Journal of Paleolimnology 6: 103–110, 1991.
© 1991 *Kluwer Academic Publishers. Printed in Belgium.*

Comments on Boyle's Acidity and organic carbon in lake water: variability and estimation of means

Pierre Legendre & Pierre Dutilleul
Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale A, Montréal, Québec, Canada H3C 3J7

Received 26 April 1991; accepted 13 July 1991

Introduction

At first glance, Boyle's (1991) paper looked simple and straightforward, and should not have generated the debate that it has between Dr. Boyle and us. The author addresses the following problem: is it better to use the original measurements, or log-transformed data, when one wants to estimate the mean of water acidity and organic content? It seemed to us, at first, that answers to this question could be found in any basic textbook of statistics and so, perhaps, a paper on the subject was not even warranted. It turns out, however, that while part of the answer is found indeed in most textbooks of statistics, another part is found almost nowhere – and in particular NOT in Dr. Boyle's paper – because of the very nature of the data subjected to analysis. So, we felt that it was appropriate to expound on the underlying statistical assumptions and procedures of the computations presented by Dr. Boyle, and on their effects on the very answer that he offers to this palaeolimnologically important question.

Our comments are not specifically directed at Dr. Boyle's work; this has been but a good opportunity to make these results known to the community of ecologists who often have to estimate parameters for variables measured over surfaces, or forming time series. We will discuss in particular: (1) the choice of the best normalizing transformation, and the generality (or lack) of the answer presented by Dr. Boyle; and (2) the correct estimation of the confidence interval of a mean when the data are autocorrelated.

Choice of the best normalizing transformation

A chi-square test, such as used by Boyle, is not the best method to test for normality, since it does not take into account the ordered nature of the data. With small sample sizes, the validity of the chi-square test is questionable; it is in any case less powerful than a Kolmogorov-Smirnov or a Shapiro-Wilk test for any sample size when the distribution under study is continuous and completely specified (David & Johnson, 1948; Massey, 1951; Stephens, 1974). The original Kolmogorov-Smirnov (*K-S*) test of normality, however, should not be used when the mean and variance are estimated from the very data that are subjected to testing. The original *K-S* table assumes that the mean and variance are known *a priori*. Using it when the mean and variance are estimated from the data leads to conservative results in the sense that too many distributions are found to be normal, because the effective probability of a type I error (declaring the distribution not normal when it actually is) is overestimated in the table. The danger with such a conservative test is its lack of power, that is characterized by failing to reject the null hypothesis when it should be rejected; remember that in this case, H_0 states that the population distribution is normal. Lilliefors (1967) has proposed a table of corrected values, that takes into account the estimation of unknown parameters from the sample. Boyle, however, does not tell us whether he used that corrected table, or not. Among the commonly used statistical packages, SPSS, for instance, does not make that correction, while SAS does. Finally, the W statistic proposed by Shapiro & Wilk (1965) for testing normality has good power properties against a wide range of alternative distributions, when the data are independent from one another and represent a random sample (Royston, 1982). For sample sizes up to 2000, SAS Version 6 computes the W statistic. For larger sample sizes, the K-S statistic is linearly interpolated within the range of simulated critical values given by Stephens (1974); see the SAS Institute Inc. (1985) manual. Both of these tests of normality, the K-S as well as the W statistics, suffer from a lack of robustness against autocorrelation in the sample data (Dutilleul & Legendre, submitted). With autocorrelated data, we recommend to carry out the tests of normality after prewhitening, *i.e.*, after adjusting for autocorrelation, instead of using the raw data; procedures for detecting and modeling temporal autocorrelation are given in Box & Jenkins (1976).

In our view, a major problem in Boyle's methodological section is that the author is not looking for the best normalizing transformation of the data; he simply compares the original metric to the log transformation. What if some other transformation turned out to be more adequate? Actually, there are reasons to believe – and we provide examples below – that with some data sets for the same type of variables (acidity, organic carbon content), some other transformations may well be more appropriate: for instance, data from a survey covering a wide diversity of lakes, or time series obtained at another time scale. There are standard methods for finding the best normalizing transformation for a data set, and these are described in basic statistics texts.

