A Visual Analytics Approach to Understanding Spatiotemporal Hotspots

R. Maciejewski, S. Rudolph, R. Hafen, A.M. Abusalah, M. Yakout, M. Ouzzani, W.S. Cleveland, S.J. Grannis, and D.S. Ebert

IEEE Symposium on Visual Analytics Science and Technology (VAST), 2009

Eric Shook (eshook2@uiuc.edu)

Outline

- Introduction and Background
  - Time series analysis methods
- Spatial methods
  - Spatial aggregation methods
  - Heatmaps (Kernel Density Estimation)
  - Coloring maps
- Temporal methods
  - Time series analysis (Cumulative Summation)
  - Temporal contours
  - Multivariate views

Introduction

What is a spatiotemporal hotspot?

Introduction

What is a **spatiotemporal** hotspot?

A phenomena that is both
spatially and temporally explicit

Introduction

What is a spatiotemporal **hotspot**?

An aberration or irregularity

Introduction

What is a **spatiotemporal hotspot**?

An aberration in space and/or time

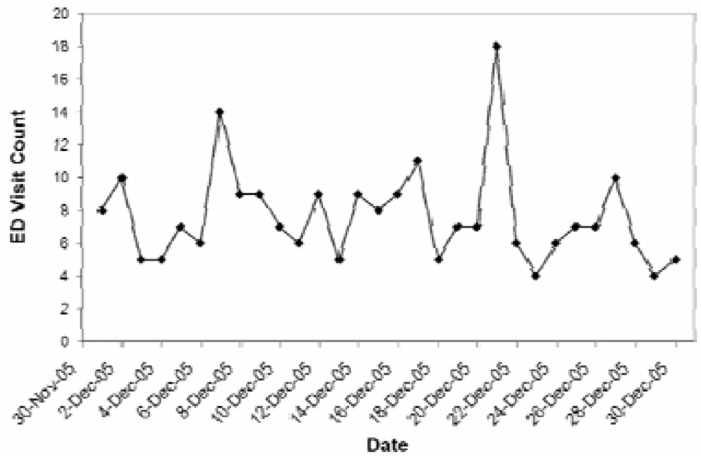
2 Spatiotemporal Case Studies

- Law enforcement data (Crime)
  - Traffic violations, criminal activities, etc.
    - categorized data
  - location, time, and description of incidence
  - (Source: West Lafayette Police Department)
- Syndromic surveillance data (Health)
  - respiratory, gastrointestinal, rash, fever, etc.
    - categorized data
  - hospital, date, and category of complaint
  - (Source: Indiana State Department of Health)

