Identifying the Significant Factors in Treated Water Using Binary Logistic Regression Model: A Case Study in Perlis, Malaysia.

by

Fatin Munawwarah Bt Aziz
(1632121978)

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science (Engineering Mathematics)

School of Engineering Mathematics
UNIVERSITI MALAYSIA PERLIS

2017

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my lead supervisor Prof. Dr. Amran Bin Ahmed on his patience guidance, enthusiastic encouragement besides many useful critiques for my research improvement. I would also like to thank Prof. Madya Dr. Nasrul Bin Hamidin, my co – supervisor who helped me a lot during the introduction to water quality data and some useful briefing on the treated water distribution system in Perlis state. Besides that, I want to express my gratitude to both panels in charged during my proposal defense session, Dr. Zainor Ridzuan Yahya and Dr. Mohamad Fadzli Ramli who gave me a lot of motivations, advices and supports for my future research progress and improvement. Also thanks to Dr. Ahmad Kadri Junoh as his commitment and hard work in make sure that the postgraduate students are following the thesis writing progress and submission due date. Last but not least, I would like to appreciate and thanks to all lecturers and staffs at the Institute of Engineering Mathematics (IMK) who had contribute and bring motivation for me to finish my research study.

TABLE OF CONTENTS

	PAGE
THESIS DECLARATION	i
ACKNOWLEDGMENT	ii
TABLE OF CONTENTS	iii
LIST OF TABLES	vi
LIST OF FIGURES	viii
LIST OF ABBREVIATIONS	ix
LIST OF SYMBOLS	x
ABSTRAK	xi
ABSTRACT	xii
CHAPTER 1 INTRODUCTION	
1.1 Overview	1
1.2 Problem Statement	2
1.3 Objectives	3
1.4 Scope of Study	4
1.5 Significance of Study	4
CHAPTER 2 LITERATURE REVIEW	
2.1 Introduction	5
2.2 Disinfection and Disinfectants	5
2.3 Types of DBP	6
2.3.1 .Trihalomethane (THM)	6

2.3.2 Haloacetic Acid (HAA) 7

2.4 Previous Study 7

2.4.1 Study On Interaction Of Dbps With Water Quality Parameters 7

2.4.2 Application of Logistic Regression Model in Water Quality Research 9

2.5 Summary 11

CHAPTER 3 METHODOLOGY

3.1 Introduction 12

3.2 Data Acquisition 13

3.3 Statistical Analysis 13

3.3.1 Correlation Analysis 14

3.3.2 The Analysis of Variance (ANOVA) 16

3.3.3 Binary Logistic Regression 18

CHAPTER 4 RESULT AND DISCUSSION

4.1 Introduction 20

4.2 Result of Correlation Analysis 22

4.2.1 Assumption Testing And Result 23

4.2.2 Correlation analysis in district Arau 28

4.2.3 Correlation analysis in district Jejawi 30

4.2.4 Correlation analysis in district Kangar 32

4.2.5 Correlation analysis in district Kuala Perlis 34

4.2.6 Correlation Analysis in Overall 34

4.3 Result Of The Analysis Of Variance (ANOVA) 36

4.3.1 Assumption Test And Result 37

4.4 Result of Binary Logistic Regression Analysis 40

4.4.1 Assumption Testing And Result 40

CHAPTER 5 CONCLUSION

5.1 Conclusion 44

5.2 Limitation and Recommendation 46

REFERENCES 47

LIST OF TABLES

NO		PAGE
3.1	Range of Spearman correlation coefficient and respective description.	16
4.1	Summary of statistics for overall data in Perlis.	21
4.2	The data summary of parameters in four districts of Perlis state	22
4.3	Result of normality test on different variables.	27
4.4	Spearman's correlation analysis result in Arau.	28
4.5	Spearman’s correlation analysis result in Jejawi	30
4.6	Spearman’s Correlations Analysis Result in Kangar	32
4.7	Spearman’s Correlations Analysis Result in Kuala Perlis	34
4.8	Correlations result on overall	35
4.9	Means and standard deviations of TCM scores by district.	36
4.10	Result of normality test among four districts in Perlis, Malaysia.	38
4.11	Result of equal variance test.	38
4.12	Welch’s F statistic test result.	38
4.13	Post hoc results for TCM scores by district	39

