

A Spatial Scan Statistic for Ordinal Data

Inkyung Jung ^{1,*}, Martin Kulldorff ¹, Ann Klassen ²

Addresses:

¹ Department of Ambulatory Care and Prevention

Harvard Medical School and Harvard Pilgrim Health Care

133 Brookline Ave. 6th Floor, Boston, MA 02215

² Department of Health, Behavior and Society

Johns Hopkins Bloomberg School of Public Health

624 N Broadway, Room 745, Baltimore, MD 21205

* Corresponding author

Email: inkyung_jung@harvardpilgrim.org

Telephone: 1-617-509-9925

Fax: 1-617-509-9846

Sponsors:

Centers for Disease Control and Prevention (CDC)

Association of American Medical Colleges (AAMC)

Grant number: MM-0870

Summary

Spatial scan statistics are widely used for count data to detect geographical disease clusters of high or low incidence, mortality or prevalence and to evaluate their statistical significance. Some data are ordinal or continuous in nature, however, so that it is necessary to dichotomize the data to use a traditional scan statistic for count data. There is then a loss of information and the choice of cut-off point is often arbitrary. In this paper, we propose a spatial scan statistic for ordinal data, which allows us to analyze such data incorporating the ordinal structure without making any further assumptions. The test statistic is based on a likelihood ratio test and evaluated using Monte Carlo hypothesis testing. The proposed method is illustrated using prostate cancer grade and stage data from the Maryland Cancer Registry.

Key words: clusters, geographical disease surveillance, prostate cancer

1 Introduction

Spatial scan statistics based on the likelihood ratio test are widely used for detecting geographical clusters. Such analyses have been available for count data using either a Poisson or Bernoulli model [1]. Bernoulli-based scan statistics are used when we have dichotomous variables such as cases and non-cases (e.g. people with and without a certain disease) or cases with two different types of disease characteristics (e.g. patients with late/early stage of cancer). Examples of use of the Bernoulli-based scan statistic are found in studies by Han et al. [2] who searched for geographical clusters of high rates of breast cancer cases in western New York and by Brooker et al. [3] who identified significant spatial clusters of malaria cases in a highland area of western Kenya. Poisson based scan statistics are used for the number of cases compared to the underlying population at risk such as disease mortality or incidence data. Jennings et al. [4] used the Poisson based scan statistic to

determine whether statistically significant geographic clusters of high prevalence gonorrhea cases can be located after controlling for race/ethnicity in Baltimore, Maryland. Other examples include studies of childhood mortality in rural Burkina Faso (West Africa) [5] and of regional variation in the incidence of symptomatic pesticide exposure in Oregon [6].

Some data are ordinal in nature. For example, in cancer registry data the stage of breast cancer is often classified as local, regional or distant, based on the extent of disease presence at the time of diagnosis. We may then be interested in geographical areas with high (or low) rates of cases with more extensive disease. One approach to analyze such data is to dichotomize and use a Bernoulli-based scan statistic, which may be desired if one has a specific (e.g. clinically determined) cut-off point in mind. Rushton et al. [7] used a dichotomized cancer stage (early/late) to identify areas with high rates of late stage colorectal cancer cases in Iowa. Sheehan and DeChello [8] also dichotomized the breast cancer stage to examine spatially the proportion of breast cancer cases diagnosed at late stage in Massachusetts from 1988 through 1997. However, there may be some loss of information in dichotomizing such data. Therefore, it may be better to use a probability model that incorporates the ordinal structure of the data rather than dichotomizing.

In this paper we propose a spatial scan statistic for ordinal data, which allows us to investigate such data without dichotomization. The proposed method is based on the likelihood ratio test and the significance of the test statistic is evaluated using Monte Carlo hypothesis testing [9]. In section 2, the spatial scan statistic for ordinal data is introduced and inference on the detected clusters is discussed. In section 3, the proposed method is illustrated with an example of prostate cancer data in Maryland on stage and grade of disease. The results from the proposed procedure are compared with those from the Bernoulli-based scan statistic using different cut-off points. This paper ends with a discussion in section 4.

2 Scan Statistic for Ordinal Data

2.1 Test Statistic

The spatial scan statistics are based on the likelihood ratio test. The most likely cluster is the area associated with the maximum value of the likelihood ratio test statistic. To find such a cluster, we first construct a collection of zones (scanning windows) and each of the zones is a candidate for the most likely cluster. We compute the value of the likelihood for every zone, from which we find the maximum. In other words, the collection of zones is a parameter space for the cluster, over which the likelihood ratio is maximized. We consider circular shape scanning zones of variable size. For exact point data, the collection of zones is all possible circular areas centered at every point. For aggregated data into sub-regions of a study area such census block groups, which are more commonly available, a candidate cluster is a set of sub-regions whose centroids are within the scanning circle. We include all potential scanning zones of any size at every location, but we may want to limit the maximum cluster size since a too large cluster would not be useful. For our analyses presented in section 3, the maximum cluster size was set to 50% of the total population.

