The Korean Communications
in Statistics Vol. 12 No. 1, 2005
pp. 207-216

Simple Recursive Approach for Detecting Spatial Clusters

Jeongjin Kim[1], Younshik Chung[2], Sungjoon Ma[3], and Tae Young Yang[4]

Abstract

A binary segmentation procedure is a simple recursive approach to detect clusters and provide inferences for the study space when the shape of the clusters and the number of clusters are unknown. The procedure involves a sequence of nested hypothesis tests of a single cluster versus a pair of distinct clusters. The size and the shape of the clusters evolve as the procedure proceeds. The procedure allows for various growth clusters and for arbitrary baseline densities which govern the form of the hypothesis tests. A real tree data is used to highlight the procedure.

Keywords : Bayesian information criterion; binary segmentation procedure; rectangular cluster

1. Introduction

A binary segmentation procedure is a recursive binary partitioning tool. This paper focuses on the procedure to detect multiple clusters and inferences of spacial problems, when we don't have knowledge about the study space covered by the clusters and the number of clusters. One of the characteristics of the procedure is that the size and the shape of the cluster vary as the procedure proceeds. Therefore the procedure can detect clusters of any size, located anywhere in the space. Our procedure also locates clusters of high and low rate simultaneously.

The proposed binary segmentation procedure for spatial data involves a sequence of nested hypothesis tests of a single cluster versus a pair of distinct clusters. For each test, the null hypothesis of a single cluster implies that the data within the region arise from a common density. For the alternative hypothesis, we split the region into the two 'most distinct' clusters and assume distinct densities for each. If the test suggests the alternative hypothesis, the region is split accordingly. For each resulting cluster, splitting and testing continue until no

1) Professor, Department of Mathematics, Myongji University, Yongin, Kyonggi, 449-728, Korea
E-mail : jjkim@mju.ac.kr
2) Professor, Department of Statistics, Pusan National University, Busan 609-735, Korea
3) Graduate student, Department of Mathematics, Myongji University, Yongin, Kyonggi, 449-728, Korea
4) Professor, Department of Mathematics, Myongji University, Yongin, Kyonggi, 449-728, Korea

more clusters are found. If at any stage, the test suggests the null hypothesis, we then estimate the density in that cluster. The manner in which a region is split is somewhat flexible; it is dictated by the specification of a growth cluster. The size and the shape of the clusters evolve as the procedure proceeds. The procedure also allows for arbitrary baseline densities which govern the form of the hypothesis tests. In this paper, we illustrate the use of rectangular growth clusters and Poisson densities.

In the testing step, we consider a finite number of rectangular clusters, and choose the two distinct clusters for which the likelihood is maximized. Testing can be carried out using the Bayesian information criterion (BIC) (Schwarz 1978). However, as derived by Raftery(1995), the asymptotic theories for the BIC do not apply when discrete parameters are considered as here. Instead, we provide the pseudo-BIC for accounting this problem, and it works well in practice even if it is not theoretically justified. We provide a supporting evidence through a simulation study. Once the clusters are obtained, the pseudo-BIC is calculated. If the pseudo-BIC is positive, then the null hypothesis is rejected and the region is split accordingly.

An alternative general procedure for classifying data into categories is the method of classification and regression trees (CART); see Breiman et al. (1984). In the tree-based approach, it is generally considered appropriate to first partition the data completely, and second, to prune segmentations based on some cost-complexity measure. However, the binary segmentation procedure terminates the partitioning when the partition process fails to attain a threshold of some target criterion. An advantage of the binary segmentation approach over CART is its simplicity with respect to computation. Yang (2005) provides a tree-based method for grouping multinomial data according to their classification probability vectors. The tree-based model is illustrated on grouping many DNA sequences.

Another general approach which can be used in partitioning problems is mixture modelling. Mixture modelling requires the specification of parametric models whereas the recursive approaches considered in this paper are often described as nonparametric. When the number of components is unknown (which is the case in the problems considered here), mixture modelling becomes more challenging and often requires Markov chain Monte Carlo (MCMC) methods for parameter estimation. An introduction to mixture modelling is given by Titterington, Smith and Makov (1985), Kim and Mallick (2002), van Dyk and Hans (2002) and Shlattmann, Gallinat and Bohning (2002) provide examples of mixture modelling approaches.

