Chapter 4

Simulations

In this chapter, we conduct a series of simulations to examine the performance of method introduced in Chapter 3. With these simulations, we plan to achieve the following goals:

1. Compare the results from the test of clustering with which illustrated by Choynowski (1959) by using aggregate data.
2. Determine if the shape, total population size and location of the clusters will influence the performance of our proposed method.

4.1 Introduction and Background

In the following simulations, we consider two kinds of models: one contains no cluster and the other contains one cluster. For the model of no cluster, we assume that tracts can have either constant or non-constant population size. In the case of constant population size, we simulate three models; regions in each model have identical number of cases within the whole study area. On the other hand, in the case of non-constant population size, in addition to models with identical number of cases, we simulate three models with different intensity rates. Based on these models, we can then compare the results from the real world case and the ideal case, and thus find out the performance of the clustering test corresponding to the influences of unequal variance in population at risk in each region.

On the other hand, for the simulation settings containing one cluster, we only consider the non-constant population strategy since this is closely linked with the real world situation. Three intensity rates similar to the no cluster model are assumed. As mentioned in the previous section, our suggested method does not require circular

cluster assumption. Therefore, in the simulated cluster model, we purposely designed synthetic data sets with long and sinuous clusters (each includes nine townships), to test the susceptibility of the test. Meanwhile, we also test our proposed method on synthetic cluster with other characteristics: shape — circular and location — populous regions, along riverbank.

Using the following steps to generate simulated data sets:

1. Construct neighborhood information for 350 townships or cities in Taiwan (peninsula). Assign two sets of population strategy to each township: constant – 1000 for each township; non-constant – real population at risk. According to the "2000 Population and Housing Census", approximate total population in whole Taiwan area is 22,000,000.
2. For the no cluster models with constant population size, we select three background number of cases: 50, 100 and 200; for the no cluster models with non-constant population size, the intensity rates under considerations are 1/500, 1/1000 and 1/2000. Furthermore, since we would like to measure the Type II error rate in the test of clustering, we consider models with 50 clustering regions and 100 clustering regions for each intensity rate; the relative risk for those clustering regions are 1.2, 1.6 and 2.0 respectively. Then, we use the Poisson random number generator in S-plus to simulate 10000 data sets for each null model.
3. For the models with one largest cluster, we assign four different characteristics for the synthetic cluster: Similar to no cluster models with non-constant population strategy, the intensity rates in consideration are 1/500, 1/1000 and 1/2000. The relative risk for the synthetic cluster is an additional parameter needed for the cluster models. So, we then select the values of 1.2, 1.6 and 2.0 as relative risks. For these models, we simulated 1000 data sets each.

From the procedures above, we finally simulate 6 models with no cluster (2 levels of population size strategy, 3 levels of number of cases and 3 levels of intensity

rate), 18 models with cluster (2 levels of population size strategy, 3 levels of number of cases and 3 levels of intensity rate) and 36 models with one largest cluster (3 levels of intensity rate, 3 levels of relative risk and 4 levels of cluster characteristics). In Appendix A, we list the no cluster models and one cluster models simulated.

In order to investigate our proposed methods' performance to detect the "real" cluster, we are interested in the Type I error, Type II error and the corresponding statistical power of the each simulation model. When we report the "false" cluster, we make the Type I error; when we miss the "true" cluster, we make the Type II error. The higher the probability of making Type I error is, the higher the statistical power and the less in the probability of making Type II error will be.

For the no cluster models' simulation, we would like to confirm the fact that the probability of the existence of clustering is consistent with the significance level used. Next, we then figure out the Type I error made based on the maximum number of regions linked and on the maximum size of population at risk.

For the simulations in models with one cluster, based on these synthetic data, we measured the Type II error made for each model. However, we should note in advance that in some situations, it is unreasonable to judge the performance of the method by using the Type II error; for an example, a "false" cluster may be detected but regions from the "real" cluster probably contained in it. So, we introduce another measurement for the error rate: Suppose the "real" cluster contains r regions and we successfully detect a "false" cluster with f regions in which a of them are from the "real" cluster, c of them are excluded and finally we let b be the number of regions in the "real" cluster that are not detected (as illustrated in Figure 4.1); so, we define the error rate as $(b+c)/(a+b+c)$ for each simulation and an average error rate can be calculated by each simulation model.

