Multivariate Visualization of Geographical Data

By
Muniba Hafeez
NUST201464112MSEECS61314F

Supervisor
Dr. Muhammad Muddassir Malik
Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science (MS CS)

In
School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(September 2016)

Approval

It is certified that the contents and form of the thesis entitled "**Multivariate Visualization of Geographical Data**" submitted by **Muniba Hafeez** have been found satisfactory for the requirement of the degree.

Advisor: **Dr. Muhammad Muddassir Malik**
Signature: _____

Date: _____

Committee Member 1: **Dr. Sohail Iqbal**

Signature: _____
Date: _____

Committee Member 2: **Dr. Asad Anwar Butt**

Signature: _____
Date: _____

Committee Member 3: **Ms. Farzana Ahmad**

Signature: _____
Date: _____

Dedicated to my Father

Muhammad Hafeez (Deceased)

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Muniba Hafeez**
Signature: _____

Acknowledgment

This thesis would not have been possible without the continuous support of my supervisor Dr. Muhammad Muddassir Malik. I would also like to thank my committee members, Dr. Sohail Iqbal, Dr. Asad Anwar Butt and Ms. Farzana Ahmad for their guidance and support.

I want to especially thank my family (my sister and mother) and friends for their constant moral and spiritual support.

List of Figures

List of Tables

Abstract

One of the prime motives of information visualization is the analysis of time oriented data. During last few years, a number of proposed solutions have been published for the visualization of such data. There are other techniques published for the multidimensional data as well as for large datasets. The representation of multidimensional data is an ambitious task although there have come a lot of methods for visualizing such data. Geographical data is mostly represented using choropleth maps, but these maps are helpful if data comprises of uni-variate or bi-variate data. For multivariate data these maps are not helpful, the other representation techniques have to be considered e.g. parallel coordinates, glyph based techniques, icon based techniques, texture representation and motion charts etc. There can be number of techniques for each dimensionality of data that represents data in an appropriate way. In this thesis, a new method for the visualization of time series multivariate geographical data is proposed. The multivariate data is based on some scenarios in which data is divided into input and output streams. The visualization is done using the size factor in which each attribute is represented with this parameter; changes in value represent the change in size accordingly. This technique is implemented using JavaScript library i.e. d3.js and compare it with its established counterpart. An improved usability of 59% is noticed after the experiment. A concise analysis of the proposed technique is presented to open door for more effective visualizations.

Table of Contents

Chapter 1

Introduction

1.1 Data Visualization

Visualization of data is the graphical representation of abstract information. Data is displayed or represented by creating diagrams, drawing images or making animations. The purpose of data visualization is to perform data analysis and communication. The term data visualization describes any endeavors which help people to understand the importance of data by displaying it in ocular context. Every data has important stories, and to understand, discover and exhibit to others, complex stories can be told using graphic designs. Data visualization plays an important role for this purpose. Hidden or unnoticeable information can be better represented by visualizing data using images or diagrams in order to make interpretation of data easier. Visualization is a domain of computer graphics.

Visualization of data is an old term and it has been used for many years to present information. Visualization is used in maps to get the overview of different countries related to that data. One of the best examples of visualizing data is rose chart (Nightingale statistical graph) (Friendly, 2008) which overturned nursing and healthcare. Abstract data means that it has quantitative or statistical information; it means that the things that are described are not physical. Whether it refers to sales, athletics performance, students record or anything else, if it does not have-to do with the physical world, the data can be displayed visually. Abstract data is translated into physical variables of vision which may include length, size, orientation, color, texture and shape. Abstract data can only got success if a little bit about visual perception is understood by us. This helps the users to efficaciously detect and observe the unexpected to attain penetration into data.

Since the 2nd century scientists have been visualizing data by arranging them into columns and rows, but the intention of graphical representation of

quantitative data did not come up till the 17th century. Then, a Mathematician and a French philosopher Rene Descartes developed a 2D coordinate system for showing values, which shows one variable on horizontal axis and other variable on vertical axis. For strong data visualization, a group of skills (mathematical, artistic, statistical) are required, may be this was the reason that late in 18th century the first multivariate graphics appeared. The first person who used a line which showed changes in the values with passage of time when line is moving in forward direction was William Playfair (Spence, 2006). The bar charts and pie charts were also invented by him. The usage of data visualization has turned more popular over time because there is an insistent necessity to convey information effectively and rapidly. It also helps us make data easily accessible, understandable and also usable for other purposes. Some of the applications of data visualization are scientific visualization and information visualization.

1.2 Motivation

Visualization plays an important role in understanding data. There are number of ways to represent data visually depending on the content of data. Data can be of many types uni-variate data (having one attribute to represent), bi-variate data (having two attributes to represent), and multivariate data (having three or more attributes to visualize). For efficacious data visualization, data has to be understood first, and visualized accordingly. Some data has better visualization patterns using bar chart, pie chart and parallel coordinates by (Chan, 2006), while some data has to be displayed on map using different color schemes or by placing different icons on each location for visual representation. These different charts, diagrams help the users in understanding their data and getting outcomes from the data. Some companies visualize their data and find the areas where the company is performing better and where it is lacking. The comparison of companies (how they are performing) to their contenders using visualization is helpful to identify the dimensions.

There are number of methods to represent each dimensionality of data, uni-variate data can be represented using bar chart, pie chart in which there is only one attribute which has to be represented although it may have multiple records, each record is represented with one bar; for example a list of students marks can be visually represented using bar chart in which each bar shows the marks of one student. Besides univariate data, there comes approaches to represent bi-variate data such as scatter plots by (Chan, 2006) in which attributes are represented using x-y coordinates, one attribute is on x-axis

while other is on y-axis and small circles represents the values. Geographical data having two attributes can be represented using choropleth maps in which color selection is based on correlation between these two attributes (Zhu, 2013). Moving forward there comes a time when data becomes multivariate and such data can be visualized using different techniques including parallel coordinates, circular parallel coordinates, glyph based techniques and different textures representation (Chan, 2006). Although multivariate geographical data is also represented using different icons to be placed on map. But only placing the icons on map is not an effective way to represent data having multiple dimensions, representation of multivariate data is still an ambitious task to represent data effectively. Similarly, there exists time series data as well which may contain different dimensionalities of attributes. Spirals can be used to represent time series data (Weber et al., 2001). The techniques to represent such data on map are still limited and there is a need to represent such data in a better way.

The multivariate data can be of many types, some datasets only consists of attributes and there is no relation in between those attributes. Such variables can be represented using the above mentioned approaches. While, there are some cases in which data is based on some scenario, if data consists of such attributes in which there is a relation between those attributes that one attribute is increasing the other is decreasing correspondingly; there is a need to represent such type of data. Such data attributes are classified into input and output terms in which one attribute shows the input value while some attributes show the output values. The data show that if input attribute value is increasing or decreasing how it affects output values, how the patterns of the output variables are going to be changed when the data is of some specific time period. As, the output values show the possible outcomes of data in which there exists success and failure factors. Lets consider an example which clearly shows the scenario based data which is divided into input and output category. Health expenditure is announced by government every year and if government spends enough on hospitals and are maintained properly there can be less chances of deaths in hospitals, or if government spends very little on hospitals and they will not be maintained properly, this causes the mortality rate higher. So, health expenditure causes the mortality rate lower or higher. In this case health expenditure lies in input category and death rate (in hospitals) lies in output category.

Thus scenario based data in which there are multiple attributes, the relation between these attributes have to be understood. If input value is changing with the passage of time how we can display the results of output values so the user can easily understand that if input variable have such value, the success factor in output variable have attained this value and so

user may conclude the better possible results that how the success factor can be increased and how to reduce the failure factor. The attributes which are divided into input and output streams how they can be displayed on map, so different countries can be compared easily based on such time series geographical data. When data is displayed geographically, the approaches of other countries help to improve that situation in ones country or help to identify other dimensions as well. Geographical data should be represented in such a way that user can understand the results and easily make comparisons between different countries; this helps the users in making decisions or to make new strategies for the betterment of their country. This idea leads towards the novel approach of visualization which is helpful for such scenario based data. Although, the use of color scheme in geographical data is an old method in which an attribute is assigned a color. The values of that attribute are represented using darker and lighter shades of the color in each geographical area; such representation is done through choropleth maps. Many other techniques were also proposed to represent one dimensional, two dimensional or multi dimensional data, but still there is a need to propose solution for multivariate geographical time series data which is based on some scenarios (Tominski et al., 2005).

1.3 Proposed Methodology

The research focuses on visualization techniques to provide a novel way to represent geographical data which have multiple attributes in which data is represented in input and output streams. Categorization of data into input and output shows that data is based on some scenario and is not just a collection of multiple attributes. Visualization of data is not just important for the users it should also be interactive as well; interactivity helps the user to understand that visualization in a better way. The different techniques that can be used to represent different kinds of data are explained. A new technique is proposed for visualization of such data which contains data as input and output streams. The analysis of such data through this technique will be helpful to get results and interpret new information. The output attributes are categorized into success and failure factors in which success is represented in upward direction while the failure is represented in downward direction and the value which lies in between success and failure factor moves in horizontal direction. For example: government spent on education sector and we want to know how many people are completing their education, how many of them never enrolled in a school and how many get out of school during their studies. The government expenditure is acting as an input while

the remaining attributes are represented as an output in which education completion shows the success factor and it will move in upward direction and those who have never enrolled in a school shows the education failure and it will be represented in downward direction. This data can be analyzed in such a way that how the government expenditure is affecting the education of students. The analysis of such data can be done through this approach which will be helpful to know that which attributes are affecting to each other. This technique is compared with other techniques to identify the differences in visualization results.

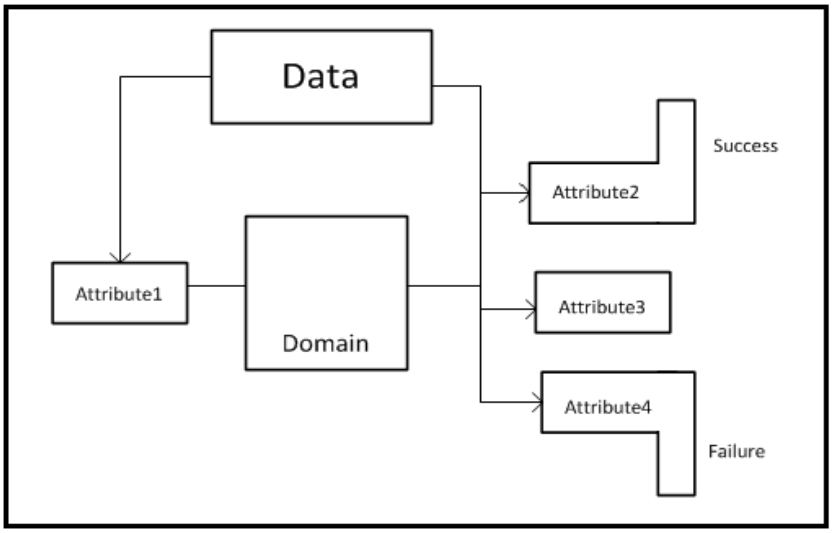


Figure 1.1: Wireframe of Proposed Solution

1.4 Organization of Thesis

The formation of residue of thesis is as follows: Chapter 2 presents state-of-the-art of already existing techniques for visualization of multivariate data. The proposed methodology is discussed in Chapter 3 in detail. Chapter 4 provides the statistical analysis of usability study when proposed solution is compared with motion charts. Chapter 5 finally concludes the thesis along with possible future directions.

Chapter 2

Literature Review

This chapter discusses some of the existing solutions for representation of data. These include uni-variate, bi-variate, and multivariate data as well as geographical data representation using different techniques. The various techniques proposed by different researchers are discussed for visualizing different types of data.

2.1 Concepts and Nomenclature

In information visualization the problem of dimensionality adverts to number of variables or number of attributes that is inside the data which has to be visualized (Spence, 2000). The data usually comprises of a set of records and each record comprises of a number of dimensions or variables. The number of variables can vary in every dataset. For example some datasets may contain four variables while others may contain fifty variables depending on the experimental requirement. Datasets can be uni-variate, bi-variate and multivariate.

- Uni-variate data
  Uni-variate data, also known as one-dimensional data, comprises of one attribute only, e.g. accumulation of cars characterized by monetary value. This kind of data can be represented effectively using histograms.

- Bi-variate data
  Bi-variate data, also known as two-dimensional data, consists of two attributes. The data can be represented in the x-y coordinate system in such a way that one variable is represented against the other variable, also called the representation in scatter plots, e.g. geographical data consists of longitude and latitude as two discrete dimensions plotted in x-y plots.

- Multivariate data
  A third term, multivariate data, also known as multidimensional data comprises of three or more variables. It is difficult to represent multivariate data in single ocular display as compared to the uni-variate or bi-variate data because it consists of large data set variables and it is difficult to show the multidimensional data along with all of its attributes. In (Keim, 2002) parallel coordinates are the best way of representing multivariate data.

- Static vs. Dynamic
  The visual representation of time dependent data is based on two cases (Muller and Schumann, 2003) i.e. static representation and dynamic representation. The representation is static when it does not change automatically over time. When the visualization changes over a period of time then it is called dynamic representation.

- Abstract vs. Spatial
  Abstract data and spatial data can be differentiated as: abstract data that is not connected to any spatial layout, and spatial data comprises of an inbuilt spatial layout (Aigner et al., 2007). Graph visualization, information visualization are more pertained with abstract data. Scientific visualization and geographic information system are concerned with spatial data.

- Linear vs. Cyclic
  Time interval and time points are temporal primitives. Linear time is a time period that proceeds from past to future; it is a temporal primitive with ordered collection (Aigner et al., 2007). Whereas, cyclic time is a time period with bounded set of recurring (e.g. yearly seasons).

2.2 Data Visualization

Visualization of data is the graphical representation of abstract information. The purpose of data visualization is data analysis and communication. Every data has important stories and to understand, discover and exhibit to others, data visualization plays an important role.

2.2.1 History

Human mind is designed to distinguish ocular patterns to get abstract data. The history of data visualization and statistical graphs goes back to map

making when Egyptian surveyors used it for exploration. In 18th century carte chronographique was created by J. Barbeu-Dubourg which was the modern timeline (Fergusoni, 1991). That timeline was created by using sheets of papers glued together. Statistical data was first visually represented by Michael Florent (Tufte, 1997). The father of statistical graphs is William Playfair as he invented statistical graphs which included bar graphs, pie charts and line plots. Today, time interval is represented by horizontal line but it was not common when William created his timeline. In 1861 an informative graph was created by Charles Joseph, he also created the Napoleons Russian campaign graph in 1812 (Tufte and Graves-Morris, 1983). It is an extraordinary graph that makes information of six attributes in 2D as seen in Figure 2.1. This graph shows losses, army position and weather (temperature).

