Communications for Statistical Applications and Methods
2013, Vol. 20, No. 1, 53–62

DOI: http://dx.doi.org/10.5351/CSAM.2013.20.1.053

County Level Clustering on Alcohol and HIV Mortality

Byeonghwa Park[1,a]

[a]Department of Business Administration, Keimyung University

Abstract

This study focuses on spatial/temporal relationship deaths caused by Human Immunodeficiency Virus (HIV) and Alcohol Use Disorder (AUD). Several studies have found links between these two diseases. By looking for clusters in mortality of Alcohol and HIV related deaths this study contributes to the field through the identification of exact spatial/temporal time of high and low occurrence risks based on the observed over the expected number of deaths. This study does not provide political or social interpretations of the data. It merely wants to show where clusters are found.

Keywords: Circular window, geographic clusters, likelihood ratio, scan statistics, space-time.

1. Introduction

Many people at risk for or Human Immunodeficiency Virus (HIV) positive tend to be heavy drinkers (Meyerhoff, 2001). In addition, people with Alcohol Use Disorders (AUD) are more likely than the general population to contract HIV (NIAAA, 2002). There are numerous studies that have been done on HIV and AUD since they are major public health problems. The use of spatial statistical methods has gained attention to different areas in order to evaluate the differences in rates observed from different geographic areas, identify disease clusters, and assess the significance of potential exposure (Waller and Gotaway, 2004). Among spatial statistic techniques spatial and space-time scan statistics are commonly used for geographical disease cluster detection in various disciplines such as epidemiology, psychology, criminology, geography, and statistics (Kulldorff, 1997). Spatial distributions of health-related factors are useful to identify where services are most needed and for making public health policy.

This study selects alcohol related deaths and HIV related deaths from the mortality database provided by the Virginia Center for Health Statistics and detects statistically significant spatial clusters. This study explores clusters for each type of mortality rates by scan statistics, and examines similarities and differences in both clusters. For mortality data, a Poisson model is used when the number of cases is compared to covariates of an underlying population at risk derived from the census. The count data used for this study is suitable for the Poisson model. SaTScan performed the scan statistics and the results were mapped out with ArcMap. SaTScan (developed by Kulldorff and Williams) is a free software to implement scan statistics to find clusters by analyzing spatial, temporal and space-time data (Kulldorff and Williams, 1997).

This research was supported by the Bisa Research Grant of Keimyung University in 2012.

[1] Corresponding author: Assistant Professor, Department of Business Administration, Keimyung University, 1095 Dalgubeoldaero, Dalseo-Gu, Deagu 704-701, Korea. E-mail: bhpark@kmu.ac.kr

2. Background and Data

The data used for this study were extracted from the Automated Classification of Medical Entities (ACME) for the Virginia Center for Health Statistics. This dataset was stored in an Access database format, as two split tables that contained continuous entries (each entry represents one death). In those tables there are many attributes related to the traits of the deceased individual, location, and cause of death. The obtained dataset spanned from 1999 to 2005. For unknown reasons the subset of data that spanned the year 1999 did not contain the exact day of death, and only contained the year; subsequently, this part of the dataset were omitted. The data that spanned 2005 were also omitted because the data was incomplete (*i.e.* this period contained many data gaps). The period for the study therefore spanned from 2000–2004. During this time period there were a total of 284,029 deaths in the state of Virginia.

Even though the mortality records contained a zip code field, county and city codes were instead selected as the spatial boundaries due to the small and irregular shape of the zip code areas. Seven small independent towns were merged into corresponding counties so that we have 135 unique locations in total in the state of Virginia.

The next step was to find the data associated with the boundaries of the counties and cities. For the study two forms were required: One (the shape files) was used to display the results from SaTScan analysis and the second (the longitude and latitude position for the centroids of each county and city) as the required input into SaTScan. Both data sets were obtained from the U.S. Census Bureau website.

