Stat

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.161

Uncertainty of a detected spatial cluster in 1D: quantification and visualization

Junho Lee[a], **Ronald E. Gangnon**[b,c], **Jun Zhu**[d,e] **and Jingjing Liang**[f,g]

Received 21 August 2017; Accepted 1 September 2017

Spatial cluster detection is an important problem in a variety of scientific disciplines such as environmental sciences, epidemiology and sociology. However, there appears to be very limited statistical methodology for quantifying the uncertainty of a detected cluster. In this paper, we develop a new method for the quantification and visualization of uncertainty associated with a detected cluster. Our approach is defining a confidence set for the true cluster and visualizing the confidence set, based on the maximum likelihood, in time or in one-dimensional space. We evaluate the pivotal property of the statistic used to construct the confidence set and the coverage rate for the true cluster via empirical distributions. For illustration, our methodology is applied to both simulated data and an Alaska boreal forest dataset. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: change-point analysis; confidence set; spatial cluster detection; spatial cluster model; uncertainty quantification; uncertainty visualization

1 Introduction

Cluster detection in spatial data is the identification of spatial units, possibly during a time period, which show distinctive patterns. For count data, spatial clusters have distinctive risks that are typically elevated, but possibly reduced, relative to background variation. For continuous data, spatial clusters show substantially higher or lower mean values than the background does.

Approaches to cluster detection in space and/or time have been discussed under frequentist framework as well as Bayesian framework. Spatial scan statistic (Kulldorff & Nagarwalla, 1995; Kulldorff, 1997), spatio-temporal scan statistic (Kulldorff et al., 1998; Kulldorff, 2001) and their many variants (e.g. Duczmal & Assunção, 2004; Gangnon & Clayton, 2004; Tango & Takahashi, 2005; Assunção et al., 2006; Kulldorff et al., 2006; Takahashi et al., 2008; Kulldorff et al., 2009; Gangnon, 2010a; Neill, 2012; Shu et al., 2012; Xu & Gangnon, 2016; Lin et al., 2016) are popular approaches to cluster detection within a frequentist hypothesis testing framework. Alternatively, Gangnon & Clayton (2000, 2003, 2007), Lawson (2000), Clark & Lawson (2002), Yan & Clayton (2006), Gangnon (2010b) and Wakefield & Kim (2013) used Bayesian models for the underlying event rates that incorporate explicit spatial or

[a]CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia
[b]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI 53726, USA
[c]Department of Population Health Sciences, University of Wisconsin, Madison, WI 53726, USA
[d]Department of Statistics, University of Wisconsin, Madison, WI 53706, USA
[e]Department of Entomology, University of Wisconsin, Madison, WI 53706, USA
[f]School of Natural Resources, West Virginia University, Morgantown, WV 26505, USA
[g]School of Forestry, Beijing Forestry University, Beijing 100083, China
*Email: junho.lee@kaust.edu.sa

spatio-temporal clusters associated with distinctive risks. Most recently, Lee et al. (2017) proposed a spatial clustering method for spatial regression coefficients, which enables the detection of an unknown number of spatial clusters in the regression coefficients via hypothesis testing and carried out spatially varying-coefficient regression based on spatial clustering. None of the existing methods for cluster detection, however, quantifies the uncertainty associated with a detected cluster.

Here, we consider a new problem, namely, the quantification and visualization of the uncertainty associated with a detected cluster. This is a challenging problem because there seems to be very little literature on this topic. Thus, we restrict our attention to the one-dimensional (1D) case, either in time or in space. We develop a likelihood-based approach to define a confidence set of the cluster, and to visualize the confidence set. Empirical distribution of the null distribution shows that a pivotal property holds, which enables a confidence set for an unknown cluster be constructed based on the null distribution. Our proposed visualization of a confidence set provides new insight into the location and extent of the cluster. We believe that this method is the first of its kind to express the uncertainty of a detected cluster in space. Further, we compare our method with some of the existing change-point analysis by simulated and real data examples (Scott & Knott, 1974; Sen & Srivastava, 1975; Killick et al., 2012).

The remainder of this paper is organized as follows. In Section 2, we introduce a spatial scan statistic for cluster detection based on Gaussian data in the 1D space. In Section 3, we develop a confidence set for the spatial cluster and its visualization. The proposed methodology is evaluated by a simulation study. An forest ecological dataset is analysed for illustration in Section 4.

2 Cluster detection

Let $\mathcal{D}$ denote a spatial domain of interest. Let N denote the number of cells that partition the spatial domain $\mathcal{D}$ and form a spatial lattice. For cell $i = 1, \ldots, N$, let y_i denote the observation in cell i. Let $\mathcal{C} = \{C_1, C_2, \ldots\}$ denote the set of all candidate clusters, where each cluster C_j, $j = 1, 2, \ldots$, is a subdomain of $\mathcal{D}$ and defined as a set of adjacent cells.

Kulldorff & Nagarwalla (1995) and Kulldorff (1997) proposed the spatial scan statistic, which is a maximum likelihood ratio test (LRT) statistic over all candidate clusters, for count data, and Kulldorff et al. (2009) developed a scan statistic for continuous data. Here, we will focus on continuous data and assume that y_i follows a Gaussian distribution. The methodology developed here could be adapted to count data or continuous data that are not necessarily Gaussian.

For an unknown cluster $C \in \mathcal{C}$, we model the ith observation as

$$y_i = \mu + \theta \cdot \mathcal{I}\{i \in C\} + \varepsilon_i, \tag{1}$$

where μ is the mean value for the background (i.e. the spatial units not in the cluster C), θ is the cluster effect associated with C, $\mathcal{I}(\cdot)$ is the indicator function and the random error ε_i's are assumed to be iid $N(0, \sigma^2)$ with a variance component $\sigma^2 > 0$.

To detect a cluster, we first consider the null hypothesis H_{0_k} versus a cluster-specific local hypothesis H_{C_k} as the alternative for $C_k \in \mathcal{C}$, $k = 1, 2, \ldots$:

$$H_{0_k} : \theta_{C_k} = 0 \quad \text{versus} \quad H_{C_k} : \theta_{C_k} \neq 0, \tag{2}$$

where θ_{C_k} is the cluster effect associated with the kth candidate cluster C_k. For a given cluster C_k, we define an LRT statistic as

$$\lambda(C_k) = \frac{\mathcal{L}_{C_k}(\hat{\mu}_{C_k}, \hat{\theta}_{C_k}, \hat{\sigma}^2_{C_k})}{\mathcal{L}_{0_k}(\hat{\mu}_{0_k}, \hat{\theta}_{0_k}, \hat{\sigma}^2_{0_k})} = \left(\frac{\hat{\sigma}^2_{C_k}}{\hat{\sigma}^2_{0_k}}\right)^{-N/2},$$

where $\mathcal{L}_{0_k}$ and $\mathcal{L}_{C_k}$ are the Gaussian likelihood function evaluated at the maximum likelihood estimates (MLEs) of (μ, θ, σ^2) in (1), $(\hat{\mu}_{0_k}, \hat{\theta}_{0_k}, \hat{\sigma}^2_{0_k})$ and $(\hat{\mu}_{C_k}, \hat{\theta}_{C_k}, \hat{\sigma}^2_{C_k})$, under the H_{0_k} and H_{C_k} in (2), respectively.

Next, we consider the null hypothesis H_0 versus a global hypothesis H_C as the alternative for an unknown generic cluster C in the candidate set $\mathcal{C}$:

$$H_0 : \theta_C = 0 \quad \text{versus} \quad H_C : \theta_C \neq 0. \tag{3}$$

From all the possible local hypotheses given in (2), we define a global test statistic for testing the H_0 versus H_C in (3) to be

$$\nu = \max_{C \in \mathcal{C}} \{\lambda(C)\}. \tag{4}$$

The test statistic (4) is, among all the candidate clusters in $\mathcal{C} = \{C_1, C_2, \ldots\}$, the largest. The candidate cluster that corresponds to the test statistic ν in (4) is the cluster estimate and is denoted by $\hat{C}$. That is,

$$\hat{C} = \arg\max_{C \in \mathcal{C}} \{\lambda(C)\}. \tag{5}$$

Further, to compute a p-value based on the test statistic (4), we adopt a Monte Carlo method in the spirit of a parametric bootstrap. First, we compute $\hat{\mu}_0$ and $\hat{\sigma}^2_0$, the MLEs of the parameters, under the H_0. Second, we generate Monte Carlo samples $y_i^{\text{new}} = \hat{\mu}_0 + \varepsilon_i^{\text{new}}$, where $\varepsilon_i^{\text{new}} \sim$ iid $N(0, \hat{\sigma}^2_0)$ for $i = 1, \ldots, N$. Third, we compute the test statistic (4) for each Monte Carlo sample. Suppose there are S random Monte Carlo samples. The p-value is $R/(S+1)$, where R is the rank of the test statistic (4) for the original dataset in comparison with all the Monte Carlo samples, and the largest test statistic gets a rank of 1.

