A Visual Analytics Approach to Understanding Spatiotemporal Hotspots

Ross Maciejewski, *Student Member*, *IEEE*, Stephen Rudolph, Ryan Hafen, Ahmad M. Abusalah, Mohamed Yakout, Mourad Ouzzani, William S. Cleveland, Shaun J. Grannis, and David S. Ebert, *Fellow*, *IEEE*

Abstract—As data sources become larger and more complex, the ability to effectively explore and analyze patterns among varying sources becomes a critical bottleneck in analytic reasoning. Incoming data contain multiple variables, high signal-to-noise ratio, and a degree of uncertainty, all of which hinder exploration, hypothesis generation/exploration, and decision making. To facilitate the exploration of such data, advanced tool sets are needed that allow the user to interact with their data in a visual environment that provides direct analytic capability for finding data aberrations or *hotspots*. In this paper, we present a suite of tools designed to facilitate the exploration of spatiotemporal data sets. Our system allows users to search for hotspots in both space and time, combining linked views and interactive filtering to provide users with contextual information about their data and allow the user to develop and explore their hypotheses. Statistical data models and alert detection algorithms are provided to help draw user attention to critical areas. Demographic filtering can then be further applied as hypotheses generated become fine tuned. This paper demonstrates the use of such tools on multiple geospatiotemporal data sets.

Index Terms—Geovisualization, kernel density estimation, syndromic surveillance, hypothesis exploration.

1 Introduction

Health reports, terrorism alerts, criminal activities, and numerous other incidents need to be analyzed and evaluated, often within the context of related data sets. It is no longer efficient for a single analyst to pull files, take notes, form hypotheses, and request data from different sources. Instead, tools need to be developed that bring varying data sources into a unified framework assisting analysis and exploration. These needs are being addressed by the emergence of a new scientific field, visual analytics.

Visual analytics is the science of analytical reasoning assisted by interactive visual interfaces [40]. Major challenges in this field include the representation and linkage of large-scale multivariate data sets. In order to facilitate enhanced data exploration and improve signal detection, we have developed a linked geospatiotemporal visual analytics tool designed for advanced data exploration. This paper presents a set of extensions to our previous suite of visual analytics tools [31] for the enhanced exploration of multivariate geospatiotemporal data. Our system features include

- a new kernel density estimation that works for both urban and rural populations;
- dually linked interactive displays for multidomain/multivariate exploration and analysis;
- novel data aggregation for effective visualization and privacy preservation;
- control charts for identifying temporal signal alerts;
- demographic filter controls that enable database querying and analysis through a simple graphical interface;
- spatiotemporal history via contour line ghosting;
- bivariate exploration combining contours and color;
- multivariate exploration combining height maps, contours, and color;
- thresholding data to analyze specific trends;
- interactive color mapping tools for enhanced data contextualization;
- region selection tools for analyzing area specific hotspots.

Our work focuses on advanced interactive visualization and analysis methods providing linked environments of geospatial data and time series graphs. Hotspots found in one display method can be selected and immediately analyzed in the corresponding linked view. Furthermore, our work focuses on the early detection and analysis of hotspots facilitated through the use of control charts for

- *R. Maciejewski and D.S. Ebert are with the Purdue University Regional Visualization and Analytics Center, Purdue University, Potter Engineering Center, Room 134, 500 Central Drive, West Lafayette, IN 47906. E-mail: {rmacieje, ebertd}@purdue.edu.*
- *S. Rudolph is with the Purdue University Regional Visualization and Analytics Center, 8550 E McDowell Rd., Apt. 279, Scottsdale, AZ 85257-3902. E-mail: stephen.rudolph@gmail.com.*
- *R. Hafen and W.S. Cleveland are with the Department of Statistics, Purdue University Regional Visualization and Analytics Center, Purdue University, 150 N. University Street, West Lafayette, IN 47906. E-mail: rhafen@purdue.edu.*
- *A.M. Abusalah is with Purdue University, 1160 Cushing Cir, Apt 315, Saint Paul, MN 55108. E-mail: aabusala@purdue.edu.*
- *M. Yakout is with the Department of Computer Sciences, Indiana Center for Database Systems (ICDS), Purdue University, 305 N. University Street, West Lafayette, IN 47907-2107. E-mail: myakout@cs.purdue.edu.*
- *M. Ouzzani is with the Cyber Center, Purdue University, 155 S. Grant St., West Lafayette, IN 47907. E-mail: mourad@cs.purdue.edu.*
- *S.J. Grannis is with the Regenstrief Institute, Inc., Indiana University School of Medicine, 410 W. 10th St., Suite 2000, Indianapolis, IN 46202. E-mail: sgrannis@regenstrief.org.*

Manuscript received 30 Jan. 2009; revised 8 May 2009; accepted 16 July 2009; published online 19 Aug. 2009.
Recommended for acceptance by T. Ertl.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCGSI-2009-01-0020.
Digital Object Identifier no. 10.1109/TVCG.2009.100.

alert detection. Alerts generated in the temporal realm can be quickly analyzed in the geospatiotemporal interface, helping users find patterns simultaneously in the spatial and temporal domains. Concurrently, we have also applied statistical modeling techniques to estimate event distributions in the spatial realm. Users may select hotspots from the generated heatmaps and analyze historical time series data in the area to look for unusual trends or potential areas of interest. Such doubly linked views allow users to quickly form and test hypotheses, thereby reducing the time needed to reject false positives and confirm true alerts.

We have also extended our previous system [31] to include the spatiotemporal history, bivariate and multivariate exploration, thresholding, and color mapping tools. Contour histories provide users with geospatiotemporal views of current and past data trends, allowing them to track hotspot movement across time, or look for correlations between multiple variables. We also allow interactive range selection and thresholding to allow users to focus directly on hypothesis specific information. Furthermore, we demonstrate the flexibility of such tools by providing example applications in the domain of law enforcement data analysis and syndromic surveillance.

1.1 Law Enforcement Data

One data source we focus on is law enforcement data. These data come in the form of traffic violations, misdemeanors, criminal activities, etc. These data are typically spatiotemporal, containing the location of the incident, the time, and some description allowing the data to be classified into various categories. Such data can be analyzed for trends, enabling agencies to better manage their resources and deploy officers to potential problem areas. Our work utilizes data from the West Lafayette Police Department, and through contacts with former State Highway Patrol officers, we are able to tailor our tools toward officer specific needs.

1.2 Syndromic Surveillance Data

Another data source that we explore is syndromic surveillance data. Recently, the detection of adverse health events has focused on prediagnosis information to improve response time. This type of detection is more largely termed *syndromic surveillance* and involves the collection and analysis of statistical health trend data, most notably symptoms reported by individuals seeking care in emergency departments. Currently, the Indiana State Department of Health (ISDH) employs a state syndromic surveillance system called Public Health Emergency Surveillance System (PHESS) [19], which receives electronically transmitted patient data (in the form of emergency department *chief complaints*) from 73 hospitals around the state at an average rate of 7,000 records per day.

These complaints are then classified into nine categories (respiratory, gastrointestinal, hemorrhagic, rash, fever, neurological, botulinic, shock/coma, and other) [11] and used as indicators to detect public health emergencies before such an event is confirmed by diagnoses or overt activity. Unfortunately, detection of events from these indicators is an extremely challenging issue. Fig. 1 shows a typical month of emergency department visits for those

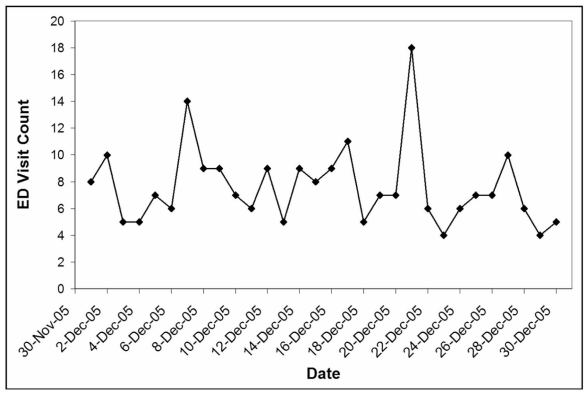

Fig. 1. A sample syndromic surveillance signal containing a carbon monoxide poisoning event.

complaints classified as neurological syndromes. During this time period, there was one event of carbon monoxide poisoning which happened to coincide with the largest peak on December 21; however, this peak is not significantly higher than any other peak during this month. Obviously, the detection of such a small signal deviation can be extremely difficult.

Again, it is important to note that the tools being developed are supervised by our partners in the Indiana State Department of Health. Prior to tool development, we meet with our partners and discuss the needs and functionalities of the tools and work to build them in such a manner as to enhance their work flow. We have found that tools developed for syndromic surveillance have translated well to the analysis of law enforcement data, and feedback from both agencies has been valuable in creating appropriate tools.

2 Related Work

As previously stated, visual analytics is the science of analytic reasoning facilitated by interactive visual interfaces [40]. In order for these interfaces to be effective, they need to integrate not only data exploration and visualization tools, but also human factors such as interaction, cognition, perception, collaboration, presentation, and dissemination. In order to create an effective visual analytics systems, methods from a variety of backgrounds must be merged together in a simple, yet effective framework. This section covers relevant topics in the areas of crime analysis, syndromic surveillance, multivariate interaction techniques, time series visualization, and geographical visualization.

2.1 Crime Analysis

In order to improve public safety and prevent crimes, law enforcement agencies need to analyze the volumes of data from multiple systems, search for trends, and deploy services appropriately. As such, many packages exist for studying spatial relationships between crime and area demographics. Work by Messner and Anselin [34] uses exploratory spatial data analysis to visualize spatial distributions and suggest clusters and hotspots. Specifically, they look at spatial autocorrelation and box maps. Other work includes WebCAT by Calhoun et al. [10] which

focuses on enhanced data sharing and crime data analysis tools via the web. Their tools include chloropleth mapping and capabilities to export records to excel. Our work presents similar capabilities to both Messner and Anselin [34] and Calhoun et al. [10]; however, we also include dynamically linked views and advanced hotspot detection tools not found in either of these works.

2.2 Syndromic Surveillance Systems

Data from public health surveillance systems have long been recognized as providing meaningful measures for disease risks in populations [25], [39]. As such, many disease modeling packages, outbreak alert algorithms and data exploration systems have been developed to aid epidemiologists in identifying outbreaks within their data. Some of the most popular of these systems are the Early Aberration Reporting System (EARS) [22], the Electronic Surveillance System for the Early Notification of Community-based Epidemics ESSENCE [27], and Biosense [28]. Unfortunately, all of these systems offer limited data exploration tools and little-to-no interactive geospatial support. Furthermore, many detection algorithms employed by these systems generate a large amount of false positives for epidemiologists to analyze. While creating algorithms to reduce false positives is important, our work focuses on creating advanced visual analytics tools for more efficiently exploring these alerts and hypotheses.

