James Magombo (South Africa), Bloodless Dzwairo (South Africa), Sibusiso Moyo (South Africa), Mendon Dewa (South Africa)

Data pre-processing for process optimization at a drinking water treatment plant in Ugu District Municipality, South Africa

Abstract

When testing and recording water quality data from treatment plants, errors arise. The errors are in the form of recordings left blank (missing values), obvious errors in writing or typing, or they can be as a result of values being very small to detect and are therefore censored. The censored values are known to be below the limit of detection (LOD). In statistical analysis, the blank cells can be filled with a certain value. Censored values are often corrected by substituting with a constant value throughout. This value will be a fraction of the limit of detection and most commonly used fractions are, half the limit of detection, the limit of detection divided by the square root of 2, or multiplying the limit of detection by 0.75. The direct substitution method for handling missing and values below the limit of detection results in a uniform distribution for values below the limit of detection, and a true distribution for those above. As a result, treatment of the values below the limit of detection is dependent upon their percentage in the sample size. An alternative method used will mimic the characteristic of the distribution pattern of the values above the limit of detection to estimate the values below it. This can be done with an extrapolation technique or maximum likelihood estimation.

In this study, data from the Umzinto Water Treatment Plant was used to develop a data pre-processing program using Visual Basics for Applications (VBA) and Microsoft Excel 2013. The procedure involved 4 stages: data preparation, data pre-processing for blanks and non-detects, data pre-processing for the censored values and finally the identification of the outliers. The developed program was then used to pre-process raw water quality data, which resulted in satisfactory process time and data conversion. The methodology used can be borrowed for the pre-processing of data driven environmental models and hence it has a great influence on sustainability of water treatment plants.

Keywords: optimization, data pre-processing, water treatment, visual basics, sustainability.
JEL Classification: N57, Q53, Q56.

Introduction

The quality and quantity of drinking water is greatly dependent on the sources from which it is drawn. Throughout the world, the state of raw water sources have been deteriorating (Schutte, 2006; Mosley et al., 2012; O'Reilly and Bezuidenhout, 2013; Saeed and Hashmi, 2014). Decrease in quality generally means that the water is becoming more expensive to treat, thus more sophisticated technologies will be required in the future to make it suitable for drinking (Lange and Hassan, 2006; Dearmont et al., 1998; Netshidaulu, 2007; Dzwairo, 2011). Decrease in quantity results in underutilization of the treatment plants and in most cases failure to meet demand (Chang et al., 2013). This is a challenge to most treatment plants as there are a lot of uncertainties involved in trying to determine the quality and quantity of both raw water at the abstraction point and treated water (Loucks et al., 2005).

© James Magombo, Bloodless Dzwairo, Sibusiso Moyo and Mendon Dewa, 2015.
James Magombo, B.Eng., Honors Industrial Engineering, Postgraduate Masters Student, Department of Industrial Engineering, Faculty of Engineering and the Built Environment, Durban University of Technology, South Africa.
Bloodless Dzwairo, D.Tech., Research Fellow, Institute of Systems Science, Durban University of Technology, South Africa.
Sibusiso Moyo, Ph.D., Director, Institute of Systems Science, Durban University of Technology, South Africa.
Mendon Dewa, M.Sc., Manufacturing Systems, Lecturer, Durban University of Technology, South Africa.

Even a small water treatment plant can be very complex compared to a manufacturing plant, which receives raw water and other inputs of almost the same quality. Water treatment plants receive water of variable quality because of various activities which cause pollution (Gertsen and Sønderby, 2009). Although the raw water quality often varies, drinking water quality is still required to meet specifications as stipulated in SANS 241:11 (DWAF, 1996; SABS, 2011). This scenario then calls for implementation of optimization techniques during water treatment.

Water treatment optimization generally involves capacity increment, pollution reduction, and energy and cost reduction. With the large variety of optimization methods available, Breese (2006) suggested that the key and first stage of optimization was trending. Because raw data from treatment plants inevitably contains errors, trending is done in order to analyze it and present the results in a usable format. Errors include, among others, blanks, censored values and outliers (Peng, 2010; Shumway et al., 2002).

It is important that as demand for good quality water increases, it also is available to consumers. This is in line with the global theme of sustainability, which is the baseline of the Brundtland Report (Brundtland, 1987). In this report, sustainable development is defined as "development that meets the needs of the present without compromising the

ability of future generations to meet their own needs" (Brundtland, 1987). Sustainability therefore carries with it the responsibility of paying attention to everyone's needs including the world's poor and other vulnerable groups. It also takes into consideration, limitations of technology and ability of the environment to meet today's needs and those for the future.

Data pre-processing has an effect on environmental models that are data driven, hence it has a great influence on sustainability of water treatment plants (Brown and Kros, 2003; Zhu et al., 2011). Before model development, data should be pre-processed so that it represents the accurate processes happening. Therefore the primary objective of this study was to develop a "Visual Basics for Applications" program that would be used to pre-process data from the Umzinto Water Treatment Plant to eliminate all the possible errors in the data set. The secondary objective was to pre-process data for the Umzinto Water Treatment Plant.

The common problems that were found in the Umzinto Water Treatment Plant data are similar to those found in most treatment plants. The data problems can be divided into basically three types:

- blanks or obvious errors;
- censored values;
- outlying values.

Obvious errors usually result from a recording error. They are generally fixed by consulting the recorder or operator to fill in the correct figure. A problem arises when the person responsible has forgotten the correct figure. For blanks, the general notion taken is that a blank is not equal or similar to a zero. Usually blanks are as a result of a test not being done or when the operator decides that the value is too insignificant to be recorded.

Censored data are usually defined by the "less than" or "greater than" sign before a number in a result. Usually they are a result of an observation that is above or below a detectable level. Detection levels are usually a result of two components, either an instrument's detection limit or the method's detection limit. In this article the focus is on the method's limit of detection.

Some researchers noted that outliers are usually defined by graphical means (Rousseeuw and Leroy, 2005; Hodge and Austin, 2004). Additionally, Grubbs (1969) defined outliers as observations that appear to be distant or to deviate from all the other observations. In a distribution system, outliers can occur by chance while in some cases they can be an indication of a measurement error. Outliers can also be a characteristic of a heavy-tailed distribution. It is therefore important to identify the outliers in order to make correct decisions in treating them.

Taylor (1987), defined the limit of detection as the level at which a measurement has a 95% chance of being any value other than zero. There is no universal procedure for determining a limit of detection for a given method, but in most cases the mean blank response (that is, the mean response produced by blank samples), the standard deviations of the blank response and some confidence factor, are used (MacDougall and Crummett, 1980; McNaught and Wilkinson, 2000). It should, however, be noted that values calculated using this procedure are statically not different from zero hence are recorded or reported as being less than the limit of detection or "non-detect."

