Exploring Clusters in Geospatial Datasets

Darya Filippova
University of Maryland
College Park
dfilippo@umd.edu

Joonghoon Lee
University of Maryland
College Park
jhl@cs.umd.edu

Andreea Olea
University of Maryland
College Park
avolea@cs.umd.edu

Michael VanDaniker
University of Maryland
College Park
mvandani@umd.edu

Krist Wongsuphasawat
University of Maryland
College Park
kristw@cs.umd.edu

ABSTRACT

Analyzing multivariate geospatial data is a non-trivial task. Geographical location is a variable in itself that might be correlated with other fields in the data. When looking at the variable distribution in a geographical region, users often need to identify areas where the variable is overrepresented. We call such local maxima "hotspots" and present Fervor: a hotspot explorer. Fervor visualizes hotspots as a continuous heat map and offers a range of tools to investigate relationships within the underlying data. We adopted and expanded the rank-by-feature framework [12] to quantify the strength of relationships between the different fields describing the data. In this paper, we use Fervor to investigate traffic incident data in the state of Maryland.

1. INTRODUCTION

Traffic Management Centers log hundreds of traffic incidents each day, and exploring the data to garner any significant meaning can be a daunting task. The usual process of determining high accident locations begins with a request from a citizen who has already noticed a number of accidents in a particular location. The traffic engineers query all accidents associated with that location and determine a course of action. Such an approach is rather passive; conditions must be bad enough that citizens complain to the agency. Clearly, traffic engineers need a special method to proactively search for accident hotspots.

We propose Fervor, an application that helps users explore the geospatial data. To accomplish this task, Fervor provides a set of visualizations, each of which brings the clusters forward and separates them from background noise. In the context of traffic management, clustered data points are indicative of high accident locations. Users can filter the data to focus on a reduced set of accidents. As users eliminate data, Fervor keeps all visualizations synchronized. The primary method for exploring the data is through an interactive map. By default, the dataset is rendered as a heat map on top of the map, but users may switch to an icon mode where each icon corresponds to a data point. The heat map draws attention to hotspots by rendering locations with many data points in intense colors while displaying locations with low data counts in dim colors. Points on the heat map bleed into each other, so contiguous data points have more emphasis. While the heat map is sufficient to find locations with the most accidents, Fervor's statistical analysis tools help deduce the likely cause of accidents. Parallel coordinate plots draw attention to data elements with similar properties (whether an accident occurred at night or in the rain, etc.), and a rank-by-feature toolbox provides a systematic method for exploring the data. The ranking criteria in the current literature are suitable only for numerical variables. Since traffic accident data consists mainly of categorical variables, this paper proposes some categorical data ranking criteria.

The remaining sections discuss Fervor in more detail. Section 2 discusses related work. Section 3 presents Fervor's interface; the subsections detail the different components, outlining its utility and value. We discuss future work and potential enhancements in section 4, and conclude in section 5. All examples use the Maryland traffic incidents dataset.

2. RELATED WORK

GIS tools, like ArcGIS, are highly specialized to work with topographic data, street and historical maps, but their primary goal in not data analysis. Other tools require data to be aggregated by a geographical area such as district, county, state, country. Tools like CommonGIS [2] and GeoVista [8] provide visualizations that range from simple state to state comparisons to statistical data analysis. However, aggregating accident counts by higher level regions like state and county would not reveal any information about the most dangerous intersections and, therefore, would not help traffic engineers make their roads safer.

Commercial products such as Spotfire [14] and Tableau [15] include mapping capabilities. Spotfire generates maps from a raster image overlaid with data points. An advantage of such approach is that users can map any dataset onto any image, however, manually aligning a thousand data points can be tedious task. Tableau uses a built-in map and users have to include latitude and longitude in their data to gen-

erate a map chart. However, both tools show data points as single icons which results in occlusion and overcrowding for large datasets. For traffic incidents data, locations with few accidents would look the same as locations with many overlapping accidents.

With the release of public APIs by Google, MSN, and Yahoo, mapping resources have become available to every aspiring programmer. Wood, et al. explore population density data using Google Earth in conjunction with other open-source products [4]. The Google Earth API represents a single entry in the data as a pushpin on the map. To deal with overcrowding, Google Earth by default collapses coincident pushpins into one icon, but provides no a visual indication the pushpin is really a cluster. The user should click the pushpin to reveal the underlying data points. This obstacle can be avoided by changing color and size for collapsed pushpins, however, the users are stuck with the discrete picture. At the same time, large pushpins, indicative of a cluster, tend to overshadow nearby smaller pushpins and obscure the view of the rest of the map.

2.1 Heat maps

Heat maps are an ideal solution for representing dense spatial data: they reveal the high-occurrence areas without obscuring the general view. Heat maps are not a novel idea and are used in cartography ubiquitously. Terrain elevation data, for example, is traditionally rendered as a heat map. Elevation is a continuous value, so it maps nicely to the continuous gradients that underlie heat map algorithms. Recently, there has been a push towards the use of heat maps to represent other sorts of spatial information [5][10] [6]. Hotmap [5] visualizes geographical database logs: the records contain information about the map tiles requested by MSN Maps users from all over the world; the tiles that were requested more often than others are assigned a brighter shade of red in Hotmap. Tiles that are requested the most were of more interest, therefore, the satellite images for those locations have to be updated more often reflecting any changes in the physical appearance. Fabien Girardin's project is similar to Hotmap: using GeoIQ and the Flickr API, he built a heat map of Barcelona's geotagged images [6]. Mehler, et al. use heat map to track the news origin of stories published on the Internet [10]. Universal Mind's LaunchPad [7] visualizes crime activity using heat maps.