2.3 Multivariate Interaction Techniques

When creating an interactive framework for data exploration and hypothesis testing/generation there are a variety of interactive techniques that can be applied. The majority of techniques utilized in our work focus on the probing, brushing, and linking of data in order to help analysts refine their hypotheses. These methods emphasize the interaction between human cognition and computation through dynamically linked statistical graphs and geographical representations of the data (e.g., [13], [7], [4]).

Examples of recent work in spatiotemporal interaction include VIST-STAMP by Liao [26], FemaRepViz by Pan and Mitra [35], and LAHVA by Maciejewski et al. [32]. VIST-STAMP supports the overview of complex patterns through a variety of user interactions. Specifically, this work focuses on visualizing multivariate patterns using parallel coordinate plots and self-organizing maps. FemaRepViz provides a display of Federal Emergency Management Agency (FEMA) reports on a globe and dynamically determines where each report should be placed based on the text of the report. It also allows the user to navigate through time; displaying only the relevant reports for that period. And finally, LAHVA looked at using multiple data sets (pet and human health data) with similar properties to enhance disease surveillance. This system provided a geospatiotemporal interface with limited interaction among different view windows. Our current system is similar in that it allows users to explore both spatially (through panning and zooming) and temporally through interactive time sliders and history filter/aggregation controls.

Examples of recent work in multivariate data exploration through linked views and probing includes work by Weaver [44], who created a system for interactively expressing sequences of multidimensional set queries by cross-filtering data values across pairs of views. Another example is work by Stasko et al. [38] which introduced the Jigsaw system which provides a series of visual interfaces that deal with identifying linkages between entities within a data set. Other work includes the use of data probes by Butkiewicz et al. [9]. This work noted that when analysts are zoomed out of their data, local trends are suppressed, and when zoomed in, spatial awareness and comparison between regions is limited. Our current system uses similar modalities in that users can selectively filter data through query command and through interaction between the linked interfaces. Furthermore, when zooming into the data, we allow users to manipulate rendering parameters (such as color) in order to help better contextualize and explore hotspots in their local surroundings.

2.4 Time Series Visualization

One of our key linked views is a time series visualization view. The analysis of time series data is one of the most common problems in any data domain, and the most common techniques of visualizing time series data (sequence charts, point charts, bar charts, line graphs, and circle graphs) have existed for hundreds of years. Recent work in time series visualization has produced a variety of techniques, an overview of which can be found in [2]. Of the more modern techniques, some of the most commonly applied are the theme river [21], the spiral graph [45], and the time wheel [41].

Work on event prediction [8] and pattern recognition [8] was done by Buono et al. for time series data. This work presented users with a tool to explore multivariate time series for common patterns, and extended this approach for predicting future events. Other techniques of interest include the visualization of queries on databases of temporal histories [12], and novel glyphs for representing temporal uncertainties [3]. Unfortunately, most temporally oriented visualization techniques are not suited to represent branching time, or time with multiple perspectives. As such, the modification of existing techniques is necessary in order to more adequately analyze multivariate data from varying sources.

Our work focuses primarily on line graphs showing event counts. These graphs are then statistically analyzed and plotted as control charts in order to quickly provide the analysts with contextual information about the significance on an event. We allow for the plotting of multiple series on a single graph, as well as interactive selection tools for area/region specific plots.

2.5 Geographical Visualization

Another key linked view is the geographical visualization component. Geographic visualization is a field focused on displaying data with a geographic context such as a map. In more recent years, it has ballooned to include increasingly complex data, other spatial contexts, and information with a temporal component.

Several current systems exist that leverage advanced geographical visualization techniques for various health data. MacEachren et al. [30] presented a system designed to facilitate the exploration of time series, multivariate, and georeferenced health statistics. Their system employed

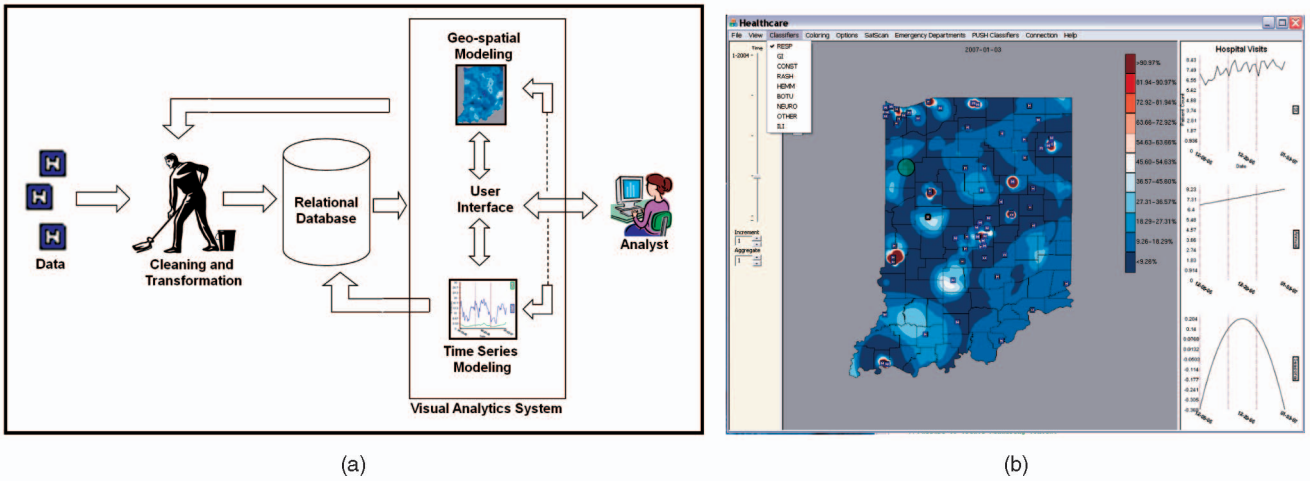

Fig. 2. The visual analytics system. (a) The conceptual diagram of our visual analytics system. Observe the interaction between the analyst and the system as well as the modeling components of the system. (b) Our visual analytics system. The left portion of the screen represents the interactive temporal tools. We include time aggregation tools, pull down menus for data classifier selections, keyword searches for chief complaint text, and demographic filtering for age and gender. The main viewing area is a geospatiotemporal view that has pan and zoom controls in the upper left corner. Hospitals and regions of the map may be selected with a circular query tool for interactive time series generation. The rightmost windows are the temporal views, showing selected time series plots broken down into their relevant components. Users may select points or regions of time to interactively manipulate the geospatial temporal window. For analyzing crime data, the interface is modified only slightly to reflect the relevant categories.

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. [16]. Here, the use of interactive parallel coordinate plots was used to explore mortality data as it relates to socioeconomic factors. Other work includes Dang et al. [15] and Zhao et al. [46] which utilized dynamic queries and brushing for creating choropleth map views, and Tominski et al. [42] 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 [43] employed the use of intuitive 3D pencil and helix icons for visualizing multiple dependent data attributes and emphasizing the type of underlying temporal dependency.

Work by Hargrove and Hoffmann [20] used multivariate clustering to characterize ecoregion borders. Here, the authors select environmental conditions in a map's individual raster cells as coordinates that specify the cell's position in environmental data space. The number of dimensions in data space equals the number of environmental characteristics. Cells with similar environmental characteristics will appear near each other in data space.

3 Visual Analytic Environment

Our system adopts the common method of displaying georeferenced data events on a map and allowing users to temporally scroll through their data. However, such exploration only provides slices of spatial data at a given time or an aggregate thereof. In order to understand these slices, users need to know the trends of previous data (and, if possible, model future data trends). Furthermore, a limiting factor in using mapping as a tool for syndromic surveillance and crime analysis is that aggregation of data can lead to unreliable estimates of the true measure of the event impact. Fortunately, the data used in our visual analytics system provide georeferenced locations, allowing us to either aggregate the data on a spatial level, or employ statistical methods to model the data over arbitrarily sized georegions. As such, our system employs advanced statistical models for data exploration, enabling new visualizations, analyses, and enhanced detection methods.

3.1 System Features

Fig. 2a provides a conceptual overview of our visual analytics system, and Fig. 2b provides a screenshot of the system modified specifically for syndromic surveillance data. Note that all features described in this section are available for both crime analysis and syndromic surveillance. Data entering our system first undergo a cleaning and transformation process. This process is then refined through feedback from our visual analytics system. Furthermore, the user may report data errors as well, allowing for data correction. Finally, frequently accessed time series models of the data are also stored in the database for future use after initial modeling is done via our visual analytics system.

Further interaction is performed within the different viewing and modeling modalities of the system. As shown in Fig. 2b, the main viewing area is the geospatiotemporal view, and the three graphs on the right allow users to view a variety of data sources simultaneously for a quick comparison of trends across varying hospitals/precincts or data aggregated over spatial regions. Both the geospatial and time series viewing windows are linked to the time slider at the left side of the screen. This allows users to view the spatial changes in the data as they scroll across time. Additionally, temporal controls are also employed. These controls are denoted as "aggregate" and "increment" in the scroll bar window. The aggregate function allows the user to show all data over a period of x days. The increment

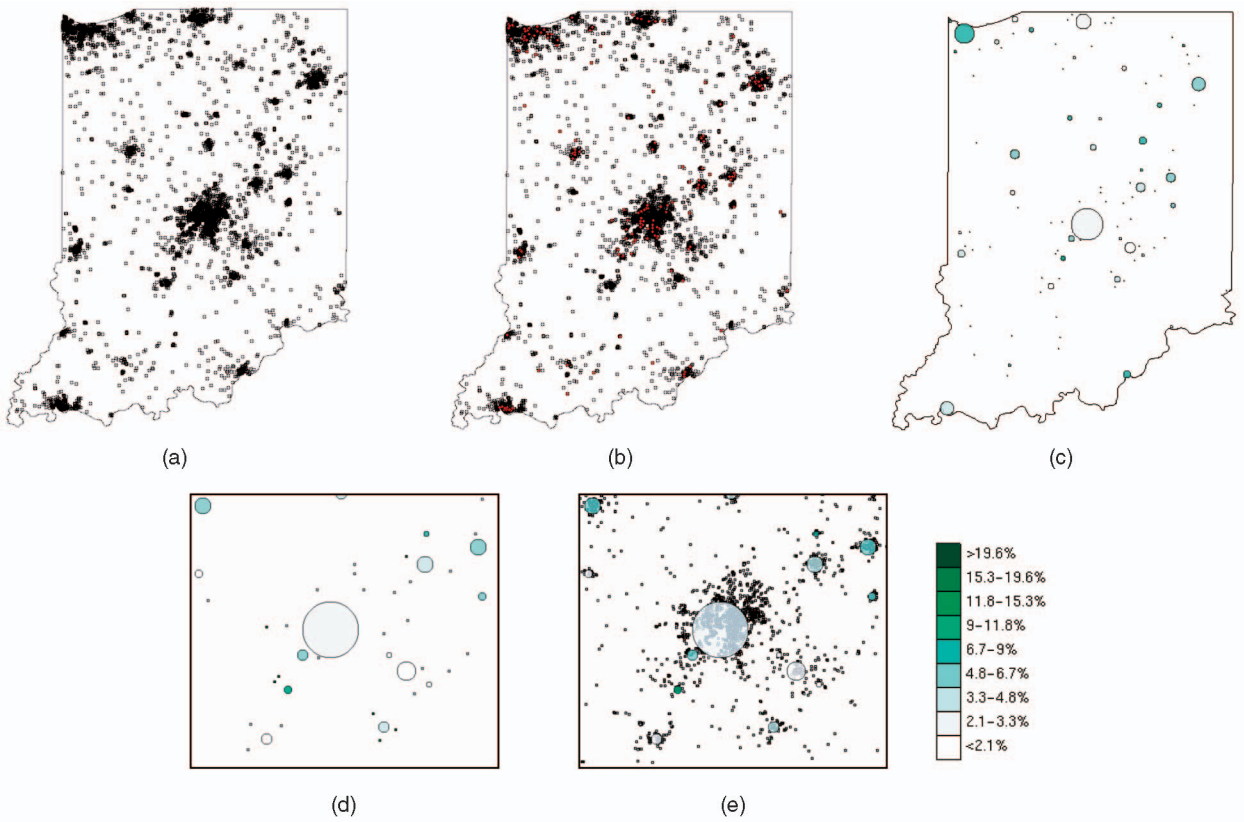

Fig. 3. Data aggregation and privacy preservation. (a) Georeferenced syndromic surveillance data as small additive opacity circles. (b) Georeferenced data overlaid with red circles representing syndromic patients. (c) Data aggregation for enhanced visualization. (d) High-resolution zoom of an area of interest. (e) Actual patient locations at a high-resolution zoom overlayed with our data aggregation method.

