Published in final edited form as:

J Geogr Syst. 2005 May ; 7(1): 67–84.

Space–time visualization and analysis in the Cancer Atlas Viewer

Dunrie A. Greiling¹, Geoffrey M. Jacquez, Andrew M. Kaufmann, and Robert G. Rommel
BioMedware, Inc. Ann Arbor, MI 48104, USA

Abstract

This article describes the Cancer Atlas Viewer: free, downloadable software for the exploration of United States cancer mortality data. We demonstrate the software by exploring spatio-temporal patterns in colon cancer mortality rates for African-American and white females and males in the southeastern United States over the period 1970–1995. We compare the results of two cluster statistics: the local Moran and the local G^* , through time.. Overall, the two statistics reach similar conclusions for most locations, although where they disagree reveals some interesting patterns in the data. There are only two persistent clusters of colon cancer mortality, and these are clusters of low values.

Keywords

spatio-temporal; temporal GIS; clustering; animation; cancer

1 Introduction

This article describes the Cancer Atlas Viewer: free, downloadable software for the exploration of data in the National Cancer Institute's *Atlas of Cancer Mortality in the United States 1950–1994* (Devesa *et al.* 1999). This software helps users avoid the bandwidth constraints and delays inherent in web-based GIS, goes beyond current GIS technology to enable true spatio-temporal visualization, and provides cluster analysis statistics.

For chronic diseases such as cancer, which have long latency and can display significant spatial pattern, atlases of health data are an important resource. Atlases allow researchers and the public alike to evaluate hypotheses about geographic variation, such as clustering, and to formulate new hypotheses (Jacquez 1998, Moore and Carpenter 1999, Rushton *et al.* 2000, Jacquez and Greiling 2003). The identification of spatial pattern in mortality has stimulated research to elucidate causative relationships such as the association between snuff dipping and oral cancer (Winn, Blot *et al.* 1981); the association between shipyard asbestos exposure and lung cancer (Blot, Morris *et al.* 1980) and others.

Mortality atlases are available in print form, such as the *Atlas of United States Mortality* (Pickle *et al.* 1996) and the *Atlas of Cancer Mortality of the United States 1950–94* (Devesa *et al.* 1999) and in web format (see Table 1). Web atlases are increasingly available at the state and national levels as the technology for online mapping has matured. Both print and web atlases provide a considerable amount of data and statistics in an easy to understand visual format, but online atlases offer a level of interactivity not available in printed books, as the user can change the colors, zoom and pan, click through to data tables, and customize the maps to address a question or purpose not envisioned by the print map's creators.

¹ Author for correspondence: Dunrie@biomedware.com, voice 734.913.1098, fax 734.913.2201.

While online atlases provide greater flexibility and customization than print atlases, their use is hindered by performance limitations that result from Internet communication between the user's computer and the mapping engine. For example, if someone is using the Cancer Mortality Maps and Graphs website and wants to change the cancer site displayed in the map, the time period, or the color scheme, the user must send a request for the change to the map server, which then sends a new image of the map to the user's computer (Figure 1). The resulting delay can diminish the user's interest in interacting with the data. This delay is especially long when using slower dial-up rather than broadband connections.

An alternative to web mapping is to download a local version of the data for mapping and interaction on a desktop computer (similar to Figure 2). Once downloading has occurred, the time for map rendering is minimal and the user can explore the data more quickly. The Cancer Atlas Viewer is an example of a new type of information system, the Space-Time Information System, or STIS (described in Jacquez et al, this issue). Andrienko *et al.* (2003) review recent work in exploratory spatio-temporal visualization, with new software tools that incorporate time as a dimension of the data, going beyond traditional GIS approaches. This article describes a new software product for spatio-temporal data exploration and analysis, a specific viewer developed for the Cancer Atlas data. The advantages of this system include the fact that it is built on the STIS architecture described in Jacquez *et al.* (this issue), and this architecture supplies animated, interactive maps that allow researchers to visualize and explore the dynamic nature of cancer mortality patterns. The STIS approach enables exploratory spatio-temporal data analysis that may cue researchers to formulate new explanatory and other hypotheses regarding temporally dynamic, georeferenced health data.

In this article, we demonstrate the STIS approach by exploring colon cancer patterns for African-American and white females and males using NCI data. Among cancers, the highest mortality for men is from lung, prostate, and colon cancers respectively; for women it is lung, breast, and colon cancers, all of which demonstrate spatial pattern (Devesa *et al.* 1999). One challenge in exploring patterns for multiple groups is that there are low populations of African-Americans in rural areas of the midwest and western states. Because of low population numbers, the counts used to calculate the mortality rates are based on small samples and are therefore unstable—subject to fluctuations that may be due to chance. The NCI print and online atlas masks data based on few counts (< 6 deaths in the 5 year time period). We focus on the southeastern United States and Gulf Coast, including part of eastern Texas, Mississippi, Louisiana, Alabama, Georgia, Florida, South Carolina, and North Carolina. This region has high enough populations of African-Americans to avoid most rural areas becoming masked out, as geographies with a lot of missing (masked) data are unsuitable for spatial analysis. The southeastern US has been identified as a region of persistently high mortality (Cossman *et al.* 2003), though it is not the highest mortality region for colon cancer in the U.S. For colon cancer mortality rates, the southeast is exceeded by the northeastern states (Devesa *et al.* 1999).

Specifically, we assess the spatial pattern of mortality from colon cancer in the Southeast, using descriptive data visualization of the cancer data and cluster analysis, Moran's I , local Moran (I_i), and local G^* analyses, for state economic areas (SEAs). We assess the changes in spatial pattern by examining trends in I_i statistics and the persistence of clusters over time as well as the concordance of the two cluster detection methods (I_i and G^*). While the I_i and G^* statistics are related, we expected that their differences in form could lead to different findings, specifically that the G^* statistic would be less sensitive to the value at the cluster center ("ego") than the I_i and more susceptible to a few strong neighbor values.

