Introduction to Point Pattern Analysis

Erum Tariq

Math 402

Spring 2004

Abstract

Point Pattern Analysis is a class of techniques that endeavor to identify patterns in spatial data. We utilize the Quadrant Count Method as an introductory algorithm of point pattern analysis. While this algorithm is very simplistic we do uncover a few mathematical and statistical subtleties. We will apply this method to the tree data that is taken from a survey carried out by D.J. Gerrard on a 19.6 acre square plot in Lansing Woods, Michigan. We attempt to classify clustering, or lack of clustering, of hickory tress in the survey area.

Introduction

The solutions to many mathematical questions, both pure and applied, rely on the ability of the investigator to uncover a pattern. In basic terms, Point Pattern Analysis is an investigation focused on finding patterns in data comprised of points in a spatial region. One common application of Point Pattern Analysis is epidemiology. The medical community is often interested in the spread of infectious disease such as: SARS, chicken pox, and West Nile virus among others. It is possible to identify pattern to the spread of infection then this might lead to an understanding of how the spread of an illness is related to social behavior, environmental factors, genetic susceptibility, or many other health care factors.

In general, a spatial data set takes the form: $X=\{x_k \mid x_k \ \varepsilon \ R^{m,} \ m \in N\}$. However, it is possible for the data to contain spatial location plus additional information. For example, earthquake data typically gives the location of earthquakes along a fault line and will often have the size and the time of each earthquake. Data that contains spatial data plus additional information is often referred to as *marked spatial data*. In our analysis, we will be concerned with only the spatial information and we will disregard any additional information associated with the data. Moreover, the examples we will work with are limited to two-dimensional data.

Our interest will lie in quantifying the dispersion of objects within a confined geographical area. We try to understand the interaction of pattern and process and use point pattern analysis as a mechanism for detecting patterns associated as compared to random processes. The random process that will serve for our comparison will be the homogenous Poisson process, which will be described in more detail in section 2.

D.J Gerrard describes an investigation of a 19 .6 acre square plot in Lansing Woods, Michigan [3]. This data includes hickories, maples and oaks grown on a square plot. The data for hickories is given in Cartesian coordinates, that is, $(x_i, \ y_i)$ form, where $x_i \ and \ y_i \in R$. Also, the points are plotted on a unit square region.Our main goal of Point Pattern Analysis is to find out whether the distribution of the hickory trees is random, clustered or regularly dispersed. The kind of pattern involved would further our understanding of the behavior of the hickory trees and thus can be of great use to ecologists and biologists. For example, if the pattern is clustered, the biologists may conclude that natural factors encourage the hickories to cohabitate and promote tree growth.

There are several methods and algorithms that endeavor to describe pattern for a collection of points. The most common methods discovered for spatial pattern analysis are as follows: [1]

1. Quadrant Count Method
2. Kernel Density Estimation (K means)
3. Nearest Neighbor Distance
    a. G function
    b. F function
    c. K function

The above list of techniques for Point Pattern Analysis is among the most popular and best established mathematical and statistical methods used in the literature [1,2,4,5]. Since Point Pattern Analysis can take several forms and can be applied in a variety of settings we will present a list of criteria in order to determine if a data set is suitable for our Point Pattern Analysis.

The criteria we will use to determine if a data set is appropriate for our type of point patter analysis is given by the following:

- Spatial data must be mapped on a plane; both latitude and longitude coordinates are needed.
- The study area must be selected and determined prior to the analysis.
- Point data should not be a selected sample, but rather the entire set of data to be analyzed.
- There should be 1-1 correspondence between objects in study area and events in pattern.
- Points must be true incidents with real spatial coordinates.

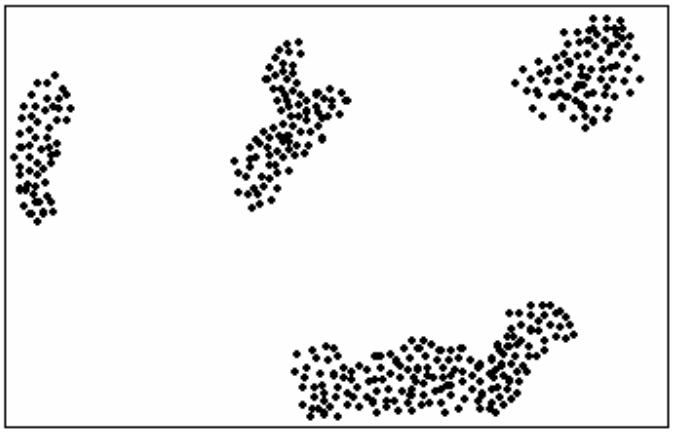
Since this is an introductory venture into the subject of Point Pattern Analysis we have selected the Quadrant Count Method for our analysis. While the other techniques listed in this section can be more descriptive and more accurate it is also true that many of these other methods are more complicated and difficult to implement. We have found that Quadrant Count Analysis is relatively easy method to implement and it has provided several opportunities to apply basic mathematical and statistical concepts.

