Data Analysis with Visualization for a Geographic Information System of Schistosomiasis Community Health Data

J. M. D. Acibar, L.N. Aguanta, J. D. E. Gomora, and †L.C. P. Velasco

Mindanao State University-Iligan Institute of Technology,
Iligan City, 9200, Philippines
†lemuelclark.velasco@g.msuiit.edu.ph
www.msuiit.edu.ph

Schistosomiasis is a disease common in geographic areas with bodies of water infested with the larval forms of parasitic blood flukes. Altair Geographic Information System (GIS) was developed to monitor the schistosomiasis cases in Salvador, Lanao del Norte. Although the system provide appropriate map visualizations, the system does not offer data analysis and visual analysis of data which would be vital to the understanding of relationships between survey variables of the GIS. This research developed a data analysis and visualization model by studying the existing features and data of Altair GIS. Developed data analysis model was then integrated to the visualizations of the GIS survey data. Google Charts Visualization API was used in producing charts for the determined statistical tools. The data analysis with visualization module helped the MHO of Salvador in making appropriate interventions and action plans to address the schistosomiasis problem of their locality.

Keywords: Geographic Information System, Data Visualization, Schistosomiasis

1. Introduction

Schistosomiasis is a diseases endemic in tropical and sub-tropical areas in which bodies of water are infested with schistosomes having freshwater snails as host that can infect the urinary tract and intestines of the people in the community [1, 2, 3]. A certain geographic information system (GIS) called Altair GIS was developed to monitor patients and visualize community health survey data gathered from four endemic barangays of Salvador, Lanao del Norte, Philipiines. The database of the GIS was designed based on the community health data provided by the Municipal Health Office (MHO) of Salvador. The integration of user interface and the database model allowed the management of the gathered data and the mapping model utilized the

coordinates of households from the database, visualizing community health data through maps using Google Maps API.

Making sense of the data is an important purpose of data management and data visualization, especially in the field of health. Public health workers monitor the health of a community by collecting and analyzing data to detect new health threats, plan public health programs, and evaluate their success. Descriptive statistical measurements and inferential statistics are used in medical literature to manage, monitor, and evaluate health services and make inferences about the sample data [4, 5]. Communicating and analyzing information through visual representations have their own advantages as they are easier for people to grasp meaning, interpret data, and share ideas easily. These are powerful means to discover relations and apprehend things which could not be seen right away. Data analysis tools connected to the database should be accessible to the user while an ideal visualization should not only communicate clearly, but stimulate viewer engagement and attention through graphs, maps, tables, and even text. Well-crafted analyses and visualizations help uncover trends, realize insights, explore sources, tell stories and communicate information clearly and efficiently to support analysis, reasoning and decision-making [4, 6, 7, 8]. The Altair GIS enabled the management of schistosomiasis community health data and provided map visualizations which assist the monitoring of households that need intervention. However, the system does not offer analysis of data which would be vital in determining the factors that affect the compliance of respondents with the prescribed schistosomiasis medications. This study aimed to integrate applicable data analysis methods and data visualization tools which would aid MHO in increasing the rate of schistosomiasis medicine intake in the community. The objective of this study is to develop a data analysis with data visualization module that can analyze and visualize the community health data. Specific objectives include data preparation by determining community health data variables for analysis and visualization. A chart models using Google Visualizations API was also integrated to the database along with the creation of descriptive and inferential visualizations for the community health data.

With the existing schistosomiasis community health data and system features of Altair GIS, this study attempted to enhance the system by integrating available tools to produce meaningful analyses and visualizations which will further contribute to the strategies and solutions of the MHO in improving the compliance rate of schistosomiasis prescribed medications in Salvador, Lanao del Norte, Philippines.

