Contextualizing Syndromic Hotspots - A Visual Analytics Approach

Ross Maciejewski Stephen Rudolph George Tebbetts
David S. Ebert

Purdue University Regional Visualization and Analytics Center (PURVAC)

ABSTRACT

In analyzing syndromic surveillance data, epidemiologists are faced with various data restrictions due to issues with both data collection and privacy preservation. Often time, the data must be viewed as an aggregate of spatial components. When aggregating data by regions, both sparse and dense population regions can create confusing mappings. Aggregation over sparsely populated regions may result in visualizations that show the area as having a high concentration of patients with a certain syndrome, while aggregation over a densely populated region may lose specificity as the signal is drowned in noise. As such, we have created a suite of visual analytic tools which can be used to view aggregate syndromic data in terms of their spatial and temporal contexts. Our system provides a kernel density estimate of spatially distributed syndromic data, creating a color map of the estimated percentage of the population with a specified syndrome. In order to map this distribution to a context an epidemiologist desires, we provide both automatic and interactive range adjustment tools which allows users to adjust the data color-mapping into their model assumptions for what is abnormal in a given syndrome. Our tool set also includes hotspot selection in which users may further refine their color range selection to look for peaks within the hotspots. Furthermore, our selection tool also provides direct links to time series views of the data, providing a historical context for the data. Such tools allow epidemiologists to quickly understand the context in which data is being presented thereby improving their ability to form and test complex hypotheses.

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) [5], 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) [2] and used as indicators to detect public health emergencies before such an event is confirmed by diagnoses or overt activity.

The work presented in this paper focuses on advanced interactive visualization and analysis methods for contextualizing hotspots. A screen shot of this system is shown in Figure 1. 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. We apply statistical modeling techniques to estimate syndrome distributions in the spatial realm. Spatially located syndromic hotspots can be selected and immediately analyzed in the corresponding linked temporal view. Further, through a data aggregation method, patient distributions can be overlaid on the density distribution to provide hints as to which areas should be further explored, and which areas may simply be false positives. Concurrently, user's may also interactively explore the data ranges through color mapping tools in order to explore the varying dimensionalities of the hotspot under analysis. Such tools allow users to quickly form and test hypotheses, thereby reducing the time needed to reject false positives and confirm true outbreaks.

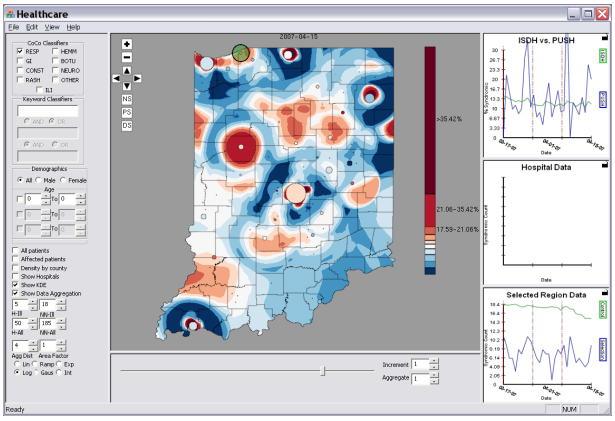

Figure 1: A screen shot of the visual analytics system.

To summarize, novel system features include:

- A new kernel density estimation that works for both urban and rural populations
- Dually linked interactive displays for multi-domain/multivariate exploration and analysis
- Novel data aggregation for effective visualization and privacy preservation
- Interactive color mapping tools for enhanced data contextualization
- Region selection tools for analyzing area specific hotspots

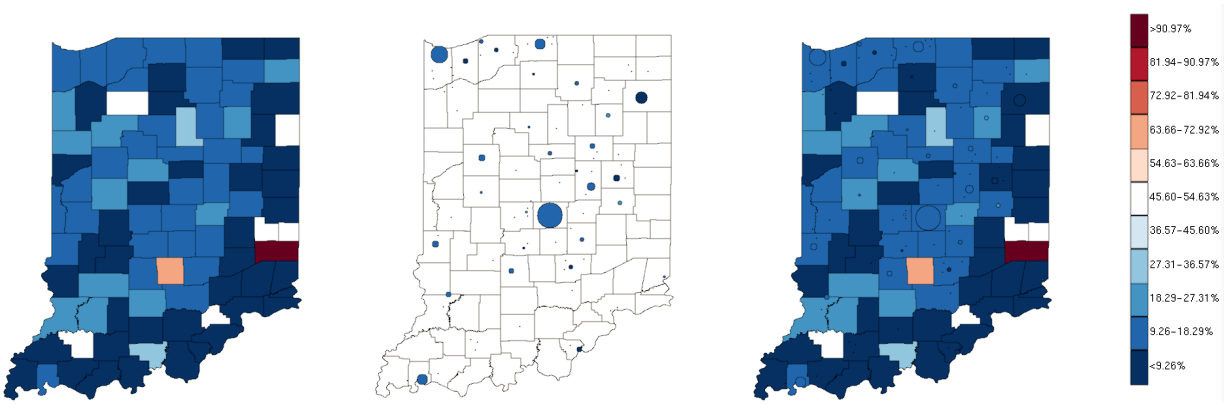

Figure 2: Data aggregation and privacy preservation visualized as a percentage of syndromic population over the total population seen. (Left) Data aggregated by county. (Middle) Data aggregated through nearest neighbor groupings. (Right) A combination of data aggregation to enhance contextual visualization.

