Visualization Based Approach for Exploration of Health Data and Risk Factors

Xiping Dai and **Mark Gahegan**
Department of Geography & GeoVISTA Center
Pennsylvania State University
University Park, PA 16802, USA
Telephone: +1-814-865 2612
FAX +1-814-863-7943
xpdai@psu.edu, mng1@psu.edu

Abstract

Developing categories for health data is a crucial step for health researchers to explore, analyze, display and disseminate information about and relationships between health data and related socio-economic factors. The process for developing categories and exploring the relationships between health data and risk factors involves three (traditionally separate) aspects that encompass: (i) mentally-held concepts, (ii) the data, and (iii) the available categorization methods. Current approaches to exploratory analysis do not integrate these three aspects well, leading to difficulties and inertia during the process. This paper describes our efforts to build a system that encompasses the full extents of the categorization problem as it applies to the analysis of health data and risk factors. The system employs a range of visual and computational components, including a concept browser, a variety of exploratory visualization tools and many different classifiers. It is recognized specifically that successful analysis involves the dynamic interaction between all of these tools within a cycle of scientific investigation. In our example, we explore the relationships between cancer mortality and the risk factors by taking into account both the abstract conceptualization of the relationships and the emergent properties, which the dataset and classifiers provide. The tools in the system, ranging from visualization to classification components, are shown in the case study, along with how they might be used.

1. Introduction

With the increasingly large and complex health data and their risk factors information, there is a new emphasis in epidemiology research that encourages exploration of health data to generate new hypotheses. Developing categories is a crucial step for health researchers to explore, analyze, display and disseminate information about health, social and economic factors, and the relationships between them. In general, categories are constructed by specific categorization methods based on analyzing data and / or utilizing an expert's domain knowledge; so categories are not only determined by the underlying conceptualization, but also by a combination of the categorization methods chosen and data used in their construction, and the domain knowledge and experience that are brought to bear during construction. Categorization and exploration of relationships between health data and their risk factors are iterative, learning processes, which are full of trading off between the data, methods and conceptual knowledge to reach a (hopefully) stable equilibrium.

However, current approaches cannot provide an efficient iterative environment to users, since the three aspects, i.e. human concept, data, categorization methods, are not well integrated in any single computational system. These tools representing the three aspects are largely separated from each other, with no means of interaction; in fact they typically reside in quite distinct software products with poor integration between them. As a result, many categorization methods are heavily computational and opaque to analysts, who can only examine whether the categorization is a success or failure by examining the results. They have no effective approach to understand whereabouts the methods are failing, if the process itself fails. Also, there is no expression of the conceptual component within the system, it typically remains locked in the head of the analyst.

A model is need that can integrate the three spaces of categorization, i.e., data, methods, and human concepts, in the exploration process, and enable analysts to move seamlessly between the three spaces until interesting patterns and relationships between health data and risk factors are identified.

2. Related Works

Within the geographic domain, there are two distinct sets of techniques to encode and depict conceptual structures. One set of techniques involves research on ontologies and concept maps (Smith and Mark, 1998; Rodriguez and Egenhofer, 2003). The other set of tools support exploration and knowledge discovery activities, such as geovisualization and exploratory spatial data analysis (ESDA) (DiBiase, 1990; Dykes, 1997; MacEachren *et al.*, 1999; Gahegan 2001; MacEachren *et al.*, 2003). These two types of tools represent both ends of a continuum. The first set of tools, ontologies and concept maps, employs a top-down view of the world, and consider mentally held concepts[1] and their relationships. The second set of tools uses a bottom-up view of the world, i.e., the actual data to be analyzed and its emergent properties. However, these sets of tools are largely separated from each other, and typically reside in quite distinct software products with poor integration between them. But activities at either end of this continuum of science activities should not be artificially isolated by the systems we use because they are intimately connected in a conceptual sense.

The literature (DiBiase, 1990; MacEachren *et al.*, 1999) indicates that visualization has potential to help analysts to iteratively explore data samples, incorporate their knowledge, display classification, and identify problems. Lucieer and Kraak (2002) implemented visualization tools together with a supervised fuzzy classification algorithm to improve a geoscientist's insights into uncertainty in remotely sensed image classification. Functionalities such as dynamically linked views and geographic brushing are emphasized in the visualization tools to explore data in a set of multiple data views (Monmonier, 1989; Carr, 1987). Thus, visualization has the potential advantage to be an interface between human and computational components since it creates graphical images of data, helps humans to explore, reason and learn effectively, and usually enables an interactive visual exploration of the data.

[1] Here we use the word 'concept' to indicate a mental notion of some set of like entities. Examples might be the mental idea of cancer mortality or the positive relationship between cancer mortality and poverty.

Visualization has been used in the health data research since the seminal work by John Snow (1855) on spatial analysis of cholera epidemiology. These days, atlases of diseases, such as cancer mortality, heart disease mortality, and so on, are effective in supporting research and disseminating results (Pickle *et al*. 1999). Edsall (1999; 2003) created a geospatial data exploration system, including a choropleth map, a parallel coordinate plot, and a scatterplot, to analyze health statistics data. He argued that the multidimensional nature of health statistics and their analysis called for the integrated approach for geovisualization. Carr *et al*. (2000; 2005) developed a linked micromap template to display maps with boxplots, dotplots, and other statistical graphics. These linked micromaps are implemented to disseminate health data for public use in an easily interpretable format. Carr *et al*. (2005) also designed CCMaps, a conditioned choropleth mapping tool. The CCMaps tool provides a matrix of conditioned choropleth maps to facilitate exploration of two variables to a third variable, on which users are able to condition by using sliders. MacEachren *et al*. (2003) explored the trend in lung cancer mortality for white females with the support of visualization tools in ESTAT. Anselin *et al*. (2004) developed the GeoDA spatial analysis toolkit to explore patterns of colon cancer incidence in parts of Appalachia.

3. Our Solution

This study introduces an iterative approach to explore health data and their socio-economic factors from various integrated perspectives, such as data, classification methods, and human conceptualization of relationships (shown as a concept map). This integrated approach is enabled by the visualization interface, via which the methods drawn from data exploration, statistical summaries, classification methods and conceptualization of relationships are linked and able to interact with each other. Users can move seamlessly between the data exploration, classification processes and their mental conceptualization until new and informative relationships are identified. This system, thus, will facilitate and connect together the processes of 1) creating, browsing and revising concepts in ontological and taxonomic browsers, 2) selecting concepts to use in a specific analysis exercise, 3) exploring the data to help formulate concepts from emergent structures, and 4) dynamically modifying the concepts according to the relationships emerged from the data (i.e. relationships that do not align well with mental concepts).

Specifically, the initial relationships can be constructed by users based on their domain knowledge or data and visualized in conceptVISTA, a kind of dynamic concept map based on the TouchGraph graph visualization package (www.tuuchtraph.com). Data, on the other hand, will be explored in various multivariate exploratory graphs, such as parallel coordinate plots, scatterplot matrices, and choropleth maps, in both attribute and geographical space. At this step, data are preprocessed by choosing informative variables, and rejecting outliers based on the data patterns and statistical summaries over all of the attributes. The intermediate steps of the exploration process and relationships between variables can also be examined in the visualization tools interactively, including parallel coordinate plots, scatterplot matrices, and choropleth maps, since the data distributions and bivariate relationships can be visualized and mapped based on the criterion that observations with similar patterns are close to each other. Potentially new and improved relationships, then, can be achieved by allowing users to choose the right variables, compare relationships of variables between different sub regions, and choose appropriate methods between different types of classifiers, or refine their conceptualization of the problem via concept map.

4. Case study

4.1. Data and study area

Appalachia's mountainous region is selected as one of the study areas in this research. In general, Appalachian states exhibit cancer mortality rates that are greater than the national average, with particularly high mortality rates of lung, cervical and colorectal cancers (Table **1**).

Table **1**: Appalachia Cancer Network Age-adjusted Cancer Mortality Rates per 100,000, 1991-1995.

