Spatial scan statistics with a restricted likelihood ratio for ordinal outcome data

Myeonggyun Lee

The Graduate School

Yonsei University

Department of Biostatistics and Computing

Spatial scan statistics with a restricted likelihood ratio for ordinal outcome data

A Master's Thesis

Submitted to the Department of Biostatistics and Computing

and the Graduate School of Yonsei University

in partial fulfillment of the

requirements for the degree of

Master of Science

Myeonggyun Lee

December 2016

This certifies that the Master's thesis of Myeonggyun Lee is approved.

_____

Thesis Supervisor. Prof. Inkyung Jung

_____

Prof. Chung Mo Nam

_____

Prof. Sohee Park

The Graduate School

Yonsei University

December 2016

Acknowledgement

지난 2년간의 긴 여정을 마무리하는 석사학위논문을 제출하게 되어 감회가 새롭습니다. 제가 하고 싶은 공부를 할 수 있었던 의학전산통계학 협동과정 석사 기간을 한 단어로 함축하면 '행복' 인 것 같습니다. 2년간 이 행복을 계속 누릴 수 있게 도움을 주시고 힘이 되어주신 많은 분들께 감사의 말씀을 전하고자 합니다.

먼저 부족한 저의 학위논문을 지도해주시고 이끌어주신 정인경 교수님께 깊은 감사를 올립니다. 저의 학문적 성장과 다양한 경험은 모두 교수님의 격려와 배려 덕분에 이룰 수 있었던 결실이라 생각합니다. 또한 항상 저를 아껴주시고 위해주셨던 남정모 교수님, 감사합니다. 교수님의 은혜 덕분에 제가 좀 더 넓은 세상을 바라볼 수 있고 시야를 넓힐 수 있었습니다. 그리고 논문지도와 더불어 많은 경험을 알려주시며 제가 가고자 하는 길에 용기를 복 돋아 주신 박소희 교수님과 다방면의 지식과 세심함을 가르쳐 주신 송기준 교수님께 감사의 말씀 드립니다. 아울러 지금의 저를 있게끔 초석을 다져주신 이주호 교수님, 이민구 교수님 감사합니다.

석사과정 동안 함께 지낸 선배, 동기, 후배들에게도 감사의 마음을 전합니다. 자주 뵐 순 없었지만 만날 때마다 밝은 모습으로 맞이해 주셨던 박사과정 선배님들께 감사드립니다. 의학정보통계학과를 행복한 추억으로 남을 수 있게 도와준 지유, 경민, 해린, 근우, 그리고 많은 시간을 같이 지내진 못했지만 함께 석사과정을 보낸 지윤, 지현, 하영, 선민, 하나, 종수, 호재에게도 고맙습니다.

비록 잠깐 동안 머물렀지만 1년간 공대생으로써 많은 경험을 함께하고 저의 결정을 누구보다 응원해준 멋진 인연 국방 석진이형, 장군 상순이형, 에이스 동하형, 뉴요커 재현이형, 정우, 고3 수험생활부터 지금까지 만날 때마다 사건사고를 만드는 평생지기 태경, 민기, 안수, 영훈 그리고 구미고 3학년 6반 동창친구들에게도 감사합니다. 학부생활을 하면서 20대 청춘의 절반을 함께 지내온 선규형, 근태형, 주성이형, 승현, 시형, 영훈, 명언, 재연, 문수, 동훈, 승범, 성규에게도 정말 고맙습니다.

마지막으로 부족한 저를 위해 매일 기도하시고 올바른 길로 자랄 수 있도록 키워주신 아버지, 어머니, 그리고 동생이자 가장 오랜 나의 친구 승균이에게 감사하고 사랑합니다. 그간 받은 은혜에 조금이나마 갚을 수 있도록 더욱더 자랑스러운 맏이가 되는 모습 보여드리겠습니다. 감사합니다.

2016년 12월

이명균 올림

Contents

List of Figures

List of Tables

Abstract

Spatial scan statistics proposed by Kulldorff are widely used as a technique to detect geographical disease clusters for different types of data such as Bernoulli, Poisson, ordinal, normal, and survival. The spatial scan statistic for ordinal data can be used to detect clusters indicating areas with high rates of more serious stages compared with the surrounding areas.

However, it has been pointed out that the Poisson-based spatial scan statistic tends to detect the most likely cluster much larger than the true cluster by absorbing insignificant neighbors with non-elevated risk. We suspect that the spatial scan statistic for ordinal data might also have the similar undesirable phenomena. Tango (2008) proposed to modify the spatial scan statistic using a restricted likelihood ratio for scanning only the regions with elevated risk. The method worked well for preventing over-detection but was evaluated only in the Poisson model.

In this paper, we propose to apply a restricted likelihood ratio into two spatial scan statistics to circumvent such a phenomenon in ordinal outcome data. Through a simulation study we compare the performance of the proposed methods with original spatial scan statistics. We calculate sensitivity, positive predicted

value (PPV), usual power and bivariate power distribution as performance measures.

The simulation study results show that the proposed spatial scan statistics with a restricted likelihood ratio have a reasonable or better performance compared with original ones. The original methods for ordinal data tend to detect larger clusters than the true cluster, and our approach seems to reduce the undesirable property. We illustrate the proposed methods using a real data set of the 2014 Health Screening Program of Korea with the diagnosis results of normal, caution, suspected disease, and diagnosed with disease as an ordinal outcome.

Keywords: Spatial scan statistics; ordinal data; cluster detection; restricted likelihood ratio

1. Introduction

For geographical surveillance, spatial scan statistics are widely used to detect spatial disease clustering in different types of data such as Bernoulli (Kulldorff and Nagarwalla, 1995), Poisson (Kulldorff, 1997), ordinal (Jung *et al.*, 2007), survival (Huang *et al.*, 2007), normal (Kulldorff *et al.*, 2009), and multinomial (Jung *et al.*, 2010). The method proposed by Kulldorff (1997) is based on the likelihood ratio test to find the area with the maximum value of test statistics as a most likely cluster. The spatial scan statistics as a cluster detection tool have been applied in various fields in order to identify geographical patterns with high or low rates for a range of diseases such as the study for birth defects (Ozdenerol *et al.*, 2005), detecting high-risk areas for leprosy in Bangladesh (Fischer *et al.*, 2008), as well as identifying spatial cluster for cancer incidence, prevalence, and mortality (Michelozzi *et al.*, 2002; Gregorio *et al.*, 2006; Alvares *et al.*, 2009; Amin *et al.*, 2014).

Even though the spatial scan statistics are commonly utilized, it has been known that this approach detects a much larger cluster within insignificant regions. Through a simulation study Tango (2007) pointed out that the Poisson-based spatial scan statistic proposed by Kulldorff (1997) tends to detect an unrealistically larger cluster than the expected true cluster by absorbing adjacent

regions with irrelevant risks. Furthermore, in the case of detecting irregular shaped clusters, the spatial scan statistic using the circular scanning window has difficulty and thus the over-detection phenomenon can occur (Tango and Takahashi, 2005). To resolve the undesirable trend, Tango (2008) proposed a Poisson-based spatial scan statistic by modifying the likelihood ratio. The Monte Carlo simulation study showed that the proposed spatial scan statistic worked well for preventing such undesirable phenomena in detecting the true cluster compared with the original spatial scan statistic.

In this paper, we focus on spatial cluster detection for ordinal data. In the medical field, ordinal scaled data are often obtained in nature such as cancer stage or grade. At this time, we are interested in geographical cluster detection of high rates of more severe categories (e.g., later stage or higher grade). There are two spatial scan statistics for ordinal outcome data. One spatial scan statistic was proposed by Jung *et al.* (2007). They assumed the alternative hypothesis based on likelihood ratio ordering (LRO) and thus showed that their method had good performances to detect spatial clusters for ordinal data. However, this approach has somewhat restricted probabilities, which are higher when the disease categories are more serious. Therefore, Jung and Lee (2011) developed another spatial scan statistic to alleviate order restriction based on stochastic ordering

(STO) as an alternative hypothesis. Thus, in the case that the true area (hot spot) has a stochastic-based hypothesis, Jung and Lee (2011) showed that the STO-based method performs better than the LRO-based method. But, we suspect that two proposed spatial scan statistics for ordinal data also tend to find a larger cluster than the true cluster.

The purpose of this study is to propose two modified spatial scan statistics using a restricted likelihood ratio to circumvent the over-detection problem in ordinal outcome data. In our simulation study, we assume both irregular and circular shaped true clusters and compute the performances, including sensitivity, PPV, usual power, and bivariate power distribution in order to evaluate the proposed spatial scan statistics compared with original spatial scan statistics. In chapter 2, we briefly review two spatial scan statistics for ordinal data and propose spatial scan statistics with a restricted likelihood ratio. In chapter 3, we conduct a Monte Carlo simulation study and evaluate the performance of the proposed approach to compare with the original methods. We illustrate the application to real data examples in chapter 4 and present discussion and conclusion of our study in chapter 5.

2. Methods

The spatial scan statistic is based on the likelihood ratio test. We first construct a large number of scanning windows of different sizes at each region on the whole study area as a candidate cluster. We compute the likelihood ratio test statistic for each candidate and the scanning window with the maximum value of the likelihood ratio test statistics is defined as the most likely cluster. Circular and elliptical shaped scanning windows are mostly used, as well as a flexible shape. . In this section, we simply review two approaches and propose spatial scan statistics with a restricted likelihood ratio to avoid over-detection phenomena. And then, we explain how to conduct statistical inference.

2.1 Likelihood ratio ordering-based approach

Suppose that a study area is composed of I sub-regions and the ordinal outcome variable has K categories. Let c_{ik} be the number of cases in the *i-th* region and the *k-th* category, where $i = 1,\ldots, I$ and $k = 1,\ldots, K$. Since the categories are ordinal scale in nature, for example, a larger k reflects a more severe disease stage. The null hypothesis that there is no clustering can be expressed as $H_0 : p_k = q_k$, for all k = 1,…, K and all scanning windows z. In

other words, the probability of being in category *k* within the scanning window is equal to the probability of being in category *k* outside the scanning window. Note that $\sum_k p_k = 1$ and $\sum_k q_k = 1$. By Jung *et al.* (2007), an alternative hypothesis was considered as follows

$$H_a : \frac{p_1}{q_1} \leq \frac{p_2}{q_2} \leq \cdots \leq \frac{p_K}{q_K}, \quad \text{for some } z \tag{1}$$

with at least one inequality being strict. This type of order restriction is called by likelihood ratio ordering (LRO) according to Dykstra *et al.* (1995).

For the ordinal model, the likelihood function is written as

$$\mathrm{L}(\mathrm{Z}, p_1, \ldots, p_K, q_1, \ldots, q_K) = \prod_k \left(\prod_{i \in Z} p_k^{c_{ik}} \prod_{i \notin Z} q_k^{c_{ik}} \right), \tag{2}$$

where p_k is the unknown probability that an observation within the scanning window *z* belongs to category *k* and q_k is also the unknown probability that an observation outside the scanning window *z* belongs to category *k*. The likelihood ratio test statistic can be expressed as

$$\lambda = \frac{\max\limits_{Z,H_a} \mathrm{L}(\mathrm{Z}, p_1, \ldots, p_K, q_1, \ldots, q_K)}{\max\limits_{Z,H_0} \mathrm{L}(\mathrm{Z}, p_1, \ldots, p_K, q_1, \ldots, q_K)} = \frac{\max\limits_{Z} L(Z)}{L_0}, \tag{3}$$

with

$$L_0 = \prod_k \prod_i \hat{p}_{0_k}^{c_{ik}} = \prod_k \left(\frac{C_k}{C}\right)^{\sum_i c_{ik}} = \prod_k \left(\frac{C_k}{C}\right)^{C_k}, \qquad (4)$$

where C_k $(= \sum_i c_{ik})$ is the sum of observation in category k, C $(= \sum_k \sum_i c_{ik})$ is the total number of observations in the whole study area and $\hat{p}_{0_k}$ $(= \hat{q}_{0_k}) = C_k/C$ is MLE of p_k $(= q_k)$ under the null hypothesis, and with

$$L(Z) = \prod_k \prod_{i \in Z} \hat{p}_k^{c_{ik}} \prod_{i \notin Z} \hat{q}_k^{c_{ik}}, \qquad (5)$$

where $\hat{p}_k$ and $\hat{q}_k$ are the MLEs of p_k and q_k under the alternative hypothesis (1). For $\hat{p}_k$ and $\hat{q}_k$, Dykstra *et al.* (1995) proved the mathematical expressions, and both can be explicitly calculated using the 'Pool-Adjacent-Violators' algorithm as described by Barlow *et al.* (1972). Jung *et al.* (2007) explained the details of how to obtain the MLEs under the LRO-based alternative hypothesis.