Related work in time series visualization

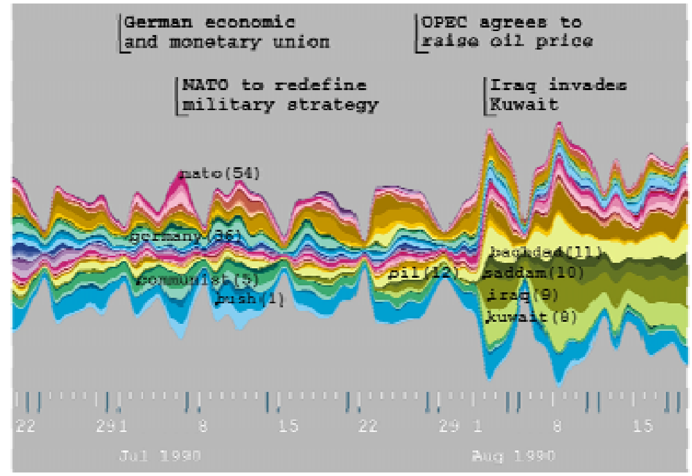
Difficulties in time series analysis

Time series data is often most difficult to analyze



December 21 peak is significant, but often difficult to distinguish

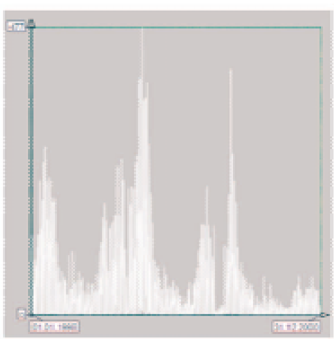
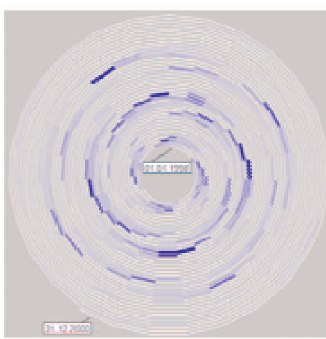
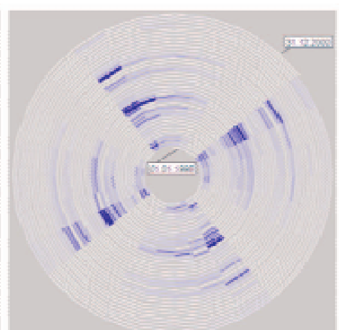
Theme River



100,000 documents from Associated Press newswire

(Havre, et al. 2000)

Spiral Graph

Time Series

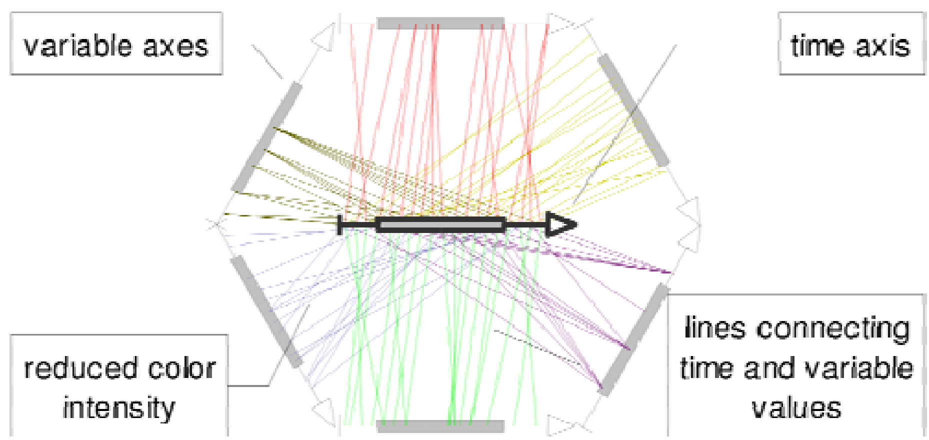
Spiral Graph
27 day cycle

Parameterized Spiral Graph
28 day cycle

Number of influenza cases over 3 years

(Aigner, et al. 2008)

Time Wheel



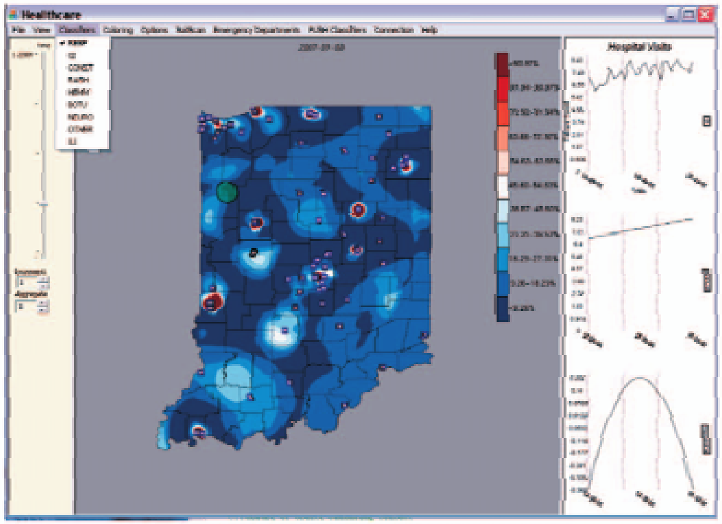
(Tominski, et al. 2004)

Visual Analytic Environment

Features

- Dually linked interactive displays for multivariate exploration and analysis
- Novel data aggregation for privacy preservation
- Kernel density estimation method
- Interactive color mapping tools to enhance data contextualization
- Control charts for identifying temporal signal alerts
- Multivariate exploration combining contours and colors
- Spatiotemporal history via contour line ghosting
- Demographic filter controls to enable database query and analysis through a simple graphical interface
- Thresholding data to analyze specific trends
- Region selection tools for analyzing area specific hotspots