	Number of Appalachian Counties	All Cancers	Lung	Breast	Cervical	Colorectal	Prostate
United States		171.4	49.8	26.0	2.8	17.8	26.1
Tennessee	50	180.3	59.2	24.6	3.0	16.6	25.7
Kentucky	49	197.6	71.2	24.2	4.5	18.1	25.3
Virginia	23	176.0	56.5	25.9	2.9	15.7	22.3
West Virginia	55	189.5	61.8	23.9	3.7	18.8	24.8
Ohio	29	184.7	58.7	25.7	3.5	19.6	23.2
Pennsylvania	52	177.0	47.6	27.6	2.7	20.1	25.5
Maryland	3	172.6	49.0	22.1	2.9	21.7	25.7
New York	14	174.9	49.7	27.2	2.8	19.4	26.7

Source: National Center for Health Statistics. State rates include only Appalachian counties. Cited from http://www2.kcr.uky.edu/acn/pdf%20files/mortality.pdf.

Generally speaking, the relatively high rates in cancer mortality are thought to be related to the lack of knowledge about cancer prevention, detection, and treatment. Moreover, the lack of knowledge about cancers can be linked to limited access to health information and health care services, which can further due to poverty caused by low incomes, low education levels, high unemployment, and other socio-economic factors that negatively impact public health in this region. Many Appalachian counties have higher poverty rates, lower education levels, and lower income as compared to the nation as a whole. While poverty might not be directly related to higher mortality rates of cancers, poverty and cancer mortality rates are certainly correlated.

Based on the discussion above, a possible general relationship between cancer mortality rates and the socio-economic risk factors, especially the poverty and poverty related status, are constructed as a concept map (Figure **1**). In this initial hypothesis, the cancer mortality rates are correlated with socio-economic status, with high cancer mortality rates relating to poverty generally. This hypothesis is the starting position for the analysis that follows, using breast cancer and cervical cancer mortality as the example.

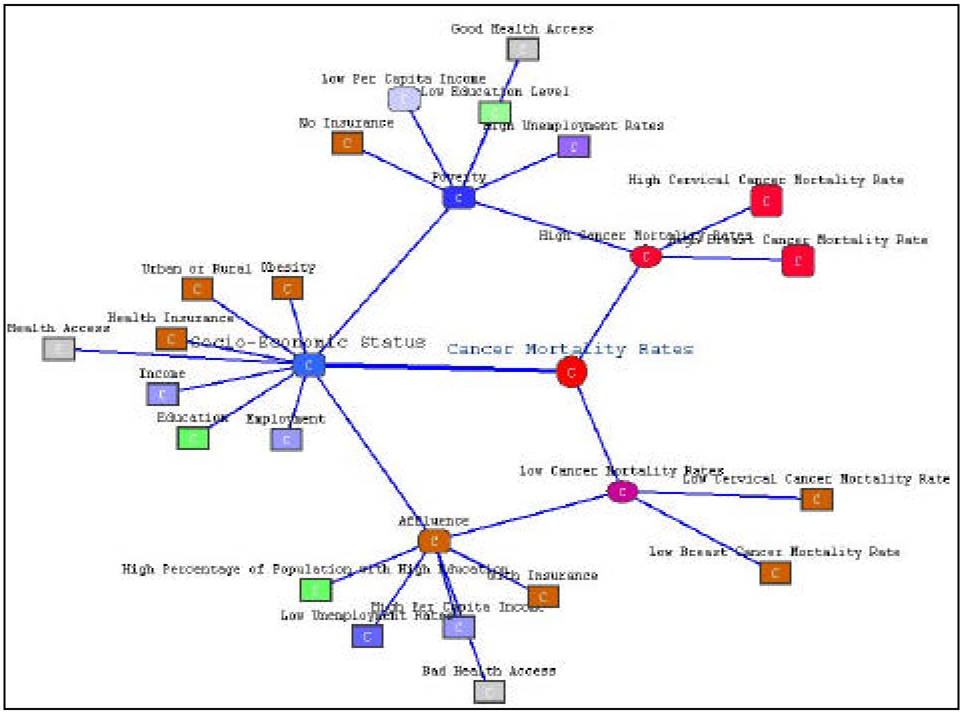


Figure **1**: The initial concept map showing relationships between cancers and poverty, and the socio-economic factors which contribute to the poverty.

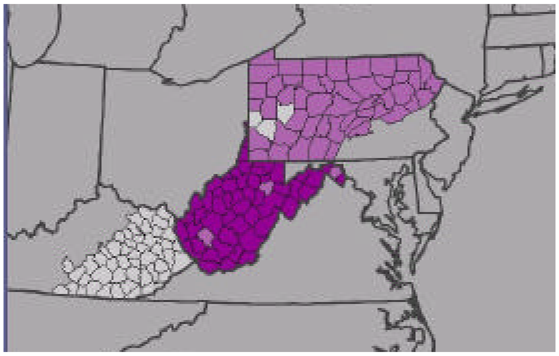


Figure **2**: Study area covers part of the states of Kentucky, West Virginia, and Pennsylvania.

This exploration of cancer and socio-economic factors involves cancer registry and demographic data, gathered at the county level for Appalachian states. The study area is focused in the 156 counties known as the NAACCR (North American Association of Central Cancer Registries) gold states (those states with good cancer registry data) within the Appalachian Cancer Network, which covers part of the states of Kentucky and Pennsylvania and all counties in West Virginia (Figure **2**). The cancer data used in this study are obtained from NAACCR. The census data can be downloaded from several websites, such as www.esri.com and http://nationalatlas.gov. Demographic datasets typically include attributes such as population, race, gender, housing, education, health care, and income information.

4.2. Category Exploration for Cancer Data and Risk Factors

The exploration process for cancer and risk factors data is one of iterative learning, involving exploration of three aspects, i.e. human concepts, data, categorization methods. The integrated system allows users to take advantages of both knowledge driven and data driven approaches, and enables analysts to move seamlessly between the three spaces.

The general relationships between cancer mortality rates and the socio-economic factors are illustrated in Figure **1**. High cancer mortality rates are usually associated with poverty and other socio-economic factors relating to poverty. Researchers tend to group data samples (counties here) according to the rates of different cancer mortality and values of socio-economic variables, and then explore the patterns and relationships between counties with different cancer mortality rates and counties in different groups for socio-economic status. For example, one might select only the breast cancer mortality rate in Figure **1**, and group the counties into three categories according to this rate, i.e. high, middle and low mortality rates (Figure **3**). Counties can then be grouped into the three categories with various exploratory tools, such as scatterplots and choropleth maps (Figure **4**). Spatial pattern can be explored via the map where we see that most of the counties in Pennsylvania are colored using dark and mid magenta colors, and many counties in Kentucky and West Virginia are shown as gray or magenta. This pattern illustrates that the counties with relatively high breast cancer mortality rate for white females for the time period 1970-1994 are clustered in Pennsylvania, and the counties with low or intermediate breast cancer mortality rates are mostly in Kentucky and West Virginia.

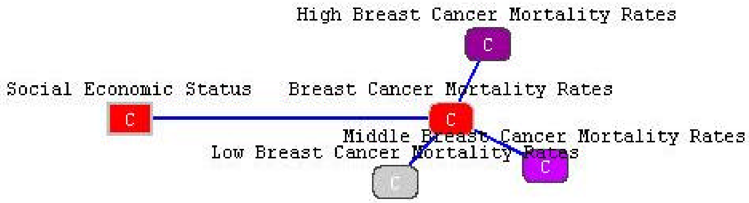

High Breast Cancer Mortality Rates
C
Social Economic Status
C
Breast Cancer Mortality Rates
C
Middle Breast Cancer Mortality Rates
C
Low Breast Cancer Mortality Rates
C

Figure **3**: Initial concept map derived from Figure **1** based on the rate of breast cancer.

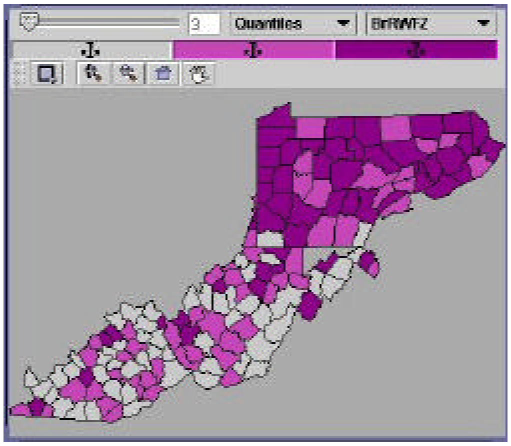


Figure **4**: Choropleth map with categories divided by low, middle, high rates in breast cancer mortality for white females 1970-1994.

