Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 05-031

Spatial Clustering Of Chimpanzee Locations For Neighborhood Identification

Sandeep Mane, Carson Murray, Shashi Shekhar, Jaideep Srivastava, and Anne Pusey

September 15, 2005

Spatial Clustering Of Chimpanzee Locations For Neighborhood Identification

Sandeep Mane[†] *, Carson Murray[¶§], Shashi Shekhar[†], Jaideep Srivastava[†] and Anne Pusey[¶§]
[†] Department of Computer Science,
[¶] The Jane Goodall Institute's Center for Primate Studies,
[§] Department of Ecology, Evolution and Behavior,
University of Minnesota,
Minneapolis, USA.

Abstract

Since 1960, the chimpanzees (Pan troglodytes) of Gombe National Park, Tanzania, have been studied by behavioral ecologists, including Jane Goodall. Data has been collected for the last 40 years and it is now being further analyzed by researchers in order to increase our understanding of the social structure of chimpanzees. In this paper, we consider the following question of interest to behavioral ecologists –"Does clustering exist among female chimpanzees in terms of the spatial locations visited by them?" The analysis of this question will help behavioral ecologists to learn about the space use and the social interactions between female chimpanzees. The data collected for this analysis are marked spatial point patterns over the park. Current spatial clustering methods lack the ability to handle such marked point patterns directly. This paper presents a novel application of spatial point pattern analysis and data mining techniques to the ecological problem of clustering female chimpanzees. We studied various spatial analysis techniques and found that the Ripley's K-function provides a powerful tool for evaluating clustering behavior among spatial point patterns. We then proposed two clustering approaches for marked point patterns based on this widely-used statistical K-function. Experimental results using the proposed clustering methods provide significant insight into the dynamics of female chimpanzee space use and into the overall social stucture of the species. In addition, the methods proposed here can be extended to also include temporal information.

Keywords

Spatial point patterns, spatial clustering, Ripley's K-function, complete-link clustering, reverse Cuthill-Mckee ordering.

*Primary contact: smane@cs.umn.edu

1 Introduction

In 1960, Jane Goodall began the first long-term field study of chimpanzees, at the Gombe National Park, Tanzania. This study continues today and has greatly increased our understanding of chimpanzee behavior and evolution of their social structure. Since chimpanzees are our closest living relatives, it also provides insight into how human societies have evolved in the past. One of the main aims of behavioral ecology is to understand how ecology influences the social structure exhibited by an animal species. This question is particularly important in chimpanzees because they have an unusual social structure. Although they live in permanent social groups, they have a fission-fusion society in which groups are transient and range from solitary individuals to larger groups. This pattern appears to be a result of differences in individual space use. It is therefore critical to measure space use in order to understand the ecological factors that influence the overall social structure. Thus, finding methods by which to assess space usage is of primary interest to chimpanzee researchers and behavioral ecologists in general.

Chimpanzees show a distinctive grouping with males often traveling in larger groups while females exhibit a more solitary behavior. The latter are studied in this paper. The data collected for the chimpanzees are a set of marked point patterns over a spatial region (chimpanzee community range). Much less research has been done to understand the interaction among such spatial point patterns for individuals or groups of individuals. Thus, in this paper, we apply data mining and spatial statistical techniques to study clustering of female chimpanzee locations. The challenge lies in the clustering of marked point patterns where the amount of overlap among the different point patterns is very pronounced. The aim here is to achieve an ecologically-meaningful clustering of the marked point patterns (for female chimpanzees). This paper shows two approaches for clustering these point patterns. The first approach uses the

Ripley's K-function with the complete-link clustering algorithm (hierarchical clustering) while the second approach uses the Ripley's K-function along with the reverse Cuthill-Mckee (RCM) algorithm, a matrix block diagonalization technique. These approaches provide a behavioral ecologist with an easy ecologically-meaningful, statistical interpretation of clustering among female chimpanzees. The techniques shown here for female chimpanzees can be applied to study spatial clustering among other species as well.

The remainder of this paper is organized as follows – section 2 summarizes the main contributions of research in this paper. Section 3 provides domain background and explains the main hypothesis of interest to behavioral ecologists, which is addressed by this paper. Section 4 summarizes related work. Section 5 explains two proposed approaches to address the problem of spatial clustering of point patterns. These approaches combine the existing spatial statistics and data mining techniques in a novel fashion. Section 6 shows some of the (ecologically) interesting experimental results. Section 7 draws conclusions and identifies future research directions.

