# Analyzing Statistical Relationships between Global Indicators through Visualization

Prabath Gunawardane*, Erin Middleton†, Suresh Lodha*, Ben Crow† and James Davis*
prabath@soe.ucsc.edu, emiddlet@ucsc.edu, lodha@soe.ucsc.edu, bencrow@ucsc.edu, davis@cs.ucsc.edu
*Department of Computer Science †Department of Sociology
University of California Santa Cruz

***Abstract*—There is a wealth of information collected about national level socio-economic indicators across all countries each year. These indicators are important in recognizing the level of development in certain aspects of a particular country, and are also essential in international policy making. However with past data spanning several decades and many hundreds of indicators evaluated, trying to get an intuitive sense of this data has in a way become more difficult. This is because simple indicator-wise visualization of data such as line/bar graphs or scatter plots does not do a very good job of analyzing the underlying associations or behavior. Therefore most of the socio-economic analysis regarding development tends to be focused on few main economic indicators. However, we believe that there are valuable insights to be gained from understanding how the multitude of social, economic, educational and health indicators relate to each other.**

**The focus of our work is to provide an integration of statistical analysis with visualization to gain new socio-economic insights and knowledge. We compute correlation and linear regression between indicators using time-series data. We cluster countries based on indicator trends and analyze the results of the clustering to identify similarities and anomalies. The results are shown on a correlation or regression grid and can be visualized on a world map using a flexible interactive visualization system.**

**This work provides a pathway to exploring deeper relationships between socio-economic indicators and countries in the hands of the user, and carries the potential for identifying important underpinnings of policy changes.**

## I. Introduction and Motivation

Visualization for the purpose of providing intuitive and deeper understanding of global inequality is an important problem. Several websites supporting these visualizations using raw data are becoming increasingly popular. However, seemingly easy to understand relationships between variables visualized using line graphs, bar graphs, and scatter plots, can sometimes provide incomplete information and may even lead to misleading or erroneous conclusions. In this work, we propose using statistical tools combined with visualization to provide a deeper and more complete understanding of relationships between the global socio-economic indicators.

There is a large amount of data collected across all countries annually over a range of socio-economic indicators by various agencies including World Bank [2], United Nations [26], UNESCO [7], and [18] . For example the World Development Indicators Database [2] has data that covers 225 countries and regions, spanning 40 years for more than 500 indicators. While having more information is definitely better, understanding and visualizing this data becomes a harder problem.

Many websites are utilizing this large collection of socio-economic indicator data to visualize global inequality. Popular websites include CISEIN [25], Gapminder [9], NationMaster [17], UC Atlas [24], and WorldMapper [5]. These websites utilize a number of classic visualization techniques including line graphs, bar graphs, scatter plots, and geographic maps to allow users to view this raw data in different ways. The temporal data is almost always visualized using animation. These visualizations take the first step to allow users to investigate a variety of questions: How does one country compare with other countries in the same geographic region or with similar GDP? How are different socio-economic indicators related to each other? What policies can be implemented to improve health nationally and globally? However, the simple indicator-wise visualization of data falls short of providing a deeper understanding of associations between various indicators and countries.

In this work, a team of computer scientists and sociologists have worked together to create a novel integration of statistical tools and visualization with a view to gain new socio-economic knowledge. Our goal is to leverage mostly the familiar and well-known statistical (correlation, linear regression, and clustering) and visualization techniques (scatter plots and geographic maps) to investigate deeper relationships between socio-economic indicators and countries. Is the intuitive understanding provided by raw indicator visualization supported by the results of correlation and linear regression analysis? Are the causality claims obtained through complex multi-regression models, often used in socio-economic literature, validated or contrasted by correlation or regression analysis? We view our system as a first step towards building a bridge between the simple approach of using a raw indicator visualization and the high-powered causality or other policy-based models.

Our system features an easy-to-use interface where the user can interactively select and visualize multiple countries and / or indicators. We have coupled it with the Globalization-Health Nexus Database [21] to analyze the relationship between various health indicators. Furthermore, we have contrasted the observations of both raw and statistical visualizations with the causal relationships between these health indicators obtained using a sophisticated econometric model by Cornia et al. [4] This integration allows us to get a much better and deeper understanding of the similarities, anomalies, and evolution of indicators and countries.

## II. Related Work

There has been considerable advances in visualizing geographic information data using a variety of novel techniques [13]. A majority of these techniques include using a combination of texture and color to create a palette that can be used to display multivariate data [14], [15], [16]. Due to challenges associated with understanding animated data, spatiotemporal geographic data has been visualized using wedges, circles, and rings [23] and mashups [27]. Distortions of geographic areas using rectangles, cartograms and a combination of cartograms with pixelmaps [19] have also been used to convey the values of socio-economic indicators. Additional efforts to visualize geographic data include geographically weighted scale varying visualization [6], diffusion-based density equalizing maps [10], and two-tone pseudo-coloring to visualize one-dimensional data [22]. [12] presents interactive feature section for identifying subspaces together with interactive hierachical clustering to assist visualization.

While many of these techniques appear promising and are very impressive from a visualization standpoint, most social scientists and users are unfamiliar with these techniques and remain wary of depending on these techniques to gain a better understanding of data.

Integrating a statistical model with visualization has been also explored in the literature. Carr et. al. presented a way to integrate statistical summaries with visualization by the use of linked micromap and conditioned choropleth maps for spatially indexed data[3]. The concept of using glyphs to visualize a correlation matrix has been explored in [8] . Andrienko et. al. use an iterative interactive approach to classify and identify patterns in spatial data, by using visualization and data mining [1]. Guo et. al. have presented an approach to cluster and sort large multivariate datasets based on self-organizing maps [11]. While these are general visualization toolkits, our application is more tailored towards the needs of our target audience, social scientists, and specifically intended to study country/indicator based patterns relative to each other.