2.2 Rank-by-feature Framework

The traffic incidents dataset is composed of multidimensional data. Dealing with multidimensionality has been a challenge to researchers in many disciplines due human's inability to sufficiently comprehend more than three dimensions while searching for relationships, outliers, clusters and gaps. This challenge is well recognized and has been dubbed "the curse of high dimensionality." Seo and Shneiderman [12] present a conceptual framework for relationship detection called the rank-by-feature framework and demonstrate its capabilities in the Hierarchical Clustering Explorer (HCE). HCE ranks all possible axis-parallel projections against the selected criterion and presents the result in a color-coded grid. However, the ranking criteria in HCE are only applicable to numerical variables. Many variables in the traffic incident dataset are categorical (for example, weather field: "dry", "rainy", "cloudy"), which indicates the need for categorical variable ranking. Continued work [11] on HCE suggests estimating the relative significance between two categorical variables. Seo and Gordish-Dressman address the problem by suggesting one ranking criterion based on chi-square test. This paper further explores relationships between categorical variables and discusses new ranking criteria.

3. INTERFACE

Fervor follows the general guidelines given by Shneiderman's mantra [13]: overview, zoom and filter, details on demand. The majority of the screen space is allotted to the map. To the left of the map are a rich set of filters that allows users to narrow down the dataset. The details panel in the bottom of the screen shows all entries in the dataset. Selecting rows in the details panel highlights corresponding data points on the map and other components and vice versa, as suggested in [9]. The map features two modes: icon mode and a heat map mode. Users can switch between the two using the controls along the top of the map. A smaller window in the lower right corner of the map shows a zoomed out version of the area, providing a global perspective of the main map's current position. The map excels at conveying geospatial clustering; other visualizations are available to explore other clustered aspects of the data set. Fervor's histograms provide traditional one-dimensional analysis but also allow for a systematic approach to drilling down on temporal clusters. Parallel coordinates plots can be used to analyze data distribution across all fields simultaneously, and the statistics panel provides a semantic method for explorer the data set based on the rank by feature framework.

3.1 The Map

Many common information visualization tools support mapping functionality, but many only provide it in a roundabout way. If users want to plot geospatial data in Spotfire DXP, for example, they must find an accurate map and specify the range of the map in latitude and longitude. For any image other than a projected globe these values can be difficult to configure because of the vast number of mapping projections available. For a local problem like accident hotspot detection, a global map is not particularly useful. Even with a properly configured image, multiple zoom levels are not supported gracefully in Spotfire. Users are required to use the standard filtering controls (sliders, checkboxes, etc.) to negotiate the map. While this does not limit users from performing any standard tasks, it is a rather clunky method for dealing with geospatial data.

Like many GIS tools, Fervor treats latitude and longitude as first class data elements. Each accident is plotted on a built in map, taking advantage of the significance of the geospatial nature of the data. Latitude and longitude sliders are available in the filters panel for clipping unnecessary data. The primary method for exploring interesting locations is through zooming and panning. Points can be rendered as icons, making details about each element available at the click of a mouse. Clicking on an icon brings up a details window displaying values for each of the data item's fields. Users can also choose to display the dataset in heat map mode, giving the impression of the volume of the data in the region. An opacity slider allows users to configure their

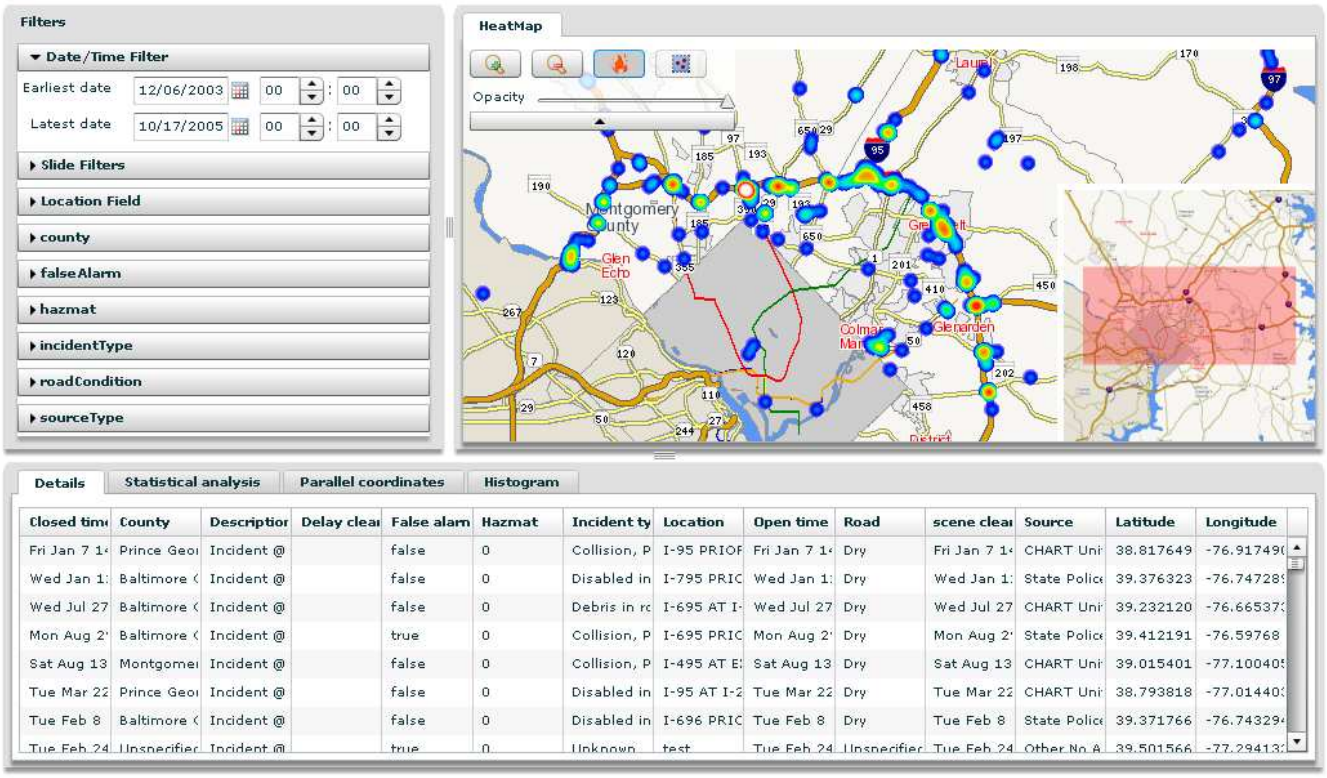

Figure 1: Fervor Interface

display so they can see underlying roads while still benefiting from either the icons or the heat map.