function allows the user to step through the data by increments of 1, 2, 3, ... days. All temporal views also provide a locking mechanism in which the user can choose to freeze the data window(s) while exploring changes across time in other views. This allows users to explore data while keeping a reference point to the time-varying trend(s) under inspection.

Another key feature of our system is the interactive demographic and category filtering. Users interactively generate database search queries through the use of check boxes and edit controls to find specific categories, keywords, and gender and age demographics from the data. Such work furthers hypothesis generation and exploration as users can now quickly filter signals by demographic constraints in order to search for correlations. The choices of filters affect both the geospatiotemporal viewing area and all unlocked temporal plots.

3.2 Data Aggregation and Privacy Preservation

Our system also provides multiple views for enhanced visualization and analysis. One simple, yet key view for this data set is showing georeferenced data locations on the map in order to provide analysts with a quick overview of statistics across the area of interest. Unfortunately, in both syndromic surveillance and crime data, showing exact event locations on a map is encumbered by privacy issues. Previous work in visualizing health statistics bypasses these concerns by showing data spatially aggregated over geographical areas such as zip code or county. While such visualizations are useful, there are times when it may be of interest to analysts to simply see a plot of event locations on a smaller level of data aggregation. Unfortunately, not all software users have the same level of permissions for viewing this data.

A naive visualization method would be to zoom out of the map at such a level that a pixel would represent a large enough region that it would be difficult to extract any private information about event mapped on a transformed geolocation to pixel basis. Unfortunately, as the data set becomes arbitrarily large, the visual clutter cannot be reduced in such a manner, see Fig. 3a, and it becomes clear that a visualization of every record at a high spatial zoom level is not effective for analysis. Furthermore, simple methods, such as using additive opacity to demonstrate event density, Fig. 3c, are inadequate as the number of events makes it impossible to readily distinguish density levels between areas. This is further complicated when the events are then highlighted with regard to their locations. Fig. 3b shows the syndromic patients mapped in red. In order to alleviate this problem, we have employed a method of data aggregation for enhanced visualization at low resolution views, which also acts as a privacy preserving technique at low zooms.

Our data aggregation method finds sets of event locations where each member is at most a set distance from at least one other member. The group is then represented by a circle at the set's geographic center that has an area proportionate to the size of the set. This allows us to

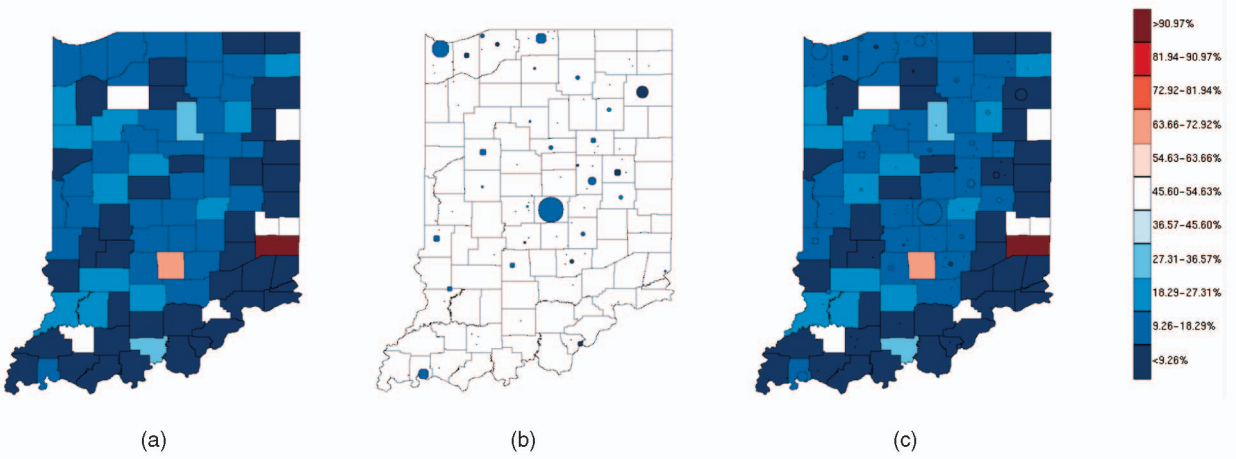

Fig. 4. Data aggregation and privacy preservation visualized as a percentage of syndromic population over the total population seen. (a) Data aggregated by county. (b) Data aggregated through nearest neighbor groupings. (c) A combination of data aggregation to enhance contextual visualization.

successfully aggregate data around major cities while preserving the autonomy of smaller sets in rural areas. This method is derived from the idea of connected components in graph theory, where patients are connected if and only if they are within the threshold distance from another patient in the graph [14]. The generated circles are then colored using a color map [6], where the color represents the percent of events within a given category found within this geographical centroid. This method operates under the assumption that the data are clumped in certain locations, otherwise it is possible to have an aggregation that hides too much of the actual data. Furthermore, as this method groups data at its geographic center of mass, it preserves the data context and helps alleviate privacy concerns.

Fig. 3c shows the low resolution aggregation of our syndromic surveillance data across the state of Indiana. Fig. 3d shows the zoomed in region, and Fig. 3e represents where the actual patient locations would be with respect to their representation as a geographic centroid.

3.3 Heatmaps

While such data aggregation can be useful for an overall view of event distribution, it is also useful to model the event distribution across the entire area of interest in order to approximate trends where little or no data values exist. Therefore, our system provides a geospatial heatmap [17] view which employs a diverging color map (or any other Color Brewer scheme) [6] to represent the percentage of a given event category over the total events seen on a given day.

The georeferenced data contain a set of observations in which an event (crime/syndrome) occurs at location X_i associate at time t with a hospital/precinct and is classified with a particular category. Such data are often aggregated by county or zip code and then shown to the user. This type of aggregation can be thought of as a histogram or box-plot of the data, and while a spatial histogram can be useful, such a visualization does not provide any hints as to what may be occurring in areas with little-to-no event data. Furthermore, areas with a small number of events may stray toward a high percentage of the total events in the category under question. In those cases, visual alerts may be triggered that would clearly appear as false positives once the individual records were analyzed. Fig. 4a demonstrates the problems with visualizing such histogram distributions. The national baseline influenza-like-illness (ILI) percentage during flu season is 2.1 percent [1] for the 2006-2007 season. Note in Fig. 4a that many counties seem to be visually displaying an extremely high level of ILI, where if we compare this to the overlaid data aggregation circles, these counties actually have very few patients contributing to the aggregations' center of mass as seen in Figs. 4b and 4c.

To overcome these issues, our system estimates the probability density function of all the recorded events using the georeferenced locations and produces a heatmap visualization of the area. To this end, we employ a kernel density estimation [37]. Kernel density estimation takes a collection of sample points and fits a weighting function at each point. A kernel is a nonnegative real-valued integrable function that integrates to one over all real values, and is symmetrical about the origin. The bandwidth of the kernel can be fixed or dynamic depending on the method employed. The bandwidth of the kernel influences the magnitude of the kernel, i.e., kernels with large bandwidths have a smaller height.

Equation (1) defines the multivariate kernel density estimation, and this method has been used in other works [23], [29], [18]. To reduce the calculation time, we have chosen to employ the Epanechnikov kernel, (2).

$$\hat{f}_h(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\mathbf{h}^d}K\left(\frac{\mathbf{x}-X_i}{\mathbf{h}}\right), \tag{1}$$

$$K(\mathbf{u}) = \frac{3}{4}(1-\mathbf{u}^2)1_{(\|\mathbf{u}\|\leq 1)}. \tag{2}$$

Here, $\mathbf{h}$ represents the multidimensional smoothing parameter, N is the total number of samples, d is the data dimensionality, and the function $1_{(\|\mathbf{u}\|\leq 1)}$ evaluates to 1 if the inequality is true and zero for all other cases. We calculate both the density estimation for the event category of interest as well as the density estimation of all categories in our system using an appropriately chosen h for each data set.

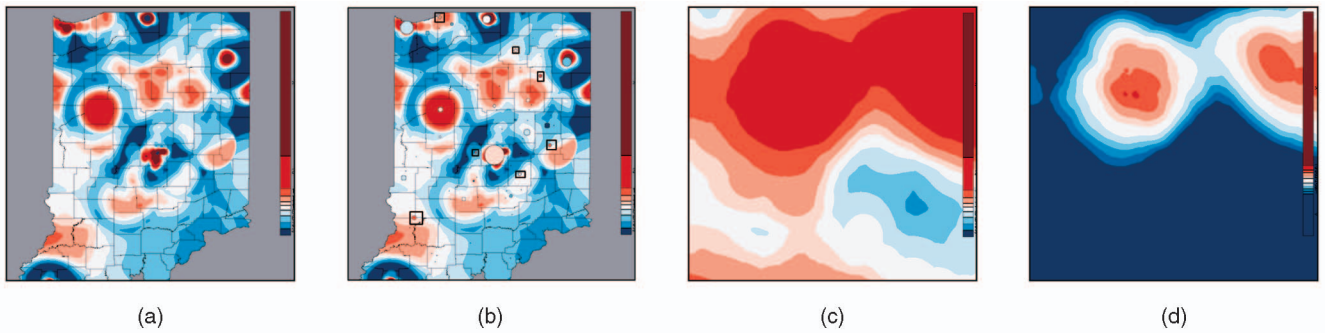

Fig. 5. Kernel density estimate (KDE) heatmaps visualized as a percentage of syndromic population over the total population seen. (a) KDE heatmap. (b) Contextualizing the KDE heatmap by overlaying patient data aggregated through nearest neighbor groupings. (c) A zoomed in view of a local hotspot. (d) Contextualizing a hotspot through interactive coloring.