	"Truth"		
	Cluster	Not Cluster	
Screening Test Positive	*a* (True positive)	*c* (False positive)	PPV *a*/(*a*+*c*)
Screening Test Negative	*b* (False negative)	*d* (True negative)	NPV *d*/(*b*+*d*)
	Sensitivity *a*/(*a*+*b*)	Specitivity *d*/(*c*+*d*)	

Figure 4.1: Evaluating Screening and Diagnostic Tests

4.2 Procedures for Simulations

We explain the procedures for our simulations in the following sections.

4.2.1 Simulations for Models with No Cluster

In this section, we will first figure out the number of regions with high occurrences in each data set. Then, we check whether any cluster is detected in those regions and record the number of times that the cluster found in each model.

For the simulated models with no cluster and clustering regions, procedures are as follows:

1. By using the data sets simulated, check whether the case simulated in each region is significantly larger than the critical value. Record the number of regions with clustering in each data set.
2. Draw out data sets with significant clustering; find out the size and population at risk of each set of regions linked.
3. Based on the results Monte Carlo simulation mentioned in Section 3.3, determine whether there exists a significant cluster by using either size or population at risk of the set found in step 2.
4. Calculate the Type I error.

For the simulated models with no cluster but clustering regions, procedures are as follows:

1. By using the data sets simulated, check whether the case simulated in each region is significantly larger than the critical value. Record the number of regions with clustering in each data set.
2. Calculate the Type II error.

4.2.2 Simulations for Models with One Cluster

For these models, we record the number of times the specially designated cluster is successfully detected in each model.

1. By using the data sets simulated, check whether the numbers of cases simulated in each region is significantly larger than the critical value assumed.
2. Find out a set of regions with maximum number of regions linked; take down the size of the set. Check whether the designated cluster is successfully detected.
3. Calculate the Type II error and average error rate.

4.3 Simulation Results

In this section, we illustrate our simulation results accompanied by some maps showing the location of the synthetic cluster.

4.3.1 Simulations for Models with No Cluster

From table 4.3.1(a), we notice that the simulated Type I errors of clustering are fairly close to those of true significance level in most cases. However, the results from significance level 0.01 have largest deviations probably because the expected mean (3.5) for regions having clustering is too less. Generally speaking, the results from

simulation closely match to the theoretic results, regardless the population sizes and intensity rates. On the other hand, table 4.3.1(b) and 4.3.1(c) show simulated Type II errors of clustering. In models either 50 or 100 clustering regions are assumed, both simulated Type II errors are very large, as expected. These results indicate that the idea from Choynowski can detect clustering effectively in most of the occasions.

Strategy in Population Size	Number of Cases/ Intensity Rate	Significance Level		
		0.01	0.05	0.10
Constant	50	0.0131	0.0548	0.0973
	100	0.0190	0.0497	0.1196
	200	0.0296	0.0457	0.1047
Non-constant	1/500	0.0211	0.0424	0.0846
	1/1000	0.0263	0.0505	0.1137
	1/2000	0.0183	0.0644	0.0986

Table 4.3.1(a): Simulated Type I Errors in Detecting Clustering (10,000 Simulation Runs)

Incidence Rate	RR	0.01	0.05	0.10
1/500	1.2	0.9911	0.9532	0.8932
	1.6	0.9971	0.9561	0.8711
	2	0.9991	0.9342	0.8702
1/1000	1.2	0.9942	0.9430	0.8925
	1.6	0.9875	0.9500	0.9000
	2	0.9923	0.9291	0.8731
1/2000	1.2	0.9904	0.9532	0.9028
	1.6	0.9861	0.9554	0.8967
	2	0.9915	0.9433	0.9171

Table 4.3.1(b): Simulated Type II Errors in Detecting Clustering in study area with 50 clustering regions (10,000 Simulation Runs)