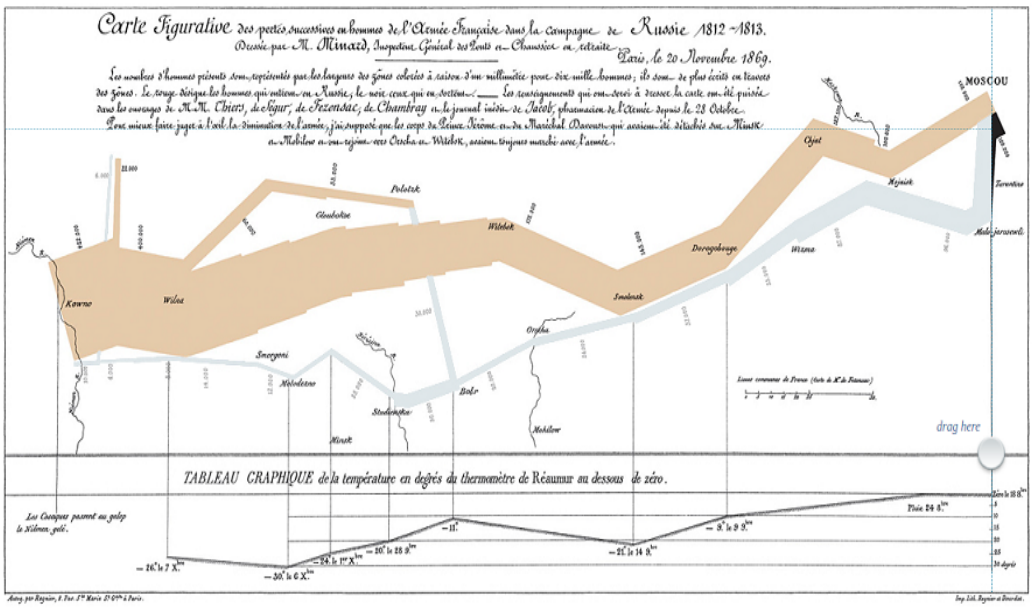


Figure 2.1: Napoleons Russian Campaign Graph. Available at: (http://www.masswerk.at/minard/)

In 19th century, another remarkable work by Florence Nightingale is statistical graph (Friendly, 2008) that overturned nursing and healthcare. She made crushing discovery that in the battlefield many lives of soldiers were taking by due to infectious disease instead of wounds. Data of (deaths due to diseases) for two years was recorded by her and she invented a chart called the rose chart which conveys her findings, seen in Figure 2.2.

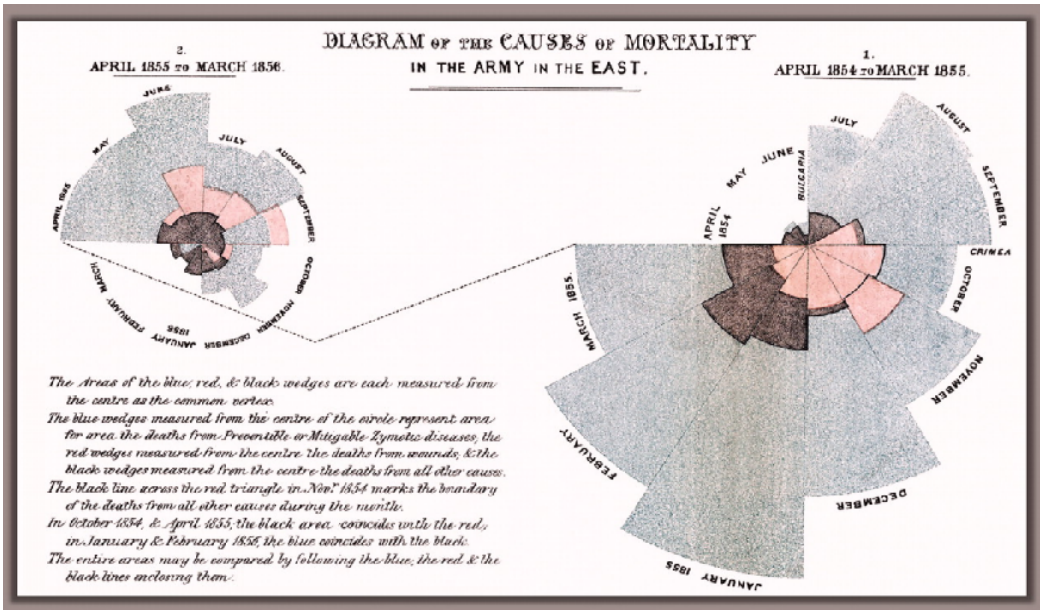


Figure 2.2: Mortality Causes Rose Chart by Florence Nightingale Available at: (http://alainapincus.com/wamfa2014/wp-content/uploads/2014/02/F1.large.jpg.)

Laurence Gantt invented Gantt charts and the timeline graphs which he invented are still in use today with some minor changes. Phillips shows the unemployment vs. inflation rate of Japan in very interesting way from time period of 1960 to 1991 (Tufte, 1997). Both of these variables are represented in x-y plot and dependent on time, where the time is represented using textual labels instead of representing on x-y axis, seen in Figure 2.3 showing the comparison between inflation and unemployment rate.

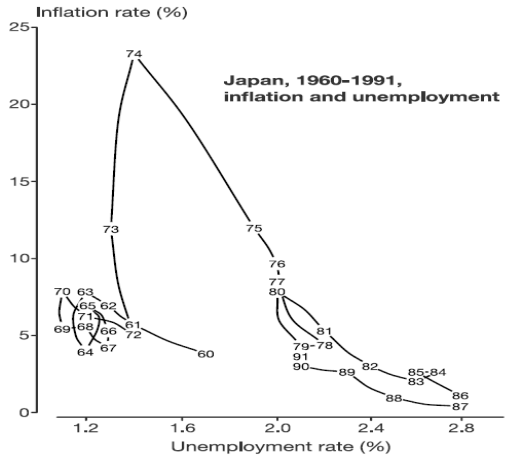


Figure 2.3: Japan, Inflation Rate vs. Unemployment Rate (Tufte., 1997)

2.2.2 Information Visualization

Information visualization is ocular representation of data which is abstract to make human cognition stronger; data may contain both numerical as well as non-numerical data. This is focused on production of methods for conveying information which is abstract in intuitive ways. The field of information visualization has resulted due to research in many other fields such as graphics, human computer interaction (HCI), visual design and computer science. As the amount of information is increasing rapidly it will become a challenge in the next decade to show such a huge data. Computers are powerful and help us collect and store large amount of data but our ability to understand such data or information is still slow (E. Morse and Olsen, 2002).

2.2.2.1 Focus & Context Visualization

Data is just not to be only visually presented but it should be interactive as well. In (Tominski et al., 2005) some of the techniques to interactively represent data in better way are:

- Overview & Detail
- Focus & Context

Overview & Detail has two different views; Overview is showing context information while Detail means detailed information. Overview provides representation of data and then allows the user to zoom in for data analysis in detail. In Overview & Detail, representation of detailed views is separate

from overview and user has to integrate both views. While in Focus & Context technique, both context information and detail are combined in common display (focus within context). Without losing context user can analyze details in single view. The example of Fisheye view shows Focus & Context visualization in which area of interest is shown quite large while the surrounding information is small showing less details.

2.3 Data Visualization Techniques

2.3.1 Multivariate Data Visualization

Different techniques of representing multidimensional data have been discussed by (Chan, 2006). These techniques are divided into different groups namely geometric projection, icon based, pixel oriented, and hierarchical display. Some of the techniques which belong to these groups are discussed in detail.

2.3.1.1 Scatter Plot Matrix

Scatter plot is usually used for two-dimensional data in which one variable is plotted against the other variable in the x-y Cartesian coordinate system. As scatter plot is for the representation of bi-variate data, the extended form of scatter plot is scatter plot matrix which is used for representation of multidimensional data. Scatter plot matrix is a collection of multiple scatter plots placed in a matrix helpful to give correlation information between the attributes where each scatter plot represents a combination of two variables data; this correlation information can be seen in Figure 2.4. This technique is helpful for limited variables of data because if the data set consists of huge variables the visualization becomes chaotic.

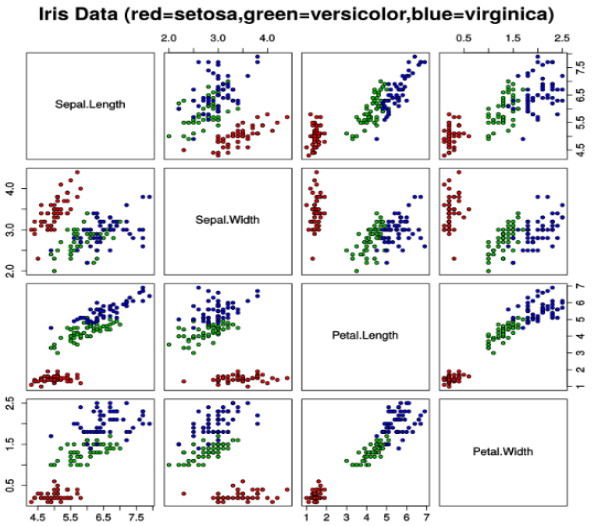


Figure 2.4: 4-Dimensional Scatter Plot Matrix of Iris Data. Available at: (https://en.wikipedia.org/wiki/Iris_flower_data_set#/media/File: Iris_dataset_scatterplot.svg.)

2.3.1.2 Parallel Coordinates

In (Chan, 2006), parallel coordinates for each attribute are represented by vertical lines which are parallel to each other. A polygonal line is used to encode each data item according to its respective value by intersecting each axis, seen in Figure 2.5. Correlation among the variables can be studied using parallel coordinates. If number of variables is too large, the representation of data on axis becomes closely packed making the space between each axis limited.

The best way to represent multivariate data is using parallel coordinates. For representation of extremely large data, an event based approach was introduced (Tominski and Schumann, 2004). In event based approach, only the informative and required data is represented while the rest is discarded. For this purpose, the user has to specify an event. Events are calculated using formulae and detected by event detectors. If the event is detected, that tuple is visualized. This approach is applied on tuples, for example setting up a threshold for an attribute, those tuples that satisfy this action will be visualized. All data is represented using parallel coordinates, when user specifies some event that data is brushed up. This approach improves the efficiency and flexibility of the visualization. The event based model can be seen in Figure 2.6.

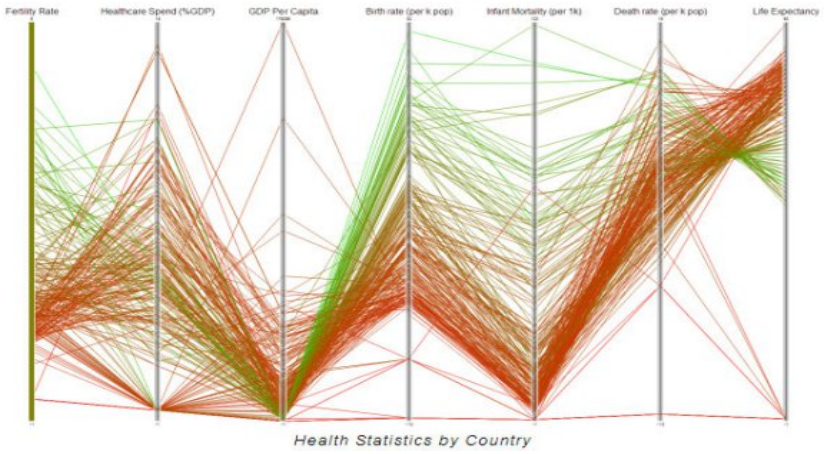


Figure 2.5: Parallel Coordinates of Health Statistics. Available at: (https://anonymousbi.files.wordpress.com/2013/01/6-parallel-coordinates.jpg.)

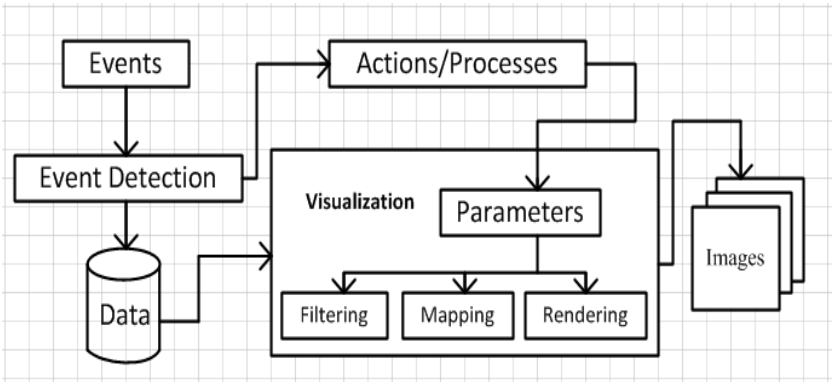


Figure 2.6: Event Based Model (Tominski et al., 2004)

2.3.1.3 Circular Parallel Coordinates

Another way of representing data in parallel coordinates system is circular parallel coordinates in which axis have radial arrangement (Hoffman, 1999). All attributes are equally distributed in circle, the outer part having longer line segments so the data having higher values can be mapped there, while the values in inner part of the circle are more jumbled. Representation of circular parallel coordinates of iris flower dataset is seen in Figure 2.7.

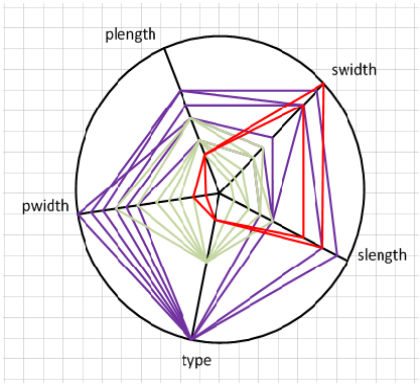


Figure 2.7: CPC of Iris Dataset (Hoffman., 1999)

2.3.1.4 Table Lens

A table consists of number of columns and rows. In table lens (Fao and Card, 1994), number of columns comprises the attributes whereas each row represents the value of that corresponding attribute. The values inside the columns are viewed as a plot or histogram. In traditional tables, rows or columns store information and this information can be rendered as logical, same is the case in table lens though they were actuated by traditional tables. It helps users distinguish relationships; examine trends in data making the entire dataset easily viewable.

2.3.1.5 Icon Based Technique (Chernoff Faces)

Multidimensional data can be mapped using iconography in such a way that each data item is represented using an icon or glyph by (Chan, 2006). In iconography, the most famous visualization is by chernoff face. A face has different properties; each attribute is mapped to one of its properties (mouth, shape of nose, eyes and shape of the face etc.) as seen in Figure 2.8. Each attribute is assigned to one of the properties but the shortcoming of this technique is that chernoff faces visualize only a limited amount of attributes.