In the study two specific causes of death were concentrated; those related to alcohol and those related to HIV. The total numbers of deaths for these causes were 936 and 1,331 for Alcohol and HIV, respectively. Depending on the type of model applied, different input files are required by SaTScan, therefore, the Poisson model was used for this study. The Poisson model requires three files such as case, population, and coordinate file. Case file contains the number of occurrences of an event, or death due to a particular cause, for each spatial location and period of time. If no event occurred for a particular time period it is excluded. In addition, the case file includes the covariates in the study that are added as additional columns. The second file needed for the Poisson input is the population file which contains the population for each permutation of covariate combination. The third file is the coordinate file that contains the centroid location of each spatial boundary in unit of latitude and longitude. Age and sex are known risk factors in most situations (Green *et al.*, 2003; Kulldorff, 1999); subsequently, two covariates (age and sex) were included in this study. The analyses are adjusted for age applying indirect standardization with 13 age groups: 0–24, 25–29, . . . , 75–79, and over 80 years. Note that the changes in population over the 5 year study period were ignored because the U.S. Census Bureau collects data every 10 years. According to U.S. Census Bureau, several counties such as Loudon, Spotsylvania, and Culpeper are the fastest growing counties in the nation in terms of population. Thus, the same population was applied for each county over 5 years.

Cases are Alcohol or HIV related death occurrences while population is the number of people that lived in that ZIP Code in a certain time given each permutation of the covariates. In this case the time will be during the same month and year. However, fixed population data was used for every month during the whole time period of the study due to the census only reporting the population size every ten years.

3. Methodology

The test statistic is based on a likelihood ratio test. Under the Poisson assumption, the likelihood ratio

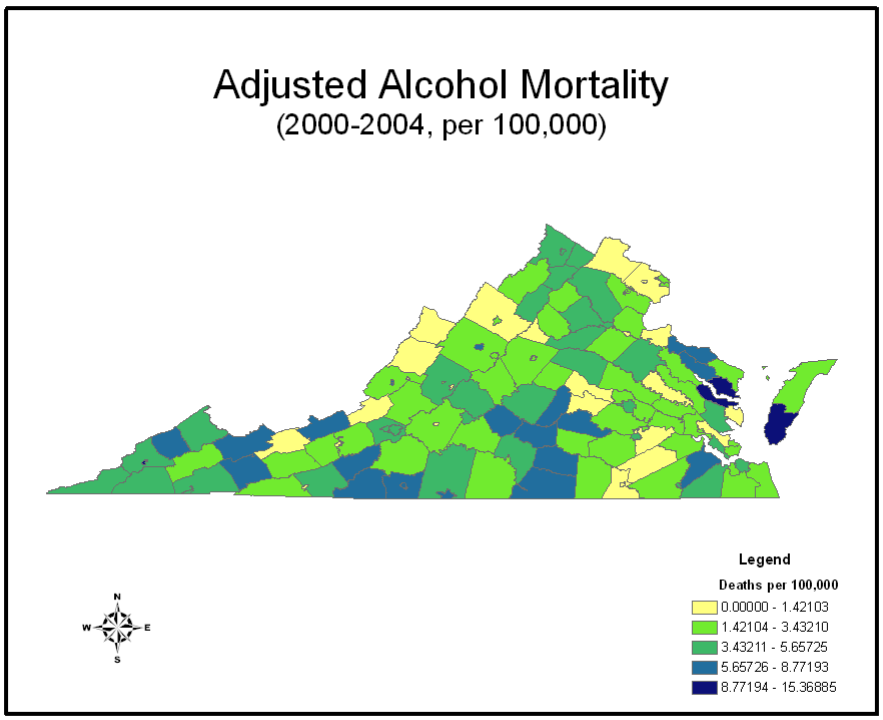

Figure 1: *Adjusted Alcohol Mortality*

is calculated as follow (Kulldorff, 1997):

$$\left(\left(\frac{c}{E[c]}\right)^{c}\left(\frac{C-c}{C-E[c]}\right)^{C-c}\right)I,$$

where $E[c]$ is expected number of events within the window under the null hypothesis. C is the total number of cases. I is an indicator function which is equal to 1 if observed number of cases c is greater than expected number of cases $E[c]$ and 0 otherwise. $c/E[c]$ and $(C-c)/(C-E[c])$ are proportional to the event ratios within and outside the window, respectively.

