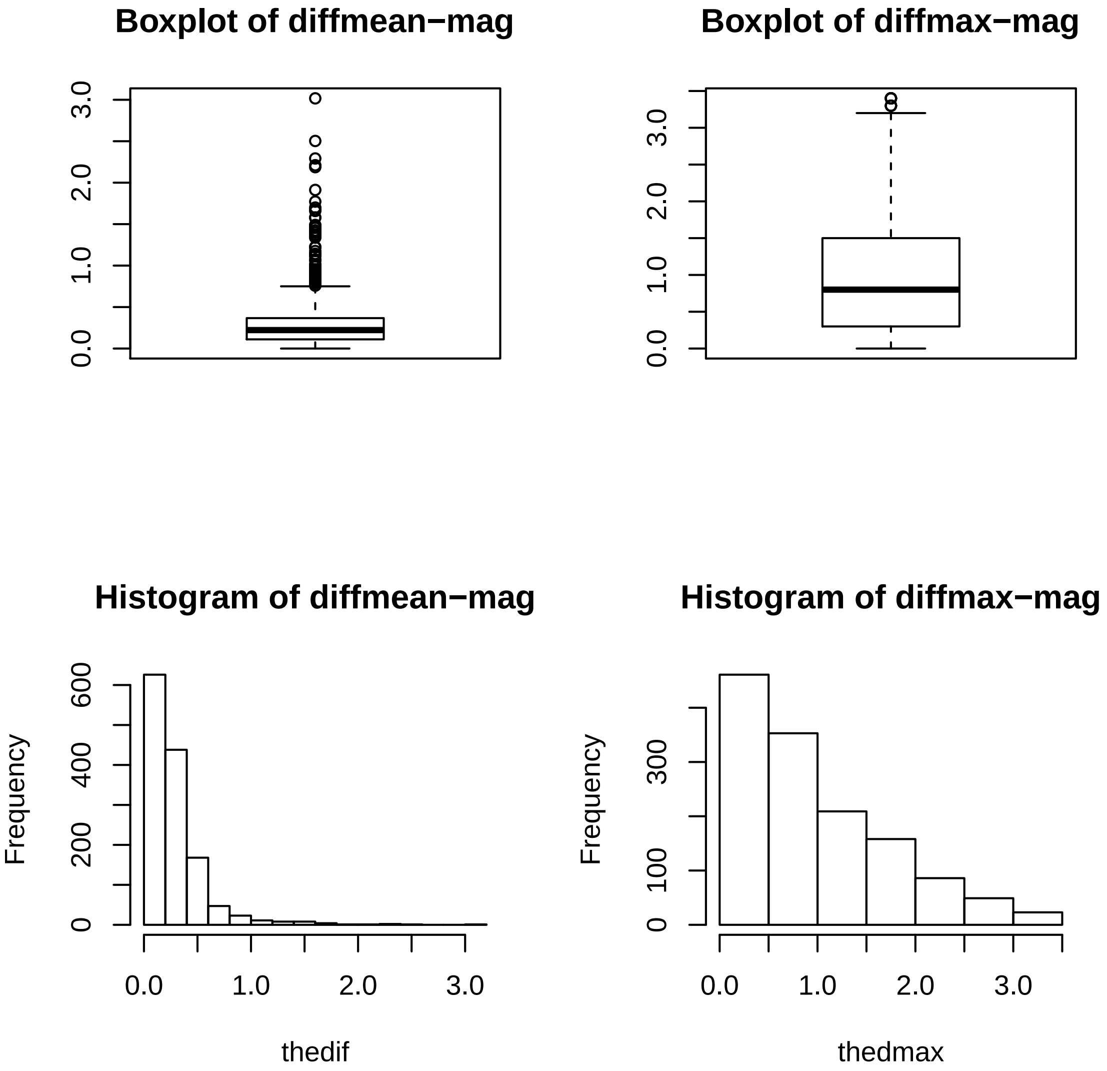

Figure 48: Boxplots and histograms of differences from mean and max for magnitude for Case 4

Based on Figure 48, we can see there is a number of outliers identified with the associated variable magnitude for Case 4 with regard to both criteria.

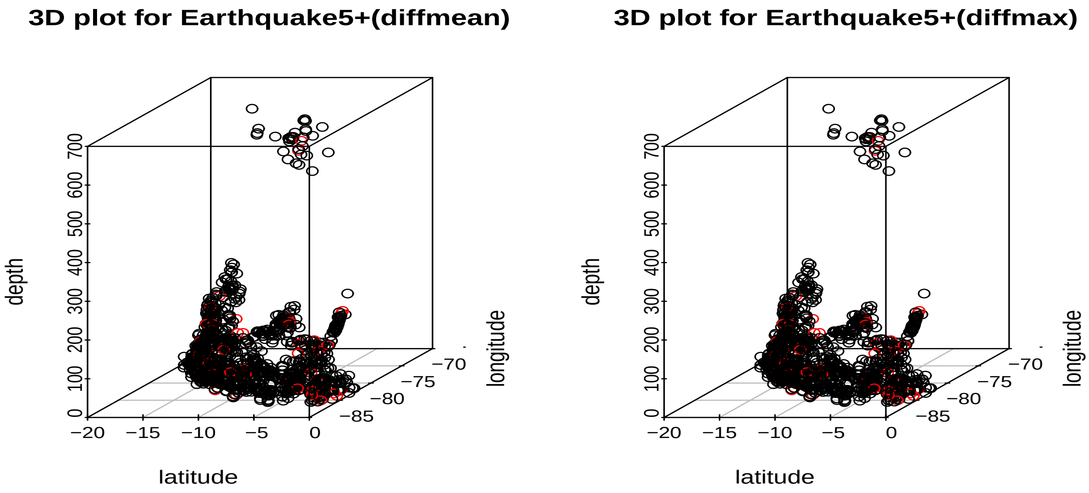

Figure 49: 3D-plot with outliers identified based on differences from mean and max for depth for Case 4