Now we introduce the spatial scan statistic for ordinal data. Suppose that we have a study area composed of I sub-regions and an outcome variable of interest recorded in K categories. Let c_{ik} be the number of observations in location i and category k , where $i = 1, \dots, I$ and $k = 1, \dots, K$. The categories are ordinal in nature, so that for example a larger k reflects a more serious disease stage. Let $c_i (= \sum_k c_{ik})$ be the total number of observations in location i , $C_k (= \sum_i c_{ik})$ be the total number of observations in category k , and $C (= \sum_k \sum_i c_{ik})$ be the total number of observations in the whole study area. We can write down the likelihood function for the ordinal model as

$$L(Z, p_1, \dots, p_K, q_1, \dots, q_K) \propto \prod_k \left(\prod_{i \in Z} p_k^{c_{ik}} \prod_{i \notin Z} q_k^{c_{ik}} \right) \quad (1)$$

where p_k is the unknown probability that an observation within the scanning window Z belongs to category k and q_k is the unknown probability of an observation outside the scanning window Z belongs to category k . Note that $\sum_k p_k = 1$ and $\sum_k q_k = 1$. The null hypothesis is that the probability of being in category k within the scanning window is the same as outside the scanning window ($H_0 : p_1 = q_1, \dots, p_K = q_K$). We consider

$$H_a : \frac{p_1}{q_1} \leq \frac{p_2}{q_2} \leq \dots \leq \frac{p_K}{q_K} \quad (2)$$

as an alternative hypothesis, with at least one inequality being strict. This ensures that detected clusters represent an area with high rates of higher stages than the surrounding area. This type of order restriction is called a likelihood ratio ordering [10]. We also can consider the opposite type of clusters with lower stages by changing the less than or equal signs to greater than or equal signs. To scan for both types of clusters simultaneously, we should include both options, but not sequences of p_k/q_k that are neither consistently increasing nor decreasing. For $K = 2$, we get the existing Bernoulli-based spatial scan statistic as a special case [1].

The likelihood ratio test statistic is expressed as

$$\lambda = \frac{\max_{Z, H_a} L(Z, p_1, \dots, p_K, q_1, \dots, q_K)}{\max_{Z, H_0} L(Z, p_1, \dots, p_K, q_1, \dots, q_K)} = \frac{\max_Z L(Z)}{L_0} \quad (3)$$

with

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

where $\hat{p}_{o_k} = C_k/C (= \hat{q}_{o_k})$ is the maximum likelihood estimator (MLE) of $p_k (= q_k)$ under the null hypothesis, and with

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

where \hat{p}_k and \hat{q}_k are the MLEs of p_k and q_k under the alternative hypothesis (2). Note that

we condition on the total numbers observed in each category (C_1, \dots, C_K) , which ensures that our analysis is based on the spatial distribution of observations, not on the total number observed. This will become clearer when the randomization procedure is described in the next section. Notice that L_0 depends only on (C_1, \dots, C_K) and hence L_0 is a constant.

The next step is to obtain the MLEs of p_k and q_k under (2). Let $W_k = \sum_{i \in Z} c_{ik}$, $U_k = \sum_{i \notin Z} c_{ik}$, $W = \sum_k W_k$ and $U = \sum_k U_k$. Note that $C_k = W_k + U_k$ and $C = W + U$. Dykstra et al. [10, Theorem 2.1] have shown that the MLEs of p_k and q_k are given by

$$\hat{p}_k = \left(\frac{W_k + U_k}{W} \right) E_{(\mathbf{w}+\mathbf{u})} \left(\frac{\mathbf{W}}{\mathbf{W} + \mathbf{U}} | \mathcal{T} \right)_k \quad (4)$$

and

$$\hat{q}_k = \left(\frac{W_k + U_k}{U} \right) E_{(\mathbf{w}+\mathbf{u})} \left(\frac{\mathbf{U}}{\mathbf{W} + \mathbf{U}} | \mathcal{A} \right)_k \quad (5)$$

where $\mathcal{T} = \{(\theta_1, \dots, \theta_K); \theta_1 \leq \dots \leq \theta_K\}$ and $\mathcal{A} = \{(\theta_1, \dots, \theta_K); \theta_1 \geq \dots \geq \theta_K\}$ with $\theta_k = W p_k / (W p_k + U q_k)$ for $k = 1, \dots, K$. $E_{\mathbf{v}}(\mathbf{B} | \mathcal{C})$ denotes the isotonic regression of $\mathbf{B} = (B_1, \dots, B_K)$ with weights $\mathbf{v} = (v_1, \dots, v_K)$ onto the cone \mathcal{C} .