The pattern of breast cancer mortality rate in these counties is apparently not random, so we can investigate reasons underlying why the cluster of counties with high breast cancer mortality rate occurs in Pennsylvania, leading us to examine the risk factors that might relate to this rate. One hypothesis is that the high cancer mortality rates in the Pennsylvania are related to the low socio-economic status, including poverty, in this area. So the next step adds one more variable, per capita income, into the analysis. A bivariate map with variables of breast cancer mortality rates and per capita income is generated to explore the spatial patterns between these two variables (Figure **5**). In Figure **5**, the variable breast cancer mortality rate is represented in magenta, and the variable per capita income is represented in green. The 156 counties are divided into nine classes based on these two variables, that is low-low, low-middle, low-high, middle-low, middle-middle, middle-high, high-low, high-middle, and high-high for the two variables. The counties colored with dark gray are those counties high in both breast cancer mortality rate and per capita income. Apparently, the counties with high per capita income are also clustered in Pennsylvania and the counties with low income and breast cancer mortality rate (shown in light gray) are mostly in West Virginia and Kentucky. This pattern indicates that the breast cancer mortality rate is positively related to income status. In other words, the phenomenon suggests that the hypothesis that high cancer mortality rates in Appalachian region result from to poverty in this region is questionable, for the breast cancer at least.

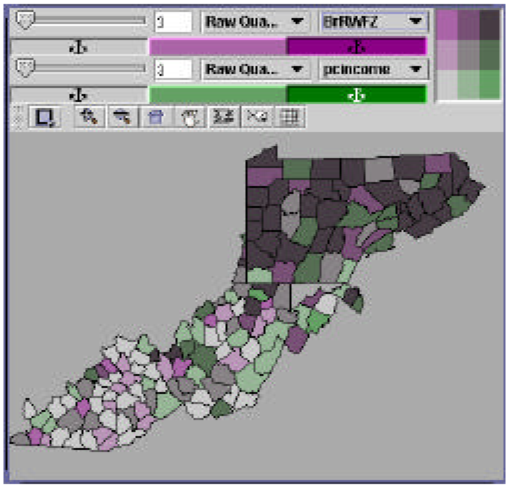
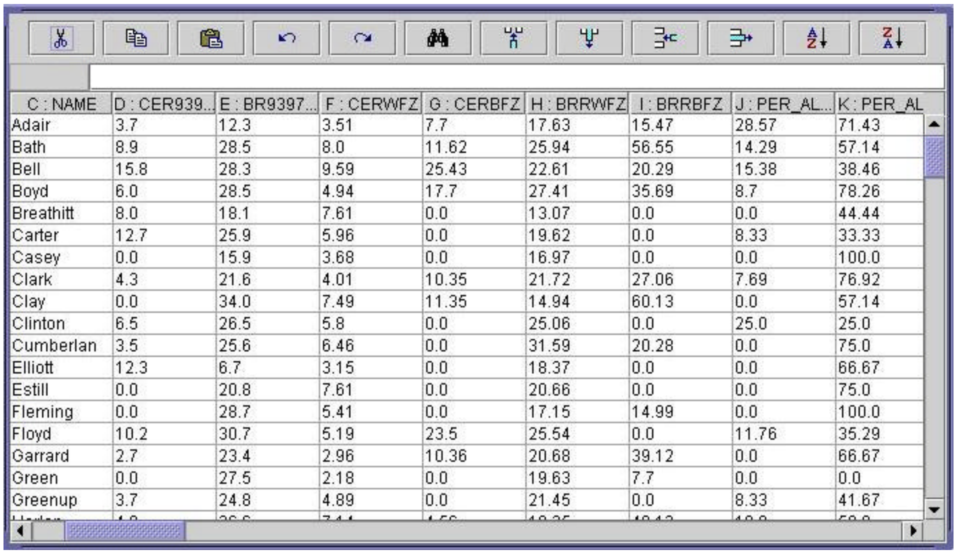


Figure **5**: A choropleth map with categories derived from breast cancer mortality rates and per capita income.

The categories generated by the bivariate classification method are easy to visualize and understand, and useful in exploring data patterns over one or two variables, as well as correlations between two variables. In Figure **5**, the observations (counties) are classified using the quantiles classification method. The relationship between breast cancer mortality rate and per capita income can be explored as a spatial pattern. However, only two variables are displayed at once. There is no direct method to include additional variables into these univariate and bivariate representations. Thus, there is no direct means to explore the relationships between the two displayed variables and any other potential risk factors. But, categories embedded in the datasets that might explain breast cancer mortality are typically not determined by only one or two variables, but rather they are related to many features or characteristics, for example, other socio-economic variables including education, health services and so forth.

A variety of exploratory data analysis tools are implemented in the system, including the spreadsheet, parallel coordinate plot (PCP), scatterplot, choropleth map, and matrix of scatterplots and/or maps. These exploratory analysis tools contribute to the data visualization and analysis from different perspectives. For instance, the spreadsheet can list all of the data in numeric format; the PCP shows the values of all of the variables in parallel axes; the scatterplot shows values of a pair of variables into an attribute space; and the choropleth map displays categories of observations geographically. Since different visualization tools have different advantages, we use a combination of several tools to take advantage of the useful features they each offer.

There are forty-six variables in this experiment, from mortality rates of various cancers for the time period 1070-1994 to social and economic information, such as population, per capita income, rent, education, smoking history, obesity, and so forth, for 156 counties in the study area. The dataset is quite large in terms of the number of variables and observations (counties). The spreadsheet (Figure **6**) can list all of the dataset, but it is difficult to find patterns through the plain numbers in the table.



C : NAME	D : CER939...	E : BR9397...	F : CERWFZ	G : CERBFZ	H : BRRWFZ	I : BRRBFZ	J : PER_AL...	K : PER_AL
Adair	3.7	12.3	3.51	7.7	17.63	15.47	28.57	71.43
Bath	8.9	28.5	8.0	11.62	25.94	56.55	14.29	57.14
Bell	15.8	28.3	9.59	25.43	22.61	20.29	15.38	38.46
Boyd	6.0	28.5	4.94	17.7	27.41	35.69	8.7	78.26
Breathitt	8.0	18.1	7.61	0.0	13.07	0.0	0.0	44.44
Carter	12.7	25.9	5.96	0.0	19.62	0.0	8.33	33.33
Casey	0.0	15.9	3.68	0.0	16.97	0.0	0.0	100.0
Clark	4.3	21.6	4.01	10.35	21.72	27.06	7.69	76.92
Clay	0.0	34.0	7.49	11.35	14.94	60.13	0.0	57.14
Clinton	6.5	26.5	5.8	0.0	25.06	0.0	25.0	25.0
Cumberlan	3.5	25.6	6.46	0.0	31.59	20.28	0.0	75.0
Elliott	12.3	6.7	3.15	0.0	18.37	0.0	0.0	66.67
Estill	0.0	20.8	7.61	0.0	20.66	0.0	0.0	75.0
Fleming	0.0	28.7	5.41	0.0	17.15	14.99	0.0	100.0
Floyd	10.2	30.7	5.19	23.5	25.54	0.0	11.76	35.29
Garrard	2.7	23.4	2.96	10.36	20.68	39.12	0.0	66.67
Green	0.0	27.5	2.18	0.0	19.63	7.7	0.0	0.0
Greenup	3.7	24.8	4.89	0.0	21.45	0.0	8.33	41.67

Figure **6**: Spreadsheet showing cancer and risk factor data.

The parallel coordinate plot (PCP) maps the values of each variable to the projected locations along the axis representing that particular variable (Figure **7**). Patterns or distributions over variables can be identified by the positions of polylines, which represent counties in this case. Relationships between variables can be identified according to trends in the polylines, for example, a parallel polyline trend indicates a positive relationship and a cross trend suggests a negative relationship between the two neighbor variables. The polylines between the breast cancer mortality rate and the cervical cancer mortality rate in Figure **7** have a cross trend, which indicates a negative relationship between them. The trends of polylines can be visually observed by users, but the quantitative values of the trends need to be further examined in a detailed view of relationships between variables, such as in a scatterplot with a regression line and value (r^2) calculated.

