Forecasting Hotspots - A Predictive Visual Analytics Approach

Ross Maciejewski [*] Ryan Hafen[*] Stephen Rudolph[*] William S. Cleveland[*]
David S. Ebert[*]

[*]Purdue University Regional Visualization and Analytics Center (PURVAC)

Abstract

Current visual analytics systems provide users with the means to explore trends amongst their data. Linked views and interactive displays provide insight into correlations between space, time, events, people and places. Analysts search for events of interest through statistical tools linked to visual displays, drill down into the data, and form hypotheses based upon the available information. However, current systems stop short of predicting events. In spatiotemporal data, analysts are searching for regions of space and time with unusually high incidences of events (hotspots). In the cases that hotspots are found, analysts would like to predict how these regions may grow in order to plan decision support and preventative measures. Furthermore, analysts would also like to predict where future hotspots may occur. To facilitate such forecasting, we have created a predictive visual analytics toolkit that provides analysts with linked geo-spatiotemporal and statistical analytic views. Our system models spatiotemporal events through the combination of kernel density estimation for event distribution and seasonal trend decomposition by loess smoothing for temporal predictions. We provide analysts with estimates of error in our temporal modeling, along with temporal alerts to indicate the occurrence of hotspots. Spatial data is distributed based on a modeling of previous event locations, thereby maintaining a temporal coherence with past events. Such tools allow analysts to perform real-time hypothesis testing, plan intervention strategies, and allocate resources to correspond to perceived threats.

Keywords: Predictive analytics, visual analytics, syndromic surveillance.

1 Motivation

Visual analytics has been defined as the science of analytical reasoning assisted by interactive visual interfaces [29]. Recently, visual analytics systems (e.g., [23, 27, 31]) have been developed that allow users to interactively explore their data through linked windows, temporal histories, document aggregations and numerous other views. Such systems allow users to find correlations between events and begin forming hypotheses about what events may be occurring in the future; however, the primary use of such systems tends to be reactive, meaning that analytic systems are typically used in the context of alert generation. As event data is captured, algorithms and analysts search for unexpected events, and these unexpected events then trigger an alert. Analysts react by drilling down into the data to confirm the alert, redistributing resources to control the problem, etc. Unfortunately, in a reactive situation, events have already occurred that are negatively affecting the population under analysis. In this work, we propose the addition of a suite of predictive analytics tools as a means of enhancing current analysis systems, thereby moving from a solely reactive paradigm to a proactive paradigm.

Our predictive analytics system focuses on categorical geo-spatiotemporal event data (e.g., financial data, crime reports, emergency department logs). In such data, events consist of locations in time and/or space, and each event fits into a hierarchical categorization structure. These categories can be filtered by linked data (e.g., demographic information), and the events may be mapped to a particular spatial location. Data categories are typically processed as either time series aggregated over some spatial location (e.g., county, zip code, collection station), or spatial snapshots of a small time aggregate (e.g., day, week). These aggregations are then analyzed for counts that exceed the expected value by some threshold. There are also systems that allow for spatiotemporal alert detection (e.g., [17]), but such systems become intractable as the data set becomes large. As previously stated, these types of alert systems force analysts into a reactive paradigm. As such, tools are needed that not only perform these alert calculations based on current events, but also perform alert calculations on predicted data.

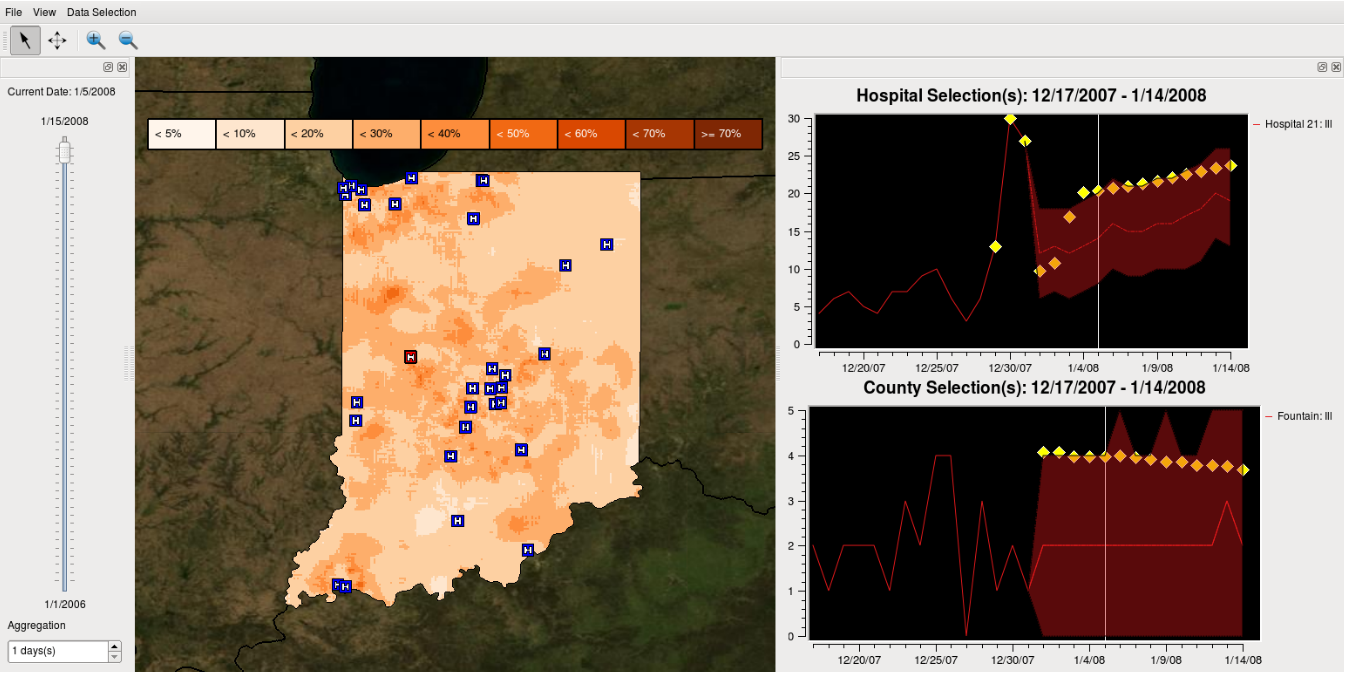
To this end, we have developed a series of novel predictive analytics tools. We base our extensions on the framework developed by Maciejewski et al. [23, 24] to move from an understanding of spatiotemporal alerts (hotspots) to a predictive modeling of such alerts; however, it is important to note that such methods are readily transferable to any similar data analysis systems. Novel system features include the following:

- The application of seasonal trend decomposition by loess for time series prediction
- Scalable 3D kernel density estimation for spatiotemporal prediction to maintain temporal coherence
- Multiple spatial aggregation schemes for hotspot analysis and forecasting
- Linked spatial and temporal views for analysis and forecasting

In order to demonstrate the impact of such tools, we focus our discussion on a representative categorical geo-spatiotemporal data set, syndromic surveillance data. Syndromic surveillance is an area of healthcare monitoring that focuses on the detection of adverse health events using pre-diagnosis information from emergency departments. Such data has long been recognized as providing meaningful measures for disease risks in populations [18, 28], and provides a solid base for discussing the impact of our methods. In this paper, we utilize data provided by the Indiana State Department of Health (ISDH) through their Public Health Emergency Surveillance System (PHESS) [13], which provides electronically transmitted patient data (in the form of Emergency Department *chief complaints*) from 77 hospitals around the state at an average rate of 7500 records per day. These complaints are classified into nine categories (respiratory, gastro-intestinal, hemorrhagic, rash, fever, neurological, botulinic, shock/coma, and other) [6] and used as indicators to detect public health emergencies before such an event is confirmed by diagnosis or overt activity. Further, to demonstrate our event detection and prediction methods, we employ the synthetic disease injection tools developed by Maciejewski et al. [22].

Our work focuses on advanced interactive visualization and analysis methods providing linked environments of geosaptial data and

[*]e-mail: {rmacieje|rhafen|srudolph|wsc|ebertd}@purdue.edu



Current Date: 1/5/2008

Aggregation: 1 days(s)

Figure 1: Our interactive predictive visual analytics environment. In this case, the interface has been customized for exploring syndromic surveillance data. The user is analyzing respiratory syndrome counts across the state at a county level aggregation. In the time series window yellow diamonds indicate temporal alerts, the white line represents the current day, and the transparent polygon represents the bounds of the time series prediction. Note that predicted values are only displayed when data is not available.

time series graphs. Time series events are forecasted for a range of spatial aggregations, providing a context in which to explore potential future events. Time series alerts are generated for current and predicted event thresholds, and analysts can explore future event bounds for resource management and response scenarios. Alerts generated in the temporal realm can be quickly analyzed in the geospatiotemporal interface, helping users find patterns simultaneously in both the spatial and temporal domain. Event distributions are generated based on population density distributions with respect to the Emergency Department (in a more generic sense, this can be viewed as any central data collection location, for example, a police department, financial institution, etc.), and these distributions are then used in conjunction with the historical data to model the expected population density of events in order to maintain temporal coherence amongst hotspots. Such methods can provide insight into the ongoing impact of current events, and provide advanced warning for future events, thereby improving interdiction and response.

2 Related Work

Recently, the development of visual analytics systems for data analysis and exploration has been rapidly growing (e.g., [4, 19, 27, 31]). These systems incorporate a variety of visualization techniques from traditional, widely used methods, such as scatterplots or parallel coordinate plots to more recently developed tools (e.g., spiral graphs [5], theme river [15]). Techniques common across these systems include the probing, brushing and linking of data in order to help analysts refine their hypotheses, and these systems emphasize the interaction between human cognition and computation through dynamically linked statistical graphs and geographical representations of the data. However, while these systems allow users to explore their data and form hypotheses, it has been only recently that visual analytics systems have begun progressing towards predictive analytics (e.g., [32, 33]).

Analytic systems in the realm of syndromic surveillance include the Early Aberration Reporting System (EARS) [16] and the Electronic Surveillance System for the Early Notification of Community based Epidemics ESSENCE [20]. Unfortunately, all of these systems offer limited data exploration tools and analytic capability is limited to reactive alerts. Furthermore, these systems tend to generate a large amount of false positives for epidemiologists to analyze. Work in the geographical and visual analytics communities have attempted to improve healthcare data analysis and exploration through a variety of systems (e.g., [21, 23, 30]) using linked views and interactive plotting; however, these systems also stop short of predicting future events.

The concept of predictive analytics is found widely across financial services (e.g. credit scoring), retail sales, and health care. With respect to categorical geo-spatiotemporal event data, we focus our discussion solely on time series modeling and spatiotemporal modeling for forecasting events. A summary of time series analysis in biostatistics can be found in [10]. Common time series modeling techniques include the use of auto regressive moving average (ARMA) models (e.g., [1]) which describe stationary time series, and auto regressive integrated moving (ARIMA) average models (e.g., [25]). While such models are useful, we focus on the application of a non-parametric method, seasonal decomposition of time series by loess (STL) [7]. This method allows for flexible modeling of the differing onsets and shapes of seasonal peaks, and allows us to account for other components of variation also, see Section 3.2 for details.

Along with time series modeling and prediction, our work also focuses on spatiotemporal predictions. A summary of spatial modeling and geostatistical methods can be found in [11] and [12]. Diggle noted that typically, the temporal prediction should take precedence over the spatial. As such, we utilize temporal modeling for predicting events, and employ the use of kernel density estimation for creating a probability distribution of patient locations. This probability distribution is then used to place the number of predicted patients into geospatial locations, see Section 3.3 for more details.