2. Methodology

2.1. *Data Preparation*

Altair GIS has an existing community health data that was utilized in this study as shown in Figure 1. The data went through thorough preparation to be able to return reliable analyses and visualizations. With the user requirements given by the MHO of Salvador, the researchers were able to identify the needed analysis and visualization features to be implemented to improve the system. After the requirements analysis phase, the database was redesigned and structured based on the needs for the data analysis and visualization features. New entities, attributes, and relationships among them were added. The next phase concerns with cleaning, verifying, and transforming the schistosomiasis community health data. This means that all but the data of interest were removed. Lastly, the cleaned data was used in the data classification phase. The data was organized based on its variable type, whether it belongs to descriptive or inferential statistics, for data analysis and visualization.

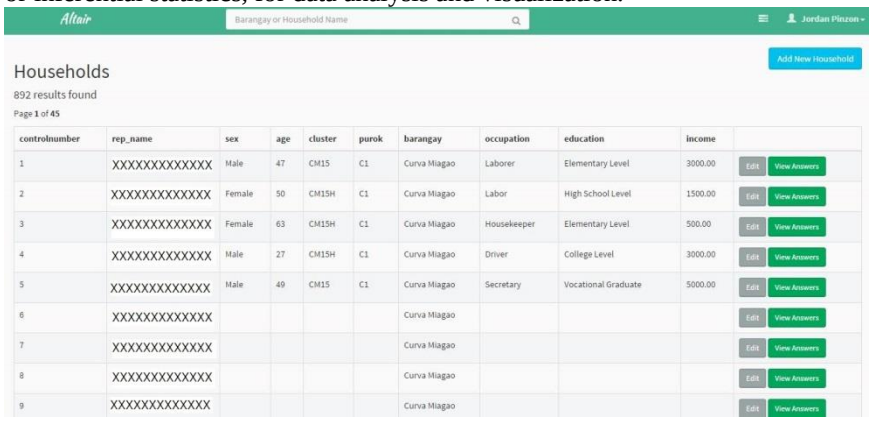

Fig. 1. The schistosomiasis community health of Altair GIS.

2.2. *Data Analysis Model*

There are many different types of correlation coefficients that reflect association between factors which include Pearson, Kendall, Spearman, and Point-biserial correlation coefficients, along with the test for independence between variables, the Chi-squared test for independence [9]. The data analysis model design starts with the determination of applicable statistical methods. The MHO of Salvador provided analysis requirements and possible statistical methods to be integrated in Altair GIS. The determination of applicable statistical methods involved finding supporting claims from related literatures and consultation of a statistician. The next phase is the design of data analysis model which consists

of the identified components and the processes between these components. It focuses on the integration of the data and determined statistical methods. The determined statistical methods are the backbone of the data model and were written in PHP scripts.

2.3. *Data Visualization Design*

The design of the data visualization involved determination of the data to be visualized, specification of visualization outputs, and implementation using scripts. The first phase of the methodology is the determination of the data to be visualized. These data are determined based on the requirements of the MHO of Salvador. The next phase focuses on the specification of visualization outputs. This includes the analysis and specification chart types to be used, designing the final outputs for descriptive reports, and transforming data analysis results into something comprehensible by the user. Once the visualization outputs were designed, client side script such as JavaScript and server side script such as PHP were used for implementation. Inferential results such as data correlation and independence made use of the data analysis model while descriptive results such as percentages of male/female, age groups, income and educational level per cluster, purok, barangay or municipality made use the charting model.

3. Results and Discussion

3.1. *Data Preparation Results*

For a better analysis and visualization, data must be clean and raw as possible so that they can be processed by any software [10]. Thus, determination of Schistosomiasis community health data variables is important before utilizing them for analysis and visualization processes. The goal of the user requirements is to check the relationship between medicine intake and other variables within the schistosomiasis community health data. New system specifications were provided but no new batches of data were given for analysis. To address this issue, the researchers need to generate data in preparation for incoming batches of respondent answers. The generation of data is for simulation purposes only and does not intend to alter the actual results of the schistosomiasis survey conducted by the MHO. There are numerous tools that could generate random data suitable for databases. Of these, the researchers used Mockaroo, an online tool for generating random data in generating of respondent answers for incoming batches.