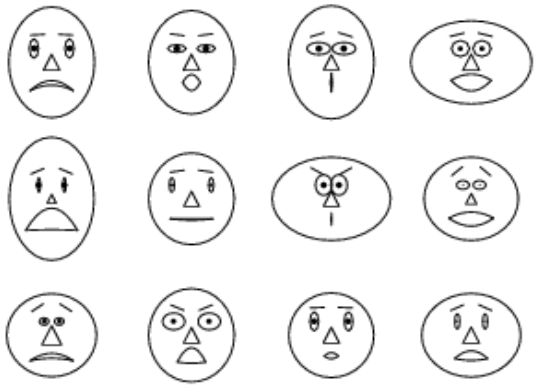


Figure 2.8: Chernoff Faces. Available at:
(http://mathworld.wolfram.com/ChernoffFace.html)

2.3.1.6 Star Plot

There are different kinds of glyphs that can be used to represent multidimensional data; one of the commonly used glyph is the star plot (Chan, 2006). In star plot, a circle is divided into equal angular axis to represent each attribute (Hoffman, 1999) and the data points are connected through a line cutting each axis, seen in Figure 2.9. As the data set contains the number of data items, each value is displayed by its own star glyph. The more the number of records in the dataset, higher the number of star glyphs. This visualization is also helpful for moderate sized data because large data items make the representation intense.

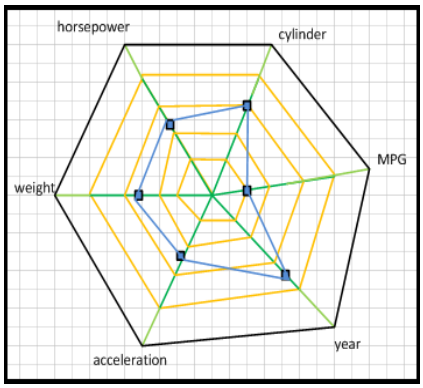


Figure 2.9: Star Plot of Dataset (car) (Chan., 2006)

2.3.2 Bi-variate Data Visualization

In visualization, what are the things that should be focused upon considering data can be of any dimensionality? What methods can be used for quantitative data to encode on a map? Bar charts and line charts provide the best representation of data in graphs but not in maps. Data representation on map using bars is not appropriate because all bars are not placed on the same baseline, this causes the comparison of values vitiate (Few and Edge, 2009). Two approaches that can be pertinent to represent quantitative data are: changes in color intensity or variation in size or both.

Data can be represented using the color or size of any shape but which color scheme is more attractive? Mostly, bi-variate data is visualized through scatter plots in which one variable is on x-axis and other variable is on y-axis, both are against each other. In (Brewer, 2006) prostate cancer mortality rate was discussed in which they showed the ratio of cancer in black males vs. white males using multivariate symbols. Co variation of these two attributes has been represented using the color scheme. Number of deaths of white males data is divided into two columns while the black male data is divided into three rows. Making a correlation between these two attributes, color scheme is settled. Lighter colors are used representing lower mortality rate while darker colors represent high mortality rate. This data can be represented using a combination of size and color both as an example of quantity/quantity symbols. Death vs. death rate data has been represented using this method in which one quantity represents size and other quantity represents hue, seen in Figure 2.10. Rows represent number of deaths while columns represent death rates. If both are low, color is light and size of the shape is small but if death rate is high as well as number of deaths, size of the shape is huge and if the number of deaths is high along with a high death rate, the intensity of color is dark.

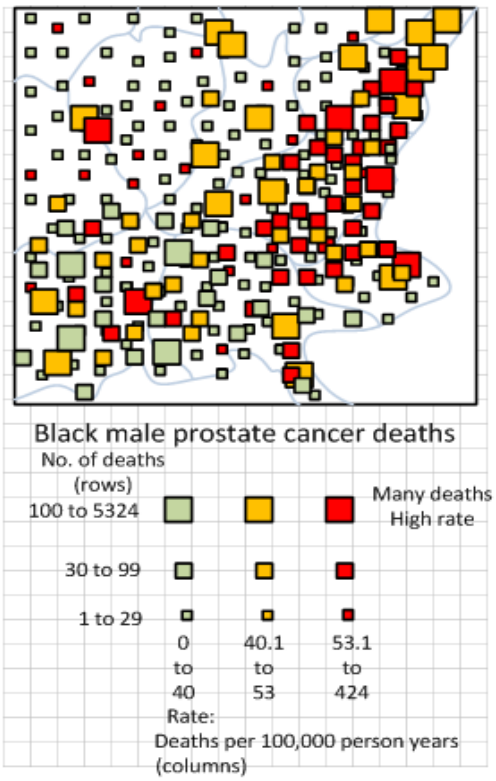


Figure 2.10: Cancer Mortality Rate (Brewer., 2006)

Uni-variate data is mostly shown on map using choropleth maps in which there is only one attribute which is represented by color scheme. Color range is divided into high and low values of that attributed data. Choropleth maps are also used in bi-variate data where the color selection depends on correlation of these attributes. Data can also be presented on map using pie charts technique; each countrys data is plotted on pie chart and these pie charts are placed on the centered location of each country on the map. Similarly, the population of world is also presented on map using choropleth or by using the size of the circle; size of circle shows how much the country is populated and these circles are placed on each countrys centered location. Choropleth maps are also referred as geographic heat-maps (Zhu, 2013). Population density of world using choropleth map is shown in Figure 2.11.

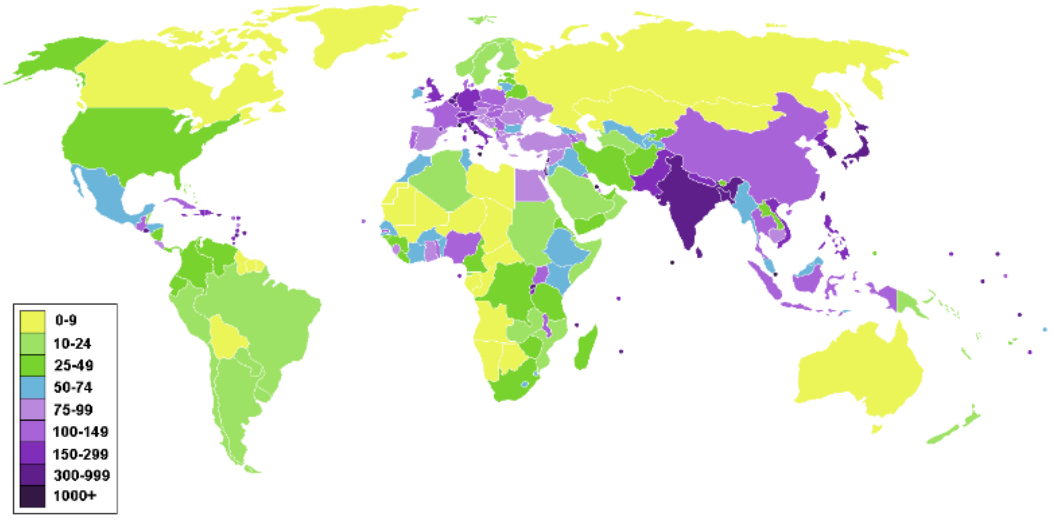


Figure 2.11: Choropleth Map of World Population Density. Available at: (https://commons.wikimedia.org/wiki/File:World_ population_ density_map.png)

2.3.3 Time Series Data Visualization

The most appearing problem in engineering, science and business is the analysis of data which contains time series data. The analysis of time series data helps to find underlying process, to anticipate future developments and to identify trends. The analysis of data shows a periodic behavior, a model to judge such trends. Visualization through line graphs has been proven effective for the time series data analysis. A new approach of Spiral Graph was introduced by (Weber et al., 2001) for the time series data visualization. Instead of circles, trends in dataset and periodic behaviors can be detected effectively using spirals; it also determines the intensity and the length of cycles. Spirals have circular structure; this makes detection of cycles and periodic dataset comparison easier. The time series data shows that data is continued from a specific time period and this continuity is expressed in spirals clearly rather than in circles. Different datasets can be compared by using multiple spirals; they are differentiated with different ocular encoding, such as texture, line style or color. Spiral graph can be seen in (Weber et al., 2001) which is representing intensity of sunshine and this visualization makes the comparison between different days intensity easy. The spiral effect can be seen in Figure 2.12.

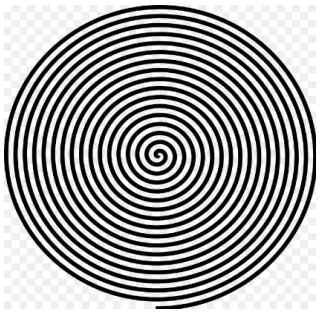


Figure 2.12: Spiral view

It is difficult to visualize complex datasets which comprises time series, geographic locations and multidimensional variables but such datasets have great potential to give valuable information that is helpful in understanding complicated systems and phenomena (Andrienko et al., 2003). Visualization of spatio-temporal data is always an ambitious problem because it is difficult to find complex patterns when the dimensionality of dataset is large. Development of an approach is required for such datasets to find unknown complex patterns and show them in such an easy form which supports human interpretation and decision making. Large datasets may cause computational efficiency and ocular effectiveness problems, so techniques should be computationally efficient. When dataset is too large or is to be visualized on scatter plot or parallel coordinates some of the values can overlap during the visual display. To resolve this issue, two directions have been proposed; one way to resolve overlapping is by repositioning the data points and the other way is to minimize the size of the data by clustering. In (Guo et al., 2006), the clustering approach is used to handle such large dataset having multivariate attributes. SOM self-organizing map is used to perform coloring and clustering on multivariate. Clusters are arranged in 2D layout and there are more chances to use many colors for the representation of each cluster. The data is represented in the form of data cube, seen in Figure 2.13.

SOM (Self Organizing Map) is used to cluster the multivariate profiles in a year/state combination. Two dimensional layouts are used by SOM to order the clusters; likewise color scheme is also in 2D so that each SOM node is assigned a unique color. Each cluster has different size which is represented by a circle and it comprises number of data values; the darker colors shows the dissimilarity between the nodes. Parallel coordinate plot (PCP) shows the representation of each data item of multivariate profiles. Using this plot, one can identify which states are laying in the same clusters of that time period. The thickness of the lines in PCP shows the size of cluster as seen

in (Guo et al., 2006). Visualization is shown on map in which each state is assigned a color of the cluster in which it lies. The visualization is also static because all the years visualization is shown in one single image, if there is more number of years to visualize, the map becomes smaller and it is difficult to see the smaller portions inside the map.

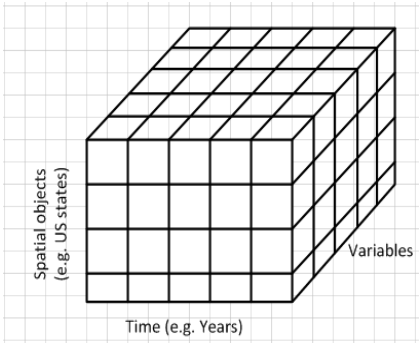


Figure 2.13: Data Cube (Guo et al., 2006)

SOM is used to represent multidimensional data through a coloring scheme in which distance between the neurons are marked; the structure of cluster cannot be manifest and their shape is often deformed. This problem is resolved by technique ViSOM (Yin, 2002) in which map resolution is controlled by ordering the distance between the neurons.

The term visualization represents the phenomena of presenting data in the form of drawings or some kind of shapes. The methodology to represent multivariate data on map consists of 3D information visualization and Information hiding (Tominski et al., 2005). The data can be spatial-temporal data and to map temporal dependencies, 3D icons can be utilized. Spatial dependencies can be represented by placing 3D icons on the map. 3D icons can be represented on map using pencil icons or helix icons. Multivariate data consists of more than three attributes and these attributes can be visualized using the pencil or helix icon. As the shape of pencil has a clear effect, it has a lot of faces and each face can individually represent single attribute. The attributes are differentiated using different colors so none of the attributes can be mixed up to another attribute. This icon is representing temporal data so each face is cut into its time series to give the clear picture of the entire dataset. These icons are then placed on the map on the centre region of every area of the country, seen in Figure 2.14.

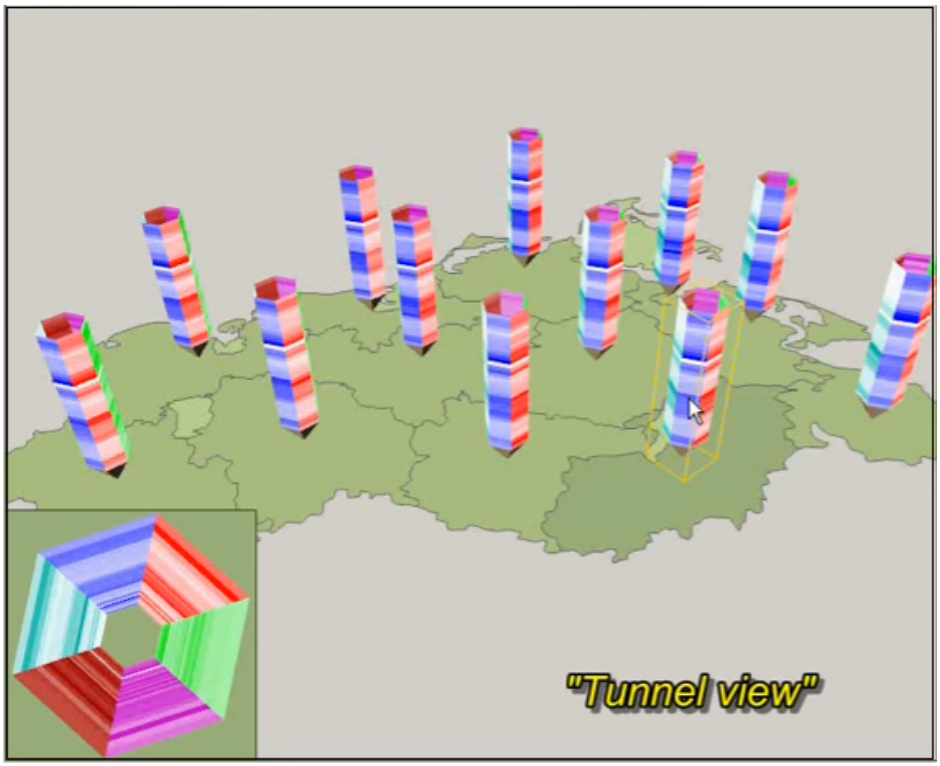


Figure 2.14: 3D Pencil Icon shows representation of six diseases, each disease is assigned a unique color. Available at: (https://i.ytimg.com/vi/wtFLqdKOImk/maxresdefault.jpg)

Helix icons have been used to represent cyclic time data and each ribbon of different color represents a single attribute, seen in Figure 2.15 (Tominski et al., 2005). These icons are placed on the map and give the overview of the data. If the user wants to get more detailed information, these icons can be zoomed in as well as can be rotated but there are chances of missing data when they rotated so they can be placed according to the view of the user. If data set is too large to visualize, information hiding is utilized. The irrelevant information can be hidden and only the interesting data can be visualized.

However, (Tominski et al., 2005) focused on 3D icons to show how the data can be presented using these icons but in order to get the analysis of data, its complicated to know in which time period the value becomes higher because there are no time stamp on these icons and are only divided into time series. Secondly in data hiding, there can be more chances of important data being hidden causing loss of important information.

