Opinion is intended to facilitate communication between reader and author and reader and reader. Comments, viewpoints or suggestions arising from published papers are welcome. Discussion and debate about important issues in ecology, e.g. theory or terminology, may also be included. Contributions should be as precise as possible and references should be kept to a minimum. A summary is not required.

When is a correlation between non-independent variables "spurious"?

Michael T. Brett, Dept of Civil & Environmental Engineering, Box 352700, Univ. of Washington, Seattle, WA 98195, USA (mtbrett@u.washington.edu).

Correlations which are artifacts of various types of data transformations can be said to be spurious. This study considers four common types of analyses where the X and Y variables are not independent; these include regressions of the form X/Z vs Y/Z, X × Z vs Y × Z, X vs Y/X, and X+Y vs Y. These analyses were carried out using a series of Monte Carlo simulations while varying sample size and sample variability. The impact of disparities in variability between the shared and non-shared terms and measurement error for the shared term on the magnitude of the spurious correlations was also considered. The accuracy of equations previously derived to predict the magnitude of spurious correlations was also assessed. These results show the risk of producing spurious correlations when analyzing non-independent variables is very large. Spurious correlations occurred in all cases assessed, the mean spurious coefficient of determination (r^2) frequently exceeded 0.50, and in some cases the 90% confidence interval for these simulations included all large r^2 values. The magnitude of spurious correlations was sensitive to differences in the variability of the shared and non-shared terms, with large spurious correlations obtained when the variability for the shared term was larger. Sample size had only a modest impact on the magnitude of spurious correlations. When measurement error for the shared variable was smaller than one half the coefficient of variation for that variable, which is generally the case, the measurement error did not generate large spurious correlations. The equations available to predict expected spurious correlations provided accurate predictions for the case of X × Z vs Y × Z, variable predictions for the case of X vs Y/X, and poor predictions for most cases of X/Z vs Y/Z, and X+Y vs Y.

It is common practice to conduct statistical analyses of non-independent variables in the scientific literature. This was first commented on by Pearson (1897) who discussed cases where ratios with identical denominators are analyzed. When commenting on published examples of this type, Pearson stated "a part of the correlation he discovers... is solely due to his arithmetic, and as a measure of organic relationship is spurious." Pearson defined spurious correlations to be correlations caused solely by data transformations which do not reflect meaningful properties of the underlying data. Pearson (1897) also derived a general equation to describe the spurious correlation between any two ratio statistics where components of these ratios were themselves correlated with each other. Pearson went on to show how using certain simplifying assumptions his general equation could be modified to an equation describing the spurious correlation expected when ratios with identical denominators are analyzed, e.g. X/Z versus Y/Z. Reed (1921) followed up on Pearson's study by deriving additional equations, using Pearson's general equation, describing the spurious correlation for other specific non-independent ratio statistics. Reed also showed how Pearson's general equation could be used to describe the spurious correlation for analyses of non-independent products, e.g. X × Z versus Y × Z. Chayes (1949) derived a wide range of new equations from Pearson's original equation which cover most types of spurious correlations of general interest.

Pearson (1897), Reed (1921), Chayes (1949), and more recently Bensen (1965) and Kenney (1982), were careful to point out how widespread these types of spurious correlations were in the literature of their day. More recently, however, Prairie and Bird (1989) suggested that the spurious correlation problem is over-inflated, and they further argued misunderstanding about the true extent of this problem can be attributed to the failure of modern ecologists to keep up with the relevant statistical literature. Peters (1991) also claimed the risk of "self-correlations" was overstated. Other authors have suggested the risk of generating spurious correlations when analyzing non-independent variables is so large that such analyses should be avoided whenever possible (Raubenheimer 1995). Several researchers have suggested some of modern ecology's most celebrated relationships may in fact be spurious correlations (Atchley et al. 1976, Kenney 1982, Weller 1987, Packard and Boardman 1988, Jackson et al. 1990, Jackson and Somers 1991, Raubenheimer 1995, Berges 1997, Knops et al. 1997, Jasienski and Bazzaz 1999). It has also been argued that regressing non-independent variables against each other can obscure real correlations (Packard and Boardman

1988, Beaupre and Dunham 1995). Despite the long history of these warnings, statistical analyses of non-independent terms are still widespread in the modern literature.

