Spatial Clustering Of Chimpanzee Locations For Neighborhood Identification *

Sandeep Mane[‡] †, Carson Murray[¶§], Shashi Shekhar[‡], Jaideep Srivastava[‡] and Anne Pusey[¶§]
[‡] Dept. of Computer Science, [§] Dept. of Ecology, Evolution & Behavior,
[¶] The Jane Goodall Institute's Center for Primate Studies,
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 more than 40 years and is being 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 their spatial locations ?" 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 found that Ripley's K-function provides a powerful statistical tool for evaluating clustering behavior among spatial point patterns. We then proposed two clustering approaches for marked point patterns using the 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 proposed methods can be extended to also include temporal information.

1 Introduction

In 1960, Jane Goodall began the first long-term field study of chimpanzees, at Gombe National Park, Tanzania. This study continues today and has greatly increased our understanding of chimpanzee behavior and the evolution of their social structure. 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 result from 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.

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. 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 K-function with the complete-link clustering algorithm (hierarchical clustering) while the second approach uses the 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 remainder of this paper is organized as follows – section 2 summarizes the main contributions of this paper. Section 3 provides domain background and explains the main hypothesis of interest to behavioral ecologists. Section 4 summarizes related work. Section 5 explains two proposed approaches to address the problem of spatial clustering of point patterns. Section 6 shows some (ecologically) interesting experimental results. Section 7 concludes this paper.

*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.

†Primary contact: smane@cs.umn.edu

2 Key contributions of this paper:

The main contributions to data mining are:

(i) *Clustering techniques for marked spatial point processes:* The 'marked spatial point pattern' characteristic of this dataset, makes it difficult to directly apply traditional spatial clustering methods (e.g. K-means) and to the best of our knowledge, no prior work has been done to address clustering in such datasets.

(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.

(iii) *Groundwork for extending to spatio-temporal analysis:* This paper lays the groundwork for future work of spatio-temporal clustering of spatial point processes.

3 Domain background

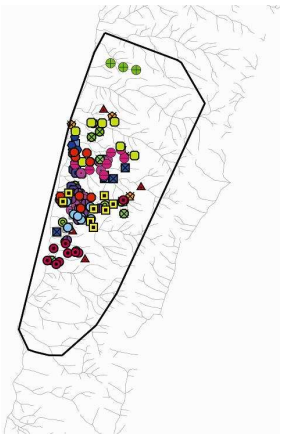
3.1 Study site and data collection:

Jane Goodall began the Gombe Stream chimpanzee research project in 1960. Gombe is a small park (35 km^2) in which there are currently 3 chimpanzee communities. 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 [3]). They note point samples at 15-minute intervals, and record information like group composition, sexual states for all females, location, and feeding of the focal. Ranging and association patterns are derived for these individuals based on where they are encountered during a follow (Williams et al. [6]). From 1974-2002, the central Kasekela community contained between 20-36 adult chimpanzees with 12-24 adult females and 7-18 adult males.

3.2 Chimpanzee social structure & space use

Chimpanzees are a highly gregarious species with transient fission-fusion groupings within a permanent community (Goodall [3]). The sociability and ranging patterns of male and female chimpanzees differ distinctively. Females are less social than males, concentrate their use in subsets of the community range, and have a subtler dominance hierarchy (Goodall [3]). Females can adopt three different space use strategies: immigration into a community, remaining within her 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. These core areas overlap substantially (Williams et al. [6]). Figure 1 illustrates the high degree of overlap of core area points for females between 2001-2002. Once they have established a core area, females demonstrate a high level of site fidelity (Williams et al [6]).

Figure 1. Core area points for adult female chimpanzees - 2001-2002



Previous research from Gombe has reported that female core areas are clustered into neighborhoods (Williams et al., [6]). "*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. Understanding such female space use is particularly crucial because it influences reproductive success (Williams et al. [6]) and because it is thought to determine male distribution, intergroup aggression, and mating systems.

4 Related work:

Analysis of spatial point patterns is an active field of study in spatial statistics. However, to the best of our knowledge, little research has been done in clustering marked spatial point pattern analysis (Han et al. [4]). Chimpanzee studies have often categorized females as "northern", "southern", etc. While these terms imply a neighborhood distribution, they were based on visual estimates of data. Clustering of chimpanzees into neighborhoods was first defined mathematically by Williams et al. [6]. However, their research did not allow for the study of how clustering varies with distance, i.e. global vs. local interactions among chimpanzees ? Also, the use of dendrograms is of concern due to the inherent sensitivity of ordering 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 neighborhoods (stable clusters) exist 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 ?"

5.2 Dissimilarity measure

We use cell-count statistics approach (Cressie [1]), since it provides a good means by which to assess the variation of interaction effects with distance. Also, statisticians generally believe that it provides a better, rigorous statistical analysis. Second-order properties of spatial point pattern describe the covariance (or correlation) between values of the spatial point pattern at different regions in space. The K-function is one such measure which 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 and it allows one to describe the characteristics of point processes at many different scales.