The entity relationship diagram of the existing system, as shown in Figure 2, is composed of seven entities from doctor, staff, household, answer, choice, question, and group. The entity relationship catered to the needs of the existing system but with the additional features added to the system, additional entities

were created to the ERD of Altair GIS. From the requirements analysis, the researchers have come up with a database design. In order to avoid data update, insertion, and deletion anomalies in the design, normalization was observed throughout the database design process. Third Normal Form was practiced as the minimum requirement of normalization. It was applied to the database to achieve atomicity of data. The existing database of Altair GIS was assessed and redesigned to cater to the needs of the new system functionalities. In comparison, as shown in Figure 3, it shows the new entity relationship diagram of Altair GIS. Batch entity was created to aid the new system feature which allows the user to add a new set of survey answers. The Batch entity contains the name attribute and is linked to the Doctor entity. This was done to keep track of the assigned doctor. Moreover, the Group Occupation entity was created to cater to the new feature that allows the user to classify occupation into groups and to consider them as a single group when displaying. With this new ERD, the researchers were able to properly analyze and display data visualizations.

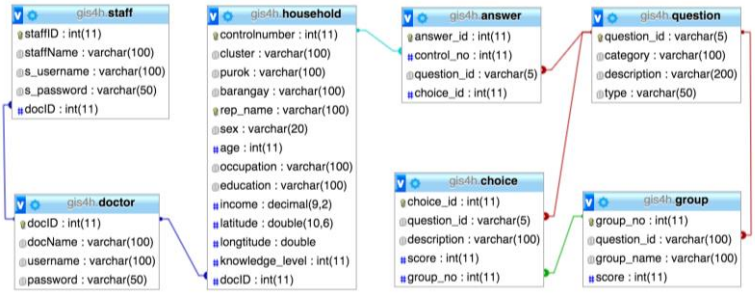

Fig. 2. Existing Entity Relationship Diagram of Altair GIS Before the Redesign.

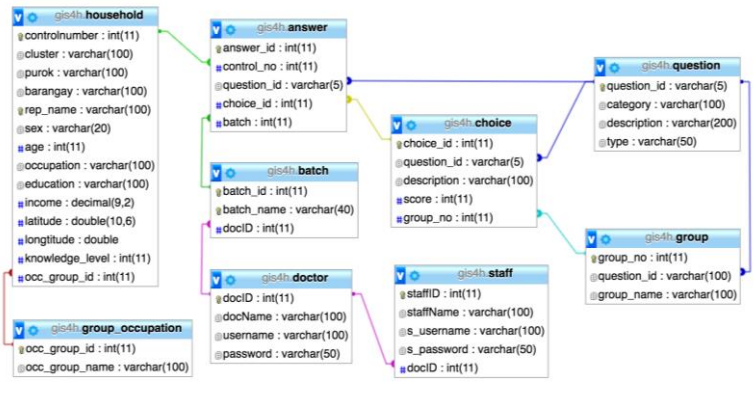

Fig. 3. New Entity Relationship Diagram After the Redesign.

The Schistosomiasis community health data provided by the Municipal Health office consists of 892 records. These include records that are incomplete, noisy, and inconsistent. The researchers have the option to either ignore, fill in the missing values, or remove the incomplete records if they do not affect the outcome of the data analysis. In the Schistosomiasis community health data, the incomplete records contain only basic information which includes the name, address, latitude, and longitude of the respondent and nothing else. Other information such as sex, age, occupation, education, and answers to questions concerning knowledge, attitude, practices, and Schistosomiasis compliance are missing. The researchers removed the incomplete records since they would not be significant during the analysis and visualization processes of the system. Only 558 records were left after the removal of incomplete records. There were no cases of filling in missing values.