The density estimation for the event category of interest is then divided by the density estimation for the total events to provide a percentage count for the expected number of events within a given area.

Unfortunately, a fixed bandwidth kernel turns out to be inappropriate for our data due to sparse data counts in rural counties and high data counts in large urban areas. A large fixed bandwidth over smoothes the data while trying to accommodate for the sparse data regions, and a small fixed bandwidth is unable to handle data in sparse regions, creating visual alerts in a similar fashion.

To overcome these issues, we employ the use of a variable kernel method [37], (3). This estimate scales the parameter of the estimation by allowing the kernel scale to vary based upon the distance from X_i to the kth nearest neighbor in the set comprising $N-1$ points.

$$\hat{f}_h(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{d_{i,k}}K\left(\frac{\mathbf{x}-X_i}{d_{i,k}}\right). \tag{3}$$

Here, the window width of the kernel placed on the point X_i is proportional to $d_{i,k}$ (where $d_{i,k}$ is the distance from the ith sample to the kth nearest neighbor) so that data points in regions where the data are sparse will have flatter kernels. Unfortunately, our data sets also exhibit problems with this method. In healthcare data, a primary recipient of emergency care are patients of long-term healthcare facilities (for example, nursing homes). As such, the use of the k nearest neighbors may result in a $d_{i,k}$ of 0 as many patients visiting emergency rooms may report the same address. This concept can be extended to large apartment complexes generating many noise complaints in the crime data, as well as data uncertainty (for example, many hospitals report unknown patient addresses as the hospital address). To overcome these issues, we slightly modify the variable kernel estimation to force it to have a minimum fixed bandwidth of h as shown in (4).

$$\hat{f}_h(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\max_{(h,d_{i,k})}}K\left(\frac{\mathbf{x}-X_i}{\max_{(h,d_{i,k})}}\right). \tag{4}$$

In the case of our modified variable kernel estimation, we calculate the kernel only spatially as opposed to both spatially and temporally as was done in the fixed bandwidth method. Future work will include extending our modified density estimation into the temporal domain. Results from our variable kernel estimation can be seen in Fig. 5. Slight problems in the estimation can be found near the state borders due to the abrupt cutoff of data in those areas. Future work will address these issues through more advanced spatial modeling.

3.4 Context through Color Exploration

Of key importance in all the previously presented data aggregation methods is the choice of coloring. In coloring our maps, data ranges get binned to a certain color. Clearly, the choice of bins can be based on model assumptions of the expected percentage of events within an area. However, each category of event will have varying model assumptions. Furthermore, the distribution of the data can also play a key role in placing hotspots into the proper context. For example, if the data are binned such that the maximum value covers a large range of variation, it is possible that such a mapping could hide hotspots within hotspots.

As such, we have created an interactive color widget for exploring data ranges. This widget allows users to modify the color scale either interactively or through a set of mathematical binning functions. We provide functions for linear, ramp, exponential, and logarithmic binning.

In linear binning, the points on the map are first binned across a large histogram. The histogram is then divided such that each color represents an equal number of points within the data. For ramp binning, the histogram is divided such that each color represents an increasingly larger number of points, following along the line $y = x$. This idea is then extended for both exponential and logarithmic curves. Future work will include binning the data to a Gaussian distribution. The distribution functions were based on observations that the mathematical distributions are often able to automatically highlight various properties of the data with little user interaction. Further research into this connection is left to future work.

In exploring the data through contextual color clues, our domain experts typically employed the use of either a default mathematical binning, or arranged the color bins manually such that the last bin began at some data threshold of interest. In Fig. 5, the data have been mapped using a logarithmic binning. Both the data aggregation and the kernel density estimation tools can be used in conjunction for contextualizing hotspots. Here, we find several hotspots in the state. When placed in the context of the data aggregation overlay (Fig. 5b), we begin to develop hypotheses of key places that need further exploration. These places are marked by the black squares in Fig. 5b.

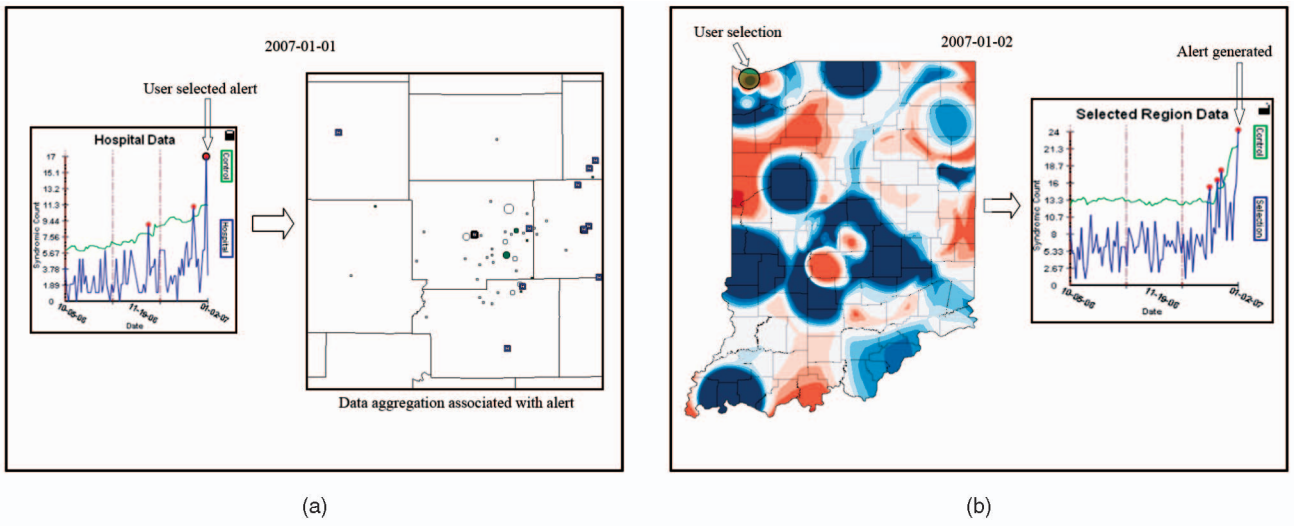

Fig. 6. Exploration using linked views. (a) Images taken from our system illustrating the linked temporal analysis to the geospatial filtering. Here, a user has selected the alert occurring on 01-01-2007. The geospatial viewing window then opens that day's data corresponding to the alert allowing for further investigation. (b) Images taken from our system illustrating the linked views in selecting geospatial areas and seeing temporal plots. Here, a user has selected an area in north-west Indiana (the green circle). This selection brings up the time series graph and our alert detection algorithm finds an unusual event in that area on that day.

Further, we see the dense hotspot centered in the middle of the state. To further explore this hotspot, users may zoom into the map. The zoom results in a recalculation of the kernel density estimate as the latitude/longitude point space mapping to the grid changes. Fig. 5c provides a zoomed in view of the state's central hotspot. Notice that this heatmap is dominated by a singular range of red. In Fig. 5b, the user interactively adjusts the color scale to provide more binning across that particular data range. Through this interaction, the user is now able to find several previously undetectable peaks within this region that may warrant further investigation.

3.5 Time Series Analysis

While the spatial visualizations employed in our system are useful for detecting hotspots, it is also helpful for an analytics system to provide hints as to where outbreaks may be occurring. To this end, we have employed the use of a standard epidemiological algorithm for time series analysis, the cumulative summation (CUSUM) [22]. The CUSUM algorithms provide alerts for potential outbreaks in the temporal domain, and users of our system may then select these alerts for further exploration in the geospatio-temporal viewing window.

$$S_t = max\left(0, S_{t-1} + \frac{X_t - (\mu_0 + k\sigma_{x_t})}{\sigma_{x_t}}\right). \quad (5)$$

Equation (5) describes the CUSUM algorithm, where S_t is the current CUSUM, S_{t-1} is the previous CUSUM, X_t is the count at the current time, μ_0 is the expected value, σ_{x_t} is the standard deviation, and k is the detectable shift from the mean (i.e., the number of standard deviations the data can be from the expected value before an alert is triggered). We apply a 28 day sliding window to calculate the mean, μ_0, and standard deviation, σ_{x_t}, with a 3 day lag, meaning that the mean and standard deviation are calculated on a 28 day window 3 days prior to the day in question. Values chosen were based on the EARS C2 alert method detailed in [22]. Such a lag is used to increase sensitivity to continued outbreaks while the 28 days provides a month worth of baseline data to test against while minimizing long-term historical effects. Fig. 6 shows the application of the CUSUM algorithm to the temporal plot of ILI counts during peak flu season. An alert is represented by a large red circle, which is generated if S_t exceeds the threshold (for a point of reference the threshold is typically set at three standard deviations from the mean in the Early Aberration Reporting System and is shown as the green line in Fig. 6).

3.6 Exploration with Linked Views

While the alerts generated from aberration detection algorithms may produce a useful starting point for exploration, they may also be eliciting false alarms. Furthermore, analysts may want to explore areas where information may be unknown, for example, visual hotspots generated in our heatmap approach may contain only sparse data points. Ideally, analysts would like to dynamically query and select elements on the visual display in order to see how selections update related views. This type of selection is commonly referred to as *brushing* [5] and it is used in many interactive visualization environments [33], [36].

For our implementation, we use only the *highlight* operation over the time dimension of our temporal view and the spatial region of our main viewing window. In the temporal view, the highlighted region is shown in red and once the mouse button is released, all other information displays are updated to reflect the selection. Because the individual plots are interrelated, only one may be brushed at a time. The principal purpose of this feature is to allow selection of the current day and the number of days being aggregated together from the plot windows based on a region of interest in the plotted data. In Fig. 6a, we see a series of hospital generated alerts (the red marks) in the middle temporal viewing window. In this figure, a user has clicked on an alert, causing the temporal window to lock in place, while scrolling the geospatial window back in time to the alert on that day. Notice that the events associated with the hospital/precinct and category are now exclusively shown on the map.

