Biometrical Journal **47** (2005) 4, 489–501 DOI: 10.1002/bimj.200310136

Applications of Binary Segmentation to the Estimation of Quantal Response Curves and Spatial Intensity

Tae Y. Yang[*,1] and **Tim B. Swartz**[2]

[1] Department of Mathematics, Myongji University, Yongin, Korea 449-728
[2] Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby BC, Canada V5A1S6

Received 7 September 2003, revised 14 October 2004, accepted 28 October 2004

Summary

This paper explores the use of binary segmentation procedures in two applications. The first application is concerned with the estimation of nonparametric quantal response curves. With Bernoulli data and an assumed monotone increasing curve, this gives rise a change-point model where the change points are determined using a sequence of nested hypothesis tests of whether a change point exists. The second application concerns cluster identification and inference for spatial data where the shape of the clusters and the number of clusters is unknown. The procedure involves a sequence of nested hypothesis tests of a single cluster versus a pair of distinct clusters. Examples of both applications are provided.

Key words: Akaike information criterion; Bioassay; Circular growth clusters.

1 Introduction

Without being too precise, there are various estimation problems in statistics where the goal is to identify 'commonality'. For example, in epidemiological studies, it is important to identify regions of safety and regions of risk. These problems are typically challenging due to the multivariate nature of the data which leads to complex and highly parametrized likelihoods.

Binary segmentation is a recursive partitioning tool which may lead to simple solutions for such problems. Roughly speaking, a binary segmentation procedure begins by tentatively dividing data into two parts. The hypothesis of commonality between the two parts is tested and the procedure terminates if commonality exists. If commonality does not exist, then the division takes place and the procedure is continued on each of the two parts. It is the simplification of the problem to nested hypothesis tests that circumvents computational difficulties related to a possibly large and unknown number of change points.

An alternative general procedure for classifying data into categories is the method of classification and regression trees (CART); see Breiman et al. (1984). An advantage of the binary segmentation approach over CART is its simplicity with respect to computation and the lack of tuning parameters. In CART models, there are many variations for the growth and pruning of trees. Also, in the spatial problems considered here, the CART approach does not have a natural way to take into account the spatial structure so that adjacent regions are more likely to be grouped together. CART is primarily intended for problems with many covariates where prediction is the ultimate goal.

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

[*] Corresponding author: e-mail: tyang@mju.ac.kr

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 in spatial statistics.

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) and Yang (2004) developed binary segmentation procedures for locating change points with respect to DNA sequencing, homogeneous Poisson processes and sporting performances. 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.

We continue the development of binary segmentation procedures in two applications. We first consider the estimation of nonparametric quantal response curves. When a stimulus is applied or observed at various levels with respect to a biological organism, we have what is known as the *bioassay* problem. Often, the observed response to the stimulus is binary where the binary response may be the occurence of death, the presence of tumours or any number of meaningful outcomes. In the case of a binary response, a primary objective is the fitting of the *quantal response curve* which is the probability of the binary response expressed as a function of the level of the stimulus. The fitting of quantal response curves has been the focus of considerable statistical activity. In fact, the importance of fitting quantal response curves extends beyond biological settings and includes problems in various domains including sample surveys, econometrics and industrial statistics (Morgan, 1992; Agresti, 2002).

A basic starting point for the quantal response problem involves independent data $y_1, \ldots, y_n$ where the response $y_i \sim \text{binomial}\,(m_i, p_i)$ is associated with the covariate x_i, $i = 1, \ldots, n$. For example, the covariate x_i may be the dosage of a toxic compound given to an experimental animal. We assume that $x_1 < x_2 < \cdots < x_n$. This leads to the n-dimensional likelihood

$$L(p) \propto \prod_{i=1}^{n} \binom{m_i}{y_i} p_i^{y_i} (1 - p_i)^{m_i - y_i}. \tag{1}$$

Early methods which are still in use today and form the backbone for many modern methods rely on a parametric specification of the probabilities $p_1, \ldots, p_n$ (Finney, 1978). For example, one might express $\text{logit}\,(p_i)$ as a linear function of x_i, $i = 1, \ldots, n$. In this case, an advantage of the linear logistic model is the reduction of the dimensionality of the likelihood from n to 2. This renders a convenient likelihood analysis where maximum likelihood estimates are numerically obtained. These classical parametric models have Bayesian analogues where in addition to the parametric form imposed on $p_1, \ldots, p_n$, prior distributions are assigned to the shape parameters (e.g. the slope and intercept). Typically, MCMC algorithms are used to obtain approximate samples from posterior distributions (Gilks, Richardson and Spiegelhalter, 1996).

The offspring of the aforementioned parametric models belong to the extensive class of generalized linear models (McCullagh and Nelder, 1989). These models abandon the binomial assumption and allow for robust inference under a wider class of distributions. Generalized linear models (GLMs) based on quantal data also allow for the possibility of extra-binomial variation (Dean 1998). With good software available, GLMs seem to be the preferred choice in the analysis of categorical data.

There are approaches that do not rely on the parametric form in (1). These include the pool-adjacent-violators algorithm (Ayer et al., 1955) which provides estimates of $p_1, \ldots, p_n$ under the monotonicity constraint $p_1 \leq \cdots \leq p_n$, simple distribution-free procedures (Morgan, 1992), generalized additive models (Hastie and Tibshirani, 1990), methods based on splines (Thisted, 1988) and methods based on the Dirichlet process (Dey, Muller and Sinha, 1998). The related problem of estimating join points in segmented linear regressions has been studied by various authors including Hudson (1966) and Gallant and Fuller (1973). They provide least squares estimates of the join points where continuity at the join points is assumed and the number of join points is specified in advance.

