An Application of Spatial Point Process

Kalanka Jayalath[*] Richard F. Gunst [†] David J. Meltzer [‡]

Abstract
Identifying spatially distributed point patterns plays an important role in many scientific areas including pattern recognition, computer vision, image processing and some geological applications. This research is focused on applying spatial point process theories to analyze suspected prehistoric house structures belongs to Pleistocene people. Various statistical methods such as quadrat method, cluster indices and Ripley's K function are used to test the complete spatial randomness of rock locations of the excavated sites. A variation of the K-function known as the L-function is used to compare the clustering structures identified in few different sites and the existence of small scale regularities of those sites is also discussed.

Key Words: Cluster Indices, K-function, L-function, Quadrat

1. Spatial Point Processes

Spatial data are important in a wide range of scientific areas, including agriculture, biology, epidemiology, astronomy, physics and geology. Much of the emphasis in these fields concentrated on the analysis of spatially correlated measurements and the development of statistical methods for such analysis. Spatial point processes are different from measurements taken at locations in a spatial domain. Spatial point processes are locations that are characterized as realizations of a random process. Common examples of spatial point processes are locations of trees in a forest, ant nests in a field, and copper sites at a mineralogical site. A realization of a point process is an unordered set of locations.

1.1 Archeological Mountaineer Site

Anthropology Professor David Meltzer and his research team excavated an archaeological site on the western slope of the Rocky Mountains. They selected three different archaeological sites, one of which is suspected of being the ruins of a late Pleistocene age (10,400 years BP) house. The rocks pattern in the main site, Block C, is the main interest of the study. The other two are control sites, named Block X and Block Y. Block C has 3762 rocks in a rectangular sampling window $13 \times 9\ m^2$. Rocks at these sites were categorized as large or small (this categerization was done by Prof. David Meltzer) depending on the longest surface length. If the longest surface length was greater than 30 cm, then a rock was categorized as large and otherwise as small. In Block C there are 282 large and 3480 small-rocks available for analysis.

The same categorization was applied to the control sites. In Block X there are 17 large-rocks and 776 small-rocks within a rectangular sampling window $5 \times 3\ m^2$. In Block Y there are 33 large-rocks and 231 small-rocks within a rectangular sampling window $4 \times 3\ m^2$. The spatial distributions of the rocks at these sites are shown in Fig. 1.

[*] Department of Mathematics and Statistics, Stephen F Austin State University, P.O. Box 13040 SFA Station, Nacogdoches, TX 75962-3040, USA. Fax +1 936 468 1669, Email: jayalathk@sfasu.edu

[†] Department of Statistical Science, Southern Methodist University, P.O. Box 750332, Dallas, TX 75275-0332, USA

[‡] Department of Anthropology, Southern Methodist University, P.O. Box 750336, Dallas, TX 75275-0336, USA