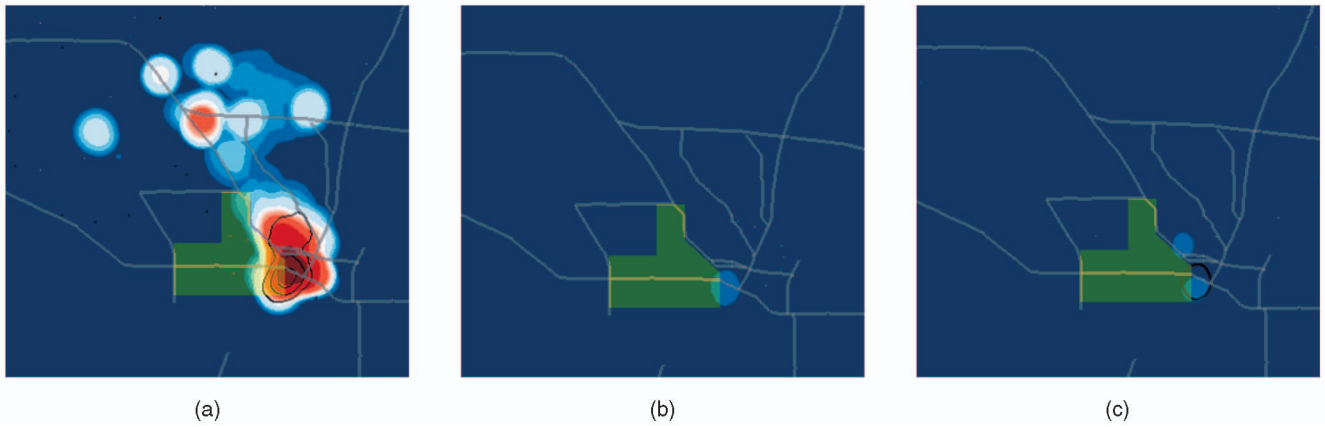

Fig. 7. Contour mapping for contextual cues. (a) The analyst has created a heatmap of one year's worth of crime data in West Lafayette Indiana with the yellow polygon representing Purdue University Campus. The contours overlaid represent the noise complaints aggregated over the past 100 days. (b) The analyst plots only the noise complaints aggregated over the past 50 days as a heatmap. (c) The analyst plots the noise complaints aggregated over the next 50 day period. The contours in this figure represent the previous noise complaint map outline.

In the geospatial view, highlighting is performed through a circular selection of an area. This circular selection allows users to select multiple geographic regions and view their temporal history. In Fig. 6b, we see a heatmap of the state. In this figure, note that the circled area represents a user selection. Here, the user has chosen a region of the state that appears to currently be a syndromic hotspot. A linked time series analysis view plots the data from that area in the lower right window. Here, we see that an alert (small red circle) is found for that area on the day in question. A user can then further explore these alerts by clicking on the alerts in the time series window to find the patients associated with this alert in the geospatial window.

3.7 Temporal Contours

Along with exploring data context through color adjustments, we also provide contour line options for preserving the temporal history of the data. In the case of using temporal contours, data shown in the heatmap mode can be overlaid with the past x-days worth of contour lines. The x-days is a user defined parameter, and the color date aggregation is shown as a label on the maps, providing users with the appropriate temporal context. This allows users to view shifting hotspots across time and analyze the movements of trends and patterns over days.

In Fig. 7a, the analyst has visualized a heatmap of all criminal reports in West Lafayette, Indiana, over the past year. The analyst then overlays a 100 day aggregate of the noise complaint data from the beginning of the semester as a contour in order to see which hotspots on the map are most related to noise. Here, we can see that the noise complaints are directly correlated with an area of West Lafayette nearest Purdue's Campus (the yellow polygon). Next, the analyst changes mode to only view the heatmap noise complaints from this time period (Fig. 7b), and then uses the contour history mode to compare the next 50 day period aggregate of noise complaints to the previous map (Fig. 7b).

3.8 Multivariate Views

Our system also provides a series of complex viewing modalities for searching for correlations between multiple variables. In a two-dimensional view of the geographic area, one can map the density estimated heatmap color to variable x, and then create another heatmap for variable y, thus providing a multivariate view of the geographical location and the x and y variables. Variable y can then be displayed as contours overlaid on the heatmap of variable x. Users can then look for places of high contours and high colors to search for correlations between data variables. Furthermore, one can create a view for multiple variables by assigning a third variable as height. The data can then be viewed in three dimensions, and users can search for correlations between three variables simultaneously. The heatmaps, contours, and height are all calculated based on the kernel density estimation described earlier in Section 3.3.

In Fig. 8a, the analyst is searching to see if there are any hotspots showing a correlation between rash cases (the contours) and shock/coma cases (the color). Here, the analyst finds that there are large concentrations of cases in several overlapping areas. The analyst then also chooses to look for cases associated with respiratory illness, and enters 3D mode. In 3D mode, height now represents the magnitude of respiratory illness cases. Fig. 8 shows the 3D view with color, contour, and height mapping.

3.9 Interactive Thresholding

Our system also provides users the ability to interactively select data threshold values that they are interested in. In order to better focus attention on areas of interest, users may choose to only look at event values, where the percentage of events occurring in an area is greater than some threshold, t. Fig. 9a illustrates an analyst searching for gastrointestinal hotspots in Indiana. In Fig. 9b, the user has thresholded the data such that only higher values will appear. The user then moves forward in time by 10 days (Fig. 9b) using the temporal contour ghosting to track the movement of hotspots across the state. Contours are displayed such that the most recent days are drawn with a higher opacity, thus, creating the temporal contour ghosting. However, as the number of historical days being viewed increases, the number of colors that can be distinguished in this manner reaches its maximum. As such, contour ghosting is only effective for a limited historical basis. Note that the thresholding is also applied to the contour history as well, where again, the lighter the contours, the further in the past they have occurred.

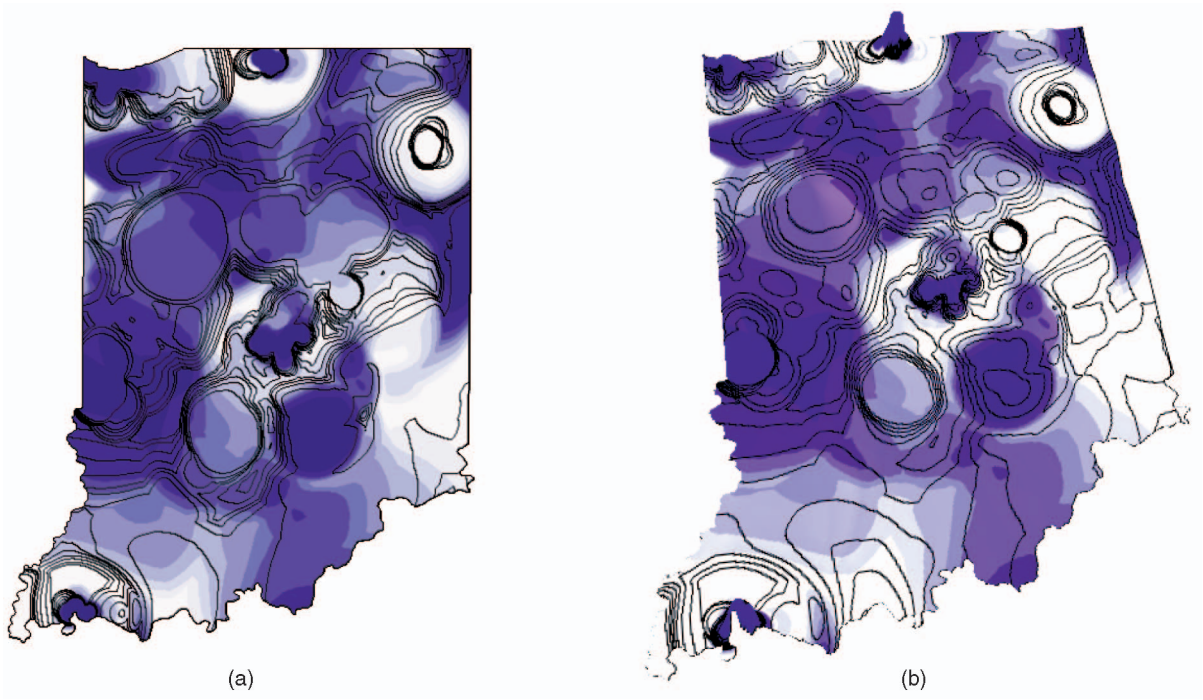

Fig. 8. Multivariate views. (a) The analyst has created a heatmap of shock/coma cases overlaid with contours for rash. (b) The analyst adds the category respiratory as the height dimension.

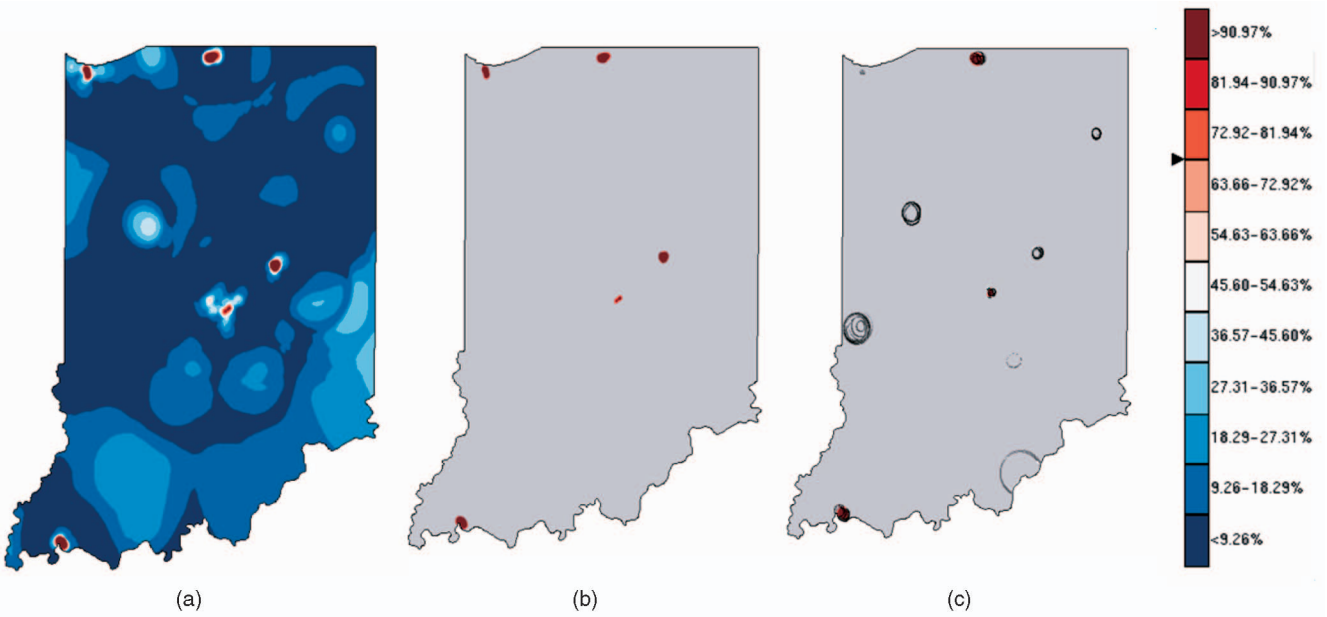

Fig. 9. Interactive thresholding. (a) The analyst searches for gastrointestinal hotspots. (b) The analyst uses the thresholding capability to filter the data. (c) The analyst moves forward in time viewing movement trends among the data.

4 Understanding Hotspots

By using a combination of geospatial and temporal visualization and analytics tools, our system provides analysts with tools for real-time hypothesis generation and exploration. Here, we present two example cases of using such a system to analyze data.

4.1 Syndromic Hotspots

To better illustrate the hypothesis generation/exploration phase, we conducted an informal interview with an Indiana State Department of Health (ISDH) syndromic surveillance epidemiologist. During this interview, we discussed how an epidemiologist would search for syndromic hotspots, creates an initial hypothesis, and what steps are taken in an attempt to confirm or deny this hypothesis.