The approach proposed in this paper begins with (1) but avoids parametric assumptions associated with the quantal response curve. We assume a monotone increasing response curve which implies a change-point model. The approach used to determine the change points is based on a binary segmentation procedure that involves a sequence of nested hypothesis tests of a single change point. Thus the procedure circumvents computational difficulties related to a possibly large and unknown number of change points. If a nested test suggests a single change-point model, the change-point is located and the data is split accordingly. For each sub-segment of data, testing continues until no change points are found. Testing is carried out using a modification of the Akaike (1973) information criterion. Other related work based on the detection of change points includes Cochran (1954), Armitage (1955), Tarone (1982), Dempster, Selwyn and Weeks (1983), and Ma et al. (2002).

Our second application concerns the use of binary segmentation to detect clusters and provide inferences for spatial data when the shape of the clusters and the number of clusters are unknown. Unlike the procedure proposed in this paper, most statistical methods for detecting clusters are restricted in the following senses; (1) they detect clustering but do not determine the location of clusters, (2) they provide inferences but do not detect clustering, and (3) they detect clusters of a high incidence rate without detecting clusters of a low incidence rate, or vice versa. Examples of (1) include Whittemore et al. (1987), Diggle and Chetwynd (1991) and the graphical analysis machine (GAM) of Openshaw et al. (1988). As examples of (2), Cuzick-Edwards (1990) k nearest neighbors (kNN) test and the methods of Stone and Diggle (1990) provide inferences about risks associated with prespecified clusters. Examples of (3) include Besag and Newell's (1991) test for the detection of clusters and Turnbull et al. (1990). Other relevant references include Kulldorff and Nagarwalla (1995), Kulldorff (1997) and Gangnon and Clayton (2001). Cressie (1993) is a comprehensive source on the statistical analysis of spatial data and Kulldorff (www.satscan.org) provides free software for the analysis of spatial data.

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 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 circular growth clusters and binomial densities. In the testing step, we identify the two distinct clusters for which the likelihood is maximized. Once the clusters are identified, a modification of the Akaike (1973) information criterion is calculated. If the criterion is positive, then the null hypothesis is rejected and the region is split accordingly. Our approach is similar to the Bayesian partition model of Holmes, Denison and Mallik (1999). In their model, they split the study space into a series of regions, and assume that the data arises from a common density within each region. Related approaches using circular clusters have been developed by Gangnon and Clayton (2001) and Kulldorff and Nagarwalla (1995). Turnbull et al. (1990) and Besag and Newell (1991) proposed circular clusters of fixed population and case radius respectively. Other related work on partition models for spatial data includes Gangnon and Clayton (2000), Knorr-Held and Rasser (2000) and Denison and Holmes (2001).

In Section 2, the binary segmentation procedure is developed with respect to fitting nonparametric quantal response curves. An example is presented based on insurance mortality data and a simulation study is provided to indicate the performance of the procedure. In Section 3, the binary segmentation procedure is developed with respect to cluster detection and inference using circular growth clusters with binomial data. The approach is illustrated using sudden infant death syndrome data (Cressie, 1993) where the data are aggregated according to districts. Some concluding remarks are given in Section 4.

2 Estimation of Quantal Response Curves

We begin with the likelihood (1) and assume that the binomial probabilities p_i, $i = 1, \ldots, n$ are monotone increasing. In other words, we assume that $p_1 \leq p_2 \leq \cdots \leq p_n$. We let the null hypothesis H_0 denote the constant model with no change points (i.e. $\theta_0 = p_1 = \cdots = p_n$). Under H_0, the likelihood is proportional to

$$L_0(\theta_0) = \theta_0^{\sum_{i=1}^{n} y_i} (1-\theta_0)^{\sum_{i=1}^{n} (m_i - y_i)}$$

which is maximized at $\hat{\theta}_0 = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$.

Let H_1 denote the single change-point model with the change point given by the parameter c. This implies $\theta_1 = p_1 = \cdots = p_c < p_{c+1} = \cdots = p_n = \theta_2$ where $c = 1, \ldots, n-1$. Under H_1, the likelihood is proportional to

$$L_1(c, \theta_1, \theta_2) = \theta_1^{\sum_{i=1}^{c} y_i} (1-\theta_1)^{\sum_{i=1}^{c} (m_i - y_i)} \theta_2^{\sum_{i=c+1}^{n} y_i} (1-\theta_2)^{\sum_{i=c+1}^{n} (m_i - y_i)}$$

which is maximized for fixed $c = 1, \ldots, n-1$ via

$$\left(\hat{\theta}_1(c), \hat{\theta}_2(c)\right) = \begin{cases} \left(\dfrac{\sum_{i=1}^{c} y_i}{\sum_{i=1}^{c} m_i}, \dfrac{\sum_{i=c+1}^{n} y_i}{\sum_{i=c+1}^{n} m_i} \right) & \text{if } \dfrac{\sum_{i=1}^{c} y_i}{\sum_{i=1}^{c} m_i} < \dfrac{\sum_{i=c+1}^{n} y_i}{\sum_{i=c+1}^{n} m_i} \\ (\hat{\theta}_0, \hat{\theta}_0) & \text{otherwise} \end{cases} .$$

The fully maximized likelihood under the single change-point model $L_1(\hat{c}, \hat{\theta}_1(\hat{c}), \hat{\theta}_2(\hat{c}))$ is then obtained by maximizing $L_1(c, \hat{\theta}_1(c), \hat{\theta}_2(c))$ over the finite set $c = 1, \ldots, n-1$. When $\hat{\theta}_2 > \hat{\theta}_1$, we assume that the maximum likelihood estimate $(\hat{c}, \hat{\theta}_1, \hat{\theta}_2)$ is unique.

