Communication papers of the Federated Conference on Computer Science and Information Systems, pp. 95–102
DOI: 10.15439/2017F206
ISSN 2300-5963 ACSIS, Vol. 13

# Binary Segmentation Methods for Identifying Boundaries of Spatial Domains

Nishanthi Raveendran
Department of Statistics
Faculty of Science and Engineering
Macquarie University, Sydney, Australia
Email: nishanthi.raveendran@students.mq.edu.au

Georgy Sofronov
Department of Statistics
Faculty of Science and Engineering
Macquarie University, Sydney, Australia
Email: georgy.sofronov@mq.edu.au

***Abstract*—Spatial clustering is an important component of spatial data analysis which aims in identifying the boundaries of domains and their number. It is commonly used in disease surveillance, spatial epidemiology, population genetics, landscape ecology, crime analysis and many other fields. In this paper, we focus on identifying homogeneous sub-regions in binary data, which indicate the presence or absence of a certain plant species which are observed over a two-dimensional lattice. To solve this clustering problem we propose to use the change-point methodology. We develop new methods based on a binary segmentation algorithm, which is a well-known multiple change-point detection method. The proposed algorithms are applied to artificially generated data to illustrate their usefulness. Our results show that the proposed methodologies are effective in identifying multiple domains and their boundaries in two dimensional spatial data.**

## I. Introduction

IDENTIFYING homogeneous domains is of particular interest in spatial statistics. It is often the case that spatial data have pre-defined subdivisions of interest. For example, data are often collected on non-overlapping administrative or census districts and these districts are often irregular in shape; see [1]. As a part of statistical modelling, spatial clustering is also an important component of spatial data analysis since spatial data may be heterogeneous and difficult to understand. However, if we cluster the data into homogeneous domains, then we can construct appropriate statistical models for each cluster. The problem of finding regional homogeneous domains is known as segmentation, partitioning or clustering. The two main problems in spatial clustering are identifying the number of domains, which is usually not known in advance, and estimating the boundaries of such domains.

Many clustering algorithms have been developed in the literature, ranging from hierarchical methods such as bottom-up (or agglomerative) methods top-down (or divisive) methods, to optimization methods such as the $k$-means algorithm [2]. The algorithms have numerous applications in pattern recognition, spatial data analysis, image processing, market research; see [3]. Spatial clustering covers enormous practical problems in many disciplines. For example, in epidemiological studies and public health research, it is known that the disease risk varies across space and it is important to identify regions of safety and regions of risk. A model using Bayesian approach for spatial clustering was discussed in [4]. Recently, a two-stage Bayesian approach for estimating the spatial pattern in disease risk and identifying clusters which have high (or low) disease risks was proposed in [5].

The homogeneity changes in space is an important research subject in ecology. In a large area, the spatial distribution of plant or animal species is never homogeneous. Studying these kinds of changes is important in several ways. For example, detecting early changes in vegetation improves productivity. A class of Bayesian statistical models to identify thresholds and their locations in ecological data was introduced in [6]. A method to estimate the change-point distribution between two patches was presented in [7].

Studies of weather and climatic systems at a global scale have become a prime area of research for a number of reasons; one of these is the concern about global climatic change. Mann-Kendall trend test, Bayesian change point analysis and a hidden Markov model to find changes in the rainfall and temperature patterns over India are used in [8].

There has also been extensive literature on image recognition with some articles presenting statistical approaches to the boundary identification in statistical imaging. For example, [9] presented a Markov chain Monte Carlo (MCMC) method to identify closed object boundaries in gray-scale images. Change curve estimation problem is also referred as multidimensional detection problem or boundary estimation problem. A wavelet method to estimate jumps and sharp curves in the plane was proposed in [10].

Even though there is wide range of applications to spatial clustering, many statistical methods for detecting clusters have some limitations: either they detect the number of clusters and do not determine their locations, or they provide the inference with no clustering. In this study, we are interested in identifying the boundaries of domains and their number with applications to an ecological landscape. In general, these problems are typically challenging due to the multivariate nature of the data which leads to complex and highly

parametrized likelihoods. We use binary data indicating the presence or absence of plant species, which are observed over a two-dimensional lattice. We consider our problem as a change-point detection problem, which is commonly used in analysing time series to detect changes and their locations. We develop new algorithms based on a binary segmentation algorithm, which is a well-known recursive partitioning tool in change-point literature and it leads to simple solutions for such problems and it has an advantage on simplicity and less computational cost compare to other methods.

Binary spatial data are commonly involved in various areas such as economics, social sciences, ecology, image analysis and epidemiology. Also, such data frequently occur in environmental and ecological research, for instance, when the data correspond to presence or absence of a certain invasive plant species at a location or, when the data happen to fall into one of two categories, say, two types of soil. The general overview of spatial data can be found in [11–13].

This study aims to develop effective procedures based on the binary segmentation method for estimating both the number of domains and their locations in spatial data. This paper is organized as follows. Section II describes the multiple change-point problem. We provide the mathematical model for our problem in section III. We explain both general binary segmentation and new algorithms in section IV and section V, respectively, and provide the numerical results in section VI. Section VII gives a discussion. Section VIII concludes the paper with the future directions.

