

Visualizing Disease Incidence in the Context of Socioeconomic Factors

Jared Shenson
Vanderbilt University
2201 West End Avenue
Nashville, TN 37212
jared.shenson@vanderbilt.edu

Alark Joshi
Boise State University
1910 University Drive
Boise, Idaho 83725
alarkjoshi@boisestate.edu

ABSTRACT

Certain biological factors such as genetics, physical fitness, and lifestyle have been shown to influence an individual's risk of acquiring disease. But are there other socioeconomic factors that influence disease incidence as well? In this paper, we introduce a visualization tool called *DiseaseTrends* that explores the associations and possible correlations between specific economic (personal income per capita), educational (percentage of adult population with a four year college degree), and environmental (air pollution level) factors with diabetes prevalence and cancer incidence rates across counties throughout the United States. It is structured as an interactive geographical visualization that displays disease incidence data as an interactive choropleth map and connects it with coordinated views of the socioeconomic variables for each county as the user scrolls over it. Additionally, the ability to compare and contrast counties as well as to interactively specify a region for comparison allows further examination of the data. This results in an informative overview of disease incidence trends that allows users to spot areas of interest and potentially pursue these areas further with more scientific research.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces graphical user interfaces, interaction styles, user-centered design.

Author Keywords

Geographic Visualization, Graphical User Interfaces, Evaluation/Methodology

INTRODUCTION

It is widely known that for any given individual, biological and personal health factors such as genetics, physical fitness, and lifestyle play a role in the possible onset of a disease. For example, the risk of Type 2 diabetes exhibits strong hereditary links and further increases with a person's age, obe-

sity, and alcohol intake [13, 21]. An interesting question however, is if there are socioeconomic factors that forecast disease incidence as well. Is being economically well-off related to a person's risk of acquiring diabetes? What about receiving a four-year college education or being exposed to high levels of air pollution? While we recognize it is difficult to give definite answers to these questions, let alone determine any causal relationships between these socioeconomic factors and disease incidence, we attempt to provide hints at connections among the variables to motivate further research in the form of an interactive visualization tool.

We present *DiseaseTrends*, an interactive geographic visualization tool that allows the examination of these variables with respect to disease incidence. Our visualization tool displays and helps users explore any possible trends as well as correlation among these. We identified a number of socioeconomic factors that we thought were potentially linked to disease incidence. On the economic side, we have personal income per capita; on the educational side, the percentage of the population holding a four-year college degree; and lastly, on the environmental side, the estimated levels of air pollution. All the variables are observed at the county level. As for the disease variable, we decided to concentrate on the incidence rates for two well-known diseases: diabetes and cancer. The reason for selecting these two came down to the fact that there is much scientific and medical literature hypothesizing links between these diseases and some of the socioeconomic factors that we have chosen. *DiseaseTrends* allows a user to filter the data, brushing the data to highlight specific counties, pick specific counties for compare-and-contrast type tasks, or examine user defined regions to include a cluster of counties. Once a user selects a county, we compute a similarity metric based on our socioeconomic factors and show counties that are mostly similar but may have varying disease incidence rates. This may spur researchers to further examine those counties for other factors causing varying disease incidences.

We believe our visualization will be most beneficial for (1) those in the public health research community who would like to use our visualization as a launch pad for ideas or areas for deeper, more thorough investigation, and (2) lay people who are interested in seeing how their county's disease incidence rates compare to those of others, and whether their relative economic, educational, and environmental factors might be related to the incidence rates. It is important

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

VINCI 2012, September 27–28, 2012, Hangzhou, China.

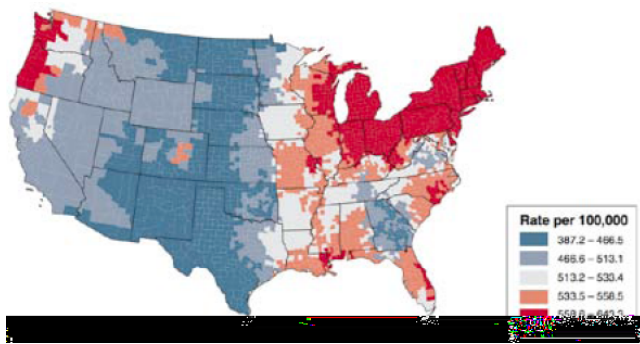


Figure 1. The National Cancer Institute’s presentation of their statistically-modeled county-level cancer prevalence estimations. This static choropleth map incorrectly uses a diverging scale, typically used to show values diverging away from a mean or median of the data. The map is, however, readily understandable, but it is static and does not invite user participation with the data.

to note that our visualization does not provide any concrete statistics or direct answers to questions about the correlation of socioeconomic factors with disease incidence; rather, we believe it will function as a potential foundation for further research.

Our approach in building DiseaseTrends was to base the tool on a geographical model and build on top of this platform to introduce new visualization techniques to improve data exploration. Geographic visualization, often presented in cartographic form, is a proven technique employed by many fields to easily visualize geocoded data. Presentation in this form is common to many users, from domain experts to the lay individual. Further discussion on the merits and challenges of geographic visualization is presented in the sections that follow.

RELATED WORK

Geographical visualization has been in existence for centuries, starting with the earliest cartographic maps [27, 16]. Geographic visualization has evolved over time, as MacEachren [29] describes the initial focus of cartographic visualization as being for “personalized, highly interactive tools that facilitate a search for the unknown”. But now geographic visualization can be used for exploration [26, 9], where hypotheses emerge from observed *visual trends*, rather than using geographic visualization to answer a hypothesis. In order for techniques to be successful in facilitating data exploration, Tomaszewski et al. [37] suggests that the implementation of highly interactive tools is critical. Further defining the role of geographic visualization, in *How Maps Work* [27], MacEachren says that the “goal of a map is to stimulate a hypothesis rather than communicate a message.” We believe that our interactive visualization tool stimulates a hypothesis regarding the prevalence of a disease and its possible correlations to other socioeconomic factors, which could further be studied using research in public policy and other related fields.

Let us now examine a couple of visualizations that are closely related to our work. Several years ago, the National Can-

cer Institute’s (NCI) Statistical Research and Applications Branch used geographical visualization to display U.S. cancer incidence rates for the year 1999, predicted by a statistical model (see Figure 1) [17]. We believe their use of geographical visualization was largely a success, but note how their visualization merely provided a static overview of the data. The emphasis of their work was on the accuracy of the statistical predicting model. In our work, we opted for a more interactive, engaging visualization that would commence a line of thought rather than end it.

