Graph analysis for space-time scan statistics

Marcelo A. Costa
Universidade Federal de Minas Gerais, Minas Gerais, Brazil
e-mail: macosta@ufmg.br

The space-time scan statistic is a widely-used method for cluster detection in which both the geographic locations and the temporal length of the cluster are unknown. It relies on a cylinder scanning window in which the base represents geographic locations and the height represents the time component, simultaneously. Due to the strict shape of the scanning window, different geometries and graphical representations have been proposed to generate irregular cluster shapes. However, creating irregular cluster shapes from graph structures is not trivial. In addition to increased computational cost, detected clusters are normally very large and oddly shaped. Alternatives, such as growing clusters based on most connected vertices have improved detection and delimited the cluster shape. Nevertheless, graph statistics such as in-degree, betweenness centrality, etc., can be explored as potential measures for growing clusters. Furthermore, if the graph structure represents flow of populations in space and time, then dynamic graph statistics can be applied to grow cluster candidates.

Keywords: scan statistics, graph analysis, space-time scan statistics

1. Introduction

Spatial cluster techniques [Costa and Kulldorff, 2009] delineate boundaries around geographical areas in which the relative risk is higher as compared to the entire region under study. Therefore, these studies are useful for public health professionals to prioritize and optimize resources to act against disease outbreaks. The purely spatial scan statistic is a statistical cluster technique which assumes that the events follow either a Poisson or Bernoulli process. In addition, the spatial scan statistic usually assumes a fixed geometry for the cluster shape. By doing so, the cluster detection algorithm is improved. For example, the circular scan statistic requires the user to select only one parameter, which is the maximum window size. The algorithm modifies the center and the radius of the circle in order to scan the geographical region under study. For each different center and radius, a likelihood statistic is calculated. The circle with the highest statistical value represents the cluster candidate.

Similarly, the space-time scan statistic [Kulldorff et al., 1998] assumes a cylinder cluster shape in which base of the cylinder represents space, typically centered at the centroids with variable radii, and the height of the cylinder represents time. However, in order to detect irregular clusters, most successfully applied approaches rely on geographical graphs. For instance, computational heuristics for irregular cluster detection using a scan-based algorithm were proposed by Duczmal and Assunção (2004) using simulated annealing. Patil and Taillie (2004) introduced cluster detection using tessellation techniques, while Tango and Takahashi (2005) proposed an

exhaustive search within pre-sized circular clusters. Assunção et al. (2006) proposed a growth technique based on likelihood maximization in a graph structure, among others.

Nevertheless, most graph based cluster techniques assume that the underlying graph is not dynamic. That is, the graph is built using geographical adjacency information which does not change over time. Furthermore, graph based cluster techniques have a higher computational cost than standard methods.

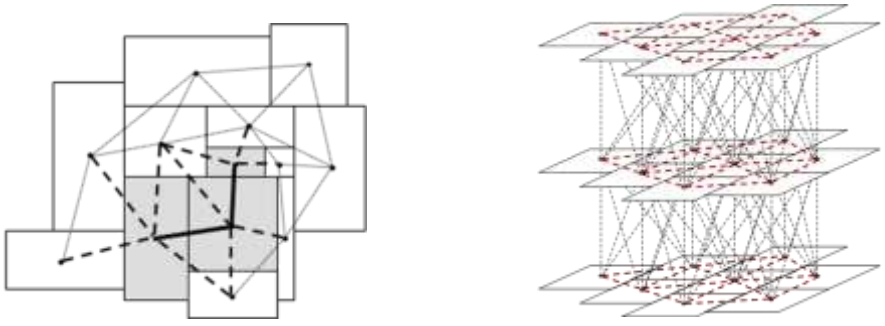
This work aims at introducing graph based cluster models for dynamic graph structures. We are particularly interested in developing an irregular scan statistic that accounts for the flow of populations within the graph. One suggested approach is to calculate statistical measures for the vertices, such as vulnerability statistics, and then use these statistics as random variables for the likelihood scanning model. By doing so, we first incorporate the dynamic information into a static graph structure and then apply irregular cluster techniques.

2. The graph based scan statistic

Graph based scan statistics usually generate cluster candidates using a graph structure built using geographical adjacency information, as show in Figure 1a. Different heuristics for building irregular cluster candidates are found in the literature. Most of these techniques aim at finding the cluster candidate that maximizes the likelihood ratio statistic. For example, if the Poisson process is assumed then the likelihood ratio test statistic is

$$\kappa(\hat{z}) = \sup_{z}\left(\frac{c_z}{\mu_z}\right)^{c_z}\left(\frac{C-c_z}{C-\mu_z}\right)^{C-c_z} \quad (1)$$

where z is the cluster candidate, C is the total number of cases in the studied region, c_z is the number of cases in cluster z, and μ_z is the expected number of cases under the null hypothesis of no cluster in the studied region. Thus, the cluster candidate $\hat{z}$ is the solution of Equation 1.