Assuming the Poisson model the expected number of cases can be calculated in a spatial location under the null-hypothesis, of no cluster. The expected number of observed cases $E[c]$, where c is the number of observed cases, is calculated using indirect standardization with the following equation;

$$E[c] = p * \frac{C}{P},$$

where C is the total number of cases and P the total population of all locations such that the summation of the populations of each individual area p is equal to P. Using the Poisson and the space-time permutation models the expected number can be adjusted to accommodate for covariates such as age, gender, and social class. To find the covariate adjusted expected number of observed cases there is a need to sum over the expected number for each covariate i.

$$E[c] = \sum_i E[c_i] = \sum_i p_i * \frac{C_i}{P_i}.$$

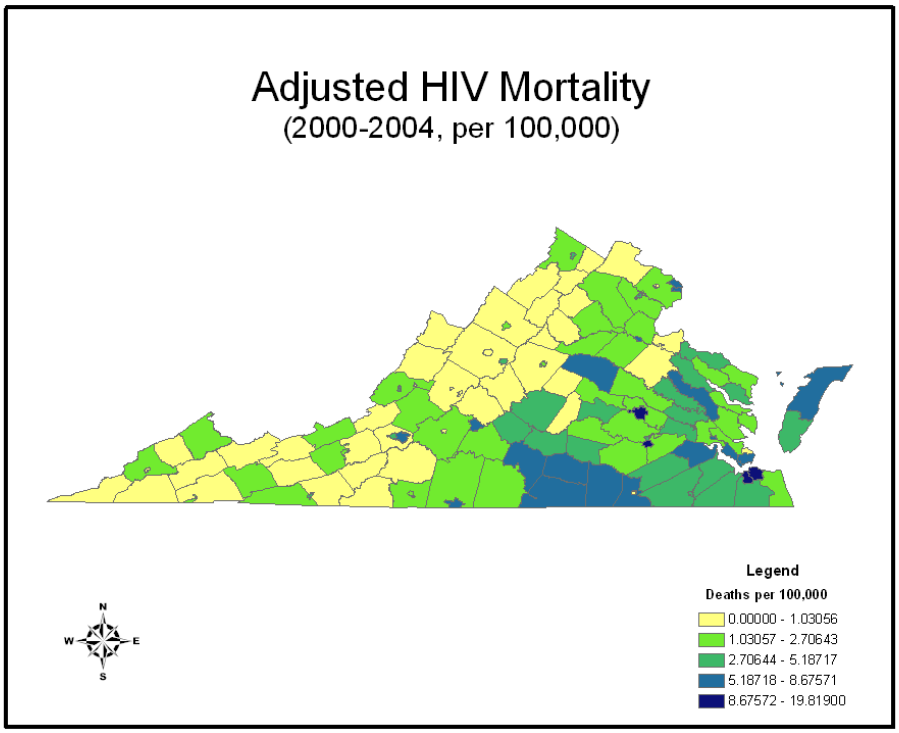

Figure 2: *Adjusted HIV Mortality*

In this study, age and gender are extracted and used as covariates for the Poisson model. The spatial scan statistic uses a circular moving window on the study area and allows its center to move over the area. A circular window of variable size detects clusters without prior knowledge of location or size, moving across the whole geographical area under the study (Donnan *et al.*, 2005; Li *et al.*, 2011). The likelihood ratio in every circular window location and size is calculated and the highest value chosen as the most likely cluster which is the least likely to have occurred by chance. The window radius is increased from zero to a maximum of 50% of the underlying population, and 999 replications were used in the Monte Carlo simulation for this study. *P*-value estimates for clusters are determined by the Monte Carlo simulation (Dwass, 1957). In SaTScan, at least 999 replications were recommended to guarantee excellent power for all types of datasets. The varying size of the moving window allows testing without presuppositions of cluster size or location (Hanson and Wieczorek, 2002). The p-value of 0.05, as a cut-off value, was used to reject or not reject the null hypothesis.

4. Results and Discussion

The patterns of alcohol and HIV mortality rates by county or city are shown in Figure 1 and Figure 2, respectively. The adjusted number of alcohol and HIV related deaths was calculated as follows:

$$\text{Adjusted Rate} = \left(\frac{100{,}000}{P_i}\right) * C_i,$$

where P_i is the population of a county or city and C_i is the number of deaths in the county or city.