2 Key contributions:

2.1 Data Mining domain:

The main contributions of this paper to data mining are:

(i) *Clustering techniques for marked spatial point processes:* Spatial clustering is currently a highly active research topic in data mining. Han et al. [8] provide a nice survey of spatial clustering methods used in data mining. However, the 'marked spatial point pattern' characteristic of this dataset, makes it difficult to directly apply traditional spatial clustering methods (partitional-based, hierarchical-based, density-based, or grid-based) in order to address our spatial clustering problem. To the best of our knowledge, no prior work has been done to address clustering in such datasets, wherein we have distinct point processes (marks i.e., female chimpanzees) with an observed spatial distribution over a given spatial region. This paper thus provides two novel approaches to cluster spatial point processes by using existing spatial statistical and data mining methods.

(ii) *Application of these techniques to a real-world dataset:* This paper shows a real-world problem of interest to researchers viz, clustering of spatial point process to identify neighborhoods. We show the application of proposed methods to a real-world ecological dataset. The paper also demonstrates that these approaches help in understanding the clustering changes over distance. Such information is useful for studying the interactions among spatial point patterns.

(iii) *Groundwork for extending to spatio-temporal analysis:* Extensions of K-function to spatio-temporal domain will be useful to domain scientists. This paper lays the groundwork for future work of spatio-temporal clustering of spatial point processes.

2.2 Behavioral Ecology Domain:

The main contributions of this paper to behavioral ecology are:

(i) *Female neighborhood identification:* These results contribute to chimpanzee research by assisting researchers in the identification of female neighborhoods or lack thereof. It also helps to identify females who use a "loner" strategy (outlier).

(ii) *Ranging patterns for males vs. females:* These techniques allow chimpanzee researchers to rigorously examine the ranging patterns for males since some studies have suggested they also have core areas comparable to those for females.

(iii) *The big picture:* While contributing greatly to chimpanzee research, these findings are also applicable to the larger behavioral ecology domain. In general, ranging and grouping patterns are of primary interest to behavioral ecologists. There is often a great deal of overlap between individual or group ranges. The techniques we develop here help to identify clusters within a species having a highly pronounced overlap between individual/group space usage. We are therefore confident that it can also be applied to species with similar or lower degrees of spatial overlap.

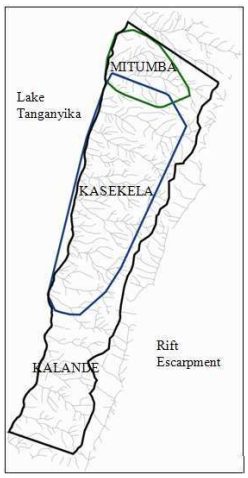
3 Domain background

3.1 Study site and data collection:

Jane Goodall began the Gombe Stream chimpanzee research project in 1960. Gombe is a small park (35 km^2) located on the eastern border of Lake Tanganyika, and its habitat ranges from semi-deciduous forests in the valleys to grasslands on the ridges. The park includes 12 main valleys along approximately 14km of shoreline. There are currently three communities in the park, two of which are habituated (Fig. 1). The community ranges for these communities are given by 99% minimum convex polygons (MCP) of all chimpanzee follow locations for a specific time interval, as shown in Fig. 1. The central Kasekela community has been studied since 1960 while the northern Mitumba community has been studied since 1983.

Since 1973, observers have followed one chimpanzee in the Kasekela community (hereafter the "*focal*" of a "*follow*") for an entire day (as described in Goodall [7]). They

Figure 1. Gombe National Park, Tanzania.

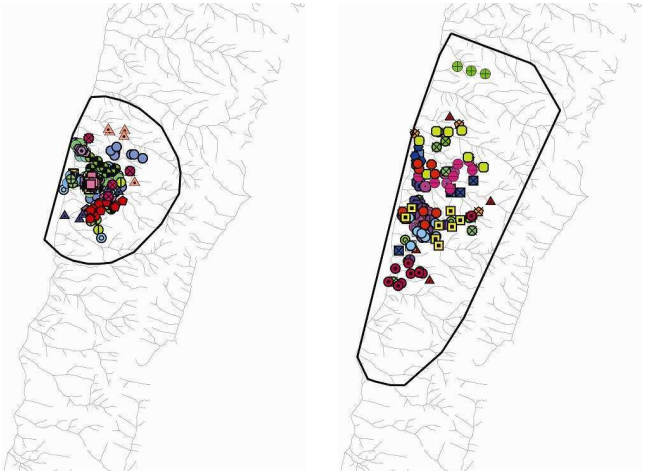


note point samples at 15-minute intervals, and record information like group composition, sexual states for all females, location, and feeding of the focal. It should be noted that only a subset of community members are followed since some females in the community are still anxious in the presence of human observers and can not be followed for an entire day. Ranging and association patterns are derived for these individuals, however, based on where they are encountered during a follow (Williams et al. [18]). All of these data have been computerized and are housed at the University of Minnesota. We have access to all long-term records, and subsequent analyses in this paper are based on data from the Kasekela community. From 1974-2002, the Kasekela community contained between 20-36 adult chimpanzees with 12-24 adult females and 7-18 adult males.

