Forecasting Hotspots - A Predictive Analytics Approach

Ross Maciejewski, Member, IEEE, Ryan Hafen, Stephen Rudolph, Stephen G. Larew, Michael A. Mitchell, William S. Cleveland, David S. Ebert, Fellow, IEEE

Abstract—Current visual analytics systems provide users with the means to explore trends in their data. Linked views and interactive displays provide insight into correlations among people, events, and places in space and time. 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 where hotspots are found, analysts would like to predict how these regions may grow in order to plan resource allocation 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 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 modeling, along with spatial and temporal alerts to indicate the occurrence of statistically significant hotspots. Spatial data are 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.

Index Terms—Predictive analytics, visual analytics, syndromic surveillance.

1 INTRODUCTION

VISUAL analytics has been defined as the science of analytical reasoning assisted by interactive visual interfaces [1]. Recently, visual analytics systems (e.g., [2–4]) 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 are 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, or other such actions. 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 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, 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 (county, zip code, collection station), or spatial snapshots of a small time aggregate (e.g., day, week). These aggregations are then analyzed using some control chart method (e.g., cumulative summation, exponential weighted moving average). There are also systems that allow for spatiotemporal alert detection ([5]), but such systems become intractable as the data set becomes large and current implementations provide little to no geographical mapping tools. 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. [2, 6] 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. In our previous work, hotspots were noted solely as areas of the map where events are occurring at a higher normalized magnitude with respect to the entire area of study. As we move to predictive analytics, the previous means of visualizing hotspots can still be applied (depending on the chosen visualization modality); however, we now extend the definition of hotspots to also include areas where the current magnitude of events in an area is larger than the predicted (or expected) number of events. As such, our system now provides users with the ability to compare the current time period’s event magnitude to past historical events; furthermore, the integration of our system with the analytical capabilities of the Kulldorff’s spatial scan statistics [5] allows users to
determine the statistical significance of the events in that area (i.e., determine whether or not the increased magnitude was due to chance or not). For event prediction, we utilize seasonal trend decomposition by loess and a scalable kernel density estimation to aid in spatially distributing the event prediction. Finally, we utilize texture mapping to provide users with information on the variability of the prediction.

In our current predictive scheme, each component of the underlying dataset is treated as a separate time series variable, and cross-correlative effects are disregarded. Future work will focus on visual analytics schemes for correlative analysis; however, current syndromic surveillance work focuses on separating syndromes and often ignores correlative effects. While our time series prediction work follows this trend, we introduce a novel visualization scheme in which we apply a k-means clustering to the data events and project these clusters into the geographic space, thereby visualizing healthcare ecotopes.

To summarize, novel system features include:

- Linked spatial and temporal views for analysis/forecasting
- The application of seasonal trend decomposition by loess for time series prediction
- Scalable 3D kernel density estimation for spatiotemporal prediction to maintain temporal coherence
- Integrated spatial scan statistics
- Comparative views for temporal changes
- Multiple spatial aggregation schemes for hotspot analysis and forecasting
- Texture overlays for uncertainty visualization
- Multivariate clustering for interactive ecotope creation

In order to demonstrate the impact of such tools, we focus our discussion on a representative categorical 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 [7, 8], 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) [9], 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, botulinc, shock/coma, and other) 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. [10].

Our work focuses on advanced interactive visualization and analysis methods providing linked environments of geospatial data and time series graphs. Time series events are forecast 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 spatiotemporal interface, helping users find patterns simultaneously in both the spatial and temporal domain. Event distributions are generated based on all events recorded 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 or financial institution). This event distribution function can then be used to place predicted events, and these predicted events are then correlated with the historical data to model the expected geospatial density of events thereby maintaining 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., [3, 4, 11–14]). 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 [15], theme river [16]). 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., [17, 18]).