NO		PAGE
4.14	The VIF statistic for each variables.	40
4.15	Result of logistic regression omnibus test of model coefficient.	41
4.16	Result for percentage correct prediction	41
4.17	Wald criterion result for predicting a purified water.	42

LIST OF FIGURES

NO		PAGE
3.1	Methodology flowchart.	12
3.2	Monotonically increasing pattern	15
3.3	Monotonically decreasing pattern	15
4.1	Scatterplot for water temperature against TCM.	23
4.2	Scatterplot for residual chlorine against TCM	24
4.3	Scatterplot of water pH against TCM.	24
4.4	Boxplot for water temperature.	25
4.5	Boxplot for residual chlorine.	25
4.6	Boxplot of water pH	26
4.7	Boxplot of water TCM	27
4.8	The scatterplot residual chlorine against TCM in Arau district.	29
4.9	The scatterplot of water pH against residual chlorine in Jejawi district.	31
4.10	The scatterplot of water pH against residual chlorine in Kangar district.	33
4.11	Boxplot of TCM by district	37

LIST OF ABBREVIATIONS

ANOVA	Analysis of Variance
BrAA	Bromo acetic acid
Br_2AA	Dibromo acetic acid
CH	Chloral hydrate
DBCM	Dibromochloromethane
DBP	Disinfection by-product
DCBM	Dichlorobromomethane
DCAA	Dichloroacetic acid
EPA	Environmental Protection Agency
HAA	Haloacetic acid
MCAA	Monochloroacetic acid
NOM	Natural organic matter
TBM	Tribromomethane
TCAA	Trichloroacetis acid
TCM	Trichloromethane
THM	Trihalomethane
TTHA	Total trihaloacetic acid
TTHM	Total trihalomethane
1,1-DCP	1,1-Dichloropropane
1,1,1-TCP	1,1,1-Trichloropropane

LIST OF SYMBOLS

r	Pearson correlation coefficient
p	Probability
p	p value
df	Degree of freedom
H_o	Null hypothesis
H_1	Alternative hypothesis
α	Alpha significant level
n	Sample size
s^2	Sample variance
x	Independent variable
y	Dependent variables
$\bar{x}$	Sample mean
SD	Standard deviation
σ^2	Variance
χ^2	Chi square
$Exp(B)$	Odds ratios
B	Predictors coefficient for logistic regression model
$S.E$	Standard error
ρ	Spearman's rho correlation coefficient
e	Exponential function

Mengenali Faktor – Faktor Signifikan Air Terawat dengan Menggunakan Model Regresi Logistik Binari: Kajian Kes di Perlis, Malaysia.

ABSTRAK

Penghasilan bahan sampingan dari proses rawatan air yang dikenali sebagai bahan sampingan disinfeksi (DBP) adalah berpunca daripada tindakan antara bahan kimia yang digunakan dalam rawatan air dengan kandungan organik sedia ada dalam air. Trihalomethane (THM) adalah salah satu daripada kumpulan DBP yang sering dikesan di dalam air terawat. THM terdiri daripada beberapa komponen iaitu trichloromethane (TCM), dibromochloromethane (DBCM), dichlorobromomethane (DCBM) and tribromomethane (TBM). Melalui kajian ini, data air terawat bagi negeri Perlis yang direkod pada tahun 2015 telah digunakan untuk tujuan analisis data. Bagi titik pensampelan, negeri Perlis telah dibahagi kepada empat buah daerah utama iaitu Kangar, Jejawi, Kuala Perlis dan Arau. Antara pembolehubah yang terdapat ialah nilai pH air terawat, suhu air, bacaan baki klorin serta bacaan TCM. Satu pembolehubah dikotomi direkod sebagai air ditapis serta air yang tidak ditapis. Kaedah regresi logistik binari telah dipilih bagi mengenal pasti pembolehubah yang signifikan dalam meramal peratusan air ditapis di negeri Perlis. Manakala, kajian korelasi antara kandungan TCM dan pembolehubah yang lain turut dijalankan bagi mengenal pasti pembolehubah yang mempengaruhi pembentukan TCM dalam air terawat. Hasil kajian kolerasi Spearman, terdapat hubungan antara kandungan baki klorin dengan pembentukan TCM di daerah Arau ($\rho = –0.537,\ p < 0.05$). Di samping itu, terdapat kolerasi negatif antara pH air dengan kandungan baki klorin dalam air terawat. Hasil yang sama didapati dari daerah Jejawi ($\rho = –0.603,\ p < 0.01$) dan Kangar ($\rho = –0.722,\ p < 0.01$). Tambahan pula, apabila kajian kolerasi dijalankan ke atas data keseluruhan, hasil mendapati wujudnya interaksi antara suhu air dengan TCM. Namun begitu, tiada sebarang hubungan wujud antara pembolehubah yang lain terhadap TCM. Hasil analisis varians mendapati terdapat perbezaan TCM yang signifikan antara daerah – daerah dalam kajian ini. Daerah Kangar ($\bar{x} = –0.189, SD = 0.171$) memiliki perbezaan ketara daripada Kuala Perlis ($\bar{x} = –0.0478, SD = 0.0667$) dengan perbezaan purata TCM sebanyak 0.14095. Manakala bagi daerah Jejawi ($\bar{x} = 0.1852, SD = 0.0816$) dengan Kuala Perlis adalah sebanyak 0.1374. Di samping itu, terdapat perbezaan ketara antara Jejawi dengan Arau iaitu sebanyak 0.09335. Hasil analisis regresi logistik mendapati, TCM serta baki klorin adalah dua oembolehubah yang signifikan dalam model bagi meramal peratusan air ditapis. Secara keseluruhannya, purata bacaan TCM bagi tahun 2015 adalah 0.13 mg/L dan masih di bawah kawalan iaitu tidak melebihi 0.20 mg/L. Manakala bacaan THM yang lain seperti DBCM, DCBM, dan TBM adalah masih terlalu rendah dan tidak mampu dikesan.

