

A method for testing low-value spatial clustering for rare diseases

by

Ge Lin¹ and Tonglin Zhang²

RESEARCH PAPER 2003-9

September 4, 2003

Abstract: We propose a method to test for the existence of low-value spatial clustering while accounting for the influence of high-value clustering and outliers. Although the method is in reference to the Tango test, it can be extended to other testing methods. The simulation results show that the proposed method can effectively detect low-value clustering with substantially lower rates of type I errors than those of the Tango test, while maintaining statistical power that is comparable to that of Tango test. A case study of leukemia in Minnesota shows that there is an overall tendency of low-value clustering of male leukemia mortality, and that the evidence for females is inconclusive.

Key words: Bias, low-value clustering, Tango index, trimmed mean

¹ Assistant Professor, West Virginia University and ²Assistant Professor, Purdue University

1 Introduction

A spatial cluster, according to Marshall is a foci of particularly high incidence or a hot spot that is unlikely to happen by chance [11]. Marshall also pointed out that a low-value foci or a cool spot should be included in this definition. In spatial statistics, when a cluster exists, there is an overall tendency to cluster, which is normally a prerequisite for further study [3]. Many spatial epidemiologists have stressed the importance of identifying and quantifying spatial clustering of elevated risks, as they often provide etiological clues for disease treatment and prevention [4]. However, we can also look at the issues from the other side and identify low-value clustering. Questions relating to the existence of low-value clustering or cool spots may be: What makes people in a particular clustered area less likely to have a certain disease? Is it because the geographic area is immune to a specific disease? Is it because people in the community are genetically endowed with resistance? Properly detecting low-value clustering can also reveal that a prevention program may be at work in a set of communities [7], which may provide lessons for other communities seeking to eliminate geographic inequality in health.

Even though testing for the existence of low-value clustering seems a straightforward application of general clustering tests, few spatial epidemiologists have undertaken it, and there are several conceptual and testing issues that need to be addressed. First, some spatial events may not necessarily have low-value clustering. The West Nile virus, for instance, was originally concentrated along the coastal regions of the mid-Atlantic states in the U.S., but we cannot presume that other parts of the U.S. were all cool spots with low-value clustering. In this case, low-value clustering inherently does not exist. Second, some clustering methods have been expressly designed for identifying high value clustering, and, consequently, are not suitable for testing low-value clustering. The scan test of Kulldorff [6], which tests the existence of one (high-value) cluster within a study area, is an example. Third, some general spatial clustering and autocorrelation tests, such as Moran's I , Tango's C_G , cannot distinguish the existence of low-value clustering from that of high-value clustering. In this case, a one-sided clustering test would supplement these test statistics.

In this paper, we propose a one-sided testing method for detecting low-value spatial clustering. We first discuss some aspects of general low-value clustering and then provide a testing method based on the Tango test for rare diseases [15]. In Section 3, we use simulated data to compare type I errors and the statistical

powers between the proposed test method and the Tango test using simulated data. We then provide a case study of leukemia mortality in Minnesota in Section 4, which is followed by concluding remarks in the final section.

2 General Tests for High and Low-value Clustering

Unlike high-value clustering, which has no upper (value) limit, the lower limit of the disease rate is 0 or no occurrence. This difference has several implications. First, it is generally unnecessary to consider low-value outliers because there is no negative value involved. If there is a statistically significant cluster with a low disease rate, it is unlikely to be attributed to outliers, especially when dealing with a rare disease. Second, to avoid the situation in which there is no inherent low-value clustering, rare infectious diseases are not appropriate candidates for testing because they often have no chance of occurring in most study areas where people are not infected. Third, low-value disease clustering tends to be more sensitive to the influence of spatial outliers and high-value clustering. Consider a study area found to have a general tendency to low-value clustering. Upon further examination, however, it is found that a hot spot with a relative risk around 0.15% has a significant leverage on the mean or the average disease risk of 0.05%. If the hot spot, which certainly has not happened by chance, is deleted or replaced with values around the mean, the tendency of low-value clustering would be reduced to the point of statistical insignificance. This is essentially the problem of comparing means for overlapping groups, which also arises in a high-value clustering test. However, the potential impact of a low-value cluster on high-value clustering is generally less severe than that of a high-value cluster on low-value clustering. To develop a robust low-value clustering test, it is necessary to first reduce the potential bias resulting from the influence of high-value clustering.