Linked spatial and temporal views



Privacy Issues

Health and Crime data = Sensitive Information

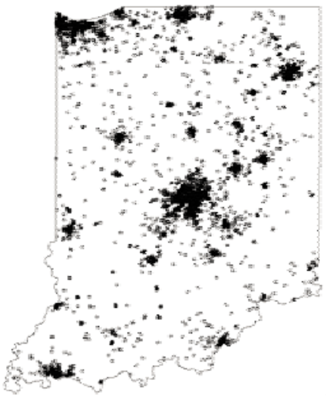
Cannot always show exact location

Spatial aggregation is a solution to this problem

Spatial Aggregation

3 Methods

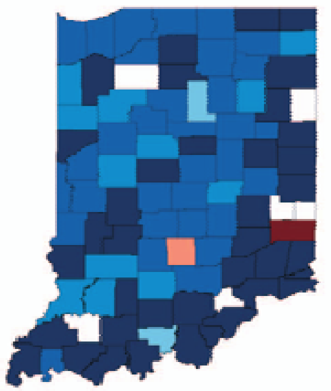
Additive opacity circles



Additive opacity circles

- Aggregation method
  - Semi-transparent circle
- Color
  - Black
- Advantage
  - High resolution
- Disadvantage
  - Cannot distinguish density levels between areas

County

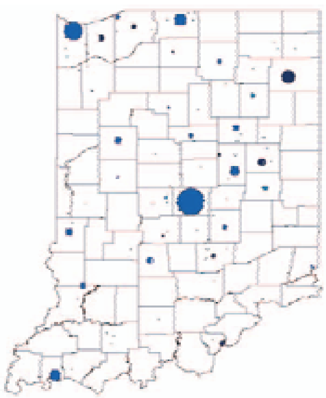


County

- Aggregation
  - County
- Color *
  - Ratio mapped to a color:
    % syndromatic pop
    total pop
- Advantage
  - Complete spatial coverage
- Disadvantages
  - Low population can cause false positives
  - Low resolution

* The mapping of colors will be discussed later

Nearest neighbor grouping

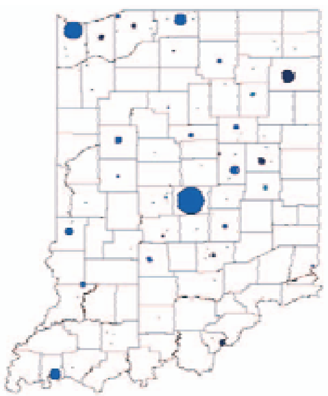


Nearest neighbor grouping

- Aggregation
  - Nearest neighbor grouping
- Color *
  - Ratio mapped to a color:
    % syndromatic pop
    total pop
- Advantage
  - Easy to identify problem areas at finer spatial resolution
- Disadvantage
  - Assume data is clumped

* The mapping of colors will be discussed later

Nearest neighbor grouping



Nearest neighbor grouping

Operation

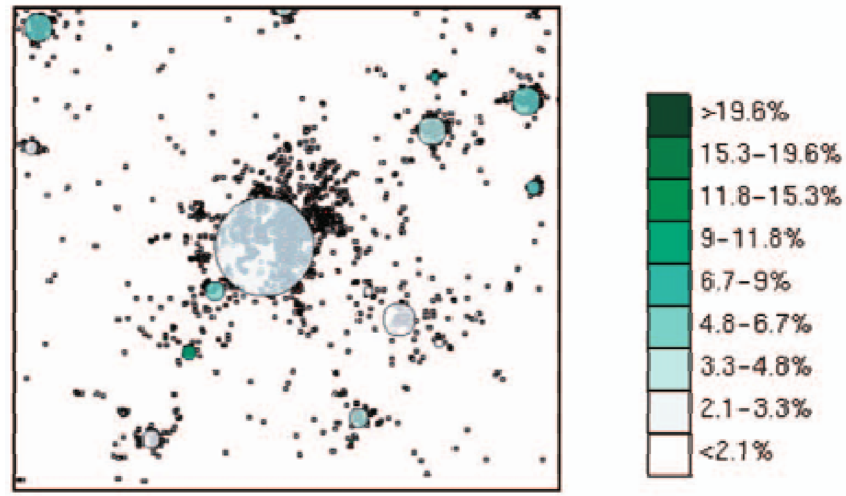
1. Define distance (d)
2. Find all sets of events that are less than distance (d) apart

foreach set (s) {

- calculate centroid (c) of s
- set circle size proportionate to size of set s
- color s based on percentage of syndromic/total population
- place circle on map at centroid c

