A New Collaborative Tool for Visually Understanding National Health Indicators

Songhua Xu, Brian Jewell, Chad Steed, Jack Schryver

Oak Ridge National Laboratory
Oak Ridge, TN, USA 37831
{xus1, jewellbc, steedca, schryverjc}@ornl.gov

ABSTRACT

We propose a new online collaborative tool for visually understanding national health indicators, which facilitates the full spectrum of investigation of indicators, from an overview of all the correlation coefficients between variables, to investigation of subsets of selected variables, and to individual data element analysis. This tool is publicly accessible at http://cda.ornl.gov/heat/heatmap.html. In this paper, we discuss the key issues regarding the interface design and implementation. We also illustrate how to use our interface for analyzing the health indicator dataset by showing some key system views. In the end, we introduce and discuss some ongoing research efforts extending this work.

Keywords: National health indicators warehouse, visual analytics, heatmap, scatterplot, online collaborative analysis platform

1 INTRODUCTION

The National Center for Health Statistics develops the Health Indicators Warehouse (HIW) website, http://healthindicators.gov/, as a platform for publishing thousands of national health indicators and their values. Some indicators only provide national level data while others provide detailed individual Hospital Referral Regions (HRR), county and state level data for specific demographic groups of the population as characterized by people's age, race, and many additional sociodemographic factors. Some indicators are also dedicated to profiling patients of specific diseases. Overall, these indicators provide a rich source of data for informing the public about the general health care quality and population health conditions in a certain region. These data resources can also support health policy decision making and analysis on various levels, ranging from counties, states, to the federal administration.

Despite the promising value potentially offered by the warehouse, a major problem remains in effectively discovering knowledge from such a comprehensive data source. Some key challenging tasks include: 1) how to examine values of these indicators and comparatively study value differences across multiple states, counties, and HRRs; 2) how to understand value correlations between multiple indicators and detect regions in the country where such correlations behave abnormally; 3) how to understand health condition and healthcare quality, as characterized by multiple indicators, for the same region as well as these indicators' relative standing in the country.

Drawing answers to the above questions can effectively support health policy makers and legislators to change existing healthcare regulations and amend the healthcare laws for providing more affordable healthcare with better care outcome. Unfortunately, existing mainstream business intelligence tools failed to provide user friendly solutions for policy researchers and makers to freely and effectively explore answers for the above questions and many variants of these questions.

2 BACKGROUND AND RELATED WORK

Continuing technological advances have resulted in increasingly complex multivariate data sets which, in turn yield information overload when explored with conventional visual analytics techniques. The ability to collect, model, and store information is growing at a much faster rate than our ability to analyze it. However, the transformations of these vast volumes of data into actionable insights is critical in many domains, such as health care. Without proper techniques, analysts are forced to discard layers of information in order to fit the tools; therefore, new approaches are necessary to turn today's information deluge into opportunity.

One promising solution for this challenge lies in the continued development of techniques in the realm of intelligent user interfaces. The intelligent user interface adapts and learns from the user's interaction with the data to adjust levels of detail and highlight potentially significant associations among sets of interrelated variables. Like the related field of visual analytics, intelligent user interfaces combine the strengths of humans with those of machines. While methods from knowledge discovery, statistics, and machine learning drive automated analytics and augment the display, human capabilities to perceive, relate, and conclude strengthen the iterative process.

Over the years, there have been many approaches to the visual analysis of multivariate data (see Wong and Bergeron, 1997). However, the techniques employed in most operational systems are generally constrained to non-interactive, basic graphics using methods developed over a decade ago; and it is questionable whether these methods can cope the with complex data of today. For example, analysts often rely on simple scatter plots and histograms which require several separate plots or layered plots to study multiple variables in a data set. However, the use of separate plots is not an ideal approach in this type of analysis due to perceptual issues described by Healey et al. (2004) such as the extremely limited memory for information that can be gained from one glance to the next. These issues are illustrated through the so-called change blindness phenomenon (a perceptual issue described by Rensink (2002)) and they are exacerbated when searching for combinations of conditions.

