DiviVis: Exploration into Socio-Economic Factors that Can Potentially Affect Individual Internet Use with Visualization

Peyvand Forouzandeh, Graduate Student, University of British Columbia email: peyvand@alumni.ubc.ca
Shirlett Hall, Graduate Student, University of British Columbia email: shirlett.hall@alumni.ubc.ca

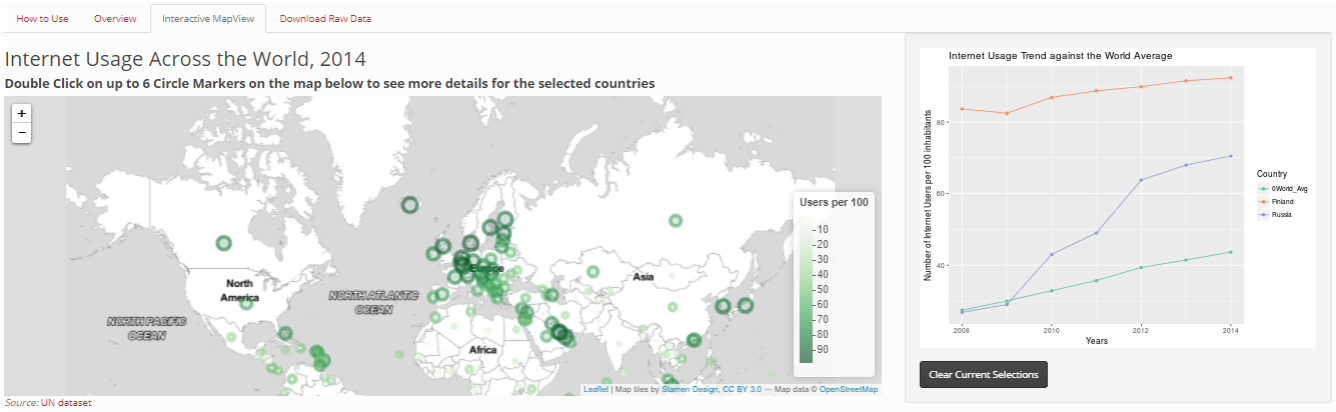

Fig. 1. Interactive map view for Internet usage map in the world in 2014 (left); Internet usage trend against the world average (right)

Abstract— This design study (problem driven programming project) provides illustrations and graphs, which enable us to compare a few social and economic factors with Internet use on a world scale. Through this hypothesis generation process, we aimed to provide a visualization tool for the user who is interested in the investigation of multiple social and economic indicators in relation to the rate of Internet usage in the world in order to find possible relationships between Internet use and geographic boundaries: Are there specific relations or differences between the concentration of Internet use among geographic boundaries including multiple continents or regions? By using visualizations that show the increase or decrease in Internet use over the terms of the seven years from 2008 to 2014, can we identify which social or economic factors from the list selected show potential related effects on the growth and increased demand. To address these questions, we took multiple steps and developed different visualization tools to enable the user to test and compare social and economic factors in relation to the Internet usage for up to 6 countries and target specific patterns or relationships between factors and the rate of Internet use.

Index Terms— Internet, Internet usage, digital divide, social elements, economic, access to the Internet, Worldwide, data visualizarion, design study, correlation analysis, data visualization

1 Introduction

As a means for spreading information at low-cost, the Internet, like other communications technologies, may have a wide political impact. It is also an important element in the current globalization process that is linking countries ever more tightly to a global economy. But the Internet is contributing to widening the gap between the better-off and worse-off parts of the world because it has enabled some nations to create new sources of wealth and political power relative to others [1].

In the past years there has been rapid growth in access to and use of information and communication technologies (ICTs) throughout the world. Given that the technology is widely known and has large benefits, why have some countries not adopted it as fast as others? A digital divide and technology diffusion refer to an economic and social inequality with regard to access to, use of, or impact of ICT. This disparity in information resources is emerging between the global North and global South [2]. This gap between rich and poor is also mirrored in the new information economy. Industrialized economies in the global North are moving towards greater dependence on and access to increasingly complex information technologies. Yet, as of June 2017 data from the International Telecommunication Union show that while 84 per cent of the world's population live where mobile-broadband services are offered, only 47 per cent of world population is using the Internet [3].

The disproportionate number of users is concentrated in the developed countries. By contrast to the progress in the developed world, the digital divide is widening (fewer active users of the Internet) and deepening (greater consequences for not being online) within developing countries, in spite of efforts at bridging it [4]. Generally, two major types of divides exist: an access divide, and skills divide [5]. However, there are multiple barriers identified for disadvantaged communities using computers and the Internet that prevent people from entering the digital world. These barriers include but are not limited to infrastructure deployment, telecommunication policies, English language dominance, and high prices of the service and the device [6]. Globally, the prices for fixed and mobile communication services, which is a critical determinant for the use of ICT has continued to fall over the past few years [3]. Still, domestic business sectors, national governments, and international organizations such as the United Nations, find availability and affordability of high-speed Internet a constraint in developing countries. The challenge is to accelerate the achievement of international goals such as Sustainable Development Goals (SDGs) as well as national objectives to harmonize development

policy approaches in order to create an enabling collaborative environment for everyone. In other words, enabling more people to have access to the Internet can help national and international organizations leverage the full potential of ICT for the achievement of socio-economic development for all. At the same time, the level of socio-economic development within a country may be a precursor to high levels of Internet adoption. It may be a mandate to improve socio-economic development by increasing access to the Internet, but exactly which aspects of the socio-economic development is an area worthy of intense study.

Although there have been many individual-country analyses of the digital divide and some international statistical series -such as those issued by The Organisation for Economic Co-operation and Development (OECD)- there has been little research to compare, synthesize, analyze, and interpret the worldwide digital divide [7].

Unfortunately, there is no reliable data on the exact size of the world's online population either. Most of the international datasets focus on a very limited number of countries, and multiple factors in relation to the divide between nations is not easy to explore and find in those datasets. Therefore, finding a global and authentic source for our analysis was one of the main challenges in this research study. We eventually decided to use multiple datasets provided by the United Nations Data Retrieval System and make a richer dataset for the purpose of our research, which integrates both Internet usage and a few social and economic factors. There are still many pieces of missing data in that dataset, and surprisingly the majority of which are related to the global North.

In this project, we started by analyzing the quality of the data in the dataset and cleaning the data. The data was retrieved primarily from the United Nations Data Retrieval System [7]. It is a central repository of data from other sources including the International Labor Organization (ILO), the World Health Organization (WHO) and the International Telecommunications Union (ITU). All four entities provide different components of the complete dataset used for this project. The major attribute of our dataset is the number of Internet users per 100 inhabitants, on a country level. These attributes were found in separate tables within the dataset owned by ITU.