2 Methods

2.1 Data description

The National Cancer Institute has released age-adjusted cancer mortality rates for U.S. counties, state economic areas (SEAs), and states, for 40 site-specific cancers, 4 groups (African-American females, African-American males, white females, white males), and for several time periods from 1950-1994. The rates are the number of cancers per 100,000 person-years, age-adjusted to the 1970 U.S. population standard age distribution. We focus here on the SEA datasets. SEAs are aggregations of counties within state boundaries that were similar according to 1960 socioeconomic data (U.S. Bureau of the Census 1966). The SEA data has better temporal resolution than the counties (counties have 20 or 25 year times only) and finer spatial resolution than the state datasets. Data for African-American males and females starts in 1970, while data for white males and females begins in 1950. More information on this data is available from the National Cancer Institute cancer mortality maps and graphs website (<http://www3.cancer.gov/atlasplus/>) and in the printed atlas (Devesa *et al.* 1999). The National Atlas has compiled metadata for this dataset, available at <http://www.nationalatlas.gov/cancerm.html>. We focus on colorectal cancer rates for African-American and white females and males for SEAs in 5-year time intervals from 1970 through 1994. We use the age-adjusted rates produced by the NCI: these rates are for 100,000 person-years and are adjusted to the 1970 age-classes (Devesa *et al.* 1999). We repeated this analysis for the county-level rates data, but did not include it in the write-up because of space constraints. The comparison between I_i and G^* conclusions for the counties geography was similar to the SEA results we present.

2.2 Software Description

The Cancer Atlas Viewer is the first implementation of the more general Space-Time Information System described in Jacques *et al.* (this issue). Although it has a unified graphical user interface, the software is built of several subcomponents, as shown in Figure 3. The STIS architecture consists of an event handler, data management, spatio-temporal data management (indexed for quick data queries), and a methods component. Currently, the Cancer Atlas software loads text files downloaded from the NCI website. This version of the graphical user interface works on Windows operating systems, although the underlying architecture is cross-platform and can be compiled for other operating systems. At the time of this writing, the Cancer Atlas Viewer software can be accessed through the Internet (<http://www.terraeer.com/products/atlasviewer.html>) free of charge. A more general version of this software that accepts other datasets is available from TerraSeer, called the TerraSeer Space-Time Intelligence System (STIS).

The Cancer Atlas Viewer and STIS software contain several statistical methods, from data transformations such as the Z-score standardization, to the creation of difference datasets, to the calculation of Moran's I , local Moran, and local G^* clustering statistics. The cluster statistics are evaluated with Monte Carlo randomization-based hypothesis testing.

The Cancer Atlas Viewer (and its STIS counterpart) has time as a dimension of the data. The spatial relationships among the observed objects (whether point objects or polygons) and the attribute data can be brought in as separate pieces. For instance, in the case of the NCI Atlas data, there is only one geography of the polygons (at the county, SEA, or state level) for the entire analysis. Although the outline of some US counties has changed over time, the NCI standardized it as one static geography for representation in GIS. Morphing polygons are easily represented in the STIS (Meliker *et al.* this issue), with counties appearing, disappearing, splitting and merging. Once the geography (e.g. county shapefile) is imported, the user then can import attribute data that is joined into a complete space-time dataset. The attribute data

can be imported as a time series (where the variables are valid for times specified as fields in the database file) or as a time slice (where the data is stored in a series of files or database records, each valid for a time interval specified on import). The latter is especially useful for importing static layers from a conventional GIS.

2.3 Z-score

The Cancer Atlas Viewer uses Z-score standardization to prepare the data for the Moran analysis. The Z-score standardizes the mortality rates by taking the observed rate, subtracting the mean rate for the entire region, and then dividing by the standard deviation. The Z-score is only one of several possible epidemiologically relevant standardizations of mortality data, including the standardized mortality rate or ratio (observed cases/expected). It is a required step for Moran's I and I_i analyses. A Z-score standardizes the mortality rate for area i , $m_{i,t}$, by its mean and standard deviation, creating a new variable $m^{\wedge}_{i,t}$

$$\widehat{m}_{i,t} = \frac{m_{i,t} - \bar{m}_t}{s_{m,t}} \quad (\text{Equation 1})$$

where $\bar{m}_t$ is the mean mortality at time t , and $s_{m,t}$ is the standard deviation of at time t . After Z-score transformation, all variables in a larger dataset have equal means (transformed mean = 0) and standard deviations (transformed $s = 1$), but different ranges. Negative Z-scores indicate the location is below the mean of the data, positive that it is above the mean. The magnitude of the Z-score is the distance in standard deviation units away from the mean.

