Communications for Statistical Applications and Methods
2018, Vol. 25, No. 4, 373–383

https://doi.org/10.29220/CSAM.2018.25.4.373
Print ISSN 2287-7843 / Online ISSN 2383-4757

Optimizing the maximum reported cluster size for normal-based spatial scan statistics

Haerin Yoo[a], Inkyung Jung[1,a]

[a]Division of Biostatistics, Department of Biomedical Systems Informatics,
Yonsei University College of Medicine, Korea

Abstract

The spatial scan statistic is a widely used method to detect spatial clusters. The method imposes a large number of scanning windows with pre-defined shapes and varying sizes on the entire study region. The likelihood ratio test statistic comparing inside versus outside each window is then calculated and the window with the maximum value of test statistic becomes the most likely cluster. The results of cluster detection respond sensitively to the shape and the maximum size of scanning windows. The shape of scanning window has been extensively studied; however, there has been relatively little attention on the maximum scanning window size (MSWS) or maximum reported cluster size (MRCS). The Gini coefficient has recently been proposed by Han *et al.* (*International Journal of Health Geographics*, **15**, 27, 2016) as a powerful tool to determine the optimal value of MRCS for the Poisson-based spatial scan statistic. In this paper, we apply the Gini coefficient to normal-based spatial scan statistics. Through a simulation study, we evaluate the performance of the proposed method. We illustrate the method using a real data example of female colorectal cancer incidence rates in South Korea for the year 2009.

Keywords: spatial cluster detection, Gini coefficient, maximum scanning window size, weighted normal model

1. Introduction

Spatial scan statistics have been widely used as a useful technique for cluster detection in different fields such as disease surveillance and spatial epidemiology. This method identifies statistically significant spatial clusters with higher or lower rates than other regions. It has been developed for various types of data such as Poisson (Kulldorff, 1997), ordinal (Jung *et al.*, 2007), survival (Huang *et al.*, 2007), normal (Kulldorff *et al.*, 2009; Huang *et al.*, 2009) and multinomial data (Jung *et al.*, 2010). There is freely available software called SaTScan^{TM} (Kulldorff and Information Management Services, 2018) for the method.

Spatial cluster detection using the spatial scan statistics is conducted based on the likelihood ratio test. In this process, a very large number of candidate areas (scanning windows) with a pre-defined shape and varying sizes are assumed in order to explore the whole study region. The likelihood ratio test statistic is calculated for each scanning window to compare inside versus outside the window; subsequently, the window which maximizes the test statistic is determined as the most likely cluster. As a result, the results of the cluster detection respond sensitively to the shape and the maximum size

[1] Corresponding author: Division of Biostatistics, Department of Biomedical Systems Informatics, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-Gu, Seoul 03722, Korea. E-mail: ijung@yuhs.ac

Published 31 July 2018 / journal homepage: http://csam.or.kr

of scanning windows. Several previous studies focused on the shape of scanning windows such as circular (Kulldorff, 1997), elliptic (Kulldorff *et al.*, 2006), or irregular shapes (Patil and Taillie, 2004; Duczmal and Assunção, 2004; Tango and Takahashi, 2005). However, there has not been significant interest in the maximum scanning window size (MSWS) or maximum reported cluster size (MRCS).

The MSWS is generally defined as less than 50% of total population at risk or the number of geographical study areas. The SaTScan^{TM} software also uses 50% as the default setting. However, a higher value of MSWS can lead to the detection of spatial clusters larger than the true clusters, including less informative surrounding areas. On the contrary, larger clusters are not considered in the analysis process and cannot be found when a smaller value of MSWS is used. Ribeiro and Costa (2012) found that the performance of the spatial scan statistic is sensitive to different values of MSWS. Therefore, they suggested selecting the optimal value of the maximum cluster size rather than using the commonly used 50%. However, Han *et al.* (2016) emphasized we should never conduct spatial cluster detection analyses repeatedly using different values of MSWS and select the result of the lowest p-value because it can cause a multiple testing problem. Alternatively, we can rerun the analysis with a fixed larger value of MSWS (e.g., 50%) and with different MRCS (e.g., 5, 10, 15, 20, and 50%) to find an optimal cluster reporting size. Then, we can report clusters smaller than the MRCS, at the same time adjusting for the multiple testing problem using a fixed larger MSWS. Han *et al.* (2016) proposed an effective criterion, the Gini coefficient, to determine the optimal MRCS. The Gini coefficient represents the degree of heterogeneity of the disease clusters. The simulation study indicates the Gini coefficient can identify the best collection of non-overlapping clusters to report and those clusters tend to have a similar size with true clusters. This method was developed only for the Poisson model in the study by Han *et al.* (2016). Recently, Kim and Jung (2017) employed the Gini coefficient for the ordinal model to show that it can also be successfully used to optimize the MRCS in the spatial scan statistic for ordinal data.

This paper applies the Gini coefficient to normal-based spatial scan statistics and evaluate it through simulations. The normal-based spatial scan statistic can be used for continuous data at the individual level (Kulldorff *et al.*, 2009) or at some aggregated level with a heterogeneous population (Huang *et al.*, 2009). The idea of using the Gini coefficient may be similar to the work by Han *et al.* (2016); however, the method should be clearly defined for the specific model and should be fully evaluated for real use. In the Section 2, we define the criterion for the weighted normal spatial scan statistic to optimize the MRCS. The standard normal model is a special case of the weighted normal model; therefore, it is straightforward to reduce the method to the standard normal model. In Section 3, we evaluate the applicability of the Gini coefficient through simulations under various scenarios. We illustrate the application of the proposed method to a real data example in Section 4. We then discuss our results and the conclusion in Section 5.