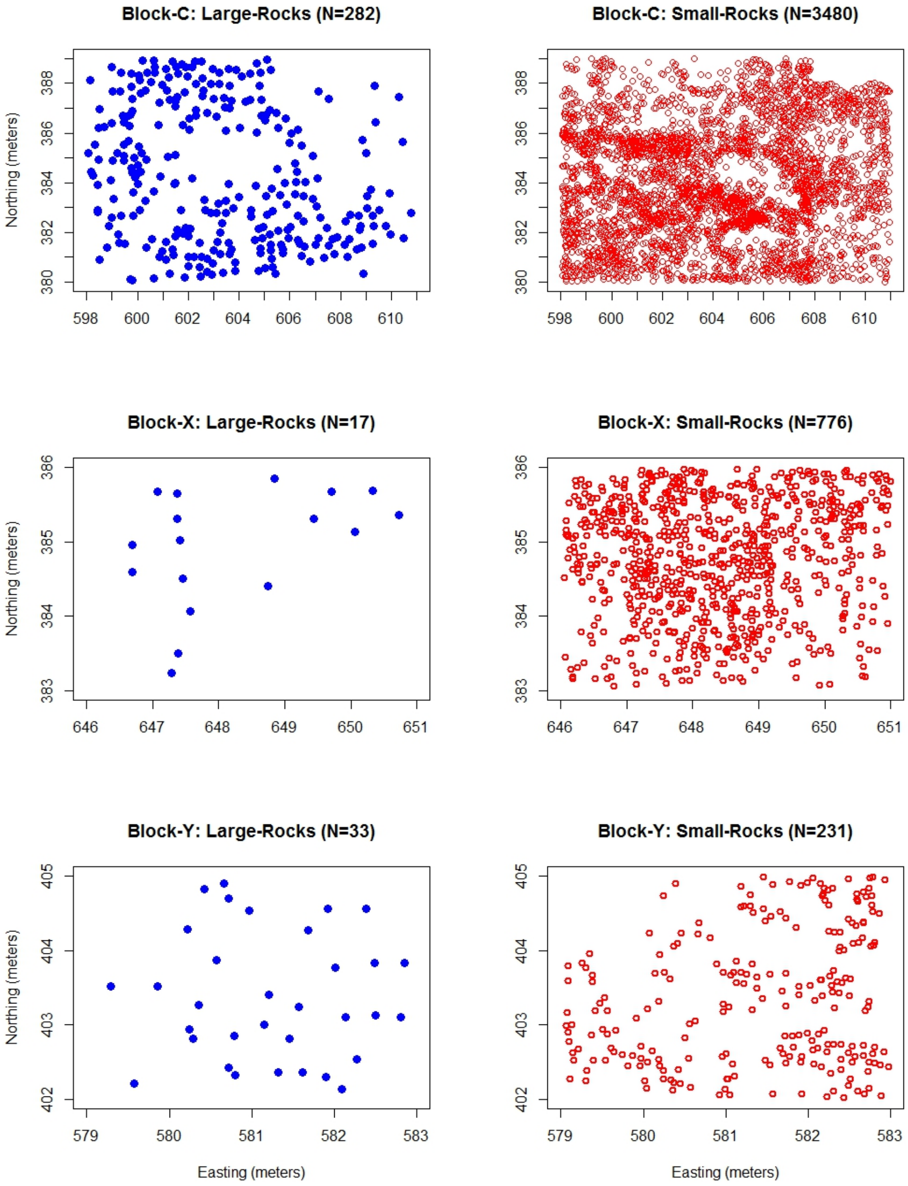


Figure 1: Large- and Small-Rocks in Mountaineer Site

One hypothesis of interest in this investigation is whether the locations of rocks in a supposed region of human activity would form an identifiable pattern whereas rocks in regions not believed to be influenced by recurring human activities would not form an identifiable pattern. In particular, a hypothesis of interest in whether the large-rock locations in Block C might form a pattern due to the supposed existence of a prehistoric house. A secondary hypothesis is that the pattern of the large-rock locations within this block is circular. In this research statistical methods are used to investigate whether the locational pattern in the large-rock locations is not reasonably attributable to spatial randomness.

1.2 Intensity Function Estimates

Intensity is the average density of events (e.g., rocks in the Mountaineer Site data), or the expected number of events per unit area in the study region, and it is denoted by λ. Intensity characterizes homogeneous and inhomogeneous spatial point processes depending on whether intensities are constant or non-constant from location to location, respectively.

Let E(X) be the expectation of a random variable X, $N(W)$ the number of points in a region W, $|W|$ the area of W, and dx an infinitesimal region which contains the point x. Then a first-order property of the point process, the intensity function, is defined as,

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

A stationary spatial point process has a constant intensity and second-order (covariance) moment properties that depend only on the distance between locations, not on the locations themselves. For a stationary point process $\lambda(x)$ is a constant value λ and can be estimated (Diggle, 1983) as,

$$\hat{\lambda} = \frac{N(W)}{|W|}. \tag{2}$$

The second order intensity function is also defined in a similar manner,

$$\lambda_2(x, y) = \lim_{|dx|, |dy| \to 0} \frac{E[N(dx)N(dx)]}{|dx||dy|}. \tag{3}$$

For a stationary process $\lambda_2(x, y) = \lambda_2(x - y)$. For a stationary isotropic spatial point process $\lambda_2(x, y)$ reduces to $\lambda_2(h)$, where h is the distance between x and y; i.e. $\lambda_2(h)$ does not depend on the direction of location x relative to location y. Then $\frac{\lambda_2(h)}{\lambda^2}$ is defined (Diggle, 1983) as the radial distribution function and $\gamma(h) = \lambda_2(h) - \lambda^2$ is defined as the covariance density. Covariance density measures spatial dependence of events for locations as a function of the distance h.

A point process that exhibits no dependence among events is called an independent process. Independent spatial point processes including, homogeneous Poisson process, have $\lambda_2(h) = \lambda^2$ and $\gamma(h) = 0$ for all $h > 0$. If the events have regular spacing, as indicated by a constant covariance density, then it is called a regular process. If similar events are clustered together with a higher positive covariance in some locations than at other locations in a region, then the process is called a clustered process. Illustrations of independent, regular and clustered spatial point processes are shown in Fig. 2.