3.1.1 Heat Map Mode

Rendering each accident on the map as an icon is reasonable in certain domains but introduces several problems when the task at hand involves looking for the densest regions of the map. In icon mode, each point is exactly the same size as every other point, so isolated icons have an equal emphasis when compared to icons that contribute to hotspots. Occlusion is a major issue when dealing with nearby points, especially when the map is at the farthest zoom levels. A region with only a few points can look like it has the same number of points as a dense region. Heat mapping solves this problem by assigning each point a sphere of influence that dissipates when moving away from the point. Spheres of influence have an additive effect on each other; the more influence any single pixel is assigned, the more intense the pixel will be colored. Figure 3 shows two maps with identical datasets rendered in icon mode (left) and heat map mode (right). Because of occlusion in icon mode it is difficult to tell which of the four circled regions has the most points. With heat mapping enabled it is clear that the region on the top right has more points than the other three regions.

3.1.2 The Heat Map Algorithm

Our heat map algorithm is inspired by the work of Corunet [3], a software company which uses heat maps to inform clients about which areas of their webpages receive the most clicks. The procedure they outline generates images on the server side, and then uses JavaScript to place the images on top of the target webpages. We adopted this methodology to work on the client machine in the Flash Player runtime environment.

The initial step in the algorithm is to build an array of 256 RGBA values that will serve as the heat map's color scheme. "Hot" locations on the map are colored with values towards the end of the array while "cooler" locations receive their colors from the beginning. The image shown in figure 4 is a rendering of that array.

Fervor generates the color scheme array by transitioning through a gradient consisting of the distinguishable colors in figure 4 (blue, cyan, green, yellow, orange, red, and white). These colors were selected because their order is associated with increasing temperature, and the relatively large number of distinct colors allows for pre-attentive differentiation between regions on the map. Highly saturated colors were selected because they contrast well with the relatively pale background colors of the underlying map.

The latitude and longitude values for each point p in the dataset are converted to screen coordinates (s_p) using Chuck Taylor's geographic/UTM conversion algorithm [16]. Depending on the zoom level, this may map nearby p's to the same s_p values. The maximum number of identical s_p values, $maxcount$, is computed and stored for later use.

Each point is drawn onto a bitmap as a circle of radius r where r is linearly dependent on the zoom level. The color of the circle fades along a gradient from black to blue when

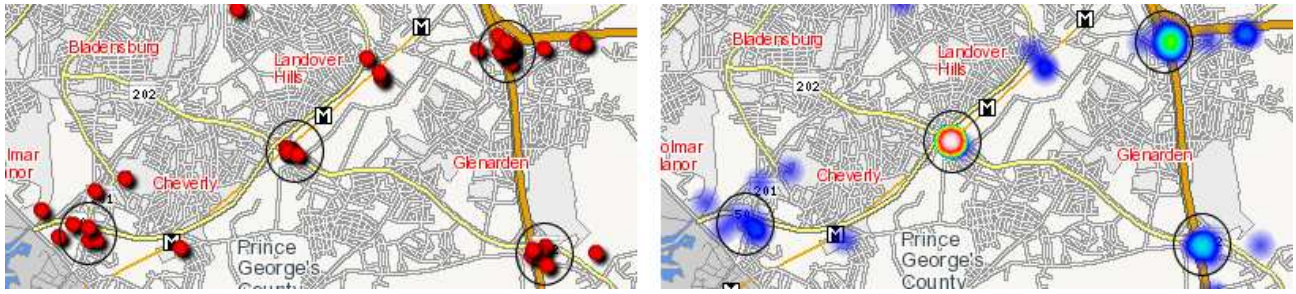

Figure 2: In the heat map on the right the circled region in the center of the map has many more points than any other region. Due to occlusion, the icon map on the left does not make this immediately obvious, and users may even be misled to believe that one of the other circled regions contains the most points.

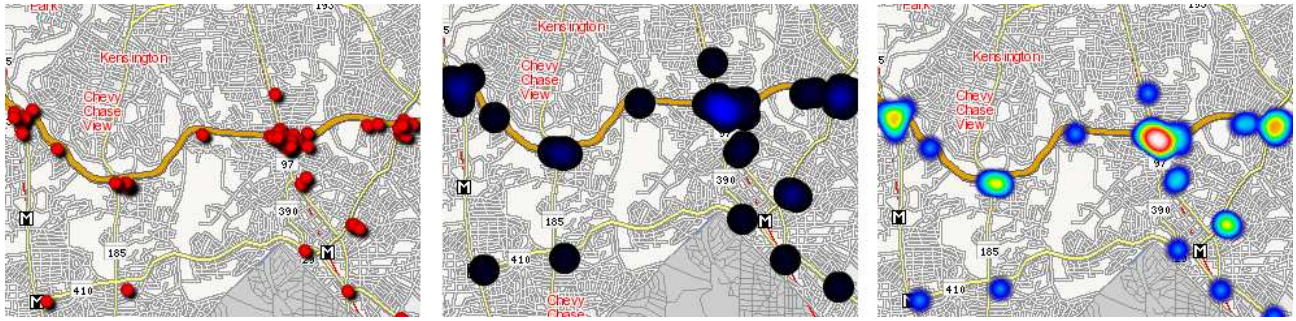

Figure 3: From left to right, a map rendered in icon mode, the same map rendered in heat map mode without the colorization process, and the fully rendered heat map. Note that the center image cannot be generated in the actual application – it is shown here for explanatory purposes only.