Our decision to select the non-null model H_1 over H_0 involves choosing H_1 if the modified Akaike criterion (Akaike, 1973)

$$A = \log L_1(\hat{c}, \hat{\theta}_1(\hat{c}), \hat{\theta}_2(\hat{c})) - \log L_0(\hat{\theta}_0) - 1.5(q_1 - q_0) > 0 \tag{2}$$

where the third term in (2) is a penalty function which adjusts for the difference in dimensionality between the two models. In this application, $q_1 = 3$ corresponding to the parameters c, θ_1 and θ_2, and $q_0 = 1$ corresponding to θ_0. If A in (2) is negative, the decision is to accept the constant model H_0. We have used the terminology "modified Akaike criterion" for two reasons. First, we have used the coefficient $\alpha = 1.5$ from the range of coefficients $1 \leq \alpha \leq 2.5$ suggested by Akaike (1973). Second, the parameter c is discrete yet we have set the degrees of freedom q_1 as though it were continuous. Since the parameter c can be discretized arbitrarily fine, there is an intuition behind the choice. The simulation study in Section 2.2 suggests that with this penalty term, correct decisions are being made with high probability. The two diverse examples that we have considered in this paper provide inferences that are in agreement with previous work; this also suggests that the penalty term is reasonable. There are many penalty terms that have been proposed in the literature (Gelfand and Dey, 1994), and in practice, a user may want to tinker with the penalty term to suit their application.

At this point, the binary segmentation procedure readily presents itself. If H_0 is accepted, the algorithm terminates. However, if H_1 is selected, the data set is divided into the two sub-segments given by $\{y_1, \ldots, y_{\hat{c}}\}$ and $\{y_{\hat{c}+1}, \ldots, y_n\}$. The test of a single change point is then carried out on each of

the two sub-segments. The algorithm continues in this fashion and terminates when no more splitting takes place. We note that the order in which subsegments are divided does not affect the subsequent inference. We also note that the algorithm may be thought of as a forward selection procedure where partitioning continues until the algorithm terminates. Interestingly, if one applied a backwards step, where given a model, the one with the lowest modified Akaike criterion were chosen, we would always choose the same model from which we came. Hence, a backward elimination procedure or a stepwise procedure based on the same decision criterion would provide the same inferences as our algorithm. This is discussed in more detail in the Appendix.

We emphasize that the proposed binary segmentation procedure is simple to carry out and avoids traditional difficulties associated with an unknown and possible large number of change points. Note also that unlike some approaches (for example, Cochrane (1954), Armitage (1955) and Tarone (1982)), the binary segmentation procedure does not rely on the covariates $x_1, \ldots, x_n$. Therefore the method presented here is also applicable when the levels of the stimulus are qualitatively labelled (e.g. control, low, medium, high).

Table 1 Insurance data recording client mortality.

i	Age (x_i)	Number Insured (m_i)	Deaths (y_i)	Death Rate (per 1000)
1	35	1172	3	2.6
2	36	2127	1	0.5
3	37	2744	3	1.1
4	38	2766	2	0.7
5	39	2463	2	0.8
6	40	2368	4	1.7
7	41	2310	4	1.7
8	42	2307	7	3.0
9	43	2060	5	2.4
10	44	1917	2	1.0
11	45	1931	8	4.1
12	46	1747	13	7.4
13	47	1580	8	5.1
14	48	1580	2	1.3
15	49	1468	7	4.8
16	50	1516	4	2.6
17	51	1372	7	5.1
18	52	1343	4	3.0
19	53	1304	4	3.1
20	54	1233	11	8.9
21	55	1205	11	9.1
22	56	1114	13	11.7
23	57	1048	12	11.5
24	58	1155	12	10.4
25	59	1019	19	18.6
26	60	945	12	12.7
27	61	853	16	18.8
28	62	750	12	16.0
29	63	693	6	8.7
30	64	594	10	16.8

Table 2 Results of the binary segmentation procedure applied to the insurance mortality data using the modified Akaike criterion A. The final state of the binary segmentation procedure is marked with an asterisk.

Step	Data Split in Years	A
0	[35, 64]	
1	[35, 53] ⋃ [54, 64]	68.3
2	[35, 44] ⋃ [45, 53] ⋃ [54, 64]	8.5*
3	[35, 39] ⋃ [40, 44] ⋃ [45, 53] ⋃ [54, 64]	−0.9
4	[35, 44] ⋃ [45, 46] ⋃ [47, 53] ⋃ [54, 64]	−1.5
5	[35, 44] ⋃ [45, 53] ⋃ [54, 58] ⋃ [59, 64]	−0.1

2.1 Insurance mortality rate example

A problem for insurance companies offering life insurance is the estimation of mortality rates. In Table 1, we present data obtained from Broffitt (1988) which has been subsequently analyzed by Liu (2000) and Carlin (1992). It provides the number of deaths for insured clients ranging from 35 to 64 years of age.