The objective of this study is to demonstrate the conditions under which "spurious correlations" are likely to arise, their expected magnitudes, and to show how the statistical significance of analyses of non-independent variables can be directly assessed. This was done using a Monte Carlo simulation approach, under a range of conditions ecologists are likely to encounter when conducting biometric analyses. These simulations were run for four common types of non-independent statistical analyses; i.e. regressions of the form X/Z vs Y/Z, $X \times Z$ vs $Y \times Z$, X vs Y/X, and $X+Y$ vs Y. The impact of measurement error, sample size, sample variability, and differences in numerator and denominator variability on the magnitude of spurious correlations was examined. Since Pearson neglected third or higher order variables when deriving his general equation, several authors (Reed 1921, Chayes 1949, Bensen 1965, Kenney 1982) have suggested the analytical solutions for the various forms of spurious correlations may only apply when the sample CV is small, which is generally not the case for ecological data. The validity of these analytic solutions was assessed using the results of the Monte Carlo simulations to show whether these equations can be used to estimate the expected magnitude of spurious correlations. This analysis will also present a simple approach for independently assessing the magnitude and statistical significance of expected spurious correlations using an example from a classic ecological hypothesis, i.e. the "nutrient-use efficiency" hypothesis (Vitousek 1982).

Methods

All random numbers were generated using the random number generator function of Microsoft Excel®, with a different random seed used for each simulation. The mean and standard deviations of all generated series were checked against the originally specified distribution. In about half the cases this procedure detected 1 or 2 values per 10 000 observations which were obviously not within the originally specified distribution. For example, a value of $\approx 50\,000$ when the specified distribution was mean = 25, SD = 10, and the second highest observation in the generated series was 62. Distributions with implausible values were not used for the Monte Carlo simulations. Once the appropriate random files were generated, correlation coefficients were calculated for the appropriate sample size and then squared to obtain the coefficient of determination. McCullough and Wilson (1999) pointed out that there are a variety of problems with the statistical packages in various Microsoft Excel® products. Because of the previously mentioned anomalous random number results, and McCullough and Wilson's (1999) warnings, both random number distributions and regression results were carefully screened. The validity of the correlation coefficients obtained using Excel® was assessed by comparing 100 correlation coefficients generated using Excel® and the statistical package Statview® using the same data-set. In the 100 cases assessed the correlation coefficients provided by the two software packages agreed perfectly to nine significant figures.

After three columns of 10 000 observations were generated, a test of the spurious correlation problem for regressions of the form X/Z vs Y/Z was conducted. For each simulation a mean of 25 and a constant SD was used when generating X, Y and Z. This process was repeated ten times (SD = 2.5, 5, 7.5, 10, 12.5, 15, 17.5, 20, 22.5, and 25) to cover a range of CVs relevant to biometry. To create X/Z and Y/Z variables, two new columns of data were created by dividing the 1st (X) and 2nd (Y) columns by data from the corresponding 3rd (Z) column. For the case where CV = 0 the X variable was compared directly to the Y variable, with a mean and SD of 25 ± 12.5. In cases where sample sizes of 20 or 100 were used, 500 or 100 simulations were conducted, respectively.

To test for spurious correlations for regressions of the form $X \times Z$ vs $Y \times Z$, new data were randomly generated, and 4th and 5th data columns were created by multiplying the 1st (X) and 2nd (Y) columns by the corresponding 3rd (Z) column. These simulations used the same range of sample variability, sample sizes, and number of simulations as for the X/Z vs Y/Z simulations.

To test for spurious correlations of the form X vs Y/X, two columns with 10 000 observations were generated, and a 3rd set of observations was created by dividing the 2nd column by the 1st (Y/X). The 1st column was then correlated with the 3rd column. These simulations used the same range of sample variability, sample sizes, and number of simulations as the preceding simulations. For these simulations, it was necessary to first transform all random numbers to their absolute value in order to eliminate negative values. This was done because results for negative numbers had the mirror image of results for positive numbers. These results were log transformed, i.e. $\log(1+X)$ vs $\log(1+Y/X)$, before coefficients of determination were calculated.