3.2. *Data Analysis Model Results*

Data analysis model architecture design is shown in Figure 4. It involves retrieving data from the database, integrating the data into the statistical methods, and integrating the analysis results into the user interface. The data about the household and/or their answers were retrieved from the database using SQL database queries and PHP functions. The retrieved data were then analyzed with the statistical methods. The data analysis returns results which were encoded into JSON format for lightweight data transfer. Through AJAX requests, the statistical data were loaded into the web application. The analysis results were passed to Javascript conditional statement which contains the string conclusions and symbol outputs. The researchers used HTML DOM to get the element with id and innerHTML to set the conclusions.

The development of the data analysis model allows the discovery of relationships that lie behind the schistosomiasis community health data. Possible statistical methods to be integrated in Altair GIS for data analysis was provided by the MHO of Salvador. Chi-square test for independence and Pearson's r correlation coefficient were among the statistical methods. After the data preparation, data was sent to a statistician for evaluation and approval of the initial statistical methods. A study supports the use of the two statistical methods[11]. The summary of appropriate statistical tests based on data type confirms that in order to test the dependence of two variables, chi-square test for independence should be used. Furthermore, the use of Pearson's r and Spearman's rho correlation coefficients are needed to test the correlation between two variables. It was also found out that point-biserial correlation coefficient which is more applicable for Altair GIS data. Pearson's r solves the correlation between two continuous data while Point-biserial solves the correlation between a continuous data. Moreover, Spearman's rho is also similar

to Pearson's r but solves the correlation between two ordinal data. All correlation coefficients vary in magnitude from 0, meaning no correlation at all, to 1, meaning perfect correlation.

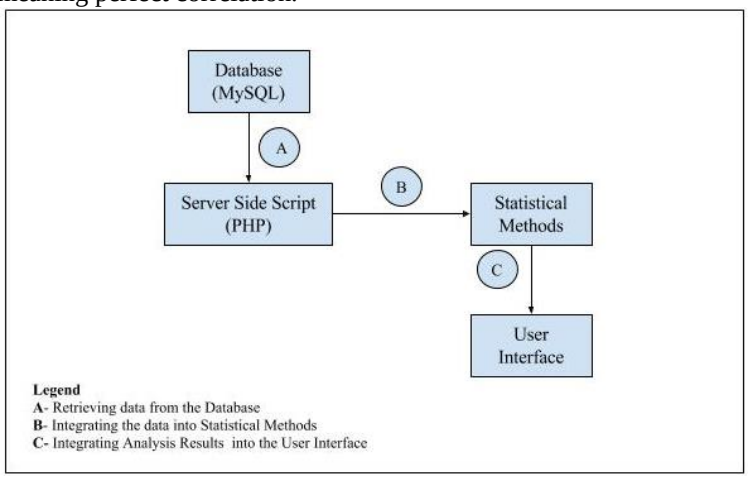

Fig. 4. Data Analysis Model Architecture Design

As shown in Table 1, the descriptive classification of the schistosomiasis community health data are the respondents' sex, age groups, monthly income, educational attainment, and attitude towards schistosomiasis. It was initially identified that these descriptive variables can be visualized using line, bar and pie charts. With these, the MHO can easily monitor the population distributions in a specific description. It was found out that through descriptive and inferential statistics, the MHO wants to discover the contributing factors to the community's participation to the schistosomiasis mass drug administration. That is why, the inferential data are the variables that are needed to be investigated whether they have strong correlation and dependence with the medicine intake. These pairs went through initial data analysis using the applicable statistical methods.