3 Confidence set

In Section 2, the H_C in (3) is considered as the alternative hypothesis for spatial cluster detection. However, in this section, we consider this H_C as the null hypothesis to draw inference about the unknown cluster C. For a given cluster C under the H_C, an LRT statistic can be defined as

$$\Lambda(C) = \frac{\max_{\theta} \mathcal{L}(C, \theta)}{\max_{C^*} \left\{ \max_{\theta^*} \mathcal{L}(C^*,\ \theta^*) \right\}} = \frac{\max_{\theta} \mathcal{L}(C, \theta)}{\max_{C^*,\ \theta^*} \mathcal{L}(C^*,\ \theta^*)}, \tag{6}$$

where $\mathcal{L}(C, \theta)$ is the likelihood evaluated at θ. Further, for the Gaussian data, the log-likelihood function is

$$\Phi(C \mid N) = -(2/N) \log \Lambda(C) = \log \hat{\sigma}^2_C - \log \hat{\sigma}^2_A, \tag{7}$$

where $\hat{\sigma}^2_A = \min_{C^*,\ \theta^*} \hat{\sigma}^2\{C^*,\ \theta^*\}$, $\hat{\sigma}^2_C = \min_{\theta} \hat{\sigma}^2\{C, \theta\}$ and $\hat{\sigma}^2\{C, \theta\}$ is the MLE of σ^2 given (C, θ). Further, $\hat{\sigma}^2_A$ is equivalent to $\hat{\sigma}^2_C$ evaluated at $\hat{C}$, where $\hat{C}$ is the cluster estimate in (5).

(wileyonlinelibrary.com) DOI: 10.1002/sta4.161

The ISI's Journal for the Rapid
Dissemination of Statistics Research

3.1 Pivotal property of the null distribution

The distribution of (7) satisfies the pivotal property if the distribution is identical for all $C \in \mathcal{C}$ (Lehmann, 1986). That is,

$$-(2/N)\log\Lambda(C_{k_1}) \overset{d}{=} -(2/N)\log\Lambda(C_{k_2}) \quad \text{for} \quad \forall\, C_{k_1}, C_{k_2} \in \mathcal{C},\ C_{k_1} \neq C_{k_2}, \tag{8}$$

where $\overset{d}{=}$ denotes the equivalence in distribution of two random variables. If condition (8) holds, then general inference about the unknown cluster C can be made based on the pivotal property. As an analytic distribution is not available for the statistic $\Phi(C\,|\,N)$ in (7), we examine the empirical distributions via a set of simulations.

In particular, we obtain the empirical distributions of $\Phi(C\,|\,N)$ in (7) for different clusters on a $1 \times N$ 1D lattice, where N is set to be 100, 200 or 300. We consider 12 clusters for the three lattice sizes N, respectively. These clusters are defined as

$$\begin{aligned}
N = 100: \quad & C_1 = \{i \mid |i-50| \leqslant 10\}, \quad C_2 = \{i \mid |i-50| \leqslant 12\}, \quad C_3 = \{i \mid |i-50| \leqslant 14\}, \\
& C_4 = \{i \mid |i-50| \leqslant 16\}, \quad C_5 = \{i \mid |i-50| \leqslant 18\}, \quad C_6 = \{i \mid |i-50| \leqslant 20\}, \\
& C_7 = \{i \mid |i-75| \leqslant 10\}, \quad C_8 = \{i \mid |i-75| \leqslant 12\}, \quad C_9 = \{i \mid |i-75| \leqslant 14\}, \\
& C_{10} = \{i \mid |i-75| \leqslant 16\}, \quad C_{11} = \{i \mid |i-75| \leqslant 18\}, \quad C_{12} = \{i \mid |i-75| \leqslant 20\}, \\
N = 200: \quad & C_1 = \{i \mid |i-100| \leqslant 10\}, \quad C_2 = \{i \mid |i-100| \leqslant 16\}, \quad C_3 = \{i \mid |i-100| \leqslant 22\}, \\
& C_4 = \{i \mid |i-100| \leqslant 28\}, \quad C_5 = \{i \mid |i-100| \leqslant 34\}, \quad C_6 = \{i \mid |i-100| \leqslant 40\}, \\
& C_7 = \{i \mid |i-150| \leqslant 10\}, \quad C_8 = \{i \mid |i-150| \leqslant 16\}, \quad C_9 = \{i \mid |i-150| \leqslant 22\}, \\
& C_{10} = \{i \mid |i-150| \leqslant 28\}, \quad C_{11} = \{i \mid |i-150| \leqslant 34\}, \quad C_{12} = \{i \mid |i-150| \leqslant 40\}, \\
N = 300: \quad & C_1 = \{i \mid |i-150| \leqslant 10\}, \quad C_2 = \{i \mid |i-150| \leqslant 20\}, \quad C_3 = \{i \mid |i-150| \leqslant 30\}, \\
& C_4 = \{i \mid |i-150| \leqslant 40\}, \quad C_5 = \{i \mid |i-150| \leqslant 50\}, \quad C_6 = \{i \mid |i-150| \leqslant 60\}, \\
& C_7 = \{i \mid |i-225| \leqslant 10\}, \quad C_8 = \{i \mid |i-225| \leqslant 20\}, \quad C_9 = \{i \mid |i-225| \leqslant 30\}, \\
& C_{10} = \{i \mid |i-225| \leqslant 40\}, \quad C_{11} = \{i \mid |i-225| \leqslant 50\}, \quad C_{12} = \{i \mid |i-225| \leqslant 60\}.
\end{aligned}$$

Further, we consider two different signal-to-noise ratios (SNRs: θ/σ) as 1 or 2. We obtain the distribution of $\Phi(C\,|\,N)$ in (7) from 1000 simulations assuming that each cluster is the true cluster for each combination of N and SNR. The set of candidate clusters $\mathcal{C}$ is predefined for each N as

$$\mathcal{C} = \left\{ \{i \mid |i-c| \leqslant r\} \;\middle|\; c = 1, \ldots, N,\ \ r = 0, 1, \ldots, R_{\max} \right\}, \tag{9}$$

where we set $R_{\max\,|\,N=100} = 24$, $R_{\max\,|\,N=200} = 49$ and $R_{\max\,|\,N=300} = 74$ for $N = 100$, 200 and 300, respectively. After obtaining these empirical distributions, we examine the Q–Q plots of $\Phi(C_{k_1}|N)$ versus $\Phi(C_{k_2}|N)$ for the two different clusters C_{k_1} and C_{k_2}. Figure 1 shows the results for $N = 100$ and $\text{SNR} = 1$. The other cases yield similar results, and the figures are omitted. These Q–Q plots support the pivotal property of the null distribution, based on which inference can be drawn about the unknown cluster C.