Analytic systems in the realm of syndromic surveillance include the Early Aberration Reporting System (EARS) [19] 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 has attempted to improve healthcare data analysis and exploration through a variety of systems (e.g., [2, 21, 22]) 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 healthcare. With respect to categorical 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 [23]. Common time series modeling techniques include the use of auto regressive moving average (ARMA) models (e.g., [24]) which describe stationary time series, and auto regressive integrated moving average (ARIMA) 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) [26, 27]. This method allows flexible modeling of the differing onsets and shapes of seasonal peaks, and allows us to account for other components of variation also, see Section 3.1.2 for details. Machine learning methods could also be applied (e.g., [28]), however, we leave such methods for future work.

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 [29] and [30], and the authors note that, typically, the temporal prediction should take precedence over the spatial. As such, we utilize temporal modeling for predicting events, and employ 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 3.2 for more details. We also incorporate the spatial scan statistic developed by Kulldorff [5] into our work utilizing the batch mode interface of the SatScan system. Previous work on visualizing the output of the SatScan has been done by Boscoe et al. [31] using nested circles to visualize the various cluster outputs, and Chen et al. [32] which analyzed the sizes of the clusters output for enhancing the resulting visualization with reliability information.

Given the sparsity of our data in certain geographical areas, our system also provides visualizations of data reliability/uncertainty. Due to the fact that low counts in spatially aggregated areas can lead to skewed statistics [33, 34], we utilize the variance of the data to represent the level of uncertainty. Previous work in uncertainty visualization has utilized glyphs in flow visualization [35], point-based probabilistic surfaces [36], procedural annotations [37], and the use of the HSI color model [38]. Our current system follows the convention of MacEachren et al. [21]. Their work utilized adjacent maps to show data reliability, and compared this to maps utilizing textures (hashing) overlaid on color to display reliability. We incorporate a similar texture overlay to represent data reliability within our system.

3 PREDICTIVE ANALYTIC ENVIRONMENT

Our current work extends the system developed by Maciejewski et al. [2, 6]. 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.

This work is motivated by the fact that in analyzing event driven data, analysts are often creating plans for future resource usage and need. In regards to syndromic surveillance, the prediction of potential disease outbreaks allows better resource management and response, thus, the need for analyzing and predicting spatiotemporal events necessitates the creation of such a system. The syndromic surveillance community has done much work in categorizing and analyzing events based on emergency department data, and we utilize this data set in order to demonstrate the functionality of our tools.

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 divergent color map [39] illustrating the number of standard deviations the data are above or below the expected average. Results from Kulldorff’s spatial scan statistic [5] are overlaid on the map, providing cues for the user to explore certain areas and are linked in the lower righthand window. 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 upper and lower bounds of prediction through an overlaid transparent polygon. The white vertical line serves as a reference for the geospatial data shown above the slider. Note that predicted data are only displayed when actual data are not available.

The contributions of our new system include methods for spatiotemporal prediction and methods for the interactive visualization of these predictions. We employ several time series modeling techniques for data forecasting, and use the results of these predictions in a spatial modeling scheme to represent event distributions. Visualization schemes include a default view of predicted hotspots using the EARS CUSUM2 model [19] overlaid with semi-transparent circles of Kulldorff’s spatial scan statistic [5] for results with $p < .1$, a view showing the magnitude of the change across time, county level views of STL predicted hotspots, and ecotope views for multivariate analysis. Note that the predictive power (false positive rates) of the models are discussed in the respective papers. The goal of this work is to demonstrate how these methods can be augmented through linked visuals and interactive displays, thereby creating a predictive visual analytics environment.

### 3.1 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 both a moving average and a seasonal-trend decomposition model [26]. 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).

For time series prediction, we utilize both the standard event detection algorithms utilized by the Center for Disease Control, as well as a seasonal trend decomposition method that we have previously developed [27]. In the work by Hafen et al., the seasonal trend decomposition method is compared to other standard methods of syndromic modeling in order to demonstrate the efficacy of our techniques. While other methods can be successfully applied to syndromic surveillance data (e.g., [40]), our previous work in modeling syndromic surveillance data shows that this method currently outperforms other suggested techniques.