}

Example of nearest neighbor grouping



Aggregation Method Comparison

Good overall view, but cannot approximate trends

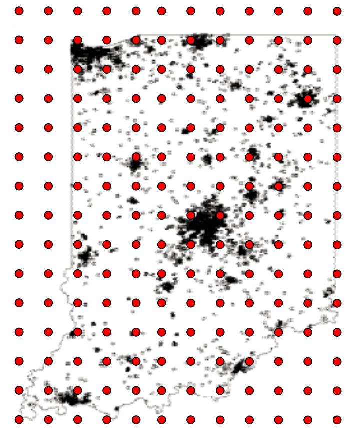
Heatmaps

Kernel Density Estimation

- Fit a weighting function (kernel) to a collection of sample points
- A kernel is a symmetric integrable function
- Tunable bandwidth parameter adjusts the magnitude of the kernel

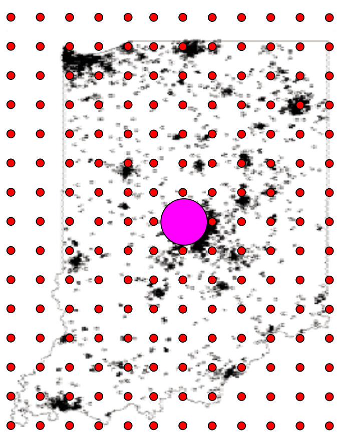
(Silverman, B.W., 1986)

Kernel Density Estimation Operation



Lay fine grid of sample points

Kernel Density Estimation Operation

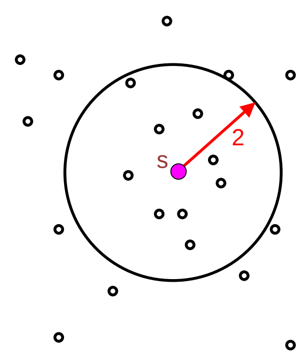


Lay fine grid of sample points

foreach sample point (s) {
- search for points within the bandwidth area
- apply kernel function to all points within bandwidth
- color point according to value calculated by kernel

}

Basic uniform kernel example



$$\lambda(s) = \frac{N(s,\tau)}{\pi \tau^2}$$

$$\lambda(s) = \frac{9}{\pi 2^2} = 0.716$$

= 0.716

(Slide material used with permission from Dr. Sara McLafferty)

Simple Kernel Function

$$\lambda(s) = \sum_{d_j \le \tau} \frac{1}{N} \frac{1}{\tau^2} \mathrm{k}(d_j/\tau)$$

Where:

λ(s) = estimate density at grid point s
d_j= distance from point s_j to grid point s
τ = bandwidth (kernel size, search radius)
k( ) = kernel weighting function
N = total number of samples

(Slide material used with permission from Dr. Sara McLafferty)

Kernel function comparison

1. Uniform

$$k() = \frac{1}{2}$$

2. Quartic

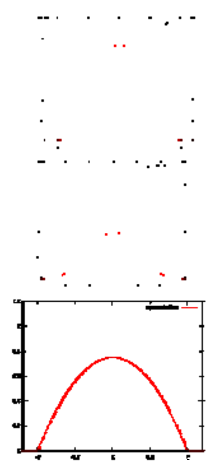
$$k() = \frac{3}{\pi\tau^2}\left(1 - \frac{d_i^2}{\tau^2}\right)^2$$

3. Gaussian

$$k() = \frac{1}{\sqrt{2\pi}} e^{-\frac{d_i^2}{2\tau^2}}$$

4. Epanechikov

$$\mathrm{k}() = \frac{3}{4}\left(1 - \frac{d_i^2}{\tau^2}\right)$$



(Slide material used with permission from Dr. Sara McLafferty)

