USE OF LOGISTIC REGRESSION MODEL TO DETERMINE THE KEY CONTAMINANTS OF WATER IN YATTA DISTRICT

By

NZIOKI NICHOLAS MUSIU

School of Mathematics

College of Biological and Physical Sciences

University of Nairobi

A project submitted in partial fulfilment of the requirement for the Degree of Master of Science in Social Statistics

Declaration

Candidate:

I, the undersigned, hereby declare that this project report is my own original work and not a duplication of similarly published work of any school for academic purpose and has not been submitted to any other institution of higher learning for the award of certificate, diploma or degree.

I further declare that, all materials cited in this paper which are not my own have been duly acknowledged.

Signature____________________ Date 26th July 2012

Nzioki, Nicholas Musiu

Registration number: I56/76585/2009

Supervisor:

This project has been submitted for examination with my approval as supervisor

Signature____________________ Date 30-7-2012

Dr. Kipchirchir Isaac Chumba

School of Mathematics

University of Nairobi

Acknowledgement

Special thanks go to the various individuals who supported me during the time of my study. First I recognize the advice given to me by my Supervisor, Dr. Kipchirchir and my Lecturers, Prof. Manene, Dr. Achia, Prof J. A. M. Otieno, the late Dr. Nguti, the late Prof. Owino, Mr. Nderitu, Mr. Obudho, Dr. Okech and Dr. Muhua among others in the Department of Statistics, School of Mathematics, University of Nairobi.

I also appreciate the support given to me by my classmates, Happiness, Wamiti, Matiri, Simba, Collins, Kimote and many others.

The input by several other people made my work at the University to run smoothly including the computer lab technical person, Jane Chelule.

The many friends who assisted me in my work should feel appreciated for their effort of lifting my spirit when it was down.

Dedication

This project report is dedicated to my lovely wife Susan Wayua for accepting to remain behind and take care of our children as I undertook my further study. I also dedicate it to my children Carol, Cathy, John, Devis and Mary who sometimes did without some essential comfort and love when I was away.

The support they gave me was of paramount importance. May God provide unto them in abundance.

Abstract

Over one sixth of the world's population lacks safe drinking-water sources. The water crisis in Kenya is due not only to the wave of droughts but also to poor management of the water sources, pollution of water supplies by untreated sewage and a huge population explosion.

In the study, the main contaminants that could make drinking-water unsafe in Yatta district are determined. The data used to analyse the contamination levels of water were collected during a survey conducted by Kenyatta National Hospital (KNH), Department of Public Health (DPH). The survey covered water sources in Kithimani area of Yatta district between January and December 2009.

The survey was based on physio-chemical parameters on the degradation in the level of the quality of drinking-water. The parameters tested were pH, color, sulphates, lead, nitrate and turbidity. Water samples from different sources within the area were collected including both surface and ground water. Logistic regression model was used to determine the key contaminants of the water.

The study shows that there was significant level of contamination of the drinking-water on most of the sources. pH was noted to be the predictor which was the most significant on the non-safety of the water.

Abbreviations and Acronyms

DPH - Department of Public Health

EPA – Environmental Protection Authority

IWRM –Integrated Water Resources Management

KEBS – Kenya Bureau of Standards

KNH – Kenyatta National Hospital

LRA – Linear Regression Analysis

ML – Maximum Likelihood

MRA – Multiple Regression Analysis

NEMA – National Environmental Management Authority

N.T.U – Nephelometric Turbidity Units

ppm – parts per million

WHO – World Health Organization

WRMA – Water Resources Management Authority

Table of Contents

List of Tables

List of Figures

CHAPTER 1: INTRODUCTION

1.1 Background

1.1.1 Definition of safe drinking-water

Drinking-water or potable water is water of sufficiently high quality that can be consumed or used with low risk of immediate or long term harm. In most developed countries, the water supplied to households, commerce and industry is all of drinking-water standard, even though only a very small proportion is actually consumed or used in food preparation. Typical uses include washing and landscape irrigation, Brisbane City Council Information (2005).

1.1.2 Current position for drinking-water

Over large parts of the world, humans have inadequate access to potable water and use sources contaminated with disease vectors, pathogens or unacceptable levels of toxins or suspended solids. Such water is not wholesome and drinking or using such water in food preparation leads to widespread acute and chronic illnesses and is a major cause of death and misery in many countries.

Actually, over one sixth of the world's population lacks safe drinking-water sources. Unsafe water supplies, along with deficient sanitary infrastructure and inadequate personal hygiene, contribute substantially to the burden of 2.2 million annual deaths from diarrhoeal diseases. Although then definitive solution to the problem of access to safe drinking-water is the universal provision of piped and treated water, this option remains elusive because of the enormous expenditure of money and time that is required (WHO, UNICEF, 2005).

The water crisis in Kenya is due not only to the wave of droughts but also to poor management of the water sources, pollution of water supplies by untreated sewage and a huge population explosion.

Kenya is limited by an annual renewable fresh water supply of only 647 m^3 per capita and is classified as a water scarce country. The situation in Yatta is typical of what Kenya is. Yatta is in a semi-arid region and the little water which is available is highly precious. The inhabitants of this place rely mostly on water sources upstream. If there is any instability upstream, then they are obviously affected.

Water consumption is increasing twice as fast as the number of people consuming it. A key issue is that the true value of water is not recognised. Most people take it for granted, and it does not add up to much on their consumption bill. Germany has one of the highest average cost of water, but still no one ever complains about water prices; it is electricity or gas prices that they are concerned with. Water tends to be undervalued around the world. Therefore, large returns will not be seen coming from the pure natural source.

1.2 Statement of the problem

Many Kenyans receive high quality drinking-water every day from public water systems (which may be publicly or privately owned). Nonetheless, drinking-water safety cannot be taken for granted. Lack of access to safe drinking-water and sanitation continues to be a major problem in both rural and urban Kenya. There are a number of threats to drinking-water, that is, improperly disposed of chemicals, animal wastes, pesticides, human threats, wastes injected underground and naturally occurring substances. Therefore, there was need to determine whether the chemical content of the water exceeds the KEBS standards laid down for safe drinking-water. However, even when water is safe for drinking at the source, it is commonly re-contaminated during collection, storage and use at home.

1.3 Objectives

The main objective is to assess water quality in Yatta district.

The specific objectives are;

1. To use the KEBS drinking-water bench marks to establish whether the contaminants exceed the safe levels.
2. To use logistic regression model to determine;
    a) the key contaminants of drinking-water.
    b) the level of safety of drinking-water in particular sources.

1.4 Significance of the study

The study is meant to determine whether the drinking-water in Kithimani area of Yatta district is safe or unsafe. It is also meant to enlighten people on the sources of pollution of the water and how to protect these sources to remain safe. This knowledge would make the water users to be aware of the dangers of contaminated water, that is, health effects on the people. It would reduce the health problems and expenditure on medication. The residents of Yatta district would concentrate on development activities instead of health deficits.

