Improving Public Health with the CHSI Correlation Browser

Bryce Cronkite-Ratcliff  
Stanford University,  
Department of Computer Science  
brycecr@stanford.edu

Ruby Lee  
Stanford University,  
Department of Computer Science  
rubylee@stanford.edu

ABSTRACT

Summary: The Community Health Status Indicators (CHSI) data set is collected by the US Department of Health and Human Services and presents county-level data on population demographics and health factors. While this rich data set has promise to accelerate data-driven policymaking, working with such complex, high-dimensional data requires novel tools for exploratory analysis and correlation finding. We developed the CHSI Correlation Browser (CHSICB), a web-based software program for visually exploring CHSIs. The CHSICB allows users to select and explore several data attributes simultaneously via choropleths, parallel coordinates charts, a data table, and a scatterplot matrix. We have developed several case studies that used the CHSICB to reveal correlation between attributes at the national level.

Availability: The CHSI Correlation Browser is freely available open-source at http://github.com/brycecr/chsi. CHSICB was implemented in HTML, JavaScript, and PHP, and is platform-independent.

INTRODUCTION

The Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease, and Cancer is a data set compiled by the United States Department of Health and Human Services (HHS) to assist public health professionals and other interested analysts in investigating the factors that shape public health. The data set contains information for each of the 3,141 U.S. counties for over 200 attributes. The attributes are classified into eight categories: Demographics, Leading Causes of Death, Measures of Birth and Death, Preventive Services Use, Relative Health Importance, Risk Factors and Access to Care, Summary Measures of Health, and Vulnerable Populations and Environmental Health. Due to the wide range of health indicators it covers, this data set has potential to reveal correlations that could inform future efforts to improve public health. We have developed the CHSI Correlation Browser (CHSICB), a web-based platform to facilitate visualization and exploratory analysis of the CHSI data set. We present several case studies that used the CHSICB to reveal correlation between attributes at the national level.

RELATED WORK

Visualization software has been used to address the CHSI data set previously. CHSI Geographic Information System (GIS) was developed by the US Department of Health and Human Services and several other institutions to implement geospatial exploration of the CHSIs [1,2,3]. CHSI GIS was developed specifically to allow users to obtain images that highlight a specific county for inclusion in media materials. However, because only one attribute and one state can be viewed at a time, it did not focus on enabling correlation browsing between attributes.

The Community Health Map was designed for Medicare data browsing [4,5]. This application allows for visual comparisons between attributes in two ways. First, the
program can filter county visibility in the map with threshold sliders for four specific pre-selected attributes (median income, poverty rate, percent bachelor's degree, and percent population over 65 years of age). Second, the program can present of any number of data columns in a numerical table. Correlation browsing is limited without additional tools to brush and link data.

**METHODS AND RESULTS**

The CHSI dataset used in this project was obtained from http://www.communityhealth.hhs.gov/ in November of 2012. The CHSI Correlation Browser is a web application written in HTML, JavaScript (JQuery), and PHP. It uses several JavaScript libraries: d3.js, d3.parcoords.js, underscore.js, colorbrewer.js, divgrid.js, and blockUI.js. These libraries depend on browser support for SVG and HTML5 Canvas elements, so a recent browser is required (such as the latest version of Chrome or Firefox).

**Navigational Tools**

The CHSICB interface is designed to explore correlations in the CHSI data set. Thus, the interface encourages the selection of multiple attributes from the CHSI, and then presents that selection of attributes in four different sets of visual abstractions. Because each type of visualization has unique strengths and weaknesses, the collection of visualizations provides a critical set of views to examine correlations. Each of the five sections of the page, including the attribute navigation, can be displayed in expanded or hidden mode.

Users may select up to six attributes to visualize concurrently. Once an attribute slot is highlighted by clicking on one of the six choropleth maps, an attribute from the data set can be selected from a menu organized hierarchically according to the attribute taxonomy used by the HHS. Each attribute tile is labeled with its CHSI code and the provided description. For existing CHSI analysts, maintaining the organization of coding of attributes used in the CHSI is natural and useful. The coded names are generally both intelligible and concise. Additionally, the organization of attributes is intuitive, consistently designed, and relatively uniform across categories.