A more recent, closely-related predecessor we would like to draw attention to is the set of visualizations by Bill Davenport from his TEDMED talk, “Your health depends on where you live” [15]. In this talk, Davenport argued that the places you have lived have a significant influence on your health. He proposed an interactive application for portraying this data to patients and doctors and exhibited two geographical visualizations: one with the distribution of heart attack rates across the country, and another with the location of toxic waste facilities monitored by the EPA. Both illustrated how easily we can see and link geographical factors to our health risks with the use of visualization. Davenport’s visualizations were more of the interactive, engaging type. It called upon users to spot trends, form hypotheses, and test conclusions. The visualizations, however, focuses significantly on the presentation aspect, and not on user interaction as a real-time data exploration tool.

Below we will discuss some work from the data visualization community that has formed our approach. Frederikson et al. [19] provided focus and context in cartography through facilitated drill-down coordination and aggregation for geographic visualization of events. They maintained a Main View that contains markers, each representing an aggregation of values at a more local zoomed in level. Markers were coded by size and color and positioned on map by location. Other approaches to display bivariate data include work by Nelson et al. [31], which identified three different types of user interaction with bivariate symbol design: separable symbols, integral symbols and configural symbols.

Most research for public health data visualization has concentrated on extending capabilities of choropleth maps by dynamic linking in a dashboard setup [34], interactive filtering/dynamic queries [33], or integration of statistical methods, computational clustering and pattern recognition. Carr et al. [11] developed “linked micromaps” similar to those shown on the NCI website. Their approach provides a method for a user to filter a choropleth map (“conditioned choropleth map”) but does not allow interactive examination of individual counties. Wessel et al. [39] discuss the prevalence of linked views for simultaneous comparison of data, and then propose an interesting method of conveying context to focus with an altitude-based focus region. A more statistics oriented approach was taken by Chen et al. [12], who investigated the application of clustering / smoothing algorithms of large datasets and visualizing results on choropleth maps. They used the Kulldorf spatial scan statistic to cluster their data. They combined the computed “reliability”

index with cluster data and provided visual representations in the form of choropleth maps. We agree with them regarding the inherent problems with static choropleth maps which are (i) data classification (ii) choosing an appropriate color map is extremely crucial (See Mersey [30] for details) (iii) small numbers problem (small changes in absolute number can have large effects on rate and thus strongly influence appearance on map) and (iv) visual bias problem (high-risk, small, densely populated regions that are not easy to see but represent real problems).

Of the tools described above, only some supported high levels of user interactivity, a feature described by many to be of strong value in geographic visualization. MacEachern et al. [28] share some of their experience with designing a linked-dashboard style web-based interface for exploring cancer incidence. They encourage developers to “keep features to a minimum for non-expert users especially since there is very limited training time.”

Current GIS Tools

Current GIS tools allow users to explore geographic data. ArcGIS [32] facilitates exploration of maps in correlation with geographic information. It is robust in its ability to handle a large variety and formats of datasets, but it lacks the flexibility of interactively defining regions of interest and comparing them across a geographical region. TIBCO Spot-Fire [36] and Tableau Software [35] can both handle geographic data, but they employ basic visualization capabilities such as coloring a region based on a quantity or placing a scaled bubble on the region to communicate that quantity. Basic “filter” type tasks as defined by Amar et al. [7] are possible, but “correlation” and the ability to “characterize distribution” and so on are not possible with these generic business intelligence tools.

The U.S. Census Bureau’s Small Area Income and Poverty Estimates (SAIPE) [6] tool provides basic geographic visualization (choropleth map) that communicate “annual estimates of income and poverty statistics for all school districts, counties, and states.” It allows interactive exploration of the value for a state, county or school district which is useful for basic “retrieve value” type tasks [7]. Unfortunately, basic tasks such as “filter, find extremum, determine range” too are not available. “Correlation” tasks are particularly hard to conduct using their interface. The Geovisual Analytics Visualization (GAV) Flash Tools [23] as created by the Swedish National Center for Visual Analytics (NCVA) is a web-enabled tool [22] that brings the power of visual analytics [8] to geographic visualization. It facilitates data exploration through the use of map layers, choropleth maps, information visualization techniques such as line graphs, parallel axes charts, scatter plots, treemaps and many others [24]. Among the kinds of tasks it seems to facilitate “filter, retrieve value, determine range, find anomalies” type tasks [7]. Interaction capabilities for comparison “correlate, characterize distribution” type tasks does not seem possible and could not be verified since the tool does not seem to be live any more.

METHODOLOGY

The methodology we followed in our project can be broken down into four major stages: (i) data collection, (ii) designing and building the visualization (iii) obtaining feedback via expert evaluation.

Data Collection

Our first step was the data collection process. We relied on online search engines and reliable websites, such as those of government agencies, to gather any economic, educational, or environmental data that seemed relevant and promising. The disease data was easy to locate because it is recorded by many organizations within the government and medical research community. We gathered data on the age-adjusted estimates of the prevalence of adults with diagnosed diabetes, compiled in 2007, from the website for the Centers for Disease Control and Prevention (CDC) [1], and the 2006 annual cancer incidence rates (which is an aggregate of all types of cancer) from the website for the National Cancer Institute [2]. Both variables were reported on the county level. For data on our *economic factors*, we looked to the Bureau of Economic Analysis. We decided to use the comparable statistics of personal income per capita, which was provided on the county level [4]. Our *education* data came from the Economic Research Service (ERS), which sourced data from the 2000 U.S. Census [3]. Although the data may be a bit dated, we chose to use it because, first, the Census is highly reliable, and second, with our visualization being geared towards highlighting long-term trends (the influences of air pollution may not be seen in disease incidence rates for years), the data is still appropriate. In the data set, there were a number of different variables shown: the number of people who have less than a high school education, only a high school education, some college education, a four-year college education, and the corresponding percentages with respect to the county populations. Lastly, we collected our *environmental* data from the Environmental Protection Agency (EPA). In 2002, the EPA created computer models to describe the estimated total concentration of toxic air pollutants by county [5], which we used for our analysis. It should be noted that the EPA warns of its data that “results are most meaningful when viewed at the state or national level, and should not be used to draw conclusions about local exposures or risks”. However, because our visualization is meant to suggest areas for further investigation, rather than draw definite conclusions, we continued to proceed with the dataset.