## II. Multiple Change-point Problem

Let us formulate the general multiple change-point problem in mathematical terms.

Let $y_n = (y_1, \ldots, y_n)$ be a sequence of observations of length $n$, $y_1, y_2, \ldots, y_n$ be independent random variables with the probability distribution functions $F_1, F_2, \ldots, F_n$. Let $\tau_1, \tau_2, \ldots, \tau_m$ be unknown positions of $m$ change-points, where $\tau_1 < \tau_2 < \cdots < \tau_m$. We define $\tau_0 = 0$ and $\tau_{m+1} = n$. The sequence of observations is divided into $m+1$ segments based on $m$ change-points. In general, the multiple change-point problem involves the following null hypothesis,

$$H_0 : F_1 = F_2 = \cdots = F_n$$

versus

$$H_1 : F_1 = \cdots = F_{\tau_1} \neq F_{\tau_1+1} = \cdots = F_{\tau_2} \neq F_{\tau_2+1} = \cdots = F_{\tau_m} \neq F_{\tau_m+1} = \cdots = F_n.$$

## III. Mathematical Model

Let us assume that we have independent binary observations on an $n \times m$ lattice. We assume the observations at each cell are univariate. Let $M$ be the number of domains and $p = (p_1, \ldots, p_M)$ be the parameters of Bernoulli distribution for the domains. The likelihood function is given by:

$$L(X|p) = \prod_{j=1}^{M} p_j^{I_{D_j}} (1-p_j)^{O_{D_j}}, \quad j = 1, 2, \ldots, M,$$

$X$ is the data (a matrix of zeroes and ones),
$M$ is the number of domains,
$D_j$ is the $j$-th (rectangular) domain,
$p = (p_1, \ldots, p_M)$ is the vector of probabilities,
$I_{D_j}$ is the number of ones in $D_j$,
$O_{D_j}$ is the number of zeroes in $D_j$.

In order to estimate the boundaries of domains and their number, we maximize the log-likelihood function

$$l(X|p) = \sum_{j=1}^{M} I_{D_j} \log p_j + O_{D_j} \log(1-p_j).$$

### A. Maximum Likelihood Framework

Let $X$ is an $n \times m$ matrix. A natural approach to split a domain into homogeneous sub-domains is to view it as the following hypothesis testing:

$$H_0 : \text{No sub-domains;} \qquad \text{Vs} \qquad H_1 : \text{Two domains.}$$

Under the null hypothesis, the log-likelihood function for the entire domain is given as

$$l(X|\hat{p}),$$

where $\hat{p}$ is the maximum likelihood estimate of the $p$. Under the $H_1$, the log-likelihood function given a change-point (in our case, cut or boundary) $c$, which divides the domain into two homogeneous domains $D_1$ and $D_2$, is

$$P(c) = l(D_1|\hat{p}_1) + l(D_2|\hat{p}_2),$$

where $\hat{p}_1$ and $\hat{p}_2$ are the maximum likelihood estimates of the parameters for the first and the second domain, respectively. To estimate the location of the change-point, the log-likelihood function under $H_1$ is maximized.

Test statistic:

$$\lambda(X) = 2[\max_c P(c) - l(X|\hat{p})],$$

where a threshold $\beta$ is chosen such that if $\lambda(X) > \beta$, the null hypothesis is rejected. The threshold could be based on the use of an information criterion: AIC, $\beta = 2k$, and SIC, $\beta = k \log n$, where $k$ is the number of extra parameters as a result of adding another domain.

The likelihood ratio test statistic can be extended to multiple change-point detection by summing the likelihood for the $m$ data segments.

One way to detect multiple change-points is to minimize

$$\sum_{i=1}^{m+1}[C(y_{(\tau_{i-1}+1):\tau_i})] + \beta f(m), \quad (1)$$

where $C$ is a cost function for a segment and $\beta f(m)$ is a penalty term in order to avoid overfitting.

## IV. The Binary Segmentation Method

The binary segmentation is a well-known multiple change-point method and has been studied by various authors. It was first introduced in [14] in the context of cluster analysis. The concept of binary segmentation in detecting changes in mean was proposed in [15]. Later, this procedure was extended to detect the number of change-points in a multidimensional random process and proved the consistency of the estimates produced by binary segmentation under mild conditions, the first of which is based on the minimal distance between change-points; see [16]. Similar results when the change-points are allowed to approach one another are achieved in [17]. Recent studies include many applications and it can be found in [24–32]. Thus, this method is now the most understood and widely cited search algorithm used within the multiple change-point literature.

Binary segmentation can be used to extend any single change-point method to multiple change-points. In the early works, binary segmentation was performed using a simple CUSUM test. It starts with applying the chosen single change-point detection method to the entire data set $y_{1:n}$, sequence of observations of length $n$. If no change-point is found, then the algorithm stops. If a change-point is detected, say, $\tau$, then the data set is split into two separate segments, $y_{1:\tau}$ and $y_{\tau+1:n}$. The single change-point method is applied to two segments and the procedure is repeated iteratively. Finally, we stop when no more change-points are detected. The generic binary segmentation algorithm [18] is given below.