Although statistically different from zero, observations that are close to the limit of detection will lack accuracy and precision (that is, they are less reliable) compared to values that are larger than the limit of detection. In most laboratories they define the smallest amount that can be considered as reliably quantifiable. This is known as the limit of quantification (LOQ). The limit of quantification is a multiple of the limit of detection. Values that are greater than the limit of detection but less than the limit of quantification are less reliable compared to those above the limit of quantification (Bergstrand and Karlsson, 2009).

The most common approaches in dealing with these errors are: (1) substitution, (2) imputation, and (3) maximum likelihood estimation. The most common and easiest of these is substitution, where censored values are replaced with zero, with some fraction of the detection limit (usually either 1/2 or $1/\sqrt{2}$), or with the detection limit itself (Hatvani et al., 2014). This is not a very accurate approach and various researchers have called it fabrication (Weiss and Indurkhya, 2000; Kotsiantis et al., 2006).

Imputation approaches use regression or probability plotting techniques to calculate the mean and standard deviation, based on the distribution of the values above the limit of detection. This means that the missing or censored values are replaced by an estimate that is calculated from a regression line equation of the observed values and their rank scores (Baraldi and Enders, 2010).

A third strategy, maximum likelihood estimation (MLE), is also a probabilistic method. It assumes that the observed sample follows a Gaussian distribution and goes on to estimate the mean and standard deviation of the sample. The MLE achieves a function by establishing parametric equations that make the observed data most probable (White, 1982).

In this study a program was developed using Visual Basics for Applications and Microsoft Excel. The program was then used to pre-process raw water quality data from the Umzinto Water Treatment Plant.

1. Study area

The data used in this study were provided by the Umzinto Water Treatment Plant. This treatment plant is owned by the Ugu District Municipality and managed by the Umgeni Water Board, which is amongst the leading water utilities in South Africa. In a case study by Ramjatan et al. (2007), the water utility was benchmarked against other utilities in Africa. The benchmarking initiatives employed were Deloitte and Touché Best Company to Work for Survey, the South African Association of Water Utilities (SAAWU), and the Water Utility Partnership (WUP) survey, amongst others. The results showed that the utility was among the leading African utilities.

Umzinto Water Treatment Plant is located in Umdoni Local Municipality, in Ugu District Municipality of South Africa (Figure 1). It has a population density of more than 260 persons per km^2. The population of the region has been increasing at a rate of 2.8% per year since 1996 and in 2007 it surpassed 70000 (City Population, 2012).

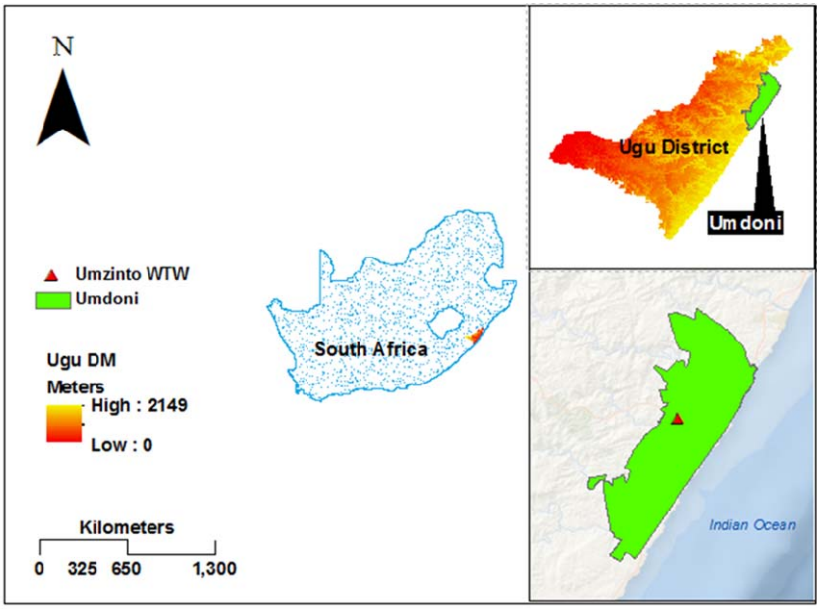

Fig. 1. Location of the Umdoni area within the Ugu District

Figure 2 shows the location of the Umzinto Water Treatment Plant relative to its raw water sources. The raw water sources are the EJ Smith Dam in the Mzimayi River and the Esperanza weir in the Umzinto River. The raw water abstraction from the EJ Smith Dam has a capacity of 5.0 Ml/d (1.8 million m^3/a) and usually is of poor quality with the main contaminants being Manganese, Iron and *Escherichia coli*. The capacity from the Esperanza weir is 10.0 Ml/d (3.6 million m^3/a) its quality is better compared to that of EJ Smith (Ugu District Municipality, 2011).

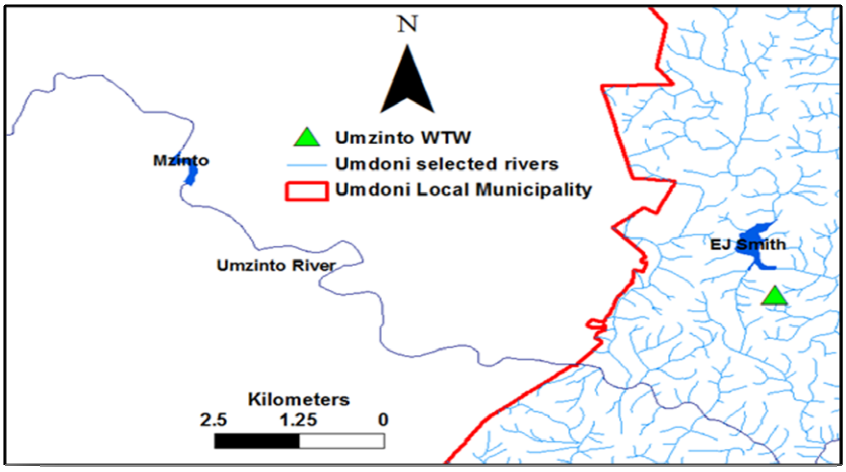

Fig. 2. Location of the Umzinto Water Treatment Plant and its raw water sources

2. Methods

The quality data collected were for the Algalcount, Manganese (Mn), Iron (Fe), *Escherichia coli* (*E.coli)*, Coliforms, Color, Conductivity, Turbidity, Suspended Solids (SS), pH, Temperature, Total Organic Carbon (TOC), Alkalinity, andTotal Hardness. With the collected data, a four stage approach is taken to preprocess it. The stages are as shown in Figure 3.

