RC23398 (W0408-068) August 13, 2004
Computer Science

IBM Research Report

Space-Time Clusters with Flexible Shapes

Vijay S. Iyengar
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Space-Time Clusters With Flexible Shapes (Extended Abstract)

Vijay S. Iyengar
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218, Yorktown Heights, NY 10598, USA
vsi@us.ibm.com

ABSTRACT

Detection of space-time clusters has an important role in epidemiology and public health. Here, we focus on the retrospective clustering analysis that is performed possibly triggered by an alarm from a surveillance system. Various approaches for detecting space-time clusters have been proposed and implemented. Many of these are based on the spatial scan statistic formulation. In this paper we present the issues to consider when choosing the shape of the cluster in such analyses. One criterion is the flexibility of the shape and its ability to model the phenomenon being studied. Another subtle and related factor is that with a more flexible shape clusters can appear more by chance. This will be reflected in the p-value obtained through Monte Carlo hypothesis testing. Choosing more complex cluster shapes can impact the computational requirements and also constrain the cluster detection approaches that could be applied. Importantly, the approach and heuristics used can impact key aspects of the results and their interpretation (e.g., p-value estimate). We use the New Mexico brain cancer data set to illustrate these tradeoffs. Clusters with two different shapes (cylinder, square pyramid) are detected in this data and compared. The results show the insights that can be gained from these shapes, individually and when put together.

1. INTRODUCTION

The *spatial scan statistic* developed by Martin Kulldorff [2, 3] has been applied to both retrospective and prospective applications in the domain of epidemiology and public health. A family of analysis methods have been developed for various models of the underlying phenomenon (e.g., Bernoulli model, Poisson model). We will utilize examples using the Poisson model in this paper to illustrate the concepts being presented. For the Poisson model, events are allowed to be generated by an inhomogenous Poisson process (e.g., number of disease events in a region over a time interval can be expected to be proportional to the corresponding population assuming no other factors).

These models have been implemented in a system for detecting space-time clusters (SaTScan) [4]. SaTScan detects space-time clusters using cylindrical windows (see Figure 1) with a circular geographic base and the height of the cylinder corresponding to some interval in time. Geographical locations are specified discretely (e.g., centers of counties) to SaTScan. Input data to SaTScan includes the number of cases and population information at these discrete locations at various times. SaTScan evaluates a set of cylindrical windows by considering all those spatially centered at any point in a user-specified grid and exhaustively varying the cylinder's radius and time duration. The evaluation computes the likelihood ratio of the alternative hypothesis that there is an elevated event rate within the cylindrical window and the null hypothesis that the rate is the same inside and outside the window. For the Poisson model, this likelihood function [2] is proportional to

$$LR = (c/n)^c([C-c]/[C-n])^{(C-c)}I() \qquad (1)$$

where C is the total number of cases over the entire space and time, c is the number of cases within the window, and n is the expected number of cases within the window under the null hypothesis. The indicator function, $I()$, is 1 when the window has more cases than expected under the null hypothesis and is 0 otherwise. The cylindrical window with the largest value of the likelihood function is the resulting cluster R. The multiple hypothesis testing problem is overcome in SaTScan using Monte Carlo methods by generating synthetic datasets for the entire space-time region in which the event counts are independently generated conforming to the Poisson model for each location and time. Each of these synthetic datasets is analyzed to determine its most dominant cluster and its likelihood function value. Using these Monte Carlo experiments one can determine the likelihood that the cluster R could have occurred by chance under the null hypothesis (p-value).

The use of cylindrical space-time windows for the clusters examined can limit the fit to the phenomenon being analyzed. For example, the cylindrical shape cannot model growth or shrinkage over time nor can it model movement over time. The square pyramid shape was proposed in [1] as an approach to overcome these limitations. Figure 2 illustrates this cluster shape using 2D and 3D views. The 3D view in Figure 2 shows a cluster growing with time. The axis of the pyramid need not be orthogonal to the two spatial axes allowing the cluster to model movement of the phenom-