(a) Cluster growing process using geographical adjacency information

(b) A space-time graph structure built using geographical and temporal adjacency information.

Figure 1. Purely spatial (a) and Space-time (b) graph structures built using adjacency information.

One major advantage of a graph based scanning statistic relies on its ability to model both space and time information in one graph, as shown in Figure 1b. In this case, temporal adjacency information is used to connect the vertices in

different temporal layers. As a result, the graph structure can be manipulated as if it were generated using purely geographical information.

3. Graph statistics

Let the vertices of the graph represent geographical locations and the edges be movements of individuals between the edges at time t. In this case, due to the temporal component of the movements, the graph structure changes over time. Furthermore, the edge between any two vertices has a direction. A proper summary of the graph is the adjacency matrix, $\mathbf{A}$, which represents connections between vertices. Briefly, if $\mathbf{[A]}_{ij} = 1$ then there is a connection between vertex i and j. In symmetric graphs, $\mathbf{[A]}_{ij} = \mathbf{[A]}_{ji}$. Graphs that use only one adjacency matrix over a period of time are named static graphs. A common approach to deal with graphs with different adjacency matrices over time is to aggregate all movements over a fixed period of time, build a global static graph and then analyze its properties (Bigras-Poulin et al., 2006; Christley et al., 2005). An alternative approach is to repeat this process for a sequence of graphs, each one for a different period of time, and then check for common patterns among the resulting graphs (Robinson et al., 2007).

In practice, it is of interest to investigate properties of adjacency matrix $\mathbf{A}$ mostly correlated to possible epidemics within the vertices. According to Vernon and Keeling (2009) a global static graph often fails to capture the dynamics of epidemics. However, statistics related to the graph topology are known to be correlated to the vulnerability of vertices to disease. For instance, Nöremark et al. (2011) evaluated several different statistical measures for risk analysis. The most predictive statistical measures are: (a) *in-degree* of vertex j which is a measure of the number of contacts from other vertices, or simply $\Sigma_i \mathbf{[A]}_{ij}$. (b) The *betweenness* is a centrality measure of a vertex within a graph. The *betweenness* of vertex j is the proportion of the shortest paths between all pairs of vertices that pass through vertex j. It is expected that a vertex with a high *betweenness* value would have a higher vulnerability to disease. (c) The *total number of received individuals* is a variation of the *in-degree* measure. It accounts for the total number of individuals entering vertex j within the studied period of time. (d) The *static neighborhood* is the total number of vertices connected to vertex j, directly or indirectly. This measure is calculated using the static adjacency matrix $\mathbf{A} = \Sigma_t \mathbf{A}_t$ for the entire temporal period of the analysis. (d) The *ingoing infection chain* counts all direct and indirect contacts to vertex j. The temporal sequence by which the edges occur is taken into account. This measure was proposed by Nöremark et al. (2011) and it was found to be the measure most correlated to risk in graphs.

4. The Likelihood Model

Either the *in-degree*, or *betweenness*, or *static neighborhood*, or *ingoing infection chain* statistics represent a positive integer number, or a counted quantity. Thus, it may be reasonable to choose a Poisson likelihood model. Nevertheless, the distribution of the number of contacts within a graph might follow a Potential distribution, which is more asymmetric than the Poisson

distribution.

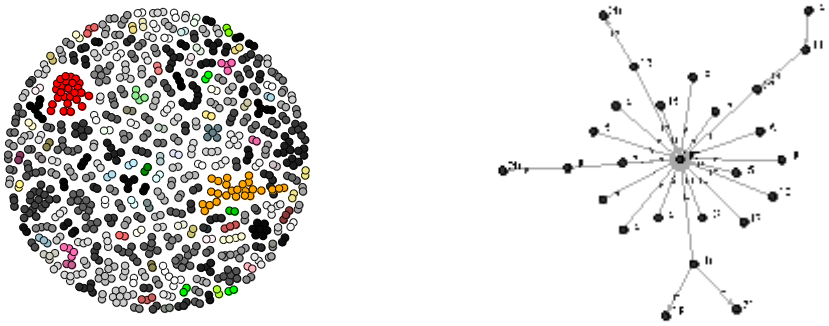
In space and space-time scan statistics, the commonly applied statistical models are Poisson and Bernoulli. These distributions represent random variables of counts of infected individuals and, in addition, the underlying population can also be used. Recently, different distributions such as Exponential, Multinomial, Normal, among others, have been proposed. However, an adequate likelihood model applied to scan statistics on graph structures was not found in the literature.

5. The application

The movements of cattle in Brazil are currently recorded in electronic format through the issuance of animal movement permit (GTA). A single GTA stores information about the movements of animals between two farms, such as the farm where the movement starts (origin), the destination farm, the number of transported animals, the date of the issued of the GTA and the reason of transportation, among others. As previously stated, the GTAs can be seen as connections or edges in a directed graph structure in which the vertices are the farms. Using graph theory, measures of vulnerability, connectivity, and other statistical measures can be assigned to vertices or subgraphs of the graph. Specifically, it is of interest to identify vertices with greater vulnerability, that is, vertices that are more susceptible to highly infectious diseases or that accelerate the spread of infection. Nevertheless, a simple statistical analysis of the structure of the network might not adequately represent the true vulnerability of a vertex or a subgraphs because information such as the population at each vertex, the number of transported animals and the rate of infection of the disease should also be considered.