Figure 4: The array of 256 colors to be used for the heat map color scheme. The checkered pattern on the left side of the image indicates that the colors in that region are not fully opaque.

moving from the edge of the circle towards the center.

The exact shade of blue used in the center is computed as $blue_{max}$ by the formula

$$blue_{max} = \max(15, 255/\max(5, maxcount))$$

The value of the blue channel at each pixel is used as a weight later in the algorithm where 255 is the highest possible weight. The formula for $blue_{max}$ is designed so that no single point receives too much weight while enforcing a limit on the minimum. The screen blend mode [1] is used when drawing circles to the bitmap.

The final step in the process is to recolor each pixel in the bitmap with a value from the heat map color scheme. The value of the blue channel at each pixel (ranging between 0 and 255) is used as an index into the color scheme array, and the extracted value replaces the old color. Figure 3 shows the various stages of heat map generation.

3.2 The Histogram Panel

A histogram is a graphical way to visualize a series of labeled numerical frequencies. While the heat map allows users to explore clusters in the latitude and longitude fields of their dataset, the histogram panel gives users the ability to explore temporal and categorical clusters in the data.

This is particularly useful in the task of inferring the cause of traffic accident clusters, because traffic patterns are correlated with conditions in the area and time of day. One might expect more accidents from 8:00 to 9:00 AM than in the early hours of the morning, or more accidents during rainy weather conditions rather than during sunny days. The histogram panel enables users to explore variations and clusters in the number of incidents. For example, one can examine the change in incident numbers over time by displaying incident occurrences in a monthly, weekly, daily or hourly fashion. Figure 5 illustrates a histogram containing accidents spread over the course of an entire year.

3.2.1 Histogram Zooming

Temporal data is represented as a hierarchy of values (minutes, hours, days, months, years), and Fervor histograms support a logical way for interacting with this hierarchy. For viewing temporal incident information, the initial view displays the correlation between incident numbers and the month in which those incidents occurred. Clicking on a particular month zooms in on the information pertaining to that particular month and displays it categorized by which day of the week it occurred in. Users can further zoom in

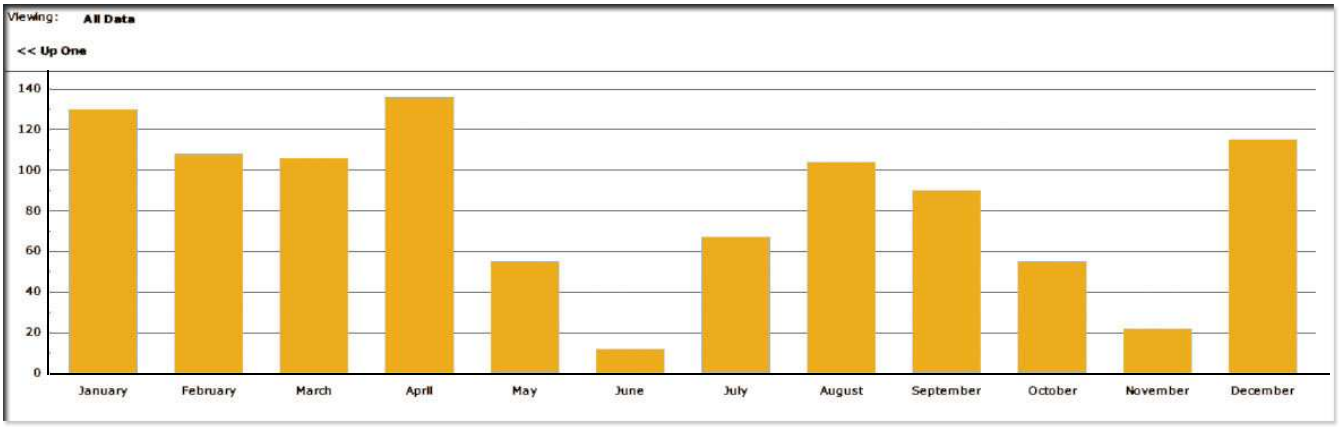

Figure 5: Fervor's histogram panel showing yearly accident information

on a particular day and view the number of incidents spread over a 24-hour period. Furthermore, users have the capability to zoom out at any time and select a different temporal category to focus on.

Zooming can sometimes become confusing as users might forget which category they had previously zoomed in on, so we incorporate drill-down and drill-up animation effects. For instance, if the user exploring weekly data zooms in on Tuesday information, the histogram will first only show the Tuesday column placed in its slot among the entire week, to indicate which day is currently being zoomed into, and then the information about Tuesday accidents will be spread across the entire chart and subdivided into hourly intervals to further examine the different times of the day when accidents may have occurred. Going back, the hours of the week columns will compress together to build one total column in the Tuesday slot, and only after that the other days will reappear.

3.2.2 *Breadcrumb Navigation*

Breadcrumbs are a navigation technique for keeping track of one's location within a series of views. Displayed as a horizontal list of labels, breadcrumbs link back to previous views users have explored, ending with the current display. Since temporal data has hierarchical significance, with multiple levels of zooming in or out corresponding to the varying granularity of temporal categorization, users need a simple way to keep track of what data is currently shown as well as an easy way to navigate back and forth through their trail of histograms. For this purpose, Fervor's histograms use a breadcrumb trail to display the path down to the category that is presently being shown (figure 6). A text field at the top of the histogram shows what information is being viewed together with the zooming path that lead to that particular view. For instance, if the view went from the entire dataset to April and then further focused on Tuesday, the breadcrumb trail will display the path "All Data : April : Tuesday", where clicking on any one labeled category in the breadcrumb trail will zoom out all the way to that category. When zooming in, the selected category will be appended to the breadcrumb trail.

