Journal of Science Research (2013) Vol. 12: 179-188

Data Visualization and Change-point detection in Environmental Data: The Case of Water Pollution in Oyo State Nigeria

Obisesan, K.O.[1], Lawal, M.[2], Bamiduro, T.A.[1], and Adelakun, A.A.[1]

Abstract
Environmental Statistics or Environmetrics is the application of statistical methods to problems concerning the environment. A large number of professionals and public office holders in many countries especially the developing countries have not taken enough cognisance of the use of statistical methodologies in understanding and explaining environmental processes. In this work, an attempt is made to use environmental pollution data to explain data visualization with respect to the standards by the World Health Organization and the Standards Organization of Nigeria. We also discuss exploratory data analysis and change-point detection. A water pollution example is given to show the importance of environmental statistics. Statistical models are used to explain the change-point problem. This effort is aimed at achieving some methodological changes in the approach to analysing environmental data such that statistical methodologies could be enhanced for sustainable development.

Introduction

Environmental Statistics or Environmetrics simply refers to the application of statistical methods to problems concerning the environment. This area of statistics has been developing rapidly over the past 20-25 years in response to growing concerns among individuals, organizations and governments, for protecting the environment. Unfortunately, practitioners in many countries especially the developing countries have not taken enough cognisance of the use of statistical methodologies in understanding and explaining many environmental processes. The practical application of modern statistical approaches to research and development (R&D) will improve the design stage of a development process and help to discover and unravel the fundamental processes essential to R&D. Thereafter, a more or less detailed model can be routinely built to explore alternative scenarios [1-5], provide some approaches to statistical modelling.

Obisesan, K.O.[1], Lawal, M.[2], Bamiduro, T.A.[1], and Adelakun, A.A.[1]
[1]Department of Statistics University of Ibadan, Nigeria
[2]Mathematical and Computer Science Department, Fountain University, Osogbo
Correspondence author: [1]email:obidairo@gmail.com

Human activities introduce contaminants of many kinds into the environment. As noted by Gilbert [6] and Ott [1], these contaminants appear in such a way that what is thrown away does not actually go away even after several years of its introduction into the environment. As a result, many pollutants stay in the environment and develop to cause very serious and hazardous effects on humans. Current environmental concerns include the response of climate to greenhouse gas emissions and the knock-on effects in areas such as water resources, agriculture and human health, the effects of industrial activity upon the quality of air and drinking water; the implications of intensive agricultural practices for biodiversity; and the sustainability of industrial scale fishing operations.

Many monitoring researches are currently being conducted to quantify the level and amount of pollutants entering the environment and to monitor ambient level for trends and potential environmental problems. Other studies seek to determine how pollutants distribute and persist in air, water, soil and biota, and to determine the effects of pollutants on humans and the environment.

Broadly, to attain or provide the information needed to reduce and control environmental pollution, it is essential we

ISSN 11179333

subject the studies into statistical scrutiny. This is one of the main objectives of environmetrics, as it has become a significant, highly recognised field in solving and studying broad fields such as environmental conservation, pollution and evaluation and control, monitoring of ecosystems, and management of resources.

Pollution is the introduction of contaminants into a natural environment. It causes instability, disorder, harm or discomfort to the ecosystem i.e. physical systems or living organisms. Pollution can occur in the form of chemical substances or energy, such as noise, air, water and heat or light. Pollutants, the elements of pollution, can be either foreign substances, energies or naturally occurring contaminants. Water pollution is the discharge of waste water from commercial and industrial waste (intentionally or through spills) into surface water; discharges of untreated domestic sewage, and chemical contaminants, such as chlorine from treated sewage; release of waste and contaminants into surface run-off owing to surface waters (including urban run-off and agricultural run-off, which may contain chemical fertilizers and pesticides), waste disposal and leaching into groundwater, eutrophication and littering [7].

The main objective of this paper is to use statistical methodologies to explain environmental processes with a view of achieving some methodological changes in the approach to environmental data analysis, and also to investigate change-point in water pollution data, and use environmental pollution data to explain data visualization.