CHAPTER 2: LITERATURE REVIEW

2.1 Poor water pollution control

In a study conducted in Nakuru district health facilities, in October 2001 by the department of public health, faculty of medicine, University of Nairobi, about 50 out of every 1000 patients were found to have water-borne diseases (typhoid, amoebiasis and dysentery). Outbreaks were observed to be seasonal and most of these occurred between March and May each year. The burden of these diseases was high and water borne diseases could be prevented at community level instead of managing them at a tertiary hospital.

In another study conducted by the centre for diseases control and prevention in Kenya (2004), fluoride levels in the river water samples showed small zonal variations. The highest level was 0.85ppm in Laikipia district and the lowest was 0.08ppm in Murangá district.

Metal and nutrient content in Genesee were typical for a non-polluted environment except for a moderate enrichment of phosphorus and a slight enrichment of lead. The phosphorus enrichment in the sediment arises both from agricultural activities and municipal wastes. Lead enrichment in the predominantly non-urban setting may be due to a diffuse atmosphere input (anonymous).

Storm water systems in urban areas are sometimes combined with sanitary sewer systems enroute to sewage treatment plants (WHO, 2003). Excessive storm water can cause this joint system to overflow. In this event, excess flow will be directed into water ways untreated, resulting in sewage contamination. According to Environmental Protection Agency (EPA), approximately 20% of the population is served by combined systems and 46% is served by separate systems. Effluent that leaks from sewer lines is generally untreated raw sewage. It may contain industrial waste chemicals

In 2004, the National Environment Management Authority (NEMA) established a law to govern the management of waste and effluent discharges to river bodies. However, the effluent discharge standards were not specified. Additionally, there is a non-point source pollution from land-husbandry, sanitation and liquid and solid waste-disposal practices (Agresti And Barbara; 1997). For example, studies carried out on pollution in

Nairobi river by NEMA found that, sewage, nutrients, toxic metals, human waste, solid waste dumping, industrial and agricultural chemicals were main pollutants. Then, due to poor management enforcement, the river may not be a safe source of drinking-water.

2.2 The current situation

Overall water coverage is declining in terms of quality, quantity and reliability due to the aging of existing infrastructure. Most of Kenya's population (in 2001) lived in rural areas (24 million), while 9 million lived in urban areas out of which more than half lived in informal settlements. Access to safe drinking-water was estimated at 68% in urban areas and only 49% in rural areas (WHO; 2010). According to a study carried out by the Institute of Economic Affairs (2009), urban per capita water consumption almost doubled that of rural areas.

Domestic sewage and industrial effluent were the main causes of pollution to the water sources. Raw or partially treated sewage, when discharged into water bodies or the environment may pollute water sources and harm public health (Dziubbane; 2006). In 2001, raw sewage discharge from a sewage treatment work from Embu town into a nearby river killed several people who used the water downstream for domestic purpose (source, daily newspaper).

2.3 Traces of drugs in drinking-water

A vast array of pharmaceuticals – including antibiotics, anti-convulsants, mood stabilizers and sex hormones – have been found in the drinking-water supplies of at least 41 million Americans, an Associated Press (AP) investigation shows. To be sure, the concentrations of these pharmaceuticals are tiny, measured in quantities of parts per billion or trillion, far below levels of a medical dose. Also, utilities insist their water is safe (Hardalo and Edberg; 1997).

But the presence of so many prescription drugs – and over-the-counter medicines like acetaminophen and ibuprofen – in so much of our drinking-water is heightening worries among scientists of long-term consequences to human health. Water providers rarely disclose results of pharmaceutical screening, unless pressed, the AP found.

In general, there is poor water and sewage service delivery in Kenya and where facilities exist, their standards of operation and maintenance are the responsibilities of newly formed water and sanitation companies (though a number are unaware of this).

The coverage is however very low compared with water supply coverage. Sewerage services cover only 14% of urban centres.

2.4 Logistic Regression Analysis

Logistic Regression Analysis was used by Mitchell (1992) who used data based on 50 premature infants, some of whom did not survive infancy. In the study, the conclusion was that birth weight is a useful predictor of infant survival. In the case, there was more interest in deciding whether or not the three predictors, in addition to birth weight, improved the predictive efficiency of the model. Given that the other predictors remained in the model, removing weight as a predictor resulted in significantly poorer predictive efficiency, although removing any of the other predictors did not have a significant impact.

Chao-Ying, et al. (2000), constructed a hypothetical data set to which logistic regression was applied, and interpreted the results. The hypothetical data consisted of reading scores and genders of 189 inner city school children. Of these, 59 (31.22%) were recommended for remedial reading classes and 130 (68.78%) were not. A legitimate research hypothesis posed to the data was that, "likelihood that an inner city school child was recommended for remedial reading instruction was related to both their reading score and gender".

A two-predictor logistic model was fitted to the data to test the research hypothesis regarding the relationship between the likelihood that an inner city child was recommended for remedial reading instruction and their reading score and gender. The result showed that, it may be inferred that a given score (e.g. 60 points), the probability of a boy being recommended for remedial reading programs is higher than that of a girl.

CHAPTER 3: METHODOLOGY

3.1 Data

The data used to analyse the contamination levels of water was collected during a survey conducted by Kenyatta National Hospital (KNH), Department of Public Health (DPH). The survey covered water sources in Kithimani area of Yatta district between January and December 2009.

The survey did not include the different activities undertaken by the residents of Kithimani like irrigation farming, fishing, cattle keeping and general uses of water. As a result it was not possible to apply the effects after use approach to measure the safety of the water.

Samples of water were collected in sterilized bottles from the different sources and taken directly to the government laboratory. All the predictors, that is, colour, nitrate, lead, sulphates, turbidity and pH, were analysed on Dionex DX-120 ion chromatograph.

3.2 Logistic Regression Model

Historically, one of the first uses of regression-like models for binomial data was bioassay results (Finlay, 1997). Responses were the proportions or percentages of success; for example, the proportion of experimental responses to given doses of drugs to treat a certain condition. Such data are sometimes called quantal responses. The aim was to describe the probability of responding positively, p, of a patient as a function of the drug administered, X, for example.

To test the hypothesis about the determinants of unsafe drinking-water, a logistic regression model is used. The predictor variables considered are colour, lead, nitrate, pH, sulphates and turbidity. In this model, the response variable is binary, taking only two values 1 or 0.

The probability of a source being safe or unsafe depends on a set of predictor variables x_j, j= 1, 2, ..., p. The source is viewed as a random variable Y that can take the values 1 or 0 with $P(Y=1/\underline{x}) = \pi(\underline{x})$ if safe and $P(Y=0/\underline{x}) = 1-\pi(\underline{x})$ if unsafe, $\underline{x} = (x_1, x_2, ..., x_p)'$. Y is a Bernoulli random variable with parameter $E(Y) = \pi$.