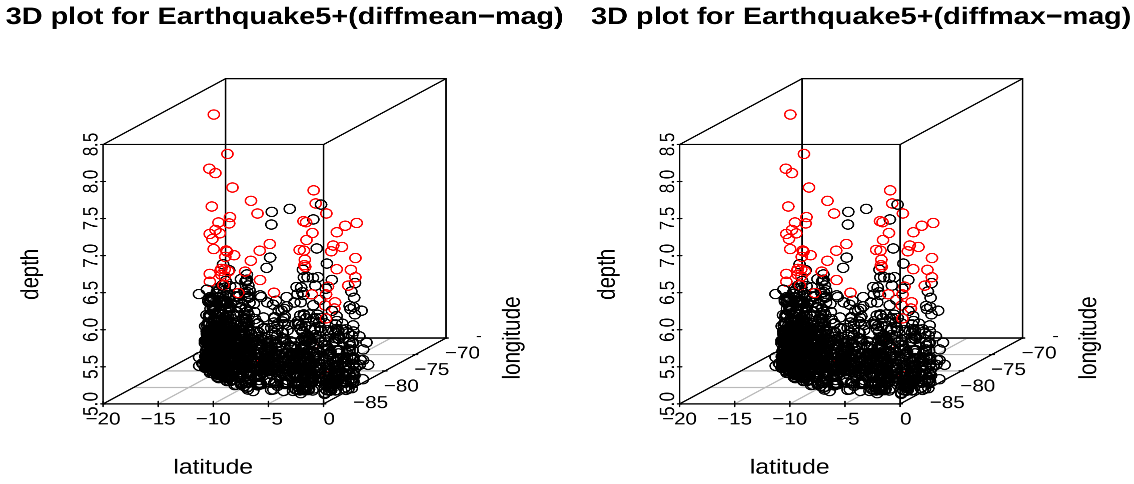

Figure 50: 3D-plot with outliers identified based on differences from mean and max for magnitude for Case 4

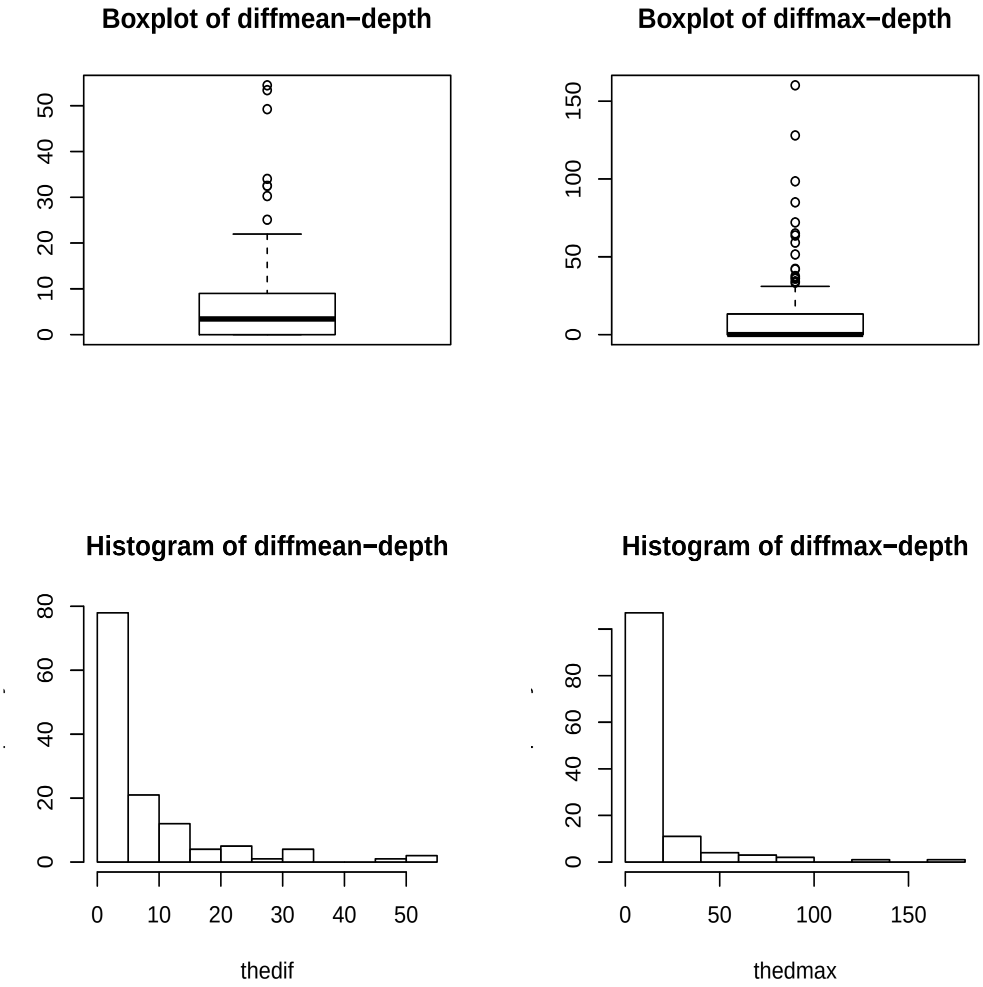

Figure 51: Boxplots and histograms of differences from mean and max for depth for Case 5

In Figure 49, 3D plots for both criteria in terms of the associated variable depth show where the local outliers are located. The local outliers are identified by both criteria are exactly the same events.