It is possible to calculate (4) and (5) explicitly using the ‘‘Pool-Adjacent-Violators’’ algorithm as described by Barlow et al. [11]. If the observed rate $(W_k/W)/(U_k/U)$ is nondecreasing for $k = 1, \dots, K$, then \hat{p}_k and \hat{q}_k are the same as the unrestricted MLEs $\tilde{p}_k = W_k/W$ and $\tilde{q}_k = U_k/U$. With the alternative hypothesis (2), we want to detect clusters with high rates of higher stages. This does not necessarily mean that detected clusters should have rates for all categories exactly in increasing order. For example, suppose that there are 4 disease stages. There could be a significant cluster with a high rate of stage 4 compared to stages 1, 2, and 3 combined, even though the rate of stage 3 is not so high as that of stage 1 or 2. There also could be a significant cluster with a high rate of stages 3 and 4 combined compared to stage 2 or 1, etc. We should be able to detect such clusters as well as clusters with the sequence of observed rates for all categories exactly in increasing order. The MLEs in equations (4) and (5) are based on those combined categories. The

algorithm for this procedure is provided in the Appendix.

2.2 Inference

Once we have the most likely cluster associated with the maximum value of the likelihood ratio, we need to evaluate the significance. If the distribution of the test statistic can be found, at least asymptotically, we can obtain a critical value at a given significance level. However, this is usually not available for spatial scan statistics. An alternative way is to use Monte Carlo hypothesis testing [9]. A large number of simulated data sets are then generated under the null hypothesis and the test statistic is computed for each simulated data set and compared to the value of the test statistic from the real data. The Monte Carlo based p-value is defined as $p = R/(\#SIM + 1)$ where R is the rank of the test statistic from the real data set among all data sets and $\#SIM$ is the number of simulated data sets that were generated. We usually use 999, 9999 or some number ending 999 for $\#SIM$ to obtain ‘nice-looking’ p-values. In our analyses, 9999 Monte Carlo replications were run.

To generate random data sets under the null hypothesis, we condition on the total numbers observed in each category. C_1 individuals are first randomly chosen and assigned to category 1, C_2 of the remaining individuals are then randomly chosen and assigned to category 2, and so on. After C_{K-1} individuals have been randomly assigned to category $K - 1$, the remaining C_k are assigned to category K .

In addition to the most likely cluster, it can be useful to examine secondary clusters with high values of the likelihood ratio. We report secondary clusters when those clusters have no geographical overlap with another reported cluster with higher likelihood ratio. The statistical significance of a secondary cluster is evaluated irrespectively of any other clusters, by comparing its likelihood ratio value with the maximum likelihood ratio from the generated random data sets.

3 Maryland Prostate Cancer Data

The data are explained in detail elsewhere [12, 13] and are reviewed here. We obtained records for all incident prostate cancer cases reported to the Maryland Cancer Registry for the years 1992-1997. There were 24189 registry cases in total during that period. Cases with verified Maryland residential addresses were geocoded to latitude and longitude. Cases that could not be successfully geocoded were assigned to a coordinate location within their zip code using a weighted population algorithm, based on 1990 United States Census race, age, and gender specific block and block group distributions within the zip codes. 23993 cases were geographically referenced and were aggregated into census block groups. Geographical locations of each centroid of census block groups containing cases were used in the analyses.

The outcome measures are histological grade of tumor and the Surveillance, Epidemiology and End Stage (SEER) summary stage of disease at diagnosis [14], with grade recorded as 1, 2, 3, or 4 and stage as 0, 1, 2, 3, 4, 5, or 7. Cases with missing information on age, race and census block group, and the cases with stage 0 were excluded. Cases missing stage information were retained in the grade analyses and vice versa, so that 19223 cases were used for the stage analyses and 18947 cases for the grade analyses.

[Tables 1 and 2 around here.]

3.1 Cluster Detection Analyses using the Spatial Scan Statistic for Ordinal Data

Using the spatial scan statistic for ordinal data, we have scanned for areas with high or low rates of higher grade prostate cancer cases in Maryland, and the detected clusters are displayed in Map 1 in Figure 1 and described in Table 3. Cluster 1 is the most likely cluster and 2-5 are the secondary clusters ordered by their statistical significance. Clusters 1, 3 and 4 are areas with high rates of higher grade and clusters 2 and 5 are areas with low rates. More precisely, the most likely cluster is an area with low rates of grades 2-4 compared to

grade 1 and cluster 2 is an area with high rates of grades 2-4 compared to grade 1. For the other clusters, no categories were combined. Clusters 3 and 4 are areas with rates for grade 1 to 4 in decreasing sequence and cluster 5 in increasing sequence.

[Table 3 and Figure 1 around here.]

