UNIVERSITY OF KWAZULU NATAL

Time Series Analyses of Microbiological Counts Associated with Treated Water Quality Data from Umgeni Water

By

Thabo Lephoto

A Thesis Submitted in Fulfilment of

The Requirements of The Degree for

Master of Science

in

Statistics

under the supervision of

Mr. O. Bodhlyera

and

Prof. H. Mwambi

Faculty of Agriculture, Engineering and Science

School of Mathematics, Statistics and Computer Science

Pietermaritzburg Campus

June 2017

Declaration of Authorship

I, Thabo Lephoto, declare that this thesis titled, 'Time Series Analyses of Microbiological Counts Associated with Treated Water Quality Data from Umgeni Water', contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any degree or diploma in any university or tertiary institution without the prior approval of the University of KwaZulu Natal.

With such exceptions, this thesis is entirely my own work.

Signed: _____ Date: _____

Approved by Supervisors:

1. Mr. Oliver Bodhlyera Signed: _____ Date: _____

2. Prof. Henry Mwambi Signed: _____ Date: _____

Acknowledgements

To Almighty God, I am grateful for His guidance and strength throughout the two years of a good experience working on my Master's thesis.

I would like to thank my supervisors, Mr. Oliver Bodhlyera and Prof. Henry Mwambi for their support in making the work of my Masters degree fruitful and enjoyable. The doors to your office were always open whenever I ran into a trouble spot or had a question about my research or writing. You consistently allowed this thesis to be my own work, but steered me in the right direction whenever you thought I needed it. I will always be grateful to have you both by my side at all times. Gratitude to Mrs Christel Barnard for her administrative support throughout the years.

I would also like to thank the experts from Microbiology department at Umgeni Water who were involved in the validation survey for this research project and their financial support: Mrs Debbie Trollip, Mr. Steve Terry, Dr. Lakesh Maharaj, Mrs Kim Hodgson, Mrs Nirosha Reddy. Without their passionate participation and input, the validation survey could not have been successfully conducted. I am gratefully indebted to their very valuable comments on this thesis.

Finally, I must express my very profound gratitude to my late father, Tumelo Lephoto, my mother Alinah Lephoto and to all my sisters and bothers for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Abstract

Water quality variables change continually through time. The main focus of this study is to study the low-frequency occurrence of microbiological organisms that may affect treated water quality. The study, using long-term data sets, also seeks to understand water quality predictor variables that correlate significantly with key microbiological determinants. The microbiological organisms studied are Total Coliforms (TC) and Heterotrophic Plate Counts at a temperature of 37°C (HPC37) analyzed at different frequencies in final water samples from three water treatment plants. The covariates or predictor variables that were studied in relation to the occurrence of microbiological determinants of water quality are total chlorine, free chlorine, pH, temperature (in the form of seasons, and probably indicative of rainfall or other related variables) and turbidity. Two dummy variables, site, and seasonal time variables were also introduced to capture the variability of HPC37 and TC counts among sites and over time. Distributed Lag (DL) quasi-Poisson and negative binomial generalized linear models (GLM) and generalized linear mixed model (GLMM) were developed based on the assumption of independence or dependence among TC or HPC37 observations coming from the same sampling point. Results show that temperature, turbidity, and chlorine correlate significantly with the occurrence of microbiological determinants. Increased temperature and turbidity levels appear to be linked to increased HPC37 and TC detection. Whereas in some instances increased total chlorine levels link with HPC37 detection. However, as expected from its use as a disinfectant in the treatment process, free chlorine has a negative association with HPC37. In fact, free chlorine showed delayed effects on bacterial counts when on the other side total chlorine showed immediate effects on bacterial counts. Relative to winter,

HPC37 counts seemed to higher during the other seasons and TC counts seemed to be lower during the other seasons. Such statistical assessments of relatively large data sets should prove a useful additional tool in assessing water quality risks.

Key phrases: Time Series Analysis of Water Quality Assessment; Generalized Linear Models; Generalized Linear Mixed Model; Distributed Lag Models.

Acronyms

AIC	Akaike Information Criteria
AV	Assumed Variance
DH	Durban Heights
DHF	Durban Heights Final
DL	Distributed Lag
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
HPC37	Heterotrophic Plate Counts at 37°C
ICC	Intraclass Correlation Coefficient
MQL	Marginal Quasi-Likelihood
PQL	Penalized Quasi-Likelihood
TC	Total Coliforms
WTP	Water Treatment Plant

Contents

List of Tables **ix**

List of Figures **x**

1 Introduction **1**

1.1 Background 1

1.2 Literature Review 4

1.3 Motivation of the Study 8

1.4 Problem Statement 8

1.5 Research Aims and Objectives 9

1.6 Research Methodology 9

1.7 Thesis Layout 10

2 Exploratory Data Analysis **11**

2.1 Introduction 11

2.2 Data Description 11

2.3 Results of the Exploratory Data Analysis 13

3 Generalized Linear Models **21**

3.1 Linear Models 21

3.2 Generalized Linear Models 22

3.2.1 Parameter Estimation 24

3.2.2 The Poisson Regression Model 29
3.2.3 Parameter Estimation for Poisson Log-linear Model 31
3.3 Overdispersion 31
3.3.1 The Negative Binomial Regression (NB) Model 32
3.3.2 Quasi-Likelihood Estimation 35
3.4 The Distributed Lag Model 38
3.5 Application of GLM Quasi-Poisson Distributed Lag Model to Water Quality Data 39
3.5.1 Model Validation in Quasi-Poisson GLM 42
3.6 Application of GLM Negative Binomial Distributed Lag Model to Water Quality Data 48
3.6.1 Model Validation in Negative Binomial 49

4 Generalized Linear Mixed Models **52**

4.1 Linear Mixed Models 52
4.2 Generalized Linear Mixed Models 53
4.3 Maximum Likelihood Estimation 55
4.3.1 Laplace Approximation 56
4.3.2 Penalized Quasi-Likelihood 57
4.3.3 Marginal Quasi-Likelihood 58
4.4 Prediction of Random Effects 59
4.5 Application of the GLMM Poisson Distributed Lag Model for Heterotrophic Plate Counts (HPC37) Predictions 59
4.6 Application of the GLMM Poisson Distributed Lag Model for Total Coliforms Predictions 62
4.7 Intraclass Correlation Coefficient 63
4.8 Model Validation 65
4.9 Random Effects Model Validation 65

5 Model Comparison and Discussion **67**

6 Conclusion and Further Research **71**

6.1 Conclusion 71

6.2 Further Research 73

References **75**

Appendices **81**

A R Code **81**

A.1 Data Preparation in R 81

A.2 R Code Used For Exploratory Data Analysis 82

A.3 R Code Used for Model Fitting 84

A.4 Output ready for Latex in R 84

B Results for Total Coliforms occurrence **86**

List of Tables

2.1 Water quality variables and the units of measurement. 12

2.2 A site and time dummy variables and respective levels. 13

2.3 Summary statistics of the data according to sites, 1991-2015. 13

2.4 Distribution of samples per TC class according to seasons, for TDH007, TDH008 and TDH010, 1991-2015. 17

3.1 Results for a Quasi-Poisson Distributed Lag Model Fitted to HPC37 counts, Durban-Heights, 1991-1997. 40

3.2 Results for a Quasi-Poisson Distributed Lag Model Fitted to TC counts, Durban-Heights, 1991-1997. 47

3.3 Results for a Negative Binomial Distributed Lag Model Fitted to HPC37 counts, Durban-Heights, 1991-1997. 48

4.1 Results for a GLMM Poisson Distributed Lag Model Fitted to HPC37 counts, Durban-Heights, 1991-1997. 60

5.1 A comparison of distributed lag count regression models for HPC37 occurrence. 69

B.1 Regression models 1 and 3 for Total Coliforms occurrence, DHF 1, 2 and 3: 1991-1997. 86

List of Figures

2.1 Data distributions for all variables, for site TDH010, 1991-2015. 14

2.2 Effect of time (seasons) on HPC37 levels, Durban-Heights, 1991-2015. 16

2.3 HPC37 levels according to water-quality parameters, TDH007, 1991-2015. . . . 18

3.1 Mean-Variance Relationship. 42

3.2 The residuals-squared plot versus the predicted mean of the quasi-Poisson model. 46

3.3 Mean-Variance Relationship. 50

3.4 Graphical validation tools for the Negative Binomial. 51

4.1 Graphical validation tools for the Poisson GLMM. 65

4.2 Within site graphical validation tools for the Poisson GLMM. 66

Chapter 1

Introduction

1.1 Background

Healthy human life requires the availability of adequate and safe drinking water to sustain life. That is, water is a vital natural resource since it is fundamental to any form of life. Therefore a satisfactory (adequate, accessible and safe) supply must be available to all at all times. Safe drinking water is water that does not pose any significant risk to health over a lifetime of consumption, including any sensitivities that may occur during the life stages (WHO, 2004). The guidelines for safe drinking water can be found in WHO's editions of the *Guidelines for Drinking-water Quality*. They provide reasonable minimum requirements for safe practice to protect the health of consumers or derive numerical guideline values for constituents of water or indicators of water quality.

This research project focuses on the microbial safety of drinking water. It seeks to find the association between microbiological out-of-range data and water-quality data. Findings of this study, will hopefully assist water agencies to statistically better understand data on the link between water quality variables and the incidence of microbial in water. Multiple barriers are used to secure the microbial safety of drinking water supplies, that is, from catchment to consumer, to prevent the contamination of drinking-water or to reduce contamination to levels

not harmful to health. Safety is guaranteed if barriers are employed, including protection of water resources, proper selection, and operation of series of treatment steps and management of distribution systems, that is, piped or otherwise to maintain and protect treated water quality. The quality of water can be described as the suitability of water to sustain various uses or processes (Bartram and Ballance, 1996).

Any use of water will have certain requirements for physical, chemical and or biological characteristics. The quality of water depends on various factors that could limit or reduce the use of water. That is, a certain amount of bacterial counts such as heterotrophic plate counts at 37°C (HPC37) or total coliforms (TC) in drinking-water might pose a health risk to the public and hence limit the people from using drinking-water as a source of life. Human and natural influences affect the quality of water. Natural influences can be geological, topological, meteorological, hydrological and biological as these affect the quantity and quality of water quality variables. These have a great impact when available water quantities are low and the use of the limited resource is at maximum levels (Bartram and Ballance, 1996).

In ensuring that water is microbially safe for drinking and for any other use in general, bacterial indicators of water quality are measured. These are HPC37 and TC. The HPC37 bacteria are suspected to signal an increased health risk when elevated quantities are present in drinking water. However, even though the literature document the universal occurrence of HPC37 bacteria in soil, food, air and water, there is insufficient clinical and epidemiological evidence to conclude that HPC37 in drinking water poses a health risk. Heterotrophic bacteria comprises of all bacteria that use organic nutrients for growth and these bacteria are found in all types of water, food, soil, vegetation and air. These broad definitions include primary and secondary bacterial pathogens known as coliforms (*Escherichia*, *Klebsiella*, *Enterobacter*, *Citrobacter*, *Serratia*) (Allen et al., 2004).

Total coliforms are defined as a group of bacteria commonly found in the environment, for example in soil or vegetation, as well as the intestines of mammals, including humans. These are non-spore-forming bacilli capable of growing in the presence of high concentrations of bile

salts with the fermentation of lactose and production of acid or aldehyde within 24 hours at 35-37°C. They are comprised of *Escherichia coli* (*E.coli*) and thermotolerant coliforms and can ferment lactose at high temperatures.

Water distribution systems are as important as the water resource itself and treatment facilities, in ensuring the supply of safe drinking water, but they also provide a habitat for microorganisms which are sustained by organic and inorganic nutrients present on pipe and in the conveyed water. Apparently, water distribution systems are difficult to maintain and operate (Payment and Robertson, 2004). This is emphasized by Reilly and Kippin (1983) in their study *"Relationship of bacterial counts with turbidity and free chlorine in two distribution systems"*.

At Umgeni Water there has been incidences where TC and HPC37 would exceed limits set by South African National Standards (SANS241-1, 2015). Since there is non-compliance results in the distribution systems, there are five independent water quality variables or factors that will be considered and hence relate to the microbiological out-of-range variables. These include turbidity which can be thought of as the amount of cloudiness or haziness of water, the free- and total chlorine which are both used as disinfectants, the pH-level of water which is a measure of acidity or alkalinity of water, and the temperature of water. All these water quality variables are measured from water samples that are routinely taken to monitor water quality.

The next section reviews the literature in microbiological organisms found in water and the water quality variables; both microorganisms and water quality variables discussed in this section. The motivation for this study, the problem statement and the section on research objectives and aims follows the section on literature review.

1.2 Literature Review

This section reviews the related work done by other researchers. The review starts by mentioning a few health and development conferences held around the world as a way of informing people and protecting them from water-related diseases. It then reviews the relationships between microbiological determinants and water quality data from other countries before looking at those who were found in South Africa. Common techniques used by others to find these relationships are discussed. It is important to understand how and why the techniques were used when different applications were done. Through such an approach new methods and other techniques can be introduced and the gap in the research filled. Lastly, the significance of this study is stated and explained in terms of methods that are used and why they are used.

The importance of water, sanitation and hygiene for health and development has been emphasized by many international policy forums decades ago. These included health-oriented conferences such as the International Conference on Primary Health Care in 1978, held in Alma-Ata, Kazakhstan, the 1977 World Water Conference in Mar del Plata in Argentina which launched the water supply and sanitation decade of 1981-1990, the Millennium Declaration goals adopted by the General Assembly of the United Nation (UN) in 2000 and the Johannesburg World Summit for Sustainable Development in 2002. Recently, the UN General Assembly declared the period from 2005 to 2015 as the International Decade for Action,"Water for Life" (WHO, 2004).

On the other hand, research on drinking water quality and their association with microbiological out-of-range variables has been an ongoing process in many countries around the world. A correlation between lack of disinfection and increase in anti-diarrheal drugs was found in the city of Le Havre, France, during the study that was done by Beaudeau et al. (1999). This means that lack of chlorine in the system was linked with the diarrheal-causing pathogens in drinking water. "These viruses are responsible for a wide range of acute and chronic illnesses, with the most common being diarrhea" (Shaban and Malkawi, 2007). In Jaipur city, India, the presence of free residual chlorine in drinking water was found to be correlated with the absence

of disease-causing organisms (Chandra et al., 2016). Some related studies have shown that disinfection either by chlorine or chloramine plays a major role in reducing levels of bacteria in water (Zhang and DiGiano, 2002; Power and Nagy, 1999).

In a study in Quebec city, Canada, a study that was conducted by Francisque et al. (2009) showed that Heterotrophic plate counts or heterotrophic bacteria have an inverse relationship with free chlorine residuals, and a positive relationship with water temperature. The study showed that below the value of 0.3 mg/L of free chlorine, a standard target at the exit of the water treatment, according to the United States Environmental Protection Agency (1989), HPC37 exceeded the threshold of 100 and 500 CFU/mL. The results showed that residual chlorine levels above the 0.3 mg/L favored HPC37 levels lower than 100 CFU/mL of free chlorine. The study also showed that water temperature under winter conditions (water temperature $\leq 4°$) had HPC37 levels lower than 500 CFU/mL. This means that the higher the temperature the higher the HPC37 bacteria in drinking water (Zhang and DiGiano, 2002). HPC37 bacteria has a positive correlation with water residence time, that is, the time it takes for water to travel from the water treatment plant to a given sampling point, and a positive correlation with water temperature and rainfall (LeChevallier et al., 1981; Zhang and DiGiano, 2002).

LeChevallier et al. (1981), in their study conducted in Oregon in the United States, found turbidity to be negatively correlated with the $\log_{10}$ decrease in coliform numbers due to seasonal changes, chlorine demand and the initial coliform level. Hsieh et al. (2015) in their study conducted in New York city in the United States of America, found turbidity to be related to diarrheal illness. It should be noted that diarrhea is an illness resulting from elevated bacteria such as *E.coli* in drinking water (Levine, 1987; Sack et al., 1997). Turbidity is said to have an impact in protecting microorganism from inactivation by disinfection and hence its removal is needed in drinking water (Hendricks, 1978).

In South Africa, research linking microbiological out-of-range organisms to treated water quality data is still lacking. In the Eastern Cape province, a study was conducted where disinfection

practices and their effect on the quality of drinking water were examined. Out of the 55 plants that were surveyed, 55% had the turbidity values within the acceptable South African Bureau of Standards (SABS) limits and only 18% complied with the limits in terms of the microbiological quality. The high bacterial numbers of total and fecal coliforms in the other plants were associated with high turbidity, which was said to be a result of heavy rains, and inefficient chemical dosing, and hence led to low chlorine residuals (Momba et al., 2006). Similar studies were conducted in Limpopo and KwaZulu-Natal provinces of South Africa, but results were based on untreated water from the river (Bezuidenhout et al., 2002; Lin et al., 2004; Germs et al., 2004).

The literature on the relationship between microbiological out-of-range data and water quality and/or other measurable variables of interest document Spearman's rank correlation coefficient and Pearson's correlation coefficient as common statistical techniques used to measure the degree of correlation (Carter et al., 2000; Zhang and DiGiano, 2002; Momba et al., 2006; Wilkes et al., 2009). The Spearman's rank correlation coefficient is a nonparametric technique for evaluating the strength of association between two independent variables (Hauke and Kossowski, 2011). Since this technique is nonparametric, it is unaffected by the distribution of the data. It is insensitive to outliers because it operates on the ranks and not actual values of the data and does not require the data to be collected at regularly spaced intervals. However, this technique has the disadvantage of losing information when the data are converted to ranks and, if the data are normally distributed, the Spearman's correlation coefficient tends to be less powerful than the Pearson's correlation coefficient (Gautheir, 2001). Pearson's correlation coefficient can be defined as a measure of the strength of the linear relationship between two variables such as HPC37 or TC and water quality variables in this case (Hauke and Kossowski, 2011).

This study is focused on time series analysis of microbiological out-of-range counts and water quality variables. It is worth noting that the outcome variables are bacterial counts variables which consist of many zeros. Consequently, it is not an ideal approach to use Spearman's

or Pearson's correlation coefficient techniques because the data is zero inflated. Therefore, models for count data and, incorporating time series techniques are considered in this study. Beaudeau et al. (1999), in their study modeling the number of medication sales bought in a pharmaceutical shop due to diarrhea illness, had suggested the use of a linear model. To model counts, the commonly used model is the log-linear Poisson model (Peng and Dominici, 2008). However, Bolker et al. (2009) point out that transforming non-normal data to achieve normality and homogeneity of variance is not ideal because random effects may be ignored or treated as fixed effects and thus committing pseudoreplication, violating statistical assumptions or limit the scope of inference.

The literature on time series studies relating microbiological counts with water quality variables is scarce. A number of studies which related bacterial counts in water to the environmental factors, as well as other water quality variables, have assumed that the effect of a water quality variable or any environmental exposure takes effect over a single day (Zhang and DiGiano, 2002; Allen et al., 2004; Chandra et al., 2016) among others. However, research relating microorganism found in drinking water samples, to the external factors as well as other related variables, particularly those that are believed to have influence in the quality of water, have been an ongoing process. Francisque et al. (2009) in their study used similar methods used in this research project, but made the assumption that the effect of a unit increase in a water quality variables manifests over a single day. This project uses distributed lag model to determine the effect of a unit increase in a water quality variable on the HPC37 and TC count outcomes. Distributed lag models assume that the effect of a unit increase in a water quality variable on a given day is spread out over K days into the future (Francisque et al., 2009). Alternatively, this also means the effect of a water quality variable at a given day is a result of an action taken K days ago.