According to Jung and Lee (2011), although an alternative hypothesis of LRO (1) surely ensures that clusters are detected when an area has more serious disease stages than the adjacent area, it does not incorporate all situations in which the probabilities of more severe disease categories are higher. For instance, with four disease categories, an area which has the probabilities of 0.15, 0.15, 0.45, and 0.25 seems to have high rates of a worse disease outcome compared with an area with probabilities of 0.25 for all four categories. However, the spatial scan statistic

based on the LRO approach tends to fail to detect such an area as a cluster.

2.2 Stochastic ordering-based approach

Jung and Lee (2011) proposed an alternative hypothesis to the LRO hypothesis (1). The considered alternative hypothesis is

$$H_a : \sum_{k=1}^{j} p_k \leq \sum_{k=1}^{j} q_k \text{ , for all j = 1, …, K and some } z \qquad (6)$$

with at least one strict inequality. This order restriction is called stochastic ordering (STO) by Robertson and Wright (1981). Compared with the LRO hypothesis (1), the STO hypothesis can include more general situations in which the higher the rate, the more severe the disease categories. The LRO hypothesis is a special case of the STO hypothesis.

Even though the likelihood ratio test statistic is the same as equation (3), the MLEs of p_k and q_k are attained under the alternative hypothesis (6) and thus they are used to calculate the value of the test statistic. Note that $\tilde{p}_k$ and $\tilde{q}_k$ are the MLEs of p_k and q_k under the STO hypothesis, respectively, c $(= \sum_k \sum_{i \in Z} c_{ik})$ is the total number of cases inside the scanning window *z*, and the total number of cases in the whole study area is C $(= \sum_k \sum_i c_{ik})$. According to

Robertson and Wright (1981), $\tilde{p}_k$ and $\tilde{q}_k$ can be obtained as

$$(\tilde{\mathbf{p}},\tilde{\mathbf{q}})=\mathbf{w}\mathrm{E}_{\mathbf{w}}(\mathbf{h}|\mathbf{B}), \qquad (7)$$

where $\mathbf{w} = (c\bar{p}_1, c\bar{p}_2, \dots, c\bar{p}_K, (C-c)\bar{q}_1, (C-c)\bar{q}_2, \dots, (C-c)\bar{q}_K)$ with $\bar{p}_k = \sum_{i\in Z} c_{ik}/c$ and $\bar{q}_k = \sum_{i\notin Z} c_{ik}/(C-c)$ for $k = 1, \dots, K$,

$$h_k = \begin{cases} C^{-1} + \dfrac{C-c}{cC}\dfrac{\bar{q}_k}{\bar{p}_k}, & k = 1,2,\dots,K, \\ C^{-1} + \dfrac{c}{(C-c)C}\dfrac{\bar{q}_k - K}{\bar{p}_k - K}, & k = K+1, K+2, \dots, 2K \end{cases}$$

and

$$\mathbf{B} = \{\mathbf{x} \in R^{2K}; x_1 \geq x_2 \geq \cdots \geq x_K, x_{K+1} \leq x_{K+2} \leq \cdots \leq x_{2K}\}.$$

In equation (7) $\mathrm{E}_{\mathbf{w}}(\mathbf{h}|\mathbf{B})$ denotes the weighted least square projection of $\mathbf{h}$ onto $\mathbf{B}$. For each category k, both $\bar{p}_k$ and $\bar{q}_k$ are assumed to have positive values and, in practice, it could happen that $\bar{p}_k = 0$ or $\bar{q}_k = 0$ for some k. Dykstra *et al.* (1996) discussed that those coordinates can be set as a very small positive number.

The most likely cluster can be detected from the maximum value of the logarithm of the likelihood ratio test statistic (3) based on the STO and it can be expressed as

$$\log\lambda = \max_{Z}\sum_{k}\left\{\sum_{i\in Z} c_{ik}log\tilde{p}_k + \sum_{i\notin Z} c_{ik}log\tilde{q}_k\right\} - L_0. \qquad (8)$$

2.3 Restricted likelihood ratio test statistics

To circumvent or rescale the over-detected phenomenon, Tango (2008) proposed a Poisson-based restricted likelihood ratio test statistic by taking each individual region's risk into account. The proposed scan statistic detects only the regions with elevated risk by modifying the likelihood ratio. Individual region's risk is obtained from the *p*-value of the statistical test under the assumption of Poisson distribution (Tango, 2008).

We apply the restricted likelihood ratio test to two scan statistics for ordinal outcome data. The concept of the restricted likelihood ratio is to use the indicator function on the significance of each region as a screening criterion. For instance, given the pre-specified significance level (α_1) for the individual region, the restricted likelihood scan statistic for LRO is considered as

$$\lambda_{re} = \max_{Z}\left(\prod_{k}\prod_{i\in Z}\hat{p}_k^{c_{ik}}\prod_{i\notin Z}\hat{q}_k^{c_{ik}}\right)\left(\prod_{k}\left(\frac{C_k}{C}\right)^{C_k}\right)^{-1}\prod_{i\in Z} I(p\text{-}value_i < \alpha_1), \qquad (9)$$

where the $p\text{-}value_i$ is the *p*-value of the Pearson chi-square test for $H_{i0}: \mathbf{p_i} =$

$\mathbf{p_0}$. An STO-based scan statistic with a restricted likelihood ratio can be considered using $\tilde{p}_k$ and $\tilde{q}_k$ instead of $\hat{p}_k$ and $\hat{q}_k$. Introducing $I(p\text{-}value_i < \alpha_1)$ as a screening criterion for the ordinal scan statistics does not mean that we are performing multiple hypothesis tests.

For the ordinal outcome, we calculate the $p\text{-}value_i$ from the Pearson chi-square (χ^2) test to compare the proportions of cases in each response category at each region with the whole study area. The Pearson chi-square test is

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}, \qquad (10)$$

which has asymptotically a chi-square distribution with (*K*-1) degrees of freedom. Under the null hypothesis, the expected frequencies are found by multiplying each region size (c_i) by the proportions specified in the whole study area $(\hat{p}_{10}, \hat{p}_{20}, \ldots, \hat{p}_{K0})$. If the screening level of α_1 is equal to 1, the proposed spatial scan statistic is equivalent to the original scan statistic. Even though χ^2 cannot completely reflect the ordinal scale, it is possible to distinguish the distinct regions compared with the whole area and thus the ordinal scan statistic is conclusively able to detect clusters.

2.4 Statistical inference

When the most likely cluster having the maximum value of the likelihood ratio test statistics is detected, we need to conduct a statistical test. Since it is hard to find an asymptotic distribution of the test statistic, we can evaluate statistical significance by using Monte Carlo hypothesis testing (Dwass, 1957). First, we generate a large number of random data sets under the null hypothesis. Then, the maximum value of the test statistic is calculated for each data set. The upper 5% of the calculated maximum test statistics can be a critical value under a significance level of 0.05. We are able to assess the significance of the test statistic for the most likely cluster based on the critical value. In order to calculate Monte Carlo based p-values, the p-value can be expressed by $p = R/(\#sim + 1)$ where R is the rank of the test statistic from the original data set and #sim is the number of generated data sets. Generally, the p-value can be calculated with 99, 999, and 9,999 replications as the number of random data sets under the null hypothesis.

3. Simulation study

3.1 Simulation data and setting

In order to compare the performances between the original and the proposed spatial scan statistics in both the LRO and the STO hypotheses, we performed a simulation study under several scenarios. The area of Seoul in South Korea is considered as an entire study area, which consists of 25 districts at the “Si-gun-gu” level (city-county-district). All districts are geographically represented by a centroid coordinate. We assumed six different true cluster models with 3, 5, and 7 districts in a circular or irregular shaped true cluster. According to the number of districts in the true cluster, we set 140, 280, and 440 cases in the true cluster and 1200, 1400, and 1600 cases in the whole study area. Table 1 and Figure 1, respectively, provide the details of the six true cluster models.

Table 1. Detailed information for simulated cluster models A-F.

Cluster model	Number of districts in true cluster	Number of cases in true cluster	Total number of cases
A	3	140	1200
B	3	140	1200
C	5	280	1400
D	5	280	1400
E	7	440	1600
F	7	440	1600

Cluster model A (circular)

Cluster model B (irregular)

Cluster model C (circular)

Cluster model D (irregular)

Cluster model E (circular)

Cluster model F (irregular)

Figure 1. True cluster models A-F in the whole study area of Seoul.

While we assumed four disease categories, we considered three different null hypotheses and four alternative hypotheses which have the LRO and the STO hypotheses in each null hypothesis in equal parts. In other words, for all true cluster models, we assumed two different alternative LRO hypotheses and two different alternative STO hypotheses against each of three null hypotheses (see Table 2). Assumed null hypotheses are based on the general situation in ordinal level diseases. Although Jung and Lee (2011) compared two spatial scan statistics for ordinal data based only on the STO hypothesis, we evaluated the performances based on both the STO and the LRO hypotheses.

Table 2. The scenario details of assumed hypotheses.

Null hypothesis	Alternative hypothesis
$H_0 : p = q = (0.25,0.25,0.25,0.25)$	H_{1a} : p $= (0.20, 0.10, 0.40, 0.30)$ H_{1b} : p $= (0.15, 0.15, 0.45, 0.25)$ H_{1c} : p $= (0.10, 0.20, 0.30, 0.40)$ H_{1d} : p $= (0.05, 0.25, 0.25, 0.45)$
$H_0 : p = q = (0.30,0.20,0.30,0.20)$	H_{1a} : p $= (0.25, 0.05, 0.45, 0.25)$ H_{1b} : p $= (0.25, 0.05, 0.50, 020)$ H_{1c} : p $= (0.15, 0.20, 0.30, 0.35)$ H_{1d} : p $= (0.10, 0.15, 0.25, 0.50)$
$H_0 : p = q = (0.40,0.20,0.20,0.20)$	H_{1a} : p $= (0.35, 0.05, 0.35, 0.25)$ H_{1b} : p $= (0.35, 0.05, 0.40, 0.20)$ H_{1c} : p $= (0.25, 0.15, 0.25, 0.35)$ H_{1d} : p $= (0.10, 0.10, 0.40, 0.40)$

H_{1a}, H_{1b}: STO-based alternative hypotheses and H_{1c}, H_{1d}: LRO-based alternative hypotheses

First, we generated 10,000 random data sets under each null hypothesis to estimate the critical values at a significance level of 0.05. The 500 highest values of the test statistics in the STO and the LRO methods were the critical values in each model. Also, we generated 1,000 random data sets for 12 different hypotheses in each true cluster model and searched for clusters with high rates of high categories with a circular scanning window. The significance levels (α_1) assumed were 0.10, 0.20, and 0.40. Based on the critical value at the level of $\alpha_0 = 0.05$, we computed the number of rejected data sets out of 1000 which is the estimated power of the tests. To assess the accuracy of detected cluster locations and sizes, we considered sensitivity and the positive predicted value (PPV) as

$$\text{Sensitivity} = \frac{number\ of\ districts\ correctly\ detected}{number\ of\ districts\ in\ the\ true\ cluster} \qquad (11)$$

$$\text{PPV} = \frac{number\ of\ districts\ correctly\ detected}{number\ of\ detected\ districts}. \qquad (12)$$

We calculated the average of proportions only for rejected data sets at a significance level of 0.05 in both sensitivity and the PPV. A larger value of these measures means that the method is more precise for detecting the true cluster. For example, a lower value of PPV means that the method tends to detect larger clusters than true clusters. In case of a lower sensitivity, the method may miss more regions in detecting the true cluster.