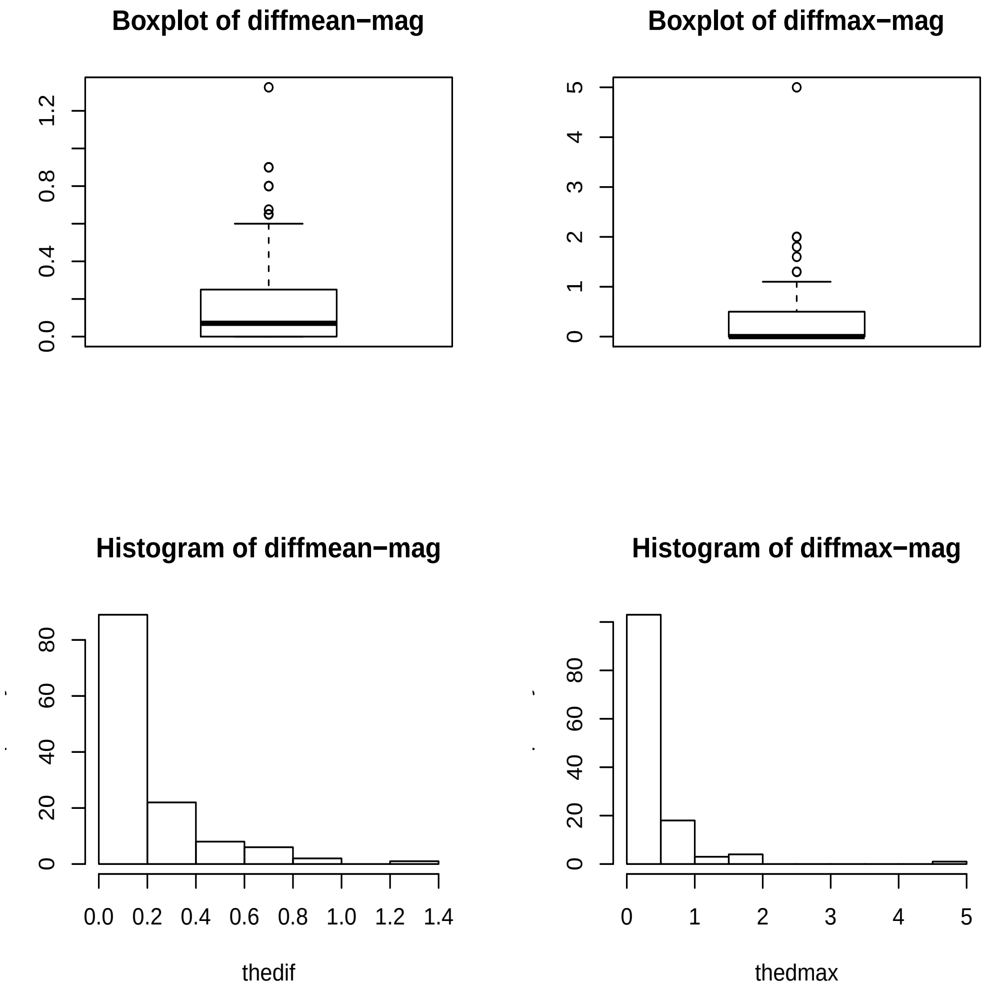

Figure 52: Boxplots and histograms of differences from mean and max for magnitude for Case 5

In Figure 50, 3D plots for both criteria in terms of the associated variable magnitude show where the local outliers are located. The local outliers are identified by both criteria are exactly the same events.

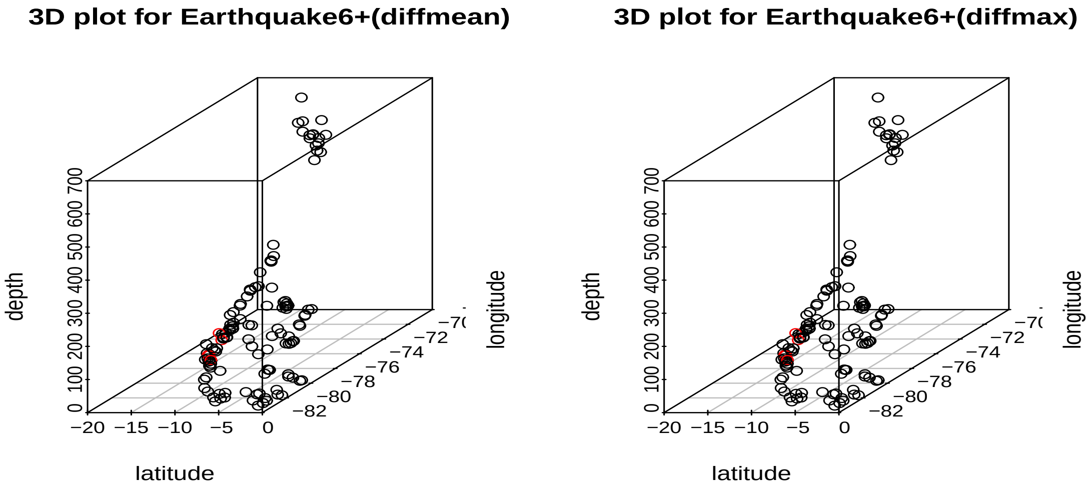

Figure 53: 3D-plot with outliers identified based on differences from mean and max for depth for Case 5