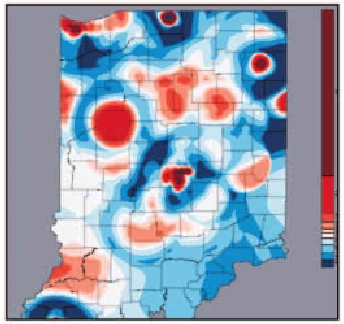
Multivariate kernel density estimation

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

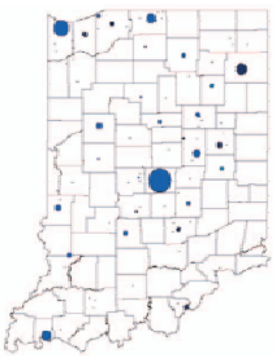
$$K(\mathbf{u}) = \frac{3}{4}(1 - \mathbf{u}^2)1_{(\|\mathbf{u}\| \le 1)}.$$

h bandwidth or multidimensional smoothing parameter
N total number of samples
d data dimensionality (2)
$1_{(||u||<=1)}$ evaluates to 1 if inequality is true, 0 otherwise (if distance < bandwidth)

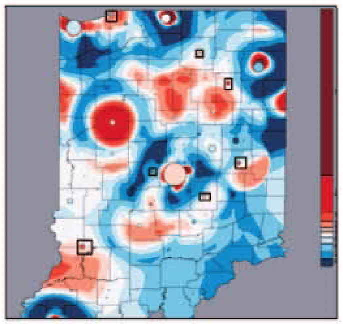
Heatmap and Aggregation Comparison



Heatmap

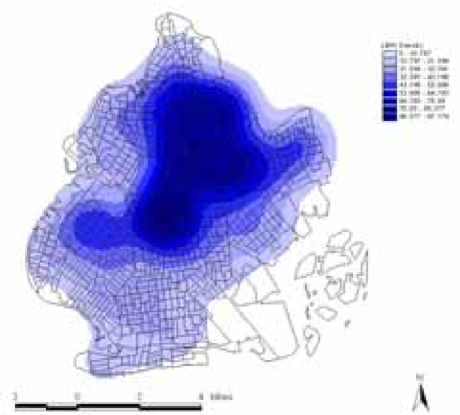


Nearest Neighbor Grouping

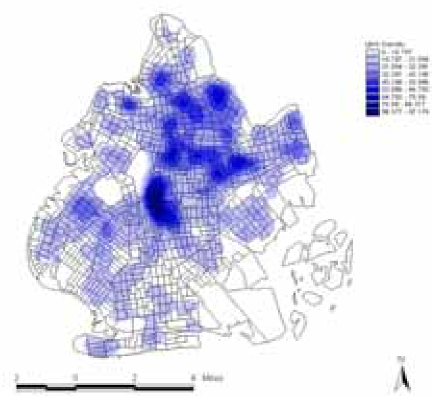


NN Grouping over heatmap

Bandwidth Sensitivity Comparison



2.0 mile bandwidth



1.0 mile bandwidth

Example is low-birth weight density and is not from the paper

(Slide material used with permission from Dr. Sara McLafferty)

Fixed vs. variable kernel

- Fixed Kernel
  - Large bandwidth over smoothes data while trying to accommodate sparse data regions
  - Small bandwidth unable to handle data in sparse regions
- Variable Kernel
  - Adjust bandwidth size according to sparsity of the local data
  - Use k nearest neighbor distance

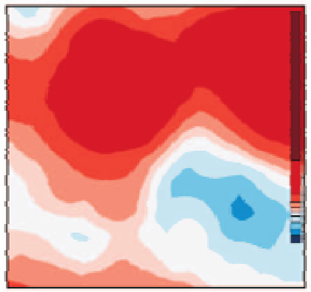
Variable kernel

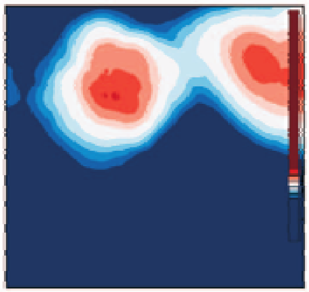
$$\hat{f}_h(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\max_{(h,d_{i,k})}} K\left(\frac{\mathbf{x} - X_i}{\max_{(h,d_{i,k})}} \right)$$

Guarantee a minimum bandwidth of h

Importance of Coloring

- Data ranges mapped to color bins
  - Choice of bins ideally based on model assumptions, but can be difficult to quantify
  - Make it easy or difficult to identify hotspots or nested hotspots
- Different binning techniques
  - User selected
  - *Mathematical*