3.2 Bivariate power distribution

Because the usual power estimates rejected data sets under the null hypothesis of no clustering, it does not reflect the precision of correctly detecting a cluster when data sets are rejected. Tango and Takahashi (2005) proposed a bivariate power distribution based on Monte Carlo simulations in order to compare the power performance of spatial scan statistics. The bivariate power distribution of $P(l,s)$, which is defined by the length l of the significant clusters and the number of s of the regions identified out of the assumed s^* regions in a true cluster, can be expressed by

$$P(l,s) = \frac{\#\{significant\ clusters\ have\ length\ l\ and\ include\ s\ true\ regions\}}{trials\ for\ each\ simulation} \quad (13)$$

where $1 \leq l, 0 \leq s \leq \min\{l, s^*\}$. When we are interested in the power of exact detection, the probability of exact detection is estimated as $P(s^*, s^*)$. The usual power can be defined as the sum of $P(l,s)$:

$$P(+,+) = \sum_{1 \leq l} \sum_{0 \leq s} P(l,s). \quad (14)$$

The bivariate power distribution can be a good measure to compare the performance of power in terms of the over-detection problem. For instance, if the number of regions in true cluster is 3, $P(l > 3, s \geq 3)$ might be over-detected

because those cases include the regions with non-elevated risk. We can identify the accuracy of cluster detection from the bivariate power distribution. For each simulation with the use of 1,000 replications at significance levels of 0.05, we presented the power distribution $P(l,s) \times 1000$ for a more intuitive understanding.

3.3 Results

Tables 3 through 5 show the estimated power, sensitivity, and PPV in each cluster model with different hypotheses for the original and our proposed methods using the STO-based and the LRO-based methods. Also, the bivariate power distribution $P(l,s) \times 1000$ in one scenario for each model is shown in Tables 6 through 11 as an example. The rest of the simulation results showed the same pattern and thus we omitted those results due to the limited space.

As we expected, in the case of the original spatial scan statistics, the STO-based method seems to perform better when compared with the LRO-based method under the STO hypothesis, while both methods have similar capacities of sensitivity, PPV and usual power under the LRO hypothesis. In particular, when the true cluster is of an irregular shape in the cluster models B (see Table 3), D

(see Table 4) and F (see Table 5), the STO-based scan statistic showed higher values of power and sensitivity than the LRO-based approach under the STO hypothesis. However, there are some cases where the STO-based approach had a lower PPV than the LRO-based approach, even though there was a slight discrepancy. The overall pattern showed similar results to Jung and Lee (2011).

Moreover, the original spatial scan statistics tended to detect larger clusters than the true cluster in both STO-based and LRO-based approaches. As we see the PPV in Table 5, the original scan statistics always have lower PPV than the restricted scan statistics regardless of their conditions, in particular, in the case of the irregular true cluster in cluster models B, D, and F. This can be interpreted that the original spatial scan statistics seem to over-detect. For more details, we can identify the over-detected phenomena from the bivariate power distribution in Tables 6 through 11. Although the original methods showed relatively higher powers, the estimated bivariate power distribution had a long tail which is an undesirable phenomenon of over-detection. For example, in Table 6, there are many cases in the original spatial scan statistic to include four or more regions in the detected cluster, while the true cluster consists of only three regions. In the case of an irregular shaped true cluster in Table 7, specifically, the original spatial scan statistics detected spatial clusters including between 4 and 13 regions.

Compared with the original spatial scan statistics, our proposed spatial scan statistics with a restricted likelihood ratio seem to alleviate the undesirable property. Even if our restricted spatial scan statistics tends to have lower sensitivity and power, we can resolve this by adjusting the screening value (α_1). We find that sensitivity and power increase as the screening value increases, but the PPV is still higher in the restricted methods. Even though our proposed method tends to overlook some true regions when the screening value is very low, it scans only the regions with elevated risk.

Table 3. The estimated power of STO-based and LRO-based methods in cluster models A - F.

Cluster model	Null hypothesis	Alternative Hypothesis	STO				LRO			
			Original (%)	Restricted (%)			Original (%)	Restricted (%)		
				0.1	0.2	0.4		0.1	0.2	0.4
A	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	100.00	86.40	96.00	99.50	100.00	74.60	90.90	98.00
		P=(0.15,0.15,0.45,0.25)	100.00	92.60	98.00	99.40	100.00	87.30	94.20	98.40
		P=(0.10,0.20,0.30,0.40)	100.00	86.60	93.80	99.10	100.00	97.80	99.00	100.00
		P=(0.05,0.25,0.25,0.45)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	100.00	96.50	99.40	100.00	100.00	86.10	95.00	99.40
		P=(0.25,0.05,0.50,0.20)	100.00	99.90	100.00	100.00	100.00	90.20	96.40	99.40
		P=(0.15,0.20,0.30,0.35)	100.00	70.20	80.70	94.10	100.00	85.70	92.70	99.10
		P=(0.10,0.15,0.25,0.50)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	100.00	100.00	100.00	100.00	100.00	91.50	96.70	99.70
		P=(0.35,0.05,0.40,0.20)	100.00	100.00	100.00	100.00	100.00	95.10	98.10	99.80
		P=(0.25,0.15,0.25,0.35)	100.00	77.40	89.50	96.20	100.00	90.40	96.50	99.40
		P=(0.10,0.10,0.40,0.40)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
B	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	96.30	87.40	94.70	95.50	86.00	69.50	80.50	85.40
		P=(0.15,0.15,0.45,0.25)	97.50	93.80	97.90	97.50	85.20	85.10	89.40	84.70
		P=(0.10,0.20,0.30,0.40)	97.50	88.00	92.90	94.80	99.80	97.90	99.00	100.00
		P=(0.05,0.25,0.25,0.45)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	100.00	97.60	99.90	100.00	85.90	73.70	80.80	76.60
		P=(0.25,0.05,0.50,0.20)	100.00	100.00	100.00	100.00	92.70	90.10	92.00	91.60
		P=(0.15,0.20,0.30,0.35)	99.40	74.70	79.60	82.80	95.50	86.70	92.90	94.80
		P=(0.10,0.15,0.25,0.50)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	100.00	99.50	100.00	100.00	85.10	76.90	82.50	80.70
		P=(0.35,0.05,0.40,0.20)	100.00	100.00	100.00	100.00	86.10	77.80	82.20	75.90
		P=(0.25,0.15,0.25,0.35)	83.50	62.20	69.40	71.60	96.70	82.70	91.80	94.20
		P=(0.10,0.10,0.40,0.40)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Continued

Table 3. The estimated power of STO-based and LRO-based methods in cluster models A - F. (Continued)

Cluster model	Null hypothesis	Alternative Hypothesis	STO				LRO			
			Original (%)	Restricted (%)			Original (%)	Restricted (%)		
				0.1	0.2	0.4		0.1	0.2	0.4
C	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	100.00	99.70	100.00	100.00	100.00	98.80	99.90	100.00
		P=(0.15,0.15,0.45,0.25)	100.00	100.00	100.00	100.00	100.00	99.10	99.90	100.00
		P=(0.10,0.20,0.30,0.40)	100.00	99.70	100.00	100.00	100.00	100.00	100.00	100.00
		P=(0.05,0.25,0.25,0.45)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	100.00	100.00	100.00	100.00	100.00	99.30	100.00	100.00
		P=(0.25,0.05,0.50,0.20)	100.00	100.00	100.00	100.00	100.00	99.90	100.00	100.00
		P=(0.15,0.20,0.30,0.35)	100.00	97.90	99.60	100.00	100.00	99.50	100.00	100.00
		P=(0.10,0.15,0.25,0.50)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	100.00	100.00	100.00	100.00	100.00	99.80	100.00	100.00
		P=(0.35,0.05,0.40,0.20)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		P=(0.25,0.15,0.25,0.35)	100.00	98.10	99.50	100.00	100.00	99.70	100.00	100.00
		P=(0.10,0.10,0.40,0.40)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
D	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	100.00	99.30	100.00	100.00	100.00	95.70	98.90	99.70
		P=(0.15,0.15,0.45,0.25)	100.00	99.40	99.80	100.00	100.00	97.20	98.60	99.50
		P=(0.10,0.20,0.30,0.40)	100.00	99.60	100.00	100.00	100.00	100.00	100.00	100.00
		P=(0.05,0.25,0.25,0.45)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	100.00	100.00	100.00	100.00	100.00	99.30	99.70	100.00
		P=(0.25,0.05,0.50,0.20)	100.00	100.00	100.00	100.00	100.00	99.30	99.90	100.00
		P=(0.15,0.20,0.30,0.35)	100.00	94.40	99.40	99.70	100.00	98.80	99.90	100.00
		P=(0.10,0.15,0.25,0.50)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	100.00	100.00	100.00	100.00	100.00	98.20	99.70	100.00
		P=(0.35,0.05,0.40,0.20)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		P=(0.25,0.15,0.25,0.35)	100.00	94.50	98.60	99.50	100.00	99.00	99.80	100.00
		P=(0.10,0.10,0.40,0.40)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Continued

Table 3. The estimated power of STO-based and LRO-based methods in cluster models A - F. (Continued)

Cluster model	Null hypothesis	Alternative Hypothesis	STO				LRO			
			Original (%)	Restricted (%)			Original (%)	Restricted (%)		
				0.1	0.2	0.4		0.1	0.2	0.4
E	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	100.00	100.00	100.00	100.00	100.00	99.10	99.60	100.00
		P=(0.15,0.15,0.45,0.25)	100.00	100.00	100.00	100.00	100.00	99.80	100.00	100.00
		P=(0.10,0.20,0.30,0.40)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		P=(0.05,0.25,0.25,0.45)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	100.00	100.00	100.00	100.00	100.00	99.90	100.00	100.00
		P=(0.25,0.05,0.50,0.20)	100.00	100.00	100.00	100.00	100.00	99.90	100.00	100.00
		P=(0.15,0.20,0.30,0.35)	100.00	98.60	99.70	100.00	100.00	99.70	100.00	100.00
		P=(0.10,0.15,0.25,0.50)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	100.00	100.00	100.00	100.00	100.00	99.80	100.00	100.00
		P=(0.35,0.05,0.40,0.20)	100.00	100.00	100.00	100.00	100.00	99.90	99.90	100.00
		P=(0.25,0.15,0.25,0.35)	100.00	97.80	99.70	100.00	100.00	99.70	100.00	100.00
		P=(0.10,0.10,0.40,0.40)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
F	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	100.00	99.80	99.90	100.00	100.00	98.50	99.50	99.80
		P=(0.15,0.15,0.45,0.25)	100.00	99.40	99.40	100.00	100.00	98.40	99.50	100.00
		P=(0.10,0.20,0.30,0.40)	100.00	99.80	100.00	100.00	100.00	100.00	100.00	100.00
		P=(0.05,0.25,0.25,0.45)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	100.00	99.90	100.00	100.00	100.00	98.70	99.70	100.00
		P=(0.25,0.05,0.50,0.20)	100.00	100.00	100.00	100.00	100.00	99.70	100.00	100.00
		P=(0.15,0.20,0.30,0.35)	100.00	95.90	98.70	99.90	100.00	99.10	99.70	100.00
		P=(0.10,0.15,0.25,0.50)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	100.00	100.00	100.00	100.00	100.00	99.60	100.00	100.00
		P=(0.35,0.05,0.40,0.20)	100.00	100.00	100.00	100.00	100.00	99.40	100.00	100.00
		P=(0.25,0.15,0.25,0.35)	100.00	100.00	100.00	100.00	100.00	99.40	100.00	100.00
		P=(0.10,0.10,0.40,0.40)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 4. The estimated sensitivity of STO-based and LRO-based methods in cluster models A - F.