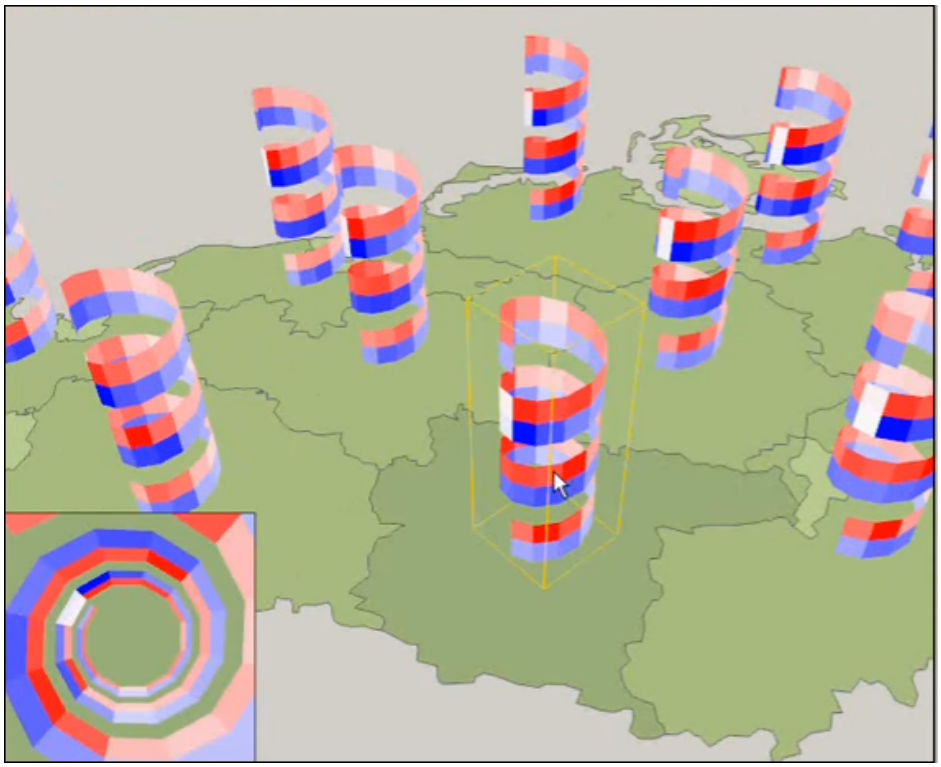


Figure 2.15: Helix Icon show the cyclic properties of two diseases. Available at:
(https://i.ytimg.com/vi/wtFLqdKOImk/maxresdefault.jpg)

In this chapter, some of the important terms related to data visualization have been discussed. The data to be visualized can be of many different types. Different types of data were also discussed. Researchers have presented some of the techniques to present data having different dimensionalities; these techniques are further helpful to present data on map but not every technique is suitable for data representation. Furthermore, visualization history is also discussed which was an initiative of this graphics field or the initiative of data representation. Initially, the data was focused on 1 or 2 dimensionalities but representation of multidimensional data on map is a challenging task. Some techniques are helpful to visualize data on map, but there is a need for more techniques in order to present geographically multivariate data effectively.

Chapter 3

Proposed Methodology

This chapter discusses the novel approach to represent multivariate data based on some scenario on map. As in previous chapter some of the techniques were represented for data visualization by different researchers. The methodology to represent data will be discussed in detail in this chapter.

3.1 Data Dimension

Different types of data have different dimensionalities; such data is to be visualized. Three different dimensionalities of data have been discussed in previous chapter. This methodology will take multivariate data (having three or more than three attributes) to visualize. The data is based on some scenario in which attributes are divided into input and output streams. One attribute is represented as an input; second attribute is represented as any domain or area on which input attribute is playing an important role. Further three attributes are represented as an output which have different meanings. Out of three output attributes one will show the success factor while the other shows the failure factor and third attribute lies in between the success and failure factor. So, multivariate data is divided into input and output streams where there is only one input variable and three output variables. The data which is based on such scenario where the success and failure factors are present will be visualized using this technique and this technique will help to gain new information through analysis of data.

3.2 Visualization Technique

The visualization of such data use HTML, CSS, SVG along with JavaScript library such as d3.js for drawing and displaying data interactively. D3.js

plays an important role for visualizing data interactively. The visualization technique is discussed in detail:

The visualization is created using d3.js JavaScript library in which data is represented using pipeline flow in such a way that one input arrow is coming towards the domain and it excludes the output arrows in three different directions. The image of such visualization is shown in Figure 3.1. Data is displayed in such a way that input attribute is represented by an arrow; the size of the arrow may vary in vertical direction which shows changes in the values. Second attribute which is represented as a rectangle in which vertical lines show the value of that attribute, the more the vertical lines, higher is the value of that attribute. The length and width of the rectangle are constant, just number of lines inside the rectangle may vary. Output attributes are also represented by arrows which are excluded from the rectangle of second attribute. The three arrows are ejected in such a way that one arrow which shows the success factor will move in upward direction while the failure factor will move in downward direction and the medium value which lies in between these two attributes will move in horizontal direction. The width of the arrows may vary due to change in values. All of the attributes have assigned different colors to make attributes distinguishable. For example: green color is selected for success factor which shows clearly the positivity in the data, while red color shows the failure factor as red shows negativity.

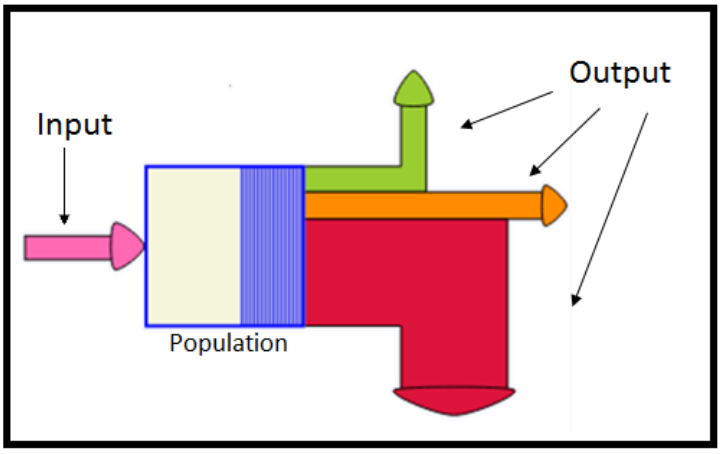


Figure 3.1: Analysis of Data through Pipeline

3.3 Time Series Data

The dataset contains data of different countries to visualize. The data is also based on some time period that is the time series data which contains time range to visualize data of that specific time. This will be helpful for analysis of data to get change in patterns if it occurs. If there is no change in patterns it means there is no progress in that attribute and that attribute requires a lot of improvement. Secondly, the attributes of data comprises of success and failure, analysis of data will improve these factors that what are the basic constraints which lead towards the success or failure.

Time series bar is shown in Figure 3.2 which helps to see the visualization of specific time period by moving the cursor towards the right side. The data over the past few years helps us to conclude new results. The visualization is done on the map using the time series data so the results from the last few years can be clearly seen. This visualization is not static because every time when the cursor is moved in forward direction a new image of that year will be visible which makes this visualization dynamic.

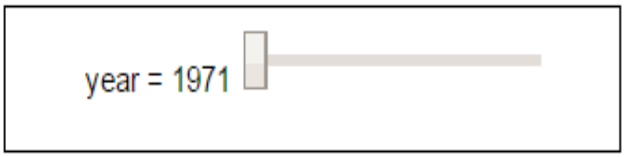
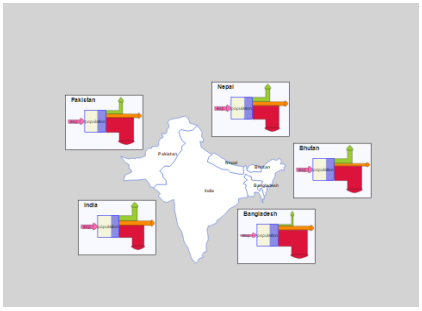


Figure 3.2: Time Series Bar

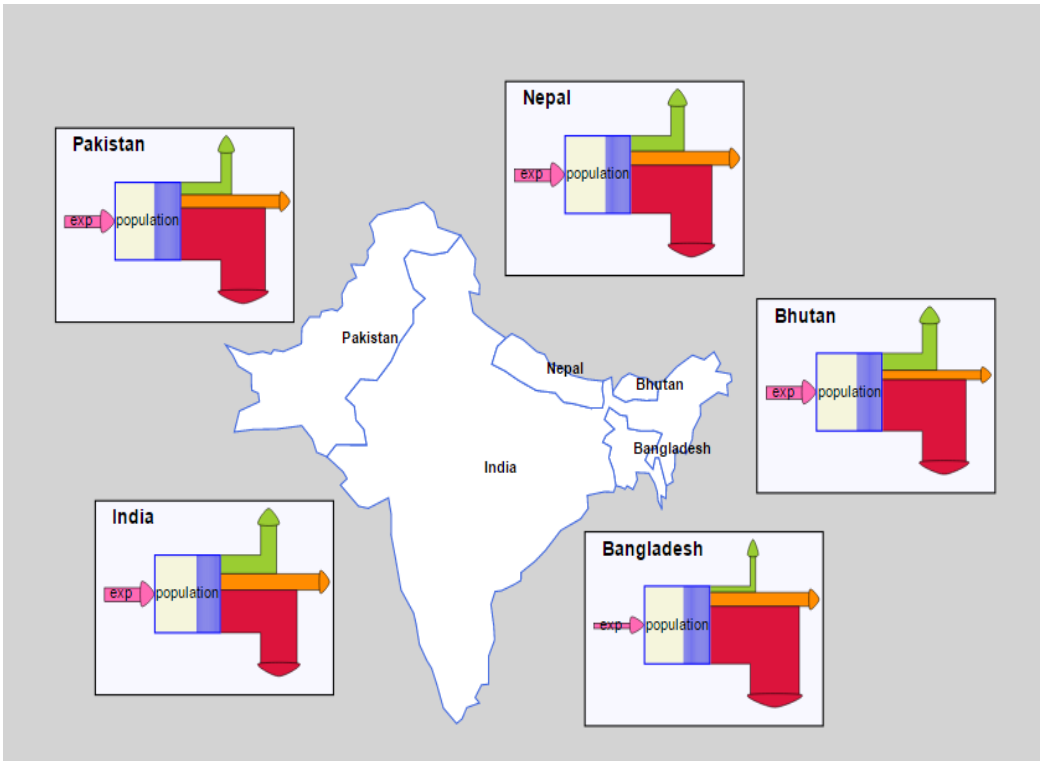
3.4 Zoom in/out

A visual representation of any data such as photograph or map is easier to use as compared to textual description that is why a picture is worth a thousand words. Data can be conveyed easily when it is described visually instead of by a spoken report or textually.

When the data is visually displayed it shows the overview of whole collection. The entire collection of each data type is viewed through zoomed out function. Zoom in function is used on items which are of some interest. If user is interested in any portion to see the minor details zoom in helps to make that area clearly visible. Visualization does not mean representation of just an image; user can also see image details by zoom in, as shown in Figure 3.3.



(a)



(b)

Figure 3.3: (a) shows the enabling of zoom out function, (b) shows the enabling of zoom in functionality.

3.5 Visual Representation Using Different Datasets

Visualization can be done when data is available. Lets consider an example of visualization with different datasets to understand that how the data is visualized using this approach and how the animation is performing its task and how the comparison of two countries can be done.

One dataset is based on educational data in which educational data was focused to get insight into the data. The education system all over the world is different. Literacy rate is somewhere increasing while somewhere is decreasing. What are the main causes, how can we find such constraints which help us to make improvements in education. Analysis of data helps us to gain usable information.

The attributes which are used to fit in this approach are multivariate variables which include:

- Government Expenditure on Education (% of GDP)
- Total Population of Age 0-14
- Primary Completion Rate (%)
- Dropout Rate (%)
- Out of School Children (%)

Other dataset is based on Energy sector which shows the total energy production by different ways. The attributes which are used in this dataset may include:

- Government Expenditure on Energy (in Million Dollars)
- Total Energy Production (in Billion BTU)
- Renewable Energy Production
- Nuclear Power Production
- Fossil Fuel Production
- Total Population

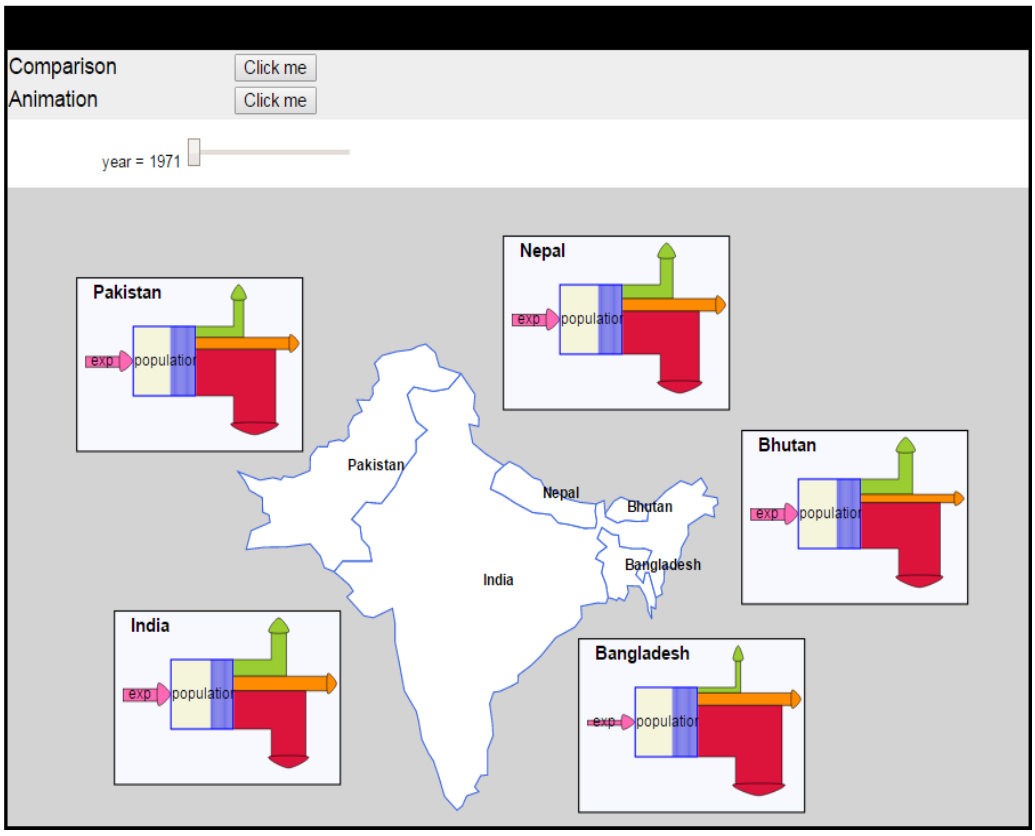
These two datasets are visualized using the pipeline approach. The visualization, animation and comparison will be shown by using these datasets.