Following from Marshall's definition, the existence of geographically cool spots can be viewed as abnormally low-values clustered somewhere within the study area. If there was no abnormally high-value cluster or hot spot, the null hypothesis of no low-value clustering (or a constant disease risk across the study area) would be appropriate for a low-value clustering test. However, hot and cool spots often co-exist within a single study area, which may lead to an upward bias of the relative risk from the null hypothesis of a spatially constant mean. Ord and Getis noticed this problem[12], and they followed a common method of partitioning

means into nonoverlapping groups [2, 10]. In this case, the partitions are made within a distance range and the rest of the study area, but the influence of hot spots remains a potential problem. The unresolved issue is how to properly define the null hypothesis such that it incorporates both spatially constant risk and the potential existence of hot spots. Rather than partitioning means, we propose to “partition” the null hypothesis that distinguishes the null hypothesis of (a) no spatial clustering (or a spatially constant risk) from the null hypothesis of (b) the existence of a hot spot without a cool spot.

If we simply use the spatially constant mean or a as the null hypothesis, the existing test methods, such as Whittemore W [17], Getis G [5], Tango C_G , may not be appropriate for detecting low-value clustering, because the mean events in a particular region may not be proportional to the population in the region. With this distinction, the null hypothesis for testing cool spots becomes $a \cup b$. When the null hypothesis of $a \cup b$ is rejected, for instance, the null distributions under $a \cup b$ should be adjusted for potential upward bias. If there were only a few outliers, it simply may be easier to delete them. If there were hot spots, we may not know about them until a cluster test is implemented, and we normally cannot delete hot spots from testing anyway. Here, we propose a conditional replacement method to reduce the potential impact of hot spots or outliers. Although the proposed method is generally applicable to several testing methods, our discussion is based on Tango C_G , as it is a general test that encompasses several test methods (e.g., Whittemore W , Oden I^* , and Rogerson R) [13].

Tango’s spatial clustering test is an extension of his one-dimensional time-series clustering test [14]. Given a population size of ξ_i and disease prevalence of N_i (with observation n_i for a total of n cases) for a study area with m regions indexed by i , where N_i independently follows Poisson distribution, the random variable \mathbf{r} is the m -dimensional vector of $r_i = N_i/N$, where $N = \sum_{i=1}^m N_i$, and the nonrandom variable \mathbf{p} can be expressed by the m -dimensional vector of $p_i = \xi_i/\xi$, where $\xi = \sum_{i=1}^m \xi_i$. Under Tango’s null hypothesis of constant risks, we have

$$\sqrt{n}(\mathbf{r} - \mathbf{p}) \sim N_m(0, V_p), \quad (1)$$

where $V_p = \Delta(\mathbf{p}) - \mathbf{p}^t \mathbf{p}$, and $\Delta(\mathbf{P})$ is the $m \times m$ diagonal matrix based on the vector \mathbf{p} . Tango provides the general test statistic as

$$C_G = (\mathbf{r} - \mathbf{p})^t A (\mathbf{r} - \mathbf{p}), \quad (2)$$

where $A = (a_{ij})$ is an $m \times m$ weight matrix, in which $a_{ij} = e^{-d_{ij}/\tau}$, d_{ij} is the distance between region i and region j , and τ is a constant. Tango also provides the null distribution to determine the critical value for the test-statistic by

$$P\{C_G > c\} = 1 - I\left(\frac{\nu + T_G\sqrt{2\nu}}{2}, \frac{\nu}{2}\right), \quad (3)$$

where the incomplete Gamma $I(x, \phi)$ is defined by

$$I(x, \phi) = \int_0^x \frac{t^{\phi-1}}{\Gamma(\phi)} t^{\phi-1} e^{-t} dt,$$

and Tango T_G , the standardized C_G , is defined by

$$T_G = \frac{C_G - EC_G}{\sqrt{Var(C_G)}},$$

with

$$E(C_G) = \frac{1}{n} tr(AV_p), \quad Var(C_G) = \frac{2}{n} tr(AV_p)^2, \quad \nu = \frac{[tr(AV_p)^2]^3}{[tr(AV_p)^3]^2}.$$

Like a Chi-squared test or a test similar to Moran's I , C_G is a two-sided test. It can be used to detect both high- and low-value clustering, but there is no way to tell if a detected clustering tendency is attributable to a hot spot or a cool spot, or both [9]. To make it a one-sided test, we need to reduce any potential effect from the other side, in this case, potential hot spots. Assuming that the relative risks of normal regions are λ_0 , then, the relative risk in a hot spot is greater than λ_0 , and the expected value within the hot spot is greater than its population times λ_0 . Since the null hypothesis for low-value clustering does not exclude the presence of hot spots, it is necessary to eliminate the influence of hot spots when testing for the existence of cool spots.

