STATISTICAL ANALYSIS AND QUALITY ASSURANCE OF MONITORING DATA

Iris Yeung
City University of Hong Kong, Kowloon, Hong Kong

Keywords: AIC, ARIMA model, BIC, cluster analysis, discriminant analysis, factor analysis, field sampling, intervention model, laboratory analysis, monitoring data, multidimensional scaling, multivariate analysis, nonparametric test, principal components analysis, quality assurance, seasonal kendall slope estimator, seasonal kendall test, serial correlation, shewhart control chart, statistical analysis, time series model, transfer function model, trend detection

Contents

Summary

In environmental monitoring, many variables are measured at various stations over a long period of time. If analyzed properly, this large bulk of data can detect trends or changes in the environmental conditions, determine the effectiveness of pollution control actions, and reveal the interrelationships between the variables and the monitoring stations. The statistical techniques which are appropriate for doing such analysis are the seasonal Kendall test and slope estimator, ARIMA, and transfer

function models (for detecting trends or changes); intervention models (for determining the effectiveness of control actions); and multivariate analysis, respectively. To have meaningful statistical results, the data used for analysis must be of high quality. So, in addition to describing the use of statistical analysis, this paper also discusses quality assurance procedures in environmental monitoring. Control charts are an important component of quality assurance and this paper also examines how various types of Shewhart control charts are constructed.

1. Introduction

Many government agencies have established environmental monitoring programs. These monitoring programs generate a large bulk of data which can provide a great deal of information about pollution, trend, effectiveness of policy, and so on, with suitable statistical analysis. However to have meaningful statistical results, the data used for analysis must be accurate, reliable, and meet certain data quality objectives. So statistical analysis and quality assurance are two essential aspects of monitoring.

Many statistical procedures have been used to analyze environmental data. For example, descriptive summary measures; graphical techniques; confidence intervals and hypothesis testing using the normal, *t*, chi-squared, and *F* distributions; analysis of variance; regression analysis; nonparametric statistics; time series models; and multivariate analysis are used. As the first five statistical techniques are commonly seen in many environmental statistics books, this paper concentrates on the last three statistical techniques that appear to have great potential for environmental applications.

Quality assurance (QA) refers to all measures that are taken to achieve quality. It is often used in industry for the purpose of getting high quality products at lower costs. In environmental monitoring, QA is applied to ensure that the data generated meet defined standards of quality. Different monitoring programs have different QA systems. This paper considers only the generalized structure of a QA system in field sampling and laboratory analysis. Control charts play a major role in QA and so this paper also describes how they are constructed.

This paper is organized into three sections beyond the introduction. The first section is devoted to the discussion of statistical analysis. The second section examines QA and control charts. The last section describes the computer programs that are available to carry out the statistical analysis and quality assurance. Throughout this paper, focus is given to statistical analysis and QA of water quality data. However, many of the principles given here can be used for other types of monitoring data, such as air quality data.

2. Statistical Analysis

The three classes of statistical analysis that are described in this paper are the seasonal Kendall test and slope estimator; time series models, and multivariate analysis. Table 1 lists some important statistical methods within each class along with brief descriptions of their purposes. As summarized in this table, the seasonal Kendall test and slope estimator, the ARIMA model, and the transfer function model can be used to detect monotonic trends and estimate the magnitude of a trend. The intervention model, which is a special case of the transfer function model, is often used to measure the impact of the interventions on the mean level of a time series. Multivariate analysis is used to discover the relationships that exist between the environmental variables and the monitoring stations from which the sample measurements are taken.

Analysis	Description
Nonparametric method	
Seasonal Kendall test and slope estimator	A nonparametric test used to detect a trend in a seasonal time series and to estimate the magnitude of the trend.
Time-series model	
ARIMA model	A type of time-series model in which the series to be forecast is expressed as a function of both previous values of the series (AR terms) and previous error values from forecasting (MA terms). The model can be used to indicate whether there is a trend in the data and whether the future values of the series comply with the environmental standards.
Transfer function model	A type of time-series model that relates the value of a time series to other related series. The model can be used to estimate the magnitude of the trend.
Intervention model	A special type of transfer function model used to determine the effects of the interventions on the mean level of a time series.
Multivariate analysis	
Principal components analysis	Multivariate technique used to reduce many original variables to a few linear combinations of them called principal components that best represent the original data. The plot of component scores is used to see whether the observations can be grouped into clusters.
Factor analysis	Multivariate technique used to analyze the interrelationships among a large number of variables and then explain these variables in terms of a few common, underlying dimensions called factors.
Discriminant analysis	Multivariate technique used to classify observations to one of a set of *a priori* defined groups on the basis of their values on a set of independent variables.
Cluster analysis	Multivariate technique used to classify observations into several mutually exclusive groups based on their similarities and differences.
Multidimensional scaling	Multivariate technique used to construct a map showing the relative location of the

	objects from a table of distances between them.