3.5.1 Visualization

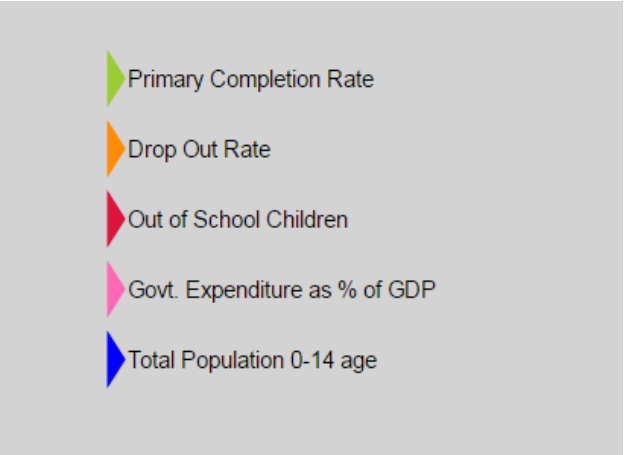
The visualization using the educational data in which government expenditure is acted as an input that how much government spent on education sector so the students can avail opportunities related to education. In every country the government expenses are different on education sector; some governments spent more to make their education system better as they can.

Primary completion rate, dropout rate and out of school children rate are acted as an output values because when children enter into education system they have to complete their education or some of them have to leave due to different reasons, while most of the students are even unable to take admission in the school. These attributes have some relation which is divided into input and output stream and this data is visualized to gather information which will be helpful to improve the education system so maximum children at least complete their primary education.

The visualization is performed on some specific countries to visualize attributes of education of these countries. The data contains attributes values from year 1971 to 2015 to see changes in that time period. The visualization using educational data is shown in Figure 3.4 (a).



(a) Pipeline flow of Educational Data (1971 to 2015)
Available at: http://muneeba.comxa.com/edu/update6.html



(b) Representation of Attributes through Colors

Figure 3.4

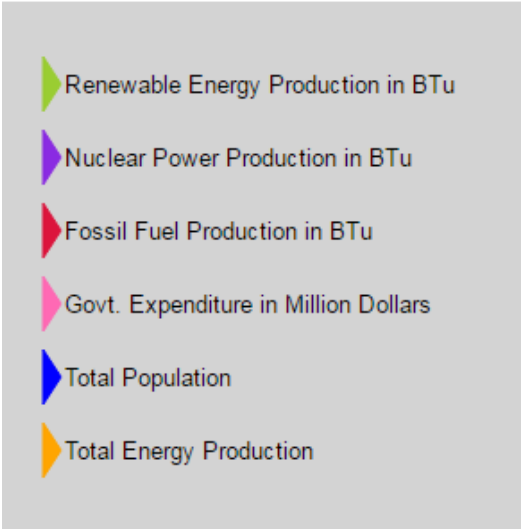
Some important points that depicting this visualization are:

- The map is placed which shows the countries whose data is going to be visualized.
- The values of attributes are changed in respect to change in size in vertical direction.
- The time bar is present on top, by moving the cursor in forward direction the visualization of that year will be shown accordingly.
- Input variable is represented by an arrow which is moving in horizontal direction showing the government expenditure as % of GDP. The width of the arrow in vertical direction may vary because minimum width shows less % of GDP whereas increase in width shows more % of GDP.
- The blue lines inside the rectangle are showing total population (up to age 0 to 14) of that country. When population is increased, there will be more vertical lines then in this box.
- Three arrows are ejected from this rectangle moving in different directions. Those who have completed their primary education moves in upward direction because this shows the success that people are getting educated and this arrow is filled with green color. The width of the arrow in vertical direction shows the percentage of students who have completed their primary education.

- Those who did not complete their education and left during their studies are represented through horizontal arrow and it is filled with orange. The width of the arrow in vertical direction shows the percentage of those students.
- The percentage of children who have never enrolled in school are represented by an arrow which is moving in downward direction and is filled with red shows the failure factor in education.
- All attributes are represented through different colors, can be shown in Figure 3.4 (b). By understanding the color of that attribute user can easily get the results by looking only to that color of the attribute which he wants to gather information.

Similarly, visualization of energy dataset can be seen in Figure 3.5 (a) and each attributes is represented by some color can be seen in Figure 3.5 (b).

(a) Pipeline Flow of Energy Dataset (1970 to 2013)
Available at: http://muneeba.comxa.com/energy/usmap.html



(b) Representation of Attributes through Colors

Figure 3.5

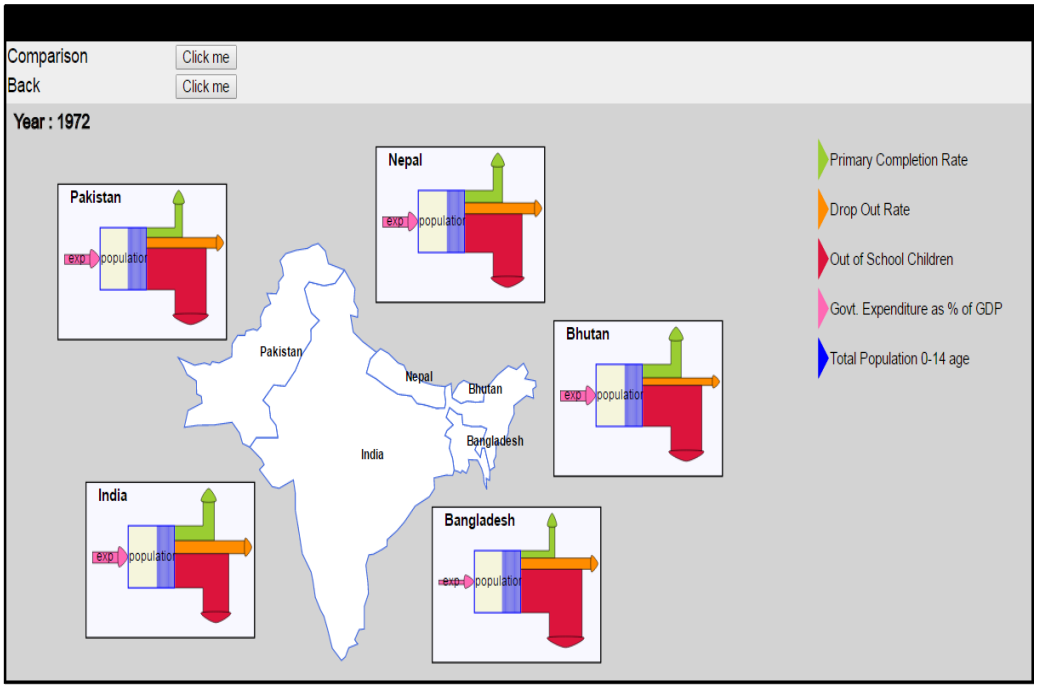
- Energy dataset is visualized of prominent US states which shows production of energy by different ways.
- Input arrow shows the government expenditure on energy sector and it is filled with pink color. Blue lines inside the rectangular box shows the population of that state, the least populated state is Alaska and the most populated state is California in the year 1970.
- Vertical bar outside the box shows the total energy production and it is divided into three kinds of production. The one which is moving in upward direction and is filled with green is the renewable energy production and in 1970 Washington is producing most of the energy from renewable energy. Fossil fuel energy production is represented by an arrow which is moving in downward direction and is filled with red. California is producing most of the energy from fossil fuel in 1970 as shown in Figure 3.5 (a). The arrow which is moving in a horizontal direction shows the nuclear power energy production. The cursor on the top can be moved in forward direction to see the visualization of next years.

3.5.2 Animations

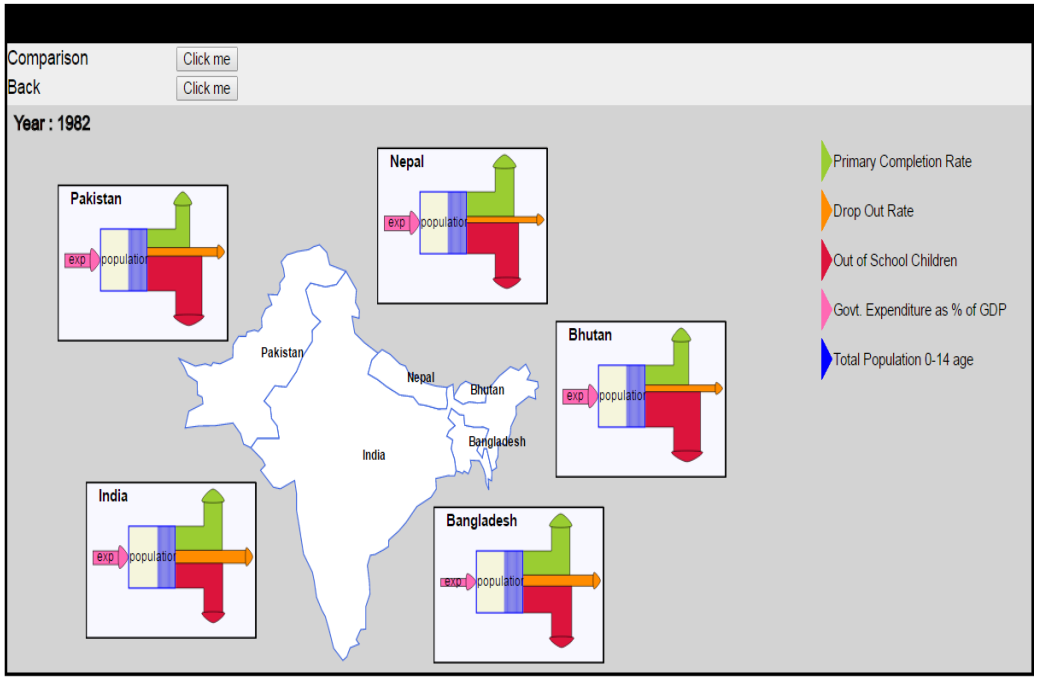
Animation plays an important role in data visualization to make data more attractive or interactive. For the time series data animation is applied on

cursor bar to run the time series automatically. As the data which has to be visualized is time series data, that data can be seen by moving the cursor bar towards right side which changes the time period through this bar manually. Instead of using this bar there is an option available of animation which shows the change in data automatically instead of moving bar in any direction. By clicking on this option the data over the time period is displayed automatically with the interval of specified milliseconds. The visualization of every year is shown automatically with the time interval of specified milliseconds so the user can see the image in that time duration. This animation helps the user to focus on just data and find out the new information instead of moving the cursor back and forth.

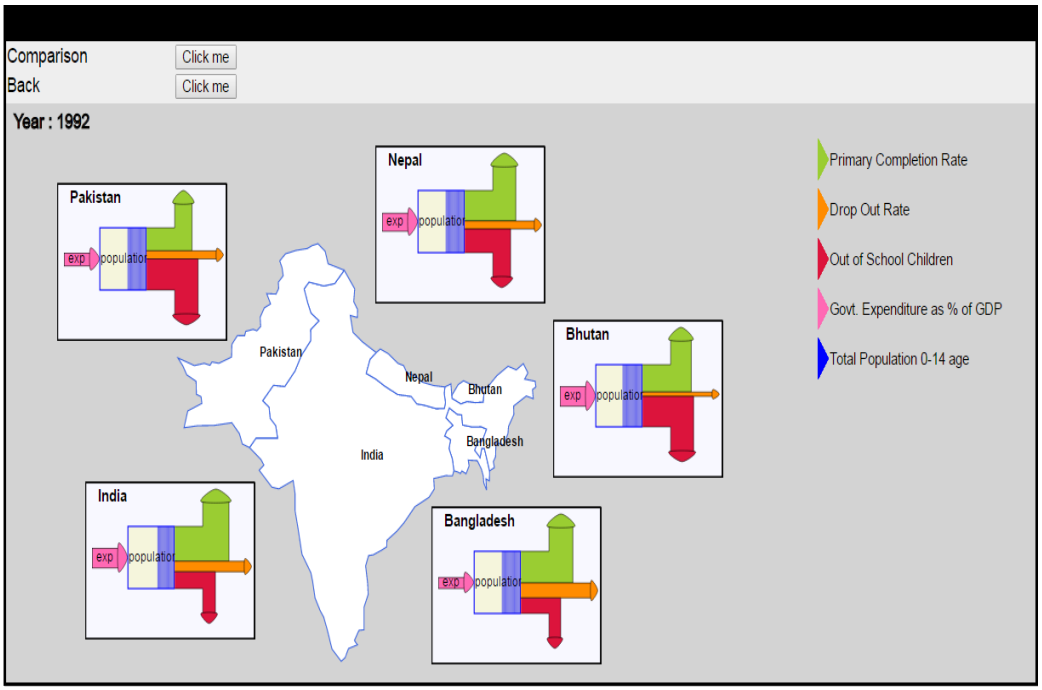
Visualization of educational data through animation can be seen by clicking on the animation button and animation will be played automatically. In Figure 3.6, animation images are displayed of some years that show the automatic change in time.


Comparison
Click me
Back
Click me
Year : 1972
Pakistan
Nepal
Bhutan
India
Bangladesh
exp
population
Pakistan
Nepal
Bhutan
India
Bangladesh
Primary Completion Rate
Drop Out Rate
Out of School Children
Govt. Expenditure as % of GDP
Total Population 0-14 age

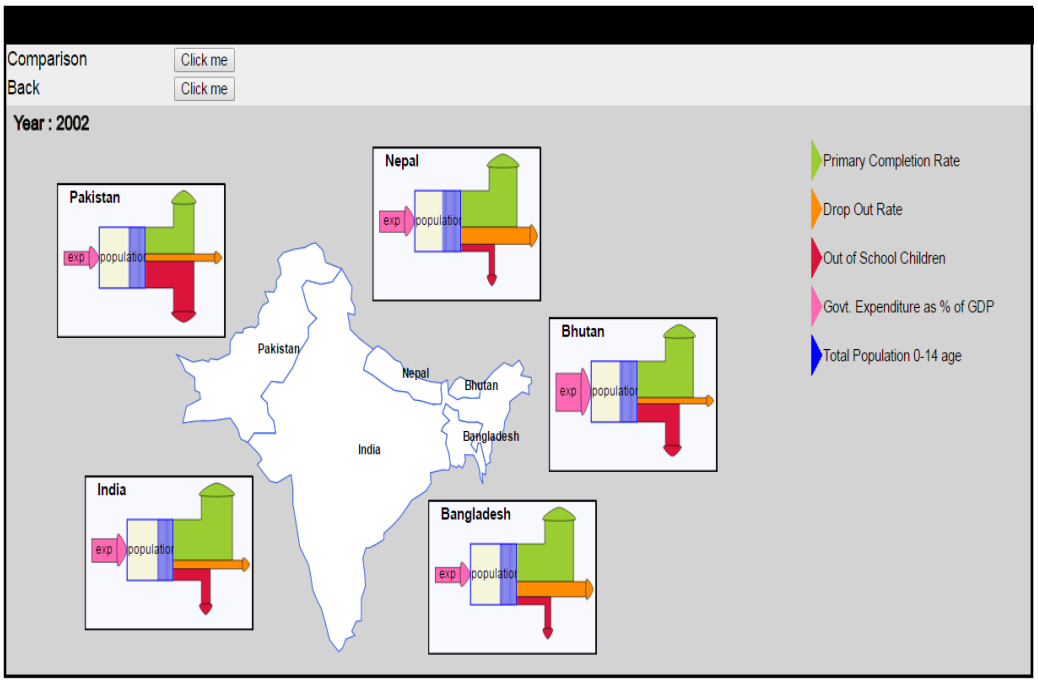
(a)


Comparison
Click me
Back
Click me
Year : 1982
Pakistan
Nepal
Bhutan
India
Bangladesh
exp
population
Pakistan
Nepal
Bhutan
India
Bangladesh
Primary Completion Rate
Drop Out Rate
Out of School Children
Govt. Expenditure as % of GDP
Total Population 0-14 age

(b)



(c)



(d)

Figure 3.6: Automatic Representation of Data by Years, Image of Year (a) 1972; (b) 1982; (c) 1992; (d) 2002

3.5.3 Comparison

The data of different countries can be visualized at the same time but if user wants to make comparison between two countries, to see the differences in values in between these two countries the comparison option is available. Comparison can be done by placing the images side by side or by using checkerboard technique (Stokking et al., 2003), in which one image is represented in blocks of one color while the other image is represented in the blocks of second color. Such visualization in which two images are used namely SPECT and MR, can be seen in Figure 3.7. There can be different techniques to make comparison of two images. But in this methodology the comparison can be done using the mirror images. For example: if a straight horizontal line is placed on right side its mirror image will be in its opposite direction that is on the left side. The two images are placed in such a way so the comparison can be done easily because visualization is done using the pipeline and height of pipes may vary due to difference in the values.

