Optimizing maximum window size in spatial scan statistic for ordinal data

Sehwi Kim

The Graduate School
Yonsei University
Department of Biostatistics and Computing

Optimizing maximum window size in a spatial scan statistic for ordinal data

A Masters 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

Sehwi Kim

December 2015

This certifies that the masters thesis of Sehwi Kim is approved.

_____

Thesis Supervisor: Prof. Inkyung Jung

_____

Prof. Chung Mo Nam

_____

Prof. Sohee Park

The Graduate School
Yonsei University
December 2015

Contents

List of Figures

List of Tables

Abstract

Spatial scan statistics are widely used in spatial epidemiology to identify areas with high or low rates of outcome. This scan-based method needs a scanning window, which is defined by its shape and maximum size. When deciding on the upper limit of the window size, 50% of the total population is often used. However, there is no rationale and the reported clusters could be too larger than the true ones.

Recently, Han et al. (2011) proposed using the Gini coefficient as a measure to assess the degree of heterogeneity of the cluster models. They also considered another measure called the Cluster Information Criterion (CLIC) similar to Akaike's Information Criterion (AIC). The two measures were evaluated for the Poisson model only and applicability to other models has not been proved.

In this study, we adapt the two measures applicable to the ordinal model proposed by Jung, Kulldorff, and Klassen (2007). Through a simulation study and real data examples, we show that the two measures give consistent results except when the true clusters are irregular-shaped or located slightly apart from each other. In these cases, the Gini coefficient picks a smaller window size as an

optimal maximum than CLIC. In doing so, it reflects a tendency to detect the clusters that are more close to true ones by detecting a set of several small clusters.

The results of this study demonstrate the necessity of optimizing the maximum window size in spatial scan statistic for ordinal data as well as for the Poisson model. Further, we believe that the two measures can be useful to optimize the maximum scanning window size in spatial scan statistic for ordinal data.

Key words: Spatial scan statistic; ordinal data; maximum window size; Gini coefficient; Cluster Information Criterion

1. Introduction

The spatial scan statistic based on the likelihood ratio test has been widely used in many fields, such as epidemiology and disease surveillance. The purpose of this method is to detect any statistically significant spatial cluster where the distribution of events (e.g., disease prevalence, incidence, and mortality) differs from that of other regions.

In this process, the candidate areas (scanning windows) are created at the centroids across the study region in varying pre-defined shapes and sizes. Numerous studies have been made on comparing methods for scan-based cluster detection methods in different shapes (e.g., circular, elliptic) (Goujon-Bellec et al., 2011; Grubesic, Wei, and Murray, 2014; Huang, Pickle, and Das, 2008). On the other hand, the subject of scanning window sizes has received relatively less attention.

The scanning window size is usually set to a maximum 50% of the total population, as in the case of many researches. However, it may draw an exaggerated conclusion. That is, with a larger scanning window size, the most likely cluster will potentially include several secondary clusters and less informative areas. Furthermore, Ribeiro and Coasta (2012) have mentioned that cluster detection results can be sensitive to the maximum size.

Recently, Han et al. (2011) suggested using the Gini coefficient and Cluster Information Criterion (CLIC) to determine the optimal maximum window size. However, this research only evaluated for a Poisson model and the applicability of other probability models has not been proven yet.

In this paper, we propose the application of the two measures for an ordinal model put forward by Jung, Kulldorff, and Klassen (2007). In chapter 2, we briefly review the spatial scan statistic for count and ordinal data and provide descriptions of the Gini coefficient and Cluster Information Criterion (CLIC) required for optimizing maximum window size for count data. From there, the application of two optimization criteria for ordinal data is proposed and in chapters 3 and 4, the performance of the criteria is evaluated via simulation study and real data examples. We discuss our findings and present the conclusion in chapter 5.

2. Method

2.1 Spatial scan statistic

The spatial scan statistic based on the likelihood ratio test is one of the cluster detection methods. Starting with the Bernoulli and Poisson model (Kulldorff, 1997), methods for various models have been developed, such as ordinal, exponential, multinomial, and normal (Cook, Gold, and Li, 2007; Huang, Kulldorff, and Gregorio, 2007; Jung, Kulldorff, and Klassen, 2007; Jung, Kulldorff, and Richard, 2010; Kulldorff, Huang, and Konty, 2009). These are used to detect any statistically significant spatial cluster where the distribution of event (e.g., disease prevalence, incidence, and mortality) differs from that of other regions. For each centroid of the study region, the candidate areas (scanning window Z) are formed as pre-defined shapes with maximum window size, over which likelihood ratio test statistics are calculated. The candidate area with the maximum likelihood defines the most likely cluster and those that are able to reject the null hypothesis on their own strength define the secondary clusters. This process is represented as follows:

$$\lambda = \frac{max_{Z,H_a} L(Z,\theta)}{max_{Z,H_0} L(Z,\theta)} = \frac{max_Z L(Z,\hat{\theta})}{L(\widehat{\theta_0})}$$

In most cases, the maximum cluster size is selected to be less than, or equal to, 50% of the total population with circular or elliptic shape. The spatial scan statistics for several models with these two shapes can be implemented using the SaTScan (www.satscan.org). The elliptic version of the spatial scan statistic uses the elliptic-shaped scanning window with three options (shapes, angles, and non-compactness) (Kulldorff et al., 2006). The shape of the ellipse is defined by the ratio of the longest to the shortest axis of the ellipse. The default values of shapes provided by SaTScan software are 1 (= circle), 1.5, 2, 3, 4 or 5. Each shape has the angle between the horizontal line and the semi-major axis of the ellipse (4, 6, 9, 12, and 15). Further, we can customize the option for non-compactness penalty in the form of $[4s/(s+1)^2]^a$, where s is the shape parameter and a is the non-compactness penalty parameter (a $= 1$: strong penalty; a $= 1/2$: medium penalty (default); a $= 0$: no penalty). It multiplies the log likelihood ratio and the ellipses with the larger penalty are better-fitted for compact clusters.

2.1.1 Scan statistic for count data

The scan statistic for count data assumed a Poisson distribution. The null hypothesis is that the incidence rates of events are the same within and outside the

scanning window, while the alternative hypothesis is that the rate inside the scanning window is higher (or lower) than outside:

$$H_0 : p = q \ \ vs. \ H_a : p > q \ (or \ p < q)$$

where p is the incidence rate of events within scanning window Z and q is the incidence rate outside it. If c_Z and n_Z represent the number of cases and populations in scanning window Z, then $\mathrm{C} = \sum_z c_Z$ and $\mathrm{N} = \sum_z n_Z$ will be the total number of cases and populations in the study area. The likelihood ratio test statistic with scanning window Z in a Poisson model (Kulldorff, 1997) is given by

$$\lambda_Z = \frac{\left(\frac{c_Z}{n_Z}\right)^{c_Z}\left(\frac{C - c_Z}{N - n_Z}\right)^{C - c_Z}}{\left(\frac{C}{N}\right)^{C}} I\left(\frac{c_Z}{n_Z} > \frac{C - c_Z}{N - n_Z}\right)$$

and Z with the maximum likelihood ratio test statistic being the most likely cluster. If the cluster area has a lower incidence rate of events, the indicator function is replaced by $I\left(\frac{c_Z}{n_Z} < \frac{C - c_Z}{N - n_Z}\right)$.

2.1.2 Scan statistic for ordinal data

Jung, Kulldorff, and Klassen (2007) proposed the spatial scan statistic for ordinal data, such as education level and cancer stage. In the ordinal model, an alternative hypothesis for detecting clusters with high rates of higher-valued category has an order restriction called the likelihood ratio ordering (LRO) (Dykstra, Kochar, and Robertson, 1995).