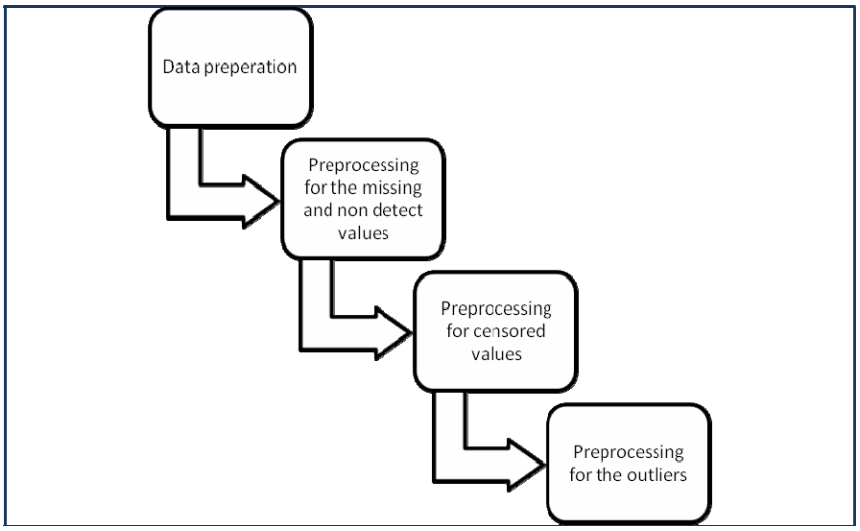

Fig. 3. Methodology to be used in pre-processing and developing a data pre-processing program using Visual Basic for Applications

2.1. Data preparation. The collected raw water quality data were in an excel format but the layout made it complex to use for analytical purposes. To make it usable the first step in pre-processing it required the use of a conditional database. In this study the researcher used the Microsoft SQL Server Management Studio 2012. An empty table was created in the management studio and then data from an excel file was imported into the table. The import was conducted using the Microsoft SQL import wizard that supports imports of excel data sheets into Microsoft SQL tables. In the SQL database the data would still have the same layout but the management studio enabled the extraction to be done in the preferred format.

The data were extracted from the conditional database into comma separated value files. Each comma separated value file contained data for a single parameter. The script which was used for the extraction is as follows:

```
SELECT * FROM [dbo].[waterparam] where Determinand = pH;
OUTPUT TO 'c:\\test\\pH.csv'
   FORMAT TEXT
   QUOTE ""
   WITH COLUMN NAMES;
SELECT * FROM [dbo].[waterparam] where Determinand = conductivity;
OUTPUT TO 'c:\\test\\conductivity.csv'
   FORMAT TEXT
   QUOTE ""
   WITH COLUMN NAMES;
SELECT * FROM [dbo].[waterparam] where Determinand = nitrates;
OUTPUT TO 'c:\\test\\nitrates.csv'
   FORMAT TEXT
   QUOTE ""
   WITH COLUMN NAMES;
SELECT * FROM [dbo].[waterparam] where Determinand = Total phosphate;
OUTPUT TO 'c:\\test\\Total phosphate.csv'
   FORMAT TEXT
   QUOTE ""
   WITH COLUMN NAMES;
SELECT * FROM [dbo].[waterparam] where Determinand = temperature;
OUTPUT TO 'c:\\test\\temperature.csv'
   FORMAT TEXT
   QUOTE ""
   WITH COLUMN NAMES;
SELECT * FROM [dbo].[waterparam] where Determinand = turbidity;
OUTPUT TO 'c:\\test\\turbidity.csv'
   FORMAT TEXT
   QUOTE ""
```

```
   WITH COLUMN NAMES;
SELECT * FROM [dbo].[waterparam] where Determinand = algae;
OUTPUT TO 'c:\\test\\algae.csv'
   FORMAT TEXT
   QUOTE ""
   WITH COLUMN NAMES;
SELECT * FROM [dbo].[waterparam] where Determinand = Ecoli;
OUTPUT TO 'c:\\test\\Ecoli.csv'
   FORMAT TEXT
   QUOTE ""
   WITH COLUMN NAMES;
SELECT * FROM [dbo].[waterparam] where Determinand = DO;
OUTPUT TO 'c:\\test\\DO.csv'
   FORMAT TEXT
   QUOTE ""
   WITH COLUMN NAMES;
SELECT * FROM [dbo].[waterparam] where Determinand = NH3;
OUTPUT TO 'c:\\test\\NH3.csv'
   FORMAT TEXT
   QUOTE ""
   WITH COLUMN NAMES;
```

Using visual basics for applications, the extracted comma separated value files were combined into 1 spreadsheet. This was done by running a code that identified the location of the folder containing these comma separated value files.

```
UserForm1-1
PrivateSubCommandBut-
ton2_Click()DimkwairiAsString
Dimkwairi2AsString
WithApplica-
tion.FileDialog(msoFileDialogFilePicker)
.Title = "Selectrequiredfiles"
.AllowMultiSelect = False
.InitialFileName = "C:"
.Filters.Clear
.Filters.Add"ExcellDocuments","*.csv",1If.ShowThen
SelectedFile =.SelectedItems(1)kwairi=SelectedFile
kwairi2=InStrRev(kwairi,"\")Me.TextBox1=Left(kwai
ri,kwairi2)
ElseEndIf
EndWith
EndSub
PrivateSubCommandBut-
ton1_Click()DimmekuisaAsString
DimMasterfileAsString
Masterfile=TextBox1.Value
DoUntilMasterfile<>""
If-
Trim(Masterfile)=""ThenMsgBox"PleaseFindFilePa
th"
ExitSub
EndIfLoop
DoUntilTextBox2.Value<>""
If-
Trim(TextBox2.Value)=""ThenMsgBox"Specifyfileto
mergeto"
ExitSub
EndIf
Loop
mekuisa=TextBox2.Value
Path=Masterfile
Filename=Dir(Path&"*.csv")
DoWhileFilename<>""
Work-
books.OpenFilename:=Path&Filename,ReadOnly:=
True
ForEachSheetInActiveWorkbook.Sheets
Sheet.CopyAfter:=Workbooks(mekuisa).Sheets(1)
NextSheet
Workbooks(Filename).CloseFilename=Dir()
Loop
MsgBox"WorkbooksSuccessfullyMerged"
UserForm1.Hide
EndSub
PrivateSubCommandBut-
ton3_Click()TextBox1.Value=""
TextBox2.Value=""EndSub
```

2.2. Pre-processing for the missing and non-detect values. For missing values the simple replacement technique is used. In this technique if a value is missing, the blank cell is filled with the average of filled values for the same month for the period between 1995 and 2013. An alternative is to

choose one definite value to fill all the blank cells but this is not recommended statistically. The code for filling missing values and non-detects with one definite value is as follows:

```
Sub Missing Data()
'Replace all blanks in selection
'with a constant value specified in InputBox
Dim cell As Range
Dim InputValue As String
On Error Resume Next
InputValue = InputBox ("Enter value that will fill empty cells in selection", _ “Replace Missing Values”)
'Test for empty cell. If empty, fill cell with value given
For Each cell In Selection
If IsEmpty(cell) Then
cell.Value = InputValue
End If
Next
End Sub
```

2.3. Pre-processing for censored values. In pre-processing the censored values, the simple replacement technique was used. The censored values were replaced by a value equal to the cut off value divided by the square root of two. The replacement would depend on the determinand and the type of censoring being associated with the values. An extract of the code used to achieve this is as shown below:

```
Private Sub UserForm_Initialize()
'Empty TextBox
TextBox1.Value = ""
'Empty CensoringListBox
CensoringListBox.Clear
'Fill CensoringListBox
With CensoringListBox
    .AddItem "<"
    .AddItem ">"
  End With
'Empty DeterminandListBox
DeterminandListBox.Clear
'Fill DeterminandListBox
With DeterminandListBox
    .AddItem "pH"
    .AddItem "electrical conductivity (EC) "
    .AddItem "nitrates (NO3,)"
    .AddItem "Total phosphate (TP)"
    .AddItem "temperature"
    .AddItem "turbidity"
    .AddItem "total algae count"
    .AddItem "Escherichia. Coli (E. coli)"
    .AddItem "Dissolved Oxygen (DO)"
    .AddItem "ammonia (NH3)"
End With
End Sub
```

The method applied is applicable to pre-process datasets with the same determinands as those from the Umzinto water treatment data.