2.4 Difference datasets

Cancer Atlas also calculates difference datasets, to allow the user to view change maps. Absolute change in cancer mortality, m_i for area i between times t and $t+1$ is calculated as:

$$\Delta a_{i(t,t+1)} = m_{i,t+1} - m_{i,t}. \quad (\text{Equation 2})$$

2.5 Moran's I

Moran's I (Moran 1950) is a spatially weighted correlation coefficient used to detect spatial pattern such as clustering (Equation 4)

$$I_t = \frac{N \sum_{i=1}^N \widehat{m}_{i,t} \sum_{j=1}^N w_{ij} \widehat{m}_{j,t}}{W \sum_{i=1}^N \widehat{m}_{i,t}^2} \quad (\text{Equation 4})$$

$$W = \sum_{i=1}^N \sum_{j=1}^N w_{ij}, i \neq j \quad (\text{Equation 5})$$

Here N is the number of regions, w_{ij} is a weight denoting the strength of the connection between areas i and j , drawing on equation 1 to calculate Z-scores. W is the sum of the weights (equation 5). We used first-order queen neighbors with row-standardization. Weights are $w_{ij} = 1/(\# \text{ neighbors})$ for first order queen neighbors, 0 for all other locations. Hence the weights for each location sum to 1 and the sum of the weights, W , is equal to the number of locations, N . It is a global statistic, in that there is one value for an entire geography. The range of the statistic is usually between $(-1, 1)$, but its range depends on the characteristics of the spatial weights set used (Cliff and Ord 1981). Positive spatial autocorrelation means that surrounding areas have similar mortality rates, negative values indicate surrounding areas have different rates. Because

we repeated the Moran's I calculation for 5 SEA time intervals for each group, we used a Bonferroni adjustment for multiple comparisons, lowering the alpha level for significance to 0.01 ($\alpha = 0.05/5$).

2.6 Local Moran

The local Moran test (Anselin 1995) detects local spatial autocorrelation in group-level data. The local Moran decomposes Moran's I into contributions for each location (each i), termed Local Indicators of Spatial Association (LISA). These indicators detect clusters of high and low values as well as anomalies, also called spatial outliers. The sum of I_i s for all observations is proportional to Moran's I . Anselin (1995) defined a local Moran statistic for an observation i :

$$I_{i,t} = \widehat{m}_{i,t} \sum_j w_{ij} \widehat{m}_{j,t} \quad (\text{Equation 6})$$

In equation 6, $\widehat{m}_{i,t}$ is the Z-score for the cancer mortality rate at location i at time, $\widehat{m}_{j,t}$ at location j at the same time. The Z-scores are calculated as in Equation 1. w_{ij} is a weight denoting the strength of connection between areas i and j , defined as for Moran's I .

2.7 Local G^*

The local G^* test (Getis and Ord 1992; Ord and Getis 1995) is also a LISA statistic. Like the local Moran, it detects clusters of extreme values (high and low). Unlike the local Moran, it is not designed to detect anomalies or outliers. Ord and Getis (1995) defined a local G^* statistic for an observation i :

$$G_{i,t}^* = \frac{\sum_j w_{ij,t}^* m_{j,t} - W_{i,t}^* \overline{m}_t}{s_{m,t} \sqrt{\frac{(n_t S_{1i,t}^* - W_{i,t}^{*2})}{n_t - 1}}} \quad (\text{Equation 7})$$

One difference between the I_i and the G^* is that the centering location is treated the same as the neighbors, its value enters the calculation when $j = i$. So, the weights sets for the two statistics are different: $w_{ij,t}^*$ in Equation 7 includes a weight between i and i . Another difference between I_i and G^* is that the G^* does not use Z-scores. Instead, the statistic itself is similar in form to Z-score, where a difference is taken between a local mean (the sum of the local values multiplied by the weights) and the predicted local mean (the global average multiplied by the sum of the weights). Then the difference is divided by the standard deviation, as in the Z-score, multiplied by a weighting factor. The weighting factor includes the sum of the weights squared

$$W_{i,t}^{*2} = \left(\sum_{j=1}^N w_{ij,t}^* \right)^2 \quad (\text{Equation 8})$$

and the squared sum of the weights

$$S_{1i,t}^* = \sum_j w_{ij,t}^{*2} \quad (\text{Equation 9})$$

In the context of interpreting the G^* by evaluating its Monte Carlo randomization p-value, the denominator of the G^* is unimportant, as it is constant for a location and all of its conditional randomizations.

2.8 Significance and Multiple Testing

We calculate p-values for I_i and G^* using 999 conditional Monte Carlo permutations of the data values. I_i and G^* statistics calculated for a given study area are not independent of one another, and hence their p-values (depending on one's philosophical inclinations) should be corrected for multiple testing. We now consider sources of this lack of independence, two approaches for accomplishing the adjustment, and a proposed solution.

Lack of independence arises in two places: Monte Carlo distributions under the null hypothesis, and in the test statistics. A conditional randomization approach is used when calculating the reference distribution of the test statistic within a Monte Carlo framework. Assume an area i for which we wish to evaluate the significance of the LISA statistic I_i , and that area i has neighbors j and k . Further assume N is the number of areas on the map. Conditional randomization repeatedly assigns new values to these 2 neighbors by drawing 2 new values at random from the $N-1$ areas surrounding location i , and calculates and records a new value of I_i to construct the reference distribution. The reference distributions for two different areas i and j thus are not independent since they will be constructed from repeated drawings from the same population. Lack of independence also arises in the test statistics for two areas that are neighbors of one another. Because they are neighbors, the I_i statistics for i , j and k will each use the values associated with one another when calculating I_i , I_j , and I_k . The test statistics therefore are correlated, and their p-values should be adjusted accordingly.

Two approaches may be used to accomplish this adjustment. We can correct the significance or alpha level of the test, or we can adjust the p-values themselves. If we correct the alpha level we then compare the unadjusted p-values to the corrected alpha level. If we adjust the p-value we compare the adjusted p-value to the unadjusted alpha level ($\alpha = 0.05$ in this case). We propose to use the second approach – adjusting the p-value and using the same alpha level for evaluating the $N I_i$ statistics on the map.

