Visual Analytics of Spatial Scan Statistic Results

Jin CHEN[1], Alan MacEachren[1], Eugene Lengerich[2]

[1]GeoVISTA center, Geography Dept, Pennsylvania State University
jxc93@psu.edu, maceachren@psu.edu

[2] Epidemiology Division, Department of Public Health Sciences,

Pennsylvania State University
elengerich@psu.edu

INTRODUCTION

Kulldorff's scan statistic[1] is a spatial scan statistics method for detecting and evaluating statistically-significant, spatial clusters (e.g. disease, crime, etc). The method and its software implementation – SaTScan – is used widely in an increasing number of applications including epidemiology and other research fields. Here, we abbreviate the method as SaTScan method. Many researchers have effectively applied SaTScan to small or medium size sets of geographically-referenced data (e.g. point data for cases in a city, counties within one or a few states). However, the method is sensitive to user-controlled parameter choices. Our research to address this problem prompted a broader question on the consistency of SaTScan results. This research employs visual analytics methods to (1) find and illustrate some limitations of SaTScan method, (2) facilitate tuning of SaTScan parameters to meet the needs of different categories of users, (3) enhancing the effectiveness of the method, particularly for relatively large datasets. The proposed methods are implemented in a software system called the Visual Inquiry Toolkits (VIT). We demonstrate our research by analyzing cervical cancer mortality data aggregated by county in the U.S. from 2000 to 2004.

BACKGROUND AND PROBLEM

SaTScan method assumes that disease events are randomly distributed under null hypothesis. The alternative hypothesis claims elevated risk inside a region as compared to the outside region. Originally developed for point data, the method scans the studied region with a large number of circles, and detects the most likely, significant cluster(s) represented by the circle(s). When applied to aggregated data, SaTScan results are sensitive to the parameter configuration, as demonstrated in Figure 2. A critical, user-controlled parameter is the maximum spatial cluster size (short as "*max-size*").The parameter sets an upper bound on the percentage of the total population at risk that can be contained by an identified cluster. The default *max-size*, 50% of the total population, seldom produces informative results with U.S. by county data. The task of determining the most

appropriate setting is challenging because a too-large *max-size* could hide small *core* clusters, and a too-small *max-size* could miss significant clusters in a larger size. SaTScan method provides no guidance setting the *max-size*. On the other hand, clusters in different size could meet different needs. For example, policy makers are often interested in larger clusters that have important implications for region-wide, policy-related initiatives; while epidemiologist are more interested in smaller clusters for disease prevention initiatives or etiologic investigations.

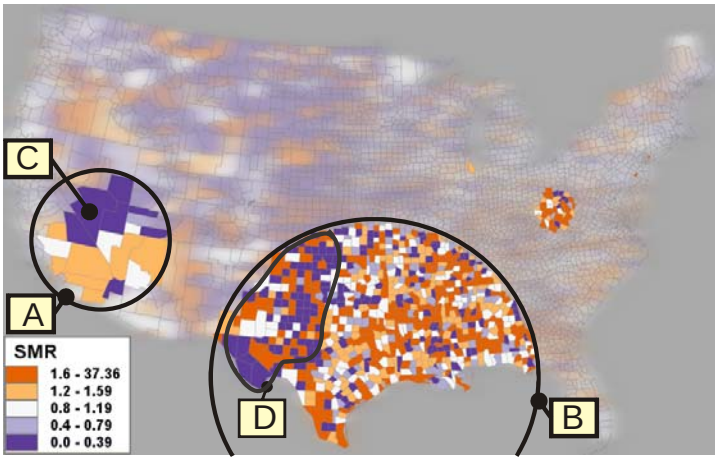

Figure 1 **High risk – orange, SMR ≥1.2; normal risk – white, SMR = 0.8 - 1.2; low risk – blue, SMR<0.8. A, B are two hetergeneous high risk clusters contain many low risk counties (e.g. C). The maximum cluster size is 40%, which means a cluster contains up to 40% total population in risk.**

Furthermore, SaTScan method trends to identify large-population clusters but low elevation in risk, and ignore small-population clusters contained within the area that have higher elevations in risk[2]. Hence, when applied to a large spatial dataset, SaTScan often reports heterogeneous clusters (e.g. A, B in Figure 1) – a cluster contains not only high risk sub-regions, but also a considerable proportion of low risk sub-regions (e.g. C, D in Figure 1). The choropleth map (Figure 1) displays standardized mortality ratio (SMR) of US cervical cancer. The SMR is expressed as the ratio of observed/expected deaths and reflects the relative risk of each county. Some of the high risk sub-regions within the cluster are more homogeneous and can reject the null hypothesis on their own strength; we refer to them as *core* clusters. A *core cluster* is statistically powerful enough to not only reject the null hypothesis on its own strength, but also enable the corresponding circle to expand to an extent large enough to encapsulate neighboring, low risk sub-regions(e.g. C, D in Figure1) while still rejecting the null hypothesis. The research reported here focuses on use of interactive geo-visualization to properly configure SaTScan to achieve more valid, consistent results, and to highlight the *core* clusters.

