Journal of International Development
J. Int. Dev. **21**, 1152–1166 (2009)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/jid.1652

UNDERSTANDING RELATIONSHIPS BETWEEN GLOBAL HEALTH INDICATORS VIA VISUALISATION AND STATISTICAL ANALYSIS

SURESH LODHA[1]*,†, PRABATH GUNAWARDANE[1‡], ERIN MIDDLETON[2§] and BEN CROW[2†]
[1]*Department of Computer Science, University of California, Santa Cruz, CA, USA*
[2]*Department of Sociology, University of California, Santa Cruz, CA, USA*

Abstract: Several agencies such as World Bank, United Nations and UNESCO are disseminating a large amount of socio-economic data at national level. Various websites such as UC Atlas, Gapminder, CIESIN and NationMaster are attempting to provide general users visualisation tools to display this data. Typical visualisation methods include line graphs, bar graphs, scatter plots, colour-coded glyphs (such as circles) and world maps. In addition to the general public, there is great interest in educational, research and public policy institutes to try to understand the relationships between these socio-economic indicators. In this paper, we juxtapose two techniques to investigate the relationships between global health indicators. The first approach employs sophisticated statistical techniques to develop a causality model between various global health indicators. The second approach, typically employed by the visualisation users of the various websites mentioned above, is to utilise a bivariate display between the health indicators in order to discover relationships between these variables. This visualisation approach is perhaps closest to a bivariate regression or correlation. Therefore, we employ these simple statistical techniques and associated visualisations as well. In this work, we analyse the two approaches using two specific examples related to health indicators. We find that the two approaches sometimes agree strengthening the conclusions or may provide different perspectives that require more careful analysis of the conclusions and need for further research. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: global health; statistical analysis; visualisation; indicators; regression; correlation

*Correspondence to: Suresh Lodha, Mail Stop: SOE 3, 1156 High Street, University of California, Santa Cruz, CA 95064, USA. E-mail: lodha@soe.ucsc.edu
†Professor.
‡Graduate Student.
§Research Assistant.

1 INTRODUCTION AND MOTIVATION

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 (2008), United Nations (UNSCB, 2008), UNESCO (2008) and OECD (2008). For example the World Development Indicators Database (World Bank, 2008) has data that cover 225 countries and regions, spanning 40 years for more than 500 indicators. While having more information is definitely better, understanding and visualising this data becomes a harder problem.

Many websites are utilising this large collection of socio-economic indicator data to visualise global inequality. Popular websites include CIESIN (2008), Gapminder (2008), NationMaster (2008), UC Atlas (2008) and WorldMapper (Dorling *et al*., 2006). These websites utilise a number of classic visualisation 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 visualised using animation. These visualisations 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 health indicators related to each other? What policies can be implemented to improve health nationally and globally?

Driven by these questions, various users including general public, students, researchers and public policy analysts employ these websites typically for bivariate data display, using two indicators at time. These visual approaches can be construed as attempts to decipher patterns that can perhaps be most closely related to bivariate regression and correlation analysis. However, the simple indicator-wise visualisation or bivariate statistical analysis of data may fall short of providing a deeper understanding of associations between various indicators and countries.

Therefore, we contrast this simple popular easy-to-use visualisation based bivariate approach with the sophisticated causality model to explain relationship between global health indicators, as proposed by Cornia *et al.* (2007). This approach investigates five different impact pathways for health—material deprivation, progress in health technology, acute psychological stress, unhealthy lifestyle pathways and socio-economic hierarchy-disintegration—utilising a large suite of socio-economic indicators and makes interesting observations and compelling conclusions about these indicators.

In this work, a team of computer scientists and sociologists have worked together to create a novel integration of statistical tools and visualisation with a view to gain new socio-economic knowledge. Our goal is to leverage the simple visualisation techniques (scatter plots and geographic maps) with the familiar and well-known statistical techniques (correlation and linear regression) in an attempt to understand relationships between socio-economic indicators and how different countries are situated with respect to each other regarding these socio-economic relationships or trends. Is the intuitive understanding provided by raw indicator visualisation supported by the results of correlation and linear regression analysis? How do the causality claims obtained through complex multi-regression models, often used in socio-economic literature, compare with correlation or simple regression analysis visually available through popular visualisation web sites? We view our system as a first step towards building a bridge between the simple approach of using a raw indicator visualisation and the high-powered causality or other policy-based models.

DOI: 10.1002/jid

2 RELATED WORK ON VISUALISATION

We are aware of a wider set of criticisms of the visualisation of social data. We do not address them in this paper. They constitute three goals for future work. First, aggregate national data are, of course, deeply problematic. Most egregiously, they conceal within-country inequalities. So, the goal here would be to find ways of representing differences within countries, including those of class, space and gender, while retaining a comparative global focus. More generally, ways of representing household and individual data are desirable.

Second, there is a critique of what data are gathered by national governments and international institutions. Crow (2006) suggests that governments prioritise economic activity, particularly that which generates revenue, and corporate facts in their data collection, while being slow to measure women's work, peasant production and awkward facts—including poverty and inequality. This second criticism suggests the need for data collection institutions which are independent of government.