3.2 Chimpanzee social structure & female space use

Chimpanzees are a highly gregarious species with transient fission-fusion groupings within a permanent community (Goodall [7]). The sociability and ranging patterns of male and female chimpanzees differ distinctively. Males are more social than females, use the entire community range, and exhibit pronounced dominance hierarchies (Bygott [2]; Goodall [7]). Females, by comparison, are less social than males, concentrate their use in subsets of the community range, and have a subtler dominance hierarchy (Nishida [11]; Wrangham et al. [19]). While females typically disperse from their natal community, individuals can adopt three different space use strategies: immigration into a community, remaining within the natal community, or occupying a peripheral range. Regardless of her position within/around a community, each female is associated with a particular area ("*core area*") in which she spends most of her time. It should be noted that female locations and core areas overlap substantially (Goodall [7]; Hasegawa [9]; Williams et al. [18]). Figure 2 illustrates a high degree of overlap of core area points for females in two distinct year intervals (1979-1982 and 2001-2002). Once they have established a core area, females demonstrate a high level of site fidelity even in the face of shifting male territories (Williams et al [18]). It should be noted females often increase their sociability and range well outside their core areas when sexually receptive.

Figure 2. Core area points for adult female chimpanzees. (Note that the community range size and the degree of overlap changes substantially. In 1979-1982, the community range (99% MCP) covered 8.1 km^2. In 2001-2002, it covered 18.6 km^2.)



(a) 1979-1982 (b) 2001-2002

Previous research from Gombe has reported that female core areas are clustered into neighborhoods (Williams et al., [18]). "*Neighborhoods*" are defined as distinct, stable clusters of the females' spatial point patterns. During their study period (1975-1992), Williams et al. reported that Gombe females were clustered into two neighborhoods with a few females occupying a peripheral core area. They also found that female space had implications in terms of reproductive success. Northern neighborhood females had higher reproductive success than southern neighborhood females while females who switched neighborhoods did poorly. Peripheral females either had very high or very low reproductive success. Understanding female space use is particularly crucial because of these implications for reproductive success and because it is thought to determine male distribution, intergroup aggression, and mating systems.

4 Related work:

The ranging information for female chimpanzees is a spatial marked point pattern over the community range. Analysis of spatial point patterns is an active field of study in the spatial statistics. However, to the best of our knowledge, little research has been done in clustering marked spatial point pattern analysis (see Han et al. [8]; Shekhar et al. [17]). In the past, chimpanzee studies have often categorized females as "northern", "southern", "central", or "peripheral" (e.g. Nishida [11]). While these terms may im-

ply a neighborhood female distribution, they were based on visual estimates of data or a general sense of female space use from field observations. Clustering of chimpanzees into neighborhoods was first defined mathematically by Williams et al. [18]. They showed the importance of ranging behavior of females and increased our understanding of the social structure of chimpanzees. However, this research did not allow for the study of how clustering varies with distance i.e. are there any global and/or local ranging patterns which influence the interactions among different chimpanzees ? In addition, the use of dendrograms is of concern due to the inherent sensitivity of ordering within dendrograms with respect to the dissimilarity measure.

5 Clustering spatial point patterns

5.1 Problem definition

The location data for female chimpanzees in Gombe Park is a marked spatial point process where each female represents a unique mark. From a behavioral ecology point of view, we want to determine whether there exist any neighborhoods (stable clusters) among female chimpanzees. Hence, from a spatial data mining perspective, the main problem addressed by this paper is –
" Given a marked spatial point process, is there a spatial clustering among the different marked processes ?"

This research aims to use unsupervised learning to study clustering among marked spatial point patterns. For this, we combine "data mining + spatial statistics" techniques to provide methods to study clustering at different distances. In spatial statistics, there are two main alternative methods (Cressie [3]) used for analyses of spatial point pattern data. One is based on determining the attraction and repulsion effects by observing cell-count statistics – e.g. using the second-order moment function, Ripley's K-function (Ripley [14]). The other is based on the use of nearest-neighbor information (Diggle and Cox [5], Diggle [4]). For our analysis, we consider the use of the former since it provides a good means by which to assess the variation of interaction effects with distance. Furthermore, statisticians generally believe that the former provides a better, rigorous statistical analysis. In this paper, we use two different methods for clustering marked spatial point pattern using a spatial measure (K-function) as the dissimilarity measure viz, the first uses complete-link (hierarchical) algorithm while the second uses reverse Cuthill-Mckee ordering for block diagonalization of matrices.