2004 Syndromic Surveillance Conference, Boston, MA
....

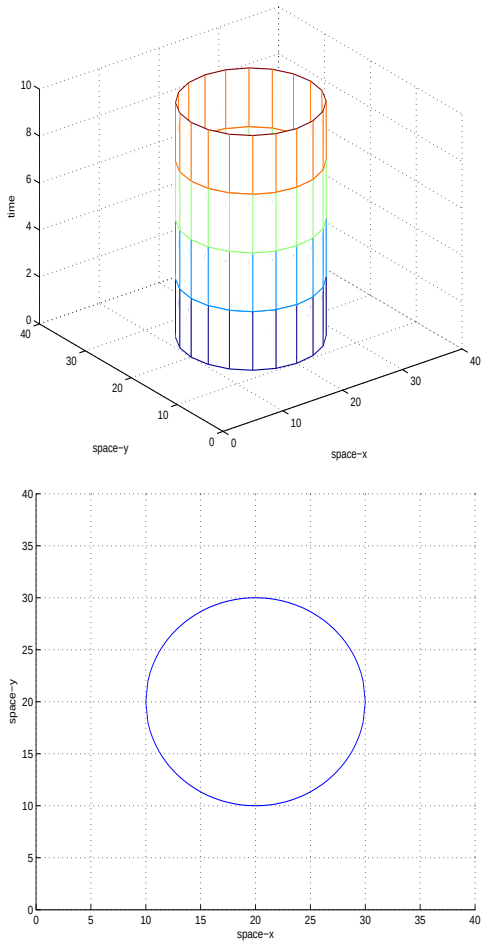

Figure 1: Cluster with a cylindrical shape (3D and 2D views)

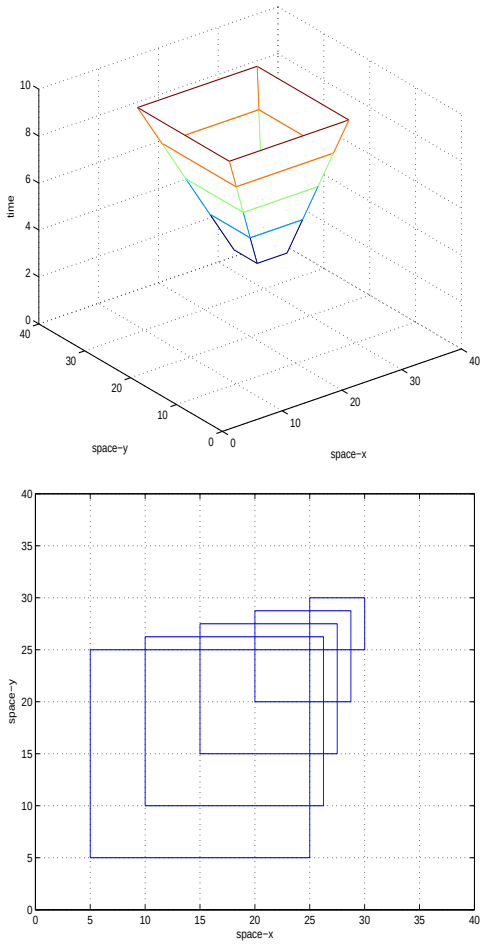

Figure 2: Cluster with a square pyramid shape (3D and 2D views)

ena. This is clear from the 2D view of Figure 2 where the squares represent the geographical extent at discrete times in the cluster time interval. The use of this flexibile shape results in greatly increased computational requirements. The computational issue is addressed in [1] by the use of a randomized search heuristic for the strongest cluster instead of the grid based pseudo-exhaustive approach used in SaTScan.

This paper explores the issues related to the choice of the shape used for space-time clusters.

2. CHOICE OF CLUSTER SHAPE

The first criterion to consider when choosing the cluster shape is the fit to the phenomenon being modeled. All the available information about the phenomenon can be used to determine which characteristics are important to model. For example, if modeling the growth of the phenomenon over time is important, it is preferable to use a shape that can represent this behavior. Note, choosing a shape arbitrarily that allows more flexibility than is needed has shortcomings also. As discussed above, the goal is to detect strong clusters that are also significant when compared to those that can occur by chance under the null hypothesis. An arbitrarily complex shape will increase the chances that the detected cluster is not significant since the chance of finding strong clusters in the synthetic data of the Monte Carlo experiments also increases.