We use the Simes (1986) adjustment, which is not as conservative as the Bonferroni correction. The Simes adjustment is calculated as in Equation 10. Assume 3 p-values p_i , p_j , and p_k – suppose they are (0.002, 0.001, 0.036). Rank the p-values from lowest to highest, obtaining the vector (0.001, 0.002, 0.036). We wish to calculate the “Simes” p-value for $p_i = 0.002$, the second element in this vector. This is

$$p'_i = (n + 1 - a)p_i \quad (\text{Equation 10})$$

Here n is the number of p-values being considered (3), and a is the index (starting at 1) indicating the location in the sorted vector of p_i (2). The Simes p-value is then (2) $0.002 = 0.004$.

2.9 Classification

After the software calculates the G^* and I_i statistics for each location, it classifies all of the SEAs in the geography. For the I_i analysis, it classifies all of the SEAs in the geography as being the center of low-low clusters, high-high clusters, a significant high outlier (high-low), a significant low outlier (low-high), or nonsignificant. It compares the Simes p-values to a prespecified alpha level, in this case $\alpha = 0.05$, and then assigns the classes based on the sign of the I_i (positive indicates cluster, negative indicates outlier) and its Z-score (high or low). This treatment is parallel to the treatment of the local Moran in other software products, such as ClusterSeer, the SpaceStat extension for ArcView, and GeoDa. For the G^* , the software compares Simes p-values to a pre-specified alpha level, in this case $\alpha = 0.05$, and then assigns the classes based on the sign of the G^* (positive indicates high cluster, negative indicates low cluster).

We then scored each set of maps for similarity of classifications and for cluster persistence. For each race-gender subgroup, we examined cluster classifications resulting from the I_i and the G^* analyses at a particular time interval (for example, African-American female mortality rate 1990-1994). We considered an I_i high-high cluster equivalent to a high G^* cluster, similarly matching an I_i low-low cluster with a low G^* cluster, and a finding of nonsignificant in both analyses matched. All matching classifications were considered concordant. Non-concordant situations occurred when the outcomes differed from this matched pairing.

For cluster persistence, we compared sets of maps over time within one classification, for example white male G^* mortality rate classes in 1985-1989 and 1990-1994. Clusters identified in the last time period could not be scored for persistence as we had no information about clustering after 1994. As most of the locations were classed as non-significant and stayed that way, we did not count transitions from the non-significant class, just transitions from a cluster class to another class (such as outlier or non-significant).

3 Results

In this section, we describe the Atlas software and a comparison of the two clustering statistics' conclusions about the patterns in colon cancer mortality from 1970 through 1994 rates in 5-year time intervals for SEAs. The data for SEAs are efficiently represented as time slices in 5-year time intervals (1970-4, 1975-79, etc) with a static geography. Thus, what changes when data are animated in a map, graph, or table are only the attributes, rather than the shapes and positions of the geographic units. Figure 4 is a screenshot illustrating the software's time-enabled views. Notice that each view (maps and graphs as shown, but also scatterplots, boxplots, and tables) have time sliders and media buttons that allow the viewer to pan through time and animate the data. The software also has linked views, a feature common to spatio-temporal visualization software (Haslett *et al.* 1991, recently reviewed in Andrienko *et al.* 2003). In Figure 4, SEAs in Florida have been selected on the map. Since all data views are linked together, the selected items in the map are also selected (highlighted in grey) in both histograms. The bottom histogram displays the distribution of 1970-74 African-American male colon cancer mortalities. The top histogram and the map illustrate 1990-94 rates.

From early 2004 to July 2004, over 200 people downloaded the Cancer Atlas Viewer. BioMedware staff have also watched individuals use the software at our project planning meetings (described <http://www.biomedware.com/innovations/atlas.html>) and in exploratory spatial data analysis courses. The response from users has been positive: they appreciate the ease of visualizing the time series data as animated maps and the interactivity that brushing linked windows provides. Our user observation sessions fed back into software design, helping us to improve parts of the interface that were difficult to use and to uncover new, required features.

For all but a few SEAs, the count of deaths from colon cancer has increased since 1970 for all gender-race combinations. Similarly, the mortality rates from colon cancer increased for African-American females and males and white males. The differences in rates are largest for African-American males, who experienced the greatest increase in mortality rates from 1970 to 1990. White females, however, experienced decreasing rates in most of the study area. The differences are illustrated for all four groups in Figure 5, and the distribution of rates for African-American males is illustrated in the histograms in Figure 4. The rates in 1990-94 follow a similar pattern to the differences (Figure 6), with the rates for African-American males being highest and white females lowest, with African-American females and white males intermediate.

The global spatial pattern in these rates has been somewhat variable over time, but typically increasing. Figure 7 shows the significant Moran's I statistics for these datasets, while Table 2 shows the I statistics and p-values. White males show the strongest patterns, with significant global autocorrelation in four of five SEA time periods. The white male SEA pattern is variable, with the I statistic between 0.14 and 0.21. For the other groups, there are few SEA time intervals with significant spatial correlation with only one or two time periods showing any significant pattern (with a significant Moran's I). While like values are clustered near each other, these patterns are not consistently strong through time.

Moran's I is a global test; it does not detect localized pattern. Yet, because there are 4 race-gender combinations and five time intervals by two cluster tests (that is forty cluster maps to compare), we will not detail the results of any single cluster test at any particular time. Instead, we present the pattern of results across all race-gender subgroups over all time intervals. Table 3 compares the results from both tests. Overall the two local statistics were in concordance. Over 97% of the time, the two statistics agreed on the status of a location. Both agreed that there were twenty significant clusters of high values, sixteen significant clusters of low values, and 1,901 non-significant areas. There is no case where the Local Moran finds a cluster of high values and the G^* finds a cluster of low values, or the reverse. They appear to be drawing similar conclusions about these data.