#### 3.1.1 Event Detection

In terms of outbreak detection through time series analysis, one of the standard epidemiological algorithms employed is the EARS C2 alert algorithm [19] based on the cumulative summation (CUSUM).

$$S_t = \max \left(0, S_{t-1} + \frac{X_t - (\mu_0 + k\sigma_0)}{\sigma_t}\right)$$  \hspace{1cm} (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, $\mu_0$ is the expected value, $\sigma_0$ 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, $\mu_0$, and standard deviation, $\sigma_0$, 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. All model parameters utilized in our system are based on recommendations by [19] and other syndromic surveillance literature.

Given the 3 day lag, we 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 moving average 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 [19] (approximately two standard deviations).

#### 3.1.2 Prediction with Seasonal-Trend Decomposition Based on Loess

Unfortunately, the moving average 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 [27]. Seasonal-trend decomposition based on loess (locally weighted regression) [26] 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 [41], with a moving window bandwidth in days. The degree of the polynomial is 0 (locally constant), 1 (locally linear), or 2 (locally quadratic).

Here, it is important to note that in order to appropriately model the time series using STL, the mean and variance of the data needs to be independent. To accomplish this, a power transformation is applied to the data. In time series analysis, the logarithm transformation is widely applied when the mean is proportional to the standard deviation [42]. In cases where the data consists of counts following a Poisson distribution a square root transformation will make the mean independent of the standard deviation. Analysis of the syndromic surveillance data showed that the square root transformation stabilizes the variability and yields a more Gaussian distribution of the time series residuals. As such, the STL modeling is performed on the square root scale of the original series in order to remove the dependence of a signal’s variance on its mean.

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:

\[ \sqrt{T_t} = T_t + S_t + D_t + r_t \]  

(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 \( \sqrt{T_t} \) minus the low-middle-frequency component. Next, the current \( D_t \) is subtracted from the \( \sqrt{T_t} \) 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. It is important to note that these parameters should be modified to work best within the confines of a given data set; however, the application of STL for modeling and prediction is not relegated only to syndromic surveillance data. Details of the methodology behind the parameter choices can be found in [27] and is not the main focus of this work. However, the extension of this method to include data prediction is novel.

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

\[ x_t = \sum_{i=1}^{n} h_{ij} w_j. \]  

(3)

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

\[ x = Hw. \]  

(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 [41], 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

\[ H = H_D + H_T(I - H_D) + H_S(I - H_D - H_T(I - H_D)), \]  

(5)

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

\[ \hat{Y} = HY. \]  

(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

\[
\text{Var}(\hat{Y}_i) = \hat{\sigma}^2 \sum_{j=1}^{n} H_{ij}^2,
\]

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, \( x \) days, we append the operator matrix \( H \) with \( x \) 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

\[
\hat{Y}_{n+1} = \sum_{j=1}^{n} H_{n+1,j} Y_j
\]

and

\[
\text{Var}(\hat{Y}_{n+1}) = \hat{\sigma}^2 \left(1 + \sum_{j=1}^{n} H_{n+1,j}^2\right),
\]

so that an approximately 95% prediction interval will be calculated as

\[
\hat{Y}_{n+1} \pm 1.96 \sqrt{\text{Var}(\hat{Y}_{n+1})}.
\]

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. Comparable results were found for all other hospitals in our system.

### 3.2 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.

#### 3.2.1 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. On any given day, we have either the number of events that occurred and the predicted number of events, or only the predicted number. In Figure 3 the analyst is scrolling across time through the predictive models. The analyst can visualize the data aggregated at the county level utilizing the STL predictive model (Figure 3 (Left)) which provides a stronger means of prediction at a larger computational cost, or the analyst can utilize a finer spatial granularity view (as shown in Figure 3 (Right)) at the cost of a less accurate moving average prediction.