One approach often used by statisticians to overcome this issue is to use the scatterplot matrix (SPLOM), which represents multiple adjacent scatterplots for all the variable comparisons in a single display with a matrix configuration (Wong and Bergeron, 1997); but the static SPLOM requires a large amount of screen space and forming multivariate associations is still challenging.

3 APPROACH

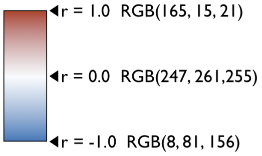
To address the limitations of using existing methods for analyzing the national health indicators, we designed and developed a novel collaborative tool for visually understanding values and relationships of indicators in HIW. In the current work, we have enhanced the traditional SPLOM by providing cues that guide and refine the analyst's exploration of the information space. This approach is akin to the concept of the scented widget described by Willett et al (2007). Scented widgets are graphical user interface components that are augmented with an embedded visualization to enable efficient navigation in the information space of the data items. The concept arises from the information foraging theory described by Pirolli and Card (1999) which models human information gathering to the food foraging activities of animals. In this model, the concept of information scent is identified as the "user perception of the value, cost, or access path of information sources obtained by proximal cues" (Pirolli and Card, 1999).

Our approach is perhaps closest to that proposed by Friendly (2002), who subsumed his suggestions to visualization of correlation matrix data in color- and shape-coded formats under the name "corrgram." Their study introduced several interesting alternatives to the display of sign and magnitude of correlation coefficients, and novel visualization techniques for summarizing higher-order shape or trend information in scatterplots.

In the current work, the focus is on a particular set of health care indicators from the Health Indicators Warehouse (HIW). The HIW hosts a number of indicators for national, state, and community level statistics and serves as a data repository for the Department of Health and Human Services (HHS) Community Health Data Initiative. We sampled a collection of variables from this repository to provide an effective visual interface to explore and discover significant associations between health indicators. We focused our efforts on a set of CMS indicators from HIW that contain values at the HRR level, augmented by multiple CMS component cost variables downloaded from the Institute of Medicine website http://iom.edu/Activities/HealthServices/GeographicVariation/Data-Resources.aspx.

Our system, which is publicly accessible at http://cda.ornl.gov/heat/heatmap.html, facilitates the full spectrum of investigation of indicators, from an overview of all the correlation coefficients between variables, to investigation of subsets of selected variables, and to individual data element analysis. This capability is realized through a level-of-detail algorithm that includes more detail as the analyst zooms into the display. Initially, the analyst is given an overview of the variable correlations in the form of a heatmap visualization of the correlation matrix. Correlation mining is an important data mining technique due to its usefulness in identifying underlying dependencies between variables. The correlation mining process attempts to estimate the strength of relationships between pairs of variables to facilitate the prediction of one variable based on what is known about another. The linear relationship between two variables *X* and *Y* can be estimated using a single number, *r*, that is called the sample correlation coefficient (Walpole and Myers 1993). Our correlation matrix display uses the Pearson product-moment correlation coefficient to measure the correlation between two variables. Given a series of n measurements of *X* and *Y* written as x_i and y_i where $i = 1, 2, \ldots, n$, *r* is given by

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}}.$$



For each pair of variables in the display, the system computes *r* which results in a correlation matrix. The correlation matrix is a symmetric *n* by *n* matrix where each *i, j* element is equal to the value of *r* between the *i* and *j* variables. The intersection of the variables is represented graphically as a color-filled square.