In the integrated geographic statistical-visualization system that we have built, we are investigating relationships between causality, simple statistical relationships between indicators and countries, and intuitive understanding as obtained through simple visualization. We have chosen to use the causality model for global health indicators recently proposed by Cornia et al. [4], that we describe in further detail in Section IV-C. In order to contrast our results with those obtained by Cornia, we have integrated our system to draw data from the GHND database [21] that has been used by Cornia et al. in their study. We have also integrated other databases including the World Bank indicators. We have introduced a visualization of correlation and regression matrix (left diagram of Figure 2), that has been used mostly by computer scientists; However, the sorted correlation matrix (right diagram of Figure 2) evokes interest by social scientists, and the resulting mapping of one of the columns of the correlation matrix on to the geographic map (Figure 3) is of great interest to all. We have also developed a user interface that allows easy selection of indicators and countries from a variety of databases and visualizations to create customized visualizations (including zooming and data mining features that allow users to gain access to detailed underlying raw or computed data) that may be helpful in analyzing the data at hand.

Our main focus is to investigate whether the integrated statistical-visualization system can provide any new socio-economic knowledge or insights. We applied our system to investigate deeper questions regarding health variables. In Section V, we present three examples of the results of our investigation. Due to simple and familiar visualizations, social and computer scientists could share and understand the results equally well to create a meaningful dialogue. Many of these investigations validated the understanding obtained through simple means, but the system produced some new and surprising results and is also helpful in quantifying the intuitive understanding.

## III. Visualization

Global socio-economic indicators can be captured in a 3D volume as illustrated in Figure 1. Although one can attempt to view all the data in 3D, social scientists are much more accustomed to familiar 2D visualizations. In this work, we first describe the typical visualizations associated with the 1D and 2D of this 3D volume.

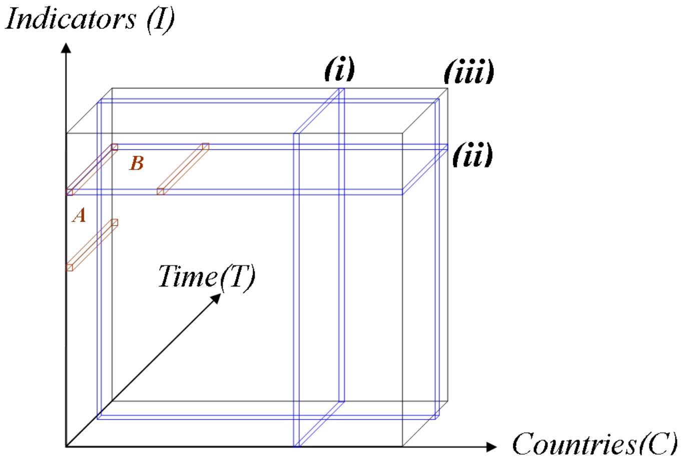

Fig. 1. *The 3-dimensional volume of indicator (I), country (C)and time (T) data, with (i) a vertical 2D slice highlighted which shows times series data for all indicators for a specific country, (ii) a horizontal 2D slice showing time series data for a single indicator over all countries, and (iii) a vertical 2D slice for all indicators for all countries at a specific time.*

The three 1D slices of the 3D volume of global socio-economic data are C-slice (one indicator, one time, all countries), T-slice (one indicator, once country, all time), and I-slice (one country, one time, all indicators). The C-slice is typically visualized by mapping the indicator values on a geographic map using pseudo-color and is perhaps one of the most popular geographic visualizations. The T-slice is commonly visualized as a time-series line graph. The I-slice is usually presented as a table.

In addition to these 1D slice visualizations, visualization of $1\frac{1}{2}$D data, that is 2 or more layers of these 1D slices, is very common. We will refer to two layers of C-slice as a 2C-slice. A 2C-slice may represent two indicators, one time, and all countries or one indicator, two times, and all countries. In the first case, the data is ideal for visualization on a geographic map using bivariate display techniques, although there is no one commonly accepted technique except perhaps for side-by-side display of two geographic maps. In the second case, although animation is commonly accepted, technique of small multiples is often employed in practice where two static images are displayed side-by-side. Gapminder [9] has developed a technique of visualizing a 4C-slice of 4 indicators using a scatter plot where 2 variables are mapped on the x-axis and y-axis and two additional variables are depicted through glyph size and glyph colors.

A 2T-slice may represent two indicators, all times, for one country or one indicator, all times, for two countries. This 2T slice is typically visualized using line graphs or bar graphs. An NT-slice can also be visualized in similar ways within the space constraints.

A 2I-slice may represent one time, two countries and all indicators or two times, one country and all indicators. The purpose of these 2 slices are very different. In the first case, the goal is to compare the two countries, while in the second case, the goal is to examine all the indicator trends for the same country. This data is typically presented in a tabular format or if a subset of indicators is chosen, then this subset can be visualized using classical visualization techniques including line or bar graphs.

### A. Slice Visualization

*1) TC Slice for an Indicator:* The Time-Country slice (Figure 1 (ii)) represents the data for a single indicator spanning all countries over a period of time. Most websites visualize this slice using an animation of a world map where countries are pseudo-colored based on indicator values for that point in time.

*2) TI Slice for a Country:* This slice (Figure 1 (i)) is useful in understanding the evolution of socio-economic trends within a country. The full 2D slice includes all indicators and is difficult to visualize. A useful task in this case would be to reduce the dimensionality of socio-economic indicator space by identifying a subset of key indicators for a chosen country. This would mean restrict oneself to a few rows of the TI slice, typically by choosing specific indicators. These indicator trends are then visualized with line graphs and bar graphs.

One can also investigate relationships between a pair of indicators for a specific country by employing a scatter plot by graphing one along the x-axis and the other one along the y-axis at different time periods. In this work, we have used this type of scatter plots (Figures 6, 7, and 8) for individual countries to support or contrast the findings based on statistical or regression analysis.

In addition, we have also supported multiple overlaid scatter plots (see Figure 9) where users are allowed to pick individual or some group of countries.

*3) IC Slice for a Time Period:* The Indicator-Time slice represents all the indicators across all the countries for a given year (Figure 1 (iii) ). We are not aware of any effective way of visualizing the whole 2D slice of this data. Again, it will be useful to reduce the dimensionality of indicator space.

As we will see soon, statistical tools allow us to quantify the relationship between two rows or columns of the given volume of data and visualize them providing us with better understanding of relationships between the indicators or the countries.