3 Predictive Analytic Environment

Our current work extends the system developed by Maciejewski et al. [23, 24]. As in many visual analytic and information visualization systems, we utilize dually linked interactive displays for multi-domain/multivariate exploration and analysis as well as interactive filter controls for variable selection. We extend both the spatial and temporal viewing windows to incorporate spatiotemporal predictions for enhanced data analysis and exploration, moving from a visual analytics environment to a predictive visual analytics environment.

Figure 1 presents a screenshot of our system. In this example, the user is exploring potential future outbreaks of respiratory syndromes across the state of Indiana. The data displayed in the geospatial window utilizes a color mapping based on the percentage of patients that went to an emergency department classified as having respiratory syndromes using a sequential color scheme [2]. Users may interactively view other syndromes, or filter the data by age, gender and/or keyword in order to perform more complex analyses. Selection of counties and/or hospitals are displayed in the time series windows on the right. The time series plots provide an upper and lower bounds of prediction through an overlaid transparent polygon. The white vertical line serves as a reference for the geospatial date shown on the left. Note that predicted data is only displayed when actual data is not available. The contributions of our new system include methods for spatiotemporal prediction and methods for the interactive visualization of these predictions. We employ the use of several time series modeling techniques for data forecasting, and use the results of these predictions in a spatial modeling scheme to represent event distributions.

3.1 Time Series Power Transformation

One method we have previously applied [24] to simplify temporal analysis was the application of a power transformation to bring the data more in line with model assumptions [8]. In time series analysis, the logarithm transformation is widely applied when the mean is proportional to the standard deviation [3]. In cases where the data consists of counts following a Poisson distribution a square root transformation will approximately make the mean independent of the standard deviation. In each case, the transformations are necessary to simplify the modeling procedure. We examined our data under both a logarithmic and square root power transformation. The use of $\log(x)$ failed to eliminate the skewness on the right tail of the distribution for the number of observations; however, experimental results show that the $\sqrt{x}$ stabilizes the variability and yields a skew-free distribution of the time series. As such, all time series analyses are performed on the square root scale of the original series in order to remove the dependence of a signal's variance on its mean.

3.2 Time Series Prediction

In our predictive analytics environment, time series models are used for forecasting the future behavior of events. Our temporal modeling is performed over a spatial aggregation of data, meaning the collection of all event records over the state, county, or data collection agency make up the time series (in the case of syndromic surveillance the collection agency would be the Emergency Department). For multivariate data, we model each event category as a separate time series signal. Future work will focus on more robust models to capture correlations between signals. We employ the use of both a cumulative summation [16] and a seasonal-trend decomposition model [7]. These predictions can then be used for supply management (e.g., insuring enough antibiotics are available) and outbreak preparedness (i.e., if an outbreak has occurred or is expected to occur, staff members may be informed of the predicted models and can look for specific symptoms).

3.2.1 Prediction with Cumulative Summation

In terms of outbreak detection through time series analysis, one of the standard epidemiological algorithms employed is the cumulative summation (CUSUM) [16].

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

Equation 1 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. Such a lag is used to increase sensitivity to continued outbreaks.

Given the 3 day lag, we can use the CUSUM method to extend the current time series into the future by simply calculating the mean of the sliding window. This method allows us to provide the analyst with both an expected value for the next 3 days (note the 3 could be modified depending on the chosen lag) and an alert threshold. While this prediction is limited, the CUSUM method is useful for providing a quick look at the expected average number of incoming patients, and thresholds can quickly be set to determine various alert levels. For syndromic surveillance data, we utilize the threshold values of the EARS CUSUM2 model [16] (approximately two standard deviations).

3.2.2 Prediction with Seasonal-Trend Decomposition Based on Loess

Unfortunately, the CUSUM model fails to take into account some important characteristics of chief complaint count data, such as the day-of-the-week. Furthermore, using a 28 day sliding average is not ideal for time series with components that evolve over the course of a month. In order to more accurately model the data, we employ a different strategy in which the time series is viewed as the sum of multiple components of variation [14]. Seasonal-trend decomposition based on *loess* (locally weighted regression) [7] is used to separate the time series into its various components. STL components of variation arise from smoothing the data using moving weighted-least-squares polynomial fitting, in particular *loess* [9], with a moving window bandwidth in days. The degree of the polynomial is 0 (locally constant), 1 (locally linear), or 2 (locally quadratic).

For a given hospital, we decompose our daily patient count data into a day-of-the-week component, a yearly-seasonal component that models seasonal fluctuations, and an inter-annual component which models long term effects, such as hospital growth:

$$\sqrt{Y_t} = T_t + S_t + D_t + r_t \qquad (2)$$

where for the t-th day, Y_t is the original series, T_t is the inter-annual component, S_t is the yearly-seasonal component, D_t is the day-of-the-week effect, and r_t is the remainder.

The procedure begins by extracting the day-of-the-week component, D_t. First, a low-middle frequency component is fitted using locally linear fitting with a bandwidth of 39 days. Then D_t is the result of means for each day-of-the-week of the $\sqrt{Y_t}$ minus the low-middle-frequency component. Next, the current D_t is subtracted from the $\sqrt{Y_t}$ and the low-middle-frequency component is re-computed. This iterative process is continued until convergence. After removing the day-of-the-week component from the data, we use loess smoothing to extract the inter-annual component, T_t, using local linear smoothing with a bandwidth of 1000 days. Finally, we apply loess smoothing to the data with the day-of-week and inter-annual components removed, thereby obtaining the yearly-seasonal component, S_t, using local quadratic smoothing with a bandwidth of 90 days. After removing the day-of-week, inter-annual, and yearly-seasonal components from the time series, the remainder is found to be adequately modeled as independent identically distributed Gaussian white noise, indicating that all predictable sources of variation have been captured in the model. All parameters were chosen after extensive data modeling, and details may be found in our previous work [14]. It is important to note that these parameters should be modified to work best within the confines of a given data structure; however, the application of STL for modeling and prediction is not delegated only to syndromic surveillance data.

For prediction using the STL method, we rely on some statistical properties of loess, namely that the fitted values $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)$ are a linear transformation of the observed data. $Y = (Y_1, \ldots, Y_n)$. Each step of the STL decomposition involves a linear filter of the data. In other words, an output time series $x = \{x_1, \ldots x_n\}$ is produced by an input time series $w = w_1, \ldots, w_n$ through a linear combination

$$x_i = \sum_{i=1}^{n} h_{ij} w_j. \tag{3}$$

If we let H be a matrix whose (i, j)-th element is h_{ij}, then we have

$$x = Hw. \tag{4}$$

We will refer to H as the operator matrix of the filter. Now, let H_D, H_S, and H_T denote the operator matrices of the day-of-week, yearly-seasonal, and inter-annual filters, respectively. All of these matrices are $n \times n$. H_S and H_T are straightforward to calculate [9], but H_D is more difficult to calculate as it is the result of an iteration of smoothing. Once all of these have been calculated, the operator matrix for the entire procedure, H can be written as

$$H = H_D + H_T(I - H_D) + H_S(I - H_D - H_T(I - H_D)), \tag{5}$$

where I is the $n \times n$ identity matrix. Now, the fitted values are obtained by

$$\hat{Y} = HY. \tag{6}$$

To make better sense of Equation 5, the day-of-the-week smoothing, H_D, is applied to the raw data, while the inter-annual smoothing, H_T, is applied to the raw data with the day-of-week removed, and finally, the yearly-seasonal smoothing, H_S, is applied to the raw data with the day-of-week and inter-annual removed.

Now, the variance of the fitted values is easily obtained

$$Var(\hat{Y}_i) = \hat{\sigma}^2 \sum_{j=1}^{n} H_{ij}^2, \tag{7}$$

where $\hat{\sigma}^2$ is the variance of Y, and can be estimated from the remainder term r_t.

Now, if we wish to predict ahead, a days, we append the operator matrix H with a new rows, obtained from predicting ahead within each linear filter and use this to obtain the predicted value and variance. For example, if we wish to predict the value for day $n+1$, we would obtain

$$\hat{Y}_{n+1} = \sum_{j=1}^{n} H_{n+1,j} Y_j \tag{8}$$

and

$$Var(\hat{Y}_{n+1}) = \hat{\sigma}^2 (1 + \sum_{j=1}^{n} H_{n+1,j}^2), \tag{9}$$

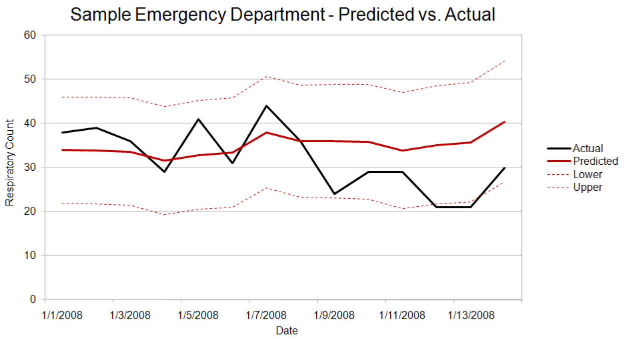


Figure 2: Seasonal-trend decomposition based on loess prediction compared to actual measurements.

so that a 95% prediction interval will be calculated as

$$\hat{Y}_{n+1} \pm 1.96\sqrt{Var(\hat{Y}_{n+1})}. \tag{10}$$

To demonstrate our prediction model, we have utilized data from PHESS from January 1, 2006 through December 31, 2007 for a single emergency department. Our STL modeling and prediction method is applied to this data to predict January 1, 2008 through January 14, 2008. We then compare this prediction to the actual data in Figure 2. Here we can see that the model is able to capture properties of the signal; however, it is important to note that as the predictions move further into the future, the accuracy decreases. Comparable results were found for all other hospitals in our system.

3.3 Geospatial Prediction

While the temporal prediction provides a forecast of the number of expected events, we are also interested in providing analysts with a means to analyze the expected spatial distributions. As stated above, our time series prediction can be performed over a variety of spatial aggregations. As such, we allow users to choose between several granularities of the spatial prediction using various data aggregations in the time series prediction as a basis for our event distribution.

3.3.1 Geographically Aggregated Distribution

The simplest means for spatial prediction is to utilize the time series counts based on an arbitrary geographic boundary (e.g., state, county, zip code), and visualize this information in the form of a color map. In the case of our system, a time series is associated with each geographically defined region (e.g., county, zip code, state). On any given day, we have either the number of events that occurred and the predicted number of events, or only the predicted number. Such values allow for several different comparison methods. In Figure 3 the analyst is scrolling across time through the predictive models, comparing the STL prediction (Figure 3 (Left)) of respiratory illness versus total illness with the CUSUM prediction (Figure 3 (Right)). By comparing two different predictive models, the analyst can see where the models disagree and flag those counties as regions to explore in the coming days.

3.3.2 Spatiotemporal Distribution

While it is useful to compare temporal predictions across an arbitrary geographical boundary (e.g., county, zip code), it is also useful to have a model which can incorporate a finer granularity of event distribution. As such, we expand on our previous use of density estimation [23] and model the spatiotemporal distributions of patients

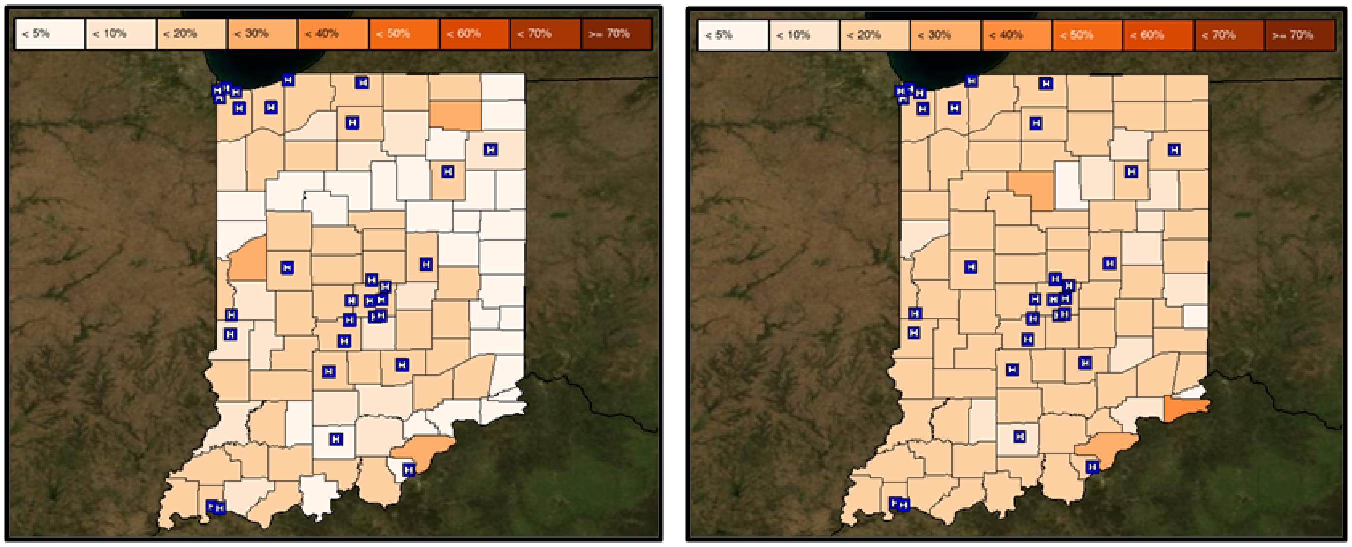


Figure 3: Spatiotemporal prediction comparing (Left) the STL prediction and (Right) the CUSUM prediction.

based on their ED visits. We employ the use of a modified variable kernel method [26] which scales the parameter of the estimation by allowing the kernel width to vary based upon the distance from X_i to the k-th nearest neighbor in the set comprising $N-1$ points.

$$\hat{f}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\max_{(h,d_{i,k})}} K\left(\frac{\mathbf{x}-X_i}{\max_{(h,d_{i,k})}}\right) \tag{11}$$

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, and h is the minimum allowed kernel width.

In order to reduce the calculation time, we have utilize the Epanechnikov kernel, Equation 12:

$$K(\mathbf{u}) = \frac{3}{4}(1-\mathbf{u}^2)1_{(||\mathbf{u}||\leq 1)} \tag{12}$$

where the function $1_{(||\mathbf{u}||\leq 1)}$ evaluates to 1 if the inequality is true and zero for all other cases.

Given the predicted number of events from the time series modeling, we want to distribute these events given the probability density function of event locations with respect to some shared geographic location (in the case of syndromic surveillance, we model the population distribution served by a given Emergency Department). For each Emergency Department, we know each patient's home address. These addresses are mapped to a grid centered around the hospital, and we employ Equation 11 to create a distribution function representing the probability that a patient will come to the hospital from a given (latitude, longitude) pair. We then randomly distribute the n predicted events according to this distribution. This is done for each emergency department to simulate patient distributions across the state.

Once the events are distributed, we create a three-dimensional array, consisting of a grid of patient locations across the predicted day being visualized and the previous t days. We then perform a three-dimensional kernel density estimation to maintain the temporal coherence of previous hotspots in order to analyze if such locations could be persistent across time under the assumption that patients will visit the Emergency Department based only on its service area distribution. Finally, the estimated density of the current days events (with the incorporated temporal history) is plotted as a ratio of the number of events under analysis versus the total number of events (also calculated to incorporate temporal history). Examples of the use of such modeling are provided in Section 4.

4 Hotspot Analysis and Forecasting

By using a combination of geospatial and temporal visualization and analytics tools, our system provides epidemiologists with tools for real-time hypothesis testing and event prediction. In order to demonstrate the strengths of our modeling tools, we utilize synthetic syndromic surveillance data [22] with known outbreaks. This section provides both a retrospective case analysis and a prospective case analysis. We utilize two years worth of synthetic syndromic surveillance data (January 1, 2006 through December 31, 2007) with 33 emergency departments across the state of Indiana [22] with two known outbreaks. We utilize synthetic data as opposed to actual data in order to remove privacy concerns; however, we have found the results to be comparable across synthetic and actual data.

4.1 Retrospective and Reactive Analysis

In order to illustrate our alert generation using CUSUM and spatial modeling effects, an outbreak containing patients presenting signs of respiratory illness was introduced beginning on July 18, 2007 and ending July 22, 2007. The injection of patients followed a log-normal distribution such that the number of excess patients showing respiratory syndrome symptoms were 1 on July 18, then 18, 8, 5, 3, and 2 for each subsequent day. In Figure 4 (Top-Left) the user is looking at a typical geospatial view at a county level aggregation. Note that in the disease injection area, the counties are only showing a slightly higher percentage than their neighbor; however, if the analyst were to instead look at a comparative view, Figure 4 (Top-Right), of the actual patient counts versus the predicted patient counts, a different story unfolds. The analyst can immediately identify a cluster occurring on July 19th in the area where an outbreak was injected. Furthermore, the time series alerts from the CUSUM algorithm, yellow diamonds in Figure 4 (Bottom), further corroborate this outbreak.