2. Initial Investigations of Complete Spatial Randomness

The first stage of an analysis of a spatial point process often focuses on identifying whether the point process exhibits complete spatial randomness (CSR). Complete spatial random-

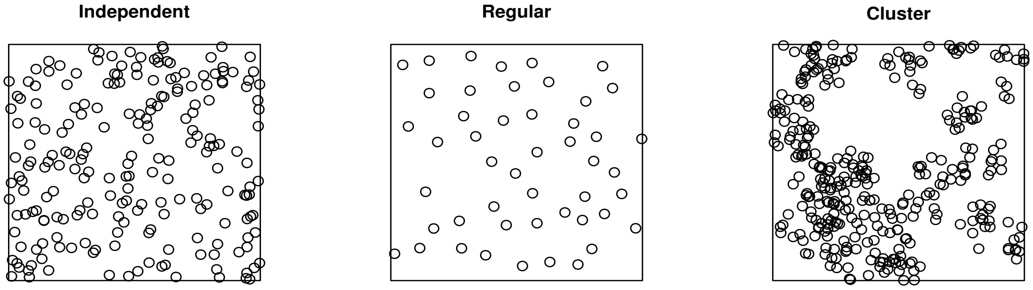


Figure 2: Plots of Independent, Regular and Clustered Point Processes in Rectangular Windows.

ness indicates that there is no specific underlying spatial pattern and that locations are simply random over the study area. If the process is not a CSR process then identification of a pattern or patterns in locations is undertaken. Identifying the spatial structure can aid the comprehension of underlying geographical processes and their relationships with the phenomenon under investigation. In spatial point processes, completely spatial randomness is characterized by locations following a homogeneous Poisson process.

2.1 Quadrat Analysis of Block C Large-rocks

The quadrat method is a frequently used technique for testing the hypothesis of complete spatial randomness. This method is based on a goodness-of-fit test that locations follow a homogeneous Poisson process. The numbers of events are counted in sub regions called quadrats. It is convenient to have quadrats with the same shape and size. Even though any geometrical shape can be used, usually quadrats are rectangles. Quadrats can be placed randomly in the study area or made contiguous, depending on the nature of the region.

Swindel (1983) discussed the importance of optimal quadrat size to obtain the maximum information from the data. The optimum quadrat size for randomly located quadrats in a study region was defined as $\frac{1.6}{\hat{\lambda}}$, the maximum likelihood estimate of quadrat size, for a Poisson process. Here $\hat{\lambda}$ is the estimated intensity (see eqn. (2)). As an example, for Block C, the large-rocks optimum quadrat size for randomly located quadrats is $0.815 \times 0.815m^2$. There are 176 quadrats of this size as shown in Fig. 3. In subsequent quadrat analysis the number of rocks in each quadrat is enumerated based on each block's optimum quadrat size, even though the quadrat will comprise a grid of contiguous quadrats. Table 1 shows the frequency distribution of the number of rocks per quadrat. The expected frequency of rocks in the 176 quadrats has been obtained under the hypothesis of a homogeneous Poisson distribution with a mean $\lambda \times |quadrat|$. Pearson's Chi-square goodness-of-fit test was performed to test the hypothesis that the process is completely spatially random (CSR). On the basis of the testing procedure for Block C large-rocks locations, the observed Chi-square value was 31.42 and the resulting p-value was 7.74×10^{-6}. Hence, reject the hypothesis of CSR. Rocks appear to be clustered.

2.2 Cluster Indices

Since the preliminary hypothesis of CSR for Block C large-rocks is rejected, it would be beneficial to identify the degree of the departure. There are several statistics and cluster indices available to quantify the departure from CSR. A few important such cluster indices

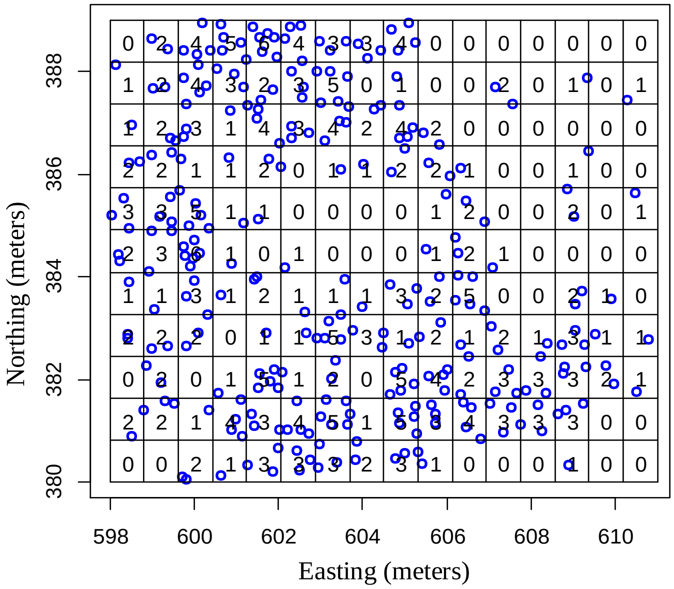


Figure 3: $0.815\ m \times 0.815\ m$ quadrats and quadrat counts for Block C large-rock locations.

Table 1: Frequency distribution of the number of rocks per quadrat in a sample of 176 quadrats of size $0.815\ m \times 0.815\ m$. See Fig. 3. for the number of large-rocks in each Block C quadrat.