3.2 Confidence set for the cluster

With the set of candidate clusters $\mathcal{C}$, we define a $(1-\alpha)100\%$ confidence set for a given "true" cluster C_0 as

$$\Psi_{1-\alpha}(C_0) = \left\{ C \in \mathcal{C} \;\middle|\; \Phi(C\,|\,N) \leqslant \mathrm{P}_{100\times(1-\alpha)}(\hat{C}) \right\}, \tag{10}$$

(wileyonlinelibrary.com) DOI: 10.1002/sta4.161

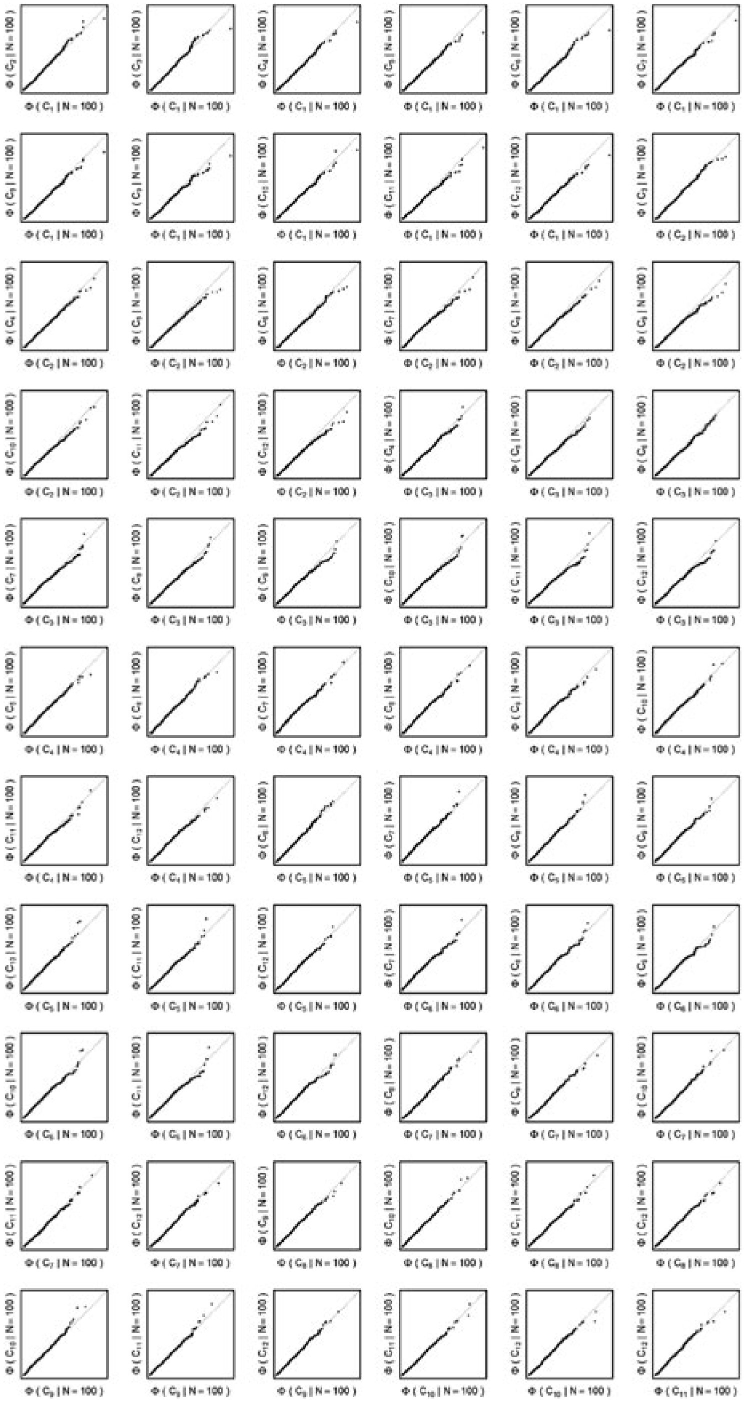


Figure 1. Q–Q plots of the empirical distribution of $\Phi(C\,|\,N) = -(2/N)\log\Lambda(C)$ from the 1000 simulations when $N = 100$ and the SNR (θ/σ) is 1. $C_1 = \{i \mid |i-50| \leqslant 10\}$, $C_2 = \{i \mid |i-50| \leqslant 12\}$, $C_3 = \{i \mid |i-50| \leqslant 14\}$, $C_4 = \{i \mid |i-50| \leqslant 16\}$, $C_5 = \{i \mid |i-50| \leqslant 18\}$, $C_6 = \{i \mid |i-50| \leqslant 20\}$, $C_7 = \{i \mid |i-75| \leqslant 10\}$, $C_8 = \{i \mid |i-75| \leqslant 12\}$, $C_9 = \{i \mid |i-75| \leqslant 14\}$, $C_{10} = \{i \mid |i-75| \leqslant 16\}$, $C_{11} = \{i \mid |i-75| \leqslant 18\}$ and $C_{12} = \{i \mid |i-75| \leqslant 20\}$.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.161

where $\hat{C}$ is the cluster estimate given in (5), and $\mathrm{P}_{100\times(1-\alpha)}(\hat{C})$ is the $100 \times (1-\alpha)$th percentile of the distribution of $\Phi(C \mid N)$ in (7) assuming that $\hat{C}$ is the true cluster.

To evaluate the coverage rate of the confidence set $\Psi_{1-\alpha}(C_0)$ given in (10), we consider three different $1 \times N$ 1D lattices with $N = 100, 200, 300$. Further, for all the three lattices, we set the true cluster as $C_0 = \{i \mid |i - 50| \leqslant 20\}$ and three different SNRs of 2, 1 or 1/2. We simulate 100 datasets for each combination of N and SNR and construct a 95% confidence set $\Psi_{0.95}(C_0)$, with 1000 null simulations to obtain the 95th percentile $\mathrm{P}_{95}(\hat{C})$, for each simulated dataset. The set of candidate clusters $\mathcal{C}$ is predefined as in (9) with $R_{\max} = 24$ for all Ns.

The empirical coverage rate, which is the proportion of the simulations in which the 95% confidence set $\Psi_{0.95}(C_0)$ contains the true cluster C_0, is around 95% with SNR $= 2$, 93% with SNR $= 1$, and 90% with SNR $= 1/2$, respectively. As the SNR decreases, the coverage rate of $\Psi_{0.95}(C_0)$ tends to decrease but is close to 95% with SNR $= 2$. This result shows that the confidence set $\Psi_{1-\alpha}(C_0)$ in (10) is suitable for quantifying the uncertainty associated with the estimated cluster $\hat{C}$ in (5) at least when the cluster effect is relatively strong.

3.3 Relation to the p-value

The cluster estimate $\hat{C}$ in (5) plays an important role in constructing the confidence set $\Psi_{1-\alpha}(C_0)$ in (10). Next, we explore the connection between the *p*-value computed in Section 2 and the number of clusters in the confidence set, developed in Section 3.2 where the confidence set is defined based on the LRT statistic $\Lambda(C)$ defined in (6), while the *p*-value is based on the statistic ν defined in (4). The sampling distribution for $\Lambda(C)$ and that for ν are unknown. We expect that the confidence set would have a smaller number of clusters if the *p*-value is smaller. Thus, here, we make an attempt to explain the confidence set in relation to the *p*-value, conceptually.

For the Gaussian data, we have

$$-(2/N)\log\nu = \log\hat{\sigma}^2_{\hat{C}} - \log\hat{\sigma}^2_0, \tag{11}$$

$$-(2/N)\log\Lambda(C) = \log\hat{\sigma}^2_C - \log\hat{\sigma}^2_{\hat{C}} = \left(\log\hat{\sigma}^2_C - \log\hat{\sigma}^2_0\right) - \left(\log\hat{\sigma}^2_{\hat{C}} - \log\hat{\sigma}^2_0\right), \tag{12}$$

where $\hat{\sigma}^2_0$ is the MLE of σ^2 under the assumption that there is no cluster. With the cluster estimate $\hat{C}$ in (5), the term in (11) would have a negative value $\log\hat{\sigma}^2_{\hat{C}} - \log\hat{\sigma}^2_0 < 0$. In addition to $\hat{C}$, there may be some similar clusters that overlap substantially with $\hat{C}$. Define a set $\mathcal{S}(\hat{C})$, which consists of such "similar" clusters that $\hat{\sigma}^2_C/\hat{\sigma}^2_{\hat{C}} \approx 1$ for $C \in \mathcal{C}$. Then, the value of $\log\hat{\sigma}^2_C - \log\hat{\sigma}^2_0$ in (12) can be expected to be $\log\hat{\sigma}^2_C - \log\hat{\sigma}^2_0 < 0$ for $C \in \mathcal{S}(\hat{C})$ and $\log\hat{\sigma}^2_C - \log\hat{\sigma}^2_0 \approx 0$ for $C \in \mathcal{C} \setminus \mathcal{S}(\hat{C})$. Then, the terms in (12) become $\log\hat{\sigma}^2_C - \log\hat{\sigma}^2_{\hat{C}} \approx 0$ for $C \in \mathcal{S}(\hat{C})$ and $\log\hat{\sigma}^2_C - \log\hat{\sigma}^2_{\hat{C}} = \left(\log\hat{\sigma}^2_C - \log\hat{\sigma}^2_0\right) - \left(\log\hat{\sigma}^2_{\hat{C}} - \log\hat{\sigma}^2_0\right) > 0$ for $C \in \mathcal{C} \setminus \mathcal{S}(\hat{C})$.