As mentioned earlier, when the observed value n_i at region i is greater than the known $\lambda_0 \xi_i$, it could be due either to a random high within normal regions or to the clustered high within a hot spot. Since the clustered high is not normal, its effect should be reduced before testing for low-value clustering. It is difficult, however, to determine beforehand if high-value regions in a study area are clustered or not. One way to deal with this uncertainty is to use the known λ to generate random numbers to replace high-value regions. Theoretically, we can always generate a disease distribution that resembles the real disease pattern by using this λ . It turns out that if randomly high-values are replaced with a set of randomly high numbers,

the overall effect in normal regions will remain the same; if the numbers in the clustered high regions are replaced with a set of randomly high numbers, the bias caused by high-value clustering could be reduced. It is, therefore, reasonable to blindly replace all the values above the known risk λ_0 with random numbers greater than λ_0 . This strategy can be implemented by replacing all the regions with $n_i > \lambda_0 \xi_i$, with a Poisson random variable P_i (with parameter $\lambda_0 \xi_i$) that is also greater than $\lambda_0 \xi_i$, where λ_0 is the real disease rate for the normal regions. With this replacement scheme, potential bias due to spatial outliers or hot spots can be reduced while retaining randomly low-values in other regions. A less biased estimator can, therefore, be obtained with the risks for all regions becoming closer to λ_0 .

3 Simulation

Since Tango's spatial clustering test originated from Tango's one-dimensional clustering test, it is worthwhile to start with one-dimensional simulation to demonstrate the importance of distinguishing between the null hypothesis of no clustering and no low-value clustering. Let the disease rate λ_0 for normal regions be fixed at 10^{-4} . The population at region i ($i = 1, \dots, 100$) is generated independently from the closest integer of the $\Gamma(10^4, 0.1)$ distribution. Hence, the mean of the population of any region is 10^5 with the standard error of 10^3 . The distance between regions i and j is based on a straight line, or $d_{ij} = |i - j|$. Without loss of generality for the test results, let $\tau = 1$ or the average distance between two adjacent area unit, thus $a_{ij} = e^{-|i-j|}$. We start by examining type I errors of the Tango and the low-value clustering tests against the null hypothesis of no low-value clustering in the presence of hot spots only. We then evaluate the statistical power of the two tests when there is at least one cool spot. Finally, we further compare the powers by combining the spatial structures from the previous simulations.

The upper panel of Figure 1 displays the values of relative risks in the simulations with one, two or three hot spots being generated over a range of δ values from 0 to 1 in increments of 0.01. When $\delta = 0$, the highest relative risk is identical to λ_0 ; when $\delta = 1$, the highest relative risk is twice as much as λ_0 . In this way, type I errors for the low value clustering test are based on the presence of one, or two, or three hot spots, and the relative risk of the control point within a hot spot increases gradually from 100% to 200% of the relative risk $R = 1$. In the case of one hot spot, for instance, we can insert a hot spot right at the midpoint as shown in

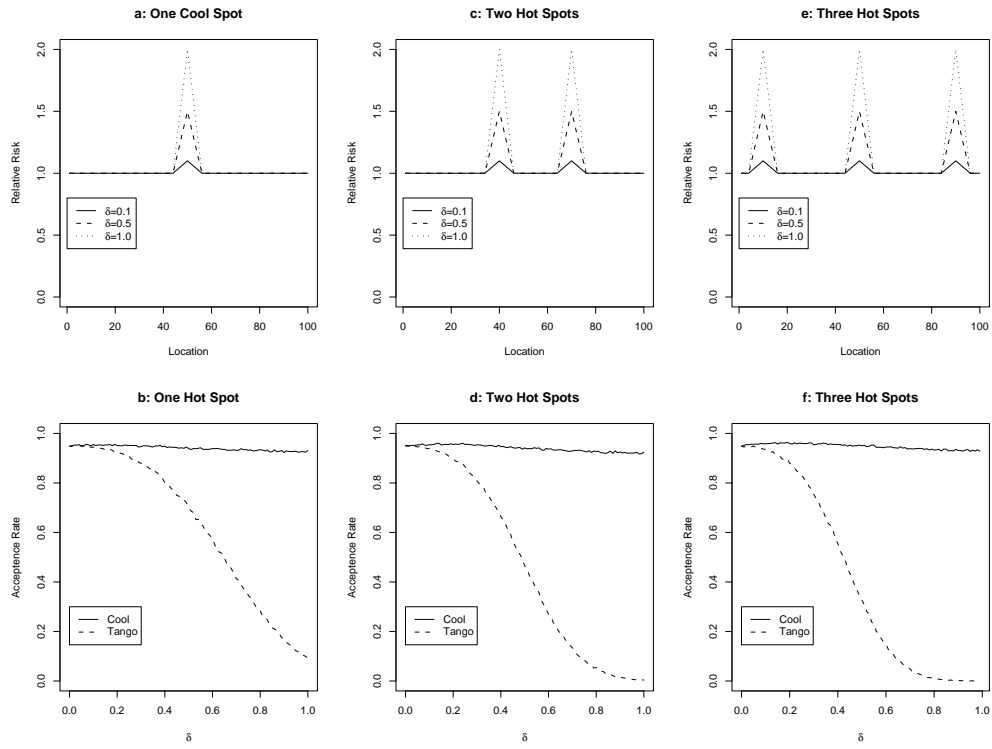


Figure 1: Acceptance rates for low-value clustering in the presence of hot spots.