2. Methods

2.1. Gini coefficient for Poisson-based spatial scan statistic

The Poisson-based spatial scan statistic compares cases against the underlying population at risk. The null hypothesis is written as $H_0 : p = q$ for all $z \in Z$ and the alternative hypothesis is $H_a : p > q$ for some z, where p and q are the intensities of the outcome variable inside and outside scanning window z, and Z denotes the collection of all scanning windows. For a given scanning window z, the

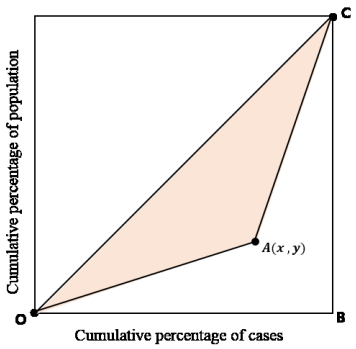

Figure 1: *Illustration of Lorenz curve and Gini coefficient.*

likelihood ratio test statistic LR(z) is expressed as

$$\mathrm{LR}(z) = \frac{\left(\frac{c_z}{n_z}\right)^{c_z}\left(\frac{C-c_z}{N-n_z}\right)^{C-c_z}}{\left(\frac{C}{N}\right)^{C}}$$

if $c_z/n_z > (C - c_z)/(N - n_z)$, and LR($z$) = 1 otherwise. Here, c_z and n_z are the number of cases and population within z, C and N are the total number of cases and population in the whole study region, respectively. The area with the maximum value of LR(z) over $z \in Z$ is the most likely cluster. Monte Carlo hypothesis testing (Dwass, 1957) is often used to obtain a p-value for the most likely cluster.

There can be significant secondary clusters with a high likelihood ratio test statistic values and we often report secondary clusters that do not overlap with a more likely cluster. However, this collection of clusters may not be the best one to report. There could be more meaningful (i.e., higher relative risk) but smaller clusters than the most likely cluster that are hidden within the most likely cluster or overlapped with. Han *et al.* (2016) proposed a criterion, the Gini coefficient, to find a suitable and informative collection of non-overlapping clusters to report.

The Gini coefficient was originally developed to represent the degree of heterogeneity of wealth distribution in economics (Gastwirth, 1972). It is a summary measure from a Lorenz curve (Lorenz, 1905) and a larger value indicates more heterogeneous distribution. Han *et al.*(2016) applied the concept of the Gini coefficient to evaluate the degree of the heterogeneity of the collection of clusters for count data. The x-axis of the Lorenz curve can be defined as the cumulative percentage of the number of disease cases (e.g., number of deaths or events of disease) and the y-axis is the cumulative percentage of the population. Figure 1 describes the corresponding Lorenz when there is only one significant cluster. The reference line OC indicates that the number of cases is proportional to the population for each region. The Gini coefficient is two times the area between the reference line and the Lorenz curve. As more cases are concentrated to the cluster, the point A becomes further away from the reference line and the value of the Gini coefficient increases. When there are I multiple clusters, the coordinates of each cluster (x_i, y_i) $(i = 1, \ldots, I)$ are defined using the cumulative cases and population by the order of relative risk of the cluster. More specifically, $x_i = (1/C)\sum_{k=1}^{i} c_k$ and $y_i = (1/N)\sum_{k=1}^{i} n_k$ where c_k and n_k are the number of cases and population in k^{th} cluster, respectively. The Gini coefficient can be calculated as $G = \sum_{i=1}^{I+1}(y_i x_{i-1} - y_{i-1}x_i)$ where $x_0 = y_0 = 0$ and $x_{I+1} = y_{I+1} = 1$ (Han *et al.*, 2016). The value of the Gini coefficient is between 0 and 1. The highest Gini coefficient value indicates the best collection of clusters to report among several competing collections of clusters. Refer to the study by Han *et al.* (2016) for detailed information.

2.2. Spatial scan statistics for continuous data

There are two types of spatial scan statistics based on the normal probability model. One is to detect spatial clusters of individuals or locations with high or low values of some continuous data attribute (Kulldorff *et al.*, 2009) and the other is to detect clusters of geographic units with continuous regional measures (Huang *et al.*, 2009). The former concerns with individual level data while the latter is used for continuous data at some aggregate level with heterogeneous population such as mortality rates, incidence rates and average survival at district level. The two models were proposed separately in two different articles, but here we briefly review the weighted normal model for aggregate level data because the standard normal model is a special case of the weighted normal model with homogeneous weight. The null and alternative hypotheses are written as $H_0 : \mu_z = \mu_{z^C} = \mu_G$ and $H_a : \mu_z \neq \mu_{z^C}$ for some z, where μ_z. μ_{z^C}, and μ_G are the means of regional summary measurements w_j's for inside and outside scanning window z, and the whole study area G. Maximizing the likelihood ratio test statistic given z is equivalent to maximizing

$$-\sum_{j\in G}\delta_j w_z^2 + \frac{(\sum_{j\in z}\delta_j w_j)^2}{\sum_{j\in z}\delta_j} + \frac{(\sum_{j\in z^C}\delta_j w_j)^2}{\sum_{j\in z^C}\delta_j} \tag{2.1}$$

assuming $w_j|\delta_j \sim N(\mu_z, \sigma_G^2/\delta_j)$ when $j \in z$ and $w_j|\delta_j \sim N(\mu_{z^C}, \sigma_G^2/\delta_j)$ when $j \in z^C (= G - z)$, where δ_j is the weight, associated with w_j and σ_G^2 is the variance of w_j's in the whole study region G. σ_G^2/δ_j is the variance of w_j after adjusting the local weight δ_j. The weight δ_j is assumed to be a known measure proportional to the inverse of the uncertainty in each j ($j \in G$). For example, in the case of mortality rates data at the district level, the inverse of the associated variances for each district can be δ_j or we may use population size at each district as a substitute for δ_j. The standard normal model can be described using $\delta_j = 1$, $j \in G$. One can refer to the article by Huang *et al.* (2009) for the derivation of the test statistic (2.1) and more detailed information on the weighted normal model.

Inference procedure for the normal model is the same as the Poisson model. However, there is no criterion to select the optimal collection of clusters to report. In the next section, we propose a Gini coefficient specifically designed for the normal model and to evaluate the performance through a simulation study in the subsequent section.