2.4. Pre-processing for outliers. In dealing with the outliers, it was important to understand the pollution problems associated with the study area. For example, in the rainy season there is high runoff which can result in some parameters like turbidity increasing. A direct discharge of the sewage effluent into raw water sources can also result in high values of *E.coli*. To identify the outliers a graphical method was used. This involved the plotting of box and whisker diagrams in Microsoft Excel 2013 (Fu and Wang, 2012; Mozejko, 2012). For the box and whisker plots in excel one needs to determine the minimum (min), first quartile (q1), median, third quartile (q3), and maximum (max). These statistical parameters are illustrated in Figure 4 and Table 1 shows the excel functions used to obtain them.

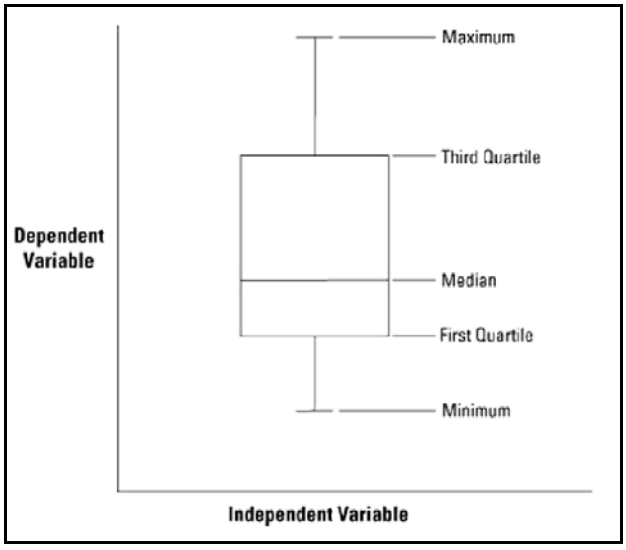

Fig. 4. Statistical parameters on a box and whisker plot

Table 1. Excel functions used to determine parameters for the box and whisker plots

Statistical parameter	Excel function
Minimum	Min (Range)
First quartile	Quartile (Range,1)
Median	Median (Range)
Third quartile	Quartile (Range,3)
Maximum	Max

3. Results

The main output generated in this study was a Visual Basics for Application program, which was then loaded into Microsoft Excel 2013 as an Excel Add-In. The graphic user interface for the Add-In is shown in Figure 5.

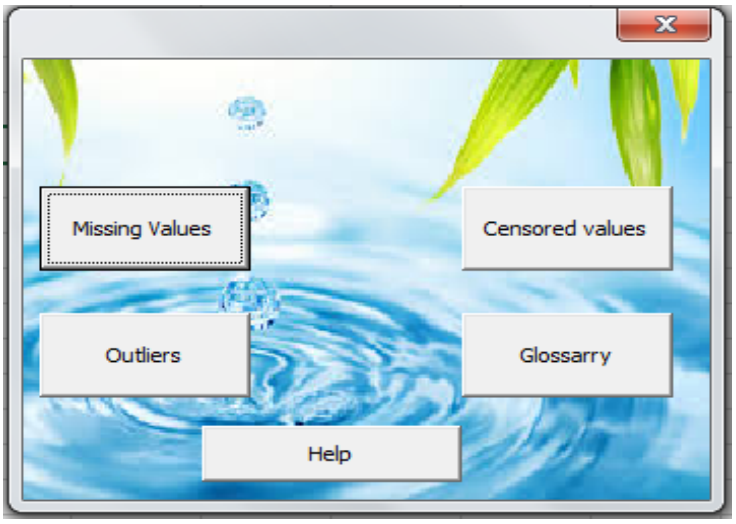

Fig. 5. Graphic user interface for the data pre-processing add-in

The Add-In is started by a click of a button and the graphic user interface shown in Figure 5 appears. This offers the user with four data pre-processing options and the help button evokes information on how to use the add-in.

3.1. Missing and non-detect values. The missing and non-detect values can be fixed using the Missing Values button on the graphic user interface: The Add-in will allow selection of either the imputation method or the simple replacement technique. If the latter is chosen, the program allows selection of the data range to be pre-processed and then after selection the screen shown in Figure 6 appears.

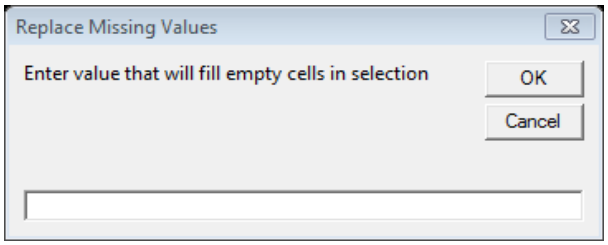

Fig. 6. Input box for replacing empty cells with one constant value

On the screen shown in Figure 6 the replacement value is entered and upon pressing the ok button the replacement value fills all the blank cells in the selected range.

3.2. Censored values. Hitting the censored values button will allow for the pre-processing for censored data. The determinand being corrected will have to be selected as shown in Figure 7.

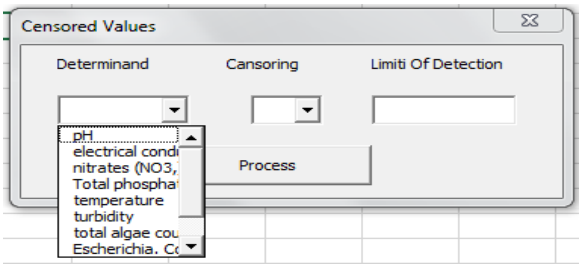

Fig. 7. Selection of determinand during censored value pre-processing

Selection of the censoring type allows the program to identify whether the data is right or left censored as shown in Figure 8.