Based on the nature and quality of data, we eventually decided to identify 6 different sub-categories for social and economic indicators for 203 countries from years 2008 to 2014: Median Life Expectancy, Gross National Income per capita (000s), Percentage of Completed Primary Education, Percentage of Adult Unemployment, percentage of Population in Urban Areas, percentage with Access to Electricity. These indicators are compared against our main attribute: Internet Usage per 100 inhabitants.

Finally, we planned a design-study project that can help us identify any potential relationships between social and economic indicators and Internet usage in the world. By developing this as a problem-driven programming project, we aim to enable users to find answers to the following questions:

1. Are there specific relationships or differences between the concentration of Internet use among geographic boundaries including multiple continents or regions?
2. Using visualizations that show the increase or decrease in Internet use over the terms of the seven years from 2008 to 2014, can we identify which social or economic factors from the list selected show potential related effects on the growth and increased demand?

The application does not seek to assert any causal relationships between Internet usage and any of the selected factors under study. It can serve only as a basis for further research to investigate the digital divide among various countries.

2 Related Work

We divide this section into works aimed at similar problems and works that employed a similar solution but to a slightly different problem:

2.1 Similar problems

While efforts have been made previously to plot the topological structure in terms of the connections between Internet nodes– computer networks or Internet Service Providers, Internet maps (connectivity data) and visualizations to record the route of the information– most previous research and visualization efforts only considered the number of connections as an indicator of nodes [8]. While there are benefits in improving the efficiency of the Internet, and improving the Internet in the future, both in the global South and global North, little attention has been paid to the structure of the demand side and interlinks between socio-economic drivers within the society.

Most of the examples we found about Internet use and its relation to social and economic elements, illustrated the trends in separate visualizations [9], or had only one specific theme such as growth, usage or security [10]. Trend lines, bar charts and world maps are dominant in illustrations around this topic. However, the authors found limitations with single attribute visualizations, like charts and trend lines. They make comparison and identification of patterns and correlations very hard when the researcher has to investigate large datasets, as is the case for 200 countries.

Some of the examples that we specifically found inspiring as case studies are listed below. However, during multiple stages of the design and development of our visualization solution, we also looked at many other types of idioms that may be related to our goals.

- Global Web Index visualized the penetration of social networks for twenty-two countries and a global average percentage. This illustration uses bar charts where hue changes from yellow to red to emphasize the higher density of users with red color. They also distinguish the global average bar with a dark blue color, which is not part of the yellow to red spectrum [11].

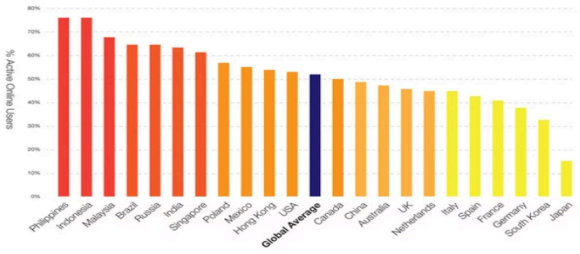

Fig. 2. Global social networks penetration [11]

- Rosling presents the way technology plays a role in the economy of the country with emphasis on the developing world with scatterplots [8]. The color channel is used for the categorical country attribute and the size channel for quantitative population attribute. One of his illustrations shows the relations between world projections on population and GDP per capita and another illustration shows fertility and life expectancy with software that gradually shows the change. The size of the circles emphasizes the population, and color separates different regions in the world.

- International Telecommunication Union illustrates individual Internet use by showing trend over time and in multiple regions of the world in both of their 2016 and 2017 reports [10] [12]. Their use of bar charts, stacked bar charts and line graphs as well as the use of color in the reports gave us a clear image about how simplicity in the visualization of large datasets can help us target specific trends and find patterns in the data presented. Fig 3. is an example of their charts.

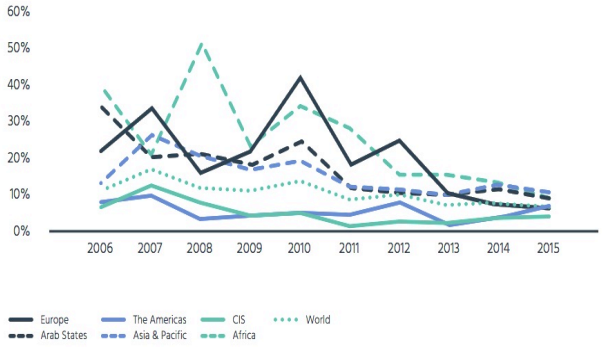

Fig. 3. Regional individual Internet user growth rate [10]

- Information Geographies from Oxford Internet Institute presented multiple examples for Internet penetration and population with multidimensional attributes [13]. The dataset is originally from the World Bank and the distortion in the map paints a picture about human activity on the Internet. Looking at the size of the countries with space-filling layout density shows the rise of specific regions in the world as the main contributors of Internet use in the world. The map shows the geographic pattern of Internet penetration in addition to revealing the difference between the world's largest Internet countries versus smallest Internet countries.

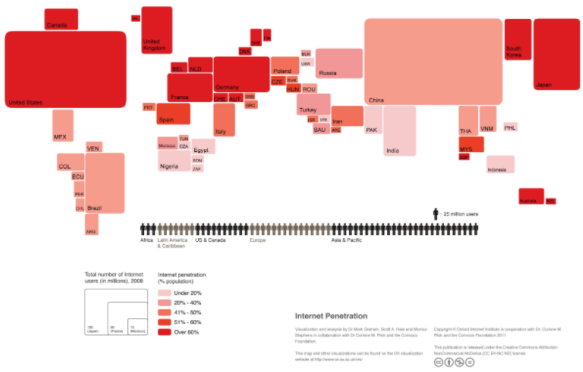

Fig. 4. Internet penetration and population [12]

The examples all have some elements of what we wanted to accomplish, however, what we believe was still missing, was the integration of multiple attributes and the possibility of exploring and discovering data in multiple levels and stages. For instance, Graham and Sabbato [12] visualized Internet population for each country in a map by using the categorical attribute of size (spatial region) for each country in geometric shapes of rectangles. At the same time, they used the ordered attribute of color saturation to show five ranges of Internet penetration for each country. In order to understand the interplay between potentially related factors to the Internet use and to predict future paths and trends, we needed to build a visual system which could help us understand which factors matter most, find similarities, correlations, and make projections.