In the regression context, it is assumed that the set of predictor variables, $x_1, x_2, ..., x_p$, are related to Y and, therefore, provide additional information for predicting Y. For theoretical (or Mathematical) reasons, LRA is based on a linear model for the **natural logarithm of odds** (that is, log-odds) in favour of Y = 1.

For the binary response variable Y and p explanatory variables (predictors), $x_1, x_2, \ldots, x_p$,

Let $\pi(\underline{x}) = P(Y = 1/\underline{x}) = 1 - P(Y = 0/\underline{x}), \quad \underline{x} = (x_1, x_2, \ldots, x_p)'$ (3.1)

The Multiple Logistic Regression (MLR) model is

$$\pi(\underline{x}) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)} \quad (3.2)$$

Equivalently, the log-odds, called the logit, has the linear relationship

$$\text{logit}[\pi(\underline{x})] = \log\left[\frac{\pi(\underline{x})}{1 - \pi(\underline{x})}\right] = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \quad (3.3)$$

This equates the logit link function to the linear predictor. The parameter β_j refers to the effect of x_j on the log-odds, that is, Y = 1, controlling the other, x_j. For instance, $\exp(\beta_j)$ is the multiplicative effect on the odds of a unit increase in x_j. At fixed levels it can be qualitative using dummy variables for categories.

Similarly, if all predictors are set equal to 0, the predicted odds are $\exp(\beta_0)$.

Finally, the results can be expressed in terms of probabilities by use of the logistic function.

3.3 FITTING THE LINEAR REGRESSION MODEL (BINARY RESPONSES)

With N subjects, we treat the N binary responses as independent.

Let $\underline{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ denote setting i of values of p explanatory variables, that is, each response has a different covariate pattern and using (3.2), we have for setting i,

$$\pi(\underline{x}_i) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}, \qquad i = 1, 2, \ldots, N \quad (3.4)$$

Now $P(Y=y) = [\pi(\underline{x})]^y [1 - \pi(\underline{x})]^{1-y}, \quad y = 0, 1; \quad \underline{x} = (x_1, x_2, \ldots, x_p)'$ (3.5)

Such that $E(Y) = \pi(\underline{x})$

and the Likelihood function is

$L(\underline{\beta}) = \prod_{i=1}^{N} P(Y_i = y_i)$, $\underline{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_p)'$

$$= \prod_{i=1}^{N}[\pi(\underline{x}_i)]^{y_i}[1-\pi(\underline{x}_i)]^{1-y_i}$$

$$= \prod_{i=1}^{N} exp\left[log\left(\frac{\pi(\underline{x}_i)}{1-\pi(\underline{x}_i)}\right)^{y_i}\right]\prod_{i=1}^{N}[1-\pi(\underline{x}_i)]$$

$$= exp\left[\sum_{i=1}^{N} y_i log\left(\frac{\pi(\underline{x}_i)}{1-\pi(\underline{x}_i)}\right)\right]\prod_{i=1}^{N}[1-\pi(\underline{x}_i)]$$

$$= exp\left[\sum_{i=1}^{N} y_i \left(\beta_0 + \sum_{i=1}^{p} \beta_j x_{ij}\right)\right]\prod_{i=1}^{N}\left[\frac{1}{1+exp\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)}\right]$$

$$= exp\left[\beta_0 \sum_{i=1}^{N} y_i + \sum_{j=1}^{p} y_i x_{ij}\beta_j\right]\prod_{i=1}^{N}\left[\frac{1}{1+exp\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)}\right] \qquad (3.6)$$

and the log-likelihood function is

$$\log L(\underline{\beta}) = \beta_0 \sum_{i=1}^{N} y_i + \sum_{j=1}^{p} y_i x_{ij}\beta_j - \sum_{i=1}^{N}\left[1 + exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right] \qquad (3.7)$$

The Maximum Likelihood Equations (MLEs) result from setting

$$\frac{\partial logL(\underline{\beta})}{\partial \beta_0} = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N}\left[\frac{exp\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)}{1+exp\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)}\right] = 0 \qquad (3.8)$$

$$\frac{\partial logL(\beta)}{\partial \beta_j} = \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} x_{ij} \left[\frac{exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}{1 + exp\left(\beta_0 + \left(\sum_{j=1}^{p} \beta_j x_{ij}\right)\right)} \right] = 0 \quad (3.9)$$

Thus the MLEs are

$$\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \pi(\underline{x}_i) = 0 \quad (3.10)$$

$$\sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} \pi(\underline{x}_i) x_{ij} = 0, \quad j=1, 2, ..., p \quad (3.11)$$

These equations are nonlinear and we require iterative solution.

As in Multiple Regression Analysis (MRA), there are two important stages in the analysis of data;

- Estimates for the parameters in the model must be obtained and,
- Some determination must be made of how well the model actually fits the observed data.

3.4 Goodness of fit

Hosmer-Lemeshow Chi-square test - This test divides the data into several groups based on *p*-values, then computes a chi-square from observed and expected frequencies of subjects falling in the two categories of the binary response variable within these groups. Large chi-square values (and correspondingly small p-values) indicate a lack of fit for the model.

3.5 Analysis software

Excel is used for the exploration of the data. Then SPSS and R are used to analyse the data to come up with the results.

CHAPTER 4: DATA ANALYSIS AND RESULTS

Detailed data on the water contents is presented in appendices I to VI.

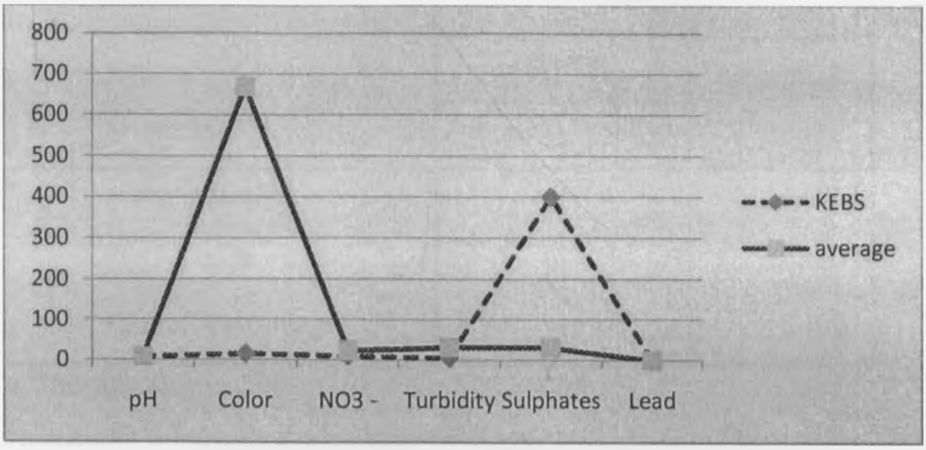

Figure 4.1: Athi R water data compared to KEBS bench marks

Table 4.1: Variables in the Equation for Athi R water source

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	pH	-28.066	7712.617	.000	1	.997	.000
	color	.027	3.315	.000	1	.993	1.028
	nitrate	-3.434	787.815	.000	1	.997	.032
	turbidity	.258	174.182	.000	1	.999	1.294
	sulphates	.175	437.463	.000	1	1.000	1.192
	lead	51.462	219960.851	.000	1	1.000	2.237E22
	Constant	233.445	61122.601	.000	1	.997	2.420E101

a. Variable(s) entered on step 1: pH, color, nitrate, turbidity, sulphates, lead.