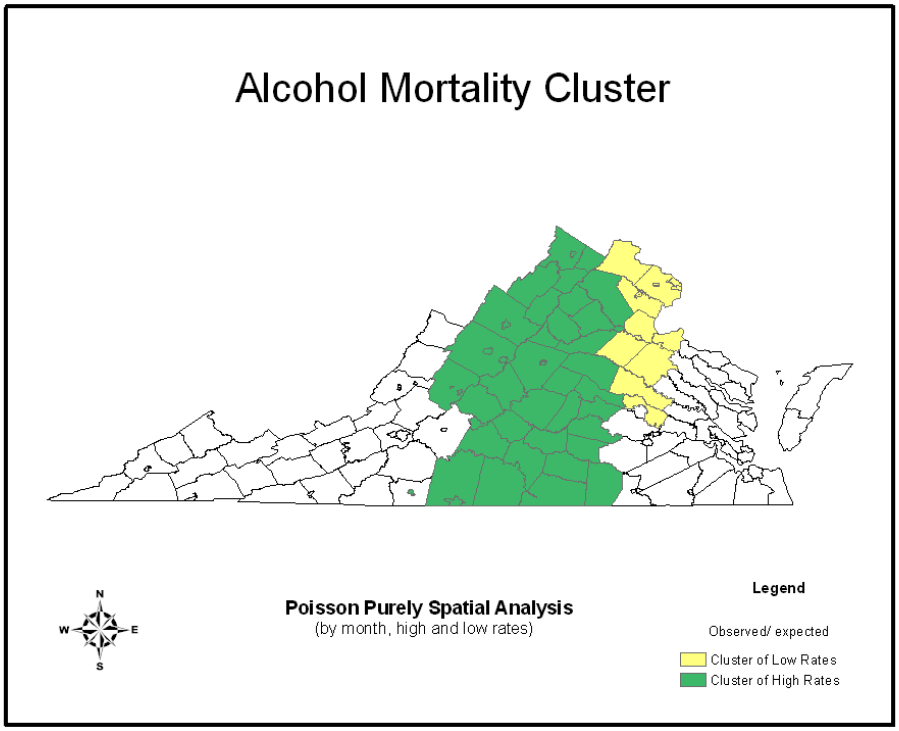

Figure 3: *Alcohol Mortality Cluster*

With adjusted rates per 100,000, simple comparison of mortality rates with each other can be shown.

Compared to HIV mortality, alcohol mortality rate spreads out over the state of Virginia. The high alcohol mortality counties (such as Lancaster, Middlesex, and Northampton) are in dark navy color. Lancaster County is bounded by the Rappahannock River to the south. Middlesex County is situated on the south side of the Rappahannock River. Northampton County is located on the southern half of the Delmarva Peninsula in Virginia, shown as a far right island in Figure 1.

The adjusted HIV mortality in Figure 2 shows that east part of Virginia has higher HIV mortality rates. Alcohol and HIV related mortality were explored to detect the most likely clusters using spatial and space-time scan statistics under the Poisson model. In order to see the similarities and differences between the two causes of deaths, it was used that different time precision and scanned areas with different rates such as having high and low rates, only high rates, and only low rates. A high cluster or high rate has more actual mortality cases than expected the mortality cases. A low cluster or low rate has less actual mortality cases than expected mortality cases. Therefore, the most likely cluster could be either a high rate or low rate.

Figure 3 depicts high and low rate clusters identified by purely spatial analysis with the time precision of month and scan cluster of high and low rates. Dark and light areas were identified as high rate and low rate clusters. Three clusters found by SaTScan are described in Table 1.

Martinsville was identified as the most likelihood ratio with 72.203702 and observed over expected value of 16.830. The secondary cluster (which includes Fairfax County) has the second highest value of likelihood ratio; however, it has different characteristics compared to the most likely cluster

Table 1: Characteristics of Alcohol Mortality Cluster by Poisson Purely Spatial

Cluster level	County/City	Likelihood ratio	Observed/expected	*p*-value
Most likely	Martinsville	72.203702	16.830	0.001
Secondary	Fairfax, Loudon, Prince William, Stafford, King George, Caroline, Hanover, Henrico ⋯	54.279056	0.545	0.001
	Fauquier, Pittsylvania, Louisa, Brunswick, Amelia ⋯	7.481943	1.261	0.067