Data Visualization

Almost all environmental data seem to have errors associated with them and therefore need a lot of careful checks. The process of checking the data is described as the visualization process. All classical statistical analyses are designed to assist in the interpretation and decision-making processes. The first stage in any statistical analysis is to clean-up the data; this process is known as data visualization (DV) and also sometimes referred to as data mining (DM) or exploratory data analysis (EDA). Data visualization or EDA can be described as the simple procedures that may be applied or carried out on some data-sets so as to see the secrets that may seem hidden in the data-sets. In this way, the data will be allowed to speak for themselves. However, this serves only as a prelude before carrying out a more sophisticated analysis.

Data visualization process has been neglected in many surveys and environmental audit reports, yet impact assessments are being reported without this very important process. Data visualization is a highly recommended procedure before any survey is carried out. In this way, it is advisable to involve statisticians at each level of intended environmental surveys. The sources of error in environmental data include the measurement processes; transcription errors; observations below the limit of detection and undocumented infilling of missing data with functional values among others. In this regard, data visualization can be employed to identify possible data quality problems and if possible correction can be made at the outset before employing sophisticated statistical modelling.

Chandler and Scott [2] pointed out many advantages from data visualization. It can lead to detection of outliers. It allows investigators to determine data features that require further investigations. We can also see clearly if anything is wrong with the data from the beginning. However, an observation may not necessarily be erroneous simply because it is an outlier. The visualization process can be used to suggest assumptions that may plausibly be held in the initial stages of subsequent analysis, to guide the choice of analysis strategy. It should be emphasised that assumptions made on the basis of a preliminary exploratory exercise should be regarded as indicative only. A more detailed analysis can reveal more secrets and features that were not apparently discovered. Therefore, the use of a preliminary analysis to suggest plausible assumptions does not replace the need for further checks.

Data Visualization: Water Pollution Example

Environmental data-sets in developing countries are difficult to gather because governments, policy and decision makers are yet to appreciate the importance of data to development. Data on the environment are usually difficult to collect and collection may be guided by law. They are also expensive to collect. In this section, Table 1 shows units of some water pollution variables. Table 2 shows the physico-chemical properties of sixty water samples taken from Eleyele reservoir as compared with permissible standards [8-9]. On this occasion, samples in respect of Eleyele reservoir have been extracted from records of a coordinating office as shown in Figure 1. Table 2 also shows the safe column referring to cases, where the chemical concentrations from the water samples do not go above, either of the SON or WHO standards. The polluted column indicates cases where water samples possess concentrations slightly above the approved standards, while the fourth column indicates the contaminated cases. This is used to represent cases where concentrations from water samples are well above the maximum standards. Other similar data available for explanation are water pollution data for Asejire reservoirs also on monthly concentrations of some physico-chemical variables monitored between 2003 and 2007. These are shown in Figures 2 and 3 respectively. Figure 4 compares the pollution levels in the two reservoirs under discussion.

Table 1: Variables and Corresponding Units of Measurement

Variables	Units of Measurements
Turbidity (Tur)	Nephelometric Turbidity Units (NTU)
Colour (Col)	Hazen Units (HU)
PH (pH)	Logarithmic Units (LU)
Dissolved Oxygen (DO)	milligram per litre (mg/l)
Alkalinity (Alk)	milligram per litre (mg/l)
Total Hardness (TH)	milligram per litre (mg/l)
Calcium Hardness (CaH)	milligram per litre (mg/l)
Chloride (Cl)	milligram per litre (mg/l)
Iron (Fe)	milligram per litre (mg/l)
Silica (Si)	milligram per litre (mg/l)
Total Solids (Sol)	milligram per litre (mg/l)
Dissolved Solids (DS)	milligram per litre (mg/l)
Total Suspended Solids (SS)	milligram per litre (mg/l)
Chemical Oxygen Demand (COD)	milligram per litre (mg/l)
Biochemical Oxygen Demand (BOD)	milligram per litre (mg/l)
Nitrate (NITRT)	parts per million (ppm)

Table 2: Maximum Chemical Concentrations based on Standard Organisation of Nigeria [8] and World Health Organisation [9] Reports as Compared with Water Samples