## IV. Statistical Tools

### A. Correlation

In this work, we utilize correlation in at least two ways – to compute correlation coefficient between two indicator trends for a given country (shown as A in Figure 1) and to compute correlation coefficient between two countries for a given indicator trend (shown as B in Figure 1). The first approach is useful if we want to analyze how two given socio-economic indicators have varied over time with relation to each other. We compute their correlation for each country, which gives us a single correlation value per country which is visualized on a geographic map. This analysis can be used to answer questions such as 'Is an increase in immunization always correlated with a decrease in infant mortality?'

In the second analysis, we compute the correlation trends (over a period of time) between countries for a single indicator. Since each country pair would have a correlation value, we visualize these results using a correlation matrix (see Figure 2). A $cell_{row,column}$ in the matrix represents the correlation between the indicator trend of a $country_{column}$ and the indicator trend of a $country_{row}$. A single column of this matrix corresponds to correlation of the indicator trends of a specific country with the indicator trend of all the other countries. This analysis can be used to cluster countries based on indicator trends. We can get a better understanding of how close those countries match up by sorting and visualizing the correlation matrix by that particular country (right diagram of Figure 2). This analysis can also be used to determine whether a particular indicator trend, for example, the increase in life expectancy, has been uniform in all parts of the world, and allows easy identification of anomalies (see Figure 3).

We use Pearsons product-movement correlation coefficient [20] as our correlation estimator. The correlation coefficient gives a measure of positive and negative linear correlation, ranging from +1 to -1.

### B. Regression

In addition to computing the correlation between two indicator trends for a country, we have also computed the linear regression fit for these indicator trends by taking one of the

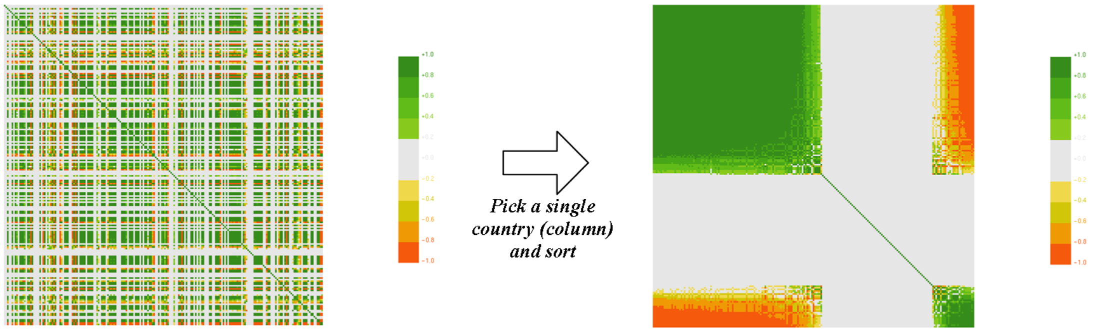

Fig. 2. *(left)Correlation matrix for LEB (Life expectancy at birth) for years 1980-2005 between 207 countries; (right) The same correlation matrix sorted by the column for Sweden, indicating correlation of other countries with Sweden for LEB from 1980-2005.*

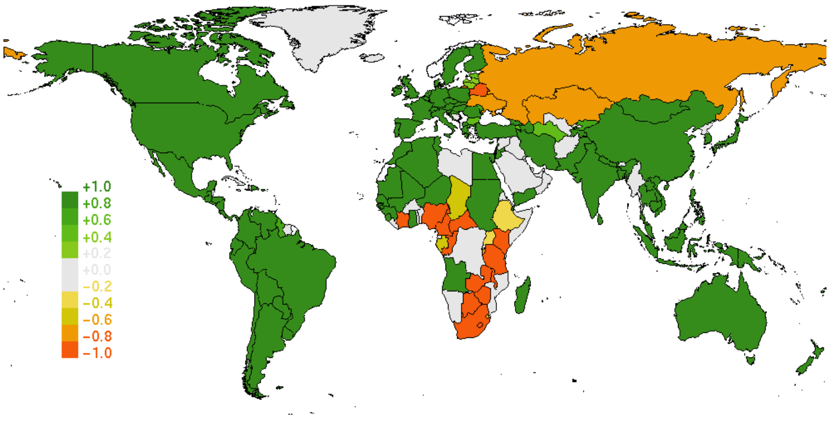

Fig. 3. *Correlation of other countries with Sweden for LEB from 1980-2005 shown on a world map.*

indicators to be the independent and the other the response variable.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where $y_i$ is the dependent or response variable, $x_i$ is the independent variable and $\varepsilon_i$ is the residual. One would expect highly correlated indicators to lead to a good linear regression fit and the regression coefficients $\beta_1$ and $\beta_0$ (which is also referred to as the intercept and the slope in the case of linear regression) can be used to understand the relationship between the two indicators. Together with the correlation visualization for the two selected indicators, we also provide a visualization of the regression coefficients on a geographic map.

### C. Causality

Recently, Cornia et al. [4] proposed a causality model for global health indicators investigating five different impact pathways for health. These pathways are material deprivation, progress in health technology, acute psychological stress, unhealthy lifestyle pathways, and socio-economic hierarchy-disintegration. Each of these pathways are measured by a cluster of socio-economic indicators that include income, income inequality, unemployment rate, inflation rate, illiteracy rate, health expenditure, number of physicians, alcohol consumption, smoking rates, unbalanced diet, migration rate, DPT immunization rate, wars, disasters, etc. Impact of these independent variables are studied on a cluster of health variables including u5MR (infant mortality under 5), IMR (infant mortality rate), and LEB (life expectancy at birth).

To improve the goodness of fit, improve the robustness of the estimates, and avoid multi-collinearity problems, some variables were dropped, normalized or modified. One such variable is log (physicians/1000 people) which was divided by log (GDP per capita) to obtain an index of availability of distribution of health personnel relative to the GDP/c norm.

The estimation was carried out for all the countries together, and also for four different groupings of countries – high income, middle income, low income, and transitional countries and for two different time periods, 1960-2005 and 1980-2005. Obtained results were examined for their statistical significance better than 1%, between 1 to 5%, between 10 to 15%, and not significant.