As the percentile value $\mathrm{P}_{100\times(1-\alpha)}(\hat{C})$ in (10) would be around 0, the clusters in $\mathcal{S}(\hat{C})$ are more likely to belong to $\Psi_{1-\alpha}(C_0)$, while those not in $\mathcal{S}(\hat{C})$ are less likely to belong to the confidence set. That is, the number of clusters in $\Psi_{1-\alpha}(C_0)$, denoted as $|\Psi_{1-\alpha}(C_0)|$, is expected to be proportional to the number of clusters in $\mathcal{S}(\hat{C})$, denoted as $|\mathcal{S}(\hat{C})|$.

Now, we consider two extreme cases when *p*-value is nearly 0 or nearly 1. If *p*-value ≈ 0, then the effect of cluster $\hat{C}$ is expected to be strong and (11) will have a large negative value (i.e. $\log\hat{\sigma}^2_{\hat{C}} - \log\hat{\sigma}^2_0 \ll 0$). Thus, the value in (12) will have a large positive value (i.e. $\left(\log\hat{\sigma}^2_C - \log\hat{\sigma}^2_0\right) - \left(\log\hat{\sigma}^2_{\hat{C}} - \log\hat{\sigma}^2_0\right) \gg 0$) for $C \in \mathcal{C} \setminus \mathcal{S}(\hat{C})$ as those clusters are unlikely to be in the confidence set $\Psi_{1-\alpha}(C_0)$. On the other hand, if *p*-value ≈ 1, the effect of $\hat{C}$ is expected to be negligible, and the value in (11) will be close to 0 (i.e. $\log\hat{\sigma}^2_{\hat{C}} - \log\hat{\sigma}^2_0 \approx 0$) and $\log\hat{\sigma}^2_C - \log\hat{\sigma}^2_0 \approx 0$ for all $C \in \mathcal{C}$.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.161

Thus, the value in (12) will be also close 0 as $\left(\log \hat{\sigma}^2_{C} - \log \hat{\sigma}^2_0\right) - \left(\log \hat{\sigma}^2_{\hat{C}} - \log \hat{\sigma}^2_0\right) \approx 0$, and almost all clusters in $\mathcal{C}$ are likely to belong to the confidence set $\Psi_{1-\alpha}(C_0)$.

Therefore, we expect that, as the *p*-value approaches 1, the number of clusters in $\mathcal{S}(\hat{C})$ approaches the size of $\mathcal{C}$, $|\mathcal{S}(\hat{C})| \to |\mathcal{C}|$, and the number of clusters in $\Psi_{1-\alpha}(C_0)$ approaches the size of $\mathcal{C}$, $\left|\Psi_{1-\alpha}(C_0)\right| \longrightarrow |\mathcal{C}|$. This relation between the *p*-value and the confidence set will be further investigated in Section 3.4 in a simulation study.

3.4 Visualization

In Section 3.2, we quantified the uncertainty of a detected cluster $\hat{C}$ by a $(1-\alpha)$ confidence set $\Psi_{1-\alpha}(C_0)$ base on likelihood. The confidence set consists of spatial clusters, and each cluster in turn consists of contiguous spatial units. Thus, it is a challenge to illustrate the confidence set that consists of sets of spatial units. Here, we propose a way to visualize the confidence set and illustrate it by a 95% confidence set $\Psi_{0.95}(C_0)$ (Figure 2). Two datasets are simulated on a $1 \times N$ 1D spatial lattice, where $N = 100$, with the true cluster $C_0 = \left\{i \;\middle|\; |i - 75| \leqslant 10\right\}$ and the SNR θ/σ is set to be 1 or 0.5. The 95th percentile $\mathrm{P}_{95}(\hat{C})$ is obtained from 1000 simulations assuming that $\hat{C}$ is the true cluster for each dataset. The set of candidate clusters $\mathcal{C}$ is predefined as (9) with $R_{\max} = 24$.

A 95% confidence set $\Psi_{0.95}(C_0)$ is illustrated in Figure 2a for SNR $= 1$ and in Figure 2b for SNR $= 0.5$, respectively. In each visualization, the top horizontal axis represents the 1×100 1D space, and the segments underneath are the clusters that belong to the confidence set and are ordered vertically by the log-LRT statistic $\log \Lambda(C)$ with $\hat{C}$ at the top. Each cluster is represented as a line segment between two spatial units on the boundary of the cluster. The top most segment in black is the cluster estimate $\hat{C}$, and the other clusters in the confidence set are displayed in grey. The cluster estimate is close to the true cluster $C_0 = \left\{i \;\middle|\; |i - 75| \leqslant 10\right\}$ as $\hat{C} = \left\{i \;\middle|\; |i - 75| \leqslant 11\right\}$ in Figure 2a, while it is far from C_0 in Figure 2b as $\hat{C} = \left\{i \;\middle|\; |i - 42| \leqslant 0\right\}$. At the bottom of each visualization in Figure 2, the first greyscale bar shows the frequency of each spatial unit that is contained in the clusters in the confidence set. The second greyscale bar is a weighted frequency based on the LRT statistic $\Lambda(C)$, weighted by $w_i = \sum_{C \in \Psi_{0.95}(C_0)} \Lambda(C) \cdot \mathcal{I}\{i \in C\} \;/\; \sum_{C \in \Psi_{0.95}(C_0)} \Lambda(C)$, for the *i*th spatial unit, where $i = 1, \ldots, N$.

The number of candidate clusters is $|\mathcal{C}| = 2500$. There are much more clusters in the confidence set in Figure 2b than in Figure 2a. Specifically, the *p*-value is 0.001 and $|\Psi_{0.95}(C_0)| = 70$ in Figure 2a, while, in Figure 2b, the *p*-value is 0.588 and $|\Psi_{0.95}(C_0)| = 1613$. This is as expected owing to the findings in Section 3.3.

The visualization we propose here enables us not only to see the number of clusters in a confidence set but also to identify where these clusters are located in the spatial domain. Further, the vertical arrangement of the clusters according to the log-LRT statistic and the greyscales provide information about each spatial unit in terms of its chance to belong to the true cluster.

4 Numerical examples

4.1 Simulation study

We simulate two datasets on a 1×100 spatial lattice with the true cluster defined to be $C_0 = \left\{i \;\middle|\; |i - 50| \leqslant 10\right\}$ for $(\mu, \theta, \sigma^2) = (0, 1, 1)$ or $(\mu, \theta, \sigma^2) = (0, 2, 1)$. That is, the 40th and 60th cells are the spatial units on the boundary of the true change set. In addition, the SNR $= 1$ is for the first dataset and SNR $= 2$ is for the second one. We apply the approach developed in Sections 2 and 3.