5.2 Dissimilarity measure

The first-order properties of a spatial point pattern are described by the variation of the expected value (mean or average) across space (e.g. the intensity of the spatial point pattern). Such properties are usually estimated using kernel estimation techniques. Second-order properties describe the covariance (or correlation) between values of the spatial point pattern at different regions in space. The K-function provides one such measure for the second-order properties over a range of distances.

The K-function provides a powerful spatial statistical approach to study both local as well as global interactions among point patterns. The K-function is an isotropic measure i.e. it is independent of the direction. It can describe the characteristics of point processes at many different scales, which is a characteristic that most other summary functions (like mean nearest-neighbor distribution and the cumulative distribution function) lack. The K-function at a distance 'r' gives a measure of difference between the observed number of pairs of points within distance 'r' of each other and the expected number of pairs of points within distance 'r' of each other. For the univariate case, the pair of points have same marks while for the bivariate case, the pair of points have different marks. The K-function tests a point pattern for attraction, repulsion or complete spatial randomness (CSR) among points. Note that for spatial point processes, CSR is equivalent to the assumption that the underlying point process is a homogeneous Poisson process. Formally,

Definition 1 *The K-function (Ripley [13]), at a distance 'r' for a univariate spatial point pattern is defined as –*
K(r) = λ^{-1} E(number of extra events within distance r of a randomly chosen event)
where, λ is the intensity of that spatial point process.

Mathematically, suppose that we observe a point process over a plane region D of area A and that the observed x-points are $x_1, \ldots, x_n$, then the unbiased estimate for univariate K-function (Lotwick and Silverman [10]), without any edge corrections, is –

$$\widehat{K}_{xx}(r) = \frac{A}{n^2} \sum_{i \neq j} \sum \frac{I_r(d_{ij})}{w_{ij}} \tag{1}$$

where $I_r(d_{ij})$ is an indicator function which is zero if distance between i and j (both have same marks) is greater than r. Otherwise, it is the reciprocal of the proportion of the circumference of the circle centered at i with radius d(x_i,x_j) that is lying within the sampling window D. In case of CSR, the expected value $\widehat{K}(r)$ is πr^2. If the observed $\widehat{K}(r) > \pi r^2$, then the point process is said to be clustered at distance r. If $\widehat{K}(r) < \pi r^2$, then the point process is said to show repulsion among the points.

Definition 2 *The definition of K-function is extended to the bivariate case, $K_{ij}(r)$, as,*
$K_{ij}(r)$ = λ_j^{-1} E(number of type j events within distance r of a randomly chosen event of type i)

where, λ_j is the intensity of that spatial point process with marks j.

The bivariate K-function is a symmetric measure, unless edge-corrections are used. Edge corrections are required if a number of points of interest are close to the boundary of the study area. In our methods, the indicator function $I_r(d_{ij})$ used in the unbiased estimate of K-function capture the edge corrections required, if any. Mathematically, suppose that we observe y-point process $y_1, \ldots, y_m$ over the same plane region D, then the unbiased estimate for K-function without any edge corrections (Lotwick and Silverman [10]) is –

$$\widehat{K}_{xy}(r) = An^{-1}m^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{I_r(d_{ij})}{w_{ij}} \quad (2)$$

where $I_r(d_{ij})$ is an indicator function is defined similar to univariate case. Since the scale for $\widehat{K}_{xy}(r)$ is not linear in r, Besag's L-function (Besag [1]) is usually used to provide an easy linear interpretation of the interactions among point patterns. An important observation is that the variance of the function $\sqrt{\frac{K_{xy}(r)}{\pi}}$ is almost constant.

Definition 3 *The L-function is thus defined as –*

$$L_{xy}(r) = \sqrt{\frac{K_{xy}(r)}{\pi}} - r \quad (3)$$

If the estimate $\widehat{L}_{xy}(r)$ is positive (greater than zero) then it indicates that there is attraction between x-points and y-points at distance less than or equal to r, while a negative value for $\widehat{L}_{xy}(r)$ indicates repulsion. If $\widehat{L}_{xy}(r)$ equals zero, then it indicates complete spatial randomness.

5.3 Spatial Point pAttern ClustEring algorithm-1 (SPACE-1)

The first algorithm uses the MAX or complete-link clustering algorithm (Han et al. [8]). The estimate of L-function, computed using the unbiased estimator of K-function, was used as the dissimilarity measure. The MAX clustering technique defines the proximity of two clusters as the maximum distance (minimum similarity) between any two points in the different clusters. The reason for using this technique was because it was less susceptible to noise, which was a problem in our dataset due to some outlier points for some marks (females) in the spatial point pattern. The disadvantage of this technique is that it favors globular shapes. This was not a problem in our case, since most of the core areas for the females were globular. The algorithm 1 shows the steps for this approach. This algorithm outputs a hierarchical clustering of the point patterns which is plotted as a dendrogram, as shown in figure 3(a).