Variables	Safe	Polluted	Contaminated	SON	WHO
Turbidity	8	31	11	5NTU	5NTU
Colour	57	2	1	15HU	15HU
Chloride	60	0	0	250mg/l	250mg/l
Total Hardness	60	0	0	150mg/l	200mg/l
Iron	0	0	60	0.3mg/l	0.3mg/l
Total Dissolved Solids	60	0	0	500mg/l	600mg/l
pH	-	-	-	-	-

The values on safe, polluted and contaminated refer to number of samples out of a total of sixty months considered. This is used to represent cases where concentrations from water samples are well above the maximum standards. For example a total of 42 samples out of 60 samples from Eleyele have been found to go out of standards. Also, three samples violate standards of colour while all sixty samples taken violate iron standards. This is based on Standard Organization of Nigeria [8] and World Health Organization [9]. With regard to colour we can say it shows that Eleyele water was colourless and met the standards for safe water by SON and WHO.

Data Visualization: Water Pollution Results

Figure 1 shows some unexpected and irregular behaviours for some variables such as Biochemical Oxygen Demand (BOD). Some extreme values can also be noticed in the variables. In particular, the values the Biochemical Oxygen Demand (BOD) and the Chemical Oxygen Demand (COD) series show some seasonality behaviour over the same period. Nitrates series also exhibit a significant change towards the 25th month with a sudden increase in concentration. However, these behaviours suggest abrupt changes that must be investigated by the scientist. This can lead to change-point detection which can help in determining the location and position of changes for appropriate statistical modelling.

Change-point as defined by Chandler and Scott [2] is a point in time at which the properties of a process changes abruptly. The shift may include sudden changes in mean, variance or autocorrelation structure. Chandler and Scott [2] also gave a practical example to understanding change-points. They indicated that change-point can be understood from known events such as damming of a river, changes in instrument-ation and strange flooding. Such events usually bring significant changes in the system.

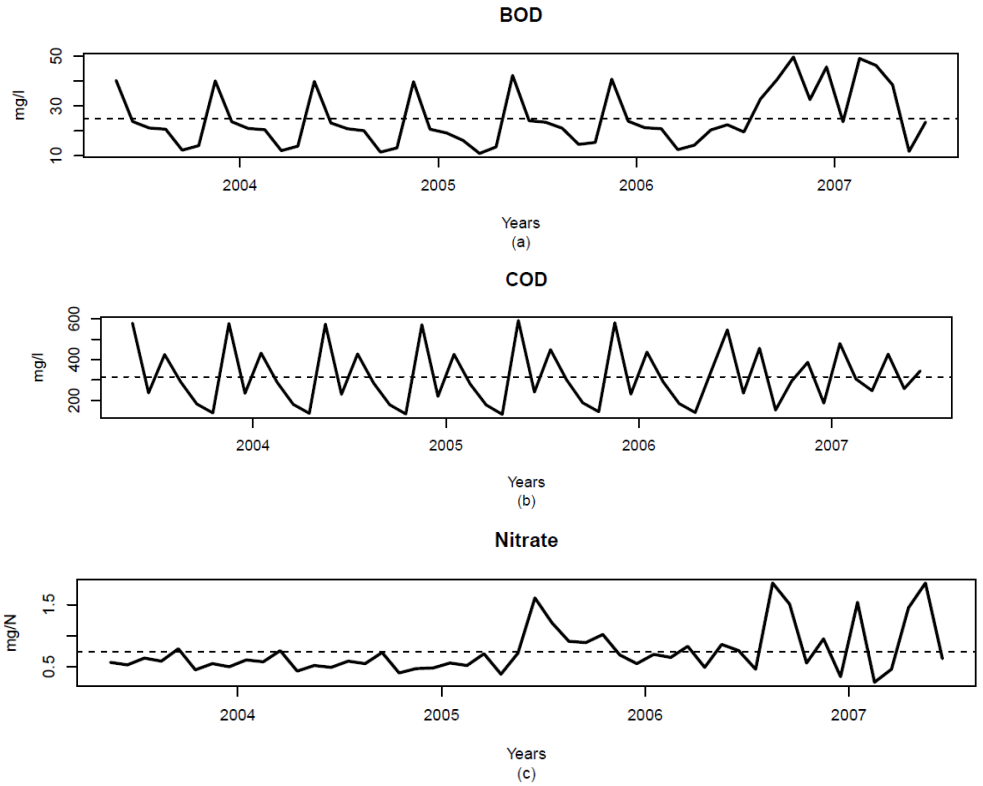

Fig. 1: Irregular time plots of monthly data from Eleyele reservoir.