If people rely so heavily on the log transformation, it is because it is widely available in computer packages, not because it is necessarily the best – except for biological variables (species abundances), where there are reasons in the ecological theory to believe that the process of population growth is exponential in some cases. For physical and chemical variables, the best normalizing transformation depends on the variable under study and on the sampling scale. For instance, in a study involving one of us (Legendre & Troussellier, 1988), variables POC (particulate organic carbon) and DOC (dissolved organic carbon) were not normal after log transformation, but requested transformations by the exponents −0.0918 and −3.78401 respectively, using the Box-Cox method (below). The best normalizing transformation may also depend on the spatial scale (a single lake, or several lakes in the same region, or else lakes covering several geological regions) or the temporal scale (a few years, or geological times) of the study.

Here are a few more examples. The James Bay area in northwestern Québec (about 54°N) is climatically and geologically somewhat similar to the region of Norway where Boyle's data come from. Schetagne & Roy (1985) report on pH data collected in reservoirs and rivers. Some of these data are reanalyzed here for normality and for the best normalizing transformation. No attempt was made to select data that fitted our argument. First, the data were back-transformed to the H^+ scale, noted 10^{pH}. From the back-transformed data, the best normalizing transformation was found using the Box-Cox method (Box & Cox, 1964; see also Sokal & Rohlf, 1981); the method allows to empirically determine the most appropriate exponent λ for the transformation $y' = (y^\lambda - 1)/\lambda$ when $\lambda \neq 0$, while the log transformation is used when $\lambda = 0$; the method works quite well for unimodal data distributions. At each step, the data were tested for normality using the Kolmogorov-Smirnov test, including the Lilliefors correction described above. The results are reported in Table 1.

Notice for instance that while the Eastmain-Opinaca confluent pH data nicely fit a normal distribution ($p > 0.20$), such is not the case with the Delorme or the Eastmain station data for instance ($p < 0.05$). Comparing the pH scale to the H^+ concentration scale (noted 10^{pH}), pH is

Table 1. Tests of normality for five pH data sets, collected between June 1980 and November 1981 (L stations) and between February 1981 and October 1982 (R stations); data from Schetagne & Roy (1985). The best result (closer to normality) is identified by a star (*).

Reservoir (L) or river (R) station	Number of samples	Range of pH values	pH *K-S* prob.	10^{pH} *K-S* prob.	Box-Cox λ for 10^{pH}	Box-Cox-transf. *K-S* prob.
Caniapiscau (L)[1]	18	5.7–6.4	$0.10 > p$[6]> 0.05	$0.01 > p$	− 0.091	$0.10 > p$[6]> 0.05*
Delorme (L)[2]	20	5.6–6.5	$0.05 > p > 0.01$	$0.15 > p > 0.10$	0.402	$p > 0.20$*
Vermeulle (L)[3]	19	5.4–6.5	$0.10 > p > 0.05$	$p > 0.20$*	0.440	$0.10 > p > 0.05$
Eastmain-Opinaca (R)[4]	27	5.9–7.2	$p > 0.20$	$0.01 > p$	0.170	$p >> 0.20$*
Eastmain (R)[5]	26	5.7–7.0	$0.01 > p$	$0.01 >> p$	− 0.239	$0.05 > p > 0.01$*

[1] Multi-modal pH distribution, relatively flat.

[2] pH distribution skewed to the left.

[3] Two modes present; this is why Box-Cox does not find a better normalization than the 10^{pH} scale.

[4] Well-balanced pH distribution.

[5] Several modes in the pH distribution; somewhat skewed to the right.

[6] For the Caniapiscau data, the probability after Box-Cox transformation is closer to 10% than in the pH metric.

better – in the sense that the distribution is closer to normality – in three instances, while the 10^{pH} scale is better in two instances. Notice also that in all instances, it is possible to find a better transformation than the pH scale.

Our conclusion is that Boyle's statement, to the effect that the pH scale (i.e., log) is the very best scale for acidity data in all cases, is an overgeneralization. Each particular frequency distribution should be studied on its merits, the result depending to a large extent on the sampling scale.