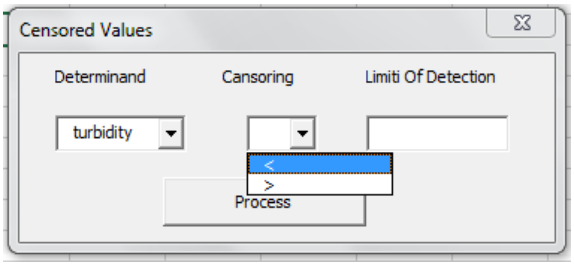

Fig. 8. Censoring type selection

3.3. Outliers. The outliers were determined by pressing the outlier button on the graphic user interface. This loads a Microsoft Excel user interface where pre-processed datasets will have been analyzed and for each determinand. The analysis is presented in the form of a statistical table and box and whisker plots. For the data used in this study the statistical parameters shown in Table 2a and 2b were produced.

Table 2a. Statistical parameters used to determine outliers

Statistic	Algal count (cells/mL)		Coliforms (CFU/100mL)		Conductivity (mS/m)		TOC (mg C/L)		*E.coli* (CFU/100mL)	
	UMZ	EJS	UMZ	EJS	UMZ	EJS	UMZ	EJS	UMZ	EJS
q1	34.4	551.0	2326.0	1583.3	29.1	24.7	6.3	4.1	27.8	31.3
min	0.0	0.0	252.7	128.3	21.1	8.2	5.0	2.5	5.0	2.0
median	81.3	3324.9	3774.5	2793.0	31.5	27.7	7.2	5.2	63.8	121.5
max	478.0	54593.9	22025.0	49230.7	44.1	48.4	15.6	13.0	1071.0	3026.7
q3	163.5	6877.9	4838.0	4843.3	33.4	29.7	8.4	5.7	125.8	527.3
mean	114.3	6368.5	4277.2	4761.6	31.2	27.7	7.5	5.1	116.7	448.8

Table 2b. Statistical parameters used to determine outliers

Statistic	Hardness (CFU/100mL)		pH		SS (mg/L)		Temperature (°C)		Turbidity (NTU)	
	UMZ	EJS	UMZ	EJS	UMZ	EJS	UMZ	EJS	UMZ	EJS
q1	60.0	57.7	7.6	7.5	3.0	7.4	17.5	18.2	3.2	11.3
min	45.7	33.7	7.1	7.2	3.0	4.0	13.1	14.3	0.7	3.2
median	65.1	61.7	7.7	7.6	5.5	12.3	20.9	21.2	6.1	25.9
max	194.9	126.8	8.2	8.3	67.4	218.8	28.0	27.8	207.7	1694.0
q3	69.0	65.7	7.8	7.8	11.6	22.5	23.2	23.9	13.9	62.4
mean	67.0	64.4	7.7	7.7	10.4	25.5	20.2	21.0	13.5	70.3

The Visual Basics for Application program then used the parameters in Table 2a and 2b to create the box and whisker plots shown from Figure 9 up to Figure 17. It should be noted that these plots were done before any changes had been done to the outliers and the datasets for the two abstraction points are represented.

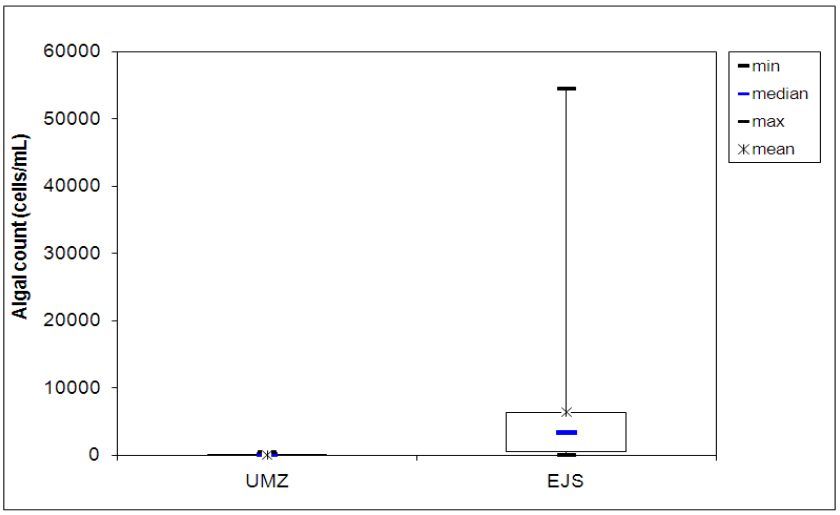

Fig. 9. Box and whisker plot for algal count

The box and whisker plot, Figure 9, shows that there is a great difference between the levels of algal count from the two sources of raw water. For the raw water from the Umzinto River (UMZ), the algal count ranges from 0 to 478 cells/mL. The levels from the EJ Smith Dam's (EJS) raw water are very high ranging from 0 to 54593.9 cells/mL. The difference between the two algal count levels should be as a result of the activities close to the abstraction points. The high monthly mean value of algal count observed in the raw water from Ej Smith Dam suggests there is need to use more coagulants and disinfectants during treatment. A concern is that when chlorine is used as a disinfectant, it can react with the algae to produce disinfection by products. Another concern is that the reaction tends to increase the amount of chlorine being used and the treatment cost. Figure 10 shows the box and whisker plot for the *Escherichia coli* at the two abstraction points of the Umzinto Water Treatment Plant.

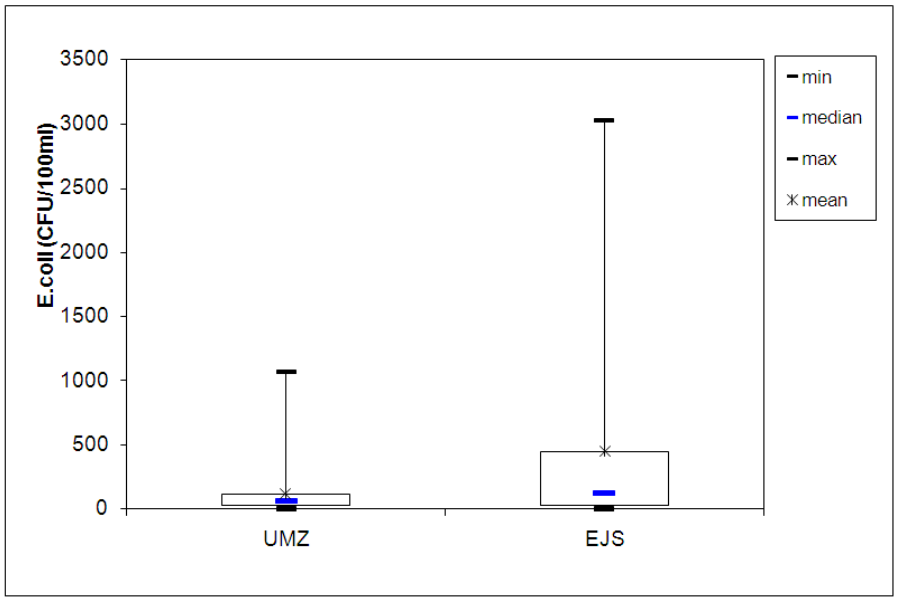

Fig. 10. Box and whisker plot for *E.coli*

Figure 10 shows that the *E.coli* can reach levels as high as 3000 CFU/100 mL. South African guidelines stipulate that in drinking water there should be no detection of E.coli per 100 ml of water. This means the *E.coli* count range for both abstraction points makes the water unsafe for drinking without proper disinfection. But particular attention needs to be given to the activities around the Mzimayi River which supplies the EJ Smith Dam. Figure 11 shows the box and whisker plot for the turbidity at the two abstraction points of the Umzinto River.