5.3.1 Time complexity of SPACE-1 algorithm

For the marked spatial point pattern S, let $M = \{m\}$ be the set of all marks and let $|M| = n$. Let $p_1, \ldots, p_n$ be the number of points of each respective mark $m_i \in S$, $1 \leq i \leq n$.

Algorithm 1 SPACE-1

Input:
- A bounding polygon for spatial region.
- A marked spatial point pattern within the spatial region.
- Distance 'r' for clustering.

Output:
- A hierarchical clustering of marks.

Pseudo-code:
1. Initialize the set S = $\{m | m \text{ is a mark }\}$.
2. For each $m_i, m_j \in$ S,
3. Compute $\widehat{L}_{m_i m_j}(r)$ at distance r between each pair of the marks m_i and m_j.
4. End for.
5. Create a dissimilarity matrix $M_{\widehat{L}(r)} = [l_{ij}]$, l_{ij} = $\widehat{L}_{m_i m_j}(r)$ (L-function estimate for marks m_i and m_j).
6. Using complete-link clustering algorithm and dissimilarity matrix $M_{\widehat{L}(r)}$, obtain a hierarchical clustering for marks in S.
7. Plot the hierarchical clustering of marks as a dendrogram.

Let $k = \max_n(p_i)$ be the maximum number of points for a mark in S. Then the worst-case time complexity for computing the value of K-function (and hence $\widehat{L}_{m_i m_j}(r)$) for any pair of marks (unique or same mark) is $O(k^2)$. Thus, the worst-case time complexity of computing all n^2 $\widehat{L}_{m_i m_j}(r)$ values in the first three steps of the algorithm is $O(k^2 n^2)$. For the complete-link clustering, the time required to initially sort the distances is $O(n^2 \log n)$. After each merge iteration, the distance metric can be updated in O(n). Thus, the overall worst-case time complexity of the algorithm SPACE-1 is $O(\max(k^2 n^2, n^2 \log n))$.

5.4 Problems in clustering using dendrograms

Initial work by Williams et al. for neighborhood detection motivated us to use dendrograms for visualization of spatial clustering of point patterns. However, several questions have been raised with respect to the applicability of dendrograms to clustering (Shroeder et al. [16]). Dendrograms can show instability or sensitivity for minor variations in dissimilarity values. In addition, dendrograms require researchers to use domain knowledge in order to identify the correct number of clusters. Finally, they assume a hierarchical structure to the dataset and will impose such a structure even in non-hierarchal systems. In this particular dataset, there is no evidence that female chimpanzee space use is hierarchical in nature. These concerns motivated us to investigate the limitations of dendrograms and also use matrix block diagonalization techniques for clus-

Algorithm 2 SPACE-2

Input:
- A bounding polygon for spatial region.
- A marked spatial point pattern within the spatial region.
- Distance 'r' for clustering.

Output:
- A block diagonalized matrix of dissimilarity indices.

Pseudo-code:
1. Initialize the set S = $\{m|m \text{ is a mark }\}$
2. For each $m_i, m_j \in$ S.
3. Compute $\widehat{L}_{m_i m_j}(r)$ at distance r between each pair of the marks m_i and m_j.
4. End for.
5. Create a dissimilarity matrix $M_{\widehat{L}(r)} = [l_{ij}]$, $l_{ij} = \widehat{L}_{m_i m_j}(r)$ (L-function estimate for marks m_i and m_j)
6. Using the reverse Cuthill-McKee algorithm, block diagonalize matrix $M_{\widehat{L}(r)}$ to obtain $M'_{\widehat{L}(r)}$.
7. Plot the block diagonalized matrix $M'_{\widehat{L}(r)}$.

tering $\widehat{L}_{m_i m_j}(r)$ estimates.

