Modelling of Water Quality Index for drinking water in South Africa

Minenhle Sethebe

UNIVERSITY OF
KWAZULU - NATAL

INYUVESI
YAKWAZULU-NATALI

UNIVERSITY OF KWAZULU-NATAL

SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE

WESTVILLE CAMPUS, DURBAN, SOUTH AFRICA

Modelling of Water Quality Index for drinking water in South Africa

By

Minenhle Sethebe

Student Number: 218002141

Date: 26 November 2023

The report submitted to the University of KwaZulu-Natal in partial fulfilment

of the requirements for the Honours degree in Statistics.

ACKNOWLEDGEMENTS

First and foremost, I want to express my heartfelt appreciation to the divine powers and my forefathers for the wisdom, knowledge, and understanding that they have bestowed upon me, enabling me to successfully complete this research project. I am also grateful for the strength they have given me, which has enabled me to overcome the various obstacles that I have faced in recent years.

Secondly, I extend my sincerest thanks to my esteemed supervisors, Prof Retius Chifurira, Prof Knowledge Chinhamu, Dr Nombuso Zondo, and Dr Danielle Roberts, for their valuable time and invaluable insights. Their contributions to this project have been nothing short of amazing and highly beneficial.

Thirdly, I am deeply grateful to CSIR for affording me the incredible opportunity to explore and draw conclusions from their data.

Lastly, I want to express my profound gratitude to my wonderful mother, Mrs Thandokuhle Sethebe, and my dear family for their unwavering love and encouragement throughout this journey.

ABSTRACT

Access to safe drinking water is a fundamental requirement for maintaining good health and a basic human right. Developing countries, like South Africa, have suffered from a lack of access to safe drinking water from improved sources and to adequate sanitation services. Water contamination is a serious concern in South Africa.This study aims to investigate the Water Quality Index for drinking water across different catchment areas and explore dams and lakes in South Africa. The objective is to build a model that can estimate the Water Quality Index for drinking water, based on the values of various water quality indicators as well the catchment areas. The Water Quality Index changed over the years, from an unsuitable Water Quality Index to a good Water Quality Index. The catchment areas are significantly different from each other, even the amounts of chemical variables present differ with every catchment area (or even dam/lake). There is a great degree of association between the Water Quality Index and variables in water. The fitted multinomial logistic regression model is adequate for Water Quality Index predictions and forecasting for Water Quality index in South Africa.

TABLE OF CONTENT

Page

Chapter 1

INTRODUCTION

Access to safe drinking water is a fundamental requirement for maintaining good health and a basic human right (Jackson et al. 2001). Sadly, fresh water is already a scarce commodity in several regions globally, and this trend is set to worsen in the coming years due to factors such as population growth, urbanization, and climate change (Jackson et al. 2001).

According to Chen (2019), numerous dams have been constructed in arid and semi-arid regions to guarantee a consistent water supply for various purposes, including drinking, recreation, ecology, and irrigation. Preserving the sustainability of these dams requires addressing the significant concern of water quality (Chen, S.K et al 2019). If the water quality in the dam is subpar, it can impact human health, biodiversity, recreational use, and increase the cost of water treatment (Chen, S.K et al 2019). Unfortunately, due to the high levels of agriculture and urbanization around the dam, large amounts of wastewater, sewage, and agricultural pollutants are discharged into the reservoir systems (Chen, S.K et al 2019). This is causing the deterioration of water quality and restricting the potential uses of water (Chen, S.K et al 2019).

Having sufficient access to safe drinking water and sanitation is a global concern. However, developing countries, like South Africa, have suffered from a lack of access to safe drinking water from improved sources and to adequate sanitation services (WHO 2006).
Water contamination is a serious concern in South Africa. A report by the Human Rights Commission has highlighted the issue of raw sewage being discharged into the Vaal Dam, which supplies water to almost 20 million people across three provinces, including Johannesburg - the commercial heart of the country (International Trade Administration, 2022). Currently, only 64 percent of households in South Africa have access to safe and reliable water. Moreover, around 9 percent of the population still draws water from polluted sources, which is a worrying trend (International Trade Administration, 2022).

In order to effectively monitor water quality and establish management strategies, various physicochemical parameters are sampled at different times and locations (Ahmed A.N et al, 2019). However, collecting and analysing large data sets is resource-intensive and may obscure critical information. Due to limited funding for water quality monitoring, it has become necessary to explore cost-effective and thoughtful approaches to managing water quality (Ahmed A.N et al, 2019). Therefore, modelling water quality is crucial for accurately predicting future water quality trends (Ahmed A.N et al, 2019).

The Water Quality Index (WQI) is a straightforward and well-organized system for measuring water quality that is easy for non-technical individuals, policymakers, and water scientists to understand (Banda T, 2020). The index is presented as a relative scale ranging from zero (good water quality index) to greater than 75 (worst water quality index), making it a useful tool for communicating water quality data to a wide variety of audiences (Banda T, 2020).

Water Quality Index (WQI) models are essential for measuring contamination levels and determining necessary restoration efforts (Banda T, 2020).

This study aims to investigate the Water Quality Index for drinking water across different catchment areas and explore dams and lakes in South Africa. The objective is to build a model that can estimate the Water Quality Index for drinking water, based on the values of various water quality indicators as well the catchment areas.

Chapter 2

DATA ANALYSIS

2.1 Data Description

The data used in this study was provided by the Council for Scientific and Industrial Research (CSIR). It consisted of two separate Microsoft Excel datasets, one covering samples from 1970 to 1998, and the other from 1999 to 2012. The data included information about South African dams and lakes, as well as the variables (chemicals) found in the water. Overall, there were approximately 90103 observations and 48 variables, including numerical and categorical ones. The data included both measured and calculated variables. The sample dictionary provided brief descriptions of localities, coordinates (latitude and longitude), and sample type (dam, lake, spring, fountain, etc.) to provide more information on the sample station ID.

2.2 Data Preparation

From the original dataset, the complete duplicated rows (rows that had the exact same values for columns) were removed and they were three in total. The observation with the pH of 60 was removed, this is because pH scale ranges from 0 to 14 and it's impossible to have a pH of 60. The observations with unknown values (-9999 and zeros) were replaced with empty space (blank). The imputation by stratification method (grouping by year and sample station id) was used to impute for the empty spaces (missing values/blanks) with the aid of the mean. The observations from the same sample station ID, point ID and sampled on the same day were aggregated and averaged to a single observation.

The clean data was then merged with the dictionary to provide more information about the sample station location. The observations with the sample station id not on the dictionary (e.g., A9R004Q01) and the observations found to be rivers (e.g., C2H164Q01 Mooi River) on the Dams and Lakes data were removed.

There is 75872 approximately observations left after data cleaning

Table 1 consists of electrical conductivity statistical measures before and after imputation. The mean, variance, and standard deviation before and after imputation are not far apart, they are almost similar. The median, mode, range and interquartile range remained the same before and after imputation. Therefore, we can conclude that imputation did not change the structure of the data.

Table 1: The Electrical Conductivity Statistical Measures Before and After Imputation

EC BEFORE IMPUTATION Basic Statistical Measures			
Location		Variability	
Mean	126.5289	Std Deviation	528.66732
Median	35.2000	Variance	279489
Mode	8.0000	Range	12599
		Interquartile Range	39.50000

EC AFTER MEAN IMPUTATION Basic Statistical Measures			
Location		Variability	
Mean	126.4024	Std Deviation	528.29756
Median	35.2000	Variance	279098
Mode	8.0000	Range	12599
		Interquartile Range	39.50000

2.3 Exploratory Data Analysis

The primary aim is to gain a comprehensive comprehension of the dataset in question, by discovering patterns and identifying trends and potential issues or anomalies. The crucial objective for data exploration is to comprehend the data structure, review fundamental summary statistics that describe the central tendency, spread, and shape of the measured variables, investigate correlations between variables, and examine temporal patterns and trends.