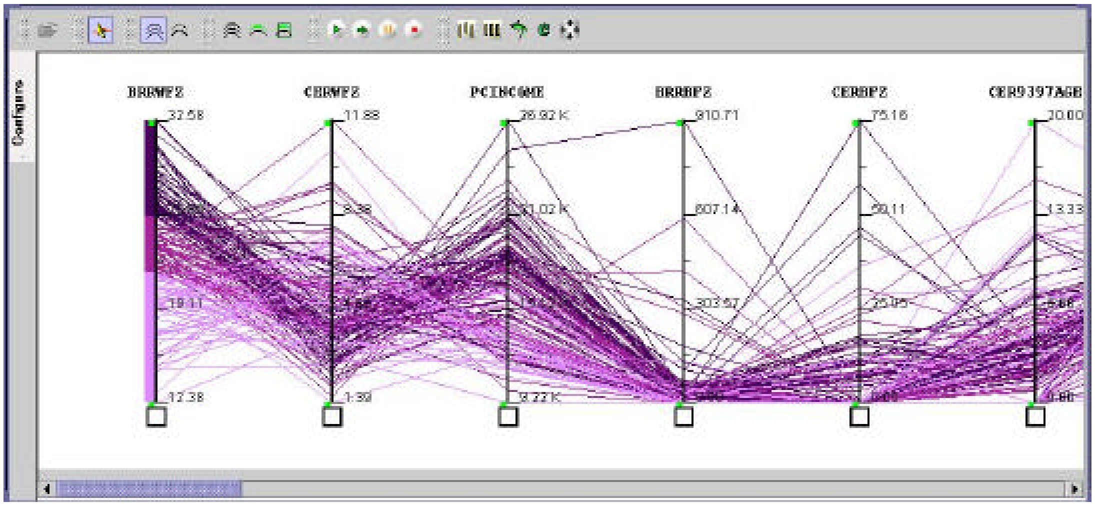
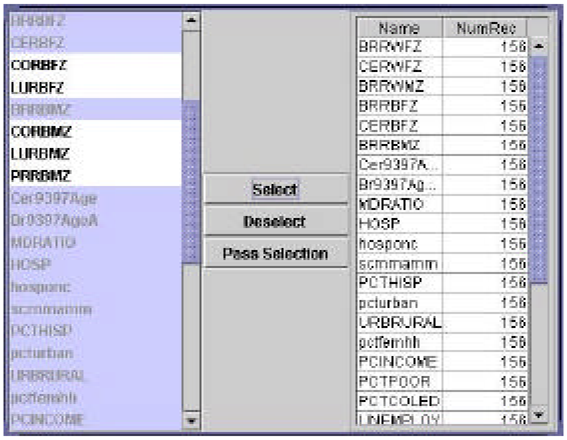


Figure **7**: Cancer and risk factor data shown in a parallel coordinate plot.

Furthermore, only a small portion of variables and data samples for counties can be effectively displayed at the same time on the screen in spreadsheets and PCPs, even though a scroll bar is provided (in the PCP) for users to browse the whole dataset, and the order of variables can be changed dynamically. A variable chooser tool is provided so that users can select any combination of variables to be displayed in all of the components in the system, and this selection can be changed interactively. The combination of the most informative variables, then, can be explored and examined together to improve the efficiency and effectiveness of the category development process.

In order to explore the mortality rates for breast cancer and cervical cancer, and their risk factors, subsets of variables can be selected out of forty-four, using the tool shown in Figure **8**. The variables shown have standardized names inherited from the creators of this dataset (National Cancer Institute); they are breast cancer rate for white females for time period 1970-1994 (BRRWFZ), cervical cancer rate for white females for time period 1970-1994 (CERWFZ), cervical cancer mortality rate for all ages from 1993 to 1997 (Cer9397Age), breast cancer mortality rate for all ages from 1993 to 1997 (Br9397AgeA), physicians per 1000 population (MDRATIO), hospitals per 1000 population (HOSP), hospitals with oncology service per 1000 population (hosponc), screening mammography facilities per 1000 population (scrnmamm), percentage of Hispanic origin (PCTHISP), percentage of urban (pcturban), USDA urban/rural code (0=most urban, 9=most rural) (URBRURAL), percentage of households headed by female (pctfemhh), per capita income (PCINCOME), percentage of adults over 25 with 4+ years of college education (PCTCOLED), unemployment rate (UNEMPLOY), percentage of women ages 50-64 who had a mammogram in past 2 years (mammog2ysm), percentage of persons ages 18+ who do not have a health plan or health insurance (NOINS), and so forth. This dataset was merged from two separate datasets, one of which encodes the variable name in capital letters, and the other has names in lower case.


BRRBFZ
CERBFZ
CORBFZ
LURBFZ
BRRBMZ
CORBMZ
LURBMZ
PRRBMZ
Cer9397Age
Br9397AgeA
MDRATIO
HOSP
hosponc
scrnmamm
PCTHISP
pcturban
URBRURAL
pctfemhh
PCINCOME

Name	NumRec
BRRWFZ	156
CERWFZ	156
BRRWMZ	156
BRRBFZ	156
CERBFZ	156
BRRBMZ	156
Cer9397A...	156
Br9397Ag...	156
MDRATIO	156
HOSP	156
hosponc	156
scrnmamm	156
PCTHISP	156
pcturban	156
URBRURAL	156
pctfemhh	156
PCINCOME	156
PCTPOOR	156
PCTCOLED	156
UNEMPLOY	156

Figure **8**: Thirty variables are selected out of the whole forty-four variables using the dataset in variable chooser tool.

Four out of the thirty chosen variables, i.e., breast mortality and cervical mortality rates, and their socio-economic risk factors, are selected for display in a scatterplot and map matrix (Figure **9**). The visual correlations between each pair of variables can be explored in both attribute space (scatterplots) and geographic space (maps). The red lines in the scatterplots are regression lines, and correlation and r^2 values are displayed at the top of each scatterplot panel. The colors for data samples (counties) are determined by the bivariate classification results based on the pair of variables displayed in the scatterplot. As before, there are nine color shades, and each represents one bivariate category. The counties in light magenta color have relatively high values for the variable represented by X-axis and low values for the Y-axis variable. The counties in light green have relatively high values for the variable represented by the Y-axis and have low values for the corresponding X-axis variable. Counties with relatively high values for both variables are shown using a dark gray color and counties with relatively low values for both variables are light gray in color, as the bivariate map in Figure **5** shows. Variable names can be found at the top row and the left most column.

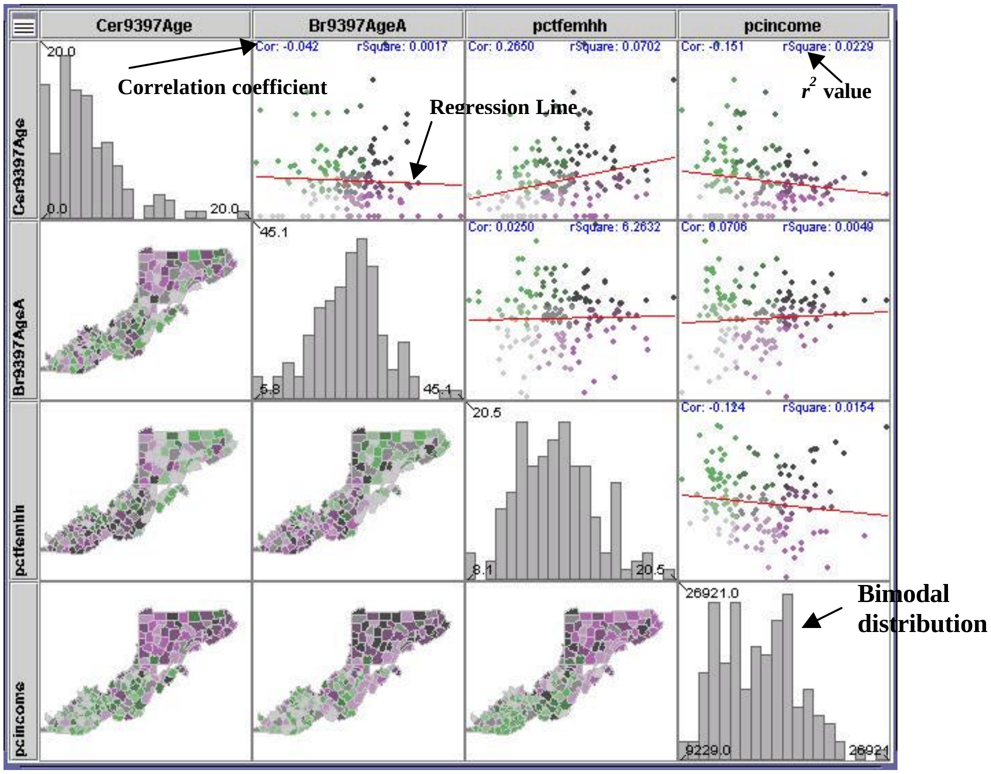


Figure **9**: Scatterplot and map matrix displaying four variables.