One of the good features of the binary segmentation algorithm is that it detects the number of change-points and their locations simultaneously. It can be seen as an approach with $f(m) = m$ to minimize (1) by iteratively deciding whether a change-point should be added or not. It is a fast algorithm and saves lots of computational time and it can be implemented with the computational complexity $O(n)$.

There are some exact methods to minimize (1) but at a higher computational cost (for example, see [33]). It is clear that in many situations the number of change-points increases as we collect more data and the computational burden increases as well. Therefore, many authors are working on developing new algorithms which are fast and exact. Recently, a dynamic programming technique called PELT (Pruned Exact Linear Time) which is $O(n)$ under certain assumptions such as the number of true change-points being linear with the data length was introduced in [19]. Still, this method has $O(n^2)$ complexity at the worst case.

**Algorithm** The Generic Binary Segmentation Algorithm

**Input:** A set of data of the form, $(y_1, y_2, \ldots, y_n)$.
A test statistic $\lambda(\cdot)$ dependent on the data.
An estimator of change-point position $\hat{\tau}(\cdot)$.
A rejection threshold $\beta$.

**Initialise:** Let $C = \emptyset$, and $S = [1, n]$

**Iterate:** while $S \neq \emptyset$
1. Choose an element of $S$; denote this element as $[s, t]$.
2. If $\lambda(y_{s:t}) < \beta$, remove $[s, t]$.
3. If $\lambda(y_{s:t}) \geq \beta$:
   (a) remove $[s, t]$ from $S$;
   (b) calculate $r = \hat{\tau}(y_{s:t}) + s - 1$, and add $r$ to $c$;
   (c) if $r \neq s$ add $[s, r]$ to $S$;
   (d) if $r \neq t - 1$ add $[r + 1, t]$ to $S$.

**Output:** The set of change-points recorded $C$.

## V. The binary segmentation method for spatial clustering

We would like to identify the number of homogeneous domains and their boundaries in binary lattice data. In this case, the change-point locations are the points which is used to draw a horizontal or vertical line to divide the domain into two homogeneous rectangular segments. Here we present three algorithms. Our proposed algorithms use maximum likelihood test as described in the previous section.

### A. Algorithm 1

The algorithm searches every column and row to detect the change-point and selects the maximum test statistic for the optimum cut. If the test statistic is greater than a threshold value, it splits the domain according to the index (row or column) and stores the obtained domains. Otherwise, the algorithm stops. This procedure is repeated until a stopping criterion is met. In this study, we consider rectangle shaped domains.

We also propose two more algorithms with modifications. In general, our method can be summarized by a three-step iterative procedure (given in Algorithm 1).

### B. Algorithm 2

We introduce a modification of Algorithm 1. It follows the similar structure but at each iteration it identifies two change-points and three domains. Here, all three segments have different means.

### C. Algorithm 3

Algorithm 3 is a modified version of Algorithm 1. The main difference is that at each iteration it selects the bigger domain for next iteration assuming that bigger domain has a higher

**Algorithm 1** Main Algorithm

**Step 1:** Given the data, search the change point column-wise and find the optimal cut which maximizes the test statistic. Repeat this procedure row-wise.

**Step 2:** Select the maximum of the two test statistics for the optimal column and row cuts and compare with the threshold value. If the test statistic is greater than the threshold value, then split the data in two domains.

**Step 3:** Repeat steps 1 and 2 for each domain until no new domains are identified.

chance to be split at the next iteration. Here, the "bigger" means the area of the rectangle. This algorithm performs well compare to Algorithm 1 and Algorithm 2. The great advantage of this algorithm is that it performs fast because at each iteration it selects only one domain to split. But in Algorithm 1, at each iteration it considers two segments in parallel. This algorithm would be useful when we need to split the data into major domains (few number of domains).

### D. Model selection

Our objective is to estimate both the number of domains and their boundaries. Thus, it can be formulated as a model selection problem, which is usually done by using a specific criterion. There are several popular model selection criteria that have been proposed in different contexts; see [34–36]. The model selection criteria are mainly used for two different purposes: first, to choose a model that well approximates the true model; second, to find the true model in a list of candidate models [20]. In this study, we use AIC [21], BIC [22] and mBIC (modified BIC, defined for change-point problems) [23]. The AIC, BIC and mBIC for our model can be described as below:

$$\begin{aligned} \text{AIC}(k) &= -2\log L(\hat{\Theta}_k) + 2k, \\ \text{BIC}(k) &= -2\log L(\hat{\Theta}_k) + k\log n, \\ \text{mBIC}(k) &= -2\log L(\hat{\Theta}_k) + 2(k+1)\log n, \end{aligned}$$

where $L(\hat{\Theta}_k)$ is the maximum likelihood for the model with $k$ parameters, $k = 1, 2, \ldots, M$, and $n$ is the sample size. A model that minimizes a criterion (for example, AIC) is considered to be the most appropriate model.

### E. Stopping criteria

In the binary segmentation, one has to define a stopping criterion to terminate the iterative procedure. We use one of two methods:

1) The algorithm is reiterated while we have significant cuts based on the results of a hypothesis testing. Let us define that number of cuts $c = C$, the process is stopped and the corresponding solution is considered as the optimal solution for the problem.
2) The decision to stop the algorithm is based on an information criterion.