Results most relevant to our work include statistically significant dependence of u5MR on DPT immunization rate for all the countries as well as for all the four subgroupings of the countries mentioned above and the dependence of LEB on Log (Physicians/1000 people)/ Log(GDP/c). In this work, we chose

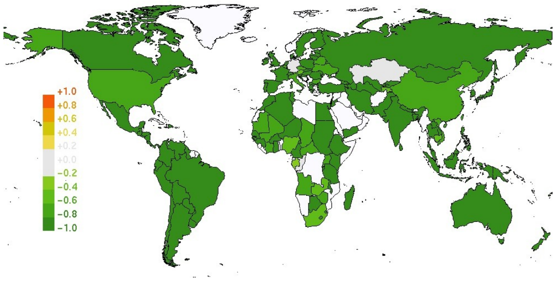

Fig. 4. *Correlation coefficients between U5MR (under 5 mortality rate) and DPT immunization rate for years 1960-2005. This world map depicts that U5MR is negatively correlated with DPT imm. for most countries as expected. Anamolous countries, such as Germany, Kazakhastan, and Congo are easily detected in this visualization*

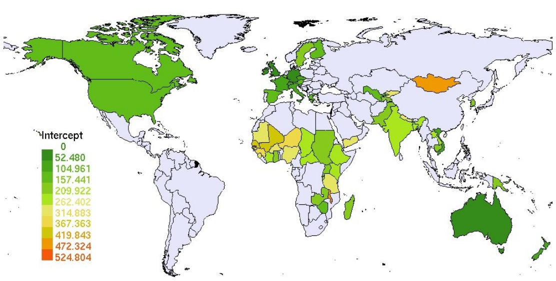

Fig. 5. *Under 5 mortality rate at zero intercept, after linear regression between U5MR and DPT immunization rate shown only for high and low income countries from 1960 - 2005. This map brings out relatively high u5MR for low income countries at comparable level of DPT immunization.*

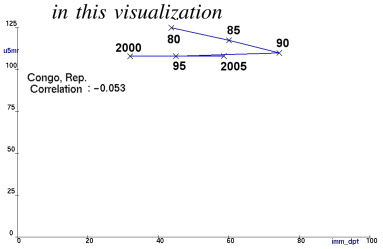

Fig. 6. *Scatter plot between U5MR vs DPT for Congo, from 1960 to 2000. Deviation from the norm is due to war.*

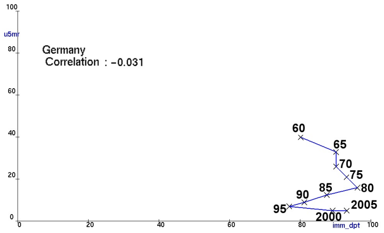

Fig. 7. *Similar scatter plot for Germany. Deviation from the norm is due to variation in health polivy during 1980-2005.*

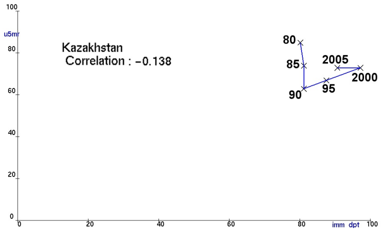

Fig. 8. *Similar scatter plot for Kazakhstan. Deviation from the norm is due to political changes.*

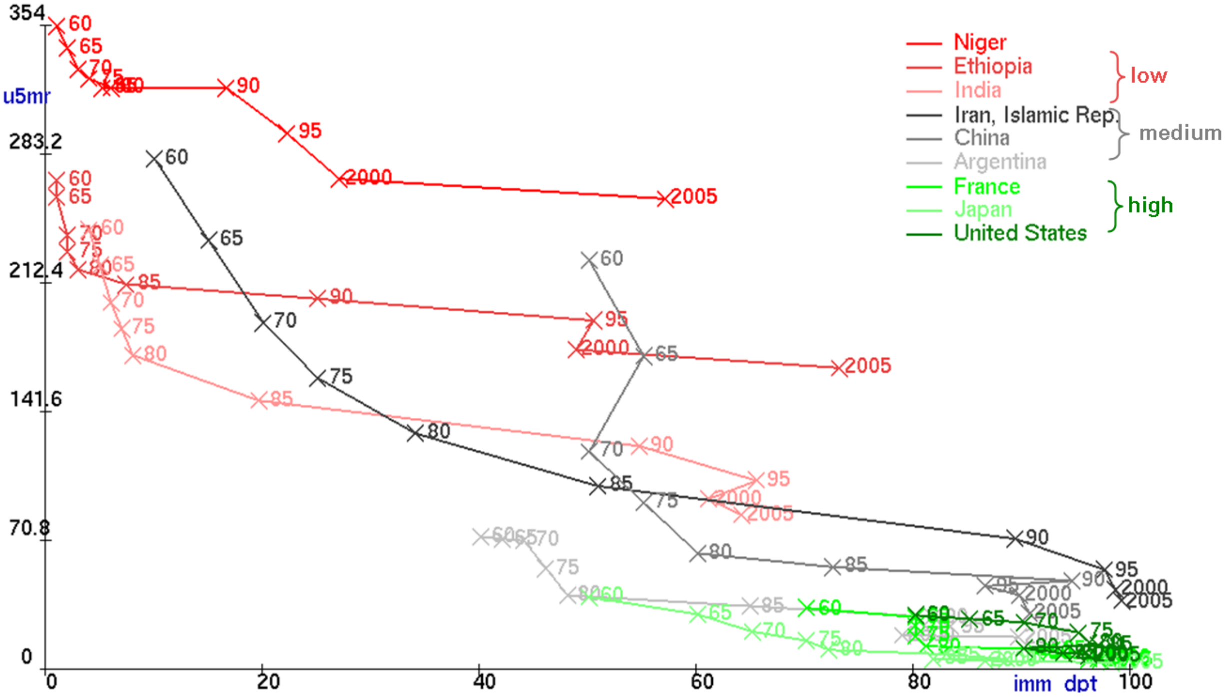

Fig. 9. *Scatter plot of u5MR vs. DPT showing 9 countries, 3 from each income group. These scatter plots reaffirm the general clustering og high, middle, and low income countries into three separate clusters, characterized by low, middle, and high u5MR at comparable DPT levels.*