Same example with different coloring

Mathematical Color Binning Methods

- Linear
  - Each color represents an equal number of points
- Ramp
  - Each color represents an increasingly larger number of points
  - Follows: y=x
- Exponential
  - Follows: $y=e^x$
- Logarithmic
  - Follows: y= log (x)

Temporal

Time Series Analysis

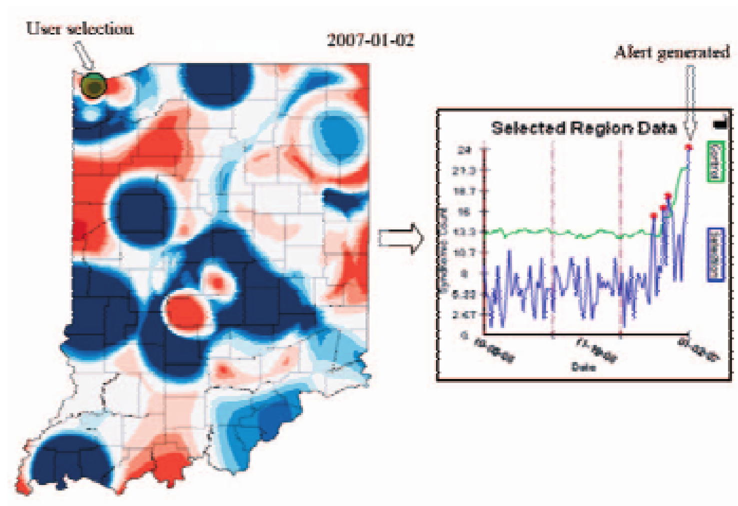
- Cumulative Summation (CUSUM)
  - Standard epidemiological algorithm
  - Provides alert for potential outbreaks

$$S_0=0$$
$$S_t = max\left(0, S_{t-1} + \frac{X_t - (\mu_0 + k\sigma_{x_t})}{\sigma_{x_t}}\right)$$

X_t count at time t
μ_0 expected value (mean)
k detectable shift from mean
σ_{xt} standard deviation from mean

(Hutwagner, L.C., et al. 2003)