2.3. Optimizing the maximum reported cluster size for normal-based spatial scan statistics

Assuming that there is only one significant cluster z^* (Figure 1), we define the x- and y-coordinates of the point A for the cluster z^* in Lorenz curve in the weighted normal model as

$$x = \frac{\sum_{j\in z^*}\delta_j w_j}{\sum_{j\in G}\delta_j w_j} = \frac{\hat{\mu}_{z^*}\sum_{j\in z^*}\delta_j}{\sum_{j\in G}\delta_j w_j} \tag{2.2}$$

and

$$y = \frac{\sum_{j\in z^*}\delta_j}{\sum_{j\in G}\delta_j}, \tag{2.3}$$

where $\hat{\mu}_{z^*} = \sum_{j\in z^*}\delta_j w_j / \sum_{j\in z^*}\delta_j$. Here, the x-coordinate (2.2) represents the weighted sum inside the cluster to the total weighted sum of the regional measures and the y-coordinate (2.3) represents the proportion of weights (or population) inside the cluster to the whole study region. The Gini coefficient

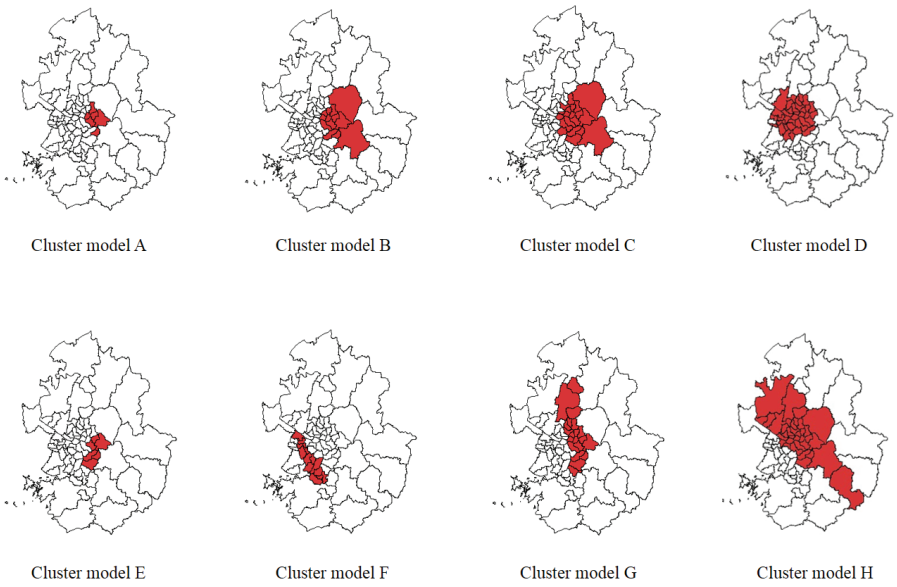

Figure 2: *True cluster models used in the simulation study.*

is the two times of the area between the reference line and the Lorenz curve. As the weighted sum of the regional measures increases, point A becomes further away from the reference line and the value of the Gini coefficient increases. When there are I multiple clusters, we define the x- and y-coordinates for each cluster (x_i, y_i) $(i = i, \ldots, I)$ using the cumulative sum in the numerators in (2.2) and (2.3) by the order of statistical significance of the clusters. That is, $x_i = \sum_{k=1}^{i} \sum_{j \in z^{(k)}} \delta_j w_j / \sum_{j \in G} \delta_j w_j$ and $y_i = \sum_{k=1}^{i} \sum_{j \in z^{(k)}} \delta_j / \sum_{j \in G} \delta_j$, where $z^{(1)}, \ldots, z^{(I)}$ are the detected clusters ordered by their statistical significance. Then the Gini coefficient can be calculated in the same way as the Poisson model as $G = \sum_{i=1}^{I+1} (y_i x_{i-1} - y_{i-1} x_i)$, where $x_0 = y_0 = 0$ and $x_{I+1} = y_{I+1} = 1$. We can select the collection of the clusters with the highest value of the Gini coefficient as the optimal one to report among several competing cluster models. The Gini coefficient for the standard normal model can be defined using $\delta_j = 1$ without any modification.

3. Simulation study

In order to evaluate the performance of the Gini coefficient in the weighted normal model, we conducted a simulation study under various scenarios. We used the area of Seoul and Gyeonggi province in South Korea as the whole study region consisting of 69 districts at the "Si-gun-gu" (district) level. We assumed 8 different true cluster models (Figure 2). Four models (A–D) are circular-shaped clusters and include 6, 13, 20, and 34 districts in the true clusters, respectively. The number of districts in the true clusters in A, B, C, and D accounts for 10%, 20%, 30%, and 50% of the whole study region, respectively. The other four models (E–H) are elliptic-shaped clusters, which also account for 10%, 20%, 30%, and 50% of the study region, respectively. For each true cluster model A–H, we assumed three different values for the mean inside the true cluster, $\mu_z = 1$, 2, and 3. For the districts outside the true cluster, we assumed $\mu_{z^C} = 0$. For all settings, we used identical variance of the whole region $\sigma_G^2 = 1$ and the true population in each district for the year 2010 for δ_j.