Designing the Interactive Data Visualization

Based on our design goals, we wanted to develop a geographic visualization with the ability to maintain interactivity, allow users to examine specific data points, facilitate easy comparison between counties, allow for filtering and highlighting data to form hypotheses. Figure 2 shows an overview of our interactive geographic visualization with a choropleth map showing incidence rates of diabetes. A user can switch between the diabetes data and the cancer data by clicking on the leftmost button in the bottom panel. As a user mouses over a county, we show a tooltip with the data for each county along with a linked bullet graph on the top right showing the other variables such as population, income, ed-

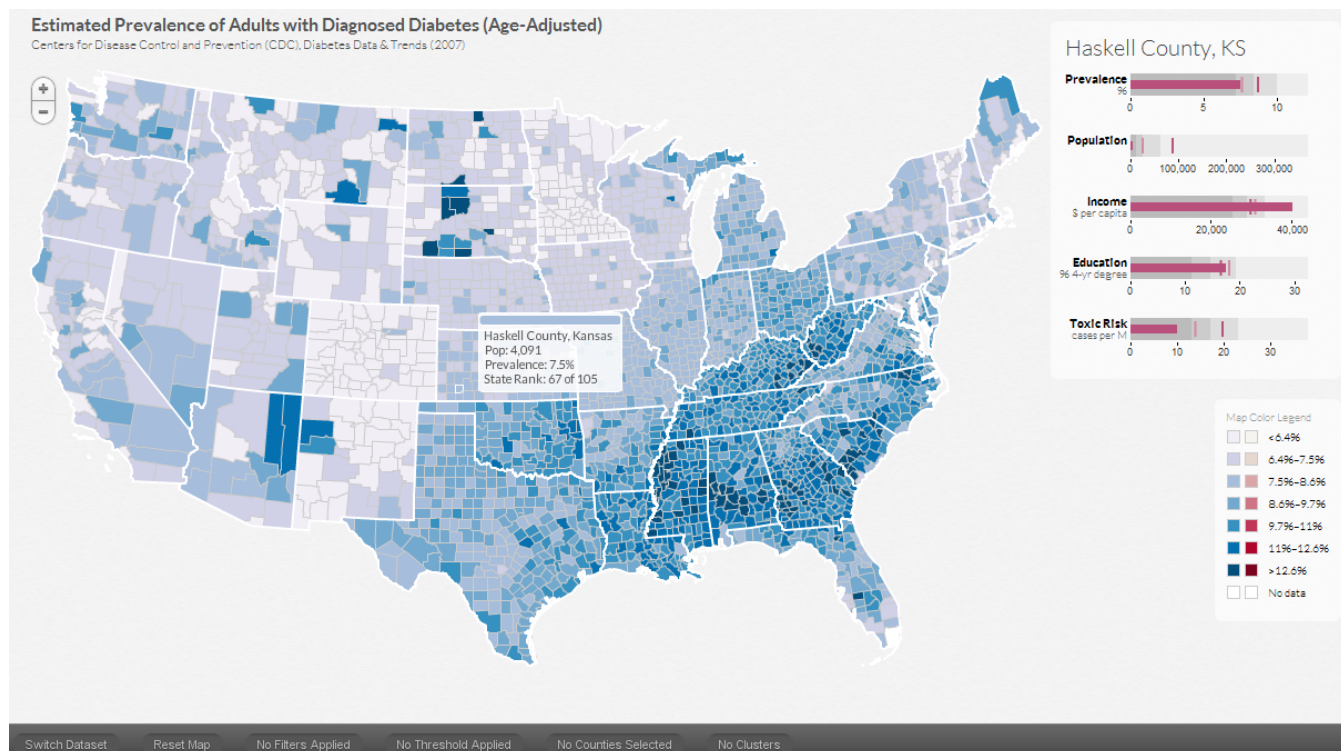


Figure 2. This figure shows an overview of DiseaseTrends. An interactive choropleth map is used to indicate incidence rates of a particular disease (diabetes or cancer) for the counties. We use the *Blues* color scheme obtained from Color Brewer for the choropleth map (shown on the right side of the figure). A user can mouse over any county to see the values for prevalence of the disease, the population of the county, the income per capita of the county, the education levels and the toxic risk in the form of air pollution for that county (shown on the right in the form of a bullet graph per variable). Other buttons in that row allow a user to filter, select specific counties for compare-and-contrast type tasks and also to select a cluster of counties.

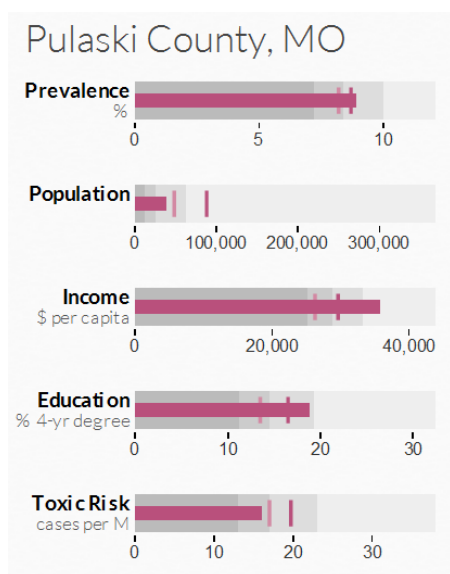


Figure 3. Bullet Graph visualization of the disease prevalence and socioeconomic factors for a county. For each variable, the value is shown as a bar and the two vertical lines perpendicular to the bar indicate the national average (dark vertical line) and the state average (light vertical line). The background represents quartile data distribution.

education and toxic risk. The Color Legend used for the map is shown below the bullet graphs. The colors for the color legend were obtained from ColorBrewer [20]. The structure of our large dataset with five unique variables and no hierarchical relationship gave us pause in our initial considerations for our visualization. We decided that it would be futile to overlay all variables on the same geographic map/frame, since it would cause visual clutter and make it particularly difficult for users to explore the data. We concluded that two separate visualizations would be more suitable to show diabetes and cancer incidence rates, as the two diseases were unlikely to show any correlation and would only serve to occlude the viewing of the other data.