CUSUM Example

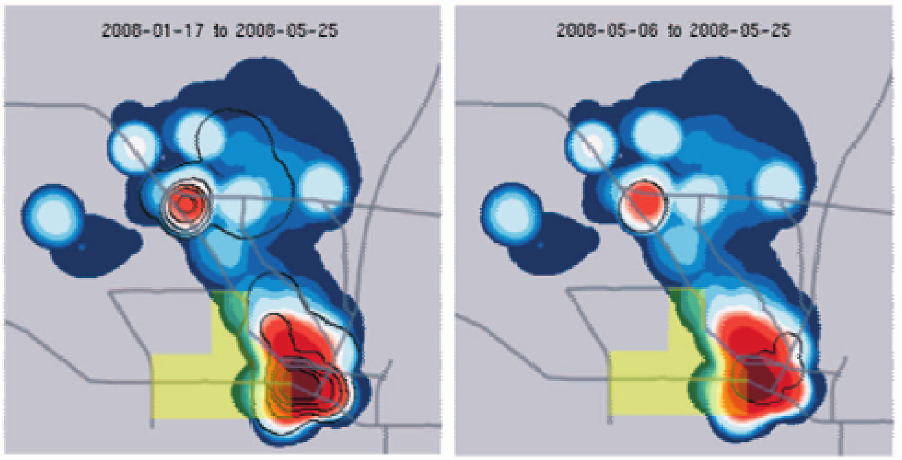


μ_0 Expected value (Control)
s_t CUSUM

Temporal Contours

Contour lines preserve temporal history

Color: all crimes for a year



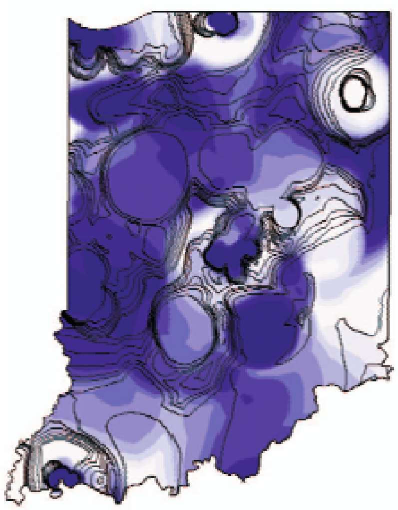
Contour: Spring semester thefts

Contour: Last 20 days of spring semester thefts

Multivariate View

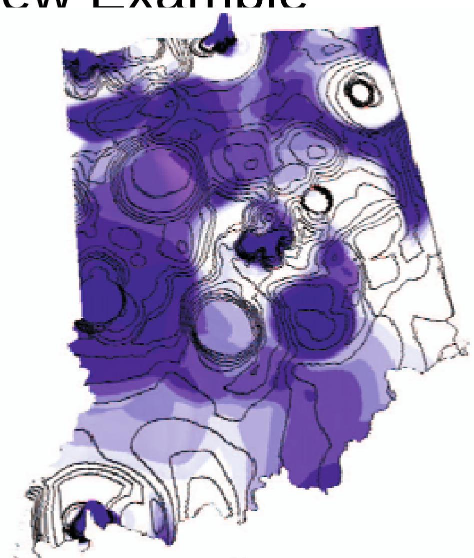
- Multivariate views enable a series of complex viewing modalities
  - Up to 3 variables
    - Color
    - Contours
    - Height
- Users can search for correlations between these variables simultaneously

Multivariate View Example



(a)

Color: Shock/Coma cases
Contours: Rash cases



(b)

Height: Respiratory cases

User Interactivity

- Brushing (Becker, R.A. & Cleveland, W.S., 1987)

  Limit analysis and/or visualization to user selected information
- Highlight

  Select information in one view and highlight it in another view
- Interactive Thresholding

  Only display data that meets a threshold value or limit
- Temporal Contour Ghosting

  Track movement of hotspots across time

Interactive Thresholding Example

Original Heatmap | Filter using threshold | View trends while filtering using temporal ghosting

Conclusions

Many benefits to using visual analytics for understanding syndromic hotspots

- Linking geospatiotemporal view with traditional time series views
  - Enhances exploration and data analysis
  - Enhances hypothesis generation
- Kernel density estimation
  - Provides finer granularity heatmap compared to spatial histogram

Conclusions

- Spatial aggregation methods
  - Provide a solution to privacy concerns
  - Provide a good overall view
  - But cannot approximate trends
- Kernel Density Estimate
  - A solution to visualizing hotspots
  - Sensitive to bandwidth parameter
  - Fixed and variable kernels can be used to fit a variety of problems and spatial contexts
- Color provides context for data

Conclusions

- Cumulative Summation (CUSUM) method
  - Useful for identifying potential outbreaks
  - User configurable alert mechanism
- Temporal contours
  - Preserve temporal history while color map provides context
- Multivariate views
  - Add a third dimension to search for correlations or “cosyndromes”
- User interactivity

References & Acknowledgements

- Hutwagner, L.C., et al. "The bioterrorism preparedness and response early aberration reporting system (EARS)", J. Urban Health, vol 80, no. 2, 2003.
- Silverman, B.W., *Density Estimation for Statistics and Data Analysis.* Chapman and Hall/CRC. 1986.
- W. Aigner, S. Miksch, W. Muller, H. Schumann, and C. Tominski, "Visual Methods for Analyzing Time-Oriented Data," IEEE Trans. Visualization and Computer Graphics, vol. 14, no. 1, pp. 47-60, 2008.
- S. Havre, E. Hetzler, and L. Nowell, "ThemeRiver: Visualizing Theme Changes over Time," Proc. IEEE Symp. Information Visualization (InfoVis '00), pp. 115-123, Oct. 2000.
- C. Tominski, J. Abello, and H. Schumann, "Axes-Based Visualizations with Radial Layouts," Proc. ACM Symp. Applied Computing (SAC '04), pp. 1242-1247, 2004.
- R.A. Becker and W.S. Cleveland, "Brushing Scatterplots," Technometrics, vol. 29, no. 2, pp. 127-142, 1987.

- Some slide materials were used with permission from Dr. Sara McLafferty of the Department of Geography, University of Illinois at Urbana-Champaign.

Thank you

Questions?