We also searched for clusters with high or low rate of later stage prostate cancer. The most likely cluster and secondary clusters are listed in Table 4 and shown in Map 1 in Figure 2. There are 4 clusters detected and cluster 2 is the only one with low rates of later stage. Some of the 6 categories for stage were combined in every cluster; for example, cluster 3 is an area with high rates of stages 2-7 compared to stage 1.

[Table 4 and Figure 2 around here.]

3.2 Ordinal model versus Bernoulli model

The Bernoulli-based scan statistic could be used as an alternative to search for clusters with high or low rates of higher grade or later stage cancer cases by dichotomizing grade and stage. Since there are 4 categories for grade and 6 for stage, 3 different cut-off points are possible for grade and 5 for stage. Using the Bernoulli-based scan statistic we analyzed Maryland prostate cancer data for all possible cut-off values.

Clusters detected using the Bernoulli model for prostate cancer grade are presented in Map 2-4 in Figure 1 and Table 5. To compare the results from the ordinal model, clusters detected using the spatial scan statistic for ordinal data are also displayed in Figure 1. For grade 2-4 (vs 1), there are 3 clusters detected and for grade 3 or 4 (vs 1 or 2), 7 clusters detected. Clusters based on these two different dichotomization are very different in terms of cluster size and location. Since there are not so many cases with grade 4, there are only two small clusters detected for grade 4 (vs 1-3) and they are also very different from the other results. Compared with results based on the ordinal model, clusters 1, 2, and 3 in

Map 1 (ordinal model) are exactly the same as in Map 2 (Bernoulli model for grade 2-4). Cluster 4 in Map 1 is approximately the same as cluster 2 in Map 3 and cluster 5 in Map 1 is part of cluster 1 in Map 4. Most of the most significant clusters from the Bernoulli models with different cut-off points are detected with the ordinal model.

[Table 5 around here.]

The results for stage using the Bernoulli-based scan statistic are shown in Map 2-6 in Figure 2 and Tables 6. The results for stage 3-7 (vs 1-2), 4-7 (vs 1-3), 5-7 (vs 1-4), and 7 (vs 1-5) are very similar since around 80% of the total cases are stage 1 with very few stages 3-5. Most clusters from the ordinal model shown in Map 1 coincide with a cluster in one of the other maps generated by the Bernoulli models. The most likely cluster in Map 1 is very similar to the most likely cluster in Maps 3-6. Cluster 3 in Map 1 is very similar to cluster 2 in Map 2 and cluster 4 in Map 1 is partly overlapping with cluster 1 in Map 2 and cluster 4 in Map 3-6. Cluster 2 in Map 1 is partly overlapping with clusters 3 and 4 in Map 2, and clusters 2 and 3 in Map 3-6.

[Table 6 around here.]

4 Discussion

We have proposed a spatial scan statistic for ordinal data. An alternative to analyzing such data is to dichotomize and use the Bernoulli-based scan statistic. As seen in the prostate cancer data example, the cluster detection results depend significantly on the cut-off points. The problems using the Bernoulli-based scan statistic for ordinal data are that the cut-off points should be determined first and that it may not be clear where to cut. Using the spatial scan statistic for ordinal data, we can detect clusters and make inferences on the detected clusters without dichotomization or any further assumptions.

This approach has been applied here to prostate cancer data on grade and stage of dis-

ease. The application of the method should not be limited to disease outcome, but rather it is applicable to any type of ordinal data in diverse fields. For example, in survey questionnaires, attitudes and preferences are often measured using a Likert scale. Answers for demographic questions such as age and income are usually classified into ordered categorical variables. In other settings, it may be of interest to use a categorized variable of continuous outcome. Suppose that we are interested in geographical variation of babies' birth weight in certain region. Then, analyses based on categorized values such as very low, low, and normal may be more appropriate than using the continuous values, when it is only the low birth weights that are of concern.

The spatial scan statistic for ordinal data proposed in this paper has been developed for spatial data, but it can also be used for temporal data using a one-dimensional scanning window. It can also be directly extended to a space-time setting using a cylindrical window in three dimensions with the base of the cylinder representing space and the height representing time, for either retrospective [15] or prospective analyses [16]. Besides the circular shape of scanning window, we can also consider other shapes such as elliptical [17] or irregular shapes [18,19,20].

The ordinal scan statistic for temporal, spatial, and space-time data has recently been implemented into the freely available SaTScanTM software version 6.0, which can be downloaded from www.satscan.org.

Acknowledgments

This research was made possible through a Cooperative Agreement between the Centers for Disease Control and Prevention (CDC) and the Association of American Medical Colleges (AAMC), award number MM-0870; its contents are the responsibility of the authors and do not necessarily reflect the official views of the CDC or AAMC. We thank Scott Hostovich at Information Management Services, Inc. for computer programming support. We also acknowledge the assistance and support of the Maryland Cancer Registry of the Maryland

Appendix

Using the “Pool-Adjacent-Violators” algorithm, the MLEs of p_k and q_k for $k = 1, \dots, K$ under (2) can be obtained explicitly. Here we explain how to calculate the MLEs for a particular zone Z .