Traditionally, the first items examined when identifying potential syndromic problem areas are the spatial alerts generated for a given syndrome. Based on the epidemiologist's experience, certain alerts will be immediately resolved as false positives, and others will be moved to the top of the queue. From the alerts the epidemiologist identifies as potential problems, a hypothesis is formulated

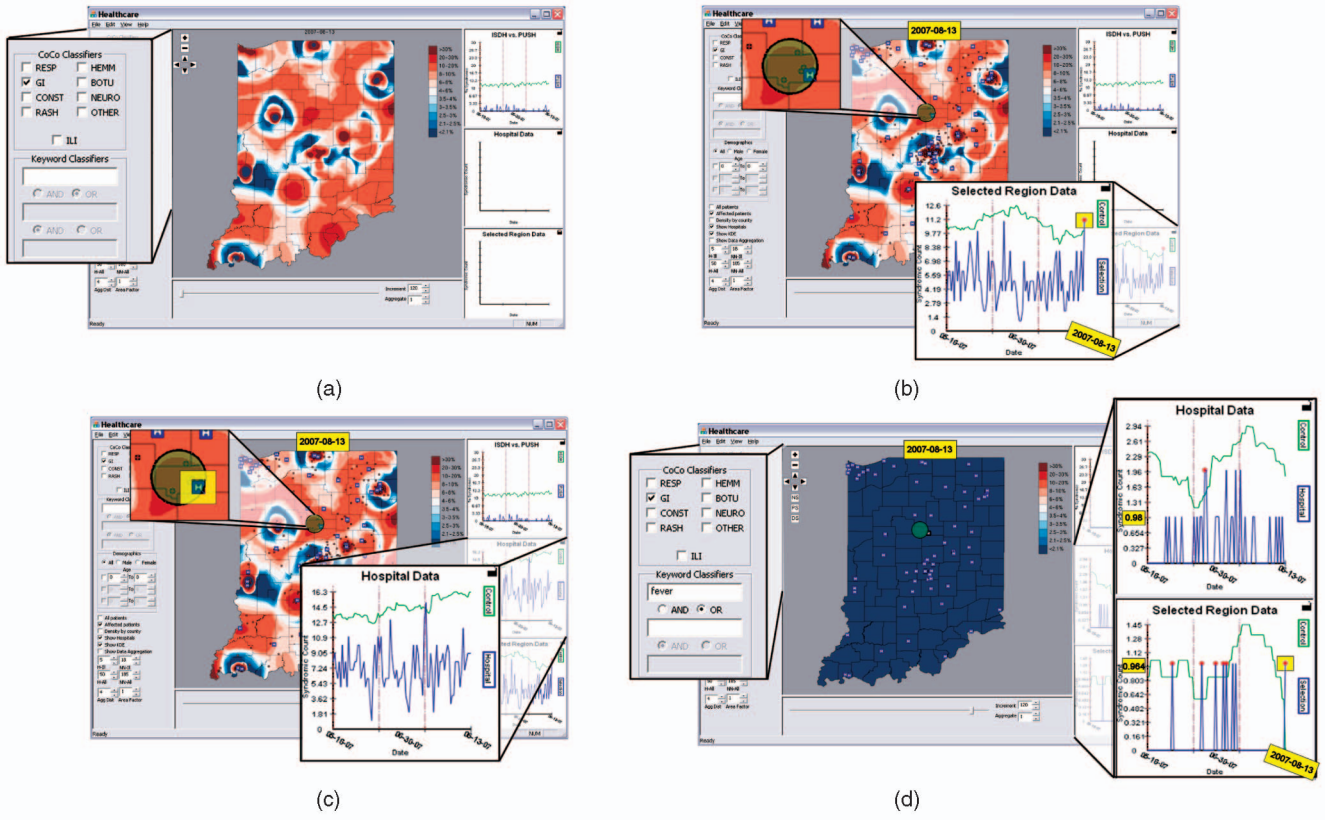

Fig. 10. Using visual analytics for hypothesis exploration in syndromic surveillance. (a) The user observes a heatmap for a given syndrome, in this case, gastrointestinal. (b) Next, the user selects an area of interest, generating a time series plot for that region. Note that in the time series plot generated, an alert is occurring on the day of interest. (c) The user then drills down to the hospital level by selecting the neighboring hospital and generating a time series plot for that emergency department. Here, we see that there is no hospital level alert for gastrointestinal syndromes. (d) Finally, the user looks for correlating symptoms and filters by the keyword fever. New time series plots are generated. While an alert still exists for the selected area, the user can now see that this alert was generated by only one individual, meaning an outbreak is unlikely.

stating that a problem with syndrome X is occurring in patients found at location Y. These alerts are aggregated by zip code level, meaning that zip codes A, B, C, etc., contribute to the alert. From this step, the epidemiologist would look at the time series data for all zip codes contributing to the alert in order to gain a better understanding of where the baseline lies. In contrast, our visual analytics tool allows users to select an arbitrary region to view the time series data, providing a baseline for the overall area, potentially allowing quicker comparison.

Often, the next step taken would be to further corroborate the geospatial area of the alert by looking at the counties involved and pulling up county level alerts, their corresponding time series plots, and county maps down to the zip code level. Similarly, our tool provides both heatmaps at the reduced levels of granularity, as well as a finer, smoother granularity heatmap option that the epidemiologist thought may add value. If, from the heatmap, the hypothesis cannot be rejected, the next step is to drill down to patient level data in order to assess the actual chief complaints. For example, if (in the case of a gastro-intestinal problem) a patient's "vomiting" is related to pregnancy, then it is less likely to be part of the gastrointestinal outbreak being considered in this hypothesis. As such, sometimes potential clusters then fall apart. Next, the epidemiologist would look at the patient level data to assess time stamps and actual chief complaints for clustering which may lead to filtering by ages for clustering and gender for skew if clues exist that lead the hypothesis refinement in those directions. If there was a string of elevated days, then the analyst would group these elevated days and do the same type of descriptive analysis. Our dual linked views provide advanced tools for such an operation, aiding in the overall hypothesis exploration.

During this process, the epidemiologist also searches for potential "cosyndromes" in the same geography, such as fever, to see if it is somehow linked to the gastrointestinal problem. Again, the linked views and filter options of our system allow the user to easily look at multivariate time series components. If concurrent syndromes are found, this potentially strengthens the hypothesis and may lead to a follow-up with the actual emergency department(s) involved. Fig. 10 illustrates the use of our system during the hypothesis exploration phase.

First, in Fig. 10a, the user has selected the syndrome he/she is interested in analyzing, in this case, gastrointestinal. This generates a query to the database, and the epidemiologist can now look at the patient distribution with either an additive opacity for all patients that visited an emergency department, or as an aggregate of the data. Next, the user visually searches for unusual hotspots using a combination of the kernel density estimation and the patient overlay. The

Date	Data	Expected	Detection
31-Dec-08	149	149.929	0.253
30-Dec-08	143	149.643	0.114
29-Dec-08	216	149.571	0.001
28-Dec-08	185	148.357	0.012
27-Dec-08	170	147.786	0.079
26-Dec-08	200	148.607	0.008
25-Dec-08	128	148.714	0.893
24-Dec-08	116	149.607	0.96

Fig. 11. Sample ESSENCE output showing daily values for a single Indiana county.

user may select multiple areas for testing; however, if the area selected shows no temporal alert for the day in question, then it is likely that the hypothesis of area X being problematic is rejected.

In Fig. 10b, the user has selected an area of the map in central Indiana, and the corresponding time series graph that was generated indicates that the selected area is showing an alert on the day in question. The next step in analyzing this alert is to look at data from the nearby emergency departments. In this case, there is only a single emergency department. The user clicks on the hospital glyph on the map, and the time series plot for this emergency department is generated, see Fig. 10c. In this time series plot, there is no alert generated for this emergency department for the day in question. This weakens the hypothesis that there is an outbreak in the area; however, the user may still want to take further steps to confirm/deny the hypothesis.

The next step taken is to look for corresponding symptoms. In this case, the user looks for patients with gastrointestinal syndromes that also reported signs of fever. Fig. 10d shows this filter query. Note that the heatmap and time series plots are automatically updated from the query. We can see now that there are no visual hotspots occurring on the map; however, there is still a time series alert for that area. Further investigation of the time series alert shows that the expected number of patients was slightly less than one, and one patient came in on that day, thereby generating an alert. It is now unlikely that an outbreak is occurring in this area, and the hypothesis can be denied after a brief analysis of the patient record.

While it may seem odd that one case can cause an outbreak alert, this is quite a common occurrence in all current systems. For example, the carbon monoxide case shown in Fig. 1 contains only three emergency department complaints. Therefore, the high sensitivity is necessary to avoid missing small cluster cases.

4.2 Syndromic Hotspots in Current Systems

In order to further evaluate our system, we would need to include experts within the field of syndromic surveillance in order to reduce training time and develop meaningful test scenarios. Currently, the Indiana State Department of Health employs only a few epidemiologists (who we have consulted). While it is possible that other counties in the state employ epidemiologists with knowledge of syndromic surveillance, the pool of available subjects is quite limited. In regards to system evaluation, we provide a comparison to the current tools employed at the State Department of Health.

The current tool used by the Indiana State Department of Health is ESSENCE [27]. The alerts are displayed in a line listing (Fig. 11) that is reviewed every day with color codes representing serious and mild alerts. There are region alerts for counties, hospital alerts, and spatial alerts for detecting clusters. Selecting an alert from the line listing will bring up a time series of 90 days, from here you can drill down to the details of the alert, including the ability to map the patients by zip code.

As discussed in the previous section, our program is also able to generate temporal alerts based on any level of spatial aggregation for counties and hospitals. Future work will include the introduction of spatial alerts through the use of SatScan [24]. Our advantages over the ESSENCE system includes the ability to interactively aggregate data over a variety of temporal ranges, as well as providing a variety of spatial aggregation methods (as opposed to only plotting data by zip code). We also provide enhanced interactive filtering for multivariate data exploration as well bivariate views through coloring and contouring and multivariate views through the addition of height fields.

4.3 Crime Hotspots

Our second expert is an independent evaluator with experience as an Indiana State Police Commanding Officer. He is an accomplished security professional having a 25 year operations background in public safety, corporate security, and the military. He is retired from the Indiana State Police reaching the rank of Regional Commander. In this capacity, he was responsible for overseeing the investigative mission within a 12,000 square mile, 21 county region in northwest Indiana. His authority covered all criminal and civil investigations, including death investigations, violent crimes, thefts, burglaries, public corruption, and internal investigations. Again, many police departments do not employ crime analysts as their budgets are limited. As such, the use of a single expert for feedback and evaluation seems to be the most appropriate. The scenario presented in this section is based on questions the evaluator wished to ask of the data.

