We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter 4

Validity and Errors in Water Quality Data — A Review

Innocent Rangeti, Bloodless Dzwairo,
Graham J. Barratt and Fredrick A.O. Otieno

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/59059

1. Introduction

While it is essential for every researcher to obtain data that is highly accurate, complete, representative and comparable, it is known that missing values, outliers and censored values are common characteristics of a water quality data-set. Random and systematic errors at various stages of a monitoring program tend to produce erroneous values, which complicates statistical analysis. For example, the central tendency statistics, particularly the mean and standard deviation, are distorted by a single grossly inaccurate data point. An error, which is initially identified and is later incorporated into a decision making tool, like a water quality index (WQI) or a model, could subsequently lead to costly consequences to humans and the environment.

Checking for erroneous and anomalous data points should be routine, and an initial stage of any data analysis study. However, distinguishing between a data-point and an error requires experience. For example, outliers may actually be results which might require statistical attention before a decision can be made to either discard or retain them. Human judgement, based on knowledge, experience and intuition thus continue to be important in assessing the integrity and validity of a given data-set. It is therefore essential for water resources practitioners to be knowledgeable regarding the identification and treatment of errors and anomalies in water quality data before undertaking an in-depth analysis.

On the other hand, although the advent of computers and various software have made it easy to analyse large amounts of data, lack of basic statistical knowledge could result in the application of an inappropriate technique. This could ultimately lead to wrong conclusions that are costly to humans and the environment [1]. Such necessitate the need for some basic understanding of data characteristics and statistics methods that are commonly applied in the water quality sector. This chapter, discusses common anomalies and errors in water quality

data-sets, methods of their identification and treatment. Knowledge reviewed could assist with building appropriate and validated data-sets which might suit the statistical method under consideration for data analysis and/or modelling.

2. Data errors and anomalies

Referring to water quality studies, an error can be defined as a value that does not represent the true concentration of a variable such as turbidity. These may arise from both human and technical error during sample collection, preparation, analysis and recording of results [2]. Erroneous values can be recorded even where an organisation has a clearly defined monitoring protocol. If invalid values are subsequently combined with valid data, the integrity of the latter is also impaired [1]. Incorporating erroneous values into a management tool like a WQI or model, could result in wrong conclusions that might be costly to the environment or humans.

Data validation is a rigorous process of reviewing the quality of data. It assists in determining errors and anomalies that might need attention during analysis. Validation is crucial especially where a study depends on secondary data as it increases confidence in the integrity of the obtained data. Without such confidence, further data manipulation is fruitless [3]. Though data validation is usually performed by a quality control personnel in most organisations, it is important for any water resource practitioner to understand the common characteristics that may affect in-depth analysis of a water quality data-sets.

3. Visual scan

Among the common methods of assessing the integrity of a data-set is visual scan. This approach assists to identify values that are distinct and, which might require attention during statistical analysis and model building. The ability to visually assess the integrity of data depends on both the monitoring objectives and experience [4]. Transcription errors, erroneous values (e.g. a pH value of greater than 14, or a negative reading) and inaccurate sample information (e.g. units of mg/L for specific conductivity data) are common errors that can be easily noted by a visual scan. A major source of transcription errors is during data entry or when converting data from one format to another [5, 6]. This is common when data is transferred from a manually recorded spreadsheet to a computer oriented format. The incorrect positioning of a decimal point during data entry is also a common transcription error [7, 8].

A report by [7] suggested that transcription errors can be reduced by minimising the number of times that data is copied before a final report is compiled. [9] recommended the read-aloud technique as an effective way of reducing transcription errors. Data is printed and read-aloud by one individual, while the second individual simultaneously compares the spoken values with the ones on the original sheet. Even though the double data-entry method has been described as an effective method of reducing transcription errors, its main limitation is of being

laborious [9-11]. [12], however, recommended slow and careful entry of results as an effective approach of reducing transcription errors.

While it might be easy to detect some of the erroneous values by a general visual scan, more subtle errors, for example outliers, may only be ascertained by statistical methods [13]. Censored values, missing values, seasonality, serial correlation and outliers are common characteristics in data-sets that need identification and treatment [14]. The following sections review the common characteristics in water quality data namely; outliers, missing values and censored values. Methods of their identification and treatment are discussed.

3.1. Outliers (extreme values)

The presence of values that are far smaller or larger than the usual results is a common feature of water quality data. An outlier is defined as a value that has a low probability of originating from the same statistical distribution as the rest of observations in a data-set [15]. Outlying values should be examined to ascertain if they are possibly erroneous. If erroneous, the value can be discarded or corrected, where possible. Extreme values may arise from an imprecise measurement tool, sample contamination, incorrect laboratory analysis technique, mistakes made during data transfer, incorrect statistical distribution assumption or a novel phenomenon, [15, 16]. Since many ecological phenomena (e.g., floods, storms) are known to produce extreme values, their removal assumes that the phenomenon did not occur when actually it did. A decision must thus be made as to whether an outlying datum is an occasional value and an appropriate member of the data-set or whether it should be amended, or excluded from subsequent statistical analyses as it might introduce bias [1].

An outlying value should only be objectively rejected as erroneous after a statistical test indicates that it is not real or when it is desired to make the statistical testing more sensitive [17]. In figure 1, for example, simple inspection might mean that the two spikes are erroneous, but in-depth analysis might correlate the spikes to very poor water quality for those two days, which would make the two observations valid. The model, however, does not pick the extreme values, which negatively affects the R^2 value, and ultimately the accuracy and usefulness of the model in predicting polymer dosage.