We also need to balance the fit of the shape to the phenomenon with the computational need for the shape being considered. This second criterion, namely the computational need, has to be considered in conjunction with the search algorithm used. A search algorithm could be exhaustive by considering all possible clusters of the chosen shape. We will elaborate on this using an example. Consider the retrospective analysis of the New Mexico brain cancer data [5] using cylindrical space-time clusters. The disease occurences and population counts are provided for each of the 32 counties included in the data. An exhaustive analysis of cylindrical time-space clusters would have to consider all possible circular cross-sections, each being represented uniquely by the subset of counties included. Note that using a regular grid for the centers of the circular cross sections and then exhaustively considering all possible radii may or may not be exhaustive depending on the positions of the county centers and the choice of the grid. For cylindrical clusters it may be practical to choose a grid fine enough to be spatially exhaustive for a given data set. However, this approach may not be computationally effective for all data

sets and efficient algorithms that guarantee exhaustive exploration by cylindrical clusters need to be developed. Exhaustive methods may not be practical for more complex shapes. For example, the computational need is significantly higher for the square pyramid shape used in [1]. The heuristic search using the randomized algorithm proposed in [1] is a practical solution to detect square pyramid clusters in retrospective analysis. However, further work is needed to characterize the p-value computed by any method that is not guaranteed to be exhaustive.

We will illustrate the use of these criteria by evaluating the applications of two different cluster shapes to the New Mexico brain cancer data [5]. This data set was analyzed using cylindrical clusters with cross-sections restricted to have one of the 32 county locations as its center [3]. Suppose we want to extend the analysis using a more complex shape that can model both growth (or shrinkage) and movement over time. We restrict our consideration to convex 3D shapes since allowing non-convex shapes is overkill for our modeling goal. The 3D convex hull would be the least restrictive convex shape but it is still too general for the goal at hand (e.g., we need to model either growth or shrinkage but not both). Truncated pyramids are adequate to model growth (or shrinkage) over time. The pyramid can model movement if its axis is not restricted to be orthogonal to the spatial plane. We can limit the degrees of freedom by choosing a regular polygon for the pyramid cross-section. While we will use the square cross-section as an example in this paper, similar analysis can be performed with other regular polygons for the cross-section. We could have increased the flexibility by allowing irregular polygons for the cross-sections. However, our attempt at using an irregular polygon (rectangle) for the cross-section of the pyramid was not successful. It was significantly harder to get good convergence behavior for the randomized search algorithm with this extra degree of freedom. As mentioned before, we would also expect some impact on the p-values if the cross-sections were not restricted to regular polygons. We also considered the truncated cone as another cluster shape candidate. A regular polygon was chosen over the circle for the cross-section since the computations with planes in the case of the polygon was simpler involving linear equations.

In the next section, we will use the New Mexico brain cancer data [5] to compare the results of the analyses using two shapes, the cylindrical and square pyramid clusters.

3. EXPERIMENTAL RESULTS

The data set [5] contains brain cancer occurrences in 32 counties in New Mexico for the period 1973 to 1991. Occurrences are aggregated at the temporal granularity of a year. Population information is provided for each year. Three covariates are provided: *age group*, *gender* and *ethnicity*. First, we will consider only the first two covariates in Section 3.1. Then we will add the third covariate (*ethnicity*) and discuss the impact of this addition in Section 3.2.

3.1 Considering Covariates: Age Group and Gender

The Poisson formulation for the spatial scan statistic provides adjusting for covariates using indirect standardization [2]. In this section, we adjust for the two covariates, age group and gender. Hence, we are assuming that both these covariates are relevant to the disease being analyzed and the analysis in this subsection is intended to find clusters that cannot be explained by these two covariates.

Log likelihood ratio	13.70
Number of cases	265 (195.33 expected)
Overall relative risk	1.357
p-value	0.004
Centroid coordinates	(89, 81)
Cross-section radius	50.25
Time frame	1985-1989

Table 1: Cylindrical cluster results using fine grid for centroids

First, we will present results for the cylindrical clusters detected by using the SaTScan system [4]. We use a fine grid of size 1 Cartesian coordinate to perform the analysis. The characteristics of the strongest cluster detected are given in Table 1. This cluster extends for 5 years from 1985 to 1989 and includes 12 counties. Note that we use a fine grid in this SaTScan application to better approximate an exhaustive analysis for cylindrical clusters. For example, the cluster detected using the default mode when no grid is specified is weaker (Log likelihood ratio = 11.07, p-value = 0.013, and includes 16 counties over the same 5 year period) since it misses analyzing many potential cylindrical clusters.