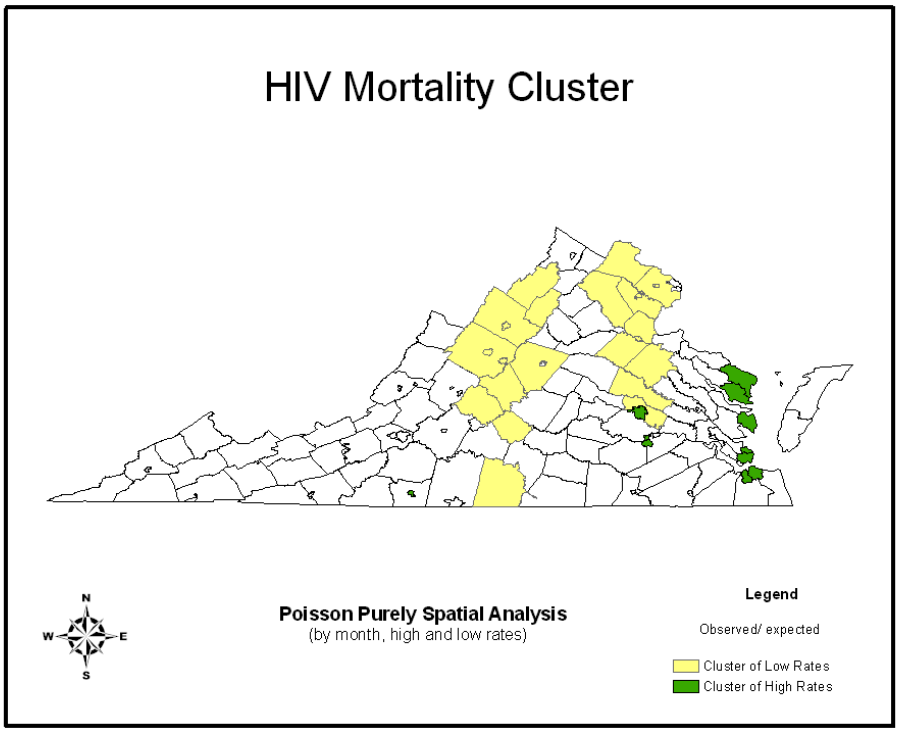

Figure 4: *HIV Mortality Cluster*

since the value of observed over expected cases for secondary cluster is below 1. The secondary cluster can be called a low rate cluster which means the number of observed death cases is less than the number of expected cases.

HIV mortality clusters (Figure 4) show that three counties (Colonial Heights, Petersburg, and Richmond) were identified in the most likely cluster. Martinsville was noticeable in both alcohol and HIV related mortality. Table 2 describes scan statistics result in detail.

Every county and city in the third secondary cluster containing Fairfax County in Table 2 and corresponding to Figure 4 completely overlap those in the first secondary cluster in Table 1 and corresponding Figure 3. The cluster with low rates of HIV mortality is likely to be a low cluster of alcohol related mortality since both the secondary clusters are low clusters.

The following figures are generated by the Poisson Purely Spatial model with the time precision of month and scan cluster of only high rates. This test is able to show which cluster is the most likely cluster among the high rate clusters.

Table 2: Characteristics of Alcohol Mortality Cluster by Poisson Purely Spatial

Cluster level	County/City	Likelihood ratio	Observed/expected	p-value
Most likely	Colonial Heights, Petersburg, and Richmond	204.694818	5.330	0.001
Secondary	Martinsville	142.414361	24.460	0.001
	Hampton, Lancaster, Mathews, Norfolk, Northumberland, Poquoson, Portsmouth	139.898791	3.687	0.001
	Fairfax, Frauquier, Loudon, Prince William, Spotsylvania $\cdots$	91.071836	0.483	0.001
	Nelson, Page, Shenandoah	34.015304	0.249	0.001