Both observational (graphical) and statistical techniques have been applied to identify outliers. Among the common observational methods are the box-plots, time series, histogram, ranked data plots and normal probability plots [18, 19]. These methods basically detect an outlier value by quantifying how far it lies from the other values. This could be the difference between the outlier and the mean of all points, between the outlier and the next closest value or between the outlier and the mean of the remaining values [20].

3.2. Box-plot

The box-plot, a graphical representation of data dispersion, is considered to be a simple observation method for screening outliers. It has been recommended as a primary exploratory tool of identifying outlying values in large data-sets (15). Since the technique basically uses the

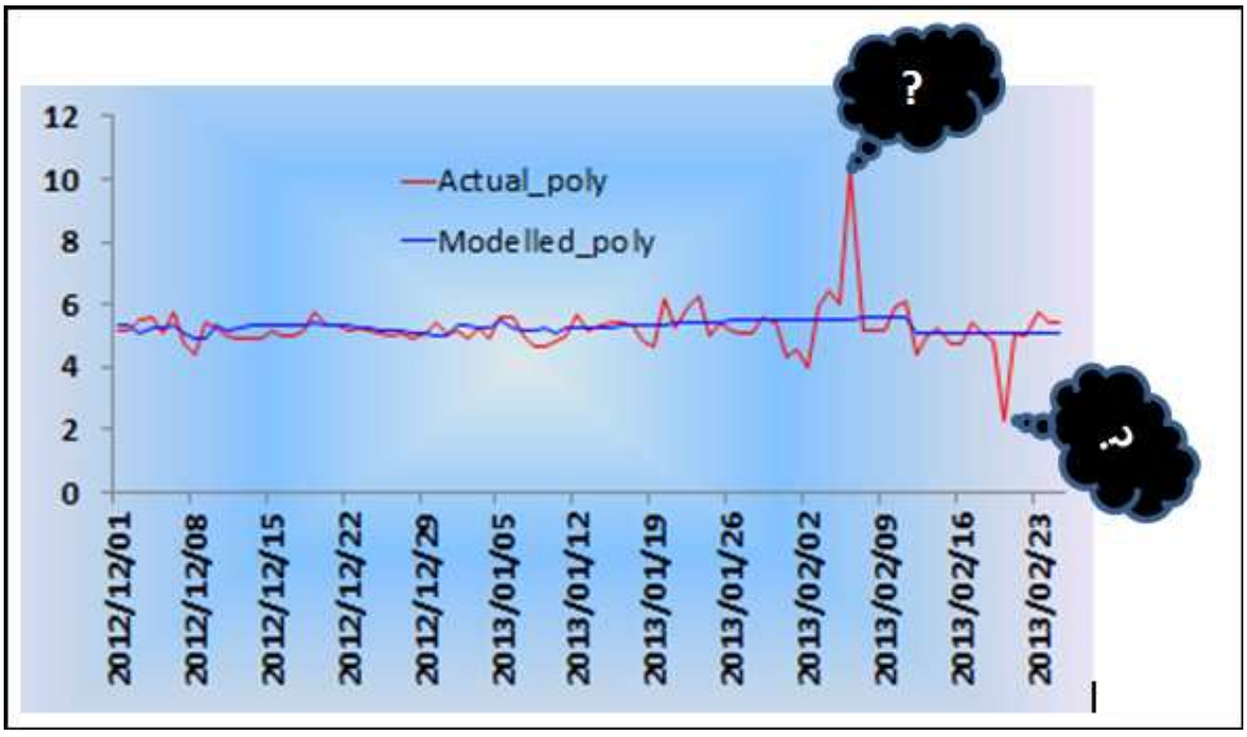

Figure 1. Data inspection during validation and treatment

median value and not the mean, it poses a greater advantage by allowing data analysis disregarding its distribution. [21] and [22] categorised potential outliers using the box-plot as:

- data points between 1.5 and 3 times the Inter Quantile Range (IQR) above the 75th percentile or between 1.5 and 3 times the IQR below the 25th percentile, and
- data points that exceed 3 times the IQR above the 75th percentile or exceed 3 times the IQR below the 25th percentile.

The limitation of a box plot is that it is basically a descriptive method that does not allow for hypothesis testing, and thus cannot determine the significance of a potential outlier [15].

3.3. Normal probability plot

The probability plot method identifies outliers as values that do not closely fit a normal distribution curve. The points located along the probability plot line represent 'normal',observation, while those at the upper or lower extreme of the line, indicates the suspected outliers as depicted in Figure 2.