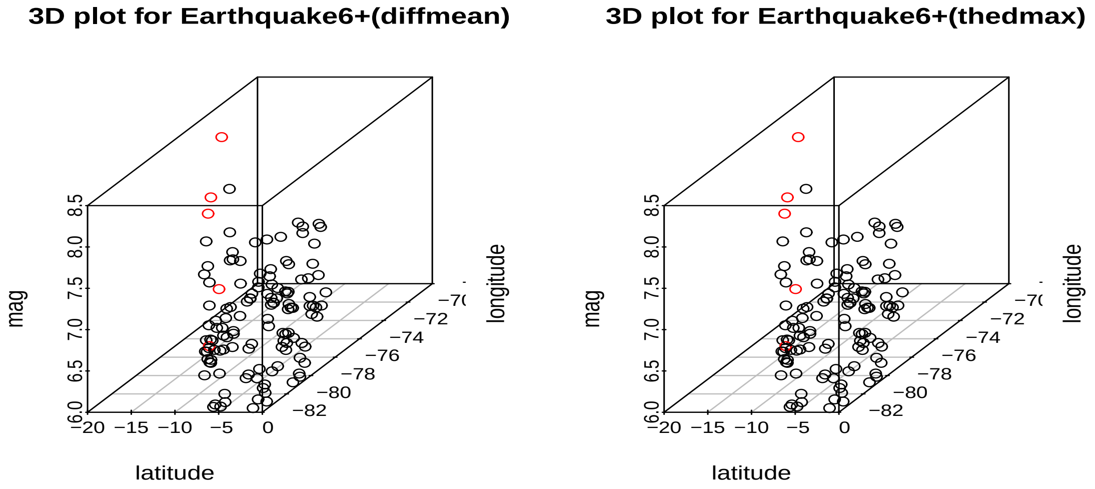

Figure 54: 3D-plot with outliers identified based on differences from mean and max for magnitude for Case 5

Case 5. Earthquakes of magnitude 6 or more within certain longitude (-68°, -83°) and latitude (0°, -18°)

Similarly to Case 4, two criteria are applied, the first one is the difference compared to the mean, and the second one is the difference compared to the maximum. Figure 51 and Figure 52 display the boxplots and histograms for the two different criteria for depth and magnitude for Case 5. A number of outliers are identified based on both criteria with regard to the associated variable depth in Figure 51. Also, there is a number of outliers identified based on both criteria with regard to the associated variable magnitude. They are displayed in Figure 52.

In Figure 53 and Figure 54, 3D plots based on both criteria with regard to the associated variable depth and magnitude are displayed. 3D plots give us a better view of where the outliers are located. The two criteria identify exactly the same outliers.

4.2 Working with Two Associated Variables

According to the reference [11], there are several attributes for each point, they should be first analyzed separately and then combined on a single deviation value. Following that suggestion we combined magnitude and depth in the following way. First the values of each variable were standardized because they are in different units and they have a different range of values, then we plot their absolute standardized values in one plot. We decided to use the criterion 'more than 3 standard deviations far from the mean' to identify outliers. That criterion could be applied in different ways. Here are three of them:

- One of the standardized values is greater than three.

- Both of standardized values are greater than three.

- The sum of the standardized values is greater than three or another arbitrary quantity.