The significance of each predictor is measured using Wald statistic. The test shows that, at p = 0.1, all the predictors are significant.

Table 4.2: Classification Table[a] for Athi R water

Observed			Predicted		
			safety		Percentage Correct
			unsafe	safe	
Step 1	safety	unsafe	13	0	100.0
		safe	0	1	100.0
	Overall Percentage				100.0

a. The cut value is .500

The overall accuracy of this model to predict a source being safe (with a predicted probability of 0.5 or greater) is 100% (table 4.2). The sensitivity is given by 1/1 = 100% and the specifity is 13/13 = 100%. The positive predictive value = 1/1 = 100% and the negative predictive value = 13/13 = 100%.

When we have a new source, we can use the logistic regression model to predict its probability of being safe.

Using the set of six predictors, pH, color, nitrate, turbidity, sulphates and lead, the logistic regression equation for the log-odds in favour of safety is estimated to be (from table 4.1);

$$log_e\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)= 233.4-28.1*pH+0.027*color-3.4*nitrate+0.3*turbidity+0.2*Sulphates+51.5*lead$$

The odds are;

$$\frac{\hat{\pi}}{1+\hat{\pi}} = e^{233.4-28.1*pH+0.027*color-3.4*nitrate+0.3*turbidity+0.2*Sulphates+51.5*lead}$$

Finally, the probability of safety is obtained by applying the logistic transformation

$$\hat{\pi} = \frac{1}{1+e^{-(233.4-28.1*pH+0.027*color-3.4*nitrate+0.3*turbidity+0.2*sulphates+51.5*lead)}}$$

Using the average measurement values for pH = 7.3, color = 668.5, nitrate = 24.3, turbidity = 32.179, sulphates = 30.8 and lead = 0.1

$$\hat{\pi} = \frac{1}{1+e^{-(233.4-28.1*7.3+0.02*668.6-3.4*24.3+0.3*32.2+0.2*30.8+51.5*0.1)}} = 0.000;$$

Goodness of fit test

This tells us how closely related the observed values match with the predicted values.

Hypothesis: Ho: the model fits

H1: the model does not fit

Significance level: $\alpha = 0.05$

Rejection region: Reject the null hypothesis if the p-value $\leq 0.05 = \alpha$

Table 4.3: Hosmer and Lemeshow Test for Deviance (Athi R.)

Step	Chi-square	df	Sig.
1	1.619	8	.991

Large Chi-square values (and correspondingly small p-values) indicate a lack of fit for the model. In table 4.3, we see that the Hosmer-Lemeshow Chi-square test for the final warranty model yields a p-value of 0.991 thus suggesting a model with good predictive value.

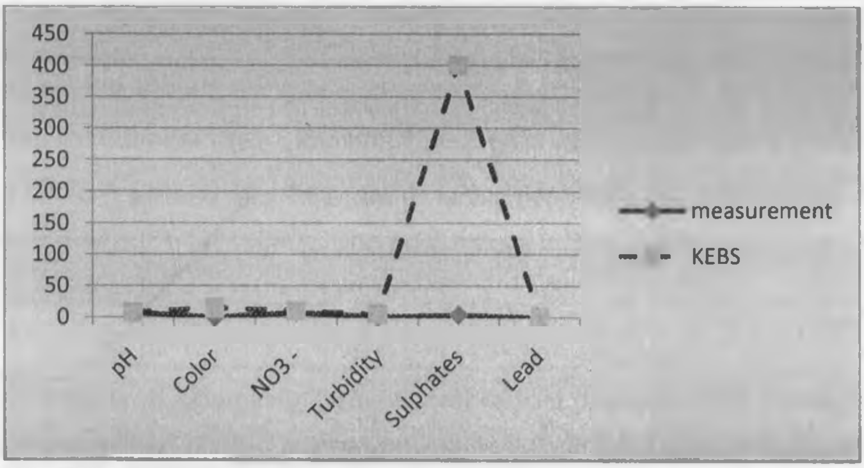

Figure 4.2 Borehole water data compared to KEBS bench marks

Table 4.4: Variables in the Equation for Borehole water source

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	pH	-23.292	2766.092	.000	1	.993	.000
	nitrate	-10.338	224.028	.002	1	.963	.000
	turbidity	39.213	675.144	.003	1	.954	1.072E17
	sulphates	33.217	1057.973	.001	1	.975	2.667E14
	Constant	135.461	24418.734	.000	1	.996	6.763E58

a. Variable(s) entered on step 1: pH, nitrate, turbidity, sulphates.

b. The variable colour is constant for the selected cases. Since a constant term was specified, the variable will be removed from the analysis.
c. The variable lead is constant for the selected cases. Since a constant term was specified, the variable will be removed from the analysis.

Using Wald statistic, the test shows that, at p = 0.1, all the four predictors are significant.

Table 4.5: Classification Table[a] for Borehole

Observed			Predicted		
			safety		Percentage Correct
			unsafe	safe	
Step 1	safety	unsafe	1	0	100.0
		safe	0	13	100.0
	Overall Percentage				100.0

a. The cut value is .500

The overall accuracy of this model to predict a source being safe (with a predicted probability of 0.5 or greater) is 100% (table 4.5). The sensitivity is given by 13/13 = 100% and the specifity is 1/1 = 100%. The positive predictive value = 13/13 = 100% and the negative predictive value = 1/1 = 100%.

When we have a new source, we can use the logistic regression model to predict its probability of being safe. Using the set of four predictors, pH, nitrate, turbidity and sulphates, the logistic regression equation for the log-odds in favour of safety is estimated to be (from table 4.4);

$$log_e\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 135.5\text{-}23.3*pH\text{-}10.3*nitrate+39.2*turbidity+33.2*Sulphates$$

The odds are;

$$\frac{\hat{\pi}}{1+\hat{\pi}} = e^{135.5-23.3*pH-10.3*nitrate+39.2*turbidity+33.2*Sulphates}$$

Finally, the probability of safety is obtained by applying the logistic transformation

$$\hat{\pi} = \frac{1}{1+e^{135.5-23.3*pH-10.3*nitrate+39.2*turbidity+33.2*Sulphates}}$$

Using the average measurement values for pH = 7.3, color = 668.5, nitrate = 24.3, turbidity = 32.179, sulphates = 30.8 and lead = 0.1

$$\hat{\pi} = \frac{1}{1+e^{-(135.5-23.3*7.4-10.3*7.6+39.2*0.8+33.2*3.9)}} = 1.000;$$

<u>Goodness of fit test</u>

This tells us how closely related the observe values match with the predicted values.