The approach assumes that if an extreme value is removed, the resulting population becomes normally distributed [21]. If, however, the data still does not appear normally distributed after the removal of outlying values, a researcher might have to consider normalising it by transformation techniques, such as using logarithms [21, 23]. However, it should be highlighted that data transformation tends to shrink large values (see the two extreme values in Figure 1, before transformation), thus suppressing their effect which might be of interest for further

http://dx.doi.org/10.5772/59059

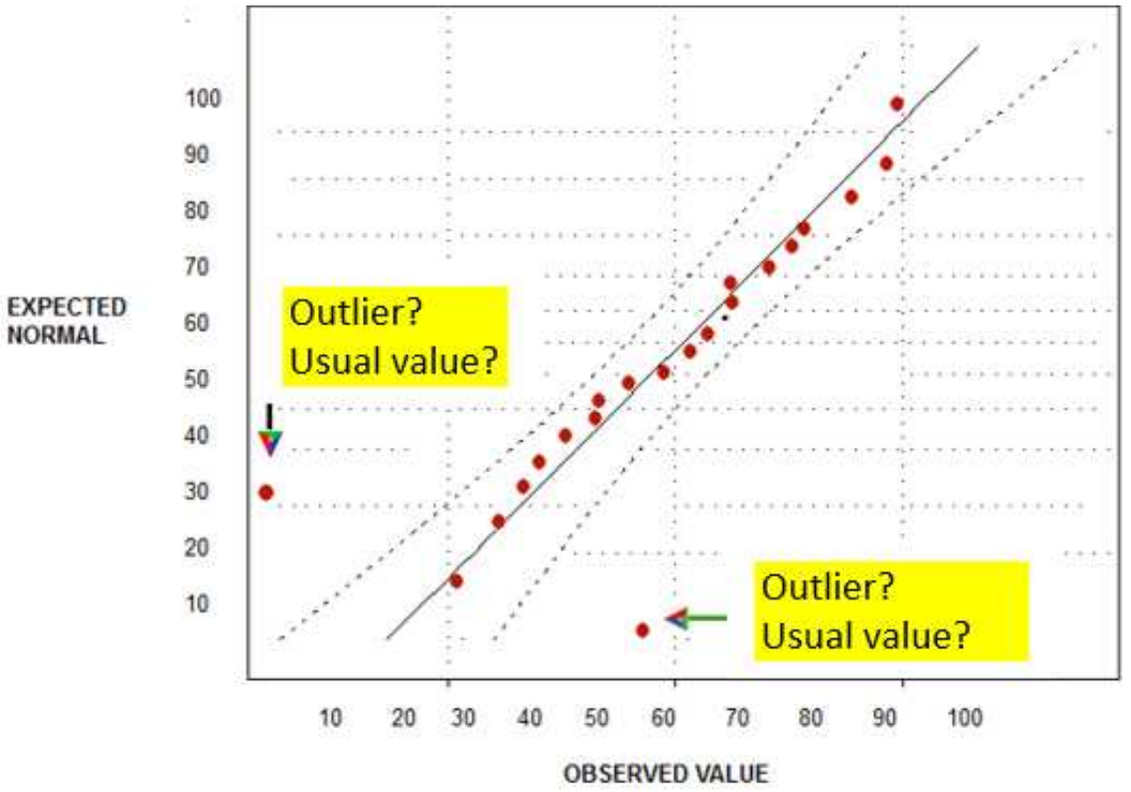

Figure 2. Normal probability plot showing outliers

analysis [23, 24]. Data should thus not be simply transformed for the sole purpose of eliminating or reducing the impact of outliers. Furthermore, since some data transformation techniques require non-negative values only (e.g. square root function) and a value greater than zero (e.g. logarithm function), transformation should not be considered as an automatic way of reducing the effect of outliers [23].

Since observational methods might fail to identify some of the subtle outliers, statistical tests may be performed to identify a data point as an outlier. However a decision still has to be made on whether to exclude or retain an outlying data point. The section below describes the common statistical test for identifying outliers.

3.4. Grubbs test

The Grubb's test, also known as the Studentised Deviate test, compares outlying data points with the average and standard deviation of a data-set [25-27]. Before applying the Grubbs test, one should firstly verify that the data can be reasonably approximated by a normal distribution. The test detects and removes one outlier at a time until all are removed. The test is two sided as shown in the two equations below.

1. To test whether the maximum value is an outlier, the test:

$$G_{max} = \frac{X_n - X_{mean}}{S}$$

2. To test whether the minimum value is an outlier, the test is:

$$G_{min} = \frac{X_{mean} - X_i}{S}$$

Where X_1 or X_n=the suspected single outlier (max or min)

s=standard deviation of the whole data set

X=mean

The main limitation of Grubbs test is of being invalid when data assumes non-normal distribution [28]. Multiple iterations of data also tends to change the probabilities of detection. Grubbs test is only recommended for sample sizes of not more than six, since it frequently tags most of the points as outliers. It suffers from masking, which is failure to identify more than one outlier in a data-set [28, 29]. For instance, for a data-set consisting of the following points; 3, 5, 7, 13, 15, 150, 153, the identification of 153 (maximum value) as an outlier might fail because it is not extreme with respect to the next highest value (150). However, it is clear that both values (150 and 153) are much higher than the rest of the data-set and could jointly be considered as outliers.

3.5. Dixon test

Dixon's test is considered an effective technique of identifying an outlier in a data-set containing not more than 25 values [21, 30]. It is based on the ratio of the ranges of a potential outlier to the range of the whole data set as shown in equation 1 [31]. The observations are arranged in ascending order and if the distance between the potential outlier to its nearest value (Q_{gap}) is large enough, relative to the range of all values (Q_{range}), the value is considered an outlier.

$$Q_{exp} = \frac{Q_{gap}}{Q_{range}} \quad (1)$$

The calculated Q_{exp} value is then compared to a critical Q-value (Qcrit) found in tables. If Q_{exp} is greater than the suspect value, the suspected value can be characterised as an outlier. Since the Dixon test is based on ordered statistics, it tends to counter-act the normality assumption [15]. The test assumes that if the suspected outlier is removed, the data becomes normally distributed. However, Dixon's test also suffers the masking effect when the population contains more than one outlier.