Table 2 displays the correlation coefficients between the variables. The values represent the correlation coefficients between pairs of variables. the diagonal always contains one's as the variable is perfectly correlated with itself. Electrical conductivity has a positive correlation with the other variables but is strongly correlated to Calcium, Chlorine, Potassium, Magnesium, Sodium, Sulphate, and Total Dissolved Solids. Electrical Conductivity has a better correlation with Fluorine as compared to pH and Total alkalinity.

Table 2: The Spearman correlation matrix for variables used Water Quality Index and Model building.

Variables	EC	CA	CL	F	K	MG	NA	PH	SO4	TAL	TDS
EC	1										
CA	0,8356	1									
CL	0,9879	0,8350	1								
F	0,5331	0,5281	0,5133	1							
K	0,9696	0,8433	0,9773	0,5266	1						
MG	0,9755	0,8661	0,9852	0,5309	0,9791	1					
NA	0,9856	0,8495	0,9962	0,5179	0,9770	0,9854	1				
PH	0,0289	0,1158	0,0090	0,2684	0,0354	0,0373	0,0149	1			
SO4	0,9316	0,8493	0,9329	0,5261	0,9208	0,9464	0,9389	0,0400	1		
TAL	0,0963	0,1577	0,0551	0,3880	0,0956	0,0879	0,0841	0,4698	0,1489	1	
TDS	0,9882	0,8535	0,9977	0,5301	0,9794	0,9891	0,9982	0,0287	0,9487	0,1015	1

The Water Quality Index was calculated using R.M Brown et.al (1972) formula and the rough calculation is demonstrated on Table 3 below.

Table 3: The calculation of Water quality Index for drinking water.

Parameter	Sn=BIS Standard (required amount)	1/Sn	Sum(1/Sn)	K=1/Sum(1/Sn)	Weight Wn=K/Sn	Bn=Ideal Value	Pi (observed)	(Pn)/(Sn)	Quality Rating Qn=(Pn/Sn)*100
PH	8.5	0,117647059	0,836980392	1,194771119	0,140561308	7	5,82	-0,786666667	-78,66666667
Electrical Conductivity(EC)	250	0,004	0,836980392	1,194771119	0,004779084	0	6,8	0,0272	2,72
Total Dissolved Solids(TDS)	500	0,002	0,836980392	1,194771119	0,002389542	0	39	0,078	7,8
Calcium(Ca)	75	0,013333333	0,836980392	1,194771119	0,015930282	0	2,7	0,036	3,6
Magnesium(Mg)	30	0,033333333	0,836980392	1,194771119	0,039825704	0	1,9	0,063333333	6,333333333
Sodium(	1,5	0,666666667	0,836980392	1,194771119	0,79651408	0	0,17	0,113333333	11,33333333
		Sum(1/Si)= 0,836980392							

According to the information in Table 4, when the calculated Water Quality Index value is between 0and 50 that indicates a good water quality index for drinking water, when the calculated Water Quality Index value is between 50 and 75 that indicates poor water quality index for drinking water and when the calculated Water Quality Index value is greater than 75 that indicates unsuitable water quality index for drinking water.

Table 4: The interpretation of Water Quality Index for drinking Water

WQI status	WQI value
Good	0-50
Poor	50-75
Unsuitable	>75

The line graph in Figure 1 illustrates a sharp decrease in Water Quality Index as from year 1970 to year 1976. After year 1976, there was fluctuation in Water Quality Index and was between 20 and approximately 57.

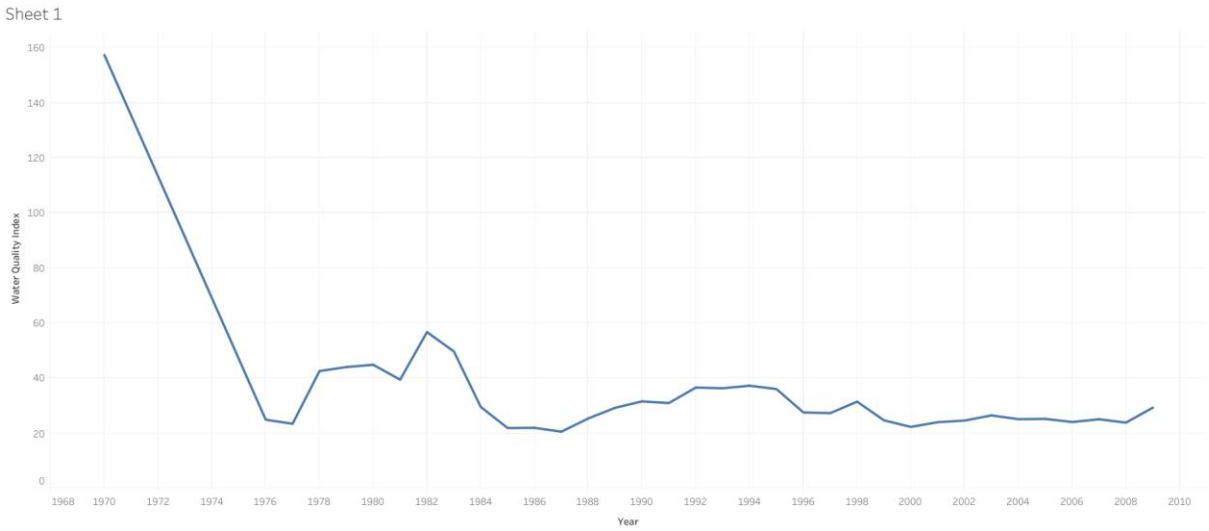


Figure 1: The Water Quality Index trend over the years.

The pie chart in figure 2 consists of information on Water Quality Index for all observations in the data. About 89% of the observations had good water quality index, 6% of the observations had poor water quality index and 4% of the observations had unsuitable water quality index.

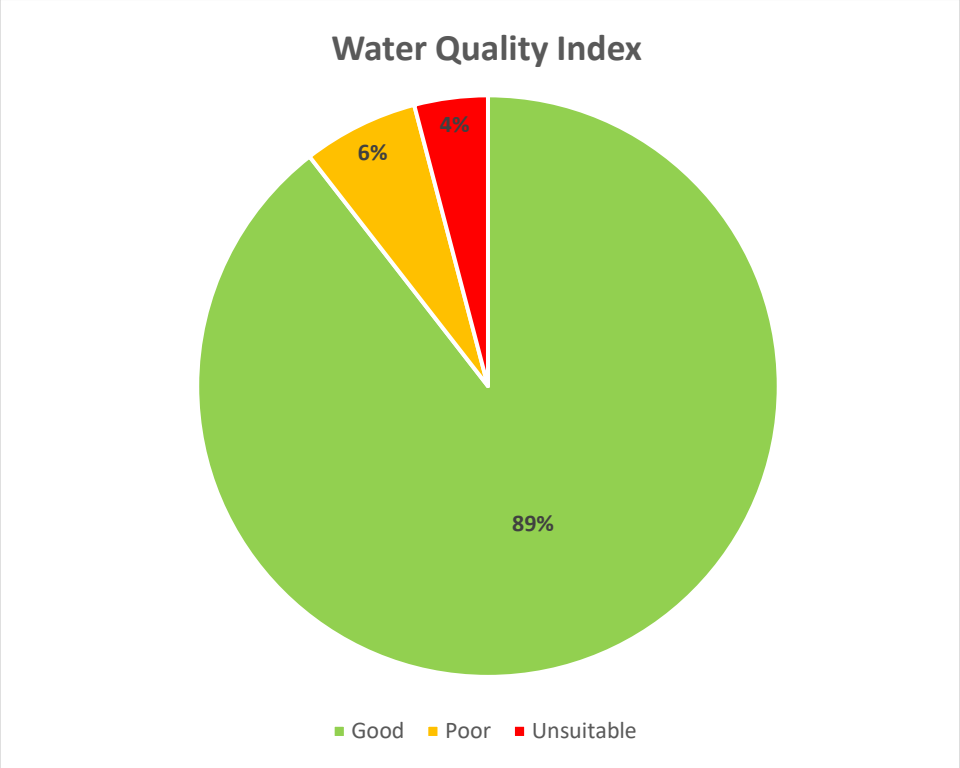


Figure 2: Water Quality Index distribution.