For Asejire, Figure 2 shows different values for many variables but noticeably repeated colour concentrations of 5 Harzen Units (HU) suggesting that the colour of the water samples may not change over time. In Figure 3 there are irregular patterns for variables from Eleyele and occasional extreme values of Dissolved Oxygen (DO) and colour concentration going outside set standards (may be dangerous). All this also supports the need to investigate change-points. In respect of the water samples from Eleyele and considering polluted and contaminated cases it is important to say that Table 2 indicates a total of 42 polluted samples of Turbidity going outside 5NTU specified by the SON and WHO. Also all the sixty samples of Iron seem to have gone outside the specified 0.3mg/l. Unless purified, the water samples being discussed in this case indicate that the water may not be safe for drinking.

Although there are other extreme values evident in other variables in the data set, colour and DO are important measures of pollution. Colour refers to physical appearance while DO supports aquatic organisms in water. In general, the analysis suggests the need to study abrupt changes in the water quality data from the two reservoirs. The point of change is very important as it can lead to detection of errors so that solutions to problems can be provided. It can also serve as a strong prelude to more advanced statistical modelling approaches. In the analysis described above, we can say therefore that the data have suggested the use of change-point to discover points of appropriate changes. In Figure 4, a bar-plot compares the average chemical profile from the two reservoirs. The average had been obtained from the raw and final water samples taken from the two reservoirs over the same period of sixty months. The comparison shows that the values such as

turbidity, total solids, colour, alkalinity and many more variables at Eleyele are higher than those from Asejire reservoir. This may be because Eleyele is in the city center and is more polluted than the Asejire reservoir. Therefore, Eleyele is prone to receiving more pollution from major sources such as industries in the city.

There is a need for creating a statistical model to capture and explain the structure of any environmental series. This will hopefully throw more light on the need to incorporate statistical methodologies at the design stage of a research or study to enable more meaningful analysis and interpretation especially of environmental data.

This is between 2003 and 2007 for Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), and Nitrate respectively. Data extracted from records of Ministry of Environment Ibadan, Nigeria.