Correct estimation of the confidence interval of a mean for autocorrelated data

One of the problems often encountered with ecological data is that they are almost always autocorrelated (Legendre, 1991). We give an illustration in Fig. 1 for pH data collected in Lake Mývatn (Sweden) and published by Ólafsson (1979); the significant and positive autocorrelations at the first distance classes are characteristic of a first-order autoregressive process. There is a large body of statistical literature, dealing with time series and spatial analysis, where explanations of the autocorrelation phenomenon can be found; see for instance Box & Jenkins (1976), Cliff & Ord (1981), or Legendre & Fortin (1989). In presence of positive autocorrelation – as it is the case in Fig. 1 and certainly also in Boyle's data, see below – the confidence intervals of parameters, such as the mean, that one can compute using the ordinary formulas is too narrow. The reason for this is that non-independent data do not provide one full degree of freedom each, so that the value *n* used in the formula is too large. Correcting *n*, as well as the estimate of the variance, is not dealt with yet in introductory textbooks of statistics because it is no simple task (Dutilleul & Legendre, 1991). A basic reference in the geostatistical literature is Isaaks & Srivastava (1989); in time series analysis, see Anderson (1971). In the meantime, approximate confidence intervals can probably be computed using *ad hoc* methods such as the Jackknife procedure, but even these alternatives have problems of their own.

We do not agree when Boyle writes 'Given that the effect of the autocorrelation, even where present, is so small that the results of this work are not affected, and that classical statistical methods are simpler and more robust than the other more complex methods (de Gruijter & ter Braak, 1990), the use of conventional standard deviations is justified'. First, the reference to the de Gruijter & ter Braak's paper is unjustified because taken out of its own context. Secondly, Dr. Boyle announces an underestimation of the variance that he estimates to be close to 1% (we

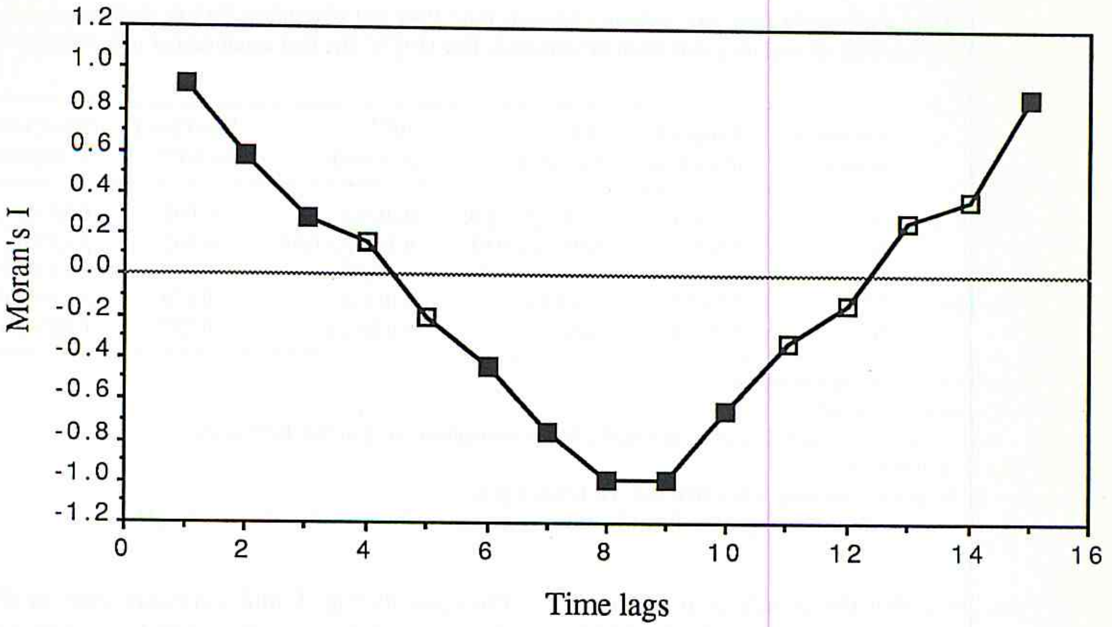

Fig. 1. Moran's *I* correlogram computed for pH data collected in Lake Mývatn (Sweden) and published by Ólafsson (1979). Open squares, nonsignificant autocorrelation; dark squares, significant autocorrelation at level $\alpha = 0.05$.