Developing a visual blueprint

Our goal in designing the data visualization has been to enable user interaction with disease incidence data as well as all three factor variables simultaneously. This design called for displaying a choropleth color map showing the disease incidence data and implementing a new linked visualization with the other factors. Since we needed to show the value of a variable in conjunction with the state and the national averages for that variable, we chose *Bullet Graphs*, as introduced by Stephen Few [18]. Bullet graphs show a quantitative scale with the current value displayed as a bar, while vertical markers perpendicular to it serve as a comparative measure for quantities such as averages, median etc. In

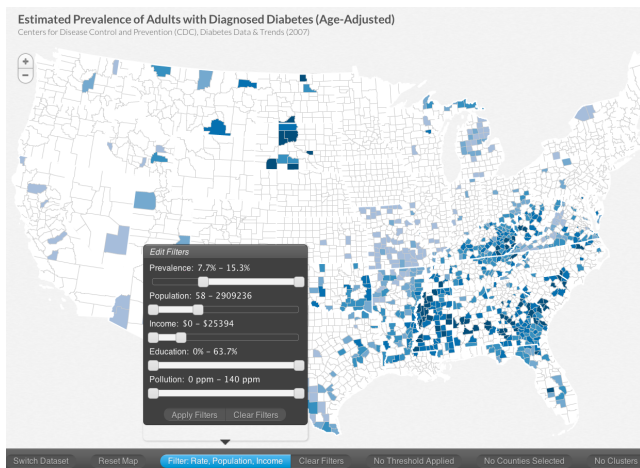


Figure 4. Filtering is enabled for all the variables: prevalence, population, income, education and toxic risk. The figure shows the user filtering out high income counties to only display counties with high prevalence, low population and low income. The tab representing filtering at the bottom is highlighted in blue and shows the filters being currently applied.

our case, we use two such vertical markers - one indicating the state average (light vertical marker) for that variable and one indicating the national average (dark vertical marker). This allows a viewer to compare the value of that variable with the state and national averages. The background of the bullet graph encodes the data distribution in four shades of gray, each shade representing a quartile of the distribution of the variable being visualized. For example, in Figure 3 the distribution of prevalence for the entire country is towards the lower quartile. Therefore, Figure 3 shows that Pulaski County, MO has an average prevalence with a population that is below the state and national average for a county, an income that is higher than the state and national average income for a county, higher education levels than state and national levels and lower toxic risk as compared to the state and national averages.

The bullet graphs are linked live to the mouse-over actions of the user interacting with the map. Hovering over a particular county causes the bullet graph to be updated with the particular county and state data. Additional information is provided to the user in the form of a tool tip (see Figure 2). The tool tip reveals the county name, population, disease incidence/prevalence rate, and state rank based on that rate (where rank 1 is equal to the highest rate in that state). The top colored strip of the tool tip box encodes the same color as the county in the map, reinforcing the user's color association with the value of the county's disease incidence and overcoming the chance that, at the lowest zoom level, the selected county (and its color) is completely obscured by the mouse cursor. At any time, the user has the ability to perform other functions to interact with the map including zooming in and out, moving the map within the display window, and switching the active disease dataset. The user can interactively select counties for compare-and-contrast type operations, draw clusters of counties around the cursor, filter the data to eliminate the display of counties that fall outside

the chosen filter values and brushing the data to highlight specific counties.

Filtering

While much can be learned about the data based on the tools described above, we believe it is important to allow users of DiseaseTrends to filter the entire dataset by any combination of the variables (the active disease data, the population and the three factor variables). The filtering feature is accessed through a button in the toolbar, which displays a drop-down box featuring sliding scales for each of the factors (Figure 4). Users can click and drag the end points of the sliders to adjust the filtering range. Once the user clicks on the Apply Filters button, the map is updated to only color counties whose data meets the filters set by the user. All counties that do not meet the filtering criteria are shown in white. Figure 4 shows an example of filtering. In the figure, the data is filtered to only preserve low-income counties with high prevalence and low population. This results in a set of counties and seems to highlight a large set of counties in the south-eastern part of the United States.

Brushing

In addition to the filtering options provided, a user can highlight certain counties on top of the entire map of the United States by using brushing. brushing allows a user to draw a viewer's attention to a specific region or set of counties that meet the brushing criteria. As compared to filtering, where we only show counties that match the user's selection, in brushing we show the counties matching the brushing criteria in red with the unaltered map of the country (shown with the blue color scale). Figure 8 shows an example where counties with a high prevalence of diabetes are highlighted.

County Selection

In addition to getting a sense of the overall distribution of a disease, we provide the ability to compare-and-contrast various counties. A user can select specific counties, which then show up on the bottom panel, as shown in Figure 5. As subsequent counties are selected, we display their details in the bottom panel and place a number in their place for easy reference of the selected counties for comparison. Since these counties are selected for comparison purposes, we compute a running maximum and minimum incidence value among the selected counties. On closer examination of the panels in Figure 5 showing the selected counties, one can notice that the lowest prevalence amongst the five counties is highlighted in green (Big Horn County, Montana) and the highest prevalence is highlighted in red (Buffalo County, South Dakota). This provides a quick snapshot of the selected counties and allows for easy comparison.

County Correlation

Since our goal has been to "stimulate a hypothesis", we use the County Selection process described above as a cue regarding the interest a user has in further examining a county. Once a user selects a county, we compute a similarity metric using a combination of the population, income, education levels and the toxic risk of the county. Based on the metric, we identify up to a maximum of five "similar" counties