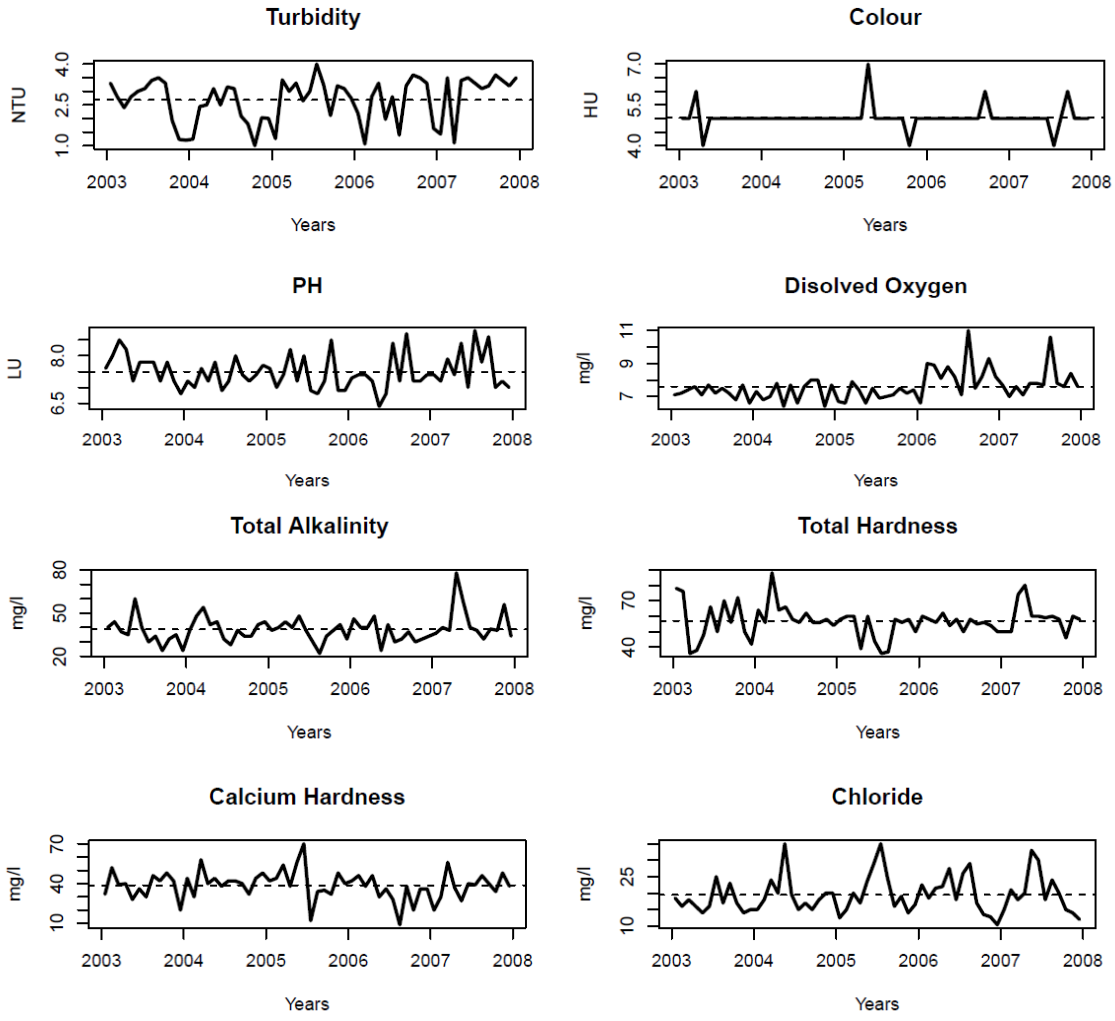

Fig. 2: Time plots of monthly concentrations between 2003-2007 for Asejire reservoir: Data supplied by the Water Corporation of Oyo State.