Table 1. Summary of Schistosomiasis Community Health Data provided by the MHO

Descriptive Statistics	Inferential Statistics
● Sex ● Age Groups ● Income ● Educational Attainment	● Sex vs. Intake ● Age vs. Intake ● Occupation vs. Intake ● Income vs. Intake ● Knowledge vs. Intake ● Attitude vs. Intake ● Practice vs. Intake

The test for the initial data analysis returns values that are essential for concluding whether these pairs are correlated and dependent. The MHO of Salvador also provided analysis requirements and possible statistical methods namely, Chi-square test for independence and Pearson's r correlation coefficient, to be used for data analyses of Altair GIS. But these statistical methods are to be investigated whether they are fit to the variable types of the schistosomiasis community health data. The summary of statistical methods that were used for the analysis of schistosomiasis community health data is shown in Table 2.

Table 2. Summary of Statistical Methods to be Used

Chi-square Test	Point-biserial	Spearman's Rho
• Sex vs. Intake • Occupation vs. Intake • Educational Level vs. Intake	• Age vs. Intake • Monthly Household Income vs. Intake	• Knowledge vs. Intake • Practice vs. Intake • Attitude vs. Intake

3.3. *Data Visualization Results*

The answers provided by the respondents in the schistosomiasis community health data was utilized for visualization. Four batches of survey data were generated to allow the researchers to design for reports that involve periods. The visualizations were divided into descriptive and inferential reports based on the user requirements provided by the MHO of Salvador. Descriptive reports visualize the basic information of the respondents in Salvador, Lanao del Norte. This includes sex, age, educational attainment, monthly income, medicine intake, and the reasons why respondents were not able to comply with the medication prescription. Using Google Visualization API, these are presented using pie, bar, and line graphs with appropriate considerations on labels, axes, colors, chart types, and the overall appearance[12, 13].

In the earlier phase of this study, a grouping mechanism was created to control the noise and inconsistencies of the schistosomiasis community health data. A scoring system was also implemented to determine the level of acceptability of each answer. Using pie and bar charts to the knowledge, attitude, and practice answers would only show their frequency, disregarding the assigned score to that answer. To resolve this, the researchers considered using a treemap chart. A treemap from Google Chart libraries is a visual representation of a data tree. Where each node can have zero or more children, and one parent,

in this case the parent nodes are the groupings and the children are the choices. The size of the rectangle indicates the population or the total size of the containing element. The size of the boxes represents the counts or population, as shown in Figure 5. The measures would sum up along the hierarchical structure of the data. Color of the boxes corresponds to the measure of scores. Choices with positive score have boxes colored green while those with negative scores were colored red, and gray for neutral scores [13]. The default behavior is to move down the tree when a user left-clicks a node, and to move back up the tree when a user right-clicks the graph.