For comparison, we apply a binary segmentation (BinSeg) and pruned exact linear time (PELT) methods for each simulated dataset. Spatial cluster detection can be seen as a kind of spatial change-point detection. In the 1D space,

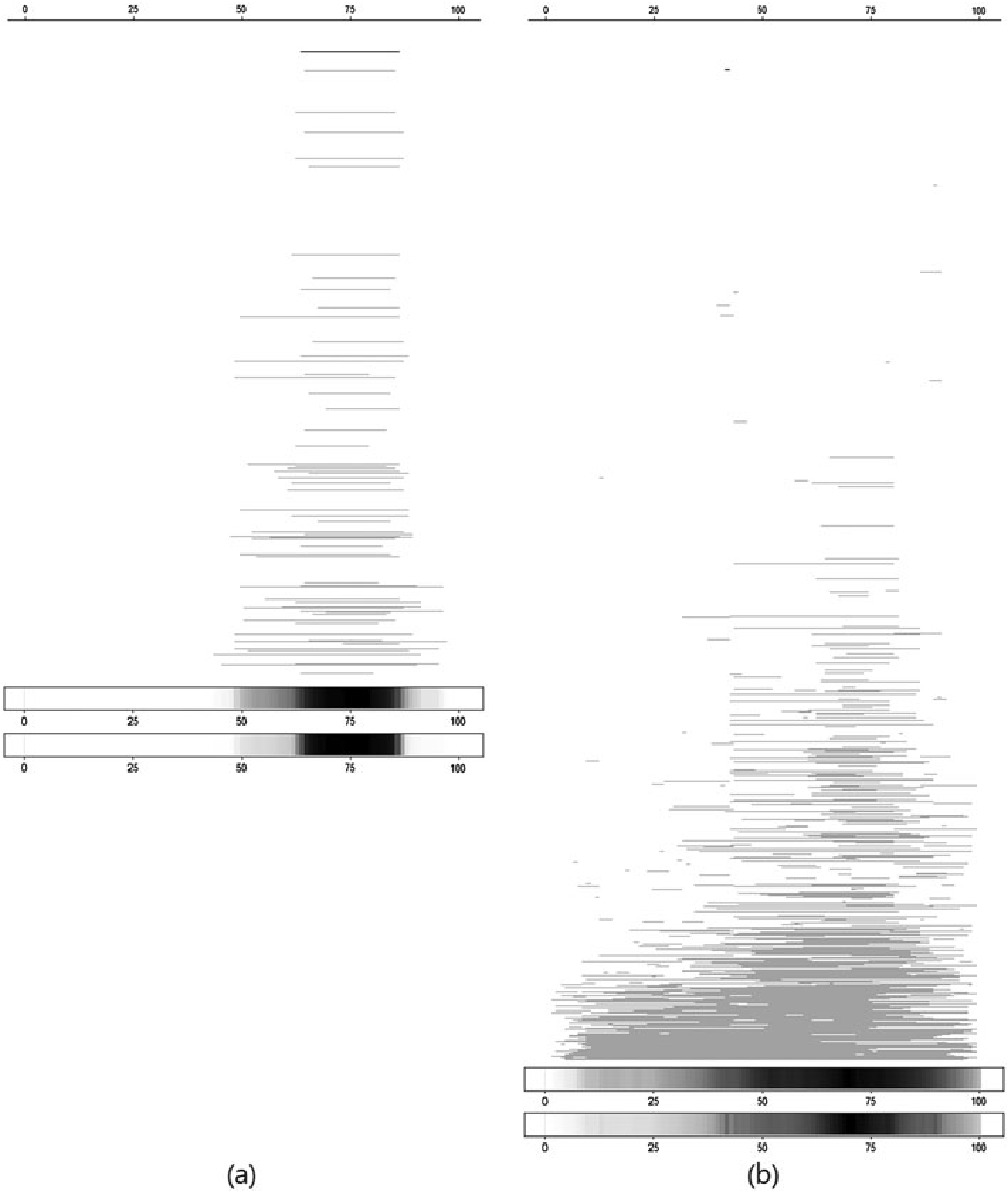


Figure 2. The 95% confidence sets $\Psi_{0.95}(C_0)$ from a simulated datasets, where $N = 100$, the true cluster is set to be $C_0 = \{i \mid |i - 75| \leqslant 10\}$ and SNR (θ/σ) is set to be 1 in (a) or 0.5 in (b), respectively. The set of candidate clusters $\mathcal{C}$ is predefined as (9) with $R_{\max} = 24$. The number of candidate clusters is $|\mathcal{C}| = 2500$. The size of confidence set is $|\Psi_{0.95}(C_0)| = 70$ in (a) and $|\Psi_{0.95}(C_0)| = 1613$ in (b), respectively.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.161

a cluster is expressed as a line segment between the two spatial units on the boundary. The two boundary points (or endpoints) can be viewed as two spatial change-points. For example, if the true cluster is set to be $C = \{i \mid |i-c| \leqslant r\}$ on the $1 \times N$ spatial lattice, then the $(\max\{1, c-r\})$th cell and the $(\min\{N, c+r\})$th cell can be considered as the first and second change-points, respectively.

Let $y_{a:b} = (y_a, y_{a+1}, \ldots, y_b)$ denote an ordered sequence of responses, where a and b are positive integers, and $1 \leqslant a \leqslant b \leqslant N$. Let τ_j denote the jth change-point, where $\tau_j \in \{1, \ldots, N-1\}$ for $j = 1, \ldots, m$, and m is the number of change-points. One approach to detecting multiple change-points is to minimize

$$\sum_{j=1}^{m+1} \left\{ G(y_{(\tau_{j-1}+1):\tau_j}) \right\} + m\xi, \tag{13}$$

where G is a cost function for a segment (e.g. negative log-likelihood), ξ is a penalty, and τ_0 and τ_{m+1} are set to be $\tau_0 = 0$ and $\tau_{m+1} = N$.

The BinSeg change-point analysis finds a single change-point (i.e. $m = 1$) that minimizes equation (13) (Scott & Knott, 1974; Sen & Srivastava, 1975). Next, the single change-point detection is repeated on each segment, which is before or after the change-point detected previously. In the BinSeg method, this procedure continues with new segments until there is not any more change-point detected. The PELT change-point analysis (Killick et al., 2012) minimizes equation (13) using dynamic programming (Bellman & Dreyfus, 1962) and finds the optimal $m + 1$ change-points based on the information obtained for m change-points. The PELT method achieves the computational efficiency using pruning. Both change-point analyses, BinSeg and PELT, are provided in the R package *changepoint* (Killick & Eckley, 2014).

In the following, we will compare our spatial cluster detection method with BinSeg and PELT using both simulated data and a real data example.

The results are illustrated in Figures 3 and 4.

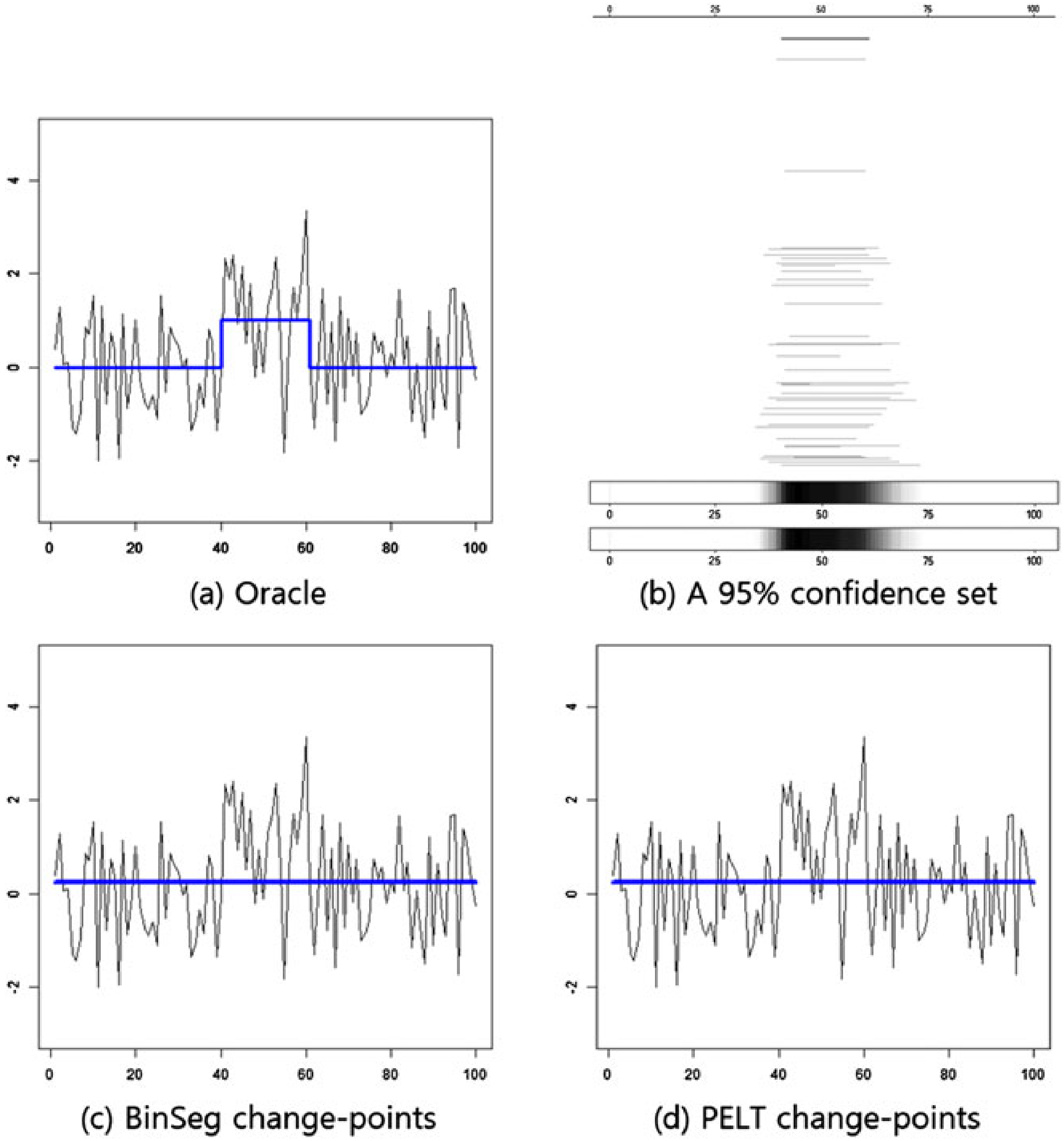
Figure 3a shows the plot of the first simulated dataset along with horizontal lines for two underlying means where the SNR is set to be 1 $(\theta/\sigma = 1)$. In Figure 3b, the 95% confidence set $\Psi_{0.95}(C_0)$ is visualized based on 1000 simulations. Plots of the simulated dataset along with horizontal lines for the fitted mean from the BinSeg and PELT change-point analyses are provided in Figure 3c and d. While both change-point analyses fail to find any change-point, our cluster detection method successfully estimates the cluster as $\hat{C} = \{i \mid |i - 51| \leqslant 10\}$, which is close to the true cluster. The p-value is 0.003 based on 1000 Monte Carlo simulations.