Incidence Rate	RR	0.01	0.05	0.10
1/500	1.2	0.9818	0.9532	0.8890
	1.6	0.9872	0.9345	0.8971
	2	0.9974	0.9592	0.9182
1/1000	1.2	0.9861	0.9421	0.9054
	1.6	0.9880	0.9332	0.8957
	2	0.9932	0.9610	0.9206
1/2000	1.2	0.9813	0.9380	0.8941
	1.6	0.9901	0.9404	0.8832
	2	0.9897	0.9512	0.9130

Table 4.3.1(b): Simulated Type II Errors in Detecting Clustering in study area with 100 clustering regions (10,000 Simulation Runs)

Table 4.3.1(e) and (e) show the simulated Type I errors under two criterions, stated in section 3.3. In the model with constant identical population size in each tract, the simulated Type I error for both criterions are the same. From these numbers, surprisingly, we find that the simulated Type I errors under both criterions are very low. Since Type II error is inversely related to Type I error, we may expect that the Type II errors would likely to be large.

Note that we gain similar results from the two settings of population assumption and the non-constant population assumption reflect the real world case; so, it seems to be more reasonable that we simply apply the non-constant population size assumption in the models with one cluster.

Strategy in Population Size	Number of Cases/ Intensity Rate	Significance Level		
		0.01	0.05	0.10
Constant	50	0.0000	0.0034	0.0074
	100	0.0004	0.0022	0.0071
	200	0.0004	0.0026	0.0070
Non-constant	1/500	0.0007	0.0033	0.0060
	1/1000	0.0008	0.0036	0.0074
	1/2000	0.0001	0.0030	0.0100

Table 4.3.1(d): Simulated Type I Errors in Cluster Detection Based on the Maximum Number of Regions Linked (10,000 Simulation Runs)

Strategy in Population Size	Number of Cases/ Intensity Rate	Significance Level		
		0.01	0.05	0.10
Constant	50	0.0000	0.0034	0.0074
	100	0.0004	0.0022	0.0071
	200	0.0004	0.0026	0.0070
Non-constant	1/500	0.0018	0.0012	0.0020
	1/1000	0.0011	0.0010	0.0016
	1/2000	0.0007	0.0018	0.0012

Table 4.3.1(e): Simulated Type I Errors in Cluster Detection Based on the Maximum Size of Population at Risk (10,000 Simulation Runs)

4.3.2 Simulations for Models with One Cluster– Specified Shapes

In this section, we will evaluate the performance of our proposed method by showing both Type II error rate and average error rate defined earlier. Locations for the specified clusters will be drawn and corresponding population sizes of the regions within the clusters are given.

From both table 4.3.2(a) and table 4.3.2(b), we notice that when the larger the relative risk, the lesser the probability of making Type II error and the average error rate, just as what we expected.

For the models with long and sinuous cluster, when we set a higher threshold to detect the regions with significant high occurrences, the prespecified cluster can more accurately be detected and then reduce the Type II error rate and average error rate. For an example, from table 4.3.2(a), when the mortality rate is 1/500 and the relative risk is 2, under significance level of 0.10, the corresponding Type II error rate and average error rate are almost twice as error rate under significance level of 0.05.

For the models with circular cluster, from table 4.3.2(b), both the Type II error rate and average error rate are getting lesser than in table 4.3.2(a), these results show that the precision of cluster detected will increase when the cluster has regular shape, as expected.

By using aggregate data, under the assumption of only one cluster exists, these results indicate that our suggested method can detect cluster in either circular or irregular shape especially when the relative risk is large.

Intensity Rate	Simulation Model	Significance Level					
		0.01		0.05		0.10	
		Type II Error Rate	Average Error Rate	Type II Error Rate	Average Error Rate	Type II Error Rate	Average Error Rate
1/500	1.2	1.000	0.756	0.999	0.562	0.997	0.517
	1.6	0.548	0.078	0.684	0.116	0.869	0.206
	2.0	0.173	0.019	0.542	0.080	0.821	0.172
1/1000	1.2	1.000	0.900	1.000	0.785	1.000	0.724
	1.6	0.877	0.196	0.855	0.183	0.952	0.268
	2.0	0.429	0.058	0.663	0.106	0.886	0.204
1/2000	1.2	1.000	0.959	1.000	0.896	1.000	0.850
	1.6	0.994	0.450	0.965	0.328	0.982	0.356
	2.0	0.772	0.139	0.806	0.159	0.932	0.258