The detailed view of the scatterplot for cervical cancer and breast cancer rate variables is displayed in Figure **10**. The correlation between these two variables is very low at -0.042, but, there is a weak negative trend in the scatterplot, which indicates that counties with relatively high mortality rate for cervical cancer might have a lower mortality rate for breast cancer and vice versa.

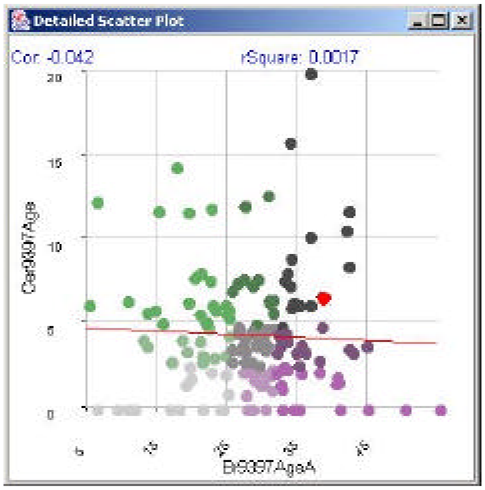


Figure **10**: A scatterplot with variables breast cancer rate and cervical cancer rate.

Observing the same pair of counties in a choropleth map in detail (Figure **11**), the geographic locations for counties in different categories derived by the bivariate quantiles classification are displayed. Counties with magenta color, which indicates they have relatively high rates in breast cancer mortality and low rates in cervical cancer mortality, are mostly clustered in Pennsylvania. Counties with light green color, which means they have relatively high rates in cervical cancer mortality and low rates in breast cancer mortality, are mostly in West Virginia and Kentucky.

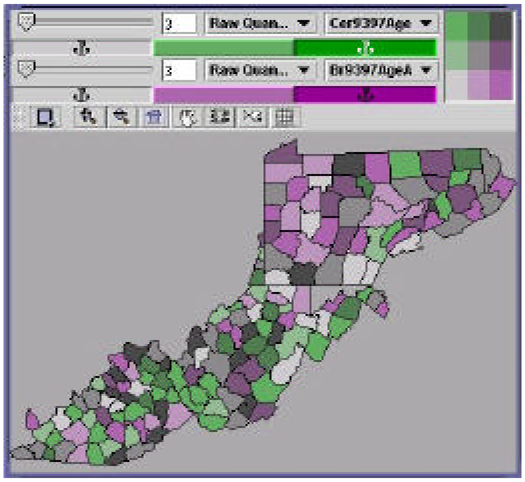


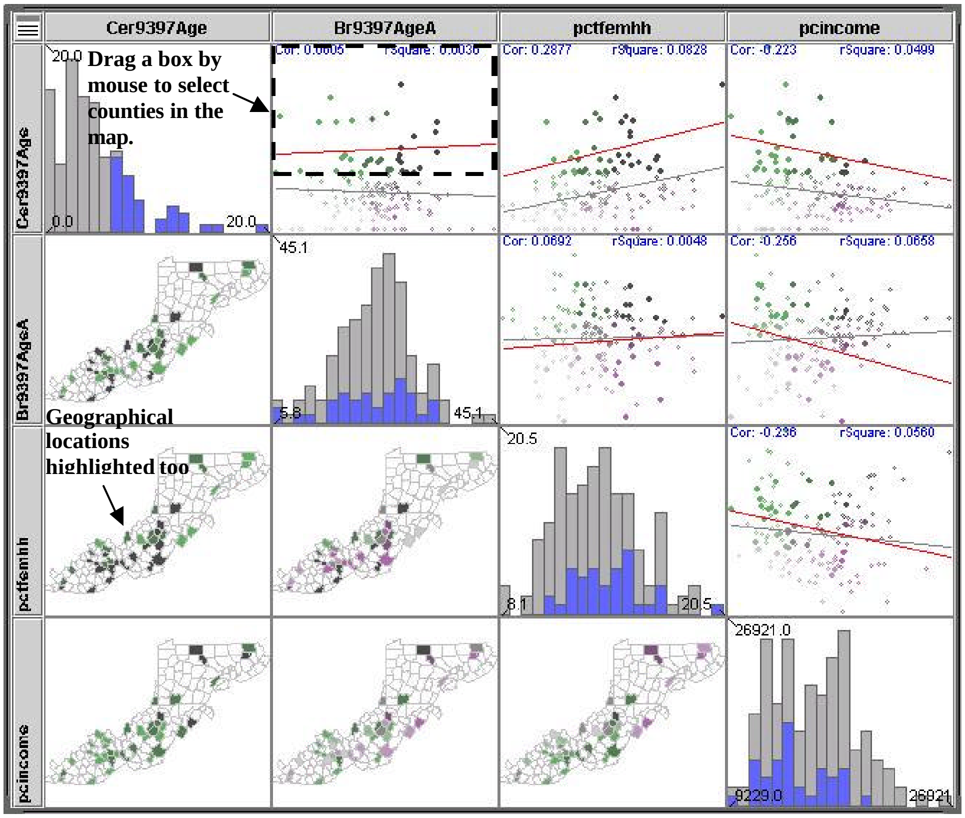
Figure **11**: Choropleth map with counties divided into nine categories according to quantiles classification on each of the two variables

In the scatterplot and map matrix in Figure **9**, pair-wise relationships can be observed and spatial patterns for pairs of variables can be identified. There is a relatively strong positive correlation (0.265) between cervical cancer mortality rate and the percentage of households with a female head. In the scatterplot, plotting this variable against cervical cancer mortality rate (Figure **9**), an increase in percentage of female head household, predicts an increase in cervical cancer mortality rate. In other words, there is a visually significant trend that the counties with a relatively high percentage of a female head of household show a high cervical cancer mortality rate. Geographically, counties with high values in both cervical cancer rate and percentage of female head of household cluster in West Virginia and Kentucky.

Apparently, the female head of household variable is negatively correlated with per capita income. Therefore, it is logical that the cervical cancer mortality rate decreases as per capita income increases. Counties with high per capita income and low cervical cancer mortality rate are mostly located in Pennsylvania and conversely, those with low per capita income and high cervical cancer mortality rate are clustered in West Virginia and Kentucky.

By contrast to cervical cancer mortality rate, breast cancer mortality rate shows a weak trend positively correlated with income, though statistically no strong correlation can be proved between these two variables. The counties with high breast cancer mortality rate and high income are mostly in Pennsylvania (the counties marked in dark gray or mid magenta in the choropleth map at the intersection of variables Br9397AgeA and pcincome, Figure **9**). Breast cancer mortality rate is not correlated with percentage of female head of household, since the corresponding regression line in the scatterplot is almost horizontal, and the correlation coefficient for these two variables is close to zero.