If an ordinal variable has K categories ($k = 1, \dots, K$), the probability of being in k of inside and outside the scanning window Z denote p_k and q_k, respectively. The likelihood ratio test statistics in sub-regions i $(i = 1, \dots, I)$ for testing $H_0: p_1 = q_1, \dots, p_K = q_K$ against $H_a: \frac{p_1}{q_1} \leq \frac{p_2}{q_2} \leq \cdots \leq \frac{p_K}{q_K}$ is

$$\lambda = \frac{max_{Z,H_a} L(Z, p_1, \dots, p_K, q_1, \dots, q_K)}{max_{Z,H_0} L(Z, p_1, \dots, p_K, q_1, \dots, q_K)} = \frac{max_Z L(Z)}{L_0}$$

with

$$\mathrm{L(Z)} = \prod_k \left(\prod_{i \in Z} \hat{p}_k{}^{c_{ik}} \prod_{i \notin Z} \hat{q}_k{}^{c_{ik}} \right)$$

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

where c_{ik} is the number of observations in region i and category k, $\hat{p}_{0k} = C_k/C$ is the maximum likelihood estimator (MLE) of p_k under the null hypothesis, and $\hat{p}_k$ and $\hat{q}_k$ are MLEs of p_k and q_k under the alternative hypothesis. The MLEs of p_k and q_k are expressed by Dykstra, Kochar, and Robertson (1995) as follows:

$$\hat{p}_k = \left(\frac{W_k + U_k}{W}\right) E_{(W+U)}\left(\frac{W}{W+U}\middle|\Gamma\right)_k = \tilde{p}_k \left(\frac{W_k + U_k}{W_k}\right) E_{(W+U)}\left(\frac{W}{W+U}\middle|\Gamma\right)_k$$

$$\hat{q}_k = \left(\frac{W_k + U_k}{U}\right) E_{(W+U)}\left(\frac{U}{W+U}\middle|\mathrm{A}\right)_k = \tilde{q}_k \left(\frac{W_k + U_k}{U_k}\right) E_{(W+U)}\left(\frac{U}{W+U}\middle|\mathrm{A}\right)_k$$

where $W_k = \sum_{i\in Z} c_{ik}$, $U_k = \sum_{i\notin Z} c_{ik}$, $W = \sum_k W_k$, and $\mathrm{U} = \sum_k U_k$. Total number of observations in category k $(= C_k)$ is the sum of W_k and U_k; thus, the total number of observations in study region $\mathrm{C} = \mathrm{W} + \mathrm{U}$. Each isotonic regression on $\Gamma = \{(\theta_1, \dots, \theta_k); \theta_1 \le \cdots \le \theta_k\}$ or $\mathrm{A} = \{(\theta_1, \dots, \theta_k); \theta_1 \ge \cdots \ge \theta_k\}$ with $\theta_k = W_{p_k}/(W_{p_k} + U_{q_k})$. When the ratio of the unrestricted MLEs $\tilde{p}_k/\tilde{q}_k$ is non-decreasing for all $k\ (= 1, \dots, K)$, $\tilde{p}_k$ and $\tilde{q}_k$ are the MLEs under the H_a. If $\tilde{p}_k/\tilde{q}_k$ is not satisfied with $H_a: \frac{p_1}{q_1} \le \frac{p_2}{q_2} \le \cdots \le \frac{p_K}{q_K}$, the 'Pool-Adjacent-Violators' algorithm (Brunk et al. 1972) works to update $\hat{p}_k$ and $\hat{q}_k$ until $\hat{p}_k/\hat{q}_k$ does not decrease. Thereafter, the final updated estimates are MLEs of p_k and q_k.

2.2 Optimizing maximum window size for count data

The scanning window size is one of the parameters that should be selected by the researcher in cluster detection. Once this is determined, the result reports a cluster of closer size to maximum window size rather than smaller sub regions. With a larger scanning window size, the most likely cluster has the potential to exaggerate the conclusion. The cluster formed from combinations of small clusters in close could have the largest likelihood ratio test statistic, although it includes some areas with few events.

To optimize maximum window size, Han et al. (2011) proposed two measures for count data: the Gini coefficient and the Cluster Information Criterion (CLIC). Each criterion offers optimal cluster size for detecting a collection of non-overlapping clusters.

2.2.1 Gini coefficient

The Gini coefficient is a measurement of income distribution inequality developed by Gini (1912). It is based on the Lorenz curve, which consists of the percentage of population (x-axis) and the proportion of the total income of the bottom $x\%$ of the population (y-axis). Using a 45 degree line ($y = x$) to denote

perfect income distribution equality, the ratio of the area between the line of equality and Lorenz curve (A) to the area under the 45 degree line ($A + B$) is defined as the Gini coefficient (Figure 1). It ranges from 0 to 1, with a value of 0 and 1 corresponding to complete equality and inequality, respectively. Higher values indicate a higher income distribution disparity.

Han et al. (2011) applied this concept to describe the distribution of events (e.g., death from cancer). In Figure 1, x-axis is the same as above, with the cumulative percentage of event plotted along the y-axis. The line at 45 degrees means that the events are randomly distributed–that is, that the number of events is proportional to the population of each region. If there is a significant cluster in the study region, it means that the distribution of events is not random but a biased state.

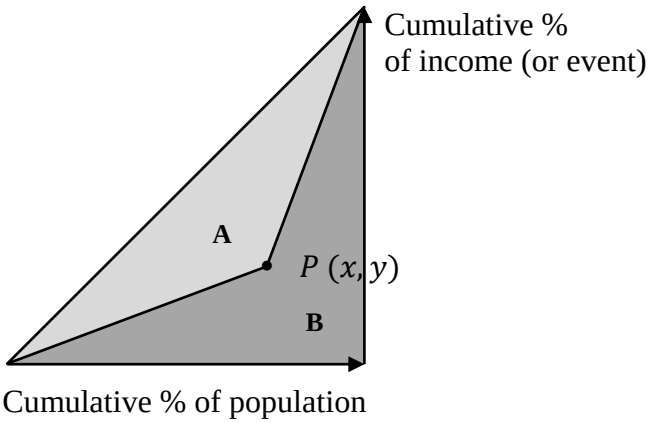


Figure 1. Graphical representation of the Gini coefficient

2.2.2 Cluster Information Criterion (CLIC)

The scan-based cluster detection methods are a likelihood-based test. Therefore, Han et al. (2011) also proposed the Cluster Information Criterion (CLIC) similar to the Akaike's Information Criterion (AIC) (Akaike, 1974). AIC as the good criterion of model selection is

$$\text{AIC} = -2(log\ likelihood) + 2(number\ of\ parameters\ in\ the\ model)$$

It consists of two terms; one represents the goodness-of-fit and the other functions as a penalty for a model having too many parameters. The model with the minimum AIC value is the better model.

Similarly, the CLIC for the Poisson model M is defined as

$$\text{CLIC}(M) = -2\sum_{i=1}^{m} LLR(z_i) + mlog(p)$$

where $z_1, z_2, \dots, z_m$ represent the significant clusters (zones) in cluster model M, m the number of significant clusters in the model and p the total population in those clusters. The sum of the log likelihood ratio of the significant clusters in the model represents the goodness-of-fit. $mlog(p)$ is regarded as the penalty term, reflecting that the number of significant clusters (m) and the total population in those clusters (p) have an inverse relationship. In other words, p decreases as m

increases (more significant clusters) and p increases as m decreases (less significant clusters). Like AIC, the model with the lowest CLIC is preferred.

2.3 Optimizing maximum window size for ordinal data

Two measures, the Gini coefficient and Cluster Information Criterion (CLIC) proposed by Han et al. (2011), are evaluated for the Poisson model only. Here we adapt the two criteria applicable to the spatial scan statistic for ordinal model proposed by Jung, Kulldorff, and Klassen (2007).

2.3.1 Gini coefficient

In ordinal data, the Lorenz curve represents the distribution of higher order categories according to cumulative percentages of total cases. Therefore, if a detected cluster is significant, there are areas with high rates of higher-valued categories than others.