1.3 Motivation of the Study

In 2002 South Africa had estimated that 84.5% of its population had access to piped and tap water and this increased to 89.3% in 2010 (Luyt et al., 2012). Water agencies are responsible for an independent and periodic review of all aspects of water quality and public health safety. That is, they are responsible for quality and safety of the water that they produce and distribute. The task of water-supply agencies includes carrying out routine testing and monitoring of the quality of the water it produces. However, tests have shown that pathogens may be present in water after treatment. At Umgeni Water, the department of microbiology had non-compliance results for treated water samples in almost all of its sample sites in the past years. But how this was possible is an interesting question of this study. Consequently, there is a need to investigate the relationships between treatment and outcomes. That is, validation is needed in terms of statistical methods as to what is the relationship between microbiological out-of-range outcomes and the associated water quality outcomes in treated water samples. Such historical data, if analyzed and interpreted correctly, will add value in improving the understanding of the efficacy of the water treatment process and the risk to public health.

1.4 Problem Statement

In this research-study our focus is on treated water samples for microbial organisms data. The focus of this study is what appropriate statistical analysis techniques can tell us about the relationships between microbiological out-of-range data (Total Coliforms, *E. coli* and Heterotrophic Plate Counts (HPC at 37°C)) and other measurable variables at Umgeni Water. The study also explores the relationship between microbiological exceedances and associated water quality data (free and total chlorine, pH, turbidity, sample temperature) in treated water samples. The study also probes if the extensive historical microbiological and other water quality database be can used to improve our understanding of the efficacy of the water

treatment process and the risk to public health.

1.5 Research Aims and Objectives

Aims

The main aim of the project is to link the microbiological out-of-range data to the water quality data using time series statistical methods. The study also seeks to find the best model for the analysis of these non-Guassian data sets.

Objectives

The study seek to relate the microbiological out-of-range data to the water quality data using time series statistical models. Finding these relationships will then make it simple for determining if the available historical microbiological data and other water quality database can be used to improve understanding of the efficacy of the water treatment process and the risk to public health.

1.6 Research Methodology

In this research project the data used was provided by the Umgeni Water agencies. Statistical modeling will be done using generalized linear models (GLM) and extension to generalized linear mixed model (GLMM). The quasi-Poisson model for count data will be used. A Negative Binomial model is also employed in the process of analyzing the data and also as one of the count data model for over-dispersed data.

1.7 Thesis Layout

In Chapter 2 we begin by describing the data, and thereafter explore it in terms of plots. In Chapter 3, we look at the theory of GLM incorporating distributed lag model (DLM) and the application to real water quality data. Chapter 4 is focused on the extension of GLM to GLMM DLM and application to real water quality data. In Chapter 5 we compare models and provide the discussion based on the information criteria. The conclusion and suggestions for further research are in Chapter 6.

Chapter 2

Exploratory Data Analysis

2.1 Introduction

There are a number of interesting features of the data that need to be looked at before actual modeling, such as how the time series plots of the variables look like and the relationships between time and the bacteria in drinking water. The purpose of this chapter is to provide a number of different answers to these issues. Graphical and regression analyses were carried out using the R statistical software (R Core Team, 2015).

2.2 Data Description

This research project uses microbiological data that was collected by the Umgeni-Water agency for more than 20 years. Umgeni-Water supplies water to a large area of the KwaZulu-Natal (KZN) province in South Africa and has more than 100 sampling sites. However, this project uses three sampling sites located in the city of Durban, namely Durban-Heights 1, Durban-Heights 2 and Durban-Heights 3. The acronyms or codes for the three sites are TDH007, TDH008 and TDH010 respectively. It is worth noting that the data is from final water samples taken daily since 1991 through 2015. Each site had more than 8000 observations

recorded with some values missing. Table 2.1 shows variable names and their corresponding units of measurements. Water samples in this study were taken from the tap and some quality measurements are taken on site. For representative samples, initially water was allowed to run for 5 minutes, then the tap was burnt for 30 seconds to kill pathogens, and then samples were taken. The pH and temperature of water were measured immediately at the site while other variables were measured at the laboratory. When sampling is done, samples are then stored in a cooler box filled with ice to slow down reactions.

Table 2.1: Water quality variables and the units of measurement.

Variable	Unit(s)
Free Chlorine	Milligrams per litre (mg/L)
Total Chlorine	Milligrams per litre (mg/L)
Total Coliforms	Colony Forming Units per 100 millilitres (CFU/100mL)
HPC37	Colony Forming Units per millilitres (CFU/mL)
pH	Moles per litre of hydrogen ions (mol/L)
Temperature	Degree Celsius (°C)
Turbidity	Nephelometric Turbidity Unit (NTU)
Rainfall	Millimetre (mm)

As displayed in 2.1, temperature was measured in degree Celsius (°C), turbidity in nephelometric turbidity units (NTU), pH is measured in moles per litre of hydrogen ions (mol/L), total and free chlorine in milligrams per litre (mg/L) over time in days. To aid in the exploratory data analysis two dummy variables were created namely, the seasonal time variable and the site variable. The site variable consist of three levels namely TDH007, TDH008 and TDH010, while the time variable is seasonal with four levels namely, summer , autumn , winter and spring respectively. The levels of time variable take the months as shown in Table 2.2. The seasonal variable was introduced to capture seasonal changes whereas the site variable was introduced to capture variability among the three sites in order to give appropriate inference about relationships.

The data can be used by both microbiologists to determine the microbiological quality of water and statisticians to determine trends and correlations of microbiological pathogens with water quality determinants. Thus sound statistical analysis are key to the current project.

Table 2.2: A site and time dummy variables and respective levels.

Dummy Variable	Levels
Site	TDH007
	TDH008
	TDH010
Time	Summer
	Autumn
	Winter
	Spring

2.3 Results of the Exploratory Data Analysis

In this section a general description of the data is shown in terms of tables and graphs. The summary description of the data is shown in Table 2.3. It can be seen from Table 2.3 that each variable consists of more than 8000 observations in each site.

Table 2.3: Summary statistics of the data according to sites, 1991-2015.

Site	Variables	N	Mean	St. Dev.	Min	Max
TDH007	Free Chlorine (mg/L)	8,656	0.939	0.169	0.100	1.600
	Total Chlorine (mg/L)	8,658	1.169	0.205	0.100	2.100
	Total Coliforms (CFU/100mL)	8,634	0.050	2.429	0	201
	HPC (CFU/10mL)	8,637	1.091	20.911	0	1,000
	pH	6,391	7.829	0.133	7.200	9.000
	Temperature (°C)	5,902	21.292	3.332	13.000	29.100
	Turbidity (NTU)	8,655	0.217	0.115	0.010	5.090
	Rainfall (mm)	8,473	2.386	9.182	9.000	275.500
TDH008	Free Chlorine (mg/L)	8,705	0.989	0.171	0.100	1.700
	Total Chlorine (mg/L)	8,706	1.226	0.204	0.500	2.400
	Total Coliforms (CFU/100mL)	8,684	0.008	0.250	0	18
	HPC (CFU/10mL)	8,686	0.815	13.631	0	936
	pH	6,401	7.834	0.125	7.400	9.100
	Temperature (°C)	6,028	21.160	3.381	10.000	29.000
	Turbidity (NTU)	8,694	0.201	0.103	0.000	3.170
	Rainfall (mm)	8,519	2.403	9.194	9.000	275.500
TDH010	Free Chlorine (mg/L)	8,447	1.013	0.131	0.000	1.500
	Total Chlorine (mg/L)	8,448	1.257	0.161	0.000	3.000
	Total Coliforms (CFU/100mL)	8,425	0.004	0.112	0	6
	HPC (CFU/10mL)	8,424	0.596	9.451	0	575
	pH	8,443	7.826	0.133	7.300	8.900
	Temperature (°C)	8,146	21.283	3.226	11.900	29.100
	Turbidity (NTU)	8,434	0.191	0.092	0.000	1.850
	Rainfall (mm)	8,254	2.427	9.285	9.000	275.500

However, other statistics in the table suggest slight differences in means, standard deviations, minimum values and maximum values among variables across the sites. It is evident from the descriptive statistics that HPC37 counts are more variable than total coliforms. The time series plots of these variables for site TDH010 are shown in Figure 2.1 for the period from 1991 through 2015. In Figure 2.1 time series of variables shown in Table 2.3 is illustrated for the years 1991 to 2015. Figure 2.1a show a time series of free chlorine, and it can be seen that there is a step up movement in the series as time goes. In Figure 2.1b the movement of total chlorine with time is shown and the series plot also shows a step up movement between 2000 and 2005.

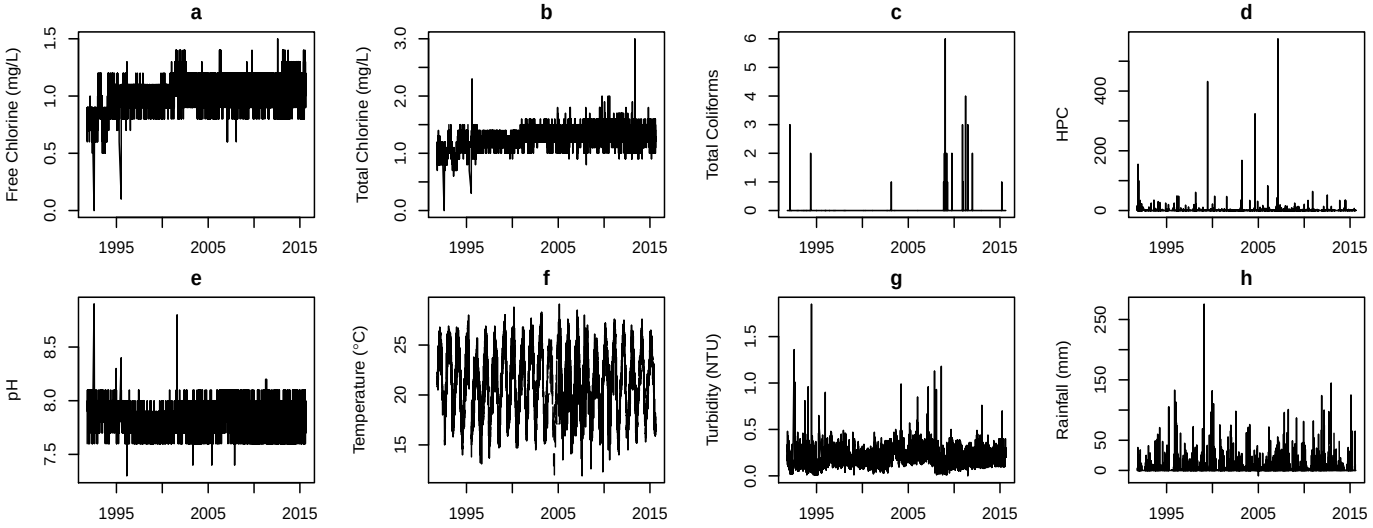

Figure 2.1: Data distributions for all variables, for site TDH010, 1991-2015.

According to South African National Standards (SANS 241) and the Umgeni Water internal standards, chlorine levels met compliance throughout the years in all the sites. The total coliforms (TC) and heterotrophic plate counts at 37°C (HPC37) counts are illustrated in Figure 2.1c and Figure 2.1d. The pH levels of the water samples does not show too much acidity or too much alkalinity but fall between the acceptable limits of 5.0-9.7 in most instances (Figure 2.1e). The water temperature shows strong seasonality throughout the years with peaks in summer and troughs in winter months as expected, see Figure 2.1f. Turbidity shows higher spikes during the first 5 years and between 2005 and 2010 when compared to other years, see Figure 2.1g. Figure 2.1h shows rainfall data. Rainfall is also seasonal with high levels in

warmer months than in colder months. It is worth noting that the rainfall data is the same for the three sites. This is because the three sites are all situated in Durban and therefore their rainfall data is assumed to be the same. There are fewer TC spikes than HPC37 spikes (see Figures 2.1c and 2.1d respectively) and the two variables differ in their characteristics in terms of their harmfulness hence their recommended standard limits are not the same.

In Figure 2.2, the distributions of HPC37 levels by seasons are illustrated using stacked bar plots. It is interesting to note that winter has the highest percentage of zero counts of HPC37 in all sites compared to other seasons. Thus temperature which is season depended, among other parameters, has an impact on HPC37 bacteria in treated water.

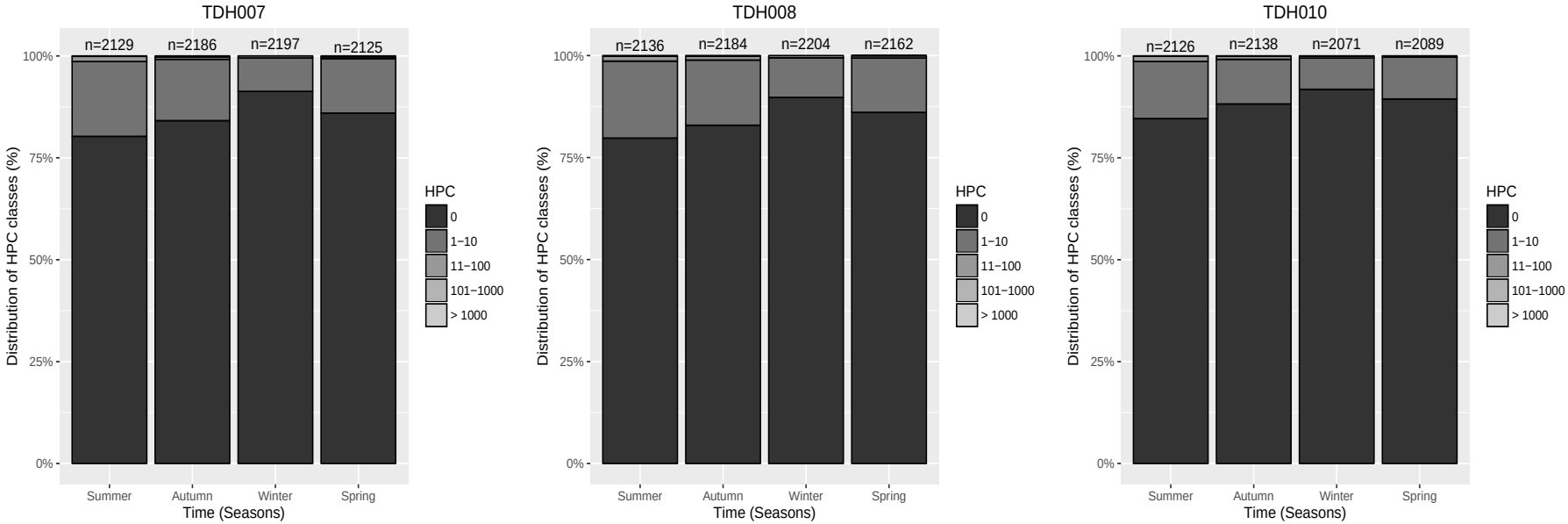

Figure 2.2: Effect of time (seasons) on HPC37 levels, Durban-Heights, 1991-2015.

The distribution of samples per TC class according to seasons across the three site is shown in Table 2.4. The corresponding percentages are given next to each of the TC count values in brackets. The percentage of TC zero counts are almost 100% in the three sites and hence are shown in the table rather than plots. Also, Table 2.4 suggest that summer, autumn and spring have high percentages of positive TC counts relative to winter across the three sites.

Table 2.4: Distribution of samples per TC class according to seasons, for TDH007, TDH008 and TDH010, 1991-2015.

Site	TC classes (NTU/100mL)	Autumn (%)	Spring (%)	Summer (%)	Winter (%)
TDH007	0	2176 (99.68)	2116 (99.58)	2117 (99.53)	2196 (99.86)
	1-10	5 (0.23)	8 (0.38)	8 (0.38)	2 (0.09)
	11-20	1 (0.05)	0 (0.00)	0 (0.00)	0 (0.00)
	More than 20	1 (0.05)	1 (0.05)	2 (0.09)	1 (0.05)
TDH008	0	2172 (99.63)	2158 (99.77)	2130 (99.67)	2202 (99.91)
	1-10	7 (0.32)	5 (0.23)	7 (0.33)	2 (0.09)
	11-20	1 (0.05)	0 (0.00)	0 (0.00)	0 (0.00)
	More than 20	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
TDH010	0	2134 (99.77)	2087 (99.86)	2118 (99.67)	2070 (99.95)
	1-10	5 (0.23)	3 (0.14)	7 (0.33)	1 (0.05)
	11-20	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
	More than 20	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)

Durban Heights 1 (TDH007) is the only site where more than 20 TC positive counts were observed during the period of 1991 to 2015. It is simple to see that the three sites have similarities in terms of their behavior and the distribution of these bacterial counts. Apparently, the maintenance of keeping low levels of HPC37 and TC under warm conditions is similar across the sites. According to Francisque et al. (2009) summer is the season during which higher levels of UV-254 nm (indicator of organic matter in water) are observed.

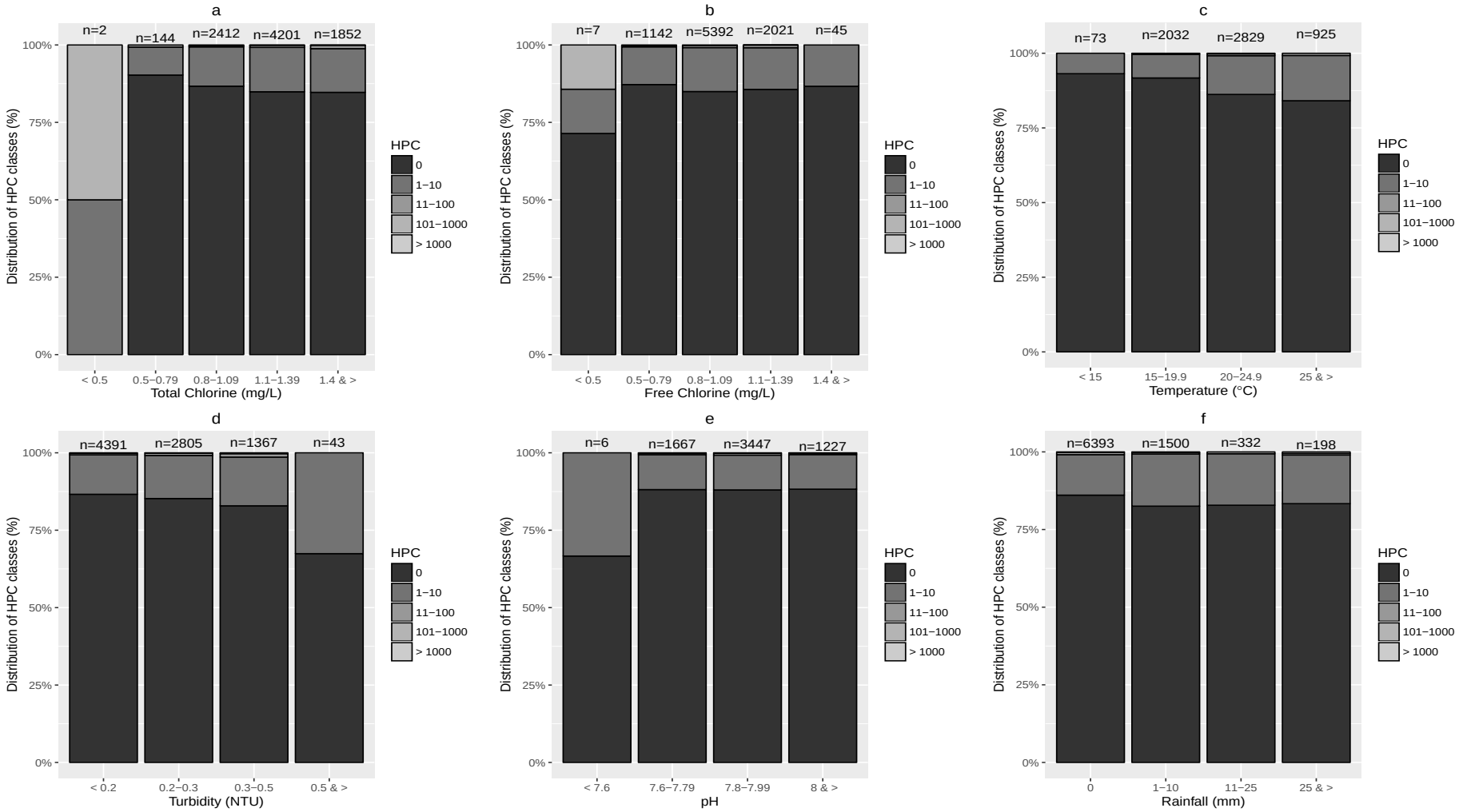

Figure 2.3: HPC37 levels according to water-quality parameters, TDH007, 1991-2015.