Rocks per Quadrat	Observed Quadrat Frequency	Expected Quadrat Frequency
0	55	35.45
1	42	56.81
2	32	45.51
3	25	24.31
4	11	9.74
5	9	3.12
6	2	0.83

and their realizations for Block C large-rock locations are given in Table 2.

Fisher et al. (1922) derived a statistic I, that can be used to identify the clustering and regularity based on I being greater than 1 or not. This statistic is derived based on the theoretical relationship of the ratio of the variance to the mean of a Poisson distribution. If the point process is clustering, the variance of quadrat counts increases relative to the mean; hence, the resulting ratio I will be greater than 1. Similarly for regular point processes quadrat counts are uniform and the variance decreases relative to the mean of the Poisson process and the resulting ratio I tends to be less than 1. The statistical significance of this test can be assessed using a Chi-square distribution. For Block C large-rocks, Fisher's I indicates significant clustering with a p-value $= 3.33 \times 10^{-8}$ based on its Chi-square test.

Loyd's X^* (Lloyd, 1967) estimates the average number of events sharing a quadrat with another event. Loyd's IP measures the mean crowding relative to the mean density. The average number of neighboring rocks within a quadrat is 2.086 and the mean crowding is 1.302 times as great as the mean density. Both of these quantities provide extra information about the underlying clustering structure tested by the Fisher's I. More details of these quantities can be found in the Ripley (1981) and Cressie (1993).

Table 2: Cluster indices for Block C large-rock locations. $\bar{X}$ is the sample mean and S^2 is the sample variance of the quadrat counts. Q is the total number of quadrats in Block C.

Index	Estimator	Realization
Fisher et al. (1922) I: Relative Variance	$\frac{S^2}{\bar{X}} \sim \frac{\chi^2_{Q-1}}{Q-1}$	1.484
Lloyd (1967) X^*: Mean Crowding	$\bar{X} + \frac{S^2}{\bar{X}} - 1$	2.086
Lloyd (1967) IP: Index of Patchiness	$\frac{X^*}{\bar{X}}$	1.302

2.3 Agglomerative Approach

Random quadrat techniques fail to account for the spatial locations in the analysis. For example, the Pearson Chi-square test is a goodness-of-fit test for the null hypothesis that the process is CSR. The alternative hypothesis is simply the negation of null. The test might reject the null hypothesis because of many reasons, such as non-uniformity of the point process or events being dependent in a variety of possible ways. Hence, it would be useful to consider alternative analysis methods that can account for the spatial dependence and non-uniformity of the processes.

Methods for the analysis of grids of contiguous quadrats have the ability to include spatial locations of the point process. In this section, a method proposed by Greig-Smith (1952), the agglomerative approach, is applied to the Block C large-rocks. The agglomerative approach requires that the Block C region be re-divided into rectangular quadrats so that the total number of quadrats is a power of 2. Consequently, $Q = 32 \times 32 = 1,024$ quadrats (Fig. 4), each of size .28 $m \times .41$ m, were selected for analysis. The method begins by combining pairs of adjacent quadrats into blocks of size two. Mean squares of block rock counts are calculated. Then, sequentially, adjacent blocks are combined and mean squares calculated.

The between-blocks mean square, MS_r, is calculated as in eqns (4) and (5). Figure 4 illustrates the sequential combining of blocks.

$$SS_r = \frac{1}{2r}\sum_{i=1}^{m}(A_i - B_i)^2, \tag{4}$$

$$MS_r = \frac{SS_r}{m}, \tag{5}$$

where $m = \frac{Q}{2r}$ denotes the number of pairs of blocks, each containing r quadrats, and A_i and B_i denote the number of events in the i^{th} pair of blocks. A plot of the mean squares versus block size is used to identify clustering and regularity of the point process. Figure 5

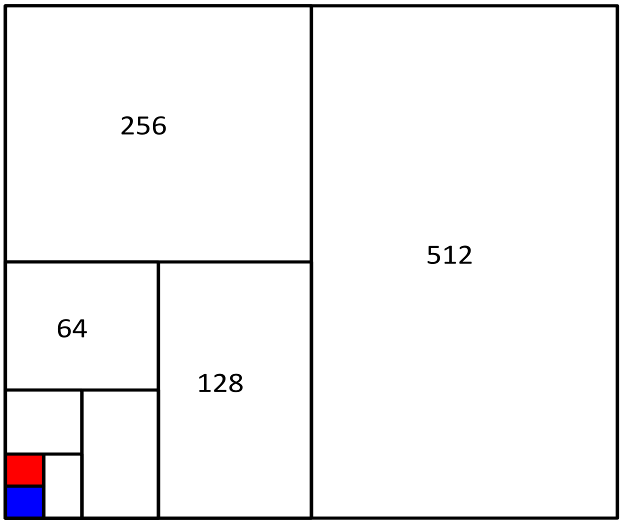


Figure 4: Agglomerative quadrats of Block C; sequentially double the size of blocks.

shows the plot of the mean squares versus block size for Block C large-rock locations. A simulation envelope is obtained by randomly rearranging the rock locations and calculating mean squares, as before, a large number of times (200). Upper and lower limits are the 97.5^{th} and the 2.5^{th} percentile of the mean squares of simulated realizations.