Cluster model	Null hypothesis	Alternative Hypothesis	STO				LRO			
			Original (%)	Restricted (%)			Original (%)	Restricted (%)		
				0.1	0.2	0.4		0.1	0.2	0.4
A	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	98.33	64.62	81.28	93.63	97.83	67.74	82.54	93.50
		P=(0.15,0.15,0.45,0.25)	97.43	67.13	80.92	92.45	96.00	68.19	81.71	91.50
		P=(0.10,0.20,0.30,0.40)	97.77	63.32	79.03	92.77	97.80	63.26	78.89	93.07
		P=(0.05,0.25,0.25,0.45)	99.83	96.17	99.03	99.77	99.87	96.33	99.07	99.80
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	98.83	72.71	88.53	97.57	96.57	73.98	87.72	95.07
		P=(0.25,0.05,0.50,0.20)	98.80	82.48	92.00	97.97	96.37	83.63	90.63	95.64
		P=(0.15,0.20,0.30,0.35)	95.53	56.51	72.49	86.89	96.10	54.18	69.15	85.20
		P=(0.10,0.15,0.25,0.50)	99.80	98.73	99.53	99.73	99.80	98.73	99.53	99.73
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	98.90	81.17	93.43	98.17	96.67	81.35	92.24	95.65
		P=(0.35,0.05,0.40,0.20)	99.40	91.03	96.53	99.07	97.53	89.27	94.70	97.09
		P=(0.25,0.15,0.25,0.35)	95.67	61.46	76.91	88.60	95.80	58.92	74.65	87.49
		P=(0.10,0.10,0.40,0.40)	99.93	99.93	99.93	99.93	99.93	99.93	99.93	99.93
B	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	82.66	54.50	61.21	66.67	82.52	53.62	58.72	64.56
		P=(0.15,0.15,0.45,0.25)	80.89	56.43	61.59	67.11	80.79	54.80	59.62	65.01
		P=(0.10,0.20,0.30,0.40)	87.38	51.67	58.16	66.21	87.68	55.77	64.34	72.70
		P=(0.05,0.25,0.25,0.45)	91.80	89.07	90.40	91.20	93.73	92.67	94.63	94.03
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	88.27	58.81	67.63	73.43	85.37	53.82	58.58	65.75
		P=(0.25,0.05,0.50,0.20)	83.97	67.90	72.97	75.53	81.70	58.86	62.72	67.79
		P=(0.15,0.20,0.30,0.35)	80.92	51.41	56.37	63.29	79.86	51.75	57.23	64.17
		P=(0.10,0.15,0.25,0.50)	92.73	88.13	87.93	87.37	90.50	91.10	91.13	91.20
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	84.20	66.16	73.70	77.93	83.86	55.44	59.72	65.96
		P=(0.35,0.05,0.40,0.20)	90.13	77.83	83.53	84.50	84.67	54.93	59.85	64.87
		P=(0.25,0.15,0.25,0.35)	84.55	48.12	55.04	62.43	85.59	48.00	55.66	61.89
		P=(0.10,0.10,0.40,0.40)	98.43	99.23	99.23	99.03	99.23	99.87	99.83	99.70

Continued

Table 4. The estimated sensitivity of STO-based and LRO-based methods in cluster models A - F. (Continued)

Cluster model	Null hypothesis	Alternative Hypothesis	STO				LRO			
			Original (%)	Restricted (%)			Original (%)	Restricted (%)		
				0.1	0.2	0.4		0.1	0.2	0.4
C	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	97.92	67.04	82.76	93.56	96.70	64.92	80.98	92.02
		P=(0.15,0.15,0.45,0.25)	98.26	67.30	80.82	92.64	97.44	64.00	77.84	91.04
		P=(0.10,0.20,0.30,0.40)	98.40	68.00	82.72	94.08	98.40	69.76	83.56	94.34
		P=(0.05,0.25,0.25,0.45)	99.74	94.38	98.44	99.64	99.74	94.72	98.48	99.64
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	98.94	77.92	90.32	96.98	97.30	73.43	87.74	95.10
		P=(0.25,0.05,0.50,0.20)	99.14	84.26	93.94	98.40	97.40	80.32	91.46	96.38
		P=(0.15,0.20,0.30,0.35)	97.12	56.63	73.82	88.00	97.24	58.15	75.20	88.98
		P=(0.10,0.15,0.25,0.50)	99.92	98.26	99.20	99.82	99.92	98.30	99.20	99.84
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	98.68	82.76	92.72	97.66	96.38	78.74	89.52	95.18
		P=(0.35,0.05,0.40,0.20)	98.94	87.84	94.88	97.94	96.14	84.38	92.00	95.24
		P=(0.25,0.15,0.25,0.35)	97.20	55.88	70.57	85.78	97.20	58.44	72.78	86.90
		P=(0.10,0.10,0.40,0.40)	99.98	99.94	99.98	99.98	99.98	99.94	99.98	99.98
D	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	76.40	53.90	60.42	63.28	77.16	49.63	57.01	61.00
		P=(0.15,0.15,0.45,0.25)	84.10	53.02	58.36	64.10	81.28	50.45	54.83	60.96
		P=(0.10,0.20,0.30,0.40)	78.26	54.78	60.32	63.42	78.78	58.72	63.36	66.44
		P=(0.05,0.25,0.25,0.45)	80.48	80.46	79.92	79.08	83.08	86.48	86.14	84.00
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	77.58	62.96	67.20	68.34	77.96	53.64	57.97	61.24
		P=(0.25,0.05,0.50,0.20)	78.56	66.18	67.92	69.34	79.34	56.92	60.46	63.56
		P=(0.15,0.20,0.30,0.35)	77.28	46.25	53.82	60.54	77.76	49.33	57.76	63.74
		P=(0.10,0.15,0.25,0.50)	91.66	89.74	89.62	89.20	93.80	93.14	93.22	92.68
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	79.10	66.22	69.12	69.74	79.06	55.03	59.16	61.50
		P=(0.35,0.05,0.40,0.20)	98.94	87.84	94.88	97.94	96.14	84.38	92.00	95.24
		P=(0.25,0.15,0.25,0.35)	80.48	46.86	54.22	60.02	80.58	49.68	57.21	62.10
		P=(0.10,0.10,0.40,0.40)	93.56	97.42	96.82	96.24	96.18	98.50	98.40	97.92

Continued

Table 4. The estimated sensitivity of STO-based and LRO-based in cluster models A - F. (Continued)

Cluster model	Null hypothesis	Alternative Hypothesis	STO				LRO			
			Original (%)	Restricted (%)			Original (%)	Restricted (%)		
				0.1	0.2	0.4		0.1	0.2	0.4
E	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	99.19	57.93	73.11	89.47	98.97	53.41	70.08	87.51
		P=(0.15,0.15,0.45,0.25)	99.10	58.21	71.40	86.36	98.46	55.60	69.11	84.60
		P=(0.10,0.20,0.30,0.40)	99.19	58.09	72.36	87.90	99.19	61.70	75.07	88.90
		P=(0.05,0.25,0.25,0.45)	99.87	93.63	98.07	99.67	99.89	94.07	98.21	99.69
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	99.57	69.87	84.97	95.34	98.50	62.26	80.09	92.89
		P=(0.25,0.05,0.50,0.20)	99.67	76.21	88.96	97.60	98.79	69.53	85.54	96.13
		P=(0.15,0.20,0.30,0.35)	98.67	47.62	62.26	80.99	98.76	50.34	64.57	82.13
		P=(0.10,0.15,0.25,0.50)	99.94	97.16	98.79	99.79	99.94	97.26	98.83	99.79
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	99.60	75.74	88.16	96.74	98.07	69.50	84.34	94.71
		P=(0.35,0.05,0.40,0.20)	99.83	83.80	93.11	98.14	98.44	75.75	89.02	96.03
		P=(0.25,0.15,0.25,0.35)	98.47	46.33	60.21	77.79	98.47	50.29	63.79	79.66
		P=(0.10,0.10,0.40,0.40)	99.99	99.89	99.94	99.97	99.99	99.89	99.94	99.97
F	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	77.37	48.34	57.67	62.96	76.01	42.64	48.79	56.01
		P=(0.15,0.15,0.45,0.25)	78.23	45.56	53.31	60.14	74.67	41.87	48.20	55.26
		P=(0.10,0.20,0.30,0.40)	75.80	48.41	56.61	62.96	77.50	54.26	63.73	69.31
		P=(0.05,0.25,0.25,0.45)	85.69	86.13	87.70	87.03	88.41	88.79	90.47	90.80
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	75.74	60.60	67.93	70.43	71.94	46.40	50.87	56.06
		P=(0.25,0.05,0.50,0.20)	82.30	65.54	72.29	74.37	76.24	49.29	53.26	57.81
		P=(0.15,0.20,0.30,0.35)	71.70	40.01	48.05	54.10	72.56	43.68	52.83	58.39
		P=(0.10,0.15,0.25,0.50)	90.49	90.44	91.07	90.61	92.03	93.10	93.50	93.29
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	78.10	65.17	71.59	73.91	73.07	46.26	51.86	56.89
		P=(0.35,0.05,0.40,0.20)	81.50	72.87	76.74	77.90	72.71	50.34	53.47	58.80
		P=(0.25,0.15,0.25,0.35)	81.50	72.87	76.74	77.90	72.71	50.34	53.47	58.80
		P=(0.10,0.10,0.40,0.40)	97.44	98.96	98.64	98.34	98.43	99.50	99.39	99.14

Table 5. The estimated PPV of STO-based and LRO-based methods in cluster models A - F.