3.3 Parallel coordinates

Parallel coordinate plots (Figure 7) are a way to visualize multi-dimensional data on a plane. The plots have as many vertical axes as the dataset has different fields. A record is represented as a line that goes across all axes; on each axis the line goes through the point corresponding to the record's value for that field. Because overcrowding may become an issue in large datasets, Fervor gives users control over the line thickness and transparency.

Using parallel coordinate plots, it is easy to see the clusters in the data: if many lines converge one point on an axis, then many data points in the dataset have that particular value in common for that variable. On the other hand, if there are only a few lines going through other points on the axis, the values corresponding to these lines are outliers in the original dataset. Examining outliers and clusters may open new insights into the data.

Parallel coordinate plots in Fervor can display both numerical and categorical fields as well as dates and timestamps. During a session with Fervor, users will most likely want to select items on the map and then investigate the items on the parallel coordinate plots. Because of brushing and linking, selected items on the map will be selected on the plot as well. If the plot has a high concentration of points in a particular column, it would be reasonable to assume those accident locations and that property are correlated. This brings users one step closer to deducing the cause of accident hotspots.

3.4 The Statistics Panel

Facing "the curse of high dimensionality", Fervor adopts the idea of rank-by-feature framework from the Hierarchical Clustering Explorer (HCE) [12]. HCE provides several methods for ranking numerical-numerical variables relationships. The traffic incidents dataset consists of numerical (N), date-time (D) and categorical (C) variables. Hence, possible relationship types are N-N, D-D, N-D, C-D, C-N, and C-C variables; most of the relationships being of type C-C. The HCE ranking criteria does not provide meaningful results when applied to categorical variables. Thus, we propose two approaches to deal with categorical variables. First, we suggest converting the data to numerical values and applying methods for ranking N-N relationships. The

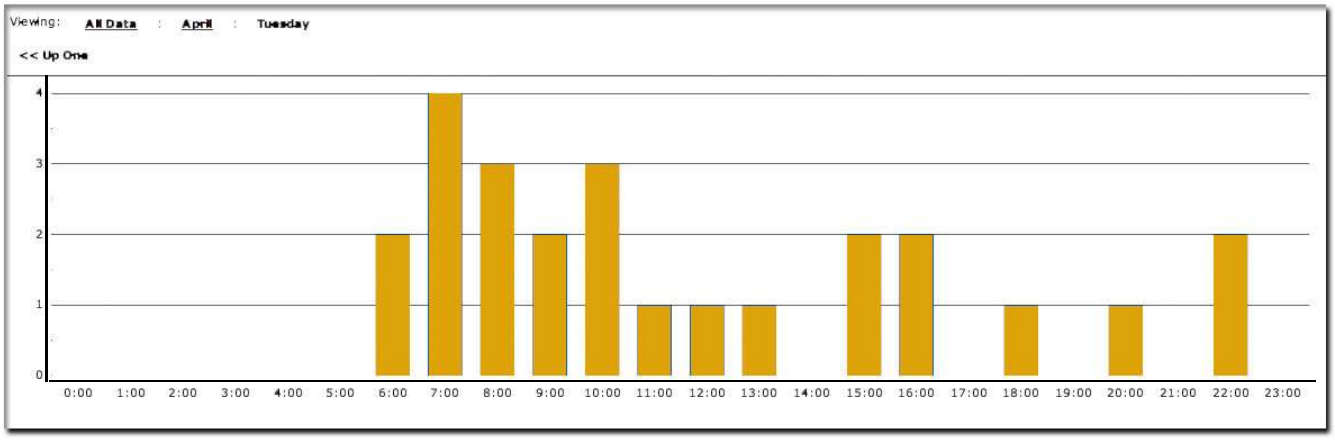

Figure 6: Fervor's Histogram panel zoomed in up to hourly information. Notice breadcrumb trail at the top provides links to day of the week, month name and full dataset

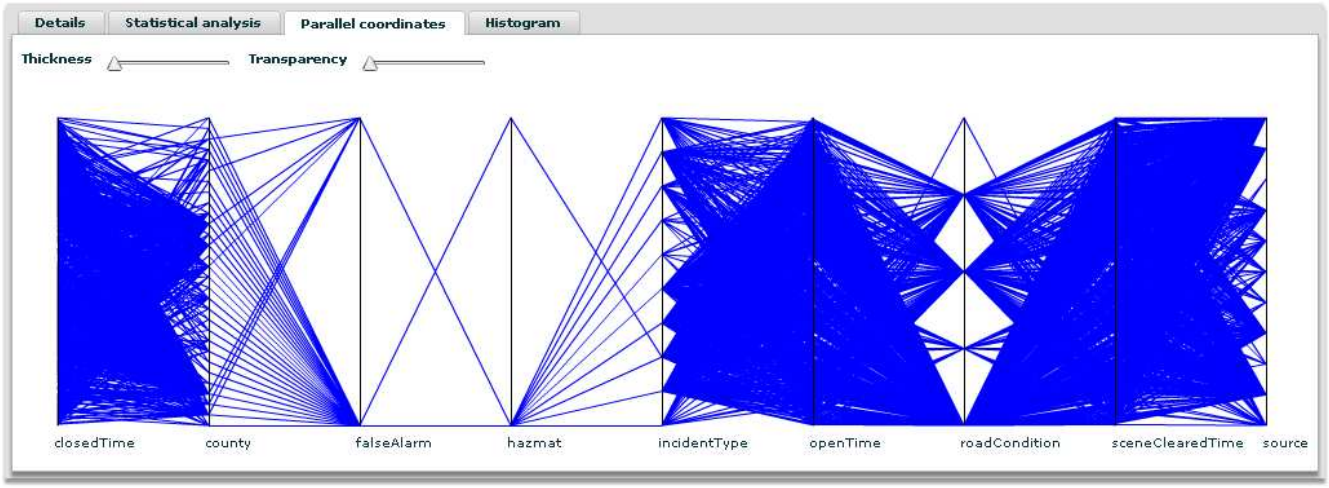

Figure 7: Parallel coordinates plot

following section describes our attempts to do this. Second, we propose ranking criteria for C-C and C-N relationships, which will be described in section 3.4.2.