In Figure 4a, the second simulated dataset with SNR $= 2$ $(\theta/\sigma = 2)$ is plotted along with horizontal lines for two underlying means. The 95% confidence set $\Psi_{0.95}(C_0)$ based on 1000 simulations is illustrated in Figure 4b. The cluster estimate is $\hat{C} = \{i \mid |i - 51| \leqslant 11\}$ with p-value $= 0.001$ from 1000 Monte Carlo simulations. Plots of the simulated dataset along with horizontal lines for the fitted means from the BinSeg and PELT change-point analyses are provided in Figure 4c and d. Both change-point analyses successfully detect change-points for the second dataset simulated with the strong signal as SNR $= 2$. That is, if the signal is strong enough, the change-point analysis works as well as the spatial cluster detection does. However, by applying our method, we are able to not only provide the cluster estimate but also quantify and visualize the uncertainty associated with the cluster estimate.

4.2 Alaska boreal forest data

We illustrate the methodology developed in Section 3 by a dataset in forest ecology. The Alaska boreal forest, the largest forest component of the Alaska landscape, occupies a vast area from the wet Pacific coast to the dry inland

(wileyonlinelibrary.com) DOI: 10.1002/sta4.161



(a) Oracle

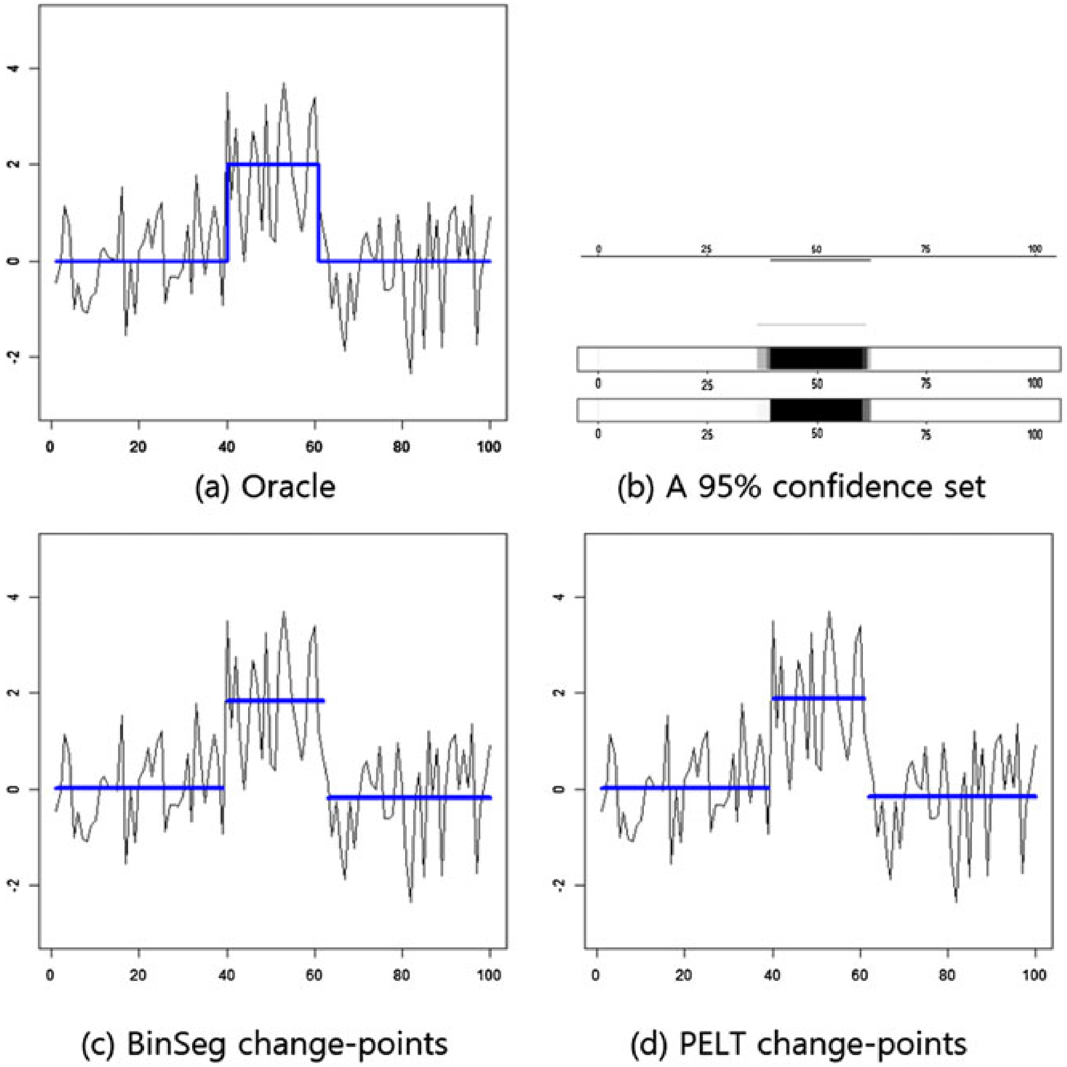
(b) A 95% confidence set

(c) BinSeg change-points

(d) PELT change-points

Figure 3. Plots of the simulated dataset along with horizontal lines for the underlying means in (a) and the fitted means in (c)–(d). A 95% confidence set for the cluster is in (b) with $\hat{C} = \{i \mid |i - 51| \leqslant 10\}$, $\hat{\mu} = -0.008$ and $\hat{\theta} = 1.206$. The true cluster is $C = \{i \mid |i - 50| \leqslant 10\}$, and $(\mu, \theta, \sigma^2) = (0, 1, 1)$.

near the Canadian border. This type of forest grows under the most severe climate conditions in the world, including an air temperature as low as -70°C and an annual precipitation that rarely exceeds 50 cm (Malone et al., 2009; Liang, 2010). The data consist of 1242 plot–year combination where permanent sample plots distributed widely across the interior, south-central and far-north Alaska from year 1994 to year 2017. The sample area stretches from about $-152°$ to $-146°$ in longitude and from about $60°$ to $68°$ in latitude. Of interest is stand basal area and its pattern across the latitude. Stand basal area (SBA) is defined as the total cross-sectional area of all live trees at breast height per unit area of land (m^2 ha^{-1}) and is commonly used in forest science as a key biotic factor of productivity and a good proxy for forest resource acquisition and competition (Liang et al., 2016).