Table 4.3.2 (a): Simulated Type II Errors in Detecting Long and Sinuous Cluster (1,000 Simulation Runs)

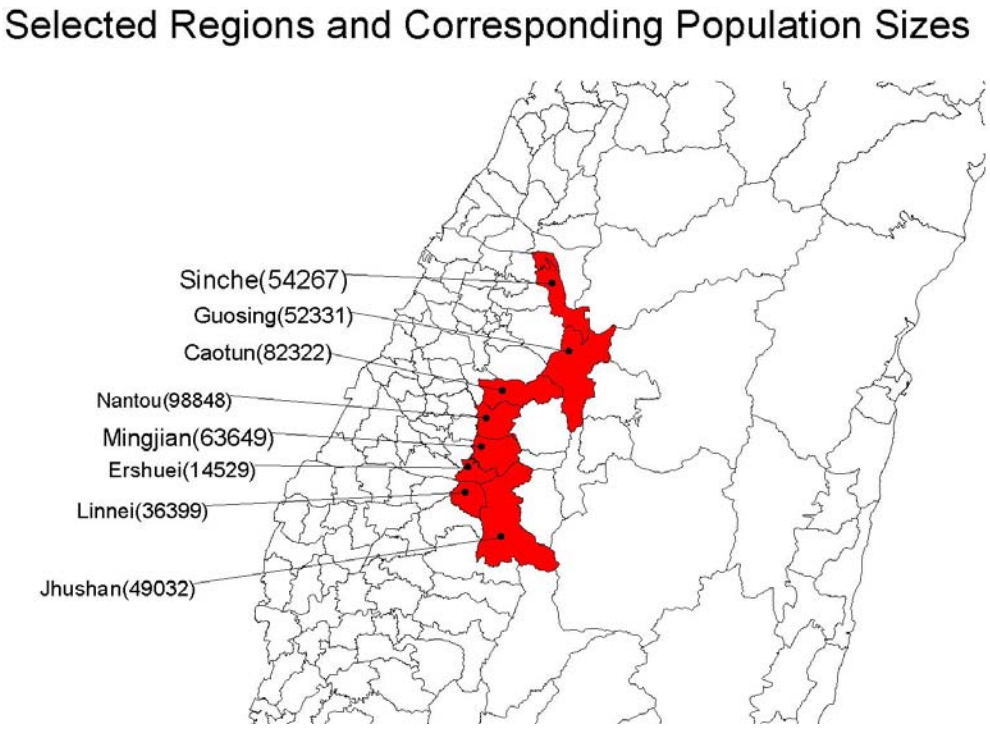

Figure 4.3.2(a): Location of Long and Sinuous Cluster for Simulation Model 25 to 33.

Intensity Rate	Simulation Model	Significance Level					
		0.01		0.05		0.10	
		Type II Error Rate	Average Error Rate	Type II Error Rate	Average Error Rate	Type II Error Rate	Average Error Rate
1/500	1.2	1.000	0.738	0.996	0.524	0.985	0.440
	1.6	0.233	0.032	0.458	0.065	0.693	0.134
	2.0	0.090	0.009	0.404	0.054	0.634	0.114
1/1000	1.2	1.000	0.902	1.000	0.759	0.999	0.686
	1.6	0.742	0.152	0.668	0.120	0.812	0.181
	2.0	0.151	0.017	0.467	0.063	0.692	0.132
1/2000	1.2	1.000	0.955	1.000	0.880	1.000	0.829
	1.6	0.984	0.429	0.904	0.251	0.919	0.275
	2.0	0.552	0.091	0.596	0.096	0.795	0.178

Table 4.3.2 (b): Simulated Type II Errors in Detecting Circular Cluster (1,000 Simulation Runs)