The nature of human mortality in adults is such that we might expect a strictly increasing and convex response curve. Liu (2000) fits a strictly increasing and convex response curve using a poisson distribution which is well-known as an approximation to the binomial distribution.

However, despite our beliefs about the actual shape of the response curve, this may be a case where the change-point approach of Section 2 is preferable to insurance companies. In other words, companies may prefer grouping people into age groups with common mortality rates. Otherwise, with a strictly increasing response curve, one could imagine the absurd situation where a client is quoted a premium, and upon phoning back the next day, is quoted a higher premium as he/she is one day older. Using the change-point approach, premiums change only on birthdays.

In Table 2, we present the step by step results of the binary segmentation procedure using the modified Akaike criterion for splitting the data. The procedure begins in step 1 by identifying the first candidate change point. The value is 53 years of age and this tentatively divides the full data $[35, 64]$ into two subsegments $[35, 53]$ and $[54, 64]$. The calculated A for this split is 68.3, and since this is positive, the split is accepted. In step 2, the first subsegment $[35, 53]$ is further divided according to the candidate change point of 44 years of age. The corresponding A is 8.5 and the split is again accepted. The data has now been divided according to $[35, 44]$, $[45, 53]$ and $[54, 64]$. In step 3, the new subsegment $[35, 44]$ is further divided according to the candidate change point of 39 years of age. This time, the A is -0.9 and the split is rejected. We continue in this fashion until no more splits are accepted. At the completion of the algorithm, the age groupings in years are $[35, 44]$, $[45, 53]$ and $[54, 64]$ with estimated death probabilities 0.0015, 0.0041 and 0.0126 respectively.

It is interesting to compare the results of binary segmentation with estimates from other techniques. Figure 1 provides a plot of the observed mortality rates together with the fitted lowess curve (Cleveland, 1979). There is considerable variability in the observed mortality rates which results in only three groupings using the binary segmentation procedure. If the observed mortality rate for 64 year olds had been a little higher, the procedure would have detected four groupings. For comparison purposes, the pool-adjacent-violators algorithm gives change points 39, 40, 41, 44, 53, 54, 55 and 59 whereas the binary segmentation procedure gives change points 44 and 53.

2.2 Simulation results

We consider the performance of the binary segmentation procedure in the estimation of quantal response curves via simulation. We generate independent variates $y_1, \ldots, y_{100}$ with change points

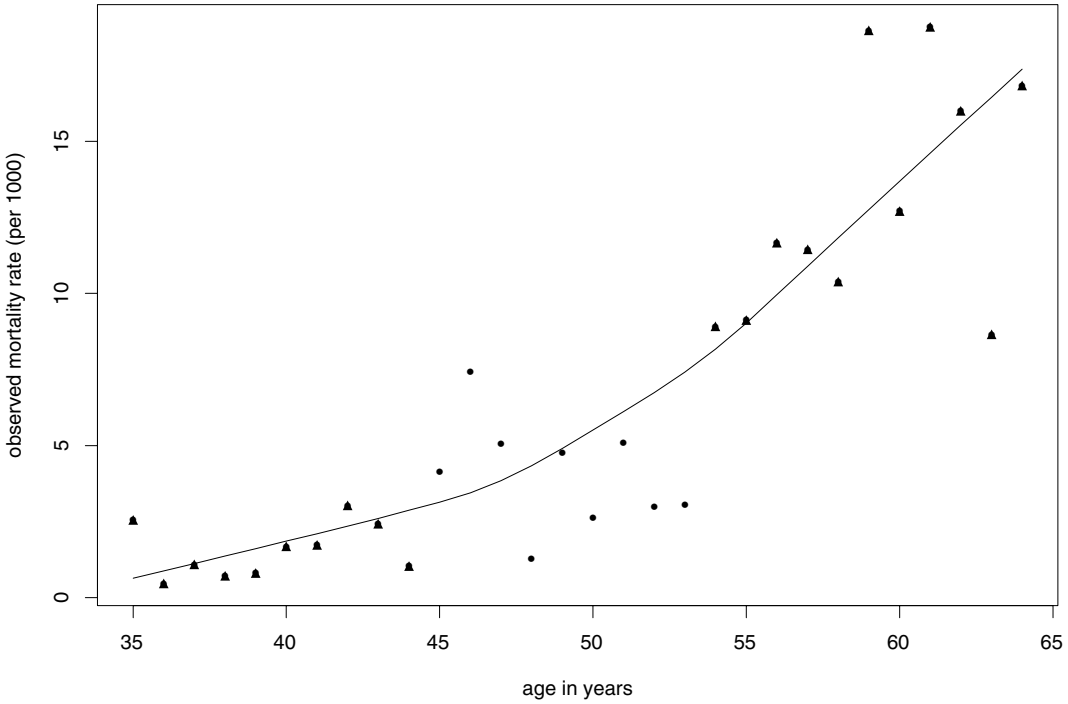

Figure 1 Observed mortality rates (per 1000) from Section 2.1 together with the fitted lowess curve. The alternating sequence of triangles and dots differentiate the three groupings obtained by the binary segmentation procedure.