The bar graph in figure 3 illustrates average Water Quality Index for different catchment areas. The catchment area Kromme and Mfolozi had an average of poor water quality index. The catchment areas with an average of good water quality index are Sundays, Berg, Fish, Limpopo, Orange, Gourits, Vaal, Kei, Olifants (B), Bushmans, Keishkamma, Gamtoos, Nkomazi, Mzimvubu, Komati, Tugela, Swartkops, Breede and Olifants (e)

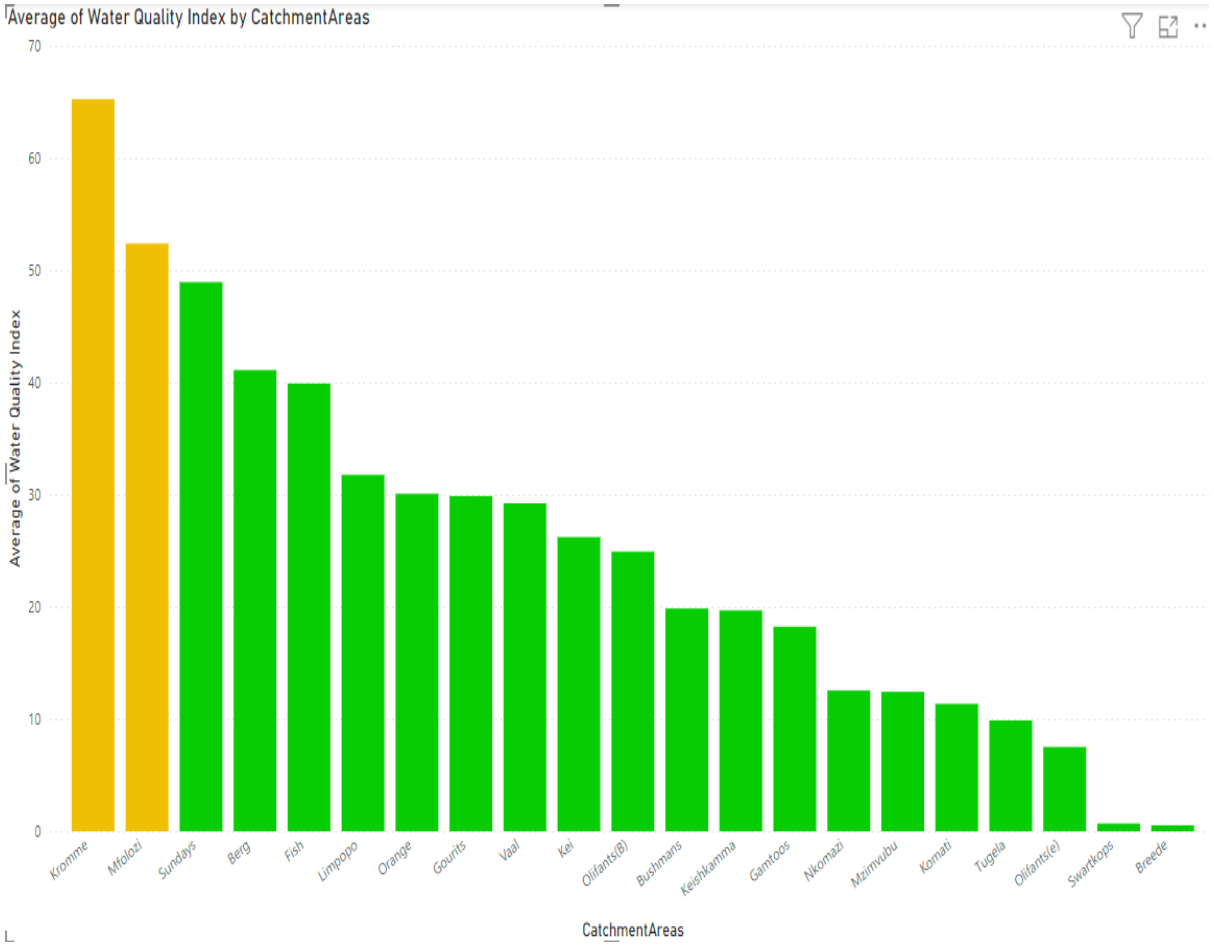


Figure 3: The Average Water Quality Index for different catchment areas.

The line graph in Figure 4 illustrates a sharp decrease in fluoride as from year 1970 to year 1976. After year 1976, there was fluctuation in fluoride and was between 0.2 and approximately 0.57 (mg/L)

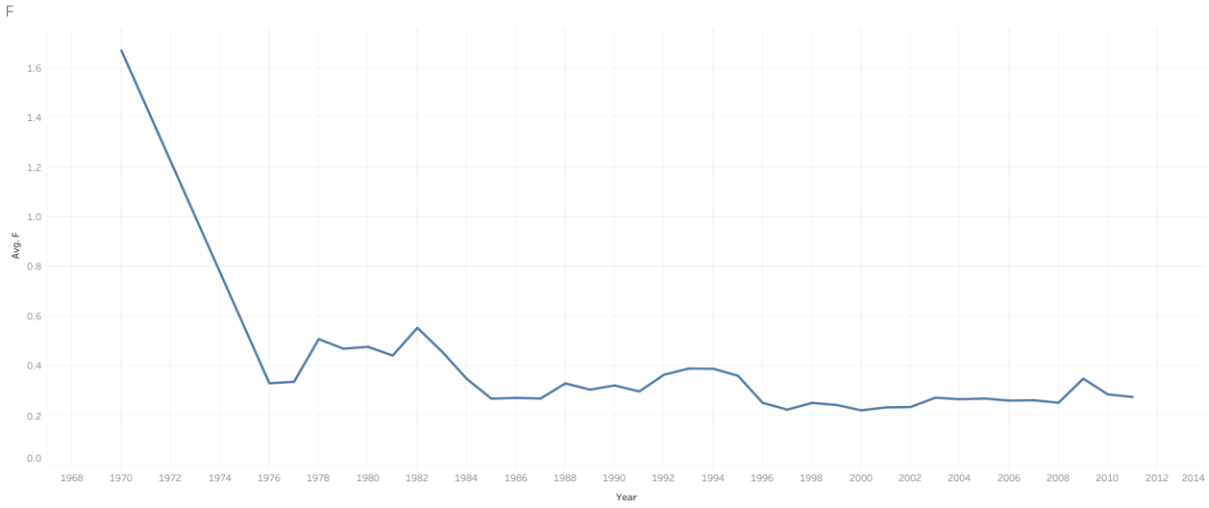


Figure 4: Historical Analysis of Fluoride trends (1970 to 2011).

The Map on figure 5 shows average concentration of fluoride for different sample stations (dams and lakes). The bubbles represent the different sample stations. The green bubble symbolizes that fluoride is between 0 and 0.5 (mg/L) indicating good water quality index. The yellow bubble symbolizes that fluoride is between 0.5 and 1.5 (mg/L) indicating poor water quality index. Many of the sample stations have good water quality index as compared to poor water quality index and there's no sample station with unsuitable water quality.