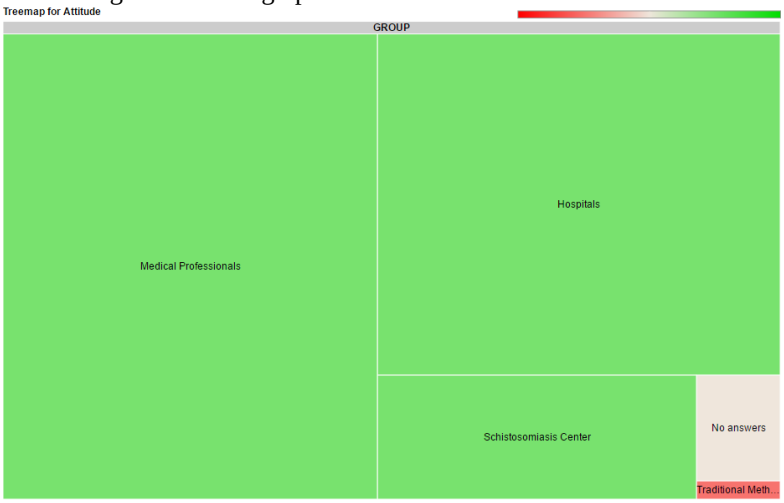

Fig. 5. Treemap Chart

Inferential reports visualize the statistical test outputs between medicine intake versus sex, age, educational attainment, monthly income, occupation, knowledge, attitude, and practice scores. The inferential visualizations consist of interpretations of the data analysis results along with the supporting descriptive visualizations. Interpretation in simplest form will be displayed in an analysis summary panel. If the user wishes to check further details, a 'View Details' button is provided. The button triggers a modal which contains raw statistical values along with an inferential visualization. On functions involving chi-square, value will then utilized by a function containing a formula which was used to solve the chi-square distribution. As shown in Figure 6, an area chart will then be used to graph the curve. An area chart was used since highlighting the area below the curve is needed to indicate the chi value area in blue and p-value area in red. A small area was shaded blue and red indicating the chi and p-value areas. The chi-square distribution skewness is determined by the degrees of freedom of the test.

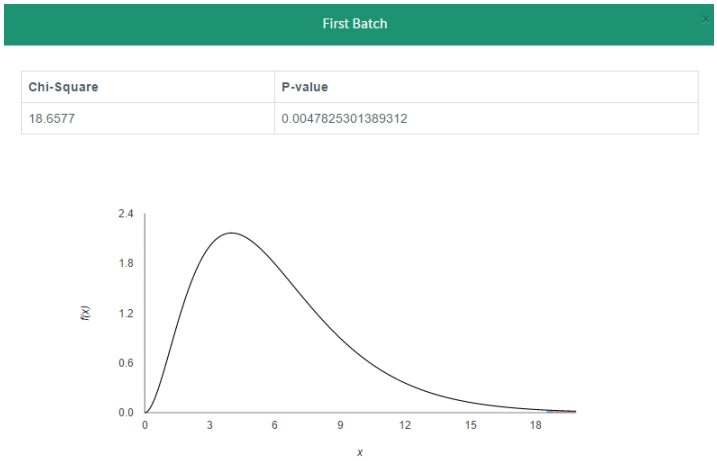

Fig. 6. Summary of Raw Values for Chi-square Test

Similar to the analysis summary of chi-square test, interpretation in simplest form will be returned to the analysis summary panel for age, income, knowledge, attitude and practice. A 'View Details' button is also provided which contains the raw statistical value and scatter plot. Scatter plots have been widely used to represent the correlation between variables. In this study's case, one array contains either the knowledge, attitude, practice, age, or income while the other array contains the medicine intake of the respondents. These two arrays were used for the scatter plot. The medians of 'did take' and 'did not take' plots were used to create a line, which indicates a positive, negative or no correlation at all. As shown in Figure 7, the downward slope indicates that there is a negative correlation between the variables.

In this study, the schistosomiasis community health data were prepared and classified for data analysis and visualization. Applicable statistical methods for data analysis were also determined. The data analysis results helped the MHO of Salvador in making judgments about the difference between groups of data. Texts, glyphicon symbols, distribution curves, and scatter plots allowed the user to see the results derived from statistical values effectively. Chart reports use pie, bar, line graphs and treemaps to visually summarize descriptive data. With these, the MHO can easily view the respondent profiles. Pie and bar allowed the user to view information such as percentage of male and female, age ranges, the income groups, work, and educational level. Linear graphs allow the user to oversee the answers of the respondents through periods. Treemaps simultaneously show the big picture, comparisons of related choices or answer, and allow easy navigation to the details. Distribution curves and scatter plots further support the analysis results.