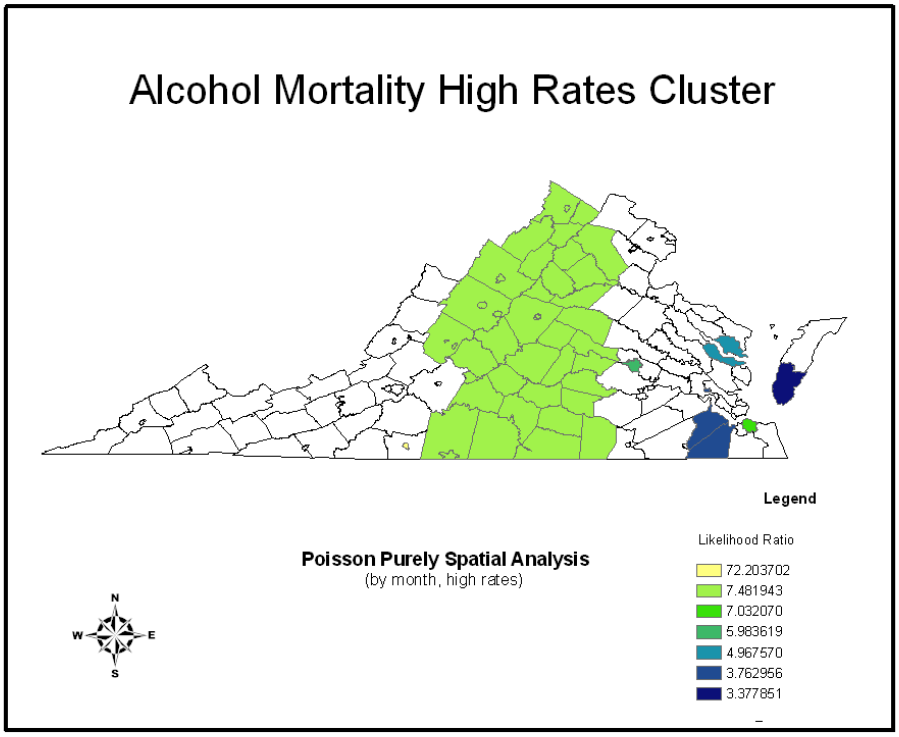

Figure 5: *Alcohol Mortality High Rates Cluster*

The most likelihood cluster is Martinsville in Figure 5. This result is very consistent with the Poisson Purely Spatial model of high and low rates. Unlike the Poisson Purely Spatial model of high and low rates, in the high rate clusters Richmond is a 4^{th} secondary cluster.

Figure 6 shows HIV mortality with only high rates and is interested in identifying which cluster is the most likely since all of clusters here are high rate clusters. From the result, Colonial Heights, Petersburg, and Richmond are detected as the most likely clusters with a likelihood ratio of 204.694818. Compared to other clusters, this cluster's likelihood ratio is two orders of magnitude larger than smallest scales. Overall, Martinsville, Lancaster, Northumberland, and Richmond were found as high rate clusters for both the Alcohol and HIV scans.

The next scan conducted was Poisson Space-Time in year intervals. Figure 7 illustrates alcohol mortality clusters with high and low rates. All detected clusters during the study period of 2000 to