2.2 Similar solutions

Some examples that we found close to our intentions and goals in visualizing Internet data are listed below.

- Roberston et al. use the same encoding technique with scatterplots where each point mark represents a country, with horizontal and vertical spatial position encoding the primary quantitative attributes of life expectancy and infant mortality [14]. They used an animation tool for the study where each frame shows a year. In the tool, the image smoothly animates between images at one-year intervals. The convergence of most countries happens toward the bottom right corner with high life expectancy, low infant mortality in Fig 5.

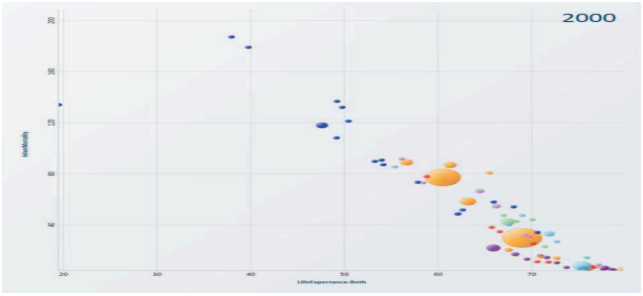

Fig. 5. Animation tool showing life expectancy and infant mortality [14]

- Vismon [15] is a visual tool for fisheries data analysis. There are two input parameters (the management options) and several output indicators. The purpose of Vismon is to help decision makers to quickly narrow down all possible management options to only a few that are agreeable to all stakeholders. Then, a detailed trade-off among the few chosen management options can be performed. This example was specifically informative for us because the user is able to target a specific event by comparing multiple catches (median, temporal and average). In addition, from a drop-down menu, there is the possibility for the user to choose a group of indicators or sort the scenarios.

3 Data and Task Abstractions

3.1 Dataset

The data for DiviVis was collected from an open dataset website named United Nations Data Retrieval System [7] with downloadable format in CSV for 203 countries. The annual data for each country is collected between 2008 and 2014. The original dataset is a flat table with 1,582 items in 3 categorical key attributes and 10 ordered attributes.

3.2 Domain Data

Instead of choosing all categories and items we chose the ones with more valid data. Missing data in our dataset was a big constraint in developing a coherent visualization. Therefore, we had to choose specific categorical attributes with the most number of valid data in order to avoid significant real-time processing delay as well as negative perceptions about the value of the tool. We chose 1,422 items in a table format. The list of the attributes are as follows:

Field Name	Field Description	Categorical, Quantitative, Ordinal	Count/ Range
Country	Abbreviated and most recognizable name of each country.	Categorical	203
Year	Calendar Year that applies to the coverage period for an observation	Ordinal - Integer	2008 to 2014
Internet_Users_per_100	Number of Internet Users for every 100 inhabitants.	Quantitative	0 to 100
Tot_pop	Total Population	Quantitative	-
Percent_rural	The percentage of all inhabitants that live in rural locations.	Quantitative	0 to 1
Percent_urban	The percentage of all inhabitants that live in urban locations.	Quantitative	0 to 1
GNI_per_cap	The Gross National Income per Capita	Quantitative	-
Median_Life_Exp	The median life expectancy. the middle number for the time of age of death	Quantitative	-
Primary_Compl_Rate	The percentage of the population that complete a primary level of education.	Quantitative	0 to 1
Per_Access_Electricity	The percentage of the population that have access to a source of electricity.	Quantitative	0 to 1
Per_Adult_Unemployment	The percentage of adults 25 and over who are unemployed.	Quantitative	0 to 1

Table 1. Selected data from Dataset

3.3 Abstract Data

With respect to attributes, we wanted to visualize country-specific data by year (2 categorical attributes), as well as 6 ordered attributes Median Life Expectancy, Gross National Income per capita, Percentage of Completed Primary Education, Percentage of Adult Unemployment, percentage of Population in Urban Areas, percentage with Access to Electricity. The main advantage of choosing these attributes in addition to having more valid data for all countries was that they were all ordinal attributes that allowed for direct comparisons within the same idioms. This issue specifically was a constraint in terms of choosing categorical attributes such as Language in the dataset, which needed a separate type of idiom for visualization.

Our key attribute is Internet Usage per 100 inhabitants. Except for Gross National Income (GNI) and Median Life Expectancy (MLE), other attributes are all percentages of population. In order to show all attributes in one graph, we had to normalize GNI to fit between 0 to 100 scales, by dividing by 1000. The range for MLE within the dataset was already 0 to 100. Longitudinal and Latitudinal coordinates for each country were also added to the dataset.

3.4 Abstract Tasks

DiviVis is intended to be used by those who are enthusiastic about global Internet use and its possible relationship with social and economic situations in the world but have no idea how to compare the data between countries and explore United Nations datasets. The target audience for this project are College students, researchers, and data analysts in government or Telecom. It can also serve policy makers by showing them potential relationships, historical trends, and help them predict the future trends to make policies around Internet access and use, specifically in the developing world.

In many instances where the researchers do not know exactly what question they need to ask, visualization allows them to explore the data. In this regard, we need to process the information to some level but also provide the basis for the user to see the dataset in detail. Additionally, finding patterns, judging whether the statistical model fits the data will be facilitated by the visualization tool.

Summary List of Abstract Tasks:

- Understanding trends across collection of time-varying tabular data
- Understanding relationships between variables
- Looking at the distribution of the variable across geographic regions
- Comparisons between countries and attributes
- Measuring the effect of the relationship between the main attribute and specific attributes

3.5 Domain Specific Tasks

In order to accomplish the abstract tasks described in Section 3.4, the main tasks that DiviVis supports are:

1. Read the introduction and how to use the website
2. **Level 1:** Browse through the dataset and compare the maximum and minimum and the distribution of data for all countries for the key attribute (Internet usage)
3. Explore the geographic distribution of data and the density and accumulation of the high volumes of the key attribute rates on a map
   Observe the key attribute value for each country on click
4. Select a number of countries by double-clicking on their respective marks
5. **Level 2:** Compare Internet usage trend from 2008 to 2014 for the same selection of countries
6. Differentiate between countries by the use of appropriate information visualization technique
7. **Level 3:** For the selected set of countries, compare multiple social and economic factors and attributes with the key attribute.
8. Compare the countries with each other and try to find patterns, correlations or similarities between different countries and different attributes
9. Clear the selection and try a new group of countries and test your hypothesis
10. **Level 4:** Select a specific attribute from the list and see the extent of the relationship between the key attribute and this selected attribute
11. Compare the trends in selected factors from 2008 to 2014 for all the selected countries
12. Read more specific statistic data about coefficients of the key attribute with other factors

13. Download raw data in tables format with all the attributes and values for the selected number of countries from 2008 to 2014.