In the case of our crime data, events occur less frequently than in the case of syndromic surveillance data. As such, analysts often wish to look at the last x-days or weeks of data and begin planning new patrol routes based on previous trouble locations. For our example, the analyst first pulls up all criminal activities for the 2007-2008 Purdue University school year (Fig. 12a), and finds two major areas of activities, one nearby campus, and one near the intersection of two major cross-streets. In this case, two major hotspots are evident, one near campus, and one located at an intersection some distance from the main campus.

Here, the analyst chooses to investigate the causes leading to the secondary hotspot. First, the analyst searches for what types of crimes are occurring near the secondary hotspot by filtering the region by various crime categories and finds that theft is the leading crime in that area. Next, the analyst compares the thefts from fall semester in West Lafayette to the overall criminal activities of the 2007-2008 school year. Fig. 12a shows the overall crimes from the

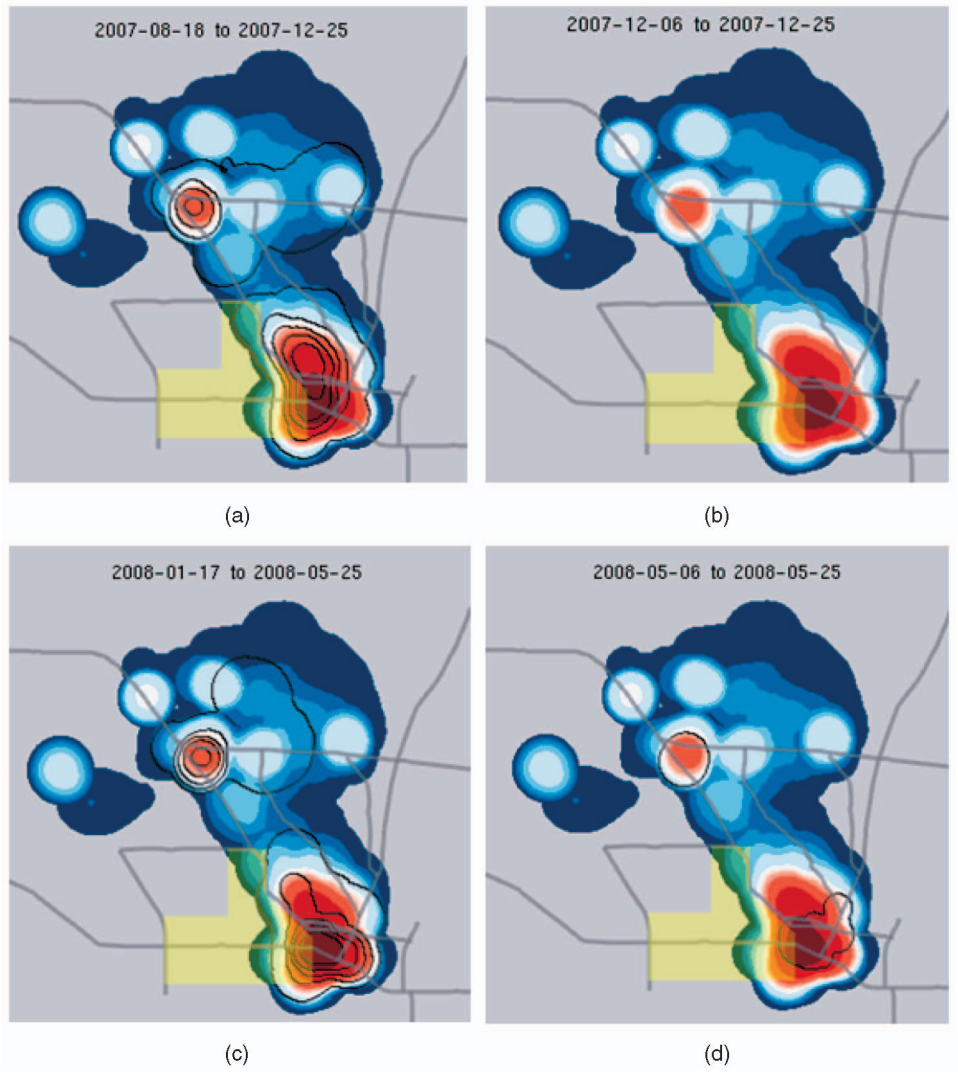

Fig. 12. Using visual analytics for hypothesis testing in crime analysis. The user is analyzing thefts (as contours) versus all crimes (as color) for a given school year (2007-2008). (a) The user analyzes fall semester thefts (contours) compared with the over all school year crimes. (b) The user analyzes the last 20 days of the fall semester for thefts. (c) The user analyzes spring semester thefts. (d) The user analyzes the last 20 days of spring semester.

2007-2008 school year overlayed with contours from the 2007 Fall semester theft reports. Next, the analyst investigates if the end of the fall semester (the week of finals and the week directly after finals) indicates a high rate of thefts, and shifts the contours to only map to the last 20 days of the fall semester (Fig. 12b). The analyst finds that in the fall semester, no thefts are occurring during this time period.

Next, the analyst chooses to compare with events from the spring semester. Fig. 12c shows the overall crimes from the 2007-2008 school year overlayed with contours from the 2008 Spring semester thefts. Next, the analyst investigates if the end of the Spring semester (finals week and the week after finals) indicates a high rate of thefts, and overlays the theft contours from the last 20 days of the semester, Fig. 12d. Here, the analyst finds that a large number of thefts are taking place during this time period. The analyst may then begin forming hypotheses about why this occurs at the end of the Spring semester (more students moving out of town, warmer weather) as opposed to at the end of Fall semester when more houses are empty for the holidays.

Based on this tool, the analyst found that he was easily able to test hypotheses that he had about the nature of criminal activities in the West Lafayette community. Currently, with departments that have no crime analysts, he felt that our tool provided a simple and intuitive means of analyzing that data the department would benefit from. Being able to visually understand what the data represent is important at the tactical level for the street officer to anticipate problems before they occur. For the Police Chief, visualization provides a strategic advantage in deploying resources, managing budgets, and developing strategies for crime reduction and predictive analytics. Our expert found this to be the next generation of crime mapping technology that along with its mobile capabilities and ease of use, could be easily integrated into current law enforcement products and techniques.

5 Conclusions and Future Work

Our current work demonstrates the benefits of visual analytics for understanding syndromic hotspots. By linking a variety of data sources and models, we are able to enhance

the hypothesis generation and exploration abilities of our state partners. Our initial results show the benefits of linking traditional time series views with geospatiotemporal views for enhanced exploration and data analysis. Our system also moves away from traditional spatial histogram visualizations, providing a finer granularity of heatmap for more accurate hotspot detection.

Other future work includes advanced modeling of geospatiotemporal data for enhanced data exploration and hotspot detection. Furthermore, we plan to include a suite of aberration detection algorithms and their corresponding control charts for enhanced alert detection in the temporal domain. We also plan on employing spatiotemporal clustering algorithms (such as SatScan [24]) for hotspot detection as well as other correlative analysis views within the temporal domain (scatter plots, calendar view, etc.). Furthermore, we plan to enhance our system from a visual analytics system to a predictive analytics system, creating views to allow for event planning, prediction and interdiction. Once these features are implemented, we plan to deploy our system with our state partners for further evaluation.

ACKNOWLEDGMENTS

The authors would like to thank the Purdue University Student Health Center, the Indiana State Department of Health, and the Police Department of West Lafayette, Indiana, for providing the data. This work has been funded by the US Department of Homeland Security Regional Visualization and Analytics Center (RVAC) Center of Excellence and the US National Science Foundation (NSF) under Grants 0811954, 0328984, and 0121288.

REFERENCES

[1] L. Blanton et al., "Update: Influenza Activity—United States and Worldwide, 2006-07 Season, and Composition of the 2007-08 Influenza Vaccine," *Morbidity and Mortality Weekly Report,* vol. 56, pp. 789-794, 2007.

[2] W. Aigner, S. Miksch, W. Muller, H. Schumann, and C. Tominski, "Visual Methods for Analyzing Time-Oriented Data," *IEEE Trans. Visualization and Computer Graphics,* vol. 14, no. 1, pp. 47-60, Jan./Feb. 2008.

[3] W. Aigner, S. Miksch, B. Thurnher, and S. Biffl, "Planninglines: Novel Glyphs for Representing Temporal Uncertainties and Their Evaluation," *Proc. Ninth Int'l Conf. Information Visualization (IV '05)* 2005.

[4] L. Anselin, I. Syabri, and O. Smirnov, "Visualizing Multivariate Spatial Correlation with Dynamically Linked Windows," *Proc. Workshop New Tools for Spatial Data Analysis,* CD-ROM, 2002.

[5] R.A. Becker and W.S. Cleveland, "Brushing Scatterplots," *Technometrics,* vol. 29, no. 2, pp. 127-142, 1987.

[6] C.A. Brewer, *Designing Better Maps: A Guide for GIS Users.* ESRI Press, 2005.

[7] A. Buja, D. Cook, and D. Swayne, "Interactive High Dimensional Data Visualization," *J. Computational and Graphical Statistics,* vol. 5, pp. 78-99, 1996.

[8] P. Buono, A. Aris, C. Plaisant, A. Khella, and B. Shneiderman, "Interactive Pattern Search in Time Series," *Proc. Conf. Visualization and Data Analysis,* pp. 175-186, 2005.

[9] T. Butkiewicz, W. Dou, Z. Wartell, W. Ribarsky, and R. Chang, "Multi-Focused Geospatial Analysis Using Probes," *IEEE Trans. Visualization and Computer Graphics,* vol. 14, pp. 1165-1172, Nov./Dec. 2008.

[10] C.C. Calhoun, C.E. Stobbart, D.M. Thomas, J.A. Villarrubia, D.E. Brown, and J.H. Conklin, "Improving Crime Data Sharing and Analysis Tools for a Web-Based Crime Analysis Toolkit: Webcat 2.2," *Proc. 2008 IEEE Systems and Information Eng. Design Symp.,* 2008.

[11] W.W. Chapman, J.N. Dowling, and M.M. Wagner, "Classification of Emergency Department Chief Complaints into 7 Syndromes: A Retrospective Analysis of 527,228 Patients," *Annals of Emergency Medicine,* vol. 46, pp. 445-455, Nov. 2005.

[12] L. Chittaro and C. Combi, "Visualizing Queries on Databases of Temporal Histories: New Metaphors and Their Evaluation," *Proc. IEEE Symp. Information Visualization (INFOVIS '01),* p. 159, 2001.

[13] W.S. Cleveland, *Visualizing Data.* Hobart Press, 1993.

[14] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms.* The MIT Press, 2001.

[15] G. Dang, C. North, and B. Shneiderman, "Dynamic Queries and Brushing on Choropleth Maps," *Proc. Fifth Int'l Conf. Information Visualization (IV '01),* pp. 757-764, 2001.

[16] R.M. Edsall, A.M. MacEachren, and L. Pickle, "Case Study: Design and Assessment of an Enhanced Geographic Information System for Exploration of Multivariate Health Statistics," *Proc. IEEE Symp. Information Visualization (INFOVIS '01),* pp. 159-162, 2001.