We analyzed records of animal movements occurring in the municipalities of Mato Grosso do Sul (Brazil) issued within the first 28 days of August 2009. During this period, 2,052 GTAs were issued concerning cattle movements between 1,149 properties, and a total of 52,038 animals were moved among the farms. Figure 2a shows the static graph for the studied period. The circles represent farms (vertices). Vertices grouped closely together represent subnetworks. It is evident from Figure 2a that movements between farms is very sparse. Nevertheless, two major subgraphs can be identified.

By means of simulation studies, the ingoing infection chain statistic was found to be strongly associated to vulnerability of farms to diseases, as shown in figure 3. The higher the ingoing infection chain, the higher the vulnerability of a farm. Therefore, the ingoing infection chain statistic can be used as a preliminary statistic to investigate clusters of critical vertices.



(a) Graph of movements of cattle for the month of August 2009. Circles represent farms (vertices). Circles grouped closely together represent interconnected farms.

(b) Subgraph with the most vulnerable vertex and the largest number of vertices. Numbers close to the vertices represent simulated vulnerability. Numbers on the edges indicate infection transmission rate among two vertices.

Figure 2. Graph of movements of cattle for the month of August 2009 (a), and subgraph with the most vulnerable vertex and the largest number of vertices (b).

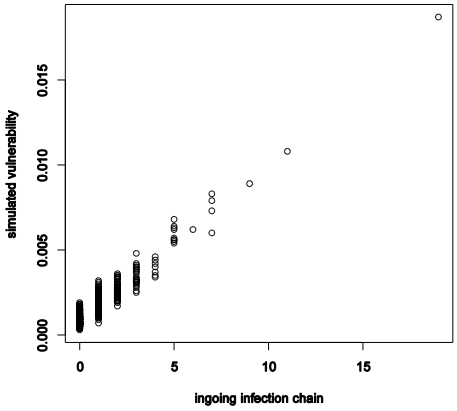


Figure 3. The *ingoing infection chain* statistics of the graph versus simulated vulnerability.

6. Discussion and conclusion

In this work we introduce the methodologies of space and space-time scan statistics using graph structures. This methodology allows the detection of irregular clusters in space and time. We propose to extend this method to the detection of cluster candidates in which the vertex information, aside from number of cases and populations, are graph statistics (i.e., the vertex vulnerability index). By doing so we aim at detecting cluster candidates in which the vulnerability of the vertices inside the cluster is higher than the vulnerability of the vertices outside the cluster. The final application is the analysis of dynamic graphs of animal movement in Brazil.

Acknowledgements

The author thanks CNPq, CAPES and FAPEMIG for financial support.

References

Assunção, R.M., Costa, M.A., Tavares, A., Ferreira, S., 2006. Fast detection of arbitrarily shaped disease clusters. Statistics in Medicine 25, 723–742.

Bigras-Poulin, M., Thompson, R. A., Chriel, M., Mortensen, S., Greiner, M., 2006. Network analysis of danish cattle industry trade patterns as an evaluation of risk potential for disease spread. Preventive Veterinary Medicine 76, 11–39.

Christley, R. M., Robinson, S. E., Lysons, R., French, N., 2005. Network analysis of cattle movement in Great Britain. Proceedings of the Society of Veterinary Epidemiology and Preventive Medicine, 234–243.

Costa MA, Kulldorff M. Scan statistics: methods and applications. Birkkäuser: Statistics for Industry and Technology; 2009. p. 129–52 [chapter 6].

Duczmal, L., Assunção, R.M., 2004. Simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. Computational Statistics and Data Analysis 45, 269–286.

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; 88:1377-1380.

Nöremark, M., Hakansson, N., Lewerin, S. S., Lindberg, A., Jonsson, A., 2011. Network analysis of catlle and pig movements in Sweden: Measures relevant for disease control and risk based surveillance. Preventive Veterinary Medicine 99, 78–90.

Patil, G.P., Taillie, C., 2004. Upper level set scan statistic for detecting arbitrarily shaped hotspots. Environmental and Ecological Statistics 11, 183–197.

Robinson, S. E., Everett, M. G., Christley, R. M., 2007. Recent network evolution increases the potential for large epidemics in the British cattle population. Journal of the Royal Society Interface 4, 669–674.

Tango, T., Takahashi, K., 2005. A flexibly shaped spatial scan statistic for detecting clusters. International Journal of Health Geographics 4.

Vernon, M. C., Keeling, M. J., 2009. Representing the UK's cattle herd as static and dynamic networks. Proceedings of the Royal Society B 276, 469–476.