Table 1. Statistical analyses and their descriptions

Environmental data are often “messy.” They may be non-normally distributed, have missing values, possess outliers, and contain censored values which are reported as less than the limit of detection (LD). Before applying a statistical method, one should check whether the method can still be used with data having these characteristics. If not, these problems have to be handled first before using a statistical method.

2.1. Seasonal Kendall Test and Slope Estimator

2.1.1 Seasonal Kendall Test

The seasonal Kendall test is a nonparametric test used to detect monotonic trends especially in seasonal data. Like the other nonparametric tests, the seasonal Kendall test does not depend on the data being normally distributed. Suppose there are K years, each having s seasons. Let x_{ij} denote the measurement for the ith season in the jth year and x_{ik} denote the measurement for the ith season in the kth year, where $k > j$. To carry out the test, the following Mann-Kendall test statistic is calculated for season i of the year

$$S_i = \sum_{j=1}^{n_i - 1} \sum_{k=j+1}^{n_i} \operatorname{sgn}(x_{ik} - x_{ij}) \qquad (1)$$

where n_i is the number of nonmissing measurements in season i; and

$$\operatorname{sgn}(x_{ik} - x_{ij}) = \begin{cases} 1 & \text{if} \quad x_{ik} - x_{ij} > 0 \\ 0 & \text{if} \quad x_{ik} - x_{ij} = 0 \\ -1 & \text{if} \quad x_{ik} - x_{ij} < 0 \end{cases} \qquad (2)$$

Under the null hypothesis of no trend, S_i is approximately normally distributed with the mean and variance as

$$E(S_i) = 0$$
$$\operatorname{Var}(S_i) = \frac{1}{18}\left[n_i(n_i - 1)(2n_i + 5) - \sum_{p=1}^{g_i} t_{ip}(t_{ip} - 1)(2t_{ip} + 5) \right] \qquad (3)$$

where g_i is the number of groups of values tied in season i and t_{ip} is the number of ties in the pth group for season i. If, for example, there are three groups of ties each of size two and one group of ties of size three, then $\sum_{p=1}^{g_i} t_{ip}(t_{ip} - 1)(2t_{ip} + 5)$ = 3(2)(1)(9) + 1(3)(2)(11) = 120.

A single value of S_i indicates whether there is a trend in season *i*. In order to know whether there is an overall trend for the entire environmental time series, all the seasonal test statistics are combined into a summary test statistic.

$$S' = \sum_{i=1}^{s} S_i \tag{4}$$

As the sum of *s* normal distributions is still normal, S' must also be normally distributed in the limit with the mean and variance as:

$$\begin{aligned} E\left(S'\right) &= \sum_{i=1}^{s} E\left(S_i\right) = 0 \\ \mathrm{Var}\left(S'\right) &= \sum_{i=1}^{s} \mathrm{Var}\left(S_i\right). \end{aligned} \tag{5}$$

Using a continuity correction of one unit, the quantity

$$Z = \begin{cases} \dfrac{S'-1}{\left[\mathrm{Var}\left(S'\right)\right]^{1/2}} & \text{if} \quad S' > 0 \\ 0 & \text{if} \quad S' = 0 \\ \dfrac{S'+1}{\left[\mathrm{Var}\left(S'\right)\right]^{1/2}} & \text{if} \quad S' < 0 \end{cases} \tag{6}$$

follows a standard normal distribution. Let $z_{\alpha/2}$ denote the value of the standard normal distribution with an area to the right of this value equal to $\alpha/2$. The null hypothesis of no trend will be rejected at a significance level of α if $Z > z_\alpha$ (test of an upward trend) or $Z < -z_\alpha$ (test of a downward trend). The decision rule for a two-tailed test is $Z > z_{\alpha/2}$ or $Z < -z_{\alpha/2}$.