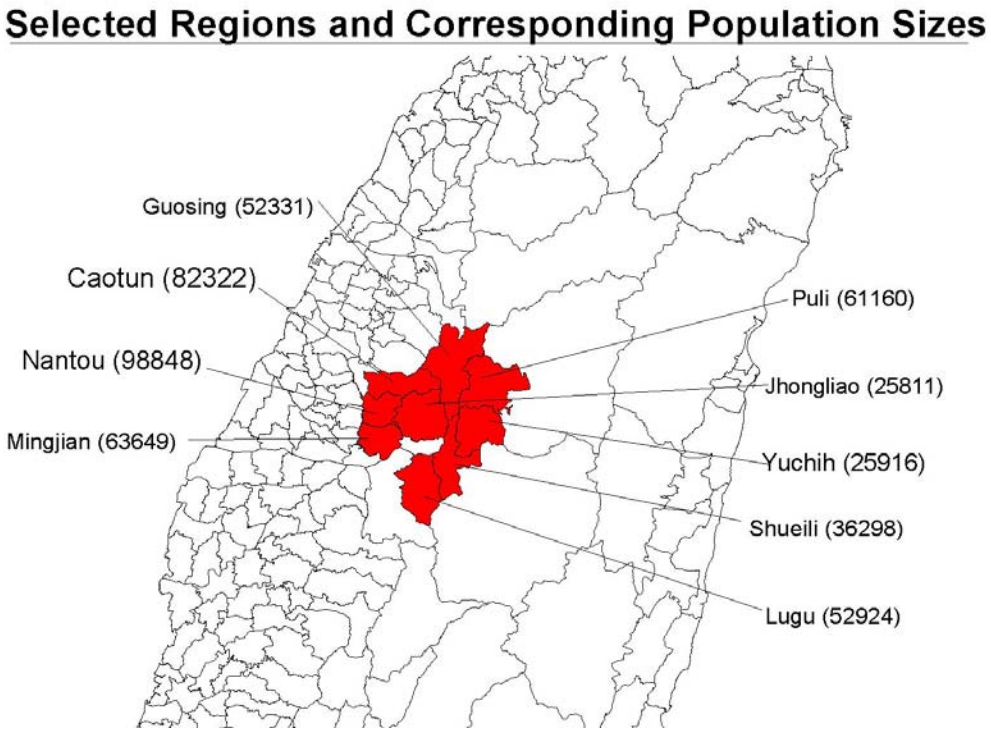

Figure 4.3.2(b): Location of Circular Cluster for Simulation Model 34 to 42.

For further investigation, we may note that the population size for each township may influence the performance of the test to detect true cluster, as we can see in table 4.3.2(a), at the significance level of 0.01, relative risk of 2.0, the average error rate of mortality rates 1/2000 is seven times the error rate at mortality rate 1/500. In other words, the discontinuity of the population size may exert an influence especially in those regions with sparse population density. From the Taiwan's census data, we found out that there are 4 townships with population size less than 2500, 21 townships with population size less than 5000 and 51 townships with population size less than 10000 (figure 4.3.2(d)); these townships may also have effect upon the probability that clustering exists (figure 4.3.2(b)) and the detection of cluster.