Hypothesis: Ho: the model fits

H1: the model does not fit

Significance level: $\alpha = 0.05$

Rejection region: Reject the null hypothesis if the p-value $\leq 0.05 = \alpha$

Table 4.6: Hosmer and Lemeshow Test (Borehole)

Step	Chi-square	df	Sig.
1	.000	8	1.000

Large Chi-square values (and correspondingly small p-values) indicate a lack of fit for the model. In table 4.6, we see that the Hosmer-Lemeshow Chi-square test for the final warranty model yields a p-value of 1.000 thus suggesting a model with good predictive value.

Figure 4.3 Ianguni river water data compared to KEBS bench marks

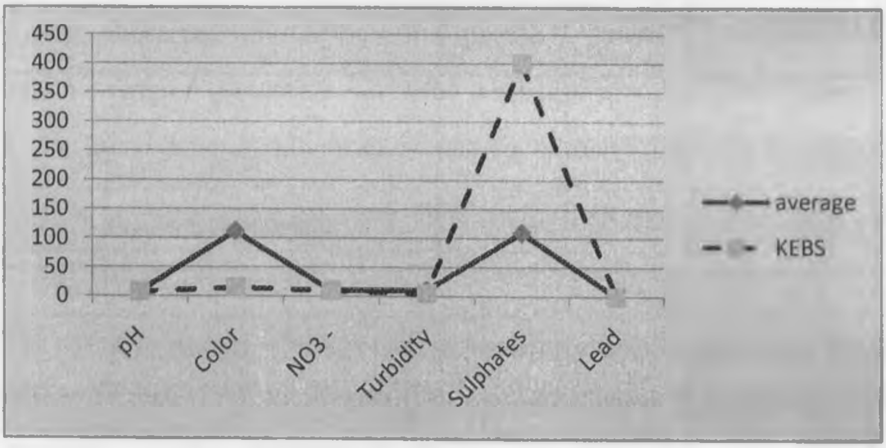

Table 4.7: Variables in the Equation for Ianguni river water source

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	pH	-98.641	21430.971	.000	1	.996	.000
	Color	-.030	28.767	.000	1	.999	.970
	Nitrate	-4.371	820.129	.000	1	.996	.013
	Turbidity	-3.987	656.315	.000	1	.995	.019
	Sulphates	.233	123.078	.000	1	.998	1.262
	Constant	812.834	187247.154	.000	1	.997	.

a. Variable(s) entered on step 1: pH, Color, Nitrate, Turbidity, Sulphates.

The variable lead is constant for the selected cases. Since a constant term was specified, the variable will be removed from the analysis.

Using Wald statistic, the test shows that, at p = 0.1, the five predictors are significant.

Table 4.8: Classification Table[a] for Ianguni

Observed			Predicted		
			Safety		Percentage Correct
			unsafe	safe	
Step 1	Safety	unsafe	13	0	100.0
		safe	0	1	100.0
	Overall Percentage				100.0

a. The cut value is .500

The overall accuracy of this model to predict a source being safe (with a predicted probability of 0.5 or greater) is 100% (table 4.8). The sensitivity is given by 1/1 = 100% and the specifity is 13/13 = 100%. The positive predictive value = 1/1 = 100% and the negative predictive value = 13/13 = 100%.

Using the set of five predictors, pH, color, nitrate, turbidity and sulphates, the logistic regression equation for the log-odds in favour of safety is estimated to be (from table 4.7);

$$log_e\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 812.8-98.6*pH-0.03*color-4.4*nitrate-4.0*turbidity+0.2*Sulphates$$

The odds are;

$$\frac{\hat{\pi}}{1+\hat{\pi}} = e^{812.8-98.6*pH-0.03*color-4.4*nitrate-4.0*turbidity+0.2*Sulphates}$$

Finally, the probability of safety is obtained by applying the logistic transformation

$$\hat{\pi} = \frac{1}{1+e^{812.8-98.6*pH-0.03*color-4.4*nitrate-4.0*turbidity+0.2*Sulphates}}$$

Using the average measurement values for pH = 7.3, color = 668.5, nitrate = 24.3, turbidity = 32.179, sulphates = 30.8 and lead = 0.1

$$\hat{\pi} = \frac{1}{1+e^{-(812.8-98.6*7.7-0.03*111.8-4.4*10.4-4.0*10.9+0.2*109.5)}} = 0.000$$

Goodness of fit test

This tells us how closely related the observe values match with the predicted values.

Hypothesis: Ho: the model fits

H1: the model does not fit

Significance level: $\alpha = 0.05$

Rejection region: Reject the null hypothesis if the p-value $\leq 0.05 = \alpha$

Table 4.9: Hosmer and Lemeshow Test (languni)

Step	Chi-square	df	Sig.
1	.000	8	1.000

Large Chi-square values (and correspondingly small p-values) indicate a lack of fit for the model. In table 4.9, we see that the Hosmer-Lemeshow Chi-square test for the final warranty model yields a p-value of 1.000 thus suggesting a model with good predictive value.

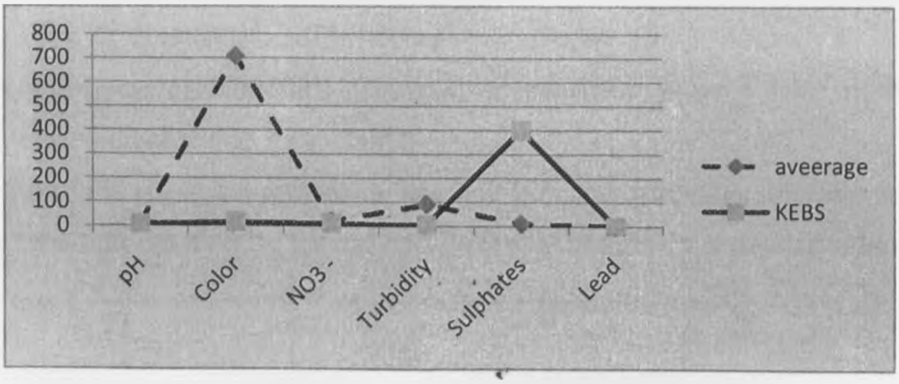

Figure 4.4 Kambi ya ndeke river water data compared to KEBS bench mark

Table 4.10: Variables in the equation for Kambi ya ndeke river source

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	pH	112.522	2947.224	.001	1	.970	7.370E48
	Color	-.058	1.448	.002	1	.968	.944
	Nitrate	1.218	89.592	.000	1	.989	3.380
	Turbidity	.326	9.012	.001	1	.971	1.386
	Sulphates	.423	503.767	.000	1	.999	1.526
	Lead	4885.535	187100.741	.001	1	.979	.
	Constant	-910.345	24829.030	.001	1	.971	

a. Variable(s) entered on step 1: pH, Color, Nitrate, Turbidity, Sulphates, Lead.