Binary segmentation procedures have been considered by various authors. Scott and Knott (1974), and Chen and Gupta (1997) developed methods to split normal data into homogeneous groups. Subsequently, Braun and Müller(1998), Yang and Kuo (2001), Yang (2004), and Yang and Swartz (2005) developed binary segmentation procedures for locating change points with respect to DNA sequencing, homogeneous Poisson processes, sporting performances, spatial intensity and quantal response curves respectively. Consistency issues related to binary segmentation have been studied by Vostrikova (1981) who proved consistency for locating the number of change points in a multi-dimensional random process under mild conditions. Venkatraman (1992) addressed consistency issues for the procedure when the change points

are allowed to approach one another.

In Section 2, the procedure is developed using rectangular growth clusters with Poisson data. In Section 3, the approach is illustrated using the longleaf-pine data (Cressie, 1993). A simulation study is also carried out to investigate the performance of the pseudo-BIC.

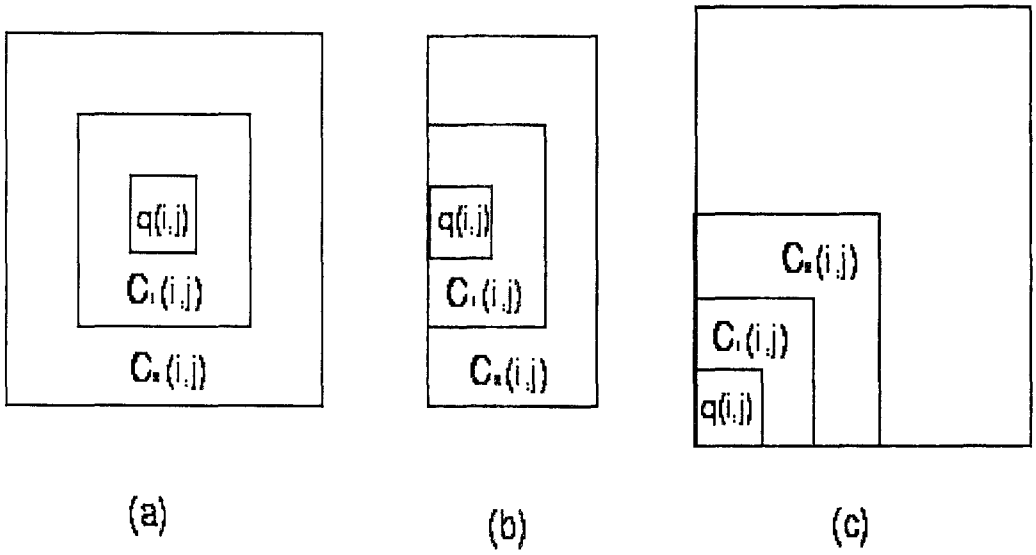

Figure 1 : Agglomeration methods for combining quadrats into rectangle clusters.

2. Binary Segmentation Procedure

We divide the study space into an $I \times J$ grid of quadrats $q(i,j)$ where $i = 1, \cdots, I$ and $j = 1, \cdots, J$. The quadrats serve as the building blocks which are combined and ultimately form the various clusters in the study space. Let $C_r(i,j)$ denote the rectangular growth cluster of size r centered at $q(i,j)$. More precisely, $C_r(i,j)$ is the collection of quadrats

$$C_r(i,j) = \{q(s,t) : \max(1, i-r) \le s \le \min(I, i+r) \cap \max(1, j-r) \le t \le \min(J, j+r)\}.$$

Graphs (a) to (c) in Figure 1 illustrate $C_1(i,j)$ and $C_2(i,j)$ centered at several locations $q(i,j)$. Note that when the boundary of a region is irregular (i.e. not a rectangle), growth clusters for the region are restricted to quadrats that lie within the region. Hence, when the region is irregular, some growth clusters will not be rectangular. Also, since there are only a finite number of quadrats, there are only a finite number of growth clusters.