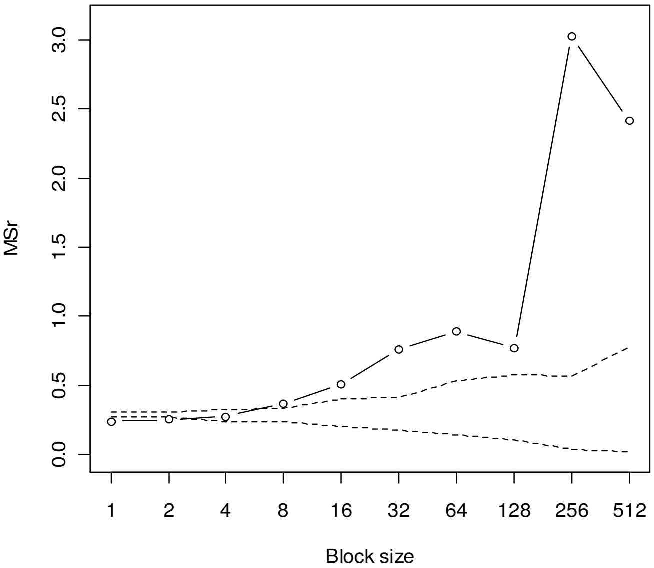


Figure 5: Plot of mean squares MS_r versus block size r for the Block C large-rock locations. The solid line is the mean squares of the block counts. Dotted lines provide upper and lower limits of simulated envelopes under the assumption of complete spatial randomness.

It is seen in Fig. 5 that mean squares fall below the lower acceptance envelope for the two smallest block sizes, indicating that adjacent quadrat counts are more homogeneous than would be expected from randomly arranged quadrat counts. For most of the larger block sizes, the mean squares fall above the upper acceptance envelope. Peak mean squares occur in blocks of sizes 64 and 256 quadrats, indicating patches of size $[3.25\ m, 2.25\ m]$ and $[6.50\ m, 4.50\ m]$. Both small-scale regularity and large-scale clustering are indicated in the realization of the Block C large-rock locations.

3. The K-function

The K-function plays an important role in spatial point process analysis because of its demonstrated ability to capture the spatial dependence between different regions of a point process. There are several variations of the K-function available in the literature. Most of them differ depending on how border (i.e., edge) corrections are applied. The choice of the edge correction does not appear to be very important as long as some edge correction is applied (Baddeley, 2010). Ripley (1976) defined the K-function for a stationary point process so that $\lambda K(h)$ is the expected number of other points of the process within a distance h of any point of the process. Ripley (1976) estimated $K(h)$ as

$$\hat{K}(h) = \frac{1}{N\hat{\lambda}} \sum_{\substack{\text{i=1} \\ i \neq j}}^{\text{N}} \sum_{\text{j=1}}^{\text{N}} w(s_i, s_j)^{-1} I(||s_i - s_j|| \leq h), \tag{6}$$

where the weight $w(s_i, s_j)$ is the proportion of the circumference of a circle centered at s_i passing through s_j, N is the number of events, $I = 1$ if $||s_i - s_j|| \leq h$ and 0 otherwise, and $\hat{\lambda} = \frac{N}{|A|}$ is the estimated intensity.

For a homogeneous Poisson process, the expected number of events within a distance h of a specified event is $\pi h^2 \lambda$ and hence the K-function value is πh^2. If the process is clustering, then $\hat{K}(h)$ tends to be greater than πh^2 and if the process is regular then $\hat{K}(h)$ tends to be less than πh^2.

The K-function is an alternative characterization of the second-order intensity function for a stationary isotropic process. Diggle (1983) showed that the relationship between the second-order intensity function and the K-function is

$$\lambda_2(h) = \lambda^2 \frac{K'(h)}{2\pi h}. \tag{7}$$

In practice, it is more convenient to work with the K-function than λ_2 because of its empirical behavior. A plot of the K-function generally provides meaningful insight about the underlying process. Figure 6 shows the K-functions for both large and small-rocks in all three Mountaineer study sites. This figure and all the K-function and related calculations and the graphs in this dissertation are obtained from the "spatstat" package in R (Baddeley and Turner, 2005).