In order to obtain a more localized view of where the outbreak is occurring, the analyst may switch over to the density estimate view, Figure 5. The analyst can scroll back in time prior to when the first alert was generated in the hospital's time series and begin looking

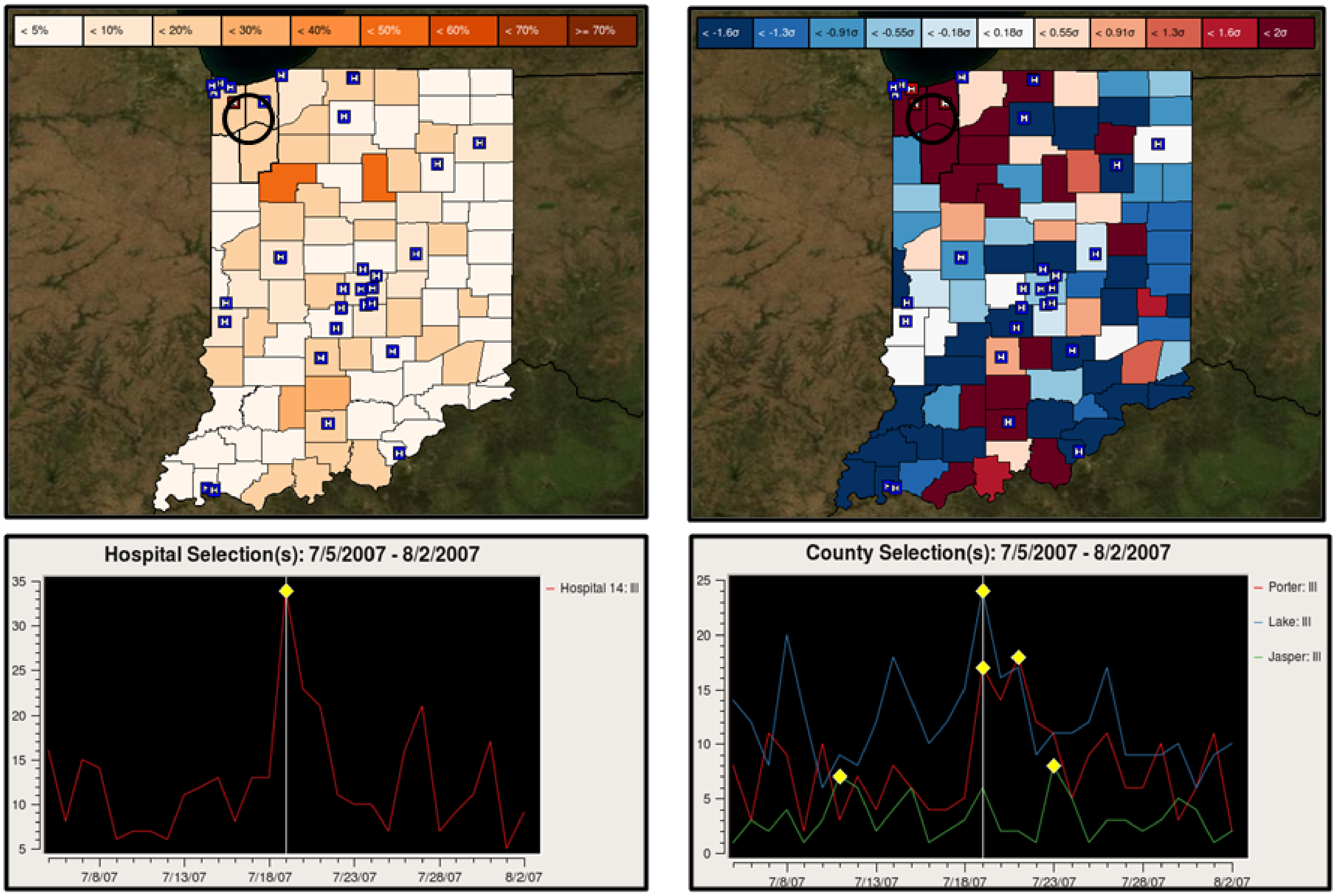


Figure 4: An outbreak has been injected beginning on July 18, 2007 and ending on July 22, 2007. The black circle represents the injection. (Top) Images show July 19, 2007. (Top-Left) A percentile view of patients with respiratory illness. Note that no outbreak is readily observable in this view. (Top-Right) A comparative view of actual patients versus the number of predicted patients. Here the analyst can quickly see which counties have exceeded the predicted values.(Bottom) Time series views of the nearby hospital and counties, yellow diamonds indicate alerts. We see that on July 19th, alerts were generated for the three dark red counties overlapped by the injection circle.

at the estimated patient density. Notice that the estimated density remains consistent in the circled area until the outbreak reaches its peak on July 19th. On that day, the spatial model shows a higher concentration of patients across a very specific geographic region, and the analysts may then focus their attention on that particular region as opposed to a multiple county alert that may have been issued if only the views in Figure 4 (Top) had been utilized. However, such a view is only applicable in a reactive manner (the analyst is comparing what happened on July 19th to what was expected to happen).

4.2 Forecasting and Proactive Analysis

In a proactive analysis, the analyst would be watching for alerts, or analyzing potential future spreading of already confirmed alerts. In order to illustrate the use of our tools for analyzing future events, an outbreak containing patients presenting signs of respiratory illness was introduced on December 29, 2007 and is still peaking on December 31, 2007. The injection of patients was such that 6 patients were injected on December 29th, then 15 and 21 patients on each of the following days. The analyst has no data past this point.

In Figure 6 (Top) the analyst is viewing December 29th, 2007 for Hospital 21 located in Montgomery County. An alert was generated based on that hospitals time series for three consecutive days (December 29th - 31st), and the analyst is concerned that the outbreak is not subsiding. In the time series window, the analyst notes that over the following two days, the predicted number of patients visiting the hospital is expected to be larger than the generated alert threshold. Further, the alert threshold within Montgomery County is far less than the upper bounds of the prediction.

With this information, the analyst may also wish to corroborate these findings with the spatial models. Figure 6 (Bottom) illustrates the hotspots across Indiana from December 31, 2007 through January 3, 2008. The analyst can immediately see a hotspot in an area near the hospital. When tracking the hotspot into the future, the levels remain high on January 1st and 2nd and begin to taper off on the third. With this information, the analyst may assume that the outbreak seems to be persistent into the future, and could choose to issue a health alert localized to that area if the threat was severe enough.