In Figure 2.3 the HPC37 percentage according to water-quality parameters is shown. Figure 2.3a shows that when total chlorine level is below 0.5 mg/L more positive counts are observed, and in fact in site TDH007 only positive counts were observed throughout the years 1991-2015 for total chlorine below 0.5 mg/L. When the level of total chlorine was between 0.5 and 0.79 mg/L a lower percentage of HPC37 positive counts is observed. Further, levels of total chlorine above 0.79 mg/L increases the percentage of positive counts (see Figure 2.3a). Figure 2.3b shows that when levels of free chlorine are below 0.5 mg/L more positive counts are encountered, but when they range between 0.5 and 0.79 mg/L positive counts decrease and increase when free chlorine levels are between 0.8 and 1.09 mg/L and after that the percentage of positive counts decrease with an increase in free chlorine levels indicating an effective killing of HPC37 bacteria in drinking water.

Increased water temperature increases percentages of positive counts in the distribution systems, this is illustrated in Figure 2.3c and this confirms the relationships shown earlier in Figure 2.2. Also, in Figure 2.3d it is shown that increased levels of turbidity increase the percentages of HPC37 positive counts. Figure 2.3e shows percentages of HPC37 counts according to pH levels. Below the pH value of 7.6 a few samples show that percentages of HPC37 positive counts are high and above pH level of 7.6 there seems to be no change in HPC37 counts and with increasing pH levels. Lastly, HPC37 percentages of positive counts show a slightly negative relationship with rainfall (see Figure 2.3f). That is, when there are heavy rains positive counts are more likely expected than when there is no rain at all.

Total coliforms cannot be simply demonstrated using stacked bar plots because they have low percentages of positive counts and that makes it difficult to see how they are related with water quality parameters. These bacteria also behave in a similar manner as HPC37 bacteria but with less positive counts compared to the HPC37 across all the three sites.

In this chapter we looked at the descriptive statistics of the time series, namely HPC37, TC, total- and free chlorine, turbidity, pH, rainfall. We looked at how HPC37, TC and other water quality variables are distributed, their seasonal changes and the relationship of bacterial counts

with water quality parameters. More bacteria is often observed during warmer seasons and thus temperature, among other water quality parameters that influence positive counts in the reservoirs, plays an important role in the presence of bacteria in water and hence decreases the quality of drinking water. Both Total and Free Chlorine have shown slight negative relationships with bacterial (positive) counts indicating reduction. Increased turbidity levels is associated with decreased percentages of bacterial zero counts, which means that turbidity should not be in the noncompliance categories (stated in SANS 241) in drinking water as this may lead to water with unacceptable levels of microbiological contaminants. Further, low pH levels were associated with the presence of bacteria in water and therefore a pH level above 7.6 should be considered best because percentages of zero counts are increased. Lastly, higher rainfall increases the levels of bacteria in the water hence lowers the percentages of zero bacterial counts. However, it is worth noting that great impact on the quality of water is observed or expected when a large number of samples is in the non compliance category. Non-compliance in this project refers to water quality variables that do not fall within the numerical limits specified in SANS 241-1. The significance of these variables will be demonstrated by means of model fitting in the chapters that follow.

Chapter 3

Generalized Linear Models

3.1 Linear Models

A linear model for an $n \times 1$ response variable $y_{n\times 1} = (y_1, y_2, \ldots, y_n)'$ is given by

$$y = X\beta + \epsilon, \tag{3.1}$$

where X is an $n \times (p+1)$ design matrix whose i^{th} row is $(1, x_{i1}, x_{i2}, \ldots, x_{ip})$ with $i = 1, 2, \ldots, n$, β is a $(p+1) \times 1$ vector of parameters $(\beta_0, \beta_1, \ldots, \beta_p)'$ and In linear models, parameter estimation is often done using the method of least-squares (Olsson, 2002). The least-square estimator $\hat{\beta}$ of β, which is also the maximum likelihood estimator (MLE) if the independent errors assumptions hold, is given by

$$\hat{\beta} = (X'X)^{-1}X'y \sim N_{p+1}(\beta, \sigma^2(X'X)^{-1}),$$

where $(X'X)^{-1}$ is the inverse of $X'X$ and if $(X'X)$ is not of full rank this inverse is replaced by a generalized inverse $(X'X)^{-}$.

The sampling distribution of $\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(X'X)^{-1})$, is used to test the hypotheses about β. In the case where the normal errors assumption and the central limit theorem conditions are

not satisfied, linear models are not applicable, so generalized linear models are used instead to model the data.

3.2 Generalized Linear Models

In this section we look briefly at the theory of generalized linear models (GLMs). According to Olsson (2002), generalized linear models refers to a generalization of linear models whereby assumptions made in linear models are relaxed to allow for any distribution that is a member of the exponential family of distributions such as the Poisson, the Normal, the Binomial and the Gamma. That is, in linear models of the form

$$\mathrm{E}(Y_i) = \mu_i = x_i'\beta; Y_i \sim N(\mu_i, \sigma^2),$$

we assume that the Y_i's are random variables that are independent normal and form the basis of most analyses of continuous data, and x_i' is the i^{th} row of a design matrix X. Advancement in computing and the statistical theory allow us to use methods comparable to those developed for linear models in the following more general situations:

1. Outcome variables have other distributions beside the Normal distribution, and can be categorical rather than continuous.

2. The relationship between the outcome variable and the predictor variables can be of any form other than the simple linear form in (3.1) with the identity link.

The theory of generalized linear models is discussed in more details by Dobson and Barnett (2008), McCullagh and Nelder (1989), Cantoni and Ronchetti (2001), and Olsson (2002) among others.

Generalized linear models are a class of models that comprise many well known distributions that fall in the exponential family of distributions. Probability density functions in the

exponential family can generally be expressed as

$$f(y_i, \theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right), \quad i = 1, 2, \ldots, n, \tag{3.2}$$

where $a_i(\phi) = \frac{\phi}{w_i}$ which a ratio of the dispersion parameter ϕ and w_i the weight specific for observation y_i, θ_i is the natural parameter and $b(\theta_i)$ is the normalizing function. The function $b(\theta_i)$ describes the relationship between the mean value and the variance in the distribution. That is, the mean $\mathrm{E}(y_i)$ and the variance $\mathrm{var}(y_i)$ can be obtained by finding the first and the second derivatives of $b(\theta_i)$ with respect to θ_i. According to the likelihood theory, it follows that

$$\mathrm{E}\left(\frac{\partial l}{\partial \theta}\right) = 0, \tag{3.3}$$

and that

$$\left(\frac{\partial^2 l}{\partial \theta_i^2}\right) + \mathrm{E}\left[\left(\frac{\partial l}{\partial \theta_i}\right)^2\right] = 0, \tag{3.4}$$

Therefore from (3.2) we get $l(\theta_i, \phi; y_i) = (y_i\theta_i - b(\theta_i))/a_i(\phi) + c(y_i, \phi)$ so that

$$\mathrm{E}\left(\frac{\partial l}{\partial \theta_i}\right) = \mathrm{E}\{[y_i - b'(\theta_i)]/a_i(\phi)\} = 0, \tag{3.5}$$

hence

$$\mathrm{E}(y_i) = \mu_i = b'(\theta_i). \tag{3.6}$$

That is, the mean is obtained from the first derivative of $b(\theta_i)$ with respect to θ_i. By definition, the variance of y_i is

$$\mathrm{var}(y_i) = \mathrm{E}[(y_i) - \mathrm{E}(y_i)]^2.$$

Therefore, the variance can be obtained from (3.5) and using the result in (3.6) to get

$$-\frac{b''(\theta_i)}{a_i(\phi)}+\frac{\text{var}(y_i)}{a_i^2(\phi)}=0, \tag{3.7}$$

so that

$$\text{var}(y_i)=a_i(\phi).b''(\theta_i). \tag{3.8}$$

The variance is a product of two terms $a_i(\phi)$, with ϕ being the dispersion parameter, and $b''(\theta_i)$ called the variance function. The linear predictor is given by is

$$\eta_i=x_i'\beta=(1,x_{i1},x_{i2},\ldots,x_p)\beta,\quad i=1,2,\ldots,n,$$

where x_i' is the i^{th} row of the design matrix.
Therefore $\eta_i=g(\mu_i)$ links the mean $\mu_i=\text{E}(y_i)$ to the linear predictor $x_i'\beta$ as follows

$$g(\mu_i)=x_i'\beta,\quad i=1,2,\ldots,n.$$

3.2.1 Parameter Estimation

In generalized linear models, the estimation of parameters can be done by the maximum likelihood approach or by the method of least squares (Dobson and Barnett, 2008); but often done by the method of maximum likelihood if the response distribution is known (Myers et al., 2012). The **likelihood function** $L(\theta;y)$ with θ and $y=[Y_1,\ldots,Y_n]'$ is algebraically the same as the joint probability function $f(y;\theta)$, that is Equation 3.2, with the change in notation reflecting a shift of emphasis from the random variables y, with the θ fixed, to the parameters θ with y fixed (Dunteman and Ho, 2006). The likelihood function is given by

$$L(\theta; y) = f(y; \theta)$$
$$= \prod_{i=1}^{n} \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right), \tag{3.9}$$

and therefore the log-likelihood function of (3.9) is

$$\ell = \log L(\theta; y) = \ell(\theta; y) = \sum_{l=1}^{n} \ell_i, \tag{3.10}$$

where

$$\ell_i = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi). \tag{3.11}$$

Estimates of β can be obtained by differentiating the log-likelihood function with respect to β_j and equating the derivatives to zero, and then solving the system of equations simultaneously for β_j as follows;

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = 0, j = 0, 1, 2, \ldots, p. \tag{3.12}$$

By the use of chain rule of differentiation, we can obtain $\frac{\partial \ell_i}{\partial \beta_j}$ as

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}, \tag{3.13}$$

called the score function and from (3.11) it can be seen that

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)},$$

since $\mu_i = b'(\theta_i)$ from (3.6) and

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta) = V(\mu_i) \qquad \text{or} \qquad \frac{\partial \theta_i}{\partial \mu_i} = [V(\mu_i)]^{-1},$$

where $V(\mu_i) = b''(\mu_i)$ is obtained by differentiating μ_i with respect to θ_i. Now, from the linear

predictor $\eta_i = g(\mu_i) = x_i'\beta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ we have

$$\frac{\partial \eta_i}{\partial \mu_i} = g'(\mu_i) \qquad \text{or} \qquad \frac{\partial \mu_i}{\partial \eta_i} = [g'(\mu_i)]^{-1}.$$

Given

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij},$$

then $\frac{\partial \ell_i}{\partial \theta_i}, \frac{\partial \theta_i}{\partial \mu_i}, \frac{\partial \mu_i}{\partial \eta_i}$ and $\frac{\partial \eta_i}{\partial \beta_j}$ in (3.13) gives

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \beta_j} &= \frac{y_i - \mu_i}{a_i(\phi)}[V(\mu_i)]^{-1}[g'(\mu_i)]^{-1}x_{ij} \\
&= \frac{(y_i - \mu_i)x_{ij}}{a_i(\phi)V(\mu_i)g'(\mu_i)} \\
&= \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)g'(\mu_i)},
\end{aligned}$$

since $\text{var}(Y_i) = a_i(\phi)V(\mu_i)$, where $V(\mu_i) = \text{var}(\mu_i) = b''(\theta_i)$. Hence, to obtain the β_j's we solve the system of equations also called the score functions which are given as

$$U_j = U(\beta_j) = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)g'(\mu_i)}, \; j = 0, 1, \ldots, p. \tag{3.14}$$

The U_j's variance-covariance matrix has elements given by

$$\mathcal{I}_{jk} = \text{E}[U_j U_k],$$

from which we form the **information matrix** $\mathcal{I}$ which is referred to as the Fisher's information matrix $\mathcal{I}$, with $(j,k)^{th}$ elements of $\mathcal{I}$ given as

$$-\text{E}\left(\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right).$$

Now, from (3.14)

$$\mathcal{I}_{jk} = \mathrm{E}\left\{ \sum_{i=1}^{N} \left[\frac{(y_i - \mu_i)x_{ij}}{\mathrm{var}(Y_i)g'(\mu_i)} \right] \sum_{i=1}^{N} \left[\frac{(y_i - \mu_i)x_{ik}}{\mathrm{var}(Y_i)g'(\mu_i)} \right] \right\}$$
$$= \sum_{i=1}^{N} \frac{\mathrm{E}[(y_i - \mu_i)^2]x_{ij}x_{ik}}{[\mathrm{var}(Y_i)g'(\mu_i)]^2}, \tag{3.15}$$

since $\mathrm{E}[(y_i - \mu_i)(y_k - \mu_k)] = 0$ for $i \neq k$ because the y_i's are independent. By the use of $\mathrm{E}[(y_i - \mu_i)^2] = \mathrm{var}(Y_i)$, (3.15) can be simplified to

$$\mathcal{I}_{jk} = \sum_{i=1}^{N} \frac{x_{ij}x_{ik}}{\mathrm{var}(Y_i)[g'(\mu_i)]^2}. \tag{3.16}$$

Notice that the information matrix, $\mathcal{I}$ at β, has the following relationship with U at β

$$\mathcal{I}_{jk} = \mathrm{E}[U_j U_k] = -\mathrm{E}\left(\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} \right).$$

To obtain $\mathcal{I}$, the iterative Newton-Raphson formula which generalizes to

$$b^{(m)} = b^{(m-1)} + [\mathcal{I}^{(m-1)}]^{-1} U^{(m-1)}, \tag{3.17}$$

is used, where $b^{(m)}$ is the vector of estimates $\hat{\beta}_1, \ldots, \hat{\beta}_p$ at the mth iteration, $[\mathcal{I}^{(m-1)}]^{-1}$ is the inverse of the information matrix with elements $\mathcal{I}_{jk}$, and $U^{(m-1)}$ is the vector of elements given by U_j, all evaluated at $b^{(m-1)}$. If we multiply both sides of 3.17 by $\mathcal{I}^{(m-1)}$ we get

$$\mathcal{I}^{(m-1)} b^{(m)} = \mathcal{I}^{(m-1)} b^{(m-1)} + U^{(m-1)}. \tag{3.18}$$

.

Now from (3.16) $\mathcal{I}$ can be written as

$$\mathcal{I} = X'WX,$$

where W is an $N \times N$ diagonal matrix with elements

$$w_{ii} = \frac{1}{\text{var}(Y_i)[g'(\mu_i)]^2}. \tag{3.19}$$

The right-hand side of (3.18) has the expression

$$\sum_{k=1}^{p}\sum_{i=1}^{N} \frac{x_{ij}x_{ik}}{\text{var}(Y_i)[g'(\mu_i)]^2} b_k^{(m-1)} + \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)[g'(\mu_i)]^2},$$

evaluated at $b^{(m-1)}$; which follows from (3.14) and (3.16), hence the right hand-side of (3.18) is therefore

$$X'Wz,$$

where z has elements

$$z_i = \sum_{k=1}^{p} x_{ik} b_k^{(m-1)} + (y_i - \mu_i)g'(\mu_i), \tag{3.20}$$

with μ_i and $g'(\mu_i)$ evaluated at $b^{(m-1)}$. Thus the iterative equation (3.18), can be written as

$$X'WX\beta^{(m)} = X'Wz, \tag{3.21}$$

which has the same form as the normal equations for a linear model obtained by weighted least squares, but since z and W generally depend on b, the maximum likelihood estimators for generalized linear models are obtained by the **iterative weighted least squares** method. Statistical packages that have commands for fitting GLMs have an algorithm based on (3.21). First, the initial approximation $b^{(0)}$ is obtained to evaluate z and W, then (3.21) is solved to give $b^{(1)}$ which is also used to obtain better approximations for z and W and so on until convergence is achieved. That is, when the difference between the successive approximations

$b^{(m-1)}$ and $b^{(m)}$ is sufficiently small, $b^{(m)}$ is taken to be the maximum likelihood estimate. The theory above is discussed in a detailed manner by Dobson and Barnett (2008), Gill (2000) and Christensen (2006) among others.

For example if the log linear model is assumed

$$\ell(\mu, \phi; y) = \sum_{i=1}^{n} \left[\phi^{-1} y_i \log \mu_i - \mu_i - C(y_i) \right]. \tag{3.22}$$

then $\eta_i = \log \mu_i = X'\beta$, then we get $\frac{\partial \eta_i}{\partial \mu_i} = \mu_i^{-1}$, $w_i = \phi^{-1}\mu_i, \mu_j = \phi^{-1} \sum_{i=1}^{n} \mu_i x_{ji} \frac{y_i - \mu_i}{\mu_i}$ and $I = \phi^{-1} \sum_{i=1}^{n} \mu_i x_{ji} x_{ki} = \phi^{-1}(X'WX)$, where $W = \text{diag}(\mu_i)$. Thus, following the procedure discussed in 3.2.1, the Fisher scoring method for obtaining the maximum likelihood estimates of the parameters β with $\phi = 1$ is

$$\beta^{(m+1)} = (X'W^m X)^{-1} X W^m z^m. \tag{3.23}$$

In the next subsection we discuss briefly the Poisson and the Negative Binomial distributions. It is worth noting that we want to analyze microbiological counts, and therefore as a result we make use of Poisson and Negative Binomial count models.