As seen in Eqs. (1) and (3), the seasonal Kendall test can cope with missing and tied observations in the data. When there are outliers that arise from obvious mistakes, they are corrected if possible and the correct value is inserted. If the correct value is not known and cannot be obtained, the datum may be excluded and considered to be the missing value case. However if the observation is a genuine value, the outlier should be retained. As the seasonal Kendall test considers only the sign rather than the absolute magnitude, it may not be seriously affected by the presence of a few outliers.

As for "less than LD" values, they are considered to be tied with each other and lower than any numerical value at or above LD. If LD has been changed over time from LD_1 to LD_2 (where $LD_2 < LD_1$) due to the development of more sensitive instruments, then all data indicated as "less than LD_2" as well as any numerical values "less than LD_1" must be recoded to "less than LD_1," and then the test can be run as described above. Also the test can be used if there is no strong serial dependence in the data. The above formula of Var(S') assumes that each of the S_i is an independent variable. If the seasons are correlated, then the covariance terms between S_i and S_j have to be included into the formula of Var(S').

2.1.2. Seasonal Kendall Slope Estimator

The seasonal Kendall test can only determine whether there is a trend or not. To estimate the magnitude of a trend, the seasonal Kendall slope estimator is used to calculate a slope that provides a measure of the rate of change of a variable per unit of time.

To obtain the seasonal Kendall slope estimator, the first step is to compute the individual slope estimate

$$Q_i = \frac{x_{ik} - x_{ij}}{k - j} \tag{7}$$

for all (x_{ij}, x_{ik}) pairs in season i, $1 \le j < k \le n_i$. Suppose there are N_i' slope estimates that can be calculated for season i, i = 1, 2 , …, s. The seasonal Kendall slope estimator is the median of all the $N' = N_1' + N_2' + \ldots + N_s'$ individual slope estimates. The lower 100(1 − α)% confidence limit is the M_1th largest of the N' ordered slope estimates where

$$M_1 = \frac{N' - z_{\alpha/2}\left[\mathrm{Var}(S')\right]^{\frac{1}{2}}}{2} . \tag{8}$$

The upper 100(1 − α)% confidence limit is the $(M_2 + 1)$th largest value of the N' ordered slope estimates where

$$M_2 = \frac{N' + z_{\alpha/2}\left[\mathrm{Var}(S')\right]^{\frac{1}{2}}}{2}. \tag{9}$$

2.2. Time Series Models

In this section, three types of time series models are described, namely ARIMA models, transfer function models, and intervention models. Contrary to the nonparametric seasonal Kendall test that is distribution free, all the time series models considered here are based on the normality assumption of the residuals. Also, all these time series models require a sufficient number of observations collected at equal time intervals. So these models cannot be used with small data sets.

If available measurements are not evenly spaced, data filling is needed. Furthermore, the fitting of these models to the data is not straightforward as compared with the nonparametric test. However, the time series models have their own advantages. In particular, they can handle any autocorrelation that may be present in the data. Also they can provide better understanding on how some external activities affect the environment.

Before discussing each type of time series model, some usual strategies in making the data suitable for analysis are presented below. First, if the data are non-normally distributed, they are transformed to near normality. Second, if there is a small proportion of missing values, they may be replaced by the average of the observations within the respective season of the year.

The Statistical Analysis System (SAS) software package has a procedure (EXPAND) to interpolate the missing data and include these interpolated values for analysis as if they were actual data in the first place. Third, for values recorded as "less than LD," they may be replaced by a number equal to half of that LD. This procedure is found to be reasonable and may be seen as a consequence of assuming a uniform distribution in the small interval of nondetection and estimating the unknown value by the corresponding mean.

Fourth, if a seasonal time series has multiple observations in a season of the year, then the sample mean or the sample median of the observations is used to represent the value of the variable for that season so that the data set consists only of one observation for each season of the year. Fifth, if there are outliers, they may be removed and treated as missing value case because they may affect the analysis.

-
-
-

Bibliography

Alt F. and Jain K. (1996). Quality control. *Encyclopedia of Operations Research and Management Science,* (ed. S.I. Gass and C.M. Harris), pp. 536–549. Boston: Kluwer Academic Publishers. [This article gives a good introduction to the basic concepts of control charts and other quality control procedures.]