(a) Oracle

(b) A 95% confidence set

(c) BinSeg change-points

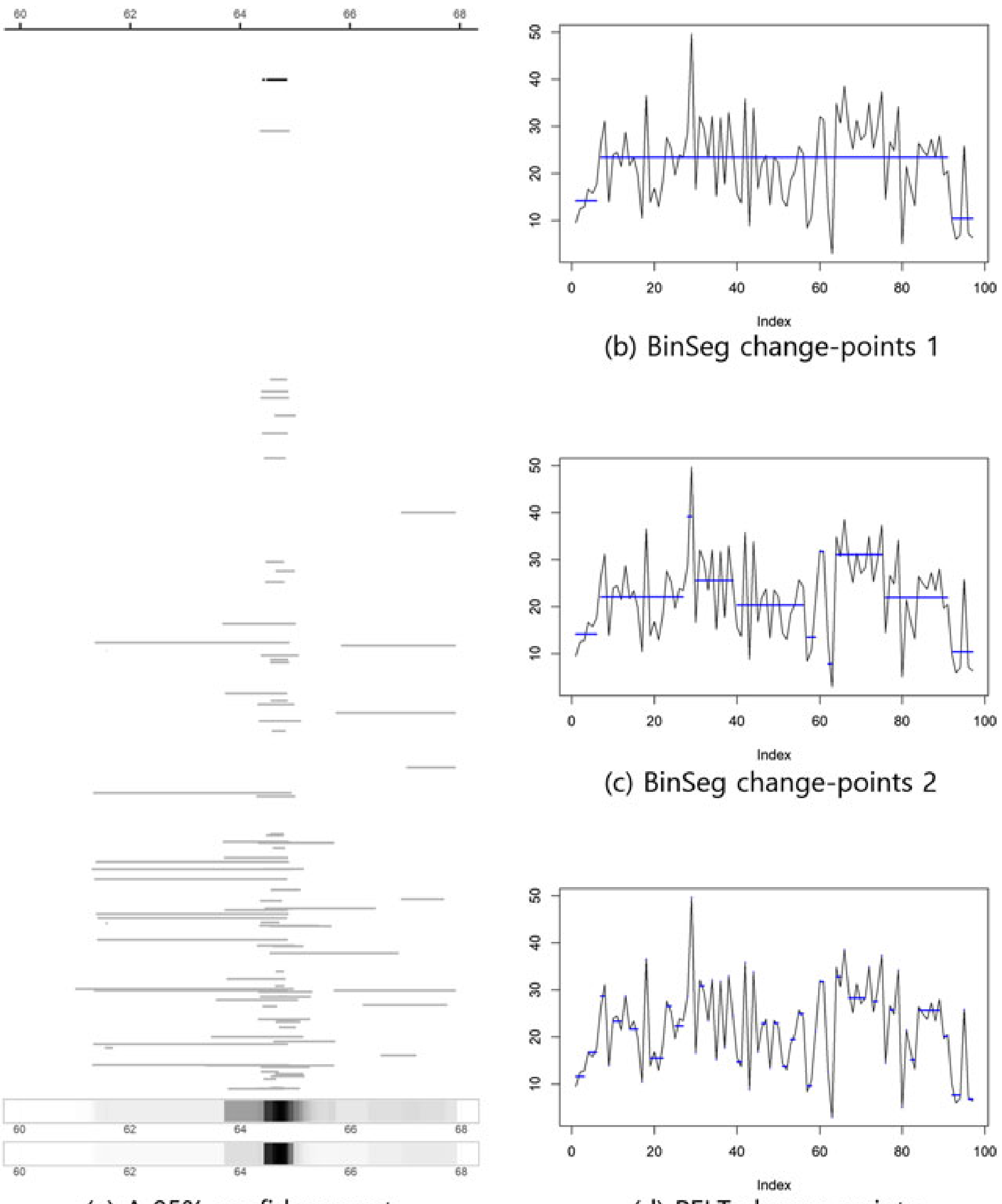
(d) PELT change-points

Figure 4. Plots of the simulated dataset along with horizontal lines for the underlying means in (a) and the fitted means in (c)–(d). A 95% confidence set for the cluster is in (b) with $\hat{C} = \{i \mid |i - 51| \leqslant 11\}$, $\hat{\mu} = -0.081$ and $\hat{\theta} = 1.909$. The true cluster is $C = \{i \mid |i - 50| \leqslant 10\}$, and $(\mu, \theta, \sigma^2) = (0, 2, 1)$.

We let y_i denote the SBA averaged across the longitude, where $i = 1, \ldots, 97$ for 97 distinct latitudes in the study area. We apply our spatial cluster detection approach and compare the results with the two change-points analyses, BinSeg and PELT. The results are illustrated in Figure 5.

Figure 5a shows a 95% confidence set $\Psi_{0.95}(C)$ based on 1000 simulations. The top horizontal axis represents the latitude from 60° to 68° north. Our cluster detection method estimates the cluster to be $\hat{C} = \{\text{latitude} \mid |\text{latitude} - 64.63| \leqslant 0.22\}$. The p-value is 0.029. The background mean SBA value is estimated to be $\hat{\mu} = 20.758$ m^{2}ha^{-1}, and the cluster effect is estimated to be $\hat{\theta} = 10.345$ m^{2}ha^{-1}. That is, the mean SBA value in this latitudinal cluster ($\hat{\mu} + \hat{\theta} = 31.103$ m^2 ha^{-1}) is much higher than the background is ($\hat{\mu} = 20.758$ m^2 ha^{-1}).

(wileyonlinelibrary.com) DOI: 10.1002/sta4.161



(b) BinSeg change-points 1

(c) BinSeg change-points 2

(a) A 95% confidence set

(d) PELT change-points

Figure 5. A 95% confidence set for the cluster is in (a) with $\hat{C} = \{\text{latitude} \mid |\text{latitude} - 64.63| \leqslant 0.22\}$, $\hat{\mu} = 20.758$ m^{2}ha^{-1}, $\hat{\theta} = 10.345$ m^{2}ha^{-1} and $\hat{\sigma} = 8.108$ m^{2}ha^{-1}. Plots of y_i's along with horizontal lines for the fitted means in (b)–(d).

This is plausible, because the latitudinal cluster window (from 64.41° to 64.85°) includes Nenana Ridge (Bonanza Creek area), which is some of the most productive forest land in interior Alaska with large trees and full stocking.

Plots of the change in the mean SBA along with horizontal lines for the fitted means from the BinSeg and PELT are provided in Figure 5b–d. Figure 5b is from the BinSeg change-point detection when the number of the change-points

m is set to be 2, while Figure 5c is from when the number of the change-points m is set to be at most 10. Figure 5d is from the PELT change-point detection where it is not an option to set the number of change-points.

The latitudinal cluster estimate $\hat{C}$ is from the 64th spatial unit (64.54°) to the 75th spatial unit (64.85°). Thus, it can be seen that our method in Figure 5a and the BinSeg method in Figure 5b provide different results from each other. In Figure 5c, we could find that the 8th and 9th change-points provide a similar result to the latitudinal cluster estimate $\hat{C}$. However, the BinSeg method in Figure 5c and the PELT method in Figure 5d seem to overfit the number of clusters.

5 Conclusions and discussion

We have developed in this paper a new methodology to quantify the uncertainty of a detected spatial cluster. We have defined a confidence set for the true spatial cluster based on a likelihood-based approach. We have also proposed a way to visualize the confidence set for the 1D case. Empirical distributions, with different cluster settings, support the pivotal property of the null distribution, which enable us to define a confidence set. The empirical coverage rate for the true cluster also suggests that the confidence set is well suited for relatively strong cluster effect. Further, visualization of the confidence set allows us to see the number of clusters in the confidence set as well as the locations of those spatial clusters. Insight can also be gained into each spatial unit in terms of its chance to belong to the true cluster.

Both the simulation study and a real data example demonstrate that our spatial approach could provide reasonable cluster estimation as well as the quantification and visualization of the uncertainty associated with the detected cluster. Our spatial cluster detection approach is more natural than are the change-point analyses to extend from the 1D space to the two-dimensional (2D) cases.