Identifying the Significant Factors in Treated Water Using Binary Logistic Regression Model: A Case Study in Perlis, Malaysia

ABSTRACT

The production of disinfection by-product (DBP) is result of reaction between disinfection agent with the natural organic matter, bromide or iodide in the water. Trihalomethane (THM) is the main group of DBP found in many chlorinated water. THM consist of trichloromethane (TCM), dibromochloromethane (DBCM), dichlorobromomethane (DCBM) and tribromomethane (TBM). In this study, data of treated water taken from year 2015 in state of Perlis, Malaysia. For sampling point, Perlis was divided into four district (Kangar, Jejawi, Kuala Perlis and Arau) and water quality variables measured are water pH, water temperature, residual chlorine, TCM readings, the purified water and non purified water. A binary logistic regression was proposed to predict the probability of purified water in housing area using all the continuous variables measured. Correlation between TCM formation and these variables were estimated using a Spearman's correlation test. A negative moderate correlation exist between residual chlorine and TCM in Arau $(\rho = -0.537, p < 0.05)$. Besides that, there exist a strong negative correlation between water pH and residual chlorine in Jejawi $(\rho = -0.603, p < 0.01)$ but no correlation exist between pH and TCM. Same result obtained for district Kangar $(\rho = -0.722, p < 0.01)$. Additionally, when correlation analysis did on the overall data, there exist a negative weak interaction between water temperature and TCM formation in Perlis $(\rho = -0.222, p < 0.05)$. However, some variables such as pH and residual chlorine did not show any influence towards TCM formation. Results of adjusted Welch's F test suggesting that there is significant mean different of TCM level in every district in Perlis. Significant pairwise difference obtained between Kangar $(\bar{x} = -0.189, SD = 0.171)$ with Kuala Perlis $(\bar{x} = -0.0478, SD = 0.0667)$ with mean difference of 0.14095. Besides that, Jejawi $(\bar{x} = 0.1852, SD = 0.0816)$ was significantly higher of TCM mean compared to Kuala Perlis with mean difference of 0.1374 and Arau $(\bar{x} = 0.092, SD = 0.063)$ with mean difference of 0.09335. From binary logistic regression model, residual chlorine and TCM have influence on the purified water prediction. On the overall, the average TCM concentration in Perlis is 0.13mg/L (130 ppb) which is under acceptable maximum value of TCM while other THM such as DBCM, DCBM, and TBM are non-quantifiable because of concentration below detection limit.