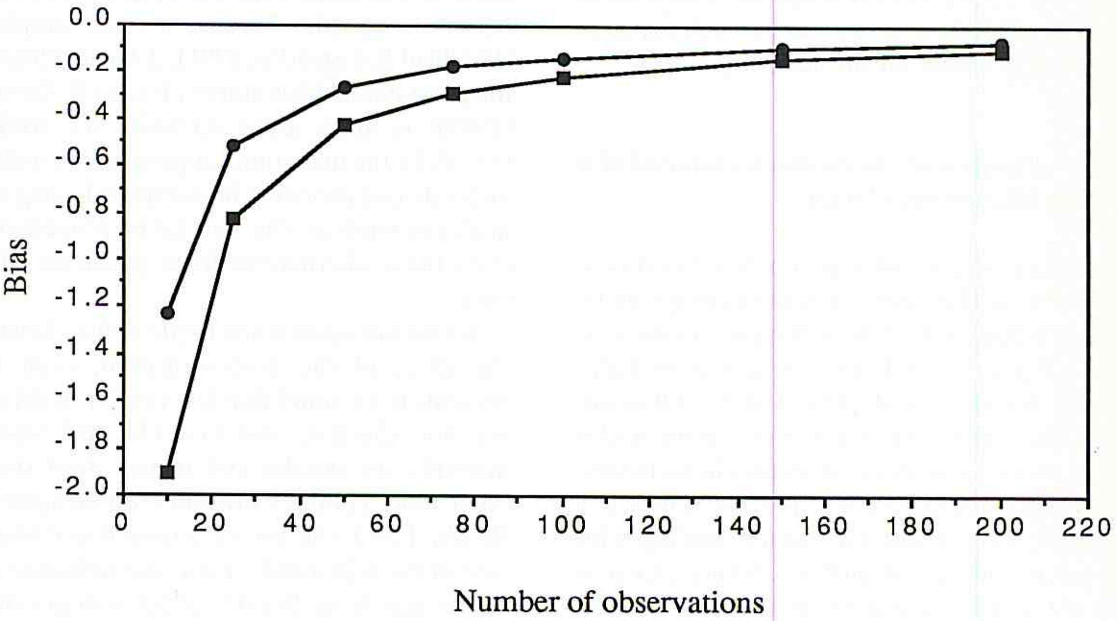

Fig. 2. Estimation of the population variance σ^2 of a first-order autoregressive process (AR(1)) with parameter ρ. Evolution of the bias of the classical sample variance S^2 in relation to the number of observations for parameter values $\sigma^2 = 10$, and $\rho = 0.4$ (circles) and $\rho = 0.52$ (squares), values presented by Boyle as first-order autocorrelation coefficients obtained with his data.

don't know the basis of this estimation). However, it is the square root of the variance, (*i.e.*, the standard deviation), and not the variance itself, which is used for computing confidence limits, that are used in turn to estimate the width of the confidence interval. The amount of underestimation of the width of the confidence interval is actually not that announced by Boyle. In conclusion, it should be obvious at this point that since the confidence intervals computed by Boyle are too narrow, then the estimated number of samples necessary to obtain a predetermined precision is affected. This factor is hard to compute because the bias due to autocorrelation in the estimation of the standard deviation is not the square root of the bias for the estimated variance (see below); this bias is not at all negligible. The effect of temporal autocorrelation on the classical sample variance is investigated below to determine whether the formula in Boyle's Results section can be directly applied to estimate the confidence interval and, conversely, to determine the minimum number of samples for a given level of precision.

Variations of pH along time can be modeled as a first-order autoregressive process, called AR(1), and characterized by an autocorrelation pattern as illustrated in Fig. 1. In such a process, for a given population variance σ^2 and an autocorrelation structure with parameter ρ, the expected value of the classical sample variance S^2 of a data series varies with the number of observations n. It can be computed as follows under the normal model $\mathrm{N}_n(\mathbf{m}, \sigma^2 \boldsymbol{\Sigma}_\rho)$ with $\mathbf{m} = m\,(1, \ldots, 1)'$, where m is the overall mean parameter and $\boldsymbol{\Sigma}_\rho$ the correlation matrix among observations along time:

$$\mathrm{E}(S^2) = \frac{n}{n-1}\,\sigma^2 \left\{ 1 - \frac{1}{n^2(n-1)}\,\mathrm{tr}(\mathbf{A}\,\boldsymbol{\Sigma}_\rho) \right\}$$

where tr denotes the trace operator in matrix algebra and $\mathbf{A}$ is a $(n \times n)$ matrix of ones (Dutilleul & Legendre, 1991). The validity of the above expression does not require normality of the population distribution.

