Data analysis

Data analysis is the component of the monitoring process that turns collected data into useful information. Analysing water/sediment quality monitoring data improves your understanding of the system being measured and drives management actions. The monitoring objectives will guide the data analysis, and they will also determine the study design, the quantity and type of data collected and sometimes the need for adequate computing power. Data analysis is quantitative and can be computationally intensive. Undertaking valid data analysis requires a good understanding of appropriate statistical methods and a strong appreciation of the context in which data were generated and the inferences that are required. In this lecture, we will understand the Water Quality Guidelines on the use of common statistical methods to analyse water/sediment quality data.

Data analysis process

In a typical data analysis process, your interpretation of the analysis outputs may refine the understanding of the system and lead to changes in monitoring design and the data that are collected in the future. Careful data preparation before any analyses should minimise the influence of anomalies or errors. Before starting a comprehensive analysis, you must enter, check and securely store the raw data. This usually involves computer database applications. The checking process will need to capture and flag data quality issues, such as missing data, detection limits and data entry errors. You should be able to retrace the steps taken back to the raw data. It is essential to perform exploratory analyses and interrogation of key variables using a variety of numerical, statistical and graphical methods. Examining the raw data can yield value, such as helping you to identify patterns for further scrutiny or raise questions for investigation. Useful modelling strategies for data analysis include model-based and probability-based analyses and inferences that may reflect how the data are collected, and the Bayesian and frequentist (classical) approaches to data analysis.

Water Data analysis

If quantifying status or change in water quality is a monitoring objective, then that may entail comparing sample data with guideline values. We describe how to derive guideline values using reference data, as well as significance testing and calculation of confidence or credible intervals to assess monitoring data against those guideline values when the sample data are spatially and temporally independent. If spatial or temporal independence is not likely, then this dependence may need to be accounted

for in the analysis. We discuss analytical options to assess temporal and spatial change, such as when a single water quality variable is considered at one site over time (temporal analysis approaches) or at multiple sites for a particular point in time (spatial and regional analysis approaches). We also introduce approaches for modelling relationships between multiple variables, including correlation analysis, multivariate and high-dimensional data analysis techniques, and regression analyses (both parametric and nonparametric approaches). After completing data analysis, interpretation and reporting of the key results and findings help to complete the cycle of meeting the set monitoring program objectives.

A useful checklist for water quality monitoring data analysis is presented in Box 1.

Box 1: Checklist for data analysis in a water quality monitoring program

1. Before commencing analysis, have you clearly identified:
    a. purpose of the data analysis exercise?
    b. parameters to be estimated or hypotheses to be tested?
    c. compatible data from different sources (levels of measurement, spatial scale, time scale)?
    d. objectives concerning quality and quantity of data?
    e. preferred methods for the statistical or data analysis?
    f. assumptions that need to be met for an appropriate application of those methods?
    g. data organisation and management considerations (storage media, layout, treatment of inconsistencies, outliers, missing observations and below detection limit data)?
2. Have data visualisation and summary methods (graphical, numerical, and tabular summaries) been applied?
3. Have data checks been performed and ‘aberrant’ observations (potential outliers) been identified?
4. Have statistical assumptions (e.g., non-normality, non-constant variance, autocorrelation) been checked?
5. Have data been suitably transformed, if necessary?
6. Have data been analysed using previously identified methods. Have alternative procedures been identified for data not amenable to particular techniques?
7. Have results of analysis been collated into a concise (statistical) summary. Have statistical diagnostics (e.g., residual checking) been used to support the appropriateness of the model approach?
8. Has the statistical output been carefully assessed and interpreted in the context of the objectives?

9. Have the objectives been addressed? If not, you may need to redesign the study, collect new or additional data, refine the conceptual models and re-analyse the data.

Planning for data analysis

Data types, quantities and methods of statistical analysis need to be considered collectively and at the early planning stages of any monitoring strategy. You must make study design decisions about measurement scales, frequency of data collection, level of replication and spatial and temporal coverage so that data of sufficient quality and quantity are collected for subsequent statistical analysis. It is important for your monitoring team to avoid the 'data rich–information poor' syndrome of collecting data that will not be subsequently analysed or that do not address the research analysis objective.