2 Previous Work

Data from public health surveillance systems has long been recognized as providing meaningful measures for disease risks in populations [7, 11]. 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) [6], the Electronic Surveillance System for the Early Notification of Community based Epidemics ESSENCE [8], and Biosense [9]. 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.

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 [12]. 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.

3 Hotspots in Context

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 clearly appear as false positives once the individual records were analyzed. In Figure 2 (Left) we present a a geospatial heatmap [4] view which employs a diverging color map [1] to represent the percentage of a given syndrome over the total patients seen on a given day. Notice that many counties seem to be visually displaying an extremely high level of respiratory syndromes; however, without other spatial or temporal contextual information it is difficult to ascertain if these areas should be investigated or if instead they should be disregarded as false positives. In this section we present several methods for contextualizing the data to enhance hypothesis generation and testing for syndromic surveillance.

3.1 Data Aggregation and Privacy Preservation

Along with aggregation by county, we also employ a data aggregation method which clusters patient location by their nearest neighbors. 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 [3]. The generated circles are then colored using the same divergent colormap [1] as the counties where the color represents the percent of patients with a given syndrome found within this geographical centroid, see Figure 2. 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.

Notice that when the data is aggregated by county, several counties immediately stand out to the user, indicating that these areas warrant investigation. However, by overlaying the data aggregated circles on the county level aggregation, we are able to represent a larger set of contextual information. In (Figure 2 (Right) we can now see that the counties showing high levels of respiratory syndromes have such a sparse distribution of patients that no aggregation levels were displayed on the map. As such, one may begin hypothesizing that these counties are merely representing false positives. Unfortunately, in the case of small and sparse populations, such levels of data aggregation often fail to represent the vital information needed. Furthermore, users could also begin looking for counties that show low levels of syndrome where their aggregate shows high levels, indicating that there is a conflict of information that needs to be further investigated.

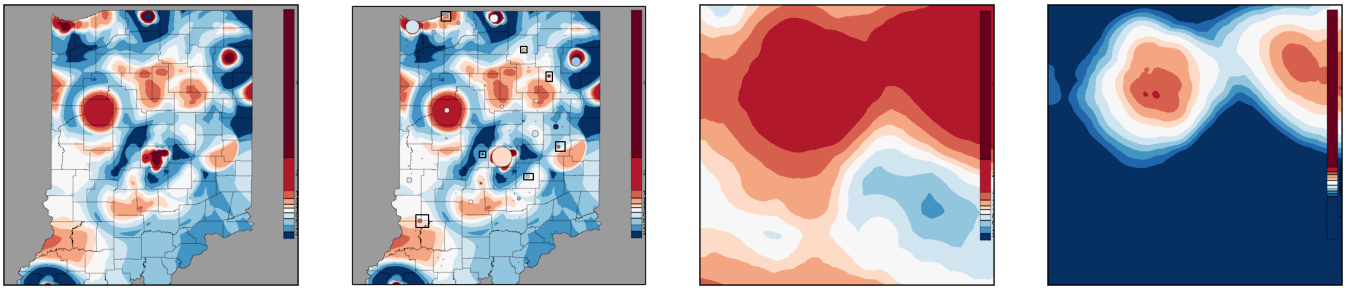

Figure 3: Kernel density estimate heatmaps visualized as a percentage of syndromic population over the total population seen. (Left) KDE heatmap. (Middle-Left) Contextualizing the KDE heatmap by overlaying patient data aggregated through nearest neighbor groupings. (Middle-Right) A zoomed in view of a local hotspot. (Right) Contextualizing a hotspot through interactive coloring.

3.2 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. To accomplish this, we employ the use of a variable kernel density estimation method [10], Equation 1. 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. 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 k for each data set. Density estimation is done only in two dimensions for the given time period aggregation. 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.

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

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

Here, N is the total number of samples and the function $1_{(||\mathbf{u}||\leq 1)}$ evaluates to 1 if the inequality is true and zero for all other cases.

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 is sparse will have flatter kernels. Unfortunately, our data set 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 the variable kernel estimation to force it to have a minimum fixed bandwidth of h as shown in Equation 3.

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

Here, $\mathbf{h}$ represents the multi-dimensional smoothing parameter. In the case of our modified variable kernel estimation, we calculate the kernel only spatially as opposed to both spatially and temporally. Future work will include extending our modified density estimation into the temporal domain. Results from our variable kernel estimation can be seen in Figure 3. 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.

