Data analysis and graphs

Exploratory data analysis

How you analyse the data will largely depend on the type of monitoring study being undertaken and the nature of the arising data (number of variables, continuous or discrete data classes, temporal and spatial attributes). Suggested analysis methods will largely have been discussed and noted as part of the study design process, based on the purpose for the analysis or your analysis profile. After preparing your data, we strongly encourage exploratory analyses and interrogation of key variables using numerical and graphical statistical methods. Simply looking at the data can yield value, identify patterns for further scrutiny or raise questions for investigation. You can use a range of analysis strategies, including model-based and probability-based analyses and inferences depending on how the data are collected, or the broader Bayesian and frequentist modelling philosophies and methods. If quantifying status or change in water quality is a monitoring objective, then comparison of sample data against guideline values will be of interest. Discuss how to derive guideline values using reference data, followed by details of significance testing and calculation of confidence intervals for testing 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 needs to be accounted for in the analysis. As an analysis project, a discussion of analysis options when a single water quality variable is considered, needs to be understood. Always revise one site over time (temporal analysis approaches) or at multiple sites for a particular point in time (spatial and regional analysis approaches). Also consider some approaches to model the relationships between multiple water quality or environmental variables, including correlation analysis, multivariate and high-dimensional data analysis techniques, and regression analyses (parametric and nonparametric approaches).

Data summary

Data summary enables you to present and summarise important features of the data. It also helps to identify anomalous observations (also known as ‘outliers’) that might be attributable to experimental or transcription errors, and to ascertain the validity of outliers and their subsequent inclusion in the analysis.

You can use a variety of numerical and graphical statistical tools to summarise data, including:

- graphs (e.g., histograms, box plots, dot plots, scatterplots)

- tables (e.g., frequency distributions, cross-tabulations)
- numerical measures (e.g., means, medians, SDs, percentiles).

Changes in water quality over time and space

Spatial or temporal components are often associated with data collection — especially in water quality monitoring studies. The larger the study region, the greater the chance that spatial dependence could occur at multiple spatial resolutions. Likewise, the longer the study, the greater the chance that temporal dependence may exhibit multiple temporal resolutions. These attributes are critical to be aware of, explore and factor into an analysis so you can make valid inferences. Here we consider analysis options when a single water quality variable is measured at one site over time or at multiple sites for a particular point in time. Spatial and temporal attributes of data are often non-separable so spatiotemporal analyses are more appropriate. In the case of multiple water quality variables measured, the appropriate analyses are even more complex, with a multivariate response being required (refer to Multivariate and high-dimensional data for a discussion of multivariate analyses).

Spatiotemporal modelling is an increasingly expanding area of statistical research with much traction due to the increased and varied spatial and temporal scales of data collection, particularly in the environmental domain. Discussion of these approaches is beyond the scope of the Water Quality Guidelines. The first step in an analysis of spatially or temporally referenced data is to establish if there is a trend. If there is a trend, then the challenge is to ascertain why there is a trend, quantify the specific nature of the trend, and ensure this relates back to the study objectives. The choice of analysis approach or model complexity needs to suit the data density and attributes. Some methods may not be appropriate for small sample sizes or for high frequency data.

Using temporal analysis approaches

One of the principal objectives of water quality monitoring is to assess changes over time. In many instances, this is driven by the need to compare water quality against a guideline value, but there is also an important need to track changes over time in response to a management intervention or other drivers, such as land use, and identify potential trajectories or trends in water quality. Here we consider approaches to the temporal analysis of water quality monitoring data under 3 subsections:

- water quality trend analysis
- time series analysis
- analysis of high frequency monitoring data.

The dependence of water quality drivers, such as flow and climate variability, appeal to related methods but are typically focused on understanding the drivers behind any of the temporal changes identified.