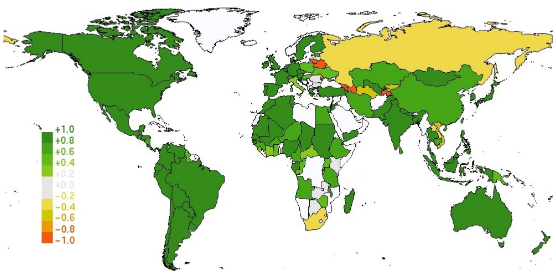

Fig. 10. *Correlation coefficients between LEB (Life expectancy at birth) and log(Physicians per 1000 people)/log(GDP per capita) for years 1960-2005. This correlation is positive for most countries including high income countries as expected.*

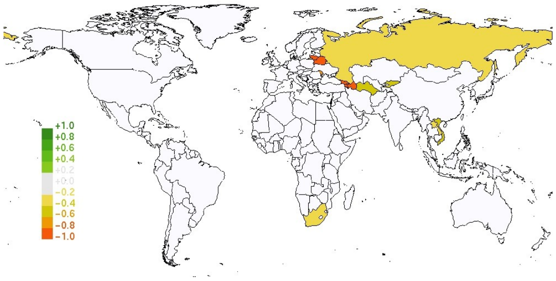

Fig. 11. *Countries with negative correlation coefficients for LEB (Life expectancy at birth) vs log(Physicians per 1000 people)/log(GDP per capita) for years 1960-2005. Most countries with negative correlation are erstwhile Russian block countries and a few African countries.*

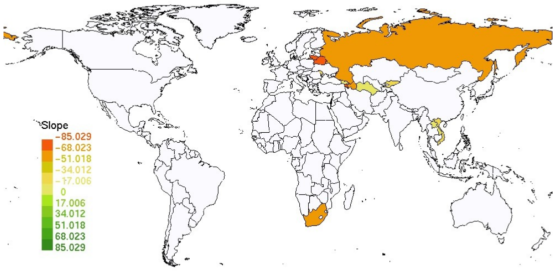

Fig. 12. *Countries with negative regression slopes for LEB (Life expectancy at birth) vs log(Physicians per 1000 people)/log(GDP per capita) for years 1960-2005. These are the same set of countries as the countries with negative correlation coefficients. Correlation and regression analysis agree with each other.*

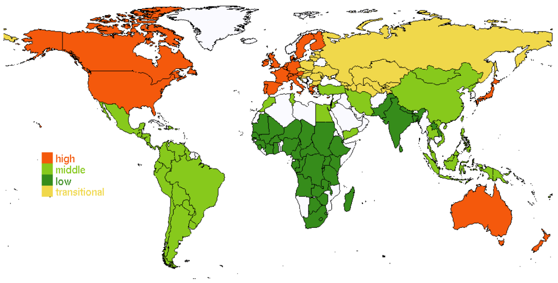

Fig. 13. *Causal coefficients between LEB and log(Physicians per 1000 people)/log(GDP per capita) obtained by Cornia et al. Causal relationship yields a surprising negative relationship for high income countries between the variables, which is counter-intuitive.*

to focus on these 4 variables – u5MR, DPT immunization rate, LEB, and Log (Physicians/1000 people)/ Log (GDP/c) together with GDP data for classification of countries.

### *D. Clustering*

Classification and clustering of countries and indicators based on similarity is a common and useful endeavor. Existing solutions deal mainly with classifying the countries based at a fixed point in time, an example would be the world bank classification of countries in to 'high','middle' and 'low' income groups. In our system, we can provide results of clustering, using K-means algorithm, on any socio-economic indicator such as life expectancy and immunization rates. Also we allow clustering over a period of time, based on the correlation results of indicator trends. For example, if we could cluster countries that had an life expectancy trend similar to Sweden for the period 1980-2000. This allows grouping of countries with similar characteristics over a period of time, as opposed to just a single year. In most of our examples discussed later, we have clustered countries into four categories, although our system allows choosing the number of clusters.

## V. Analysis and Visualization

We now present three examples to illustrate how the integration of visualization with statistical tools have provided us with valuable socio-economic insights. All our examples draw from the highly reliable GHND database of socio-economic indicators [21]. We have chosen to focus primarily on health indicators so that we can contrast or validate our results against those obtained by Cornia et al. [4], which was described previously. For this work, we have chosen a subset of these indicators and variables to illustrate the utility of our integrated visualization-statistical tool.

For health indicators, we have chosen u5MR (Infant mortality rate under 5) and LEB (Life expectancy at birth). For independent variables that impact health, we have chosen DPT imm (DPT immunization rate) and Log (Physicians/1000 people)/ Log (GDP per capita). In addition, we have also used GDP per capita. Most of the data is available for 137 countries for 204 indicators over the time period 1960 to 2005.

### *A. Statistical Visualization: Anomalies and Similarities*

In this example, we focus on validating how correlation analysis and visualization may be helpful in analyzing relationship between indicators. To this purpose, we chose to explore the relationship between u5MR and DPT imm for all the countries. Correlation between these two indicators are computed for all the countries individually for a time period of 1960-2005. This correlation coefficient is then visualized on the world map in Figure 4. This map clearly brings out that there is a strong negative correlation between the two variables, as expected, for most of the countries, with few exceptions. This figure validates the common working assumption that an increase in immunization reduces u5MR.

Anomalies in the relationship between u5MR and DPT imm is also brought out by Figure 4. These anomalies appear as negative or close to zero correlation for some countries. These countries include Congo, Germany, and Kazakhstan. Scatter plots of relationships between u5MR and DPT for these 3 countries are shown in Figures 6, 7 and 8 respectively. Reversal or decrease in DPT immunization rate in Congo from 1990 to 2005 is a result of war. Reversal of decrease in DPT imm rate between 1990 to 2000 in Germany is due to a variation in health policy that has been checked since 2000 resulting in continuance of the desirable trend. Finally, the increase in u5MR in Kazakhastan from 1990 to 2005 is due to political changes in the country. In summary, the correlation visualization on the world map quickly leads us to anomalies; supporting scatter plots quickly helps us in validating the anomalies and leads us to causes of these anomalies and points towards possible challenges or recommendations for changes in health policy.