### F. Likelihood test for spatial clustering

In this study, we use the likelihood ratio test to check whether the domains obtained by the proposed algorithms are homogeneous or not. The null hypothesis for this model is given as:

$$H_0 : \text{Domain 1 and Domain 2 are homogeneous.}$$

The alternative hypothesis is

$$H_1 : \text{Domain 1 and Domain 2 are not homogeneous.}$$

The test statistic is:

$$\text{LRT} = -2\log\left(\frac{\text{likelihood for null model}}{\text{likelihood for alternative model}}\right).$$

After the algorithm finishes, we obtain several homogeneous domains. The next step is to perform a multiple comparison test for all combinations of the domains. Further, we consider the Bonferroni correction, which is used to control the family-wise error rate when conducting multiple hypothesis tests. The Bonferroni correction adjusts $p$-values when several statistical tests are being performed simultaneously on a single data set. To perform the Bonferroni correction, divide the critical $p$-value ($\alpha$) by the number of comparisons or the number of hypothesis being made. For example, if we have $M$ domains for our data set and have to perform $N$ comparisons, then the Bonferroni correction would test each individual hypothesis at $\alpha/N$. Here we do not need to perform all comparisons since we consider only rectangular domains in this study.

## VI. Results

In this section we include all numerical results to illustrate and validate the proposed algorithms. A simulation study was carried out to demonstrate the properties of our algorithms and to analyse their segmentation capabilities. We present an example to illustrate the usefulness of our method. Finally, we compare our three algorithms using the Root Mean Square Error (RMSE) and information criteria. All proposed algorithms have been implemented using the statistical software R.

### A. Simulation study

To perform simulation study, we generate artificial matrices using a Bernoulli distribution. We apply Algorithm 1, record the position of the optimal cut and estimate the parameter of the Bernoulli distribution for each domain. Each time we calculate the RMSE and plot a kernel density estimation curve to analyze the effectiveness of the algorithm. Hereinbelow we denote the number of obtained domains by $D$.

The RMSE is calculated as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(e_{ij} - t_{ij})^2}{N}},$$

where $e_{ij}$, $t_{ij}$ denotes the estimated and the true values, respectively, for each cell of the matrix; $i, j$ indicate the corresponding rows and columns; $n \times m$ is the size of the

TABLE I
THE NUMBER OF DOMAINS WITH FREQUENCIES

| $D$ | 1 | 2 | 3 | 4 | 5 | 5+ |
|---|---|---|---|---|---|---|
| **Frequency** | 0 | 0 | 0 | 316 | 529 | 155 |

matrix.

*1) Case 1: Testing the performance of the binary segmentation algorithm:* Our aim is to find the optimal number of domains for a particular data set. We generate artificial data with four domains and run our algorithm 1000 times. We record the number of domains identified by Algorithm 1.

Table I shows that the algorithm correctly found four domains in 316 simulations (out of 1000). However, the algorithm tends to overestimate the number of domains.

*2) Case 2: Reporting the RMSE on the parameters of the domains:* We analyse how the RMSE depends on the parameters of domains, that is, the probability of "1". In this study, we generate artificial data with two domains, where $p_1$ and $p_2$ are the parameters of the Bernoulli distributions for the domains.

The algorithm performs well in identifying the correct position of the cut when the difference between $p_1$ and $p_2$ is rather large (for example, $p_1 = 0.8$ and $p_2 = 0.2$). Note that even if the difference is getting smaller, the algorithm works quite well; in this situation the RMSE is slightly higher compare to the case with the large difference of the probabilities. Figure 2 shows a kernel density estimation for three different cases. In the first row, we fix $p_1 = 0.8$ and we change $p_2$ from 0.1 to 0.9. Likewise, in the second row, we fix $p_1$ as 0.5 and in the third row, $p_1 = 0.2$. It is clear that the effectiveness depends on the difference of the probabilities.

*3) Case 3: Reporting the RMSE on the size of the data:* In this section, we analyse how the RMSE changes depending on the size of the data. It is important to test our algorithms for different sizes. We generate matrices of different sizes ($50 \times 50$, $100 \times 100$, $200 \times 200$) but with the same probabilities $p_1$ and $p_2$ for domain 1 and domain 2, respectively. Here we restrict the number of domains to two.

Figure 1 shows the plot of kernel density estimation, which illustrates that the average value of the RMSE is not significantly influenced by the size of the data, whereas it is clear that the variability in the RMSE is getting smaller when the data size is becoming larger.

## B. Example

We generate a $100 \times 100$ matrix using Bernoulli distributions with four domains (all are vertical cuts); the parameters are given in Table II. We apply our binary segmentation algorithms, record the positions of the optimal cuts and estimate the parameters of the Bernoulli distributions at each iterations. Each time we calculate the RMSE, AIC, BIC and mBIC.