Batley G.E. (1999). Quality assurance in environmental monitoring. *Marine Pollution Bulletin* 39, 23–31. [This paper describes the quality assurance for a field monitoring exercise where trace metals in aquatic ecosystems are studied.]

Bowerman B.L. and O'Connell R.T. (1990). *Forecasting and Time Series: An Applied Approach*, 726 pp. Belmont, California: Duxbury. [This book is good for beginners and practitioners because it gives a detailed treatment of the Box-Jenkins methodology and shows how to use the SAS package to analyze many real world data sets.]

Box G.E.P., Jenkins G.M., and Reinsel G.C. (1994). *Time Series Analysis, Forecasting and Control*, third edition, 598 pp. Englewood Cliffs, New Jersey: Prentice Hall. [The first edition of this book, published in 1970, has been successful in popularizing the ARIMA models through the formulation of an iterative process of model building, consisting of the stages of identification, estimation, and diagnostic checking. Although the beginner may not find this book easy to read, it is an essential reference source.]

Box G.E.P. and Tiao G.C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association* 70, 70–79. [This paper gives an initial presentation of intervention analysis and uses it to study the impact of air pollution control and economic policies on the concentration of pollutants such as ozone and carbon monoxide.]

Chapman D. (1996). *Water Quality Assessments - A Guide to the Use of Biota, Sediments and Water in Environmental Monitoring*, 626 pp. London: E & FN Spon. [While this book describes water quality assessments in all types of freshwater bodies, it contains useful materials on obtaining good quality data in Chapters 2 and 10.]

Der G. and Everitt B. S. (2002). *A Handbook of Statistical Analysis Using SAS*, 360 pp. Boca Raton, Florida: Chapman and Hall. [This book gives a good introduction to most multivariate methods discussed in this paper and describes the use of SAS Software for the methods. The presentation of discriminant analysis in this paper follows their ideas.]

Duncan A.J. (1986). *Quality Control and Industrial Statistics*, 1123 pp. Homewood, Illinois: Irwin. [This book gives a complete introduction to statistical quality control and contains some rules which are not discussed in this paper to identify a trend or nonrandom pattern on a control chart.]

Farrar A.C. (2000). Quality Assurance in Air Sampling and Analysis. Available: http://www.claytongrp.com/qualassur_art.html [This article gives a good description of quality assurance in air sampling and analysis. Few sentences in the quality assurance section of this paper are quoted from this article.]

Gilbert R.O. (1987). *Statistical Methods for Environmental Pollution Monitoring*, 320 pp. New York: Van Nostrand Reinhold. [This book gives a full description of the basic statistical techniques used in environmental pollution monitoring studies and contains Fortran program codes that estimate and test for trends over time using nonparametric methods. The topics covered in this book include sampling plans, statistical tests, and parameter estimation techniques.]

Granger C.W.J. (1989). *Forecasting in Business and Economics*, 279 pp. Boston: Academic Press. [Although this book describes many forecasting techniques used in business and economics, it contains useful concepts on forecasting and time series models.]

Hair J.F., Anderson R.E., Tatham R.L., and Black W. C. (1998). *Multivariate Data Analysis*, 730 pp. Upper Saddle River, New Jersey: Prentice Hall. [This book is easy to read and gives a good introduction to multidimensional scaling. It also contains a good source of glossaries in multivariate analysis area. A few glossary entries in this paper are adapted from this book.]

Hipel K.W., Lennox W.C., Unny T.E., and McLeod A.I. (1975). Intervention analysis in water resources. *Water Resources Research* 11(6), 855–861. [This paper applies intervention analysis to examine the effect of a dam on the mean annual flow of a river.]

Hirsch R. M., Slack J. R., and Smith R. A. (1982). Techniques of trend analysis for monthly water quality data. *Water Resources Research* 18 (1), 107-121. [This paper gives a more detailed description on seasonal Kendall test.]

Hunt W.F., Clark J.B., and Goranson S.K. (1978). The Shewhart control chart test: a recommended procedure for screening 24 hour air pollution measurements. *Journal of the Air Pollution Control Association* 28(5), 508–510. [This paper illustrates how Shewhart control charts are applied to check ambient air quality data for anomalies due to keypunch, transcription, and measurement errors.]