We now examine the relationship between the same variables, u5MR and DPT, using linear regression between the two variables. After a linear fit, we compute the y-intercept, that is, level of u5MR at a hypothetical zero DPT level. These levels of u5MR are then visualized only for the high and low income countries (excluding the middle income countries) in Figure 5. This visualization brings out the sharp contrast between the two groups of countries.

This observation is further validated by picking 3 sample countries from each of the three groups – low, middle, and high income – and then visualizing the relationship between u5MR and DPT on a scatter plot in Figure 9. This supporting visualization using raw numbers further validates the observation that the low income countries are typically clustered towards the high range of u5MR and also saturate at higher levels of u5MR than the middle or high income countries. This observation leads to the conclusion that DPT can help reduce u5MR only up to a certain point in low and middle income countries and additional health measures need to be undertaken to reduce u5MR further. Although this observation may seem obvious after these visualizations, the causality model described by Cornia et al. [4] focus mostly on the regression slope and not making any of the observations listed above since their multi-variable regression model does not accommodate the simple intercept view of linear regression. Nevertheless, it is to be noted that most users, when browsing raw data using popular websites such as Gapminder and UC Atlas are intuitively looking for simple relationships between variables and the closest statistical analogs are typically correlation and regression analysis. In the examples discussed so far, simple visualizations including scatter plot, correlation, and regression visualization go a long way to provide valuable information regarding the relationship between these variables.

### B. Correlation, Regression, and Causality

We now present a second example of relationship between LEB (life expectancy at birth per 1000 children) and Log (Physicians per 1000 people)/ Log (GDP per Capita) over the period 1960-2005.

We first discuss the derivations of the causality model regarding the relationship between these variables. Cornia et al. [4] derive that the regression coefficient between these two variables for middle, low, and transitional (Eastern block countries) are 11.2796, 14.2350, and 8.6528, being significant at 1% level for middle income countries and being significant between 1% to 5% level for low income and transitional countries. The relationship between these variables is also significant at 1% level for all the countries together with even higher regression coefficient of 36.89. Surprisingly, the regression coefficient between these two variables is *negative*, -28.9, also significant at 1% level. These regression coefficients are visualized in Figure 13, where the negative regression coefficient is mapped to the red color, while the other three coefficients are mapped to yellow, light green, and dark green in increasing magnitude of the regression coefficient. These causality results are in contrast with the correlation coefficients visualized for all the countries in Figure 10. The dark green colors in Figure 10 illustrate that the relationship between the two variables are positive, as expected, that is increasing the number of physicians (compared to GDP per capita) 'results' in an increase in LEB. While the causality model in Figure 13 points to a hypothesis that in a multi-regression model, the overall increase in LEB attributed to other factors such as Log GDP/volatility, female education, alcohol consumption, and smoking, etc. is in fact offset by physicians to bring the model in line with the rest of the countries. This example illustrated the utility of statistical visualization in bringing deeper understanding and checks against the more sophisticated but harder to understand multi-regression causality models.

Correlation and regression computations and visualizations bring further insight into the relationship between the two variables. Figures 11 and 12 show the countries with negative correlation coefficients and those with negative regression coefficients respectively. In this case, we observe that correlation and linear regression results agree with each other strongly. Furthermore, the negative relationships between the two variables are present predominantly for transitional (Eastern block countries). This is, again, a surprising result, since the causality model by Cornia et al. [4] computed a positive regression coefficient for these countries with high statistical significance. These visualization based observations lead us to believe that the relationship between these two variables is more complicated than a simple causal one and requires further investigation.

### C. Clustering

We now present our third and final example, using clustering, to illustrate the utility of integrating statistical tools with visualization. To this purpose, we classified countries into four categories using many different indicators. Figure 14 presents the visualization of countries classified into low (red), lower middle (yellow), higher middle (light green), and high income (dark green) countries based on GDP in the

| | GDP | GDP(T) | u5MR | u5MR(T) | LEB | LEB(T) | DPT | DPT(T) | Phy | Phy(T) |
|---|---|---|---|---|---|---|---|---|---|---|
| Brazil | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 4 | 1 | 1 |
| Russia | 2 | 3 | 1 | 2 | 2 | 4 | 1 | 4 | 1 | 1 |
| India | 1 | 1 | 2 | 1 | 3 | 1 | 3 | 3 | 2 | 1 |
| China | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 3 | 1 | 1 |

TABLE I
THE RESULTS OF USING KMEANS TO CLUSTER COUNTRIES FOR A SPECIFIC YEAR (2000) AND ALSO BASED ON TRENDS FOR A SPAN OF YEARS (1980-2005). CLUSTERS 1 THROUGH 4 REPRESENTS THE 'BEST' TO 'WORST' CLASSIFICATIONS RESPECTIVELY.

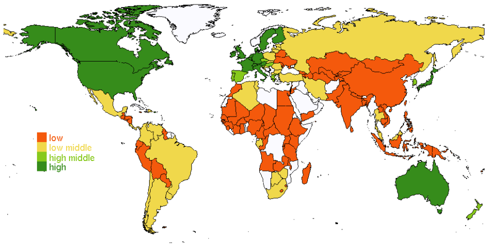

Fig. 14. *Clustering of countries based on GDP in year 2000*

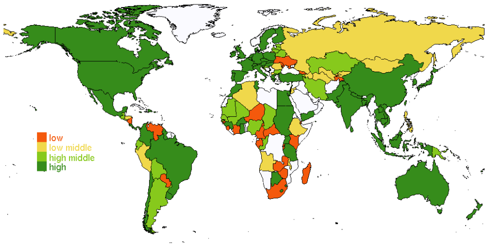

Fig. 15. *Clustering based on GDP trends between Sweden and other countries for years 1980-2005.*

year 2000. We will follow the results for the emerging BRIC (Brazil, Russia, India, and China) countries. In this case, India and China are still low income countries while Brazil and Russia are lower middle income countries. However, when we view the classification of GDP trends for the four BRIC countries over the period 1980-2005 in Figure 15, we observe that India, China, and Brazil are classified into the same (and strongest) category as most of the developed nations including USA, and most European countries, while Russian GDP growth is the next lower category. For the purposes of the trend classification, we computed the correlation between GDP trends between all the countries and Sweden. We chose Sweden because it was consistently in the top for most of the indicators that are we investigating in this study, including GDP, LEB, etc.