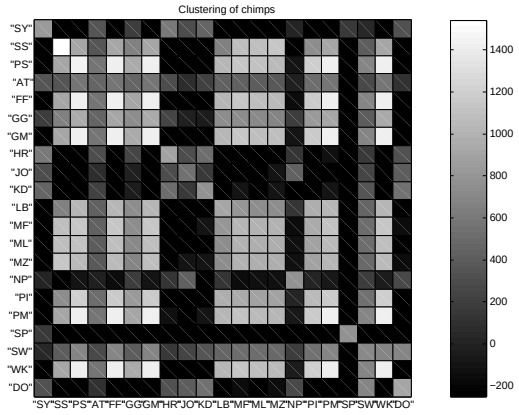
5.5 Spatial Point pAttern ClustEring algorithm-2 (SPACE-2)

This algorithm also uses $\widehat{L}_{m_i m_j}(r)$, estimated using unbiased estimator of K-function, as the dissimilarity measure. However, instead of using the complete-link algorithm, here we use the reverse Cuthill-McKee (RCM) ordering algorithm (George and Liu [6]) to block diagonalize the matrix $M_{\widehat{L}(r)}$ of the $\widehat{L}_{m_i m_j}(r)$ estimated for each pair of marks. This method also requires a subjective determination of clusters but it is more stable than dendrograms and does not assume a hierarchical nature to the dataset. Also, as a heuristic approach is used to choose the initial starting vertex in RCM technique, the current approach may require several iterations with different starting nodes in order to get a good block diagonalized matrix (clustering). Algorithm 2 shows the steps for this approach. This algorithm outputs a block diagonalized matrix for the marks of spatial point pattern, as shown in figure 3(d).

5.5.1 Time complexity of SPACE-2 algorithm

Similar to SPACE-1's analysis, the worst-case time complexity of computing all n^2 $\widehat{L}_{m_i m_j}(r)$-values for the matrix $M_{\widehat{L}(r)}$ using the first three steps of the algorithm is $O(k^2 n^2)$. For RCM algorithm, the time complexity is $O(d \log d\ |V|)$, where d = $\max\{\ degree(v) \mid v \in V\}$. Since in algorithm SPACE-2, all the cell values in the matrix may have a non-zero value, in worst-case scenario - $d = |V| = n$. Hence, the worst-case time complexity of RCM is also $O(n^2 \log n)$. Thus, the overall worst-case time complexity of the algorithm SPACE-2 is also $O(\max(k^2 n^2, n^2 \log n))$.

Figure 4. Clustering using SPACE-2 for year interval 1979-1982 and r = 250m.