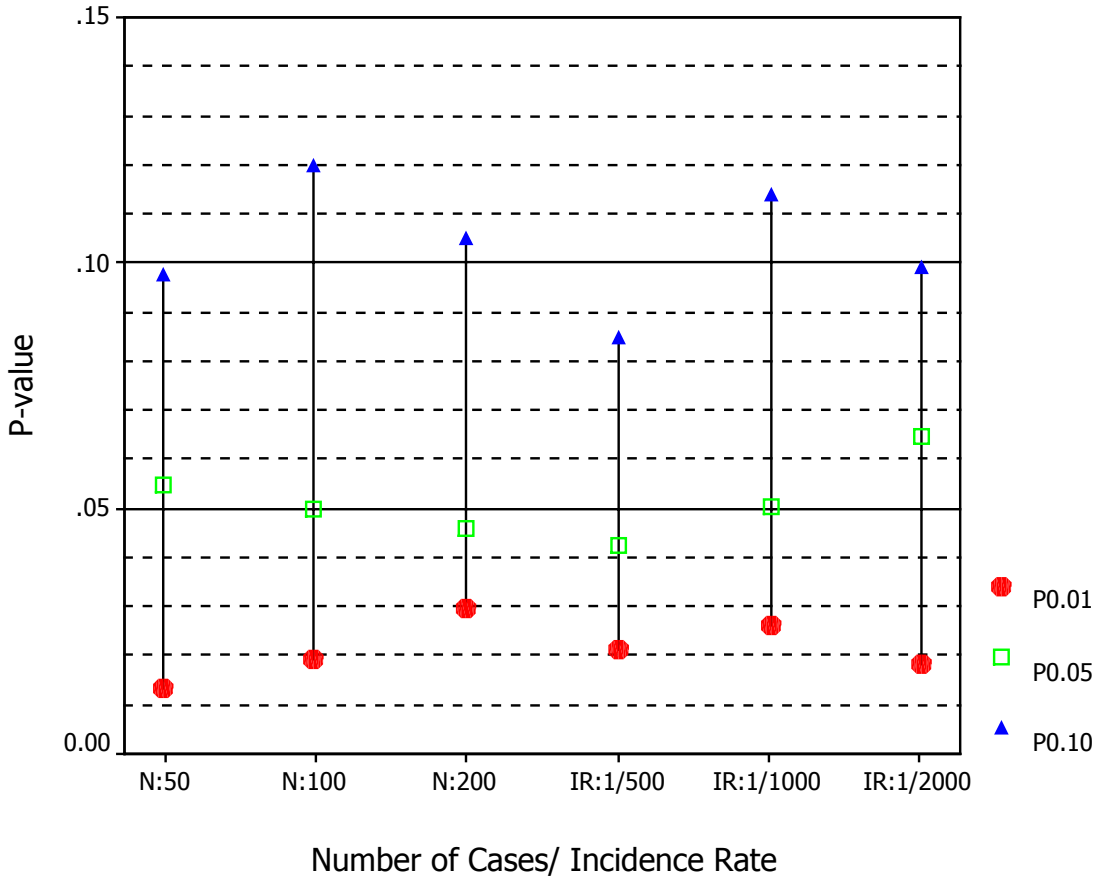

Figure 4.3.2(c): Chart Showing the Probability that Clustering Exists in Each No Cluster Model with Different Significance Level.

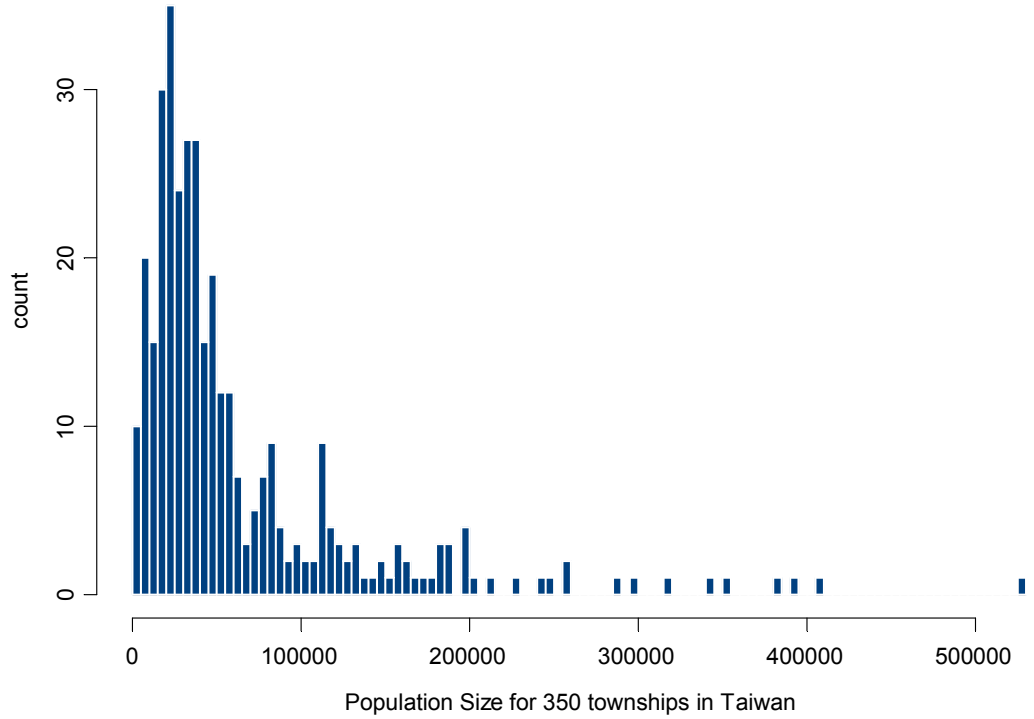

Figure 4.3.2(d): Histogram of the Population Size for 350 Townships in Taiwan.