3.4.1 Converting Categorical Data into Numerical Data

In this section, we discuss the idea of converting categorical variables into numerical variables. Depending on the data field, we applied to different methods.

1. Order categorical values

   Ordinal variables are categorical variables that can be ordered by some criteria. By ordering the possible values of a categorical variable, a numerical value can be obtained: the 'rank' of the value. Consider an example where, road condition (dry, rain, snow) can be ranked by the relative risk of driving in that condition. A dry would be safest, whereas a rainy road would be more slippery, but not as much as when it snows. Thus, one may assign increasing numerical values 1, 2, 3 to represent each road condition. Introducing a systematic or scientific method for assigning appropriate numerical values to each categorical value is an issue that is left for future work. Such ordering may not always be representative of the value's significance. For instance, there are three different types of collisions found in the dataset. It is not clear whether a "property damage" collision is considered as more severe than a "personal injury" collision, although the third level, "fatality" is clearly the most severe.

2. Derive numerical data that is inherent in existing variables

   The "duration" of an incident could be derived, as the difference between the "open time" and "clear time". Such derived numerical data could be used in the rank-by-feature framework, just as if it were in the original dataset.

   Both approaches require a certain level of domain knowledge of the original dataset. In the first approach, there must be a meaningful measure that can be used to order the variables, and in the second approach, knowledge of what can be derived from the existing data is necessary.

   It is also unclear how missing data should be interpreted. Currently, Fervor uses subjective ordering in such cases. Using statistical methods or domain expertise to assign better values and compute meaningful statistics is an issue left to explore. By converting categorical variables into ordinal variables and representing them as numerical values, ranking criteria can be applied for numerical-numerical relationships with those where at one or both of the fields is categorical. This way statistical relations such as Correlation Coefficients or Least Square Errors that use numerical values can be computed.

3.4.2 Ranking Relationship Between Categorical Variables

Continued work from HCE presents one ranking criterion for C-C relationships, a contingency coefficient based on the chi-square test. However, only one ranking criterion may be not flexible enough for all users. Allowing other options for ranking improve the flexibility of the rank-by-feature framework. Our approach is based on counting the number of incidents in each relationship first. Let X be the first categorical variable and Y be another categorical variable. X has n possible values (X_1 to X_n) while Y has m possible values (Y_1 to Y_m).

O_{ij} is the number of incidents in which $X = X_i$ and $Y = Y_j$

Calculate all O_{ij} for each pair of categorical-categorical variables.

$$O = \{O_{ij} \text{ for all } i \text{ in } [1, n] \text{ and } j \text{ in } [1, m]\}$$

We then propose the following ranking criteria based on the number of occurrences to rank categorical-categorical relationships.

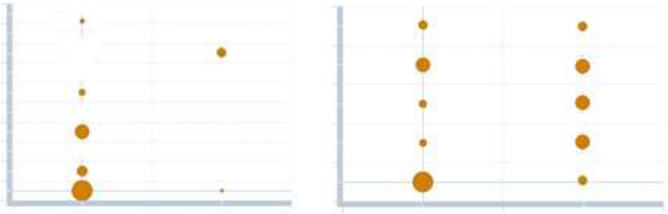

Figure 8: High vs. Low Percentage of Empty Area

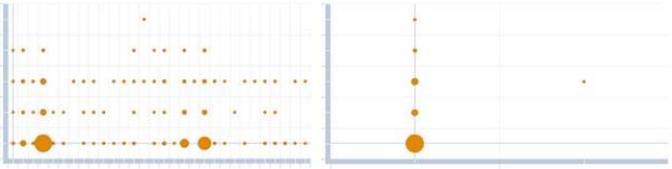

Figure 9: High vs. Low Distinct Number of Occurrences with equal percentage of empty area.

1. Percentage of Empty Area (0 to 100)

   The criterion is to sort scatter plots in terms of percentage which number of occurrences (O_{ij}) are zero. A higher percentage suggests that the scatter plots are more pruned. Users can use this criterion to easily isolate sparse scatter plots from dense scatter plots.

2. Distinct Number of Occurrences (0 to n)

   One scatter plot with 4 out of 10 pairs of relationship between possible values and another one with 25 out of 100 will have the same percentage of empty area. Hence, the second criterion is presented to distinguish them. It is the number of occurrences (O_{ij}) which are zero. A lower number suggests that there are fewer pairs of existing relationship between possible values in the two categorical variables. Relationships which have small number of existing pairs might spot some interesting points.

3. Maximum Number of Occurrences (0 to n)

   Ranking by maximum number of occurrences gives priorities to scatter plots which contain at least one spot

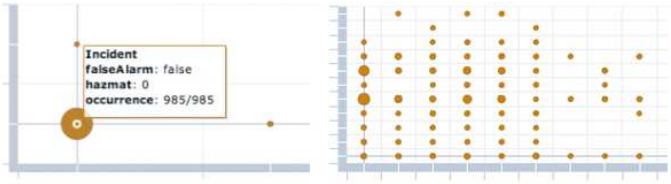

Figure 10: High vs. Low Maximum Number of Occurrences

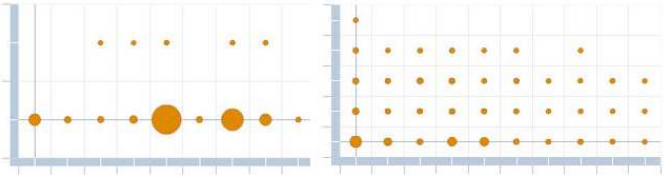

Figure 11: High vs. Low Standard Deviation of Distinct Number of Occurrences

which has high number of occurrences. This allows users to easily identify dense areas in the dataset, a key theme for Fervor.