Here we consider that there is only one significant cluster z^* in the model. The x-coordinate of point $P(x, y)$ in Figure 1 is defined as:

$$1 - \frac{\sum_k \sum_{i \in z^*} c_{ik}}{\sum_k C_k}$$

To define the y-coordinate of $P(x, y)$, we need to consider how to weight on each category k with the pool-adjacent-violators algorithm for the satisfaction of the order restriction called LRO (Jung, Kulldorff, and Klassen, 2007). We can write down the y-coordinate of $P(x, y)$ as

$$1 - \frac{\sum_k k(\hat{p}_k \sum_k \sum_{i \epsilon z^*} c_{ik})}{\sum_k (kC_k)}$$

Here, we assigned ordinal scores for the cases per category to reflect the order of categories. This idea is from the method of goodness-of-fit in ordinal response regression models (Lipsitz, Fitzmaurice, and Molenberghs, 1996). Plus, in numerator, to consider the case of the combined categories using the algorithm, we give a weighting on the total number of observations in the significant cluster z^* multiplied by MLE of p_k under the alternative hypothesis.

If more than two significant clusters exist, we can calculate each coordinates by cumulatively subtracting from 1.

2.3.2 Cluster Information Criterion (CLIC)

Let $z_1, z_2, \dots, z_m$ be significant clusters (zones) in cluster model O, m the number of significant clusters in the model and c_{ik} the number of observations in location i and category k, then the CLIC for ordinal data can be expressed as

$$\mathrm{CLIC}(O) = -2\sum_{i=1}^{m} LLR(z_i) + mlog\left(\sum_{i=1}^{m}\sum_{k=1}^{K} c_{ik}\right)$$

The only difference from the CLIC for Poisson distribution is that p is replaced by $\sum_{i=1}^{m}\sum_{k=1}^{K} c_{ik}$ as a penalty term. It considers that the number of total observations in the significant clusters are an element of complexity in the ordinal model as the total population (p) in $z_1, z_2, \dots, z_m$ in the Poisson model.

3. Simulation study

3.1 Simulation setting

In order to evaluate the performance of the two criteria in ordinal data, we conducted simulation studies using several cluster models in 25 districts (gu) of Seoul, Korea.

In the first cluster model, we set 2,000 cases in the whole study region. A true cluster with high rates of higher-valued categories comprises three regions (Seocho-gu, Gangnam-gu, and Songpa-gu) with 200, 400, and 800 cases (see Figure 2). We assumed $H_0: p = q = (0.25, 0.25, 0.25, 0.25)$ against five different alternative hypotheses meeting the LRO:

Scenario A: $p = (0.10, 0.30, 0.30, 0.30)$
Scenario B: $p = (0.20, 0.20, 0.30, 0.30)$
Scenario C: $p = (0.20, 0.20, 0.20, 0.40)$
Scenario D: $p = (0.15, 0.25, 0.25, 0.35)$
Scenario E: $p = (0.15, 0.20, 0.25, 0.40)$

For the 15 situations, we generated 1,000 random data sets and searched for the clusters with high rates of high-valued categories using the circular SaTScan.

Then, we calculated the value of the two criteria for each candidate of maximum size (1, 2, 3, 4, 5, 6, 8, 10, 12, 15, 20, 25, 30, 35, 40, 45 and 50% of the total cases) and reported the frequency of optimal maximum window size chosen by the Gini (highest value) and CLIC (lowest value) among 1,000 random data sets.

We also estimated sensitivity and positive predicted value (PPV) to evaluate the accuracy of the proposed criteria for ordinal data. The sensitivity and PPV are defined in each upper limit. In the case of the significant data sets at the $\alpha = 0.05$ level, sensitivity is the proportion of districts detected correctly among the districts in the true cluster, and PPV the proportion of districts detected correctly among the districts in the detected cluster. Larger values of these measures indicate that the result with the upper limit is more precise in detecting the true cluster.

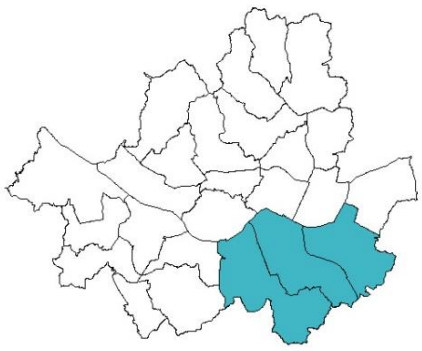
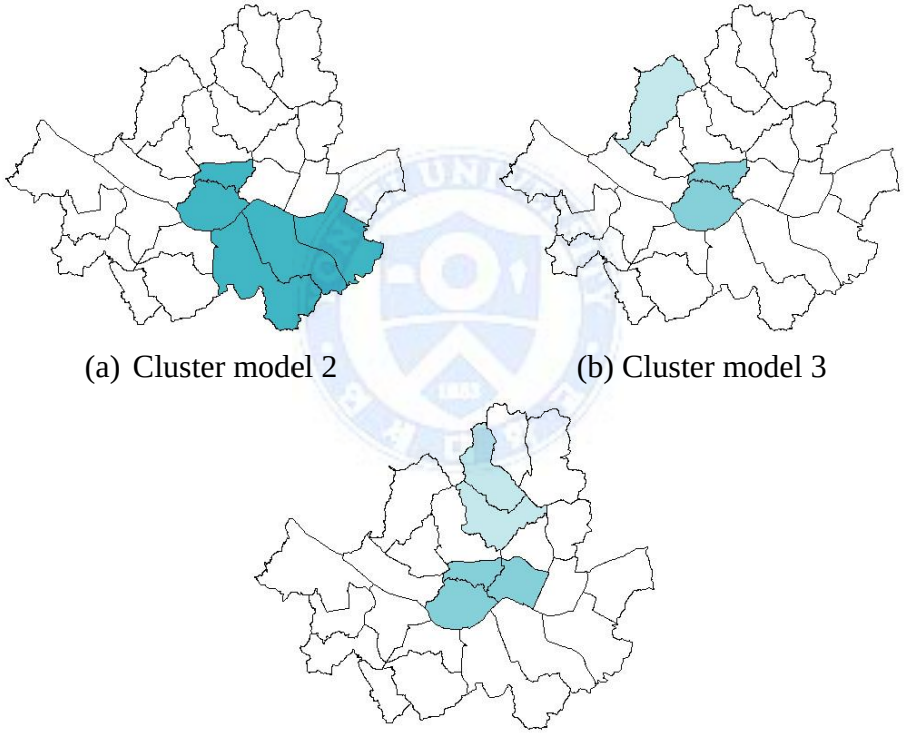


Figure 2. Study region for simulated cluster model 1

In addition, we considered three models with $H_a : p = (0.15, 0.20, 0.25, 0.40)$ and 10,000 cases using both circular and elliptic shapes with default options for the shape, angle, and non-compactness parameter (medium penalty). Figure 3 and Table 1 show the details of these cluster models.



(a) Cluster model 2 (b) Cluster model 3

(c) Cluster model 4

Figure 3. Study region for simulated cluster model 2, 3, 4

Table 1. Simulated cluster model 2, 3, 4

Cluster model	Number of clusters	Number of districts	Number of cases in clusters
2	1	5	2000, 3000, 4000
3	2	1 / 2	1500 / 2000
4	2	2 / 3	1500 / 2000

3.2 Results

Tables 2–4 show the results of the first simulated cluster model. For each scenario and criterion, the cells most chosen as the optimal maximum window size are shaded in gray (the same in Tables 5–8). The most picked upper limits are the same for both Gini coefficient and CLIC. In most cases, the best upper limit is achieved in each percentage of cases in true cluster. In addition, the sensitivity and PPV are high results in the best upper limit category in each except for scenario B with 10% of the total cases. This is because the ordinal model for $H_a: p = (0.20, 0.20, 0.30, 0.30)$ attains the lower power, sensitivity, and PPV than others according to Jung, Kulldorff, and Klassen (2007) as well as a small number of cases in true cluster. The sensitivity trend toward decrease was observed when the upper limits are lower than the best upper limit. Conversely, PPV tends to decrease when the upper limits are higher than the best upper limit. Compared to

the true cluster, the results imply that the significant clusters detected the small area with the lower upper limit and large area with the higher upper limit.