Yet, there are some differences between the results. These differences could be caused by differences in the search pattern of each statistic (the geographic alternative hypothesis to which the statistic is sensitive) or because of the random nature of Monte Carlo probability assessment. We classed the forty-three nonconcordant results into four categories for convenience: no comment on outlier, outlier disagreement, significance disagreement, and marginal significance disagreement. These categories are summarized in Table 4.

No comment on outlier was a category we expected in the beginning—that the G^* may have “no comment” on locations identified as significant spatial outliers by the I_i . Since the G^* is not designed to detect outliers, and the Local Moran is, we expected outliers by the Moran analysis not to show up as clusters in the G^* analysis. This occurred thirteen times (where the Local Moran was Low-High or High-Low and the G^* was not significant). What was unexpected, however, was the five times that the local G^* called something a cluster and the local Moran called it an outlier. We will discuss two examples, Columbus, GA and Greenville, SC.

The Columbus, GA SEA is considered the center of a cluster of low mortality rates for white males in 1970-1974 by the G^* but a high outlier among low values by I_i . As shown in the left side of Figure 8, Columbus is surrounded by several low SEAs, with Z-scores between -4.6 to -0.69. Columbus is near the dataset mean, its Z-score is 0.09, and its southern neighbor is also close to the mean. In this case, the description of Columbus as a significant spatial outlier, specifically a higher outlier among low neighbors, does not correspond to the map pattern. Columbus is an average SEA with several low neighbors. Thus, the G^* is a better descriptor of the local area—the group of SEAs is lower than the regional average. The G^* , however, does not describe Columbus itself very well, it is not low, but it does connect the low group of SEAs. This cluster of low values continues to the northwest, as the Auburn, AL SEA is classed as a significant cluster of low values by the I_i and a marginally significant low cluster by the G^* (shown in Figure 10).

The Greenville, SC SEA is considered the center of a cluster of high mortality rates for African-American females in 1980-1984 by the G^* but a low outlier among high neighbors by the I_i . Yet, as shown in Figure 9, neither classification provides an entirely adequate description of the local spatial pattern. Greenville is average, neither especially high nor low, but it does have

some very high neighbors, specifically Easley, SC and Waynesville, NC. These high neighbors seem to be driving both classifications—the high neighbors result in a high local mean which is deemed a high cluster by G^* , and they cause Greenville to be declared a low outlier by I_i . As shown in Figure 9, it is not a cluster of high values but only two locations with high values, and there is nothing particularly extreme about Greenville itself. The other location that neighbors both Easley and Waynesville (Cornelia SEA in northeastern Georgia) is also the center of a significant cluster of high values, but this time both tests agree. As its rate is also high, this result makes more sense. So both tests found a strong signal in the vicinity of Greenville, but it is not correct to say that Greenville is significantly low or even surrounded by high neighbors or part of a cluster of high values. Neither classification provides a fully accurate description of the local pattern.

The twenty-five other cases of difference between the G^* and local Moran results occurred when only one of the two tests called a location the center of a significant cluster. In all cases, the statistics agreed about the pattern, both the I_i and the G^* showed clustering of high or low values for each location, but their results disagreed about the significance of the pattern. The G^* called 4 locations clusters that the local Moran called not significant, while the local Moran called 21 locations clusters that the G^* called not significant. Overall, the local Moran finds clusters more often than the G^* does. Whether either is more accurately reporting the “true” number of clusters in the region cannot be determined with this dataset, but we can examine those cases where the two tests differ to see what triggers each statistic.

Of the twenty-five disagreements about the significance of the clustering, twelve occur when there is a marginal difference in the p-values of the two statistics. For all items in this category, the p-value for both the I_i and the G^* were < 0.10 . For example, for white males in 1970, Auburn, AL was the center of a cluster of low values according to the local Moran ($I_i = 0.67$, $p = 0.049$); its G^* was marginally significant ($G^* = -2.72$, $p = 0.056$). Both statistics are in agreement about the pattern and its strength, but the Moran statistic happens to be just below the decision criterion ($\alpha = 0.05$) and the G^* above, so they provide different answers. The mean p-value for each statistic in this class was low (mean I_i p value = 0.042, mean G^* p value = 0.059, Table 4). Hence this lack of concordance reflects the arbitrariness of the $\alpha = 0.05$ decision threshold. Because of the random nature of the conditional Monte Carlo randomization used to assess significance for both statistics, it is entirely possible that the significance for Auburn, AL would be the same for both statistics (either a significant low cluster or nonsignificant) or the pattern of significance reversed (with the G^* being below the threshold and the I_i above) if the analysis was re-run. Also, we could have chosen a higher number of Monte Carlo randomizations (such as 9,999) to get a more precise p-value from the software. More p-value precision could alleviate these minor p-value disagreements.

The other thirteen cases of disagreement about the significance of a cluster are more interesting. In these cases, the difference in the p-value is large. For example for African-American females in 1970, Sumter, SC was the center of a significant low cluster according to I_i ($I_i = 0.291$, $p = 0.040$) but not close to significant by the G^* ($G^* = -1.68$, $p = 0.500$). Similarly, Prattville, AL was the center of a local Moran cluster of low mortality for African-American females in 1970 ($I_i = 0.05361$, $p = 0.024$) but not for the G^* ($G^* = -1.83$, $p = 0.500$). In these and other cases where the Moran p-value is much lower than the G^* p-value it is often the case that the I_i statistic is low (mean of this class is 0.364, Table 4) though significant. There is a significant but weak correlation between the values in the local neighborhood. Overall, the range of I_i values was from about -2 to 7. Positive I_i values less than 0.5, while significant, do not indicate strong clusters of extreme values, and correspond instead to clusters of values slightly lower or higher than the mean. The findings for G^* and I_i differ because G^* is considering divergence from the mean rather than correlation between the values. In many of these cases, G^* provides a more reasonable interpretation of the pattern in terms of what we are looking for in the study

of cancer mortality rates—researchers are understandably more concerned about clusters of extremely high or low mortality rates than clusters of rates within 1 standard deviation of the mean.