For every quadrat $q(i,j)$, there is a count $z(i,j)$ corresponding to the number of events occurring in the quadrat. We assume that $z(i,j)$ follows a Poisson distribution. Let V be a set of all district quadrats and $n(V)$ be a number of all district quadrats within the region of interest. For each potential cluster, we consider two models $z(s,t) \sim \text{Poisson}(\lambda_1)$ if $q(s,t)$ belongs to the corresponding cluster and $z(s,t) \sim \text{Poisson}(\lambda_2)$ otherwise. We calculate the maximum likelihood under the null hypothesis $H_0: \lambda_1 = \lambda_2 = \lambda_0$ and the maximum likelihood under the alternative hypothesis $H_1: \lambda_1 \neq \lambda_2$. The likelihood under H_0 is proportional to $L_0(\lambda_0) = \lambda_0^{\sum_{(i,j)\in V} z(i,j)} \exp(-n(V)\lambda_0)$ which is maximized at

$$\hat{\lambda}_o = \sum_{(i,j)\in V} \frac{z(i,j)}{n(V)} . \tag{1}$$

Under H_1, the likelihood is proportional to

$$L_1(C_r(i,j), \lambda_1, \lambda_2) = \prod_{(s,t)\in V\cap C_r(i,j)} (\lambda_1^{z(s,t)} \exp(-\lambda_1)) \times \prod_{(s,t)\in V\setminus C_r(i,j)} (\lambda_2^{z(s,t)} \exp(-\lambda_2)).$$

For fixed $C_r(i,j)$, the profile likelihood $L_1(C_{r(i,j)}, \lambda_1, \lambda_2)$ is maximized at

$$\hat{\lambda}_1 = \frac{\sum_{(s,t)\in V\cap C_r(i,j)} z(s,t)}{n(V\cap C_r(i,j))} \quad \text{and} \quad \hat{\lambda}_2 = \frac{\sum_{(s,t)\in V\setminus C_r(i,j)} z(s,t)}{n(V\setminus C_r(i,j))},$$

where $n(V\cap C_r(i,j))$ and $n(V\setminus C_r(i,j))$ are the number of quadrats belonging and no-belonging to $C_r(i,j)$ respectively. The fully maximized likelihood $L_1(\hat{C}, \hat{\lambda}_1, \hat{\lambda}_2)$ is then obtained by maximizing the profile likelihood over $i = 1, \cdots, I$, $j = 1, \cdots, J$ and r. We have discretized indices i, j and r which yield a finite search.

The asymptotic theories for the BIC do not apply when the discretized rectangular cluster is considered as a parameter. We propose a pseudo-BIC to account for this problem;

$$\text{BIC} = \log L_1(\hat{C}, \hat{\lambda}_1, \hat{\lambda}_2) - \log L_0(\hat{\lambda}_0) - \frac{1}{2}(q_1 - q_0)\log\left(\sum_{(i,j)\in V} z(i,j)\right) \tag{2}$$

where a positive (negative) value determines $H_1(H_0)$. We provide a supporting evidence of the pseudo-BIC through a simulation study. The third term in (2) is a penalty function which adjusts for the difference in dimensionality between the two models. In this application, $q_0 = 1$ and we set $q_1 = 5$ with respect to $\hat{C}$, $\hat{\lambda}_1$ and $\hat{\lambda}_2$.

If H_0 is accepted, then a final cluster has been determined which includes all of the quadrats within V. However, if H_0 is rejected, the data set is divided into quadrats which lie in $\hat{C}$ and quadrats which lie outside of $\hat{C}$. The testing procedure is then carried out on each of the two subregions. The algorithm continues in this fashion and terminates when no more splitting takes place. Whenever a test suggests the null hypothesis, we estimate $\hat{\lambda}_0$ as

in (1).