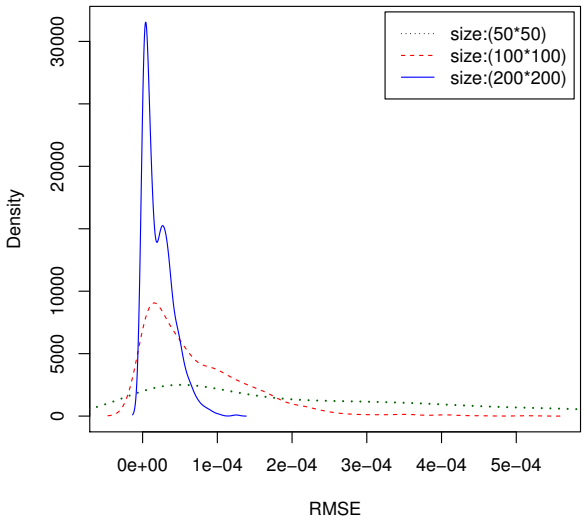


Fig. 1. Kernel density estimation of the RMSE for different sizes

TABLE II
THE PARAMETERS OF THE GENERATED DATA MATRIX

| **Domains** | **Coordinates (top left to bottom right)** | **Probability** ($p_i$) |
|---|---|---|
| Domain 1 | (1,1) — (100,20) | $p_1 = 0.1$ |
| Domain 2 | (1,20) — (100,60) | $p_2 = 0.5$ |
| Domain 3 | (1,60) — (100,90) | $p_3 = 0.9$ |
| Domain 4 | (1,90) — (100,100) | $p_4 = 0.2$ |

*1) Result on Algorithm 1:* We applied our binary segmentation algorithm to the data generated above (Table II). Table III shows that the algorithm run up to four iterations and at the end it identified seven domains. The RMSE value attains its minimum at the third iteration (Number of domains = 5), which coincides with the results given by the information criteria AIC, BIC and mBIC.

Figure 3 plots the values of the information criteria versus the number of domains; it shows that the minimal values for all three criteria correspond to five domains. Now we examine the obtained domains for their heterogeneity using the likelihood ratio test. Here, we consider only rectangle shaped domains so we do not need to check all possible comparisons. Therefore, in this example, we perform only four comparisons.

Table IV shows the obtained domains for this example. Table V illustrates the results of the likelihood ratio test. According to Table V, Domain 4 and Domain 5 can be considered as homogeneous. Thus, we combined those two

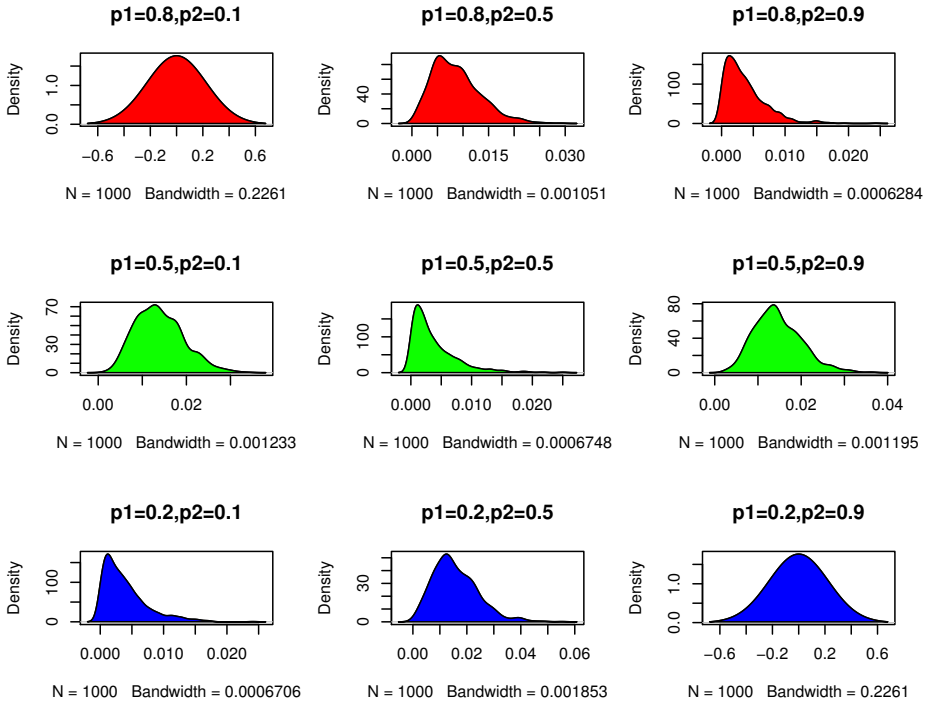


Fig. 2. Kernel density estimation of the RMSE

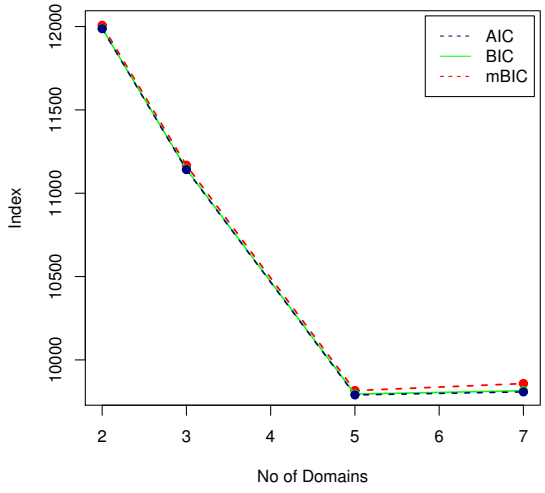


Fig. 3. The values of the AIC, BIC and mBIC

TABLE III
RESULTS ON ALGORITHM 1

| # Iterations | $D$ | RMSE | AIC | BIC | mBIC |
|---|---|---|---|---|---|
| 1 | 2 | 0.217 | 11,987.2 | 11,989.2 | 12,007.2 |
| 2 | 3 | 0.164 | 11,141.5 | 11,144.5 | 11,167.5 |
| 3 | 5 | 0.007 | 9,789.8 | 9,794.8 | 9,815.8 |
| 4 | 7 | 0.019 | 9,807.9 | 9,814.9 | 9,857.9 |