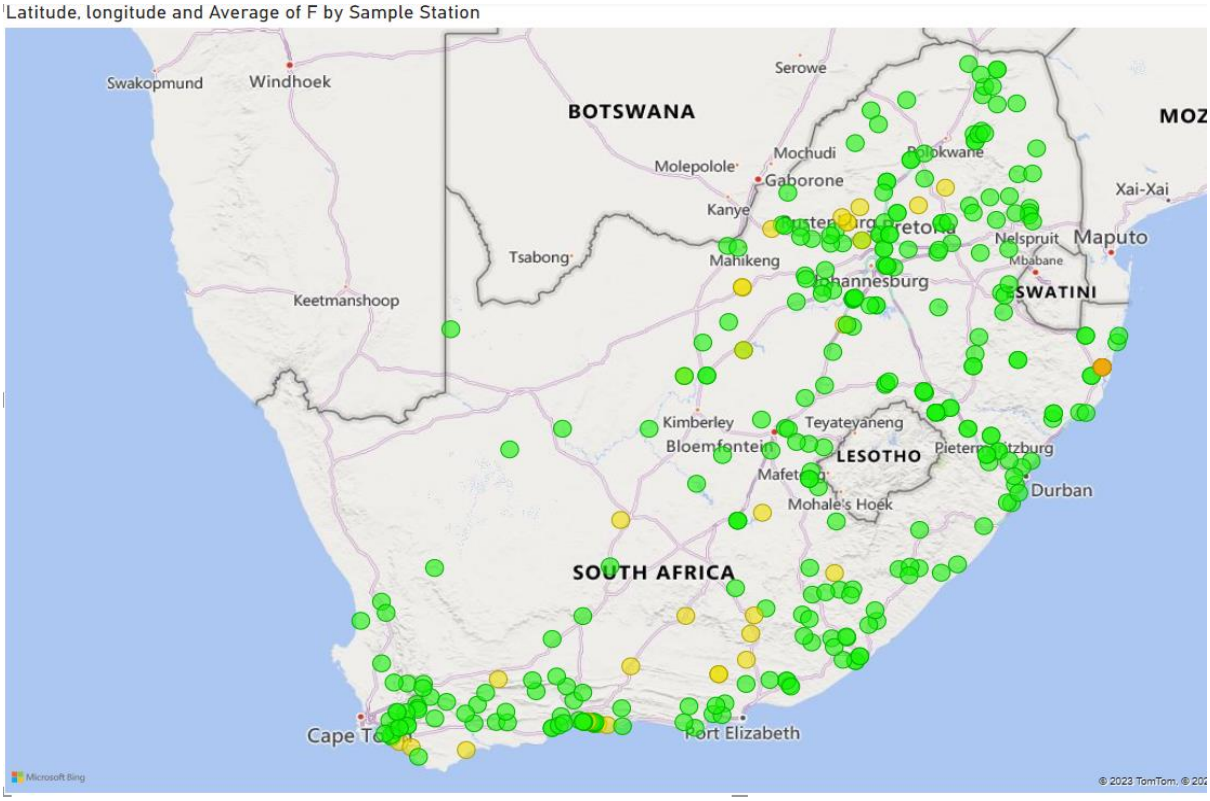


Figure 5: The Spatial Distribution of average concentration of Fluoride (F) in dams and lakes.

The line graph in Figure 6 illustrates a sharp decrease in chlorine as from year 1970 to year 1976. After year 1976, there was fluctuation in chlorine and was between 200 and approximately 1650 (mg/L). In year 1999 to 2011, chlorine stabilized in water.

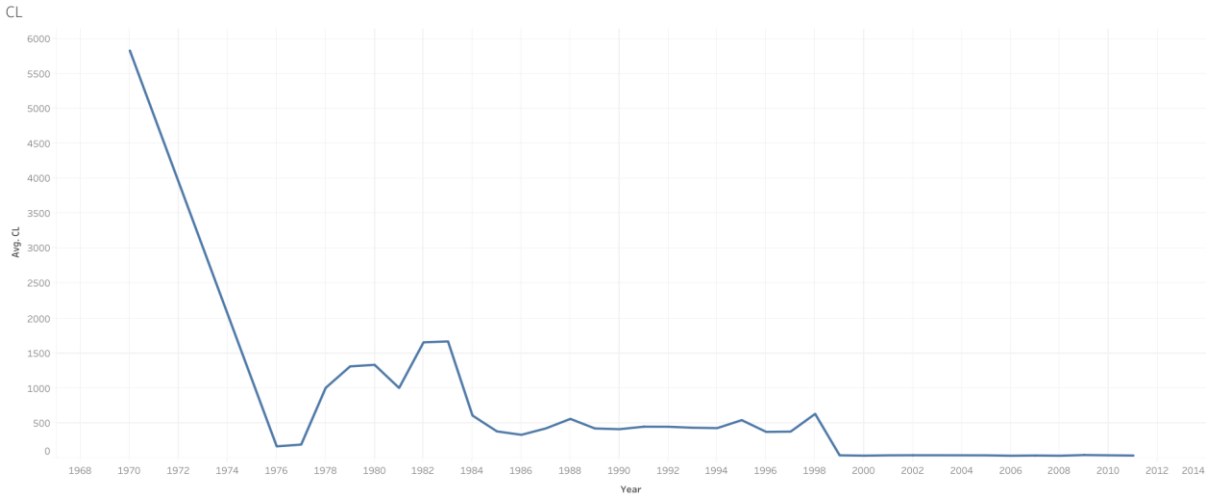


Figure 6: Historical Analysis of Chlorine trends (1970 to 2011).

The Map on figure 7, shows average concentration of chlorine for different sample stations (dams and lakes). The bubbles represent the different sample stations. The green bubble symbolizes that chlorine is between 0 and 60 (mg/L) indicating good water quality index. The yellow bubble symbolizes that chlorine is between 60 and 120 (mg/L) indicating poor water quality index. The red bubble symbolizes that chlorine is greater than 120 (mg/L) indicating unsuitable water quality index.
Majority of the sample stations have good water quality index and very few have unsuitable water quality index.

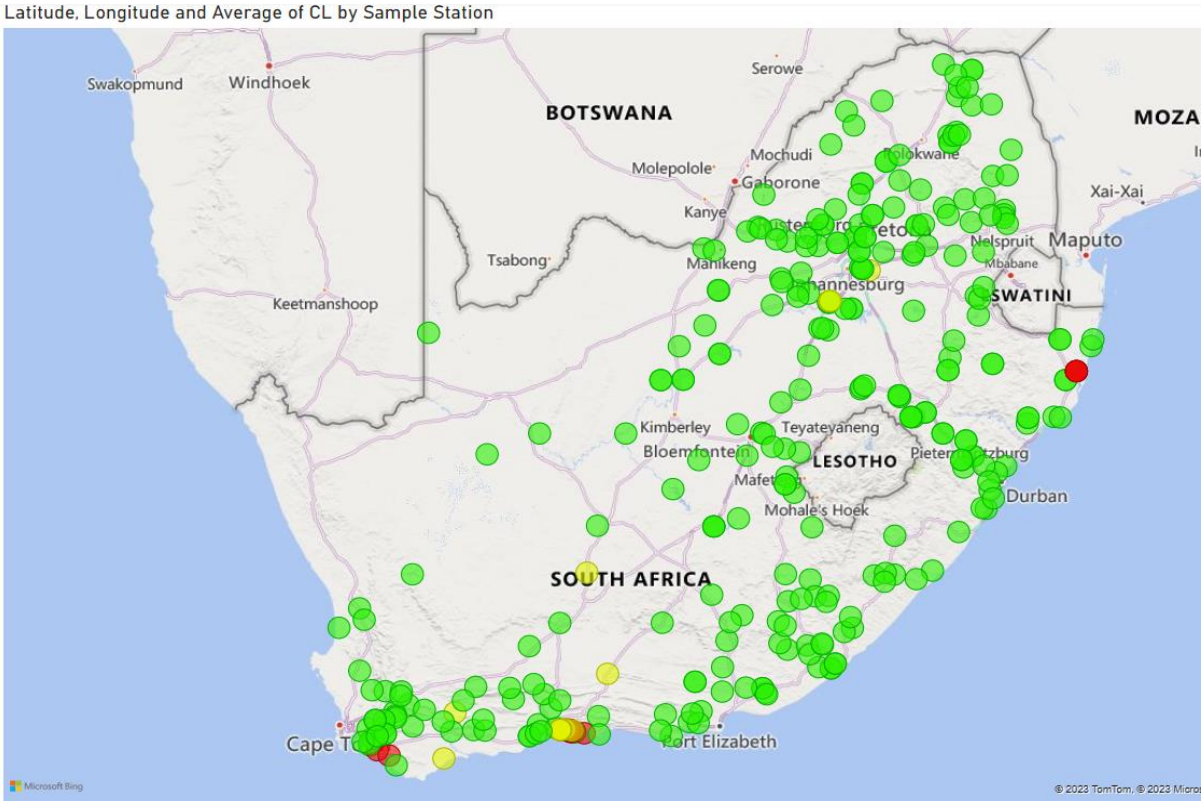


Figure 7: The Spatial Distribution of average concentration of Chlorine (Cl) in dams and lakes