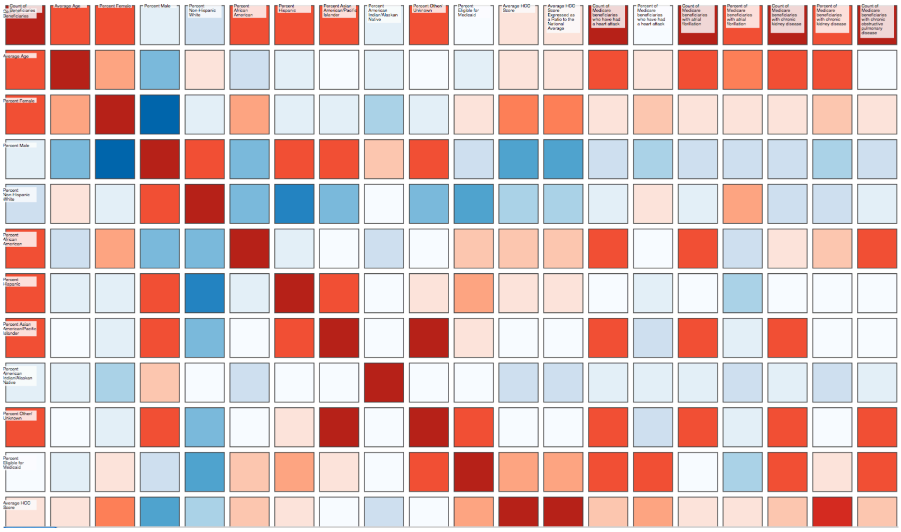

Count of Beneficiaries
Average Age
Percent Female
Percent Male
Percent Non-Hispanic White
Percent African American
Percent Hispanic
Percent Asian American/Pacific Islander
Percent American Indian/Alaskan Native
Percent Other/ Unknown
Percent Eligible for Medicaid
Average HCC Score
Average HCC Score Expressed as a Ratio to the National Average
Count of Medicare beneficiaries who have had a heart attack
Percent of Medicare beneficiaries who have had a heart attack
Count of Medicare beneficiaries with atrial fibrillation
Percent of Medicare beneficiaries with atrial fibrillation
Count of Medicare beneficiaries with chronic kidney disease
Percent of Medicare beneficiaries with chronic kidney disease
Count of Medicare beneficiaries with chronic obstructive pulmonary disease

Fig. 1 Top-level overview on pairwise health indicator correlations.

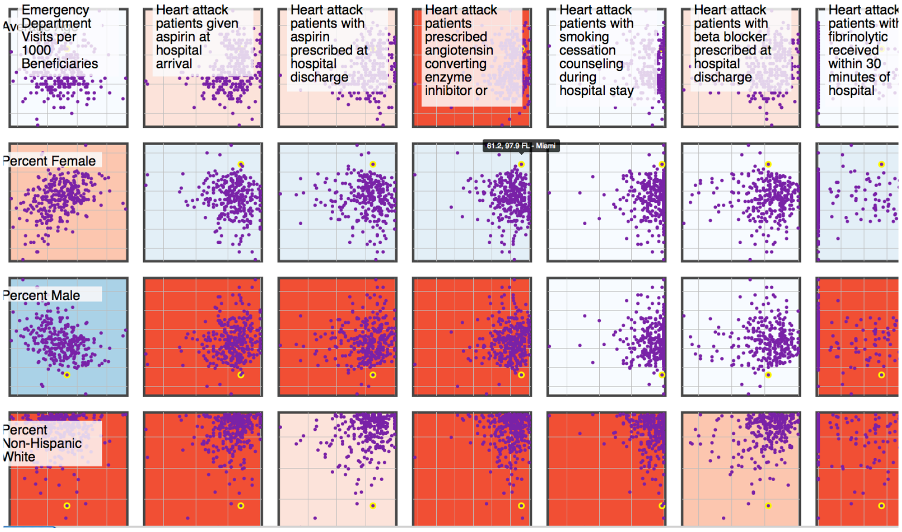
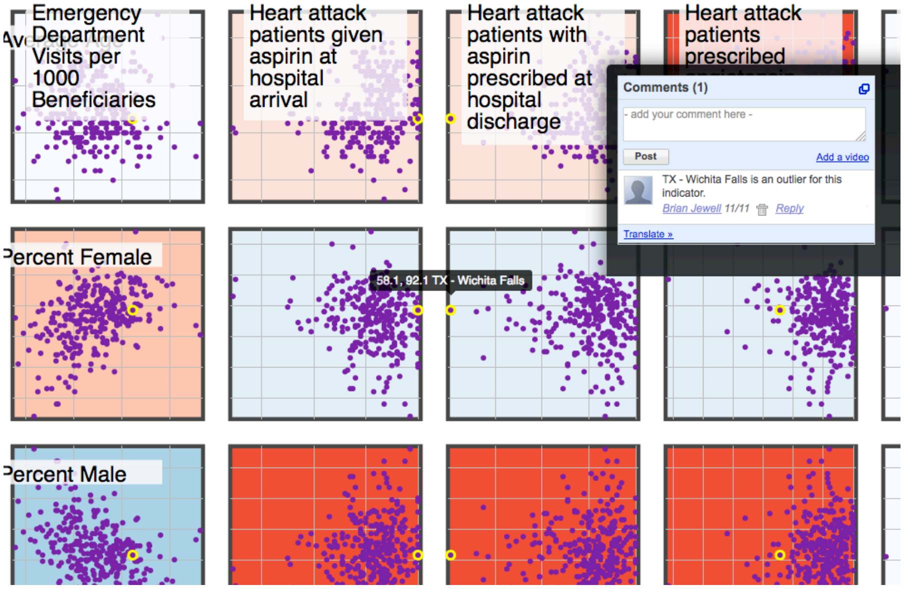