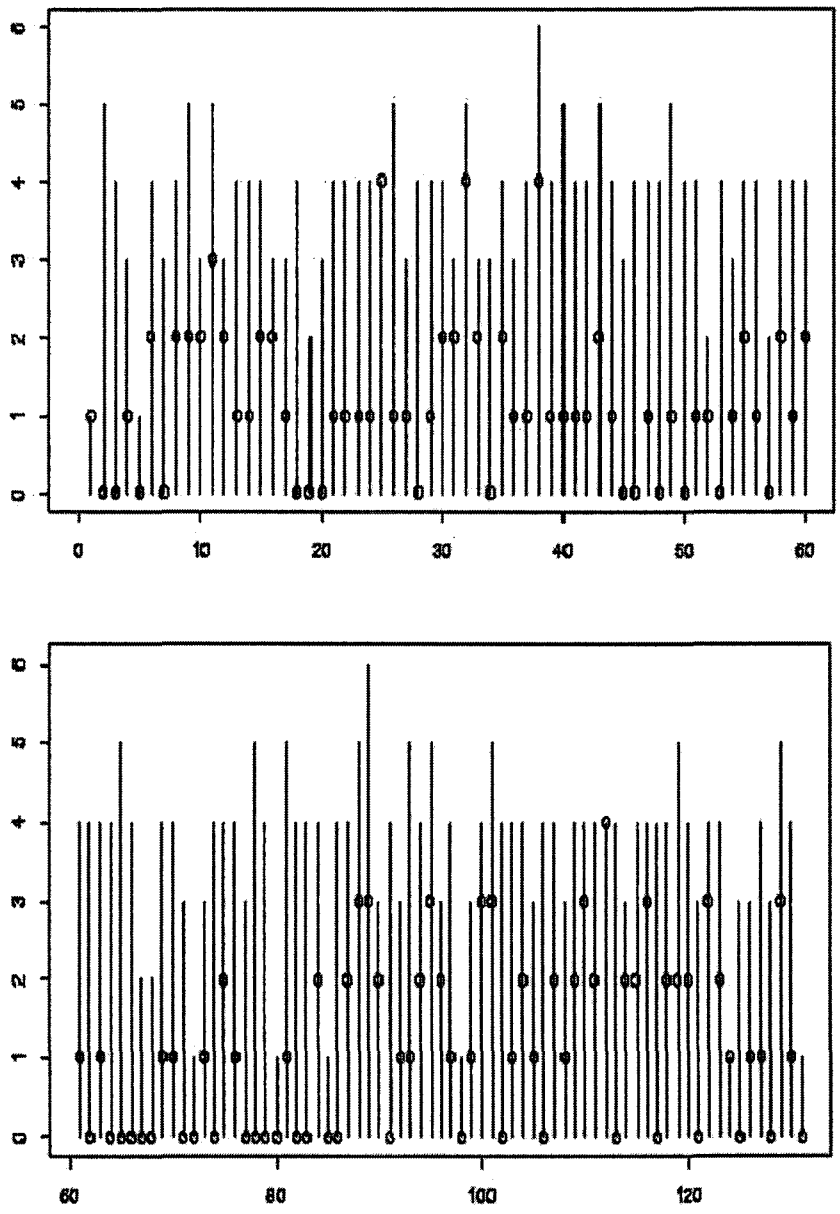

Figure 2: Simulated successes (o) with number of trials (|).

3. Numerical Examples

3.1 Simulation result for the pseudo-BIC

We note that the asymptotic theories for the BIC do not apply to discretized parameters. We provide a supporting evidence of the pseudo-BIC for a discrete changepoint case through a simulation study. We note that the pseudo-BIC in (2) works well in practice even if it is not theoretically justified.

We use the IMSL RNBIN routine to generate random successes from a binomial density with the same number of at-bats as in Javy Lopez's 1998 Major League Baseball season (Yang 2004). To investigate the performance of the binary segmentation procedure based on the pseudo-BIC, we set the success rate

$$p(t) = 0.35\ I(t \in [1, 44]) + 0.2\ I(t \in [45, 86]) + 0.4\ I(t \in [87, 131]),$$

where $I(E)$ is the indicator function of event E. There are two change points, at the 44th trial and the 86th trial, and three associated success rates; 0.35, 0.20 and 040. The simulated data are plotted in Figure 2, where the game number is plotted against the number of successes with the number of trials.