Using Wald statistic, at $p = 0.1$, all the predictors are significant.

Table 4.11: Classification Table[a] for Kambi ya ndeke

Observed			Predicted		
			safety		Percentage Correct
			unsafe	safe	
Step 1	safety	unsafe	13	0	100.0
		safe	0	1	100.0
	Overall Percentage				100.0

a. The cut value is .500

The overall accuracy of this model to predict a source being safe (with a predicted probability of 0.5 or greater) is 100% (table 4.11). The sensitivity is given by 1/1 = 100% and the specifity is 13/13 = 100%. The positive predictive value = 1/1 = 100% and the negative predictive value = 13/13 = 100%.

Using the set of six predictors, pH, color, nitrate, turbidity, sulphates and lead, the logistic regression equation for the log-odds in favour of safety is estimated to be (from table 4.10);

$log_e\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$ = -910.3+112.5*pH-0.06*color1.2*nitrate+0.3*turbidity+0.4*Sulphates+4885.5*lead

The odds are;

$$\frac{\hat{\pi}}{1+\hat{\pi}} = e^{-910.3+112.5*pH-0.06*color1.2*nitrate+0.3*turbidity+0.4*Sulphates+4885.5*lead}$$

Finally, the probability of safety is obtained by applying the logistic transformation

$$\hat{\pi} = \frac{1}{1+e^{-910.3+112.5*pH-0.06*color+1.2*nitrate+0.3*turbidity+0.4*Sulphates+4885.5*lead}}$$

Using the average measurement values for pH = 7.6, color = 708.1, nitrate = 19.8, turbidity = 91.2, sulphates = 11.3 and lead = 0

$$\hat{\pi} = \frac{1}{1+e^{-(-910.3+112.5*7.6-0.06*708.1+1.2*19.8+0.3*91.2+0.4*11.3+4885.5*0)}} = 0.000$$

Goodness of fit test

This tells us how closely related the observe values match with the predicted values.

Hypothesis: Ho: the model fits

H1: the model does not fit

Significance level: α = 0.05

Rejection region: Reject the null hypothesis if the p-value ≤ 0.05 = α

Table 4.12: Hosmer and Lemeshow Test (Kambi ya ndeke)

Step	Chi-square	df	Sig.
1	.000	8	1.000

Large Chi-square values (and correspondingly small p-values) indicate a lack of fit for the model. In table 4.12, we see that the Hosmer-Lemeshow Chi-square test for the final warranty model yields a p-value of 1.000 thus suggesting a model with good predictive value.

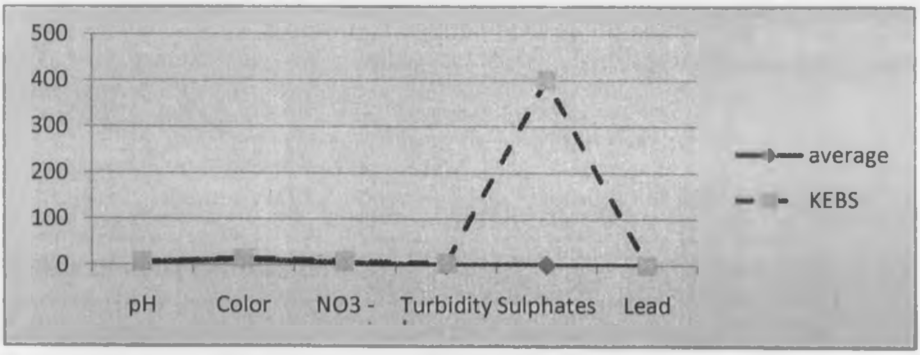

Figure 4.5 Rain water data compared to KEBS bench marks

Table 4.13: Variables in the Equation for Rain water source

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	pH	11.761	12084.223	.000	1	.999	128165.971
	Color	-1.373	57.137	.001	1	.981	.253
	Nitrate	-2.329	437.394	.000	1	.996	.097
	Turbidity	2.689	1841.543	.000	1	.999	14.713
	Sulphates	30.353	7471.279	.000	1	.997	1.521E13
	Lead	463.553	1207375.374	.000	1	1.000	2.082E201
	Constant	-75.537	86074.431	.000	1	.999	.000

a. Variable(s) entered on step 1: pH, Color, Nitrate, Turbidity, Sulphates, Lead.

Using Wald statistic, the test shows that, at p = 0.1, all the predictors are significant.

Table 4.14: Classification Table[a] for Rain water

Observed			Predicted		
			safety		Percentage Correct
			unsafe	safe	
Step 1	safety	unsafe	8	0	100.0
		safe	0	6	100.0
	Overall Percentage				100.0

a. The cut value is .500

The overall accuracy of this model to predict a source being safe (with a predicted probability of 0.5 or greater) is 100% (table 4.14). The sensitivity is given by 6/6 = 100% and the specifity is 8/8 = 100%. The positive predictive value = 6/6 = 100% and the negative predictive value = 8/8 = 100%.

Using the predictors, the logistic regression equation for the log-odds in favour of safety is estimated to be (from table 4.13);

$$log_e\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 75.5+11.8*pH - 1.4*color - 2.3*nitrate + 2.7*turbidity + 30.4*sulphates + 463.6*lead$$

The odds are;

$$\frac{\hat{\pi}}{1+\hat{\pi}} = e^{75.5+11.8*pH -1.4*color-2.3*nitrate+2.7*turbidity+30.4*sulphates+463.6*lead}$$

Finally, the probability of safety is obtained by applying the logistic transformation

$$\hat{\pi} = \frac{1}{1+e^{75.5+11.8*pH -1.4*color-2.3*nitrate+2.7*turbidity+30.4*sulphates+463.6*lead}}$$

Using the average measurement values for pH = 7.0, color = 13.1, nitrate = 6.8, turbidity = 0.6, sulphates = 0.8 and lead = 0

$$\hat{\pi} = \frac{1}{1+e^{-(-75.5+11.8*7.0 -1.4*13.1-2.3*6.8+2.7*0.6+30.4*0.8+463.6*0)}} = 0.281$$

<u>Goodness of fit test</u>

Hypothesis:

Ho: The model fits

H1: The model does not fit

Significance level: $\alpha = 0.05$

Rejection region: Reject the null hypothesis if the p-value $\leq 0.05 = \alpha$

Table 4.15: Hosmer and Lemeshow Test(Rain)

Step	Chi-square	df	Sig.
1	.000	8	1.000

Large Chi-square values (and correspondingly small p-values) indicate a lack of fit for the model.

In table 4.15, we see that the Hosmer-Lemeshow Chi-square test for the final warranty model yields a p-value of 1.000 thus suggesting a model with good predictive value.

