Data Aggregation and Analysis for Cancer Statistics - A Visual Analytics Approach

Ross Maciejewski∗ Travis Drake∗ Stephen Rudolph∗ Abish Malik∗ David S. Ebert∗

∗Purdue University Regional Visualization and Analytics Center (PURVAC)

ABSTRACT

The disparity between data collected in rural and urban counties is often detrimental in the appropriate analysis of cancer care statistics. Low counts drastically affect the incidence and mortality rates of the data, leading to skewed statistics. In order to more accurately report the data, various levels of aggregation have been used (grouping counties by population, age percentages, etc.); however, such data aggregation methods have often been ad hoc and/or time consuming. Such groupings are performed on a user defined basis; however, grouping based purely on population demographics does not take into account the spatial relationships between data. Furthermore, researchers want to search for spatiotemporal correlations within their data domain. In this work, we introduce a visual analytics system for exploring cancer care statistics in a series of linked views and interactive user interface queries. We also apply the AMOEBA algorithm [1] for clustering counties based on population demographics in a visual analytics environment. Users select the population demographics field on which they wish to cluster, and these county clusters then form the basis for the data aggregation. Such a system allows the user to group their data by fields (age, gender, income) while maintaining spatial structure and provides interactive mapping system in which to compare and explore such groupings. By utilizing such geographical groupings, we hope to better enhance the underlying structure of the data and help alleviate reporting problems associated with small area statistics.

Keywords: Spatial aggregation, cancer statistics, AMOEBA, visual analytics.

1 MOTIVATION

Relative grouping of cancer statistics for analysis and summary reporting is an important task for public health officials. Unfortunately, summary statistics of cancer data are either provided only by geographic unit (county, state, etc.), or by population demographic unit (age, ethnicity, etc.). In the case of summarizing cancer statistics on a county by county basis, the disparity between data collected in rural and urban counties is often detrimental in the appropriate analysis of cancer care statistics. Low counts drastically affect the incidence and mortality rates of the data, leading to skewed statistics. This problem is often referred to as the small area problem [5] or the small number problem [10]. One common method of handling this is to simply summarize the cancer data by population demographics within a state, ignoring the spatial data components.

In this work, we present a system that allows analysts to create demographic clusters of their data while maintaining the spatial data constraints. Our scheme increases the stability of the reported cancer rates by aggregating areas with similar demographics through interactive spatial clustering. Interactive selection of demographic groupings allows analysts to ask questions of their data and see reports displayed on an interactive map. Users may scroll through time and use a novel dual time slider control to compare changes in rates with a variable lag time. Spatial clustering is done through the AMOEBA algorithm developed by Aldstadt and Getis [1].

In this approach, we hope to reduce the issues associated with small areas by enlarging the area base over which the summary statistic may be calculated. Further, the clustering algorithm threshold can be user defined in order to enlarge clusters where the data aggregation would still fail to be sufficient. This work does not propose to seek out statistically significant cancer clusters within a population. Instead, our goal is to cluster spatial data based on relationships between underlying factors (age, demographics, income, etc.). Once a set of counties are grouped by this underlying factor, age adjusted cancer statistics may be calculated for this new region that is statistically homogenous with respect to a given population factor. Then, the counties can be clustered, and inferences can be made. In such a way, we are able to reduce the uncertainty found within small area statistics by increasing the area size (when possible).

Furthermore, our system introduces a novel dual time slider interaction technique in which users may explore various lag times and temporal aggregations amongst data resources. In this way, we encourage hypothesis generation. Linked views of the temporal domain can provide a static reference for exploration while interactive comparative maps can animate changes over time during exploration.

2 RELATED WORK

In recent years, much research has been done on determining effective means of exploring and disseminating health and cancer statistics. Previous work in the visual exploration and analysis of cancer statistics has looked at ways to visualize the uncertainty in data due to small area statistics. MacEachren et al. [13] utilized adjacent maps to show data reliability, and compared this to maps utilizing textures (hashing) overlaid on color to display reliability. MacEachren et al. [12] also presented a system designed to facilitate the exploration of time series, multivariate, geo-referenced health statistics. Their system employed linked brushing and time series animation to help domain experts locate spatiotemporal patterns. Further work in analyzing health statistics was done by Edsall et al. [6]. Here, the use of interactive parallel coordinate plots was used to explore mortality data as it relates to socio-economic factors. Other work includes Dang et al. [4] and Zhao et al. [20] which utilized dynamic queries and brushing for creating choropleth map views, and Schulze-Wollgast et al. [17] developed a system for visualizing health data for the German state Mecklenburg-Vorpommern. This system allowed users to interactively select diseases and their parameters and view the data over a specific time interval at different temporal resolutions. Further work in this system [18] employed the use of intuitive 3D pencil and helix icons for visualizing multiple dependent data attributes and emphasizing the type of underlying temporal dependency.