Cluster model	Null hypothesis	Alternative Hypothesis	STO				LRO			
			Original (%)	Restricted (%)			Original (%)	Restricted (%)		
				0.1	0.2	0.4		0.1	0.2	0.4
A	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	91.46	98.02	97.58	95.40	91.52	98.11	97.62	95.60
		P=(0.15,0.15,0.45,0.25)	90.60	97.40	97.29	95.56	90.94	97.73	97.28	95.74
		P=(0.10,0.20,0.30,0.40)	93.23	98.69	98.22	96.22	93.71	99.00	98.41	96.71
		P=(0.05,0.25,0.25,0.45)	98.72	99.55	99.12	98.71	98.96	99.73	99.28	98.96
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	98.83	72.71	88.53	97.57	90.71	98.05	97.91	95.73
		P=(0.25,0.05,0.50,0.20)	93.31	97.99	97.67	96.14	90.54	97.72	97.09	95.79
		P=(0.15,0.20,0.30,0.35)	92.20	97.97	97.41	96.50	92.39	98.58	98.01	96.59
		P=(0.10,0.15,0.25,0.50)	90.97	97.88	97.57	96.08	98.74	99.56	99.35	98.86
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	94.45	97.93	97.88	96.86	92.29	98.82	98.07	96.33
		P=(0.35,0.05,0.40,0.20)	89.82	97.97	97.22	95.51	89.98	98.31	97.54	95.34
		P=(0.25,0.15,0.25,0.35)	89.63	97.29	96.94	95.24	90.42	98.28	97.73	95.98
		P=(0.10,0.10,0.40,0.40)	99.29	99.52	99.38	99.25	99.66	99.60	99.61	99.61
B	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	76.10	96.87	95.17	91.13	76.46	97.06	94.78	91.40
		P=(0.15,0.15,0.45,0.25)	77.60	96.72	95.19	91.18	76.26	97.34	95.83	90.49
		P=(0.10,0.20,0.30,0.40)	73.61	97.68	95.65	90.82	73.89	98.53	96.48	91.76
		P=(0.05,0.25,0.25,0.45)	80.39	98.04	96.35	92.38	80.55	98.30	96.67	92.82
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	70.89	97.65	96.12	91.36	73.10	96.45	94.36	88.90
		P=(0.25,0.05,0.50,0.20)	71.82	97.49	95.75	92.37	77.60	97.17	95.51	91.56
		P=(0.15,0.20,0.30,0.35)	74.89	96.39	94.62	90.56	77.40	97.74	96.23	92.66
		P=(0.10,0.15,0.25,0.50)	82.02	97.80	96.30	93.40	84.08	98.22	96.81	94.05
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	78.59	97.41	95.99	92.06	75.47	96.97	95.34	89.64
		P=(0.35,0.05,0.40,0.20)	73.57	96.79	94.24	90.10	74.26	96.22	94.39	89.17
		P=(0.25,0.15,0.25,0.35)	73.00	96.61	93.97	89.17	73.83	98.16	96.07	91.32
		P=(0.10,0.10,0.40,0.40)	82.04	98.50	97.04	94.06	82.29	98.70	97.29	94.35

Continued

Table 5. The estimated PPV of STO-based and LRO-based methods in cluster models A - F. (Continued)

Cluster model	Null hypothesis	Alternative Hypothesis	STO				LRO			
			Original (%)	Restricted (%)			Original (%)	Restricted (%)		
				0.1	0.2	0.4		0.1	0.2	0.4
C	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	94.91	98.90	98.62	98.07	98.18	99.63	99.50	99.12
		P=(0.15,0.15,0.45,0.25)	92.39	98.78	98.43	97.74	98.32	99.23	99.13	98.99
		P=(0.10,0.20,0.30,0.40)	98.87	99.63	99.60	99.35	98.93	99.74	99.66	99.41
		P=(0.05,0.25,0.25,0.45)	99.62	99.76	99.81	99.72	99.86	99.97	99.93	99.89
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	91.26	98.63	98.63	98.06	97.78	99.67	99.58	99.09
		P=(0.25,0.05,0.50,0.20)	80.19	98.17	97.83	96.57	97.64	99.12	98.94	98.68
		P=(0.15,0.20,0.30,0.35)	98.63	99.52	99.54	99.18	98.75	99.81	99.76	99.33
		P=(0.10,0.15,0.25,0.50)	99.80	99.83	99.81	99.81	99.85	99.93	99.91	99.90
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	82.19	98.43	97.48	96.05	98.29	99.57	99.43	98.98
		P=(0.35,0.05,0.40,0.20)	60.35	97.42	96.61	94.38	97.78	99.26	99.12	98.55
		P=(0.25,0.15,0.25,0.35)	97.84	99.35	99.17	98.83	97.98	99.67	99.52	99.10
		P=(0.10,0.10,0.40,0.40)	99.95	99.90	99.88	99.88	99.98	99.98	99.97	99.97
D	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	80.08	98.57	98.19	96.69	78.54	99.13	98.60	96.88
		P=(0.15,0.15,0.45,0.25)	72.27	98.33	97.87	94.78	74.50	99.25	98.30	95.26
		P=(0.10,0.20,0.30,0.40)	80.06	99.10	99.04	97.24	80.04	99.44	99.23	97.39
		P=(0.05,0.25,0.25,0.45)	86.53	99.84	99.61	98.57	86.87	99.89	99.64	98.63
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	80.84	98.49	97.77	96.25	77.64	99.26	98.73	96.69
		P=(0.25,0.05,0.50,0.20)	79.29	97.72	97.23	94.30	77.16	99.01	98.19	95.52
		P=(0.15,0.20,0.30,0.35)	78.33	98.97	98.50	96.61	78.64	99.17	98.81	97.15
		P=(0.10,0.15,0.25,0.50)	74.70	99.55	99.05	95.43	74.98	99.70	99.15	95.68
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	79.06	98.33	97.67	95.41	76.48	99.11	98.48	95.44
		P=(0.35,0.05,0.40,0.20)	60.35	97.42	96.61	94.38	97.78	99.26	99.12	98.55
		P=(0.25,0.15,0.25,0.35)	76.11	99.14	98.63	96.31	76.49	99.38	98.85	96.64
		P=(0.10,0.10,0.40,0.40)	94.79	99.72	99.64	99.43	94.92	99.93	99.89	99.70

Continued

Table 5. The estimated PPV of STO-based and LRO-based methods in cluster models A - F. (Continued)

Cluster model	Null hypothesis	Alternative Hypothesis	STO				LRO			
			Original (%)	Restricted (%)			Original (%)	Restricted (%)		
				0.1	0.2	0.4		0.1	0.2	0.4
E	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	90.97	98.53	98.13	97.14	98.32	99.41	99.30	99.04
		P=(0.15,0.15,0.45,0.25)	89.38	98.42	97.69	96.41	97.76	99.31	99.10	98.48
		P=(0.10,0.20,0.30,0.40)	99.04	99.75	99.32	99.30	99.05	99.86	99.43	99.38
		P=(0.05,0.25,0.25,0.45)	99.88	99.89	99.87	99.87	99.89	99.91	99.89	99.88
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	77.79	98.20	97.62	95.88	97.92	99.61	99.32	98.94
		P=(0.25,0.05,0.50,0.20)	52.47	97.03	96.14	92.91	98.25	99.17	99.09	99.00
		P=(0.15,0.20,0.30,0.35)	98.51	99.27	98.88	98.51	98.58	99.43	98.99	98.62
		P=(0.10,0.15,0.25,0.50)	99.96	99.92	99.87	99.90	99.95	99.93	99.88	99.93
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	60.11	97.72	96.49	94.07	97.80	99.48	99.28	98.92
		P=(0.35,0.05,0.40,0.20)	41.18	95.13	92.85	87.65	97.14	99.13	98.75	98.19
		P=(0.25,0.15,0.25,0.35)	98.16	99.35	98.80	98.20	98.16	99.42	98.92	98.25
		P=(0.10,0.10,0.40,0.40)	99.98	99.94	99.94	99.98	99.99	100.00	99.99	99.99
F	p=q =(0.25,0.25,0.25,0.25)	P=(0.20,0.10,0.40,0.30)	77.42	98.42	97.14	93.03	77.15	99.16	97.73	93.92
		P=(0.15,0.15,0.45,0.25)	77.11	97.72	96.95	92.74	78.85	99.04	97.98	94.59
		P=(0.10,0.20,0.30,0.40)	80.06	99.32	98.26	94.40	80.34	99.43	98.34	94.66
		P=(0.05,0.25,0.25,0.45)	82.26	98.47	96.72	92.83	82.61	98.58	96.91	93.06
	p=q =(0.30,0.20,0.30,0.20)	P=(0.25,0.05,0.45,0.25)	81.51	98.12	96.55	92.86	80.43	98.86	97.33	94.12
		P=(0.25,0.05,0.50,0.20)	73.67	96.30	94.16	88.57	77.40	98.23	96.41	92.23
		P=(0.15,0.20,0.30,0.35)	80.75	99.46	98.64	95.95	81.11	99.57	98.80	96.11
		P=(0.10,0.15,0.25,0.50)	78.06	95.98	92.65	87.51	78.33	96.11	92.84	87.80
	p=q =(0.40,0.20,0.20,0.20)	P=(0.35,0.05,0.35,0.25)	78.64	96.60	94.76	90.61	79.63	98.46	97.13	93.62
		P=(0.35,0.05,0.40,0.20)	74.58	95.73	93.32	87.08	79.47	97.88	96.27	92.28
		P=(0.25,0.15,0.25,0.35)	74.58	95.73	93.32	87.08	79.47	97.88	96.27	92.28
		P=(0.10,0.10,0.40,0.40)	82.83	95.72	93.36	89.28	82.88	95.78	93.39	89.27

Table 6. Estimated bivariate power distributions $P(l,s)\times 1000$ of original and proposed STO-based methods for $H_0: p = q = (0.25,0.25,0.25,0.25)$ and $H_1: p = (0.20, 0.10, 40, 0.30)$ under the STO hypothesis in cluster model A (circular).

STO-based original scan statistic				
Length	Include *s* true regions			
l	0	1	2	3
1	0	9	0	0
2	0	0	30	0
3	0	1	0	747
4	0	0	0	95
5	0	0	0	51
6	0	0	0	31
7	0	0	0	9
8	0	0	0	5
9	0	0	0	2
10	0	0	0	4
11	0	0	0	1
12	0	0	0	2
13	0	0	0	4
14	0	0	0	1
15	0	0	0	3
16	0	0	0	3
17	0	0	0	0
18	0	0	0	0
19	0	0	0	0
20	0	0	0	0
21	0	0	0	2
		usual power = 1.000		

STO-based restricted scan statistic					
α_1	Length	Include *s* true regions			
	l	0	1	2	3
0.10	1	1	316	0	0
	2	0	7	245	0
	3	0	3	14	258
	4	0	0	3	16
	5	0	0	0	1
			usual power = 0.864		
0.20	1	0	186	0	0
	2	0	3	142	0
	3	0	2	10	561
	4	0	1	3	43
	5	0	0	0	6
	6	0	0	0	1
	7	0	0	0	2
			usual power = 0.960		
0.40	1	0	60	0	0
	2	0	0	55	0
	3	0	3	3	744
	4	0	1	0	84
	5	0	1	0	25
	6	0	0	0	13
	7	0	1	0	5
			usual power = 0.995		

Table 7. Estimated bivariate power distributions $P(l,s)\times 1000$ of original and proposed LRO-based methods for $H_0: p = q = (0.30,0.20,0.30,0.20)$ and $H_1: p = (0.15,0.20,0.30,0.35)$ under the LRO hypothesis in cluster model B (irregular).

LRO-based original scan statistic				
Length	Include s true regions			
l	0	1	2	3
1	0	98	0	0
2	0	5	296	0
3	0	1	35	0
4	0	0	28	284
5	0	0	9	34
6	0	0	1	67
7	0	0	0	24
8	0	0	0	28
9	0	0	0	18
10	0	0	0	11
11	0	0	0	7
12	0	0	0	5
13	0	0	0	4
		usual power = 0.955		

LRO-based restricted scan statistic					
α_1	Length	Include s true regions			
	l	0	1	2	3
0.10	1	0	401	0	0
	2	0	18	395	0
	3	0	0	19	20
	4	0	0	3	11
			usual power =0.867		
0.20	1	0	300	0	0
	2	0	20	511	0
	3	0	2	32	18
	4	0	1	3	40
	5	0	0	0	2
			usual power = 0.929		
0.40	1	0	199	0	0
	2	0	20	524	0
	3	1	0	44	11
	4	0	0	10	116
	5	0	0	0	17
	6	0	0	0	5
	7	0	0	0	1
			usual power = 0.948		

Table 8. Estimated bivariate power distributions $P(l,s) \times 1000$ of original and proposed LRO-based methods for $H_0: p = q = (0.30,0.20,0.30,0.20)$ and $H_1: p = (0.15,0.20,0.30,0.35)$ under the LRO hypothesis in cluster model C (circular).