Table 1: Simulation results for cluster model A (circular cluster, 10% of the whole study region)

	μ_z		Maximum reported cluster size													Default
			3%	5%	6%	8%	10%	12%	15%	20%	25%	30%	35%	40%	45%	50%	
Circular window	1	Frequency	2	19	36	32	28	26	9	11	14	8	8	5	4	3	
		Sensitivity	0.333	0.456	0.523	0.740	0.839	0.769	0.815	0.818	0.869	0.938	0.917	1.000	1.000	0.889	0.733
		PPV	1.000	0.912	0.785	0.888	0.839	0.639	0.514	0.409	0.346	0.294	0.244	0.235	0.197	0.165	0.667
	2	Frequency	10	31	39	139	378	127	22	32	25	8	2	7	2	2	
		Sensitivity	0.333	0.478	0.632	0.821	0.982	0.951	0.902	0.932	0.960	1.000	1.000	0.952	1.000	1.000	0.903
		PPV	1.000	0.957	0.949	0.986	0.982	0.786	0.557	0.460	0.385	0.325	0.261	0.220	0.207	0.182	0.883
	3	Frequency	2	6	20	126	747	71	6	4	7	3	-	1	-	-	
		Sensitivity	0.333	0.500	0.675	0.833	0.998	0.986	0.883	1.000	1.000	1.000	-	1.000	-	-	0.964
		PPV	1.000	1.000	0.986	1.000	0.998	0.823	0.528	0.483	0.417	0.316	-	0.231	-	-	0.974
Elliptic window	1	Frequency	12	16	22	13	18	29	21	12	11	9	10	3	3	4	
		Sensitivity	0.319	0.375	0.485	0.641	0.880	0.822	0.778	0.875	0.894	0.889	0.833	0.833	0.917	1.000	0.716
		PPV	0.958	0.750	0.727	0.769	0.880	0.664	0.502	0.437	0.348	0.283	0.221	0.192	0.192	0.182	0.603
	2	Frequency	16	29	59	87	248	213	59	42	22	11	2	6	5	1	
		Sensitivity	0.323	0.494	0.653	0.803	0.968	0.965	0.980	0.972	0.962	0.985	1.000	1.000	1.000	1.000	0.898
		PPV	0.969	0.989	0.979	0.963	0.968	0.783	0.633	0.497	0.385	0.313	0.267	0.231	0.203	0.176	0.833
	3	Frequency	5	4	33	108	610	176	35	12	6	2	-	-	-	-	
		Sensitivity	0.333	0.500	0.667	0.833	0.996	0.993	0.986	1.000	1.000	1.000	-	-	-	-	0.961
		PPV	1.000	1.000	1.000	1.000	0.996	0.827	0.634	0.520	0.381	0.308	-	-	-	-	0.943

PPV = positive predictive value

Table 2: Simulation results for cluster model B (circular cluster, 20% of the whole study region)

	μ_z		Maximum reported cluster size													Default
			3%	5%	6%	8%	10%	12%	15%	20%	25%	30%	35%	40%	45%	50%	
Circular window	1	Frequency	10	9	9	16	10	26	48	175	89	39	26	15	20	10	
		Sensitivity	0.154	0.231	0.291	0.361	0.408	0.538	0.646	0.894	0.918	0.890	0.893	0.933	0.950	0.931	0.795
		PPV	1.000	1.000	0.944	0.938	0.883	0.927	0.864	0.925	0.803	0.618	0.519	0.463	0.416	0.367	0.809
	2	Frequency	-	-	1	-	1	11	32	756	155	18	9	4	4	2	
		Sensitivity	-	-	0.308	-	0.462	0.580	0.702	0.985	0.989	0.953	0.966	0.962	0.942	0.923	0.969
		PPV	-	-	1.000	-	1.000	0.976	0.947	0.997	0.881	0.664	0.566	0.486	0.419	0.353	0.961
	3	Frequency	-	-	-	-	-	-	2	912	82	4	-	-	-	-	
		Sensitivity	-	-	-	-	-	-	0.692	0.994	0.999	0.981	-	-	-	-	0.994
		PPV	-	-	-	-	-	-	1.000	1.000	0.909	0.654	-	-	-	-	0.991
Elliptic window	1	Frequency	7	8	13	10	27	53	57	74	121	39	33	19	12	12	
		Sensitivity	0.143	0.231	0.278	0.346	0.427	0.515	0.657	0.824	0.888	0.931	0.925	0.903	0.936	0.955	0.745
		PPV	0.929	1.000	0.904	0.900	0.926	0.895	0.895	0.886	0.771	0.641	0.548	0.453	0.418	0.376	0.779
	2	Frequency	-	-	1	4	2	24	63	576	260	31	20	4	3	2	
		Sensitivity	-	-	0.308	0.385	0.462	0.558	0.716	0.965	0.972	0.955	0.981	0.981	0.923	1.000	0.937
		PPV	-	-	1.000	1.000	1.000	0.983	0.977	0.986	0.870	0.651	0.586	0.500	0.396	0.394	0.931
	3	Frequency	-	-	-	-	-	1	11	832	153	2	1	-	-	-	
		Sensitivity	-	-	-	-	-	0.615	0.748	0.991	0.989	1.000	1.000	-	-	-	0.988
		PPV	-	-	-	-	-	1.000	1.000	0.999	0.890	0.686	0.619	-	-	-	0.981

PPV = positive predictive value

First, we generated 1,000 random data sets under each of 24 different settings with 8 different cluster models and 3 different mean values. Then, we analyzed each data set using the weighted normal spatial scan statistic in the SaTScan software searching for clusters with a high continuous value. We used both circular and elliptic scanning windows. We fixed the MSWS as 50% and used 15 different values (3–6, 8, 10, 12, 15, 20, 25, 30, 35, 40, 45, and 50%) for the MRCS. For the detection result from each value of MRCS, we calculated the Gini coefficient and reported the frequency of each MRCS chosen as the optimal maximum reporting size among 1,000. We also evaluated the accuracy of the detected clusters using sensitivity and positive predicted value (PPV) defined as the number of districts correctly detected among districts in the true cluster and as the number of districts correctly detected among districts in the detected cluster, respectively. Higher values of these two measures reflect a more accurate detection. A smaller value of sensitivity means that detected cluster missed more districts in the true cluster. A smaller value of PPV indicates that the detected cluster includes more districts other than the true cluster. The sensitivity and PPV were estimated as average values

Table 3: Simulation results for cluster model C (circular cluster, 30% of the whole study region)