For the 2D spatial domain, a $(1-\alpha)$ confidence set for C_0 can be defined as (10) in Section 3.2 along with a set of candidate clusters $\mathcal{C}$, $\Psi_{1-\alpha}(C_0) = \left\{C \in \mathcal{C} \;\middle|\; -(2/N)\log\Lambda(C) \leqslant \mathrm{P}_{100\times(1-\alpha)}(\hat{C})\right\}$, where $\hat{C}$ is the cluster estimate in the 2D space. However, $\Psi_{1-\alpha}(C_0)$ in the 2D space is more challenging to visualize than that in the 1D space case is. Interactive maps or 3D plots may be considered to illustrate the confidence set for the spatial cluster in the 2D space. The verification of the pivotal property of the null distribution as well as the evaluation of the coverage rate for the true cluster needs to be conducted for the 2D space. We leave this for future research.

Acknowledgements

Funding has been provided by a USDA Cooperative State Research, Education and Extension Service (CSREES) McIntire-Stennis project. We thank the School of Natural Resources and Agricultural Sciences, University of Alaska Fairbanks, for the development and maintenance of the Cooperative Alaska Forest Inventory, and thank the Global Forest Biodiversity Initiative (GFBI) for establishing data-sharing protocols and standards for international collaborative research projects. We also thank Thomas Malone for his insights on Alaska boreal forest.

References

Assunção, R, Costa, M, Tavares, A & Ferreira, S (2006), 'Fast detection of arbitrarily shaped disease clusters', *Statistics in Medicine*, **25**, 723–742.

Bellman, RE & Dreyfus, SE (1962), *Applied Dynamic Programming*, *Princeton University Press*, Princeton, NJ.

Clark, AB & Lawson, AB (2002), *Spatio-temporal cluster modelling of small area health data*, Spatial Cluster Modelling in Lawson, AB & Denison, D (eds), *Chapman and Hall/CRC*, Boca Raton, FL, 235–258.

Duczmal, L & Assunção, R (2004), 'A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters', *Computational Statistics and Data Analysis*, **45**, 269–284.

Gangnon, RE (2010a), 'Local multiplicity adjustments for spatial cluster detection', *Environmental and Ecological Statistics*, **17**(1), 55–71.

Gangnon, RE (2010b), 'A model for space–time cluster detection using spatial clusters with flexible temporal risk patterns', *Statistics in Medicine*, **29**, 2325–2337.

Gangnon, RE & Clayton, MK (2000), 'Bayesian detection and modeling of spatial disease clustering', *Biometrics*, **56**, 922–935.

Gangnon, RE & Clayton, MK (2003), 'A hierarchical model for spatially clustered disease rates', *Statistics in Medicine*, **22**, 3213–3228.

Gangnon, RE & Clayton, MK (2004), 'Likelihood-based tests for detecting spatial clustering of disease', *Environmetrics*, **15**, 797–810.

Gangnon, RE & Clayton, MK (2007), 'Cluster detection using Bayes factors from overparameterized cluster models', *Environmental and Ecological Statistics*, **14**, 69–82.

Killick, R, Fearnhead, P & Eckley, IA (2012), 'Optimal detection of changepoints with a linear computational cost', *Journal of the American Statistical Association*, **107**(500), 1590–1598.

Killick, R & Eckley, IA (2014), 'changepoint: an R package for changepoint analysis', *Journal of Statistical Software*, **58**(1), 1–19.

Kulldorff, M (1997), 'A spatial scan statistic', *Communications in Statistics, Part A*, **26**, 1481–1496.

Kulldorff, M (2001), 'Prospective time-periodic geographical disease surveillance using a scan statistic', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **164**, 61–72.

Kulldorff, M, Athas, W, Feuer, E, Miller, B & Key, C (1998), 'Evaluating cluster alarms: a space–time scan statistic and brain cancer in Los Alamos, New Mexico', *American Journal of Public Health*, **88**, 1377–1380.

Kulldorff, M, Huang, L & Konty, K (2009), 'A scan statistic for continuous data based on the normal probability model', *International Journal of Health Geographics*, **8**, 58.

Kulldorff, M, Huang, L, Pickle, L & Duczmal, L (2006), 'An elliptic spatial scan statistic', *Statistics in Medicine*, **25**, 3929–3943.

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

Lawson, AB (2000), 'Cluster modelling of disease incidence via RJMCMC methods: a comparative evaluation', *Statistics in Medicine*, **19**, 2361–2375.

Lee, J, Gangnon, RE & Zhu, J (2017), 'Cluster detection of spatial regression coefficients', *Statistics in Medicine*, **36**(7), 1118–1133.

Lehmann, EL (1986), *Testing Statistical Hypotheses*, *Wiley*, New York.

Liang, J (2010), 'Dynamics and management of Alaska boreal forest: an all-aged multi-species matrix growth model', *Forest Ecology and Management*, **260**(4), 491–501.

Liang, J, Crowther, TW, Picard, N, Wiser, S, Zhou, M, Alberti, G, Schulze, E-D, McGuire, AD, Bozzato, F, Pretzsch, H, de-Miguel, S, Paquette, A, Hérault, B, Scherer-Lorenzen, M, Barrett, CB, Glick, HB, Hengeveld, GM, Nabuurs, GJ, Pfautsch, S, Viana, H, Vibrans, AC, Ammer, C, Schall, P, Verbyla, D, Tchebakova, N, Fischer, M, Watson, JV, Chen, HY, Lei, X, Schelhaas, MJ, Lu, H, Gianelle, D, Parfenova, EI, Salas, C, Lee, E, Lee, B, Kim, HS, Bruelheide, H, Coomes, DA, Piotto, D, Sunderland, T, Schmid, B, Gourlet-Fleury, S, Sonké B, Tavani, R, Zhu, J, Brandl, S, Vayreda, J, Kitahara, F, Searle, EB, Neldner, VJ, Ngugi, MR, Baraloto, C, Frizzera, L, Bałazy, R, Oleksyn, J, Zawiła-Niedźwiecki T, Bouriaud, O, Bussotti, F, Finér L, Jaroszewicz, B, Jucker, T, Valladares, F, Jagodzinski, AM, Peri, PL, Gonmadje, C, Marthy, W, O'Brien, T, Martin, EH, Marshall, AR, Rovero, F, Bitariho, R, Niklaus, PA, Alvarez-Loayza, P, Chamuya, N, Valencia, R, Mortier, F, Wortel, V, Engone-Obiang, NL, Ferreira, LV, Odeke, DE, Vasquez, RM, Lewis, SL & Reich, PB (2016), 'Positive biodiversity–productivity relationship predominant in global forests', *Science*, **354**(6309).

Lin, PS, Kung, YH & Clayton, MK (2016), 'Spatial scan statistics for detection of multiple clusters with arbitrary shapes', *Biometrics*, **72**(4), 1226–1234.

Malone, T, Liang, J & Packee, EC (2009), *Cooperative Alaska forest inventory*, Technical Report PNW-GTR-785, *USDA Forest Service, Pacific Northwest Research Station*, Portland, OR.

Neill, DB (2012), 'Fast subset scan for spatial pattern detection', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**(2), 337–360.

Scott, AJ & Knott, M (1974), 'A cluster analysis method for grouping means in the analysis of variance', *Biometrics*, **30**(3), 507–512.

Sen, A & Srivastava, MS (1975), 'On tests for detecting change in mean', *The Annals of Statistics*, **3**(1), 98–108.

Shu, L, Jiang, W & Tsui, KL (2012), 'A standardized scan statistic for detecting spatial clusters with estimated parameters', *Naval Research Logistics*, **59**, 397–410.

Takahashi, K, Kulldorff, M, Tango, T & Yih, K (2008), 'A flexibly shaped space–time scan statistic for disease outbreak detection and monitoring', *International Journal of Health Geographics*, **7**, 14.

Tango, T & Takahashi, K (2005), 'A flexibly shaped spatial scan statistic for detecting clusters', *International Journal of Health Geographics*, **4**, 11.

Wakefield, J & Kim, A (2013), 'A Bayesian model for cluster detection', *Biostatistics*, **14**(4), 752–765.

Xu, J & Gangnon, RE (2016), 'Stepwise and stagewise approaches for spatial cluster detection', *Spatial and Spatiotemporal Epidemiology*, **17**, 59–74.

Yan, P & Clayton, MK (2006), 'A cluster model for space–time disease counts', *Statistics in Medicine*, **25**, 867–881.