The results of the simulated cluster model 2 are listed in Tables 5–6. The study region of the second model is irregularly shaped as shown in Figure 3a. As a result, CLIC tends to pick a higher upper limit (close to the percentage of total cases) than the Gini coefficient. The SaTScan outputs show that the districts in true clusters are detected separately when using the scanning window size chosen by the Gini coefficient. With CLIC criteria, we identified only one (exactly the same as the true cluster) or two (making up the true cluster) significant clusters. Further, we found that the results of the CLIC using the elliptic shape are closer to the percentage of the total cases, in comparison to using the circular shape.

In the case of cluster model 3 and 4, the Gini and CLIC generally picked the upper limit being similar to 15% or 20% of the total cases as shown in Tables 7 and 8. However, when using the elliptic shape, the most chosen upper limit based on CLIC was off from our estimate. According to the SaTScan output, there were some districts in-between the true clusters.

Table 2. Simulation results of cluster model 1 (10% of the total cases in study region)

	Upper limit																
	1	2	3	4	5	6	8	10	12	15	20	25	30	35	40	45	50
Scenario A																	
Gini																	
# of OMWS	0	0	3	0	36	40	109	729	5	36	24	15	2	1	0	0	0
Sensitivity	-	-	0.667	-	0.648	0.642	0.765	0.998	0.800	0.991	0.986	1.000	1.000	1.000	-	-	-
PPV	-	-	1.000	-	1.000	0.979	0.980	0.997	0.650	0.748	0.580	0.495	0.429	0.375	-	-	-
CLIC																	
# of OMWS	0	0	7	0	6	8	138	771	5	31	21	11	1	1	0	0	0
Sensitivity	-	-	0.571	-	0.444	0.583	0.676	0.996	0.800	0.978	0.984	1.000	1.000	1.000	-	-	-
PPV	-	-	1.000	-	1.000	1.000	0.989	0.996	0.650	0.745	0.582	0.494	0.429	0.375	-	-	-
Scenario B																	
Gini																	
# of OMWS	0	26	120	17	16	15	31	2	19	20	23	16	9	11	8	11	12
Sensitivity	-	0.167	0.325	0.098	0.229	0.333	0.645	0.000	0.614	0.700	0.739	0.917	1.000	0.788	0.917	0.879	0.889
PPV	-	0.500	0.942	0.294	0.656	1.000	0.946	0.000	0.610	0.567	0.489	0.463	0.417	0.277	0.278	0.245	0.217
CLIC																	
# of OMWS	0	26	122	17	16	16	32	3	17	19	24	17	9	10	9	10	9
Sensitivity	-	0.167	0.325	0.098	0.229	0.333	0.646	0.111	0.608	0.702	0.750	0.922	1.000	0.767	0.889	0.900	0.852
PPV	-	0.500	0.939	0.294	0.656	1.000	0.948	0.111	0.603	0.570	0.494	0.465	0.417	0.271	0.269	0.252	0.209

Scenario C																	
Gini																	
# of OMWS	0	0	9	3	11	14	180	598	13	60	60	28	11	5	3	2	3
Sensitivity	-	-	0.407	0.333	0.515	0.357	0.670	0.997	0.692	0.944	0.967	1.000	1.000	1.000	1.000	0.833	1.000
PPV	-	-	1.000	0.667	1.000	0.964	0.990	0.997	0.633	0.728	0.592	0.492	0.424	0.350	0.300	0.220	0.237
CLIC																	
# of OMWS	0	0	26	3	7	16	198	600	9	55	51	19	9	3	3	0	1
Sensitivity	-	-	0.436	0.333	0.333	0.333	0.670	0.997	0.704	0.945	0.967	1.000	1.000	1.000	1.000	-	1.000
PPV	-	-	1.000	0.833	0.929	1.000	0.987	0.998	0.648	0.727	0.592	0.489	0.423	0.361	0.300	-	0.231
Scenario D																	
Gini																	
# of OMWS	0	0	23	3	14	15	214	450	26	78	71	45	24	13	8	8	8
Sensitivity	-	-	0.333	0.333	0.429	0.378	0.667	0.993	0.667	0.932	0.953	0.993	0.986	1.000	1.000	0.958	0.958
PPV	-	-	1.000	0.667	1.000	0.967	0.988	0.997	0.656	0.708	0.592	0.483	0.416	0.349	0.308	0.268	0.237
CLIC																	
# of OMWS	0	0	45	8	14	21	252	426	29	72	60	34	15	8	7	5	4
Sensitivity	-	-	0.400	0.417	0.381	0.365	0.667	0.993	0.667	0.926	0.939	0.990	0.978	1.000	1.000	1.000	1.000
PPV	-	-	0.989	0.813	1.000	0.976	0.990	0.997	0.657	0.702	0.590	0.481	0.408	0.354	0.305	0.278	0.245
Scenario E																	
Gini																	
# of OMWS	0	0	1	0	29	25	130	684	9	47	40	18	5	5	6	1	0
Sensitivity	-	-	0.333	-	0.586	0.573	0.705	0.995	0.667	0.950	1.000	1.000	1.000	1.000	1.000	1.000	-
PPV	-	-	1.000	-	0.954	1.000	0.996	0.997	0.667	0.734	0.601	0.492	0.429	0.367	0.311	0.273	-
CLIC																	
# of OMWS	0	1	5	1	12	17	170	693	9	35	35	12	4	3	2	1	0
Sensitivity	-	0.333	0.600	0.667	0.472	0.529	0.671	0.995	0.667	0.962	1.000	1.000	1.000	1.000	1.000	1.000	-
PPV	-	1.000	0.933	1.000	0.958	1.000	0.995	0.997	0.667	0.740	0.601	0.500	0.429	0.375	0.317	0.273	-

Table 3. Simulation results of cluster model 1 (20% of the total cases in study region)

	Upper limit																
	1	2	3	4	5	6	8	10	12	15	20	25	30	35	40	45	50
Scenario A																	
Gini																	
# of OMWS	0	0	0	0	0	0	0	97	76	214	549	36	22	5	1	0	0
Sensitivity	-	-	-	-	-	-	-	0.924	0.904	0.992	0.995	1.000	1.000	1.000	1.000	-	-
PPV	-	-	-	-	-	-	-	0.989	0.964	0.983	1.000	0.750	0.595	0.486	0.429	-	-
CLIC																	
# of OMWS	0	0	0	0	0	0	0	3	0	17	912	37	25	5	1	0	0
Sensitivity	-	-	-	-	-	-	-	0.556	-	0.706	0.991	1.000	1.000	1.000	1.000	-	-
PPV	-	-	-	-	-	-	-	1.000	-	1.000	0.997	0.750	0.596	0.486	0.429	-	-
Scenario B																	
Gini																	
# of OMWS	0	2	4	2	17	21	0	18	8	88	321	88	76	86	38	23	28
Sensitivity	-	0.000	0.000	0.000	0.333	0.333	-	0.333	0.333	0.659	0.849	0.939	0.996	1.000	0.991	0.971	0.988
PPV	-	0.000	0.000	0.000	1.000	0.976	-	0.894	1.000	0.978	0.940	0.694	0.568	0.477	0.399	0.314	0.280
CLIC																	
# of OMWS	0	2	4	2	25	23	0	21	10	100	311	93	75	79	33	20	22
Sensitivity	-	0.000	0.000	0.000	0.333	0.333	-	0.333	0.333	0.660	0.842	0.928	0.991	1.000	0.990	0.983	0.985
PPV	-	0.000	0.000	0.000	0.960	0.978	-	0.909	1.000	0.981	0.937	0.686	0.565	0.476	0.400	0.318	0.279