CHAPTER 1

INTRODUCTION

1.1 Overview

The quality of drinking water is a powerful environmental determinant of health. It has been an important issue worldwide especially in many developing countries. According to the U.S Geological Survey, 3% of fresh water is inaccessible and more than 68% of the fresh water on Earth found to be on ice while more than 30% is found in ground water (Mohan, et al., 2012). Gleick (1993) stated that the total fresh water resources available was around 2.5%. Amount of water used had increases year by year due to increasing human population. In Malaysia, clean water demand increase about 58% from year 2000 to 2010.

Some water sources contain disease-causing organisms which need to be removed or killed before the water can be used in drinking (Sadallah & Al-Najar, 2015). Disinfection usually achieved by the use of chlorine, ozone, chlorine dioxide, or a combination of chlorine and ammonia (chloramines) which prevents the water from containing pathogenic microorganisms (Cohn, Cox, & Berger, 1999). Chlorination as a disinfectant can remain even after the disinfecting process to provide protection of drinking water to the consumer (Khleifat, Abboud, Alshamayleh, Jiries, & Tarawneh, 2006). The remaining disinfectant usually in the form of residual chlorine and chloramine. The variation of DBP are depending on type of disinfectant used, the dose of disinfectant, the concentration of the natural organic matter in the water itself, the time since dosing, temperature, and pH of the water (Koivusalo & Vertianien, 1997).

In Malaysia, chlorine is generally used as disinfectant in municipal water supply systems (Sukiman & Pauzi, 1993) and as the result, trihalomethane (THM) are formed during chlorination at the treatment plant or in other water distribution system.

1.2 Problem Statement

The formation of DBP in water treatment process have always been discussed in many developing countries since their bad effect and risk towards human health. Chlorination is the most common disinfection method. TCM or the other name is chloroform commonly found after chlorination. The presence of TCM was easy to be detected in any water pipeline system. It is important to determine the factors or effect that relate to production of TCM. Even though there were many studies did to identify the effect on DBP formation but no study did in Perlis. Additionally, a water purification system have been widely used, therefore the relation between TCM content and purified water quality must be identified where TCM and other variables were used to predict whether water have undergo the purification process or not. In this study, the TCM with other measured variables were analysed using data in Perlis only.

1.3 Objectives

1. To identify the correlation between water quality parameters (water pH, water temperature, chlorine residual) with TCM concentration in treated water.
2. To determine the significant mean differences of TCM between four district.
3. To identify the significant predictors on the probability of purified water using `logistic regression model.

1.4 Scope of Study

This study considers the analysis of TCM formation where the effect of water quality variables towards its formation will be determined. The data of chlorinated water for year 2015 in the state of Perlis, Malaysia covers four districts which are Arau, Jejawi, Kangar and Kuala Perlis. The parameters involved in this study are the water pH, water temperature (^{o}C), locations (districts) of water samplings, the chlorine residual concentration (mg/L) and the TCM concentrations (mg/L). Besides that, there was one variables indicate the purification property of treated water. From the data, a purified water will be recorded as "Y" and "N" for non-purified water.

1.5 Significance of Study

An improved understanding on the effect towards TCM formation is important to the development of effective strategies to regulate TCM besides other THM formation and the chlorine consumptions in Perlis. The objective of this work is to develop statistical model application for the prediction of TCM formation and to predict the practice of water purification in Perlis using TCM and other measured variables. Water filtration system mainly used to filter the treated water before consumption. Through this study the ability of purification system to reduce or remove DBP can be justified by identifying the correlation between purified water and TCM concentration in treated water. Future improvement in every housing filtration system can prevent the risk of TCM and other effect of THM towards human health.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter reviews on DBP formation and the previous studies related to DBP properties and effect of water quality variables on DBP formation in certain condition. Besides that, there are some discussion on application of logistic regression model in a study of water quality.

2.2 Disinfection and Disinfectants

Disinfection is the process of killing or stop any harmful and abhorrent bacteria, cysts and pathogenic microorganisms using various agents such as chemicals, heat, ultraviolet (UV) light, ultrasonic waves, or radiation. The two main reasons to use chemical agent during disinfection are to kill or inactivate pathogens and also to prevent microbial regrowth in water distribution through the providing of disinfection residual (Rahman, 2015). Disinfectants usually made up of free chlorine, chloramine, ozone, and chlorine dioxide. They act as an oxidants efficiently (Rahman, 2015). These components are added into water during the treatment process or into the water distribution systems. Besides that, water treatment using a UV radiation is also good to oxidize organic matter that present in that water efficiently.