Figure 8 illustrates that there is a strong positive correlation between fluoride and water quality index.

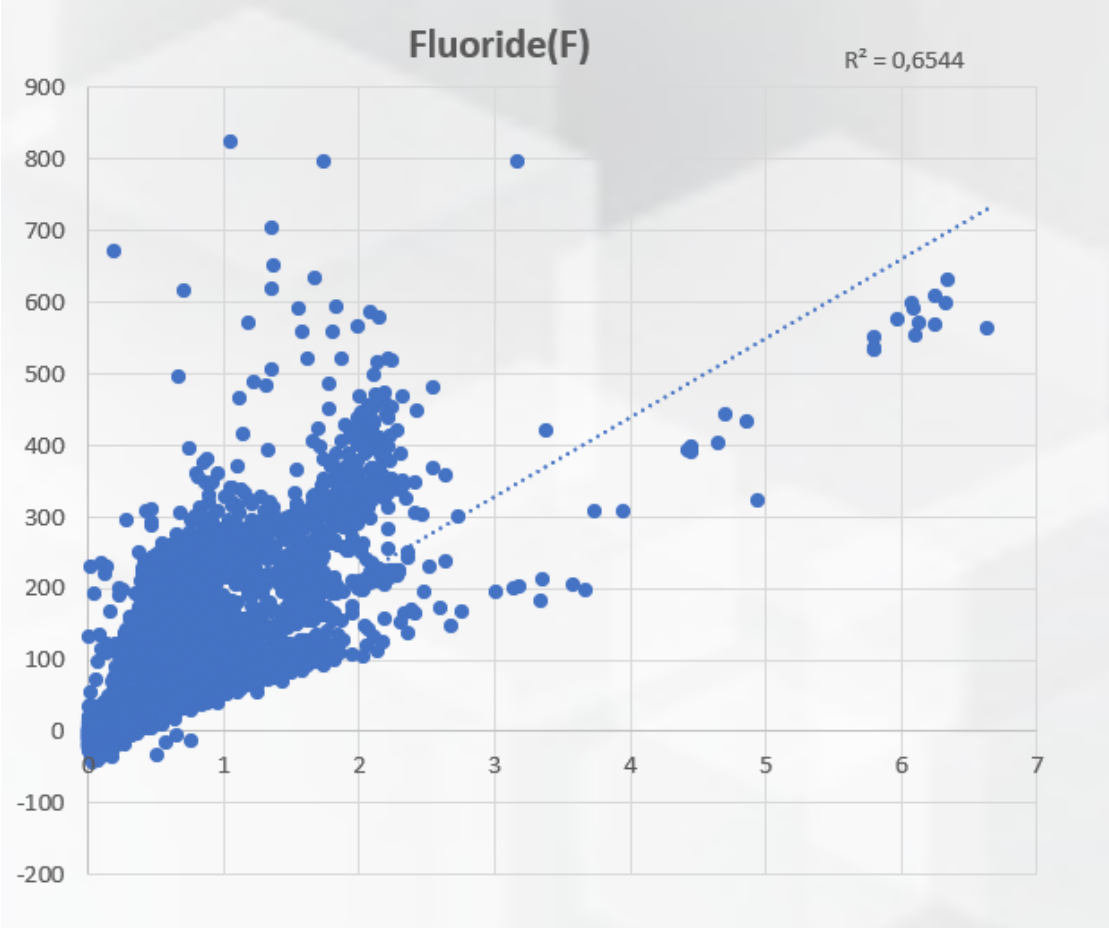


Figure 8: The Association between Water Quality Index and Fluoride

Figure 9 illustrates that there is a strong positive correlation between chlorine and water quality index.

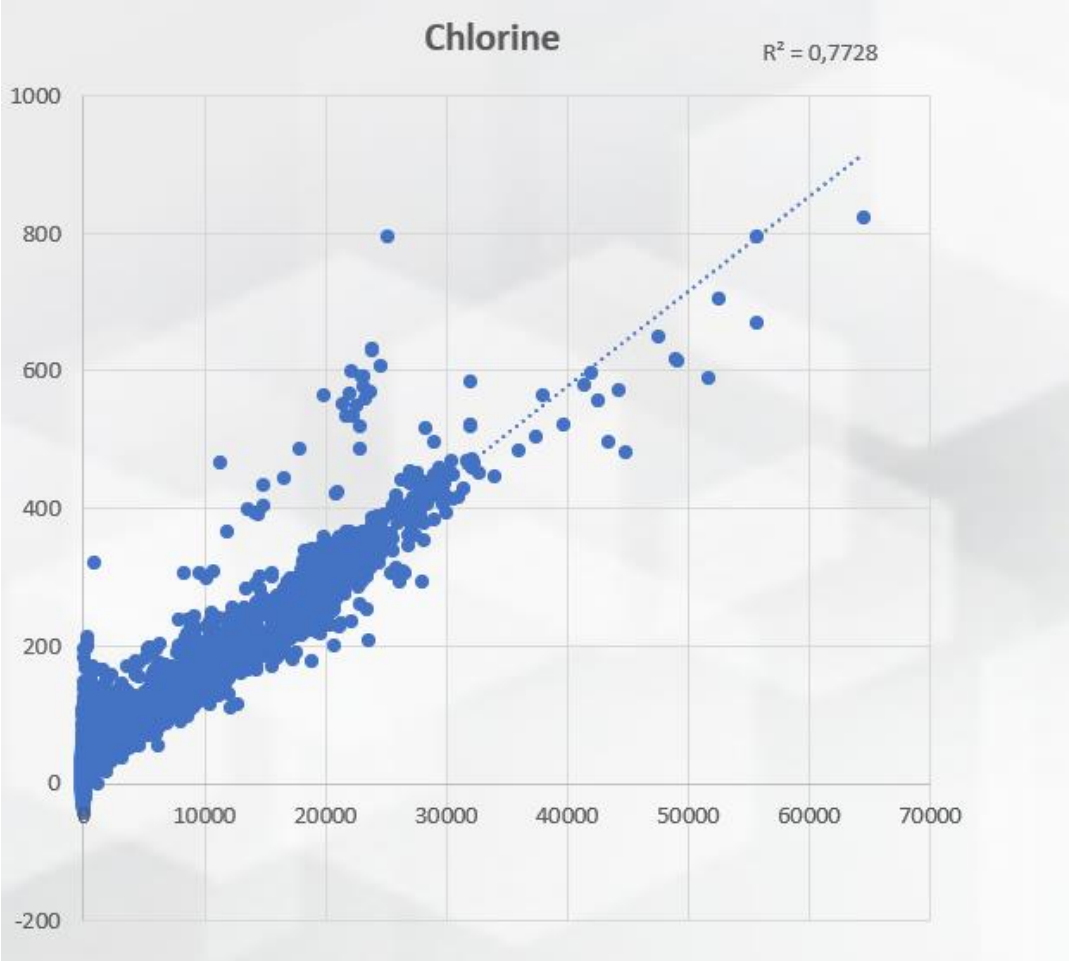


Figure 8: The Association between Water Quality Index and Fluoride

Chapter 3
METHODOLOGY

3.1) The Multinomial Logistic Regression

Multinomial logistic regression is used for categorial response variable predictions placement. The independent variables can be either dichotomous (i.e., binary) or continuous (i.e., interval or ratio in scale). Multinomial logistic regression is a an extension of binary logistic regression that cater for more than two categories in the dependent or outcome variable. The multinomial logistic regression uses maximum likelihood estimation to evaluate the probability. Multinomial logistic regression does have assumptions, such as the assumption of independence among the dependent variable choice.