3.3 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 syndromic patients within an area. However, each syndrome will have varying model assumptions. Furthermore, the distribution of the data can also play a key role in placing syndromic hotspots into the proper context. For example, if the data is 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 curve $y = x$. This idea is then extended for both exponential and logarithmic curves. Future work will include binning the data to a Gaussian distribution.

In Figure 3, the data has 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 (Figure 3 (Middle-Left)), we begin to develop hypotheses of key places that need further exploration. These places are marked by the black squares in Figure 3 (Middle-Left).

Further, we see the dense hotspot centered in the middle of the state. To further explore this hotspot, user's may zoom into the map. The zoom results in a re-calculation of the kernel density estimate as the latitude/longitude point space mapping to the grid changes. Figure 3 (Middle-Right) provides a zoomed in view of the state's central hotspot. Notice that this heatmap is dominated by the a singular range of red. In Figure 3 (Right) 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.

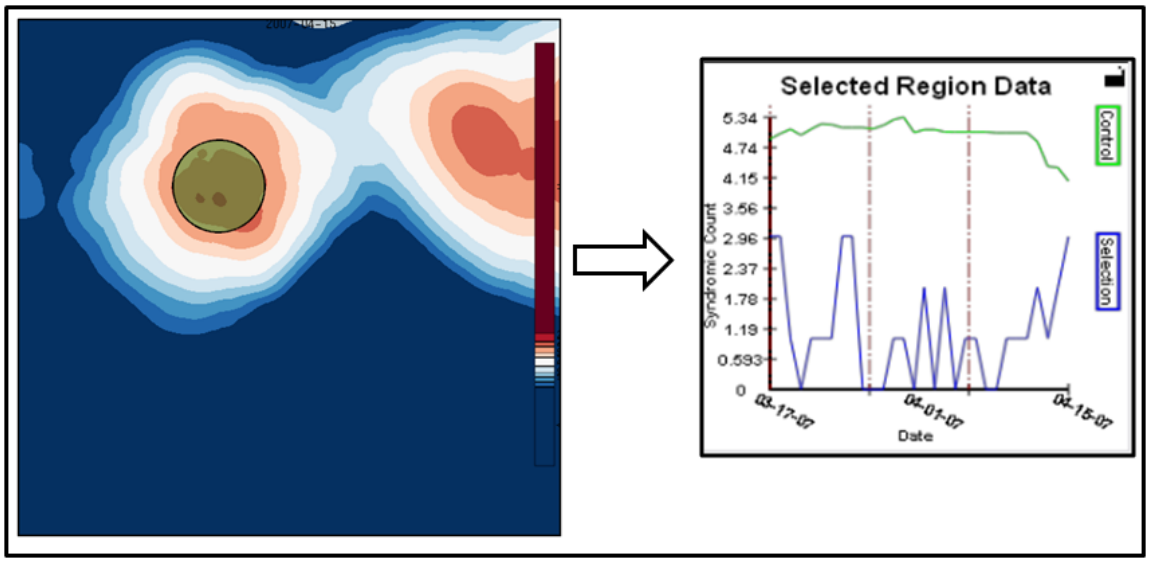

Figure 4: Contextualizing a hotspot by its temporal history. Here, a user has selected an area (the green circle) and the temporal history of that spatial region is displayed in the graph.

3.4 Temporal Context

The previous tools focus on viewing data in a spatial context; however, syndromic surveillance data is directly keyed to temporal aspects as well. To present relevant linked spatio-temporal context, our system allows user highlighting in the geospatial view through a circular selection of an area. This circular selection allows users to select multiple geographic regions and view their temporal history. In Figure 4, we see a heat map 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. Future work will allow the selection of arbitrarily shaped hotspots based on the kernel density estimation shape.

4 Conclusions and Future Work

Our current work demonstrates the benefits of visual analytics for contextualizing 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 multiple overlays of similar data, linking traditional time-series epidemiological views with geo-spatiotemporal views, and interactive color maps for 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.

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.

5 Acknowledgments

The authors would like to thank the Purdue University Student Health Center and the Indiana State Department of Health for providing the data.

References

[1] C. A. Brewer. *Designing better Maps: A Guide for GIS users*. ESRI Press, 2005.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] A. D. Langmuir. The surveillance of communicable diseases of national importance. *New England Journal of Medicine*, 268:182 – 192, 1963.

[8] 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.

[9] J. W. Loonsk. Biosense - a national initiative for early detection and quantification of public health emergencies. *MMWR*, 53:53 – 55, 2004.

[10] B. W. Silverman. *Density Estimation for Statistica and Data Analysis*. Chapman & Hall/CRC, 1986.

[11] 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.

[12] J. J. Thomas and K. A. Cook, editors. *Illuminating the Path: The R&D Agenda for Visual Analytics*. IEEE Press, 2005.