LRO-based original scan statistic						
Length	Include *s* true regions					
l	0	1	2	3	4	5
1	0	0	0	0	0	0
2	0	0	1	0	0	0
3	0	0	0	11	0	0
4	0	0	0	1	111	0
5	0	0	0	0	0	815
6	0	0	0	0	0	48
7	0	0	0	0	0	8
8	0	0	0	0	0	4
9	0	0	0	0	0	0
10	0	0	0	0	0	1
usual power = 1.000						

LRO-based restricted scan statistic							
α_1	Length	Include *s* true regions					
	l	0	1	2	3	4	5
0.10	1	0	88	0	0	0	0
	2	0	0	255	0	0	0
	3	0	0	3	349	0	0
	4	0	0	0	1	253	0
	5	0	0	0	0	3	43
	usual power = 0.995						
0.20	1	0	12	0	0	0	0
	2	0	0	99	0	0	0
	3	0	0	1	249	0	0
	4	0	0	0	3	384	0
	5	0	0	0	0	4	245
	6	0	0	0	0	0	3
	usual power = 1.000						
0.40	1	0	3	0	0	0	0
	2	0	0	21	0	0	0
	3	0	0	0	95	0	0
	4	0	0	0	4	270	0
	5	0	0	0	0	8	576
	6	0	0	0	0	0	21
	7	0	0	0	0	0	2
	usual power = 1.000						

Table 9. Estimated bivariate power distributions $P(l,s)\times 1000$ of original and proposed STO-based methods for $H_0: p = q = (0.40,0.20,0.20,0.20)$ and $H_1: p = (0.35,0.05,0.40,0.20)$ under the STO hypothesis in cluster model D (irregular).

STO-based original scan statistic						
Length	Include *s* true regions					
l	0	1	2	3	4	5
1	0	0	0	0	0	0
2	0	0	6	0	0	0
3	0	0	0	335	0	0
4	0	0	0	6	122	0
5	0	0	0	5	0	0
6	0	0	0	5	57	0
7	0	0	0	3	3	0
8	0	0	0	5	3	257
9	0	0	0	2	5	96
10	0	0	0	3	1	7
11	0	0	1	6	3	7
12	0	0	0	5	2	13
13	0	0	0	0	7	4
14	0	0	0	0	2	4
15	0	0	0	0	1	11
16	0	0	0	0	1	11
17	0	0	0	0	0	1
usual power = 1.000						

STO-based restricted scan statistic							
α_1	Length	Include *s* true regions					
	l	0	1	2	3	4	5
0.10	1	0	0	0	0	0	0
	2	0	0	18	0	0	0
	3	0	0	0	403	0	0
	4	0	0	1	19	472	0
	5	0	0	0	11	35	19
	6	0	0	0	0	16	1
	7	0	0	0	2	0	1
	8	0	0	0	0	1	1
	usual power = 1.000						
0.20	1	0	0	0	0	0	0
	2	0	0	7	0	0	0
	3	0	0	0	386	0	0
	4	0	0	0	21	477	0
	5	0	0	0	12	21	21
	6	0	0	1	8	24	2
	7	0	0	0	8	5	1
	8	0	0	0	2	2	1
	9	0	0	0	0	0	0
	10	0	0	0	0	0	1
	usual power = 1.000						
0.40	1	0	0	0	0	0	0
	2	0	0	5	0	0	0
	3	0	0	0	395	0	0
	4	0	0	0	16	427	0
	5	0	0	0	11	14	20
	6	0	0	1	13	30	0
	7	0	0	0	9	7	3
	8	0	0	0	5	12	16
	9	0	0	0	0	3	5
	10	0	0	0	1	4	2
	11	0	0	0	0	0	1
	usual power = 1.000						

Table 10. Estimated bivariate power distributions $P(l,s)\times 1000$ of original and proposed STO-based methods for $H_0: p = q = (0.30,0.20,0.30,0.20)$ and $H_1: p = (0.25,0.05,0.45,0.25)$ under the STO hypothesis in cluster model E (circular).

STO-based original scan statistic						
Length	Include *s* true regions					
l	0-2	3	4	5	6	7
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	2	0	0
6	0	0	0	0	16	0
7	0	0	0	0	1	586
8	0	0	0	0	0	25
9	0	0	0	0	1	4
10	0	0	0	0	1	7
11	0	0	0	0	0	9
12	0	0	0	0	0	4
13	0	0	0	0	0	13
14	0	0	0	0	1	13
15	0	0	0	0	0	18
16	0	0	0	0	0	23
17	0	0	0	0	2	30
18-22	0	0	0	0	4	240
usual power = 1.000						

STO-based restricted scan statistic							
α_1	Length	Include *s* true regions					
	l	0-2	3	4	5	6	7
0.10	1	0	0	0	0	0	0
	2	15	0	0	0	0	0
	3	1	70	0	0	0	0
	4	1	1	251	0	0	0
	5	0	0	17	340	0	0
	6	1	1	6	19	159	0
	7	0	0	0	8	4	90
	8	0	0	0	5	0	6
	9	0	0	0	1	0	1
	10-12	0	0	0	0	0	3
	usual power = 1.000						
0.20	1	0	0	0	0	0	0
	2	0	0	0	0	0	0
	3	0	10	0	0	0	0
	4	0	1	105	0	0	0
	5	0	0	4	212	0	0
	6	0	0	2	10	168	0
	7	0	0	1	5	10	408
	8	0	0	3	8	5	20
	9	0	0	0	3	2	8
	10-12	0	0	0	1	0	14
	usual power = 1.000						
0.40	1	0	0	0	0	0	0
	2	0	0	0	0	0	0
	3	0	1	0	0	0	0
	4	0	0	24	0	0	0
	5	0	0	0	68	0	0
	6	0	0	0	1	61	0
	7	0	0	0	1	5	701
	8	0	0	3	3	6	44
	9	0	0	0	1	4	14
	10	0	0	1	1	2	12
	11-15	0	0	1	1	5	40
	usual power = 1.000						

Table 11. Estimated bivariate power distributions $P(l,s)\times 1000$ of original and proposed LRO-based methods for $H_0: p = q = (0.40,0.20,0.20,0.20)$ and $H_1: p = (0.25,0.15,0.25,0.35)$ under the LRO hypothesis in cluster model F (irregular).

LRO-based original scan statistic						
Length	Include *s* true regions					
l	0-2	3	4	5	6	7
1	0	0	0	0	0	0
2	11	0	0	0	0	0
3	0	151	0	0	0	0
4	0	2	35	0	0	0
5	0	0	4	32	0	0
6	0	0	1	180	1	0
7	0	0	4	2	18	0
8	0	0	0	36	380	0
9	0	0	0	1	40	31
10	0	0	0	0	43	5
11	0	0	0	0	3	2
12	0	0	0	0	14	0
13	0	0	0	0	3	1
usual power = 1.000						

LRO-based restricted scan statistic							
α_1	Length	Include *s* true regions					
	l	0-2	3	4	5	6	7
0.10	1	71	0	0	0	0	0
	2	184	0	0	0	0	0
	3	1	373	0	0	0	0
	4	0	10	261	0	0	0
	5	0	0	6	71	0	0
	6	0	0	0	6	9	0
	7	0	0	0	0	0	1
	8	0	0	0	0	1	0
	usual power = 1.000						
0.20	1	26	0	0	0	0	0
	2	80	0	0	0	0	0
	3	1	300	0	0	0	0
	4	0	10	329	0	0	0
	5	0	0	9	188	0	0
	6	0	0	0	27	16	0
	7	0	0	0	2	6	0
	8	0	0	0	1	4	0
	9	0	0	0	0	1	0
	usual power = 1.000						
0.40	1	4	0	0	0	0	0
	2	34	0	0	0	0	0
	3	0	279	0	0	0	0
	4	0	8	256	0	0	0
	5	0	0	10	195	0	0
	6	0	0	0	108	20	0
	7	0	0	1	1	21	0
	8	0	0	0	6	45	0
	9	0	0	0	0	3	6
	10	0	0	0	0	3	0
	usual power = 1.000						

4. Application

4.1 Data explanation

We applied two approaches, the original and our proposed spatial scan statistics for the ordinal data, to real data in the 2014 Health Screening Program by the National Health Insurance Service (NHIS) of Korea. The data was obtained from the Korean Statistical Information Service (KOSIS). The NHIS annually offers the National Health Screening Statistical Yearbook since 2008 in order to provide basic data to be used for establishing medical and healthcare policies and health insurance policies, presenting directions for national policies for the improvement of regional health and medical service (NHIS, 2014). This program contains general health screening, life turning point health examinations, cancer screening, health screenings for infants, and other commissioned programs.

We used the data set of statistics on first diagnoses based on general health screening by district and gender in 2014 as an ordinal data (normal, caution, suspected disease, and diagnosed with diseases) only in Seoul with 25 districts (gu). Tables 12 and 13 show the criteria for determining diagnoses based on the results of general health screening and the number of cases and percentage by gender, as well as the proportion by gender in Seoul (Figure 2).

Table 12. The criteria for determining diagnoses based on the general health screening.

Division	Explanation	Criteria
Normal A	Individuals determined to be in sound health based on the results of the 1st step screening test	Normal (1)
Normal B (Cautionary)	Individuals determined to be normal health based on the results of the first step screening test but who require self-care and preventive measures through improvements in dietary habits and environmental conditions	Caution (2)
Suspected Disease – General	Individuals determined to be at risk of developing disease based on the results of the first screening test and who therefore require follow-up examinations or accurate diagnosis and treatment through a specialized medical institution	Suspected disease (3)
Suspected Disease – Hypertension or Diabetes	Individuals determined to be suspected of experiencing hypertension or diabetes based on the results of the first step screening test and who therefore require treatment and care	
Individuals Diagnosed with Disease	Individuals diagnosed with hypertension, diabetes, dyslipidemia or tuberculosis and who are currently receiving drug treatment	Diagnosed with diseases (4)

Table 13. Data on the diagnoses of general health screening in Seoul (2014).

Male	Level of diagnosis	N	%
1	Normal	55,891	4.97
2	Caution	369,425	32.84
3	Suspected disease	467,957	41.60
4	Diagnosed with diseases	231,599	20.59
	Total	1,124,872	

Female	Level of diagnosis	N	%
1	Normal	139,729	12.92
2	Caution	434,593	40.19
3	Suspected disease	302,015	27.93
4	Diagnosed with diseases	205,126	18.97
	Total	1,081,463	

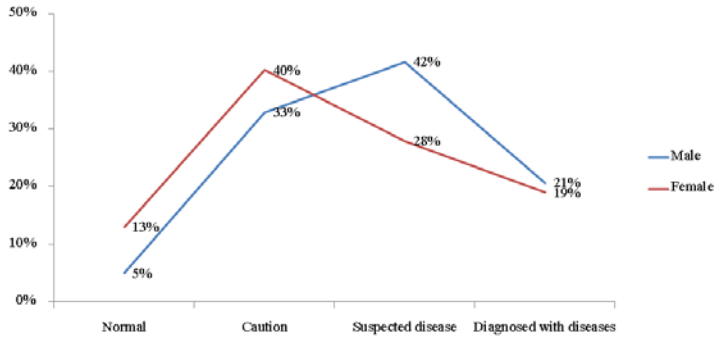


Figure 2. The distribution of the general health screening by gender in Seoul.