2.3 Types of DBP

Chlorine is the most widely used disinfectant for drinking water in many countries including Malaysia. DBP is the product from a reaction between disinfection agent and natural organic matter, bromide or iodide content in the water (Zhang, 2012). DBP usually consist of THM group, haloacetic acids (HAA) group, chlorate, and also bromate compound (Ye, Wang , Yang, Wei, & E, 2009). Richardson and Postigo (1998) compiled a review on DBP and about 600 type of DBP are listed. Some DBP occur at the very low µg/L level, sub µg/L or mg/L level (Xie, 2003). Other than that, there are also countless number of DBP which are unidentified because of their below detection limit (Zhang, 2012).

2.3.1 The Trihalomethane (THM)

In 1970s, Rook (1974) and Bellar et al. (1974) first identified the formation of THM in chlorinated drinking water. THM can be divided into four elements which are chloroform or shortly known as TCM, the tribromomethane (TBM), dichlorobromomethane (DCBM), and dibromochloromethane (DBCM).

However in Perlis, Malaysia, TCM is the main DBP found in treated water while TBM, DCBM, DBCM were non – quantifiable because of below detection limits.

2.3.2 The Haloacetic Acid (HAA)

The second major group of DBP can be formed through water disinfection is the HAA group when chlorine, chloramine, chlorine dioxide, and ultraviolet ray were used as the chemical agents. This HAA group basically consist of five elements such as monochloroacetic acid (MCAA), dichloroacetic acid (DCAA), trichloroacetic acid (TCAA), bromoacetic acid (BrAA), and dibromoacetic acid (Br_2AA). However there are many other type of HAA that probably formed depending on water disinfection process and water characteristics itself (Rahman, 2015).

2.4 Previous Study

In the following part there will be the discussion on the previous studies regarding water quality analysis on interaction between several parameters such as the water pH, water temperature, chlorine residual concentration with DBP production.

2.4.1 Study on Interaction of DBPs with Water Quality Parameters

There are a number of studies conducted to determine the effect of distance, pH, temperature and chlorine residual towards the formation of DBP. Nikolaou et al. (2004) reported that the THM concentrations increase when there was increase of the water pH. However, several type of volatile DBP, such as chloral hydrate (CH), 1,1-dichloropropanone (1,1-DCP), 1,1,1-trichloropropanone (1,1,1-TCP) were only formed at lower water pH suggested that the interaction exist between water pH and DBP are

depending on certain type of DBP. In another study by (Ye, Wang , Yang, Wei, & E, 2009) which subjected to study of factors influencing DBP formation reported that when the pH value increases, there was a significant effect on the formation of total trihalomethane (TTHM) indicating the TTHM increases as the pH increases. Through this study, Pearson correlation analysis gave result of moderate and definite relationship between TCM, DCBM, DBCM, TBM, TTHM formation with pH (r = 0.43, 0.49, 0.48, 0.38, 0.51 respectively). In year 2015, Mohammadi, Miri, Ebrahimi, Khorsandi, and Nemati conducted a same study on THM which stated that no correlation exist between pH and THM formation.

Besides that, there are several papers discussed on effect of temperature towards DBP formation. In Malaysia, Abdullah, Yew, and Ramli (2003) found no correlation between temperature on the formation of THM for the district of Sabak Bernam and Tampin. In this study, all the samples had a constant values of temperature about 28 to 32 ^{o}C. Less variation of temperature on sampling may be the cause of no correlation between formation of THM and temperature (Abdullah, Yew, & Ramli, 2003). Besides that, Ye et al. (2009) reported that there was a low but definite relationship (r = 0.18) obtained between TTHM and water temperature while there exist a moderate correlation between THAA and water temperature. In other study by Mohammadi et al. (2015), there were no specific correlation obtained between temperature and formation of THM.

The chlorine residual content was determined to have effect towards the production of THM in water distribution system (Abdullah, Yew, & Ramli, 2003). From Pearson correlation analysis, the result of a negative low but definite with small relationship (r = - 0.311) obtained. Ye et al. (2009) stated that as the residue free chlorine decreases, the concentration of THM and HAA production will increases. In another study, Mohammadi et al. (2015) reported that there was no significant relationship

between residual chlorine and THM by using Spearman non parametric correlation coefficient. Sadallah and Al-Najar (2015) tried to relate the amount of chlorine residual with the variation in distances of chlorination sources.