$c = 20$ and $c = 50$ such that $y_1, \ldots, y_{20} \sim$ binomial $(m, 0.5 - \delta)$, $y_{21}, \ldots, y_{50} \sim$ binomial $(m, 0.5)$ and $y_{51}, \ldots, y_{100} \sim$ binomial $(m, 0.5 + \delta)$. The binary segmentation procedure is applied to the generated data and we determine whether a correct decision is made (i.e. whether two change points are selected and their values are $c = 20$ and $c = 50$). We also determine whether a nearly correct decision is made (i.e. whether two change points are selected and their values are each within two units of $c = 20$ and $c = 50$). We repeat the simulation 10000 times to obtain estimates of correct and nearly correct decisions.

In Table 3 we report the results of the simulations for different values of m and δ. We observe that as the number of Bernoulli trials m increases, the performance of the procedure improves. We also notice that as the difference δ between the Bernoulli parameters increases, the performance of the procedure improves. It is also important to note what happens in cases where the data are insufficient to determine the correct change points. For example, in the case where $m = 10$ and $\delta = 0.05$, the binary segmentation procedure gives zero change points 98% of the time, and one change point 2% of the time. The conservative nature of the procedure is often viewed as a desirable property.

3 Estimation of Spatial Intensity

It is often the case that spatial data have predefined subdivisions of interest. For example, data is often collected on non-overlapping administrative or census districts, and these districts are often irregular in shape. For these problems, we let $C_r(x, y)$ denote the circular growth cluster centered at coordinate (x, y) with radius r. If the centroid of a district lies within $C_r(x, y)$, then the entire district is defined to lie within the candidate region. We investigate various regions by considering a fixed set of (x, y) values and by increasing the radius r in small fixed quantities allowing districts to enter the candidate region. This facilitates the binary segmentation procedure where the largest A is obtained through a

Table 3 Simulation results for quantal response curves. The generated data $y_1, \ldots, y_{100}$ are independent binomial (m, p) variates where $p = 0.5 - \delta, 0.5, 0.5 + \delta$ varies according to the change points $c = 20, 50$. The probabilities that the binary segmentation procedure gives the correct change points and nearly the correct change points are estimated using 10000 simulations.

m	δ	Prob (correct)	Prob (nearly correct)
10	0.05	0.00	0.00
10	0.10	0.01	0.04
10	0.20	0.12	0.45
10	0.40	0.58	0.80
20	0.05	0.00	0.01
20	0.10	0.04	0.20
20	0.20	0.26	0.63
20	0.40	0.72	0.77
50	0.05	0.01	0.08
50	0.10	0.13	0.48
50	0.20	0.51	0.74
50	0.40	0.79	0.79

systematic search over a finite number of circular growth clusters. To identify the radii, we consider $r = v, 2v, \ldots, wv$ for specified v and w.

Using an epidemiological context, let $z(x, y)$ be the number of occurrences of a particular disease within the district where the coordinate (x, y) resides. Further, let $m(x, y)$ be the population of the district. We then assume that $z(x, y)$ follows a binomial distribution with $m(x, y)$ trials. For each potential circular growth cluster, we consider the model $z(s, t) \sim \text{binomial}(m(s, t), \theta_1)$ if the district with centroid (s, t) belongs to the corresponding region and $z(s, t) \sim \text{binomial}(m(s, t), \theta_2)$ otherwise. Then, we calculate the maximum likelihood under the null hypothesis $H_0 : \theta_1 = \theta_2 = \theta_0$ and the maximum likelihood under the alternative hypothesis $H_1 : \theta_1 \neq \theta_2$. The likelihood under the null hypothesis is proportional to

$$L_0(\theta_0) = \theta_0^{\sum_V z(x,y)} (1 - \theta_0)^{\sum_V (m(x,y) - z(x,y))}$$

where V is the set of all district centroids within the region of interest and the maximum is obtained at $\hat{\theta}_0 = \dfrac{\sum_V z(x, y)}{\sum_V m(x, y)}$. Under H_1, the likelihood is proportional to

$$\begin{aligned} & L_1(C_r(x, y), \theta_1, \theta_2) \\ &= \prod_{(s,t) \in V \cap C_r(x,y)} \theta_1^{z(s,t)} (1 - \theta_1)^{(m(s,t) - z(s,t))} \\ &\quad \times \prod_{(s,t) \in V \setminus C_r(x,y)} \theta_2^{z(s,t)} (1 - \theta_2)^{(m(s,t) - z(s,t))}. \end{aligned} \tag{3}$$

For fixed $C_r(x,y)$, the profile likelihood $L_1(C_r(x,y),\theta_1,\theta_2)$ is maximized at

$$\hat{\theta}_1 = \frac{\sum_{(s,t)\in V\cap C_r(x,y)} z(s,t)}{\sum_{(s,t)\in V\cap C_r(x,y)} m(s,t)} \quad \text{and} \quad \hat{\theta}_2 = \frac{\sum_{(s,t)\in V\setminus C_r(x,y)} z(s,t)}{\sum_{(s,t)\in V\setminus C_r(x,y)} m(s,t)}\,.$$

The fully maximized likelihood $L_1(\hat{C},\hat{\theta}_1,\hat{\theta}_2)$ is then obtained by maximizing $L_1(C_r(x,y),\hat{\theta}_1,\hat{\theta}_2)$ over the discretized indices (x,y) and radii $r = v, 2v, \ldots, wv$. We select the non-null model H_1 if the modified Akaike criterion

$$A = \log L_1(\hat{C},\hat{\theta}_1,\hat{\theta}_2) - \log L_0(\hat{\theta}_0) - 1.5(q_1 - q_0) > 0 \tag{4}$$

where $q_1 = 5$ and $q_0 = 1$. A negative value of A leads us to accept H_0. Using similar logic as in the estimation of quantal response curves, we obtain $q_1 = 5$ by assigning two degrees of freedom for (θ_1,θ_2) and assigning three degrees of freedom for $C_r(x,y)$ keeping in mind that the grids can be chosen arbitrarily fine. The degrees of freedom $q_0 = 1$ corresponds to the single parameter θ_0.