	μ_z		Maximum reported cluster size													Default
			3%	5%	6%	8%	10%	12%	15%	20%	25%	30%	35%	40%	45%	50%	
Circular window	1	Frequency	4	6	12	10	7	21	26	38	162	123	123	41	44	30	
		Sensitivity	0.100	0.150	0.192	0.240	0.300	0.371	0.458	0.562	0.758	0.891	0.916	0.940	0.941	0.955	0.777
		PPV	1.000	1.000	0.958	0.960	1.000	0.986	0.963	0.912	0.954	0.939	0.824	0.724	0.636	0.573	0.873
	2	Frequency	-	-	-	-	-	1	4	8	163	590	201	17	11	5	
		Sensitivity	-	-	-	-	-	0.350	0.475	0.594	0.805	0.978	0.989	0.965	0.968	0.970	0.946
		PPV	-	-	-	-	-	1.000	0.950	0.989	0.990	0.997	0.908	0.756	0.670	0.592	0.968
	3	Frequency	-	-	-	-	-	-	-	-	42	869	89	-	-	-	
		Sensitivity	-	-	-	-	-	-	-	-	0.829	0.994	1.000	-	-	-	0.988
		PPV	-	-	-	-	-	-	-	-	1.000	1.000	0.938	-	-	-	0.994
Elliptic window	1	Frequency	4	7	7	8	6	27	28	63	120	92	112	73	46	26	
		Sensitivity	0.075	0.150	0.200	0.238	0.283	0.363	0.455	0.579	0.734	0.837	0.907	0.953	0.942	0.967	0.762
		PPV	0.750	1.000	1.000	0.950	0.944	0.974	0.952	0.950	0.930	0.879	0.806	0.741	0.643	0.585	0.848
	2	Frequency	-	-	-	1	1	2	2	20	150	525	227	60	11	2	
		Sensitivity	-	-	-	0.200	0.300	0.375	0.450	0.603	0.789	0.960	0.979	0.987	0.982	1.000	0.930
		PPV	-	-	-	0.800	1.000	1.000	1.000	0.988	0.988	0.985	0.889	0.772	0.688	0.606	0.948
	3	Frequency	-	-	-	-	-	-	-	-	28	871	99	2	-	-	
		Sensitivity	-	-	-	-	-	-	-	-	0.823	0.990	0.995	1.000	-	-	0.985
		PPV	-	-	-	-	-	-	-	-	1.000	0.998	0.928	0.770	-	-	0.991

PPV = positive predictive value

Table 4: Simulation results for cluster model D (circular cluster, 50% of the whole study region)

	μ_z		Maximum reported cluster size													Default
			3%	5%	6%	8%	10%	12%	15%	20%	25%	30%	35%	40%	45%	50%	
Elliptic window	1	Frequency	5	4	6	4	4	6	9	21	28	26	65	93	145	317	
		Sensitivity	0.059	0.081	0.118	0.147	0.176	0.201	0.288	0.356	0.446	0.528	0.643	0.740	0.831	0.937	0.777
		PPV	1.000	0.917	1.000	1.000	1.000	0.887	0.989	0.996	0.967	0.939	0.960	0.965	0.950	0.953	0.957
	2	Frequency	-	-	-	-	-	-	-	-	1	3	10	31	109	846	
		Sensitivity	-	-	-	-	-	-	-	-	0.441	0.529	0.671	0.759	0.879	0.984	0.962
		PPV	-	-	-	-	-	-	-	-	1.000	1.000	0.992	0.996	0.989	0.993	0.992
	3	Frequency	-	-	-	-	-	-	-	-	-	-	-	2	24	974	
		Sensitivity	-	-	-	-	-	-	-	-	-	-	-	0.794	0.897	0.996	0.995
		PPV	-	-	-	-	-	-	-	-	-	-	-	1.000	0.997	0.999	0.999
Elliptic window	1	Frequency	4	3	7	5	7	5	14	18	31	31	74	89	203	224	
		Sensitivity	0.059	0.078	0.118	0.141	0.172	0.206	0.273	0.342	0.449	0.535	0.641	0.728	0.821	0.913	0.744
		PPV	1.000	0.889	1.000	0.960	0.976	0.925	0.969	0.976	0.971	0.949	0.966	0.943	0.938	0.932	0.944
	2	Frequency	-	-	-	-	-	-	-	-	1	1	11	47	208	732	
		Sensitivity	-	-	-	-	-	-	-	-	0.500	0.559	0.674	0.761	0.864	0.974	0.947
		PPV	-	-	-	-	-	-	-	-	1.000	1.000	0.988	0.985	0.978	0.986	0.985
	3	Frequency	-	-	-	-	-	-	-	-	-	-	-	1	45	954	
		Sensitivity	-	-	-	-	-	-	-	-	-	-	-	0.765	0.893	0.994	0.991
		PPV	-	-	-	-	-	-	-	-	-	-	-	0.963	0.999	0.998	0.998

PPV = positive predictive value

from the rejected data sets out of 1,000. We compared the accuracy of the detected clusters based on the Gini coefficient with that from the results using the default setting with 50% of MSWS and 50% of MRCS.

Tables 1–4 show the results for cluster models A–D. When the true cluster is circular-shaped, the Gini coefficient most often selected the best MRCS the same as the size of the true cluster when $\mu_z = 2$ or 3 using either circular or elliptic windows. Compared with the default setting, both the sensitivity and PPV of the detected clusters at the most often picked optimal MRCS were higher. When $\mu_z = 1$, the most often chosen optimal MRCS did not exactly agree with the true cluster size. Still, the PPV were slightly higher than the default setting, which suggests that the default setting tended to detect larger clusters than the true clusters. However, we do not consider the results meaningful because the statistical power in that case is very low, especially for cluster model A.