Scenario C																	
Gini																	
# of OMWS	0	0	0	0	0	0	0	34	19	118	705	57	31	29	4	2	1
Sensitivity	-	-	-	-	-	-	-	0.657	0.667	0.932	0.981	0.994	1.000	1.000	1.000	1.000	1.000
PPV	-	-	-	-	-	-	-	1.000	0.921	0.986	0.993	0.746	0.584	0.485	0.388	0.333	0.273
CLIC																	
# of OMWS	0	0	0	0	0	1	0	1	0	53	827	58	29	27	3	1	0
Sensitivity	-	-	-	-	-	0.333	-	0.333	-	0.667	0.976	0.994	1.000	1.000	1.000	1.000	-
PPV	-	-	-	-	-	1	-	1.000	-	0.978	0.993	0.746	0.586	0.484	0.375	0.333	-
Scenario D																	
Gini																	
# of OMWS	0	0	0	0	1	0	0	18	14	68	701	78	57	43	9	6	5
Sensitivity	-	-	-	-	0.333	-	-	0.611	0.619	0.824	0.977	0.991	1.000	1.000	1.000	1.000	0.933
PPV	-	-	-	-	1.000	-	-	1.000	0.905	0.978	0.992	0.736	0.588	0.483	0.417	0.324	0.269
CLIC																	
# of OMWS	0	0	0	0	2	2	0	5	6	56	757	70	57	32	7	5	1
Sensitivity	-	-	-	-	0.333	0.500	-	0.467	0.500	0.679	0.969	0.990	1.000	1.000	1.000	1.000	1.000
PPV	-	-	-	-	1.000	0.750	-	1.000	0.833	0.994	0.991	0.735	0.589	0.484	0.421	0.313	0.273
Scenario E																	
Gini																	
# of OMWS	0	0	0	0	0	1	0	40	19	87	752	55	29	15	2	0	0
Sensitivity	-	-	-	-	-	0.667	-	0.858	0.825	0.946	0.991	0.976	1.000	1.000	1.000	-	-
PPV	-	-	-	-	-	1.000	-	1.000	1.000	0.981	0.998	0.732	0.593	0.486	0.402	-	-
CLIC																	
# of OMWS	0	0	0	0	0	1	0	2	1	34	866	55	28	11	2	0	0
Sensitivity	-	-	-	-	-	0.667	-	0.333	0.667	0.706	0.985	0.970	1.000	1.000	1.000	-	-
PPV	-	-	-	-	-	1.000	-	1.000	1.000	1.000	0.997	0.728	0.593	0.494	0.402	-	-

Table 4. Simulation results of cluster model 1 (40% of the total cases in study region)

	Upper limit																
	1	2	3	4	5	6	8	10	12	15	20	25	30	35	40	45	50
Scenario A																	
Gini																	
# of OMWS	0	0	0	0	0	0	0	0	0	0	134	109	189	3	492	55	18
Sensitivity	-	-	-	-	-	-	-	-	-	-	0.983	0.976	1.000	0.889	1.000	1.000	1.000
PPV	-	-	-	-	-	-	-	-	-	-	0.994	0.989	0.993	0.622	1.000	0.709	0.524
CLIC																	
# of OMWS	0	0	0	0	0	0	0	0	0	0	0	0	0	2	900	74	24
Sensitivity	-	-	-	-	-	-	-	-	-	-	-	-	-	0.667	1.000	1.000	1.000
PPV	-	-	-	-	-	-	-	-	-	-	-	-	-	1.000	1.000	0.709	0.532
Scenario B																	
Gini																	
# of OMWS	0	0	0	0	0	0	0	1	5	0	0	4	41	85	442	191	231
Sensitivity	-	-	-	-	-	-	-	0.333	0.333	-	-	0.333	0.667	0.667	0.992	0.997	0.996
PPV	-	-	-	-	-	-	-	1.000	1.000	-	-	1.000	1.000	0.857	0.988	0.677	0.513
CLIC																	
# of OMWS	0	0	0	0	0	0	0	1	8	0	0	4	49	93	452	176	217
Sensitivity	-	-	-	-	-	-	-	0.333	0.333	-	-	0.333	0.667	0.667	0.994	0.996	0.995
PPV	-	-	-	-	-	-	-	1.000	1.000	-	-	1.000	1.000	0.851	0.990	0.677	0.514

Scenario C																	
Gini																	
# of OMWS	0	0	0	0	0	0	0	0	0	0	12	10	94	11	717	111	45
Sensitivity	-	-	-	-	-	-	-	-	-	-	0.667	0.700	0.972	0.788	1.000	1.000	1.000
PPV	-	-	-	-	-	-	-	-	-	-	1.000	0.867	0.977	0.894	0.999	0.709	0.522
CLIC																	
# of OMWS	0	0	0	0	0	0	0	0	0	0	0	0	9	11	818	114	48
Sensitivity	-	-	-	-	-	-	-	-	-	-	-	-	0.667	0.667	1.000	1.000	1.000
PPV	-	-	-	-	-	-	-	-	-	-	-	-	1.000	1.000	1.000	0.712	0.521
Scenario D																	
Gini																	
# of OMWS	0	0	0	0	0	0	0	0	0	0	7	8	55	16	658	153	103
Sensitivity	-	-	-	-	-	-	-	-	-	-	0.667	0.667	0.915	0.708	0.999	1.000	1.000
PPV	-	-	-	-	-	-	-	-	-	-	1.000	0.917	0.975	0.958	0.999	0.698	0.533
CLIC																	
# of OMWS	0	0	0	0	0	0	0	0	0	0	0	0	24	24	690	160	102
Sensitivity	-	-	-	-	-	-	-	-	-	-	-	-	0.667	0.681	1.000	1.000	1.000
PPV	-	-	-	-	-	-	-	-	-	-	-	-	1.000	0.944	0.999	0.697	0.532
Scenario E																	
Gini																	
# of OMWS	0	0	0	0	0	0	0	0	0	0	36	19	37	4	767	115	22
Sensitivity	-	-	-	-	-	-	-	-	-	-	1.000	0.965	0.973	0.833	1.000	1.000	1.000
PPV	-	-	-	-	-	-	-	-	-	-	1.000	0.970	1.000	0.750	1.000	0.710	0.515
CLIC																	
# of OMWS	0	0	0	0	0	0	0	0	0	0	0	0	3	4	850	121	22
Sensitivity	-	-	-	-	-	-	-	-	-	-	-	-	0.667	0.667	1.000	1.000	1.000
PPV	-	-	-	-	-	-	-	-	-	-	-	-	1.000	0.875	1.000	0.708	0.515

Table 5. Simulation results of cluster model 2 (Circular shape)

	Upper limit																
	1	2	3	4	5	6	8	10	12	15	20	25	30	35	40	45	50
4000 cases																	
Gini																	
# of OMWS	0	0	0	0	0	0	0	29	536	231	101	57	36	0	0	0	0
Sensitivity	-	-	-	-	-	-	-	1.000	1.000	1.000	1.000	1.000	1.000	-	-	-	-
PPV	-	-	-	-	-	-	-	1.000	1.000	1.000	0.986	0.990	1.000	-	-	-	-
CLIC																	
# of OMWS	0	0	0	0	0	0	0	0	0	0	0	125	135	0	0	1	739
Sensitivity	-	-	-	-	-	-	-	-	-	-	-	1.000	1.000	-	-	0.800	1.000
PPV	-	-	-	-	-	-	-	-	-	-	-	1.000	1.000	-	-	0.667	0.714
3000 cases																	
Gini																	
# of OMWS	0	0	0	0	0	0	556	27	317	8	92	0	0	0	0	0	0
Sensitivity	-	-	-	-	-	-	0.999	1.000	0.999	1.000	0.998	-	-	-	-	-	-
PPV	-	-	-	-	-	-	1.000	1.000	1.000	0.750	1.000	-	-	-	-	-	-
CLIC																	
# of OMWS	0	0	0	0	0	0	0	0	0	1	954	0	0	0	45	0	0
Sensitivity	-	-	-	-	-	-	-	-	-	0.800	0.999	-	-	-	1.000	-	-
PPV	-	-	-	-	-	-	-	-	-	1.000	1.000	-	-	-	0.714	-	-