As discussed in the above section, there is only weak positive correlation between per capita income and breast cancer mortality rate, and between the cervical cancer mortality rate and breast cancer mortality rate for the entire study region as a whole. However, there are obvious spatial patterns for the cancer mortality rates and two risk factor variables, such that counties with high breast cancer mortality rate and per capita income are clustered in Pennsylvania, and counties with relative high cervical mortality rate and high percentage of female head household are mostly in Kentucky and West Virginia. Examining only the sub region with those counties having high rate in cervical cancer mortality, the relationship between per capita income and breast cancer mortality rate provides a different trend (the red regression line) from that of the region as a whole (the gray regression line) (Figure **12**). The counties in the sub region can be highlighted easily by dragging a box around these counties in the scatterplot of cervical cancer and breast cancer mortality rate (the dashed box in Figure **12**). And this selected subset of counties can be passed to other visualization tools, such as maps, so that the geographical locations for these counties are highlighted too. For these counties, breast cancer mortality rate decreases quickly with increase of per capita income as displayed in the scatterplot of breast cancer mortality rate and per capita income, where the solid dots represent those counties with high cervical cancer mortality rate and the red line is the regression line for only the counties selected. We see now that the correlation coefficient between breast cancer mortality rate and per capita income changes from +0.071 to –0.256. This illustrates an important point, that selected data can exhibit different patterns from those observed across all the counties, and the tools described here allow us to quickly create and test such subsets. Among the counties with high cervical cancer mortality rate, the counties with high per capita income have a lower likelihood of high breast cancer mortality.



Drag a box by mouse to select counties in the map.

Geographical locations highlighted too

Figure **12**: Scatterplot and map matrix displaying subset of counties with high cervical cancer rate selected.

Notice also that the regression lines in scatterplots are useful tools to summarize the relationships between two variables, especially the direction and strength of the trends, i.e., positive or negative correlations between them. The regression lines for the selected subset of counties are generated dynamically, as subset selected. The relationships between pair-wise variables for subsets of counties can be observed in both attribute and geographical space, thus, the variations among subsets or sub regions can be identified.

From Figure **4**, spatial patterns are observed in Pennsylvania with counties having both high per capita income and breast cancer mortality rates. These counties are selected for the following exploration in geographical space by simply dragging a box, which covers all of the counties in Pennsylvania, in the choropleth map (Figure **13**).

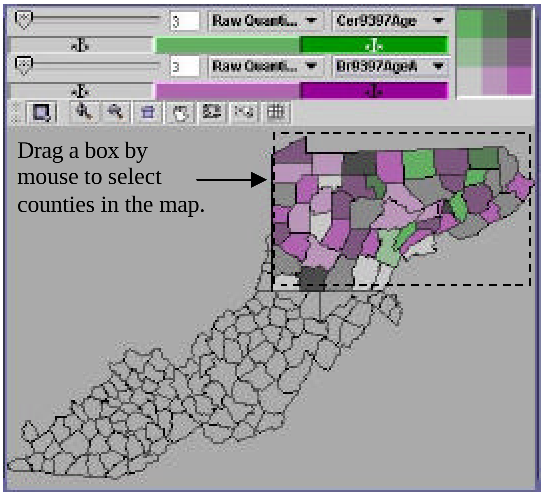


Figure **13**: Counties in Pennsylvania are selected.

The relationship between per capita income and breast cancer mortality rate changes dramatically and shows an opposite positive trend, (Figure **14**, Figure **15**) from that for the subset of counties having high cervical cancer mortality rate (Figure **12**, Figure **16**). There is now a fairly strong positive correlation, with the correlation coefficient at 0.4252, between breast cancer mortality rate and per capita income among the counties in Pennsylvania (Figure **14**, Figure **15**). In other words, in Pennsylvania, the counties with high per capita income are likely to also have a high breast cancer mortality rate. By contrast, counties with high cervical cancer mortality rate demonstrate a negative relationship, between breast cancer mortality rate and per capita income, with correlation coefficient at –0.256 (Figure **9**, Figure **16**). The difference between the relationships of breast cancer mortality rate and per capita income among different sub regions of the study area indicates that additional risk factors in these sub regions may be different.

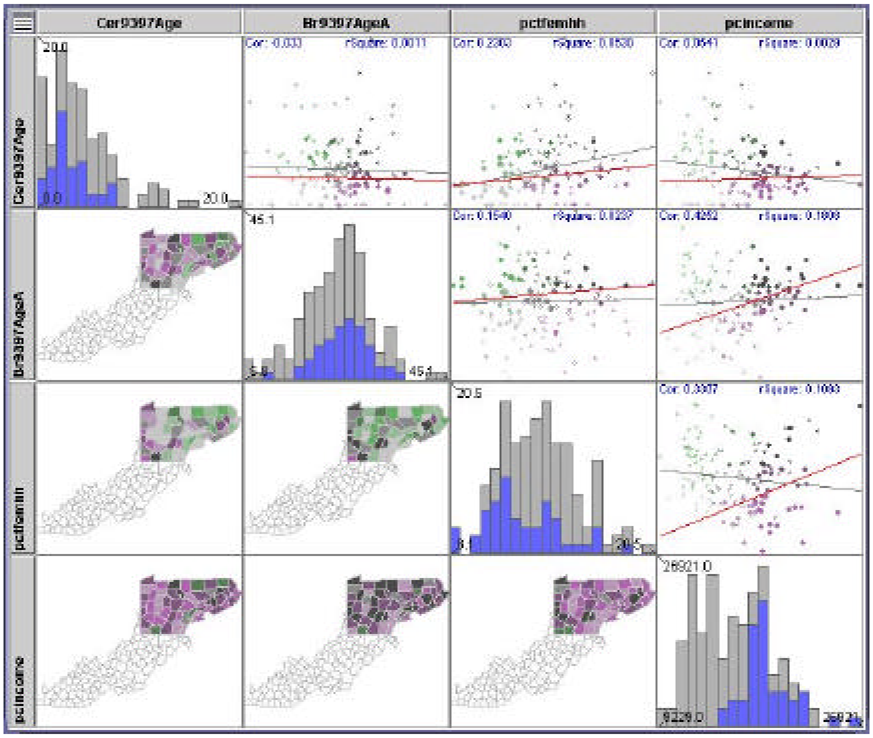


Figure **14**: Scatterplot and map matrix with counties in Pennsylvania selected.

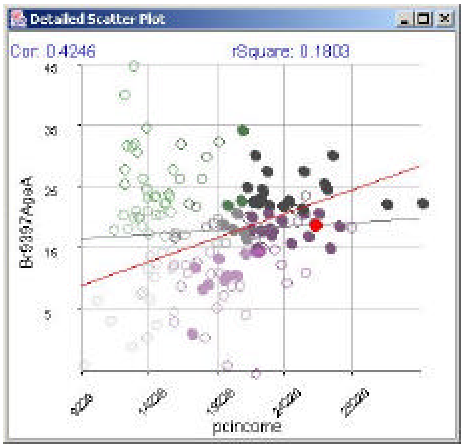


Figure **15**: Scatterplot for variables per capita income and breast cancer mortality rate with counties in Pennsylvania selected. The red line is the regression line for selected counties, and the gray line is for all counties.

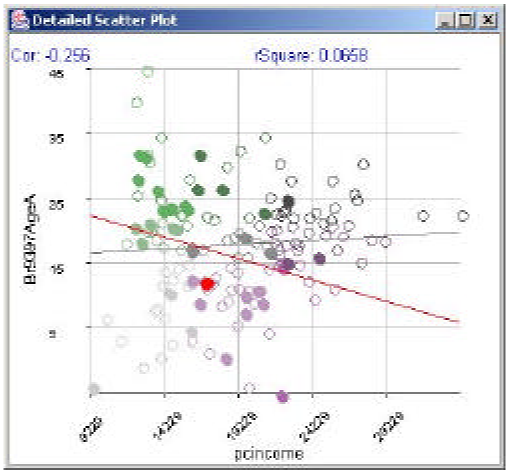


Figure **16**: Scatterplot for variables per capita income and breast cancer mortality rate with counties having high cervical cancer rate selected. The red line is the regression line for selected counties.

In contrast to breast cancer, the cervical cancer mortality rate for those Pennsylvania counties does not appear to be correlated with income (Figure **17**), which is also different from the previous observation, where cervical cancer mortality rates are negatively correlated with income visually (Figure **12**).

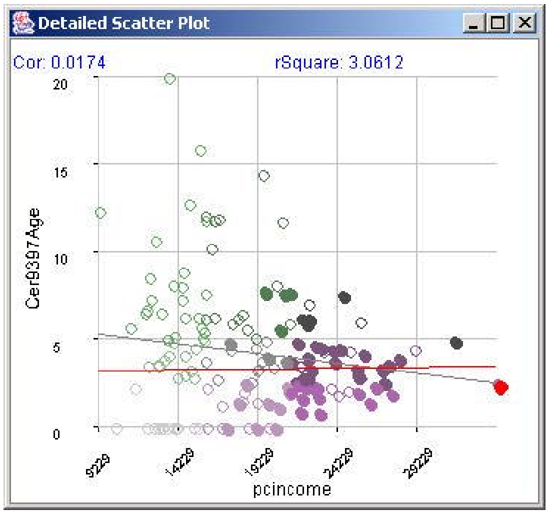


Figure **17**: Scatterplot for variables per capita income and cervical cancer rate with counties in Pennsylvania selected. The red line is the regression line for selected counties, and the gray line is the regression line for all counties.