The plots suggest that small-rock locations in all blocks and large-rock locations in Block C are clustering. Large-rock locations in Block X appear to indicate small-scale regularity and large-scale clustering while large-rocks locations of Block Y are suggesting only the regularity. These suggested behaviors of regularity and clustering await formal statistical analysis to determine whether the suggested behaviors are due to chance.

Besag (1977) recommended plotting a variation of the K-function, namely the L-function defined as

$$L(h) = \sqrt{\frac{K(h)}{\pi}} - h. \tag{8}$$

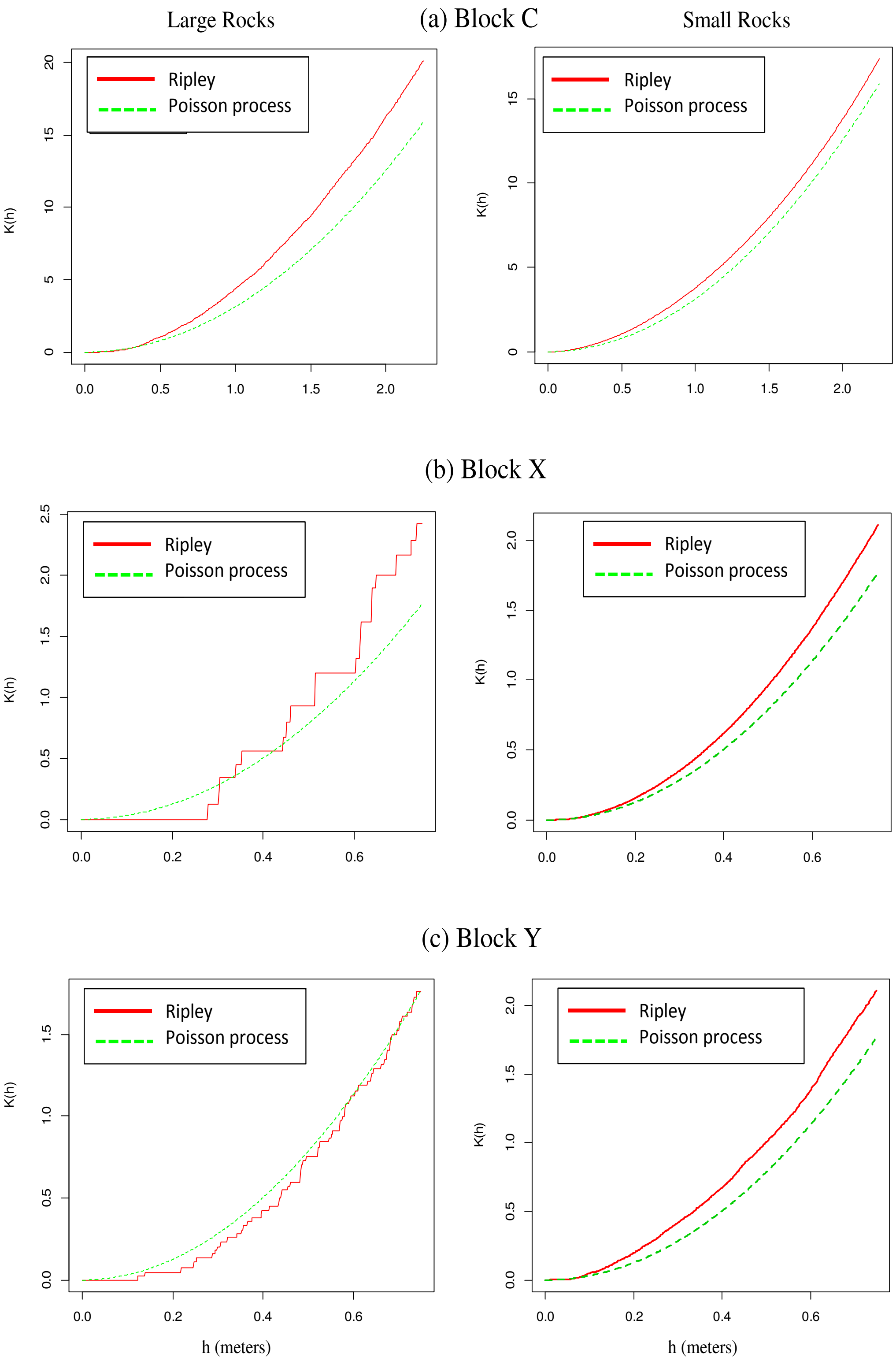


Figure 6: K- function for both large and small-rocks in Block C, Block X and Block Y. The dotted line is the theoretical K-function for a homogeneous Poisson process. The solid line is the empirical K-function of the mapped rocks.