The multinomial logistic model is an extended version of generalized linear models allowing for an estimation of nominal categorical response. The response variable should be categorical and consist of more than two categories for it to classified as multinomial.

3.1.1) The Multinomial distribution

Assume that a random variable with J categories. Let $\pi_1, \pi_2, \dots, \pi_J$ denote respective probabilities, with $\pi_1 + \pi_2 + \cdots + \pi_J = 1$. If there are n independent observations of Y which result in y_1 outcomes in category 1, y_2 outcomes in category 2, $\dots y_J$ outcomes in category J, with

The following is the multinomial distribution

$$f(y/n) = \frac{n!}{y_1!\, y_2! \dots y_J!} \pi_1^{y_1} \pi_2^{y_2} \dots \pi_J^{y_J} \qquad (1)$$

If J=2, then $\pi_2 = 1 - \pi_1$, $y_2 = n - y_1$ and is the binomial distribution. This multinomial does not satisfy the requirements for being a member of the exponential family. The association with the Poisson distribution, ensures that generalized linear modelling is appropriate.

The multinomial distribution arrives in the form above and to transform it into the exponential family distribution its relationship with the Poisson distribution is used.

$$log\frac{\pi_j}{\pi_J} = X\beta_{(j)} \qquad for\ j = 1,2,\dots J-1.$$

The $(J-1)$ logit equations are used simultaneously to estimate the parameters, $\beta_{(j)}$.
Note: We are going to estimate $(J-1)$ times $(p+1)\times 1$ βs. Because each $\beta_{(j)}$ is $(p+1)\times 1$.
That is, the effects vary according to the response paired with the baseline.

(2)

Where π_j is respective probability of the given category and π_J the probability of the reference category.
In a multinomial logistic regression model, dependent variables probability of being in category jπ_1= P(Y=j) is shown below

$$\pi_j = \frac{\exp(\sum_{k=1}^{K}\beta_{jk}x_k)}{1+\sum_{j=1}^{J-1}(\sum_{k=1}^{K}\beta_{jk}x_k)} \qquad (3)$$

3.2) Model Diagnostics and Goodness-of-Fit

All goodness-of-fit tests suggest that the model is significant and adequate. The Hosmer

Lemeshow goodness-of-fit test is used to test the adequacy of the model. Low p-values suggest rejection of the model

3.2.1) The Hosmer Lemeshow goodness-of-fit test

The Hosmer-Lemeshow test is used to determine the goodness of fit of the logistic regression model. Essentially it is a chi-square goodness of fit test for grouped data, usually where the data is divided into 10 equal subgroups. The initial version of the test we present here uses the groupings that we have used elsewhere and not subgroups of size ten.

Since this is a chi-square goodness of fit test, we need to calculate the HL statistic

$$(4)\quad \sum_{i=1}^{g}\sum_{j=1}^{2}\frac{(obs_{ij}-exp_{ij})^2}{exp_{ij}}$$

where g = the number of groups. The test used is chi-square with g – 2 degrees of freedom. A significant test indicates that the model is not a good fit and a non-significant test indicates a good fit.

3.3) For this study

Assumptions:

- Response variable is unordered
- Adequate sample size
- Independence in observations
- No multicollinearity

Good Water Quality Index vs Unsuitable Water Quality Index:

$$\text{logit(good)} = \beta_{0\text{good}} + \beta_{1\text{good}}(\text{Chlorine}) + \beta_{2\text{good}}(\text{Fluoride}) + \beta_{3\text{good}}(\text{Year}) + \beta_{\text{pgood}}(\text{Catchment Area})$$

Poor Water Quality Index vs Unsuitable Water Quality Index:

$$\text{logit(poor)} = \beta_{0\text{poor}} + \beta_{1\text{poor}}(\text{Chlorine}) + \beta_{2\text{poor}}(\text{Fluoride}) + \beta_{3\text{poor}}(\text{Year}) + \beta_{\text{p.poor}}(\text{Catchment Area})$$

Chapter 4

RESULTS AND DISCUSSION

The VIF, or Variance Inflation Factor, was the measure used to assess the severity of multicollinearity in regression analysis.

The VIF in Table 4.1 suggests that there is no multicollinearity. The variance of the estimated regression coefficient is not inflated.

Table 4.1: The Multicollinearity test using the variance inflation.

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	4.93086	0.05068	97.29	<.0001	0
CL	CL	1	0.00989	0.00001548	639.35	<.0001	1.36192
F	F	1	66.49252	0.13559	490.39	<.0001	1.36192

The joint test of the contribution of the variables to the model. The response variable was the water quality index and its reference category having an unsuitable water quality index. The exploratory variables were chlorine (Cl), Fluoride(F), year, and catchment area that had reference category of Vaal dam.

H0: Variable does not significantly improve model fit

H1: Variable significantly improves model fit

The chi-square test statistics and associated p-values (<0.05) shown in table 4.2 indicate each variable in the model significantly improve the model.

Table 4.2 The Type 3 Analysis of Effect

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
CL	2	2917.4357	<.0001
F	2	8592.1748	<.0001
YEAR	2	1423.4925	<.0001
CatchmentAreas	40	1656.8173	<.0001

With reference to Table 4.3, the Likelihood ratio test, Score test and Wald test were all significant at 5% level of significance. This means that the fitted model is better than the model with intercepts only.

Table 4.3: Testing Global Null Hypothesis: Beta=0

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	101824.727	46	<.0001
Score	125422.885	46	<.0001
Wald	9541.0121	46	<.0001

The Hosmer and Lemeshow goodness of fit test states that:

H0: The fitted model is adequate

H1: The fitted model is not adequate

In Table 4.4 the Pvalue=0.105>0.05, H0 is not rejected at 5% level of significance. Therefore, the fitted model is adequate.

Table 4.4: The Hosmer and Lemeshow Goodness-of-Fit Test

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
9.0572	2	0.0108

From Table 4.5, about 94.72% of variance in Water Quality Index could be explained in the exploratory variables. Therefore, the model is good fit.

