Chapter 5

Comparison of Cluster Detection Methods on Synthetic Data Sets

In this chapter, we compare our cluster detection method with Nagarwalla's Spatial Scan Statistic by tested on 100 synthetic data sets. We compare with the scan statistic because it can handle aggregate data and it assume only one cluster exists. In this part, we divide the data sets into two groups, Group 1 contains a cluster with sinuous shape and Group 2 contains a cluster with circular shape; the relative risk is 2.0. When applying the SaTScan software, we choose maximum spatial cluster size to be 30 percent of population at risk.

From table 5.1 (Detailed tables are given in Appendix B) where the definitions of *a*, *b*, and *c* are stated in Section 4.1, we found that the average error rate induced by using the scan statistic is very large. Under the significance level of 0.05, the average error rate is 0.8903 for Group 1 and 0.7893 for Group 2; the corresponding standard error is 0.1659 and 0.2948. By using our proposed method for cluster detection, under the significance level of 0.05, the average error rate is 0.0953 for Group 1 and 0.0825 for Group 2; the corresponding standard error is 0.0996 and 0.1002.

For both groups of synthetic data sets, our proposed method produces more satisfactory results than Nagarwalla's Spatial Scan Statistic. As pointed out by Tango (2000), if there are actually many small clusters in the study area, the SaTScan program tends to detect one unrealistically larger cluster than expected because it will detect one large cluster which encompasses the small clusters and those areas outside the clusters which do not have elevated risk.

Due to lack of time, we do not apply other methods for comparison; however, the results of Smith (2002) are given in table 5.2 to give the readers a picture on the performance of other methods. From Smith's results where individual data are analyzed, when he tried to achieve balance between Type I and Type II errors,

GAM/K has the highest sensitivity (38.81%), but the SaTScan has the highest specificity (99.1%) and PPV (56.25%). As a result, he concluded that the spatial scan statistic also failed to achieve a high sensitivity (proportion of the true positives that a screening test can successfully detect) because of the shape and the low relative risk (RR=2) of the actual cluster. Although we do not apply our proposed method on Smith's data, based on Group 1 data sets where sinuous clusters exist, our PPV is 93.09%, Specificity is 99.80% and Sensitivity is 96.67%.

	SaTScan		Our method	
	Data Sets containing a cluster with sinuous shape	Data Sets containing a cluster with circular shape	Data Sets containing a cluster with sinuous shape	Data Sets containing a cluster with circular shape
Average Significant Cluster Size	5.16	6.48	9.38	9.62
Mean of *a*	1.56	2.50	8.70	8.86
Mean of *b*	7.44	6.50	0.30	0.14
Mean of *c*	3.60	3.98	0.68	0.76
Mean of *d*	337.4	337.02	340.32	340.24
Average Error Rate	0.8903	0.7893	0.0953	0.0825
Sensitivity	16.67%	27.78%	96.67%	97.78%
Specificity	98.94%	98.83%	99.80%	99.78%
PPV	30.23%	38.58%	93.09%	92.10%

Table 5.1: Summary of Comparison Results (*a*: True positive, *b*: False negative, *c*: False positive)

Test	*a*	*c*	*b*	*d*	Sensitivity	Specificity	PPV
SaTScan	9	7	1011	7662	0.88%	99.91%	56.25%
GAM/K	397	645	626	7019	38.81%	91.58%	38.10%
Significance Map	359	663	661	7006	35.20%	91.35%	35.13%

Table 5.2: Comparison of the Three Cluster Detection Methods: Maximize Sensitivity and Positive Predictive Value, Smith (2002)