2.4.2 Application of Logistic Regression Model in Water Quality Research

Logistic regression has been used extensively in the health sciences since the late 1960s to predict a binary response from explanatory variables (Lemeshow, Teres, Avrunin, & Pastides, 1988) and more recently in the environmental sciences to identify variables that significantly affect ground water quality. Gardner and Vogel (2005) used a logistic regression model to predict ground water nitrate concentration from land use. The analysis revealed that the percentages of forest, undeveloped, and low-density residential land areas were excellent predictors of ground water nitrate concentrations in excess of 2 mg/L. The logistic model was reported as quite accurate for predicting the probability of the ground water nitrate concentrations exceeding 2 mg/L from the reliability model obtained (Gardner & Vogel, 2005).

Venkataraman and Uddameri (2012) did a study to model the simultaneous exceedance of two type of chemical which are arsenic and nitrate from agriculture activities in Ogallala aquifer of Texas, United States. Through the study, two types of logistic regression models were developed. Firstly, by treating arsenic and nitrate independently and combining the marginal probabilities of their exceedance, and second one was by treating the two exceedance together using a multinomial model. Logistic model constructed from the marginal probabilities had lower overall accuracy (59% correct classifications). In contrast, the multinomial model showed good overall accuracy (79% correct classifications), made the correct predictions 90% of the time when both

arsenic and nitrate exceedences were observed. The variable influence the soil and aquifer properties have been identified.

In year 2013, Mair and El-Kadi did a study by applying a multiple logistic regression to assess the water vulnerability in Hawaii. The main objective of the analysis was to develop a vulnerability model for prediction of groundwater contamination in drinking water sources. Input for the logistic regression models utilized explanatory variables based on hydrogeology, land use, and well geometry/location. The objective logistic regression model approach developed in this study is flexible enough to address a wide range of contaminants and represents a suitable addition to the current subjective approach.

In Malaysia, there was a study conducted in Sabah by Zin et al. (2015). The logistic regression is used as one of the methodology in predicting cholera infection and virus in water from rural area in Sabah. Zin et al. (2015) implemented a multivariable logistic regression to determine predictors for household water, sanitation, hygiene, and knowledge about cholera and diarrhoea. Statistically significant associations were obtained.

2.5 Summary

From the review of papers and studies by other researchers, a simple conclusion can be made. There are lack of study did in Malaysia which are totally focusing on the production of DBP and effect of water quality variables towards DBP formation especially in Northern Malaysia. Additionally, the logistic regression was identified to be a powerful statistical tool to predict probability of any event occurrence from continuous or categorical independent variables when there are categorical response variable involved in that data. However, there is no study that apply the use of logistic regression to predict the occurrence of purified water by using THM and other measured variables as the predictors.

CHAPTER 3

METHODOLOGY

3.1 Introduction

In this chapter, there will be more explanation on the statistical methods which have been used to analyse and interpret the data. The flow chart in Figure 3.1 shows the process of this research. Firstly, the secondary data of treated water in Perlis, Malaysia was requested followed by data screening process where any missing values in data were observed and several assumptions respective to the selected statistical analysis are checked. Finally, a binary logistic regression model developed using the variables in data.

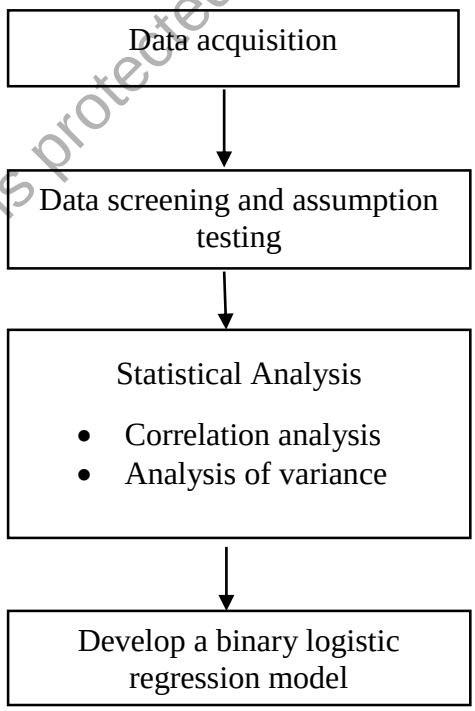

Figure 3.1: Methodology flowchart.