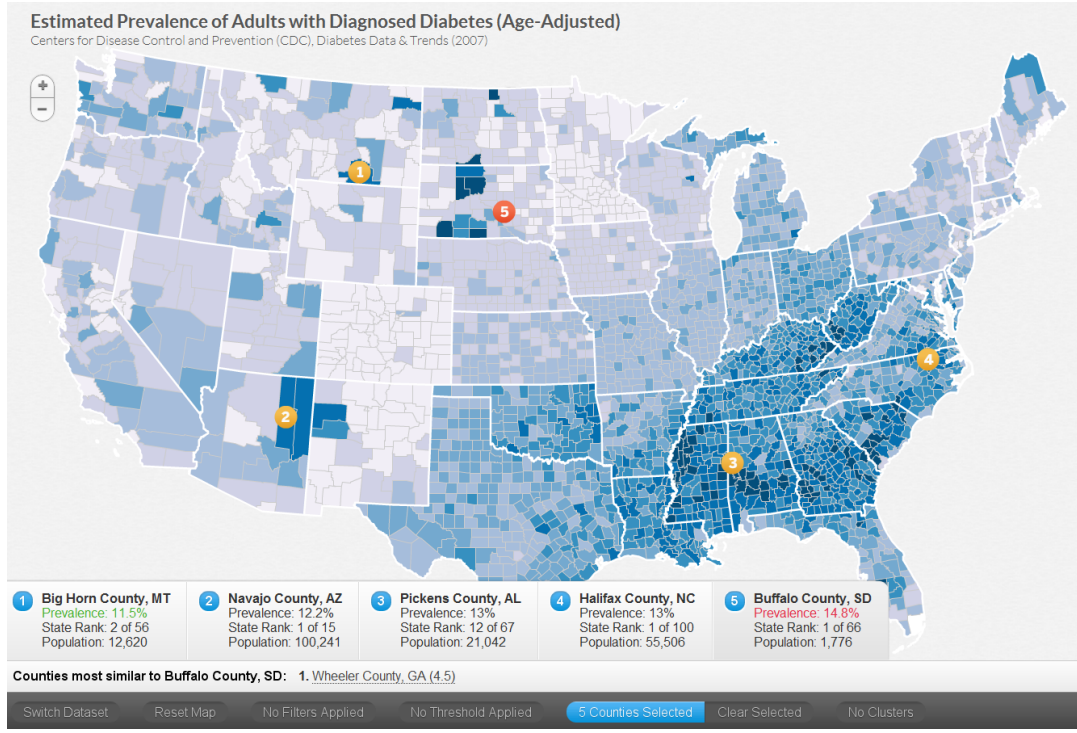


Figure 5. County selection: A user can select counties for comparison by holding down the Shift key and clicking the county. This results in the county information being displayed at the bottom with a number. A user can select multiple counties to compare counties. Since Big Horn County, MT (1) has the least prevalence its value is shown in green, whereas Buffalo County, SD (5) has the highest prevalence among the selection and is shown in red.

and display them in the bottom panel with their correlation scores. We use the following formula to compute the similarity metric:

$$S = \sqrt{\left(\frac{p_1 - p_2}{1000}\right)^2 + \left(\frac{i_1 - i_2}{1000}\right)^2 + (e_1 - e_2)^2 + (t_1 - t_2)^2} \quad (1)$$

where S is the computed similarity metric and for the two counties under consideration, p_1 and p_2 are the population values, i_1 and i_2 are the average income values, e_1 and e_2 are the education levels, t_1 and t_2 are the toxic risk levels. Based on experimental evaluation, we identified a value of $S < 5.0$ to identify meaningfully similar counties. Since population values as well as the income values were in the thousands as compared to the education levels and the toxic levels, we divided the population component and income component by 1000. This ensures similar weighting for the calculating of the similarity score when comparing two counties. Note that a smaller similarity score (S) denotes a more similar county. Additionally, we display only the top five counties in decreasing order of their similarity. For example in Figure 6, a user has selected Val Verde County, TX and based on the selection we have identified similar counties in Colorado, Oklahoma, Idaho and a couple of counties in Pennsylvania. An expert could then examine the incidence rates for these counties and further investigate reasons for differences in incidence rates despite similarities in population, income, education and toxic risk.

Selecting a Cluster of Counties

A *disease cluster* is defined as a cluster that has an unusually high concentration of disease incidence in a region that is unlikely to have occurred by chance [14]. It is often referred to as a *hot-spot cluster* [12]. In situations where a user may want to explore such a region or a cluster of counties, we provide the ability to interactively select a cluster of counties. Based on the user's selection, we draw a translucent circle representing the counties selected by the user. All the counties that lie completely within as well as those touching the drawn circle are considered as part of the cluster. For each cluster, we compute and display the average prevalence, population, income, education and toxic risk for that cluster in the form of bullet graphs, as shown in Figure 7. Additionally, for each cluster of counties we display the average, minimum and maximum prevalence. We also maintain a running minimum and maximum for the three values displayed in the bottom panel. For example, in Figure 7, the second cluster of counties has the highest average, minimum and maximum prevalence shown in red. At this point, we allow the user to only select a circular region but we are working on allowing a user to specify a region by drawing an arbitrary shape.

IMPLEMENTATION DETAILS

We decided to use a framework for ActionScript 3 called ClearMaps [25], produced by Sunlight Labs. ClearMaps provides the base foundation for rendering Shapefile (shp) data in Adobe Flash. We found it to be a good solution due to the lightweight nature of the framework, its rendering speed and its built-in mouse-over functionality.

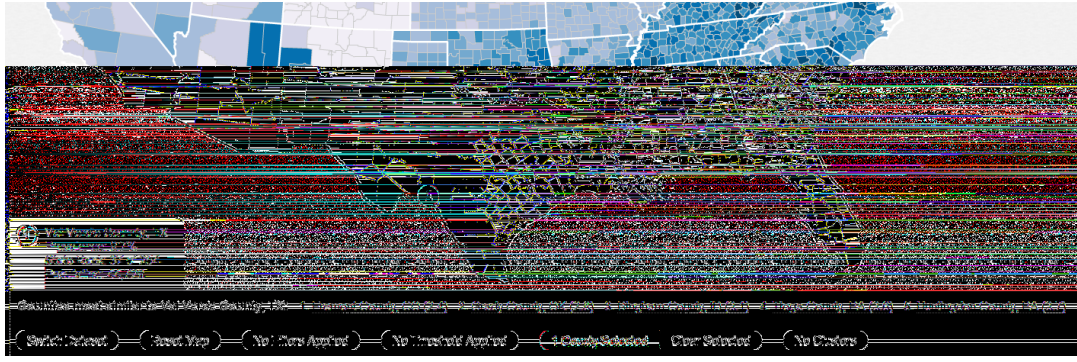


Figure 6. Similar Counties displayed once a county is selected. A similarity metric based on population, income, education and the toxic risk is computed and up to a maximum of five counties are displayed with their similarity metric shown in brackets. Here we see that a user has selected Val Verde County, TX and five other similar counties are shown below with their similarity scores.