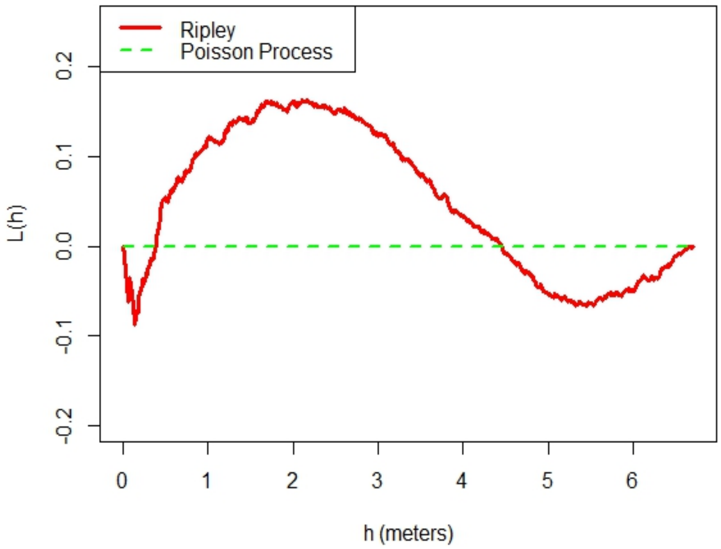


Figure 7: Empirical L-function (Ripley's edge corrected) for the Block C large-rock locations. The dotted line is the theoretical L-function for a homogeneous Poisson process.

Under an assumption of a homogeneous Poisson process, $K(h) = \pi h^2$ and therefore the empirical L-function should reasonably approximate a horizontal line centered at zero; that is, $\hat{L}(h) = 0$. A plot of the estimated L-function allows identifying the small-scale regularity and large-scale clustering in a transformed scale much more clearly.

Figure 7 shows the L-function of the large-rocks in Block C. It suggests regular rock locations for small distances ($< .40\ m$) and clustering at the larger ($0.40\ m$ to $4.50\ m$) distances.

4. Evaluating Complete Spatial Randomness by Using Simulation Envelope Approach

As it was shown in the Fig. 6, all the small-rock locations tend to provide clustering structures. However, large-rock locations in Blocks X and Y provide lack of graphical evidence to suggest any clustering structure. Ripley (1981) suggested a Monte-Carlo-based goodness-of-fit method to test the reasonableness of the fitted process by using the K-function. A thorough discussion of this method can be found in Baddeley (2010). The testing procedure is based on a simulation envelope which obtains by simulating the fitted process a large number of times using the estimated parameters of either the fitted K-function or the L-function of the clustering process.

In the case of testing completely spatial randomness, the simulation envelope of the homogeneous Poisson process based on the estimated intensity function will be obtained and compared to the empirical L-function (see Chapter 20 in Baddeley (2010)). The empirical L-functions and the corresponding simulation envelopes of all the sites are presented in Fig. 8.

5. Conclusion

As suspected, all the small-rock locations and the large-rock locations in Block C show very strong evidence of clustering. This departure from complete spatial randomness is clearly evident in Fig. 8 since the empirical L-function substantially deviates from the simulation envelope. Also, the suspected small scale regularity is significantly evident only in the large-rock locations in Block C. The apparent small dips at the beginning of the curves of the rest of the locations are not provide any significant evidence of small scale regularity. However, large-rock locations in Blocks X and Y do not show any evidence of clustering or regularity; hence, those locations will not be considered for further investigation in future studies. Rest of the blocks with clustering structures will be further studied and attempt to identify the underlying geometrical patterns. Due to the apparent shape of the clustering structure of the Block C large rocks data, it may be suggested to investigate the possibility of fitting a circular model using least squares principle. Statistical evidence of such a structure may solidify the professor David Meltzer's initial hypothesis of existence of man made prehistoric house structure on the Block C.

References

Baddeley, A. (2010), "Analyzing Spatial Point Patterns in R," in *Case Studies in Spatial Point Pattern Modelling*, Western Australia: CSIRO and University of Western Australia, Workshop Notes, pp. 23–74.

Baddeley, A. and Turner, R. (2005), "Spatstat: An R Package for Analyzing Spatial Point Patterns," *Journal of Statistical Software*, 12, 1–42.

Besag, J. E. (1977), "Comment on Modeling Spatial Point Process by B.D. Ripley," *Journal of the Royal Statistical Society*, 39, 193–195.

Cressie, N. (1993), *Statistics for Spatial Data*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, J. Wiley.

Diggle, P. (1983), *Statistical Analysis of Spatial Point Patterns*, Mathematics in Biology, Academic Press.

Fisher, R. A., Thornton, H. G., and Mackenzie, W. A. (1922), "The Accuracy of the Platting Method of Estimating the Density of Bacterial Populations," *Journal of Statistical Software*, 9, 325–359.