3.2.2 The Poisson Regression Model

Suppose that Y_i is a random variable representing counts with means μ_i, the simple Poisson distribution is the probability function given by

$$\text{p}(y_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, \tag{3.24}$$

for $\mu_i > 0$. The mean and variance of this distribution can be shown to be

$$\text{E}(y_i) = \text{var}(y_i) = \mu_i.$$

The Poisson mass function can be written as a member of the exponential family of distributions as follows

$$\begin{aligned} f(y_i;\mu_i) &= \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \\ &= \exp[y_i \log(\mu_i) - \mu_i - \log(y_i!)]. \end{aligned} \tag{3.25}$$

This expression can be compared to Equation 3.2 where $\theta_i = \log(\mu_i)$; meaning that $\mu_i = \exp(\theta_i)$;so that Equation 3.25 is as follows

$$f(y_i;\mu_i) = \exp[y_i\theta_i - \exp(\theta_i) - \log(y_i!)], \tag{3.26}$$

thus $\theta_i = \log(\mu_i)$, $b(\theta_i) = \exp(\theta_i)$, $c(y_i, \phi) = -\log(y_i!)$ and $a_i(\phi) = 1$.

The Poisson log-linear regression model, from the expression in Equation 3.25, can be obtained by taking the log as follows

$$\log f(y_i;\mu_i) = y_i\theta_i - \exp(\theta_i) - \log(y_i!). \tag{3.27}$$

In generalized linear models the function $b(\cdot)$ describes the relationship between the mean and the variance in the distributions. For a Poisson model and using Equation 3.6 the mean is

$$b'(\theta_i) = \exp(\theta_i) = \mu_i,$$

and the variance is

$$\text{var}(y_i) = a_i(\phi).b''(\theta_i) = \exp(\theta_i) = \mu_i,$$

which verifies that the mean and the variance function of the Poisson distribution are equal.

3.2.3 Parameter Estimation for Poisson Log-linear Model

Suppose that y_i's are independent and identically distributed observations from a Poisson distribution with unknown parameter μ. The log-likelihood for the Poisson regression model is

$$\ell(\mu; y) = \sum_{i=1}^{n} y_i \log(\mu_i) - n\mu_i. \tag{3.28}$$

Equating the above expression to zero and solving for μ gives the MLE for μ,

$$\hat{\mu} = \sum_{i=1}^{n} \frac{y_i}{n},$$

as the sample mean. To obtain the vector of parameters β_i's we solve the system of equations also called the score functions given in Equation 3.14.

3.3 Overdispersion

Overdispersion (underdispersion) means that the variance is greater (less) than the mean (Cameron and Trivedi, 2013). When there is greater variability in the data than would be expected by the Poisson regression model, then there is overdispersion (i.e $\phi > 1$) in the data. That is, the variances of Y_i are greater than their expected values. According to Hinde and Demétrio (1998), overdispersion might result from variability of experimental material which can be thought of as individual variability of the experimental units, or result from correlation between individual responses or cluster sampling or maybe aggregate level data which is the aggregation process and can lead to compound distributions, and/or omitted unobserved variables.

Hinde and Demétrio (1998) state that overdispersion, when ignored, can lead to incorrect standard errors obtained from the model and may be underestimated and consequently leading to incorrect assessment of the significance of individual regression parameters.

It is worth noting that empirical count data sets typically exhibit over-dispersion and/or excess number of zeros and therefore the use of classical Poisson regression model become limited. That is, the requirement that the mean and variance of the standard Poisson regression model are equal is hardly ever met. In such cases, alternative regression models such as overdispersed Poisson regression (quasi-Poisson) and negative binomial models can be used.

Overdispersion parameter ϕ can be estimated separately using the method of moments estimator as follows

$$\hat{\phi} = \frac{1}{n-p-1} \sum \frac{(y_i - \hat{\mu_i})^2}{\text{var}(y_i)},$$

and the estimated asymptotic covariance matrix of coefficients $\hat{\beta}$ is given by

$$\text{cov} = \hat{\phi}(X'WX)^{-1},$$

where W is a diagonal matrix of weights w_i.

The next subsection introduces the Negative Binomial distribution.

3.3.1 The Negative Binomial Regression (NB) Model

To model count data with overdispersion we use the NB distribution by adding a multiplicative random effect parameter to represent an unobserved variation.

Let the random variables y_i represent counts and $\text{E}(\theta_i) = \mu_i$ and $\text{var}(\theta_i) = \sigma^2$, where θ_i's are random variables. According to Hinde and Demétrio (1998), unconditionally, $\text{E}(y_i) = \mu_i$ and $\text{var}(y_i) = \mu_i + \sigma^2$ giving an overdispersed model. A common assumption is that $\theta_i \sim \Gamma(k, \lambda_i)$

distribution and leads to a negative binomial distribution. Therefore, the negative binomial distribution can be written as follows

$$f(y_i;\mu_i,k)=\frac{\Gamma(k+y_i)\mu_i^{y_i}k^k}{\Gamma(k)y_i!(\mu_i+k)^{k+y_i}},\quad y_i=0,1... \tag{3.29}$$

with mean

$$\mathrm{E}(y_i)=\frac{k}{\lambda_i}=\mu_i, \tag{3.30}$$

and the variance

$$\begin{aligned}\mathrm{var}(y_i)&=\mathrm{E}[\mathrm{var}(y_i|\theta_i)]+\mathrm{var}(\mathrm{E}[y_i|\theta_i])\\&=\mathrm{E}[\theta_i]+\mathrm{var}(\theta_i)\\&=\frac{k}{\lambda_i}+\frac{k}{\lambda_i^2}\\&=\mu_i+\frac{\mu_i^2}{k}.\end{aligned} \tag{3.31}$$

Parameter Estimation under NB

The log-likelihood under negative-binomial model has the expression

$$\begin{aligned}\ell(\mu,k;y)&=\sum_{i=1}^{n}\left\{y_i\log\mu_i+k\log k-(k+y_i)\log(k+\mu_i)+\log\frac{\Gamma(k+y_i)}{\Gamma(k)}-\log y_i!\right\}\\&=\sum_{i=1}^{n}\left\{y_i\log\mu_i+k\log k-(k+y_i)\log(k+\mu_i)+\mathrm{dlg}(y_i,k)-\log y_i!\right\},\end{aligned} \tag{3.32}$$

where $\mathrm{dlg}(y_i,k)=\log\Gamma(k+y_i)-\log\Gamma(k)$ and for fixed values of k the expression becomes a linear exponential family model and consequently a generalized linear model (Hinde and Demétrio, 1998).

The score equations for maximum likelihood estimations can be obtained by modelling the μ_i's with a linear predictor $\eta_i = x_i'\beta$ and the link function $g(\mu_i) = \eta_i$ as follows

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^{n}\left\{\frac{y_i}{\mu_i} - \frac{k+y_i}{k+\mu_i}\right\}\frac{\partial \mu_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n}\frac{(y_i-\mu_i)}{\mu_i(1+\frac{\mu_i}{k})}\frac{1}{g'(\mu_i)}x_{ij},\end{aligned} \tag{3.33}$$

and

$$\frac{\partial \ell}{\partial k} = \sum_{i=1}^{n}\left\{\mathrm{ddg}(y_i,k) - \log(\mu_i+k) - \frac{k+y_i}{k+\mu_i} + \log k + 1\right\}, \tag{3.34}$$

where $\mathrm{ddg}(y_i,k) = \frac{\partial}{\partial k}(\mathrm{dlg}(y_i,k)) = \psi(y_i+k) = \psi(k)$.

The scores $V(\mu) = \mu(1+\frac{\mu}{k})$ and $g(\mu) = \eta$ for β are the usual quasi-score equations for a generalized linear model and thus provide a simple approach for fitting a negative-binomial regression models using a Gauss-Seidel approach and iterating using the following steps

1. for fixed k, estimate β using an iterative re-weighted least squares with a variance function $V(\mu) = \mu + \frac{\mu^2}{k}$,
2. for fixed β, and hence μ, k is estimated using the Newton-Raphson iterative scheme

$$k^{(m+1)} = k^{(m)} - \left(\frac{\partial \ell}{\partial k}\Big/\frac{\partial^2 \ell}{\partial k^2}\right)\Bigg|_{k^{(m)}},$$

and continue iterating until convergence. The second order derivative with respect to k is

$$\frac{\partial^2 \ell}{\partial k^2} = \sum_{i=1}^{n}\left\{\mathrm{dtg}(y_i,k) - \frac{1}{\mu_i+k} + \frac{k+y_i}{(k+\mu_i)^2} - \frac{1}{\mu_i+k} + \frac{1}{k}\right\}, \tag{3.35}$$

where $\mathrm{dtg}(y,k) = \partial\{\mathrm{ddg}(y,k)\}/\partial k$ is a function of tri-gamma functions. The derivative

$$\frac{\partial^2 \ell}{\partial \beta_j \partial k} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{(k + \mu_i)^2} \frac{1}{g'(\mu_i)} x_{ij}, \tag{3.36}$$

and $\mathrm{E}\left(\frac{\partial^2 \ell}{\partial \beta_j \partial k}\right) = 0$, that is β and k are asymptotically uncorrelated.

The initial values for k can be obtained from fitting a Poisson model to obtain $\hat{\mu_i}$ and setting

$$k_0 = \frac{\sum_{i=1}^{n} \hat{\mu_i}(1 - h_i \hat{\mu_i})}{\sum_{j=1}^{n} \frac{(y_j - \hat{\mu_j})^2}{\hat{\mu_j}} - (n - p)}.$$

where $h_i = \mathrm{var}(\hat{\beta}' x_i)$ is the variance of the linear predictor (Breslow, 1984).

3.3.2 Quasi-Likelihood Estimation

The method of quasi-likelihood estimation is used when there is uncertainty about the distribution of the data. This uncertainty makes it impossible to directly use the techniques discussed earlier. The basic idea behind the quasi-likelihood estimation method is to use inferential methods which work as almost as well as maximum likelihood but without having to make specific distributional assumptions. That is, a likelihood that has less restrictive assumptions. Let us restate the score equation as

$$U = \frac{\partial \ell}{\partial \beta_i} = \frac{1}{a(\phi)} \sum_{i=1}^{N} \frac{(y_i - \mu_i) x_{ij}}{\mathrm{var}(Y_i) g'(\mu_i)}$$

The likelihood on the assumed distribution for y_i is constructed through μ_i and $\mathrm{var}(\mu_i)$ respectively. Notice that the choice of the distribution determines the mean-variance relationship. Also, it is worth noting that the probability distribution is not specified (unlike in full likelihood estimation methods) but only the mean and variance function are specified.

By definition, the quasi-likelihood is defined as

$$Q_i(\mu_i; y_i) = \int_{y_i}^{\mu_i} \frac{y_i - \mu_i}{\phi \text{var}(\mu_i)} \text{d}s$$

and by definition has the derivative with respect to μ_i which is equal to

$$q_i = \frac{y_i - \mu_i}{\phi \text{var}(\mu_i)}.$$

The q_i satisfies the same conditions satisfied by $\partial \ell / \partial \mu_i$, where

$$\partial \ell / \partial \mu_i = \frac{\partial \log f(y_i, \theta_i, \phi)}{\partial \mu_i}$$

for the exponential family of distributions.

The quasi-likelihood for the complete data is the sum of the individuals contributions

$$Q(\mu, y) = \sum Q_i(\mu_i, y_i),$$

since the components of Y are independent by assumption.

By analogy, notice that the quasi-deviance function for a single observation is given in the reversed order of integration as follows

$$Q(y_i; \mu_i) = -2\sigma^2 Q(\mu_i; y_i) = 2 \int_{y_i}^{\mu_i} \frac{y_i - \mu_i}{\phi \text{var}(\mu_i)} \text{d}s.$$

The total deviance, $D(y_i; \mu_i)$, is the sum of the individual components, and depends only on y and μ, but not on σ^2. It is worth noting that the quasi-likelihood has a multiplicative dependence on σ^2, and does not affect the maximum likelihood estimators of β.

Recall from earlier that the log-likelihood of the exponential family is given as

$$\ell_i = \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi).$$

It can be shown that

$$\mathrm{E}(q_i) = 0, \quad \mathrm{E}\left(\frac{\partial \ell_i}{\partial \mu_i}\right)$$

and

$$\mathrm{var}\left(\frac{\partial \ell_i}{\partial \mu_i}\right) = \frac{1}{\phi \mathrm{var}(\mu_i)}.$$

Thus the log-likelihood and $\frac{\partial \ell_i}{\partial \mu_i}$ played roles that can be taken up by Q_i and q_i respectively.

The ϕ in q_i is the constant of proportionality relating the $\mathrm{var}(y_i)$ to $\mathrm{var}(\mu_i)$. The maximum quasi-likelihood method assume that these variances are proportional, that is $\mathrm{var}(y_i) = \phi \mathrm{var}(\mu_i)$, where

$$\mathrm{var}(\mu_i) = \frac{\partial \ell_i}{\partial \mu_i}.$$

It is worth noting that the variance function $\mathrm{var}(\mu_i)$ is specified using the information about how the variance changes with the mean.

The maximum quasi-likelihood estimator of β can be obtained by solving the maximum quasi-likelihood given by

$$\frac{\partial}{\partial \beta}(\sum_{i=1}^{N} Q_i) = 0. \tag{3.37}$$

The above Equation 3.37 can be evaluated as follows:

$$\frac{\partial}{\partial \beta}\left(\sum_{i=1}^{N} Q_i\right) = \sum_{i=1}^{N} \frac{\partial Q_i}{\partial \beta} \tag{3.38}$$

$$= \sum_{i=1}^{N} \frac{\partial Q_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta} \tag{3.39}$$

$$= \sum_{i=1}^{N} \left[\frac{y_i - \mu_i}{\phi \mathrm{var}(\mu_i)}\right] \frac{\partial \mu_i}{\partial q} \frac{\partial q}{\partial \beta} \tag{3.40}$$

$$= \sum_{i=1}^{N} \left[\frac{y_i - \mu_i}{\phi \mathrm{var}(\mu_i)}\right] \frac{x_{ij}}{g'(\mu_i)} = 0 \tag{3.41}$$

In matrix notation the above expression can be presented as follows

$$\frac{1}{\phi}X'W\delta(Y-\mu)=0 \tag{3.42}$$

Equation 3.42 is score statistic U under the GLM, with $\text{var}(\mu_i)$ determined by the mean-variance relationship, not by the distributional assumptions. Therefore, the quasi-likelihood model can be fitted using exactly the same method as for fitting a GLM to obtain the estimates $\hat{\beta}$.

3.4 The Distributed Lag Model

In this section we incorporate time by introducing the distributed lag model (DLM). The DLM generally says that the effect of a unit increase in water quality determinant x_t is spread out over K number of days into the future (Peng and Dominici, 2008).

In time series studies of air pollution and health, the outcome is modeled as a time series of counts representing the number of times a particular event has occurred on a given day (Peng and Dominici, 2008). In this project, a microbiological outcome is modeled as a time series of counts representing the number of colonies found in a 100mL water sample on a given day. Therefore each observation of the outcome y_t is a count per sample on day t. However, it is worth noting that the assumption that the effect of a unit increase in the water-quality determinant only plays out over a single day is relaxed by introducing the distributed lag model. Therefore, for time series count response data, the log-linear Poisson model similar to that used by Peng and Dominici (2008) takes the form

$$y_t \sim \text{Poisson}(\mu_t),$$

where

$$\log\mu_t=\beta_0+\sum_{i=1}^{p}\sum_{\ell=0}^{1}\beta_{i\ell}X_{t-\ell}, \tag{3.43}$$

and β_0 is the model intercept, $\beta_{i\ell}$ is the i^{th} parameter at lag ℓ, $X_{t-\ell}$ is the known vector of water-quality determinants that is included at lag ℓ. Appropriate number of lags to be included in the model is, however, a problem that generally need a subject matter knowledge (Peng and Dominici, 2008).

Suppose that we observe X_t from 1991 to 2015. The lagged value of the independent variable X_t is the same value but for the previous period, that is, 1990 to 2014. Note that since 1990 would not have been observed, the start date or time for the lagged value $X_{t-\ell}$ would have to be from 1991. Lagging once means that the current X_t series will have to start at 1992 and end at 2015. In practical this means that when we lag once we loose one observation from the data, whilst on the other hand one extra parameter β is estimated with every lag. This is jeopardy with respect to loss of degrees of freedom. Apart from the loss of degrees of freedom, regressors X_t's are often highly correlated with their lagged values and thus introducing the problem of multicollinearity among the regressors X_t. The higher the multicollinearity the lower the reliability of the regression estimates (Baltagi, 2011).

The problem of multicollinearity can be dealt with using the methods used by Almon (1965). Suppose we want to fit the data using a finite DLM. A finite DLM is one in which the number of lags to be included in the model are known (Almon, 1965). According to Almon (1965), Weierstrass's Approximation Theorem can be used to approximate a continuous function defined on a closed form interval by a polynomial function of finite degree. For further details on DLM you can refer to Almon (1965); Baltagi (2011) among others.

3.5 Application of GLM Quasi-Poisson Distributed Lag Model to Water Quality Data

Quasi-Poisson models are used to allow the dispersion parameter to be estimated from the data since under the Poisson model the variance is assumed to be equal to the mean. If this assumption holds then the sample mean and variance of the dependent count variable should

be equal.

In this section the quasi-Poisson distributed lag model is fitted to the microbiological count data. When the quasi-Poisson is fitted to the data the dispersion parameter was estimated as $\hat{\phi} = 158$ from the data. This implied that the data is overdispersed and hence other models should be considered to account for overdispersion.

The distributed lag model assumes that the effect of a unit increase in a water-quality determinant on any given day is spread out over a number of days into the future. The seasonal time variable has four levels where the winter season is a reference category because it is expected to have the lowest bacterial counts than the other seasons. Also, a site variable with three levels is used, where site 3 (TDH010) is used as a reference category because it is a site which usually has the lowest counts than the other two sites.

Table 3.1: Results for a Quasi-Poisson Distributed Lag Model Fitted to HPC37 counts, Durban-Heights, 1991-1997.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	38.2022	13.7421	2.78	0.0055
Temperature at lag 0	-0.0748	0.1301	-0.57	0.5654
Temperature at lag 1	0.1285	0.1281	1.00	0.3158
Turbidity at lag 0	0.4896	0.3811	1.28	0.1989
Turbidity at lag 1	0.6036	0.2864	2.11	0.0351
pH at lag 0	-2.3822	1.4808	-1.61	0.1077
pH at lag 1	-2.1758	1.4477	-1.50	0.1329
Free Chlorine at lag 0	-4.2663	1.7789	-2.40	0.0165
Free Chlorine at lag 1	-2.6950	2.2524	-1.20	0.2316
Total Chlorine at lag 0	-0.8672	1.8323	-0.47	0.6360
Total Chlorine at lag 1	1.0353	2.0138	0.51	0.6072
Rainfall at lag 0	0.0078	0.0135	0.58	0.5620
Rainfall at lag 1	-0.0159	0.0244	-0.65	0.5147
Summer	1.1285	1.1149	1.01	0.3115
Autumn	2.1430	1.0152	2.11	0.0348
Spring	1.5488	0.9264	1.67	0.0946
TDH007	0.3753	0.4806	0.78	0.4350
TDH008	0.2663	0.4919	0.54	0.5882

Table 3.1 shows the results from the DL GLM of HPC37 counts to all water quality determinants of interest in this study.