Third, a more wide-ranging critique of available data is opened up by Amartya Sen's work on desired functionings, capabilities and freedoms (Sen, 1999), and a new literature on well being (Gough and MacGregor, 2007; Gough *et al*., 2006). This literature explores new ways of understanding human needs, agency and resources, and the opportunities and processes available to individuals. This third criticism suggests the need for visualisation techniques which go beyond measurement of easily-quantified ideas to more sophisticated concepts such as functionings, freedoms, opportunities, processes and subjective well-being.

There is a long tradition of using visualisation as a tool to investigate discrete data sets within the computer science discipline. In particular, there have been considerable advances in visualising geographic information data using a variety of novel techniques (Guo *et al*., 2005). A majority of these techniques include using a combination of texture and colour to create a palette that can be used to display multivariate data (Healey and Enns, 1999; Interrante, 2000; Hagh-Shenas *et al*., 2007). Due to challenges associated with understanding animated data, spatiotemporal geographic data has been visualised using wedges, circles and rings (Shanbhag *et al*., 2005) and mashups (Wood *et al*., 2007). Distortions of geographic areas using rectangles, cartograms and a combination of cartograms with pixelmaps (Panse *et al*., 2006) have also been used to convey the values of socio-economic indicators. Additional efforts to visualise geographic data include geographically weighted scale varying visualisation (Dykes and Brunsdon, 2007), diffusion-based density equalising maps (Gastner and Newman, 2004), and two-tone pseudo-colouring to visualise one-dimensional data (Saito *et al*., 2005). Interactive feature selection for identifying subspaces together with hierarchical clustering to assist visualisation has also been proposed (Guo *et al*., 2006).

Although these sophisticated visualisation techniques have been well received within the computer science community, their impact within the social science community is still unknown. This gap arises because although the proposed visualisation techniques appear promising and impressive from a visualisation standpoint, these general techniques must be tailored or adapted to the needs of the users of a specific application.

Integrating a statistical model with visualisation has also been explored in the literature. Carr *et al.* (2002) presented a way to integrate statistical summaries with visualisation by the use of linked micromap and conditioned chloropleth maps for spatially indexed data. The concept of using glyphs (or symbols) to visualise a correlation matrix has been explored in (Friendly, 2002). Andrienko and Andrienko (1999) use an iterative interactive

DOI: 10.1002/jid

approach to classify and identify patterns in spatial data, by using visualisation and data mining. Guo (2003) has presented an approach to cluster and sort large multivariate datasets based on self-organising maps. These general visualisation schemas have not made their way into the popular visualisation web sites, mentioned earlier, because their sophisticated nature, not only requires steep learning but also their utility in comparison to simpler but well-developed visualisation techniques remains in question.

Our application is a first step to tailor the visualisation towards the needs of our target audience—general public, students, social scientists and researchers, and specifically intended for the application of investigating country/indicator-based patterns relative to each other. This audience is turning towards previously mentioned visualisation web sites such as UC Atlas, Gapminder, CISEIN, NationMaster and WorldMapper in the hope of discovering patterns and relationships between different indicators and countries. However, most of these websites provide visualisation of raw indicators, without much underlying statistical analysis. We are leveraging and enhancing simple but powerful visualisation techniques and strengthening them with underlying statistical analysis to provide a more detailed understanding of the relationships between global socio-economic indicators.

3 STATISTICAL TOOLS AND ANALYSIS

We first present a brief description and outline of the causality model for global health indicators, recently presented by Cornia *et al.* (2007). We then describe the bivariate correlation and regression analysis that we coupled with the familiar and easy-to-understand visualisation techniques that can be easily integrated with popular web visualisation sites for socio-economic indicators.

The two approaches are different. Cornia *et al.* use panel data and take into account both time and space in their multi-regression model. In contrast, correlation and bivariate regression coefficients use 2D data as the first step, for example, computing correlations between two variables along time for one country, and then extending this analysis to every country. The difference between these two approaches by taking different slices of the three-dimensional data—space, time and indicator—is described below and is further explored in greater detail by the authors elsewhere (Gunawardane *et al.*, 2009).

There is a deeper difference between the two approaches. As is well known, bivariate correlation coefficients are useful to measure the link between two quantitative variables, without regards to the true cause–effect relation between them. If for instance, we compute the correlation coefficient between car accidents and GDP/c in the last 30 years for a developed country, it will probably be high and positive; however, it is not a good idea to think that an increase in car accidents causes an increase in GDP/c. Nevertheless, if a cause–effect relationship is assumed or proposed, it makes sense to analyse the impact of the explanatory variable (for example, DPT immunisation) on the independent one (for example, under 5 mortality rate) through various means including correlation and simple bivariate regression, as is often done visually at popular web sites.

We also note that both approaches have their own challenges, for example, simple correlation coefficient as a measure of causal relationship may generate biased estimates while the sign and the strength of an explanatory variable in a multiple regression model could change after including or dropping an additional explanatory variable.