Relationships between cancer mortality rates and other risk factors can be explored to show that there is a fairly strong positive correlation between breast cancer mortality rate and percentage of adults with college education, and most of the counties with high breast cancer mortality rate and high percentage of population with college education are in Pennsylvania (Figure **18)**. This phenomenon conforms to the previous observation that breast cancer mortality rate is positively correlated with income, since people with a college degree usually have a higher income. The negative correlations between breast mortality rate and percentage of population below poverty line, and unemployment tell the same story that, generally, people with higher economic status have a higher probability of developing breast cancer. In contrast to breast cancer, cervical cancer mortality rate is negatively correlated with economic status that counties with high mortality rate usually have high unemployment, high percentage of population below poverty line, and low percentage of population with college education, and these counties are mostly in Kentucky and West Virginia (Figure **19**).

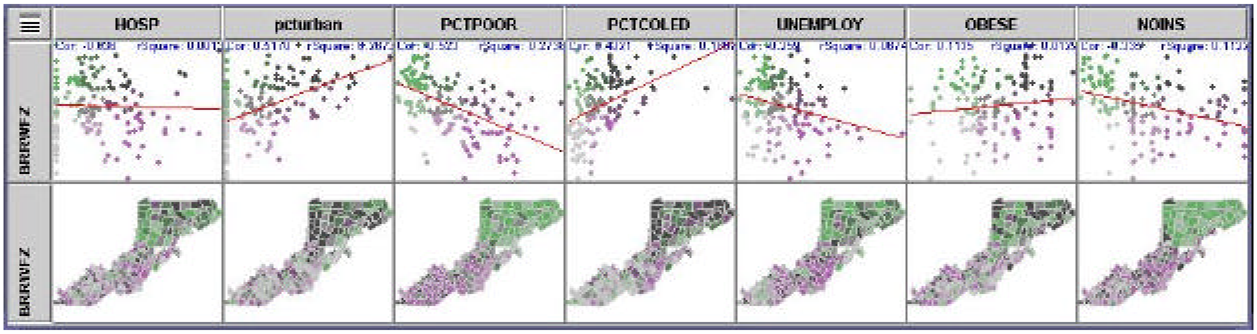


Figure **20**: The map and scatterplot matrix with the display of breast cancer mortality rate for white females during the time period 1970-1994, and hospital to population rate, percentage of urban, percentage of population below poverty line, percentage of adults with college education, and unemployment rate, % of persons ages 18+ who are >120% of the median body mass index, and % of persons ages 18+ who do not have a health plan or health insurance.

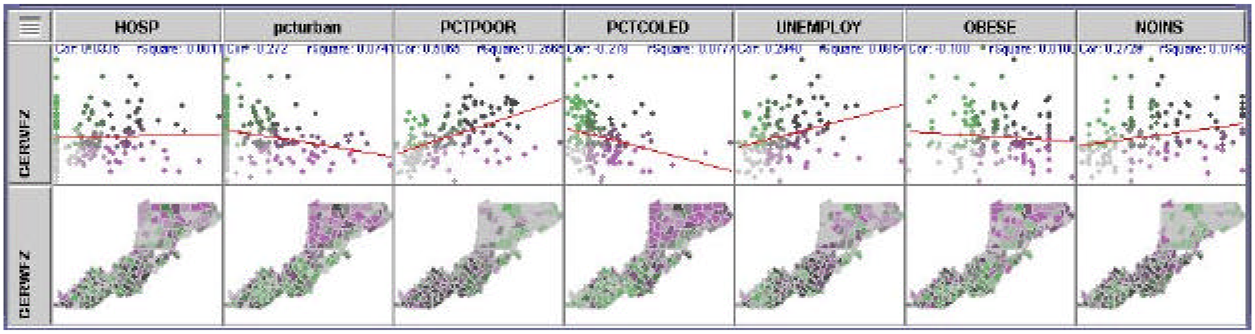


Figure **21**: The map and scatterplot matrix with the display of cervical cancer mortality rate for white females during the time period 1970-1994, and hospital to population rate, percentage of urban, percentage of population below poverty line, percentage of adults with college education, and unemployment rate, % of persons ages 18+ who are >120% of the median body mass index, and % of persons ages 18+ who do not have a health plan or health insurance.

4.3 Discussion

Exploration of the data, classification, and hypotheses surrounding cancer mortality rates and socio-economic status illustrate that there are visually significant spatial patterns in the mortality rates for both types of cancer, as well as correlations between cancer rates and socio-economic factors, in the study area. Cervical cancer mortality rate appears negatively correlated with economic status. Breast cancer mortality rate has little or no correlation with income when considering the entire study area as a whole.

The relationships between cancer mortality rates and socio-economic factors also display spatial patterns. Cervical cancer and breast cancer mortality rates have negative relationships with socio-economic status, especially per capita income, in counties in West Virginia and Kentucky. This factor confirms the suggested hypothesis that poverty, in general, has an adverse effect on cancer mortality rates. The cervical cancer mortality rate does not vary with increase of per capita income for counties with relatively high income. This phenomenon indicates that cervical cancer rate will not increase after the economic status achieves a certain level.

Breast cancer mortality rate is positively correlated with socio-economic factors, including per capita income and education, in Pennsylvania, where counties have relatively high economic status. This phenomenon indicates that different cancers can be influenced by, or related to, different risk factors in the study area, and the same cancer can also have different major risk factors or relationships with socio-economic status in different sub regions. For the counties in Pennsylvania, the relationship between breast cancer mortality rate and socio-economic status is different from the counties in the other sub regions of Appalachia. The biological factor, that women with higher socio-economic status tend to have fewer children and give birth to their first child at later ages increases the incidence of breast cancer, and overwhelms the increased risk due to poverty. Those women usually have a college education and high income. The relationships retain visually significance, though they are not proven statistically with strong correction coefficients and high r^2 values. Nevertheless, with more data over a longer time period, the positive correlation may be established as significant.

The different risk factors associated with different cancers within the study area could be overlooked if we view the study area as a single region, because the contradictory relationships in different subsets of counties are averaged, and the relationships, which indicate the local risk factors for cancer mortality rates, are lost in the analysis.

As a result of these data explorations, a new concept map showing relationships between cancer mortality rates and social economic status is suggested (Figure **22**). The initial concept map (Figure **1**), based on the hypothesis that high cancer mortality rates are associated with poor socio-economic status in the Appalachian region, is revised. It is true that the mortality rate for cervical cancer is negatively correlated with socio-economic variables, such as income and education, and positively correlated with percentage of population below the poverty line and unemployment rate The general hypothesis that cancer mortality rates show negative relationships to socio-economic status is questionable, and seems to be an oversimplification when considering the breast cancer case, since they show some positive correlation with socio-economic status. The risk factor of access to health service is also removed in the revised concept map, since the variables such as physician to population ratio, hospital to population ratio and insurance do not have obvious correlations with cancer mortality rates.