Note:

$$\beta_i = i^{th} \text{ parameter}$$
$$\beta_{i\ \ell} = i^{th} \text{ parameter at lag}\ell$$

This project uses only two lags, that is lag zero and lag one. A seasonal-time variable and a site variable are dummy variables introduced to capture time effects and site specific effects. Table 3.1 shows the results of only the first 2000 observations from each of the three Durban-Heights Final sites merged together to give a total of 6000 observations.

Turbidity at lag one was found to be significant at the 5% significance level (p-value = 0.0351), which means that the effects of turbidity measured or observed on any day is expected to be seen one day later. Therefore, on any given day of the week, a one unit increase in nephelometric turbidity units (NTU) would increase HPC37 by a factor of $\exp(\hat{\beta}_{4\ 1} = 0.603) = 1.828$ colony forming units per mL (CFU/mL) of water sample one day later. Given the sampling day, a 1 mg/L increase in free chlorine would decrease HPC37 counts by a factor of $\exp(\hat{\beta}_{7\ 0} = -4.266) = 0.0014$ CFU/mL of water sample on the same day. The quasi-Poison model in Table 3.1 also shows that, given the reference category as the winter season, the autumn season has a significantly higher HPC37 CFU/mL counts (p-value = 0.0348) than in winter season [$\exp(\hat{\beta}_{14} = 2.143) = 8.525$]. In short, this means increased levels of turbidity would results in higher HPC37 positive counts. Free Chlorine reduces the HPC37 counts. Table 3.1 also shows that there are non-significant $\exp(\hat{\beta}_{16} = 0.375) = 1.455$ and $\exp(\hat{\beta}_{17} = 0.266) = 1.305$ higher HPC37 positive counts in Durban Heights Final 1 and 2 relative to Durban Heights Final 3.

Total Coliforms Predictions

On the other hand, the quasi-Poisson model does not give any predictors for Total Coliforms (TC) due to the proportion of counts being very small compared to that of HPC37. The AIC suggest that the quasi-Poisson model is the best model, but the ICC under Poisson GLMM

suggest that TC are greatly influenced by unobserved factors at the level of the sampling point and therefore this model brings about a substantial improvement compared to the quasi-Poisson model, thus the result from Poisson GLMM are therefore interpreted. These results are discussed in Chapter 4. Before determining that the quasi-Poisson family is appropriate, we check to see if the variance of the residuals is proportional to the mean. The plot of the residuals squared is shown in Figure 3.1.

3.5.1 Model Validation in Quasi-Poisson GLM

In model validation, residuals are an important tools. They are used to check violations of assumptions such as that of homogeneous variances and provide guidance concerning the adequacy of the model.

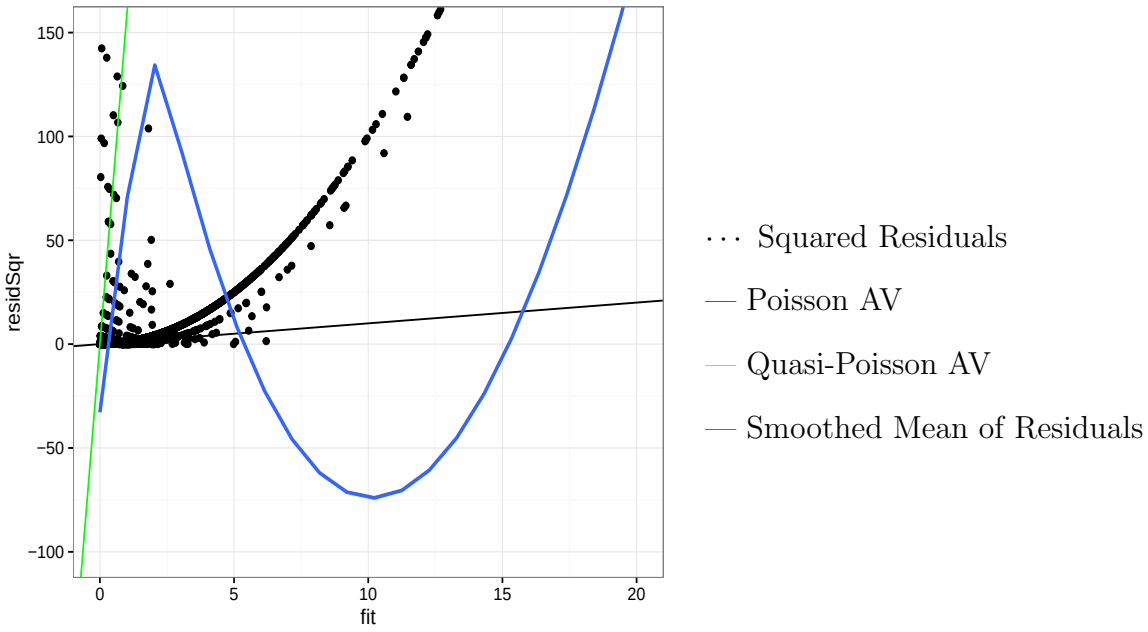

Figure 3.1: Mean-Variance Relationship.

In Figure 3.1, the black line represent the Poisson assumed variance (AV), the green line represent the quasi-Poisson AV, and the blue curve represent the smoothed mean of the residuals square.

Ideally the blue curve would be straight and it would be collinear with the green line for the quasi-Poisson variance. The greater the deviation from the green line the greater the concern is about the proportionality of the variance to the mean. Here we have some indication that the variance may not be proportional to the mean. When applying the GLMs with a fixed scale parameter, as is certainly the case for Poisson distribution where $\phi = 1$, subject to certain asymptotic conditions for a well fitting model we would expect

$$\text{residual deviance} \approx \text{residual degrees of freedom}$$

If the residual deviance is greater than residual degrees of freedom we are faced with two possible scenarios to consider. The first scenario is a bad fitting model for some reasons such as

- omitting variables in the linear predictor;
- specifying the incorrect link function between the mean and predictor variables;
- outliers.

The second scenario is when variation is greater than that predicted by the model, that is overdispersion. In essence the model is too restrictive for the data at hand. Overdispersion may result from

- individual variability of the experimental units which may give an additional component of variability which is not accounted for by the basic model;
- correlated individual responses;
- omitted unobserved variables.

It is therefore important to take overdispersion into account as this phenomenon renders the results that are not reliable and thus leading to false conclusions. The standard errors obtained from the model will be incorrect and therefore consequently we may end assessing the significance of individual regression parameters that are incorrect. Changes in the deviance become too large leading to selection of overly complex models. Lastly, the interpretation of

the model will be incorrect and thus predictions will not be too precise (Hinde and Demétrio, 1998).

To define residuals in a GLM, consider the case for linear regression in which residuals are defined as

$$\hat{\varepsilon} = y_i - \mu_i, \tag{3.44}$$

which is the vertical distance between an observation and the regression line of the predicted values. The types of residuals that are often used in a GLM used are the ordinary residuals, the Pearson residuals, and the deviance residuals.

The Pearson residuals can be written as follows:

$$\hat{\varepsilon}_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}, \tag{3.45}$$

and for overdispersed count data with ϕ estimated from data the Pearson residuals is divided by the adjusted square root of variance $\phi\mu_i$.

The deviance residuals are said to be GLM equivalent of the residual sum of squares, whereby the smaller the deviance residuals the better the model (Zuur et al., 2009). The contribution of each observation to the residual deviance explain how the model fits the data. The residual deviance are defined by

$$\hat{\varepsilon_i^D} = \text{sign}(y_i - \mu_i)\sqrt{d_i}, \tag{3.46}$$

where sign is positive when y_i is greater than μ_i and negative when y_i is less than μ_i, d_i is the contribution of each observation to the deviance.

There is no much difference between the deviance and the Pearson residuals for a Poisson

GLM, though however, this may not be the case for data with excess of zeros. Deviance residuals are recommend by McCullagh and Nelder (1989) for model checking as these have distributional properties that are closer to the residuals from a Gaussian linear regression model than Non-Gaussian models. It is worth noting that the interest is not on normality of the residuals from the Pearson and deviance residuals, but the lack of fit of the model by looking for patterns in the residuals.

To validate the model we can plot deviance residuals against (i) each explanatory variable in the model, (ii) the fitted values, and (iii) against time, if appropriate. If any patterns are detected in the graph showing residuals against each explanatory in the model, then either quadratic terms should be included, the use of GAM, or the conclusion should be that there is violation of independence. If patterns are detected in graph showing residuals against fitted values, then there is oversdispersion or the wrong use of mean-variance relationship, and if plotting the residuals against time, and there are patterns, conclude there is an assumption of independence violated. Meaning that nearly always an important covariate was excluded from the model. If non of the above seem to be solution then generalized linear mixed model should bale (Zuur et al., 2009).

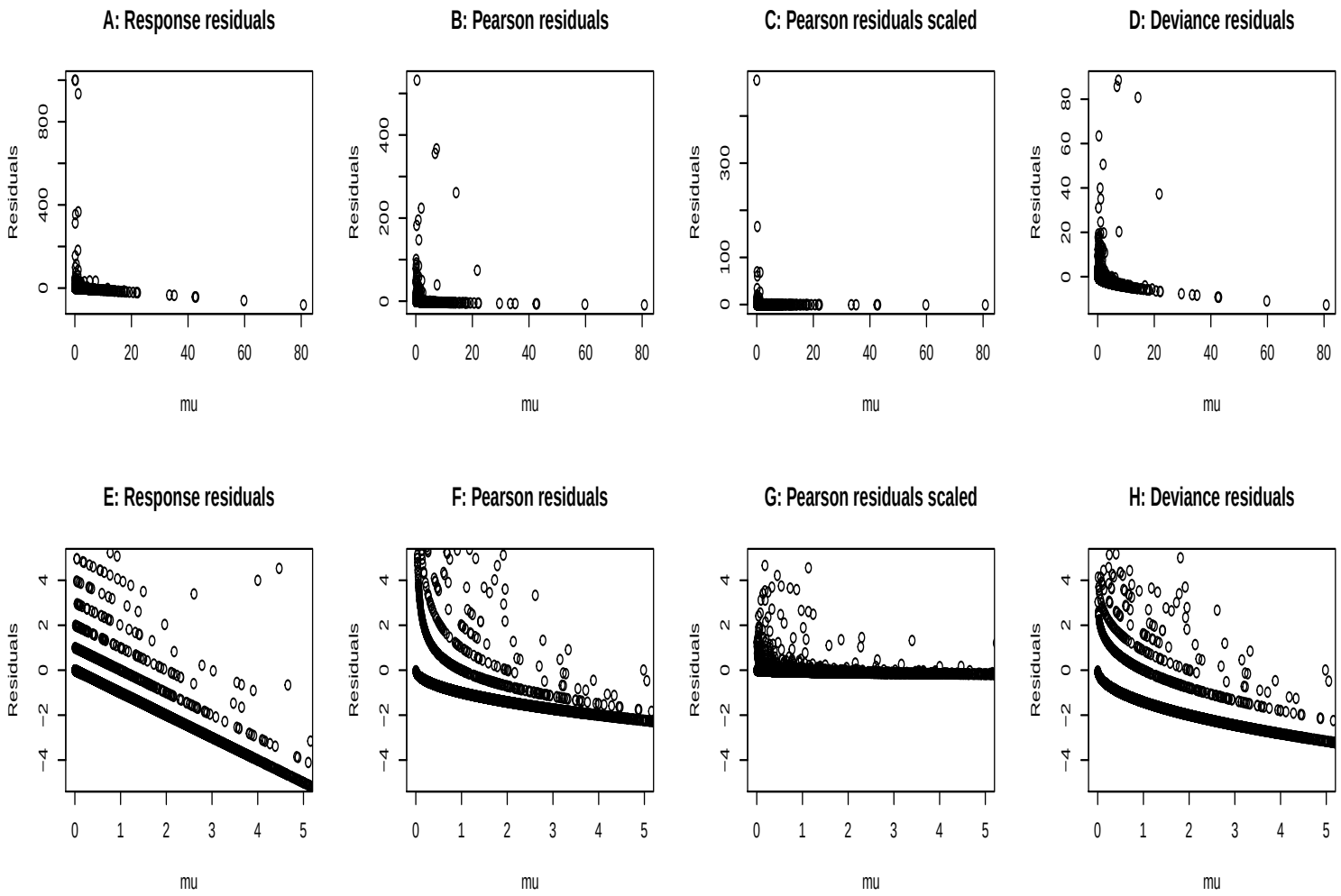

Figure 3.2: The residuals-squared plot versus the predicted mean of the quasi-Poisson model.

Figure 3.2 shows graphs of response residuals, Pearson residuals, Pearson scaled residuals, and deviance residuals, all plotted against the fitted values. The upper row shows the actual residual plots from quasi-Poisson model and the lower row shows the zoomed in plots of the upper residual plots against the fitted values since it is not simple to see if there are any trend or patterns from those in the upper row.

From the upper row, it is clear that there are very high values in the residuals from this model, indicating non-constant variance or the wrong mean-variance relationship. From the lower row, the residuals form some parallel pattern with high variation at low fitted values (mu). Therefore we conclude that there is overdispersion in the data and hence the negative binomial model is introduced. Next we consider the application of a negative binomial model which also account for overdispersion.

The results for predicting TC are shown in Table 3.2. The Quasi-Poisson model does not show any significant predictors of TC. The proportion of positive counts is lower than that of HPC37 and thus this model experiences difficulty compared to one fitted the HPC37 counts.

Table 3.2: Results for a Quasi-Poisson Distributed Lag Model Fitted to TC counts, Durban-Heights, 1991-1997.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3486	42.4353	0.01	0.9934
Temperature at lag 0	0.7266	0.4595	1.58	0.1138
Temperature at lag 1	0.0631	0.4347	0.15	0.8847
Turbidity at lag 0	-0.2425	5.8449	-0.04	0.9669
Turbidity at lag 1	0.8956	1.8268	0.49	0.6240
pH at lag 0	0.4698	4.1384	0.11	0.9096
pH at lag 1	-3.1002	4.3275	-0.72	0.4738
Free Chlorine at lag 0	3.7414	7.8255	0.48	0.6326
Free Chlorine at lag 1	-2.5416	7.4979	-0.34	0.7346
Total Chlorine at lag 0	-1.6485	6.7657	-0.24	0.8075
Total Chlorine at lag 1	0.2759	6.5552	0.04	0.9664
Rainfall at lag 0	-0.0688	0.1784	-0.39	0.7000
Rainfall at lag 1	-0.1084	0.3021	-0.36	0.7197
Summer	-3.6844	3.3676	-1.09	0.2739
Autumn	-4.1065	3.4042	-1.21	0.2277
Winter	-3.1545	3.1970	-0.99	0.3238
TDH007	2.2541	1.7390	1.30	0.1949
TDH008	0.1393	2.3361	0.06	0.9524

3.6 Application of GLM Negative Binomial Distributed Lag Model to Water Quality Data

We now apply the negative binomial model to the data to see if it might be of better fit. It is worth noting that the negative binomial model assumes independence among the observations. Table 3.3 shows the results of the negative binomial model incorporating distributed lag model. Parameter estimates for the negative binomial model are similar to those of the quasi-Poisson model in terms of their signs and their magnitude. However, standard errors in the negative binomial model have been reduced compared to those for a quasi-Poisson model. This is an indication that the negative binomial model captures the variability much better than the quasi-Poisson model. It is an issue of better precision.

Table 3.3: Results for a Negative Binomial Distributed Lag Model Fitted to HPC37 counts, Durban-Heights, 1991-1997.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	23.7794	6.7853	3.50	0.0005
Temperature at lag 0	-0.0485	0.0641	-0.76	0.4490
Temperature at lag 1	0.2526	0.0640	3.95	0.0001
Turbidity at lag 0	1.4641	0.5875	2.49	0.0127
Turbidity at lag 1	1.4128	0.5861	2.41	0.0159
pH at lag 0	-2.0028	0.7425	-2.70	0.0070
pH at lag 1	-1.3577	0.7404	-1.83	0.0667
Free Chlorine at lag 0	0.8144	1.2133	0.67	0.5021
Free Chlorine at lag 1	-2.7320	1.2182	-2.24	0.0249
Total Chlorine at lag 0	-2.5025	1.0040	-2.49	0.0127
Total Chlorine at lag 1	1.0702	1.0124	1.06	0.2905
Rainfall at lag 0	0.0207	0.0088	2.36	0.0184
Rainfall at lag 1	0.0048	0.0092	0.52	0.6011
Summer	-0.1568	0.4125	-0.38	0.7039
Autumn	1.4049	0.3604	3.90	0.0001
Spring	0.3646	0.2768	1.32	0.1878
TDH007	0.2091	0.2011	1.04	0.2985
TDH008	0.4503	0.1920	2.35	0.0190

Parameter estimates for a negative binomial model are interpreted in the similar way as those for a quasi-Poisson model. At the 5% significance level, temperature is significant at lag one (p-value = 0.0001). A one unit increase in temperature on any particular day would increase

HPC37 counts by a factor of $\exp(\hat{\beta}_{2\ 1} = 0.253) = 1.288$ CFU/mL one day later.

Turbidity is significant at both lags zero and one (p-values = 0.0127 and 0.0159). For one unit increase in NTU on any given day, HPC37 counts would increase by a factor of $\exp(\hat{\beta}_{3\ 0} = 1.464) = 4.323$ CFU/mL on the same day and increase by $\exp(\hat{\beta}_{4\ 1} = 1.413) = 4.108$ CFU/mL a day later.

At the 5% significance level pH is significant only at lag zero (p-value = 0.0070). On any given day a one unit increase in pH would decrease HPC37 counts by a factor of $\exp(\hat{\beta}_{5\ 0} = -2.003) = 0.135$ CFU/mL on the same day.

Free chlorine is significant at lag one only (p-value = 0.0249). This means that for a one unit increase in free chlorine on any given day, the HPC37 counts would decrease by a factor of $\exp(\hat{\beta}_{8\ 1} = -2.732) = 0.065$ CFU/mL a day later.

Total chlorine is significant at lag zero only (p-value = 0.0127). For a one unit increase in total chlorine on any given day total chlorine would decrease the HPC37 counts by factor of $\exp(\hat{\beta}_{9\ 0} = -2.503) = 0.082$ CFU/mL on the same day.

Rainfall is significant at lag zero (p-value = 0.0184). Therefore on any given day rainfall would increase the log HPC37 counts by a factor of $\exp(\hat{\beta}_{11\ 0} = 0.021) = 1.021$ CFU/mL on the same day.

The autumn season is highly significant at the 5% level of significance (p-value = 0.0001), relative to winter season in DH-WTP by a factor of $\exp(\hat{\beta}_{14} = 1.405) = 4.076$ CFU/mL.

Sampling point TDH008 is significant at 5% level (p-value = 0.0190), meaning TDH008 has higher HPC37 counts relative to TDH010 by a factor of $\exp(\hat{\beta}_{18} = 0.450) = 1.568$ CFU/mL.

3.6.1 Model Validation in Negative Binomial

Consider Figure 3.3, before we decide if the negative binomial is appropriate, we check if the variance of the residuals is proportional to the mean. Plotting the square of the residual to the fitted values, with a black line for Poisson, green line for quasi-Poisson, a blue line for smoothed mean of the square of the residual, and a red line for predicted variance from the negative

binomial fit, we find that the blue line is no close nor parallel to the green line indicating that the mean-variance relationship is not proportional. This means negative binomial model is also not an appropriate model. However, the negative binomial predicted variance (red curve) is close to the green line and therefore compared to the quasi-Poisson model, the negative binomial seems to be behaving better.