3.6 Encoding

Encoding is the most important part of data visualization because this step will transform the abstract data into user-friendly visualized forms. The encoding techniques of DiviVis are briefly summarized in Table 2 as follows:

Idiom	What	Why	How	Scale
Dot chart	**Data:** One quantitative value attribute, one categorical key attribute	Find cluster of countries, Identify countries, Record, Browse	**Encode**: Express value attribute with aligned vertical position and point marks, **Reduce**: Range selector, Plot size adjustment **Facet**: pop-up window on hover **Manipulate**: download	203 countries
Parallel Coordinates Chart	**Data:** Table-categorical attribute (Country name), six quantitative value attributes **Derived:** Average of all quantitative	Lookup and compare values, Record	**Encode:** Line charts, hue by country categorical attribute, hover, toggle lines **Facet**: pop-up window on hover **Manipulate**: download, select and deselect country	6 countries by 7 attributes
Geo-Spatial Bubble Map	**Data:** One quantitative value attribute – Internet Usage using geographic coordinates	Find outliers, distribution ; locate clusters	**Encode:** Express values geographic spatial position, with size of glyph and saturation of color, **Manipulate:** pan, zoom **Facet**: pop-up window	203 countries
Scatterplot	**Data:** Two quantitative value attributes	Find trends, outliers, distribution , correlation; locate clusters	**Encode:** Express values with horizontal and vertical spatial position and point marks with linear regression line, **Manipulate:** drop-down menu to select independent variable	Hundreds of data points
Multiple Line charts	**Data:** One quantitative value attribute, one ordered key attribute **Derived:** Average of quantitative value	Show and compare trends for Internet Usage and Factors	**Encode:** Dots with connection marks between dots, hue by country, animation	6 countries
Table	**Data:** List of all quantitative and categorical attributes	Find detailed information , record, search	**Encode:** Express values in rows and columns, **Manipulate:** Search, Sort, Download **Reduce:** Filter	Up to 42 items by 10 attributes

Table 2. Encoding techniques and idioms used in DiviVis

4 SOLUTIONS

Throughout the design process, we continually assessed the suitability of our choice of idioms for the task required. At times the best choices were elusive due to lack of experience with the programming solution, an R shiny web application, and the scale of the data. All these elements worked as multidimensional forces that eventually shaped our broad view of final project.

We found four levels of data abstraction required and most appropriate to visualize our rich dataset. We specifically were interested in building in flexibility in the presentation of data from an overview to detailed scale; and in facilitating basic statistical analysis for our target audience.

4.1 Interface

DiviVis tool is designed in four stages and levels. When the user clicks on the URL, they will see an introduction page, which is the first tab. The “How to Use” tab gives the user some information about design goals and how to use the application. The second tab is an Overview tab that provides a large image showing the range of the main attribute, Internet Usage, for all countries (level 1). The third tab, “Interactive MapView”, shows visualizations on the main attribute as well as 6 social and economic attributes using different idioms and three levels of detail (level 2 to 4). By clicking on the last tab, Download Raw Data, the user can see their selected data in raw format in a table.

We explain these four levels in detail as follows:

4.2 Level One

In the first level view we want to show the range and distribution of Internet usage attribute for all 203 countries and provide a large-scale picture of the most recent data available for all countries in 2014 in order to help the user see the diversity and distribution of Internet usage in the world.

Since we were looking at one ordinal attribute for 203 countries, the list of the countries is very long, and the user has to scroll down to see all the countries. Therefore, we have decided to show them the big picture but also provide the ability to adjust the range of Internet Usages (between 0 to 100 percent) and the plot height (between 0 to 2,000). This option can be specifically helpful when the user is comparing some countries with similar values or values at opposite ends of the spectrum.

By clicking and holding the mouse to select a specific area on the DotPlot, the user will see a zoomed-in view for that selection. Also by hovering the mouse on top of each Dot on this visualization, we can see the name of the country. On the top right corner of the DotPlot, the user can see different options such as zoom in and zoom out as well as a download option in png format. Using this option, the current selection of the range and height of the data will be used to make a png image. This image can be used a reference for tasks in level two.

The data is presented and ordered from the highest value for Internet Usage per 100 Habitants for Iceland with 98% to the lowest value about 3%, which is related to Burundi. A DotPlot was our choice because we had one quantitative attribute and one categorical attribute. We also needed to express the value for our ordinal attribute with aligned position and point marks. Since the number of countries was a large number - 203, we chose vertical position for the names and the Internet usage can be compared with aligned vertical point marks. It is also possible that the user can target specific countries in the list and locate them in the plot. This data is arranged from maximum to minimum usage. We can also identify possible outliers in the dataset in this stage.

4.3 Level Two

Levels two to four are comprised of a multi-form layout that allows for more in-depth analysis than level one. In the second level we want to integrate the geographic location with the Internet usage data. We started by developing a choropleth for main attribute (Fig. 6). The choropleth map shows a quantitative attribute (Internet usage) encoded as color over regions delimited as area marks, where the shape of each region is determined by using given geometry, which was in our case political boundaries between countries. However, after testing our choropleth map, we realized that the 2D size of the countries could potentially make the understanding of the distribution of our main attribute, the Internet usage, less clear. It was not easy to see how data is clustered in one area with multiple numbers of small countries such as in Europe. We were interested in being able to see the density of higher or lower levels of usage in specific regions, but at the same time we also wanted to enable the user to select any country of their choice by interacting with the map and help them to see the relative level of usage in that identified country regardless of the size of the country. Also, the choropleth only shows the amount of usage by degree by color saturation. By using a geo-spatial bubble map, in addition to ten degrees of ordered color saturation, we are able to use the size of a glyph to show Internet usage for each country (Fig. 10).

This level consists of a degree of aggregation of data where we provide the world average for each of the attributes as one of the items. Since the main intention for showing the data in the global scale is to compare multiple countries and regions, we wanted to provide the opportunity for the user to select a limited number of countries from the interactive map. The user can also see a line chart on the right side of the map to see the trend within the specified timeline (2008-2014) and compare the Internet usage data for the selected number of countries in the line chart (Fig. 11). We used world average data as the default line in our line chart (Fig. 10). Each time the user selects a country on the geospatial map, an animation or spinner appears to indicate that the client has received the input and the resulting chart is being reprocessed. This is important so that the user knows that their interaction is successful, especially in cases where their Internet connection is slow. They will not be left wondering whether they are still connected. The user also can clear their selections in the event they select a country inadvertently.