2000 cases																	
Gini																	
# of OMWS	0	0	0	0	28	554	291	10	9	98	0	0	10	0	0	0	0
Sensitivity	-	-	-	-	1.000	0.997	0.999	0.980	1.000	1.000	-	-	1.000	-	-	-	-
PPV	-	-	-	-	1.000	1.000	1.000	0.771	0.778	1.000	-	-	0.714	-	-	-	-
CLIC																	
# of OMWS	0	0	0	0	0	0	0	3	14	982	0	0	1	0	0	0	0
Sensitivity	-	-	-	-	-	-	-	0.800	0.957	0.999	-	-	1.000	-	-	-	-
PPV	-	-	-	-	-	-	-	1.000	1.000	1.000	-	-	0.714	-	-	-	-

Table 6. Simulation results of cluster model 2 (Elliptic shape)

	Upper limit																
	1	2	3	4	5	6	8	10	12	15	20	25	30	35	40	45	50
4000 cases																	
Gini																	
# of OMWS	0	0	0	0	0	0	0	33	557	217	16	73	32	0	69	3	0
Sensitivity	-	-	-	-	-	-	-	1.000	0.999	1.000	0.988	1.000	-	1.000	1.000	1.000	-
PPV	-	-	-	-	-	-	-	1.000	0.999	0.984	0.850	0.960	-	0.809	1.000	0.833	-
CLIC																	
# of OMWS	0	0	0	0	0	0	0	0	0	0	0	0	0	8	990	2	0
Sensitivity	-	-	-	-	-	-	-	-	-	-	-	-	-	0.800	0.999	1.000	-
PPV	-	-	-	-	-	-	-	-	-	-	-	-	-	1.000	1.000	0.833	-
3000 cases																	
Gini																	
# of OMWS	0	0	0	0	0	0	663	32	105	14	123	11	46	6	0	0	0
Sensitivity	-	-	-	-	-	-	0.997	1.000	1.000	1.000	0.998	1.000	0.991	1.000	-	-	-
PPV	-	-	-	-	-	-	0.998	0.995	0.856	0.782	0.958	0.779	1.000	0.833	-	-	-
CLIC																	
# of OMWS	0	0	0	0	0	0	0	0	0	0	0	0	999	1	0	0	0
Sensitivity	-	-	-	-	-	-	-	-	-	-	-	-	0.990	1.000	-	-	-
PPV	-	-	-	-	-	-	-	-	-	-	-	-	1.000	0.833	-	-	-

2000 cases																	
Gini																	
# of OMWS	0	0	0	0	37	672	104	35	5	102	40	5	0	0	0	0	0
Sensitivity	-	-	-	-	0.989	0.987	0.996	0.977	0.920	0.994	0.995	1.000	-	-	-	-	-
PPV	-	-	-	-	1.000	0.998	0.923	0.849	1.000	0.867	1.000	0.833	-	-	-	-	-
CLIC																	
# of OMWS	0	0	0	0	0	0	0	0	0	1	999	0	0	0	0	0	0
Sensitivity	-	-	-	-	-	-	-	-	-	1.000	0.963	-	-	-	-	-	-
PPV	-	-	-	-	-	-	-	-	-	1.000	1.000	-	-	-	-	-	-

Table 7. Simulation results of cluster model 3

	Upper limit																
	1	2	3	4	5	6	8	10	12	15	20	25	30	35	40	45	50
Circular shape																	
Gini																	
# of OMWS	0	0	0	0	0	0	0	0	0	942	57	0	0	0	1	0	0
Sensitivity	-	-	-	-	-	-	-	-	-	1.000	1.000	-	-	-	1.000	-	-
PPV	-	-	-	-	-	-	-	-	-	1.000	1.000	-	-	-	0.375	-	-
CLIC																	
# of OMWS	0	0	0	0	0	0	0	0	0	0	748	0	0	0	0	241	11
Sensitivity	-	-	-	-	-	-	-	-	-	-	1.000	-	-	-	-	1.000	1.000
PPV	-	-	-	-	-	-	-	-	-	-	1.000	-	-	-	-	0.500	0.429
Elliptic shape																	
Gini																	
# of OMWS	0	0	0	0	0	0	0	0	0	940	49	0	0	0	11	0	0
Sensitivity	-	-	-	-	-	-	-	-	-	1.000	1.000	-	-	-	1.000	-	-
PPV	-	-	-	-	-	-	-	-	-	1.000	1.000	-	-	-	0.682	-	-
CLIC																	
# of OMWS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	998	2	0
Sensitivity	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.000	1.000	-
PPV	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.666	0.500	-

Table 8. Simulation results of cluster model 4

	Upper limit																
	1	2	3	4	5	6	8	10	12	15	20	25	30	35	40	45	50
Circular shape																	
Gini																	
# of OMWS	0	0	0	0	0	0	0	376	407	99	101	6	11	0	0	0	0
Sensitivity	-	-	-	-	-	-	-	1.000	1.000	1.000	0.998	1.000	1.000	-	-	-	-
PPV	-	-	-	-	-	-	-	1.000	1.000	1.000	0.833	0.813	0.833	-	-	-	-
CLIC																	
# of OMWS	0	0	0	0	0	0	0	2	3	25	965	1	0	4	0	0	0
Sensitivity	-	-	-	-	-	-	-	1.000	0.933	0.976	0.998	1.000	-	1.000	-	-	-
PPV	-	-	-	-	-	-	-	1.000	1.000	1.000	0.835	0.714	-	0.714	-	-	-
Elliptic shape																	
Gini																	
# of OMWS	0	0	0	0	0	0	0	317	556	19	93	3	0	0	12	0	0
Sensitivity	-	-	-	-	-	-	-	1.000	0.933	0.976	0.998	1.000	-	-	1.000	-	-
PPV	-	-	-	-	-	-	-	1.000	1.000	1.000	0.835	0.714	-	-	0.714	-	-
CLIC																	
# of OMWS	0	0	0	0	0	0	0	0	0	0	0	0	2	2	996	0	0
Sensitivity	-	-	-	-	-	-	-	-	-	-	-	-	0.800	0.800	1.000	-	-
PPV	-	-	-	-	-	-	-	-	-	-	-	-	1.000	1.000	0.833	-	-

4. Application

4.1 Data explanation

Two real data examples were used for demonstrating the utility of the two criteria in the ordinal model. The first data set is the 2013 birth order data based on birth certificate registrations provided by the Korean Statistical Information Service (KOSIS). Birth order was recorded as first, second, and third child and over, and we used the data set of 25 districts (gu) in Seoul only. Table 9 shows the number of cases and percentage by category of the birth order.

Table 9. Data on birth order in Seoul (2013)

	Birth order	n	%
1	First child	48248	57.5
2	Second child	29656	35.4
3	Third child and over	5944	7.09

The second data set is obtained from the 2013 Korea Community Health Survey (KCHS) conducted by the Korea Centers for Disease Control and Prevention. The KCHS is an annual nationwide health survey, which uses multistage sampling design to obtain a representative sample of adults aged over 19 in 253 communities. We used KCHS data on educational levels in 25 districts (gu) of Seoul, South Korea. We classified educational level into four categories:

elementary school and under, middle school, high school, and college and above. The data set is explained in detail in Table 10.

Table 10. Data on educational levels in Seoul (2013)

	Educational level	n	%
1	Elementary school and under	2996	13.0
2	Middle school	1985	8.6
3	High school	6040	26.1
4	College and above	12067	52.3

For the two data sets, we used the spatial scan statistic for ordinal data to search the clusters with high rates of higher-valued categories, and determined the maximum scanning window size based on the Gini coefficient and the CLIC. Both circle and ellipse are used as the scanning window shape.