[32] recommended the use of multivariate techniques like Jackknife distance and Mahalanobis distance [33, 34]. The strength of multivariate methods is on their ability to incorporation of the correlation or covariance between variables thus making them more correct as compared to univariate methods. [34] introduced the chi-square plot, which draws the empirical distribution function of the robust Mahalanobis distances against the chi-square distribution. A value that is out of distribution tail indicates that it is an outlier [33].

For an on-going study, an outlier can be ascertained by re-analysis of the sample, if still available and valid. [28] and [2] advised the practise of triplicate sampling as an effective method of verifying the unexpected results. When conducting a long-term study, researchers might consider re-sampling when almost similar conditions prevail again. Nevertheless, this option might not be feasible when carrying out a retrospective study since it generally depend on secondary data from past events.

For data intended for trend analysis, studies have recommended the application of nonparametric techniques such as the Seasonal Kendal test where transformation techniques do not yield symmetric data [19]. Should a parametric test be preferred on a data-set that includes outliers, practitioners may evaluate the influence of outliers by performing the test twice, once using the full data-set (including the outliers) and again on the reduced data-set (excluding the outliers). If the conclusions are essentially the same, then the suspect datum may be retained, failing which a nonparametric test is recommended.

4. Missing values

While most statistical methods presumes a complete data-set for analysis, missing values are frequently encountered problems in water quality studies [35, 36]. Handling missing values can be a challenge as it requires a careful examination of the data to identify the type and pattern of missingness, and also have a clear understanding of the most appropriate imputation method. Gaps in water quality data-sets may arise due to several reasons, among which are imperfect data entry, equipment error, loss of sample before analysis and incorrect measurements [37]. Missing values complicate data analysis, cause loss of statistical efficiency and reduces statistical estimation power [37-39]. For data intended for time-series analysis and model building, gaps become a significant obstacle since both generally require continuous data [40, 41]. Any estimation of missing values should be done in a manner that minimise the introduction of more bias in order to preserve the structure of original data-set [41, 42].

The best way to estimate missing values is to repeat the experiment and produce a complete data-set. This option is however, not feasible when conducting a retrospective study since it depend on historical data. Where it is not possible to re-sample, a model or non-model techniques may be applied to estimate missing values [43].

If the proportion of missing values is relatively small, listwise deletion has been recommended. This approach, which is considered the easiest and simplest, discards the entire case where any of the variables are missing. Its major advantage is that it produce a complete data-set, which in turn allows for the use of standard analysis techniques [44]. The method also does not require special computational techniques. However, as the proportion of missing data increases, deletion tends to introduce biasness and inaccuracies in subsequent analyses. This tends to reduce the power of significance test and is more pronounced particularly if the pattern of missing data is not completely random. Furthermore, listwise deletion also decreases the sample size which tends to reduce the ability to detect a true association. For example, suppose a data-set with 1,000 samples and 20 variables and each of the variables has missing

data for 5% of the cases, then, one could expect to have complete data for only about 360 individuals, thus discarding the other 640.

On the other hand, pairwise deletion removes incomplete cases on an analysis-by-analysis basis, such that any given case may contribute to some analyses but not to others [44]. This approach is considered an improvement over listwise deletion because it minimises the number of cases discarded in any given analysis. However, it also tend to produce bias if the data is not completely random.

Several studies have applied imputation techniques to estimate missing values. A common assumption with these methods is that data should be missing randomly [45]. The most common and easiest imputation technique is replacing the missing values with an arithmetic mean for the rest of the data [35, 41]. This is recommended where the frequency distribution of a variable is reasonably symmetric, or has been made so by data transformation methods. The advantage of arithmetic mean imputation is generation of unbiased estimates if the data is completely random because the mean lands on the regression line. Even though the insertion of mean value does not add information, it tends to improve subsequent analysis. However, while simple to execute, this method does not take into consideration the subjects patterns of scores across all the other variables. It changes the distribution of the original data by narrowing the variance [46]. If the data assumes an asymmetric distribution, the median has been recommended as a more representative estimate of the central tendency and should be used instead of the mean.

[47], recommended model-based substitution techniques as more flexible and less ad hoc approach of estimating missing values as compared to non-model methods. A simple modelling technique is to regress the previous observations into an equation which estimates missing values [35, 48]. The time-series auto-regressive model has been described as an improvement and more accurate method of estimating missing values [25]. Unlike the arithmetic mean and median replacement methods, regression imputation techniques estimates missing values of a given variable using data of other parameters. This tends to reduce the variance problem, which is common with the arithmetic mean imputation and median replacement methods [41, 49].

On the other hand, the maximum likelihood technique uses all the available complete and incomplete data to identify the parameter values that have the highest probability of producing the sample data [44]. It runs a series of data iterations by replacing different values for the unknown parameters and converges to a single set of parameters with the highest probability of matching the observed data [41]. The method has been recommended as it tends to give efficient estimates with correct standard errors. However, just like other imputation methods, the maximum likelihood estimates can be heavily biased if the sample size is small. In addition, the technique requires a specialised software which may be expensive, challenging to use and time consuming.