DOI: 10.1002/jid

3.1 Causality

Recently, Cornia *et al.* (2007) 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 immunisation rate, wars, disasters, etc. Impact of these independent variables is studied on a cluster of health variables including u5MR (infant mortality under 5), IMR (infant mortality rate) and LEB (life expectancy at birth).

The authors examined various estimation procedures including ordinary least squares (OLS) to quantify the impact of independent variables on the health indicators using the data from a reliable database GHND (Rosignoli *et al*., 2007), which is a tri-dimensional matrix of data for 137 countries, 234 indicators, and annual data over a time period of 1960–2005. However, OLS procedure provides inefficient estimates and distorts the values of coefficients since the information on the countries' fixed effects would be neglected. Therefore, the authors use a fixed effects model for estimation and used Hausman test to confirm that this estimation procedure is indeed appropriate for the given data.

To improve the goodness of fit, improve the robustness of the estimates, and avoid multi-collinearity problems, some variables were dropped, normalised 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 health personnel relative to the GDP/c.

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 and 5%, between 10 and 15% and not significant.

Results relevant to our work include statistically significant dependence of u5MR on DPT immunisation rate for all the countries and the dependence of LEB on log (physicians/1000 people)/log (GDP/c). In this work, we chose to focus on these four variables—u5MR, DPT immunisation rate, LEB and log (physicians/1000 people)/log (GDP/c). Most of these data are available for 137 countries over the time period 1960–2005. However, note that roughly 6–7% data were added to the GHND database on the basis of interpolation or on the basis of other information at the disposal by the authors of the GHND database, except for the DPT coverage which required addition of 44% of missing data points to stabilise low levels of data available between 1960 and 1980.

3.2 Correlation

Correlation between two variables is a well-known statistical tool and measures the strength of relationship between two variables. This is perhaps the easiest attribute that stands out in a bivariate visualisation of two variables using a line graph or a scatter plot. In this work, we have primarily utilised correlation to compute correlation coefficient between two indicator trends for a given country.

We use Pearson product-moment as our correlation estimator. This is defined as, $r_{x,y} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{n\Sigma x_i^2 - (\Sigma x_i)^2}\sqrt{n\Sigma y_i^2 - (\Sigma y_i)^2}}$, where x_i and y_i are two data series of n elements with

DOI: 10.1002/jid

standard deviation of s_x and s_y. The correlation coefficient gives a measure of positive and negative linear correlation, ranging from $+1$ to -1. When we are comparing data from different indicators, it is also important that we normalise them before computing the correlation. This process simply consists of a mean shift and a division by the standard deviation. This removes the scale disparities in the data.

As a first step, correlation analysis is useful if we want to analyse the relationship between two indicators over time for the same country. This analysis can be used to answer questions such as 'Is an increase in immunisation correlated with a decrease in under 5 mortality rate for this country'? In the next step, we compute the correlation between the same two indicators for each country, which can then be visualised on a geographic map. This analysis and visualisation can be used to answer questions such as 'Is an increase in immunisation correlated with a decrease in under 5 mortality rate for *most countries*'? (Figure 1) This visualisation helps to bring out similarities and anomalies between different countries (Figures 2 and 3), discussed later in Section 5. The visualisation can further be used to examine whether similar countries, with respect to these variables, belong to the same geographic regions, or the same GDP grouping. Anomalies are easily detected, which then brings focus to questions as to why certain techniques worked in a region and failed in another. Finally, it begins to provide insight as to what further steps can be taken to remedy the situation, from the lessons that can be drawn from similar countries.

We have also utilised correlation coefficients in another way—to compute correlation coefficient between two countries for a given indicator trend. This analysis can be used to cluster countries based on indicator trends. For example, Figures 4 and 5 visualise the results of this correlation analysis to address the question: 'How do various countries compare with each other in LEB (life expectancy at birth), over the period, 1980–2005'? We chose to compare all the countries with respect to Sweden, which is one of the countries that has a desirable LEB trend over the period 1980–2005, namely steadily increasing from 75.8 years of LEB to 80.5 years over this 25 year period. By choosing a country as a

Figure 1. Correlation coefficients between U5MR (under 5 mortality rate) and DPT immunisation rate for years 1960–2005. This world map depicts that U5MR is negatively correlated (white or lighter shades of grey) with DPT immunisation for most countries as expected. Anomalous countries (shown in darker shades of grey), such as Germany, Kazakhastan and Congo are easily detected in this visualisation. These anomalies are further examined via scatter plots in Figure 2