4.2 Results

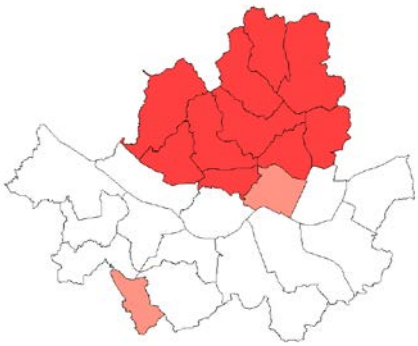
For the general health screening data set, two approaches, the original scan statistics and the proposed spatial scan statistics with a restricted likelihood ratio, are utilized. The maximum size of the scanning windows of each location in this study is set to include 50% of the total population of Seoul. Based on the null hypothesis $H_0 : p_k = q_k$, for all k = 1,…, 4 and all scanning window *z*, the LRO-based and STO-based alternative hypotheses can be defined as $H_a : \frac{p_1}{q_1} \leq \frac{p_2}{q_2} \leq \frac{p_3}{q_3} \leq \frac{p_4}{q_4}$ and $H_a : \sum_{k=1}^{4} p_k \leq \sum_{k=1}^{4} q_k$, respectively. We compare the results of original spatial scan statistics with that of the proposed spatial scan statistics using the value of $\alpha_1 = 0.10$. We evaluate the statistical significance for clusters via 9999 replications for Monte Carlo simulations at significance level $\alpha_0 = 0.05$.

Figure 3 shows the result map for spatial cluster detection on the level of first diagnoses on the general health screening in Seoul. In the case of male's diagnosis results, we identified that the original cluster detection methods detected larger clusters than our proposed methods, in particular, on the most likely cluster. Although the original scan statistic for ordinal data detected a large cluster in the north area, our method detected two or three small clusters except in "Jonglo-gu". It may be expected that the original method tends to detect larger clusters by

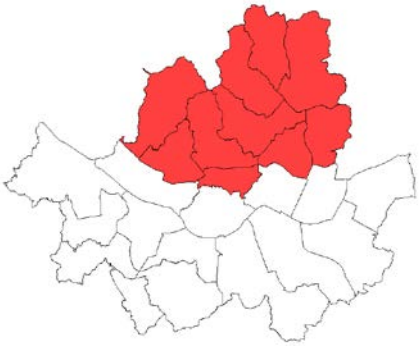
absorbing adjacent regions with irrelevant risks, “Jonglo-gu” in this case. There is no difference between the original and restricted method in the female data. Since the stochastic ordering hypothesis incorporates the likelihood ratio ordering hypothesis, the results in the STO-based approach detected more districts as a cluster than the results in the LRO-based method which, for example, did not detect the districts “Sungdong-gu”, “Kwangjin-gu”, and “Keumchun-gu”, as shown in Figure 3 (c) and (d). We represent the proportion of spatial clusters on general health screening for males and females in Figures 4 and 5. Due to the large cases in Seoul, there might be a slight difference between clusters and total population of Seoul. However, the patterns showed that spatial clusters almost have lower proportion in normal and caution and higher proportion in suspected disease and diagnosed with diseases than the total proportion in Seoul.

We illustrate the detailed information about all statistically significant clusters through the MC hypothesis testing in Tables 13 through 16. The most likely cluster that has the maximum likelihood ratio is the primary cluster level and the rest of the clusters are the secondary cluster level. The most likely cluster for the original method has ten districts based on “Gangbuk-gu”, while only six districts are belonging to the most likely cluster in spatial scan statistic with a restricted likelihood ratio in both Tables 13 and 14. “Keumchun-gu” and

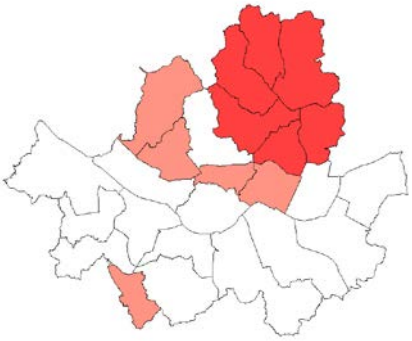
“Sungdong-gu” cannot be detected with the LRO-based alternative hypothesis in the health diagnoses for males. For females, “Dongjak-gu” also has stochastic ordering in its level of diagnosis, so that it cannot be detected using the LRO-based hypothesis.



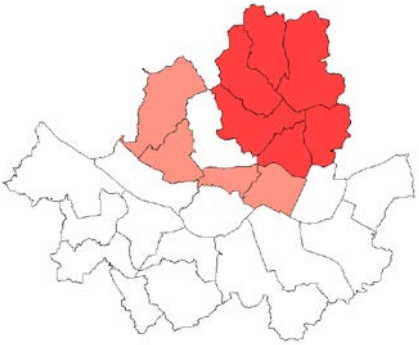
(a) STO cluster map in males



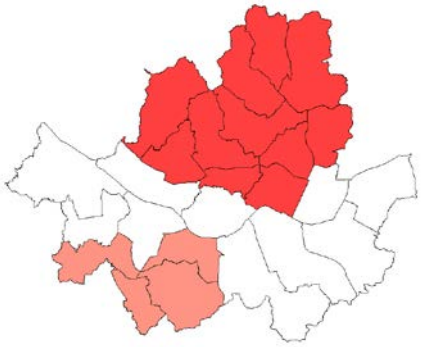
(b) LRO cluster map in males



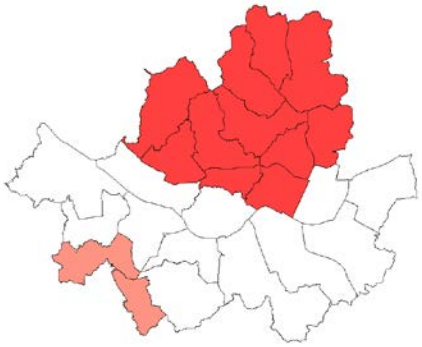
(c) Restricted STO cluster map in males



(d) Restricted LRO cluster map in males

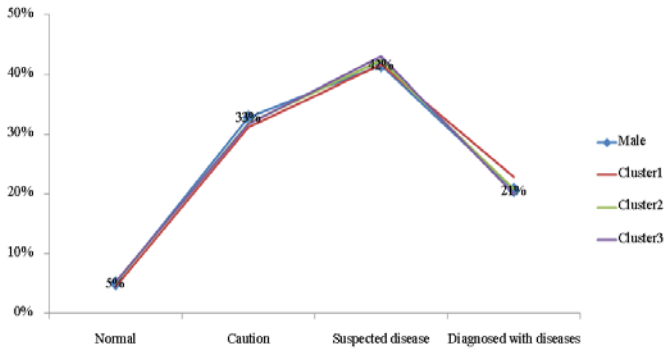


(e) STO cluster map in females

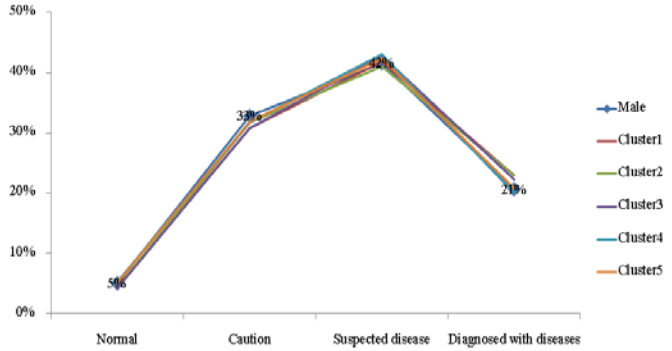


(f) LRO cluster map in females

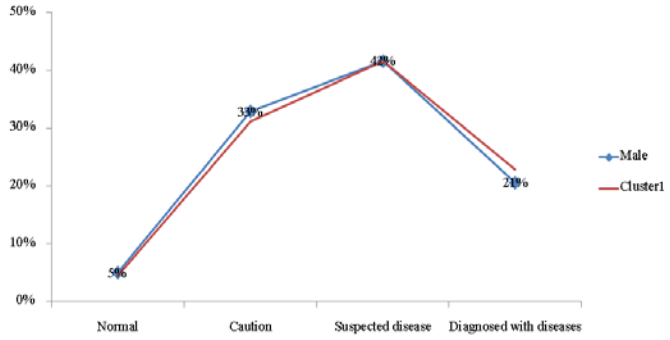
Figure 3. Spatial cluster detection results for the general health screening by gender in Seoul using original and restricted approaches in both STO-based and LRO-based methods.



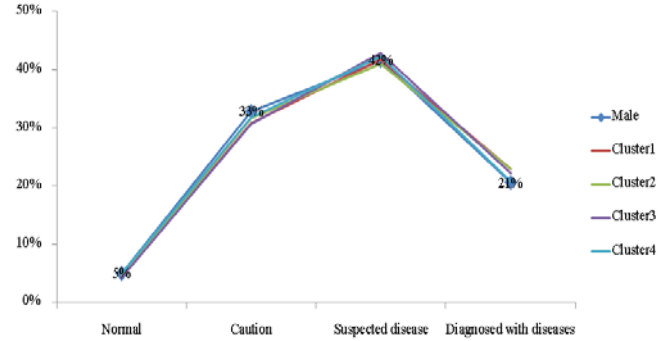
(a) STO-based results for males



(b) LRO-based results for males

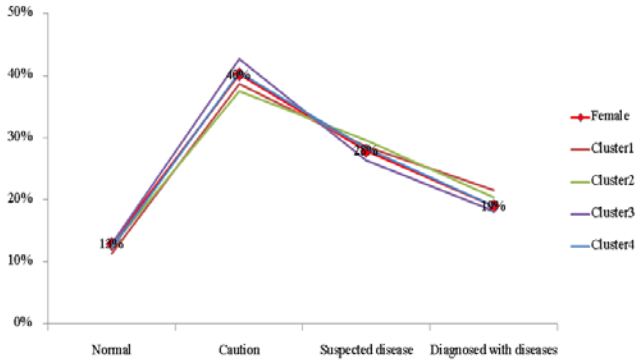


(c) Restricted STO-based results for males

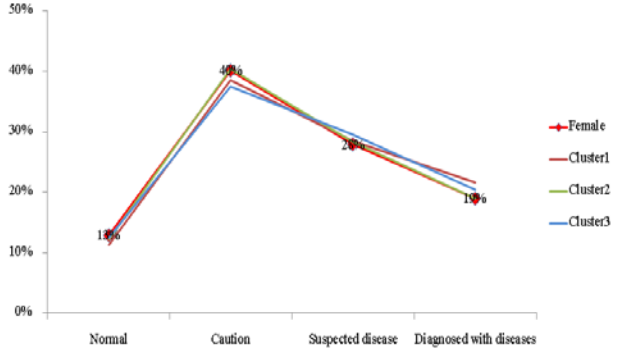


(d) Restricted LRO-based results for males

Figure 4. The comparison of the distribution between whole study area and clusters based on the results of spatial cluster detection for males.



(a) STO-based results for females



(b) LRO-based results for females

Figure 5. The comparison of the distribution between whole study area and clusters based on the results of spatial cluster detection for females.

Table 14. Spatial clusters of high rates of the diagnoses on the general health screening in Figure 3 (a) and (c).

	Cluster Level		Centroid (Gu)	No. of Districts	LLR	Total cases	Percent cases in area ([1],[2],[3],[4])
STO	Primary	1	Gangbuk	10	1239.14	399,525	(4.41,31.09,41.65,22.84)
	Secondary	2	Keumchun	1	13.18	33,399	(5.03,31.81,42.35,20.81)
		3	Sungdong	1	8.37	30,761	(5.24,31.75,43.03,19.99)
Restricted STO	Primary	1	Nowon	6	851.60	281,474	(4.45,30.80,41.79,22.96)
	Secondary	2	Eunpyung	2	188.19	89,643	(4.32,31.70,41.07,22.90)
		3	Joong	1	28.89	13,185	(4.19,30.73,42.81,22.26)
		4	Keumchun	1	13.18	30,761	(5.24,31.75,43.03,19.99)
		5	Sungdong	1	8.37	33,399	(5.03,31.81,42.35,20.81)

Table 15. Spatial clusters of high rates of the diagnoses on the general health screening in Figure 3 (b) and (d).