Table 4.5: The Adjusted R-square for model testing

Root MSE	8.49886	R-Square	0.9472
Dependent Mean	30.23443	Adj R-Sq	0.9472
Coeff Var	28.10986		

For catchment area Bushmans, the odds of good water quality index instead of unsuitable water quality index are 32.422 times of that of catchment area Vaal having good water quality index instead of unsuitable water quality index. For catchment area Fish, the odds of good water quality index instead of unsuitable water quality index are 4.130 times of that of catchment area Vaal having good water quality index instead of unsuitable water quality index. For catchment area Gamtoos, the odds of good water quality index instead of unsuitable water quality index are 4.801 times of that of catchment area Vaal having good water quality index instead of unsuitable water quality index. For catchment area Gourits, the odds of good water quality index instead of unsuitable water quality index are 0.194 times of that of catchment area Vaal having good water quality index instead of unsuitable water quality index. For catchment area Kei, the odds of good water quality index instead of unsuitable water quality index are 53.253 times of that of catchment area Vaal having good water quality index instead of unsuitable water quality index. For catchment area Kromme, the odds of good water quality index instead of unsuitable water quality index are 83.930 times of that of catchment area Vaal having good water quality index instead of unsuitable water quality index. For catchment area Limpopo, the odds of good water quality index instead of unsuitable water quality index are 3.423 times of that of catchment area Vaal having good water quality index instead of unsuitable water quality index. For catchment area Mfolozi, the odds of good water quality index instead of unsuitable water quality index are 77.559 times of that of catchment area Vaal having good water quality index instead of unsuitable water quality index. For catchment area Olifant (B), the odds of good water quality index instead of unsuitable water quality index are 438.649 times of that of catchment area Vaal having good water quality index instead of unsuitable water quality index. For catchment area Sundays, the odds of good water quality index instead of unsuitable water quality index are 2.689 times of that of catchment area Vaal having good water quality index instead of unsuitable water quality index.

For catchment area Fish, the odds of poor water quality index instead of unsuitable water quality index are 2.394 times of that of catchment area Vaal having poor water quality index instead of unsuitable water quality index. For catchment area Gourits, the odds of poor water quality index instead of unsuitable water quality index are 0.073 times of that of catchment area Vaal having poor water quality index instead of unsuitable water quality index. For catchment area Kei, the odds of poor water quality index instead of unsuitable water quality index are 11.920 times of that of catchment area Vaal having poor water quality index instead of unsuitable water quality index. For catchment area Limpopo, the odds of poor water quality index instead of unsuitable water quality index are 5.244 times of that of catchment area Vaal having poor water quality index instead of unsuitable water quality index. For catchment area Olifant (B), the odds of poor water quality index instead of unsuitable water quality index are 35.407 times of that of catchment area Vaal having poor water quality index instead of unsuitable water quality index. For catchment area Orange, the odds of poor water quality index instead of unsuitable water quality index are 3.017 times of that of catchment area Vaal having poor water quality index instead of unsuitable water quality index. For catchment area Swartkops, the odds of poor water quality index instead of unsuitable water quality index are 0.003 times of that of catchment area Vaal having poor water quality index instead of unsuitable water quality index. For catchment area Tugela, the odds of poor water quality index instead

of unsuitable water quality index are 291.748 times of that of catchment area Vaal having poor water quality index instead of unsuitable water quality index.

Table 4.6: The Analysis of Maximum Likelihood Estimates

			Good WQI			Poor WQI		
Parameter		DF	Estimate	Pvalue	Odds Ratio	Estimate	Pvalue	Odds Ratio
Intercept		1	458.7	<.0001	N/A	235.9	<.0001	N/A
CL		1	-0.00899	<.0001	0.991	-0.00287	<.0001	0.997
F		1	-41.1309	<.0001	<0.001	-18.0969	<.0001	<0.001
YEAR		1	-0.2132	<.0001	0.808	-0.1090	<.0001	0.897
CatchmentAreas	Berg	1	7.1158	<.0001	>999.999	-0.1599	0.7105	0.852
CatchmentAreas	Breede	1	9.2028	0.5724	>999.999	8.4355	0.6044	>999.999
CatchmentAreas	Bushmans	1	3.4788	0.0137	32.422	1.2151	0.2996	3.371
CatchmentAreas	Fish	1	1.4183	0.0007	4.130	0.8730	0.0303	2.394
CatchmentAreas	Gamtoos	1	1.5688	0.0041	4.801	-0.0815	0.8503	0.922
CatchmentAreas	Gourits	1	-1.6386	0.0004	0.194	-2.6147	<.0001	0.073
CatchmentAreas	Kei	1	3.9751	<.0001	53.253	2.4782	<.0001	11.920
CatchmentAreas	Keishkamma	1	-0.6830	0.8952	0.505	-2.0516	0.6911	0.129
CatchmentAreas	Komati	1	10.6216	0.9685	>999.999	8.5776	0.9745	>999.999
CatchmentAreas	Kromme	1	4.4300	<.0001	83.930	0.5062	0.1354	1.659
CatchmentAreas	Limpopo	1	1.2307	0.0003	3.423	1.6571	<.0001	5.244
CatchmentAreas	Mfolozi	1	4.3510	<.0001	77.559	0.3357	0.3719	1.399
CatchmentAreas	Mzimvubu	1	-1.6677	0.9988	0.513	-11.8612	0.9940	<0.001
CatchmentAreas	Nkomazi	1	17.6423	0.9371	>999.999	14.6190	0.9479	>999.999
CatchmentAreas	Olifants(B)	1	6.0837	<.0001	438.649	3.5670	<.0001	35.409
CatchmentAreas	Olifants(e)	1	2.7556	0.9953	42.749	-8.9149	0.9896	<0.001
CatchmentAreas	Orange	1	-0.1181	0.8086	0.889	1.1042	0.0181	3.017
CatchmentAreas	Sundays	1	0.9891	0.0129	2.689	0.2752	0.4721	1.317
CatchmentAreas	Swartkops	1	8.1240	0.9879	>999.999	0.9941	0.003	0.003
CatchmentAreas	Tugela	1	11.0131	<.0001	>999.999	5.6759	<.0001	291.748

Chapter 5
CONCLUSION

The aim was to investigate the Water Quality Index for drinking water across different catchment areas and explore dams and lakes in South Africa.

The Water Quality Index changed over the years, from an unsuitable Water Quality Index in 1970 to a good Water Quality Index. The catchment areas are significantly different from each other, even the amounts of chemical variables present differ with every catchment area (or even dam/lake).

There is a great degree of association between the Water Quality Index and fluoride. The presence of fluoride does influence the state of the Water Quality Index for drinking water. There is a great degree of association between the Water Quality Index and chlorine in water. The amount and presence of chlorine in drinking water do influence the state of the Water Quality Index.

The majority of the dams and lakes sampled had the recommended amount of fluoride, indicating a good Water Quality Index for drinking water. Similarly, the majority of dams and lakes had the recommended amount of chlorine, indicating a good Water Quality Index for drinking water. There were very few dams or lakes with chlorine above the recommended range.

Although Water Quality differs with every catchment area, the majority of the catchment areas have a good Water Quality Index only Kromme and Mfolozi have a poor Water quality index. With a very large percentage, the dams and lakes in South Africa had a good Water Quality Index.

This is no surprise because South Africa is a developing country and it has changed in the past, including its issues related to pollution, sanitation, and insufficient access to healthy drinking water.

The fitted multinomial logistic regression model is adequate for Water Quality Index predictions and forecasting in South Africa. The model also be used in overall Water Quality assessment and early detection of issues.

REFERENCES

Ahmed, A.N., Othman, F.B., Afan, H.A., Ibrahim, R.K., Fai, C.M., Hossain, M.S., Ehteram, M. and Elshafie,
A., 2019. Machine learning methods for better water quality prediction. *Journal of Hydrology*, *578*, p.124084

Banda T.D., 2020. *Development of a universal water quality index variability model for South Africa river catchments*. University of KwaZulu-Natal, South Africa.

Chen, S.K., Jang, C.S. and Chou, C.Y., 2019. Assessment of spatiotemporal variations in river water quality
for sustainable environmental and recreational management in the highly urbanized Danshui River
basin. Environmental monitoring and assessment, 191(2), p.100.

Fagerland,M (2012)A generalized Hosmer-Lemeshow goodness-of-fit test for multinomial logistic regression models, The Stata Journal,447-453

Goeman, J.J. & le Cessie (2006, December) A goodness-of-fit test for Multinomial Logistic Regression, Biometrics,980-985

Hosmer, D. W., & Lemeshow, S. (1989). Applied logistic regression. New York: Wiley.