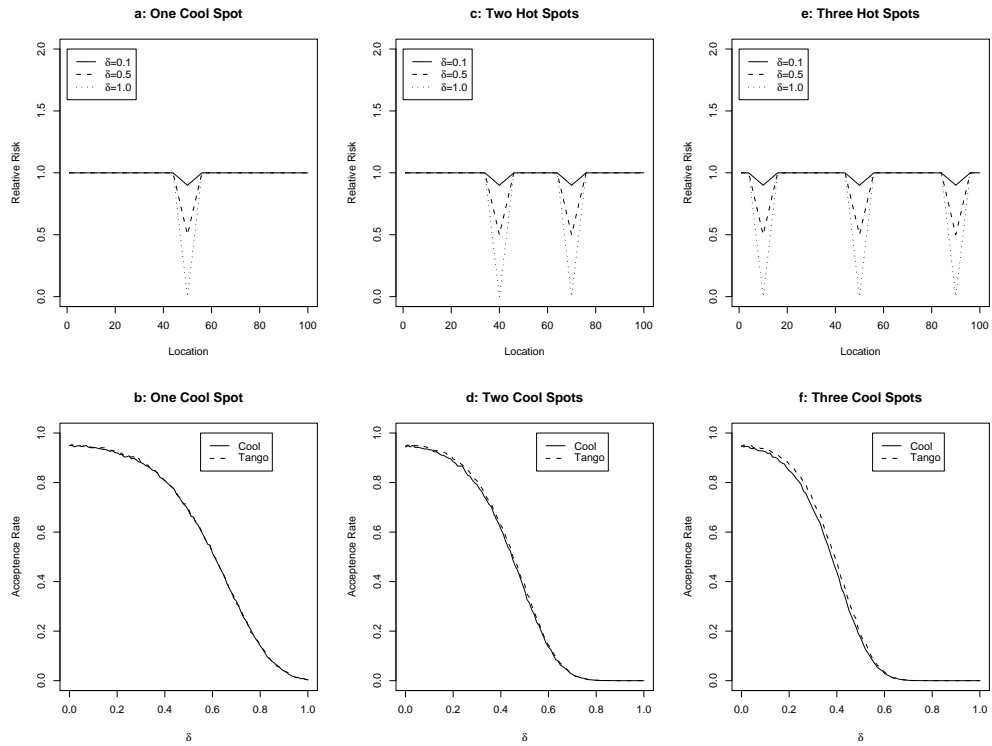


Figure 2: Acceptance rates for low-value clustering in the presence of cool spots

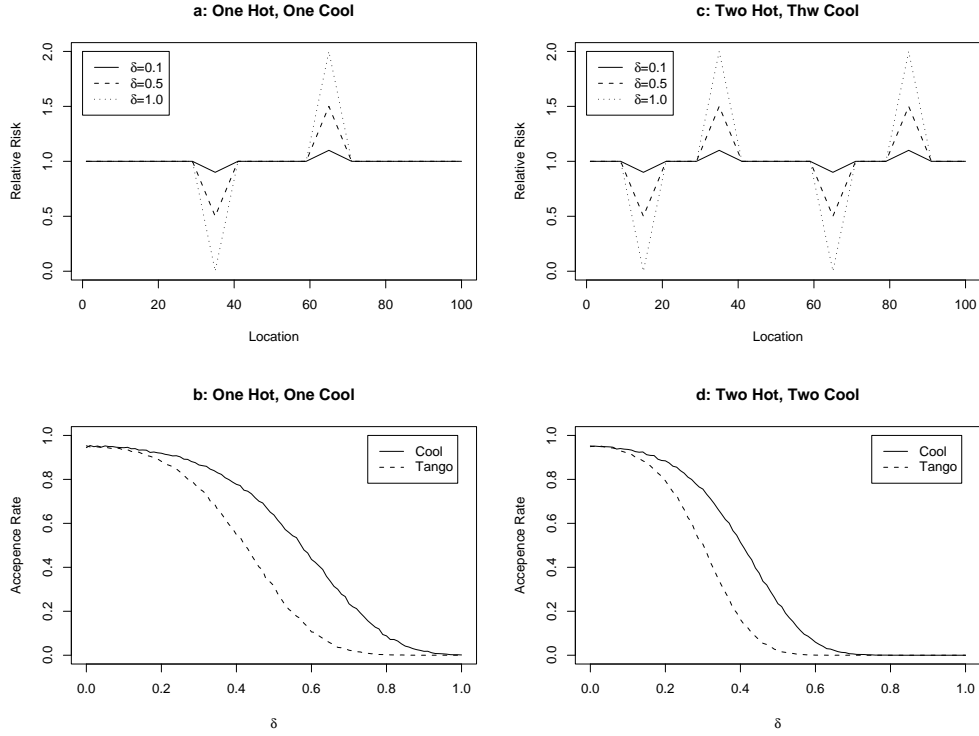


Figure 3: Acceptance rates for low-value clustering in the presence of hot and cool spots.