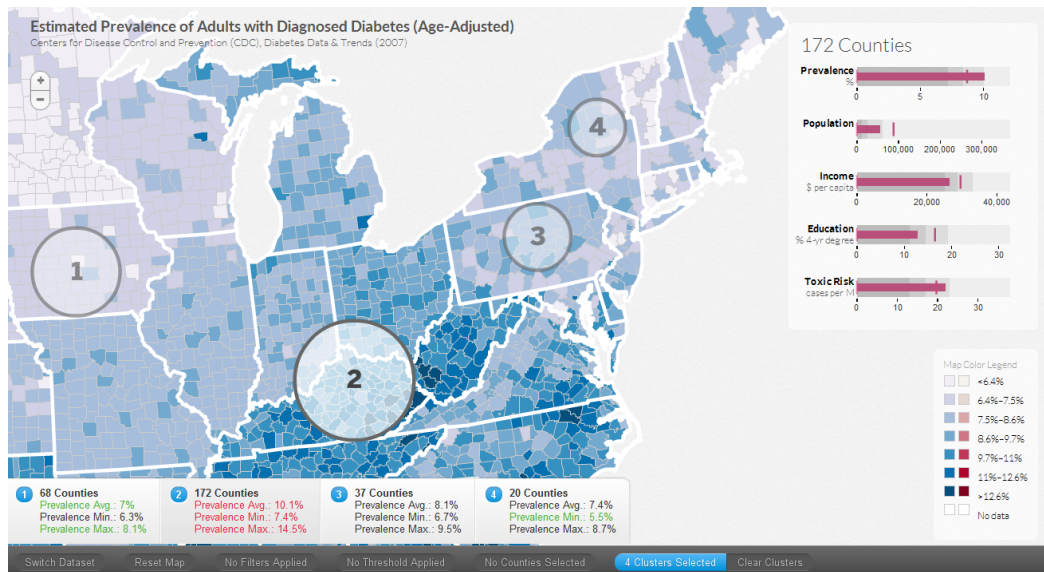


Figure 7. Selecting a cluster of counties: A user can interactively specify a region to create a cluster of counties. For each cluster, we compute and display the average, minimum and the maximum prevalence. Here we have selected four clusters of counties. In this case, cluster 2 has the highest values for the average, minimum and maximum prevalence (shown in red) and cluster 1 has lowest average and maximum prevalence (shown in green). For the current cluster (cluster 2 in this case), the bullet graphs on the right show the variables averaged for those counties. Here we only show the national average (dark line), since clusters can cross state boundaries.

The team at Sunlight Labs mentions that the tool was designed to combat two issues of geographical visualizations produced for the browser: rendering of vector data in-browser and reducing vector data size for timely loading [38]. Both of these issues are addressed by ClearMaps through the use of binary Shapefile data, in which the map features are converted into a compressed binary vector representation of original ShapeFile, significantly reducing the size of the file and making it much easier for the rendering engine to quickly rendering all 3140 counties in the U.S. Since the ActionScript 3 code is run by the Adobe Flash browser plug-in, a component installed on a wide majority of computers worldwide, our visualization can be viewed and used by anyone regardless of the computer platform (Windows, Mac, Linux) they use. The ClearMaps framework was fairly flexible and allowed us to build upon it, which we did extensively in order to create DiseaseTrends.

RESULTS

We think the best way to highlight the hypothesis generating ability in DiseaseTrends is to examine the system with the aim of exploring the data. Let us consider our diabetes data. We note that there is a very high diabetes incidence rate in the southeast, indicated by the high frequency of counties with dark red color (result of brushing high prevalence counties as shown in Figure 8). This region has recently been dubbed as part of the *Diabetes Belt* [10] and further research into the reasons for the high prevalence are being investigated. This Diabetes Belt, as identified, consists of parts of Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, North Carolina, Ohio, Pennsylvania, South Carolina, Tennessee, Texas, Virginia, West Virginia and the entire state of Mississippi. We can clearly see these regions in Figure 8, which not only helps confirm current findings in the field of preventative medicine, but also gives us confidence re-

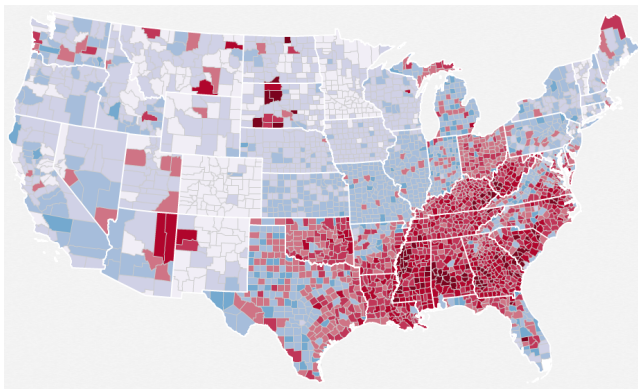


Figure 8. Visualizing the Diabetes Belt – brushing counties with high prevalence leads to a visual representation that highlights a large region in the southeastern United States. Parts of these states have been identified lately as the Diabetes Belt .

garding other unseen findings that we might unearth using DiseaseTrends.

We also notice that there are spurts of high incidence towards the top of Maine, center of South Dakota, and a couple of regions in Arizona. Upon mousing over Arizona, we see that the two counties with noticeably higher diabetes incidence rates than the rest are the Navajo and Apache Counties. With prevalence rates of 12.2% and 12.1%, respectively, they rank first and second in the state, and fall in the second highest break class of prevalence rates. We decide to focus in on Navajo County. From the coordinated views of socioeconomic factors shown on the right (Figure 9 - top), we see that Navajo County has a lower income, education, and air pollution level than the national average. It seems that there may be an association between low income, low education, and high incidence rates of diabetes, but we should take a look at a few other cases before forming these hypotheses. We further examine high prevalence counties in the western part of the United States, and select a few counties, as shown in bottom image in Figure 9. The selected counties seem to have a significantly high rate of incidence as compared to the national average. As it turns out, a Native American Reservation is situated in each of these counties. Based on conversations with our expert evaluator in the Public Policy department, it confirms the findings in their line of work regarding the high prevalence of diabetes in Native American reservations.