5 Conclusions and Future Work

Our current work demonstrates the benefits of predictive visual analytics for forecasting 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 time-series prediction views with geo-spatiotemporal views for enhanced exploration and data analysis through the use of traditional spatial histogram visualizations and finer granularity heatmaps for more accurate event detection. Further, the extension of the kernel density estimation to

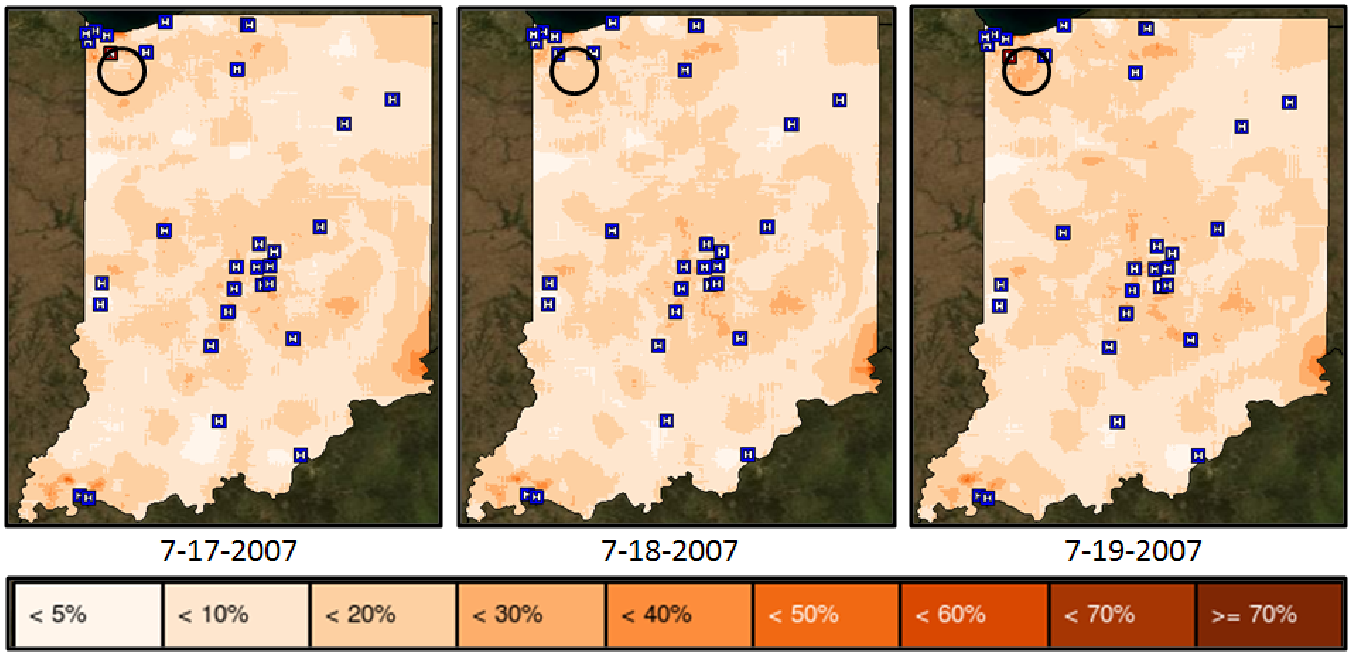


Figure 5: Detecting an injected outbreak with spatiotemporal modeling. An outbreak has been injected beginning on July 18, 2007 and ending on July 22, 2007. The black circle represents the injection. Note that on July 19th, a rise in cases is visible in that area, providing information on a more exact cluster location.

incorporate temporal modeling maintains the temporal coherency of outbreaks, enhancing analysis.

In order to ensure system usability, the development of our system has been done under the guidance of collaborators at the Indiana State Department of Health and the Indiana University School of Medicine. Further, discussions with local police departments and state law enforcement agencies indicate that such tools are able to assist other analysts whose data dimensions fall into the realm of categorical geo-spatiotemporal data.

Future work includes the introduction of advanced control chart methods for more accurate alert generation. Furthermore, it is important to note that the power transformation used can change based on the time series data under analysis. We also plan on employing spatiotemporal clustering algorithms for event detection as well as correlative analysis views within the temporal domain.

References

[1] G. Box and G. Jenkins. *Time series analysis: Forecasting and control*. Holden-Day, San Francisco, 1970.

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

[3] P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting (2nd edition)*. Springer, 2003.

[4] T. Butkiewicz, W. Dou, Z. Wartell, W. Ribarsky, and R. Chang. Multi-focused geospatial analysis using probes. *IEEE Transactions on Visualization and Computer Graphics*, 14:1165–1172, Nov. - Dec. 2008.

[5] J. V. Carlis and J. A. Konstan. Interactive visualization of serial periodic data. In *Proc. Symp. User Interface Software and Technology (UIST '98)*, 1998.

[6] 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.

[7] R. B. Cleveland, W. S. Cleveland, J. McRae, and I. Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6:3–73, 1990.

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

[9] W. S. Cleveland and S. J. Devlin. Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83: 596–610, 1988.

[10] P. J. Diggle. *Time series analysis: A biostatistical introduction*. Oxford University Press, Oxford, 1990.

[11] P. J. Diggle. *Statistical Analysis of Spatial Point Patterns*. Edward Arnold, London, 2003.

[12] P. J. Diggle and P. J. Ribeiro. *Model-based Geostatistics*. Springer, New York, 2007.

[13] 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.

[14] R. P. Hafen, D. E. Anderson, W. S. Cleveland, R. Maciejewski, D. S. Ebert, A. Abusalah, M. Yakout, M. Ouzzani, and S. Grannis. Syndromic Surveillance: STL for Modeling, Visualizing, and Monitoring Disease Counts. *BMC Medical Informatics and Decision Making*, 2009, to appear.

[15] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002.

[16] 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.

[17] M. Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26:1481–1496, 1997.

[18] A. D. Langmuir. The surveillance of communicable diseases of national importance. *New England Journal of Medicine*, 268, 1963.

[19] 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.

[20] 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.