World Health Organization (2006) In water, sanitation and health world health Organization

Jackson et al (2001) Water in changing world, Issues in Ecology. Ecol Soc Am, Washington, pp 1–16

APPENDIX

Microsoft Excel (Inc Power Query) and SQL Server.

Data Preparation was done on Microsoft excel with help of SQL Server and Microsoft Power Query.

Microsoft Power BI and Tableau

Data Exploration and Visualisation was done on Power BI and Tableau Public.

SAS Studio

The model was built in SAS using the below code.

```
/*Importing Data*/
proc import datafile="/home/u58791075/Beloved.xlsx" dbms=xlsx
out=rough.beloved;
sheet=sheet1 ;
run ;

/*Test of Multicollinearity*/
proc reg data=rough.beloved ;
model WnQn= Cl f / vif  ;
TITLE ‘Test for multicollinearity’;
run;

/*Multinomial Logistic Regression Model and Goodness of Fit Test*/
proc logistic data=rough.beloved plots=all ;
class WQI (ref="Unsuitable") CatchmentAreas(ref="Vaal") /param=ref;
 model WQI= Cl F Year CatchmentAreas / link=glogit lackfit;
run;
```

Below is the SAS output of the above mentioned code.

'Test for multicollinearity'

The REG Procedure
Model: MODEL1
Dependent Variable: WnQn WnQn

Number of Observations Read	73765
Number of Observations Used	73765

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	95667126	47833563	662235	<.0001
Error	73762	5327870	72.23055		
Corrected Total	73764	100994995			

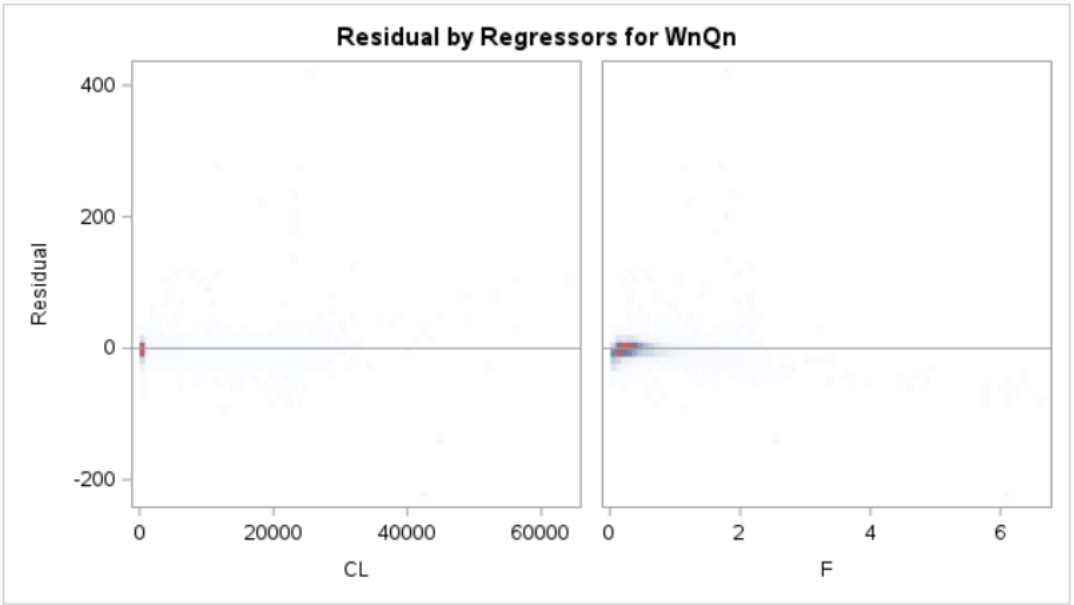
Root MSE	8.49886	R-Square	0.9472
Dependent Mean	30.23443	Adj R-Sq	0.9472
Coeff Var	28.10986		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	4.93086	0.05068	97.29	<.0001	0
CL	CL	1	0.00989	0.00001548	639.35	<.0001	1.36192
F	F	1	66.49252	0.13559	490.39	<.0001	1.36192

'Test for multicollinearity'

The REG Procedure
Model: MODEL1
Dependent Variable: WnQn WnQn



The LOGISTIC Procedure

Model Information		
Data Set	ROUGH.BELOVED	Posterior Probabilities for DATA=WORK.FILTERED_DATA.
Response Variable	WQI	WQI
Number of Response Levels	3	
Model	generalized logit	
Optimization Technique	Newton-Raphson	

Number of Observations Read	73765
Number of Observations Used	73765

Response Profile		
Ordered Value	WQI	Total Frequency
1	Good	66007
2	Poor	4744
3	Unsuitable	3014

Logits modeled use WQI='Unsuitable' as the reference category.

			Good WQI			Poor WQI		
Parameter		DF	Estimate	Pvalue	Odds Ratio	Estimate	Pvalue	Odds Ratio
Intercept		1	458.7	<.0001	0.991	235.9	<.0001	0.997
CL		1	-0.00899	<.0001	<0.001	-0.00287	<.0001	<0.001
F		1	-41.1309	<.0001	0.808	-18.0969	<.0001	0.897
YEAR		1	-0.2132	<.0001	>999.999	-0.1090	<.0001	0.852
CatchmentAreas	Berg	1	7.1158	<.0001	>999.999	-0.1599	0.7105	>999.999
CatchmentAreas	Breede	1	9.2028	0.5724	32.422	8.4355	0.6044	3.371
CatchmentAreas	Bushmans	1	3.4788	0.0137	4.130	1.2151	0.2996	2.394
CatchmentAreas	Fish	1	1.4183	0.0007	4.801	0.8730	0.0303	0.922
CatchmentAreas	Gamtoos	1	1.5688	0.0041	53.253	-0.0815	0.8503	0.073
CatchmentAreas	Gourits	1	-1.6386	0.0004	0.505	-2.6147	<.0001	0.194
CatchmentAreas	Kei	1	3.9751	<.0001	>999.999	2.4782	<.0001	11.920
CatchmentAreas	Keishkamma	1	-0.6830	0.8952	83.930	-2.0516	0.6911	0.129
CatchmentAreas	Komati	1	10.6216	0.9685	3.423	8.5776	0.9745	>999.999
CatchmentAreas	Kromme	1	4.4300	<.0001	77.559	0.5062	0.1354	1.659
CatchmentAreas	Limpopo	1	1.2307	0.0003	0.189	1.6571	<.0001	5.244
CatchmentAreas	Mfolozi	1	4.3510	<.0001	>999.999	0.3357	0.3719	1.399
CatchmentAreas	Mzimvubu	1	-1.6677	0.9988	438.649	-11.8612	0.9940	<0.001
CatchmentAreas	Nkomazi	1	17.6423	0.9371	15.730	14.6190	0.9479	>999.999
CatchmentAreas	Olifants(B)	1	6.0837	<.0001	0.889	3.5670	<.0001	35.409
CatchmentAreas	Olifants(e)	1	2.7556	0.9953	2.689	-8.9149	0.9896	<0.001
CatchmentAreas	Orange	1	-0.1181	0.8086	>999.999	1.1042	0.0181	3.017
CatchmentAreas	Sundays	1	0.9891	0.0129	>999.999	0.2752	0.4721	1.317
CatchmentAreas	Swartkops	1	8.1240	0.9879	-5.9231	0.9941	0.003	0.003
CatchmentAreas	Tugela	1	11.0131	<.0001	>999.999	5.6759	<.0001	291.748

Partition for the Hosmer and Lemeshow Test							
Group	Total	Observed WQI = Good	Observed WQI = Poor	Observed WQI = Unsuitable	Expected WQI = Good	Expected WQI = Poor	Expected WQI = Unsuitable
1	7381	670	3697	3014	693.12	3674.09	3013.78
2	5624	4657	967	0	4651.99	971.80	0.21
3	60760	60680	80	0	60661.9	98.11	0.00

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
4.5080	2	0.1050