DOI: 10.1002/jid

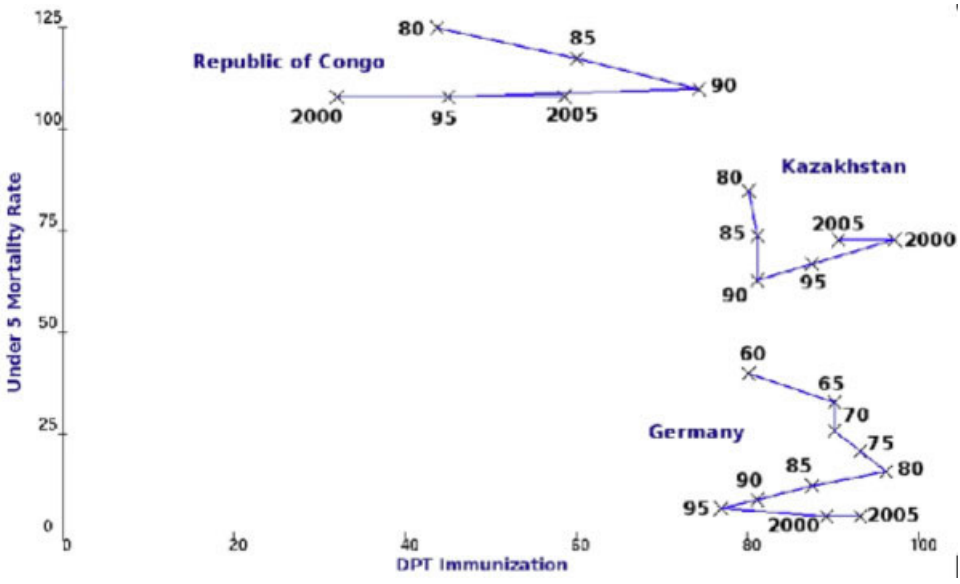

Figure 2. Scatter plots between U5MR vs. DPT for Congo, Kazakhstan and Germany show a reversal of a desirable trend from 1990 till 2000. Reversal occurs in these countries for different reasons—in Congo due to war, in Kazakhstan due to political changes and in Germany due to changes in health policy.
This figure is available in colour online at www.interscience.wiley.com/journal/jid

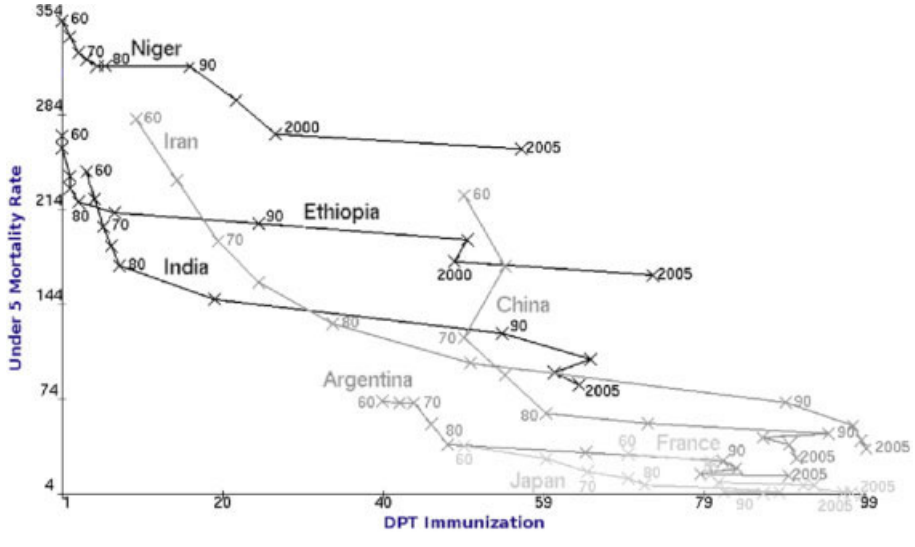

Figure 3. Scatter plot of u5MR vs. DPT showing eight countries from three different income groups (high, middle and low income). These scatter plots show that u5MR is high at comparable levels of DPT for Niger, Ethiopia, India and Iran; medium for China and Argentina and low for France and Japan. For low- and middle-income countries with high u5MR, additional steps beyond DPT immunisation need to be taken to reduce u5MR.
This figure is available in colour online at www.interscience.wiley.com/journal/jid