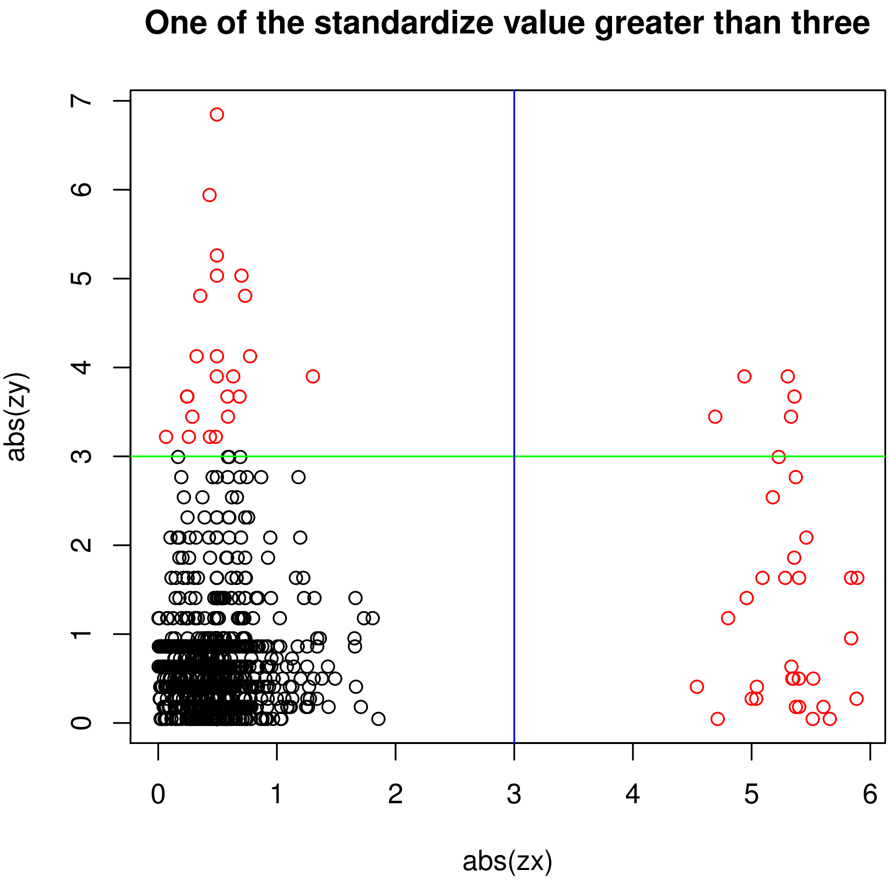

Figure 55: Plot of the standardized value of one variable >3 for Case 4

Figure 55 is an application of the first criterion for Case 4, the outliers identified would include all the points shown in red beyond either the green or the blue line.

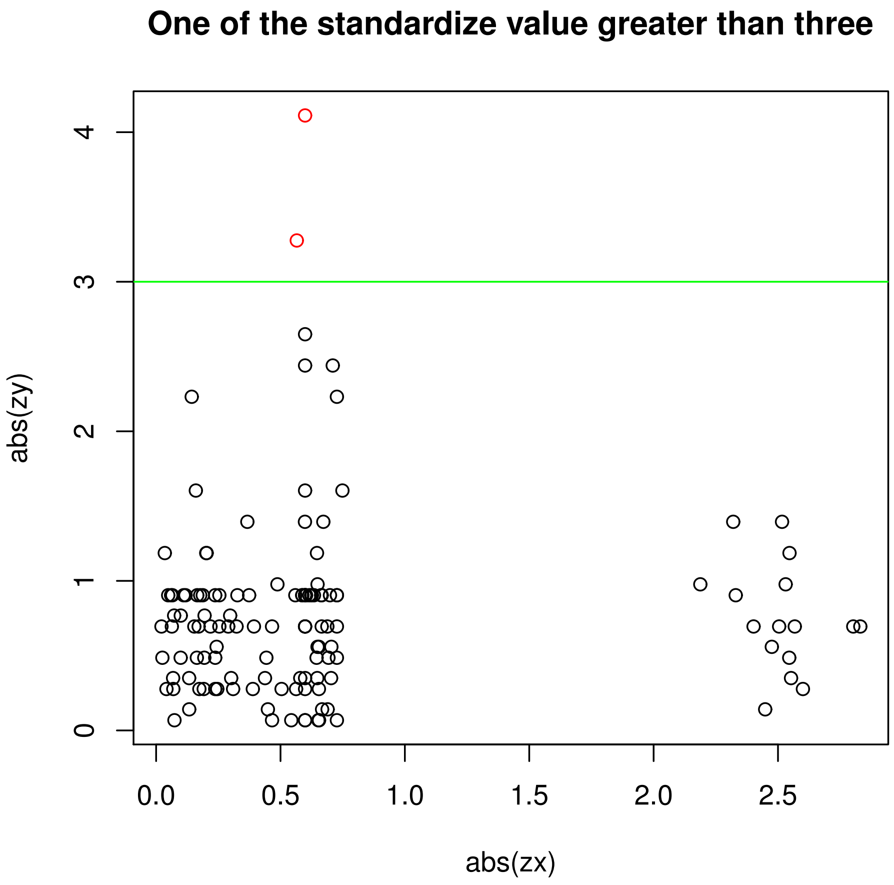

Figure 56: Plot of the standardized value of one variable >3 for Case 5

In Case 5, there is no absolute value of the standardized values of depth (zx) greater than three. Thus, only two outliers shown in red are identified in Figure 56.

