ISSN 1757- 5958

amc technical briefs

Editor: Michael Thompson Analytical Methods Committee AMCTB No 47 March 2010

How continuous should 'continuous' monitoring be?

Recent developments in measurement technology have raised the possibility of greatly increased frequencies of monitoring, at least for a limited number of determinands. Attractive as this prospect might be for gaining a better understanding of the processes under examination, monitoring at high frequencies raises a new question—at what point is the effort devoted to obtaining more data no longer justified by the increase in knowledge that it provides?

This Technical Brief discusses the reason for diminishing returns as the frequency of monitoring is increased, and the ways in which to determine the best frequency of monitoring—one that provides a balance between uncertainty arising from not collecting enough data and the information overload caused by collecting too much. The discussion below concentrates on time series data and the issue of optimising the frequency of observations of a temporal process. However, similar considerations would apply to spatial data.

Statistical Independence and Autocorrelation

One of the important (but often incorrect) assumptions about analytical data is that each measurement is independent of the previous one. However, in many cases data are not completely random, each data point to some extent being influenced by the points either side. In a time series (where any influence is necessarily from past to future) the way in which any given observation is related to the previous one is referred to as autocorrelation.

Importance of statistical independence (randomness)

The issue of independence is critically important in relation to the interpretation of data because many statistical tests depend on its assumption. The validity of many statistical inferences is directly linked to the validity of this assumption. Many commonly-used statistical formulae depend on the independence assumption, the best example being the formula for determining the standard error (se) of the sample mean, namely $\text{se}(\bar{x}) = s/\sqrt{n}$ where s is the standard deviation of the data and n is the number of observations. This formula is applied in the determination of the uncertainty associated with a mean value. It is not universally appreciated that such estimates of uncertainty can be seriously misleading if the assumption of independence does not hold. Furthermore, in data interpretation, two of the principal aims are to determine the value of the measured quantity (how much of it there is) and its variability (how much it fluctuates) and how easy it would be to detect an underlying change. This translates into the need to determine the mean and its standard error. If data are autocorrelated, and frequent observations are made in the shorter term, the standard error can be seriously underestimated, with respect to that applying to data unaffected by correlation.

When the possible sources of non-randomness are considered, it is convenient to divide them into two types: longer–term systematic changes such as trends and seasonal effects; and serial correlation – any autocorrelation that remains once trends and seasonality have been accounted for. The utility of this division is that systematic factors are often the things that we are interested in and serial correlation can lead to confusion in our estimates of what we might think we have found.

Data Redundancy

From the above discussion it can be seen that serial correlation in environmental variables can have important consequences for the interpretation of measurements – with mean values being assigned narrower confidence intervals than are truly justified and consequently inaccurate (false positive) claims concerning either the detection of trends or the statistical significance of observations.

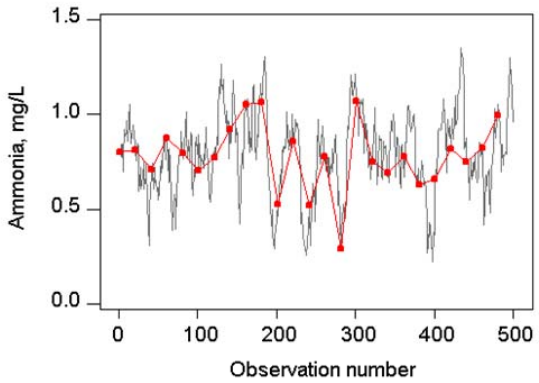

Figure 1. 'All data' (500 observations, grey line) and 'independent data' (25 observations, red points and line).

Correlation can be important in a second way because it leads to redundancy in the collection of data – obtaining data for a time-correlated variable that are too close to one another is really equivalent to a repeat measurement of the same thing and as such adds little additional information. The question this raises is that, for a given level of serial autocorrelation, what is an appropriate interval between samples or optimum frequency of samples? It is clear that in qualitative terms sampling at a lower frequency than the optimum leads to loss of information, whereas sampling at a higher frequency produces data that have no real information content and which can pose problems of data analysis, processing and storage.

To make quantitative sense of this issue we need to determine the effective number of independent observations n_e in an autocorrelated time series. This can be estimated[1] as $n_e = n(1-r)/(1+r)$, where n is the actual number of observations and r is the so-called "lag-one" autocorrelation coefficient. This is the correlation between one result and the next one; r can have any value between 0 and 1(for positive correlation). There are other lag correlations – for example the lag-3 correlation corresponds to the correlation between a result and next result but two. In simple instances, lag-k correlations decrease as r^k as k increases. So if the lag-1 r is 0.5 then the lag-2 value would be 0.25. This expresses the way in which results further down the series of values have less and less similarity to the initial one.

An example

An example will illustrate what this means. Suppose a water company sets up a continuous monitor that can provide measurements of ammonia concentration in a wastewater treatment works effluent at hourly intervals and that the aim is to establish the overall mean effluent concentration and its uncertainty. Monitoring over a two month period provides 500 data points. The table below shows summary statistics for the whole data set. Calculation of the lag-1 correlation coefficient to estimate n_e shows that out of the 500 observations the effective number of independent values is only 25.

	All data - not accounting for autocorrelation	Using only independent data (sampling at a rate of approximately once per day)
Mean	0.80	0.80
Standard deviation	0.21	0.21
Standard error	0.009	0.040
Half width of 90% confidence interval	0.015	0.069

In other words the optimum sampling frequency is just greater than once per day and approximately 20 times too much data had been collected. The second column of figures in the table shows the correct interpretation of the data – in which the confidence interval for the mean value is more than four times larger than originally (and erroneously) estimated. Missing out every 19 data points effectively reduces the lag correlation from the original value of 0.9 to less than 0.15.

Conclusions

If data are serially correlated, there is a diminishing return from increasing the frequency of sampling (or measurement) within a selected time period. It is worth assessing serial correlation as a means of determining the maximum useful frequency of monitoring. Special attention to the possibility of serial correlation is required in calculating confidence intervals for average values and correctly assessing the statistical significance of trends.

[The CORREL function in Excel can be used to provide an estimate of r. If time series data are in cells A1 to (say) A30 the command =CORREL(A1:A29,A2:A30) provides an estimate of the lag-1 r value.]

References and further reading

1. Ellis J.C., (1989) Handbook on the design and interpretation of monitoring programmes. WRc report NS29. Medmenham: WRc. ISBN 0 902156 72 1 35..
2. J C Loftis, G B McBride and J C Ellis. *Considerations of scale in water quality monitoring data and data analysis.* Water Resources Bulletin,1991, **27**, 255-264.
3. P Muskens and G Kateman. *Sampling of internally correlated lots.* Anal. Chim. Acta 1978, **103** 1-9.

This Technical Brief was drafted for the Analytical Methods Committee by M Gardner (Atkins Ltd) on behalf of the Statistical Subcommittee (Chair M Thompson).

You can help the AMC in the production of Technical Briefs by posting queries, comments and suggestions on the AMC Discussions Web-board. See www.rsc.org/amc.

CPD Certification I certify that I have studied this document as a contribution to Continuing Professional Development.

Name…….....
Signature…….…….Date…….

Name of supervisor……..
Signature…….…….Date……