[17] *Information Visualization in Data Mining and Knowledge Discovery,* U. Fayyad, G.G. Grinstein, and A. Wierse, eds. Morgan Kaufmann Publishers, Inc., 2002.

[18] M. Gibin, P. Longley, and P. Atkinson, "Kernel Density Estimation and Percent Volume Contours in General Practice Catchment Area Analysis in Urban Areas," *Proc. Geographical Information Science Research Conf.,* 2007.

[19] S.J. Grannis, M. Wade, J. Gibson, and J.M. Overhage, "The Indiana Public Health Emergency Surveillance System: Ongoing Progress, Early Findings, and Future Directions," *Proc. Am. Medical Informatics Assoc. Ann. Symp.,* 2006.

[20] Hargrove and Hoffman, "Using Multivariate Clustering to Characterize Ecoregion Borders," *Proc. Computing in Science & Eng., the AIP and the IEEE Computer Soc.,* vol. 1, pp. 18-25, 1999.

[21] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "Themeriver: Visualizing Thematic Changes in Large Document Collections," *IEEE Trans. Visualization and Computer Graphics,* vol. 8, no. 1, pp. 9-20, Jan.-Mar. 2002.

[22] L.C. Hutwagner, W.W. Thompson, and G.M. Seeman, "The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS)," *J. Urban Health,* vol. 80, no. 2, pp. i89-i96, 2003.

[23] D. Kao, A. Luo, J.L. Dungan, and A. Pang, "Visualizing Spatially Varying Distribution Data," *Proc. Sixth Int'l Conf. Information Visualization,* pp. 219-225, 2002.

[24] M. Kulldorff, "A Spatial Scan Statistic," *Comm. Statistics: Theory and Methods,* vol. 26, pp. 1481-1496, 1997.

[25] A.D. Langmuir, "The Surveillance of Communicable Diseases of National Importance," *New England J. Medicine,* vol. 268, pp. 182-192, 1963.

[26] K. Liao, "A Visualization System for Space-Time and Multivariate Patterns (VIS-STAMP)," *IEEE Trans. Visualization and Computer Graphics,* vol. 12, no. 6, pp. 1461-1474, Nov./Dec. 2006.

[27] J.S. Lombardo, "A Systems Overview of the Electronic Surveillance System for the Early Notification of Community Based Epidemics (ESSENCE II)," *J. Urban Health,* vol. 80, pp. 32-42, 2003.

[28] J.W. Loonsk, "Biosense—A National Initiative for Early Detection and Quantification of Public Health Emergencies," *Morbidity and Mortality Weekly Report,* vol. 53, pp. 53-55, 2004.

[29] A.L. Love, A. Pang, and D.L. Kao, "Visualizing Spatial Multivalue Data," *IEEE Computer Graphics and Applications,* vol. 25, no. 3, pp. 69-79, May/June 2005.

[30] A.M. MacEachren, F.P. Boscoe, D. Haug, and L. Pickle, "Geographic Visualization: Designing Manipulable Maps for Exploring Temporally Varying Georeferenced Statistics," *Proc. IEEE Symp. Information Visualization (INFOVIS '98),* p. 87, 1998.

[31] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W.S. Cleveland, S.J. Grannis, M. Wade, and D.S. Ebert, "Understanding Syndromic Hotspots—A Visual Analytics Approach," *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST),* Oct. 2008.

[32] R. Maciejewski, B. Tyner, Y. Jang, C. Zheng, R. Nehme, D.S. Ebert, W.S. Cleveland, M. Ouzzani, S.J. Grannis, and L.T. Glickman, "Lahva: Linked Animal-Human Health Visual Analytics," *Proc. IEEE Symp. Visual Analytics Science and Technology,* Oct. 2007.

[33] A.R. Martin and M.O. Ward, "High Dimensional Brushing for Interactive Exploration of Multivariate Data," *Proc. Sixth Conf. Visualization (VIS '95),* pp. 271-278, 1995.

[34] S.F. Messner and L. Anselin, "Spatial Analyses of Homicide with Areal Data," *Spatially Integrated Social Science CD-ROM,* pp. 127-144, Oxford Univ. Press, 2002.
[35] C.-C. Pan and P. Mitra, "Femarepviz: Automatic Extraction and Geo-Temporal Visualization of Fema National Situation Updates," *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST '07),* pp. 11-18, Oct./Nov. 2007.
[36] J.C. Roberts and M.A.E. Wright, "Towards Ubiquitous Brushing for Information Visualization," *Proc. Int'l Conf. Information Visualization (IV '06),* pp. 151-156, 2006.
[37] B.W. Silverman, *Density Estimation for Statistics and Data Analysis.* Chapman & Hall/CRC, 1986.
[38] J. Stasko, C. Gorg, Z. Liu, and K. Singal, "Jigsaw: Supporting Investigative Analysis through Interactive Visualization," *Proc. IEEE Symp. Visual Analytics Science and Technology,* pp. 131-138, 2007.
[39] S.B. Thacker, R.L. Berkelman, and D.F. Stroup, "The Science of Public Health Surveillance," *J. Public Health Policy,* vol. 10, pp. 187-203, 1989.
[40] *Illuminating the Path: The R&D Agenda for Visual Analytics.* J.J. Thomas and K.A. Cook, eds. IEEE Press, 2005.
[41] C. Tominski, J. Abello, and H. Schumann, "Axes-Based Visualizations with Radial Layouts," *Proc. ACM Symp. Applied Computing (SAC '04),* pp. 1242-1247, 2004.
[42] C. Tominski, P. Schulze-Wollgast, and H. Schumann, "Visual Analysis of Health Data," *Proc. Int'l Information Resource Management Assoc. (IRMA) Conf.,* 2003.
[43] C. Tominski, P. Schulze-Wollgast, and H. Schumann, "3D Information Visualization for Time Dependent Data on Maps," *Proc. Int'l Conf. Information Visualization (IV),* 2005.
[44] C. Weaver, "Multidimensional Visual Analysis Using Cross-Filtered Views," *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST),* Oct. 2008.
[45] M. Weber, M. Alexa, and W. Muller, "Visualizing Time-Series on Spirals," *Proc. IEEE Symp. Information Visualization (INFOVIS '01),* pp. 7-14, Oct. 2001.
[46] H. Zhao, B. Shneiderman, and C. Plaisant, "Improving Accessibility and Usability of Geo-Referenced Statistical Data," *Proc. 2003 Ann. Nat'l Conf. Digital Govt. Research (DGO '03),* p. 1, 2003.

Ross Maciejewski received the MS degree in electrical and computer engineering from Purdue University and the BS degree from the University of Missouri, Columbia. He is a PhD student in electrical and computer engineering at Purdue University. His research interests include non-photorealistic rendering, volume rendering, and visual analytics. He is a student member of the IEEE and the IEEE Computer Society.

Stephen Rudolph received the BS degree in computer systems engineering from Arizona State University. He is a master's student in electrical computer engineering at Purdue University. His research interests include casual information visualization and visual analytics.

Ryan Hafen received the MStat degree in mathematics from the University of Utah and the BS degree in statistics from Utah State University. He is a PhD student in Statistics at Purdue University. His research interests include exploratory data analysis and visualization, massive data, computational statistics, time series, modeling, and nonparametric statistics.

Ahmad M. Abusalah received the BS degree in mechanical engineering from Jordan University of Science and Technology, the BS degree in computer sciences and information systems from Philadelphia University-Jordan, the MS degree in computer sciences from the University of Illinois, and the MS degree in computer sciences from Purdue University in 2000, 2003, 2006, and 2009, respectively. Currently, he is pursuing the PhD degree in health informatics at the University of Minnesota. He is involved in many healthcare-related applications as database administrator or engineer. For example, health surveillance systems, cancer care engineering hub, cardiac patient management system, and chemical database. His main interests are databases and scientific data management within healthcare domain. Recently, he was awarded the Graduate School Fellowship in Health Informatics from the University of Minnesota.

Mohamed Yakout is a member of Indiana Center of Database Systems (ICDS). His research areas of interests are data quality and integration, privacy preserving data integration, data mining, and digital libraries.

Mourad Ouzzani received the PhD degree in computer science from Virginia Tech. He is a research assistant professor in the Cyber Center, Discovery Park, at Purdue University. His research interests cut across databases, Web services, and Semantic Web with an emphasis on data integration, data quality, service querying and optimization, and data/service support for life sciences. He is involved in several projects for enabling discovery in life sciences, accessing biological databases using Web services, and providing database support for public health surveillance. He has published several papers in international journals and conferences including the *IEEE TKDE*, the *IEEE Internet Computing*, the *IEEE Computer*, *Bioinformatics*, VLDB, CIDR, ICDE, and SIGMOD. He was selected as an outstanding reviewer by the *IEEE Internet Computing* in 2002. He was a recipient of the USENIX scholarship in 2000.

William S. Cleveland is the Shanti S. Gupta distinguished professor of statistics and courtesy professor of computer science at Purdue University. His areas of methodological research are in statistics, machine learning, and data visualization. He has analyzed data sets ranging from very small to very large in his research in computer networking, homeland security, visual perception, environmental science, healthcare engineering, and customer opinion polling. In the course of this work, he has developed many new methods that are widely throughout engineering, science, medicine, and business; in 2002, he was selected as a highly cited researcher by the American Society for Information Science and Technology in the newly formed mathematics category. He has carried out fundamental work in data visualization. His two visualization books, *The Elements of Graphing Data* and *Visualizing Data*, are widely read and reviewed. In graphical perception, he set out basic theory and carried out many experiments. With Richard A. Becker, he developed the trellis display framework for visualization, used by a worldwide community of data analysts via its implementation in two software systems based on the S language: S-Plus (commercial) and R (open source).

Shaun J. Grannis received the BS degree in aerospace engineering from the Massachusetts Institute of Technology (MIT), the MD degree from Michigan State University, and the MS in clinical research and informatics from Indiana University. He is an assistant professor of family medicine at Indiana University and medical informatics research scientist at the Regenstrief Institute in Indianapolis, where his interests include developing, implementing, and studying technology to overcome the challenges of integrating data from distributed systems for use in healthcare delivery and research.

David S. Ebert is a professor in the School of Electrical and Computer Engineering at Purdue University, a university faculty scholar, a fellow of the IEEE and the IEEE Computer Society, the director of Purdue University Rendering and Perceptualization Lab (PURPL), and the director of Purdue University Regional Visualization and Analytics Center (PURVAC), which is a part of the Department of Homeland Security's Regional Visualization and Analytics Center of Excellence. He performs research in novel visualization techniques, visual analytics, volume rendering, information visualization, perceptually based visualization, illustrative visualization, and procedural abstraction of complex, massive data. He has been very active in the visualization community, teaching courses, presenting papers, cochairing many conference program committees, serving on the ACM SIGGRAPH Executive Committee, serving as an editor in chief of the *IEEE Transactions on Visualization and Computer Graphics*, serving as a member of the IEEE Computer Society's Publications Board, serving on the IEEE Computer Society Board of Governors, and successfully managing a large program in external funding to develop more effective methods for visually communicating information.