Tables 5–8 provide the results for the for elliptic-shaped cluster models E–H. As expected, the elliptic spatial scan statistic using the Gini coefficient most often chose the optimal MRCS similar

Table 5: Simulation results for cluster model E (elliptic cluster, 10% of the whole study region)

	μ_z		Maximum reported cluster size													Default
			3%	5%	6%	8%	10%	12%	15%	20%	25%	30%	35%	40%	45%	50%	
Circular window	1	Frequency	6	13	23	21	13	35	18	19	12	13	7	3	5	3	
		Sensitivity	0.167	0.423	0.609	0.683	0.718	0.876	0.917	0.825	1.000	0.833	0.929	1.000	0.967	0.889	0.769
		PPV	0.500	0.846	0.913	0.819	0.718	0.724	0.611	0.409	0.382	0.266	0.245	0.228	0.201	0.162	0.616
	2	Frequency	10	41	103	131	42	318	85	27	39	8	9	2	2	2	
		Sensitivity	0.300	0.512	0.655	0.803	0.810	0.989	0.973	0.932	0.987	1.000	0.963	1.000	0.917	1.000	0.872
		PPV	0.900	0.976	0.983	0.963	0.810	0.833	0.636	0.478	0.378	0.312	0.259	0.231	0.181	0.179	0.809
	3	Frequency	2	109	63	147	13	574	48	18	12	3	3	-	-	-	
		Sensitivity	0.250	0.948	0.667	0.828	0.821	1.000	0.990	0.991	1.000	1.000	1.000	-	-	-	0.938
		PPV	0.750	1.000	1.000	0.993	0.821	0.847	0.644	0.505	0.392	0.333	0.281	-	-	-	0.859
Elliptic window	1	Frequency	15	12	14	12	17	32	15	17	14	17	6	4	4	9	
		Sensitivity	0.244	0.319	0.488	0.736	0.725	0.927	0.900	0.794	0.869	0.941	0.917	0.917	0.958	0.907	0.753
		PPV	0.733	0.639	0.732	0.883	0.725	0.756	0.581	0.404	0.344	0.300	0.247	0.214	0.204	0.162	0.565
	2	Frequency	10	24	45	113	228	258	50	43	20	6	11	1	4	2	
		Sensitivity	0.317	0.500	0.615	0.804	0.958	0.981	0.980	0.961	0.967	0.944	0.955	1.000	1.000	1.000	0.907
		PPV	0.950	1.000	0.922	0.965	0.958	0.815	0.637	0.485	0.382	0.306	0.262	0.231	0.198	0.176	0.834
	3	Frequency	-	18	12	82	585	253	31	8	2	2	1	-	1	-	
		Sensitivity	-	0.481	0.667	0.829	0.996	0.997	0.984	1.000	1.000	1.000	1.000	-	1.000	-	0.969
		PPV	-	0.963	1.000	0.995	0.996	0.831	0.631	0.513	0.376	0.325	0.286	-	0.214	-	0.934

PPV = positive predictive value.

Table 6: Simulation results for cluster model F (elliptic cluster, 20% of the whole study region)

	μ_z		Maximum reported cluster size													Default
			3%	5%	6%	8%	10%	12%	15%	20%	25%	30%	35%	40%	45%	50%	
Circular window	1	Frequency	7	8	11	13	4	18	15	27	22	15	16	14	10	10	
		Sensitivity	0.088	0.144	0.217	0.308	0.327	0.427	0.467	0.516	0.675	0.718	0.837	0.918	0.831	0.885	0.564
		PPV	0.571	0.625	0.705	0.800	0.708	0.728	0.642	0.569	0.564	0.486	0.495	0.462	0.366	0.345	0.575
	2	Frequency	7	30	44	53	30	101	33	60	43	30	54	105	21	14	
		Sensitivity	0.143	0.221	0.274	0.360	0.428	0.506	0.506	0.617	0.735	0.792	0.877	0.985	0.974	0.934	0.634
		PPV	0.929	0.956	0.892	0.925	0.928	0.792	0.694	0.675	0.627	0.529	0.525	0.500	0.428	0.367	0.689
	3	Frequency	3	35	58	85	71	144	46	96	43	12	89	227	17	9	
		Sensitivity	0.154	0.218	0.292	0.392	0.489	0.562	0.547	0.652	0.762	0.788	0.905	0.985	0.995	0.991	0.671
		PPV	1.000	0.943	0.911	0.975	0.996	0.814	0.734	0.711	0.670	0.531	0.550	0.505	0.441	0.393	0.707
Elliptic window	1	Frequency	10	4	7	3	5	26	21	42	45	31	21	15	23	8	
		Sensitivity	0.100	0.154	0.187	0.205	0.292	0.464	0.571	0.672	0.776	0.864	0.901	0.928	0.906	0.942	0.671
		PPV	0.650	0.667	0.607	0.533	0.633	0.821	0.777	0.719	0.649	0.593	0.541	0.464	0.407	0.371	0.628
	2	Frequency	1	4	5	11	7	35	89	346	258	85	58	10	10	1	
		Sensitivity	0.154	0.231	0.292	0.371	0.516	0.527	0.711	0.880	0.957	0.961	0.987	0.992	0.977	0.923	0.873
		PPV	1.000	1.000	0.950	0.964	0.988	0.938	0.962	0.919	0.809	0.674	0.591	0.506	0.443	0.353	0.840
	3	Frequency	-	1	1	2	12	16	46	538	314	38	32	-	-	-	
		Sensitivity	-	0.231	0.615	0.385	0.821	0.582	0.736	0.911	0.989	0.996	1.000	-	-	-	0.926
		PPV	-	1.000	1.000	1.000	1.000	0.992	0.993	0.941	0.839	0.694	0.604	-	-	-	0.892

PPV = positive predictive value.

to the true cluster size with a higher accuracy than the default setting. Even when the most often picked optimal MRCS did not exactly agree with the true cluster size, the accuracy of the detected clusters was still generally higher than the results from the default setting. The Gini coefficient with circular windows when the true clusters were elliptic-shaped seems not to work very well. The most frequently picked MRCS as best size was usually larger than the true cluster size, in which case the PPV was lower than the default setting despite the higher sensitivity. The frequency of each MRCS to be chosen as the optimal size was also distributed over all sizes considered, while it was usually concentrated concentrated around the size of the true cluster when using elliptic windows.