Jobson J.D. (1991). *Applied Multivariate Data Analysis,* Volume 2, 731 pp. New York, Berlin: Springer-Verlag. [This book gives a good coverage on all multivariate methods discussed in this paper.]

Keith L.H. (1991). *Environmental Sampling and Analysis: A Practical Guide*, 143 pp. Chelsea, Michigan: Lewis. [This book gives detailed information on what aspects of sampling and analytical activities can successfully obtain data of a known quality.]

Makridakis S., Wheelwright S.C., and Hyndman R. J. (1998). *Forecasting: Methods and Applications*, 642 pp. New York: John Wiley and Sons. [This book contains a good source of glossaries in statistical forecasting area. A few glossaries in this paper are adapted from this book.]

Manly B. F. J. (1994). *Multivariate Statistical Methods: A Primer*, 215 pp. London: Chapman and Hall. [This book is easy to read and gives a good introduction to multivariate analysis.]

Maurer D., Mengel M., Robertson G., Gerlinger T., and Lissner A. (1999). Statistical process control in sediment pollutant analysis. *Environmental Pollution* 104, 21–29. [This paper illustrates how Shewhart control charts are applied to identify long- and short-term trends and outliers of sediment cadmium concentration at an ocean outfall and reference station on the San Pedro Shelf, California.]

Mittag H.J. and Rinne H. (1993). *Statistical Methods of Quality Assurance*, 663 pp. London: Chapman & Hall. [This book gives a good introduction to a broad range of quality assurance procedures and provides a sound understanding of the basic statistical theory involved.]

Montgomery D.C., Johnson L.A., and Gardiner J.S. (1990). *Forecasting and Time Series Analysis*, 381 pp. New York: McGraw Hill. [This is a good reference book on time series models.]

Rheem S. and Holtzman G.I (1990). A SAS program for seasonal Kendall trend analysis of monthly water quality data. *SAS Users Group International (SUGI)* (Proceedings of the Sixteenth Annual Conference, New Orleans, LA, February 17–20, 1991), pp. 1193–1198. Cary, North Carolina: SAS Institute. [This paper gives an algorithm for the Seasonal Kendall test and slope estimator using SAS/IML.]

SAS. (2000). *SAS/ETS User's Guide, Version 8, Volumes 1 and 2*, 1532 pp. Cary, North Carolina: SAS Institute. [This document provides complete reference information on the SAS/ETS procedures used for time series modeling, simulating, and forecasting.]

SAS. (2000). *SAS/QC User's Guide, Version 8,* Volumes 1, 2 and 3, 1994 pp. Cary, North Carolina: SAS Institute. [This document provides complete reference information on the SAS/QC procedures used for statistical quality control and quality improvement.]

Sharma S. (1996). *Applied Multivariate Techniques*, 493 pp. New York: John Wiley & Sons. [This book gives a good introduction to the multivariate techniques and contains many SAS programs that can be used to analyze real data sets.]

Storey A., Briggs R., Jones H., and Russell R. (2000). Quality assurance. *Monitoring Bathing Waters: a Practical Guide to the Design and Implementation of Assessments and Monitoring Programs*, (ed. Jamie Bartram and Gareth Rees), Chapter 4. pp. 49–67. London, New York: E & FN Spon. [This article contains a detailed discussion on quality assurance, which is useful for any environmental monitoring program.]

Wei W.S. (1994). *Time Series Analysis*, 478 pp. Reading, Massachusetts: Addison-Wesley. [This book is useful for graduate and advanced undergraduate students who want more information in time series models.]

Biographical Sketch

Iris Yeung is currently an Associate Professor in the Department of Management Sciences at the City University of Hong Kong. She received a BSocSc (Hons) degree from the University of Hong Kong, a MSc degree from Imperial College, University of London, and a PhD degree from the University of Kent at Canterbury. She has published articles in the *Journal of Statistical Computation and Simulations*, *Statistica Sinica*, *Applied Statistics*, *Journal of Applied Statistical Science*, *Environmental Monitoring and*

Assessment, and *Marine Pollution Bulletin*. She has participated in the consultancy project of the Environmental Protection Department, Government of the Hong Kong Special Administrative Region. Her task in this consultancy project is to analyze water quality and sediment data for Hong Kong.