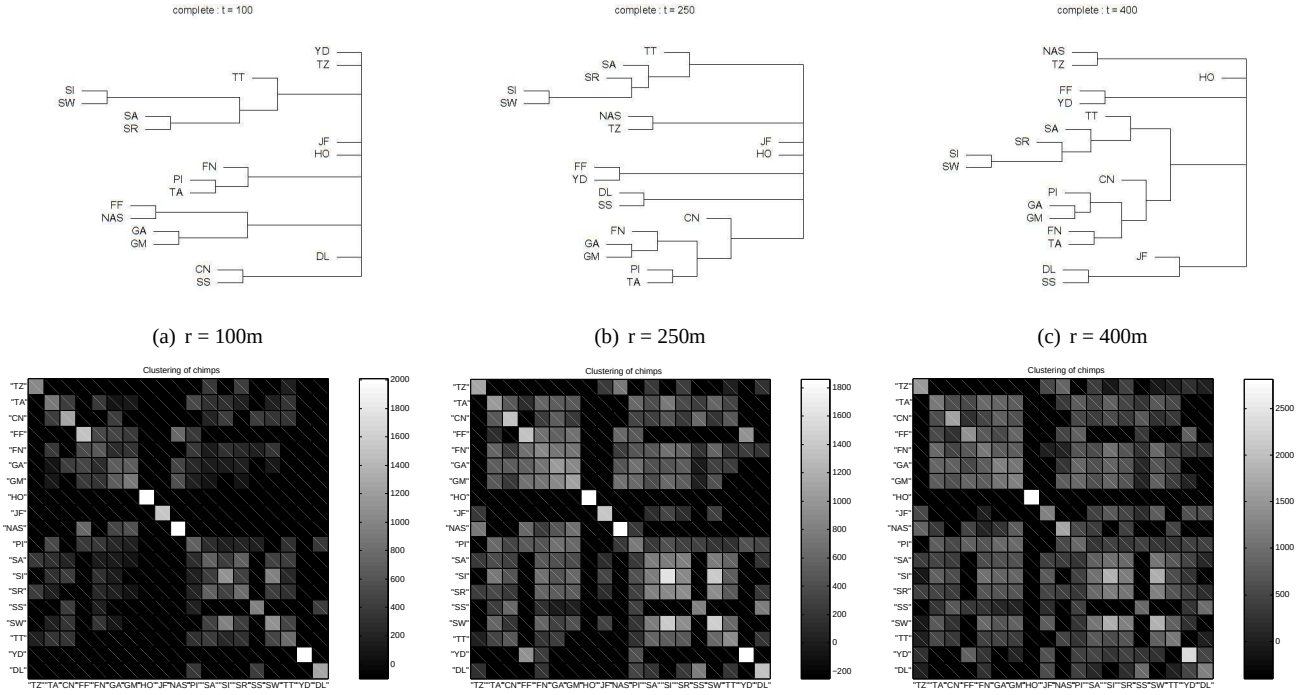


6 Experimental results

For our analysis, we use the locations for each female chimpanzee within its core area. In order to establish a core area, we use a 50% usage kernel of "alone" locations for each female (similar to the approach used by Williams et al. [18]). A female is "alone" so long as she is not sexually receptive and no other chimpanzees arrive into that follow within five minutes. Mother/daughter pairs and dependent offspring are still considered alone. The unbiased estimate of K-function (and hence $\widehat{L}_{m_i m_j}(r)$) for each pair of females is obtained using the splancs package (Rowlingson and Diggle [15]). For complete-link clustering, we use the "hclust" method in R statistical language [12] while for obtaining RCM ordering, we use "symrcm" function in MatLab.

One of the main advantages of the K-function is the ability to consider clustering at different scales. For this analysis, we choose distances (r) based on domain knowledge in terms of what distances would be biologically meaningful. The distances used in subsequent analyses are 100m, 250m, and 400m. The minimum distance (100m) reflects the inherent locational inaccuracy present in the dataset, since locations have been digitized from hand drawn maps. Given the topology of Gombe National Park, a 400m radius around a point will most often include points located in different valleys. Chimpanzees primarily communicate via vocalizations, the loudest of which is generally confined to a single valley. We therefore considered 400m to be the maximum distance at which chimpanzees can communicate or gain knowledge about the location of conspecifics. 250m

Figure 3. Clustering using SPACE-1 (dendrograms) and SPACE-2 (matrices) for year interval 2001-2002.



(a) r = 100m (b) r = 250m (c) r = 400m

(d) r = 100m (e) r = 250m (f) r = 400m

represents the midpoint between the minimum and maximum distances.

Here, we consider the prevalence of clusters or neighborhoods for two distinct time periods: 1979-1982 and 2001-2002. These intervals represent the extremes of our dataset with 1979-1982 having the smallest community range and most overlap of core area locations while the points are most spread out in 2001-2002 (refer to figure 2). While dendrograms for 1979-1982 seem to cluster females into neighborhoods, the results from our block diagonal matrix suggest that there are no distinct clusters (Fig 4). Rather most females are clustered into the center of the range except for a few outliers (loners or peripheral females) (SY, HR, JO, KD, NP, SP, DO). Interestingly, HR, JO, and KD form a spatially peripheral cluster while the other peripheral females are more solitary spatially. It seems likely that the lack of distinct clusters results from a high density of females in a small community range.

The dendrograms for 2001-2002 illustrate the difficulties of determining clusters (neighborhoods) at different distances (Fig. 3(a),(b),(c)). The ordering of individuals within clusters changes dramatically with slight variations in L-values. For example, FF belongs to different clusters at different distances. At 100m, she is clustered with NAS, GA, and GM while at greater distances she is removed from larger clusters and is grouped with YD. This example illustrates how clustering at different distances affects analyses of neighborhood (cluster)-level questions. The results of the block diagonal matrix clustering (Fig. 3(d),(e),(f)) illustrates that FF is clustered to a group of individuals (NAS, GA, GM) at all distances. It should be noted, however, that she is also clustered with TA and slightly clustered to PI. Visual inspection of the alone points confirm that FFs core area overlaps with those for both PI and TA as compared to those for other females, though the overlap is more pronounced with TA. Thus, the results illustrate the instability of dendrograms for slight changes in L-function values and the relative stability of block diagonal matrices. This stability gives behavioral ecologists more confidence when investigating cluster-level hypotheses.

The 2001-2002 matrices allow behavioral ecologists to define clusters (female neighborhoods) for further analyses. As subfigure 3(d) illustrates, there are two clearly defined neighborhoods for this era: (i) TA, FF, FN, GA, and GM and (ii) SA, SI, SR, SW, and TT. Note that PI appears to centrally located with a slightly higher degree of overlap to neighborhood (i). These neighborhoods persist at 250m but become less prominent at 400m. This shows that most females are globally clustered at that distance of 400m while they show local neighborhoods at smaller distances like 100m/250m. Similar to the 1979-1982 era, there are some peripheral females like TZ, HO, JF, and YD in 2001-2002.

HO remains peripheral at all distances which illustrates that her space use is markedly different from that for all other community females.