4. Application to real data

We applied the proposed method to a real data set of female colorectal cancer incidence rates in South Korea for the year 2009. The age-standardized incidence rates per 100,000 at the “Si-gun-gu”

Table 7: Simulation results for cluster model G (elliptic cluster, 30% of the whole study region)

	μ_z		Maximum reported cluster size													Default
			3%	5%	6%	8%	10%	12%	15%	20%	25%	30%	35%	40%	45%	50%	
Circular window	1	Frequency	1	13	17	6	18	29	41	64	77	52	59	37	52	33	
		Sensitivity	0.100	0.131	0.197	0.242	0.289	0.360	0.441	0.541	0.616	0.677	0.781	0.792	0.858	0.879	0.615
		PPV	1.000	0.872	0.985	0.967	0.963	0.945	0.934	0.912	0.804	0.712	0.680	0.609	0.580	0.528	0.773
	2	Frequency	-	1	1	-	9	18	53	210	167	105	215	62	125	23	
		Sensitivity	-	0.150	0.200	-	0.361	0.436	0.507	0.578	0.648	0.768	0.830	0.837	0.896	0.898	0.720
		PPV	-	1.000	1.000	-	0.981	0.955	0.979	0.971	0.844	0.782	0.744	0.646	0.614	0.544	0.803
	3	Frequency	-	-	1	-	5	33	18	212	158	150	264	30	123	6	
		Sensitivity	-	-	0.200	-	0.510	0.645	0.647	0.589	0.669	0.794	0.846	0.850	0.900	0.908	0.754
		PPV	-	-	1.000	-	1.000	0.982	0.969	0.990	0.886	0.798	0.760	0.654	0.618	0.551	0.816
Elliptic window	1	Frequency	4	9	8	9	15	23	28	59	101	68	69	47	35	32	
		Sensitivity	0.100	0.150	0.175	0.239	0.280	0.365	0.418	0.566	0.682	0.783	0.814	0.871	0.890	0.934	0.678
		PPV	1.000	1.000	0.875	0.956	0.933	0.954	0.877	0.933	0.883	0.830	0.729	0.666	0.603	0.568	0.806
	2	Frequency	-	1	-	-	1	6	13	62	244	372	164	94	30	10	
		Sensitivity	-	0.150	-	-	0.300	0.375	0.481	0.602	0.771	0.883	0.920	0.954	0.958	0.955	0.845
		PPV	-	1.000	-	-	1.000	0.979	0.985	0.989	0.975	0.951	0.837	0.737	0.660	0.575	0.903
	3	Frequency	-	-	-	-	-	3	-	11	146	667	120	50	3	-	
		Sensitivity	-	-	-	-	-	0.633	-	0.591	0.792	0.896	0.964	0.989	0.983	-	0.891
		PPV	-	-	-	-	-	1.000	-	1.000	0.986	0.978	0.904	0.770	0.702	-	0.956

PPV = positive predictive value.

Table 8: Simulation results for cluster model H (elliptic cluster, 50% of the whole study region)

	μ_z		Maximum reported cluster size													Default
			3%	5%	6%	8%	10%	12%	15%	20%	25%	30%	35%	40%	45%	50%	
Circular window	1	Frequency	2	1	7	5	5	6	12	24	44	43	69	82	155	160	
		Sensitivity	0.059	0.088	0.118	0.147	0.159	0.211	0.270	0.337	0.440	0.531	0.644	0.729	0.822	0.894	0.706
		PPV	1.000	1.000	1.000	1.000	0.900	0.955	0.965	0.975	0.70	0.949	0.958	0.945	0.943	0.914	0.939
	2	Frequency	-	-	-	-	-	-	1	1	1	5	24	40	254	673	
		Sensitivity	-	-	-	-	-	-	0.294	0.353	0.471	0.576	0.662	0.764	0.865	0.946	0.912
		PPV	-	-	-	-	-	-	1.000	1.000	1.000	1.000	0.982	0.989	0.986	0.970	0.972
	3	Frequency	-	-	-	-	-	-	-	-	-	-	-	2	96	902	
		Sensitivity	-	-	-	-	-	-	-	-	-	-	-	0.794	0.884	0.966	0.959
		PPV	-	-	-	-	-	-	-	-	-	-	-	1.000	0.991	0.986	0.986
Elliptic window	1	Frequency	-	3	10	5	7	7	10	16	38	33	72	90	138	183	
		Sensitivity	-	0.088	0.109	0.147	0.160	0.197	0.274	0.344	0.433	0.526	0.628	0.722	0.787	0.841	0.685
		PPV	-	1.000	0.925	1.000	0.905	0.908	0.950	0.955	0.948	0.950	0.949	0.941	0.900	0.857	0.909
	2	Frequency	-	-	-	-	-	-	-	-	4	9	43	106	186	648	
		Sensitivity	-	-	-	-	-	-	-	-	0.434	0.556	0.650	0.751	0.805	0.867	0.830
		PPV	-	-	-	-	-	-	-	-	0.926	0.994	0.969	0.977	0.928	0.875	0.901
	3	Frequency	-	-	-	-	-	-	-	-	-	1	4	28	132	835	
		Sensitivity	-	-	-	-	-	-	-	-	-	0.588	0.647	0.761	0.811	0.873	0.860
		PPV	-	-	-	-	-	-	-	-	-	1.000	0.968	0.989	0.941	0.876	0.889

PPV = positive predictive value.

(district) level were obtained from the National Cancer Center. The incidence data at the individual level were not provided due to confidentiality concerns. The age-standardized incidence rates are continuous summary measures with varying regional uncertainty. A weighted normal spatial scan statistic should be used to search for clusters of districts with unusual high incidence rates. We used the 2010 population at each district as the weight of each region. We excluded incidence rates in five districts, including four districts of Jeju Island and one of Ulleung Island, among 251 districts because these islands are far from the mainland.

To select the best MRCS based on the proposed method, we used 17 different values (1–6, 8, 10, 12, 15, 20, 25, 30, 35, 40, 45, and 50%) of MRCS and chose the MRCS with the highest values as the optimal one. The clusters at the optimal MRCS were compared with the results of the default setting (50%). We used both the circular and elliptic windows.