To test for spurious correlations of the form $X+Y$ vs Y, 10 000 X and Y observations were generated. A mean and SD of 25 ± 12.5 was used, with a second set of simulations using a mean and SD of 25 ± 2.5 also run. A 3rd set of observations was created by adding some fraction of the Y column to the X column ($X+Y$). The fraction of the Y column added to the X column was varied from $0.1 \times Y$ to $2 \times Y$ by units of 0.1, for 20 different simulations using a new set of random numbers

for each case. The 3rd (X+Y) data column was then correlated with the 2nd column (Y), using the same sample sizes and number of simulations as the preceding cases.

The impact of disparities between the shared term and non-shared term variability on the magnitude of spurious correlations for the cases considered was assessed by holding the mean and SD of the non-shared term constant at 25 ± 5 (± 1 SD), and varying the SD of the randomly generated shared term (mean = 25) from 20, 14.29, 10, 7.04, 5, 3.55, 2.5, 1.77, and 1.25. This resulted in ratios of non-shared to shared term variability of 0.25, 0.35, 0.5, 0.71, 1, 1.41, 2, 2.83 and 4, respectively. Spurious correlations were then generated for each of the four cases as described above using a sample size of 20 and repeated 500 times.

An analysis of measurement error impacts on the four non-independent regression types was conducted by calculating spurious correlations as previously described. A new variable was then randomly generated which gave a "measurement error" (± 1 SD) which corresponded to $\pm 10\%$, 3.3% or 1% of the mean. This new variable was then added to the shared term in these non-independent regressions. That is, it was added to Z for X/Z vs Y/Z and $X \times Z$ vs $Y \times Z$ regressions, to X for X vs Y/X regressions, and to Y for X+Y vs Y regressions. This procedure was repeated for a range of sample variation (i.e. SD = 2.5, 5, 7.5, 10, 12.5, and 15). This process was carried out 500 times using a sample size of 20 for each combination of sample variability and sample error.

In addition, the accuracy of the equations for predicting the expected spurious correlation was compared to the results of the Monte Carlo simulations. For a non-independent analysis of the form X/Z vs Y/Z (assuming $r_{XY} = r_{XZ} = r_{YZ} = 0$) the expected spurious correlation is:

$$r = \frac{CV_Z^2}{(CV_X^2 + CV_X^2)^{1/2}(CV_Y^2 + CV_X^2)^{1/2}} \quad (1)$$

where CV_Z, CV_X, and CV_Y equal the CVs of variables Z, X, and Y, respectively (Pearson 1897). The equation for a non-independent analysis of the form $X \times Z$ vs $Y \times Z$ (again assuming $r_{XY} = r_{XZ} = r_{YZ} = 0$) is identical to Eq. 1 above (Reed 1921). The equation for the expected spurious correlation of the form X vs Y/X (assuming $r_{XY} = 0$) is:

$$r = \frac{-CV_X}{(CV_Y^2 + CV_X^2)^{1/2}} \quad (2)$$

(Chayes 1949). The equation for the expected spurious correlation for regressions of the form X+Y vs Y (assuming $r_{XY} = 0$) is:

$$r = \frac{1}{(1 + (CV_X/CV_Y)^2)^{1/2}} \quad (3)$$

(Bensen 1965). These equations are special cases of Pearson's general equation, which assume no underlying correlations between the variables. Chayes (1949) also published these equations in their more general form which accounts for correlations between the variables.