Figure 1a (i.e., $R_i = 1 + \delta|i - 50|/6$ if $|i - 50| < 6$ and $R_i = 1$ otherwise, where $\delta \in [0, 1]$ defines the strength of the cool spot, 50 is the midpoint). For each selected δ , we repeated 10,000 runs, and the 5% level test results were computed based on the p -values of the related statistics for each run. The lower panel of Figure 1 displays the corresponding results.

In the presence of a hot spot, the test of low-value clustering consistently accepts the null hypothesis for all δ s, with the acceptance rate being around 0.95. The acceptance rates of Tango's C_G , however are at the acceptable level of 0.95 only if there is no hot spot or the strength of the hot spot is very small (δ is close to zero). As the strength of the high-value clustering increases, the acceptance rate for C_G decreases rapidly; it reaches almost 0 when δ is close to 1. There is little difference in type I errors after two or three hot spots are inserted. Hence, there is a clear inverse relationship between the strength of a hot spot and the type I error rate of C_G . Apparently, if hot spots exist, the rejection rate of Tango C_G is much greater than the test level if it were to be used to detect the low-value clustering.

Alternatively, we can evaluate the powers of these tests and their effectiveness in detecting low-value

clustering (Figure 2). The results show that all of the clustering tests are very effective, having an acceptance rate of 0.95 when there is no clustering or $\delta = 0$; the acceptance rates of the corresponding null hypotheses are all close to 0 when there is low-value clustering or δ is close to 1. The statistical powers increase slightly when two or three cool spots are inserted. In all cases, the statistical powers of C_G and the low-value clustering test are almost identical, but the conclusion from the low value clustering test represents an important information gain, which unambiguously rejects the null hypothesis of no low-value clustering and, thus, concludes its existence. The proposed low-value clustering test is indeed supplementary to C_G . Like C_G , the low-value clustering test is likely to be significant in the presence of a cool spot; unlike C_G , it is rarely significant if there is a hot spot only.

Finally, we simulated situations in which both cool and hot spots exist (Figure 3). The results from the Tango test consistently register a greater statistical power than does the low-value clustering test. This can be explained from two different perspectives. On one hand, the greater power of Tango's test is expected, because it considers the existence of either cool or hot spots as clustering while the low-value clustering test only considers the existence of cool spots. So the lower statistical power of the low-value clustering test is compensated for by greater information gain or less ambiguity. On the other hand, we can evaluate the power gap between the two tests for different δ values, and determine the area in which a false rejection of the null hypothesis may fall. In the case of one cool and one hot spot, C_G rejects the null hypothesis of no clustering greater than 95% when δ is greater than 0.72, but the low-value clustering test does not reach this level until δ is greater than 0.91. This discrepancy means that between a weaker cool spot range (i.e., between 1.72 and 1.91 times of the relative risk), C_G may signify significant clustering that may be due to the effects of high-value clustering. These two explanations together suggest that the low-value clustering test is an effective one-sided test that could be used as an alternative to C_G when there is a need to reduce abnormal effects caused by hot spots and outliers.

4 Minnesota Leukemia Case Study

We chose leukemia as an example, because its etiological causes are largely unknown and because epidemiologists can learn ecological risk factors from both cool and hot spots. Extensive studies have related environ-

mental factors and agrochemicals to leukemia incidence and mortality, but no conclusive geo-environmental leukaemogens have been reported [1, 16]. For our case study, we selected the five-year (1992-96) county-level leukemia- mortality data from the Minnesota Cancer Surveillance System, which records the number of deaths due to leukemia for males and females separately. According to the U.S. National Cancer Institute, the five year (1990-94) leukemia mortality rate for white males in Minnesota is the second highest in the U.S., or about 11% higher than the national average. The rate for white females is ranked the 23rd, or just slightly higher than the national average. The county-level populations for the state from the 1990 U.S. Census were used to analyze leukemia mortality rates for both males and females. We compared the 1990 Census populations with the 1994 county estimates (the mid year of 1992-96), and found very small changes in population for most of the counties. For this reason, we decide to use the Census data rather than the estimates. We used Euclidean distance to measure geographic proximity and set $\tau = 35$, which is the average distance between any two centroids of adjacent counties.

Figures 4 and 5 display geographic distributions of the five-year male and female leukemia-mortality rates. The mortality rates are grouped into seven percentiles, with roughly an equal number of counties within each percentile-category. The average mortality rates for males and females were 0.80 and 0.55 per 1,000, respectively with the total number of deaths during the five-year period being 1,718 for males and 1,226 for females. Among the 87 counties, 2 reported a 0 death rate for females but none reported a 0 death rate for males. The highest rate for males was 1.87 per 1,000, more than double the corresponding mean; the highest death rate for females was 0.89 per 1,000, or 1.62 times the corresponding mean. For males, there appeared to be clusters of both high and low-values. The high-value cluster seem located around the central west, and the low-value cluster seems located along the lower Minnesota River basin in the central south. The picture for females is murkier, having no apparent geographical pattern.