Figure 3 shows the detected clusters at the optimal MRCS and at the default setting when using circular and elliptic windows. The Gini coefficient selected 50% as the optimal MRCS when using

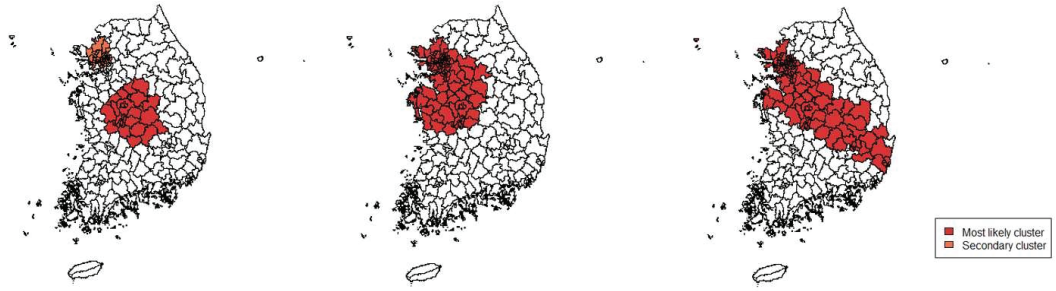

Figure 3: *Detected clusters with high values of female colorectal cancer incidence rates at the optimal MRCS based on the Gini coefficient and at the default setting. MRCS = maximum reported cluster size.*

Table 9: Detected clusters with high incidence rates of female colorectal cancer using the circular spatial scan statistic

	Cluster	Number of districts	*p*-value	Weighted mean*
Gini	Most likely	27	0.016	29.89
(20%)	Secondary	46	0.020	28.63
Default setting	Most likely	105	0.006	28.29

*Weighted mean of incidence rates per 100,000.

the elliptic window. The most likely cluster was large, covering from the northwest areas including Seoul to the southeast areas. However, 20% was selected as the best MRCS and two significant were found when using the circular windows, while only one cluster was detected at the default setting. Table 9 provides information on the detected clusters. The most likely cluster at the default setting included 105 districts, while the most likely and secondary clusters at the optimal MRCS based on the Gini coefficient included 27 and 46 districts, respectively. Similar to the simulation study, the default setting seemed to detect a larger cluster by absorbing neighboring areas with irrelevant risk. The weighted means of incidence rates for the clusters at the optimal MRCS were 29.89 and 28.63, respectively, which were higher than that for the most likely cluster at the default setting of 28.29. The weighted mean of incidence rates on the entire study region was 27.70. The clusters found at the optimal MRCS look more meaningful.

5. Discussion and conclusion

In this paper, we defined the Gini coefficient to evaluate the degree of heterogeneity of clusters for continuous data. Through the simulation study and the real data example, we showed that the proposed method can be useful to optimize the MRCS for normal-based spatial scan statistics. In the simulation study results, the Gini coefficient most often selected the optimal value of MRCS similar to the true cluster size. The accuracy of the detected clusters was also consistently higher at the most frequently chosen MRCS as best compared to the results from the default setting. Regarding the scanning window shape, elliptic windows seemed to work well regardless of the shape of true clusters. The real data example shows that it is possible to obtain a more meaningful and informative collection of clusters when using the Gini coefficient than when using the default setting.

Many users of SaTScan^{TM} are often tempted to rerun the analyses using different MSWS when they do not like the results at the default setting. However, we emphasize that this approach should

not be used because it causes a multiple testing problem. It is okay to filter clusters at different MRCS to choose the clusters to report. In the process, the Gini coefficient is a very useful criterion. The Gini coefficient has been implemented into SaTScan^{TM} for the Poisson model only. We think that it will be valuable to add the same option to the other models of spatial scan statistics such as the normal model that can help researchers find a more refined collection of clusters to report.

References

Duczmal L and Assunção R (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters, *Computational Statistics & Data Analysis*, **45**, 269–286.

Dwass M (1957). Modified randomization tests for nonparametric hypotheses, *The Annals of Mathematical Statistics*, **20**, 181–187.

Gastwirth JL (1972). The estimation of the Lorenz curve and Gini index, *The Review of Economics and Statistics*, **54**, 306–316.

Han J, Zhu L, Kulldorff M, Hostovich S, Stinchcomb DG, Tatalovich Z, Lewis DR, and Feuer EJ (2016). Using Gini coefficient to determining optimal cluster reporting sizes for spatial scan statistics, *International Journal of Health Geographics*, **15**, 27.

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

Huang L, Tiwari RC, Zou Z, Kulldorff M, and Feuer EJ (2009). Weighted normal spatial scan statistic for heterogeneous population data, *Journal of the American Statistical Association*, **104**, 886–898.

Jung I, Kulldorff M, and Klassen AC (2007). A spatial scan statistic for ordinal data, *Statistics in Medicine*, **26**, 1594–1607.

Jung I, Kulldorff M, and Richard OJ (2010). A spatial scan statistic for multinomial data, *Statistics in Medicine*, **29**, 1910–1918.

Kim S and Jung I (2017). Optimizing the maximum reported cluster size in the spatial scan statistic for ordinal data, *PLoS ONE*, **12**, e0182234.

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

Kulldorff M and Information Management Services Inc (2018). SaTScan^{TM} v9.5: Software for the spatial and space-time spatial scan statistics, from: http://www.satscan.org/

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

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

Lorenz MO (1905). Methods of measuring the concentration of wealth, *Publications of the American Statistical Association*, **9**, 209–219.

Ribeiro SHR and Costa MA (2012). Optimal selection of the spatial scan parameters for cluster detection: a simulation study, *Spatial and Spatio-Temporal Epidemiology*, **3**, 107–120.

Patil GP and Taillie C (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots, *Environmental and Ecological Statistics*, **11**, 183–197.

Tango T and Takahashi K (2005). A flexibly shaped spatial scan statistic for detecting clusters, *International Journal of Health Geographics*, **4**, 11.

Received February 28, 2018; Revised May 15, 2018; Accepted June 12, 2018