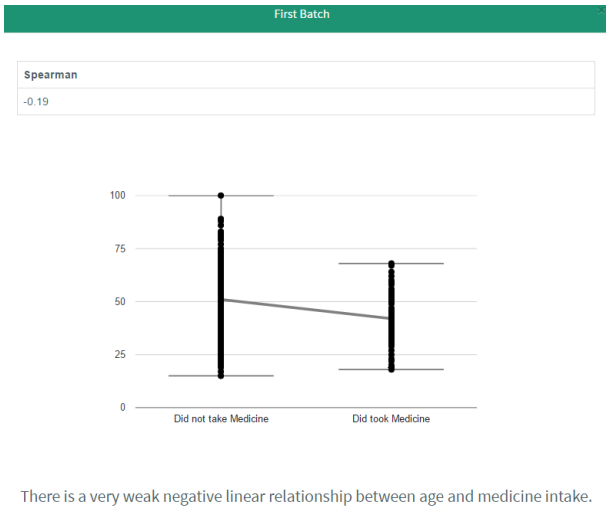

Fig. 7. Summary of Raw Values for Chi-square Test

4. Conclusion and Recommendations

This study enhanced the existing geographic information system by providing it with data analysis model with data visualization which further helped the MHO of Salvador in devising solutions through the visualizations in identifying problems and formulating solutions to increase the rate of schistosomiasis medicine intake. The researchers were able to prepare and categorize the survey data for data analysis and data visualization. The data analysis model was developed by integrating the applicable statistical methods to the schistosomiasis community health data allowing it to produce analysis results through inferential reports. Chi-square test, Point-biserial, and Spearman's rho correlation coefficients were used to investigate the relationship and correlation between medicine intake and other variables in the Schistosomiasis survey data. The determined statistical methods were implemented into scripts. Distribution curves and scatter plots were displayed to further support the analysis results. Google Charts Visualization API was used in producing charts visualizations. Chart reports used pie, bar, line graphs, and treemaps to visually summarize descriptive data. Furthermore, this research gives ideas to software developers on integrating and developing modules to a GIS. This also benefits the GIS specialists by providing them knowledge on the statistics used in the study where they can also use it in other related applications or studies. This research can be used by the future researchers as it can provide a foundation for analysis and planning for health services.

Altair GIS can be further enhanced through an additional functionality that maps specific bodies of water contaminated with schistosomiasis. To be able to

make more appropriate interventions, the researchers recommend the future developers to integrate predictive analytics onto the system to predict the future occurrence of schistosomiasis in the municipality. Lastly, the researchers recommend to broaden the system scope by including other types of diseases aside from schistosomiasis.

Acknowledgements

This research would not have been made possible if not without the assistance of Dr. Jordan C. Pinzon, the Municipal Health Officer of Salvador, Lanao del Norte.

References

1. W.H.O.,*Programmes and Projects - Schistosomiasis* (World Health Organization , 2014).
2. S. Montgomery, *Schistosomiasis* (Centers for Disease Control and Prevention, 2015).
3. L. C. Velasco, J. P. Postrano, L. C. Diaz, L. A. Catane, *A Geographic Information System Using Google Maps for Schistosomiasis Survey Data*(Asia Pacific Journal of Science, Mathematics and Engineering, 2015)
4. S. Leena, *Role of Statistics in Public Health* (NTI Bulletin, 2007)
5. I. Scott, D. Mazhindu, Statistics for Healthcare Professionals (SAGE Publications, 2005).
6. K. L. Bansal, S. Sood, *Data Visualization A Tool of Data Mining* (International Journal of Computer Science and Technology, 2011)
7. M. Friendly, *Milestones in the history of thematic cartography, statistical graphics, and data visualization* (Department of Mathematics and Statistics at York University, 2008)
8. R. White, *Interactions with Search Systems* (Cambridge University Press, 2016).
9. N. S. Chok, *Pearson's versus Spearman's versus Kendall's Correlation Coefficients for Continuous Data*, (Winona State University, 2008)
10. E E. A. *Prepare Data for Analysis and Visualizations* (European Environment Agency, 2015)
11. B. K. Nayak, A. Hazra, *How to choose the right statistical test?* (Indian Journal of Ophthalmol , 2011)
12. D. U. Libraries, *Introduction to Data Visualization: Chart Dos and Don'ts* (Duke University, 2015)

    D. Borland, R. M. Taylor II, *Rainbow color map still considered harmful* (IEEE Computer Graphics and Applications, 2007)