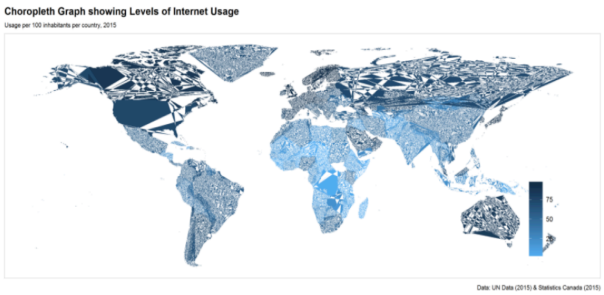

Fig. 6. Choropleth map showing levels of world Internet usage

4.4 Level Three

In this level we want to help the user see the Internet usage in comparison to the six other social and economic attributes in the dataset. This was the most challenging piece of this design because on one hand we wanted to make the comparison simple and clear in order to make the potential patterns and similarities more visible and understandable. On the other hand, looking at six different factors mentioned previously in addition to the Internet usage for multiple countries and the world average rate could easily make the graph unreadable and cluttered.

We started by encoding the attributes into bar charts (Fig. 7) to show the trend for each attribute over one year. The Internet usage rate was shown as a text on top of each bar for each country. The limitation of the bar charts is that they can only encode single attributes and we required multiple attributes and multiple countries. Having all 7 attributes in one idiom could help us find similarity of patterns and trends between countries, which was one of the main goals for this tool. So, for this level we decided to look into other options. We explored the possibility of polar area charts or radar graphs for all the attributes for a single country and comparison against the world average. The problem with the polar charts was that reading the data from this type of layout was not as clear as recti-linear charts. Since we were working on multiple attributes, we realized that comparison between countries in order to find patterns and trends is not easy with polar charts either. Some authors also agreed that in many cases parallel coordinates graph would make the message equally or even more clear than radar graphs [16].

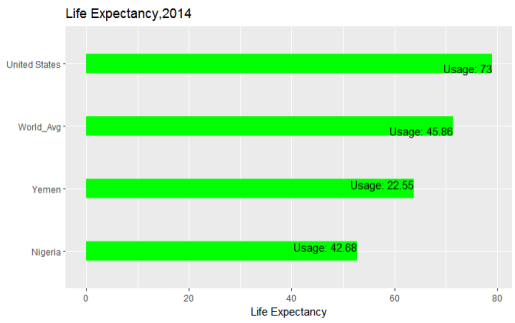

Fig. 7. Bar charts encoding single attributes for each country and the world average rate.

The third option was parallel coordinates line charts (Fig. 12). While parallel coordinates line charts encourage trend assessments, they are also used for ordered variables, which is the nature of the variables in the dataset. We continued with this idiom and investigated different ways that we could show multiple selected countries and 7 attributes in one single graph.

We started this process by testing a selection of 10 countries, however data from multiple years and multiple countries made the loading time very long (greater than 5 seconds) and issues around scalability became a major obstacle. Reducing the number of countries at this stage was a major advantage and we decided on 6 countries as well as the world average (default value) as the limit of selections. This process included some trade-offs. We thought about showing multiple years, multiple countries, and multiple attributes. However, the limitation we had for this was that first-time users do not have intuition about the meaning of the patterns they view when each point represents three different values. The combination of more familiar views such as scatterplots or line charts with a parallel

coordinates view accelerates learning, particularly since linkages reinforce the mapping between the intersection points in the lines [17]. We also had to mitigate the effects of occlusion since many countries can have similar values for an attribute. There were two methods used to reduce this effect – 1. The user can hover over each intersection point to see the exact detail of each country whether or not it is hidden from view; 2. The user can use the legend to make lines on the parallel coordinates graph disappear and reappear, so they can have more control over the view.

4.5 Level Four

In this final level we want to show derived data to help the user to get a better understanding of the level of significance of the relationship between our main data attribute, Internet usage, and other social and economic factors. After the user has navigated through all the attributes at level 2 and 3, we want to reduce the number of data points under consideration to manage the complexity of the dataset and allow the user to focus on the attributes that are most significant to the countries under study.

At this level, the user can investigate statistical indicators such as the p-value and t-statistic. A linear regression chart and a table summarizing the statistical values were our first choices for level 4. However, after looking at multiple scenarios of use and testing the application, we realized that it would also be helpful to add another multiple line chart that shows the trend from 2008 to 2014 for the selected countries and for that specific selected attribute. This can help the user see the growth and changes over time and compare the countries after having an understanding about the level of correlation between that attribute and Internet usage (Fig.13).

As we discussed earlier, the final tab gives the user an option to download the data for the selected countries in the table format as shown in Fig. 14. The table can also be sorted by any attribute or searched for any specific value.

5 Implementation

The tool is built using R [18] with a Shiny [19] Web Interface. In addition to Shiny, the other libraries used in R are built on several languages including, HTML, CSS, and Javascript. Table 3 shows a list of the libraries, and the purpose each library fulfilled in the overall design of the application.

Other Major Supporting R Libraries	Purpose
Plotly	Build Interactivity in parallel coordinates plot and dot plot to reduce the effects of occlusion and scrolling
ColorBrewer	Get consistent color scheme on all graphs for sequential ordering and categories
Leaflet	Create mobile-friendly interactive map
Markdown	Create instructions and short descriptions of variables
DT	Create sortable data frame table

Table 3. Other major supporting R libraries in DiviVis

R and Shiny were the major software used in the final design. R is a popular language and environment useful for data wrangling and data visualization with ggplots. GGplots is also a library within R and was used to create all the graphs, however, another library, Plotly was used to render the dot plot and the parallel coordinates graph on the user interface. Plotly uses a JavaScript library of functions for interactivity. Shiny is an open source R package that allows the developer to embed graphs, maps, tables and other elements on a webpage that can respond to user inputs or selections. The map on the third tab was built using Leaflet, which is also another interface to the Javascript library, useful for creating interactivity within maps. For each country in the dataset, we found a list of central longitudinal and latitudinal coordinates such that Leaflet could plot the corresponding Internet usage data as clickable glyphs on the map. We initially attempted to use ggmap but this idea was abandoned since the Google Services in R did not work well with the layers of glyphs we needed to add. This may have only been because of the incompatibility between the currently installed version of R 3.3.2 and the available version of ggmap.