Some studies have considered the relationship between parameters as an effective approach of estimating missing values [50]. For instance, missing conductivity values can be calculated

http://dx.doi.org/10.5772/59059

from the total dissolved solids value (TDS) by a simple linear regression where p-value and r-value are known to exist and the missing value lies between the two variables. Equation 2, where ***a*** is in the range 1.2-1.8, has been described as an equally important estimator of missing conductivity values [1, 51].

$$Conductivity \approx TDS \; x \; a \tag{2}$$

The constant, *a*, is high in water of high chloride and low sulphate [51]. [52] estimated missing potassium values by using a linear relationship between potassium and sodium. The relationship gave a high correlation coefficient of 0.904 ($p<0.001$).

As of late, research has explored the application of artificial intelligence (AI) techniques to handle missing values in the water quality sector. Among the major AI techniques that have been applied is the Artificial Neural Networks (ANN) and Hybrid Evolutionary Algorithms (HEA) (48, 53, 54). Nevertheless, it should also be highlighted that all techniques for estimating missing values invariably affect the results. This is more pronounced when missing values characterise a significant proportion of the data being analysed. A research should thus consider the sample size when choosing the most appropriate imputation method.

5. Scientific facts

The integrity of water quality data can also be assessed by checking whether the results are inline with known scientific facts. To ascertain that, a researcher must have some scientific knowledge regarding the characteristics of water quality variables. Below are some scientific facts that can be used to assess data integrity [1].

1. Presence of nitrate in the absence of dissolved oxygen may indicate an error since nitrate is rapidly reduced in the absence of oxygen. The dissolved oxygen meter might have malfunctioned or oxygen might have escaped from the sample before analysis.
2. Component parts of a water-quality variable must not be greater than the total variable. For example:
3. Total phosphorus ≥Total dissolved phosphorus>Ortho-phosphate.
4. Total Kjeldahl nitrogen ≥Total dissolved Kjeldahl nitrogen>ammonia.
5. Total organic carbon ≥Dissolved organic carbon.
6. Species in a water body should be described correctly with regards to original pH of the water sample. For example, carbonate species will normally exist as HCO_3^- while CO_3^{2-} cannot co-exist with H_2CO_3.

6. Censored data

A common problem faced by researchers analysing environmental data is the presence of observations reported to have non-detectable levels of a contaminant. Data which are either less than the lower detection limit, or greater than the upper detection limit of the analytical method applied are normally artificially curtailed at the end of a distribution, and are termed "censored values" [14]. Multiple censored results may be recorded when the laboratory has changed levels of detection, possibly as a result of an instrument having gained more accuracy, or the laboratory protocol having established new limits. If the values are below the detection limit, they are abbreviated as BDL, and when above the limit, as ADL [55, 56].

Various methods of treating censored values have been developed to reduce the complication generally brought about by censored values [57]. The application of an incorrect method may introduce bias especially when estimating the mean and variance of data distribution [58]. This may consequently distort the regression coefficients and their standard errors, and further reduce the hypothesis testing power. A researcher must thus decide on the most appropriate method to analyse censored values. One might reason that since these values are extra ordinarly small, they are not important and discard them while some might be tempted to remove them inorder to ease statistical analysis. Deletion has however been described as the worst practise as it tends to introduce a strong upward bias of the central tendency which lead to inaccurate interpretation of data [19, 59-62].

The relatively easiest and most common method of handling censored values is to replace them with a real number value so that they conform to the rest of data. The United State Environmental Agency suggested substitution if censored data is less than 15% of the total data-set (63, 64). [8], BDL, for example $x < 1.1$, were multiplied by the factor 0.75 to give 0.825. ADL values, for example $500 < x$, were recorded as one magnitude higher than the limit values to give 501. [65] recommended substituting with $\frac{1}{2}$ DL or $\frac{1}{\sqrt{2}}$ DL if the sample size is less than 20 and contains less than 45% of its data as censored values. [66] suggested substitution by $\frac{1}{\sqrt{2}}$DLif the data are not highly skewed and substitution by $\frac{1}{2}$ DL otherwise. [67], however, criticised the substitution approach and illustrated how the practice could produces poor estimates of correlation coefficients and regression slopes. [68] further explained that substitution is not suitable if the data has multiple detection limits [68, 69].

A second approach for handling censored values is the maximum likelihood estimation (MLE). It is recommended for a large data-set which assumes normality and contains censored results [38, 65, 70, 71]. This approach basically uses the statistical properties of non-censored portion of the data-set, and an iterative process to determine the means and variance. The MLE technique generates an equation that calculates mean and standard deviation from values assumed to represent both the detects and non-detect results [69]. The equation can be used to estimate values that can replace censored values. However, the technique is reportedly ineffective for a small data-set that has fewer than 50 BDLs [69].

When data assumes an independent distribution and contain censored values, non-parametric methods like the Kaplan-Meir method, can be considered for analysis [59]. The Kaplan-Meir method creates an estimate of the population mean and standard deviation, which is adjusted for data censoring, based on the fitted distribution model. Just like any non-parametric techniques for analysing censored data, the Kaplan-Meier is only applicable to right-censored results (i.e. greater than) [72]. To use Kaplan-Meier on left-censored values, the censored values must be converted to right-censored by flipping them over to the largest observed value [65, 71, 72]. To ease the process, [73] have developed a computer program that does the conversion. [71], however, found the Kaplan-Meier method to be effective when summarising a data-set containing up to 70% of censored results.