TABLE IV
OBTAINED DOMAINS FOR ALGORITHM 1

| Domains | Coordinates (top left to bottom right) |
|---|---|
| D1 | (1,1) — (100,20) |
| D2 | (1,20) — (100,60) |
| D3 | (1,60) — (100,90) |
| D4 | (1,90) — (6,100) |
| D5 | (6,90) — (100,100) |

TABLE V
LIKELIHOOD RATIO TEST FOR ALGORITHM 1

| Domain combinations | $p$-value | Results |
|---|---|---|
| D1 and D2 | $< 0.00001$ | Significant |
| D2 and D3 | $< 0.00001$ | Significant |
| D4 and D5 | 0.077242 | Not significant |

domains into one domain. Finally, we obtained the same domains as in our generated data matrix (see Table II).

*2) Results on Algorithm 2:* We applied Algorithm 2 for the same example described in the previous section and, as before, we recorded the positions of cuts at each iterations. It

TABLE VI
RESULTS ON ALGORITHM 2

| # Iterations | $D$ | RMSE | AIC | BIC | mBIC |
|---|---|---|---|---|---|
| 1 | 3 | 0.146 | 10,838.1 | 10,859.7 | 10,864.1 |
| 2 | 9 | 0.004 | 9,974.5 | 10,039.4 | 10,036.5 |
| 3 | 17 | 0.030 | 18,840.4 | 18,857.4 | 18,950.4 |

TABLE VII
RESULTS ON ALGORITHM 3

| # Iterations | $D$ | RMSE | AIC | BIC | mBIC |
|---|---|---|---|---|---|
| 1 | 2 | 0.217 | 11,987.2 | 11,989.2 | 12,007.2 |
| 2 | 3 | 0.164 | 11,141.5 | 11,144.5 | 11,167.5 |
| 3 | 4 | 0.000 | 9,790.9 | 9,794.9 | 9,822.9 |
| 4 | 5 | 0.017 | 9,880.5 | 9,885.5 | 9,918.5 |

TABLE VIII
COMPARISON OF ALL THREE ALGORITHMS (A1, A2, A3)

| | # Iterations | $D$ | RMSE | AIC | BIC | mBIC |
|---|---|---|---|---|---|---|
| A1 | 3 | 5 | 0.007 | 9,789.8 | 9,794.8 | 9,815.8 |
| A2 | 2 | 9 | 0.004 | 9,974.5 | 10,039.4 | 10,036.5 |
| A3 | 3 | 4 | 0.000 | 9,790.9 | 9,794.9 | 9,822.9 |

is clear from Table VI that the RMSE, AIC, BIC and mBIC values are lowest for the case when the number of domains is equal to nine. Thus, Algorithm 2 identified nine domains for the same example illustrated above. Furthermore, all obtained domains identified by Algorithm 2 are significantly different.

*3) Results on Algorithm 3:* In this section, we applied Algorithm 3 for the same example described in above section. We recorded the positions of cuts at each iterations. Table VII shows that the algorithm found five domains in four iterations. The RMSE, AIC, BIC and mBIC values are lowest for the case when number of domains equals four. Thus, Algorithm 3 identified four domains (the same domains as we expected) in three iterations.

### C. Comparison of the Algorithms

In this section, we compare our all algorithms. Final results of all three algorithms in the form of the RMSE, AIC, BIC and mBIC are given in Table VIII.

Our results show that the algorithms based on binary segmentation work well in identifying correct number of domains and their boundaries. Algorithm 2 finds more domains which are buried within larger domains. Algorithm 3 is fast and it is accurate in identifying major domains but overestimates the total number of domains.

## VII. DISCUSSION

There have been very few studies in the existing literature that focus on the development of statistical segmentation methods for spatial data. To address this issue, we have generalised the binary segmentation method for identifying the number of homogeneous domains and their boundaries in spatial data. In particular, we have applied the modified versions of the binary segmentation algorithms to binary spatial data indicating the presence or absence of a certain plant species, which are observed over a two dimensional lattice. The numerical results have illustrated that the algorithms work well under different scenarios; they accurately identify both the expected number of domains and their boundaries in few iterations.

Binary segmentation is described as "arguably the most widely used change-point search method" [19] and it is used for multidimensional data sequence. The benefits of binary segmentation include low computational complexity (typically of order $O(n)$), conceptual simplicity, the fact that it is usually easy to code, even in more complex models, and at each stage it involves one-dimensional rather than multi-dimensional optimization. On the other hand, the method is a "greedy" procedure in the sense that it is performed sequentially, with each stage depending on the previous ones, which are never revisited.