If H_0 is accepted, then a final cluster has been determined which includes all of the districts with centroids in V. However, if H_0 is rejected, the data set is divided into districts whose centroids lie in $\hat{C}$ and districts whose centroids 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 the probability of disease in the resulting cluster by $\hat{\theta}_0$.

3.1 SIDS example

We consider the spatial clustering of the incidence of sudden infant death syndrome (SIDS) in the state of North Carolina. The data are taken from chapter 6 of Cressie (1993) and consist of SIDS counts in the 100 counties of North Carolina during the years 1974 through 1978. For every county, a geological coordinate (x,y) is available that corresponds to the longitude x and latitude y in miles of the county seat from an arbitrarily chosen origin. We use the coordinate (x,y) as a surrogate for the centroid of the county. In addition, the number of live births in each county is available for the time period. The number of live births in the counties ranges from 284 to 21588, and the incidence rate of SIDS over the state is 0.00202. The data were compiled by M. Symons, D. Atkinson and the State Center for Health Statistics of the North Carolina Department of Human Resources. The data have been previously analyzed by various investigators including Cressie and Chan (1989) and Kulldorff (1997). Figure 2 is a county map of North Carolina which we have taken from Cressie (1993) and modified.

We apply the proposed binary segmentation procedure to the SIDS data of North Carolina using circular growth clusters and the binomial model. We choose $v = 3$ and $w = 100$ which determines the size of the growth clusters, and we let the county seat coordinates form the set of (x,y) values under consideration. In Figure 3, a map of the county seats is given where larger circles correspond to counties with a higher number of births and darker circles correspond to counties with higher SIDS rates. Figure 2 also displays the final clusters obtained using the binary segmentation procedure. We observe that a total of seven clusters are detected where subcluster A1 is distinguished within the first detected cluster A. It is noteworthy that the proposed binary segmentation procedure can lead to non-contiguous regions falling in the same partition. This is the case in the SIDS example where the leftover region is a cluster of its own, and consists of four geographically separate regions.

In a comparison of Figure 2 with the auto-Gaussian models proposed by Cressie and Chan (1989), the regions of high/low SIDS rates roughly correspond. Overall, the Cressie and Chan (1989) models have more clusters (e.g. 24 scales in their Model 4.4) and their focus is more concerned with white/nonwhite SIDS rates than with county rates. We note that our clusters D and E have SIDS rates that are more than double the state rate; these two regions are also detected by Cressie and Chan (1989). Our cluster C (Anson county) has the highest SIDS rate which is more than four times the state rate; it is not prominent in the Cressie and Chan (1989) map (their Figure 3) since it was excluded from their analyses.

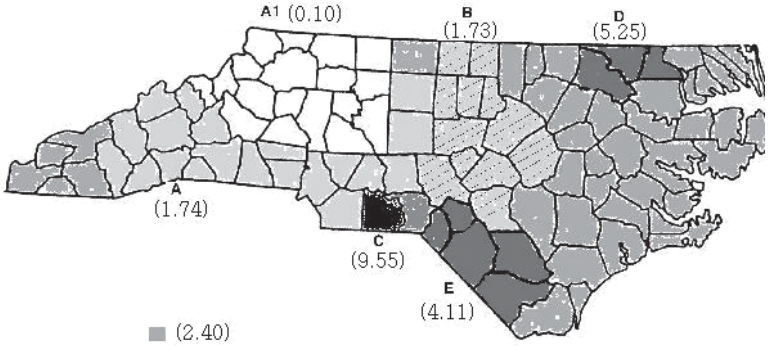

Figure 2 County map of North Carolina indicating estimated county clusters obtained from the SIDS data using the binary segmentation procedure. Darker colors correspond to areas with higher SIDS rates. The numbers inside parentheses indicate the estimated number of deaths per 1000 live births for the corresponding clusters. The left-over area corresponding to four geographically separate regions has an estimated death rate of 2.40.

4 Concluding Remarks

The methods proposed in the paper are motivated by simplicity.

We have demonstrated how quantal response curves can be easily constructed using binary segmentation by viewing the problem as a change-point model. Extensions of the approach to parametric families such as the poisson are also possible.