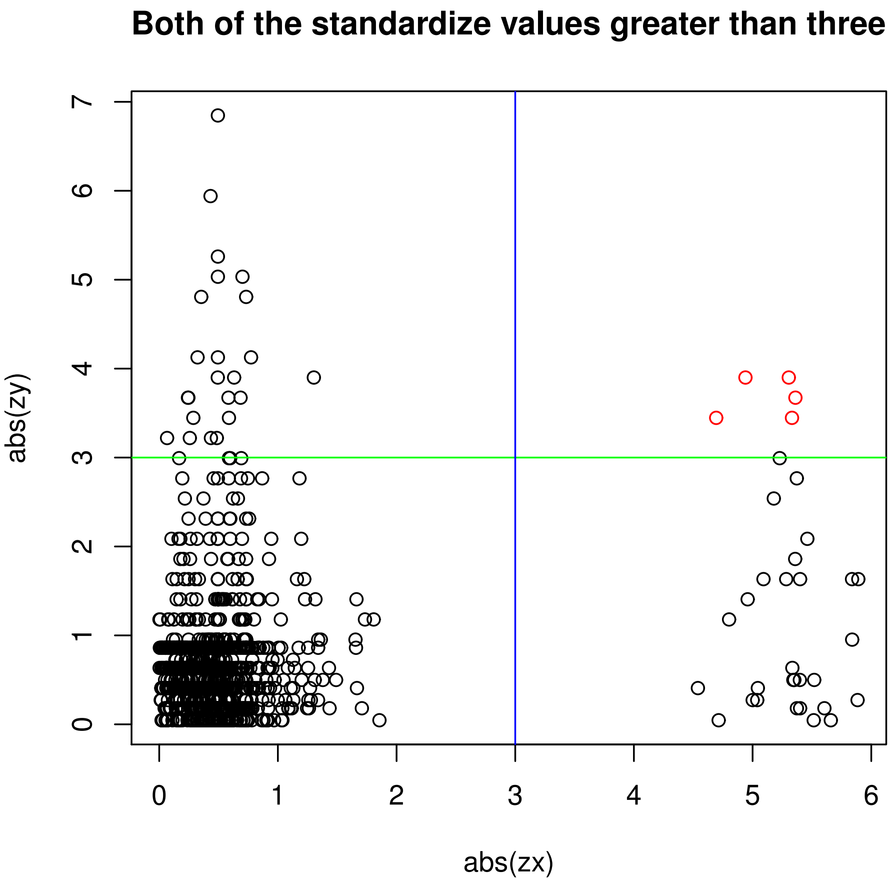

Figure 57: Plot of the standardized values of both variables >3 for Case 4

Figure 57 shows the outliers identified in red when both of the standardized values are greater than 3. In Case 4, we see there are 5 outliers shown in red on the upper right that are outliers based on this criterion.

Since there is no $abs(zx)$ for Case 5 greater than three, then there is no need to apply this criterion to Case 5.

We could define a criterion on the sum of the standardized values. One event could be considered an outlier if the sum of the absolute standardized values is greater than a quantity c. In Figures 58 and 59 $c = 3$ and $c = 6$ are used for Case 4 respectively.

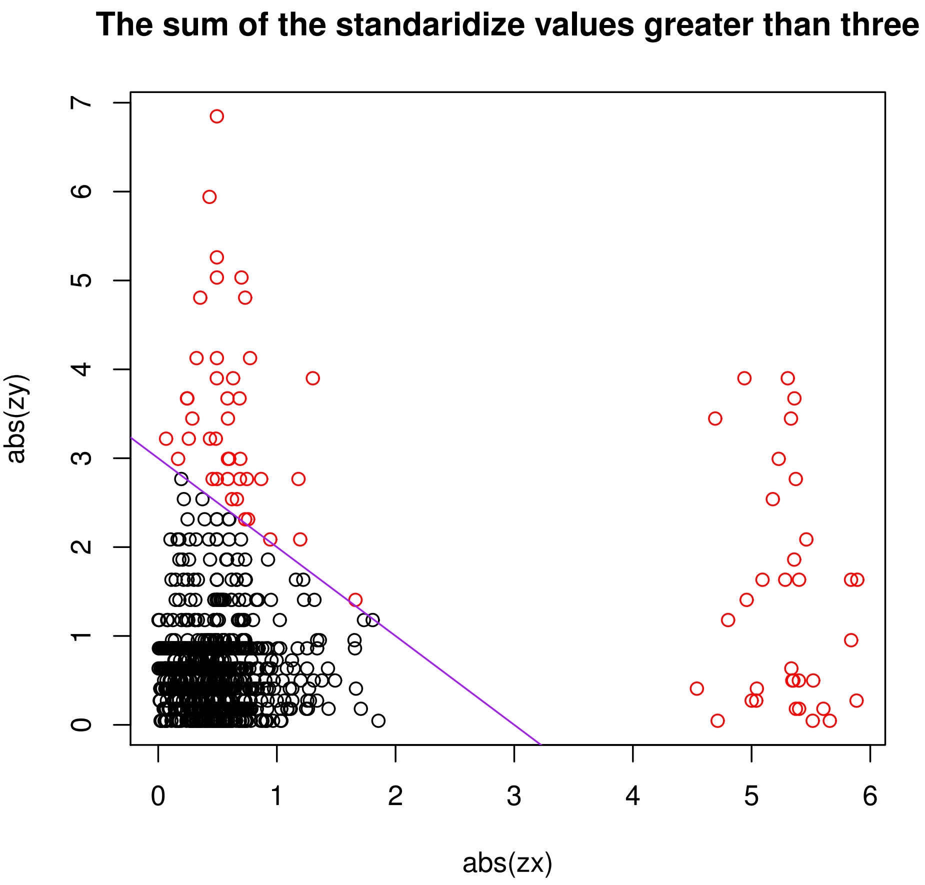

Figure 58: Plot of the sum of the standardized values >3 for Case 4

All the outliers shown in red for Case 4 identified by the criterion that the sum of the standardized values is greater than three are beyond the purple line in Figure 58.