This study will also demonstrate the ease with which the expected magnitude of spurious correlations and the statistical significance of analyses of non-independent variables can be determined using data from a classic study which regressed two non-independent variables against each other (Vitousek's 1982, 1984). To test for the expected spurious correlation in Vitousek's (1982) Fig. 4, the data from this figure were first extracted. A series of 1000 bootstrap simulations with replacement were then conducted by copying Vitousek's nitrogen and carbon data 1000 times, randomly sorting each independently of the other several times, and then calculating 1000 regressions of nitrogen vs dry weight/nitrogen using the same sample size (n = 105) as Vitousek. Because Vitousek (1997) claimed a significant positive intercept when regressing litterfall dry weight (Y) against litterfall nitrogen content (X) provided statistical support for the original "nutrient-use efficiency hypothesis", the same data were then used to test whether statistically significant intercepts resulted when regressing randomly sorted data against each other. This was tested by randomly sorting Vitousek's data and regressing litterfall dry weight against nitrogen content 100 times with replacement.

Results

X/Z vs Y/Z

This type of spurious correlation can be generated when, for example, lake primary production/surface area is compared to lake fisheries production/surface area (Jackson et al. 1990) or in the case of Sterner et al. (1997) when they compared the carbon to phosphorus ratio in epilimnetic seston to the light to phosphorus ratio in the epilimnion. Spurious correlations of the form X/Z vs Y/Z always occur, and they are moderately strong when the CV of the sample is < 0.4 (Fig. 1). When the sample CV is ≥ 0.4, the 90% confidence intervals of these distributions include essentially all large and interesting r^2 values.

X × Z vs Y × Z

This type of non-independent regression can occur when hydrologists attempt to validate sediment loading curves by comparing predicted sediment concentration (C_p) multiplied by the observed flow (Q) against observed sediment concentration (C_o) multiplied by the observed flow or $C_p \times Q$ vs $C_o \times Q$ (Cohn et al. 1992). Unlike