DOI: 10.1002/jid

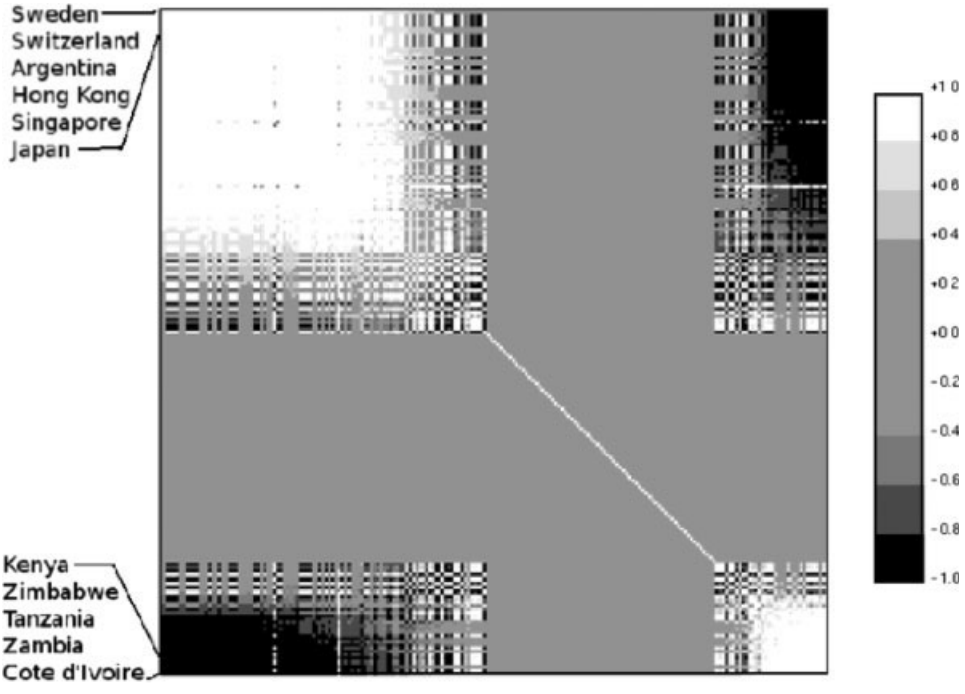

Figure 4. Correlation coefficient between 207 countries for LEB (life expectancy at birth) for years 1980–2005 are shown. Similar countries with improving LEB trends are mapped to white and dissimilar are mapped to black; most high-income countries with improving LEB trends such as Sweden, Switzerland, etc. are clustered together on the top, while most low-income countries such as Tanzania, Zambia with deteriorating LEB trends, are clustered at the bottom

comparison country and mapping the LEB trend correlation coefficients, one can quickly identify countries that are similar to the chosen country with respect to the LEB trend. This visualisation can then be used to explore policy issues implemented in these countries to investigate how to impact these trends. Figure 4 shows the correlation between Sweden and other countries for the LEB trend. In this visualisation, we have sorted the countries in the decreasing order of correlation between Sweden and other countries. This correlation is then mapped onto the world map in Figure 5. This map clearly brings out that most high-income countries have similar LEB trend as Sweden, however, many countries in sub-Saharan Africa (such as South Africa, Zimbabwe, etc.) and several transitional countries (such as Russia, Lithuania etc.) do not share this LEB trend. A further examination of these trends for these countries reveal that they were in upward trajectory for LEB from 1980 to mid-1995s but begin to falter from 1995 onwards till 2005. In the case of sub-Saharan Africa, this may be attributed to the Aids virus, while in the case of transitional countries, this may be attributed to political changes (or socio-economic hierarchy disintegration). This visualisation supported by underlying correlation analysis allowed easy identification of anomalies between countries. This example illustrates that the idea of a cluster analysis coupled with the causality model analysis is likely to be useful in grouping the countries on the basis of the level or dynamics of variables included in a database, such as GHND. A slightly more detailed application of cluster analysis for the grouping of the countries has been reported by the authors elsewhere (Gunawardane *et al*., 2009).

DOI: 10.1002/jid

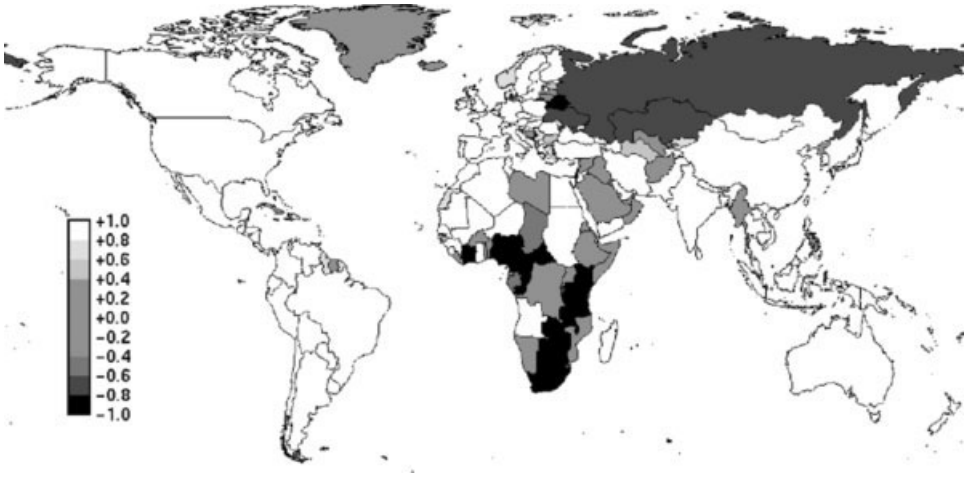

Figure 5. How does the LEB trend from 1980–2005 of a country compare with the LEB trend of Sweden during the same period? Sweden is chosen as a comparison country because it belongs to the cluster of well-correlated high-income countries that have desirable LEB trends over this period (see Figure 4). Countries well correlated with Sweden are shown in white and poorly correlated are shown in increasingly darker shades of grey. Several low-income countries in sub-Saharan Africa and several transitional countries in Eastern Europe do not have increasing LEB trends from 1980–2005

3.3 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 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 ε_i is the residual. One would expect highly correlated indicators to lead to a good linear regression fit and the regression coefficients β_0 and β_1 (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 visualisation for the two selected indicators, we also provide a visualisation of the regression coefficients on a geographic map.