4.2 Results

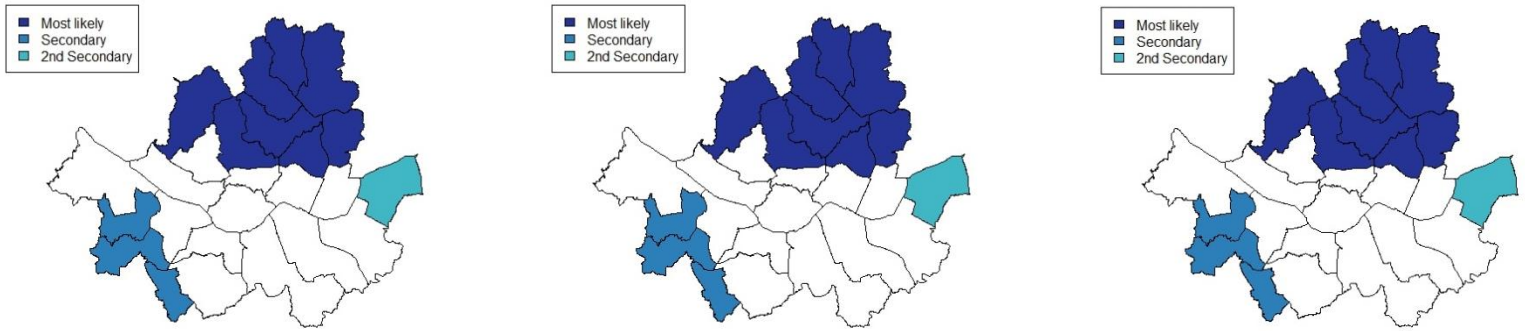
The results of the birth order data are presented in Figure 4. First, when we applied the circular shape to the data, both Gini and CLIC picked 30% as the maximum scanning window size. At the same time, the detected clusters were exactly the same as the clusters chosen the maximum cluster size of 50%. In this case, the most likely cluster contained the 29.1 percentage of the total cases. The

elliptic scan statistic, however, provided slightly different results. Gini and CLIC picked 12% and 35% as the maximum scanning window size, respectively. The detected clusters based on the CLIC result (35%) were equal to that of using 50% and 34% of the total cases in the most likely cluster. However, we found that the result of the Gini included only some parts of the most likely cluster based on default size. According to Table 11, the most likely cluster and some secondary clusters based on the Gini result (12%) can reject the null hypothesis on their own strength. This means that the most likely cluster when using the default upper limit (50%) contains some districts whose tendency to be high rates in the higher-valued category is unapparent than others. For example, the observed proportions of the three categories in the most likely clusters using 50% (default) and 12% (Gini result) as the upper limit are (0.555, 0.368, 0.077) and (0.546, 0.374, 0.080), respectively. Compare these with the observed proportions (0.575, 0.354, 0.071) in the whole study area, and the increasing order trend is more clear in the latter case, although it has the smaller log likelihood ratio (LLR) than the former.

Table 11. Some spatial clusters of high rates of later birth order appear in Figure 4(b): Most likely cluster (upper limit at 50%); most likely cluster, 2nd secondary cluster, and 3rd secondary cluster (upper limit at 12%)

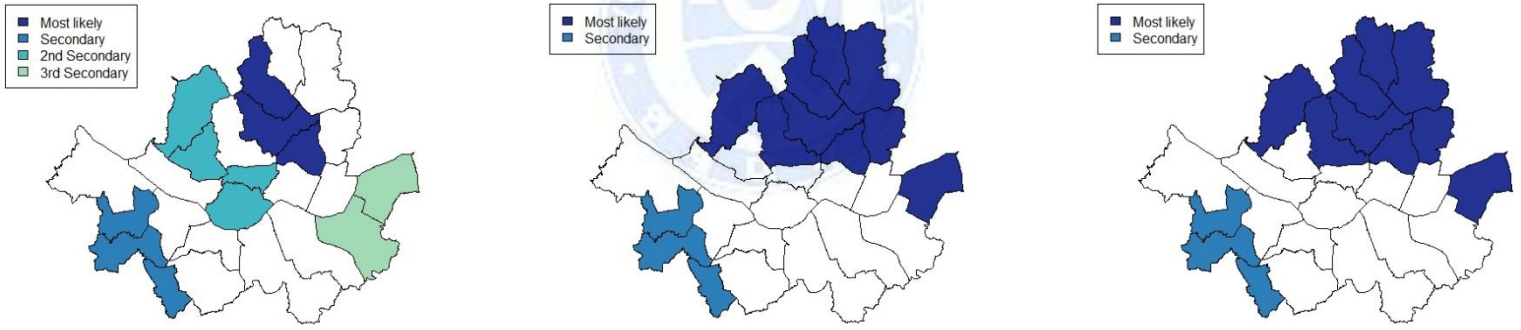
Upper limit	Cluster	# Districts	# Obs in each category	LLR	*p*-value
50%	Most likely	9	(15806, 10493, 2192)	40.00	0.001
12%	Most likely	3	(4894, 3348, 720)	19.23	0.001
	2nd Secondary	4	(5554, 3287, 794)	10.52	0.005
	3rd Secondary	2	(5565, 3719, 729)	8.95	0.008

Figure 5 summarizes the results for educational level data. In both scanning window shape, two criteria chose the same optimal maximum window sizes and CLIC picked the larger maximum window size than Gini (5% for Gini and 12% for CLIC). Further, the result in Figure 5 indicates that both scanning window shapes detected similar clusters. A characteristic feature of the results is that the detected districts in the significant clusters based on the Gini and CLIC results are more widespread than the results using the default size (50%). This indicates the possibility that the irregular-shaped clusters (not shaped in circle or ellipse) can be detected by using small maximum window size.



Gini (30%) CLIC (30%) Default (50%)

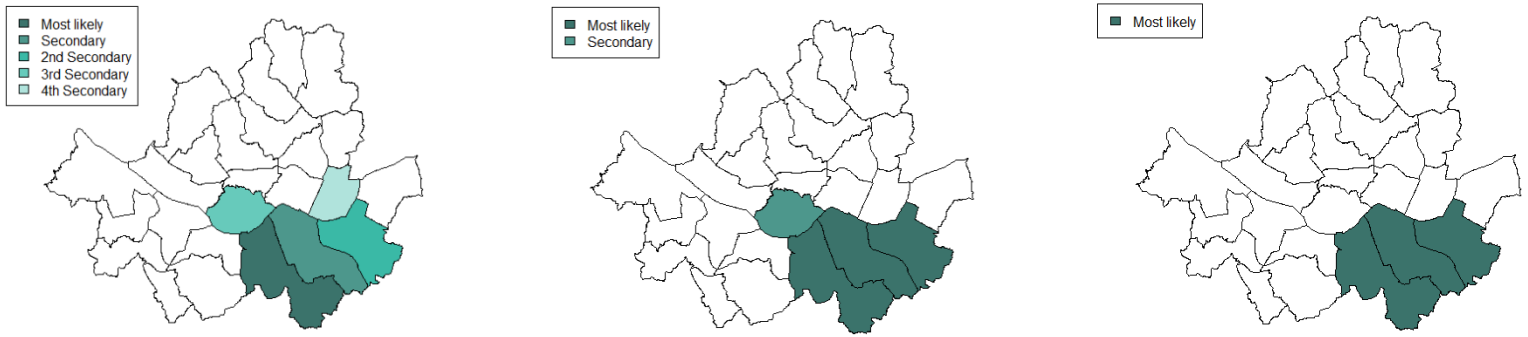
(a) Circular shape



Gini (12%) CLIC (35%) Default (50%)