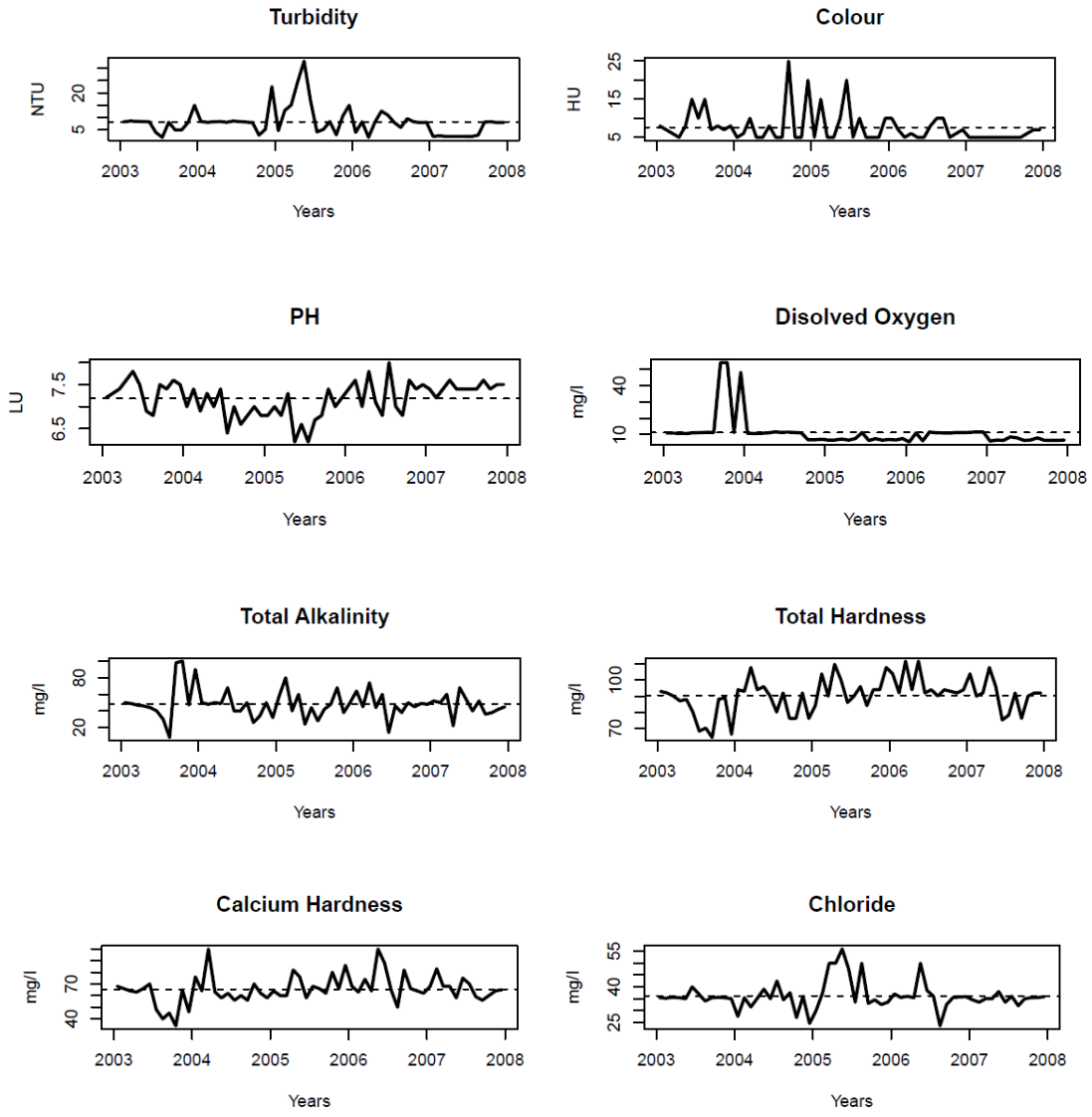

Fig. 3: Time plots of monthly concentrations between 2003-2007 for Eleyele reservoir: Data supplied by the Water Corporation of Oyo State.

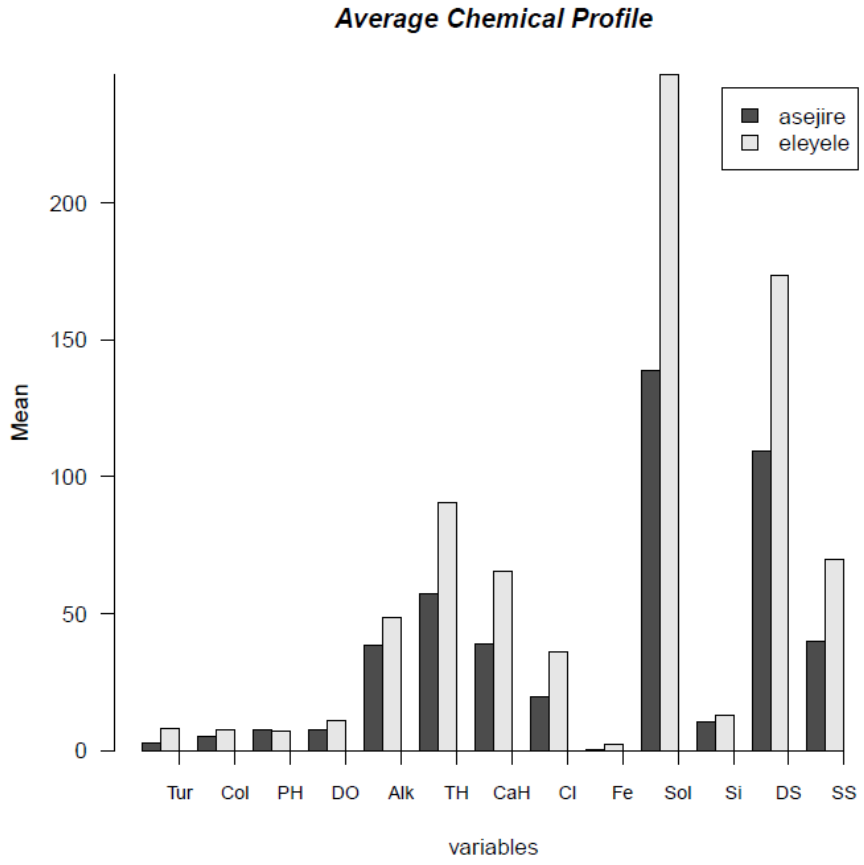

Fig. 4: Bar plot for each variable taken between 2003-2007 for Asejire and Eleyele reservoirs in Nigeria.