4. Standard Deviation of Distinct Number of Occurrences (0 to n) This criterion is calculated from standard deviation of number of occurrences (O_{ij}) which are more than zero. It represents how widely spread the numbers of occurrences in scatter plots are. A high value represents widely spread number of occurrences while a low value represents a more fairly distributed number of occurrences. Users can easily identify scatter plots where incidents are not fairly distributed.

5. Number of Potential Outliers (0 to n) Fervor's outlier detection is based on Z-Scores. We calculate Z-Scores for each number of occurrences (O_{ij}) with non-zero values using the standard deviation calculated in the same way with previous criterion. This potential outlier detection define a pair of possible values from one categorical variable to another as an outlier if the Z-score calculated from number of occurrences (O_{ij}) is lower than -1.5 or higher than 1.5.s.

3.4.3 Statistical Panel Interface

The interface is organized from left to right by using order as shown in Figure 13. Users can start by selecting a ranking criterion from the drop down menu on the left. The table in the middle then shows rankings of relationships according to the selected criterion with color labels. When a row in the grid is selected, the scatter plot on the right displays the relationship on the plot. Users can view more details of each data point by moving their cursors over each circle in the scatterplot. Users can also swap the axes by clicking on "Swap Axes" button. This is useful when the displayed data is shown better with the inverse orientation.

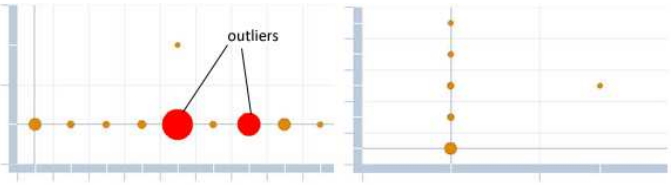

Figure 12: Two selected scatterplots order by Number of Potential Outliers - 2 (left) and 0 (right)

3.5 Sample workflow example

When the user starts the application, Fervor loads the traffic incidents dataset. The map shows a heat map of all accidents and the user can immediately identify that Washington DC and Baltimore Beltways have the most accidents. On the filters panel, the user can deselect a checkbox for disabled vehicles thus removing those incidents from the dataset. Now the user visually identifies a hotspot and zooms in on it. As the user zooms in, the hotspot visually breaks into two separate clusters. The user selects one of those clusters using the selecting tool (available along the top of the map). Clicking the "Show only" button, the user tells Fervor to dismiss all data points except those corresponding to the selected incidents. Now the user wants to explore the data points in the cluster, so he uses the parallel coordinates plot to get an overview of the data distribution. He might discover that there were more accidents on a sunny day than any other. The statistical panel will help him find out whether there are other confounded variables correlated with the weather field. During this interaction, the user may discover that the majority of these events occurred in the westbound direction. Finally, the user can analyze the temporal information for this cluster in the histogram. The user might find that the majority of the accidents he selected all happened at 6pm in the westbound direction. From the data he discovered, the user could logically infer that this hotspot is likely due to the hindered vision of drivers heading directly towards the setting sun.

4. EVALUATION

To get expert feedback on Fervor, we demonstrated the tool to Michael Pack, Center for Advanced Transportation Technology (CATT Lab) director. CATT Lab is a University of Maryland research facility that works with various incident and traffic data. CATT Lab's main accent is on visualizing the data and developing custom tools for analyzing it. Applications developed at CATT Lab are used by Maryland and Virginia state agencies responsible for road safety, traffic and incident management. One of the Fervor's authors gave Michael Pack a tour of the tool explaining the proposed workflow and introducing the features. We recorded the comments our expert user had during the meeting and had a discussion with him after the demonstration. Michael Pack's comment on the heatmap visualization was that while heatmaps represent the overral distribution of a value in the region quite well, heatmap for accidents was somewhat misleading since it often aggregated accidents on the different roads together due to the fact the accidents on the two roads were situated close together. We are working on heatmap algorithm improvements so that the user can have control of the heatmap granularity. Michael Pack was impressed with a number of supporting analytical components that the tool provided and was particular intrigued by the statistical panel since it provided a way to explore relationships within data. We are in the process of developing the ground for a more extensive user study. We would like to show the tool

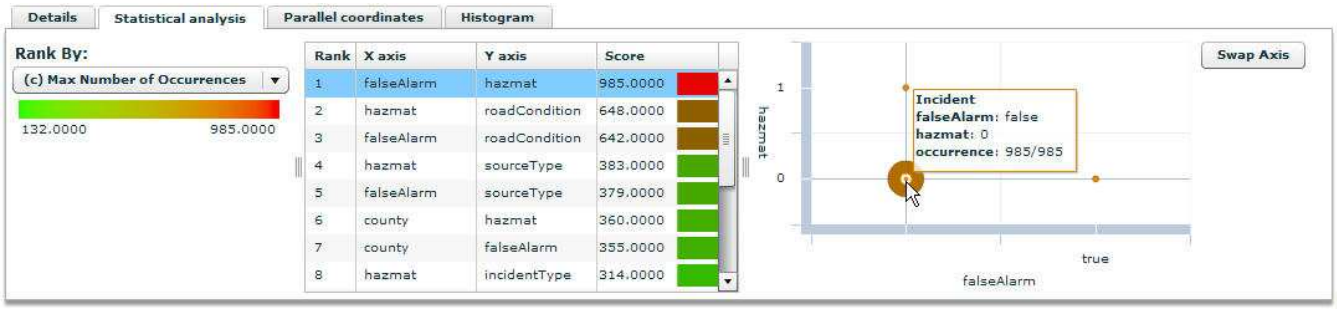

Figure 13: The Statistical Panel, ranked by maximum number of occurrences. It suggests that relationship between hazardous material and road condition has some dense areas. From the scatter plot, it shows that most of the incidents occurred when the road condition is dry and when no hazardous material was present.

to traffic engineers that have to deal with road safety on the daily basis. To explore the applicability of the tool for other datasets (for example, a dataset showing the availability of a particular service in the region), we have to make several changes in the tool and we are working on that right now. We are looking for multivariate geospatial datasets to extensively test the tool as well.