Given the costs associated with the data collection process, it is imperative for the monitoring team to use formal **quality assurance and quality control (QA/QC) procedures** to ensure the integrity of the data. These procedures should be supported by established back-up and archival processes. Seriously consider the archival medium to be used because rapid advances in computer technology tend to increase the speed at which equipment becomes obsolete.

QA/QC of any monitoring program should include:

- data analysis
- field sampling and measurement practices
- laboratory analysis.

Before statistically analysing the monitoring data, you should use standard methods of data summary, presentation and outlier checking to help identify 'aberrant' observations. If undetected, these data values can have profound effects on subsequent statistical analyses and can lead to incorrect conclusions and flawed decision-making. Develop a plan of the sequence of actions that the monitoring team will use for the statistical analysis of the water quality data. Only some of the many available statistical techniques need be used. An initial focus of monitoring might be to assess water quality against a guideline value or to detect trends. An ultimate objective of the data analysis exercise will probably be to increase your team's understanding of the natural system under investigation. Improved understanding should result in more informed decision-making, which in turn will lead to better environmental management. One of the most significant challenges for the data analysis phase is to extract a 'signal' from an inherently noisy environment.

We can categorise monitoring study designs as:

- **descriptive studies**, including audit monitoring
- **studies for the measurement of change**, including assessment of water quality against a guideline value (which can also be categorised as descriptive)
- **studies for system understanding**, including cause-and-effect studies and general investigations of environmental processes.

The statistical requirements for the descriptive studies are less complex than for the other 2 categories in which more detailed inferences are being sought.

Most of the statistical methods that we present here are based on classical tools of statistical inference (e.g., analysis of variance, *ANOVA*). These methods have served researchers from a variety of backgrounds and disciplines extremely well over many years but people have concerns about their utility for environmental sampling and assessment. When measuring natural ecosystems or processes, it is invariably hard to justify the assumptions that:

- response variables are normally distributed
- variance is constant in space and time
- observations are uncorrelated.

In these cases, remedial action (e.g., data transformations) may overcome some of the difficulties but it is more probable that an alternative statistical approach is needed. For example, generalised linear models (GLMs) are often more suitable than classical analysis of variance (ANOVA) techniques for the analysis of count data because of their inherent recognition and treatment of a non-normal response variable.

Statistical software for data analysis

So many statistical software tools are available and it is beyond the scope of the Water Quality Guidelines to review them all. There are packages available for this kind of analysis. These packages support different statistical techniques and different features, such as size of datasets handled, graphical representations, database interfaces, linkages to other software and the level of expertise needed to use them. Many software tools provide a high level of functionality and technical sophistication but they also lend themselves to abuse through blind application. It is important to scrutinise both the output and the choice of technique by asking yourself:

- Does the analysis make sense?
- Is it consistent with what has been observed in the exploratory data analysis?

If these questions are not routinely asked, then you run the risk that the 'mental models' of your monitoring team may have undue influence on the outcome. Results are more likely to be accepted — even if they occur for the wrong reasons — when they match the expectations of those who are interpreting the analysis.

Data preparation

Careful preparation of data for a water quality monitoring program will help ensure that:

- proposed analysis is feasible and runs smoothly
- valid results are obtained
- analytical results are not unduly influenced by anomalies or errors.

Data preparation should not be hurried. Often this is the most time-consuming aspect of data analysis. Check and clean electronic data before starting a comprehensive analysis. The data may need to be formatted, collated and manipulated. Make it possible to retrace your steps back to the raw data. Many numerical and graphical exploratory data analysis methods are useful to check and clean data for integrity. The choice of data formats may depend on how the data will be analysed and the software available. It's best to enter laboratory and field study data into a database or spreadsheet that is accessible for many software applications.

Only use data that are:

- deemed acceptable by field or laboratory quality assurance/quality control (QA/QC) criteria
- treated consistently (e.g., rounded off to an appropriate number of significant figures).