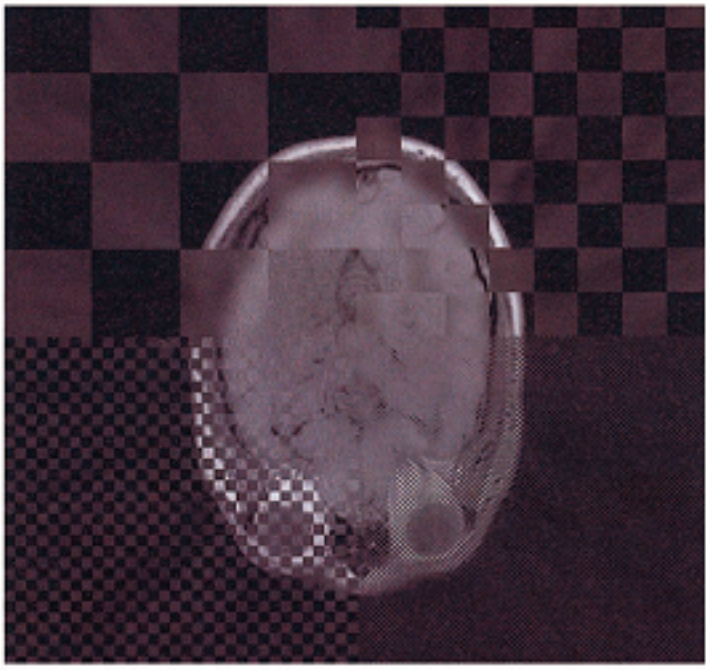


Figure 3.7: SPECT and MR images displayed in a checkerboard pattern (Stokking et al., 2003)

For this purpose user will first select two countries, which he wants to see the comparison between countries. The countries can be selected by clicking on area of that country; the map is placed on one side through which

countries can be selected. When user clicks the first option the image data of that country is placed on right side, when second option is clicked the mirror image is placed on the left side of that area. The visualization of only those two countries is available instead of all countries, and then he has to start the animation which shows the changes in data in between these two countries. The comparison is shown in such a way that one country has the real image which is placed on right side while the second country has the mirror image and is placed on left side of the screen. User can also focus the data by using zoom in option. The interface of selecting two countries for comparison is shown in Figure 3.8, when user clicks on area of that country the image will be visible in white grid portion. There is also an option provided to enable the values on each pipe which shows the textual representation as well, which will be more affecting in understanding the comparison of values.

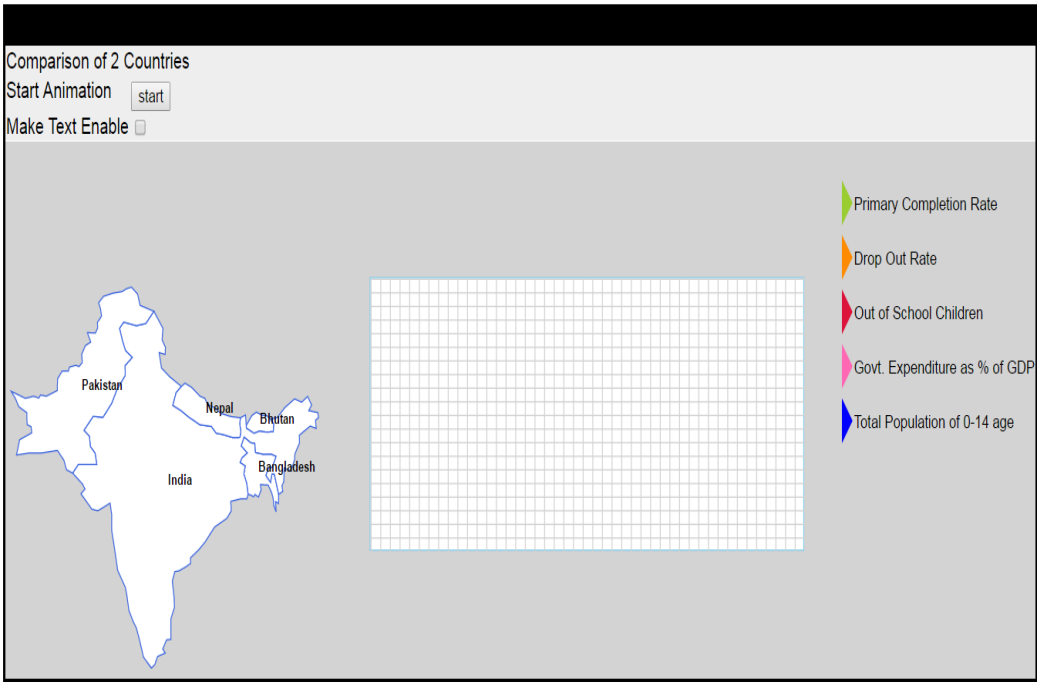


Figure 3.8: Interface to Show Comparison between 2 Images

If user wants to see comparison between two countries, to see changes in patterns in between two countries he will select comparison option where he will select two countries. The two countries can be selected by clicking on the area of countries one by one. The image of those two countries will be placed on screen which shows that you have selected these two countries.

The example of comparison between Pakistan and India in educational field, seen in Figure 3.9, these two countries are selected for comparison.

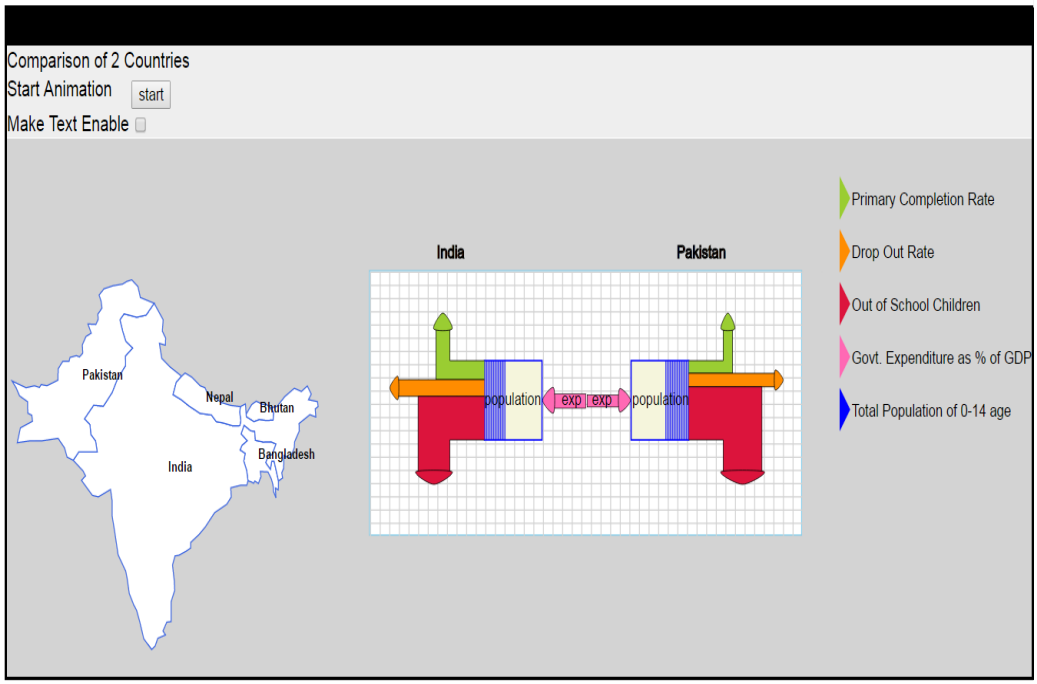


Figure 3.9: Comparison between Pakistan and India (without playing animation)

In Figure 3.9 user has selected Pakistan and India by clicking on the area of that country, the visual representation of these two countries are shown on grid. The two options are available on top, user can start the animation by clicking on start button or he can check the checkbox to make text enable on these pipes. When user will start the animation button the changes in patterns will be seen that whose expenditure is going higher or whose enrolment ratio is more than the other.

So, the user will start the animation by clicking on the animation button, the data of that specific time will start to display over that period of time. If user wants to stop that animation he will press the stop button whenever he wants to stop. During the animation if user wants to see the original values that what percent of value pipe is representing, this can be done by clicking on the checkbox which enables the text on pipe. If again user does not want to see that text further he can uncheck the checkbox. Visualization of data without enabled text can be seen in Figure 3.10.

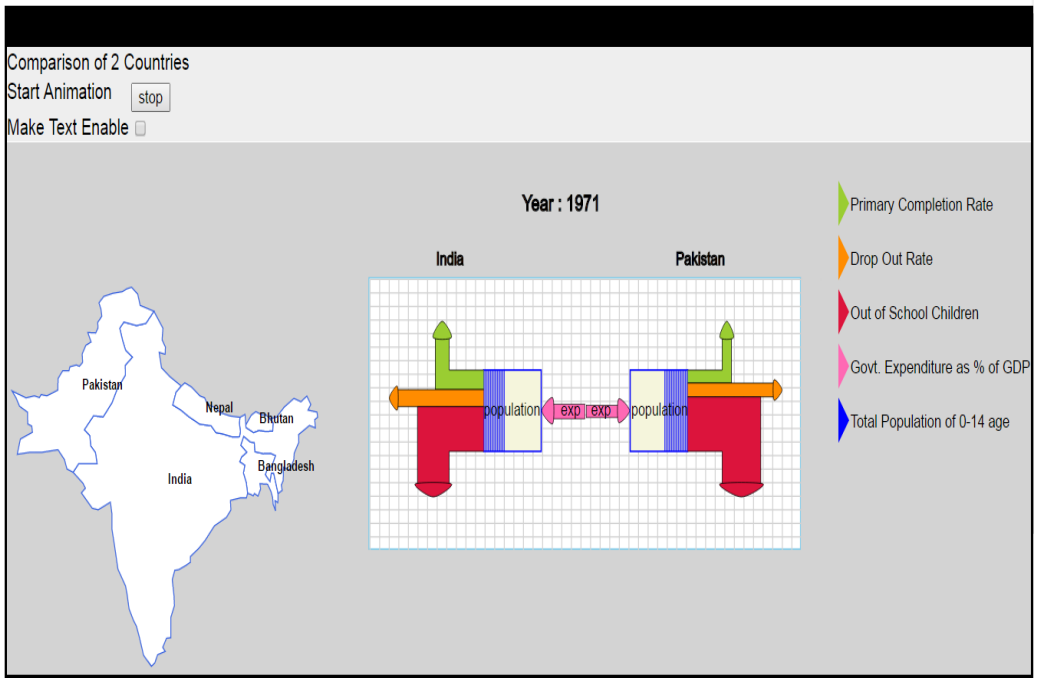


Figure 3.10: Comparison between Pakistan and India (without enable text)

The Figure 3.10 shows the comparison between Pakistan and India of the year 1971 in educational field. In 1971 India spent a little more than Pakistan on Education sector and as a result more children have completed their primary education in India as seen in upward direction arrows. Whereas, the ratio of students who have never enrolled in school is higher in Pakistan as seen in downward direction arrows. Those children who have enrolled in school but they left the school due to some reasons, this ratio is more in India as compared to Pakistan.

As the animation is playing the data will be displayed accordingly of that year. If user wants to stop that animation he will click on stop button and animation will be paused. If user wants to make text enable, he will check the checkbox, text will be displayed on each pipe can be seen in Figure 3.11.

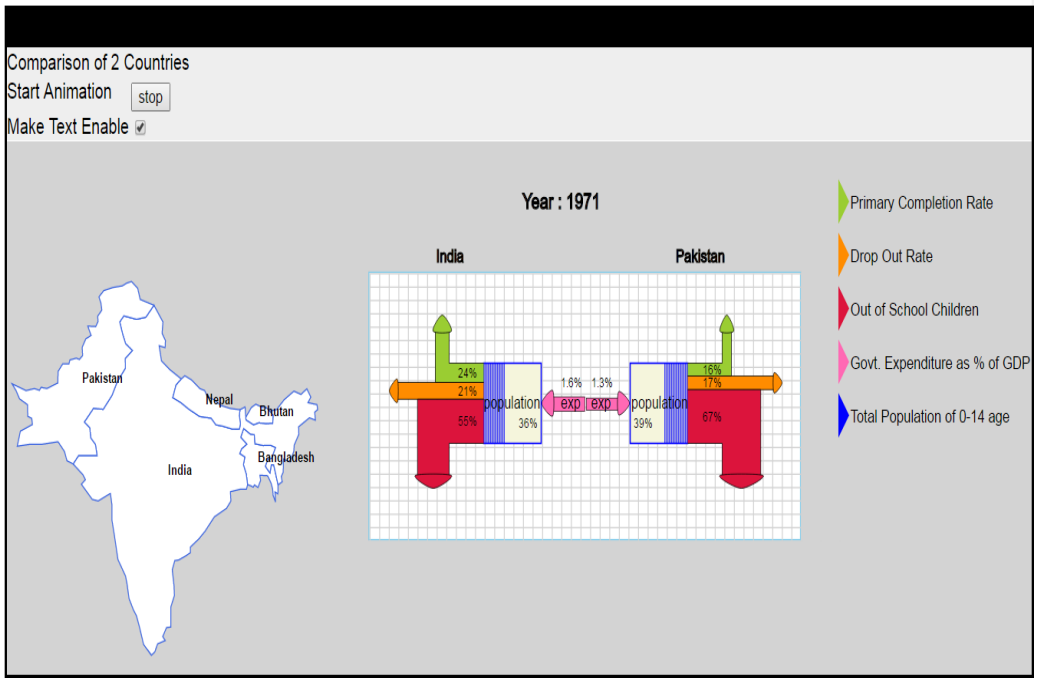


Figure 3.11: Comparison between Pakistan and India (with enable text)

This chapter discussed the visualization technique that how data can be represented if data is divided into input and output streams. This visualization is done through pipeline flow in which attributes are represented using arrows. The arrow which moves in the upward direction shows the success factor of that scenario while the arrow moving in downward direction shows the failure factor. The visualization is done using the parameter size (all attributes may vary in size depending on the value of that attribute). Whereas, all attributes have assigned different color to make distinguishable with other attributes. User can perform zoom in and zoom out functionalities, he can animate the time series data and he can also do comparison in between two countries/states to make the analysis of data or to gather new information.