It should be noted that the CHSI contains many attributes (e.g. confidence intervals) that are intended for use in statistical evaluations. We omitted these attributes from the navigation to reduce visual clutter, because they were not informative in our visualizations.

**Implementation.** The data set is stored in a MySQL relational database, along with metadata for each attribute (name, category, description). Each attribute is retrieved in Javascript Object Notation (JSON) and delivered using Asynchronous Javascript and XML (AJAX) to the browser.

**Choropleth Maps**

Each attribute data point in the CHSI is associated with a geographic county in the United States, so any attribute can be mapped onto a choropleth map, where the fill color of the region is determined by its data value. By default, we map values to colors with a simple quantile scaling of nine bins. This color scaling works well for cases without a small number of extreme values in the data set. The choropleth maps can each be expanded to visualize national distributions with high resolution, or all six can be viewed at once as a set of small multiples to examine trends between maps.

**Implementation.** The choropleth maps in the CHSICB are implemented in d3.js. Backing geographic data was originally derived from the US Census Bureau and then converted into a JSON object using GDAL and MapShaper. Each county in the map is indexed by the concatenation of
its state and county FIPS codes (each county in the United States is uniquely identified by a five-digit FIPS code). Attribute data is stored as a JavaScript object keyed by FIPS code, and d3.js maps each data value to a fill color on a scale of reds from colorbrewer.js for the corresponding county path on the map.

**Parallel Coordinates**

Parallel coordinate graphs present multivariate data that can be brushed and linked in a single diagram. Each dimension is represented by a single axis, and all axes are parallel. Each row (i.e. each county) in the data is represented by a line, where the intersection of the line with each axis indicates the value of the data point for that row along for the intersecting dimension. Our parallel coordinates support both axis reordering and brushing. Each axis can be brushed to restrict the highlighted data rows, which allows the user to focus on the relationship between attributes within a group of counties that share similar values along one dimension.

*Implementation.* The implementation of parallel coordinates in the CHSICB relies on d3.parcoords.js. This library interoperates with d3.js, but uses HTML5 Canvas elements rather than SVG to accelerate rendering of data lines. Data line color is determined by statistical binning of values for a particular dimension, selected by the user by clicking on the desired axis label. Statistical binning is based on standard deviations away from the mean for the value at the coloring attribute.

**Data Table**

The CHSICB provides a simple dynamic table to facilitate the linking of a visual trend to backing data or to specific counties of interest. Each row in the table contains the name of a county and state, with each of the county’s values for all currently selected attributes. In order to aid focused analysis, the table is linked with the parallel coordinates graph; only the data rows that satisfy the brush constraints on the parallel coordinates are present in the table. Additionally, hovering over a row in the table highlights the corresponding line in the parallel coordinates graph.

*Implementation.* The table is implemented using a div-based data grid developed by Kai Chang. Because the data sources for the parallel coordinates and data table are identical, the same key uniquely identifies the same data row in each, simplifying linking between the charts.

**Scatterplot Matrix**

Scatterplots provide visual cues to indicate the degree of correlation between a pair of attributes. The CHSICB provides a scatterplot matrix, which contains one plot for each pair of attributes selected. A region of a scatterplot can be selected to show linkage across the matrix. These are

![Figure 3: Parallel coordinates graph and corresponding data table linking the No_HS_Diploma, Unemployed, Population_Density, and Poverty attributes.](image-url)
particularly useful to supplement the choropleths and parallel coordinate graph, which provide a higher-level overview of the data.

Implementation. The scatterplot matrix is adapted from an example by Mike Bostock. The scatterplot matrix can be brushed and linked, and tooltips with the county and data value appear on hover over a data point.