1. With the same notation of W_k , U_k , W , and U in page 5, compute the unrestricted MLEs \tilde{p}_k and \tilde{q}_k ($k = 1, \dots, K$);

$$\tilde{p}_k = \frac{W_k}{W}, \quad \tilde{q}_k = \frac{U_k}{U}.$$

2. Temporarily, we set

$$\hat{p}_k = \tilde{p}_k, \quad \hat{q}_k = \tilde{q}_k.$$

If \hat{p}_k/\hat{q}_k is nondecreasing for all $k = 1, \dots, K$, then these are the MLEs of p_k and q_k ;

$$\hat{p}_k = \frac{W_k}{W}, \quad \hat{q}_k = \frac{U_k}{U}.$$

3. Otherwise, we do the following for $k = 1, \dots, K$ while $\hat{p}_l/\hat{q}_l > \hat{p}_k/\hat{q}_k$ with $l = k - 1$ for $l > 0$. For $j = l, \dots, k$

$$\hat{p}_j = \frac{\sum_{j=l}^k W_j}{W} \frac{C_j}{\sum_{j=l}^k C_j}, \quad \hat{q}_j = \frac{\sum_{j=l}^k U_j}{U} \frac{C_j}{\sum_{j=l}^k C_j}.$$

Note that \hat{p}_j/\hat{q}_j is the same for $j = l, \dots, k$, which is the observed rate for combined categories. In this step, we keep “pooling adjacent violators” of \hat{p}_k/\hat{q}_k and updating \hat{p}_k and \hat{q}_k accordingly until \hat{p}_k/\hat{q}_k becomes nondecreasing for $k = 1, \dots, K$. Finally, the likelihood ratio test statistic for the particular zone Z is computed with these updated estimates and the MLE of p_k under the null hypothesis ($\hat{p}_{ok} = C_k/C$).

For a very simple illustration, suppose that we have 4 categories and that the number observed for each category is 1,2,3,4 inside a particular zone and 1,4,3,2 outside the zone. Then, $[\tilde{p}_k] = (1, 2, 3, 4)/10$, $[\tilde{q}_k] = (1, 4, 3, 2)/10$ and $[\tilde{p}_k/\tilde{q}_k] = (1, 1/2, 1, 2)$. Since $\tilde{p}_1/\tilde{q}_1 > \tilde{p}_2/\tilde{q}_2$, we combine the categories 1 and 2 and obtain $(\hat{p}_1/\hat{q}_1 = \hat{p}_2/\hat{q}_2, \hat{p}_3/\hat{q}_3, \hat{p}_4/\hat{q}_4) = ((1 + 2)/(1 + 4), 1, 2) = (3/5, 1, 2)$ with $\hat{p}_1 = (1 + 2)/10 \cdot 2/8$, $\hat{q}_1 = (1 + 4)/10 \cdot 2/8$, $\hat{p}_2 = (1 + 2)/10 \cdot 6/8$, $\hat{q}_2 = (1 + 4)/10 \cdot 6/8$, $\hat{p}_k = \tilde{p}_k$ and $\hat{q}_k = \tilde{q}_k$ for $k = 3, 4$.

References

1. Kulldorff M. A spatial scan statistic. *Communication in Statistics - Theory and Methods* 1997; **26**:1481–1496
2. Han D, Rogerson PA, Nie J, Bonner MR, Vena JE, Vito D, Muti P, Trevisan M, Edge SB, Freudenheim JL. Geographic clustering of residence in early life and subsequent risk of breast cancer (United States). *Cancer Causes and Control* 2004; **15**(9):921–929.
3. Brooker S, Clarke S, Njagi J, Polack S, Mugo B, Estambale B, Muchiri E, Magnussen P, Cox J. Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of western Kenya. *Tropical Medicine and International Health* 2004; **9**(7):757–766.
4. Jennings JM, Curriero FC, Celentano D, Ellen JM. Geographic Identification of High Gonorrhoea Transmission Areas in Baltimore, Maryland. *American Journal of Epidemiology* 2005; **161**(1):73–80
5. Sankoh OA, Yé Y, Sauerborn R, Müller O, Becher H. Clustering of childhood mortality in rural Burkina Faso. *International Journal of Epidemiology* 2001; **30**:485–492
6. Sudakin DL, Horowitz Z, Giffin S. Regional variation in the incidence of symptomatic pesticide exposures: Applications of geographic information systems. *Journal of Tox-*