ColorBrewer is also another package that can be integrated into R and was used for both the sequential saturation of green on the leaflet map, and the categorical colors of the country in the line charts and parallel coordinates chart. The same qualitative color palette, Set 2, was used for all the country graphs so that there was consistency in the lines of each country on each idiom.

Generally, the implementation of this project was a multi-stage process that is similar to a standard system development life cycle, from the initiation of the concept to the deployment of the tool in a production environment. These tasks in the process were divided between the authors and are summarized in Table 4 below:

Task No	Task Description	% Completed by Peyvand	% Completed by Shirlett
1	Develop idea and create scope of proposal	80	20
2	Research on textbook and case studies	90	10
3	Create plan and timeline for execution of the project	60	40
4	Perform requirements analysis against the available tools	30	70
5	Create a mock-up design of the format and layout of the application	50	50
6	Installation of the required libraries and identification of source files, deployment to a Shiny server and backup of source code to Github	0	100
7	Prepare, merge and clean datasets	20	80
8	Create initial/test idioms ex choropleth and bar chart in R	0	100
9	Write the code for functionality including graphs, pop-up, menu, tabs and tables in addition to the instructions on the first tab	5	95
10	Test for expected output and sample analysis	90	10
11	Create PowerPoint presentations	80	20
12	User Testing and incorporation of feedback	60	40
13	Write-up and Report	85	15
14	Literature review, context and background studies	95	5
15	Create Demo	100	0

Table 4. Division of Tasks

6 Results

6.1 Scenario of use

This section describes a scenario where a Sociology professor needs to understand what socio-economic factors may be influencing the persistent digital divide between European and South American countries. He suspects that the difference is due to average income, but he wants to see some data that can help to support this idea and he is unsure about the countries to target for further study.
When he visits the DiviVis web application, he sees the Introduction and How to Use in the first tab with broad explanations about the interactivity of the map, adjustment options with sliders and download options (Fig. 8).

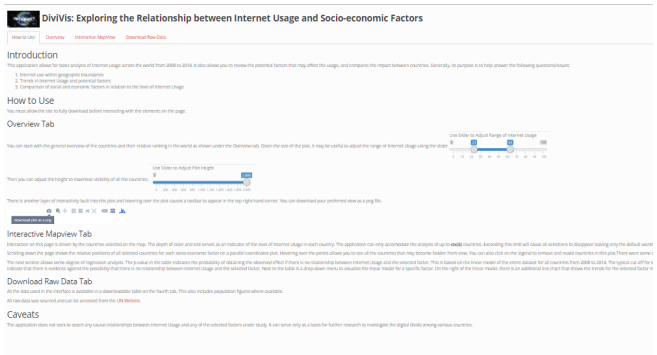

Fig. 8. Introduction and How to Use page, first tab for DiviVis

By clicking on the second tab for the application, the professor will see the first visualization in the first level of detail. Figure 9 shows the DotPlot we designed for visualizing the range of Internet usages for 203 countries. The name of the countries are listed on the left side of the page in a vertical order and the data associated with the Internet use is in a descending order from the highest percentage to the lowest. Horizontal dashed lines connect the name of the country to the green dot presenting the percentage for that country. By hovering the mouse on top of the dots, the professor will see a pop-up text box showing the percentage of usage and the name of the country. This is helpful in case he loses the ability to align the country with the dot on the graph.

Since the professor is interested in seeing only a specific range of the Internet usage, he can use the top left slider to select a specific range. To make this visualization even more adaptable and flexible to his needs, he can also select the height of the DotPlot by adjusting the slider for plot height on the top right side of the page. The professor saves about three snapshots of the resulting graphs, so he can use them as references for the next step.

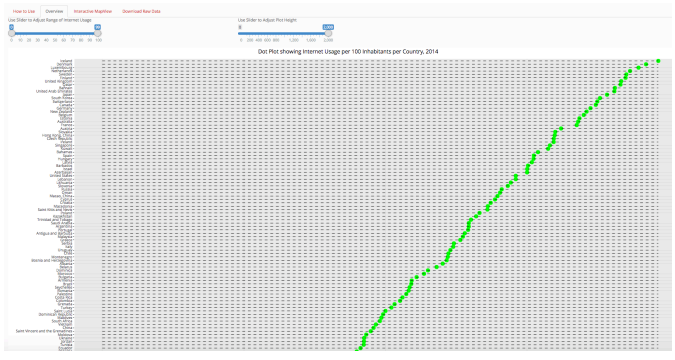

Fig. 9. DotPlot visualizing the percentage of Internet Use for all countries in the second tab and level 1

By clicking on the third tab of DiviVis, the second level of visualization is shown on the top in a Geo-Spatial bubble map and line chart (Fig. 10). The professor can select up to 6 countries by clicking on the markers representing the usage for each country. We chose the green color scheme with different saturations for the map and as a default color for the world average lines in other idioms. By zooming-out and zooming-in as well as panning, the professor can navigate through the map to find the countries he wants to study. Also, hovering on each circle on the map shows the actual value in a text box (Fig. 11). If he does select a 7th country, all previous selections will be removed from the list.

Fig. 10. Geo-spatial bubble map visualizing Internet Usage per 100 habitants on the interactive map (left) and multiple line chart visualizing Internet usage trend against world average (right)

In this scenario, the professor chooses Switzerland, Finland, Chile, Georgia, Mexico, and Kazakhstan to show a wide range of Internet usage (Fig. 11). As we can see in this image, Switzerland and Finland have higher percentages than all the other 4 countries as well as the world average. Also between 2010 and 2012, Kazakhstan and Georgia had a jump to higher rates. The professor can investigate these in the next level.

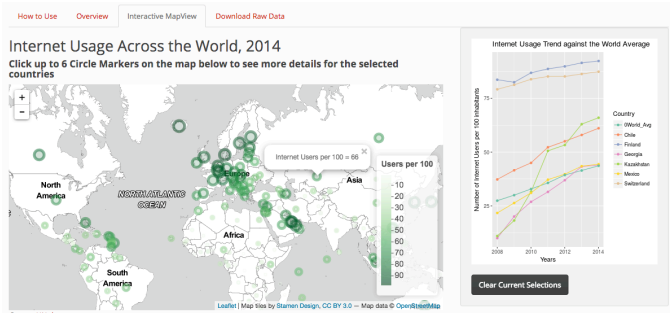

Fig. 11. Selection of 6 countries from the map and visualizing the Internet usage trend in a multiple line chart on the right side in level 2

By scrolling down to the middle of the page the third level of visualization appears in a parallel coordinates line chart (Fig. 12). The description of each attribute is shown in a table on the right side of the page. Each country is associated with a color similar to level 1 and the user can see 7 attributes on the horizontal axis. The vertical axis of the chart shows Internet usage per 100 inhabitants. By hovering over points on each of the parallel vertical lines, the name of the country, name of the factor and the attribute measurement is shown in a box.