	Cluster Level		Centroid (Gu)	No. of Districts	LLR	Total cases	Percent cases in area ([1],[2],[3],[4])
LRO	Primary	1	Gangbuk	10	1239.14	399,525	(4.41,31.09,41.65,22.84)
Restricted LRO	Primary	1	Nowon	6	851.60	281,474	(4.45,30.80,41.79,22.96)
	Secondary	2	Eunpyung	2	188.19	89,643	(4.32,31.70,41.07,22.90)
		3	Joong	1	28.89	13,185	(4.19,30.73,42.81,22.26)
		4	Sungdong	1	8.37	33,399	(5.03,31.81,42.35,20.81)

Table 16. Spatial clusters of high rates of the diagnoses on the general health screening in Figure 3 (e).

	Cluster Level		Centroid (Gu)	No. of Districts	LLR	Total cases	Percent cases in area ([1],[2],[3],[4])
STO	Primary	1	Sungbuk	11	2174.73	421,400	(11.32,38.60,28.50,21.58)
	Secondary	2	keumchun	1	57.88	27,429	(12.51,37.46,29.58,20.45)
		3	Dongjak	2	26.01	10,648	(12.85,42.62,26.31,18.23)
		4	guro	1	10.56	49,116	(12.26,40.45,28.31,18.98)

Table 17. Spatial clusters of high rates of the diagnoses on the general health screening in Figure 3 (f).

	Cluster Level		Centroid (Gu)	No. of Districts	LLR	Total cases	Percent cases in area ([1],[2],[3],[4])
LRO	Primary	1	Sungbuk	11	2174.73	421,400	(11.32,38.60,28.50,21.58)
	Secondary	2	Keumchun	1	56.05	49,116	(12.26,40.45,28.31,18.98)
		3	Guro	1	10.05	27,429	(12.51,37.46,29.58,20.45)

5. Discussion and Conclusion

The purpose of this study was to propose modified spatial scan statistics for ordinal outcome data by considering the restricted likelihood ratio in order to resolve the undesirable phenomenon. Since the spatial scan statistic by Kulldorf tends to detect much larger clusters in a Poisson-based model, we suspected that the two spatial scan statistics for ordinal data would also have that tendency. According to Tango (2008), we applied a screening criterion to the spatial scan statistics on ordinal data and compared the performance our proposed method with the original ones.

There are several findings in the simulation studies. Similar patterns have been identified in all of the simulation results regardless of the different scenarios. As we supposed, the original spatial scan statistics tended to detect clusters larger than the true clusters on ordinal outcome data. Our proposed spatial scan statistics seemed to relieve that undesirable property; they have a good performance with a high value of PPV compared with the performance of the original method. Even though sensitivity seemed to be lower in our proposed approach, it can be solved by adjusting a screening value. Sensitivity and power can be higher when we have the appropriate screening value and this can be advantageous in our method. In other words, our proposed approach gives the researcher to capability of adjusting

the screening level of α_1 in accordance with the purpose of the research.

We used the general health screening data set in 2014 from the NHIS. To establish health care and health insurance policies, it is important to understand the geographical patterns about certain risky-areas compared to surrounding areas for improving health care in the local area, such as post-management of health checkups and disease prevention. Our findings can contribute to the development of the system for promoting the public health by detecting the spatial clusters which need prevention and intervention. For instance, by adjusting the screening level of α_1, our proposed method is able to help health planners decide an appropriate range of areas for their health care program.

In conclusion, the proposed spatial scan statistic with a restricted likelihood ratio for ordinal data demonstrates a better property in detecting the true cluster compared with the original method, and the screening value of α_1 can be useful for conducting an accurate cluster detection in accordance with the purpose of cluster detection.

However, some limitations are discussed in this study. We used the circular scanning window to conduct the cluster detection. Tango and Takahashi (2012) proposed a spatial scan statistic with a restricted likelihood ratio using the flexible

scanning window in the Poisson-based model and their proposed method had better performance than the circular spatial scan statistic. Nevertheless, we did not use the flexible scanning window due to a heavy computational load and we expect that similar results may be shown in terms of the comparison between the original and restricted spatial scan statistics. Moreover, there are some methodologies for more effectively cluster detection such as CLIC and the Gini coefficient (Han *et al.*, 2016). Further studies need to compare those methods with our method in cluster detection for ordinal data. These improvements could be considered for future research building on the findings of this study.

Reference

Alvarez, P. L. L., Guerra-Yi, M. E., Faes, C., Alvarez, Y. G., & Molenberghs, G. (2009). Spatial analysis of breast and cervical cancer incidence in small geographical areas in Cuba, 1999–2003. *European Journal of Cancer Prevention*, *18*(5), 395-403.

Amin, R., Hendryx, M., Shull, M., & Bohnert, A. (2014). A cluster analysis of pediatric cancer incidence rates in Florida: 2000–2010. *Statistics and Public Policy*, *1*(1), 69-77.

Barlow RE, Bartholomew DJ, Bremner JM, Brunk HD. (1972). *Statistical Inference Under Order Restrictions*: *The Theory and Applications of Isotonic Regression*. New York: Wiley.

Dykstra, R., Kochar, S., & Robertson, T. (1995). Inference for likelihood ratio ordering in the two-sample problem. *Journal of the American Statistical Association*, *90*(431), 1034-1040.

Dykstra, R. L., Lee, C. I. C., & Yan, X. (1996). Multinomial estimation procedures for two stochastically ordered distributions. *Statistics & probability letters*, *30*(4), 353-361.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, 181-187.

Fischer, E. A. J., Pahan, D., Chowdhury, S. K., Oskam, L., & Richardus, J. H. (2008). The spatial distribution of leprosy in four villages in Bangladesh: An observational study. *BMC infectious diseases*, *8*(1), 1.

Gregorio, D. I., Samociuk, H., DeChello, L., & Swede, H. (2006). Effects of study area size on geographic characterizations of health events: prostate cancer incidence in Southern New England, USA, 1994–1998. *International Journal of Health Geographics*, *5*(1), 1.

Han, J., Zhu, L., Kulldorff, M., Hostovich, S., Stinchcomb, D. G., Tatalovich, Z., ... & Feuer, E. J. (2016). Using Gini coefficient to determining optimal cluster reporting sizes for spatial scan statistics. *International Journal of Health Geographics*, *15*(1), 27.

Huang, L., Kulldorff, M., & Gregorio, D. (2007). A spatial scan statistic for survival data. *Biometrics*, *63*(1), 109-118.

Jung, I., Kulldorff, M., & Klassen, A. C. (2007). A spatial scan statistic for ordinal data. *Statistics in medicine*, *26*(7), 1594-1607.

Jung, I., Kulldorff, M., & Richard, O. J. (2010). A spatial scan statistic for multinomial data. *Statistics in medicine*, *29*(18), 1910-1918.

Korean Statistical Information Service. (2014). National Health Insurance Corporation.(http://kosis.kr/statHtml/statHtml.do?orgId=350&tblId=DT_35007_N098&conn_path=I2).

Kulldorff, M., & Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in medicine*, *14*(8), 799-810.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods*, *26*(6), 1481-1496.

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*(1), 1.

Michelozzi, P., Capon, A., Kirchmayer, U., Forastiere, F., Biggeri, A., Barca, A., & Perucci, C. A. (2002). Adult and childhood leukemia near a high-power radio station in Rome, Italy. *American Journal of Epidemiology*, *155*(12), 1096-1103.

National Health Insurance Service. (2014). The 2014 National Health Screening Statistical Yearbook. (http://www.nhis.or.kr/menu/boardRetriveMenuSet.xx?menuId=F3328)

Ozdenerol, E., Williams, B. L., Kang, S. Y., & Magsumbol, M. S. (2005). Comparison of spatial scan statistic and spatial filtering in estimating low birth weight clusters. *International Journal of Health Geographics*, *4*(1), 1.

Robertson, T., & Wright, F. T. (1981). Likelihood ratio tests for and against a stochastic ordering between multinomial populations. *The Annals of Statistics*, 1248-1257.

Tango, T. (2000). A test for spatial disease clustering adjusted for multiple testing. *Statistics in medicine*, *19*(2), 191-204.

Tango, T., & Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics*, *4*(1), 1.

Tango, T. (2007). A spatial scan statistic scanning only the regions with elevated risk. *Advances in Disease Surveillance*, *4*, 117.

Tango, T. (2008). A spatial scan statistic with a restricted likelihood ratio. *Japanese Journal of Biometrics*, *29*(2), 75-95.

국 문 요 약

순서형 자료에서의 제한된 우도비를 이용한 공간검색 통계량 연구

공간검색통계량(spatial scan statistic)은 우도비 검정을 기반으로 특정 사건에 대한 분포가 다른 지역의 분포와 통계적으로 유의하게 다른 공간군집(spatial cluster)을 탐색하는 방법으로 여러 분야에서 이용되고 있다. 이 방법은 연구자가 사전에 각 지역의 중심점을 기준으로 형성되는 후보 군집(scanning window)의 모양과 최대 군집 크기를 설정한다. 후보 군집의 모양은 원형, 타원형, 비정형이 널리 사용되고, 최대 군집 크기는 보통 전체 인구의 50%로 설정한다.

Kulldorff (1997)에 의해 제안된 공간검색통계량이 군집 탐색을 위한 방법으로 널리 쓰이나, 이 방법이 실제 군집보다 더 넓은 범위의 군집을 도출한다는 것이 Tango (2007)에 의해 알려졌다. Tango (2008)는 모의실험을 통하여 포아송 기반의 공간검색통계량이 실제 군집 주변의 유의하지 않는 지역들을 흡수함으로써 더 넓은 지역을 군집으로 도출한다는 사실을 보였고, 이에 대한 해결책으로 포아송 기반의 공간검색 통계량에 제한된 우도비를 적용함으로써 유의하지 않는 지역들을 사전에 제거하여 관심대상의 지역들 만으로 군집을 도출하는 방법을 제안하였다. Tango (2008)가 제안한 방법이 기존의 방법보다 실제 군집을 비교적 더 정확히 찾아냄을 모의 실험을 통해 보였다.

한편 순서형 자료는 질병의 진행단계와 같은 순위 범주를 가지는 자료로 의학 분야

에서 빈번히 나타난다. 이러한 자료를 위한 공간검색통계량은 대립가설에 따라 두 가지 방법이 Jung *et al.* (2007)과 Jung and Lee (2011)에 의해 제안된 바가 있으며, 본 연구에서는 이 공간검색통계량들 또한 위와 같은 현상을 보일 것이라 예상한다. 따라서 본 연구에서는 순서형 자료를 위한 공간검색통계량에 제한된 우도비를 적용하는 방안을 제안하고, 모의실험을 통하여 기존의 방법과 비교 및 평가해 보고자 한다.

그 결과, 순서형 자료를 위한 기존의 공간검색통계량이 우리의 예상과 같이 실제 군집보다 더 넓은 지역의 군집을 도출한다는 것이 발견되었고, 제한된 우도비를 적용한 공간검색통계량 방법이 이러한 점을 어느 정도 잘 해결함을 알 수 있었다. 또한 제안된 방법을 실제 데이터에 적용함으로써 본 방법의 필요성을 제안하였다.

핵심되는 말: 공간검색통계량, 순서형 자료, 군집 탐색, 제한된 우도비