- icology - Clinical Toxicology* 2002; **40**(6):767–773
7. Rushton G, Peleg I, Banerjee A, Smith G, West M. Analyzing Geographic Patterns of Disease Incidence: Rates of Late-Stage Colorectal Cancer in Iowa. *Journal of Medical Systems* 2004; **28**(3):223–236
 8. Sheehan TJ, DeChello LM. A space-time analysis of the proportion of late stage breast cancer in Massachusetts, 1988 to 1997. *International Journal of Health Geographics* 2005; **4**:15
 9. Dwass M. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* 1957; **28**:181–187
 10. Dykstra R, Kocher S, Robertson T. Inference for likelihood ratio ordering in the two-sample problem. *Journal of the American Statistical Association* **90**:1034–1040
 11. Barlow RE, Bartholomew DJ, Bremner JM, Brunk HD. *Statistical Inference Under Order Restrictions* New York: Wiley
 12. Klassen AC, Curriero FC, Hong JH, Williams C, Kulldorff M, Meissner HI, Alberg A, Ensminger M. The role of area-level influences on prostate cancer grade and stage at diagnosis. *Preventive Medicine* 2004; **39**(3):441–448
 13. Klassen AC, Kulldorff M, Curriero FC. Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors. *International Journal of Health Geographics* 2005; **4**:1
 14. Young JL Jr, Roffers SD, Ries LAG, Fritz AG, Hurlbut AA, (eds): *SEER Summary Staging Manual-2000: Codes and Coding Instructions*, National Cancer Institute, NIH Pub. No. 01-4969, Bethesda, MD, 2001
 15. Kulldorff M, Athas W, Feuer E, Miller B, Key C. Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos. *American Journal of Public*

Health 1998; **88**:1377–380

16. Kulldorff M. Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society* 2001; **A164**:61–72
17. Kulldorff M, Huang L, Pickle LW, Duczmal L. An elliptic spatial scan statistic. *Statistics in Medicine* 2006, in press.
18. Patil GP, Taillie C. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics* 2004; **11**:183–197
19. Duczmal L, Assunção R. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis* 2004; **45**:269–286
20. Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 2005; **4**:11

	Histological Tumor Grade	n	%
1	well differentiated	2289	12.1
2	moderately well differentiated	12335	65.1
3	poorly differentiated	4199	22.2
4	undifferentiated	124	0.7
	Total	18947	100.0

Table 1: The number of diagnosed prostate cancer cases in Maryland (1992-1997), used for analyses, broken down by tumor grade

	SEER Summary Stage	n	%
1	localized	15223	79.2
2	regional, direct extension only	2190	11.4
3	regional, lymph nodes only	255	1.3
4	regional, direct extension and lymph nodes	165	0.9
5	regional, other	145	0.8
7	distant	1235	6.4
	Total	19223	100.0

Table 2: The number of diagnosed prostate cancer cases in Maryland (1992-1997), used for analyses, broken down by disease stage

	Radius(km)	Categories	#O/#E	RR	LLR	P-value
cluster 1	46.46	(1,[2-4])	(2.43,0.80)	(2.46,0.80)	136.56	0.0001
cluster 2	45.94	(1,[2-4])	(0.68,1.04)	(0.67,1.08)	135.04	0.0001
cluster 3	25.11	(1,2,3,4)	(2.87,0.80,0.59,0.38)	(2.88,0.80,0.58,0.38)	73.34	0.0001
cluster 4	53.39	(1,2,3,4)	(1.40,1.04,0.68,0.34)	(1.40,1.05,0.68,0.34)	23.01	0.0001
cluster 5	8.50	(1,2,3,4)	(0.63,0.97,1.21,4.04)	(0.63,0.97,1.21,4.05)	17.51	0.0015

Table 3: Cluster detection analysis result for prostate cancer grade in Maryland using the ordinal scan statistic; #O/#E=ratio of numbers observed versus expected, RR=relative risk, LLR=log-likelihood ratio

	Radius(km)	Categories	#O/#E	RR	LLR	P-value
cluster 1	5.30	(1,[2-4],[5,7])	(0.86,1.25,2.04)	(0.86,1.27,2.20)	56.15	0.0001
cluster 2	42.92	(1,[2-4],5,7)	(1.04,0.91,0.71,0.71)	(1.07,0.87,0.61,0.61)	46.67	0.0001
cluster 3	19.56	(1,[2-7])	(0.75,1.97)	(0.74,2.01)	44.29	0.0001
cluster 4	126.86	(1,[2-5],7)	(0.91,1.28,1.51)	(0.90,1.32,1.60)	33.90	0.0001

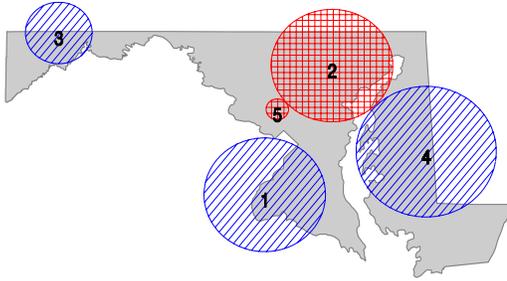
Table 4: Cluster detection analysis result for prostate cancer stage in Maryland using the ordinal scan statistic; #O/#E=ratio of numbers observed versus expected, RR=relative risk, LLR=log-likelihood ratio