Chapter 4

Results

The aim of this thesis is to make a visualization technique which shows multivariate data in such a way that data is divided into input and output streams. Data of different countries can be visualized at the same time to make comparison in between different countries that where the success factor is more and where the failure factor is more. This chapter will discuss the comparison of proposed solution to the motion charts to see the level of users satisfaction.

4.1 Usability Study

Usability is defined as how it is easy to use something, to understand, to learn and to operate. To check the efficiency, satisfaction and ease of use, this study is conducted. Efficiency of any system is measured by time completion that how much user took time to complete the task as well as how many attempts user has taken to complete this task. Satisfaction is measured by reaction of users to the given solution. These reactions are taken from any feedback questionnaire. As, the usability study has a lot of dimensions, so all the goals cannot be achieved at the same time. When data is collected some statistical tests will be performed to analyze the data. In usability study our proposed solution will be compared to motion charts (Battista and Cheng, 2011).

4.1.1 Motion Charts

Motion charts are used to visualize multivariate data, also shows dynamic visualization. It consists of bubbles like in bubble chart which permits interactive exploration. In motion chart variables are mapped into 2D coordinate axis, color and size; and time is mapped on time slider which alleviates interactive representation of temporal and multidimensional data.

In motion charts four attributes will be mapped in such a way that one attribute is on x-axis, second attribute is on y-axis, and third attribute is represented by color whereas the fourth attribute is represented by size. There is also a time slider which helps to change time span by moving the cursor towards right side. The changes can be seen on display by moving the cursor.

The dataset file of educational data is passed to this chart and four attributes are set in such a way that population variable is lying on y-axis, total expenditure is lying on x-axis, primary completion will be represented using color and dropout rates will be represented by size. By moving the slider different changes can be seen in this chart in such a way that bubbles are moving according to their values, some bubbles are moving towards right side which shows the maximum value of expenditure attribute while those which are on upward direction shows the maximum population value. The size of the bubbles may also vary shows the increase or decrease in percentage values of that attribute which is represented through size. Color scheme is represented in such a way as more the color moves in upward direction shows the maximum value of that attribute. Using these four parameters (x-axis, y-axis, color, size) variables are represented in motion charts to make the data clearly accessible and usable to all the users so users may conclude their results and interpret new information.

4.1.2 Method

Pipeline flow technique is compared with motion charts in such a way that participants were given some tasks which they have to perform on both interfaces. They have to find out the answers of those questions, in the mean while their time was calculated that how much time they took to answer that question as well as how many clicks they have used in finding the answer. Further some questionnaires were given to the participants after the completion of the tasks. They have to rate both interfaces according to their flexibility. People having different backgrounds were selected to take part in this study, numbers of participants were 30. Visualization have been shown using the educational dataset as well as energy dataset using the pipeline flow as well as motion charts and participants have to ask to perform the tasks in which they have to find answers of some questions in both interfaces. The tasks which were given to participants are:

- In 1990 which country has highest primary completion rate?
- In 1997 which state has spent more on energy and which state is producing more through renewable energy?

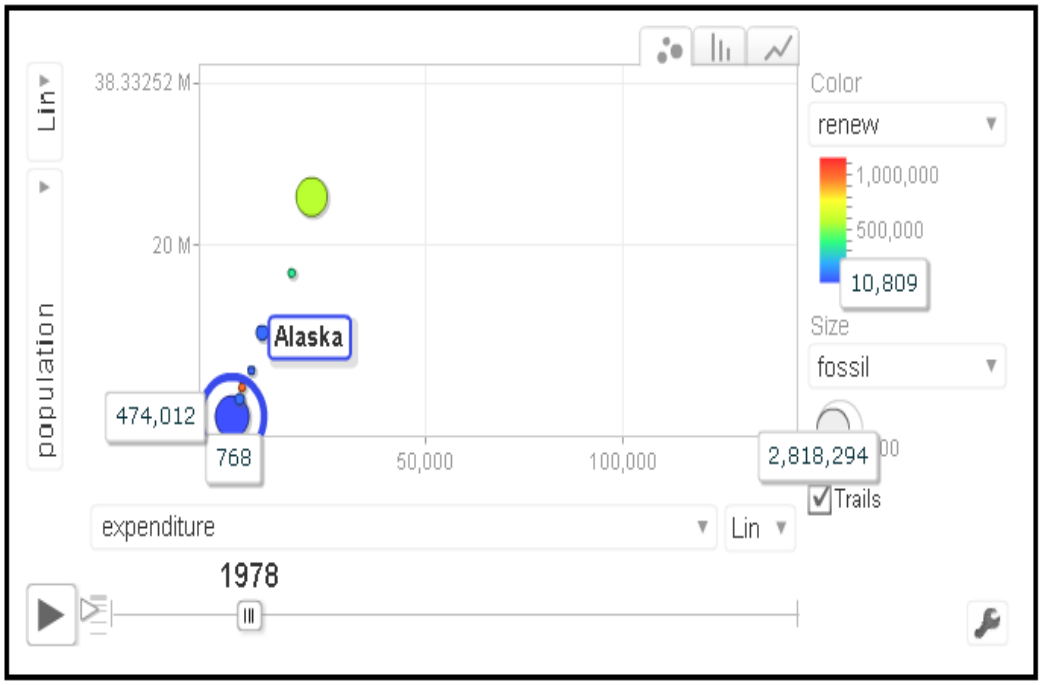
- In 1978 which state was producing more energy through fossil fuels?
- From 1990 to 2000 which state has decreased its production through fossil fuels?
- Make comparison between Pakistan and India that whose primary completion rate is increasing and whose out of school children rate is decreasing over the time period?

Firstly a demonstration was given to participants about both the interfaces that how to use these interfaces and what describes this term. After the complete description of interfaces tasks were given to the participants which they have to perform on both interfaces; half participants have to perform the task firstly using motion charts after that they have to answer the questions using pipeline flow and vice versa. The task was same for all participants but half participants have to answer first using the motion chart visualization while the remaining participants have to answer first using the pipeline flow visualization. When user gets ready to answer the question time is noted that how much time user has taken to answer that question as well as how may clicks user has used in finding the answer of that question. Time in seconds and number of clicks were counted for all the participants. Afterwards, they were given a survey questionnaire in which they have to ask to rate these both interfaces according to the given questions shown in Table 4.1.

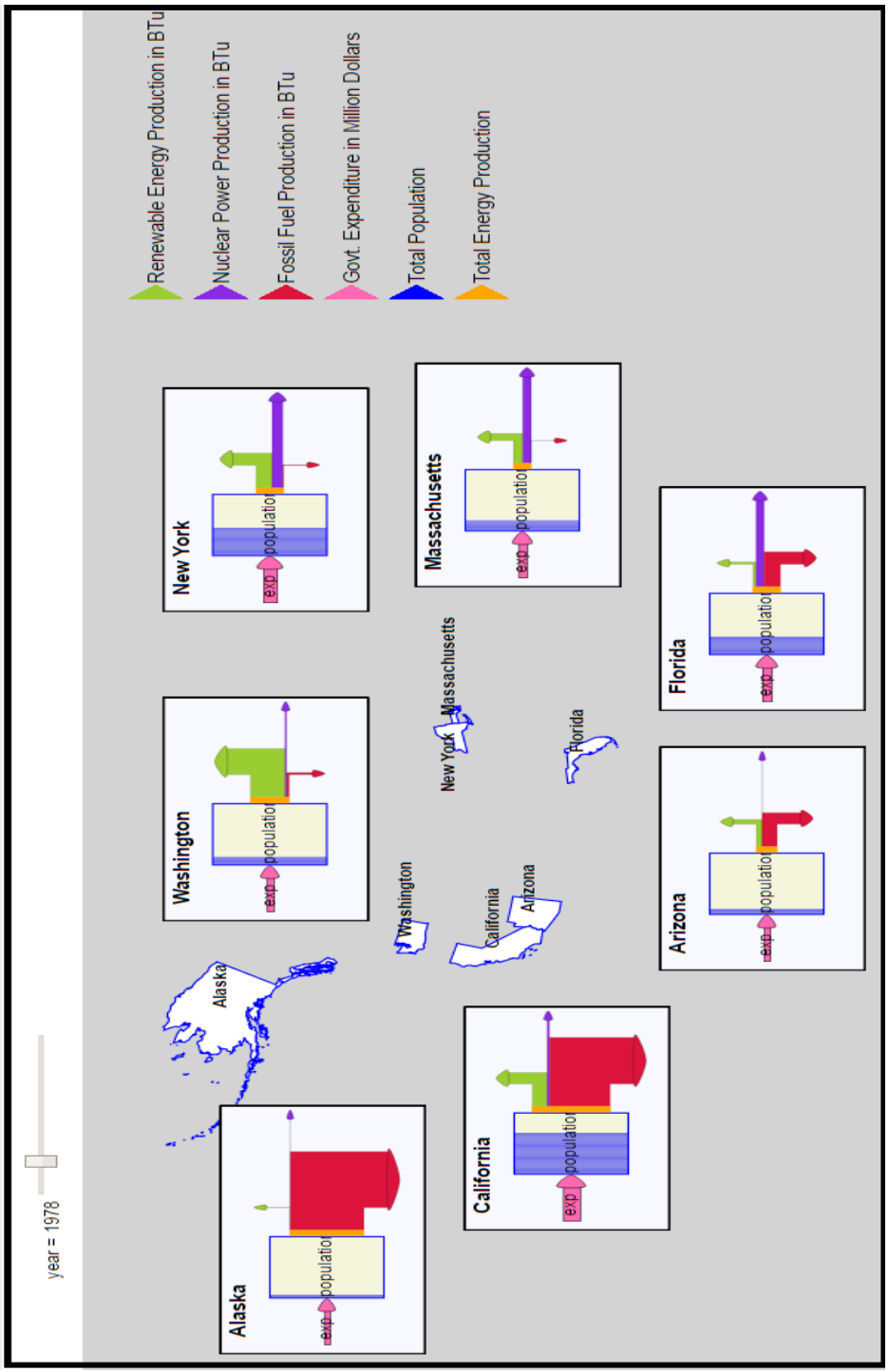
Lets consider a one question which was asked to participants to answer, the visualization results in both interfaces of that question can be seen in Figure 4.1 (a) and (b). The question was In 1978 which state is producing more energy through fossil fuels? The answer was Alaska, this can be shown clearly in pipeline flow but in motion charts size of circle represents that which state is producing more energy and users have to click the circle to get the name of that state.

Questions
1. This interface is easy to use 2. Learning to use this interface was easy 3. Using this interface I can find needed information easily 4. The information provided is easy to understand 5. I feel comfortable using this interface 6. This interface is pleasant 7. I am satisfied with this visualization technique
Ranking Scale: 1(strongly disagreed), 2, 3, 4, 5(strongly agreed) (higher the value better the results)

Table 4.1: Usability Questionnaire



(a) Motion Chart Interface



(b) Pipeline Flow Interface

Figure 4.1

4.2 Statistical Analysis

When large amount of quantitative data is collected, the next step is to perform statistical analysis to make sense of that data. As, statistics is an area of science which deals with assembling and analysis of such data. The important terms of statistics are methods which help to make judgments on that collected data although if there is any doubt and variation (Walpole et al., 2015). There are different methods for statistical analysis which help to analyze data and point out ups and downs which will be helpful to make betterment in that area. Mostly statistics are performed through mean that gives the central value actually. Mean is equivalent to find an average value, that is take the sum of all the given values of that attribute and divide this result to the total number of that values. Mean value can be represented through different symbols like here $\overline{x}$ is used to represent mean value and formula to calculate mean value is

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.1}$$

For the reliable results mean value is not enough for analysis of data in true way because there can be two samples which have similar mean value but in terms of variability they can be different completely. As the mean value is same does not assure that values inside the data sample are same, the data values can be scattered in one sample than other. This dispersion of data can be told by Standard Deviation (Walpole et al., 2015). σ is used to represent Standard Deviation and it can be calculated using the formula

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{4.2}$$

Where n shows number of samples. All the analysis will be performed using this Mean and Standard Deviation, and data results will be represented using histograms or bar charts.

4.3 Results

4.3.1 Efficiency

Participants were given some tasks to perform on two different visualization interfaces. Their time (in seconds) was recorded in which they have performed that task. The histogram of the mean value of total completion time

of all the tasks is shown in Figure 4.2. Whereas, Figure 4.3 shows the mean value of individual task completion time using the bar chart. The Mean and Standard Deviation values of time (in seconds) for all the tasks are shown in Table 4.2.