#### 3.2.2 Spatiotemporal Distribution

In order to compute predictions at the finer spatial granularity, we expand on our previous use of density estimation [2] and model the spatiotemporal distributions of patients based on their emergency department visits. We employ a modified variable kernel method [43] 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.

\[
\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\max(h, d_{i,k})} K \left( \frac{x - X_i}{\max(h, d_{i,k})} \right)
\]

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.
Fig. 4. (Left) Percentile views showing the magnitude of a syndrome count with respect to the underlying population that visited the emergency department in a given area on 12-31-2007. (Right) Temporal change from 12-31-2007 to 1-1-2008. The SatScan cluster was found on 12-31-2007 and was chosen to remain in persistent mode by the user, meaning that the cluster visualization will persist as the user scrolls through time. The user can see that the magnitude of illness in that area is predicted to remain consistent or increase.

We utilize the Epanechnikov kernel \[43\], Equation 12:

\[
K(u) = \frac{3}{4}(1 - u^2)1(|u| \leq 1)
\]  

(12)

where the function \(1(|u| \leq 1)\) 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. This allows the analysts to determine if the hotspots 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 day’s events (with the incorporated temporal history) can be plotted as a ratio of the number of events under analysis versus the total number of events (also calculated to incorporate temporal history), or each grid point can be utilized in the EARS calculations to determine hotspots as in Figure 3 (Right). Examples of the use of such modeling for hypothesis generation and testing are provided in Section 4.

3.3 Spatial Scan Statistics

Kulldorff proposed a spatial scan statistic [5] to detect the location and size of the most likely cluster of events in space or space-time data. A scanning window of variable size and shape is moved across the area of interest and for each window a likelihood value is calculated based on the contents of the window. The likelihood value is maximized over all possible windows and this maximum likelihood is called the scan statistic. Our system utilizes the available batch mode settings of the SatScan system, focusing on the Bernoulli data model [44] as patients are point based locations within our system. Figure 1 shows the SatScan control window (lower right) utilized within our system. Statistically significant events are automatically highlighted (with yellow representing significance at the .1 level and red at the .05 level). Semi-transparent circles (also colored according to significance) for statistically significant events are automatically drawn on the map (as seen in Figure 1). Circles may be toggled on and off by selecting the items from the SatScan control window (lower right controls of Figure 1).

3.4 Visualization Modalities

The application of various models and aggregations to categorical spatiotemporal data allows a variety of visualization modalities to support users views. System views include:

- Percentile views using sequential color maps to visualize which areas had the highest number of patients with a
given syndrome with respect to the number of patients seen visiting a hospital

- Temporal change views in which a divergent color map is used to represent which areas of the map saw an increase or decrease in the number of syndromic patients
- Hotspot views in which a divergent color map is used to represent the number of standard deviations an area is above or below the expected value
- Uncertainty views in which a texture is overlaid on any visualization modality representing areas with a variance above some user defined threshold
- Ecotope views in which a qualitative color map is used to represent multivariate clusters of syndromes

### 3.4.1 Percentile Views

Figure 4 (Left) demonstrates the percentile view across both county aggregation and the kernel density estimated data. Percentages are calculated as a function of the number of patients with syndrome X in an area, A, divided by the total number of patients seen in A. Such a view is useful in providing an analyst with an overall picture of the health of the state; however, it provides little statistical information about events without further exploration and interaction (such as selecting counties or hospitals and looking at the time series view).