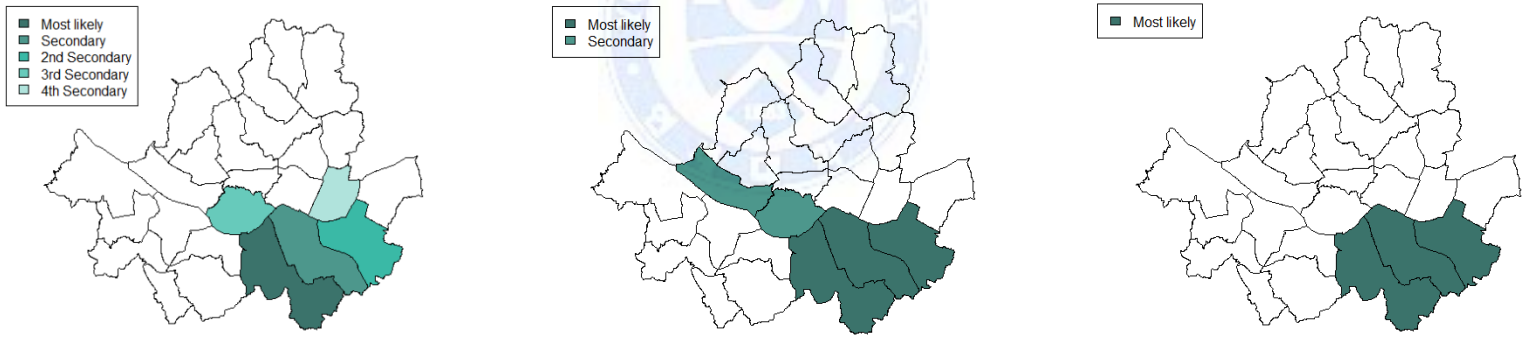
(b) Elliptic shape

Figure 4. Results of the birth order data using the maximum scanning window size chosen by Gini, CLIC, and default setting



Gini (5%) CLIC (12%) Default (50%)

(a) Circular shape



Gini (5%) CLIC (12%) Default (50%)

(b) Elliptic shape

Figure 5. Results of the education level data using the maximum scanning window size chosen by Gini, CLIC, and default setting

5. Discussion and Conclusion

The objective of this study was to examine the applicability of the Gini and CLIC to the ordinal model proposed by Jung, Kulldorff, and Klassen (2007). Through the simulation study and the real data examples, we conclude that the two criteria are proven to be useful criteria to optimize maximum window size in spatial scan statistic for ordinal data as well as for the Poisson model.

There are several findings in the simulation studies and the applications. First, we found that using the default upper limit for the scanning window size (50% of the total cases) tends to detect the unnecessarily big cluster. Second, the two measures pick the same size as the optimal maximum window size in most cases. However, when the true clusters are irregular-shaped or located slightly apart from each other, the Gini chooses a smaller window size than the CLIC, which gives several smaller clusters. Although the results show that they are all in different clusters, the clusters can be regarded as one cluster if they are contiguous. In other words, the use of the Gini for optimizing scanning window size makes it possible to detect even the irregular-shaped clusters.

In conclusion, the results of this study demonstrate the necessity of optimizing the maximum window size in spatial scan statistic for ordinal data, and

the Gini coefficient and the CLIC are applicable to the ordinal model for optimizing the maximum scanning window size.

Reference

Akaike, H. 1974. "A new look at the statistical model identification". Automatic Control, IEEE Transactions on, 19(6): 716-723.

Brunk, H., R. Barlow, D. Bartholomew and J. Bremner. 1972. "Statistical Inference under Order Restrictions.(The Theory and Application of Isotonic Regression)". DTIC Document.

Community Health Survey. 2013. Korea Centers for Disease Control and Prevention.

Cook, A. J., D. R. Gold and Y. Li. 2007. "Spatial cluster detection for censored outcome data". Biometrics, 63(2): 540-549.

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

Goujon-Bellec, S., C. Demoury, A. Guyot-Goubin, D. Hémon and J. Clavel. 2011. "Detection of clusters of a rare disease over a large territory: performance of cluster detection methods". Int J Health Geogr, 10: 53.

Grubesic, T. H., R. Wei and A. T. Murray. 2014. "Spatial Clustering Overview and Comparison: Accuracy, Sensitivity, and Computational Expense". Annals of the Association of American Geographers, 104(6): 1134-1156.

Han J, Feuer R, Stinchcomb D, Tatalovich Z, Lewis D, and L. Zhu. 2011. “Optimizing maximum window size for scan statistics”. [oral presentation]. Louisville (KY): Annual Meeting of North American Association of Central Cancer Registries [cited 2015.7.13]. <http://www.naaccr.org/LinkClick.aspx?fileticket=hR6UMTigRM4%3D&tabid=257&mid=732>.

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

Huang, L., L. W. Pickle and B. Das. 2008. "Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases". Statistics in Medicine, 27(25): 5111-5142.

Jung, I., M. Kulldorff and A. C. Klassen. 2007. "A spatial scan statistic for ordinal data". Statistics in Medicine, 26(7): 1594-1607.

Jung, I., M. Kulldorff and O. J. Richard. 2010. "A spatial scan statistic for multinomial data". Statistics in medicine, 29(18): 1910.

Korean Statistical Information Service. 2013. Statistics Korea, Vital Statistics.

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

Kulldorff, M., L. Huang and K. Konty. 2009. "A scan statistic for continuous data based on the normal probability model". International journal of health geographics, 8(1): 58.

Kulldorff, M., L. Huang, L. Pickle and L. Duczmal. 2006. "An elliptic spatial scan statistic". Statistics in medicine, 25(22): 3929-3943.

Lipsitz, S. R., G. M. Fitzmaurice and G. Molenberghs. 1996. "Goodness-of-fit tests for ordinal response regression models". Applied Statistics: 175-190.

Ribeiro, S. H. R. and M. A. Costa. 2012. "Optimal selection of the spatial scan parameters for cluster detection: a simulation study". Spatial and spatio-temporal epidemiology, 3(2): 107-120.

국 문 요 약

순서형 자료를 위한 공간검색통계량에서
최대 후보 군집 크기의 최적화

우도비 검정을 기반으로 하는 공간검색통계량은 어떤 사건에 대한 분포가 다른 지역의 분포와 통계적으로 유의하게 다른 공간 군집(spatial cluster)을 탐색하기 위한 방법으로 여러 분야에서 이용되고 있다. 이 때, 군집 탐색 결과는 사전에 설정한 후보 군집(scanning window)의 모양뿐만 아니라 최대 군집 크기에 따라 달라질 수 있다. 보통 최대 군집 크기를 보통 전체 인구의 50%로 설정하게 되는데 이를 뒷받침할 만한 연구 결과가 충분하지 않다.

최근 Han 등(2011)이 최적의 최대 후보 군집 크기를 결정하기 위한 방법으로 지니계수(Gini Coefficient)와 Cluster Information Criterion (CLIC)를 제안하였다. 하지만, 포아송 분포를 가정한 공간검색통계량에 제한하여 적용된 방법으로 다른 분포에서의 활용 가능성은 밝혀진 바 없다. 본 연구는 순서형 자료를 위한 공간검색통계량(Jung 등, 2007)에 적용할 수 있는 두 방법의 활용 방안을 제안하고, 모의실험 및 실제 자료에의 적용을 통해 적합성을 살펴보는 데에 목적이 있다.

그 결과, 지니계수와 CLIC 모두 실제 군집에 근사한 최대 후보 군집 크기를 결정함을 확인할 수 있었다. 단, 예외적으로 실제 군집의 모양이 비정형이거나 군집들이 다소 떨어져 있는 경우, 지니계수가 CLIC 보다 최대 후보 군집 크기를 작게

선택하였으며 결과적으로 작은 크기의 군집을 여러 개 탐색하게 됨으로써 실제 군집을 좀 더 정확히 찾는 경향이 있었다.

이로써 본 연구는 순서형 자료를 위한 공간검색통계량에서 최대 후보 군집 크기를 이용하면 보다 효율적인 군집 탐색이 가능함을 보였고, 이에 지니계수와 CLIC가 유용하게 활용될 수 있을 것으로 기대된다.

핵심되는 말: 공간검색통계량, 순서형 자료, 최대 후보 군집 크기, 지니계수, CLIC