The average chemical profiles are derived from raw and final samples from the two reservoirs. Eleyele reservoir seems more polluted than Asejire since it is in the city center, thereby, receiving more pollutants from many sources.

Statistical Modelling of Change-point

We now apply statistical theory to explain the change-point problem. In this case, we shall review some theory on change-point problem and then build a change-point model to be applied to the available data.

Two-mean Model

David Hinkley's contribution to change-point analysis is a central research that should be reviewed in detail. Hinkley [10] considered sequences of random variables and discussed the point at which the probability distribution changes using a normal distribution with changing mean. The asymptotic distribution of the maximum likelihood estimate discussed in this paper is particularly relevant to change-points in the study of the likelihood ratio statistic for testing hypotheses about the change-point. The author indicated the simplest model over a whole range of data as $X_t = \theta(t) + \epsilon_t$ for $t = 1, \dots T$ where $\theta(t)$ is a mean function and ϵ_t refers to error terms. Hinkley [10] computed the asymptotic distribution in the normal case when θ_0 and θ_1 are unknown. The asymptotic distribution is found to be the same when the mean levels are known.

The two-mean model to be considered supposes that there exists a mean $\theta_0(t)$ and mean $\theta_1(t)$ for $t = 1, \dots, \delta$ and $t = \delta + 1, \dots, T$ respectively. He also computed the asymptotic distribution of the likelihood ratio

statistic for testing hypothesis on the change-point δ. The maximum likelihood estimate of the change-point $\hat{\delta}$ (where θ_0 and θ_1 are known and δ is unknown) is obtained from a sample $x_1, \ldots, x_T$ by simply maximizing the likelihood function of the form.

$$L(\delta, \theta_0, \theta_1) = \prod_{i=1}^{\delta} f(x_i, \theta_0) \prod_{i=\delta+1}^{T} f(x_i, \theta_1)$$

which can be written in the form of the log-likelihood as

$$L(\delta, \theta_0, \theta_1) = \sum_{i=1}^{\delta} \ln f(x_i, \theta_0) + \sum_{i=\delta+1}^{T} \ln f(x_i, \theta_1) \quad (1)$$

Given that the variance of the independent observations is $\widehat{\sigma^2}$, consider the case where the mean levels θ_0 and θ_1 are not known. The log-likelihood of the observed sequence $(x_1, \ldots, x_T)$ is

$$L(\delta, \theta_0, \theta_1, \sigma^2 | x_1, \ldots, x_T) = -\frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left\{\sum_{i=1}^{\delta}(x_i - \theta_0)^2 + \sum_{i=\delta+1}^{T}(x_i - \theta_1)^2\right\} \quad (2)$$

If we assume that δ is known, therefore the maximum likelihood estimators of θ_0, θ_1 and σ^2 respectively are $\hat{\theta}_0 = \frac{\sum_{i=1}^{\delta} x_i}{\delta}$, $\hat{\theta}_1 = \frac{\sum_{i=\delta+1}^{\delta} x_i}{T-\delta}$ and $\widehat{\sigma^2} = \frac{\sum_{i=1}^{\delta}(x_i - \theta_0)^2 + \sum_{i=\delta+1}^{T}(x_i - \theta_1)^2}{T}$.

Particularly for convenience, Hinkley [10] substituted $\sigma^2 = 1$ as known so that equation (2) becomes

$$L(\delta, \theta_0, \theta_1 | x_1, \ldots, x_T) = -\frac{1}{2}\left\{\sum_{i=1}^{\delta}(x_i - \theta_0)^2 + \sum_{i=\delta+1}^{T}(x_i - \theta_1)^2\right\} \quad (3)$$