Water quality trend analysis

Although trend analysis is a popular tool in many fields, including environmental science, it is not a specific research topic in statistics. Instead, many statistical methods can be applied to extract an underlying pattern or behaviour by treating time as a variable. The exception is time series analysis, which was developed to deal with data collected over time (correlated data). Formal statistical analysis of temporal data can be rather complex and requires a good degree of skill in identifying suitable models. The objectives of a water quality trend analysis are often about trend detection or trend estimation.

The focus of **trend detection** is on determining whether or not there has been a departure from the background or historical conditions. This change is usually viewed as monotonic. Our interest is in establishing whether the condition is improving or deteriorating. Hypothesis testing may be used to determine if this change is statistically significant. **Trend estimation** is distinct from trend detection in that the objective is to quantify the nature of the change and investigate models that provide the interpretation of the process possibly causing them. Trend estimation is typically of interest for longer periods of data and needs to allow the possibility of nonlinear trends.

Presenting the Data

When presenting numerical data, one of your chief goals should be to maintain the attention and interest of your audience. This is very difficult using tables filled with numbers. Most people will not be interested in the absolute values of each parameter at each sampling site. Rather, they will want to know the bottom line for each site (e.g., is it good or bad) and seasonal and year to year trends. ***Graphs and charts***, therefore, are typically the best way to present volunteer data. Take care, however, that your graphs "fit" your audience and are neither too technical nor too simplistic.

Graphs and Charts

Graphs can be used to display the summarized results of large data sets and to simplify complicated issues and findings. The three basic types of graphs that are typically used to present volunteer monitoring data are:

- Bar graph

- Line graph
- Pie chart

Bar and line graphs are typically used to show results, such as bioassessment scores, along a vertical or y axis for a corresponding variable (such as sampling date or site) which is marked along the horizontal or x axis. These types of graphs can also have two vertical axes, one on each side, with two sets of results shown in relation to each other and to the variable along the x axis.

Bar Graph

A bar graph uses columns with heights that represent the value of the data point for the parameter being plotted.

Line Graph

A line graph is constructed by connecting the data points with a line. It can be effectively used for depicting changes over time or space. This type of graph places more emphasis on trends and the relationship among data points and less emphasis on any particular data point.

Pie Chart

Pie charts are used to compare categories within the data set to the whole. The proportion of each category is represented by the size of the wedge. Pie charts are popular due to their simplicity and clarity.

Graphing Tips

Regardless of which graphic style you choose, follow these rules to ensure you use them most effectively.

- *Each graph should have a clear purpose.* The graph should be easy to interpret and should relate directly to the content of the text of a document or the script of a presentation.
- *The data points on a graph should be proportional to the actual values so as not to distort the meaning of the graph.* Labelling should be clear and accurate and the data values should be easily interpreted from the scales. Do not overcrowd the points or values along the axes. If there is a possibility of misinterpretation, accompany the graph with a table of the data.
- *Keep it simple.* The more complex the graph, the greater the possibility for misinterpretation.

- *Limit the number of elements.* Pie charts should be limited to five or six wedges, the bars in a bar graph should fit easily, and the lines in a line graph should be limited to three or less.
- *Consider the proportions of the graph and expand the elements to fill the dimensions, thereby creating a balanced effect.* Often, a horizontal format is more visually appealing and makes labelling easier. Try not to use abbreviations that are not obvious to someone who is unfamiliar with the program.
- *Create titles that are simple, yet adequately describe the information portrayed in the graph.*
- *Use a legend if one is necessary to describe the categories within the graph.* Accompanying captions may also be needed to provide an adequate description of the elements.

Summary Statistics

Summary statistics can reduce a very large data set to a few numerical values that can then be easily described and analysed. Such statistics include the mean and standard deviation--two of the most frequently used descriptors of environmental data. Textbook statistics commonly assume that if a parameter is measured many times under the same conditions, then the measurement values will be randomly distributed around the average with more values clustering near the average than further away. In this ideal situation, a graph of the frequency of each measure plotted against its magnitude should yield a bell-shaped or normal curve. The *mean and the standard deviation* determine the height and breadth of this curve, respectively.