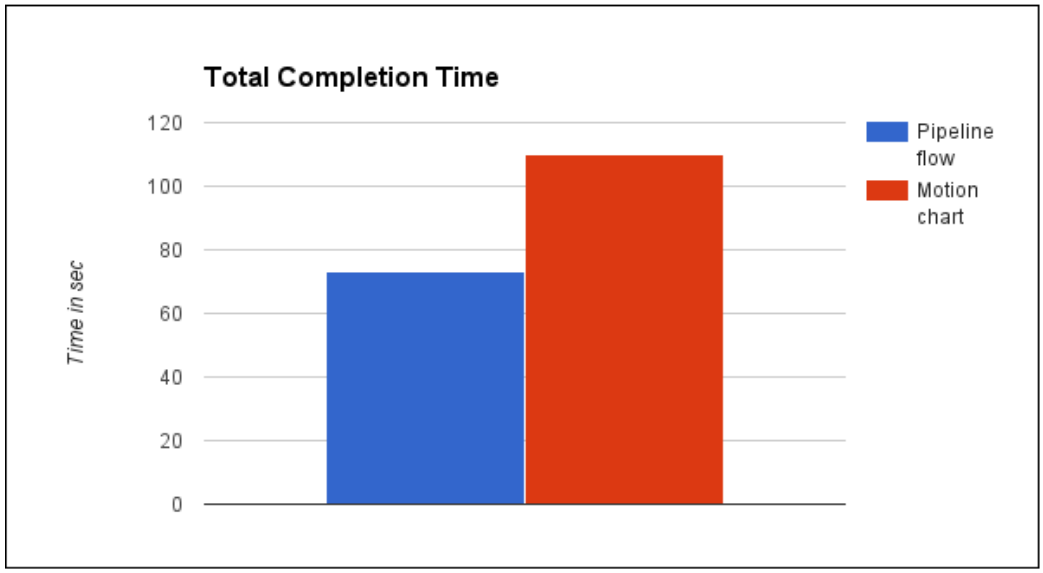


Figure 4.2: Mean Total Completion Time (in seconds)

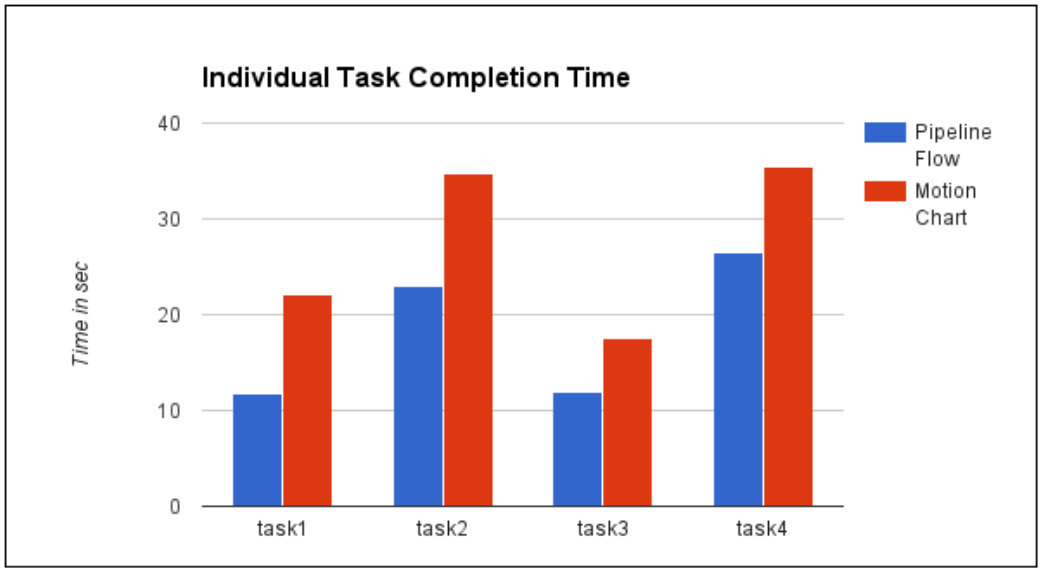


Figure 4.3: Mean of Individual Task Completion Time (in seconds)

Tasks	Motion Chart	Pipeline Flow
Task 1	22.03 (11.94)	11.7 (4.78)
Task 2	34.73 (15.98)	23.03 (8.96)
Task 3	17.63 (7.79)	11.97 (3.46)
Task 4	35.4 (14.24)	26.57 (10.52)
Total Mean Time	109.78 (36.66)	73.27 (18.94)

Table 4.2: Mean (S.D) of Time (in seconds) by Tasks

4.3.2 Number of Clicks

It was also recorded that how many clicks were used to perform each task by participants. Number of clicks was recorded for each task in both interfaces. The mean value of number of clicks is represented using histograms as shown in Figure 4.4.

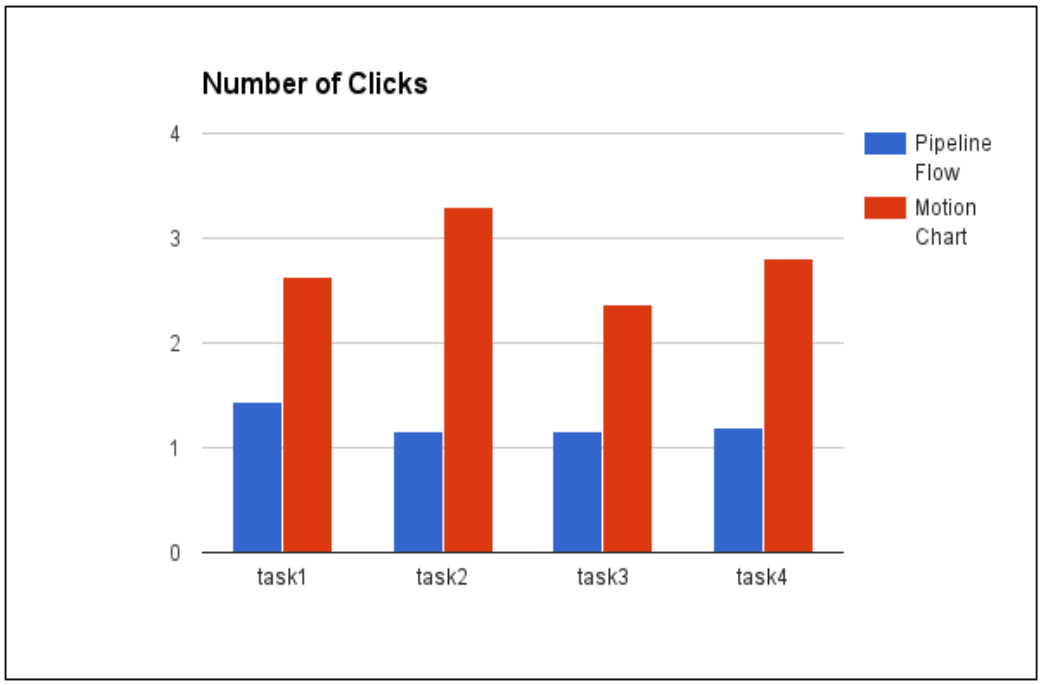


Figure 4.4: Number of Clicks Used in Each Task

4.3.3 Usability Questionnaire

Average score of 7 attributes for both interfaces were calculated to obtain overall usability score. Both interfaces were rated after performing the tasks. The mean value of all the questionnaires is represented using horizontal bars as shown in Figure 4.5.

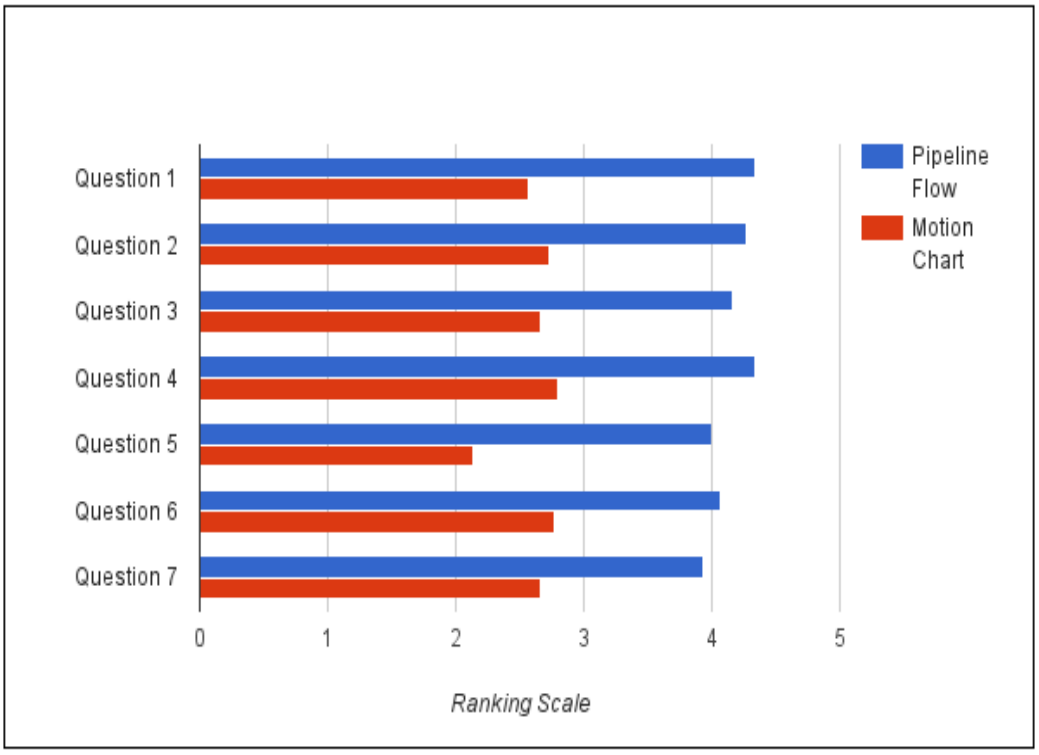


Figure 4.5: Rating of Usability Questionnaire (higher the value better the results)

4.4 Discussion

Visualization based on education as well as energy dataset is represented on both interfaces (motion chart, pipeline flow). The results showed that pipeline flow is much efficient than motion charts as well as participants used less number of clicks using pipeline flow. Pipeline flow has an improved efficiency up to 33 % as compared to motion charts. Usability questionnaire results also in the favor of pipeline flow. Pipeline flow has an improved usability of 59%. Respondents did not find any difficulty level in finding the results

of both education and energy field using pipeline flow, although visualization is same in both datasets but attributes name are different. In motion chart users face difficulty while giving answer to the question because that representation is a combination of multiple factors and people got confused that either which factor they have to follow to answer that question. Some people got confused from the color range because it is difficult to differentiate colors when placed one over the other. The comparison between two countries has to be performed in one question which was performed on pipeline flow interface only and participants show positive response while comparing the two countries. In motion charts we cannot compare two countries only; we have to see overall visualization only.

Visualization presents some of the attributes in education and energy sector because these attributes are based on some scenario in which input and output factors are important. For example in education dataset we have presented attributes: government expenditure, total population up to age 14, primary completion rate, dropout rate, out of school children rate. These attributes play an important role as the government spent money on education and how much of students have completed their primary education, if the primary completion rate is going low we can conclude that this attribute require an improvement, what are the factors that causes this attribute in such condition. As the visualization shows that how much money is actually spent on education sector but it does not show that how much money is actually wasted from that expended money. Similarly, those children who have left the school without completing their primary education, where they have gone or what they are doing either they become labor or something else does not show actually. As primary completion rate is shown but how much of them go for further studies or how many quit their studies after just completing their primary education. So, these are some of the obscured attributes which did not present in the visualization. Similarly some of the obscured attributes in energy dataset are also present.

Chapter 5

Conclusion & Future Work

5.1 Conclusion

The problem of multivariate data visualization based on some scenario was discussed in detail in this thesis. The methodology to represent scenario based data was discussed in detail. Additionally, the approaches that can be used to represent multivariate data along with time series data were also covered in detail.

This thesis suggested a solution to represent multivariate data using a pipeline flow technique. Afterwards this approach was compared with motion charts and usability study was conducted. After usability study we conclude that pipeline flow give better results than the motion charts i.e. it is efficient and utilized less number of clicks. As the pipeline flow is representing a scenario based attributes, all attributes are linked to each other which conclude that if one value is increasing then on the other side which value is decreasing. This helps to improve those parameters which are not performing in a way in which they should. These kinds of scenarios can be easily visualized using pipeline flow approach. Whereas in motion chart every single attribute is represented differently using individual factors like size, color, x-y axis. If only single attribute has to be visualized then motion chart will be helpful, but in motion charts if values are same bubbles may overlap each other which hides one value. Similarly the range of colors in motion charts is also confusing, we have to see again and again that either the color shows increase or decrease in value because darker tone of color represent both maximum and minimum value. If we want to make comparison between two countries to see that which country is performing better, we cannot find out in motion charts. Pipeline flow helps to make comparison in between two countries and new information can be interpreted then.

5.2 Future Work

Pipeline flow technique should be more attractive, interactive or friendly to user. As this technique is represented three outcome values based on some scenario, that is there exists three output values which are handled easily but if there comes more than three output attributes which is also the part of that scenario then how these attributes should be handled. The output arrows for the multiple attributes (more than three) should be emerged in a better way which represents the whole scenario. Similarly, if there exists more than one input variable then what will be the way to represent such situation? As the population is handled using the rectangle box in which vertical lines represent population. This attribute should be handled in another way which describes rate of population of different countries effectively. The proposed technique used the size factor only; which means each attribute value is represented using the change in length of arrow in vertical direction only. The multiple factors can be used to represent multivariate data; it can be a combination of both size and color or color or any other factor.

Bibliography

Aigner, W., Miksch, S., Muller, W., Schumann, H., and Tominski, C. (2007). Visualizing time-oriented data a systematic view. *Journal of Computer and Graphics*, 31(3):401–409.

Andrienko, N., Andrienko, G., and Gatalsky, P. (2003). Exploratory spatio temporal visualization: An analytical review. *Journal of Visual Languages & Computing*, 14(6):503–541.

Battista, V. and Cheng, E. (2011). Motion charts: Telling stories with statistics. In *JSM, Section on Statistical Graphics*.

Brewer, C. A. (2006). Basic mapping principles for visualizing cancer data using geographic information systems (gis). *Journal of Preventive Medicine*, 30(2):25–36.

Chan, W. W. (2006). A survey on multivariate data visualization.

E. Morse, M. L. and Olsen, K. A. (2002). Testing visual information retrieval methodologies case study: Comparative analysis of textual, icon, graphical, and spring displays. *Journal of the American Society for Information Science and Technology*, 53(1):28–40.

Fao, R. and Card, S. K. (1994). The table lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of the SIGCHI Conference on Human Factors in Computer Systems*, pages 318–322.

Fergusoni, S. (1991). The 1753 carte chronographique of jacques barbeudubourg. 52(2):190–230.

Few, S. and Edge, P. (2009). Introduction to geographical data visualization.

Friendly, M. (2008). The golden age of statistical graphics. *Journal of Statistical Science*, 23(4):502–535.

Guo, D., Chen, J., MacEachren, A. M., and Liao, K. (2006). A visualization system for space-time and multivariate patterns. In *IEEE Transactions on Visualization and Computer Graphics*, volume 12, pages 1461–1474.

Hoffman, P. E. (1999). *Table Visualizations: A Formal Model and Its Applications*. dissertation, University of Massachusetts at Lowell.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE*

Transactions on Visualization and Computer Graphics, 8(1):1–8.

Muller, W. and Schumann, H. (2003). Visualization methods for time-dependent data an overview. In *Proceedings of the Winter Simulation Conference 2003*, pages 737–745, New Orleans.

Spence, I. (2006). William playfair and the psychology of graphs. *JSM, Proceedings of the American Statistical Association, Section on Statistical Graphics*, pages 2426–2436.

Spence, R. (2000). Information visualization. page 224.

Stokking, R., Zubal, I. G., and Viergever, M. A. (2003). Display of fused images: Methods, interpretation, and diagnostic improvements. 33(3):219–227.

Tominski, C., Schulze-Wollgast, P., and Schumann, H. (2005). 3d information visualization for time dependent data on maps. In *Proceedings of the Ninth International Conference on Information Visualisation*, volume 5, pages 175–181.

Tominski, C. and Schumann, H. (2004). An event-based approach to visualization. In *Proceedings of the Eighth International Conference on Information Visualisation*, pages 101–107.

Tufte, E. (1997). Visual explanations 1997. 200:158.

Tufte, E. R. and Graves-Morris, P. (1983). The visual display of quantitative information. 2.

Walpole, E., Myers, R. H., Myers, S. L., and Ye, K. (2015). *Probability and statistics for engineers and scientists*, volume 5. Cengage Learning.

Weber, M., Alexa, M., and Muller, W. (2001). Visualizing time-series on spirals. In *Proceedings of the IEEE Symposium on Information Visualization 2001*, pages 7–13.

Yin, H. (2002). Visom a novel method for multivariate data projection and structure visualization. In *IEEE Transactions on Neural Networks*, volume 13, pages 237–243.

Zhu, N. Q. (2013). *Data Visualization with D3.js Cookbook*. Packt Publishing Ltd., Birmingham, UK.