Quadrant Count Method

The Quadrant Count Method can be described simply as partitioning the data set into *n* equal sized sub regions; we will call these sub regions quadrants. In each quadrant we will be counting the number of events that occur and it is the distribution of quadrant counts that will serve as our indicator of pattern. The choice of the quadrant size can greatly affect our analysis, where large quadrants produce a coarse description of the pattern. If the quadrant size is too small then many quadrants may contain only one event or they might not contain any events at all. We will use the rule of thumb for the area of a square is twice the expected frequency of points in a random distribution (i.e., $2\frac{Area}{n}$), where *n* is the number of points in the sample size. After partitioning the data set into quadrants, the frequency distribution of the number of points per quadrant has been constructed. The Mean and Variance of the sample are then computed to calculate the Variance-to-Mean Ratio (VTMR). The following is the way we will interpret the VTMR of a sample:

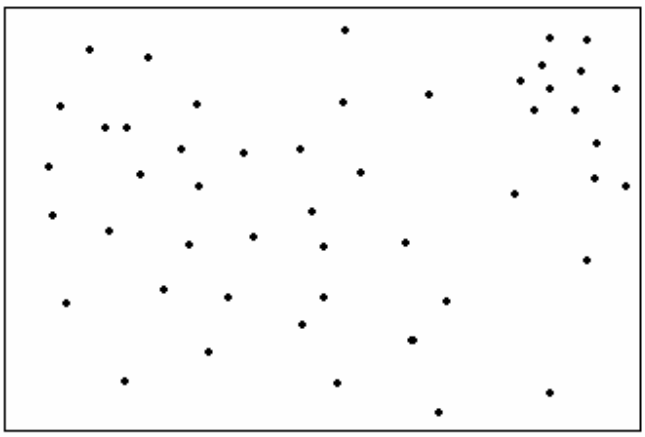
- If VTMR>1, the pattern is clustered. This implies that the data set has one or more groups of points in clusters and large areas of maps without points. The region might look like Figure 1:

Figure 1 A Clustered pattern



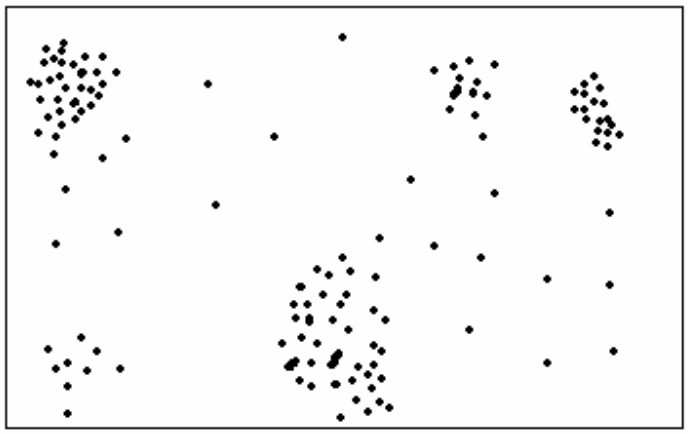
- If VTMR<1, the pattern is regularly dispersed implying the events are distributed more or less regularly over the region. A regularly dispersed area might look like Figure 2:

Figure 2 Regularly Dispersed Pattern



- If VTMR=1, the pattern is random. This implies the data set has no dominant trend towards clustering or dispersion. A random pattern may look like Figure 3:

Figure 3 Random Pattern



The random model that will serve as our standard of comparison is the *Complete Spatial Randomness* (CSR) model [1,3]. The CSR model has two basic characteristics:

(1) The number of events in any planar region *A* is with area |*A*| follows a Poisson distribution with mean λ |*A*|.

(2) Given there are n events in *A*, those events are independent and form a random sample from a uniform distribution on *A*.

The constant λ is the intensity, or the mean number of events per unit area. Also, by (1), the intensity of events does not vary over the plane. According to (2), CSR also implies, the events are independent of each other and there is no interaction between them.

The mathematical construct that we will use to simulate a CSR model is the homogenous Poisson process. The Poisson process is suitably defined by the following postulates:

(a) If λ>0, and any finite planar region *A*, *N (A)*, follows a Poisson distribution with mean λ |*A*|.