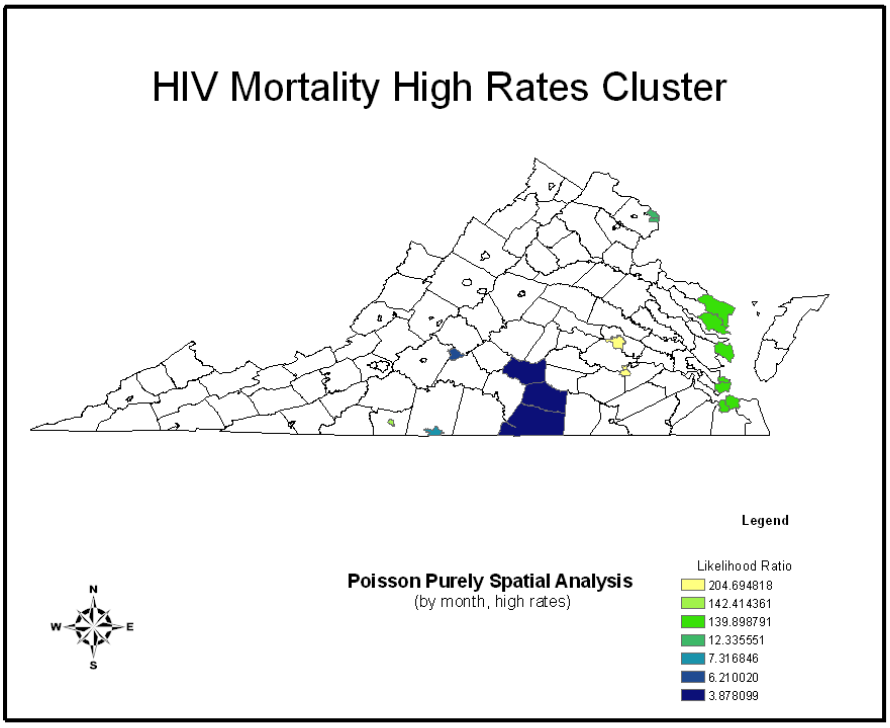

Figure 6: *HIV Mortality High Rates Cluster*

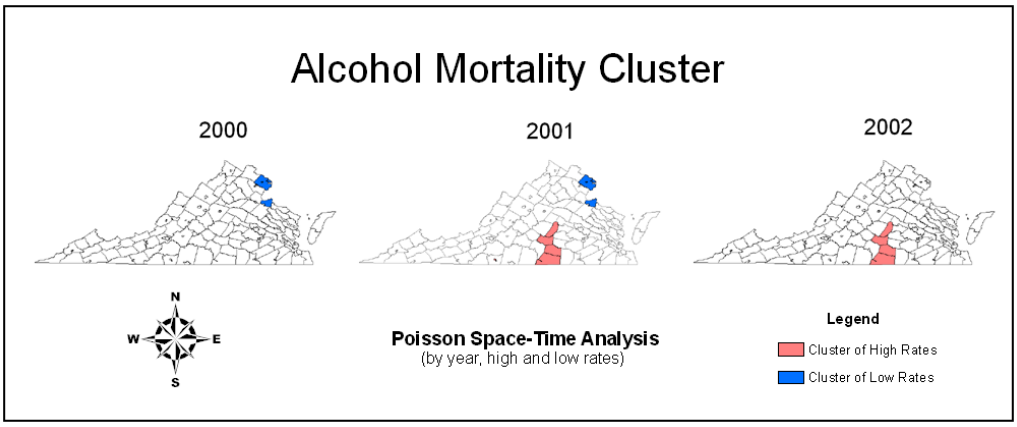

Figure 7: *Alcohol Mortality Cluster by Poisson Space-Time with year, high and low rates*

2004 had only a 2 year time interval. For example, high rate cluster, consisting of Cumberland, Prince Edward, Lunenburg, and Mecklenburg Counties was seen to span the years 2001 to 2002. This cluster had a p-value of 0.738 which is not statistically significant; however, the cluster of low rates detected had a p-value of 0.001 which is a statistically significant value.

The Poisson Space-Time clusters in the HIV dataset were identified. For this run every cluster was found to have the same p-value of 0.001. All clusters identified here were statistically significant.