We further examine the data by looking at the counties in the southeast. Mousing over some of these dark blue counties, we see that the actual diabetes incidence rates seem slightly higher here than in Navajo County, and in terms of socioeconomic factors, they tend to be similar: lower-than-national-average income, but significantly lower education and air pollution levels. So, now we can more confidently posit that there is little significance of air pollution on diabetes. Navajo County has an air pollution level that is lower than the national average, but higher than these southern counties, yet has a higher diabetes incidence rate. Negative correlation between air pollution and diabetes seems unlikely, so we lean

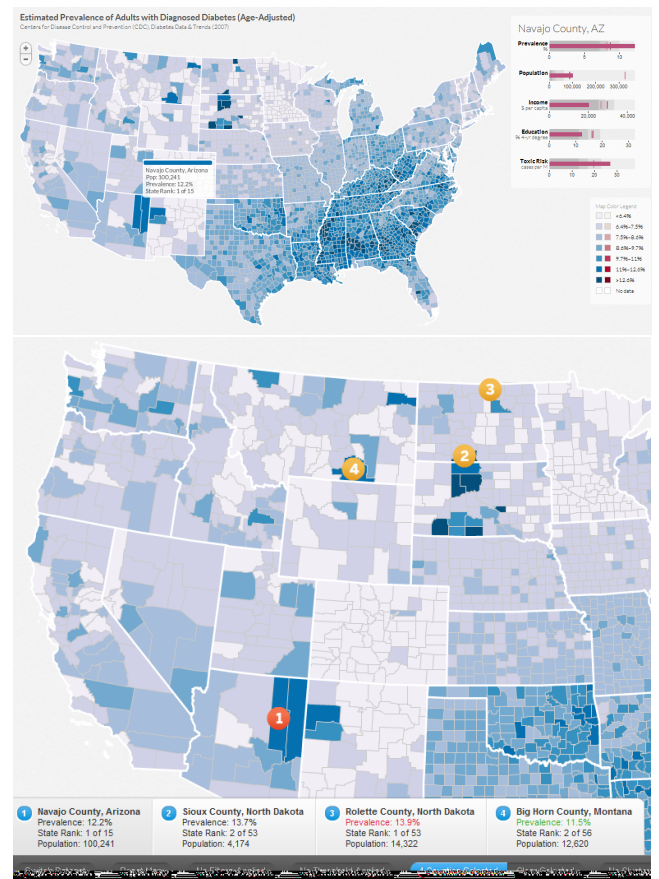


Figure 9. The top image shows the DiseaseTrends visualization when Navajo County is being explored. We can see that the county has a lower income, education and air pollution level than the national average. It is known that there is a big Native American reservation in the Navajo County. This is further explored by identifying counties with high incidence rates in the western part of the United States. The bottom image shows regions with high prevalence of diabetes and interestingly enough there a Native American reservation is situated all the counties selected here.

towards saying there is no significant relationship between the two. On the other hand, the likelihood that income and education levels are negatively correlated with diabetes incidence rate is still there, if not stronger. It is evident that the majority of the Southern counties have higher than average diabetes incidence rates, and now we see they also tend to have lower than average income and education levels.

Notice how easily we moved from one county to the next, making comparisons, spotting patterns, and forming hypotheses. At the end of the trial, we can not claim to have a clear answer on which factors affect disease incidence and to what extent, but we think this example makes clear how useful our visualization tool can be as an overview for the trends and foundation for further research.

EXPERT FEEDBACK

We obtained feedback from four experts in the fields of public policy, economics and public health. Instead of conducting a task-based evaluation, we gave each of them a demonstration and then allowed them to interact with the

data. Based on their interactions and prior knowledge, we identified certain trends as mentioned above regarding high prevalence of diabetes in Native American Reservations and regarding the Diabetes Belt.

All the experts were highly impressed with the user interface and specifically the ability to filter data, select specific counties for compare-and-contrast type tasks and the ability to specify a cluster of counties. The economist provided us with feedback regarding displaying similar counties based on the current selected counties. She said that it might be interesting to examine counties with similar population, income, education and air pollution levels but varying incidence rates. We incorporated her suggestion as shown in Section . She also mentioned that it may be interesting to visualize other variables such as racial distribution since they have seen some correlations with certain races and high incidences of diabetes. Unfortunately, they did not have any county level data for the entire country and so this could not be incorporated. We have started conversations regarding collaboration on the state level data that they have.

Based on conversations with the public policy expert, we found that they would like to further examine whether caloric intake data was available to investigate a correlation between caloric intake and diabetes. The public health researcher mentioned that it certainly could be a useful tool for public health researchers as a way to identify counties/clusters of interest for further investigation. He also felt that the tool might suggest too much causation, rather than simply association. We have been particularly cautious with not implying any causation throughout the paper and have merely hinted at possible associations with our variables. It is completely up to a researcher in the field of public policy / public health to further investigate the findings.

CONCLUSION AND FUTURE WORK

We have presented an interactive geographic visualization tool that allows for exploration, examination and the generation of hypotheses for further study. Based on our expert feedback, we believe our visualization tool serves as an informative and effective overview of trends in disease incidence and specific socioeconomic factors. It allows a user to easily drill down on any trends and aspects in the data for further examination. Our visualization can easily incorporate additional analysis, whether it is by adding other factors, looking at other diseases, or implementing different coordinated views as part of the analysis. This generalization of the DiseaseTrends program to work with new datasets, both disease as well as external factors can be accomplished with minimal additional work to our existing program. In our next version, a user will be able to select from a number of preloaded datasets or upload their own data. Additionally, we plan to undertake a detailed task-based expert evaluation comparing of our system to evaluate the effectiveness of our tool. We are also planning to integrate Geographically Weighted Regression (GWR) into the next version of our tool. We are currently working on incorporating statistical cluster detection methods into the tool.

ACKNOWLEDGEMENTS

The authors wish to thank Yoonie Hoh for her help with the project and the Sunlight Labs.