To give an illustration, if we set the first-order autocorrelation coefficient ρ to 0.4 and to 0.52 (values presented by Boyle), and the population variance σ^2 to 10, the bias, measured by $[\mathrm{E}(S^2) - \sigma^2]$ and computed for various n values (10, 25, 50, 75, 100, 150, 200), is as plotted in Fig. 2. Since Boyle's n values go from 6 to 221, they fall quite in the range covered by our results. For $\rho = 0.40$, using 25 (50, 100) data points, the bias value is equal to -0.5185 (-0.2630, -0.1324), or 5.2% (2.6%, 1.3%). For $\rho = 0.52$ the problem is more serious: using 25 (50, 100) data points, the bias value is -0.8275 (-0.4237, -0.2143), or 8.3% (4.2%, 2.1%). In both cases, the bias tends to vanish for high sample sizes.

Unfortunately, Boyle does not present his data (sampling dates and observed pH values) in such a way that we could have used them for further computations. In order to justify his procedure, he presents first-order autocorrelation coefficients. He claims that most of his data are not significantly autocorrelated, although the only two autocorrelation values reported are $\rho = 0.52$ and 0.40, which do not look negligible to us. The only alternative would be to work out a correction formula for his estimated variance, given the amount of first-order autocorrelation present in his series. Assuming for instance that pH variations along time are modeled as an AR(1) process, let $\hat{\rho}$ and $\hat{\boldsymbol{\Sigma}}_{\rho} = \hat{\boldsymbol{\Sigma}}_{\hat{\rho}}$ respectively denote the estimation of parameter ρ and of the correlation matrix $\boldsymbol{\Sigma}_\rho$; variance σ^2 may then be correctly estimated by

$$(\mathbf{x} - \bar{x}\,\mathbf{1}_n)'\,\boldsymbol{\Sigma}_{\hat{\rho}}^{-1}(\mathbf{x} - \bar{x}\,\mathbf{1}_n)$$

where $\mathbf{x}$ denotes the time series of pH variations, $\bar{x}$ the sample mean, $\mathbf{1}_n$ an $(n \times 1)$ vector of ones, and the transpose operator in matrix algebra (Cliff & Ord, 1981; Dutilleul & Legendre, 1991). If the temporal pattern of the pH variations depends on a higher number of parameters, a vector of parameters $\boldsymbol{\rho}$ may be substituted for the scalar parameter ρ and the procedure remains unchanged.

Boyle unwarrantedly uses z instead of t as the reference distribution when computing his confidence intervals. With a small number of cases, the difference is important; in the example that he describes (4 observations), the t value for a 95%