Data integrity

The integrity of water quality data can be reduced in many ways. Losses or errors can occur at any time, from sample collection and preparation through to interpretation and reporting. After the QA/QC checked data leave the field or laboratory, there is ample opportunity for accidental alteration of results to occur. Gross errors that are probably the result of data manipulations (e.g., transcribing, transposing rows and columns, editing, recoding and conversion of units) are easily overlooked unless you perform a modest level of screening. These sorts of errors can usually be detected by a scan of the raw data. Subtle erroneous effects (e.g., repeated data, accidental deletion of 1 or 2 observations, mixed scales) are more difficult to identify. If left unchecked,

anomalous readings can have a profound effect on subsequent statistical analyses and possibly lead to erroneous conclusions being drawn.

Quality control measures

A good QA/QC program for data analysis uses simple yet effective statistical tools for screening data as they are received from the field or laboratories. Data screen measures typically include a mixture of:

- graphical procedures — histograms, box plots, time sequence plots, control charts
- descriptive numerical measures of key distributional aspects — mean, standard deviation (SD), coefficient of variation (CV), measures of skewness and kurtosis.

These routine analyses often provide the first opportunity to inspect data, recognise patterns and draw preliminary understanding. Most pre-processing, manipulation and screening can be undertaken using in-built database or spreadsheet functionality.

Missing data

Missing observations are common in environmental datasets, and may occur due to:

- dropout of sites from a study
- equipment failure
- time or resource constraints
- fault of the observer.

In monitoring studies, missing data can reduce the representativeness of the sample and bias the results, which may lead to misleading inferences about the population. You should not ignore omitted observations because they can have a significant effect on the conclusions that can be drawn from data. Exploring the missing data to uncover how they were omitted can help you to proceed correctly with analysis. Rubin (1976) differentiated between 3 types of ‘missingness’ mechanisms that will have implications for your approach to analysis and the conclusions drawn from it:

- missing completely at random — the value of the missing variable is independent of both observed variables and unobserved parameters
- missing at random — missingness does not depend on some variables
- missing not at random — missingness is structured; the value of the missing variable is related to the reason it is missing.

In the presence of missing data, we recommend robust analysis methods to minimise the effects of bias on the overall conclusions. Alternative techniques for dealing with missing data include:

- data imputation or ‘filling-in’, after which standard data analysis approaches could be used
- Bayesian approaches to some popular methods, which can allow missing values to be treated as an additional parameter and estimated (or imputed)
- data reduction, which can remove those records or cases where missing data are present
- maximum likelihood estimation
- spatial modelling
- data interpolation.

Understanding how the missing data arise will help you to select an appropriate data imputation approach.

Reference

Inference and missing data by Donald B. Rubin

Aas, W. and A. Semb. 2001. Standardisation of Methods for Long-term monitoring. Water, Air, and Soil Pollution 130:1595-1600.

Hounslow, A. (1995). Water Quality Data: Analysis and Interpretation (1st ed.). CRC Press. https://doi.org/10.1201/9780203734117

Techniques of trend analysis for monthly water quality data

Robert M. Hirsch, James R. Slack, Richard A. Smith

World Bank Group. 2007. Access to Safe Water Map. Available at: http://www.worldbank.org/depweb/english/modules/environm/water/map1.html

Wilde, F.D., Radtke, D.B., Gibs, Jacob, and R.T. Iwatsubo. 1998. Preparations for Water Sampling. In National Field Manual for the Collection of Water-Quality Data: U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chap. A1, 42 p. (http://water.usgs.gov/owq/FieldManual/chapter1/html/Ch1_contents.html)

Wilde, F.D., Radtke, D.B., Gibs, Jacob, and R.T. Iwatsubo. 1999. Collection of Water Samples. In National Field Manual for the Collection of Water-Quality Data: U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chap. A4, 151 p. (http://water.usgs.gov/owq/FieldManual/chapter4/html/Ch4_contents.html)

Wiley, M.J., P.W. Seelbach, K. Wehrly, and J.S. Martin. 2003. Regional ecological normalization using linear models: a meta- method for scaling stream assessment indicators. In: T. P. Simon (ed) Biological Response signatures, pp.201-224, CRC Press, Boca Raton, FL.