4.3.3 Simulations for Models with One Cluster– Specified Locations

In this section, we will evaluate the performance of our proposed method by locating the designated cluster in a populous area and somewhere along downstream of a river. Similar to what we have done in Section 4.3.2, we show both Type II error rate and average error rate defined earlier and auxiliary maps are drawn.

From both table 4.3.3(a) and table 4.3.3(b), we notice that when the larger the relative risk, the lesser the probability of making Type II error and the average error rate, just as what we have shown in Section 4.3.2.

From table 4.3.3(a), the Type II error rate and average error rate are similar to the results from table 4.3.2(a), where a long and sinuous cluster is pre-selected. Both the error rate will get less if we have larger relative risk and lower significance level.

For the models with cluster located in populous regions, the results from Type II error rate and average error rate are encouraging. Hence, we may conjecture that the

chance to detect a cluster within populated regions is very high.

Intensity Rate	Simulation Model	Significance Level					
		0.01		0.05		0.10	
		Type II Error Rate	Average Error Rate	Type II Error Rate	Average Error Rate	Type II Error Rate	Average Error Rate
1/500	1.2	1.000	0.851	1.000	0.719	1.000	0.670
	1.6	0.522	0.148	0.630	0.124	0.853	0.185
	2.0	0.141	0.017	0.536	0.074	0.801	0.149
1/1000	1.2	1.000	0.938	1.000	0.859	1.000	0.827
	1.6	0.943	0.419	0.840	0.259	0.904	0.279
	2.0	0.350	0.088	0.597	0.102	0.834	0.173
1/2000	1.2	1.000	0.975	1.000	0.929	1.000	0.904
	1.6	0.999	0.658	0.986	0.500	0.979	0.472
	2.0	0.859	0.318	0.805	0.209	0.908	0.256

Table 4.3.3(a): Simulated Type II Errors in Detecting Cluster Located along River (1,000 Simulation Runs)