In our example, while the professor sees that the unemployment rate is quite similar in all 6 countries, Finland and Switzerland have significantly higher rates of GNI. So, he can make a hypothesis that maybe there is a significant relationship between Internet usage and GNI. Also, the chart shows that while Chile has a

higher percentage of urban density in comparison to the other countries, the percentage of Internet usage is lower than Finland, Switzerland, and Kazakhstan. Therefore, the association is probably relatively less significant between urban density and Internet usage in our second hypothesis. He can review these hypotheses in more detail in level 4.

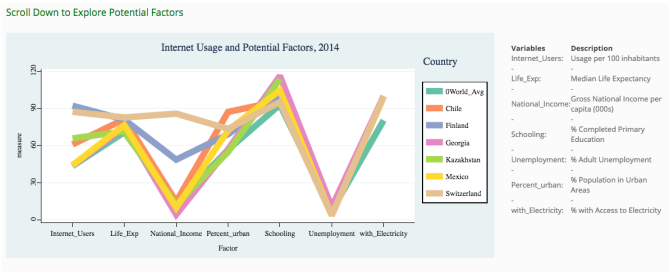

Fig. 12. Parallel coordinates line chart show 7 attributes for 6 countries and the world average in level 3

After, looking at all the factors together, the user can scroll down to level 4 (Fig. 13). In this level a linear regression chart and a table summarizing the statistical analysis for the linear regression is shown on the left and right part of the screen. The user can choose one of the socio-economic factors from a drop-down menu (in our scenario the professor chooses GNI) to test the alternate hypotheses made in level 3. Looking at the regression model and the p-value for GNI versus Internet usage shows that there is a strong positive relationship between these two factors that is not due to random chance. For every $1,000 increase in GNI per capita, the number of Internet users per 100 is expected to increase by 0.72, all else being equal. This can be confirmed by running a Pearson correlation coefficient in a statistical tool, like R, which results in 0.71. The same method shows a correlation value between urban density and Internet usage as 0.5.

Finally, the professor can look at the trends in GNI per capita from 2008 to 2014 on the right side of the page in this level and compare the countries or go back to level 3. He can see that GNI per cap in Kazakhstan has been steadily increasing and confirm that it matches a steady increase in Internet use. He can do further studies to determine whether there are other factors that will serve to bolster GNI per capita in this country and make further conclusions about the state of the social climate in the coming years.

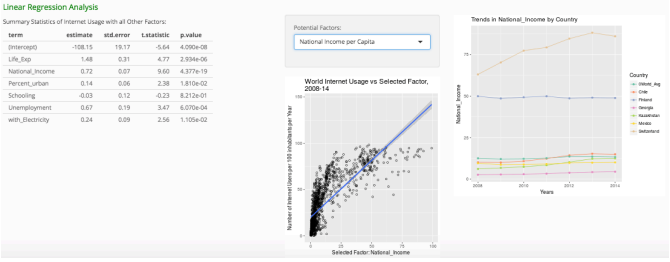

Fig. 13. Linear regression analysis between Internet usage and selected factor and historical trends

In the last tab, the user can download raw data for all the countries selected in level 2 and see the summary of all data for attributes in addition to the population for each country. This level can specifically be helpful when the researcher is comparing multiple countries and wishes to sort by factors.

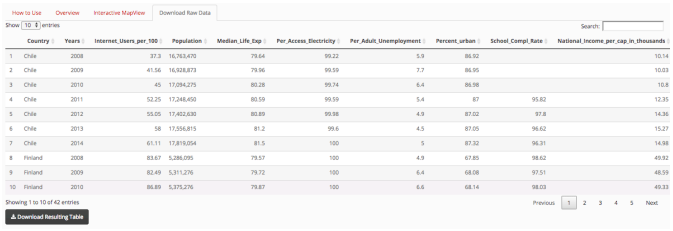

Fig. 14. Downloadable table for raw data for all attributes as well as population and for selected courtiers

6.2 Evaluation

A few users were tested to determine the extent to which the application satisfied the intended goals in a manner that was intuitive and seamless. The main modes of interaction are as follows:

- Mode 1 – the user was quick, no overview reading, some frustration in level 2, retraced, but got to the desired conclusions in about seven minutes.
- Mode 2 – the user was slower and more deliberate. He already had a few countries in mind (Finland vs. Mongolia), explored deeper, and concluded that the results confirmed certain suspicions, and provided a basis for him to research news articles about the state of the countries in present day.
- Mode 3 – the user looked at the map and chose some countries and noticed the parallel coordinate chart was changed. The user scrolled down to see the changes. However, since some of the countries had missing data, he was not able to find applicable correlations for all the countries. The user tested another set of countries. He specified that connection between different idioms could help him understand the flow of the page better. This can be an area to develop in the future. This user was not interested in level 4 and stopped at level 3.

Overall, the users enjoyed using DiviVis and believe that it is a successful and useful tool. One of them asserted that they will use this in their future studies about specific countries. One user also suggested that in addition to social and economic factors we also look at political indicators such as the measure of democracy. This is another potential area of development for the DiviVis tool.

6.3 Performance and Scale

The DiviVis application currently processes the data for over two hundred countries. The application is being hosted on a Shiny server and was tested in Google Chrome, Version 63.0.3239.84 (64-bit).

On initial load, there is a delay of 2-5 seconds until all idioms become fully visible. The responsiveness to any interaction is about 1 second on the overview tab and approximately 2 seconds on the Interactive MapView tab. This is due to the fact that each plot is not modified but is fully recreated with each change in country selection. Since the number of countries that can be potentially added to the source file is very limited, it is unlikely that this delay will increase in the future. The number of factors in the analysis on the page is fairly static, and the limit is only a function of the number of variables that can be viewed in one parallel coordinates graph. The addition of a few more factors will not cause any more delays.

7 Discussion and Future Work

All technological change creates groups that gain and those that lose from the change: its winners and losers. Governments have the

capacity to affect the adaptation of new technologies by making policies that shape the costs and benefits of its use, thus affecting both demand and supply for the technology. The goal for developing a new tool in this research was to show multiple factors from the demand side and design a system that can find answers to questions such as what factors matter most and the effect of these factors.