	Radius(km)	#O/#E	RR	LLR	P-value
<u>Grade 2-4 (vs 1)</u>					
cluster 1	46.46	0.80	0.79	136.56	0.0001
cluster 2	45.94	1.04	1.09	135.04	0.0001
cluster 3	25.11	0.74	0.74	70.84	0.0001
<u>Grade 3-4 (vs 1-2)</u>					
cluster 1	5.99	1.28	1.33	28.79	0.0001
cluster 2	44.93	0.69	0.67	22.44	0.0001
cluster 3	10.34	0.69	0.68	17.81	0.0004
cluster 4	5.93	0.52	0.52	16.30	0.0011
cluster 5	6.39	1.69	1.79	15.40	0.0019
cluster 6	40.44	0.57	0.57	14.00	0.0080
cluster 7	13.23	3.24	3.25	13.49	0.0131
<u>Grade 4 (vs 1-3)</u>					
cluster 1	12.86	3.22	3.96	17.17	0.0001
cluster 2	15.42	0.19	0.16	12.06	0.0144

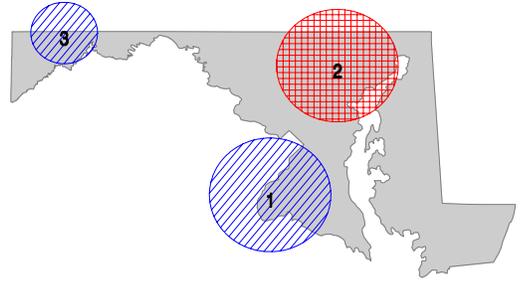
Table 5: Cluster detection analysis result for prostate cancer grade in Maryland using the Bernoulli-based scan statistic; #O/#E=ratio of numbers observed versus expected, RR=relative risk, LLR=log-likelihood ratio

	Radius(km)	#O/#E	RR	LLR	P-value
<u>Stage 2-7 (vs 1)</u>					
cluster 1	85.32	1.18	1.31	45.83	0.0001
cluster 2	20.72	1.94	1.99	45.56	0.0001
cluster 3	41.23	0.71	0.68	34.56	0.0001
cluster 4	10.61	0.71	0.69	28.70	0.0001
<u>Stage 3-7 (vs 1-2)</u>					
cluster 1	3.85	1.98	2.07	46.73	0.0001
cluster 2	42.40	0.78	0.68	32.61	0.0001
cluster 3	32.16	0.66	0.63	18.50	0.0002
cluster 4	82.88	1.51	1.56	14.46	0.0045
<u>Stage 4-7 (vs 1-3)</u>					
cluster 1	3.85	2.08	2.23	46.87	0.0001
cluster 2	42.41	0.74	0.62	41.68	0.0001
cluster 3	32.21	0.63	0.60	18.16	0.0003
cluster 4	83.24	1.56	1.62	14.88	0.0030
<u>Stage 5,7 (vs 1-4)</u>					
cluster 1	3.85	2.18	2.34	48.21	0.0001
cluster 2	49.68	0.71	0.60	42.58	0.0001
cluster 3	36.82	0.65	0.61	19.28	0.0001
cluster 4	78.06	1.48	1.54	13.03	0.0169
<u>Stage 7 (vs 1-5)</u>					
cluster 1	4.39	2.07	2.25	42.97	0.0001
cluster 2	46.30	0.70	0.59	39.76	0.0001
cluster 3	36.82	0.62	0.58	21.14	0.0001
cluster 4	55.86	1.67	1.75	17.14	0.0005

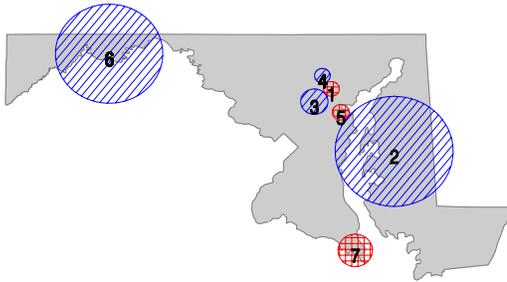
Table 6: Cluster detection analysis results for prostate cancer stage in Maryland using the Bernoulli-based scan statistic; #O/#E=ratio of numbers observed versus expected, RR=relative risk, LLR=log-likelihood ratio



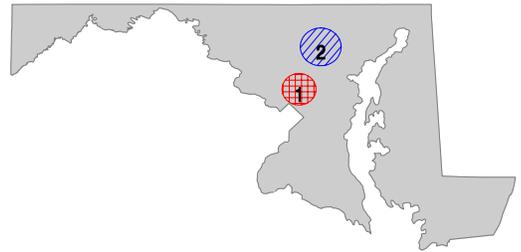
Map 1: Clusters based on Ordinal Model for Grade