Analysing literature on binary segmentation, we have found out that it has been never discussed with respect to identifying both number of domains and boundaries in spatial data. To fill this gap, we develop effective procedures for estimating both the number of domains and their locations in spatial data by modifying the binary segmentation method. The applications of the proposed procedures are not limited to analysing ecological data. They can be easily extended and applied to other spatial data. For instance, it can be applied to epidemiological and economic data.

## VIII. FUTURE DIRECTIONS

Over the last decades, spatial statistical models have been studied by many authors from different angles and the spatial clustering problem is one of main topics in spatial statistics. However, the problem has not been considered as a change-point detection problem. In this study, we have demonstrated how spatial clusters can be identified by using a new approach based on binary segmentation. At this stage, we have considered a simple model which assumes that observations are independent. However, statistical models that involve spatial dependence are more realistic. Extension to dependent data is considered as one of our future works. Moreover, we have only considered rectangular shaped domains and we plan to extend it to other more complex shapes in the future.

In this work, we have used univariate binary data. It is possible to consider multivariate case (for example, for several species) and other types of data such as count or continuous data as well. Furthermore, we have assumed that data is observed over a regular shaped lattice but it is also possible to consider a set of random points on a plane. The problem that we consider can be seen as a model selection problem and one of the major challenges is to determine the optimal number of domains. We have used well-known information criteria such as the AIC, BIC and modified BIC. The criteria may not work well for spatial cluster models because of irregularities in their likelihood functions. Our intention is to develop new modified information criteria particularly for specific spatial segmentation problems under different assumptions.

In this study, we have focused on constructing binary segmentation methods because of their simplicity and low computation cost. We plan to develop new spatial segmentations algorithms bases on well-known statistical computational methods such as Cross Entropy (CE) [37], Markov chain Monte Carlo (MCMC) [39], [40] and Sequentially Importance Sampling (SIS) [38] methods.

## REFERENCES


[1] T. Y. Yang and T.B Swartz, "Application of binary segmentation to the estimation of quantal response curves and spatial intensity," *Biometrical journal,* vol. 4, 2005, pp. 489–501. https://doi.org/10.1002/bimj.200310136

[2] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1988 IEEE International Conference on,* vol. 2, 1998, pp. 645–648. https://doi.org/10.1109/ICASSP.1998.675347

[3] A. K. Tung, J. Hou and J. Han, "Spatial clustering in the presence of obstacles," In *Data Engineering, 2001. Proceedings. 17th International Conference on,* IEEE, 2001, pp. 359–367. https://doi.org/10.1109/ICDE.2001.914848

[4] R. E. Gangnon and M.K Clayton, "Bayesian detection and modeling of spatial disease clustering," *Biometrics,* vol. 3, 2000, pp. 922–935. https://doi.org/10.1111/j.0006-341X.2000.00922.x

[5] C. Anderson, D. Lee and N. Dean, "Bayesian cluster detection via adjacency modelling." *Spatial and spatio-temporal epidemiology,* 2016, pp. 11–20. https://doi.org/10.1016/j.sste.2015.11.005

[6] B. Beckage, L. Joseph, P. Belisle, D. B. Wolfson and W. J. Platt, "Bayesian change-point analyses in ecology," *New Phytologist,* vol. 2, 2007, pp. 11–20. https://doi.org/10.1111/j.1469-8137.2007.01991.x

[7] I. López, M. Gámez, J. Garay, T. Standovár and Z. Varga, "Applications of change-point problem to the detection of plant patches," *Acta biotheoretica,* vol. 1, 2010, pp. 51–63. https://doi.org/10.1007/s10441-009-9093-x

[8] S. Tripathi and R. S. Govindaraju, "Change detection in rainfall and temperature patterns over India," In *Proceedings of the Third International Workshop on Knowledge Discovery from Sensor Data,* ACM, 2009, pp. 133–141. https://doi.org/10.1145/1601966.1601988

[9] J. D. Helterbrand, N. Cressie and J. L. Davidson, "A statistical approach to identifying closed object boundaries in images," *Advances in applied probability,* vol. 4, 1994, pp. 831–854. https://doi.org/10.1017/S0001867800026641

[10] Y. Wang, "Change curve estimation via wavelets," *Journal of the American Statistical Association,* vol. 441, 1998, pp. 163–172. http://dx.doi.org/10.1080/01621459.1998.10474098

[11] G. J. Upton and B. Fingleton, "Spatial data analysis by example. vol. 1: Point pattern and quantitative data," *Chichester: Wiley,* vol. 1, 1985.

[12] A. D. Cliff and J. K. Ord, *Spatial processes: models & applications,* Taylor & Francis, 1981.

[13] N. Cressie, *Statistics for spatial data,* John Wiley and Sons, 2015.

[14] A. J. Scott and M. Knott, "A cluster analysis method for grouping means in the analysis of variance," *Biometrics,* 1974, pp. 507–512. https://doi.org/10.2307/2529204

[15] A. Sen and M. S. Srivastava, "On tests for detecting change in mean," *The Annals of statistics,* vol. 1, 1975, pp. 98–108. https://doi.org/10.1214/aos/1176343001

[16] L. Vostrikova, "Detection of the disorder in multidimensional random-point problems," *Doklady Akademii Nauk SSSR,* vol. 2, 1998, pp. 270–274.