DISCUSSION

It has historically been challenging and time-consuming to perform exploratory analysis on the CHSI data set with a focus on cross-attribute correlations. The CHSICB provides a set of interacting visualization types in an intuitive and approachable framework to enable discovery and visual analysis of correlations in the CHSI data set.

In preliminary usage analysis, we note that users initially find the choropleth maps most engaging and most natural to interpret at first take. This may be because users generally have preconceived notions about the national distribution of a given attribute. However, providing a general coloring scheme for all attributes is difficult to automate, and interpretation of the graph lacks precision.

The parallel coordinates graph, in addition to allowing to providing a comprehensive overview of the entire dataset, allows questions about correlations to be more precisely defined and explored. A user may limit visible data to counties where a given attribute is within a specified percentage range, and see how those counties are distributed over other dimensions.

The scatterplot matrix, presented at the bottom of the CHSICB, is important in establishing that a correlation is supported by substantial data. Once a user had established an interesting correlation to examine more closely, the scatterplots were straightforward to interpret and provided a signal for a positive or negative correlation coefficient.

We have developed several case studies during cursory exploration of the CHSICB during testing:

- Exploring “Asian” attribute (the proportion of the population in a county that identifies as Asian according to the 2008 US Census) shows a small number of counties with large Asian populations. A parallel coordinates graph representing the Asian and Average Life Expectancy (ALE) attributes, with counties with >30% Asian population selected, shows that counties with high proportions of Asians also have higher than average life expectancy.

- If “Obesity” (percentage adults of in the county that are classified as obese) and “Prim_Care_Phys_Rate” (number of primary care physicians per 100,000 people) are selected, the scatterplots and parallel coordinates graph indicate that the counties with the highest obesity rates also have relatively low rates of primary care physicians.

- By selecting the “Unmarried” (number of births to unmarried women), “Late_Care” (percentage of pregnancies with no care in first trimester), and “Infant Mortality” attributes (percentage of pregnancies), we see that all three attributes are positively correlated with each other (Figure 5). This is an example in which a health policymaker might become motivated to investigate possible causational relationships between being unmarried, not visiting an obstetrician in the first trimester, and high rates of infant mortality.

FUTURE WORK

While our observations have suggested that the CHSICB enables correlation exploration, several interface adjustments could make the platform more efficient for use as a professional analysis tool and more approachable for casual users. Because these two user groups have different needs, an overall architectural shift might focus on the distinction between two slightly different interfaces, one for casual and one for professional usage. Both views would benefit from a system that allows tagging and sharing of interesting “data stories.” A “data story” consists of a set of attributes, a short description, a set of selection constraints, and a set of focus points (e.g. arrows pointing to particular counties on choropleths, highlighted data table rows, etc.).
These stories could be created by users, stored in a central bank, and shared through social media or email, providing a mechanism for collaborative visualization of CHSI data.

Advanced users may find more technical features important, such as the ability to generate new data attributes as combinations of existing attributes. The necessity of this is suggested by attributes such as Major_Depression, where the statistic is cumulative and not per-capita. For users interested in how a specific county or region compares to its peers, we would like to provide mechanisms for searching for a data row and focusing visualization to specific geographic regions of the United States. Lastly, the load time and responsiveness of the interface can also be improved. If the JSON files describing the geometry of the United States were simplified further for smaller maps, the load time for new attributes could be decreased.

ACKNOWLEDGEMENTS

The CHSI is a dataset produced by the U.S. Department of Health and Human Services. This project uses code from d3.js by Mike Bostock and contributors, and d3.parcoords.js and divgrid by Kai Chang. We are very grateful to the CS448B course staff -- Eli Marschner, Amy Jang, Megan Kanne, and Professor Jeff Heer -- for their inspiration and guidance.

REFERENCES

Institute for Health Metrics and Evaluation.
http://www.healthmetricsandevaluation.org/tools/data-visualizations


Figure 5: Screenshot of complete CHSICB view with three selected attributes: Unmarried, Late_Care, and Infant_Mortality.