In between the parametric and non-parametric methods is a robust technique called Regression on Order Statistics (ROS) [38]. It treats BDLs based on the probability plot of detects. The technique is applicable where the response variable (concentration) is a linear function of the explanatory variable (the normal quartiles) and if the error variance of the model is constant. It also assumes that all censoring thresholds are left-censored and is effective for a data-set which contains up to 80% censored values [59]. The ROS technique uses data plots on a modelling distribution to predict censored values. [59] and [68] evaluated ROS as a reliable method for summarising multiply censored data. Helsel and Cohn (38) also described ROS as a better estimator of the mean and standard deviation as compared to MLE, when the sample size is less than 50 and contains censored values.

7. Statistical methods

The success of an analysis of water quality data primarily depends on the selection of the right statistical method which considers common data characteristics such as normality, seasonality, outliers, missing values, censoring, etc., [74]. If the data assumes an understandable and describable distribution, parametric methods can be used [14]. However, non-parametric techniques are slowly replacing parametric techniques mainly because the latter are sensitive to common water characteristics like outliers, missing values and censored value [75].

7.1. Computer application in data treatment

The increase in various computer programs has made it easy to detect and treat erroneous data. Computers now provide flexibility and speedy methods of data analysis, tabulation, graph preparation or running models, among others. Various software such as Microsoft Excel, Minitab, Stata and MATLAB have become indispensable tools for analysing environmental data. These software perform various computations associated with checking assumptions about statistical distributions, error detection and their treatment. However, the major problem encountered by researchers, is lack of guidance regarding selection of the most appropriate software. Computer-aided statistical analysis should be undertaken with some understanding of the techniques being used. For example, some statistical software packages might replace

missing values with the means of the variable, or prompt the user for case-wise deletion of analytical data, both of which might be considered undesirable [52].

Lately, machine learning algorithms like the artificial neural networks (ANNs) [67, 76-78], and genetic algorithms (GA) [76, 79] have gained momentum in water quality monitoring studies. [41] pointed out that these technique generally yields the best parameter estimates in the data set with the least amount of missing data. Nevertheless, as the percentage of missing data increases, the performance of ANN which is generally measured by the errors in the parameter estimates, decreases and may reach performance levels similar to those obtained by the general substitution methods. However, in all cases the effectiveness of these methods lies on the user's ability to manipulate and display data correctly.

8. Conclusion

This chapter discussed the common data characteristics which tend to affect statistical analysis. It is recommended that practitioners should explore for outliers, missing values and censored values in a data-set before undertaking in-depth analysis. Although an analyst might not be able to establish the causal of such characteristic, eliminate or overcome some of the errors, having knowledge of their existence assists in establishing some level of confidence in drawing meaningful conclusions. It is recommended that water quality monitoring programs should strive to collect data of high quality. Common methods of ascertaining data quality are practising duplicate samples, using blanks or reference samples, and running performance audits. If a researcher is not sure of how to treat a characteristic of interest, a non-parametric method like Seasonal Kendal test could provide a better alternative since it is insensitive to common water quality data characteristics like outliers.

Author details

Innocent Rangeti[1], Bloodless Dzwairo[2], Graham J. Barratt[1] and Fredrick A.O. Otieno[3]

*Address all correspondence to: innoranger@gmail.com

1 Department of Environmental Health, Durban University of Technology, Durban, South Africa

2 Institute for Water and Wastewater Technology, Durban University of Technology, Durban, South Africa

3 DVC: Technology, Innovation and Partnerships, Durban University of Technology, Durban, South Africa

References

[1] Steel A, Clarke M, Whitfield P. Use and Reporting of Monitoring Data. In: Bartram J, Ballance R, editors. Water Quality Monitoring-A Practical Guide to the Design and Implementation of Freshwater Quality Studies and Monitoring Programmes: United Nations Environment Programme and the World Health Organization; 1996.

[2] Mitchell P. Guidelines for Quality Assurance and Quality Control in Surface Water Quality Programs in Alberta. Alberta Environment, 2006.

[3] Tasić S, Feruh MB. Errors and Issues in Secondary Data used in marketing research. The Scientific Journal for Theory and Practice of Socioeconomic Development. 2012;1(2).

[4] Doong DJ, Chen SH, Kao CC, Lee BC, Yeh SP. Data quality check procedures of an operational coastal ocean monitoring network. Ocean Engineering. 2007;34(2):234-46.

[5] Taylor S, Bogdan R. Introduction to research methods: New York: Wiley; 1984.

[6] Wahi MM, Parks DV, Skeate RC, Goldin SB. Reducing errors from the electronic transcription of data collected on paper forms: a research data case study. Journal of the American Medical Informatics Association. 2008;15(3):386-9.

[7] UNEP-WHO. Use and Reporting of monitoring data. In: Bartram J, Ballace R, editors. Water Quality Monitoring A practical Guide to the Design and Implementation of Freshwater Quality Studies and Monitoring Programmes. London: United Nations Environmental program, World Health Organisation; 1996.

[8] Dzwairo B. Modelling raw water quality variability in order to predict cost of water treatment. Pretoria: Tshwane University of Technology; 2011.

[9] Kawado M, Hinotsu S, Matsuyama Y, Yamaguchi T, Hashimoto S, Ohashi Y. A comparison of error detection rates between the reading aloud method and the double data entry method. Controlled Clinical Trials. 24. 2003:560–9.

[10] Cummings J, Masten J. Customized dual data entry for computerized analysis. Quality Assurance: Good Practice, Regulation, and Law. 1994;3:300-3.