5. FUTURE WORK

Currently Fervor is tailored to the incident dataset described throughout this paper. In the future we would like to make this tool publicly available and allow users to upload and explore their own spatial datasets.

Because Fervor is designed to investigate clusters in data, we would like to adopt features from the Hierarchical Clustering Explorer (HCE) and other similar projects. The dendrogram view and the minimum similarity bar are two such features that were out of the scope of this project but may make useful additions to the application. The spiral [17] is a useful way to explore clusters in data, especially data of a temporal nature; we will consider including a spiral explorer in future versions.

Fervor explores clusters, or local maxima. It would be interesting to investigate areas where there is a dearth of data, or local minima. Inverting the heat map, or producing a "cold map", seems an obvious solution, however, judging by the incidents dataset inverting the heat map will make a larger part of the map a hotspot. Overlaying a "cold map" with a map of locations of interest (for example, all cities) might resolve this issue.

Another interesting visualization would be a heat map difference. For example, one dataset can represent availability of the service and another dataset can contain the service demand data. Looking at the difference of the two, the users can see where the service is needed the most.

Fervor proposed the idea of ranking relationships between categorical variables and numerical variables by converting the categorical variables into values. The conversion is done by ranking the categories by an arbitrary standard. However, there can be many alternatives to rank the categories. Thus, we believe that having an interface that allows users to customize the ranking while using the tool would make the software more flexible.

The ranking criteria for categorical variables in Fervor are based on the number of the value occurrences for each pair of fields. Applying statistical inference methods and data mining algorithms to rank relationship between categorical variables may yield more accurate results.

6. CONCLUSION

Fervor is not meant to be the ultimate geospatial information visualization tool but rather a lightweight application that facilitates finding dense regions of a map and inferring the relationships within the data behind it. Heat maps allow users to explore clusters in latitude and longitude using the familiar metaphor of a map. While the proximity of points in standard icon-based maps obscures the view and hides some icons behind others, heat maps successfully overcome these problems. The drill-down approach, e.g. selecting a cluster on the map and zooming in on it, allows users to explore local data. The statistics panel provides advanced exploration of relationships between variables. The statistical panel also builds upon the rank-by-feature framework by introducing a novel method of dealing with categorical data. Breadcrumb navigation in histograms makes it easy to reveal temporal clusters at varying degrees of resolution. Our work does not suggest that Fervor can replace an actual visit to the high-accident site, however we hope that Fervor can serve traffic engineers as an accident hotspots explorer.

7. ACKNOWLEDGEMENTS

We would like to thank Michael Pack and Ben Shneiderman for their guidance, Justin Grimes, Behjat Siddiquie, Cody Dunne, and Prahalad Rajkumar for their critical eyes, and Ken Knudsen for his contributions to this project.

8. REFERENCES

[1] Adobe:master transparency and blends, 2008. `http://www.adobe.com/designcenter/indesign/articles/idsn3kbtrans/idsn3kbtrans.pdf`; Last accessed: 04-20-2008.

[2] N. Andrienko and G. Andrienko. Interactive visual tools to explore spatio-temporal variation. In *AVI '04: Proceedings of the working conference on Advanced visual interfaces*, pages 417–420, New York, NY, USA, 2004. ACM.

[3] Corunet: How to make heat maps. august 6, 2006. http://blog.corunet.com/english/how-to-make-heat-maps; Last accessed: 04-20-2008.

[4] J. Dykes, A. Slingsby, and K. Clarke. Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1176–1183, 2007.

[5] D. Fisher. Hotmap: Looking at geographic attention. *Transactions on Visualization and Computer Graphics*, 13(6):1184–1191, 2007.

[6] F. Girardin. Heat map of barcelona geotagged images, 2006. http://www.girardin.org/fabien/blog/2006/12/01/heat-map-of-barcelona-geotagged-images/; Last accessed: 03-30-2008.

[7] Universal mind, "universal mind: Demos". http://www.universalmind.com/demo/launchpad.cfm; Last accessed: 05-05-2008.

[8] A. MacEachren, X. Dai, F. Hardisty, D. Guo, and G. Lengerich. Exploring high-d spaces with multiform matrices and small multiples. *IEEE Proceedings of the Symposium on Information Visualization*, 2003.

[9] A. R. Martin and M. O. Ward. High dimensional brushing for interactive exploration of multivariate data. In *VIS '95: Proceedings of the 6th conference on Visualization '95*, page 271, Washington, DC, USA, 1995. IEEE Computer Society.

[10] A. Mehler, Y. Bao, X. Li, Y. Wang, and S. Skiena. Spatial analysis of news sources. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):765–772, 2006.

[11] J. Seo and H. Gordish-Dressman. Exploratory data analysis with categorical variables: An improved rank-by-feature framework and a case study. *International Journal of Human-Computer Interaction*, 23(3):287– 314, December 2007.

[12] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, 2005.

[13] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. *IEEE Conference on Visual Languages (VL'96)*, pages 336–343, 1996.

[14] Spotfire. http://www.spotfire.com; Last accessed: 04-20-2008.

[15] Tableau. http://www.tableausoftware.com; Last accessed: 04-20-2008.

[16] C. Taylor. Geographic/utm coordinate converter. http://home.hiwaay.net/~taylorc/toolbox/geography/geoutm.html; Last accessed: 04-20-2008.

[17] M. Weber, M. Alexa, and W. Müller. Visualizing time-series on spirals. In *INFOVIS '01: Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01)*, pages 7–13, 2001.