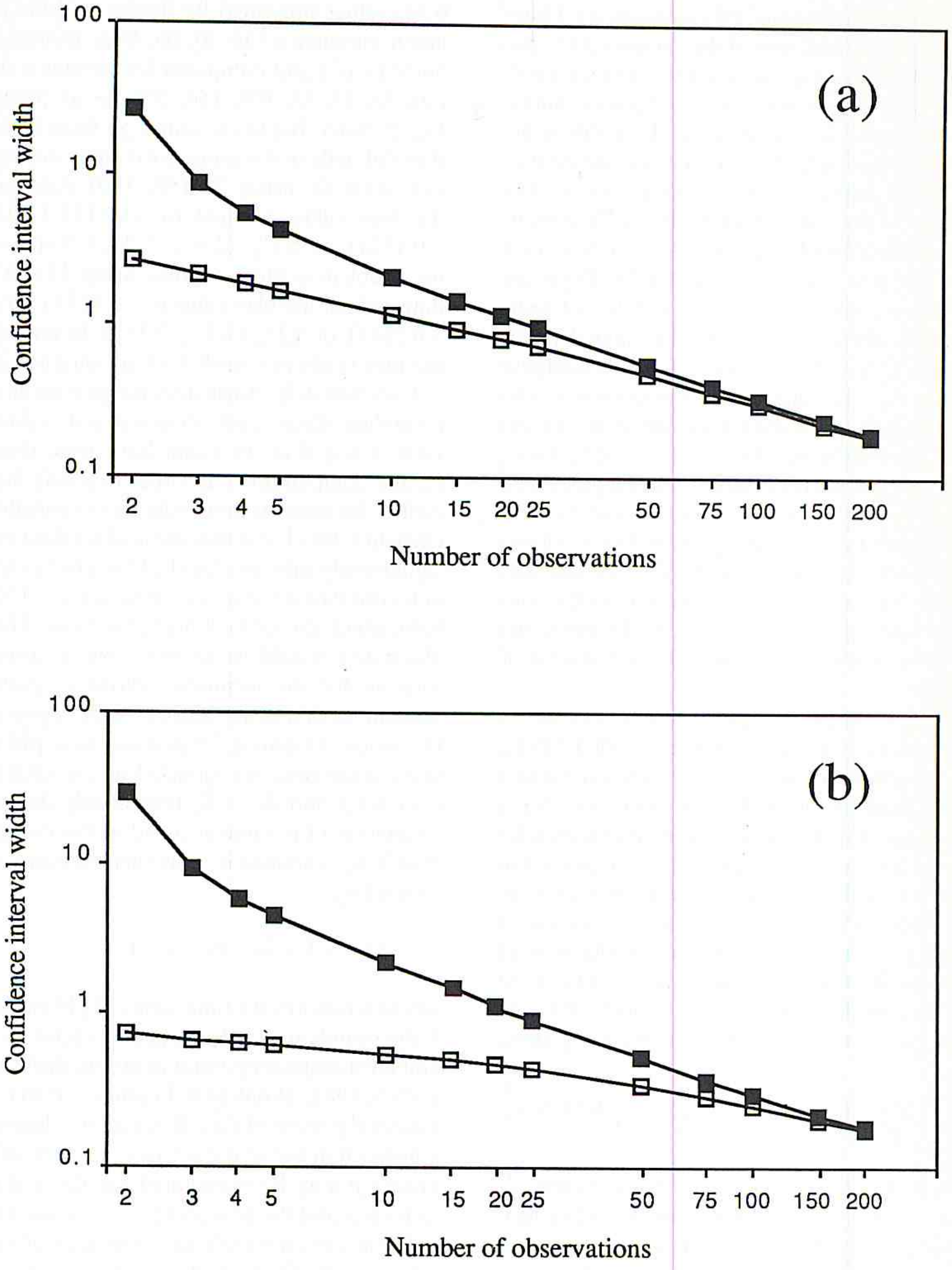

Fig. 3. Width of the confidence interval at level $\alpha = 0.05$, as a function of the number of observations. The graph shows the width calculated as recommended by Boyle (open squares) and with the appropriate corrections (dark squares), when pH varies along time as an AR(1) process with population variance $\sigma^2 = 1$, for two values of parameter ρ: (a) $\rho = 0.5$ and (b) $\rho = 0.9$.

confidence level (3 degrees of freedom) is 3.182 instead of 1.960, and it would take 60 observations for t to reach the value 2.000 that Boyle uses with 4 observations. So, with 4 observations, the confidence interval computed by Boyle would be too narrow by a factor of 1.6, assuming that his estimation of the variance was valid.

Using now the correction formula given above for estimating the variance in the presence of first-order autocorrelation, the following correction factors can be computed: for $n = 100$, $t = 1.984$ for $\alpha = 5\%$, compared to $z = 1.96$; this is not a large difference, but for $n = 25$, $t = 2.060$, which is already a more important one. Figure 3 illustrates the differences observed between confidence interval widths calculated using the classical sample variance as recommended by Boyle, and confidence interval widths computed using the appropriate corrections for pH varying along time as an AR(1) process, for two values of parameter ρ. Figure 3a ($\sigma^2 = 1$, $\rho = 0.5$) shows that when only two consecutive observations are used to compute the mean, the width of the 95% confidence interval is actually 9.2 times larger than estimated by Boyle (its value is 17.97 instead of 1.96 – notice that the scale of this graph is logarithmic). It takes at least 50 observations for Boyle's calculations to approximate correctly the width of the confidence interval. In the same way in Fig. 3b ($\sigma^2 = 1$, $\rho = 0.9$), with two observations only, the 95% confidence interval as calculated by Boyle would be 20.5 times too small.