7 Conclusions and future work

This paper shows the application of spatial statistics and data mining techniques to an ecological dataset. The main question studied in this paper requires techniques for clustering of marked spatial point patterns. Hence, two clustering algorithms, SPACE-1 and SPACE-2, were proposed. Both the approaches used the bivariate K-function (L-function) as a dissimilarity measure. Experimental results show that the former method always gives neighborhoods for female chimpanzees. However, there is a some instability in the results due to the sensitivity of the dendrograms in the former approach. The latter approach provides a better, more stable picture of clustering among the chimpanzees. Overall, these techniques enable behavioral ecologists to study the effect of distance on the spatial clustering of female chimpanzees. While these techniques were developed to analyze space use among female chimpanzees, they will be broadly applicable to other species showing territorial space usage.

Future research directions include using inhomogeneous K-function and extending the analysis to spatio-temporal domain. As evidenced by the diagonal block matrices, most females exhibit high degree of self-clustering. This suggests that they concentrate their space use to small areas of the larger community range. Further investigation of female clustering using an inhomogeneous K-function is required. Primate socio-ecological models predict that female space use is primarily determined by the distribution of food resources. We are therefore currently developing spatial covariates (like vegetation quality surface) for obtaining an intensity function for the inhomogeneous Poisson process. In addition to this extension, we are also incorporating temporal information in order to test the hypothesis that "females may overlap spatially but avoid each other by occupying temporally disjoint core areas."

8 Acknowledgments

This research was supported by NSF Grant No. IIS-0431141. The authors also thank The Jane Goodall Institute (http://www.janegoodall.org/) for the availability of data for this research.

References

[1] J. E. Besag. Comments on Ripley's paper. *Journal of the Royal Statistical Society B*, 39(2):193–195, 1977.

[2] D. Bygott. Agonistic behavior, dominance, and social structure in wild chimpanzees of the Gombe National Park. In D. Hamburg and E.McCown, editors, *The Great Apes*, pages 405–427. Benjimani/Cummings, Menlo Park, CA, 1979.

[3] N. A. Cressie. *Statistics for Spatial Data*. Wiley: New York, 1993.

[4] P. J. Diggle. *Statistical Analysis of Spatial Point Patterns, 2nd Edition*. Arnold: London, 2003.

[5] P. J. Diggle and T. Cox. On sparse sampling methods and tests of independence for multivariate spatial point patterns. *Bulletin of the International Statistical Institute*, 49:213–229, 1981.

[6] A. George and W. Liu. *Computer Solution of Large Sparse Positive Definite Systems*. Prentice-Hall series in computational mathematics, 1981.

[7] J. Goodall. *The Chimpanzees of Gombe: Patterns of Behavior*. Harvard University Press, Cambridge, MA, 1986.

[8] J. Han, M. Kamber, and A. K. H. Tung. Spatial clustering methods in data mining: A survey. In H. Miller and J. Han, editors, *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, 2001.

[9] T. Hasegawa. Sex differences in ranging patterns. In T. Nishida, editor, *The Chimpanzees of Mahale*, pages 100–114. University of Tokyo Press, Tokyo, 1990.

[10] H. W. Lotwick and B. W. Silverman. Methods for analysing spatial processes of several types of points. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(3):406–413, 1982.

[11] T. Nishida. Social interactions between resident and immigrant female chimpanzees. In P. Heltne and L. Marquardt, editors, *Understanding Chimpanzees*, pages 68–89. The Chicago Academy of Science, Cambridge, MA, 1989.

[12] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-07-0.

[13] B. D. Ripley. The second-order analysis of stationary point processes. *Journal of App. Prob.*, 13:255–266, 1976.

[14] B. D. Ripley. *Statistical inference for spatial processes*. Cambridge University Press, 1988.

[15] B. S. Rowlingson and P. J. Diggle. SPLANCS: spatial point pattern analysis code in S-Plus. *Comput. Geosci.*, 19(5):627–655, 1993.

[16] M. Schroeder, D. Gilbert, J. van Helden, and P. Noy. Approaches to visualisation in bioinformatics: from dendrograms to space explorer. *Information Sciences: an International Journal*, 139(1):19–57, 2001.

[17] S. Shekhar, P. Zhang, Y. Huang, and R. Vatsavai. Trends in spatial data mining. In H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha, editors, *Data Mining: Next Generation Challenges and Future Directions*. AAAI/MIT Press, 2003.

[18] J. Williams, A. Pusey, J. Carlis, B. Farm, and J. Goodall. Female competition and male territorial behavior influence female chimapnzees ranging patterns. *Animal Behavior*, 63, 2002.

[19] R. Wrangham, A. Clark, and G. Isagirye-Basuta. Female social relationships and social organization of Kibale Forest chimpanzees. In T. Nishida, W. McGrew, P. Marler, M. Pickford, and F. de Waal, editors, *Topics in Primatology: Human Origins*, pages 81–98. University of Tokyo Press, Tokyo, 1992.