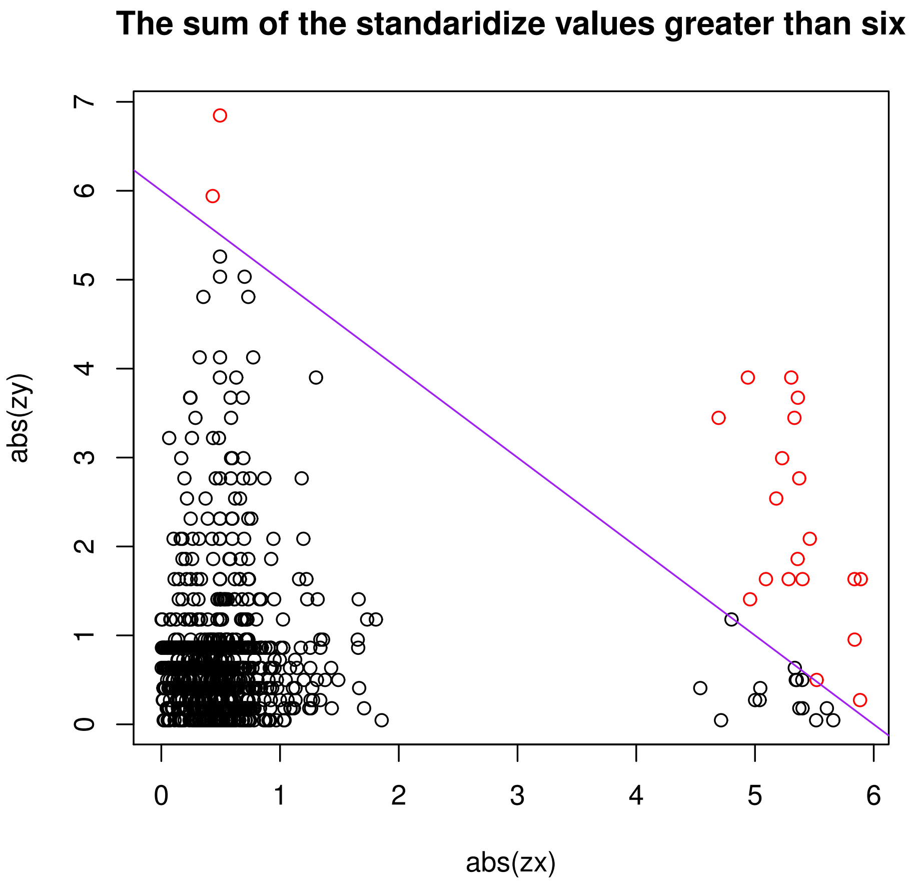

Figure 59: Plot of the sum of the standardized values >6 for Case 4

All the outliers shown in red for Case 4 identified by the criterion that the sum of the standardized values is greater than six are beyond the purple line in Figure 59. In Figure 60, the criterion ‘the sum of the standardized values greater than three’ is applied. All the outliers for Case 5 which are the events with the sum of the standardized values greater than three are beyond the purple line and shown in red.

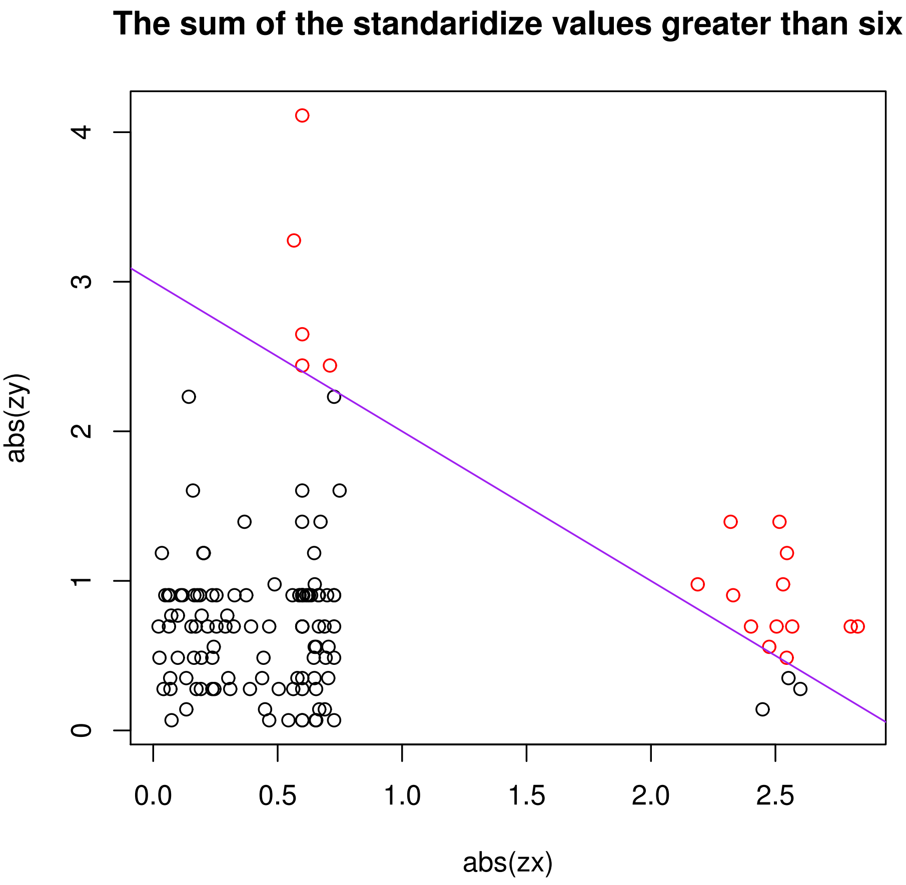

Figure 60: Plot of the sum of the standardized values >3 for Case 5

5 CONCLUSIONS

The thesis focuses on the exploration of alternative ways to detect outliers in spatial point patterns both when outliers are determined based on location alone and when they are determined based on the values of an associated variable. Several methods to identify outliers were proposed and explored for different types of spatial point patterns by using five case studies. The first two case studies are the locations of spiders in a web. The third case study is the location of waggle dances in an experimental bee hive during one day. The last two are the location of earthquakes within certain longitude (-68, -93°) and latitude (0°, -18°).