METHOD AND SOLUTION

To address the issue of sensitivity and consistency of SaTScan results, we proposed to run multiple scans with various of *max-sizes*. Specifically, we ran 50 scans using the *max- size* from 50% to 1% of population, with a step of 1%.

We first evaluate a scan result based on multiple criteria, and then select the representative results. Two important criteria are: (1) the total number of Two important criteria are: (1) the total number of counties contained in significant high risk clusters, denoted by T; and (2) the homogeneity of the clusters, measured by the overall variance in SMR of the clusters, denoted by V. If some scans of consecutive *max- sizes* (e.g. 9 sizes from 50% to 42%) produces similar T and V values, we select only one to represent the others. We also plot the 50 results in a map matrix, with each map displaying the significant, high-risk clusters reported by a result. We visually verify that the selected result can represent the others in terms of the location and shape of the clusters. Figure 2 displays the selected 12 results that reasonably represent the all 50 results. Those larger clusters (e.g. >15%) can provide policy-making initiative; while those smaller and more manageable clusters (e.g. <5%) are useful for disease prevention initiatives. We notice that some clusters are not consistent across the results (e.g., A, B in Figure 2). The phenomenon raises a question on the reliability of clusters across scale.

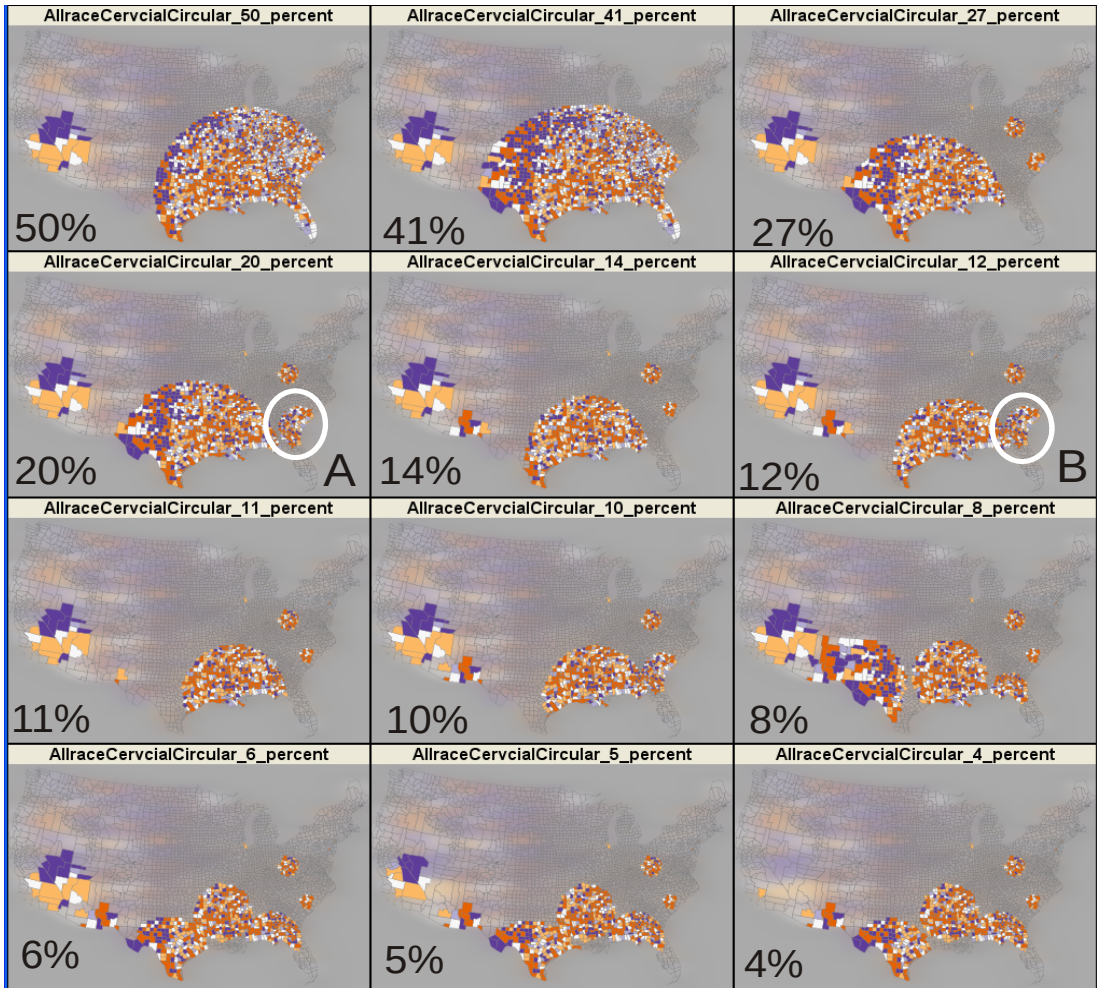

Figure 2. The maps matrix displays 12 SaTScan results. Only significant clusters (p-value <0.05) are displayed. A, B disappears in the result of 14%. Note: although the map of 41% covers a larger geogarphic area than the map of 50%, the later covers some densely-populated counties in eastern US.