In this example, λ_0 is unknown, making it necessary to estimate λ_0 as the true value in the proposed low-value clustering test. Given the emphasis of our study, we used a simple method to derive a less biased mean by a 10% trimmed mean (See Page 32, Devore [3]) Because this method could still be biased from the true λ_0 , we assess the sensitivity of the p -values by extended our testing to a range of λ values 10% below $\hat{\lambda}$. In other words, a range of downward λ values was used to correct any potential upward bias resulting

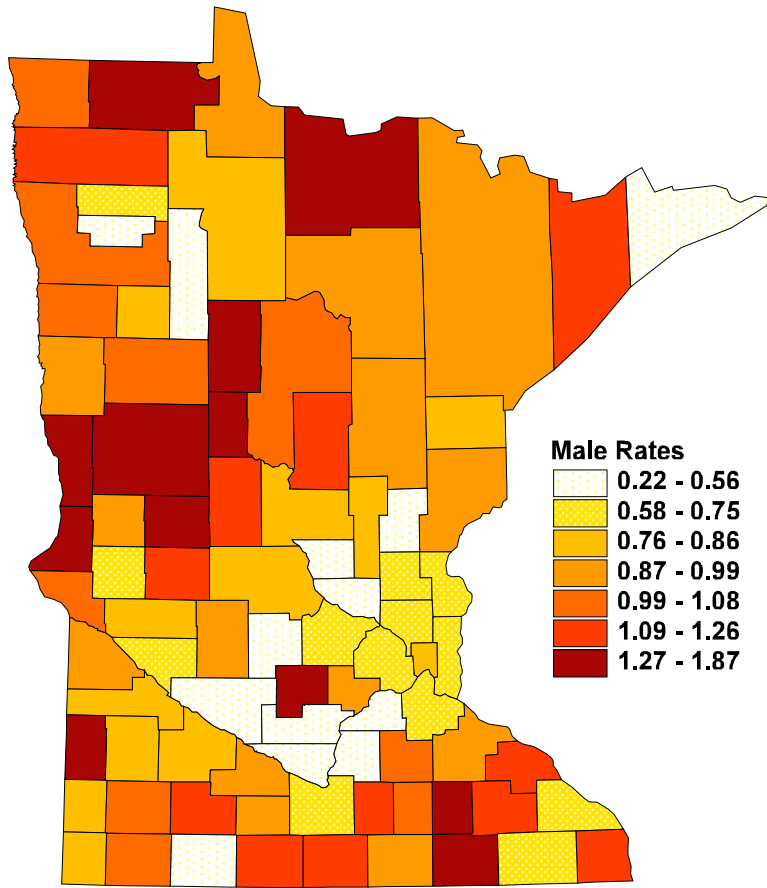


Figure 4: Male leukemia mortality rates per 1,000 in Minnesota

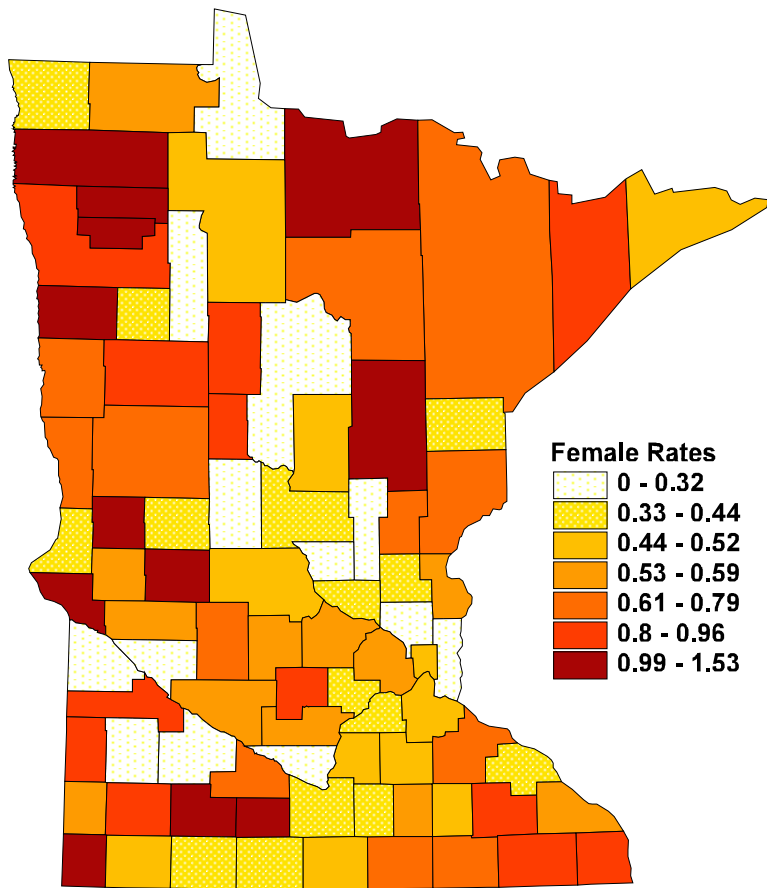


Figure 5: Female leukemia mortality rates per 1,000 in Minnesota

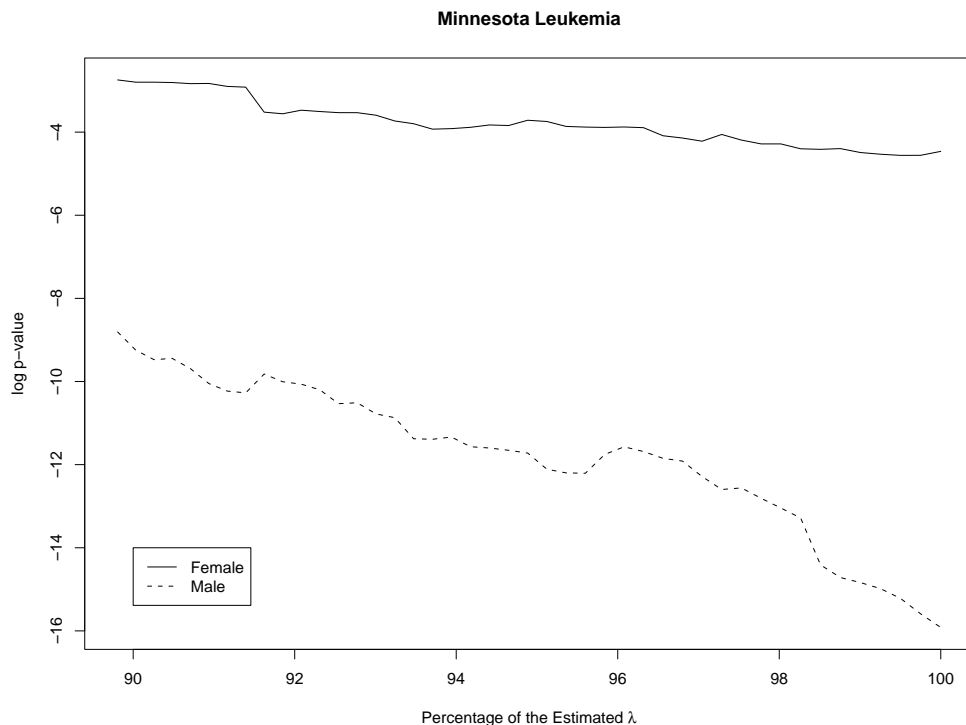


Figure 6: The p -values for testing low-value leukemia clustering in Minnesota

from the existence of high-value clusters. The $\hat{\lambda}$ values for males and females were 0.000796 and 0.000546 respectively, which are not very different from the average rates of 0.000801 for males and 0.000550 for females.

Taking the male- and female- $\hat{\lambda}$ values to be the true values, the p -values of the low-value clustering test are computed repeatedly 10,000 times. We considered the medians of those repeated p -values to be the most trustworthy, and they are displayed in Figure 6 in the natural log scale. Note that the smaller the p -value, that is, the more negative the value along the log scale, the greater is its significance. A log p -value greater than -3 would not be significant at the probability level of 0.05, and a log p -value greater than -4.6 would not be significant at the probability level of 0.01.

Overall, the p -values for males were very small, being less than 0.001 ($e^{-8} = 0.0003$) anywhere between $\hat{\lambda}$ and 10% lower. As the λ values decrease, the p -values of the low-value clustering test increase, and become less significant. Because a potential abnormally hot spot would cause an upward bias from the true λ_0 , a correction would lead to a less significant result. In any case, the influence of a potential biased mean in a

range of 10% upward, which is evaluated by a range 10% downward of λ , will not lead to a different outcome. We, therefore, concluded that there is an overall tendency of low-value spatial clustering among males, who died from leukemia during the study period.

For females, however, the results were not significant at the 0.01 probability level anywhere from the estimated λ to 10% lower. In addition, the p -values are very sensitive to the shifts of λ values. When the λ falls within 8% of the $\hat{\lambda}$ value, the results are significant at the 0.05 level. When the λ value shifts beyond 8%, the results are no longer significant at the 0.05 level. Since the sample was fairly large, and the estimated mean could easily be biased upward by 6 to 7%, we used a more conservative confidence level of 0.01. Thus, the existence of cool spots could not be concluded for females in general.