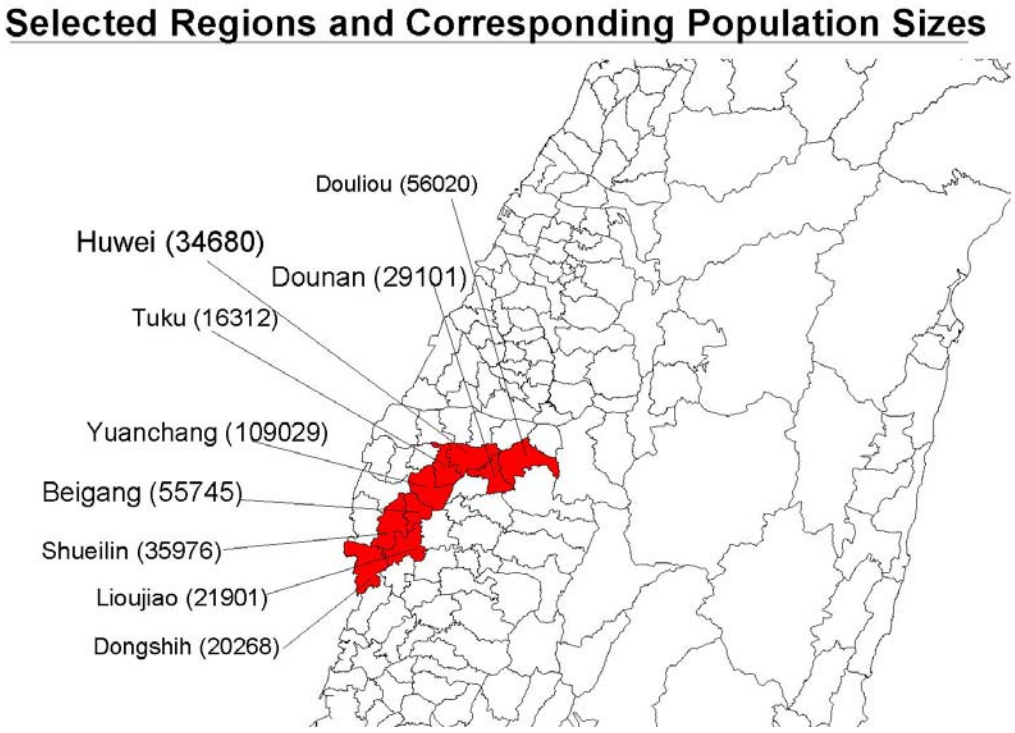

Figure 4.3.3(a): Location of Cluster along River for Simulation Model 43 to 51.

Intensity Rate	Simulation Model	Significance Level					
		0.01		0.05		0.10	
		Type II Error Rate	Average Error Rate	Type II Error Rate	Average Error Rate	Type II Error Rate	Average Error Rate
1/500	1.2	0.772	0.182	0.534	0.089	0.663	0.113
	1.6	0.019	0.002	0.115	0.012	0.263	0.032
	2.0	0.005	0.001	0.054	0.006	0.150	0.017
1/1000	1.2	0.993	0.543	0.929	0.298	0.901	0.255
	1.6	0.048	0.005	0.220	0.026	0.438	0.062
	2.0	0.013	0.001	0.123	0.013	0.241	0.028
1/2000	1.2	1.000	0.823	1.000	0.623	0.988	0.515
	1.6	0.166	0.022	0.323	0.040	0.555	0.082
	2.0	0.033	0.003	0.172	0.020	0.336	0.046

Table 4.3.3 (b): Simulated Type II Errors in Detecting Cluster Located in Populous Regions (1,000 Simulation Runs)

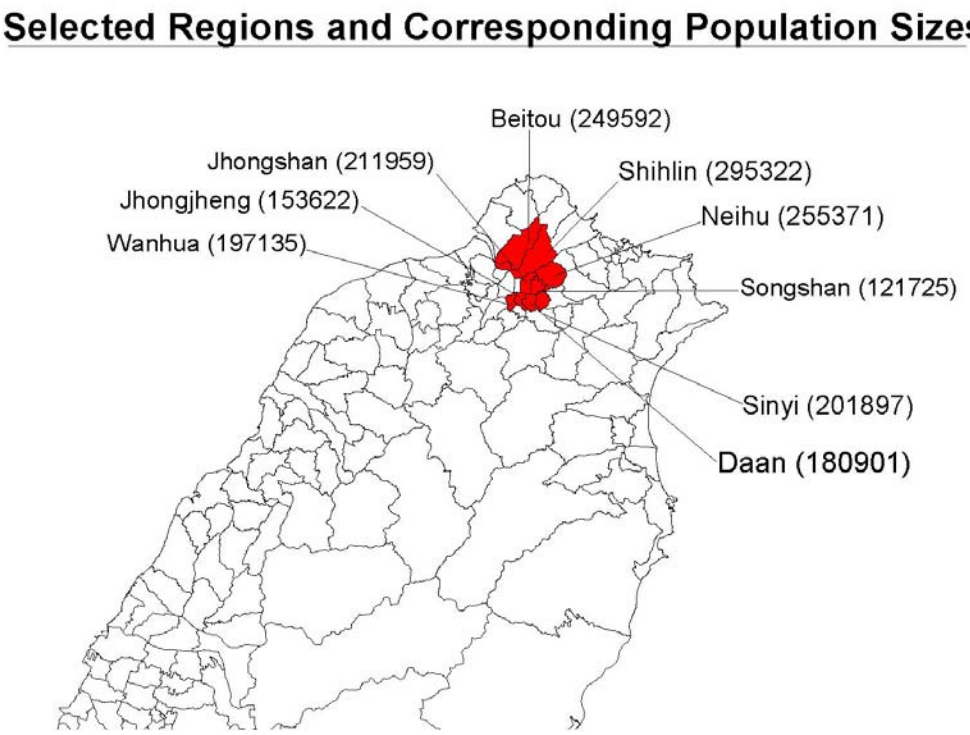

Figure 4.3.3(b): Location of Cluster Located in Populous Regions for Simulation Model 52 to 60.