In Figure 3, we present the step by step results of the binary segmentation procedure using the pseudo-BIC for splitting the data. In this case, change points are obtained at the 44th and 86th trials with associated success rates 0.35, 0.16, and 0.42. This agrees fairly well with the underlying model.

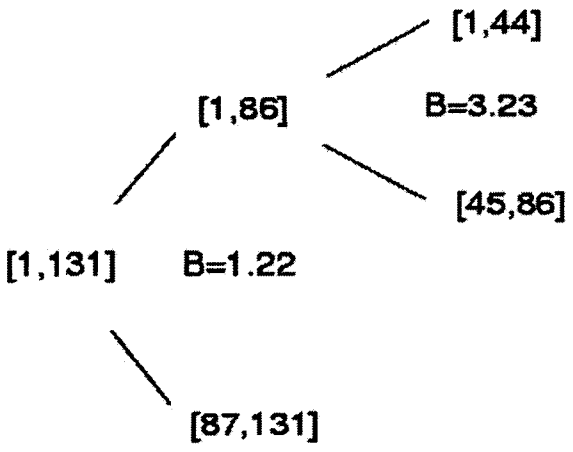

Figure 3 : Results of the binary segmentation procedure applied to the simulated data using the pseudo-BIC criterion

3.2 Real data analysis

We consider the spatial clustering of longleaf-pine trees located in a 4 hectare region (200 meters by 200 meters) in Thomas County, Georgia. The data are taken from chapter 8 of Cressie (1993). The circles in Figure 4 give the locations of 584 trees that are at least 2 cm in diameter when measured at breast height. An interesting problem is to determine clusters of these trees with respect to their frequency in the study space.

Table 1 : Clusters of quadrats for the longleaf-pine data.

	Notation of cluster	Quadrats	Shape of cluster
cluster A	$C_4(17,16)$	$\{q(i,j): 13 \le i \le 21 \cap 12 \le j \le 20\}$	square
subcluster A1	$C_2(14,13)$	$\{q(i,j): 13 \le i \le 16 \cap 12 \le j \le 15\}$	square
subcluster A2	$C_2(17,20)$	$\{q(i,j): 15 \le i \le 19 \cap 18 \le j \le 20\}$	rectangle
cluster B	$C_8(4,17)$	$\{q(i,j): 1 \le i \le 12 \cap 11 \le j \le 25\}$	rectangle
cluster C	$C_2(22,14)$	$\{q(i,j): 22 \le i \le 24 \cap 12 \le j \le 16\}$	rectangle
subcluster C1	$C_2(22,15)$	$\{q(i,j): 22 \le i \le 24 \cap 13 \le j \le 16\}$	rectangle
cluster D	$C_{10}(11,1)$	$\{q(i,j): 1 \le i \le 21 \cap 1 \le j \le 11\}$	rectangle
cluster E	$C_2(25,24)$	$\{q(i,j): 23 \le i \le 25 \cap 22 \le j \le 25\}$	rectangle

We divide the forest into a 25×25 grid of quadrats $q(i,j)$ where $i = 1, \cdots, 25$ and $j = 1, \cdots, 25$. Hence the quadrats are squares of size $8\,\mathrm{m} \times 8\,\mathrm{m}$, and the average number of trees per quadrat in the study space is 0.93. Figure 4 also displays the final clusters obtained using the proposed binary segmentation procedure. We note that we have applied the procedure to finer grids (e.g. 30×30, 40×40 and 30×40) and have obtained similar results.

The steps in the binary segmentation procedure are illustrated in Figure 5. Using rectangular growth clusters, the first cluster (cluster A) is identified with the largest BIC value (26.1) amongst all candidate clusters. Therefore the study space is tentatively divided into cluster A and the complement of cluster A within the study space. The procedure continues within each of the two subregions as follows: (i) Outside of cluster A, cluster B is identified with the largest BIC value (19.4) amongst all candidate clusters. Therefore the complement of cluster A is tentatively divided into cluster B and the remaining space outside of both cluster A and cluster B. (ii) Inside cluster A, cluster A1 is identified with the largest BIC value (5.3) amongst all candidate clusters. Therefore cluster A is tentatively divided into cluster A1 and the complement of cluster A1 within cluster A. We continue in this fashion until no more splits are accepted. Table 1 provides a detailed description of the final clusters

obtained using the binary segmentation procedure. Figure 6 displays the estimated Poisson rates for each of the final clusters. We observe considerable differences in the cluster rates when compared to the overall rate 0.93.