Map 2: Clusters based on Bernoulli Model for Grade 2-4 (vs 1)

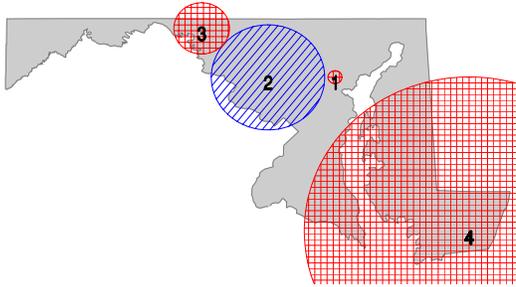


Map 3: Clusters based on Bernoulli Model for Grade 3-4 (vs 1-2)

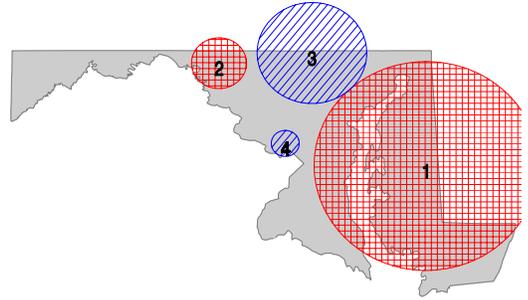


Map 4: Clusters based on Bernoulli Model for Grade 4 (vs 1-3)

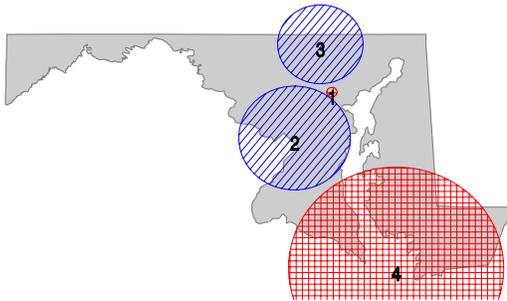
Figure 1: Cluster detection analysis results for prostate cancer grade in Maryland using the spatial scan statistic for ordinal data (Map 1) versus using the Bernoulli based spatial scan statistic (Map 2-4); red areas with crosshatched lines=clusters with high rates of higher grade, blue areas with parallel lines=clusters with low rates of higher grade



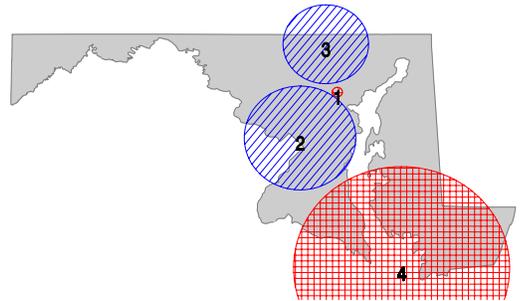
Map 1: Clusters based on Ordinal Model for Stage



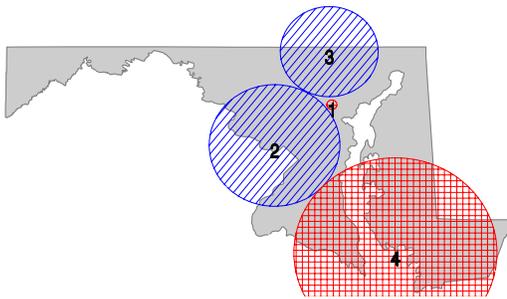
Map 2: Clusters based on Bernoulli Model for Stage 2-7 (vs 1)



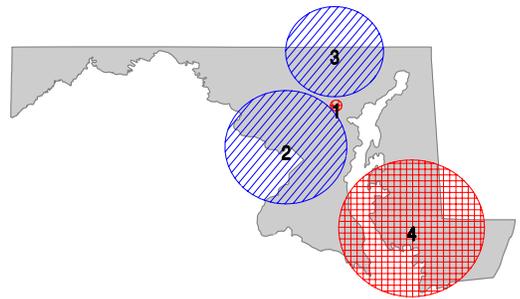
Map 3: Clusters based on Bernoulli Model for Stage 3-7 (vs 1-2)



Map 4: Clusters based on Bernoulli Model for Stage 4-7 (vs 1-3)



Map 5: Clusters based on Bernoulli Model for Stage 5,7 (vs 1-4)



Map 6: Clusters based on Bernoulli Model for Stage 7 (vs 1-5)

Figure 2: Cluster detection analysis results for prostate cancer stage in Maryland using the spatial scan statistic for ordinal data (Map 1) versus using the Bernoulli based spatial scan statistic (Map 2-6); red areas with crosshatched lines=clusters with high rates of later stage, blue areas with parallel lines=clusters with low rates of later stage