Emergency Department Visits per 1000 Beneficiaries
Heart attack patients given aspirin at hospital arrival
Heart attack patients with aspirin prescribed at hospital discharge
Heart attack patients prescribed angiotensin converting enzyme inhibitor or
Heart attack patients with smoking cessation counseling during hospital stay
Heart attack patients with beta blocker prescribed at hospital discharge
Heart attack patients with fibrinolytic received within 30 minutes of hospital
61.2, 97.9 FL - Miami
Percent Female
Percent Male
Percent Non-Hispanic White

Fig. 2 Zoomed-in view of county-level health indicator correlations.

The colors used to fill the squares are calculated based on the value of *r* between the variables using the color scale shown on the previous page. This color scale results in shades of blue for negative correlations and red for positive correlations. The color scale maps the saturation of the color to the strength of *r* so that the strongest correlations are displayed more prominently. Characteristic of correlation matrices, the correlation of a variable with itself is always a perfect positive correlation (1) and the diagonal of the matrix is therefore shown as a series of squares with the highest saturated red.

From the initial correlation matrix, the analyst can zoom into the display using the mouse scroll gesture to focus on a particular set of variables. As the analyst zooms into the display, the number of variables shown decreases. When the number of variables across the row or column dimension of the visible display is reduced to 6 or fewer variables, the display fades in the scatterplots for each variable intersection. That is, the square block for the variable intersections maintain the background fill that is associated with the correlation strength, and the individual points that contribute to the correlation measure are revealed.


Emergency Department Visits per 1000 Beneficiaries
Heart attack patients given aspirin at hospital arrival
Heart attack patients with aspirin prescribed at hospital discharge
Heart attack patients prescribed
Percent Female
Percent Male
58.1, 92.1 TX - Wichita Falls
Comments (1)
- add your comment here -
Post
Add a video
Reply
Translate »

TX - Wichita Falls is an outlier for this indicator.
Brian Jewell 11/11

Fig. 3 Sharing analysis results and findings using the comment and tagging feature built into our tool.

When the scatterplots are shown, the user can hover the mouse over individual points to see the associated values and geographic place names for the item. This information is shown as a tooltip in the display. Furthermore, subtle grid lines are shown in the scatterplots to provide reference points for comparing distributions across the SPLOM rows and columns.

Overall, our tool offers the following views to explore indicator values and their relationships: 1) a high-level graphic overview on pairwise correlations between indicator values in the form of a heatmap (Fig. 1), 2) a zoomed-in view of such correlations, where both the national average correlation strength and that of a region's local value are visually presented (Fig. 2), also upon selection of a certain region, a highlighted view on the region's relative standing in the country as characterized by multiple indicators, 3) a tagging view where an analyst can document thoughts and discoveries over a particular indicator for a particular region; since we provide the tool as an online application, a group of analysts can also collaboratively exchange their analytic findings by using our tool as a new collaboration platform for conducting social analysis (Fig. 3).

4 KEY IMPLEMENTATION ISSUES

Given the desire for the visualization to be accessible in a common format by a wide audience for collaboration a web platform was chosen. With further considerations of browser compatibility and interactivity, we chose to implement our application using the Data-Driven Documents (D3) library (Bostock11). D3 is a library that allows for the direct manipulation of the standard document object model (DOM). Using D3 we were able to bind data directly to SVG elements in order to construct a web based visualization with all the desired intractability.