Reliability is defined as the capacity of a test to give the same result - positive or negative, whether correct or incorrect - on repeated applications[3]. This research refers to reliability as the consistency that a place (e.g. a county) is reported as high risk by a set of scans. The reliability is measured by the following equation : R_i=C_i/S, where R_i is the reliability value for place *i*, *S* is the total number of scans, and C_i is the count that the place *i* is reported in a high risk region by these scans. The reliability has a value range from 0 to 1; 1 means all the scans report a place as high risk, 0 means no scan reports such. Reliability should be distinct from validity; the later refers to the probability that a cluster represents a true high risk region and is measured by the cluster's statistic significance. Reliability visualization can help to highlight the *core* high risk regions (possibly in irregular shape). In Figure 3, A and B can reject the null hypothesis for all the *max-sizes*; while those low-reliability regions (in light green and yellow) are unable to do so, as *max-size* reduced. Reliability visualization also helps to identify the tiny but important high risk clusters (e.g., cluster C in Figure 3) that can otherwise be easily hidden in the reported SaTScan clusters.

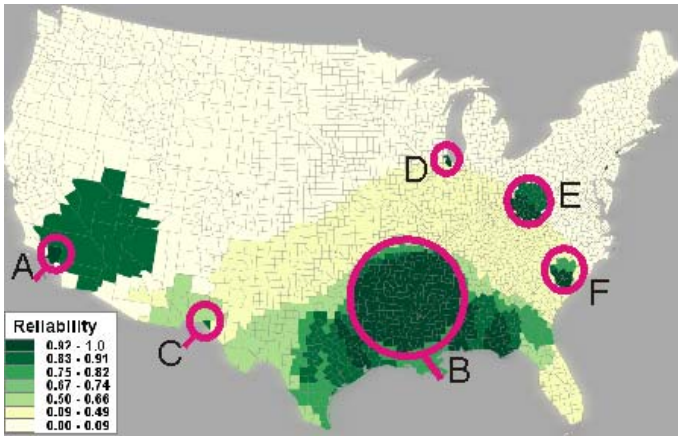

Figure 3. The map displays the reliability values calcualted from the 12 scans. The dark green regions are consistenly reported in high risk. Six *core* clusters are highlighed: A, B, C, D, E, F.

DISCUSSION

We present a visual analytics methods that help to enhance scalability, usability and effectiveness of SaTScan. The methodology presented is, however, extendable to other spatial scan statistic implementations with minor alterations. We will work on more related issues, including configuration for elliptic scanning, the visual understanding of the SaTScan result and its limitation. The research is also supported by grant CA95949 from the National Cancer Institute.

REFERENCES

1. Kulldorff M: **A spatial scan statistic.** *Communications in Statistics - Theory and Methods* 1997, **26:**1481-1496.
2. Boscoe FP, McLaughlin C, Schymura MJ, Kielb CL: **Visualization of the spatial scan statistic using nested circles.** *Health & Place* 2003, **9:**273-277.
3. Rothman KJ, Greenland S: *Modern epidemiology*. Philadelphia, PA: Lippincott-Raven; 1998.