4 INTEGRATING VISUALISATION WITH STATISTICAL ANALYSIS

We have built an integrated geographic statistical-visualisation system that allows us to investigate simple statistical relationships between indicators and countries. We are using correlation and regression analysis as described in Section 3 and simple visualisations using geographic maps and scatter plots. We are also incorporating some enhancements such as the use of correlation matrix visualisation as an intermediate step. Our objective is to contrast the findings from this statistical visualisation system with those from the causality model for global health indicators proposed by Cornia *et al.* (2007).

DOI: 10.1002/jid

Our visualisation system can draw data from various databases including World Bank and United Nations data bank. However, there is a challenge that these data may be incomplete or incompatible with other databases and may require conversion to bring them to a common datum. Data interoperability or exchange between different databases is an important area of research in its own. In this work, we focus only on health indicators and utilise the GHND (Global Health Nexus Database), that has been carefully put together by Rosignoli *et al.* (2007).

We have also developed a user interface that allows easy selection of indicators and countries from a variety of databases and visualisations to create customised visualisations (including zooming and data mining features that allow users to gain access to detailed underlying raw or computed data) that may be helpful in analysing the data at hand.

Our main focus is to investigate whether the integrated statistical-visualisation system can provide any new socio-economic knowledge or insights. We applied our system to investigate deeper questions regarding health variables.

In Section 5, we present two examples of the results of our investigation. Due to simple and familiar visualisations, 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 that emphasise the need for careful evaluation of conclusions, whether they are drawn via simple visualisation or sophisticated causal models.

5 ANALYSIS AND VISUALISATION

We now present two examples to illustrate how the integration of visualisation with statistical tools have provided us with valuable socio-economic insights.

5.1 Statistical Visualisation: Anomalies and Similarities

In this example, we focus on validating how correlation analysis and visualisation may be helpful in analysing relation-ships between indicators. To this purpose, we chose to explore the relationship between u5MR and DPT immunisation for all the countries. Correlation between these two indicators is computed for all the countries individually for a time period of 1960–2005. This correlation coefficient is then visualised on the world map in Figure 1. 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 immunisation reduces u5MR.

Anomalies in the relationship between u5MR and DPT immunisation is also brought out in Figure 1. These anomalies appear as close to zero correlation for some countries. These countries include Congo, Germany and Kazakhastan. Scatter plots of relationships between u5MR and DPT for these three countries are shown in Figure 2. Reversal or decrease in DPT immunisation in Congo from 1990 to 2005 is a result of war. Reversal of decrease in DPT immunisation between 1990 and 2000 in Germany is due to a variation in health policy that has been corrected 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 visualisation on the world map quickly

DOI: 10.1002/jid

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 bivariate linear regression between the two variables. We observed that there is a sharp contrast between high and low-income countries. This observation is validated by picking a few sample countries from each of the three groups—low, middle and high income—and then visualising the relationship between u5MR and DPT on a scatter plot in Figure 3. This supporting visualisation 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 visualisations, the causality model described by Cornia *et al.* (2007) focuses mostly on the regression slope and does not make these 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 analogues are typically correlation and regression analysis. In the examples discussed so far, simple visualisations including scatter plot, correlation and regression visualisation go a long way to provide valuable information regarding the relationship between these variables.

5.2 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 conclusions of the causality model regarding the relationship between these variables. Cornia *et al.* (2007) compute that the regression coefficient between these two variables for middle, low and transitional (Eastern block) countries are 11.2796, 14.2350 and 8.6528 respectively, being significant at 1% level for middle-income countries and being significant between 1 and 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 for high-income countries, also significant at 1% level. The question then arises: Is there a negative impact on LEB by increasing physicians in relationship to GDP for high-income countries?

The regression coefficients, derived by Cornia's model are visualised in Figure 6, where the negative regression coefficient of high-income countries is mapped to the darkest grey colour, while the other three coefficients of transitional, medium and low-income countries are mapped to successively lighter shades of grey in the increasing magnitude of the regression coefficient. These causality results are in contrast with the correlation coefficients visualised for all the countries in Figure 7 or the regression coefficients visualised in Figure 8. These two figures show that the relationship between these two variables are positive for high-income countries, as expected, that is increasing the number of physicians (compared to GDP per capita) 'results' in an increase in LEB.