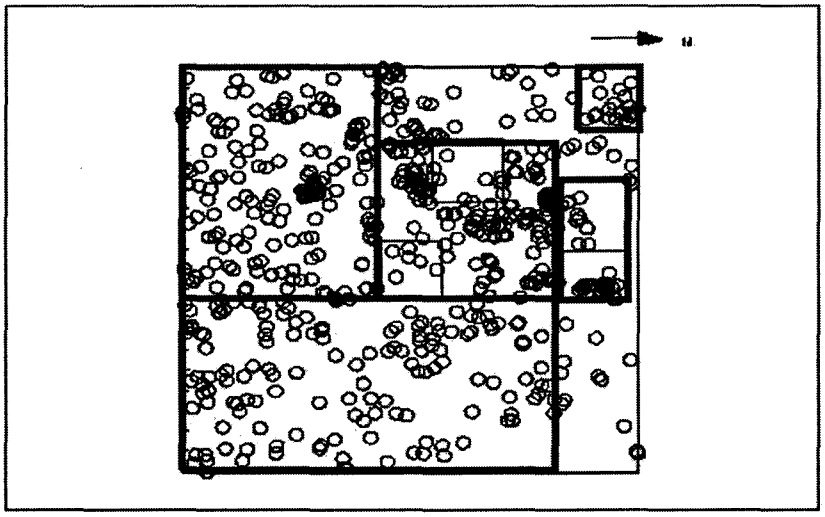

Figure 4 : Locations of 584 trees in the 4 hectare study region with the clusters and subclusters obtained by the binary segmentation procedure.

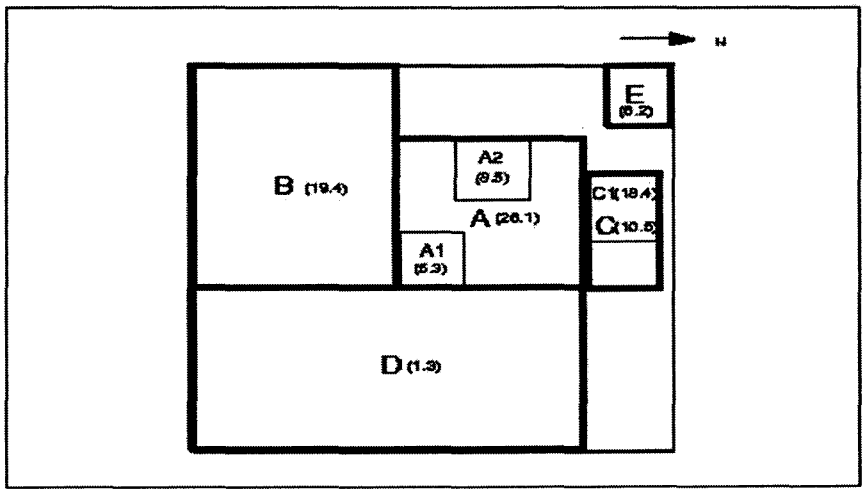

Figure 5 : Sequential clusters from A to E with the corresponding BIC values (values inside parentheses) from the binary segmentation procedure. Cluster A includes subclusters A1 and A2 sequentially, and Cluster C includes subcluster