[17] I. A. Eckley, P. Fearnhead and R. Killick, "Analysis of changepoint models," *Bayesian Time Series Models,* 2011, pp. 205–224.

[18] E. S. Venkatraman, *Consistency results in multiple change-point problems,* PhD thesis, to the Department of Statistics. Stanford University. 1992.

[19] R. Killick, P. Fearnhead and I. A. Eckley "Optimal detection of change-points with a linear computational cost," *Journal of the American Statistical Association,* vol. 500, 2012, pp. 1590–1598. http://dx.doi.org/10.1080/01621459.2012.737745

[20] W. Li, "DNA segmentation as a model selection process," In *proceedings of the fifth annual international conference on Computational biology,* ACM, 2001, pp. 204–210. http://dx.doi.org/10.1145/369133.369202

[21] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control,* vol. 6, 1974, pp. 716–723. http://dx.doi.org/10.1109/TAC.1974.1100705

[22] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics,* vol. 2, 1978, pp. 461–464. http://dx.doi.org/10.1214/aos/1176344136

[23] J. Chen, A. Gupta and J. Pan "Information criterion and change point problem for regular models," *Sankhyā: The Indian Journal of Statistics,* 2006, pp. 252–282.

[24] J. Chen and A. K. Gupta, "Testing and locating variance changepoints with application to stock prices," *Journal of the American Statistical association,* vol. 438, 1997, pp. 739–747. http://dx.doi.org/10.1080/01621459.1997.10474026

[25] T. Young Yang and L. Kuo, "Bayesian binary segmentation procedure for a poisson process with multiple changepoints," *Journal of Computational and Graphical Statistics,* vol. 4, 1998, pp. 772–785. http://dx.doi.org/10.1198/106186001317243449

[26] R. Killick, I. A. Eckley, P. Jonathan and U. Chester, "Efficient detection of multiple changepoints within an oceano-graphic time series," In *Proceedings of the 58th World Science Congress of ISI,* 2011.

[27] P. Fryzlewicz, "Wild binary segmentation for multiple change-point detection," *The Annals of Statistics,* vol. 6, 2014, pp. 2243–2281. http://dx.doi.org/10.1214/14-AOS1245

[28] W. J. R. M. Priyadarshana, *The cross-entropy method and multiple change-point detection in genomic sequences,* PhD thesis, Macquarie University, 2015.

[29] H. Cho and P. Fryzlewicz, "Multiple-change-point detection for high dimensional time series via sparsified binary segmentation," *Journal of the Royal Statistical Society: series B (statistical methodology),* vol. 2, 2015, pp. 475–507. http://dx.doi.org/10.1111/rssb.12079

[30] A. B. Olshen, E. Venkatraman, R. Lucito and M. Wigler,"Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics,* vol. 4, 2004, pp. 557–572. https://doi.org/10.1093/biostatistics/kxh008

[31] J. Chen and A. K. Gupta, *Parametric statistical change point analysis: with applications to genetics, medicine, and finance,* Springer Science & Business Media, 2011.

[32] D. V. Hinkley and E. A. Hinkley, "Inference about the change-point in a sequence of binomial variables," *Biometrika,* vol. 3, 1970, pp. 477–488. https://doi.org/10.2307/2334766

[33] B. Jackson, J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan and T. T. Tsai,"An algorithm for optimal partitioning of data on an interval," *IEEE Signal Processing Letters,* vol. 12, 2005, pp. 105–108. https://doi.org/10.1109/LSP.2001.838216

[34] J. Pan and J. Chen, "Application of modified information criterion to multiple change point problems," *Journal of multivariate analysis,* vol. 10, 2006, pp. 2221–2241. https://doi.org/10.1016/j.jmva.2006.05.009

[35] A. N. Shiryaev, "On optimum methods in quickest detection problems," *Theory of Probability & Its Applications,* vol. 1, 1963, pp. 22–46. https://doi.org/10.1137/1108002

[36] N. R. Zhang and D. O. Siegmund, "A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data," *Biometrics,* vol. 1, 2007, pp. 22–32. https://doi.org/10.1111/j.1541-0420.2006.00662.x

[37] T. Polushina and G. Sofronov, "Change-point detection in binary Markov DNA sequences by the Cross-Entropy method," In *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on,* IEEE, vol. 2, 2014, pp. 471–478. https://doi.org/10.15439/2014F88

[38] G. Yu Sofronov, G. E. Evans, J. M. Keith and D. P. Kroese, "Identifying change-points in biological sequences via sequential importance sampling," In *Environmental Modeling & Assessment,* vol. 5, 2009, pp. 577–584. https://doi.org/10.1007/s10666-008-9160-8

[39] G. Sofronov, "Change-Point Modelling in Biological Sequences via the Bayesian Adaptive Independent Sampler," In *International Proceedings of Computer Science and Information Technology,* vol. 5, 2011, pp. 122–126.

[40] M. Manuguerra, G. Sofronov, M. Tani and G. Heller, "Monte Carlo methods in spatio-temporal regression modeling of migration in the EU," In *Computational Intelligence for Financial Engineering & Economics (CIFEr), 2013 IEEE Conference on,* IEEE, 2013, pp. 128–134. https://doi.org/10.1109/CIFEr.2013.6611708