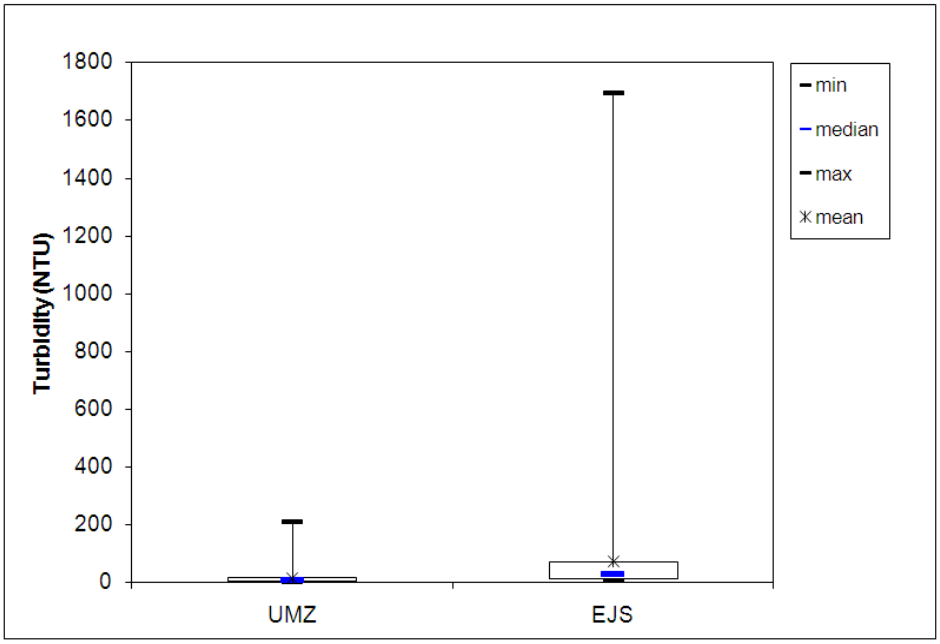

Fig. 11. Box and whisker plot for turbidity

The plot as expected shows that Turbidity is a problem mainly in the raw water from the EJ Smith Dam. The stipulated limit for drinking water turbidity is 0 to 1 NTU according to the South African standards. Reaching values as high as 1600 NTU there is need to monitor the effectiveness of disinfection. With high turbidity levels there is also an increase in the chances of disinfection products creation when using chlorine. Figure 12 shows the box and whisker plot for the suspended solid levels at the two abstraction points of the Umzinto Water Treatment Plant.

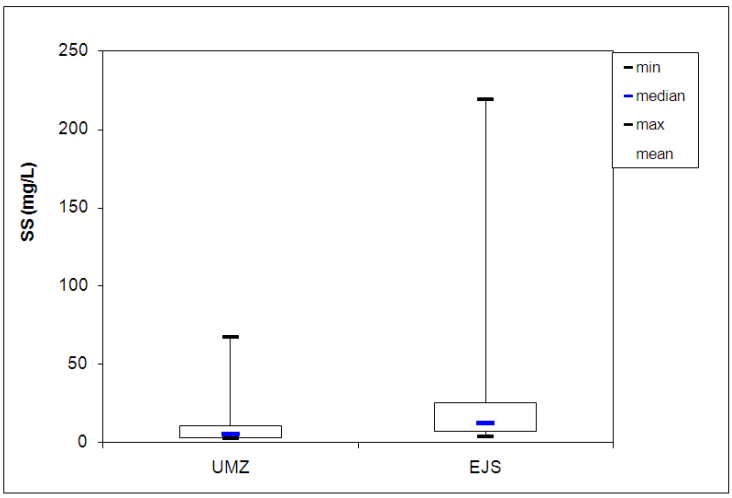

Fig. 12. Box and whisker plot for suspended solids

The box and whisker plot shows that range for Suspended Solids for the water from the EJ Smith Dam is quite wider than that from the Umzinto River. This is similar to the differences seen in the *E.coli* levels and may be due to the activities along the Mzimayi River that supply the EJ Smith Dam. Figure 13 shows the box and whisker plot for the pH level at the two abstraction points of the Umzinto Water Treatment Plant.

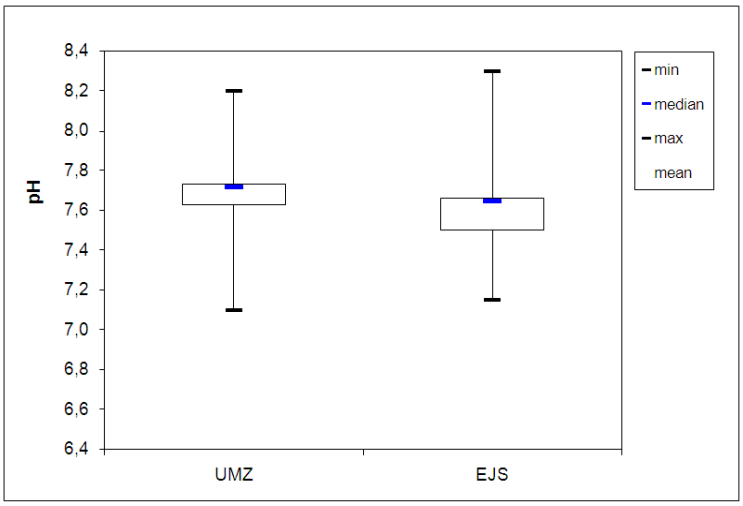

Fig. 13. Box and whisker plot for pH

As expected for raw water from South African Rivers, the two abstraction points show that the raw water has a high pH. The pH levels range from 7 to 8.4.