### 3.4.2 Temporal Change Views

Figure 4 (Right) demonstrates the temporal change view across the kernel density estimated data. The temporal change is calculated as the current time aggregation minus the previous time aggregation, thereby providing a magnitude of change as the interactively scrolls through the data. Here, a divergent color map is used to show the percentile change from day to day. Such a view can provide analysts with details about the growth of previously found clusters, allowing them to check for persistence, growth or decay. Analysts may lock the previous day’s clusters to the map using a circular drawing tool or using the SatScan clusters. Then by moving forward in time with this view, analysts can view the predicted data change in that area. Note that the magnitude of the illness is now known to be consistent (or increasing) from 12-31-2007 to 1-1-2008 as the area colored under the 12-31-2007 red cluster is either white (meaning no increase or decrease) or red (meaning the number of events increased).

### 3.4.3 Hotspot and Uncertainty Views

The hotspot view (as illustrated in Figure 3(Left)) is the analysts default view. Here the analyst can immediately see which areas of the map were statistically significant. In this mode, the analyst then forms (or generates) a hypothesis stating that areas in red may be having a disease outbreak. The overlay of the spatial scan statistics results can help strengthen this hypothesis and provide a hypothesis testing tool by returning p-values.

Furthermore, an overlay of uncertainty information through textureing based on the variability of the data can also help in hypothesis formation. The uncertainty information is calculated as the variance of the data in a given area over the last week. If the variance is above some user defined threshold, then a texture is mapped to the area. The area being analyzed is a single areal bin of the map. The sparser the data are, the less reliable and more uncertain the data will tend to be. Areas of the map with high variance tend to show up as hotspots; however, these areas may not be worth investigating as the statistical power of the data are quite low. In Figure 3(Right) we can see that many of the hot spots correspond to a high level of uncertainty, making those less reliable in the analysts mind. This is shown through the use of the angled line texture overlays.

### 3.4.4 Multivariate Clustering Ecotope Creation

Given the multivariate nature of the data, our system also provides an ecotope view in which qualitative color maps are utilized to represent data clusters within an n-dimensional space. Ecotopes are areas of similar conditions relating an environment to the organisms or inhabitants of that space. They can be thought of as the geographical component to an ecosystem in that they denote the particular area where a relation between habitats and their inhabitants exist. In this paper the idea of an ecotope is applied to health care symptom rates such that the ecotope regions are taken to mean geographical areas whose human population exhibits similar rates for any particular symptom or set of symptoms. Instead of a broad climate zone style stratification where all climate factors are taken into account to produce large generalized climate zones, the method taken in this paper is to create health ecotopes specific to certain selected symptoms. We utilize K-means clustering [45] as shown in work by Hargrove and Hoffman [46, 47] for ecotope creation.

First we place each syndrome classification to a different grid and use kernel density estimation to create a probability density function (pdf) for the syndrome for a given day. Each syndrome then becomes an axis in the n-dimensional space, where each cell from the pdf is mapped into this space. We then apply K-means clustering, placing each data cell into a group of cells with similar property rates. By mapping the found clusters of like data points onto a two dimensional geographic space health ecoregions will be delineated as colored regions. Figure 5 (Left) demonstrates this process on three dimensional data. Our system also allows the user to then sort these clusters by the magnitude of one dimension in order to provide enhanced visual exploration as shown in Figure 5 (Right).

By classifying our geography into health care ecotopes an understanding of how our environment and symptom rates relate to each other can be found. The relationship may be that the environment or location may be a strong contributing factor to the syndrome rate. In this case it is likely that environmental anomalies would be given a unique ecotope color. However, it is important to note that the relationship between the environment and inhabitants may also be weak. In the case where a region does not have any strong contributing factor it is likely that the ecotope boundary will change over time.

In Figure 5 (Right), we can see that certain areas of the state map to various combinations of respiratory, gastrointestinal,
neurological and constitutional syndromes. In this view, the analyst may first have utilized the hotspot views to find areas on the map that correspond to a significant increase in a particular syndrome. From there, the analyst might then move to the healthcare ecotope view and see what syndrome groupings are occurring in that area. In this case, Figure 5 (Right), we can see near Lake Michigan (the upper West portion of the map) a pink area surrounded by an orange area meaning that a cluster with a higher amount of respiratory illness is in that area in comparison to the surrounding. Further, we can see that the average gastrointestinal value is higher than the respiratory value, indicating that there may be a correlation between the two syndromes.