Log likelihood ratio	16.918
Number of cases	284 (204.92 expected)
Overall relative risk	1.386
p-value	0.038
Time frame	1982-1989

Table 2: Square pyramid cluster results

Next, we present the results for square pyramid cluster detected using the method described in [1]. The characteristics of the strongest cluster detected by this heuristic search are given in Table 2. The number of cases included in this square pyramid cluster is somewhat larger than the number in the cylindrical cluster above. It also extends over a longer period of time. The p-value of 0.038 computed using 999 Monte Carlo replications is much higher than the 0.004 value in Table 1, but the cluster is significant using the threshold of 0.05.

The cylindrical and square pyramid clusters can be compared using the 3D and 2D views in Figure 3. Consider the 2D view first. The county locations are marked by $*$ in this view. This view also shows the square cross-sections of the pyramid for each of the eight years (we will ignore the circle for now). For the 5 years (1985-1989) common to both clusters, the square cross-sections of the pyramid are shown using solid lines. The first 3 years (1982-1984) of the square pyramid cluster are not included in the cylindrical cluster and are marked with dashed lines. The square pyramid cluster originates with 6 counties at the start in 1982 but expands to include 15 counties at the end in 1989. In contrast, the cylindrical cluster whose spatial extent is marked by the circle covers 12 counties for the 5 year period 1985-1989. The square pyramid cluster also indicates a movement over time in addition to the growth as some counties at the right of the 2D view get dropped in the later years.

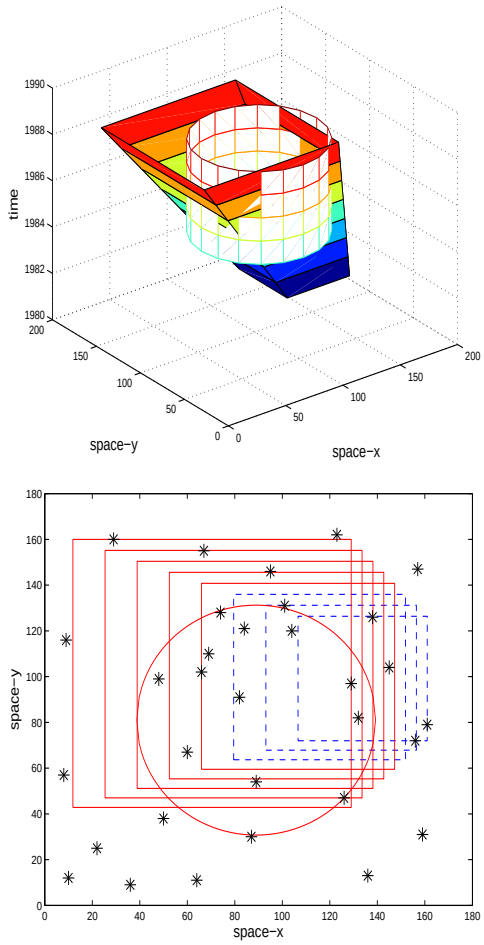

Figure 3: Comparing the cylindrical and square pyramid clusters (3D and 2D views)