[11] Brown ML, Austen DJ. Data management and statistical techniques. Murphy BR, Willis DW, editors. Bethesda, Maryland: America Fisheries Society; 1996.

[12] Rajaraman V. Self study guide to Analysis and Design of Information Systems. New Delhi: Asoke K. Ghosh; 2006.

[13] Data Analysis and Interpretation. Rhe Monitoring Guideline. Austria2000.

[14] Helsel DR, Hirsch RM. Statistical Methods in Water Resources. Amsterdam, Netherlands: Elsevier Science Publisher B.V; 1992.

[15] Köster D, Hutchinson N. Review of Long-Term Water Quality Data for the Lake System Health Program. Ontario: 2008 Contract No.: GLL 80398.

[16] Iglewicz B, Hoaglin DC. How to Detect and Handle Outliers1993.

[17] Zar JH. Biostatistical Analysis. Upper Saddle River, NJ.: Prentice-Hall Inc; 1996.

[18] Silva-Ramírez E-L, Pino-Mejías R, López-Coello M, Cubiles-de-la-Vega M-D. Missing value imputation on missing completely at random data using multilayer perceptrons. Neural Networks. 2011;24(1):121-9.

[19] USEPA., Ecology. Do. Technical Guidance for Exploring TMDL Effectiveness Monitoring Data. 2011.

[20] Stoimenova E, Mateev P, Dobreva M. Outlier detection as a method for knowledge extraction from digital resources. Review of the National Center for Digitization. 2006;9:1-11.

[21] US-EPA. Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities, Unified Guidance (Unified Guidance). US-Environmetal Protection Agency, 2009.

[22] US-EPA. Data Quality Assessment: Statistical Methods for Practitioners. In: Agency USEP, editor: United States Environmental Protection Agency; 2006.

[23] Osoborne JW. Improving your data transformations: Applying the Box-Cox transformation. Practical Assessment, Research and Evaluation. 2010;15(12).

[24] High R. Dealing with 'Outlier': How to Maintain your data's integrity.

[25] Chi Fung DS. Methods for the Estimation of Missing Values in Time Series. Western Australia: Edith Cowan University; 2006.

[26] US-EPA. Statistical Training Course for Ground-Water Monitoring Data Analysis.. Washington D.C.: Environmental Protection Agency, 1992.

[27] Grubbs FE, Beck G. Extension of sample sizes and percentage points for significance tests of outlying observations. Technometric. 1972;14:847-54.

[28] Tiwari RC, Dienes TP. The Kalman filter model and Bayesian outlier detection for time series analysis of BOD data. Ecological Modelling 1994;73:159-65.

[29] De Muth JE. Basic statistics and pharmaceutical statistical applications: CRC Press; 2014.

[30] Gibbons RD. Statistical Methods for Groundwater Monitoring. New York: John Wiley & Sons; 1994.

[31] Walfish S. A Review of Statistical Outlier Methods. Pharmaceutical Technology. 2006.

[32] Robinson RB, Cox CD, Odom K. Identifying Outliers in Correlated Water Quality Data. Journal of Environmental Engineering. 2005;131(4):651-7.

[33] Filzmoser P, editor A multivariate outlier detection method. Proceedings of the seventh international conference on computer data analysis and modeling; 2004: Minsk: Belarusian State University.

[34] Garret D. The Chi-square plot: A tool for multivariate recognition. Journal of Geochemical Exploration. 1989;32:319-41.

[35] Ssali G, Marwala T. Estimation of Missing Data Using Computational Intelligence and Decision Trees. n.d.

[36] Noor NM, Zainudin ML. A Review: Missing Values in Environmental Data Sets. International Conference on Environment 2008 (ICENV 2008); Pulau Pinang2008.

[37] Calcagno G, Staiano A, Fortunato G, Brescia-Morra V, Salvatore E, Liguori R, et al. A multilayer perceptron neural network-based approach for the identification of responsiveness to interferon therapy in multiple sclerosis patients. Information Sciences. 2010;180(21):4153-63.

[38] Helsel DR, Cohn T. Estimation of description statistics for multiply censored water quality data. Water Resource Research. 1988; 24:1997-2004.

[39] Little RJ, Rubin DB. Statistical analysis with missing data. New York: John Wiley and Sons; 1987.

[40] Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M. Methods for imputation of missing values in air quality data sets. Atmospheric Environment. 2004; 38:2895–907.

[41] Nieh C. Using Mass Balance, Factor Analysis, and Multiple Imputation to Assess Health Effects of Water Quality. Chicago, Illinois: University of Illinois 2011.

[42] Luengo J, Garcia S, Herrera F. A study on the use of imputation methods for experimentation with Radial Basis Function Network classifiers handling missing attribute values: The good synergy between RBFNs and EventCovering methodI. Neural Networks. 2010; 23:406-18.

[43] Lakshminarayan K, Harp S, Samad T. Imputation of missing data in industrial databases. Applied Intelligence. 1999;11(3):259-75.

[44] Baraldi AN, Enders CK. An introduction to modern missing data analyses. Journal of School Psychology. 2010;48(1):5-37.

[45] Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-92.

[46] Enders CK. A primer on the use of modern missing-data methods in psychosomatic medicine research. PsychosomaticMedicine. 2006;68:427–36.

[47] Fogarty DJ. Multiple imputation as a missing data approach to reject inference on consumer credit scoring. Intersat. 2006.