4 CASE STUDIES IN HOTSPOT ANALYSIS AND FORECASTING

By using a combination of geospatial and temporal visualization and analytics tools, our system provides analysts with tools for real-time hypothesis generation/testing and event prediction. In order to demonstrate the strengths of our modeling tools, we utilize synthetic syndromic surveillance data [10] 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 [10] 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 EARS C2 alerts 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 6 (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 hotspot view, Figure 6 (Top-Middle and Top-Right), of the actual patient counts versus the predicted patient counts, a different story unfolds. The analyst can immediately identify several areas where a larger number of patients were seen than expected. More information can be garnered by overlaying the SatScan cluster output onto the maps. Now the analyst can immediately find the general location of an outbreak occurring on July 19th coinciding with the area where an outbreak was injected. Furthermore, the time series alerts from the EARS C2 algorithm, yellow diamonds in Figure 6 (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 7 (Top). The analyst can scroll...
Fig. 6. 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-Middle) A hotspot view of actual patients versus the number of predicted patients. Here the analyst can quickly see which counties have exceeded the predicted values. (Top-Right) A finer granularity hotspot view with uncertainty overlaid as a texture. Both (Top-Middle and Right) also utilize the SatScan cluster overlay feature. (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 SatScan circle.

back in time prior to when the first alert was generated in the hospital’s time series and begin looking at the estimated patient density. Notice that the estimated density remains consistent in the northwest portion of the state 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 6 (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). Furthermore, the analyst could also utilize the temporal change view and observe how the outbreak began over time. Figure 7 (Bottom) shows the magnitude of temporal changes. Note that the analyst can immediately see large changes occurring in the state on any given day. It is only through a combination of linked views and interactive displays that more knowledge and confidence can be gained to allow them to accurately assess an outbreak.

In the case of retrospective and reactive analyses, our system provides an analyst with a means to form hypotheses. In this case, the analyst may look at the hotspot view and notice that several counties are indicating problems. A hypothesis is then generated by the analyst of the form “There may be a health outbreak in region X.” This hypothesis can then be tested through the use of SatScan, providing the analyst with a significance test, or through the linked views and temporal models which also provide alerts based on a significance value. Further, the analyst can scroll back in time and watch the outbreak develop, and through the use of finer granularity density estimation begin getting a picture of the exact area of infection as opposed to the purely circular region denoted by the spatial scan statistic.

The analyst may also then choose to look at the ecotope view of the data. Figure 5 shows a clustered view of the July 19th data, clustering on respiratory, gastro-intestinal, neurological and constitutional syndromes. Here, one can see a light pink ecotope grouping that would largely overlap the shapes seen in the July 19th pictures of Figure 7. This type of mapping along with the related information can enhance the information the analyst receives by providing them with information on what other syndromes may be found in that area, allowing them to refine their hypothesis from “An outbreak is occurring in area A.” To “An outbreak of disease X is occurring in area A.”
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 8 (Top) the analyst is viewing December 31th, 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. Figure 1 shows a low level SatScan alert for the area first appears on December 29th, 2007. In the time series window of Figure 8 (Top), 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. This hypothesis that the analyst has generated is further corroborated by the fact that SatScan has also detected a disease outbreak near Montgomery.
Fig. 8. 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 SatScan circle in Indiana covers the approximate area of injection. (Top) The predictive visual analytics system. Note the time series view contains an upper and lower bound 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.

County as well as an outbreak in northern Kentucky. Note that the Kentucky outbreak is a false positive in this case, demonstrating the need for investigation into these alerts.