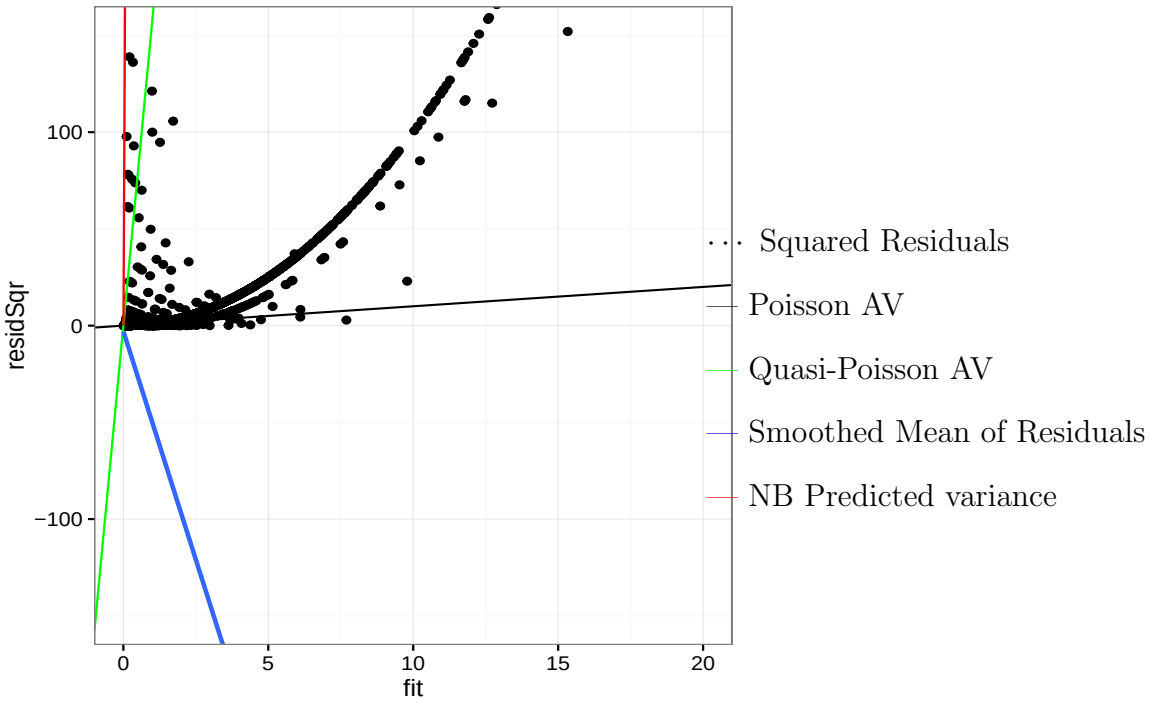

Figure 3.3: Mean-Variance Relationship.

In Figure 3.4, graphical validation plots are shown. The negative binomial residuals does not show patterns and therefore, again, this is an indication that the negative binomial model is better than the quasi-Poisson. The upper left graph (A) does not show any pattern but has extreme values.

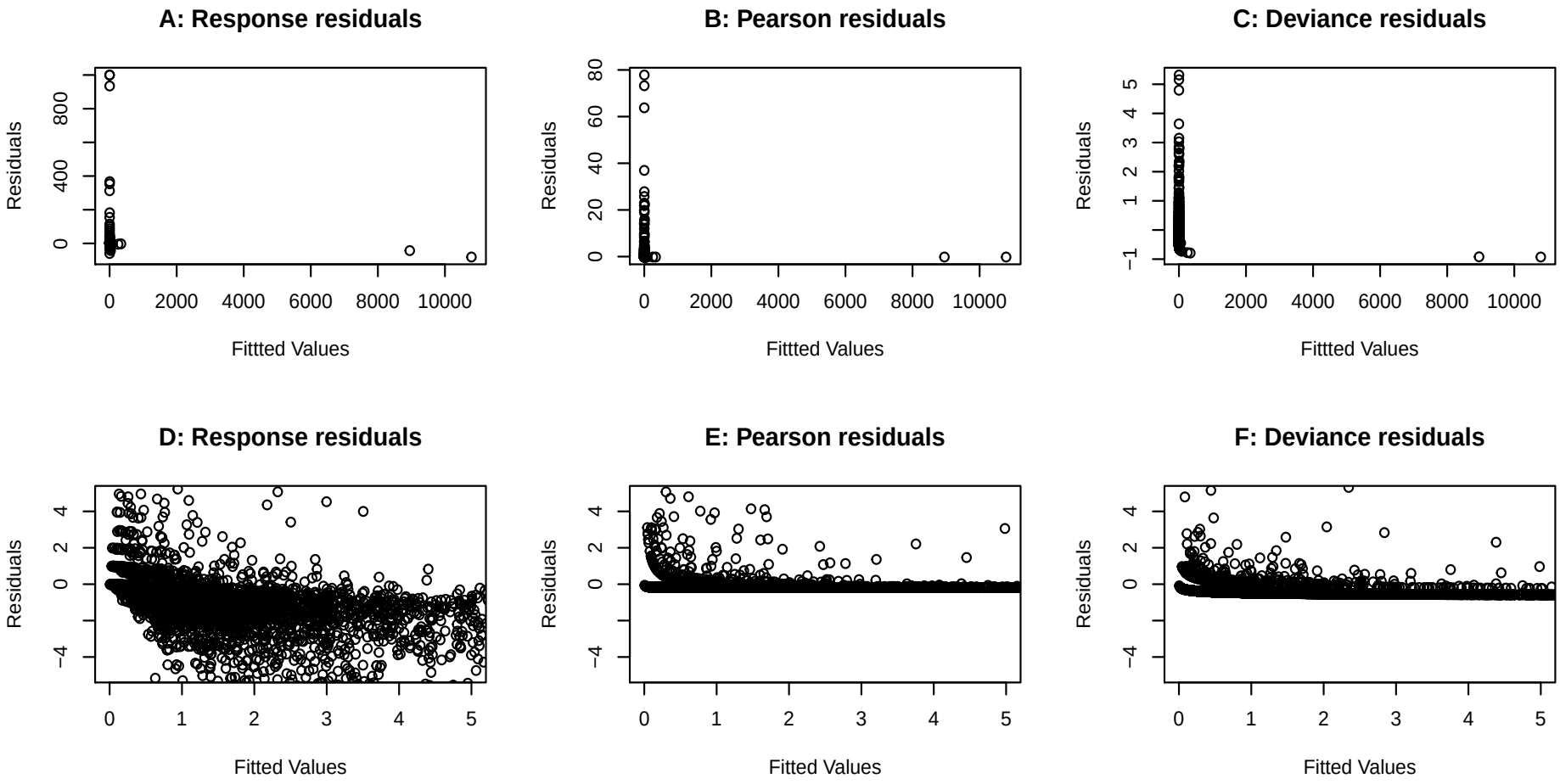

Figure 3.4: Graphical validation tools for the Negative Binomial.

Chapter 4

Generalized Linear Mixed Models

4.1 Linear Mixed Models

The linear mixed model is an extension of linear models where the vector of the random effects u is added and the response variable y which is normally distributed. The form of the linear mixed model is therefore as follows:

$$y = X\beta + Zu + \epsilon, \tag{4.1}$$

where $y_{n\times 1}$ is a vector of the response variable, $X_{n\times(p+1)}$ is the design matrix for fixed effects, $\beta_{(p+1)\times 1}$ is a vector of fixed effect parameters parameters, $Z_{n\times q}$ is a design matrix for the random effects, $u_{q\times 1}$ is a vector of unknown random effects parameters assumed to have a multivariate normal distribution with mean vector 0 and covariance matrix G, i.e $u \sim N(0, G)$ and $\epsilon_{n\times 1}$ is a vector of random errors assumed to have a multivariate normal distribution with mean vector 0 and covariance matrix R, i.e $\epsilon \sim N(0, R)$. However, for non-normal data which belong to the exponential family of distributions, generalized linear mixed models are deployed.

4.2 Generalized Linear Mixed Models

Generalized linear mixed models (GLMMs) provide an extension to generalized linear models (GLMs) by addition of random effects that are not captured in the GLM. This chapter looks at the extension of GLM and the estimation algorithm. The following theory on GLMMs is discussed in details by Jiang (2007); Littell et al. (2006); Wood (2006) among others.

A standard GLM assumes that the expectation of the response Y_{ij} can be written as a function of linear predictor,

$$\eta = x'_{ij}\beta,$$

where x_{ij} is a vector of covariates and β is a vector of fixed effects parameters respectively.

Now let u_i be a random effect parameter to account for individual variability or equivalently an individual effect parameter. Assuming observations are conditionally independent given the vectors x_{ij}, β and u_i the likelihood of the η_i observations, $Y_{i1}, Y_{i2}, \ldots, Y_{im_i}$, coming from the same cluster i, is given by

$$Pr(Y_{i1}, Y_{i2}, \ldots, Y_{im}|x, \beta, u_i) = \prod_{j=1}^{n_i} Pr(Y_{ij}|x_{ij}, \beta, u_i). \tag{4.2}$$

GLMMs are an extension of GLM to longitudinal data that accommodates correlated and over-dispersed data by adding random effects to the linear predictor η. The frequently encountered non-Gausian models for repeated or longitudinal measured outcomes have binomial or Poisson distributions. The generalized linear mixed model conditionally satisfies the exponential family of distribution structure. Therefore, given u and x_{ij} the probability density function for the observations y_{ij}'s is given by

$$f(y_{ij}|u, \beta) = \exp\left[\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right] \tag{4.3}$$

where $\mu_{ij} = \mathrm{E}(Y_{ij}|u_i)$ is modeled through a linear predictor containing a fixed regression

parameter vector β and a vector of subject specific parameters u_i as follows

$$g(\mu_{ij}) = x'_{ij}\beta + z'_{ij}u_i. \tag{4.4}$$

Assuming that the link function g and the vectors of covariate values x_{ij} and z_{ij} are known, if the natural or the canonical link function given as $g(\mu_{ij}) = \eta_{ij}$ holds, then the model becomes

$$\theta_{ij} = x'_{ij}\beta + z'_{ij}u_i.$$

The model assume that conditionally on the subject-specific effect u_i, the responses Y_{ij} are independent and u_i are normally distributed with mean zero and variance matrix G among the random effects. Notice that the function $c(y_{ij}, \phi)$ may or may not depend on Y_{ij}. By inventing the link function the conditional mean and variance are given by

$$\mu_{ij} = \mathrm{E}[Y_{ij}|u_i] = g^{-1}(x'_{ij}\beta + z'_{ij}u_i)$$

and

$$\mathrm{var}(Y_{ij}|u_i) = V(\mu_{ij})\phi$$

where g and V are the link and variance functions. Assuming that η_{ij} is the corresponding linear predictor of the form

$$\eta_{ij} = x'_{ij}\beta + z'_{ij}u_i, \tag{4.5}$$

the commonly used link function for the count Poisson model with log link is specified as

$$\log(\mu_{ij}) = \eta_{ij}$$

$$Y_{ij}|\mu_{ij} \sim \mathrm{Poisson}(\mu_{ij}).$$

When we introduce random effects in the above models we end up with the corresponding specific GLMM. It is worth noting that the dispersion parameter ϕ accounts for extra variability in the model where random effects are not included. When fitting the model, ϕ in is known prior or may require that it be estimated using methods such as moment estimation. For Poisson model, the model implied variance function may not be consistent with the actual distribution. The quasi-likelihood methods can be used to estimate the dispersion parameter, ϕ.

4.3 Maximum Likelihood Estimation

The likelihood function under GLMM is said not to have a closed-form expression when the data is non-normal (Jiang, 2007). This means that obtaining the marginal distribution is not easy if the conditional distribution of y, given u, is non-normal. Such likelihood may involve high-dimensional difficulties and hence approximation becomes one of the alternatives (Jiang, 2007). The difficulty in maximizing the likelihood is due to the presence of N integrals over the q dimensional random effects u_i. Therefore, numerical approximation methods that can be used include the approximation of the integrand, approximation of the data, and those that are based of the integral itself. For an extensive overview of these approximations, see Tuerlinckx et al. (2004); Molenberghs and Verbeke (2006) among others. The likelihood is given by

$$\begin{aligned} L(\beta, G, \phi) &= \prod_{i=1}^{N} f_i(y_i|\beta, G, \phi) \\ &= \prod_{i=1}^{N} \int \prod_{i=1}^{n_i} f_{ij}(y_{ij}|\beta, u_i, \phi) f(u_i|G) \mathrm{d}u_i. \end{aligned} \tag{4.6}$$

This expression in Equation 4.6 is the same as in the case of the marginal distribution of Y obtained by integrating the joint distribution of Y and u with respect to u. We will briefly

discuss the methods of approximations.

4.3.1 Laplace Approximation

The Laplace integrals are a well known methods of approximating the integrand and is one of the commonly used methods when the likelihood function is difficult to compute. The objective of the approximated integrands is to obtain the traceable integrals such that closed form expressions can be obtained through the feasibility of numerical maximization of the approximated likelihood. Tierney and Kadane (1986) used the Laplace method designed to approximate integrals of the form

$$I = \int \exp\{q(x)\}\mathrm{dx}, \tag{4.7}$$

where $q(x)$ is a known, unimodal and bounded function of a $q-$dimensional variable x. Let $\hat{x}$ be the value of x that maximizes q. The second-order Taylor expansion of the function $q(x)$ is of the form

$$q(x) \approx q(\hat{x}) + \frac{1}{2}q''(\hat{x})(x-\hat{x})^2 \tag{4.8}$$

where $q''(\hat{x})$ is equal to the Hessian of q, which is the matrix of the second-order derivative of q evaluated at $\hat{x}$. Notice that when we replace $q(x)$ in Equation 4.7 by its approximation in Equation 4.8 we get

$$I \approx (2\pi)^{r/2}|-q''(\hat{x})|^{-1/2}\exp\{q(x)\}.$$

The integral in Equation 4.6 is proportional to an integral of the form given by Equation 4.7 for functions of $q(x)$ given by

$$q(x) = \sum_{i=1}^{n_i} [y_{ij}(x'_{ij}\beta + z'_{ij}u) - \psi(x'_{ij}\beta + z'_{ij}u)] - \frac{1}{2}u'D^{-1}u$$

such that the Laplace method can be applied. Note that the function $q(\hat{x})$ depends on the uknown parameters β, ϕ and D such that in each iteration of the numerical maximization of the likelihood, $\hat{x}$ will be recalculated conditionally on the current values of parameter estimates.

4.3.2 Penalized Quasi-Likelihood

The theory of quasi-likelihood is a basis for the analysis of over-dispersed count data (Wedderburn, 1974). The method of penalized quasi-likelihood (PQL) may be illustrated under a general framework as an approximate quasi-likelihood estimate approach. The PQL method uses Taylor expansion around current estimates $\hat{\beta}$ of fixed effects and $\hat{u}_i$ of random effects assuming canonical or natural link. This gives

$$\begin{aligned}
Y_{ij} &= \mu_{ij} + \epsilon_{ij} = h(x'_{ij}\beta + z'_{ij}u) + \epsilon_{ij} \\
&\approx h(x'_{ij}\hat{\beta} + z'_{ij}\hat{u}) \\
&+ h(x'_{ij}\hat{\beta} + z'_{ij}\hat{u})x'_{ij}(\beta - \hat{\beta}) \\
&+ h(x'_{ij}\hat{\beta} + z'_{ij}\hat{u})z'_{ij}(u_i - \hat{u}_i) + \epsilon_{ij} \\
&= \hat{\mu}_{ij} + V(\hat{\mu_{ij}})x'_{ij}(\beta - \hat{\beta}) + V(\hat{\mu_{ij}})z_{ij}(u_i - \hat{u}_i) + \epsilon_{ij}, \qquad (4.9)
\end{aligned}$$

where $\hat{\mu_{ij}}$ equals its current predictor $h(x'_{ij}\beta + z'_{ij}u)$ of the conditional mean $\mathrm{E}[Y_{ij}|u_i]$. This expression can be written in vector form as follows

$$Y_i \approx \hat{\mu_{ij}} + \hat{V}_i X_i(\beta - \hat{\beta}) + \hat{V}_i Z_i(u_i - \hat{u}_i) + \epsilon_i, \qquad (4.10)$$

where X_i and Z_i are design matrices and V_i is a diagonal matrix with diagonal entries $V(\hat{\mu}_{ij})$. Re-ordering the above expression yields

$$Y_i^* \equiv \hat{V}_i^{-1}(Y_i - \hat{\mu}_i) + X_i\hat{\beta} + Z_i\hat{u}_i \approx X_i\beta + Z_iu_i + \epsilon_i^*, \tag{4.11}$$

where $\epsilon_i^* = V_i^{-1}\epsilon_i$ and has a mean of zero. The modified response Y_i^* allow for the approximation of the problem as a linear mixed model. The resulting estimates are called penalized quasi-likelihood estimates since they are obtained by optimizing a quasi-likelihood function which only involves first and second order conditional moments augmented with a penalty term on the random effects.

4.3.3 Marginal Quasi-Likelihood

The alternative approximation method is the marginal quasi-likelihood (MQL) which is very similar to the PQL method. The PQL approach is the most common estimation procedure for the GLMM. The difference between MQL and PQL is that the MQL does not incorporate the random effects in the linearization process, but they both have the same key idea and similar properties. The current predictor has the form $h(x'_{ij}\beta)$ which result in the expression that follow

$$Y_i^* \equiv \hat{V}_i^{-1}(Y_i - \hat{\mu}_i) + X_i\hat{\beta}, \tag{4.12}$$

which satisfy the approximate linear mixed model in Equation 4.11. More information can be found in Molenberghs and Verbeke (2006) among others.

4.4 Prediction of Random Effects

To predict the values of the of the random effects we make use of the conditional expectations of the random effects, given the observed response values y_i. The conditional expectation for u is

$$u = \mathrm{E}(u|y) = \hat{G}Z'\hat{V^{-1}}(y - X\hat{\beta}). \tag{4.13}$$

The predicted values in Equation 4.13 are the expected values of the random effects, $\mathbf{u}$, associated with the i^{th}-level of a random factor, given the observation y_i. These conditional expectations are referred to as empirical best linear unbiased predictors (EBLUPs) because they are based on the estimates $\hat{\beta}$ and $\hat{\theta}$ parameters.

The variance-covariance matrix is therefore

$$\mathrm{var}(\hat{u}) - \hat{G}Z\prime(\hat{V}^{-1} - \hat{V}^{-1}X(\sum X\hat{V}^{-1})X)^{-1}X\hat{V}^{-1})Z\hat{G}. \tag{4.14}$$

These predictors are "best" because they have a minimum variance among all linear estimators. They are "linear" in that they are linear functions of the observations, "unbiased" in that their expectation is equal to the expectation of the random effects for a single subject and that they are "predictors" based on the observed data (West et al. 2014; Jiang 2007).

4.5 Application of the GLMM Poisson Distributed Lag Model for Heterotrophic Plate Counts (HPC37) Predictions

In this section we fit a Poisson model to the Durban-Heights data but now with site as a random variable. It is worth noting that DH-WTP supply water to more than forty reservoirs, but for

this study the interest is only on the three sites selected; namely TDH007 (site 1), TDH008 (site 2) and TDH010 (site 3).