(b) Given *N (A)=n*, the *n* events in *A* form an independent random sample from the uniform distribution on *A*. (In our case, n is the number of trees in the region)

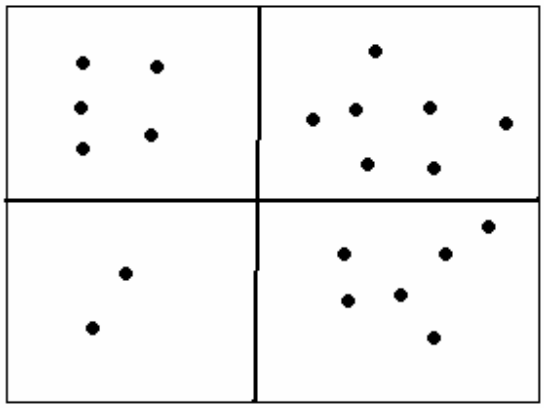
As stated above, the CSR corresponds to the homogenous Poisson Distribution. Please recall that the Poisson distribution can be used in place of the Binomial

distribution in the case of very large samples. In a binomial distribution, all eligible phenomena are studied, whereas in the Poisson distribution only the cases with a particular outcome are studied. The Poisson distribution can also be used to study how "events" are distributed on the level of a population. If having one "event" has no influence on the chance of having another accident, the "event" is put back into the population immediately after an "event"; people may have one, two, three, or more events during a certain period of time. One assumption in this application of the Poisson distribution is that the chance of having an event is randomly distributed: every individual has an equal chance. Mathematically, this is expressed in the fact that for a Poisson distribution, the variance of the sample is equal to its mean. Hence, in the QCM, the VTMR of a random sample is equal to 1, since the variance is equal to its mean.

How to apply Quadrant Count Analysis

To explain the application of QCA in detail, a small data set is plotted on a square region. The region is then divided into equally sized quadrants (squares). This is demonstrated in the following figure. For the sake of simplicity, we have chosen to divide the region into only four quadrants as illustrated in Figure 4.

Figure 4 Dividing a region into quadrants



The Mean of the sample can then be calculated as:

$$Mean = \frac{No.\ of\ \ pts.in the region}{No.of\ \ quadrants} = \frac{20}{4} = 5\,.$$

Let x_i be the frequency of points in each quadrant. Then Variance of the sample can the be calculated as

$$Variance = \frac{\sum x_i{}^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{2^2 + 5^2 + 6^2 + 7^2 - \frac{(20)^2}{4}}{4-1} = 4.5\,.$$

The Variance to Mean Ratio or VTMR is calculated as

$$VTMR = \frac{Variance}{Mean} = \frac{4.5}{5} = 0.9 .$$

According to QCM, since VTMR for this example is less than 1, our simple data set is classified as regularly dispersed. Thus the points tend to repel each other and are thought to spread evenly throughout the region.

Using QCM to analyze the hickories

To apply Quadrant Count Analysis to our data, we used a code written in C++. We experimented with different quadrant sizes, using $4^m, m \in N$ as our number of quadrants. This choice of number of quadrants made it easier to experiment with different number of quadrants. Table 1 displays the results generated with different quadrant sizes.

Table 1 VTMR for different quadrant sizes

N	Grid Size	Mean	Variance	VTMR
1	2*2	175.75	3808.33	21.669
2	4*4	43.9375	529	12.0398
3	8*8	10.9844	55.6032	5.06202
4	**16*16**	**2.74609**	**6.52941**	**2.37771**
5	18*18	2.16975	4.29721	1.98051
6	19*19	1.94837	3.57778	1.83724
7	32*32	0.686523	0.939394	1.36833
8	64*64	0.171631	0.188767	1.09884
9	128*128	0.0429077	0.0443753	1.0342
10	256*256	0.0107269	0.0109865	1.0242