The code in Fig. 4 comes directly from our implementation for creating the high-level heatmap view, and is included to demonstrate how D3 was used. On line 1 and 2, a DOM element is selected and a SVG element is appended. D3 then allows for the setting of element attributes (attr, style, property, html, text) through operators, which wrap the W3C DOM API. When data is bound a hierarchical structure of host elements is created for us to accommodate each array element. Thus, when our field array is bound on line 12, SVG elements for each column of our heatmap is created, and on line 17 and 18 the columns are then selected and row elements are created for each field in each column and data is appended appropriately creating a DOM based representation of our heatmap. We then append visual svg:rect elements to this representation colored using the data which is bound to their parent elements producing the heatmap which the user sees.

In order to optimize performance, the lower level view elements aren't added to the scenegraph until the users view is limited to a roughly 6x6 matrix. Addition of new elements is then limited to the users viewable area (viewbox). This means that the performance of the visualization is primarily bound to the *O(m*n)* (or *O(n^2)* for a symmetric representation) complexity of the matrix used for the high-level heatmap view. This does in fact affect the performance scaling of our application greatly. Complicating the further optimization is two factors 1) creating and removing DOM elements is expensive and 2) we do not wish to restrict the users ability to zoom/pan across the entire heatmap. A future implementation may be improved by rasterizing this high level view on the server side to increase scalability.

```
01    var svg = d3.select("#chart")
02    .append("svg:svg")
03    .attr("width", size * health.fields.length)
04    .attr("height", size * health.fields.length)
05    .attr("class", "BlRd")
06    .attr("pointer-events", "all")
```

```
        .call(d3.behavior.zoom()
        .on("zoom", redraw)).append("svg:g");

        // One column per field.
        var column = svg.selectAll("g")
        .data(health.fields)
        .enter().append("svg:a")
        .attr("transform", function(d, i) { return "translate(" + i * size + ",0)"; });

        // One row per field.
        var row = column.selectAll("g")
        .data(cross(health.fields))
        .enter().append("svg:g")
        .attr("x", padding / 2)
        .attr("y", padding / 2)
        .attr("width", size - padding)
        .attr("height", size - padding)
.attr("transform", function(d, i) { return "translate(0," + i * size + ")"; });

        row.append("svg:rect")
        .attr("x", padding / 2)
        .attr("y", padding / 2)
        .attr("width", size - padding)
        .attr("height", size - padding)
        .style("stroke", rectColor)
```

Figure 4 – Implementation of high-level heatmap view in d3.

5 DISCUSSION AND FUTURE WORK

In this project, we introduce a new platform for visually analyzing national health indicators. In contrast to the traditional visual analysis interfaces, our new platform offers a mixed heatmap and scatterplot view, allowing end users to hierarchically examine the correlations between multiple indicators, both on the national level and on finer geographical resolutions. Such a multi-resolution view enables policy analysts to comprehensively understand the healthcare performance and cost of a specific region in the context of the national average performance as well as that of its peer locations. Another unique function of our interface is that it allows analysts to share their discoveries through spatially anchoring their comments inside our visual interface onto individual geographic regions. By setting one's comments public, the analyst can invite other peer users in the system to participate in his/her analysis efforts over the region. With such an online comment and discussion forum feature, our visual analytics platform also serves as an online collaborative platform for uncovering hidden, underlying relationships embedded in the national health indicator warehouse through engaging social intelligence in the cloud.

Several conceptual extensions of the tool are being considered for future development. First, we conceive the analyst workflow as proceeding from the highest levels of abstraction to drill-downs to more limited areas of interest. One simple but important option to aid directed drill-down is to permit the analyst/user to select a smaller group of indicators for viewing in a smaller visual area; for example, clicking on multiple rows or columns generates a more compact matrix containing only those indicators of sufficient immediate interest.

Second, we desire to leverage human capability to detect complex color patterns within a visual matrix. The basic idea is to reorder indicators in the basis of similarity with respect to the vector of correlations. Reordering can lead to emergence of color patterns in the matrix and identification of latent factors in the form of variable clusters. Friendly (2000) illustrated the use of reordering techniques for baseball and automobile data; our approach is to use hierarchical clustering as a basis for reordering rows and columns. If the initial order is random, a clustered heatmap usually has the effect of transforming a random color matrix into a highly patterned diagram that illustrates regions of high negative and positive correlation. Although the clustered heatmap has been used to great effect in profiling gene expression data from micro-arrays (e.g., Rajaram and Oono 2010), we are not aware of any applications to visualization of health care indicators.