Figures 20 to 23 classify all the countries on the four variables – u5MR, DPT imm, LEB, and Log (Physicians per 1000 people)/ Log (GDP per Capita) using the static data from the year 2000. Figures 16 to 19 classify the countries based on the trends of these four variables over the period 1980-2005 as compared the trend of Sweden. Results of comparison of BRIC countries are presented in a table.

The table along with the visualization show that India is slightly behind other BRIC countries in u5MR while Russia is slightly behind in the u5MR trend. In LEB, again, India is slightly behind other BRIC countries while Russia is slightly behind in the LEB trend. Put together, in the two health indicator trends, u5MR and LEB, India lags behind other BRIC countries, but can catch up if it maintains its trend, while Russia is at the greatest risk of falling behind in the health indicators.

In terms of action or independent variables that impact health, DPT immunization rates for all BRIC countries are in the lowest two categories, Brazil and Russia being the lowest, and India and China next to the lowest. However, with respect to the DPT trend, Brazil, Russia, and China are in the strongest category (green), while India is somewhat lagging behind. In terms of the presence of physicians trends, all four BRIC countries are in the strongest category, while India slightly behind in the year 2000. Put together, all BRIC countries are likely to improve their health indicators by increasing the DPT immunization rate and need to maintain their strong growth trend regarding physicians.

## VI. Conclusions and Future Work

In this work, we proposed an integration of statistical computing with visualization to glean deeper understanding of global socio-economic indicators. We utilized correlation and linear regression to quantify relationships between pairs of variables and between pairs of countries. We utilized these tools to investigate static data for a fixed time period as well as dynamic trends over a large time period. Current state-of-the-art global inequality websites provide visualization support using raw data without the use of any statistical tools. Using three different examples, we demonstrate that correlation, linear regression, and causality models can bring out similarities and anomalies and provide better understanding of relationships between the variables by validating our intuitions based purely on raw data visualization or sometimes yields insights that are counter-intuitive or surprising. These observations or conclusions carry important implications in policy making both at national and global level.

This research has opened up several new exciting opportunities. Which countries can be grouped together? Based on which indicators? Which socio-economic indicators can be clustered together? Can we reduce the dimensionality of indicators so that a profile of a country is captured by some principal socio-economic indicators? What lessons can a nation learn from a similar group of nations? Ideally, we would like to build a system so that the empowered users can explore relationships between countries and between variables using appropriate statistical tools combined with visualization. We believe that this exploration can always be used to validate or contrast the proposed policy decisions and may also lead to important underpinnings of national or global policy decisions that are not immediately obvious.

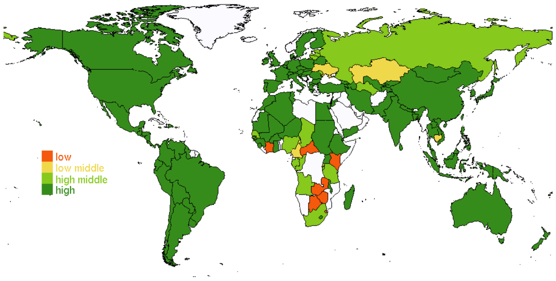

Fig. 16. *Clustering based on U5MR trends between reference country (Sweden) and other countries for years 1980-2005.*

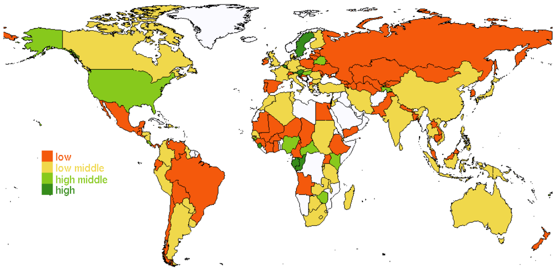

Fig. 17. *Clustering based on DPT immunization trends between reference country (Sweden) and other countries for years 1980-2005.*

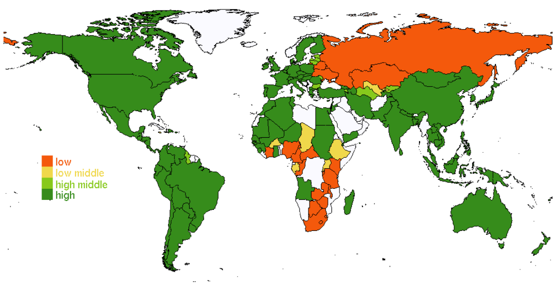

Fig. 18. *Clustering based on LEB trends between reference country (Sweden) and other countries for years 1980-2005.*

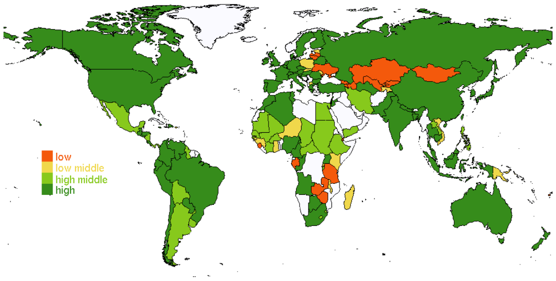

Fig. 19. *Clustering based on log(physicians per 1000 people) over log(GDP per capita) trends between reference country (Sweden) and other countries for years 1980-2005.*