[48] Smits A, Baggelaar PK. Estimating missing values in time series. Netherlands: Association of River Waterworks – RIWA; 2010.

[49] Sartori N, Salvan A, Thomaseth K. Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose. ComputationalStatistics and Data-Analysis. 2005;49(3):937–53.

[50] Güler C, Thyne GD, McCray JE, Turner AK. Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeology 2002;10:455–74.

[51] Dzwairo B, Otieno FAO, Ochieng' GM. Incorporating surface raw water quality into the cost chain for water services: Vaal catchment, South Africa. Research Journal of Chemistry and Environment. 2010;14(1):29-35.

[52] Guler C, Thyne GD, McCray JE, Turner AK. Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeology. 2002;10:455-74.

[53] Starret KS, Starret KS, Heier T, Su Y, Tuan D, Bandurraga M. Filling in Missing Peak-flow data using Artificial Neural Networks. Journal of Engineering and Applied Science. 2010;5(1).

[54] Aitkenhead MJ, Coull MC. An application based on neural networks for replacing missing data in large datasets n.d.

[55] Lin P-E, Niu X-F. Comparison of Statistical Methods In Handling Minimum Detection Limits. Department of Statistics,, Florida State University, 1998.

[56] Darken PF. Testing for changes in trend in water quality data. Blacksburg, Virginia: Virginia Polytechnic Institute and State University; 1999.

[57] Farnham IM, Stetzenbach KJ, Singh AS, Johannesson KH. Treatment of nondetects in multivariate analysis of groundwatervgeochemistry data. Chemometrics Intelligent Lab Sys. 2002;60(265-281).

[58] Lyles RH, Fan D, Chuachoowong R. Correlation coefficient estimation involving a left censored laboratory assay variable. Statist Med. 2001;20:2921-33.

[59] Lopaka L, Helsel D. Statistical analysis of water-quality data containing multiple detection limits: S-language software for regression on order statistics. Computers & Geosciences. 2005;31:1241-8.

[60] Gheyas IA, Smith LS. A neural network-based framework for the reconstruction of incomplete data sets. Neurocomputing. 2010;73(16–18):3039-65.

[61] Nishanth KJ, Ravi V, Ankaiah N, Bose I. Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts. Expert Systems with Applications. 2012;39(12):10583-9.

[62] Helsel D. Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. Chemosphere. 2006;65:2434-9.

[63] Kalderstam J, Edén P, Bendahl P-O, Strand C, Fernö M, Ohlsson M. Training artificial neural networks directly on the concordance index for censored data using genetic algorithms. Artificial Intelligence in Medicine. 2013;58(2):125-32.

[64] Environmental Protection Agency. Data Quality Assessment: Statistical Methods for Practitioners. Agency USEP, editor. Washington: United States Environmental Protection Agency, Office of Environmental Information, 2006; 2006 14 June 2013. 198 p.

[65] Hewett P, Ganser GH. A Comparison of Several Methods for Analyzing Censored Data. British Occupational Hygiene. 2007;57(7):611-32.

[66] Hornung RW, Reed LD. Estimation of Average Concentration in the Presence of Nondetectable Values. Applied Occupational and Environmental Hygiene. 1990;5(1): 46-51.

[67] Kang P. Locally linear reconstruction based missing value imputation for supervised learning. Neurocomputing. 2013(0).

[68] Shumway RH, Azari RS, Kayhanian M. Statistical approaches to estimating mean water quality concentrations with detection limits. Environmental Science and Technology. 2002;36:3345-53.

[69] Helsel D. Less than obvious-statistical treatment data below the detection limit. Environmental Science and Technology. 1990;24(12):1766-44.

[70] Sanford RF, Pierson CT, Crovelli RA. An objective replacement method for censored geochemical data. Math Geol. 1993;25:59-80.

[71] Antweiler RC, Taylor HE. Evaluation of Statistical Treatments of Left-Censored Environmental Data using Coincident Uncensored Data Sets. Environmental Science and Technology. 2008;42(10):3732-8.

[72] Fu L, Wang Y-G. Statistical tools for analysing water quality data2012.

[73] Silva JdA, Hruschka ER. An experimental study on the use of nearest neighbor-based imputation algorithms for classification tasks. Data & Knowledge Engineering. 2013;84(0):47-58.

[74] Visser A, Dubus I, Broers HP, Brouyere S, Korcz M, Orban P, et al. Comparison of methods for the detection and extrapolation of trends in groundwater quality. Journal of Environmental Monitoring. 2009;11(11):2030-43.

[75] Schertzer TL, Alexander RB, Ohe DJ. The Computer Program Estimate Trend (ESTREND), a system for the detection of trends in water-quality data Water Resources investigation report. 1991;91-4040:56-7.

[76] Recknagel F, Bobbin J, Whigham P, Wilson H. Comparative application of artificial neural networks and genetic algorithms for multivariate time-series modelling of algal blooms in freshwater lakes. Journal of Hydroinformatics. 2002;4(2):125-34.

[77] Lee JHW, Huang Y, Dickman M, Jayawardena AW. Neural network modelling of coastal algal blooms.Ecological Modelling. 2003;159(179-201).

[78] Singh KP, Basant A, Malik A, Jain G. Artificial neural network modeling of the river water quality—a case study. Ecological Modelling. 2009;220(6):888-95.

[79] Muttil N, Lee JHW. Genetic programming for analysis and real-time prediction of coastal algal blooms. Ecological Modelling. 2005;189:363-76.