As we can decipher with this tool, currently Internet usage rates are climbing in many of the developing countries and the developed world still has the most penetration on average. However, the digital divide remains substantial between developed and developing countries. The diffusion of the Internet is not merely a matter of computer technology, but has profound impacts on the continuation of social inequality. People, social groups and nations on the wrong side of the digital divide may be increasingly excluded from knowledge-based societies and economies.

During this project we learned how to work with a complex and big dataset in order to help people find meaningful information from that dataset. We understood that dividing the tasks and data into well-defined clusters could help us understand the patterns, differences and similarities in the dataset to create meaningful hypothesis and a system to test those hypothesis for further research and deliberation. Working and programming in R and multiple libraries also showed us that for different data flows and idioms we needed different tools to be able to create the best experiences for the users. And finally, we learned that after creating a tool, testing the tool with multiple users could give us knowledge to justify pieces of design in order to serve the users better. This understanding could not be achieved without observing the interactions between the tool and the user closely and carefully.

We can identify some strengths in our design and from our experience such as:

- Creating a simple and clear methodology to understand the data and use a large dataset related to many geographic locations;
- Visualizing in multiple scales and levels of details such as the world scale, group of countries and a single country;
- Analyzing the data in multiple scales of details: Internet usage trends, multiple social and economic factors, and statistical analysis with linear regression;
- Providing flexibility to see data in tables, on the map or in the charts for multiple types of users with different skills and interests;
- Creating an Interactive map and multiple plots as different pieces of a system that work together to show the stories behind the numbers using marks and channels.

One of the main weaknesses and issues around this tool is related to the number of missing data, which narrowed the options in visualization. Therefore, a main future step is to find more complete dataset about all attributes specifically in the developed countries. As another future step, we also can consider looking at other social factors such as language, as well as political indicators, such as the rate of democracy in each country.

Other possibilities for future work could be designing the break-out menu beneath the interactive map to choose country/groups of countries for comparison. This can help the user choose the countries without searching for the names and locations on the map.

Another limitation and weakness for this design is that by selecting the seventh country on the map, the tool clears the list of the countries. In the next steps we want to provide the possibility of clearing the list of the countries one by one. Also, based on the feedback from users, highlighting links between charts and maps when clicking on one country can make the connections and the information flow more clearly. And finally, making sortable tables with bars to show similarity in pattern and trend for different factors can be another type of visualization for this dataset that we wish to design in the future.

8 Conclusion

In an effort to see the Internet use within geographic boundaries, to find trends in Internet and social and economic factors, and to test and compare these factors in relation to the Internet usage in big and complex datasets, we wanted to design an interactive tool that was effective at visually communicating these goals. DiviVis is the result of this effort and allows a user to embark on an exploratory data analysis process, which is designed in four levels of information visualization detail. The user can intuitively generate hypothesis and find interplays between six social and economic factors and the Internet usage.

In this problem-driven design study, we were not seeking to isolate a single cause for the levels of Internet use but to find potential relationships between Internet usage and social and economic factors. This programming project helps the user to make guesses about stories behind the relationship between data attributes from large and complex data and use different idioms to generate scenarios, and eventually test that hypothesis by looking at multiple examples. Multiple data analysis approaches in this study provides multiple stages of interpreting data from the dataset that not only makes DiviVis a flexible tool suitable for different types of audiences, but also helps the user see patterns from an overview to a detailed level.

Acknowledgments

The authors wish to thank Dr. Tamara Munzner and Giulio Valentino Dalla Riva for their support and valuable feedback throughout this project.

References

[1] Franda, Marcus. (2002). Launching into cyberspace: Internet development and politics in five world regions. Boulder, CO: Lynne Rienner.
[2] Canadian International Development Agency (2002, p. 10).
[3] Measuring the Information Society Report 2016. International Communication Union. Geneva 2016
[4] Kubicek, H. 2004. Fighting a Moving Target. ITandSociety (This Issue)
[5] France, Belanger and Lemuria, Carter. “The Effects of the Digital Divide on E-government: An Empirical Evaluation”; Virginia Polytechnic Institute and State University (2005)
[6] Chen, Wenhong Chen and Wellman Barry. “The Global Digital Divide-Within and between countries”. ITandSociety, Volume 1, Issue 7, Summer 2004, PP. 18-25
[7] United Nations Data Retrieval System. http://data.un.org/
[8] Graham-Row, Duncan. “Mapping the Internet” 19, June 2007. MIT Technology Review. https://www.technologyreview.com/s/408104/mapping-the-internet/
[9] The Organization for Economic Co-operation and Development (OEDC). Economic and social benefits of Internet openness. 2016. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/ICCP(2015)17/FINAL&docLanguage=En
[10] Global Internet Report 2016. CIGI-Ipsos Global Survey on Internet Security and Trust. 2016. https://www.internetsociety.org/globalinternetreport/2016/wp-content/uploads/2016/11/ISOC_GIR_2016-v1.pdf
[11] How we use the Internet https://wearesocial.com/uk/blog/2011/06/world-map-global-social-media-usage

[12] Global Internet Report 2017. ICT Facts and Figures. 2017. https://www.internetsociety.org/globalinternetreport/2016/wp-content/uploads/2016/11/ISOC_GIR_2016-v1.pdf
[13] Graham, Mark and Sabbato, Stephano. Information Geographies. Oxford Internet Institure. (2013)http://geography.oii.ox.ac.uk/?page=home
[14] George Robertson, Roland Fernandez, Danyel Fisher, Bongshin Lee, and JohnStasko. “Effectiveness of Animation in Trend Visualization.” IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis 08) 14:6 (2008), 1325–1332. (pages 142, 147)
[15] Maryam Booshehrian, Torsten Möller, Randall M. Peterman, and Tamara Munzner. "Vismon: Facilitating Risk Assessment and Decision Making In Fisheries Management". Technical report TR 2011-04, School of Computing Science, Simon Fraser University, Burnaby, BC, Canada, September 2011. http://vismon.cs.univie.ac.at/
[16] Alberto Cairo. Radar graphs: Avoid them 99% of the time. The Functional Art. 2012. http://www.thefunctionalart.com/2012/11/radar-graphs-avoid-them-999-of-time.html
[17] Alberto Cairo. Radar graphs: Avoid them 99% of the time. The Functional Art. 2012. http://www.thefunctionalart.com/2012/11/radar-graphs-avoid-them-999-of-time.html
[18] R Core Team. R: A Language and Environment for Statistical Computing
[19] W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson. Shiny: Web Application Framework for R, 2017. R package version 1.0.2.
[20] Rosling, Hans. The best stats you’ve ever seen. TED video. 2006. https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen