

# Understanding Syndromic Hotspots - A Visual Analytics Approach

Ross Maciejewski\*   Stephen Rudolph\*   Ryan Hafen\*   Ahmad Abusalah\*   Mohamed Yakout\*  
Mourad Ouzzani\*   William S. Cleveland\*   Shaun J. Grannis†   Michael Wade‡   David S. Ebert\*

\*Purdue University Regional Visualization and Analytics Center (PURVAC)

†Regenstrief Institute and Indiana University School of Medicine

‡Indiana State Department of Health

## ABSTRACT

When analyzing syndromic surveillance data, health care officials look for areas with unusually high cases of syndromes. Unfortunately, many outbreaks are difficult to detect because their signal is obscured by the statistical noise. Consequently, many detection algorithms have a high false positive rate. While many false alerts can be easily filtered by trained epidemiologists, others require health officials to drill down into the data, analyzing specific segments of the population and historical trends over time and space. Furthermore, the ability to accurately recognize meaningful patterns in the data becomes more challenging as these data sources increase in volume and complexity. To facilitate more accurate and efficient event detection, we have created a visual analytics tool that provides analysts with linked geo-spatiotemporal and statistical analytic views. We model syndromic hotspots by applying a kernel density estimation on the population sample. When an analyst selects a syndromic hotspot, temporal statistical graphs of the hotspot are created. Similarly, regions in the statistical plots may be selected to generate geospatial features specific to the current time period. Demographic filtering can then be combined to determine if certain populations are more affected than others. These tools allow analysts to perform real-time hypothesis testing and evaluation.

## 1 MOTIVATION

Recently, the detection of adverse health events has focused on pre-diagnosis 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 PHESS (Public Health Emergency Surveillance System) [9], 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 7000 records per day.

These complaints are then classified into nine categories (respiratory, gastro-intestinal, hemorrhagic, rash, fever, neurological, botulinic, shock/coma, and other) [4] 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. Figure 1 shows a typical month of emergency department visits for those 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 21st; 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.

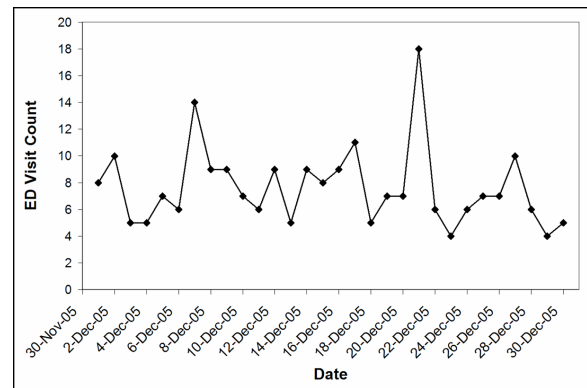


Figure 1: A sample syndromic surveillance signal containing a carbon monoxide poisoning event.

In order to facilitate enhanced syndromic surveillance and improve signal detection, we have developed a linked geo-spatiotemporal visual analytics tool designed for advanced data exploration for epidemiologists and other healthcare officials. The system was designed from its inception in collaboration with health surveillance experts, state healthcare officials and epidemiologists to address their needs. Our system features include:

- A new kernel density estimation that works for both urban and rural populations
- Dually linked interactive displays for multi-domain/multi-variate 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

Our work focuses on advanced interactive visualization and analysis methods providing linked environments of geospatial data and time series graphs. Syndromic 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 syndromic hotspots facilitated through the use of control charts for outbreak detection. Alerts generated in the temporal realm can be quickly analyzed in the geo-spatiotemporal interface, helping users find patterns simultaneously in the spatial

and temporal domains. Concurrently, we have also applied statistical modeling techniques to estimate syndrome distributions in the spatial realm. Users may select syndromic hotspots from the generated heatmaps and analyze historical time series data in the area to look for unusual trends or potential outbreaks. Such doubly linked views allow users to quickly form and test hypotheses, thereby reducing the time needed to reject false positives and confirm true outbreaks.

## 2 PREVIOUS WORK

Data from public health surveillance systems has long been recognized as providing meaningful measures for disease risks in populations [12, 23]. 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) [10], the Electronic Surveillance System for the Early Notification of Community based Epidemics ESSENCE [14], and Biosense [15]. 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.

Our work employs a variety of methods from information visualization and geographical visualization as a basis for creating an advanced visualization and analytics environment. This section briefly describes components from related work and addresses their applicability to syndromic surveillance.

A key component that must be addressed in syndromic surveillance data is the geo-spatiotemporal nature of the data. 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. [17] 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. Schulze-Wollgast et al. [25] 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 [26] employed the use of intuitive 3D pencil and helix icons for visualizing multiple dependent data attributes and emphasizing the type of underlying temporal dependency.

Many of these previous systems provided useful visualization and exploration of data, but did not support interactive analysis. To address this gap, visual analytics has emerged as a relatively new field formed at the intersection of analytical reasoning and interactive visual interfaces [24]. 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 we are most concerned for this paper is that of geospatial and temporal analytics, which applies the concepts of visual analytics to problems rooted in space and time.

A few examples of recent work in the spatiotemporal branch

of visual analytics include VIST-STAMP by Liao et al. [13], FemaRepViz by Pan and Mitra [20], and LAHVA by Maciejewski et al. [18]. VIS-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 datasets (pet and human health data) with similar properties to enhance disease surveillance. This system provided a geo-spatiotemporal interface with limited interaction amongst different view windows.

## 3 VISUAL ANALYTIC ENVIRONMENT

Our system adopts the common method of displaying geo-referenced data 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 is that aggregation of data can lead to unreliable estimates of the true measure of infection. Fortunately, the PHESS data used in our visual analytics system provides geo-referenced patient locations, allowing us to either aggregate the data on a spatial level, or employ statistical methods to model the data over arbitrarily sized geo-regions. As such, our system employs advanced statistical models for data exploration, enabling new visualizations, analyses, and enhanced detection methods.

### 3.1 System Features

Figure 2 (Left) provides a conceptual overview of our visual analytics system, and Figure 2 (Right) provides a screenshot of the system. Data entering our system first undergoes a cleaning and transformation process. This process fills in missing patient information from past visit information, normally distributes patients with unknown addresses to locations within their county, and aggregates patient visits that occur with similar syndromes multiple times on a given day. 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 Figure 2 (Right), the main viewing area is the geo-spatiotemporal view, and the three windows on the right allow users to view a variety of data sources simultaneously for a quick comparison of trends across varying hospitals or data aggregated over spatial regions. Both the geospatial and time series viewing windows are linked to the time slider at the lower portion 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 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 syndrome filtering. Users interactively generate database

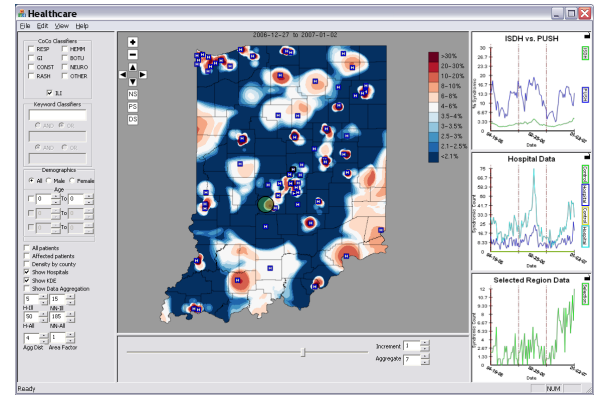
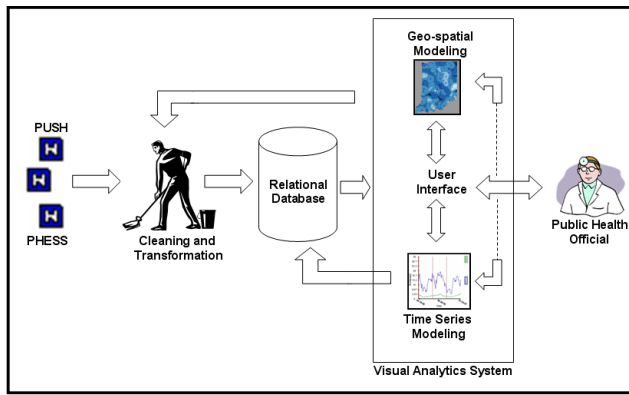


Figure 2: The visual analytics system. (Left) The conceptual diagram of our visual analytics system. Observe the interaction between the analyst and the system as well as the modeling components of the system. (Right) Our visual analytics system. The left portion of the screen represents the interactive database querying tools. We include checkboxes for classified syndromes, keyword searches for chief complaint text, and demographic filtering for age and gender. The main viewing area is a geo-spatial temporal 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. Users may select points or regions of time to interactively manipulate the geo-spatial temporal window. Finally, a time slider is included on the bottom portion of the screen, allowing users to move through time on all unlocked screens.

search queries through the use of check boxes and edit controls to find specific syndromes, keywords, and gender and age demographics amongst patients. Such work furthers hypothesis testing as users can now quickly filter signals by demographic constraints in order to see if adverse health conditions are targeting a particular segment of the population. The choices of filters affect both the geo-spatiotemporal 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 patient locations on the map in order to provide health officials with a quick overview of health statistics across the state. Unfortunately, showing exact patient 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 health officials to simply see a plot of patient 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 patients mapped on a transformed geolocation to pixel basis. Unfortunately, as the data set becomes arbitrarily large, the visual clutter can not be reduced in such a manner, see Figure 3 (Top-Left), and it becomes clear that a visualization of every patient record at a high spatial zoom level is not effective for analysis. Furthermore, simple methods, such as using additive opacity to demonstrate patient density, Figure 3 (Top-Left), are inadequate as the number of patients makes it impossible to readily distinguish density levels between areas. This is further complicated when the syndromic patients are then highlighted with regard to their locations. Figure 3 (Top-Middle) 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 patient 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 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 [5]. The generated circles are then colored using a sequential colormap [3] where the color represents the percent of patients with a given syndrome found within this geographical centroid. This method operates under the assumption that the data is 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.

Figure 3 (Top-Right) shows the low resolution aggregation of our data across the state of Indiana. Figure 3 (Bottom-Left) shows the zoomed in region, and Figure 3 (Bottom-Right) 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 patient distribution, it is also useful to model the population distribution across the state in order to approximate trends where little or no data values exist. Therefore, our system provides a geospatial heatmap [7] view which employs a diverging color map [3] to represent the percentage of a given syndrome over the total patients seen on a given day.

As previously discussed, the healthcare data provided by PHESS contains a set of observations in which an individual from location  $X_i$  arrives at time  $t$  to a hospital and is diagnosed with a particular syndrome. Such data is 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 patient visits. Furthermore, areas with a small number of patients may stray towards a high percentage of the population seen reporting the syndrome in question. In those cases, visual alerts may be triggered that would

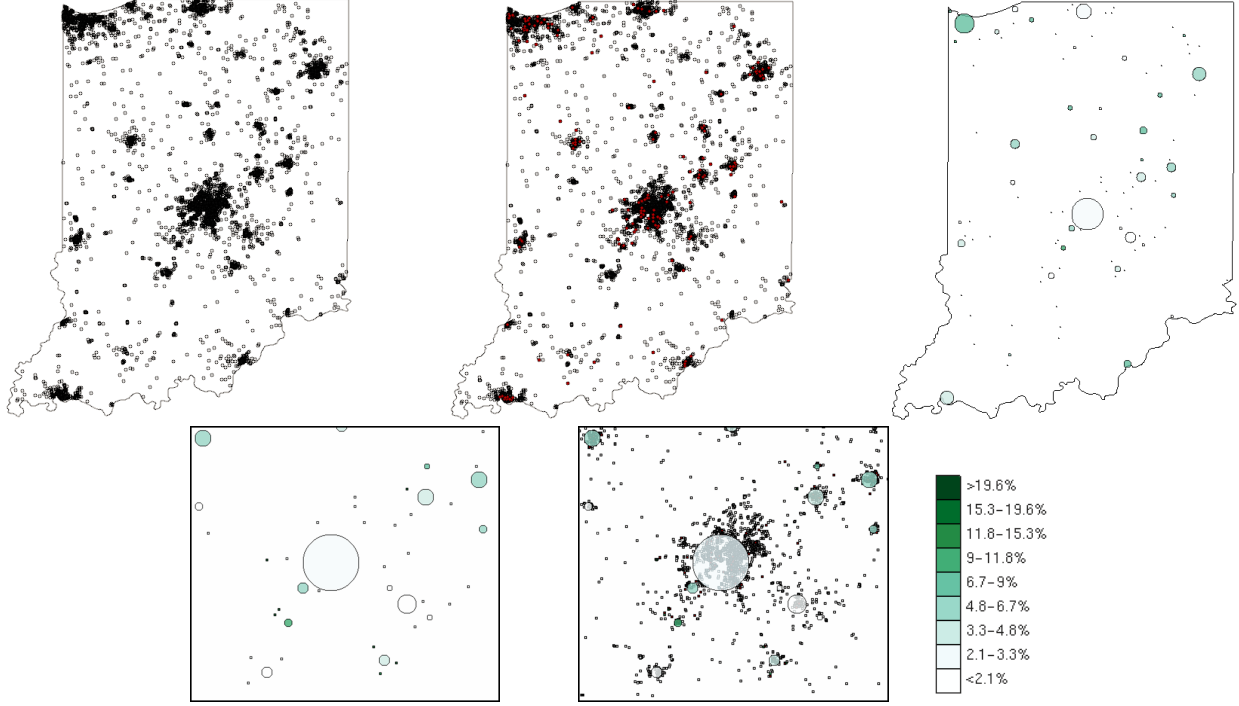


Figure 3: Data aggregation and privacy preservation. (Top-Left) Georeferenced patient data as small additive opacity circles. (Top-Middle) Georeferenced patient data overlaid with red circles representing syndromic patients. (Top-Right) Data aggregation for enhanced visualization. (Bottom-Left) High-resolution zoom of an area of interest. (Bottom-Right) Actual patient locations at a high-resolution zoom overlaid with our data aggregation method.

clearly appear as false positives once the individual records were analyzed. Figure 4 (Left) demonstrates the problems with visualizing such histogram distributions. The national baseline influenza-like-illness (ILI) percentage during flu season is 2.1% [1] for the 2006-07 season. Note in Figure 4 (Left) 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.

To overcome these issues, our system estimates the probability density function of the entire population using the known patient locations and produces a heatmap visualization of the entire state. To this end, we employ a kernel density estimation [22]. Equation 1 defines the multivariate kernel density estimation, and this method has been used in other works [11, 16, 8]. To reduce the calculation time, we have chosen to employ the Epanechnikov kernel, Equation 2.

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

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

Here,  $\mathbf{h}$  represents the multi-dimensional 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 ill patients as well as the density estimation of all patients that visited a hospital in our system using an appropriately chosen  $h$  for each data set. Density estimation for the ill patients is done only in two dimensions for the given time period aggregation. Density estimation for the total patients is done both spatially and temporally over

the last seven aggregate time periods. The density estimation for the ill patients is then divided by the density estimation for the total patients to provide a percentage count for the expected number ill of the population.

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 (as shown in Figure 4 (Middle)) 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 as Figure 4 (Left).

To overcome these issues, we employ the use of a variable kernel method [22], Equation 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  $k$ th nearest neighbor in the set comprising  $N - 1$  points.

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

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  $i$ th sample to the  $k$ th nearest neighbor) so that data points in regions where the data is sparse will have flatter kernels. Unfortunately, our data set also exhibits problems with this method. In health care data, a primary recipient of emergency care are patients of long-term health care facilities (for example, nursing homes). As such, the use of the  $k$  nearest neighbors may result in a  $d_{i,k}$  of 1 as many patients visiting emergency rooms may report the same address. This concept can be extended to large apartment complexes, as well as data uncertainty (for example, many hospitals report unknown patient addresses as the hospital address). To overcome these issues, we slightly modify

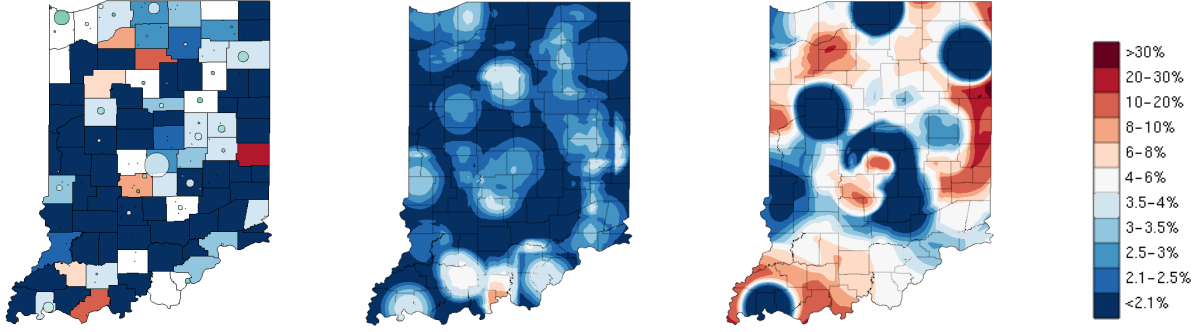


Figure 4: A variety of heatmaps. (Left) Heatmap showing percentage of patients with ILI aggregated by county and overlaid with the patient data aggregation. (Middle) Heatmap generated using a fixed bandwidth kernel density estimation. (Right) Heatmap generated using our modified variable kernel density estimation.

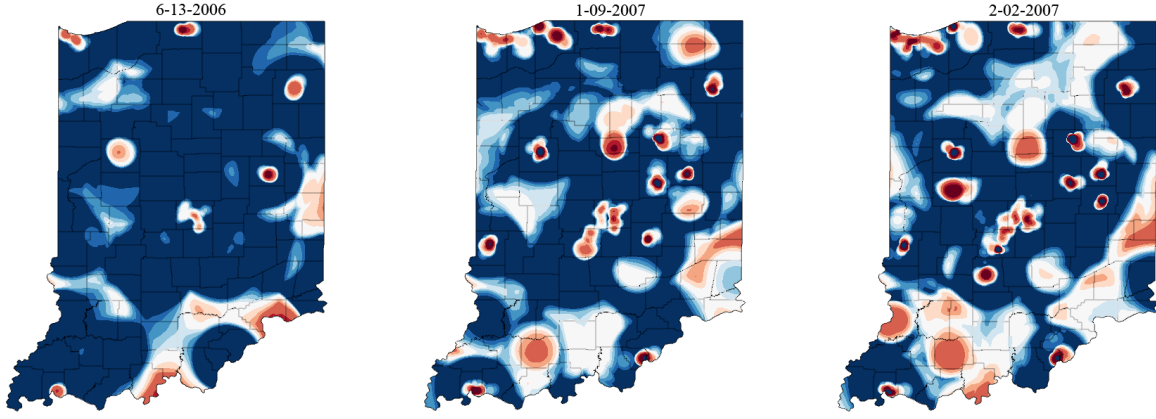


Figure 5: Heatmap plots of the 2006-2007 flu season. Time increments from left to right with exact time periods shown on each figure.

the variable kernel estimation to force it to have a minimum fixed bandwidth of  $h$  as shown in Equation 4.

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

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 Figure 4 (Right). Slight problems in the estimation can be found near the state borders due to the abrupt cut of data in those areas. Future work will address these issues through more advanced spatial modeling. Presently, we feel that our results show that our modified variable kernel method provides a better estimate than either the histogram approximation of Figure 4 (Left) or the fixed bandwidth kernel of Figure 4 (Middle).

Figure 5 further demonstrates our heatmap ability, showing a progressive set of geospatial time series snap shots from our system. Note that in the summer months the estimated statewide level of influenza falls into the lower regions of our scale. As we move into winter and towards flu season, we can visually see the percent of ILI across the state growing.

We find that in localized areas near the hospitals, the %ILI shows clear peaks of ILI, most likely due to geo-coding errors in those hospitals. Note in June that the predominant colors of the state

are the two darker blue colors showing that Indiana is residing at approximately the national average or below. As we move into December, the predominant color shifts to light blue and white, indicating flu season has started as the state exceeds the baseline percentages, moving towards the 3-8% range. Our future work in this area will address uncertainty visualization in density estimation as part of a more effective analysis.

### 3.4 Time Series Analysis

While the spatial visualizations employed in our system are useful for detecting syndromic 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) [10]. 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 spatiotemporal 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,  $\mu_0$  is the expected value,  $\sigma_{x_t}$  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,

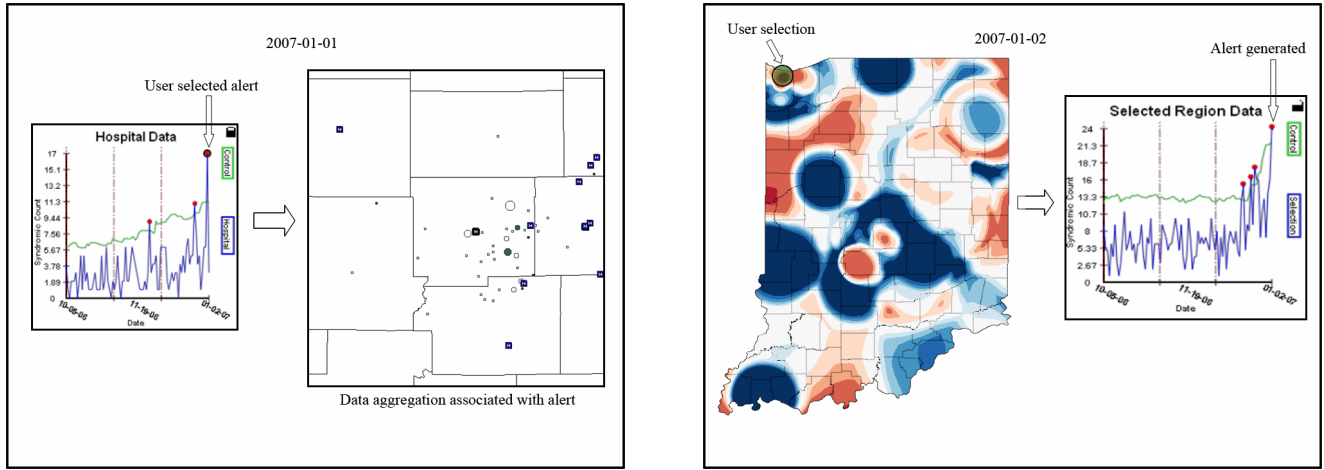


Figure 6: Exploration using linked views. (Left) Images taken from our system illustrating the linked temporal analysis to the geospatial filtering. Here, a user has selected the alert occurring on 01-01-2007. The geospatial viewing window then opens that day's data corresponding to the alert allowing for further investigation. (Right) 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.

$\mu_0$ , and standard deviation,  $\sigma_{x_i}$ , 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. Such a lag is used to increase sensitivity to continued outbreaks. Figure 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 Figure 6).

### 3.5 Exploration with Linked Views

While the alerts provided from aberration detection algorithms may provide a useful starting point for exploration, they may also be providing false alarms. Furthermore, epidemiologists 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, epidemiologists 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* [2] and it is used in many interactive visualization environments [19, 21].

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 Figure 6 (Left), we see a series of hospital generated alerts 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 patients who are associated with that hospital and syndrome are now exclusively shown on the map.

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 Figure 6 (Right), 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.

## 4 UNDERSTANDING HOTSPOTS

By using a combination of geospatial and temporal visualization and analytics tools, our system provides epidemiologists with tools for real-time hypothesis testing. To better illustrate the hypothesis testing phase, we conducted an informal interview with an Indiana State Department of Health (ISDH) syndromic surveillance epidemiologist. During this interview, we discussed how he searches 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 his experience, certain alerts will be immediately resolved as false positives, and others will be moved to the top of the queue. From the alerts he identifies as potential problems, a hypothesis is formulated 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

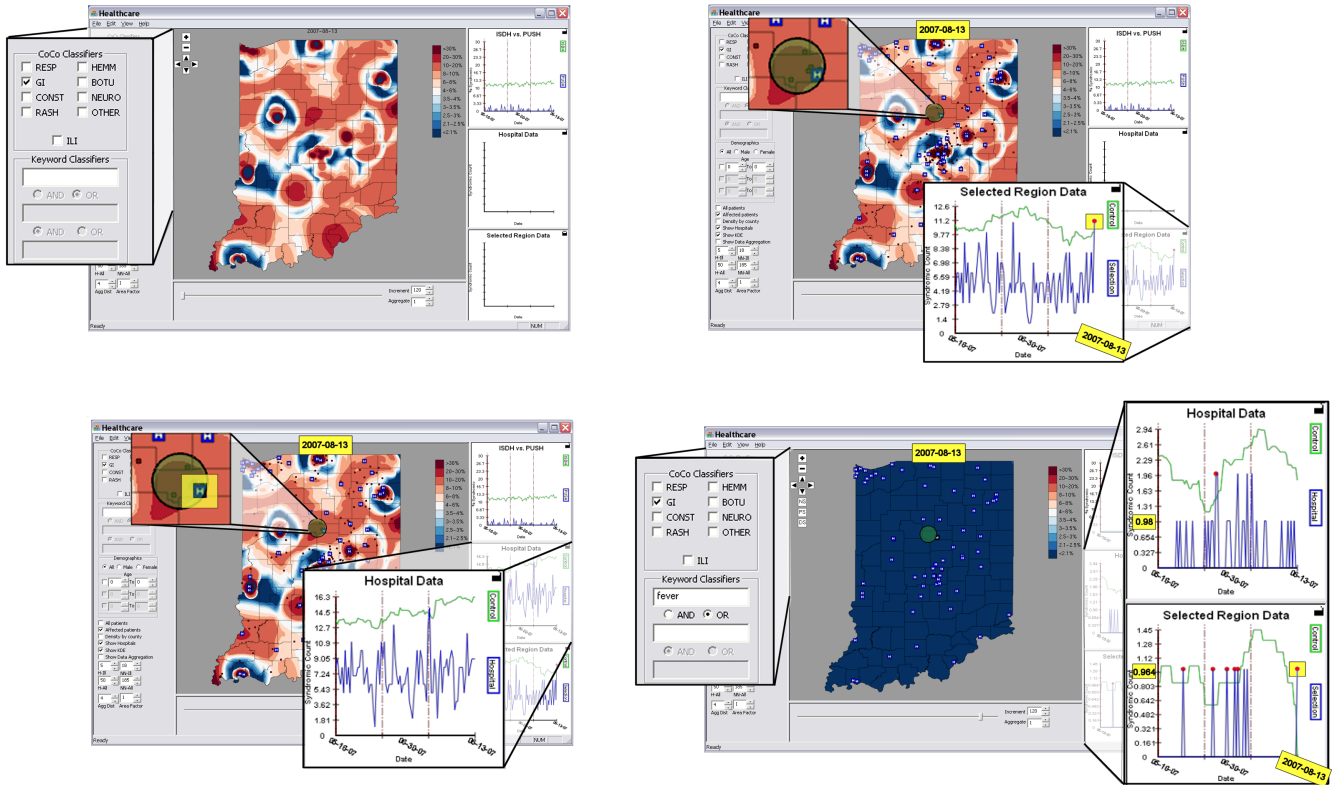


Figure 7: Using visual analytics for hypothesis testing in syndromic surveillance. (Top-Left) The user observe a heatmap for a given syndrome, in this case, gastro-intestinal. (Top-Right) Next, the user selects an area of interest, generating a time series plot for that region. Note that in the time series plot generated, an alert is occurring on the day of interest. (Bottom-Left) 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 gastro-intestinal syndromes. (Bottom-Right) 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.

can not 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 gastro-intestinal 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 timestamps 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 he 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 testing.

During this process, he also searches for potential "co-syndromes" in the same geography, such as fever, to see if it is somehow linked to the gastro-intestinal problem. Again, the linked views and filter options of our system allow the user to easily look at multi-variate 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. Figure 7 illustrates the use of our system during the hypothesis testing phase.

First, in Figure 7 (Top-Left), the user has selected the syndrome he/she is interested in analyzing, in this case, gastro-intestinal. 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 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 Figure 7 (Top-Right), 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 Figure 7 (Bottom-Left). 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 gastro-intestinal syndromes that also reported signs of fever. Figure 7 (Bottom-Right) 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 Figure 1 contains only three emergency department complaints. Therefore, the high sensitivity is necessary to avoid missing small cluster cases.

## 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 epidemiologists. Our initial results show the benefits of linking traditional time-series epidemiological views with geo-spatiotemporal 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 syndromic detection.

Unfortunately, database query performance currently reduces the interactivity of several functions of our system, most notably the keyword filtering. These queries can take seconds to minutes depending on the length of the time series being visualized. We plan to enhance our database system to optimize keyword queries on our ten million record database and achieve interactive system rates for all components in the near future.

Other future work includes advanced modeling of geo-spatiotemporal 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 for syndromic event detection as well as correlative analysis views within the temporal domain. Once these features are implemented, we plan to deploy our system with our state health partners for further evaluation.

## 6 ACKNOWLEDGMENTS

The authors would like to thank the Purdue University Student Health Center and the Indiana State Department of Health 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 0328984 and 0121288.