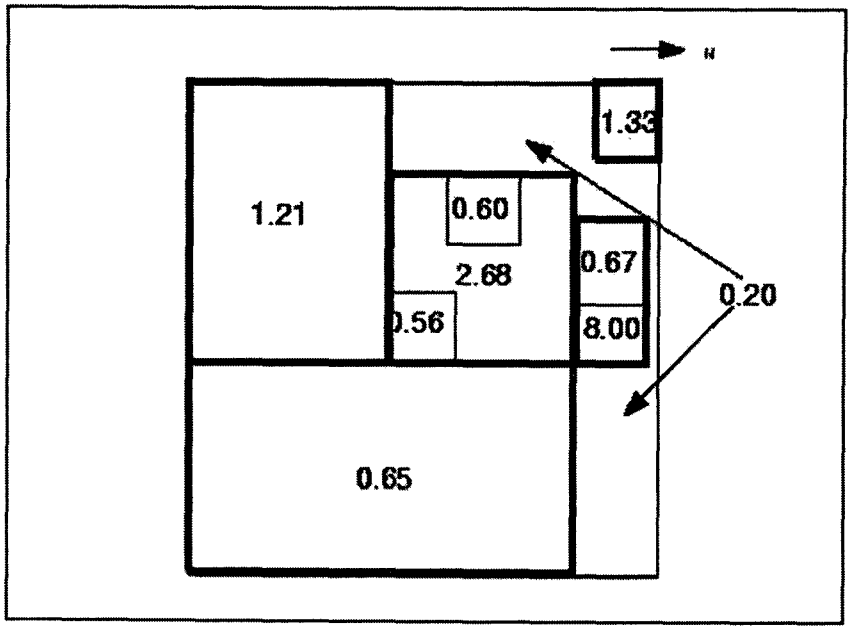

Figure 6 : Estimated Poisson rates for clusters obtained on the longleaf-pine data using the binary segmentation procedure.

Reference

[1] Breiman, L., Friedman J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. Wadworth and Brooks/Cole, Monterey.

[2] Braun, J.V. and Müller, H. (1998). Statistical methods for DNA sequence segmentation. *Statistical Science* 13, 142-162.

[3] Chen, J. and Gupta, A.K. (1997). Testing and Locating Variance Change points with Application to Stock Prices. *Journal of the American Statistical Association* 92, 739-747.

[4] Cressie, N. (1993). *Statistics for spacial data, second edition.* Wiley, New York.

[5] Kim, H. and Mallick, B.K. (2002). Analyzing spatial data using skew-Gaussian processes. In Spatial Cluster Modelling, A. Lawson and D. Denison (editors). Chapman and Hall, London, 163-173.

[6] Raftery, A. (1995). Bayesian model selection in social research. *Sociological methodology*, ed. P. Marsden, Blackwell, Oxford.

[7] Schlattmann, P., Gallinat, J., and Bohning, D. (2002). Spatio-temporal partition modelling: an example from neurophysiology. In Spatial Cluster Modelling, A. Lawson and D. Denison (editors). Chapman and Hall, London, 227-234.

[8] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6,

461-464.

[9] Scott, A.J. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics* 30, 507-512.

[10] Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley: New York.

[11] van Dyk, D.A. and Hans, C.M. (2002). Accounting for absorption lines in images obtained with the Chandra X-ray Observatory. In Spatial Cluster Modelling, A. Lawson and D. Denison (editors). Chapman and Hall, London, 175-198.

[12] Venkatraman, E.S. (1992). Consistency results in multiple change-point situations. Manuscript, Department of Statistics, Stanford University.

[13] Vostrikova, L.J. (1981). Detecting 'disorder' in multidimensional random processes. *Soviet Mathematics Doklady* 24, 55-59.

[14] Yang, T.Y. and Kuo, L. (2001). Bayesian binary segmentation procedure for a Poisson process with multiple changepoints. *Journal of Computational and Graphical Statistics* 10, 772-785.

[16] Yang, T.Y. (2004). Bayesian binary segmentation procedure for detecting streakiness in sports. *Journal of the Royal Statistical Society Series A* 167, 627-637.

[17] Yang, T.Y. (2005). A tree-based model for homogeneous groupings of multinomials. *Statistics in Medicine,* in press.

[15] Yang, T.Y. and Swartz, T. (2005). Applications of binary segmentation to the estimation of quantal response curves and spatial intensity. *Biometrical Journal,* in press.

[Received December 2004, Accepted March 2005]