DOI: 10.1002/jid

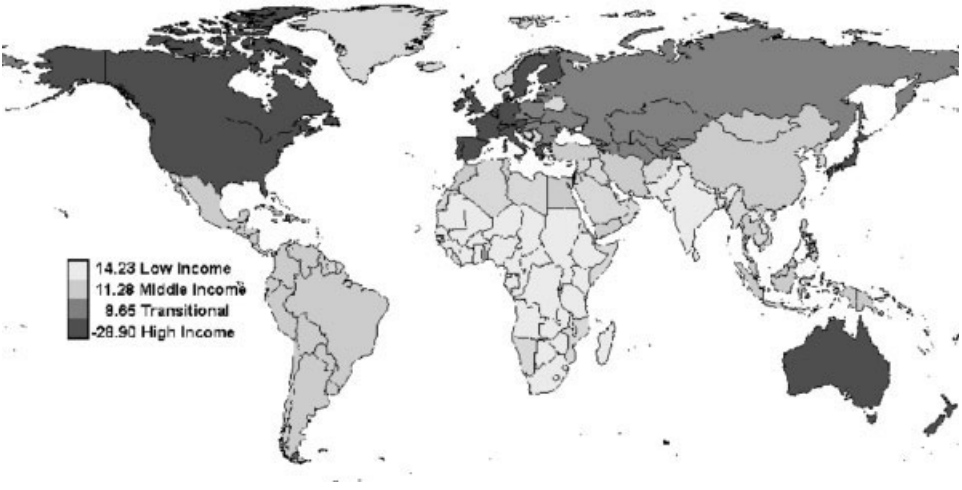

Figure 6. Causal coefficients between LEB and log (physicians per 1000 people)/log (GDP per capita) obtained by Cornia *et al.* In sharp contrast to correlation and regression coefficients shown in previous figures, causal relationship yields a surprising negative relationship between these variables for high-income countries, which is counter-intuitive

How do we reconcile these opposite conclusions between the two approaches for high-income countries? Cornia *et al.* state that of the 10 variables explaining LEB all but one (log of physicians per 1000/log GDP/c) have the right sign. Reasons for the wrong sign are not discussed in the paper by Cornia *et al.* A possible statistical explanation is perhaps that in a multi-regression causality model for high-income countries, 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. However, this explanation requires further investigation and

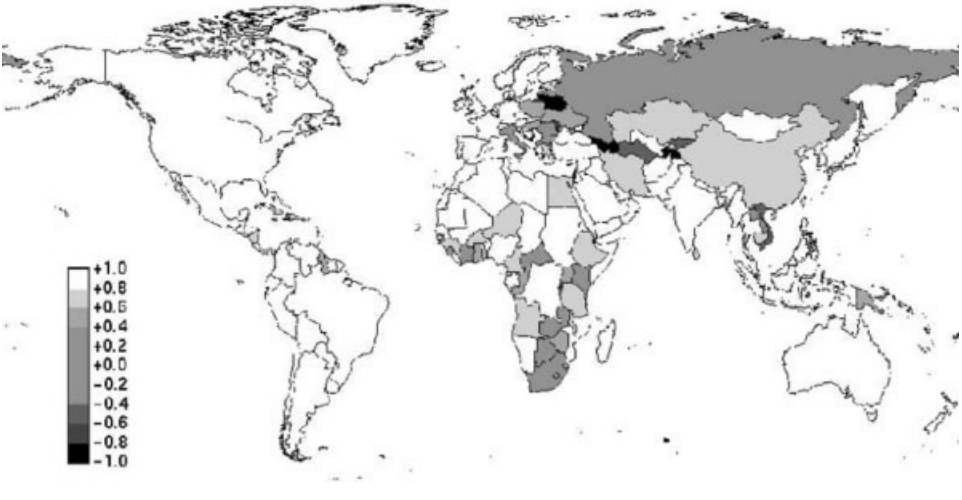

Figure 7. Correlation coefficients between LEB (life expectancy at birth) and log (physicians per 1000 people)/log (GDP per capita) for years 1960–2005. This correlation is positive (white or lighter shades of grey) for most countries including high-income countries as expected

DOI: 10.1002/jid

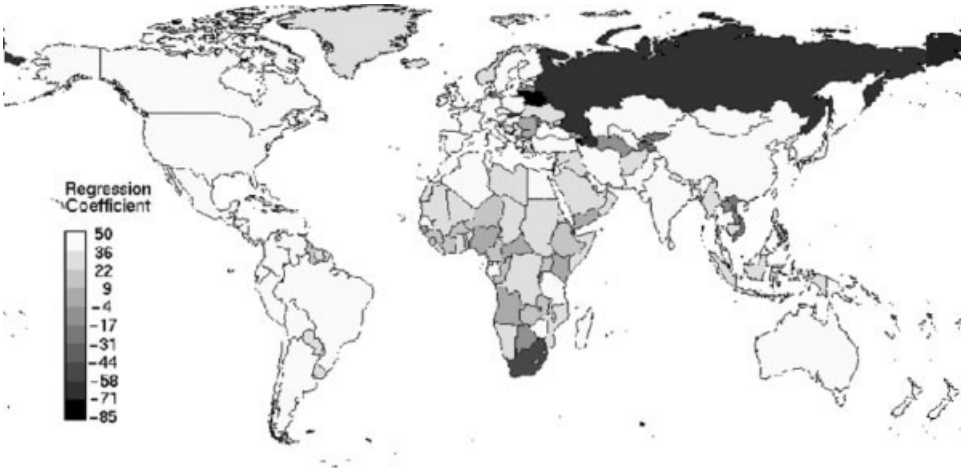

Figure 8. Regression coefficients for LEB (life expectancy at birth) vs. log (physicians per 1000 people)/log (GDP per capita) for years 1960–2005. This regression coefficient is also positive for most countries including high-income countries as expected. Correlation and regression analysis mostly agree with each other in this case