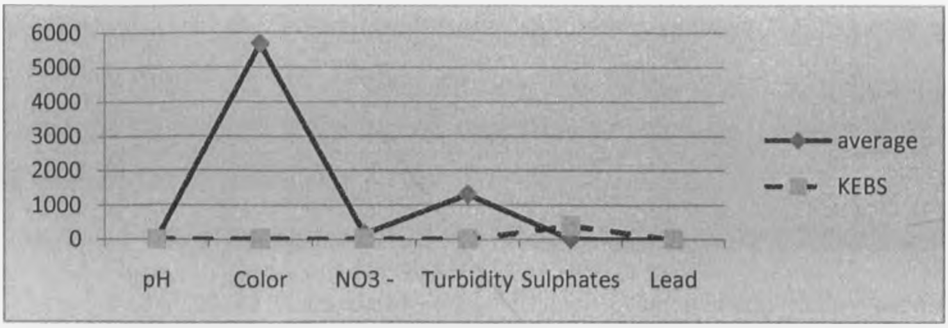

Figure 4.6 Utithini river water data compared to KEBS bench marks

Table 4.16: Variables in the Equation for Utithini river source

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	pH	.085	3822.422	.000	1	1.000	1.089
	Color	.007	81.434	.000	1	1.000	1.007
	Nitrate	-.173	47.562	.000	1	.997	.841
	Turbidity	-.002	17.513	.000	1	1.000	.998
	Sulphates	-.204	1742.016	.000	1	1.000	.816
	Lead	10.854	158958.288	.000	1	1.000	51742.942
	Constant	20.814	49685.881	.000	1	1.000	1.095E9

a. Variable(s) entered on step 1: pH, Color, Nitrate, Turbidity, Sulphates, Lead.

Using Wald statistic, the test at p = 0.1, all the predictors are significant.

Table 4.17: Classification Table[a] for Utithini river water

Observed			Predicted		
			safety		Percentage Correct
			unsafe	safe	
Step 1	safety	unsafe	11	0	100.0
		safe	0	3	100.0
	Overall Percentage				100.0

a. The cut value is .500

The overall accuracy of this model to predict a source being safe (with a predicted probability of 0.5 or greater) is 100% (table 4.17). The sensitivity is given by 3/3 = 100% and the specifity is 11/11 = 100%. The positive predictive value = 3/3 = 100% and the negative predictive value = 11/11 = 100%.

Using the six predictors, the logistic regression equation for the log-odds in favour of safety is estimated to be (from table 4.16);

$$log_e\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 20.8+0.09*pH+0.01*color-0.17*nitrate-0.00*turbidity-0.20*sulphates+10.8*lead$$

$$\frac{\hat{\pi}}{1+\hat{\pi}} = e^{20.8+0.09*pH+0.01*color-0.17*nitrate-0.00*turbidity-0.20*sulphates+10.8*lead}$$

Finally, the probability of safety is obtained by applying the logistic transformation

$$\hat{\pi} = \frac{1}{1+e^{-(20.8+0.09*pH+0.01*color-0.17*nitrate-0.00*turbidity-0.20*sulphates+10.8*lead)}}$$

Using the average measurement values for pH = 7.7, color = 5673.1, nitrate = 164.8, turbidity = 1311.8, sulphates = 13.4 and lead = 0.00

$$\hat{\pi} = \frac{1}{1+e^{-(20.8+0.09*7.7+0.01*5673.1-0.17*164.8-0.00*1311.8-0.20*13.4+10.8*0.00)}} = 0.003$$

<u>Goodness of fit test</u>

Hypothesis:

Ho: The model fits

H1: The model does not fit

Significance level: $\alpha = 0.05$

Rejection region: Reject the null hypothesis if the p-value $\leq 0.05 = \alpha$

Table 4.18: Hosmer and Lemeshow Test(Utithini)

Step	Chi-square	df	Sig.
1	.000	8	1.000

Large Chi-square values (and correspondingly small p-values) indicate a lack of fit for the model. In table 4.18, we see that the Hosmer-Lemeshow Chi-square test for the final warranty model yields a p-value of 1.000 thus suggesting a model with good predictive value.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

5.1 CONCLUSIONS

- All the predictors are significant in the contamination of water in Athi. These make the water to have a zero probability of safety for drinking.
- The significant predictors in the contamination of Borehole water are pH, nitrate, turbidity and sulphates. These do not affect the safety of the water much because the probability of its safety is one. Thus, the water is safe for drinking.
- For Ianguni river water, the significant predictors are pH, color, nitrate, turbidity and sulphates. These reduce the probability of safety for the drinking water to zero.
- All the predictors are significant in the contamination of water in Kambi ya ndeke. These make the water to have a zero probability of safety for drinking.
- All the predictors are significant in the contamination of water in Athi. These reduce the probability of safety to 0.281 for drinking-water.
- All the predictors are significant in the contamination of water in Athi. These reduce the probability of safety for drinking-water to 0.003.

Generally, pH, nitrate, turbidity and sulphates are significant in all the sources. Thus, any change in these in the drinking-water makes a significant change in its safety.

The validity of these tests rely, however, on the assumption of large data. This is not the case as the tests show unsatisfactory behaviour with sparse data.

5.2 RECOMMENDATIONS

From the results, it is clear that drinking-water reaches the users in contaminated form. It is important to make sure that the water is subjected to the right processes before it is ready for use.

- Private partnerships should be promoted to urgently fill the gaps in sewerage management. Only 14% of Kenya's urban towns have sewerage systems and disposal facilities and this calls for a paradigm shift.
- Low-cost facilities in the short term should be adopted: such facilities can include septic tanks and ecosanitation as convectional sewerage may be too expensive for many to afford. Many view sewerage and sewage disposal as service without returns unlike water.

- Appropriate and comprehensive policy and legal instruments should be enacted and enforced on sewerage and sewage disposal: the policy should assign roles and functions and recognise the importance of mechanisms and co-ordination. The key actors here are the Ministries of Water and Irrigation, Health, Local Government, WRMA and NEMA.
- Industries should be compelled to pre-treat waste: industrial waste is often more toxic than domestic sewage as it contains higher concentrations of metals, chemicals and complex organic pollutants. Removing toxic elements of any industrial effluent in the form of pre-treatment is essential to safeguard water supply sources, sewerage and sewage treatment plants. Non adherence to pre-treatment of industrial wastes will lead to a near irreversible situation like the Nairobi river.
- Strict regulations and standards should be formulated for treatment plants and enforce these by fining polluters: this is critical as a deterrent measure, and can also act as a way of financing sewerage development and management through the enforcement of polluter pays principle.
- The residents of Yatta should be enlightened on the importance of safety in drinking-water. They should also be educated on how to keep their drinking-water safe. They should be shown the methods in which they can conserve the environment suitable to conserve the water sources and keep them safe.