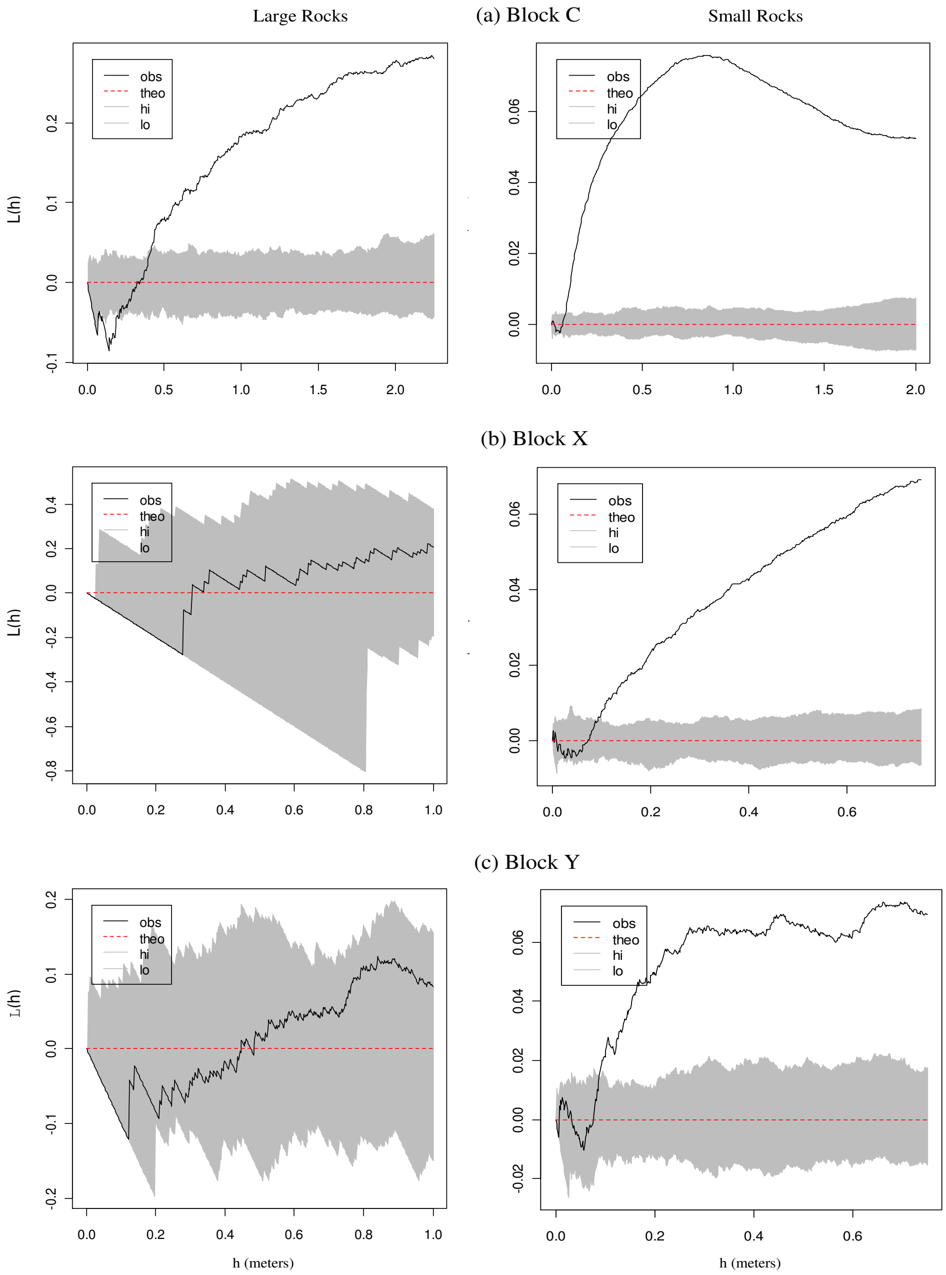
Greig-Smith, P. (1952), "The Use of Random and Contiguous Quadrats in the Study of the Structure of Plant Communities," *Annals of Botany*, 16, 293–316.

Lloyd, M. (1967), "Mean Crowding," *Journal of Animal Ecology*, 36, pp. 1–30.

Ripley, B. (1976), "The Second-Order Analysis of Stationary Point Processes," *Journal of Applied Probability*, 13, pp. 255–266.

— (1981), *Spatial Statistics*, Wiley, New York.

Swindel, B. F. (1983), "Choice of Size and Number of Quadrats to Estimate Density from Frequency in Poisson and Binomially Dispersed Populations," *Biometrics*, 39, 455–464.



(a) Block C

(b) Block X

(c) Block Y

Figure 8: Empirical L-functions and corresponding simulation envelopes for both small and large-rock locations in all blocks. The dotted red line is the theoretical L-function for a homogeneous Poisson process and the solid black curve is the empirical L-function (Ripley's edge correction). Solid shaded regions are represent the simulation envelope.