Table 4.1: Results for a GLMM Poisson Distributed Lag Model Fitted to HPC37 counts, Durban-Heights, 1991-1997.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	38.4337	1.0905	35.24	< 0.001
Temperature at lag 0	-0.0748	0.0103	-7.24	< 0.001
Temperature at lag 1	0.1285	0.0102	12.61	< 0.001
Turbidity at lag 0	0.4901	0.0303	16.16	< 0.001
Turbidity at lag 1	0.6041	0.0228	26.50	< 0.001
pH at lag 0	-2.3834	0.1176	-20.27	< 0.001
pH at lag 1	-2.1755	0.1150	-18.92	< 0.001
Free Chlorine at lag 0	-4.2700	0.1416	-30.17	< 0.001
Free Chlorine at lag 1	-2.6978	0.1792	-15.06	< 0.001
Total Chlorine at lag 0	-0.8697	0.1458	-5.97	< 0.001
Total Chlorine at lag 1	1.0333	0.1602	6.45	< 0.001
Rainfall at lag 0	0.0078	0.0011	7.29	< 0.001
Rainfall at lag 1	-0.0159	0.0019	-8.20	< 0.001
Summer	1.1296	0.0887	12.74	< 0.001
Autumn	2.1439	0.0808	26.55	< 0.001
Spring	1.5491	0.0737	21.02	< 0.001

less than 0.001 as shown in the table.

The main effects of temperature on HPC37 counts are negative at lag zero and positive at lag one. It is estimated that the mean HPC37 counts of a 1mL water sample in Durban-Heights on a given day would be decreased by $\exp(\hat{\beta}_{1\ 0} = -0.075) = 0.928$ CFU/mL and increased by $\exp(\hat{\beta}_{2\ 1} = 0.129) = 1.138$ CFU/mL a day later, adjusting for seasonal changes and all the covariates conditional on samples that share the given site.

Effects of turbidity are positive at both lags zero and one. This means on any given day, the mean HPC37 counts of a 1mL water sample in Durban-Heights would be increased by $\exp(\hat{\beta}_{3\ 0} = 0.490) = 1.632$ CFU/mL on the same day and by $\exp(\hat{\beta}_{4\ 1} = 0.604) = 1.829$ CFU/mL a day later , adjusting for seasonal changes and all the covariates conditional on samples that share the given site.

The main effects of pH are both negative at lags zero and one. It is estimated that the mean HPC37 counts of a 1mL water sample in Durban-Heights on a given day would be decreased

by $\exp(\hat{\beta}_{5\ 0} = -2.383) = 0.092$ CFU/mL on a particular day and $\exp(\hat{\beta}_{6\ 1} = -2.176) = 0.113$ CFU/mL one day later, adjusting for seasonal changes and all the covariates conditional on samples that share the given site.

The main effects of free chlorine are both negative at lag zero and lag one. It is estimated that the mean HPC37 counts of a 1mL water sample in Durban-Heights on a particular day would be decreased by factor of $\exp(\hat{\beta}_{7\ 0} = -4.270) = 0.014$ CFU/mL on the same day and $\exp(\hat{\beta}_{8\ 1} = -2.698) = 0.067$ CFU/mL a day later, adjusting for seasonal changes and all the covariates conditional on samples that share the given site.

The main effect of total chlorine is negative at lag zero and positive at lag one. It is estimated that the mean HPC37 counts of a 1mL water sample in Durban-Heights on a particular day would be decreased by $\exp(\hat{\beta}_{9\ 0} = -0.869) = 0.419$ CFU/mL and increased by $\exp(\hat{\beta}_{10\ 1} = 1.033) = 2.810$ CFU/mL on the next day, adjusting for seasonal changes and all the covariates conditional on samples that share the given site.

The effect of rainfall is positive at lag zero and negative at lag one. It is estimated that the mean HPC37 counts of a 1mL water sample in Durban-Heights on a particular day is increased by $\exp(\hat{\beta}_{11\ 0} = 0.008) = 1.008$ CFU/mL on the same day and decreased by $\exp(\hat{\beta}_{12\ 1} = -0.016) = 0.984$ CFU/mL a day later, adjusting for seasonal changes and all the covariates conditional on samples that share the given site.

The main effects of seasons (summer vs winter, autumn vs winter and spring vs winter) are all positive. It is estimated that the mean HPC37 counts in the summer, autumn and spring are $\exp(\hat{\beta}_{13} = 1.129) = 3.093$ CFU/mL, $\exp(\hat{\beta}_{14} = 2.144) = 8.534$ CFU/mL and $\exp(\hat{\beta}_{15} = 1.549) = 4.707$ CFU/mL higher that the mean HPC37 counts during winter, adjusting for all water quality variables conditional on samples that share the same site.

4.6 Application of the GLMM Poisson Distributed Lag Model for Total Coliforms Predictions

GLMM Poisson distributed lag model has shown that all water quality variables are the good predictors of Total Coliforms occurrence in the DH system, however, not at both lags in some predictors. It is worth noting that p-values in Table B.1 found in Appendix B are denoted in terms of asterisks, where *, ** and *** denote p-value that is less than 10%, 5% and 1% respectively.

All the covariates or predictor variables interpreted here are significant at 5% level of significance. The main effects of temperature on TC count are positive at lag zero. It is estimate that the mean TC counts of a 100mL water sample in DH would be increased by a factor of $\exp(\hat{\beta}_{1\ 0} = 0.727) = 2.069$ CFU/100mL on the same day, adjusting for seasonal changes and all the covariates conditional on samples that share the same site.

The main effects of turbidity on TC counts are positive at lag one. It is estimated that the mean TC counts of a 100mL water sample in DH would be increased by a factor of $\exp(\hat{\beta}_{4\ 1} = 0.896) = 2.450$ CFU/100mL one day later, adjusting for seasonal changes and all the covariates conditional on samples that share the same site.

The main effects of pH at lag one are significant at the 5% level of significant. It is estimated that the mean TC counts of a 100mL water sample in DH would be decreased by a factor of $\exp(\hat{\beta}_{6\ 1} = -3.105) = 0.045$ CFU/mL one day later, adjusting for seasonal changes and all the covariates conditional on samples that share the same site.

The main effects of free chlorine are significantly positive at lag one and negative at lag two, at the 5% level of significance. It is estimated that the mean TC counts of a 100mL water sample in DH would be increased by a factor of $\exp(\hat{\beta}_{7\ 0} = 3.744) = 42.267$ CFU/100mL on same day and decreased by $\exp(\hat{\beta}_{8\ 1} = -2.524) = 0.078$ CFU/100mL one day later, adjusting for seasonal changes and all the covariates conditional on samples that share the same site.

The main effects of total chlorine are significantly negative at lag zero. It is estimated that the mean TC counts of a 100mL water sample would be decreased by a factor of

$\exp(\hat{\beta}_{9\ 0} = -1.655) = 0.191$ CFU/100mL on the same day, adjusting for seasonal changes and all the covariates conditional on samples that share the same site.

The main effects of rainfall are significantly negative at lags one and two. It is estimated that the mean TC counts of a 100mL water sample in DH would be decreased by a factor of $\exp(\hat{\beta}_{11\ 0} = -0.069) = 0.933$ CFU/100mL water sample on the same day and $\exp(\hat{\beta}_{12\ 1} = -0.108) = 0.898$ CFU/100mL water sample a day later, adjusting for seasonal changes and all the covariates conditional on samples that share the same site.

The main effects of seasons (summer vs winter, autumn vs winter and spring vs winter) are all negative. It is estimated that the mean TC counts in the summer, autumn and spring are $\exp(\hat{\beta}_{13} = -3.688) = 0.025$ CFU/mL, $\exp(\hat{\beta}_{14} = -4.110) = 0.016$ CFU/mL and $\exp(\hat{\beta}_{15} = -3.157) = 0.043$ CFU/mL lower mean TC counts during winter, adjusting for all water quality variables conditional on samples that share the given site.

4.7 Intraclass Correlation Coefficient

The intraclass correlation coefficient (ICC) is a measure describing the homogeneity of observations (or responses) on the dependent variable within a cluster (West et al., 2014).

Examples of clustered data include:

- individuals sampled within sites. In this case the sites are the clusters.
- longitudinal data (repeated measures) where multiple observations are collected from the same individual over time. The ICC is the ratio of the between-cluster variance to the total variance. It explains the proportion of the total variance in the response that is accounted for by the clustering or the correlation among observations within the same cluster. It helps determine whether or not the use of a mixed model is necessary.

According to Francisque et al. (2009) this coefficient is equal to zero if HPC37 measures are

independent. On the other hand, it is equal to one if HPC37 measures are exactly the same and not equal to zero implies that the HPC37 measures are not independent. That is, the ICC not equal to zero means that HPC37 measures in the same sub-system (reservoir) are found in similar environment or affected by the unobserved sub-system factors; for example the location of pipe, the material of pipe, the age of the pipe, the amount of time spent by water in the pipe, the velocity and flow, etc.

Recall that under GLM, the distributed lag Poisson model for HPC37 counts y_t takes the form

$$y_t \sim \text{Poisson}(\mu_t),$$

$$\log \mu_t = \theta_t, \tag{4.15}$$

where $\theta_t = \beta_0 + \sum_{i=1}^{p} \sum_{\ell=0}^{1} \beta_{i\ell} \mathbf{x}_{t-\ell}$.

Also, recall that under GLMM, the distributed lag Poisson model can be represented as

$$\log \mu_{ts} = \theta_{ts} + \gamma_s, \tag{4.16}$$

where site $s = 1, 2$ and 3 for sampling sites, θ_{ts} and μ_{ts} are defined as in Equation 4.15, and in addition, $\boldsymbol{\gamma}_s$ is the sampling site (reservoir) random effect, that is $\gamma_s \sim N(0, \tau)$.

Now to measure the intraclass correlation, that is the correlation among HPC37 count observations within a sub-system, we use the ICC denoted by ρ and it ranges between zero and one. This correlation coefficient can be obtained as:

$$\rho = \text{corr}(y_{ts}, y_{t's}) = \frac{\tau}{\tau + \zeta}. \tag{4.17}$$

4.8 Model Validation

The distribution of residuals for the distributed lag Poisson GLMM is similar to those in Chapter 3, see Figure 4.1. These residual confirm that the data is overdispersed. It is not easy to tell if the model is adequate for this microbiological data using these plots.

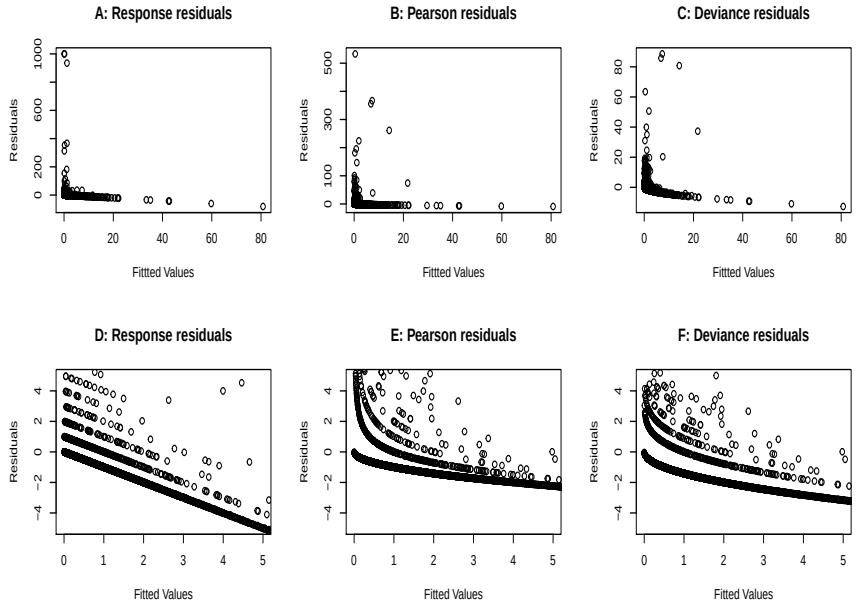

Figure 4.1: Graphical validation tools for the Poisson GLMM.

However, in comparing the fitted models in Chapter 3 with the distributed lag Poisson GLMM we make use of the information-theoretic approaches. Information-theoretic model selection allow comparison of multiple, non nested models (Bolker et al., 2009).

4.9 Random Effects Model Validation

To diagnose random effects effects, West et al. (2014) recommend the use of standard diagnostic plots like histogram, Q-Q plots, and scatterplots to investigate empirical Bayes predictors, which are also referred to as random-effects predictors or empirical best linear unbiased predictors (EBLUPs). This investigation is conducted for detection of potential outliers that may warrant further investigation. According to West et al. (2014) checking for normality is

of limited value because their distribution does not necessarily reflect the distribution of the random effect.

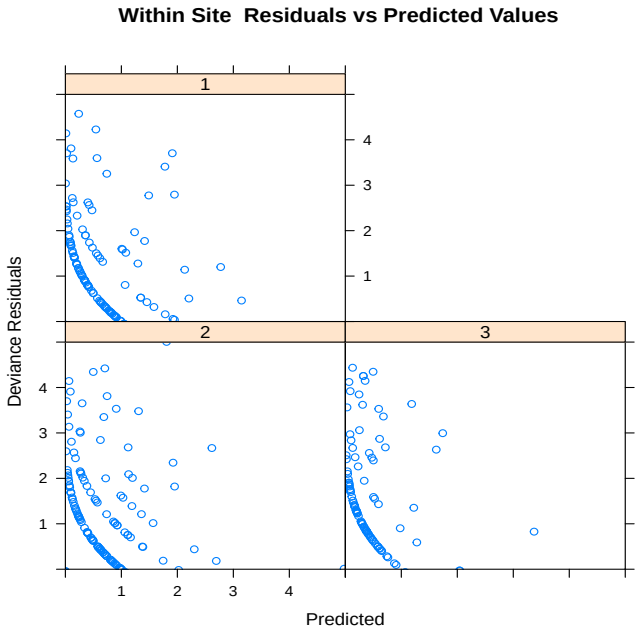

Figure 4.2: Within site graphical validation tools for the Poisson GLMM.

Figure 4.2 shows the within site deviance residuals against the predicted values of the distributed lag Poisson GLMM. The distribution of the residuals is similar across the three sites.

Chapter 5

Model Comparison and Discussion

Models for explanations and predictions of HPC37 and Total Coliforms in the water treatment plant (distribution system) were developed based on multiple regression analysis. Identification of the association between HPC37 and other water quality parameters is allowed through the regression approach. Three models were developed and fitted to the data under the assumption of independence or not about HPC37 and Total Coliform count observations. A distributed lag (DL) quasi-Poisson model and a DL negative binomial were applied in Chapter 3 and a DL Poisson GLMM was applied in Chapter 4.

After estimating parameter values, the next step is statistical inference: that is, hypothesis testing, drawing of conclusions, selection of best model and evaluation of goodness-of-fit among models.

Deviance and Pearson's chi-square statistics measure the discrepancy of fit between the maximum log-likelihood achievable and achieved log-likelihood by the fitted model Christensen 2006. However, obtaining a goodness of fit test for over-dispersed models is not as simple as fitting a Poisson model where the residual deviance or the Pearson's χ^2 can often be used (Hinde and Demétrio, 1998). It is important to know which among the fitted model best fit the data.

Deviance is two times the difference between the maximum likelihood of the saturated model and the maximum log-likelihood of the specified generalized linear model, that is,

$$D(y, \hat{\mu}) = 2\{\ell(y; y) - \ell(\hat{\mu}; y)\}, \tag{5.1}$$

where $\ell(y; y))$ is the log-likelihood under the maximum achievable model and $\ell(\hat{\mu}; y)$ is the log-likelihood under the current model. The main aim is to minimize $D(y, \hat{\mu})$, scaled by the dispersion parameter ϕ, by maximizing $\ell(\hat{\mu}; y)$ as follows

$$D^* = \frac{1}{\phi} D, \tag{5.2}$$

where D^* denote the scaled deviance. In cases where ϕ is not known

$$\hat{\phi} = \frac{D}{n - p}, \tag{5.3}$$

where n is the number of observations, and p is the number of parameters (Lindsey, 1997). The deviance $D \sim \chi^2_{n-p}$, where $n-p$ are the degrees of freedom (McCullagh and Nelder, 1989; Lindsey, 1997). This statistics is called the goodness-of-fit and test the hypothesis

H_O: Model is adequate

H_A: Model is not adequate

The null hypothesis H_O is rejected if $D > \chi^2_{n-p,\alpha}$, where α is the level of significance.

Therefore models fitted in Chapters 3 and 4 are compared using information theoretic approaches. Information theoretic approaches such as Akaike's inforamtion criteria (AIC; Akaike 1973) or Bayesian information criteria (BIC; Schwarz 1978) are often considered for selection between quasi-Poisson model and a negative binomial (Ver Hoef and Boveng, 2007). It is worth noting that these approaches depend on a distributional form and the likelihood; therefore since quasi models are characterized by their mean and variance and do not have the

distributional form Burnham and Anderson (2003) developed quasi-AIC for comparing within the class of quasi models.

The AIC can be calculated as follows:

$$\text{AIC} = -2\ell(\mu; y) + 2p,$$

where p is the number of parameters in the model.

In cases where count data are overdispersed, quasi-likelihood methods are appropriate and the theory leads to modified information criteria such as QAIC and QAIC_c (Burnham and Anderson, 2003; Kim et al., 2014). According to Burnham and Anderson (2003) the principles of quasi-likelihood suggest modification of AIC to

$$\text{QAIC} = -[2\log(\mathcal{L}(\hat{\mu}); y)\hat{\phi}] + 2p.$$

From Table 5.1, the AIC suggest that the distributed lag quasi-Poisson is a better model for the data since it has the smallest AIC value.

Table 5.1: A comparison of distributed lag count regression models for HPC37 occurrence.

	Dependent variable:	
	HPC37	
	GLM: Quasi-Poisson link = log	*Negative Binomial*
	(1)	(2)
Quasi-/Akaike Inf. Crit.	393.2104	6,193.071

The quasi-Poisson and the negative binomial models assume independence among HPC37 observations coming from the same reservoir (sampling site). The assumption holds, since on the basis of the generalized linear mixed model only about 2% of the variability was explained by the reservoir. Note that the quasi-AIC is compared to the AIC for the negative binomial model. For further details on the comparison between AIC and QAIC can be found

in Sileshi (2006) among others. According to Ver Hoef and Boveng (2007), to choose between a quasi-Poisson and the negative binomial needs a good understanding between them. Model selection using likelihood-based methods is still a problem. However, Ver Hoef and Boveng (2007) emphasize that an important way to choose an appropriate model is based on scientific reasoning; that is understanding the difference in weighting between quasi-Poisson and negative binomial. It is important to note that only models falling under the same class of models are compared. In this case we are comparing the quasi-Poisson with negative binomial which are GLMs. The Poisson under GLMM is not compared to any model because it is a model in its own class. Moreover, the sampling-point level ICC for HPC37 appears to be very weak,$\hat{\rho} = 0.002$. This result suggest that the HPC37 measures originating from the same sampling-point/tap are practically independent.