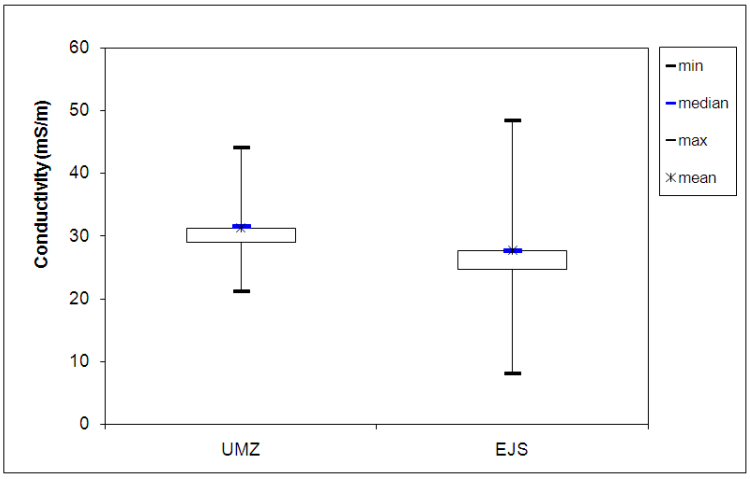

Fig. 14. Box and whisker plot for conductivity

The box and whisker plot, Figure 14, shows that the EJ Smith Dam's conductivity had a much wider range compared to that of Umzinto River. This can be due to outliers in the data sets but overall the box shows that electrical conductivity for the Ej Smith Dam's raw water was lower for the longer parts of the observations. Figure 15 shows the box and whisker plot for raw water temperature at the two abstraction points of the Umzinto Water Treatment Plant.

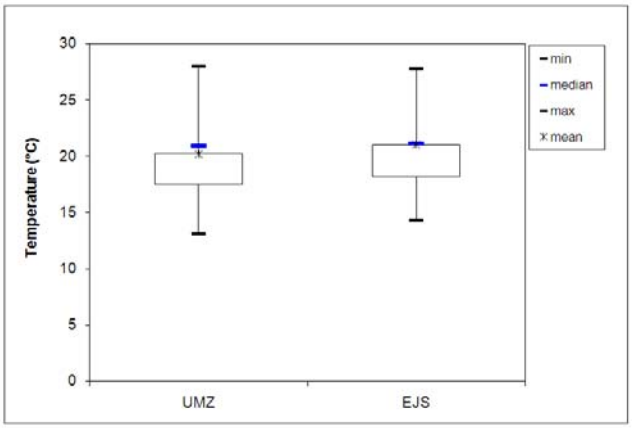

Fig. 15. Box and whisker plot for temperature

The abstraction points are within the same region and their raw water temperature as shown by the box and whisker is in the same region with almost similar mean and median values. Figure 16 shows the box and whisker plot for the total hardness at the two abstraction points of the Umzinto Water Treatment Plant.

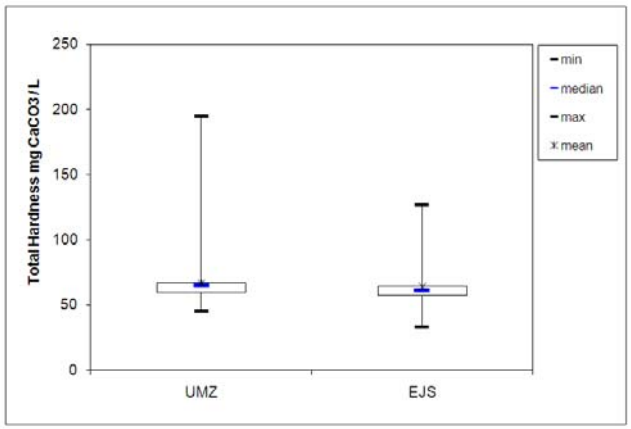

Fig. 16. Box and whisker plot for hardness

The box and whisker plot shows that the mean and median values for the two abstraction points are almost similar. The wider range for the Umzinto river water might be due to irregular peaks that can occur along its flow.

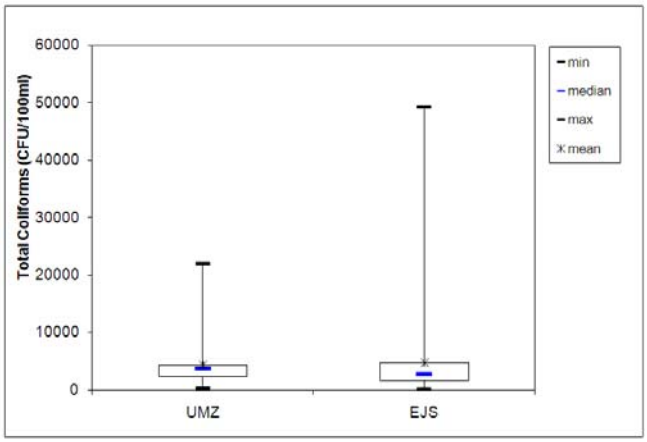

Fig. 17. Box and whisker plot coliforms

According to the South African standards, the permissible limit for coliforms is 10 CFU/100 mL. This shows that water from both sources is unsafe for drinking unless it has under gone treatment. The box plots show that the EJ Smith Dam's raw water quality is worse as it has a wider range and a higher maximum.

3.4. Pre-processed datasets. For each dataset, the mean and standard deviation were calculated without considering the missing data and the censored values. The statistical properties of the pre-processed datasets are shown in Table 3a and Table 3b.

Table 3a. Statistical properties before and after pre-processing the datasets

	Algal count (cells/mL)		Coliforms (CFU/100mL)		Conductivity (mS/m)		*E.coli* (CFU/100mL)		Hardness (mg CaCO3/L)	
	UMZ	EJS	UMZ	EJS	UMZ	EJS	UMZ	EJS	UMZ	EJS
Mean	114.3	6368.5	4277.2	4761.6	31.2	27.7	116.7	448.8	67.0	64.4
Median	81.3	3324.9	3774.5	2793.0	31.5	27.7	63.8	121.5	65.1	61.7
Std. deviation	107.1	9927.6	3391.8	6920.9	3.6	5.7	167.8	704.6	16.2	14.9
Range	478.0	54593.9	21772.3	49102.3	23.0	40.3	1066.0	3024.7	149.2	93.2
Minimum	0.0	0.0	252.7	128.3	21.1	8.2	5.0	2.0	45.7	33.7
Maximum	478.0	54593.9	22025.0	49230.7	44.1	48.4	1071.0	3026.7	194.9	126.8

Table 3b. Statistical properties before and after pre-processing the datasets

	pH		SS (mg/L)		Temperature (°C)		TOC (mg C/L)		Turbidity (NTU)	
	UMZ	EJS	UMZ	EJS	UMZ	EJS	UMZ	EJS	UMZ	EJS
Mean	7.7	7.7	10.4	25.5	20.2	21.0	7.5	5.1	13.5	70.3
Median	7.7	7.6	5.5	12.3	20.9	21.2	7.2	5.2	6.1	25.9
Std. Deviation	0.2	0.2	12.6	39.1	3.6	3.5	1.8	1.6	24.6	194.0
Range	1.1	1.2	64.4	214.8	14.9	13.5	10.6	10.5	207.0	1690.9
Minimum	7.1	7.2	3.0	4.0	13.1	14.3	5.0	2.5	0.7	3.2
Maximum	8.2	8.3	67.4	218.8	28.0	27.8	15.6	13.0	207.7	1694.0

Conclusions

A data pre-processing program was developed. The program provided a graphic user interface, which gave the user access to data pre-processing options. Most functions that would take a long time to do in excel can then be completed within a click of a button. Since most of the data manipulation is handled by the program this eliminates the risk of errors that can be established when data are pre-processed manually.