research. How will the sign and strength of this explanatory variable be impacted by addition or deletion of some other variables? In any case, this example illustrates that the use of correlation or simple bivariate regression coefficient, often utilised in various visualisation web sites, may not agree with and may lead to opposite conclusions than indicated by more sophisticated multi-regression causality models, and that further investigation is required to bring deeper understanding of the reasons for seemingly disparate conclusions and related policy issues.

6 CONCLUSIONS AND FUTURE WORK

In this work, we proposed an integration of statistical computing with visualisation to glean deeper understanding of global socio-economic indicators. We utilised correlation and linear regression to quantify relationships between pairs of variables and between pairs of countries. We utilised 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 visualisation support using raw data without the use of any statistical tools. Using two 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 visualisation or sometimes yield 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

DOI: 10.1002/jid

the empowered users can explore relationships between countries and between variables using appropriate statistical tools combined with visualisation. 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.

REFERENCES

Andrienko GL, Andrienko NV. 1999. *Data Mining with C4.5 and Interactive Cartographic Visualization: User Interfaces to Data Intensive Systems*. IEEE Computer Society: G.T. Los Alamitos, CA, pp. 162–165

Carr DB, Chen J, Bell BS, Pickle L, Zhang Y. 2002. Interactive linked micromap plots and dynamically conditioned choropleth maps. *Proceedings of the 2002 Annual National Conference on Digital Government Research*, Digital Government Society of North America, pp. 1–7

Center for International Earth Science Information Network (CIESIN) and World Bank. 2008. Global poverty mapping project. Available at: http://www.ciesin.org/povmap/atlas.html

Cornia GA, Rosignoli S, Tiberti L. 2007. Globalisation and health: impact pathways and recent evidence. *Proceedings of Conference on Mapping Global Inequality.*

Crow B. 2006. *Statistics* in *Encyclopedia of Globalization*, Roland R, Scholte JA (eds). Routledge: London.

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

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

Friendly M. 2002. Corrgrams: exploratory displays for correlation matrices. *The American Statistician* **56**: 316–324.

Gapminder, 2008. Available at: http://www.gapminder.org/

Gastner MT, Newman MEJ. 2004. Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences*, **101** (20): 7499–7504.

Gough I, MacGregor A (eds). 2007. *Well-being in Developing Countries: From Theory to Research*. Cambridge University Press: New York.

Gough I, MacGregor A, Camfield L. 2006. Wellbeing in developing countries: a conceptual approach. *ESRC Working Paper 19*, University of Bath.

Gunawardane P, Middleton E, Lodha S, Crow B, Davis J. 2009. Analyzing statistical relationships between global indicators through visualization. *Proceedings of the ICTD (International Conference on Technology and Development)*.

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

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

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

Hagh-Shenas H, Kim S, Interrante V, Healey C. 2007. Weaving versus blending: a quantitative assessment of the information carrying capacities of two alternative methods for conveying

DOI: 10.1002/jid

multivariate data with color. *IEEE Transactions on Visualization and Computer Graphics* **13** (6): 1270–1277.

Healey CG, Enns JT., 1999. Large datasets at a glance: combining textures and colors in scientific visualization. *IEEE Transactions on Visualization and Computer Graphics* **5** (2): 145–167.

Interrante V. 2000. Harnessing natural textures for multivariate visualization. *IEEE Computer Graphics and Applications* **20** November–December: 6–11

NationMaster. 2008. Nations of the world. Available at: http://www.nationmaster.com/

OECD. 2008. Organisation for economic co-operation and development. Available at: http://www.oecd.org

Panse C, Sips M, Keim D, North S. 2006. 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.

Rosignoli S, Tiberti, Cornia GA. 2007. The globalization-health nexus database (ghnd). Available at: http://www.unifi.it/ dpssec/sviluppo/database.html

Saito T, Miyamura HN, Yamamoto M, Saito H, Hoshiya Y, Kaseda T. 2005. Two-tone pseudo coloring: compact visualization for one-dimensional data. *Proceedings of Information Visualization*.

Sen A. 1999. *Development as Freedom*. Alfred A. Knopf: New York.

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

UNESCO Institute for Statistics. 2008. Global statistics. Available at: http://www.uis.unesco.org

University of California Santa Cruz. 2008. UC atlas. Available at: http://ucatlas.ucsc.edu/

UNSCB. 2008. United nations common database. Available at: http://unstats.un.org/unsd/cdb/cdb help/cdb_quick_start.asp

Wood J, Dykes J, Slingsby A, Clarke K. 2007. 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.

World Bank 2008. World development indicators. Available at: http://www. worldbank.org/data/

DOI: 10.1002/jid