## REFERENCES

- [1] Update: Influenza activity — United States and Worldwide, 2007–08 season, and composition of the 2007–08 influenza vaccine. *MMWR Morb Mortal Wkly Rep*, 56:789–794, 2007.
- [2] R. A. Becker and W. S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.
- [3] C. A. Brewer. *Designing better Maps: A Guide for GIS users*. ESRI Press, 2005.
- [4] 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*, 46:445–455, November 2005.
- [5] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, 2001.
- [6] 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. In *INFOVIS '01: Proceedings of the IEEE Symposium on Information Visualization 2001*

- (*INFOVIS '01*), page 159, Washington, DC, USA, 2001. IEEE Computer Society.
- [7] U. Fayyad, G. G. Grinstein, and A. Wierse, editors. *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002.
- [8] M. Gibin, P. Longley, and P. Atkinson. Kernel density estimation and percent volume contours in general practice catchment area analysis in urban areas. In *Geographical information science research conference*, 2007.
- [9] 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. In *American Medical Informatics Association*, 2006.
- [10] L. C. Hutwagner, W. W. Thompson, and G. M. Seeman. The bioterrorism preparedness and response early aberration reporting system (ears). *Journal of Urban Health*, 80(2):i89 – i96, 2003.
- [11] D. Kao, A. Luo, J. L. Dungan, and A. Pang. Visualizing spatially varying distribution data. In *Proceedings of the sixth international conference on information visualization*, pages 219–225, 2002.
- [12] A. D. Langmuir. The surveillance of communicable diseases of national importance. *New England Journal of Medicine*, 268:182 – 192, 1963.
- [13] K. Liao. A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1461–1474, 2006. Member-Diansheng Guo and Student Member-Jin Chen and Member-Alan M. MacEachren.
- [14] J. S. Lombardo. A systems overview of the electronic surveillance system for the early notification of community based epidemics (ESSENCE II). *Journal of Urban Health*, 80:32 – 42, 2003.
- [15] J. W. Loonsk. Biosense - a national initiative for early detection and quantification of public health emergencies. *MMWR*, 53:53 – 55, 2004.
- [16] A. L. Love, A. Pang, and D. L. Kao. Visualizing spatial multivariate data. *IEEE Comput. Graph. Appl.*, 25(3):69–79, 2005.
- [17] A. M. MacEachren, F. P. Boscoe, D. Haug, and L. Pickle. Geographic visualization: Designing manipulable maps for exploring temporally varying georeferenced statistics. In *INFOVIS '98: Proceedings of the 1998 IEEE Symposium on Information Visualization*, page 87, Washington, DC, USA, 1998. IEEE Computer Society.
- [18] 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. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, October 2007.
- [19] A. R. Martin and M. O. Ward. High dimensional brushing for interactive exploration of multivariate data. In *VIS '95: Proceedings of the 6th conference on Visualization '95*, page 271, Washington, DC, USA, 1995. IEEE Computer Society.
- [20] C.-C. Pan and P. Mitra. Femarepviz: Automatic extraction and geo-temporal visualization of fema national situation updates. *Visual Analytics Science and Technology*, 2007. VAST 2007. *IEEE Symposium on*, pages 11–18, Oct. 30 2007–Nov. 1 2007.
- [21] J. C. Roberts and M. A. E. Wright. Towards ubiquitous brushing for information visualization. In *IV '06: Proceedings of the conference on Information Visualization*, pages 151–156, Washington, DC, USA, 2006. IEEE Computer Society.
- [22] B. W. Silverman. *Density Estimation for Statistica and Data Analysis*. Chapman & Hall/CRC, 1986.
- [23] S. B. Thacker, R. L. Berkelman, and D. F. Stroup. The science of public health surveillance. *Journal of Public Health Policy*, 10:187 – 203, 1989.
- [24] J. J. Thomas and K. A. Cook, editors. *Illuminating the Path: The R&D Agenda for Visual Analytics*. IEEE Press, 2005.
- [25] C. Tominski, P. Schulze-Wollgast, and H. Schumann. Visual analysis of human health data. In *2003 IRMA International Conference*, 2003.
- [26] C. Tominski, P. Schulze-Wollgast, and H. Schumann. 3d information visualization for time dependent data on maps. In *International Conference on Information Visualization (IV)*, 2005.