For outliers with respect to the location only, the identification of outliers was done using the distances to the nearest neighbor. We developed the idea that different methods could be applied based on the shape of the distribution of the *dnn*. If the *dnn* is fairly symmetric, the usual way of identifying outliers based on the boxplot does work well. If the *dnn* is moderately skewed, the adjusted boxplot defined in the reference [9] works well because it identify the outliers that are clearly outliers. However, when the distribution of the *dnn* is highly skewed, which means the events are highly clustered, several methods are proposed and some of them give the same results identifying the most relevant outliers. For the case of the dances of the bees in which the events are highly clustered the same four outliers are identified by the first gap method, three standard deviations from the mean of the transformed distances and the adjusted boxplot based on an exponential model. For the example of the earthquakes in which the events are clustered but not as extremely as in the bee dances case, the first gap method, the boxplot of the transformed distances and the modified boxplot based on a linear model identified the most revelent outliers.

When an associated variable is available, such as depth or magnitude of earthquakes in Case 4 and Case 5, we explored ways of identifying spatial outliers with respect to one or two associated variables. Two types of outliers were defined: global and local. The comparison of the $\hat{G}$ functions for the global outliers and all the events was used to explore whether global outliers tended to be more clustered than all the events or not, meaning whether it was more likely for outliers with respect to associated variables to happen in some regions and not in others. For the earthquakes of magnitude 5 or more, we found that the outliers identified in terms of depth are clustered, but the outliers identified in terms of magnitude are much more randomly distributed. The local outliers were determined by comparing the value of the variable at each point with the values of the variable at the surrounding points and a method based on circles of radius r was applied to identify the points with which each event was to be compared.

When working with two associated variables at the same time, we standardized the values of both associated variables since they are expressed in different units and used the criterion of three standard deviations from the mean to identify outliers for each variable. Several ways of looking at both standardized variables simultaneously were explored and compared.

Searching for outliers in spatial point patterns with respect to location only is a complex issue and it is not convenient to give one single rule to be applied to all cases. The shape of the distribution of the distances to the nearest neighbor, either fairly symmetric, moderately skewed or highly skewed needs to be taken into account when deciding the method to identify outliers.

Searching for outliers in spatial point patterns with respect to an associated variable is relatively more straight forward but it should include the discussion of the spatial distribution of the outliers for which we find the G function to be useful.

BIBLIOGRAPHY

[1] Illian, J., Penttinen, A., Stoyan, H. and Stoyan, D.(2008). *Statistical Analysis and Modelling of Spatial Point Pattern*, Wiley, New Jersey.

[2] Chen, D., Lu, C., Kou, Y. and Chen, F.(2008). "On detecting spatial outliers" *Geoinformatica*, **12**, 455–475.

[3] Earthquake hazards program. Accessed March 12, 2015.
`http://earthquake.usgs.gov/`

[4] Cucala, L. and Agnan, C. T. *Test for Spatial Randomness Based on Spacings*. Accessed November, 2004.
`http://www.math.univ-montp2.fr/~cucala/testsCSR.pdf`

[5] Knorr, E. M. and Ng, R. T.(1998). *Algorithms for Mining Distance-Based Outliers in Large Datasets*. In proceedings of 24th International Conference on Very Large Data Bases, New York.

[6] Jin, W., Jiang, Y. L., Qian W. N. and Tung, A. K.(2006). *Mining Outliers in Spatial Networks*, Springer, **3882**, 156–170.
`http://www.comp.nus.edu.sg/~atung/publication/DASFAA06Outlier.pdf`

[7] R Core Team. *R: A language and environment for statistical computing*, R Foundation and Statistical Computing, Vienna, Austria, 2012. Accessed March, 2015.
`http://www.r-project.org/`

[8] Pebesma, E., *Customising Spatial data Classes and Methods.* Accessed February, 2008.

`http://cran.r-project.org/web/packages/sp/vignettes/csdacm.pdf`

[9] Vanderviere, M. and Huber, M.(2004). *An Adjusted Boxplot for Skewed Distribution*, Physica-Verlag/Springer 2004.

[10] Zimeras, S.(2008). *Exploratory Point Pattern Analysis for Modeling Earthquake data*, 1st WSEAS International Conference on ENVIRONMENTAL and GEOLOGICAL science and ENGINEERING, Malta.

[11] Aggarwal C. C.(2013). *Outlier Analysis*, Springer, New York.

VITA

JIE LIU

Education:

B.S. Life Science, ShanDong Normal University
JiNan, ShanDong, China 2013

B.S. Biological Science, East Tennessee State University
Johnson City, Tennessee 2013

M.S. Mathematical Sciences, East Tennessee State University
Johnson City, Tennessee 2015

Professional Experience:

Graduate Assistant , East Tennessee State University
Johnson City, Tennessee, 2013–2015