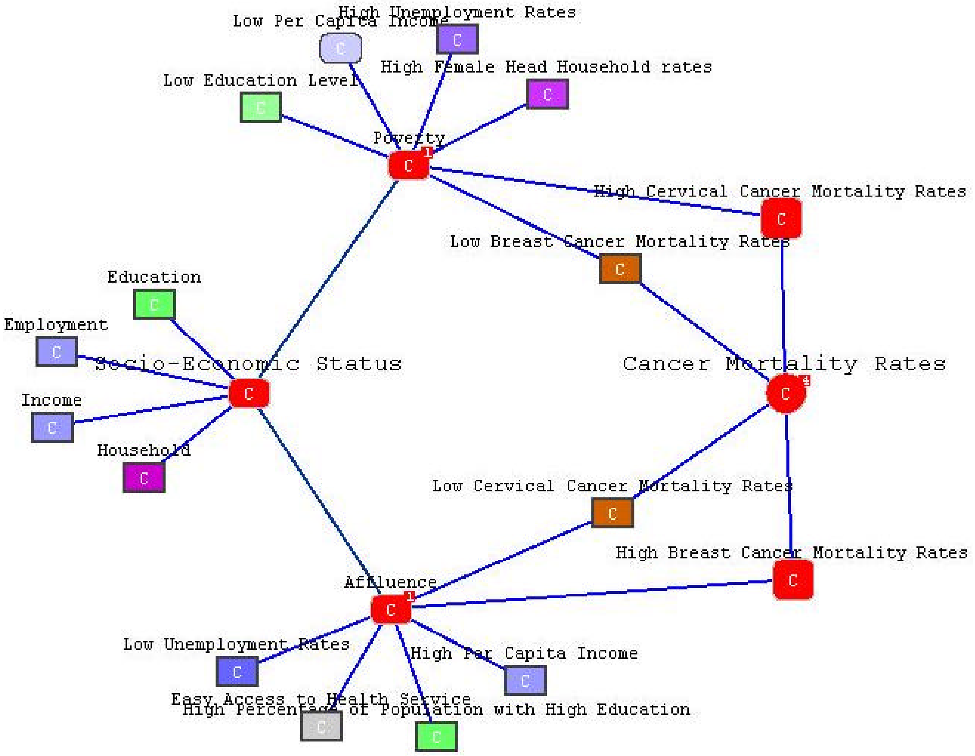

Low Per Capita Income
High Unemployment Rates
High Female Head Household rates
Low Education Level
Poverty
High Cervical Cancer Mortality Rates
Low Breast Cancer Mortality Rates
Education
Employment
Socio-Economic Status
Cancer Mortality Rates
Income
Household
Low Cervical Cancer Mortality Rates
High Breast Cancer Mortality Rates
Affluence
Low Unemployment Rates
High Per Capita Income
Easy Access to Health Service
High Percentage of Population with High Education

Figure **22**: New concepts generated according to the analysis in the integrated system.

5. Conclusion and Future Works

The case study demonstrates that the system can help to explore the relationships between cancers and their risk factors by providing interactive data visualization and classification methods for data and category exploration. The human conceptualization of the relationships between cancers and risk factors can then be revised via the progress of data exploration.

Selecting subsets of counties in attribute space (scatterplots, PCP) or geographical space (choropleth maps) is an important and useful function to explore counties within 'interesting' clusters. The linked brushing function can help to highlight the selected subsets of counties in other connected visualization tools, so that these subsets can be explored in multiple views. The operation of selection is so simple in the system that users only need to use the mouse to drag a box around the observations (counties), which they want to select. However, this simple method to sometimes cannot handle some more complex situations easily. If the observations (counties) to be select do not fall in a uniform rectangular area in some display (e.g. scatterplots or maps), selection can be more difficult. Multiple selections, by holding the "Shift" key while selecting counties, is provided in the tools, but it is a tedious procedure and requires a careful and patient approach. A function to select such groups of records, whether counties in a class or state, or some other units comprised of several composing members is desirable in future development.

Regression analysis is an effective method to explore the relationships between variables. In particular, the dynamic visualization of regression lines for the whole study area or the subset of data samples allows users to compare relationships between variables for the whole study area, with the variations among the sub regions. The regression analysis implemented in the research is linear, whereas several different distributional forms have been encountered in the analysis described here**.** For example, a bimodal distribution would be better represented by a nonlinear regression, such as the distribution for per capita income in Figure **9,** and there are clearly some variables that are gathered on the nominal and ordinal scales, for which non-parametric statistical summaries would be both more valid and more useful.

The visualization tools in the system support detections of outliers, and allow the exclusion of outliers in data analysis by linked brushing and change of data extension in scatterplots. However, the dataset underlying the visual analysis remains the same, because the data structure in the system does not easily support deletion on outliers from the dataset. Hence, the outliers still remain in the analysis unless they are deleted by some separate method outside of the system described. In addition, the data structure in the system cannot support the direct creation of new variables derived from existing variables, such as statistical z scores used for standardizing variables (though they can be calculated as required. A more flexible data structure is needed to support the above functions in the future.

In future work we will concentrate on improving the usability of the tools, making the connections between the concept map and the exploration components more naturally, and developing new visualization and classification tools as alternative to take advantages of different methods. The software to support this research is implemented as a series of visual and computational components that are all connected into a workflow design supporting direct interaction between components using GeoVISTA *Studio* (http://www.geovistastudio.psu.edu, Gahegan *et al*., 2002).

Acknowledgments

This work is partially supported by NSF Digital Government grant (#99883451), ITR Ontology in Action grant (#0219025) and CDC Association of Teachers of Preventive Medicine (ATPM), award number TS-1125.

References

Anselin, L., Kim, Y., and Syabri, I., 2004, Web-based analytical tools for the exploration of spatial data. *Journal of Geographic Systems*, **6**,197-218.

Carr, D., Littlefield, R., Nicholson, W., and Littlefield, J., 1987, Scatterplot matrix techniques for large N. *Journal of the American Stiatistical Association*, **82**(398), 424-436.

Carr, D., Wallin, J., and Carr, D., 2000, Two new templates for epidemiology applications: Linked micromap plots and conditioned choropleth maps. *Statistics in Medicine*, **19**, 2521-2538.

Carr, D., White, D., MacEachren, A., and MacPherson, D., 2005, Conditioned choropleth maps and hypothesis generation. *Annals of the Association of American Geographers*, **95**, 32-53.

DiBiase, D., 1990, Visualization in earth sciences. *Earth and Mineral Sciences, Bulletin of the College of Earth and Mineral Sciences, The Pennsulvania State University*, **59**(2), 13-18.

Dykes, J., 1997, Exploring spatial data representation with dynamic graphics. *Computers and Geosciences*, **23**(4), 345-370.

Edsall, R., 1999, Development of interactive tools for the exploration of large geographic database. In *Proceedings of the 19th International Cartographic Conference. Ed. D. P. Keller*. (Ottawa: Canadian Institute of Geomatics), pp. 14-20.

Edsall, R., 2003, Design and usability of an enhanced geographic information system for exploration of multivariate health statistics. *Professional Geographer*, **55**, 605-619.

Gahegan, M., Wachowicz, M., Harrower, M., and Rhyne, T., 2001, The integration of geographic visualization with knowledge discovery in databases and geocomputation. *Cartography and Geographic Information Science*, **28** (1), 29-44.

Gahegan, M., Takatsuka, M., Wheeler, M. and Hardisty, F., 2002, GeoVISTA *Studio*: a geocomputational workbench. *Computers, Environment and Urban Systems*, **26**, 267-292.

Lucieer, A. and Kraak, M., 2002, Interactive visualization of a fuzzy classification of remotely sensed imagery using dynamically linked views to explore uncertainty. In *Proceeding of Accuracy 2002 Symposium*, July 10-12, Melbourne, Australia, pp. 348-356.

MacEachren, A., Masters, R., Wachowicz, M., Edsall, R., and Haug, D., 1999, Constructing knowledge from multivariate spatiotemporal data: Integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Systems*, **13**(4), 311-334.

MacEachren, A., Hardisty, F., Dai, X., and Pickle, L., 2003, Supporting visual analysis of federal geospatial statistics. *Communications of the ACM*, **46**,59-60.

Monmonier, M., 1989, Geographic brushing: Enhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis*, **21**(1), 81-84.

Pickle, L., Mungiole, M., Jones, G., and White., A., 1999, Exploring spatial patterns of mortality: The new atlas of United States mortality. *Statistics in Medicine*, **18**, 3211-3220.

Rodriguez, M., and Egenhofer. M. J., 2003, Determining Semantic Similarity among Entity Classes from Different Ontologies. *IEEE Transactions on Knowledge and Data Engineering*, **15** (2), 442-456.

Snow, J., 1855. *On the mode of communication of cholera* (New York: The Commonwealth Fund).

Smith, B., and Mark, D., 1998, Ontology and geographic kinds. In Poiker, T.K., and Chrisman, N., (eds.), *Proceedings, 8th International Symposium on Spatial Data Handling*, pp. 308-320.