The sampling-point level ICC for total coliforms appears to be strong, $\hat{\rho} = 0.8$. This result suggests that the total coliforms measures coming from the same sampling-point/tap are greatly influenced by unobserved factors at the sampling point level. These include factors such as location, the age of the pipe, the material of pipe, the amount of time spent in the pipe, the velocity of water and its flow. Therefore, the quasi-Poisson GLMM approach is preferred than the quasi-Poisson GLM in predicting total coliforms positive counts. That is, the use of generalized linear mixed models for a clustering effect at the level of sampling points/subsystems is suggested for modeling total coliforms data, respectively. This is based on the fact that quasi-Poisson model violate the independence postulate of the observations.

Chapter 6

Conclusion and Further Research

6.1 Conclusion

This thesis was primarily concerned with modelling the role of the water quality and environmental factor variables on the occurrence of HPC37 and TC counts in the Durban Heights (DH) water treatment plant that supply water to Durban city in KwaZulu-Natal province in South Africa. The statistical models that were developed fall under the framework of GLMs and GLMMs which allow for predictions and investigation of count data such as HPC37 and TC.

The exploratory data analysis showed that temperature influences bacterial counts in the system. That is, as seasons change, so does the temperature and thus bacterial counts differ across the four seasons. High turbidity, rainfall and temperature levels seemed to be associated with high bacterial counts and high chlorine and pH levels seemed to be associated with low bacterial counts. These are similar to findings of a study by Francisque et al. (2009) among others.

Models showed that, on average, the effects of temperature on HPC37 can be explained or observed one day later, which means there is a time lapse of one day before the effects of

increased or decreased temperature on HPC37. Free chlorine showed consistency in bacterial reduction in all the models across the three DH sites on HPC37. On any particular day, increased levels of turbidity are related to bacterial counts in the DH system on the present day and also one day later. The pH of final water in DH has shown negative effect with bacterial counts in the system as well as rainfall. Total chlorine effects significantly reduce HPC37 positive counts on the present day but not one day later. Lastly, relative to winter, autumn is the season with the highest number of HPC37 positive counts followed by summer then spring comes last.

On the other hand, a GLMM showed that the effects of temperature on TC can be explained or observed on the same day, which means there are immediate effects of increased or decreased temperature on TC. There is a time lapse of one day before the effects of increased or decreased turbidity on TC. The effects of free chlorine can be explained on the same day and one day later on TC counts. There is a time lapse of one day before the effects of increased or decreased pH of water on TC counts. The effects of total chlorine can be explained or observed on the same day, which means there are immediate effects of increased or decreased total chlorine on TC. The effects of rainfall can be explained or observed both on the same day and one day later. According to the results of a GLMM, winter relative to summer, spring and autumn is expected to have more TC counts.

Models were compared using quasi-/AIC information criterion. Results showed that simple quasi-Poisson performed better than the negative binomial model. In fact, GLM and GLMMs are the appropriate framework of models to use when dealing with non-normal data. Thus, the Quasi-/Poisson and negative binomial model for counts data were used and relationships between microbiological-out-of-range data and measurable variables were obtained.

The ICC indicated a strong intra-site or system correlation of about 80%, meaning that, assuming water samples coming from a specific site are independent when analyzing the data is not justified. This also means that TC counts may not result from water quality or external environmental factors but may be a result of subsystem factors such a the age of the pipe,

material of the pipe, amount of time water spends in the system/reservoir, velocity and the flow of the water or other possible unobserved factors (Francisque et al., 2009). These factors need to be investigated.

The historical microbiological data is useful for understanding previous patterns and trends so that future events can be predicted. The time series approach in this project has shown how, on average, various predictors influence the TC and HPC37 bacteria over a two-day period. That is, it should be easy to control for bacteria over a certain period of time through the use of distributed lag models GLMs and GLMMs respectively.

This work is of use to microbiologists, statisticians and water supply agencies in general.

6.2 Further Research

The major challenges in statistical modeling of bacterial counts is concerned with understanding water quality dormancy and relapse and their interactions within the system. This thesis sought to find the relationship between microbiological out of range counts and related water quality including other measurable variables at Umgeni Water. Nonetheless, it is impractical to incorporate all the variables influencing the bacteria in final water. Statistical modeling of these bacterial counts therefore leave out certain aspects which when considered could give more insights concerning microbiological out of range, water quality and other measurable variables at Umgeni Water.

Accounting for missing data is important but has limitations if the amount of missing information is large, thus loss of information during data collection should be minimized. A missing rate of 5% or less is recommended, see (Schafer, 1999). Variability of microbiological occurrences located at different locations but supplied by one reservoir can be looked at. That is, water samples from different locations supplied by one reservoir can be used to test if there is variability due to different places. The model need to be enhanced in order to account for excess zeros in the data and inclusion of the capacity to handle latent effects by

using Bayesian estimation methods. Bayesian estimation methods have the potential of being used as a decision making tool by continually updating it with new information to adapt to changing operational needs. The goal should be to move towards a predictive dynamic model for monitoring and intervention purposes.

References

Allen, M. J., Edberg, S. C., and Reasoner, D. J. (2004). Heterotrophic plate count bacteria–what is their significance in drinking water? *International journal of food microbiology*, 92(3):265–274.

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica: Journal of the Econometric Society*, pages 178–196.

Baltagi, B. H. (2011). Distributed lags and dynamic models. *Econometrics*, pages 131–147.

Bartram, J. and Ballance, R. (1996). *Water quality monitoring: a practical guide to the design and implementation of freshwater quality studies and monitoring programmes*. London: CRC Press.

Beaudeau, P., Payment, P., Bourderont, D., Mansotte, F., Boudhabay, O., Laubies, B., and Verdiere, J. (1999). A time series study of anti-diarrheal drug sales and tap-water quality. *International Journal of Environmental Health Research*, 9(4):293–311.

Bezuidenhout, C., Mthembu, N., Puckree, T., and Lin, J. (2002). Microbiological evaluation of the mhlathuze river, kwazulu-natal (rsa). *Water SA*, 28(3):281–286.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3):127–135.

Breslow, N. E. (1984). Extra-poisson variation in log-linear models. *Applied Statistics*, 33(1):38–44.

Burnham, K. P. and Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Heidelberg: Springer Science & Business Media.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge: Cambridge university press.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96(455):1022–1030.

Carter, J., Rice, E., Buchberger, S., and Lee, Y. (2000). Relationships between levels of heterotrophic bacteria and water quality parameters in a drinking water distribution system. *Water Research*, 34(5):1495–1502.

Chandra, S., Saxena, T., Nehra, S., and Mohan, M. K. (2016). Quality assessment of supplied drinking water in jaipur city, india, using pcr-based approach. *Environmental Earth Sciences*, 75(2):1–14.

Christensen, R. (2006). *Log-linear models and logistic regression*. New-York, NY: Springer Science & Business Media.

Dobson, A. J. and Barnett, A. (2008). *An introduction to generalized linear models*. Boca Rota: CRC press.

Dunteman, G. H. and Ho, M.-H. R. (2006). *An introduction to generalized linear models*. Thousand Oaks: Sage.

Francisque, A., Rodriguez, M. J., Miranda-Moreno, L. F., Sadiq, R., and Proulx, F. (2009). Modeling of heterotrophic bacteria counts in a water distribution system. *Water research*, 43(4):1075–1087.

Gautheir, T. D. (2001). Detecting trends using spearman's rank correlation coefficient. *Environmental forensics*, 2(4):359–362.

Germs, W., Coetzee, M., Van Rensburg, L., and Maboeta, M. (2004). A preliminary assessment of the chemical and microbial water quality of the chunies river-limpopo: short communication. *Water SA*, 30(2):267–272.

Gill, J. (2000). *Generalized linear models: a unified approach*, volume 134. Thousand Oaks, California: Sage Publications.

Hauke, J. and Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93.

Hendricks, C. W. (1978). *Evaluation of the microbiology standards for drinking water*. Virginia: Springfield, National Technical Information Service.

Hinde, J. and Demétrio, C. G. (1998). Overdispersion: models and estimation. *Computational Statistics & Data Analysis*, 27(2):151–170.

Hsieh, J. L., Nguyen, T. Q., Matte, T., and Ito, K. (2015). Drinking water turbidity and emergency department visits for gastrointestinal illness in new york city, 2002-2009. *Public Library of Science*, 10(4):1–16.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. New York, NY: Springer Science & Business Media.

Kim, H.-J., Cavanaugh, J. E., Dallas, T. A., and Foré, S. A. (2014). Model selection criteria for overdispersed data and their application to the characterization of a host-parasite relationship. *Environmental and ecological statistics*, 21(2):329–350.

LeChevallier, M. W., Evans, T., and Seidler, R. J. (1981). Effect of turbidity on chlorination efficiency and bacterial persistence in drinking water. *Applied and environmental microbiology*, 42(1):159–167.

Levine, M. M. (1987). Escherichia coli that cause diarrhea: enterotoxigenic, enteropathogenic, enteroinvasive, enterohemorrhagic, and enteroadherent. *Journal of infectious Diseases*, 155(3):377–389.

Lin, J., Biyela, P., Puckree, T., and Bezuidenhout, C. (2004). A study of the water quality of the mhlathuze river, kwazulu-natal (rsa): Microbial and physico-chemical factors. *Water SA*, 30(1):17–22.

Lindsey, J. K. (1997). *Applying generalized linear models*. New-York, NY: Springer Science & Business Media.

Littell, R. C., Stroup, W. W., Milliken, G. A., Wolfinger, R. D., and Schabenberger, O. (2006). *SAS for mixed models*. Cary, North Carolina: SAS institute.

Luyt, C. D., Tandlich, R., Muller, W. J., and Wilhelmi, B. S. (2012). Microbial monitoring of surface water in south africa: an overview. *International journal of environmental research and public health*, 9(8):2669–2693.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. Boca Raton: CRC press.

Molenberghs, G. and Verbeke, G. (2006). Models for discrete longitudinal data.

Momba, M., Tyafa, Z., Makala, N., Brouckaert, B., and Obi, C. (2006). Safe drinking water still a dream in rural areas of south africa. case study: The eastern cape province. *Water SA*, 32(5):715–720.

Myers, R. H., Montgomery, D. C., Vining, G. G., and Robinson, T. J. (2012). *Generalized linear models: with applications in engineering and the sciences*, volume 791. Hoboken: John Wiley & Sons.

Olsson, U. (2002). *Generalized linear models: An applied approach*. LaVergne, Tennessee: Lightning Source Inc.

Payment, P. and Robertson, W. (2004). The microbiology of piped distribution systems and public health. *Safe, Piped Water: Managing Microbial Quality in Piped Distribution Systems*.

Peng, R. D. and Dominici, F. (2008). *Statistical methods for environmental epidemiology with R*. New-York, NY: Springer.

Power, K. N. and Nagy, L. A. (1999). Relationship between bacterial regrowth and some physical and chemical parameters within sydney's drinking water distribution system. *Water Research*, 33(3):741–750.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reilly, J. K. and Kippin, J. S. (1983). Relationship of bacterial counts with turbidity and free chlorine in two distribution systems. *Journal (American Water Works Association)*, 75(6):309–312.

Sack, R. B., Rahman, M., Yunus, M., and Khan, E. H. (1997). Antimicrobial resistance in organisms causing diarrheal disease. *Clinical infectious diseases*, 24(Supplement 1):S102–S105.

SANS241-1 (2015). South african national standard 241-1: Microbiological, physical, aesthetic and chemical determinands. *South African Bereau of Standards*.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, 8(1):3–15.

Shaban, A. B. and Malkawi, H. I. (2007). Rapid detection of human enteric pathogens (viruses and bacteria) in water resources from jordan using polymerase chain reaction (pcr). *Journal of Applied Sciences Research*, 3(10):1084–1093.

Sileshi, G. (2006). Selecting the right statistical model for analysis of insect count data by using information theoretic measures. *Bulletin of entomological research*, 96(05):479–488.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86.

Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D., Van den Noortgate,

W., Meulders, M., and De Boeck, P. (2004). Estimation and software. In *Explanatory item response models*, pages 343–373. New-Yok: Springer.

Ver Hoef, J. M. and Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, 88(11):2766–2772.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gaussnewton method. *Biometrika*, 61(3):439–447.

West, B. T., Welch, K. B., and Galecki, A. T. (2014). *Linear mixed models: a practical guide using statistical software*. Boca Raton: CRC Press.

WHO (2004). *Guidelines for drinking-water quality: recommendations*, volume 1. World Health Organization.

Wilkes, G., Edge, T., Gannon, V., Jokinen, C., Lyautey, E., Medeiros, D., Neumann, N., Ruecker, N., Topp, E., and Lapen, D. R. (2009). Seasonal relationships among indicator bacteria, pathogenic bacteria, cryptosporidium oocysts, giardia cysts, and hydrological indices for surface waters within an agricultural landscape. *Water Research*, 43(8):2209–2223.

Wood, S. (2006). *Generalized additive models: an introduction with R*. CRC press.

Zhang, W. and DiGiano, F. A. (2002). Comparison of bacterial regrowth in distribution systems using free chlorine and chloramine: a statistical study of causative factors. *Water Research*, 36(6):1469–1482.

Zuur, A., Ieno, E., Walker, N., Saveliev, A., and Smith, G. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Spring Science and Business Media.

Appendix A

R Code

This appendix show the R statistical software commands used.

A.1 Data Preparation in R

Given data sets TDH007, TDH008 and TDH010 with equal number of columns variables that are in the same order, the R code to column merge data is:

```
data12 <- rbind(TDH007,TDH008) # Merge site 1 with site 2
data123 <- rbind(data12,TDH010) # Merge sites 1 and 2 with site 3


# Declare a date variable:
data123$date <- as.Date(data123$date, format="%d-%m-%y")


# Convert dates in data123 to months:
data123$time <- month(as.POSIXlt(data123$date, format="%d/%m/%Y"))


# Declare a site variable and time as factors:
data123$site <- as.factor(data123$site)
```

```
data123$time <- as.factor(data123$time)


# Group months from time-variable into seasons and create a new categorical
# variable and call it Time":
data123$t <- unclass(data123$time)
data123$Time <- rec(data123$t, "5,4,3=1;1,9,8=2;6,2,7=3; 11,10,12=4")
data123$Time <- factor(data123$Time)


# Categorize observations within a variable using function "Recode":
data123$HPC37.cat <-  Recode(data123$HPC37, "0 ='0';
                             0.1:10='1-10';
                             10.1:100='11-100';
                             100.1:1000='101-1000';
                             1000.1:10000000='> 1000'",
                             as.factor.result=TRUE)
```

A.2 R Code Used For Exploratory Data Analysis

Note that R can generate Latex commands. To produce summary statistics tables for Latex in R:

```
stargazer(TDH007); stargazer(TDH008); stargazer(TDH010)
```

The plots showing the distribution of the data in R:

```
# Put multiple plots as a single figure:
# Example: for plotting in 2 rows and 4 columns use:
par(mar=c(2,4,2,2), mfrow=c(2,4))
```

```
with{data123,
    plot(date, HPC37, xlab="Time (Years)", ylab="HPC37 (Counts)",
    main="Heading of the graph")
}
```

Plotting stacked bar plots in R as follows:

```
cat1<- data123$rain.cat
HPC <- data123$HPC37.cat

df1 <-data.frame(cat1,HPC)

df3 <- df1 %>%
  group_by(cat1, HPC) %>%
  summarise(n=n()) %>%
  mutate(percent = (n / sum(n)), cumsum = cumsum(percent),
         label=ifelse(HPC=="11-100", paste0("n=", sum(n)),""))

ggplot(df3[order(df3$HPC),],aes(x=cat1, y=percent, fill=HPC,na.rm=TRUE)) +
scale_y_continuous(labels = scales::percent) +
scale_fill_grey(drop=FALSE) +
labs( y = "Distribution of HPC classes (%)") +
labs( x= "Time (Seasons)")+
geom_bar(position = 'fill',color = "black", stat="identity") +
geom_text(aes(y=cumsum + 0.02, label=label,na.rm=T), vjust=0,
           position = "identity") +
  ggtitle("TITLE")
```

A.3 R Code Used for Model Fitting

Start by setting the reference category using the commands:

```
data123 <- within(data123, site <- relevel(site,ref=3))
data123 <- within(data123, Time <- relevel(Time,ref=3))
```

For fitting distributed lag quasi-/Poisson GLM:

```
model 1 <- glm(HPC37 ~ Lag(Temperature, 0:1, Time)+Lag(Turbidity, 0:1, Time)+ ...
              +Lag(Rainfall, 0:1, Time),family="quasipoisson" data="data123")
```

For fitting distributed lag negative binomial GLM:

```
model 2 <- glm.nb(HPC37 ~ Lag(Temperature, 0:1, Time)+Lag(Turbidity, 0:1, Time)+ ...
              +Lag(Rainfall, 0:1, Time), data="data123")
```

For fitting distributed lag Poisson GLMM:

```
model 3 <- glmer(HPC37 ~ Lag(Temperature, 0:1, Time)+Lag(Turbidity, 0:1, Time)+ ...
              +Lag(Rainfall, 0:1, Time),family="poisson", data="data123")
```

A.4 Output ready for Latex in R

Regression model output codes were generated in R using:

```
xtable(summary(model 1)) # Or model 2
```

For extracting GLMM fitted model:

```
Reg_glmm <- coef(summary(model 3))
xtable(Reg_glmm)
```

Appendix B

Results for Total Coliforms occurrence

Table B.1: Regression models 1 and 3 for Total Coliforms occurrence, DHF 1, 2 and 3: 1991-1997.

	Dependent variable:	
	Total Coliforms	
	Model 1: Quasi-Poisson link = log	*Model 3: Generalized Linear Mixed-Effects*
	Estimate (Std. Error)	Estimate (Std. Error)
Temperature at lag 0	0.727	0.727***
	(0.459)	(0.048)
Temperature at lag 1	0.063	0.063
	(0.435)	(0.046)
Turbidity at lag 0	−0.243	−0.235
	(5.845)	(0.613)
Turbidity at lag 1	0.896	0.896***
	(1.827)	(0.192)
pH at lag 0	0.470	0.469
	(4.138)	(0.435)
pH at lag 1	−3.100	−3.105***
	(4.327)	(0.455)
Free Chlorine at lag 0	3.741	3.744***
	(7.825)	(0.824)
Free Chlorine at lag 1	−2.542	−2.545***
	(7.498)	(0.789)
Total Chlorine at lag 0	−1.649	−1.655**
	(6.766)	(0.712)
Total Chlorine at lag 1	0.276	0.271
	(6.555)	(0.689)
Rainfall at lag 0	−0.069	−0.069***
	(0.178)	(0.019)
Rainfall at lag 1	−0.108	−0.108***
	(0.302)	(0.032)
Summer	−3.684	−3.688***
	(3.368)	(0.354)
Autumn	−4.106	−4.110***
	(3.404)	(0.358)
Spring	−3.155	−3.157***
	(3.197)	(0.336)
DHF 1	2.254	RF
	(1.739)	
DHF 2	0.139	RF
	(2.336)	
Constant	0.349	1.189
	(42.435)	(4.500)
sampling-point ρ_s	n.a.	0.825
Quasi-/AIC	83.343	4,502.682

** significant at 5%; *** significant at 1%,
n.a.: not applicable; ρ_s: correlation within the reservoir;
RF: random factor.