We note that the correlation/scatterplot matrix is symmetric, i.e., half the information is redundant. In addition, information displayed in basic scatterplots is exclusively bivariate. In order to utilize available space more efficiently we can extend the capability of the tool to explore relationships between two indicators while controlling for the effects of one or more additional indicators. Mediating factors are commonly required for adequate interpretation of aggregated multivariate health care data. This kind of information can be represented in partial correlation coefficients and partial regression plots. A promising approach introduced by Davison and Sardy (2000) is the partial scatterplot matrix. The authors introduced a mixed partial scatterplot matrix that displays univariate histograms along the diagonal (which typically contains no quantitative information in a conventional scatterplot matrix), and partial regression residuals after adjusting for all other variables along the lower panel.

Finally, although the correlation coefficient is useful for summarizing the dependency between two variables, it is primarily a linear index. There are also many possible nonlinear relationships that can be very informative from a data mining perspective. Examples of noteworthy nonlinear relationships include, for example, the inverted U-function, exponential or logarithmic functions, sinusoidal functions, non-coexistence relationships, and composite relationships such as piecewise linear functions. The latter can arise when two or more distinct groups independently cluster in a scatterplot. Recent advances in data mining are making the goal of characterizing nonlinear relationships increasingly feasible with current computer architectures. For example, the maximal information criterion and related family of MINE statistics for nonlinear characterization (Reshef, et al., 2011) are computationally expensive but scalable to visualizations of HIW-sized datasets like those described in this paper. Algorithms of this type can complement more conventional nonlinear visualization techniques such as locally weighted regression (Cleveland and Devlin 1988). We are exploring methods of displaying nonlinear relationship information in readily consumable forms within the framework of a scatterplot matrix.

ACKNOWLEDGMENTS

This work is sponsored by U. S. Centers for Medicare & Medicaid Services (CMS). Songhua Xu performed this research as a Eugene P. Wigner Fellow and staff member at the Oak Ridge National Laboratory.

REFERENCES

Bostock, M., V. Ogievetsky, and J. Heer. 2011. D3: Data-Driven Documents, IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis):2301-2309.

Cleveland, W. S., and S. J. Devlin. 1988. Locally weighted regression: an approach to regression analysis by local fitting. Journal of the American Statistical Association 83(403): 596-610.

Davison, A. C., and S. Sardy. 2000. The partial scatterplot matrix. Journal of Computational and Graphical Statistics, 9(4): 750-758.

Friendly, M. 2002. Corrgrams: exploratory displays for correlation matrices. The American Statistician 56(4): 316-324.

Healey, C. G., L. Tateosian, J. T. Enns, and M. Remple. 2004. Perceptually-based brush strokes for nonphotorealistic visualization. ACM Transactions on Graphics. 23(1):64–96.

Pirolli, P. and S. K. Card. 1999. Information foraging. Psychological Review 106(4):643–675.

Rensink, R. A. 2002. Change detection. Annual Review of Psychology 53:245–577.

Reshef, D. N., Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, P. C. Sabeti. 2011. Detecting novel associations in large data sets. Science, 334: 1518-1524.

Rajaram, S. and Y. Oono. 2010. NeatMap - non-clustering heat map alternatives in R. BMC Bioinformatics, 11:45.

Wong, P. C. and R. D. Bergeron. 1997. 30 years of multidimensional multivariate visualization. In: Scientific Visualization - Overviews, Methodologies, and Techniques, IEEE Computer Society Press, pp 3–33.

Walpole, R. E. and R. H. Myers. 1993. Probability and Statistics for Engineers and Scientists, 5th ed., Prentice Hall, Englewood Cliffs, New Jersey.

Willett, W., J. Heer, and M. Agrawala. 2007. Scented widgets: Improving navigation cues with embedded visualizations. IEEE Transactions on Visualization and Computer Graphics 13(6):1129–1136.