The differences in the results of the two cluster detection statistics stem from two factors. The marginal significance disagreement arises from the use of the 0.05 decision criterion. The outlier disagreement, no comment on outlier, and significance disagreement arise from differences in the search pattern of each statistic. In some cases, such as Columbus, GA, G^* provides a better description of the local pattern but not the ego location, while in others, such as Greenville, AL, neither explanation fits.

For cluster persistence, the results were clear: most of the significant clusters do not persist to the next time period, as detailed in Table 5. For sixty-two clusters or outliers identified by the I_i from 1970 through 1989, sixty were no longer significant in the next time period. For thirty-six clusters identified by the G^* , thirty-five were no longer significant in the next time period. Only one I_i cluster persisted into the next time period, a cluster of low mortality around Auburn, AL in 1970-74 and 1975-79. The area around Auburn in both time intervals is illustrated in Figure 10. The G^* found this cluster to be marginally significant in 1970-74 ($p = 0.056$). The I_i classification change occurred around Columbus, GA. Columbus is shown in Figure 8. In 1970-4, Columbus was classified as a high outlier among low neighbors by the I_i , and then in 1975-79 it was classified as the center of a low cluster. The change in the data that drives this change in class seems to be that the rate in Columbus went down in this interval. So, it was more similar to its low neighbors in the second time period. Columbus was a significant low cluster according to the G^* in both time periods, and is the only persistent G^* cluster in the region in the times studied.

4 Conclusion

The lack of persistent clustering in these data suggests that the clusters detected may be ephemeral. There is not persistent clustering of high values indicating a stable environmental exposure or a stable social or genetic contributing factor to colon cancer. Ephemeral clusters can be explained in several ways: they might result from unstable mortality rates generated by small populations-at-risk, population migration, or short term factors such as geographic differences in treatment or screening that do not persist. This seems to be a positive conclusion—there is no area at high for risk colon cancer that persists through time. There are some disparities in colon cancer mortality, as shown in Figure 6, with males having higher mortality rates than females, and African-Americans having higher mortality than whites. This does not seem to be a local phenomenon, but instead applies across the study geography. And, the rates for African-American males and females are increasing more steeply than they are for whites (Figure 5). Researchers have suggested that the increasing rates of colon cancer may be due to changing diets and increasing rates of obesity (Murphy et al. 2000). Other research has shown that patterns in colon cancer mortality in African-Americans may in part be due to socioeconomic factors, specifically differential access to health care. Freeman and Alshafie (2002) found that poorer individuals are diagnosed with more advanced cancers and die more frequently.

This comparison finds that the results of the two statistics are similar, with agreement on their classifications over 97% of the time. Because of the large amount of concordance between the two cluster statistics, there seems to be little additional value gained from applying both cluster statistics to a dataset. Because of the differences in reporting of significance, with the I_i reporting some clusters quite near the mean, the G^* may be more sensitive to clusters of extreme values. The cluster classifications produced by the statistics needed to be further examined to be interpreted well. Although several software products, including the Cancer Atlas Viewer,

produce crisp maps classifying locations into clusters, outliers, and non-significant areas, the simplicity of these maps can obscure the complexity of the observed data. The differences we found in the classifications of the G^* and the I_i pointed out a few locations where the interpretation for either classification required careful examination of the mapped data. These were instances in which the local map pattern did not correspond to the search image of either cluster statistic.

The few times they disagree on the status of an outlier seem to come from limitations on the shape of the cluster to be detected imposed by the first-order neighbor relationships considered. Disagreement emerged from a situation where there was variability among the neighbor set, so the “true cluster”, if it existed at all, was likely a subset of the neighbor set, rather than the whole group of first-order neighbors. This is shown in Figures 8-10, where significant clusters or sets of outliers contain individuals that are close to the mean and those that are more extreme. These locations would be better described by a different type of cluster statistic, one that connects areas into sinuous shapes that reflect similarity of values rather than searches for clusters matching a preexisting shape pattern (such as first-order neighbors). This argument is similar to one we have made before about the limitations of centroid-based cluster statistics that search for circular clusters (Jacquez and Greiling 2003).

This analysis was performed using the free Cancer Atlas Viewer software, providing researchers with sophisticated visualization and statistics for the exploration of patterns in mortality from 40 site-specific cancers. It can act as a quicker and more interactive way to explore the data made available by the National Cancer Institute, cutting out the delays inherent in web mapping that may hinder exploration. The statistical analysis presented here may be beyond the interest and commitment of a casual user, but the software provides a means for researchers to assess cancer mortality patterns, examine these patterns over time in animated maps and graphics, and to assess the persistence of clustering.

Acknowledgements

This project was funded by grant CA92669 from the National Cancer Institute to BioMedware, Inc. The positions espoused in this article are those of the authors and do not necessarily represent the official views of the National Cancer Institute. Constructive criticism from Heidi Durbeck of BioMedware, Peter Rogerson of SUNY-Buffalo, and three anonymous reviewers helped us improve the interpretation and presentation of these results.