APPENDICES

Appendix I: Athi River source data

safety	pH	colour	nitrate	turbidity	sulphates	lead
0	7.38	850	24	32	28	0.1
0	7.37	852	24.2	32	30	0.1
0	7.37	870	24.5	31	29	0.1
1	7.35	874	24.5	30.05	32	0
0	7.4	874	25	37	33.4	0.1
0	7.39	871	25	33	27	0.1
1	7.3	871	25.3	27	28.3	0
0	7	474	24	23.1	29.1	0.1
0	7.3	801	25.7	25	32.5	0.1
0	7.2	790	26.2	43	36	0.1
0	7.05	879	22	45	36.2	0.1
0	7.5	129	22.5	43	36	0.1
0	7.4	110	21.6	26	26.2	0.1
0	7.35	115	26	23	24	0.1

Appendix II: Borehole water source data

Safety	pH	colour	nitrate	turbidity	sulphates	Lead
1	7.5	0	10	0.9	3.93	0
1	7.4	0	10.3	0.9	3.93	0
1	8.06	0	9.5	0.9	3.93	0
1	7.37	0	9.5	0.86	3.9	0
1	7.2	0	12	1.16	3.7	0
1	7.19	0	10	0.1	4.3	0
1	7.3	0	10	0.81	4.1	0
1	7.2	0	7	0	3.7	0
0	7.3	0	10.3	0	3.9	0
1	7.2	0	7	1.84	4	0
1	7.1	0	3.4	1.9	4.3	0
1	7.37	0	3.6	1.84	4	0
1	7.9	0	5	0	3.6	0
1	7.8	0	4.3	0	3.1	0

Appendix III: Ianguni River water source data

safety	pH	Colour	Nitrate	Turbidity	Sulphates
0	7.72	115	11	11.4	98
0	7.74	116	11.3	11.4	98
0	7.7	116	11.5	11.4	98
0	7.77	103.5	11.5	8.33	101
0	7.6	112	13	10.79	112
0	7.75	117.5	12	11	119
0	7.65	123	12	10.7	118.5
1	7.6	3	10	8.47	121
0	7.7	206	11.3	10.9	119.8
0	7.6	210	10	14.4	123
0	7.52	213	13.7	14.7	124.3
0	7.8	48	5	14.4	98.1
0	8	43	7	8.2	92
0	7.9	46	8	7.9	87

Appendix IV: Kambi ya ndeke River source data

safety	pH	Colour	Nitrate	Turbidity	Sulphates	Lead
1	7.7	358	18.5	151	11.8	0
0	7.48	362	19	151	11.8	0
0	7.4	1206	18	151	11.8	0
0	7.76	1256	24.5	192	10.6	0
0	7.4	1218	36	178	11.3	0
0	7.4	1209	17	123	9.9	0.01
0	7.43	370	33	113	11	0
0	7.5	353	13	21.9	11.5	0
0	7.4	360	17.9	26	12.3	0.01
0	7.4	365	14	31.6	13	0
0	7.33	1256.3	36.7	180	13.7	0
0	7.9	301	8	31.6	13	0
0	7.9	304	10	24.8	9.1	0
0	7.8	299	9	21.6	7.9	0

Appendix V: Rain water source data

safety	pH	Colour	Nitrate	Turbidity	Sulphates	Lead
1	7.1	25	8.5	0.3	0.8	0
1	7.06	26	8.4	0.3	0.8	0
1	7.04	21	7.5	0.3	0.8	0
1	7.09	0	9	0.75	0.8	0
1	6.8	0	8	1.13	0.9	0
1	7.02	0	8	0.81	0.8	0
0	7.03	17	10	0.7	0.81	0
0	6.8	34	10	0.59	1.1	0
0	7.02	26	10	0.2	0.9	0
0	7	24	10	1.63	1	0
0	6.9	34.9	8	1.65	1.1	0
1	7.11	0	0.3	0	0.6	0
1	7.2	0	0.5	0	0.5	0
1	7.1	0	0	0	0.2	0

Appendix VI: Utithini River water source data

safety	pH	Color	Nitrate	Turbidity	Sulphates	Lead
0	7.41	7450	207	1425	14.5	0
0	7.47	7454	209	1425	14.5	0
0	8.01	7452	209	1425	14.5	0
0	7.38	7456	212	1457	14.1	0
0	7.5	7441	209	1463	13.5	0
0	7.4	7443	210	1469	13.9	0.01
0	7.35	7460	209	1477	13.7	0
0	7.35	7460	209	1477	13.7	0
0	7.5	7451	209	1477	13.7	0.01
0	7.5	7451	209	1477	13.7	0
0	7.5	7451	209	1477	13.7	0
1	8.3	379	33	1324	13.2	0
1	8.1	325	29	652	11.9	0
1	7.9	317	30.2	618	10.8	0

REFERENCES

Agresti, A., and Barbara, B. (1997), Statistical Methods for the Social Sciences. 3rd Edition. London, IWA Publishing.

Bartram J., Alexander, K. L., and Dauber, S. L. (2003). Heterotrophic plate counts and drinking-water safety: the significance of HPCs for water quality and human health. WHO Emerging Issues in Water and Infectious Disease Series. London, IWA Publishing.

Brisbane City Council Information (December, 2005). Drinking water quality and the drought.

Chao-Ying, J. P., Kuk, L. L., and Gray, M. I. - Indiana University (2002)

Collett, D. (1991). Modeling binary data. Chapman and Hall, London

Daily Nation. (June 2001). page 3

Dziuban E. J., Machamer, A. M., and Gruber, E. (22 December 2006). Surveillance for Waterborne Disease and Outbreaks Associated with Recreational Water — United States (2004). MMWR Surveill Summ. 55 (12): 1–30.

Hardalo C., and Edberg S. C. (1997). Pseudomonas aeruginosa: Assessment of risk from drinking-water. Critical Reviews in Microbiology, 23:47–75.

Hosmer, D. W., and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model, communications in statistics – Theory and Methods A10, 1043 – 1069.

IPCS (2000). Disinfectants and disinfectant by-products. Geneva, World Health Organization, International Programme on Chemical Safety (Environmental Health Criteria 216).

McCullagh, P., and Nelder, J. A. (1989). Generalized linear models, 2nd edition. Chapman and Hall, London.

Mitchell D. (September 1992) Department of Measurement, Statistics & Evaluation, University of Maryland.

Report to Congressional Requesters: Hazardous Waste Issues Pertaining to an Incinerator in East Liverpool, Ohio. (September 1994).

WHO (2003). Aldicarb in drinking-water. Background document for preparation of WHO Guidelines for drinking-water quality. Geneva, World Health Organization WHO/SDE/WSH/03.04/72).

WHO, UNICEF, water supply and sanitation council - 30th Nov. 2005.

Victorica J. and Galván M. (2001). Pseudomonas aeruginosa as an indicator of health risk in water for human consumption. Water Science and Technology, 43:49–52.