With this information, the analyst may also wish to corroborate these findings with the spatial models. Figure 8 (Bottom) illustrates the temporal changes for the hotspots across Indiana from December 31, 2007 through January 3, 2008, and Figure 9 illustrates the hotspot view by county from December 31, 2007 through January 2, 2008. The analyst can immediately see a hotspot in an area near the hospital. When tracking the hotspot into the future, the levels remain consistent or even increase on January 1st and 2nd and begin to taper off on the 3rd, as can be seen by the color scale legend representing the temporal change. 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 spatiotemporal views for enhanced exploration and data analysis through the use of
Fig. 9. Here we show the county level hotspot view for analyzing the future ramifications of the December 31st outbreak. (Left) December 31, 2007. (Middle) January 1, 2008. (Right) January 2, 2008. Note that it is now the county south of Montgomery that is showing signs of the outbreak persisting.

traditional choropleth maps with various granularities for more accurate event detection. Further, the extension of the kernel density estimation to 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 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. Other needs include better modalities for uncertainty visualization, as well as enhancements to the ecotope view for better data analysis.

ACKNOWLEDGMENTS

The authors would like to thank the Indiana State Department of Health and the Regenstrief Institute for their input on system development and user feedback, particularly Shaun Grannis for his feedback in this project. The Cancer Care Engineering project is supported by the Department of Defense, Congressionally Directed Medical Research Program, Fort Detrick, MD (W81-XWH-08-1-0065) and the Regenstrief Cancer Foundation administered jointly through the Oncological Sciences Center at Purdue University and the Indiana University Simon Cancer Center. This work has been funded by the US Department of Homeland Security Command, Control and Interoperability Center of Excellence...

REFERENCES


Ross Maciejewski is a postdoctoral researcher in electrical and computer engineering at Purdue University. His research interests include nonphotorealistic rendering, volume rendering, and visual analytics. Maciejewski has a PhD in computer engineering from Purdue University. Contact him at rmacieje@purdue.edu.
Ryan Hafen is a PhD student in statistics at Purdue University. His research interests include exploratory data analysis and visualization, massive data, computational statistics, time series, modeling, and non-parametric statistics. Hafen has a masters in statistics from the University of Utah. Contact him at rhafen@purdue.edu.

Stephen Rudolph is a masters student in electrical computer engineering at Purdue University. His research interests include casual information visualization and visual analytics. Rudolph has an MS in computer engineering from Purdue University. Contact him at sru-dolph@purdue.edu.

Stephen G. Larew is an undergraduate student in electrical computer engineering at Purdue University. Research interests include GPGPU computing and visual analytics. When he is not diligently researching and studying, Stephen likes to play sports, read, and practice piano. Contact him at sglarew@purdue.edu.

Michael A. Mitchell is an undergraduate student in electrical computer engineering at Purdue University. His research interests include casual information visualization and visual analytics. Mike is a member of the National Society of Collegiate Scholars and the Alpha Lambda Delta Phi Eta Sigma honors society. Contact him at mamitch@purdue.edu.

William S. Cleveland is the Shanti S. Gupta Distinguished Professor of Statistics and courtesy professor of computer science at Purdue University. His research interests include statistics, machine learning, and data visualization. Cleveland has a PhD in statistics from Yale University. He is the author of The Elements of Graphing Data (Hobart Press, 1994) and Visualizing Data (Hobart Press, 1993). Contact him at wsc@stat.purdue.edu.

David S. Ebert is a professor in the School of Electrical and Computer Engineering at Purdue University, a University Faculty Scholar, director of the Purdue University Rendering and Perceptualization Lab, and director of the Purdue University Regional Visualization and Analytics Center. His research interests include novel visualization techniques, visual analytics, volume rendering, information visualization, perceptually based visualization, illustrative visualization, and procedural abstraction of complex, massive data. Ebert has a PhD in computer science from Ohio State University and is a fellow of the IEEE and member of the IEEE Computer Societys Publications Board. Contact him at ebertd@purdue.edu.