References

- Andrienko N, Andrienko G, Gatalsky P. Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages and Computing* 2003;14:503–41.
- Anselin L. Local Indicators of Spatial Association — LISA. *Geographical Analysis* 1995;27:93–115.
- Anselin, L. The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association. In: Fischer, M.; Scholten, H.; Unwin, D., editors. *Spatial Analytical Perspectives on GIS*. Taylor and Francis; London: 1996. p. 111-125.
- Anselin, L. *Exploring Spatial Data with DynESDA2*. CSISS and Spatial Analysis Laboratory University of Illinois; Urbana-Champaign: 2002.
- Blot WJ, Morris LE, Stroube R, Tagnon I, Fraumeni JF Jr. Lung and laryngeal cancers in relation to shipyard employment in coastal Virginia. *Journal of the National Cancer Institute* 1980;65:571–575. [PubMed: 6931936]
- Clayton D, Kaldor J. Empirical Bayes Estimates of Age-standardized Relative Risks for Use in Disease Mapping. *Biometrics* 1987;43:671–681. [PubMed: 3663823]
- Cliff, AD.; Ord, JK. *Spatial processes: Models and Applications*. Pion; London: 1981.
- Cossman RE, Cossman JS, Jackson R, Cosby A. Mapping high or low mortality places across time in the United States: a research note on a health visualization and analysis project. *Health and Place* 2003;9:361–9. [PubMed: 14499220]

- Devesa, SS.; Grauman, DG.; Blot, WJ.; Pennello, G.; Hoover, RN.; Fraumeni, JF, Jr. Atlas of cancer mortality in the United States, 1950-94. US Govt Print Off; Washington, DC: 1999. NIH Publ No (NIH) 99-4564
- Freeman HP, Alshafie TA. Colorectal carcinoma in poor blacks. *Cancer* 2002;94:2327–2332. [PubMed: 12015756]
- Getis A, Ord JK. The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis* 1992;24:189–206.
- Goovaerts PE, Jacquez GM. Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. *International Journal of Health Geographics* 2004;3:14. [PubMed: 15272930]
- Haslett J, Bradley R, Craig P, Unwin A, Wills G. Dynamic Graphics for Exploring Spatial Data with Application to Locating Global and Local Anomalies. *The American Statistician* 1991;45:234–242.
- Hornsby K, Egenhofer M. Identity-based change: A foundation for spatio-temporal knowledge representation. *International Journal of Geographical Information Science* 2000;14(3):207–224.
- Jacquez, GM. GIS as an Enabling Technology. In: Gatrell, A.; Loytonen, M., editors. *GIS and Health*. Taylor and Francis; London: 1998. p. 17-28.
- Jacquez GM. Spatial Epidemiology: Nascent Science or a Failure of GIS? *Journal of Geographical Systems* 2000;2:91–7.
- Jacquez GM, Greiling DA. Local clustering in breast, lung and colorectal cancer in Long Island, New York. *International Journal of Health Geographics* 2003;2:3. [PubMed: 12633503]
- Jacquez GM, Greiling DA, Kaufmann AM. Design and implementation of space-time information systems. *Journal of Geographical Systems* X(this issue):xx–xx.
- Langran, G. *Time in Geographic Information Systems*. Taylor and Francis; London: 1992.
- Loytonen, M. GIS, Time Geography and Health. In: Gatrell, A.; Loytonen, M., editors. *GIS and Health*. Taylor and Francis; London: 1998. p. 97-110.
- Moore DA, Carpenter TE. Spatial analytical methods and Geographic Information Systems: Use in health research and epidemiology. *Epidemiologic Reviews* 1999;21(2):143–161. [PubMed: 10682254]
- Moran PAP. Notes on continuous stochastic phenomena. *Biometrika* 1950;37:17–23. [PubMed: 15420245]
- Mungiole M, Pickle LW, Simonson KH, White AA. Application of a weighted headbanging algorithm to mortality data maps. *Proceedings of the Statistical Graphics Section of the 1996 Annual Meeting of the American Statistical Association* 1997:45–49.
- Murphy TK, Calle EE, Rodriguez C, Kahn HS, Thun MJ. Body Mass Index and Colon Cancer Mortality in a Large Prospective Study. *Am J Epidemiol* 2000;152:847–854. [PubMed: 11085396]
- Ord JK, Getis A. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis* 1995;27:286–306.
- Peuquet DJ. It's about time: A conceptual framework for the representation of temporal dynamics in GIS. *Annals of the Association of American Geographers* 1994;84(3):441–461.
- Pickle, LW.; Mungiole, M.; Jones, GK.; White, AA. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. Hyattsville, MD: DHHS Publication No (PHS); 1996. *Atlas of United States Mortality*; p. 97-1015.
- Rushton G, Elmes G, McMaster R. Considerations for improving Geographic Information System research in public health. *Journal of the Urban and Regional Information Systems Association* 2000;12(2):31–49.
- Schaerstrom, A. *Pathogenic Paths? A time geographical approach in medical geography*. Lund University Press; Lund, Sweden: 1996.
- Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrics* 1986;73:751–4.
- U.S. Bureau of the Census. *Census of Population: 1960, Number of inhabitants, United States summary, final report PC (1)-1A*. Washington, DC: US Government Printing Office; 1966.
- Wegman EJ. Visual data mining. *Statistics in Medicine* 2003;22:1383–97. [PubMed: 12704604]

Winn DM, Blot WJ, Shy CM, Pickle LW, Toledo A, Fraumeni JF. Snuff dipping and oral cancer among women in the southern United States. *New England Journal of Medicine* 1981;304:745–749. [PubMed: 7193288]

Figure 1. Schematic of web-enabled GIS. Maps and images are transmitted over the web from the GIS to the client(s). The client views the map in an internet browser.

Figure 2.

Atlas Viewer design. A local database and Cancer Atlas Viewer software (both resident on the PC) provide rapid interactive, animated visualization. The user must go onto the Internet to download the software and the cancer data, but this only happens at the beginning of the process. Subsequent data visualization and exploration occur on the user's machine.