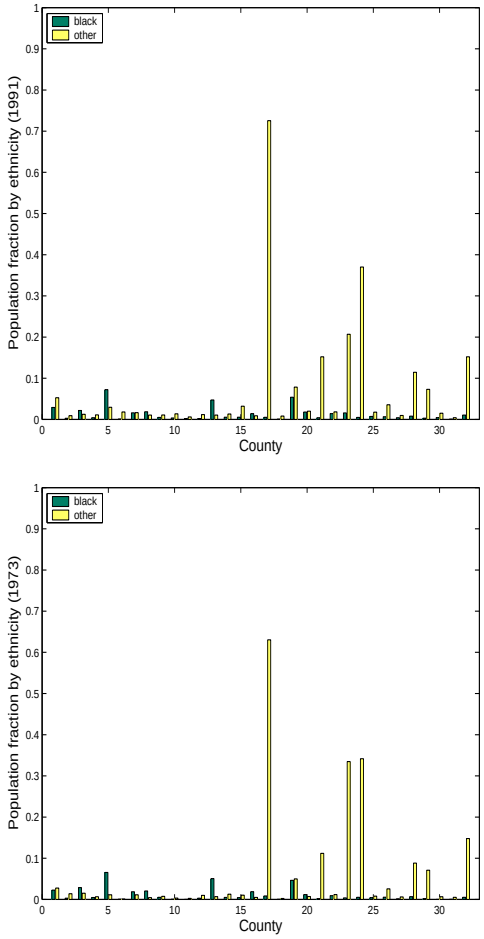

Figure 4: Distributions of the ethnicity covariate for the years 1973 (top) and 1991(bottom)

Together, the 2D and 3D views provide visualization of the cylindrical and square pyramid clusters showing key aspects like overlap. If we believe that the more flexible square pyramid cluster has indeed captured key characteristics of the phenomenon, then the visualization suggests that the detected cylindrical cluster could be construed as a reasonable approximation given its shape constraints.

3.2 Adding Covariate: Ethnicity

In this section, we add the covariate *ethnicity* to the analysis. This covariate can take one of three values: *white*, *black* or *other*. The spatial distribution of the covariate at the beginning (1973) and at the end (1991) of the time period is illustrated in Figure 4. The bar charts in Figure 4 show the population fractions for the ethnicity values, *black* and *other*, for each of the 32 counties. The figure clearly illustrates that wide variation of the ethnicities over the counties and also illustrates shifts in the distribution over time. Hence, factoring out this covariate could be expected to impact the cluster detected.

The results are interesting since there is a split based on the cluster shape. The strongest cylindrical cluster is the same as was seen in the two covariate case in Section 3.1. Its log likelihood ratio is lower (12.86) and the p-value higher (0.01), with this additional covariate factored out. In contrast, the square pyramid cluster detected in Section 3.1 is not the strongest cluster any more. Its log likelihood ratio drops to 16.05. Moreover, even the strongest square pyramid cluster detected with a log likelihood ratio of 16.208 is not significant with a p-value estimate of 0.054 (using the earlier threshold of 0.05).

4. CONCLUSION

The purpose of considering the different sets of covariates in the earlier section was to illustrate and compare the behavior of cluster detection methods with different underlying shapes. The actual set of covariates that needs to be adjusted for in any data set should be determined by the domain expert performing the analysis. The domain expert should also choose the cluster shape keeping in mind the phenomenon being modeled and analysis goals. For example, a flexible shape like the square pyramid can model growth (shrinkage) and movement of the phenomenon and maybe provide some insights on its origin. However, computational considerations may limit the analysis to utilize heuristic approaches that can only estimate the strongest cluster and more importantly its p-value. For retrospec-

tive analysis, we would argue that performing the analyses with more than one shape can lead to greater insights about the phenomenon. Moreover, we gain more confidence in these insights when the results of the analyses with different shapes support each other as illustrated in the example earlier.

5. ACKNOWLEDGMENTS

This material is based upon work supported by the Air Force Research Laboratory (AFRL) / (DARPA) Defence Advanced Research Projects under AFRL Contract No. F30602-01-C-0184. (Distribution Statement A: approved for public release, distribution unlimited). Any opinions, findings, conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the AFRL and/or DARPA.

6. REFERENCES

[1] V. Iyengar. On detecting space-time clusters. In *Proceedings of Tenth ACM International Conference on Knowledge Discovery and Data Mining*, pages 000–000, 2004.

[2] M. Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26(6):1481–1496, 1997.

[3] M. Kulldorff, W. Athas, E. Feuer, B. Miller, and C. Key. Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos. *American Journal of Public Health*, 88:1377–1380, 1998.

[4] M. Kulldorff and Information Management Services Inc. Satscan v. 3.1: Software for the spatial and space-time scan statistics. Technical report, 2002. URL=http://www.satscan.org/.

[5] National Cancer Institute. Brain cancer in New Mexico. Technical Report Data set (1973-1991), Division of Cancer Prevention, Biometry Research Group. URL=http://www.cancer.gov/prevention/bb/brain_nm.html.