Putting the maximum likelihood estimates of θ_0 and θ_1 back into the log-likelihood in Equation (3) and re-arranging the emerging sum of squares conditional on t, equation (3) becomes

$$l(t) = -\frac{1}{2}\left\{\sum_{i=1}^{T}(x_i - \bar{x}_T)^2 - t(T-t)(\bar{x}_t - \bar{x}_t^*)^2/_T\right\}, \quad (4)$$

for which $\bar{x}_T = \frac{\sum_{i=1}^{T} x_i}{T}$, $\hat{\theta}_0 = \bar{x}_t$, $\hat{\theta}_1 = \bar{x}_t^*$ and $\hat{\delta}$ is the value of t that maximises the observed value z_t^2 of Z_t^2 where

$$Z_t^2 = t(T-t)(\bar{X}_t - \bar{X}_t^*)^2/_T$$

In our own case; suppose that σ^2 is unknown, then we can use equation (2) to obtain a profile log-likelihood function for δ by substituting all maximum likelihood estimates of the other parameters except the change-point into equation (2) to obtain the profile log-likelihood function of the change-point $l_p(t)$ as in

$$l_p(t) = -\frac{T}{2}\ln\left[\frac{\sum_{i=1}^{T}(x_i - \bar{x}_T)^2}{T} - \frac{t(T-t)(\bar{x}_t - \bar{x}_t^*)^2}{T^2}\right] - \frac{T}{2} \quad (5)$$

Applying equation (5), which represents the test statistics used to obtain the change point, we obtain the results in Table 3.

Table 3: Hinkley Model Application to Turbidity Concentration from Asejire and Eleyele Reservoirs

Hinkley	δ	θ_0	θ_1	σ^2	$l_p(\delta)$	W	Year
Asejire	51	2.562	3.36	0.563	-12.77	7.97	2007
Eleyele	48	9.2	4.3	27.16	-12.9	7.77	2006

The points of change in the two reservoirs considered are close, as displayed in Table 3. Mainly the change-point, 51 for Asejire corresponds to 2007 and change-point, 48 for Eleyele corresponds to 2006. All these could have gone unnoticed and knowing these in advance could make room for planning and therefore, could enhance sustainable development.

Conclusion

This work underscores the importance of incorporating statistics into the experimental designs or the study at the outset of a development process to enable the right procedures to be employed and ensure correct analysis and interpretation of the data collected. It points out that the application of well-designed statistics in environmental data collection, processing and interpretation enriches the policy and decision-making process and improves the quality of the environment and development. The work also underlines the relevance of data visualization, a component of environmental statistics, which makes it possible to see clearly from the beginning if anything is wrong with the data. Data visualization is a highly recommended procedure before any survey is conducted. It provides a model that could be used to detect change-point in water pollution contributing to achieving sustainable development.

Acknowledgements

This work is part of a research done at the University College London (UCL). Authors are extremely grateful to the Department of Statistical Science at UCL and the World-Bank USA for providing funds at different stages. UCL provided training in Environmental Statistics and the World-Bank provided some conference grants in the year 2012.

References

[1] Ott, W.R. 1992. Environmental Statistics and Data Analysis, Lewis Publishers, New York.

[2] Chandler, R.E. and Scott, O.M. 2011. Statistical Methods for Trend Detection and Analysis in the Environmental Sciences, Wiley.

[3] Adelakun, A.A. 2009. Elements of Environmental Statistics, Unpublished (in-press), Ibadan.

[4] Obisesan, O.K. 2008. Use of Modified Lognormal Distribution for Assessing Water Pollution in Eleyele Reservoir, M.Phil Thesis, Department of Statistics University of Ibadan, Nigeria, pp 1-4.

[5] Obisesan, O.K. 2011. Change-point Detection in Time-Series with Hydrological Applications., M.Phil Thesis, University College London, London, pp 2-5.

[6] Gilbert, R.O. 1987. Statistical Methods for Environmental Pollution Monitoring, Van Nostrand Reinhold, USA.

[7] Spengler, J.D. and Sexton, K. 1983. Indoor Air Pollution: A Public Health Perspective Science (New Series), 221(4605):9-17.

[8] Standards Organisation of Nigeria. 2007. Nigerian Standards for Drinking Water Quality, Standards Organisation of Nigeria Governing Council, Lagos, Nigeria.

[9] World Health Organisation. 2008. Guidelines for Drinking Water Quality, Incorporating the first and second edition, WHO, Geneva, volume I recommendation.

[10] Hinkley, D.V. 1970. Inference about Change-point in a Sequence of Random Variables, *Biometrika*, 57(1): 1-17.