We have also demonstrated how spatial clusters can be easily detected using an approach based on binary segmentation. Although the resultant clusters can have diverse shapes, it is clear that not every

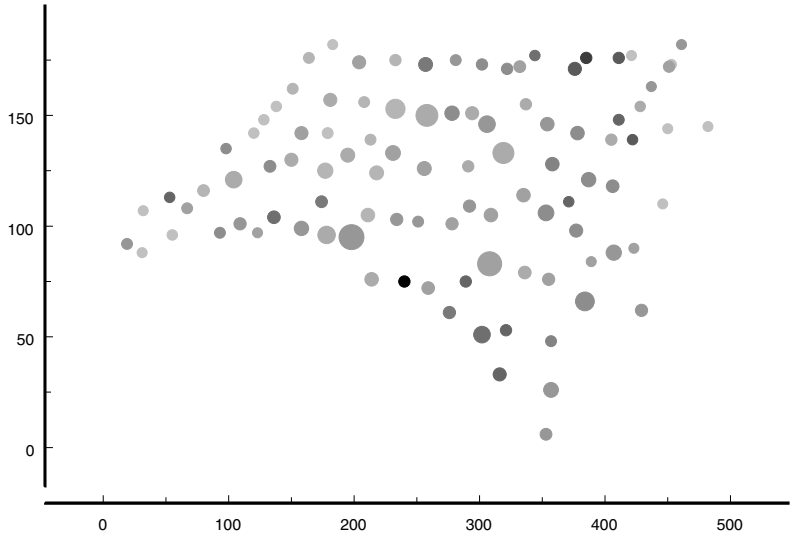

Figure 3 Map of the 100 county seats of North Carolina where larger circles correspond to counties with a higher number of births and darker circles correspond to counties with higher SIDS rates. The axes are distances labelled in miles.

shape is possible. This leads to possible extensions of the approach. When an experimenter is entertaining spatial problems, he should consider the way in which the phenomenom of interest spreads. For example, if the phenomenom spreads in an airborne manner in an area that is not excessively windy, it may be sensible to use circular growth clusters. However, there are more complex ways that phenomena may spread, and for these cases, one may prefer a growth cluster that expands in a more general fashion, perhaps in a contiguous way but in any direction. For example, we note in Figure 2 that the incidence rates in clusters A and B are similar, and these clusters might be sensibly grouped together. Such generality would necessarily compound the computations, and this is a topic of future research.

Appendix: Discussion on the Forward Selection Approach

Both of the applications proposed in the paper concern the determination of complex models. In each application, a complex model is determined by beginning with a simple null model and then adding parameters (i.e. complexity) one step at a time. The procedures are terminated according to the Akaike stopping criterion.

There is an opposing school of thought which suggests that complex models should be built and then reduced. For example, in the determination of classification and regression trees, the standard approach is to grow trees via node splitting and then prune the trees back (e.g. Breiman et al., 1984).

In other words, our approach may be generally viewed as a forward selection procedure whereas one might also consider some sort of stepwise procedure. However, it turns out that our use of binary segmentation with the Akaike criterion provides a special structure whereby a backwards step always returns to the simpler model from which it came. Thus, our forward selection approach provides the same inferences as some sort of stepwise procedure based on the Akaike decision criterion.

To see that this is the case, consider the estimation of quantal response curves and let

$$l(t_1,t_2) = \log\left[\left(\frac{\sum_{t_1}^{t_2} y_i}{\sum_{t_1}^{t_2} m_i}\right)^{\sum_{t_1}^{t_2} y_i}\left(1-\frac{\sum_{t_1}^{t_2} y_i}{\sum_{t_1}^{t_2} m_i}\right)^{\sum_{t_1}^{t_2} m_i - y_i}\right].$$

Furthermore, let the model with cutpoints $c_1 < \cdots < c_k$ be denoted $[c_1, \ldots, c_k]$. Then, in going from the null model $[\emptyset]$ to model $[c_1]$, the criterion

$$A(c_1) = l(1,c1) + l(c_1+1,n) - l(1,n) - K$$

is calculated where $A(c_1) > 0$, and in going from model $[c_1]$ to model $[c_1,c_2]$, the criterion

$$\begin{aligned} A(c_1,c_2) &= l(1,c1) + l(c_1+1,c_2) + l(c_2+1,n) - l(1,c_1) - l(c_1+1,n) - K \\ &= l(c_1+1,c_2) + l(c_2+1,n) - l(c_1+1,n) - K \end{aligned}$$

is calculated where $A(c_1,c_2) > 0$ and $K = 3.0$.

Suppose that you now prune. This means that the only model reductions that can be considered are the model reduction $[c_1,c_2] \to [c_1]$ with criterion $A(c_1,c_2)$, and the model reduction $[c_1,c_2] \to [c_2]$ with criterion

$$\begin{aligned} A(c_2,c_1) &= l(1,c1) + l(c_1+1,c_2) + l(c_2+1,n) - l(1,c_2) - l(c_2+1,n) - K \\ &= l(1,c_1) + l(c_1+1,c_2) - l(1,c_2) - K \end{aligned}$$

where we choose the reduced model $[c_1]$ over the reduced model $[c_2]$ if $A(c_1,c_2)$ is smaller than $A(c_2,c_1)$ and vice-versa.

Now, since $A(c_1) > A(c_2)$ by virtue that model $[c_1]$ was chosen in the first forward step amongst all single cutpoint models, it follows that

$$A(c_1, c_2) - A(c_2, c_1) = A(c_2) - A(c_1) < 0$$

which implies that $[c_1]$ is selected in the backward step. Therefore, in pruning, you go back to the model from which you came. This is easily seen to be true not only for the case presented above but for any stage of the algorithm. It is also seen to be true in our application concerning the estimation of spatial intensity.

Acknowledgements This work was initiated during a sabbatical visit by Yang to the Department of Statistics and Actuarial Science at Simon Fraser University. Yang and Swartz were partially supported by grants from the Korea Science and Engineering Foundation and the Natural Sciences and Engineering Research Council of Canada respectively. The authors thank two referees whose comments and suggestions lead to improvements in the manuscript.