5 Concluding remarks

In this paper, we have provided a testing method for low-value spatial clustering. When the expected disease risk is known, the proposed conditional replacement method was effective in reducing a potential overestimate of the disease rate due to the presence of structurally high value regions. Although the simulations and the data example were in reference to the Tango test, the conditional replacement method can be applied to other clustering tests. In the presence of a hot spot, the type I error rate based on the null hypothesis of the low-value clustering test was much lower than that based on the Tango test. The powers of the two tests, however, were almost identical in the presence of a cool spot only. When both hot and cool spots coexisted, the Tango test had a greater statistical power, which also came with a loss of information. In this regard, the low-value clustering test can be used not only for testing tendency for low-value clustering, but also for supplementing other general clustering tests to reduce false alarms. This is especially the case when there is a suspicion of which clustering tendency (high or low) contributes more to a significant test result.

When λ is unknown, as in the case of leukemia mortality in Minnesota, we first estimated the true disease rate for normal regions by cutting off 10% from both sides. Because hot spots may appear according to our null hypothesis, this estimated disease rate for normal regions may be biased upward. This means that the true λ_0 may be less than the estimated value. For this reason, we evaluated a range of λ values. The existence of low-value clustering was concluded for males but not for females. Based on these results, one

can further identify local clusters by using a focused test, which we currently are investigating.

Several issues warrant further investigation. First, even though the focus of our study is on low-value clustering, we can equally see its logical application in high-value clustering tests. Second, a better way to empirically derive an estimator that is sufficiently close to the true λ is needed. In addition, when the study area is part of a larger region, the risk at the regional level should not be ignored. In our empirical case, the male mortality due to leukemia is 11% higher than the national average; if there is a hot spot with a highly excessive mortality rate in the study area, we may want to deal with it first before estimating the true mean. Third, covariates may still be considered by using covariate-related statistics, such as the scan test [6]. Finally, how the critical region is bounded or affected by an estimated λ value should be investigated when extending the current method to other clustering tests.

References

- [1] Boyle P Walker AM and Alexander FE (1996) Historical aspects of leukemia clusters. In: Boyle P, and Alexander FE (eds) *Methods for Investigating Localized Clustering of Disease*, IARC Scientific Publications, Lyon, France, **No.135** 1-20.
- [2] Calinski T and Corsten LC (1985) Clustering means in ANOVA by simultaneous testing, *Biometrics*, **41**, 39-48.
- [3] Cuzick J and Edwards R (1990) Spatial clustering for inhomogeneous populations (with discussion). *Journal of Royal Statistical Society, Series B*, **52**, 73-104.
- bibitemlak:proofs Devore, Jay (1999). *Probability and Statistics for Engineering and the Sciences*, 5th edition, Duxbury, Pacific Grove, USA.
- [4] Elliott P, Wakefield J, Best N, and Briggs D (2000) *Spatial Epidemiology: Methods and Applications*. Oxford, Oxford University Press
- [5] Getis A, and Ord J (1992) The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, **24**, 189-206.

- [6] Kulldorff M (1997) A spatial scan statistic. *Communications in Statistics, Theory and Methods* **26** 1481-1496.
- [7] Lawson, A and Kulldorff M (2000) A review of cluster detection methods In: Lawson A, Biggeri A, Bohning D Lesaffre E, Viel J, and Bertollini R (eds). *Disease Mapping and Risk Assessment for Public Health*. Chapter 7 New York, Wiley.
- [8] Lawson A, and Clark A (2002) Spatial mixture relative risk models applied to disease mapping. *Statistics in medicine*, **21**: 359-370.
- [9] Lin, G, (2003) Comparing spatial clustering tests based on rare to common spatial events. *Computers, Environment and Urban Systems* forthcoming.
- [10] Looney, S. and Jones, P. (2003) A method for comparing two normal means using combined samples of correlated and uncorrelated data *Statistics in Medicine*, **22**: 1601-10.
- [11] Marshall, RJ (1991) A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society Series A*, **154**: 421-441
- [12] Ord, K, and Getis A (2001) Testing for local spatial autocorrelation in the presence of global autocorrelation. *Journal of Regional Science*, **41**: 411-432.
- [13] Rogerson PA (1999) The detection of clusters using a spatial version of the chi- square goodness-of-fit statistics. *Geographical Analysis*, **31**: 130-147.
- [14] Tango T.(1984) The detection of disease clustering in time. *Biometrics*, **40**: 15-26.
- [15] Tango, T (1995) A class of test for detecting 'General' and 'Focused' clustering of rare diseases. *Statistics in Medicine*, **14**: 2323-2334.
- [16] Wartenberg D (1998) Residential magnetic fields and childhood leukemia: A meta- analysis. *American Journal of Public Health*, **88**: 1787-94.
- [17] Whittemore A, Friend N, Brown N, and Holly E. (1987) A test to detect clusters of disease *Biometrika*, **74**: 631-635.