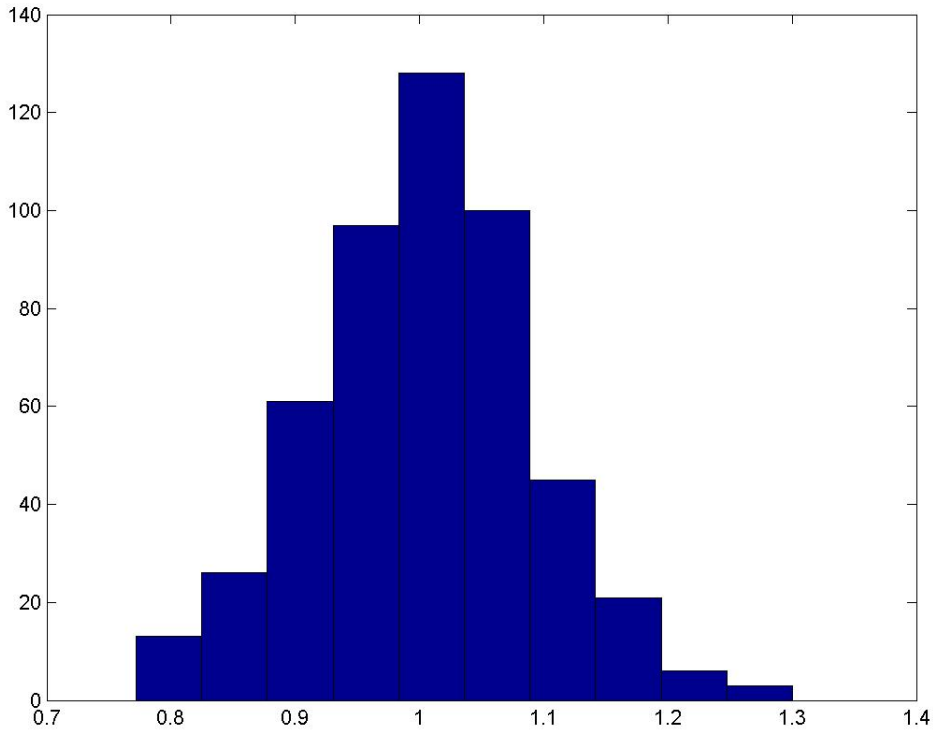
As we test different quadrant sizes, we notice the smaller number of quadrants corresponds to larger variance. But as we divide the region into smaller quadrants, the variance starts decreasing to one.

We chose to divide the hickories into 256 quadrants (16*16 grid) as our experimenting size. We select this number since it is the power of 4 that is to the rule of thumb: number of quadrants= $2\frac{Area}{n}$. For 256 quadrants the VTMR was found to be approximately 2.37771. According to the QCM, a VTMR that is greater than 1 implies that the hickories must be clustered. However, it is possible of a data set that is randomly generated would also have a VTMR that is greater than one when counting over 256 quadrants. How confident can we be that our 2.38 VTMR does indeed correspond to the clustering pattern of our hickories? The answer lies in using the homogeneous Poisson process to simulate random data for the purpose of comparison.

Simulations

To be more confident about our hickories being clustered, we generate 500 data sets of 703 random points each. Their VTMR is then calculated and a histogram of the 500 VTMR is built. This histogram is illustrated in Figure 5.

Figure 5 Histogram variance to mean ratios of the 500 random simulations



We will use the empirical p-value of our observed value of VTMR (2.37771) in correspondence with our random simulations. Please recall the empirical p-value is the percentage of the VTMR from the random samples that are greater than or equal to our observed VTMR. Hence, a small the empirical p-value implies that we can be more confident that our data set is clustered. It turns out that none of the VTMR of the random samples are equal to or greater than our calculated VTMR. Therefore, our Empirical p-value is 0 and we can be highly confident that our data set of hickories is clustered.

To illustrate the utility of the empirical p-value consider the hypothetical case that we had a set of tree data that had a VTMR is very close to 1, say 1.01. The Quadrant Count Method would classify this data set as clustered, but Figure 5 shows that random data could also generate a VTMR that is larger than one. If we utilize the same collection of 500 random data sets and calculate the empirical p-value that corresponds to an observed VTMR of 1.01 then our empirical p-value is approximately .58. This value implies more than half the random simulations have a VTMR greater than or equal to

1.01. Thus, we cannot be confident that data set with a VTMR of 1.01 is in reality clustered.

Conclusion

We have introduced and applied the Quadrant Count Method to analyze the data set of hickories for a pattern. Our analysis showed that the hickories were classified as clustered. By using the empirical p-value and the random model we were able to provide strong support that the hickories are certainly clustered. There are a few things that could be done to improve our methods and our analysis. The empirical p-value is a very simple tool, but it is not necessarily a rigorous test. Another statistical method for validating the correct classification of clustered and would be the Goodness of Fit Test. Moreover, it has been mentioned that the Quadrant Count Method is not the most accurate method for identifying clustering. The Nearest Neighbor Analysis would have been a better choice of method, but we did not have time to complete this type of analysis.

Bibliography

1. Cressie, Noel A., "Statistics for Spatial Data (revised edition)", John Wiley & Sons, 1993.

2. Dale, Mark R. T., "Spatial Pattern Analysis in Plant Ecology", Cambridge University Press, 1999.

3. Diggle, Peter J., "Spatial Analysis of Spatial Point Patterns", 2nd Edition, Arnold Publishers, 2003.

4. Stoyan, Dietrich, and Helga Stoyan, "Fractals, Random Shapes and Point Fields (Methods of Geometrical Statistics)", John Wiley & Sons, 1994.

5. Upton, Graham J.G., and Bernard Fingleton, "Spatial Data Analysis by Example (Point Pattern and Quantitative Data)", volume 1, John Wiley & Sons, 1985.