References

Agresti, A. (2002). *Categorical Data Analysis, second edition.* Wiley, New York.

Akaike, H. (1973). Information theory and the extension of the maximum likelihood principle. In B. N. Petrior and F. Csaki (Eds.), *Proc. 2nd Int. Symp. Information Theory Budapest: Akademiai Kiado*, 267–281.

Armitage, P. (1955). Test for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386.

Ayer, M., Brunk, H., Ewing, G., Reid, W., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics* **26**, 641–647.

Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A* **154**, 143–155.

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

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

Broffitt, J. D. (1988). Increasing and increasing convex Bayesian graduation. *Transactions of the Society of Actuaries* **40**, 115–148.

Carlin, B. P. (1992). A simple Monte Carlo approach to Bayesian graduation. *Transactions of the Society of Actuaries* **44**, 55–76.

Chen, J. and Gupta, A. K. (1997). Testing and locating variance change points with applications to stock prices. *Journal of the American Statistical Association* **92**, 739–747.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10**, 417–451.

Cressie, N. (1993). *Statistics for Spatial Data, revised edition.* Wiley, New York.

Cressie, N. and Chan, N. (1989). Spacial modeling of regional variables. *Journal of the American Statistical Association* **84**, 393–401.

Cuzick, J. and Edwards, R. (1990). Spacial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society, Series B* **52**, 73–104.

Dean, C. B. (1998). Overdispersion. In P. Armitage and T. Colton (Eds.), *Encyclopedia of Biostatistics* Wiley, 467–472.

Dempster, A. P., Selwyn, M. R., and Weeks, B. J. (1983). Combining historical and randomized controls for assessing trends in proportions. *Journal of the American Statistical Association* **78**, 221–227.

Denison, D. and Holmes, C. (2001). Bayesian partitioning for estimating disease risk. *Biometrics* **57**, 143– 149.

Dey, D., Muller, P., and Sinha, D. (Eds.) (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics.* Springer-Verlag, New York.

Diggle, P. J. (1990). A point process modelling appproach to raised incidence of a rare phenomenon in the vincity of a prespecified point. *Journal of the Royal Statistical Society, Series A* **153**, 349–362.

Diggle, P. J. and Chetwynd, A. G. (1991). Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics* **47**, 1155–1163.

Finney, D. J. (1978). *Statistical Methods in Biological Assay, third edition.* Cambridge University Press, Cambridge.

Gallant, A. R. and Fuller, W. A. (1973). Fitting segmented polynomial regression models whose join points have to be estimated. *Journal of the American Statistical Association* **68**, 144–147.

Gangnon, R. and Clayton, M. K. (2000). Bayesian modeling and detection of spatial disease clustering. *Biometrics* **56**, 922–935.

Gangnon, R. and Clayton M. K. (2001). A weighted average likelihood ratio test for spacial clustering of disease. *Statistics in Medicine* **20**, 2977–2987.

Gelfand, A. E. and Dey, D. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B* **56**, 501–514.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (Eds.) (1996). *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Holmes, C., Denison, G., and Mallik, B. (1999). *Bayesian partitioning for classification and regression.* Manuscript, Imperial College.

Hudson, D. J. (1966). Fitting segmented curves whose join points have to be estimated. *Journal of the American Statistical Association* **61**, 1097–1129.

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.

Knorr-Held, L. and Rasser, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **56**, 13–21.

Kulldorff, M. (1997). A spacial scan statistic. *Communications in Statistics – Theory and Methods* **26**, 1487–1496.

Kulldorff, M. and Nagarwalla, N. (1995). Spacial disease clusters: detection and inference. *Statistics in Medicine* **14**, 799–810.

Liu, C. (2000). Estimation of discrete distributions with a class of simplex constraints. *Journal of the American Statistical Association* **95**, 109–120.

Ma, Y., Guo, J., Shi, N., and Tang, M. (2002). On the use of historical control information for trend tests in carcinogenesis. *Biometrics* **58**, 917–927.

Morgan, B. J. T. (1992). *Analysis of Quantal Response Data.* Chapman and Hall, London.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, second edition.* Chapman and Hall, London.

Openshaw, S., Craft, A., Charlton, M., and Birch, J. (1988). Investigation of leukemia clusters by use of a geological analysis machine. *Lancet* **1**, 272–273.

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

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

Stone, R. A. (1988). Investigation of excess environmental risks around putative sources: statistical problems and a proposed test. *Statistics in Medicine* **7**, 649–660.

Tarone, R. E. (1982). The use of historical control information in testing for a trend in proportions. *Biometrics* **38**, 215–220.

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

Thisted, R. A. (1988). *Elements of Statistical Computing.* Chapman and Hall, New York.

Turnbull, B. W., Iwano, E. J., Burnett, W. S., Howe, H. L., and Clark, L. C. (1990). Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *American Journal of Epidemiology* **132**, S136–S148.

Venkatraman, E. S. (1992). Consistency results in multiple change-point situations. Unpublished PhD Thesis, Department of Statistics, Stanford University.

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

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

Whittemore, A., Friend, N., Brown, B. W., and Holly, E. A. (1987). A test to detect clusters of disease. *Biometrika* **74**, 31–35.

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

Yang, T. Y. and Kuo, L. (2001). Bayesian binary segmentation procedure for a Poisson process with multiple change points. *Journal of Computational and Graphical Statistics* **10**, 772–785.