All of these systems fall into the realm of visual analytics. Visual
analytics is the science of analytic reasoning facilitated by interactive visual interfaces [16]. It is primarily concerned with presenting large amounts of information in a comprehensive and interactive manner. By doing so, it is hoped that the end user will be able to quickly assess important data and, if required, investigate points of interest in detail. The branch of visual analytics with which this work is most concerned is that of geospatial and temporal analytics, which applies the concepts of visual analytics to problems rooted in space and time.

Recently, the development of visual analytics systems for data analysis and exploration has been rapidly growing (e.g., [2, 7, 15, 19]). These systems incorporate a variety of visualization techniques from traditional, widely used methods, such as scatterplots or parallel coordinate plots to more recently developed tools (e.g., spiral graphs [3], theme river [8]). Techniques common across these systems include the probing, brushing and linking of data in order to help analysts refine their hypotheses, and these systems emphasize the interaction between human cognition and computation through dynamically linked statistical graphs and geographical representations of the data.

To this end, the goal of the system presented in this paper is to provide the user with a system capable of performing advanced spatial analysis methods in an interactive environment, thereby enabling hypothesis generation and data exploration. An important aspect of analyzing these cancer datasets is comparative and correlational analysis of different data sources and different time intervals. If ones wants to find correlation among incidence rates and mortality rates, what time intervals of each dataset should they compare? Should an aggregate of 5 years of incidence data be compared to one or two years of mortality data? What is the best alignment of these time windows? These are important questions to cancer care researchers, and our system provides researchers with an interactive tool that allows flexible comparison and aggregation of various time scales and factors to enable the generation of new correlative hypotheses and enable them to be tested and validated.

3 Visual Analytics Environment

Figure 1 provides a screenshot of the system. Cancer data is collected from the Indiana State Cancer Registry [14], details of which are summarized in Section 3.1. Our system provides a simple GUI in which users can query cancer data collected in Indiana from 1997 - 2004. Users can choose to cluster data by the 2000 census statistics data. Various visualizations are shown across four maps, allowing users to compare various cluster groupings and population statistics. Users can see the unclustered cancer statistics (incidence and mortality rate), the census statistics, the county groupings based on AMOEBA clustering, and then the age-adjusted combined cancer statistics for the resultant clusters. Linked views provide historical context to view the changes of rates over time.

3.1 Cancer Statistics

Cancer data is collected by county agencies and reported to the Indiana State Department of Health. In order to ensure confidentiality and stability of rates, counts by gender are suppressed if fewer than 16 cases were reported in a specific area-sex-race category. Incidence and mortality rates are calculated based on a per 100,000 population per year metric and are age-adjusted to the 2000 US standard population by 5-year age groups. When interpreting the data provided, it is important to understand that observed groupings may be subject to ecological fallacy (correlation does not imply causation). Furthermore, data collection is such that when the population size for a denominator is small, the rates may be unstable. A rate is unstable when a small change in the numerator (e.g., only one or two additional cases) has a dramatic effect on the calculated rate. Suppression is used to avoid misinterpretation when rates are unstable, and visualization methods for expressing such instances in geographical visualization systems have been explored by MacEachren et al. [13].

In this work, we utilize the available rates as data for a proof of concept demonstration. All collected data is available, and falls under various privacy concerns due to small area statistics. Our system provides researchers with a potential means of reporting statistics over a larger areal aggregation, allowing for less suppression and higher rate stability. Further, our system could be modified to include a privacy threshold such that if data were available to the backend of the system for calculations, data reports could be restricted by the system if the questions being asked by the user would violate established privacy rules (i.e., fewer than 16 cases reported in a given area).

3.2 AMOEBA Clustering

In order to group data based on spatially similar population statistics, we utilize the AMOEBA algorithm for creating spatial weights matrix developed by Aldstadt and Getis [1]. AMOEBA (A Multi-directional Optimum Ecotope-Based Algorithm) procedure is designed to identify hot and cold spots in mapped data by assessing the spatial association of a particular mapped unit to its surrounding units. It is able to aid in the demarcation of clusters of related spatial units, and we utilize this fact to group counties based on population statistics.

AMOEBA maps clusters of high and low values by creating a spatial weights matrix based on the Getis-Ord statistic. For a given location \(i\), \(G^*_i\) is defined as

\[
G^*_i = \frac{\sum_{j=1}^{N} w_{ij}(x_j - \bar{x})(\sum_{j=1}^{N} w_{ij})}{\sqrt{\sum_{j=1}^{N} w_{ij}^2(N-1)}}
\]

(1)

Here, \(N\) is the number of spatial units, \(x_j\) is the value of interest within the areal unit at location \(j\), \(\bar{x}\) is the mean of all values, and

\[
S = \sqrt{\frac{\sum_{j=1}^{N} w_{ij}^2}{N} - \left(\frac{\bar{x}}{N}\right)^2}
\]

(2)

\(w_{ij}\) is used as an indicator function that is one if \(j\) is a neighbor of \(i\) and zero otherwise.