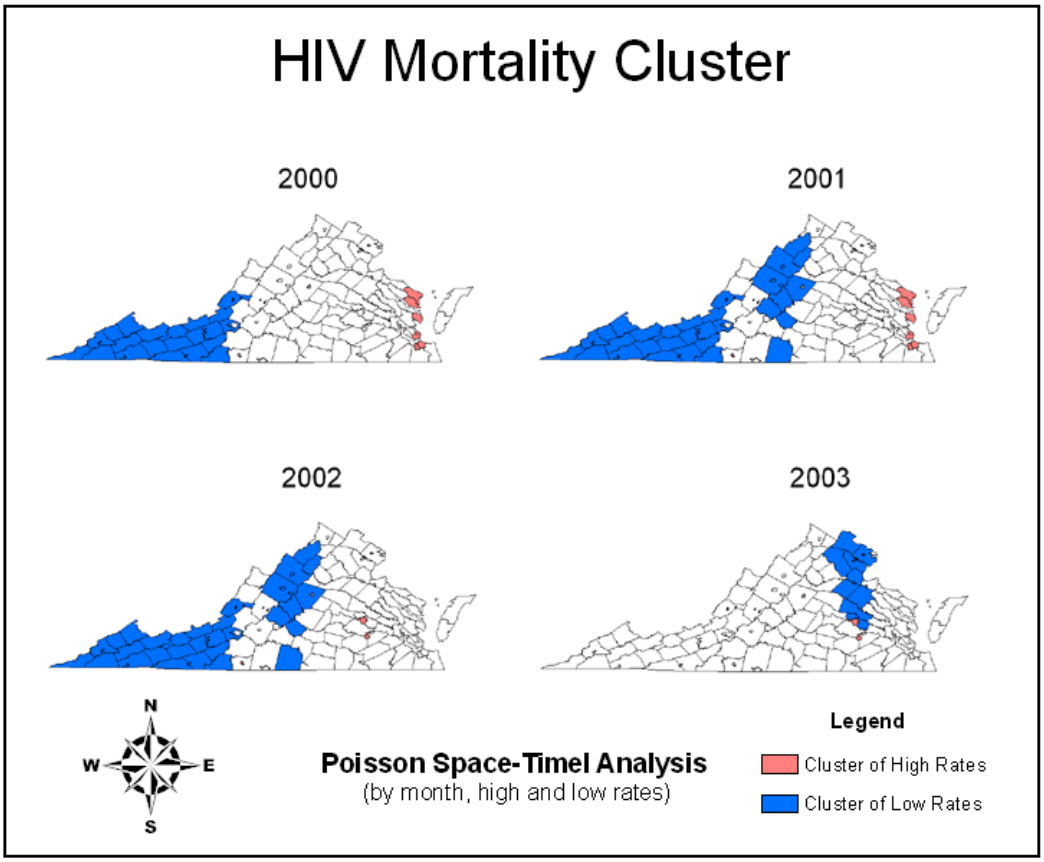

Figure 8: *HIV Mortality Cluster by Poisson Space-Time with year, high and low rates*

Figure 8 illustrates the changes of the clusters over a 4 year time span. Compared to adjusted HIV mortality rates in Figure 2, HIV mortality clusters identified in Figure 8 (from 2000 to 2004) overlapped with high valued adjusted mortality rates in the south-east. In 2003 Northern Virginia area was identified as a low rate cluster. Low rate clusters look to be moving towards the northeastern area of Virginia. As opposed to low rate clusters, high rate clusters moved inland. Clusters of high rates were detected in Richmond, Norfolk, and Martinsville. Most of the high rate clusters were found around coastal areas.

5. Conclusion

Scan statistics were conducted as a tool to find clusters within temporal and spatial data. The causes of death studied were deaths related to Alcohol and HIV, with presumed Poisson distribution, *i.e.* causes of death with low probability. With the software SaTScan Spatial, and Spatial-Temporal clusters have been found assuming the Poisson model. In addition, the results were successfully mapped out by ArcMap for further visual investigation. The clusters found are based on a ranking of likelihood ratio and can be either high or low risk. Where risk here refers to the observed over expected ratio, where if this risk is greater than one we say high risk, and for a value less than one we say low risk. This study applied the method for the state of Virginia using counties and cities as discrete spatial locations and

the temporal scale ranged from, 2000–2004. The results of this study show clustered areas of high and low risk ranked by the likelihood ratio.

It is not clear to us the exact spatial/temporal relationship between AUD and HIV and further interpretation by experts is needed to understand a more exact relationship.

References

Donnan, P., Parratt, J., Wilson, S., Forbes, R., O'Riordan, J. and Swingler, R. (2005). Multiple sclerosis in Tayside, Scotland: Detection of clusters using a spatial scan statistic, *Multiple Sclerosis Journal*, **11**, 403–408.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses, *Annals of Mathematical Statistics*, **28**, 181–187.

Green, C., Hoppa, R., Young, T. and Blanchard, J. (2003). Geographic analysis of diabetes prevalence in an urban area, *Social Science and Medicine*, **57**, 551–560.

Hanson, C. and Wieczorek, W. (2002). Alcohol mortality: A comparison of spatial clustering methods, *Social Science and Medicine*, **55**, 791–802.

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

Kulldorff, M. (1999). Spatial scan statistics: Models, calculations, and applications, In *Scan Statistics and Applications* (Eds., Glaz, J. and Balakrishnan, N.), Birkhäuser, Boston.

Kulldorff, M. (2007). SaTScan v 9.0.2: Software for the spatial and space-time scan statistics. Bethesda, MD: National Cancer Institute.

Kulldorff, M. and Williams, G. (1997). SaTScan v 1.0: Software for the space and space-time scan statistics. Bethesda, MD: National Cancer Institute.

Li, X., Wang, J., Yang, W., Li, Z. and Lai, S. (2011). A spatial scan statistic for multiple clusters, *Mathematical Biosciences*, **233**, 135–142.

Meyerhoff, D. (2001). Effects of Alcohol and HIV Infection on the Central Nervous System, *Alcohol Research and Health*, **25**, 288–298.

National Institute on Alcohol Abuse and Alcoholism (2002). Alcohol and HIV/AIDS, Alcohol Alert, 57, Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism.

Waller, L. and Gotaway, C. (2004). *Applied Spatial Statistics for Public Health Data*, Wiley, New Jersey.

Received October 31, 2012; Revised December 26, 2012; Accepted December 31, 2012