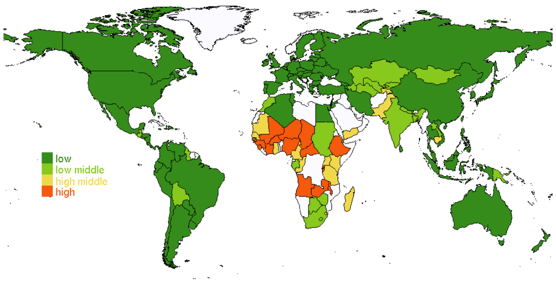

Fig. 20. *Clustering of countries based on U5MR in year 2000*

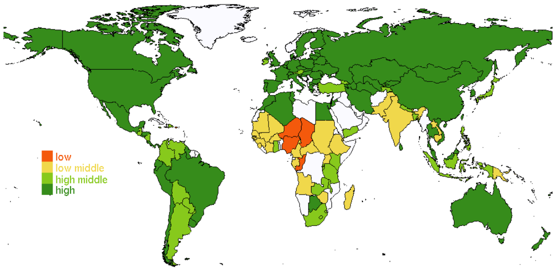

Fig. 21. *Clustering of countries based on DPT immunization rates in year 2000*

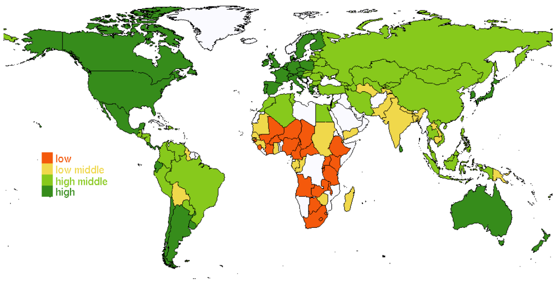

Fig. 22. *Clustering of countries based on LEB in year 2000*

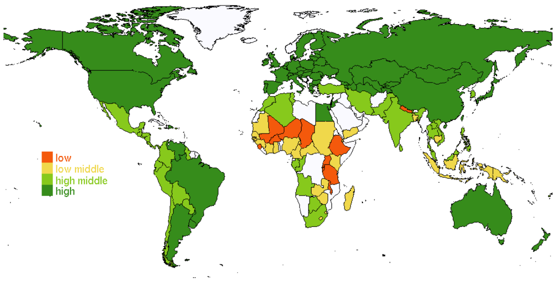

Fig. 23. *Clustering of countries based on log(physicians per 1000 people) over log(GDP per capita) for year 2000*

## VII. Acknowledgements

We would like to thank Brian Fulfrost for providing us with valuable feedback at various stages of this project.

## References

[1] G. L. Andrienko and N. V. Andrienko. Data mining with C4.5 and interactive cartographic visualization. user interfaces to data intensive systems. *G. T. Los Alamitos, CA, IEEE Computer Society*, pages 162–165, 1999.

[2] W. Bank. World development indicators. Website, 2008. http://www.worldbank.org/data/.

[3] D. B. Carr, J. Chen, B. S. Bell, L. Pickle, and Y. Zhang. Interactive linked micromap plots and dynamically conditioned choropleth maps. In *Proceedings of the 2002 Annual National Conference on Digital Government Research*, pages 1–7. Digital Government Society of North America, 2002.

[4] G. A. Cornia, S. Rosignoli, and L. Tiberti. Globalisation and health: impact pathways and recent evidence. In *Proceedings of Conference on Mapping Global Inequality*, 2007.

[5] D. Dorling, A. Barford, and M. Newman. Worldmapper: The world as you've never seen it before. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):757–764, 2006.

[6] J. Dykes and C. Brunsdon. Geographically weighted visualization: Interactive graphics for scale varying exploratory analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1161–1168, 2007.

[7] U. I. for Statistics. Global statistics. Website, 2008. http://www.uis.unesco.org.

[8] M. Friendly. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56:316–324, November 2002.

[9] Gapminder. Gapminder world 2006. Website, March 2008. http://tools.google.com/gapminder.

[10] M. T. Gastner and M. E. J. Newman. Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences*, 101(20):7499–7504, 2004.

[11] D. Guo. Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization*, 2(4):232–246, 2003.

[12] D. Guo, J. Chen, A. M. MacEachren, and K. Liao. A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1461–1474, 2006.

[13] D. Guo, M. Gahegan, A. M. MacEachren, and B. Zhou. Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach. *Cartography and Geographic Information Science*, 32(2):113–133, 2005.

[14] H. Hagh-Shenas, S. Kim, V. Interrante, and C. Healey. Weaving versus blending: a quantitative assessment of the information carrying capacities of two alternative methods for conveying multivariate data with color. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1270–1277, 2007.

[15] C. G. Healey and J. T. Enns. Large datasets at a glance: Combining textures and colors in scientific visualization. *IEEE Transactions on Visualization and Computer Graphics*, 5(2):145–167, 1999.

[16] V. Interrante. Harnessing natural textures for multivariate visualization. *IEEE Computer Graphics and Applications*, pages 6–11, November-December 2000.

[17] Nationmaster. Nations of the world. Website, 2008. http://www.nationmaster.com/.

[18] OECD. Organisation for economic co-operation and development. Website, 2008. http://www.oecd.org.

[19] M.-C. Panse, M.-M. Sips, M.-D. Keim, and S. M.-S. North. Visualization of geo-spatial point sets via global shape transformation and local pixel placement. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):749–756, 2006.

[20] J. L. Rodgers and A. W. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

[21] S. Rosignoli, L. Tiberti, and G. A. Cornia. The globalization-health nexus database (ghnd). Website, February 2007. http://www.unifi.it/dpssec/sviluppo/database.html.

[22] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya, and T. Kaseda. Two-tone pseudo coloring: Compact visualization for one-dimensional data. *Proceedings of Information Visualization*, 0, 2005.

[23] P. Shanbhag, P. Rheingans, and M. desJardins. Temporal visualization of planning polygons for efficient partitioning of geo-spatial data. *IEEE Symposium on Information Visualization*, 2005.

[24] UCSC. UC atlas. Website, 2008. http://ucatlas.ucsc.edu/.

[25] C. University and W. Bank. Global poverty mapping project. Website, 2008. http://www.cisein.org/povmap/atlas.html.

[26] UNSCB. United nations common database. Website, 2008. http://unstats.un.org/unsd/cdb/cdb_help/cdb_quick_start.asp.

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