REFERENCES

1. Centers for disease control and prevention: National diabetes surveillance system. county level estimates of diagnosed diabetes. <http://www.cdc.gov/diabetes/statistics/index.htm>.
2. National cancer institute. state cancer profiles: Incidence rates report. <http://statecancerprofiles.cancer.gov/incidencerates/index.php>.
3. United states department of agriculture: Economic research service. county-level education data. <http://www.ers.usda.gov/Data/Education/#list>, 2005.
4. Bureau of economic analysis: U.S. department of commerce. regional economic analysis: Local area personal income. <http://www.bea.gov/regional/reis/default.cfm#step2>, 2009.
5. United states environmental protection agency. technology transfer network: National scale air toxics assessment. <http://www.epa.gov/ttn/atw/nata2002/tables.html#table1>, 2009.
6. US census bureau small area income and poverty estimates (saip). <http://www.census.gov/did/www/saipe/data/maps/index.html>, 2012.
7. R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization*, 2005, pages 111–117, 2005.
8. G. Andrienko, N. Andrienko, U. Demsar, D. Dransch, J. Dykes, S. Fabrikant, M. Jern, M. Kraak, H. Schumann, and C. Tominski. Space, time and visual analytics. *International Journal of Geographical Information Science*, 24(10):1577–1600, 2010.
9. N. Andrienko and G. Andrienko. *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer Verlag, 2006.
10. L. E. Barker, K. A. Kirtland, E. W. Gregg, L. S. Geiss, and T. J. Thompson. Geographic distribution of diagnosed diabetes in the U.S.A. diabetes belt. *American Journal of Preventive Medicine*, 40(4):434–439, April 2011.
11. D. B. Carr. Designing linked micromap plots for states with many counties. *Statistics in Medicine*, 20(9-10):1331–1339, 2001.
12. J. Chen, R. E. Roth, A. T. Naito, E. J. Lengerich, and A. M. MacEachren. Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of U.S. cervical cancer mortality. *International journal of health geographics*, 7:57, Jan. 2008.

13. S. Colagiuri. Epidemiology of prediabetes. *Medical Clinics of North America*, 95(2):299 – 307, 2011. Prediabetes and Diabetes Prevention.
14. E. K. Cromley and S. L. McLafferty. *GIS and Public Health*. The Guilford Press, 2002.
15. B. Davenhall. Your health depends on where you live. http://www.ted.com/talks/bill_davenhall_your_health_depends_on_where_you_live.html, 2010.
16. J. Dykes, A. MacEachren, and M. Kraak. *Exploring geovisualization*, volume 1. Pergamon, 2005.
17. B. K. Edwards, E. J. Feuer, and L. W. Pickle. US predicted cancer incidence, 1999: Complete maps by county and state from spatial projection models. In *NCI Cancer Surveillance Monograph Series*, 2003.
18. S. Few. Bullet graph design specification. In *Perceptual Edge - White Paper*, 2010.
19. A. Fredrikson, C. North, C. Plaisant, and B. Shneiderman. Temporal, geographical and categorical aggregations viewed through coordinated displays: A case study with highway incident data. In *Proceedings of the Workshop on New Paradigms in Information Visualization and Manipulation*, pages 26–34. ACM Press, 1999.
20. M. Harrower and C. A. Brewer. Colorbrewer.org: An online tool for selecting colour schemes for maps. *Cartographic Journal*, 40(1):27–37, Jun 2003.
21. B. Herrera and C. Lindgren. The genetics of obesity. *Current Diabetes Reports*, 10:498–505, 2010.
22. Q. Ho, P. Lundblad, T. Åström, and M. Jern. A web-enabled visualization toolkit for geovisual analytics. *SPIE: Electronic Imaging Science and Technology, Visualization and Data Analysis, Proceedings of SPIE, San Francisco*, 2011.
23. M. Jern, T. Astrom, and S. Johansson. Geoanalytics tools applied to large geospatial datasets. In *Information Visualisation, 2008. IV'08. 12th International Conference*, pages 362–372. IEEE, 2008.
24. M. Jern, J. Rogstadius, and T. Astrom. Treemaps and choropleth maps applied to regional hierarchical statistical data. In *Information Visualisation, 2009 13th International Conference*, pages 403–410. IEEE, 2009.
25. S. Labs. Clearmaps [mapping framework]. <http://github.com/sunlightlabs/clearmaps/>, 2010.
26. Q. Li, X. Bao, C. Song, J. Zhang, and C. North. Dynamic query sliders vs. brushing histograms. In *CHI '03 extended abstracts on Human factors in computing systems*, CHI EA '03, pages 834–835, New York, NY, USA, 2003. ACM.
27. A. M. MacEachren. *How Maps Work: Representation, Visualization, and Design*. The Guilford Press, New York, 2nd ed. edition, 2004.
28. A. M. MacEachren, S. Crawford, M. Akella, and G. Lengerich. Design and implementation of a model, web-based, gis-enabled cancer atlas. *Cartographic Journal*, 45(4):246–260, Nov. 2008.
29. A. M. MacEachren and M.-J. Kraak. Research challenges in geovisualization. *Cartography and Geographic Information Science*, 28(1):3–12, Jan. 2001.
30. J. E. Mersey. Color and thematic map design: The role of colour scheme and map complexity in choropleth map communication. *Cartographica*, 27(3):1–167, 1990.
31. E. Nelson. The impact of bivariate symbol design on task performance in a map setting. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 37(4):61–78, 2000.
32. T. Ormsby, E. Napoleon, and R. Burke. *Getting to Know ArcGIS Desktop: The Basics of ArcView, ArcEditor, and ArcInfo Updated for ArcGIS 9*. Esri Press, 2004.
33. C. Plaisant and V. Jain. Dynamaps: Dynamic queries on a health statistics atlas. In *CHI '94 Video Program, ACM CHI '94 Conference Companion*, pages 439–440. ACM, 1994.
34. J. Symanzik, G. Klinke, S. Klinke, S. Schmelzer, D. Cook, and N. Lewin. The arcview/xgobi/xplore environment: Technical details and applications for spatial data analysis. In *ASA Proceedings of the Section on Statistical Graphics*, pages 73–78. American Statistical Association, 1997.
35. Tableau. Tableau software. <http://tableausoftware.com/>, 2012.
36. TIBCO. Tibco spotfire - business intelligence analytics software. <http://spotfire.tibco.com/>, 2012.
37. B. Tomaszewski, A. Robinson, C. Weaver, M. Stryker, and A. MacEachren. Geovisual analytics and crisis management. In *Proc. 4th International Information Systems for Crisis Response and Management (ISCRAM)*, pages 173–179, 2007.
38. K. Webb. Clearmaps: A mapping framework for data visualization. <http://sunlightlabs.com/blog/2010/clearmaps-mapping-framework/>, 2010.
39. G. Wessel, R. Chang, and E. Sauda. Visualizing GIS: Urban Form and Data Structure. In *96th Annual Conference of Association of Collegiate Schools of Architecture (ACSA)*, pages 378–384, 2008.