The AMOEBA algorithm develops a cluster from a selected seed location by evaluating \(G^*_i\) at all locations surrounding this seed location, and if the addition of a neighbor to the cluster increases the
Figure 2: Aggregating data on population groups under 18. (Left) Choropleth map of the percent of population under 18. (Right) AMOEBA clustering of population based on percent under 18.

$G_i^*$ value, then the neighbor is added. Details of this algorithm and the use of it in other visualization applications can be found in [1] and [9].

Figure 2 illustrates the application of AMOEBA clustering in our system. In Figure 2 (Left) we have the choropleth map of the percentage of the population under 18 in Indiana counties. In Figure 2 (Right) we show the results of an AMOEBA clustering. Groups are colored based on their $G_i^*$ values, and counties that connect to other counties of the same color are considered to be a cluster. These clusters make up the basis of our summary statistic groups for the cancer analysis. Furthermore, users may also utilize this clustering mechanism to group counties specifically by cancer types in order to see which counties have similar cancer statistics in a general sense.

Figure 3: Aggregating data on population groups over 65 averaging the colon cancer incidence rates. Here we see the selection of a county reveals the underlying cluster structure and provides summary statistics about the cluster.

Figure 3 shows an aggregate grouping of the percentage of population over 65 colored by their average colon cancer incidence rates. The user can select a county in the state to reveal the underlying cluster structure. The resulting summary statistics are then displayed for the cluster to provide advanced analysis and exploration. By doing this, the user can begin to see patterns emerging in the data that may be otherwise obscured. However, it is important not to fall prey to ecological fallacies in such an analysis.

3.3 Temporal Comparison Controls

Along with our clustering tools, we also provide users with a unique temporal control system in the form of a dual slider system for comparative visualization. Users may interactively scroll through time in the normal mode, using the leftmost slider (Figure 1). Time can also be aggregated by year.

In comparative mode, the user interactively chooses the time period of interest using the leftmost slider and compares it to the period selected by the rightmost slider. In this case, the choropleth maps displayed show the difference in rates between the two time periods.

Users may also select counties in this mode and display the cancer statistics for each choropleth map on a linked statistical view. Here, raw rates are shown as a time series plot for each county. This allows users to view the stability of the signal over time and jointly investigate other aspects of their data.

4 Examples in Aggregation and Exploration

Our cancer analytics system provides researchers with a mean to interactively compare cancer rates for various cancer types. These comparisons can be used to influence policy decisions and guide health care spending (for example, targeting areas with high lung cancer rates with more ads, or increasing free screenings in areas with high colon cancer mortality rates. Figure 4 illustrates a way in which researchers can search for large changes in mortality rates. In Figure 4, the upper left choropleth map shows the change in mortality rate in 2004 compared to 1999 for the aggregation of all cancer sites. Here, the analyst may begin exploring the various red groups (indicating that the rate is higher in 2004 than it was in 1999). For example, a group of four counties with higher mortality rates is found in the central east portion of the state. The analyst can select the counties, bringing up the temporal history for each choropleth map in the corresponding linked window to the right. Here, the topmost window is linked to the map in the upper left, and the following time series windows are linked in order of the maps going from left to right, top to bottom. In each window, the analyst can select the variable by which to color the map. Here, the analyst has
chosen to separate the cancers by site in each of the other three windows in order to determine which types of cancers are most likely to be influencing the high rate changes. After interactively searching through the cancer types, the analyst finds that three of the four counties rates have increased due to an increase in lung cancer, and the fourth county is due to colon cancer. In this case, policy makers could use this information to warrant more spending on an anti-smoking campaign, or increase free prostate examinations.

In Figure 5, the analyst has chosen to look at colon cancer rates. Further, the analyst has chosen to cluster the counties by those with similar populations under the age of 18. In Figure 5(Left), we see the choropleth map of colon cancer rates by county for 1999. In Figure 5(Right), a new picture emerges where the data has been smoothed due to the grouping, and new patterns emerge. The analyst can then begin analyzing these new patterns and forming hypotheses by pulling up other data attribute maps in other linked windows of the system.

5 CONCLUSIONS AND FUTURE WORK

Our current work demonstrates the benefits of visual analytics for helping analysts ask questions of their data and begin generating hypotheses. Our initial results show the benefits of linking time-series views with geo-spatiotemporal views for enhanced exploration and data analysis through the use of traditional choropleth map visualizations and advanced temporal controls allow for finer tuning of map comparisons.

Future work includes the introduction of SatScan [11] and the use of AMOEBA as means to search for statistically significant cancer clusters under user controlled guidance. We hope this preliminary work demonstrates the potential power a visual analytics system could have for cancer care engineering, and we hope that this will prompt researchers to release more detailed data to be used in this setting under a refined set of privacy controls.

REFERENCES

Figure 5: Aggregating data on percentage of population under 18. (Left) Choropleth map of the incidence rates of colon cancer. (Right) Incidence rates of colon cancer aggregated by population groups under 18.