The mean is simply the sum of all the measurement values divided by the number of measurements. This statistic is a measure of location and in a normal curve marks the highest point at the centre of the bell. The standard deviation, on the other hand, describes the variability of the data points around the mean. Very similar measurement values will have a small standard deviation while widely scattered data will have a much larger standard deviation. While both the mean and standard deviation are quite useful in describing stream data, often the actual measures do not fit a normal distribution. Other statistics often come into play to describe the data. Some data are skewed in one direction or the other. Other data may have a flattened bell shape. It is important to note that biological information often does not follow normal, bell-shaped distribution. This is because biological communities are dynamic, complex, and interdependent systems; many factors influence them, and these cannot be statistically predicted. For describing non-normally distributed data, it is best to use statistics that can convey the information for a variety of conditions and which are not overly influenced by the data points at the extremes of the distribution.

The median and the interquartile range are two statistics that are commonly used to describe the central tendency and the spread around the median, respectively. These statistics are derived by placing the data points in order of value from lowest to highest. The median is simply the value that is in the middle of the data set. The interquartile range is the difference between the value at the 75 percent level and the value at the 25 percent level. The best method for presenting this type of data is called a box and whisker plot. One simple box and whisker plot will graphically display the following information:

- Median
- Variability of the data around the median
- Skew of the data
- Range of the data
- Size of the data set

Statistical software packages for computers will easily construct box and whisker plots. You can construct these plots by following procedure shown below:

1. Order the data from the lowest to the highest.
2. Plot the lowest and highest values on the graph as short horizontal lines. These are the extreme values of the data set and represent the data range.
3. Determine the 75 percent value and the 25 percent value of the data set. These values define the interquartile range and are represented by the location of the top and bottom lines of the box.
4. The horizontal length of the lines that define the top and bottom lines of the box (the box width) can be used as a relative indication of the size of the data set. For example, the box width that describes a data set of 20 values can be displayed twice as wide as a data set of 10 values. Any proportional scheme can be used as long as it is consistently applied.
5. Close the box by drawing vertical lines that connect to the ends of the horizontal lines.
6. Plot the median inside the box.

Maps

Displaying the results of your monitoring data on a map can be a very effective way of showing the data and helping people understand what it means. A map shows the location of sample sites in relation to land features, such as cities, wastewater treatment plants, farmland, and tributaries that may have an effect on water quality. Because a map also displays the stream's relationship to neighbourhoods, parks and

recreational areas, it can help to develop concern for the stream and strengthens interest in protecting it.

Choosing a Map

It is best to have two types of maps. One should be a working map with a lot of detail. The other should be used for display purposes. The working map should include important features such as:

- Stream and its tributaries
- Wetlands
- Lakes and ponds
- Cultural features such as roads
- Rail and power lines; municipal boundaries
- Some indication of land use patterns and vegetation.

The display map is best used to illustrate your program results at public meetings or in reports. This map should be simpler than the detailed map and show only principal features such as roads, municipal boundaries, and waterways. It should have sufficient detail and scale to show the location of sample sites, and have space for summary information about each of the sample sites.

Creating a Display Map

Some suggestions for using a map to display your data include:

- Keep the amount of information presented on each map to a minimum. Do not try to put so much on one map that it becomes visually complicated and difficult to read or understand. Use another map to display a different layer or "view" of the data. For example, if there are several dates for which you wish to display sampling results, use one map for each date.
- Clearly label the map and provide an explanation of how to interpret it. If you need a long and complicated explanation, you may want to present the data differently. If you have reached a clear conclusion, state the conclusion on the map. For example, if a map shows that tributaries are cleaner than the mainstem, use that information as the subtitle of the map.
- Provide a key to the symbols that are used on the map.
- Rather than packing lots of information into a small area of the map, enlargement of the area elsewhere on the map to adequately display the information will be helpful.
- Use symbols that vary in size and pattern to represent the magnitude of results.

Reference

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.