Conclusion

The problem of correctly estimating the true mean of a time series of observations is not a simple one. Data normalization must be done with care, and no general recipe can be found that applies to all pH data series – or, for that matter, to any other variable of limnological interest. It depends on the type of variable and on the sampling scale (temporal or spatial), among other factors; each case has to be subjected anew to the search of the best normalizing transformation. Then, when estimating the confidence interval of the mean from a few observations only, the autocorrelation properties of the series must imperatively be taken into account. If they are not, the width of the confidence interval can be grossly underestimated.

References

Anderson, T. W., 1971. The statistical analysis of time series. Wiley, New York. xiv + 704 pp.

Box, G. E. P. & D. R. Cox, 1964. An analysis of transformations. J. Roy. Stat. Soc., Ser. B 26: 211–243.

Box, G. E. P. & G. M. Jenkins, 1976. Time series analysis: forecasting and control, 2nd ed. Holden-Day, San Francisco. xxi + 575 pp.

Boyle, J. F. Acidity and organic carbon in lake water: variability and estimation of means. J. Paleolimnol. 6: 95–101.

Cliff, A. D. & J. K. Ord, 1981. Spatial processes: models and applications. Pion Limited, London. xiv + 266 pp.

David, F. N. & N. L. Johnson, 1948. The probability integral transformation when parameters are estimated from the sample. Biometrika 35: 182–190.

Dutilleul, P. & P. Legendre, 1991. Spatial autocorrelation and its nuisance in statistical analysis: Outline of a global solution. Paper presented at Joint Meeting of the Psychometric Society and the Classification Society of North America, June 13–16, Rutgers University, New Brunswick, NJ, USA.

Dutilleul, P. & P. Legendre. Lack of robustness against autocorrelation in sample data for two tests of normality. J. Stat. Comput. Simul. (submitted).

Isaaks, E. H. & R. M. Srivastava, 1989. An introduction to applied geostatistics. Oxford University Press, New York. xix + 561 pp.

Legendre, P., 1991. Spatial autocorrelation: Trouble or new paradigm? Ecology (paper submitted as part of a Special Feature).

Legendre, P. & M.-J. Fortin, 1989. Spatial pattern and ecological analysis. Vegetatio 80: 107–138.

Legendre, P. & M. Troussellier, 1988. Aquatic heterotrophic bacteria: Modeling in the presence of spatial autocorrelation. Limnol. Oceanogr. 33: 1055–1067.

Lilliefors, H. W., 1967. The Kolmogorov-Smirnov test for normality with mean and variance unknown. J. Am. Stat. Ass. 62: 399–402.

Massey, E. J., 1951. The Kolmogorov-Smirnov test for goodness of fit. J. Am. Stat. Ass. 46: 68–78.

Ólafsson, J., 1979. The chemistry of Lake Mývatn and River Laxá. Oikos 32: 82–112.

Royston, J. P., 1982. An extension of Shapiro and Wilk's W test for normality to large samples. Appl. Statist. 31: 115–124.

SAS Institute Inc., 1985. SAS Procedures Guide for Personal Computers, Version 6 Edition. SAS Institute Inc., Cary, NC. xix + 373 pp.

Shapiro, S. S. & M. B. Wilk, 1965. An analysis of variance test for normality (complete samples). Biometrika 52: 591–611.

Schetagne, R. & D. Roy, 1985. Réseau de surveillance écologique du Complexe La Grande 1977–1984. Physico-chimie et pigments chlorophylliens. Annexe 2: Région d'Opinaca. Annexe 3: Région de Caniapiscau. Société d'Energie de la Baie James, Direction Ingénierie et Environnement, Montréal.

Sokal, R. R. & F. J. Rohlf, 1981. Biometry, 2nd ed. W. H. Freeman, San Francisco. xviii + 859 pp.

Stephens, M. A., 1974. EDF statistics for goodness of fit and some comparisons. J. Amer. Statist. Ass. 69: 730–737.