[21] A. M. MacEachren, F. P. Boscoe, D. Haug, and L. Pickle. Geographic visualization: Designing manipulable maps for exploring temporally varying georeferenced statistics. In *Proceedings of the 1998 IEEE*

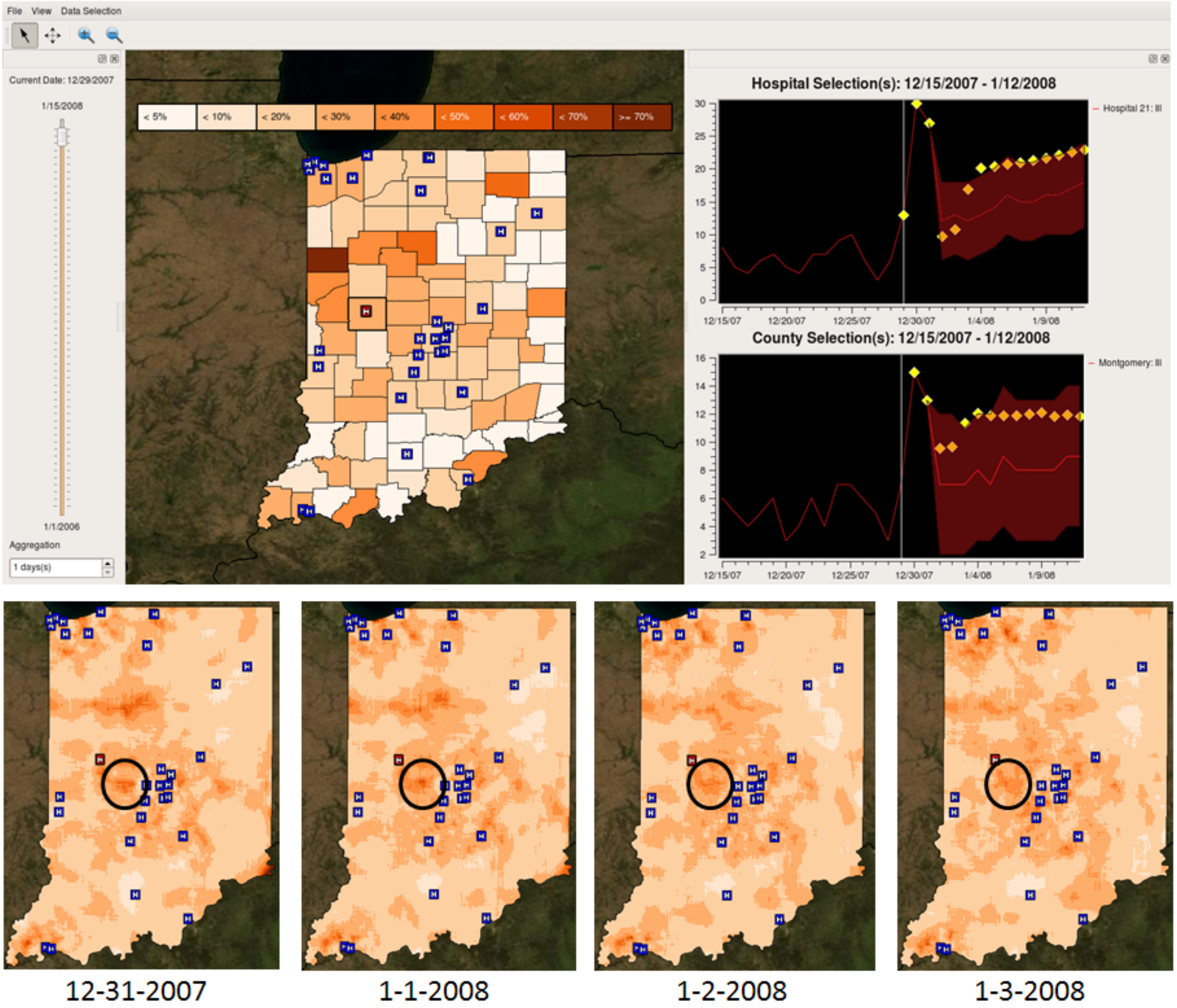


Figure 6: In this data, an outbreak has been injected beginning on December 29, 2007 and is still peaking on December 31, 2007 (the last day of available data). The black circle represents the approximate area of injection. (Top) The predictive visual analytics system. Note the time series view contains upper and lower bounds for the prediction, as well as at what range alerts would occur. Here we can see that on January 1st and 2nd, the predicted value would be higher than the alert threshold indicating the continued presence of an outbreak. (Bottom) A series of predicted geospatial data. Here we follow the outbreak in the spatial prediction model and see it begins to subside on January 3rd.

Symposium on Information Visualization, page 87, 1998.

[22] R. Maciejewski, R. Hafen, S. Rudolph, G. Tebbetts, W. S. Cleveland, S. J. Grannis, and D. S. Ebert. Generating synthetic syndromic surveillance data for evaluating visual analytics techniques. *IEEE Computer Graphics & Applications*, 2009.

[23] 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. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, October 2008.

[24] 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 (VAST)*, pages 27–34, October 2007.

[25] B. Reis and K. Mandl. Time series modeling for syndromic surveillance. *BMC Med Inform Decis Mak*, 3:2, 2003.

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

[27] J. Stasko, C. Gorg, Z. Liu, and K. Singal. Jigsaw: Supporting investigative analysis through interactive visualization. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology 2007*, pages 131–138, 2007.

[28] 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.

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

[30] C. Tominski, P. Schulze-Wollgast, and H. Schumann. 3d information visualization for time dependent data on maps. In *International Conference on Infomation Visualization (IV)*, 2005.

[31] C. Weaver. Multidimensional visual analysis using cross-filtered views. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, October 2008.

[32] P. C. Wong, R. Leung, N. Lu, M. Paget, J. C. Jr., W. Jian, P. Mackey, T. Tayler, Y. Xie, J. Xu, S. Unwin, and A. Sanfilippo. Predicting the impact of climate change on U.S. power grids and its wider implications on national security. In *Proceedings of the AAAI Spring Symposium on Technosocial Predictive Analytics*, 2009.

[33] J. Yuei, A. Raja, D. Liu, X. Wang, and W. Ribarsky. A blackboard-based approach towards predictive analytics. In *Proceedings of the AAAI Spring Symposium on Technosocial Predictive Analytics*, 2009.