Figure 3.
Architecture of the Cancer Atlas Viewer Software.

Figure 4.

Time enabled visualization in the Cancer Atlas Viewer. The SEAs in Florida have been selected on the map. The histogram views of the data are also linked, and so Florida's contribution to the histograms is also highlighted in grey. The bottom histogram displays the distribution of 1970-4 African-American male colon cancer mortalities. The top histogram and the map illustrate 1990-4 rates. Each view of the data (maps, graphs, tables) has a time slider and media play buttons, so the data can be played through time.

Figure 5. Difference maps for colon cancer mortality rates between the time intervals 1990-4 and 1970-4. The classification is diverging, with white indicating no change from 1970-4, a grayscale gradient indicating increased mortality rates from colon cancer, and hatching indicating decreased rates.

Figure 6.
Rates for colon cancer mortality in the years 1990-4.

Figure 7.

This graph shows the significant Moran's *I* statistics for SEA units over time. Non-significant *I* statistics are omitted.

Figure 8.

Statistical disagreement about Columbus, GA. This map shows the location of the SEA (outlined in black) and its neighbors in Georgia and Alabama. The SEAs shown are colored by their Z-scores for the mortality rate for white males (RWM) in the period, with SEAs within half a standard deviation of the mean shown as white and below the mean shown hatched. The I_i and G^* statistics disagree about the classification in 1970-74, but both call Columbus the center of a cluster of low values in 1975-79.

Figure 9.

Statistical disagreement about Greenville, SC. This map shows the location of the Greenville SEA (outlined in black) and its five neighbors. The disjoint polygon on the North Carolina border to the east of the group is actually part of a polygon set that does border the Greenville SEA. The SEAs shown are colored by their Z-scores for the mortality rate for African-American females (RBF) in the period 1980-1984, with darker grey indicating a higher mortality rate in the period.

Figure 10.

Persistent clustering of low white male mortality rates centered on Auburn, AL. The Auburn SEA is outlined in black, and its neighbors are shown with grey borders. The map shading is from the Z-score of the rate, with negative Z-scores shown as hatched. Auburn has three neighbors made up of more than one polygon; this is why some polygons are shown as neighbors but do not share a border with Auburn.

Table 1

A listing of a few recent online atlas projects.

Initiative	Description
Washington State's Epidemiologic Querying and Mapping System (EpiQMS) https://epiqms.doh.wa.gov/	Death certificate data (cause of death) by county along with population information. Users can map the data or prepare graphs, or view tables.
New York State's Cancer Surveillance Improvement Initiative (NY CSII) http://www.health.state.ny.us/nysdoh/cancer/csii/nyscsii.htm	Information on breast, colorectal, lung and prostate cancer diagnoses by ZIP code in New York State. Users can view prepared pdf maps, or view the data for individual ZIP codes by county.
Reproductive Health Atlas http://www.cdc.gov/reproductivehealth/GISAtlas	Information on variables such as infant mortality, pregnancy outcomes, infant health, and maternal risks by demographic groups and different geographic aggregation. No data is currently available, but the website talks about distributing starter shapefiles for policymakers and service providers.
Cancer Mortality Maps and Graphs (NCI) http://www3.cancer.gov/atlasplus/	Data on mortality for 40 site-specific cancers by county, state economic area, and states from 1950-94. The data can be mapped online or downloaded for viewing and manipulation.

Moran's *I* statistics and p-values calculated from 999 Monte Carlo randomizations. Values in bold are significantly below $\alpha = 0.01$ for SEAs.

Table 2

Group	1970-4	1975-9	1980-4	1985-9	1990-4
African-American Females	0.116988 (0.021)	-0.06699 (0.142)	0.222788 (0.001)	0.04774 (0.084)	0.064046 (0.021)
African-American Males	0.001491 (0.428)	0.100771 (0.019)	0.175791 (0.001)	0.034729 (0.196)	0.1113591 (0.010)
White Females	0.175 (0.001)	0.014104 (0.333)	0.006583 (0.370)	0.011811 (0.312)	0.174887 (0.001)
White Males	0.201898 (0.001)	0.135539 (0.008)	0.181063 (0.001)	0.07956 (0.066)	0.151863 (0.004)

Table 3

Classification of 1,980 locations (99 areas over 5 time intervals by 4 population subgroups) by the local Moran and local G^* clustering tests. The cells that hold the totals for agreement between the two clustering methods are shaded grey. Other cells show differences between the clustering results.

Local Moran results	Local G^* results			Total not concordant
	High	Low	NS	
High-high	20	0	6	6
Low-low	0	16	15	15
NS	1	3	1901	4
Low-high	2	0	4	6
High-low	0	3	9	12
Total not concordant	3	6	34	43

Table 4
Average statistic values and standard deviations in the different non-concordant classes from table 3.

Class	n	Mean I_i	I_i std dev	Mean I_i p value	Mean $ G^* $	$ G^* $ std dev	Mean G^* p value
No comment on outlier	13	-1.171	0.965	0.018	1.577	0.457	0.405
Outlier disagreement	5	-0.354	0.134	0.020	3.198	0.437	0.028
Marginal disagreement	12	1.104	0.654	0.042	2.683	0.518	0.059
Significance disagreement	13	0.364	0.284	0.038	1.895	0.222	0.284

Table 5

Cluster persistence results from comparing cluster classifications at two times. These totals are from all race-gender subgroups over all pairs of sequential times.

Change Class	I_i	G^*
Cluster -> not significant	60	35
outlier -> cluster	1	n/a
cluster -> cluster (no change)	1	1
Total	62	36