The processed data showed that the method that the developed program uses to pre-process data is very effective as it produced distributions with almost the same statistical parameters for most of the parameters. Algal count showed a relatively huge difference between the statistical parameters of pre-processed data and original data. The reason for this huge difference is not readily apparent. Outliers could be identified and the box plot produced good graphical representation of the datasets.

Acknowledgments

Financial assistance from South Africa'sDepartment of Science Technology (DST) forthe supervisor's National Research Foundation Post-Doctoral Fellowship, is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the DST. Durban University of Technology (DUT) is acknowledged for hosting and co-funding the Masters research and the supervisor's Post-Doctoral Fellowship. Umgeni Water and Ugu District Municipality are acknowledged for collaborating with DUT and for providing data during this research.

References

1. Baraldi, A.N. & Enders, C.K. (2010). An introduction to modern missing data analyses, *Journal of School Psychology*, 48, pp. 5-37.
2. Bergstrand, M. & Karlsson, M.O. (2009). Handling data below the limit of quantification in mixed effect models. *The AAPS journal*, 11, pp. 371-380.
3. Breese, S. (2006). Optimizing Conventional Water Treatment Plants. *AWWOA Annual Serminar*. Banff, Alberta.
4. Brown, M.L. & Kros, J.F. (2003). Data mining and the impact of missing data. *Industrial Management & Data Systems*, 103, pp. 611-621.
5. Brundtland, G.H. (1987). United Nations development and international economic co-operation: environment: our common future: the Brundtland report. Oslo: General Assembly.
6. Chang, J.-X., Bai, T., Huang, Q. & Yang, D.-W. (2013). Optimization of Water Resources Utilization by PSO-GA. *Water Resources Management*, 27, pp. 3525-3540.

7. Dearmont, D., Mccarl, B.A. & Tolman, D.A. (1998). Costs of water treatment due to diminished water quality: a case study in Texas, *Water Resources Research*, 34, pp. 849-854.
8. DWAF (1996). South African Water Quality Guidelines (second edition). Volume 1: Domestic Use. In: Forestry, D.O.W.A.A. (ed.). Pretoria: Department of Water Affairs and Forestry.
9. Dzwairo, B. (2011). *Modelling raw water quality variability in order to predict cost of water treatment.* Doctor Technologiae, Tshwane University of Technology.
10. Gertsen, N. & Sønderby, L. (2009). *Air, Water and Soil Pollution Science and Technology: Water Purification,* New York, NY, USA, Nova Science Publishers, Inc.
11. Grubbs, F.E. (1969). Procedures for detecting outlying observations in samples, *Technometrics*, 11, pp. 1-21.
12. Hatvani, I.G., Magyar, N., Zessner, M., Kovács, J. & Blaschke, A.P. (2014). The Water Framework Directive: Can more information be extracted from groundwater data? A case study of Seewinkel, Burgenland, eastern Austria, *Hydrogeology Journal*, 22, pp. 779-794.
13. Hodge, V.J. & Austin, J. (2004). A survey of outlier detection methodologies, *Artificial Intelligence Review*, 22, pp. 85-126.
14. Kotsiantis, S., Kanellopoulos, D. & Pintelas, P. (2006). Data preprocessing for supervised leaning, *International Journal of Computer Science*, 1, pp. 111-117.
15. Lange, G.M. & Hassan, R. (2006). *The Economics of Water Management in Southern Africa: An Environmental Accounting Approach*, Cheltenham, UK, Edward Elgar Publishing Ltd.
16. Loucks, D.P., Van Beek, E., Stedinger, J.R., Dijkman, J.P. & Villars, M.T. (2005). *Water resources systems planning and management: an introduction to methods, models and applications*, Paris, UNESCO.
17. Macdougall, D. & Crummett, W.B. (1980). Guidelines for data acquisition and data quality evaluation in environmental chemistry, *Analytical Chemistry*, 52, pp. 2242-2249.
18. Mcnaught, A.D. & Wilkinson, A. (2000). IUPAC Compendium of chemical terminology. International Union of Pure and Applied Chemistry.
19. Mosley, L.M., Zammit, B., Leyden, E., Heneker, T.M., Hipsey, M.R., Skinner, D. & Aldridge, K.T. (2012). The impact of extreme low flows on the water quality of the Lower Murray River and Lakes (South Australia), *Water Resources Management*, 26, pp. 3923-3946.
20. Municipality, U.D. (2011). First Stage Reconciliation Strategy For Umzinto Water Supply Scheme Area – Umdoni Local Municipality. In: Affairs, D.O.W. (ed.). South Africa.
21. Netshidaulu, A.E. (2007). Personal Communication. Bothaville: Sedibeng Water.
22. O'reilly, G. & Bezuidenhout, C. (2013). *Assessment of the Physico-chemical and Microbiological Quality of Household Water in the Vaalharts Irrigation Scheme, South Africa.* North-West University, Potchefstroom Campus.
23. Peng, C. (2010). Interval estimation of population parameters based on environmental data with detection limits, *Environmetrics*, 21, pp. 645-658.
24. Population, C. (2012). *City Population* [Online]. Available at: http://www.citypopulation.de/php/southafrica-admin.php?adm2id=KZN212 [Accessed 8 August 2014].
25. Ramjatan, A., Dlamini, P., Tiba, F. & Pillay, M. (2007). Water Utilities and Benchmarking A Case Study of Umgeni Water. Pietermaritzburg: Umgeni Water.
26. Rousseeuw, P.J. & Leroy, A.M. (2005). *Robust regression and outlier detection*, John Wiley & Sons.
27. SABS (2011). SANS-241:2011. In: Standard, S.A.N. (ed.). Pretoria: SABS Standard Division.
28. Saeed, A. & Hashmi, I. (2014). Evaluation of anthropogenic effects on water quality and bacterial diversity in Rawal Lake, Islamabad, *Environmental Monitoring and Assessment*, 186, pp. 2785-2793.
29. Schutte, F.E. (2006). *Handbook for the Operation of Water Treatment Works* Gezina, Water Research Commission.
30. Shumway, R.H., Azari, R.S. & Kayhanian, M. (2002). Statistical approaches to estimating mean water quality concentrations with detection limits, *Environmental Science & Technology*, 36, pp. 3345-3353.
31. Taylor, J.K. (1987). *Quality assurance of chemical measurements*, CRC Press.
32. Weiss, S.M. & Indurkhya, N. (2000). *Decision-rule solutions for data mining with missing values*, Springer.
33. White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica: Journal of the Econometric Society*, pp. 1-25.
34. Zhu, X., Zhang, S., Jin, Z., Zhang, Z. & Xu, Z. (2011). Missing value estimation for mixed-attribute data sets. *Knowledge and Data Engineering, IEEE Transactions on*, 23, pp. 110-121.