Definition 1 *The K-function (Cressie [1]), at a distance 'r' for a bivariate spatial point pattern is defined as –*
$K_{ij}(r) = \lambda_j^{-1}$ *E(number of type j events within distance r of a randomly chosen event of type i)*
where, λ_j *is intensity of spatial point process with marks j.*

Mathematically, suppose that we observe two marked point process over a plane region D having area A, with the observed x-points being $x_1, \ldots, x_n$, and the observed y-point process being $y_1, \ldots, y_m$, then the unbiased estimate for K-function without any edge corrections is –

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

where $I_r(d_{ij})$ is an indicator function which is zero if distance between i and j is greater than r. Otherwise, it is the reciprocal of proportion of circumference of the circle centered at i with radius d(x_i,x_j) that lies within the sampling window D. In case of complete spatial randomness (CSR), the expected value $\widehat{K}_{xy}(r)$ is πr^2. If the observed $\widehat{K}_{xy}(r) > \pi r^2$, then the point process are said to be clustered at distance r while if $\widehat{K}_{xy}(r) < \pi r^2$, then the point processes are said to show repulsion. 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. Here, the indicator function $I_r(d_{ij})$ used in the unbiased estimate of K-function captures the edge corrections required, if any. Since the scale for $\widehat{K}_{xy}(r)$ is not linear in r, Besag's L-function (Cressie [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 2 *The L-function is thus defined as –*

$$L_{xy}(r) = \sqrt{\frac{K_{xy}(r)}{\pi}} - r \qquad (2)$$

If the estimate $\widehat{L}_{xy}(r)$ is positive, 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. $\widehat{L}_{xy}(r)$ equal to zero represents CSR.

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

The first algorithm uses the MAX or complete-link clustering algorithm as it is less susceptible to noise (Han et al. [4]). The L-function, computed using the unbiased estimator of K-function, is used as the dissimilarity measure. Also, though it favors globular shapes, it is not a problem here, since most core areas for females are globular.

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, dendrograms can show instability or sensitivity for minor variations in dissimilarity values. Also, they assume a hierarchical structure to the dataset and 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 also use matrix block diagonalization techniques for clustering $\widehat{L}_{m_i m_j}(r)$ 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, here we use the RCM ordering algorithm (George and Liu [2]) 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. Details of above two algorithms can be found in Mane et al. [5].

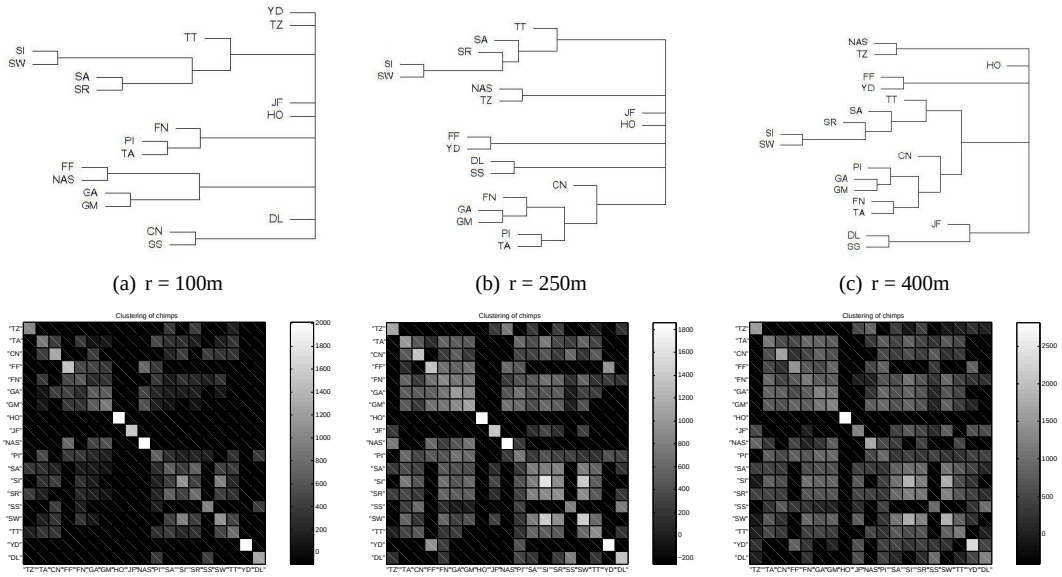
6 Experimental results

For our analysis, we use the locations for each female chimpanzee within its core area for 2001-2002. In order to establish a core area, we use a 50% usage kernel of "alone" locations for each female (similar to Williams et al. [6]). A female is "alone" so long as she is sexually non-receptive and no other unrelated adult chimpanzees arrive into the same follow within five minutes. 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. For complete-link clustering, we use the "hclust" method in R statistical language (http://www.r-project.org) 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) of 100m, 250m, and

400m based on domain (biologically meaningful) knowledge. The minimum distance (100m) reflects locational inaccuracy present in the dataset. Given the topology of Gombe National Park, we considered 400m to be the maximum distance at which chimpanzees can communicate or gain knowledge about the location of conspecifics. 250m represents the midpoint between these.

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



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

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

The dendrograms illustrate the difficulties of determining clusters (neighborhoods) at different distances (Fig. 2(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. 2(d),(e),(f)) illustrates that FF is clustered to a group of individuals (NAS, GA, GM) at all distances. 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 2(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 when most females are globally clustered.

7 Conclusions

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, using bivariate L-function as a dissimilarity measure. Experimental results show that the latter approach gives more stable results across different distances. Overall, these techniques enable behavioral ecologists to study the effect of distance on the spatial clustering of female chimpanzees. Future research directions include using inhomogeneous K-function and extending the analysis to spatio-temporal domain.

References

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

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

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

[4] 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.

[5] S. Mane, C. Murray, S. Shekhar, J. Srivastava, and A. Pusey. Spatial clustering of chimpanzee locations for neighborhood identification. Technical report, Dept. of Computer Science, University of Minnesota, Minneapolis, USA, 2005.

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