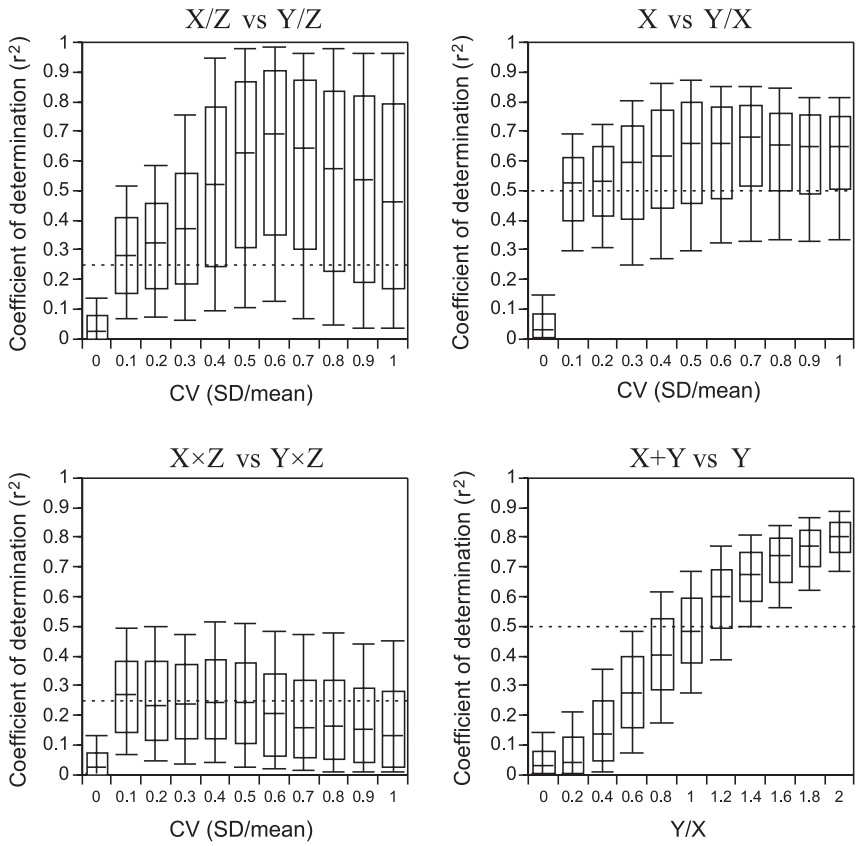

Fig. 1. The results of Monte Carlo simulations for the cases of X/Z vs Y/Z, X × Z vs Y × Z, and X vs Y/X across a gradient of equal sample CVs, and X + Y vs Y across a gradient of Y versus X. The CV = 0 results represent the case where the X and Y variables were completely independent of each other (e.g. X vs Y). The "Box and Whisker" plots show the median (inner horizontal bar), the 25th and 75th percentiles (the outer dimensions on the "boxes") and the 10th and 90th percentiles (the "whiskers") of these distributions. The sample size for these simulations was 20, and 500 simulations were run. Although not depicted, the results for a sample size of 100 had in most cases virtually identical means, but much less variability. The hatched horizontal lines show the spurious correlations predicted by Eq. 1, 2 and 3.

regressions of the form X/Z vs Y/Z, this type of non-independent regression only produced moderately large ($r^2 = 0.15–0.35$) spurious correlations (Fig. 1). These spurious correlations were not strongly dependent on sample variability.

X vs Y/X

This is a very common type of non-independent regression. Examples include regressing the ratio of leaf litter nitrogen to dry weight against leaf litter nitrogen or some weight specific function (e.g. nutrient excretion/body weight) against an organism's body weight (Berges 1997). The present analysis shows regressions of the form X vs Y/X often produce spurious correlations averaging $r^2 = 0.52–0.63$ (Fig. 1). Although not shown in Fig. 1, these regressions are almost always negative.

X + Y vs Y

Prairie and Bird (1989) used the example of regressing body mass against liver mass in vertebrates as a form of the X + Y vs Y regression. Another example of this type of non-independent regression is regressing lake water total phosphorus against particulate phosphorus. The present analysis shows that in cases where $Y \ll X$ (e.g. liver mass is a small portion of body mass), this form of non-independent regression will only produce small spurious correlations (Fig. 1). This form of spurious correlation is not notably influenced by variation for the data used to generate them. An analysis of the magnitude of spurious correlations generated using large (CV = 0.5) or small (CV = 0.1) sample variation resulted in spurious correlations with nearly identical means and distributions as those presented in Fig. 1.

Sample size impacts on spurious correlations

In almost all the cases simulated in Fig. 1, similar average spurious correlations were observed when using sample sizes of 20 and 100. However, for the cases of X × Z vs Y × Z, X vs Y/X, and X + Y vs Y only about 45% as much variability in the generated spurious correlations was observed when using the larger sample size. For the case of X/Z vs Y/Z, about 15% less variability was observed when using a larger sample size.

Disparities between shared and non-shared term CVs

The magnitude of spurious correlations was very strongly correlated with disparities in variability between the shared and non-shared terms (Fig. 2). When the shared term CV was more than 1.5 times larger than the non-shared term CV, spurious correlations were very large. Conversely, when the shared term CV was smaller, spurious correlations were generally much smaller.

Measurement error in shared terms and spurious correlations

This study also examined the impact of measurement error for shared terms on the spurious correlation problem. Measurement error refers to error due to collecting an unrepresentative sample and/or analytical error when this sample is processed in the lab. This error can be directly estimated by collecting a blind field duplicate; i.e. collecting a separate sample in the field which is processed independently of the other duplicate in both the field and the laboratory. Measurement error should be distinguished from natural variability, such as that variability normally obtained within treatments by the end of an experiment or the variability that occurs within a population of observations (e.g. between lake variation in nutrient concentrations). If, for example, measurement error for Z is large and natural variation in Z is small it is easy to envision a scenario where a regression of the form X/Z vs Y/Z would generate strong but meaningless correlations. An analysis of the measurement error problem shows that unless the measurement error is a substantial fraction of the variable variance, e.g. measurement error/CV ≥ 0.5 then spurious correlations generated by measurement error are generally small and much less than the spurious correlations introduced by the common term (Fig. 3).

Analytical versus simulation results

Equation 1 failed to predict the strong dependence of spurious correlations of the form X/Z vs Y/Z on the overall variable CV when all CVs were equal (Fig. 1). In addition, when variable CVs were equal and larger than 0.3, Eq. 1 severely underestimated the magnitude of the observed spurious correlation. Equation 1 also underestimated the magnitude of the expected spurious

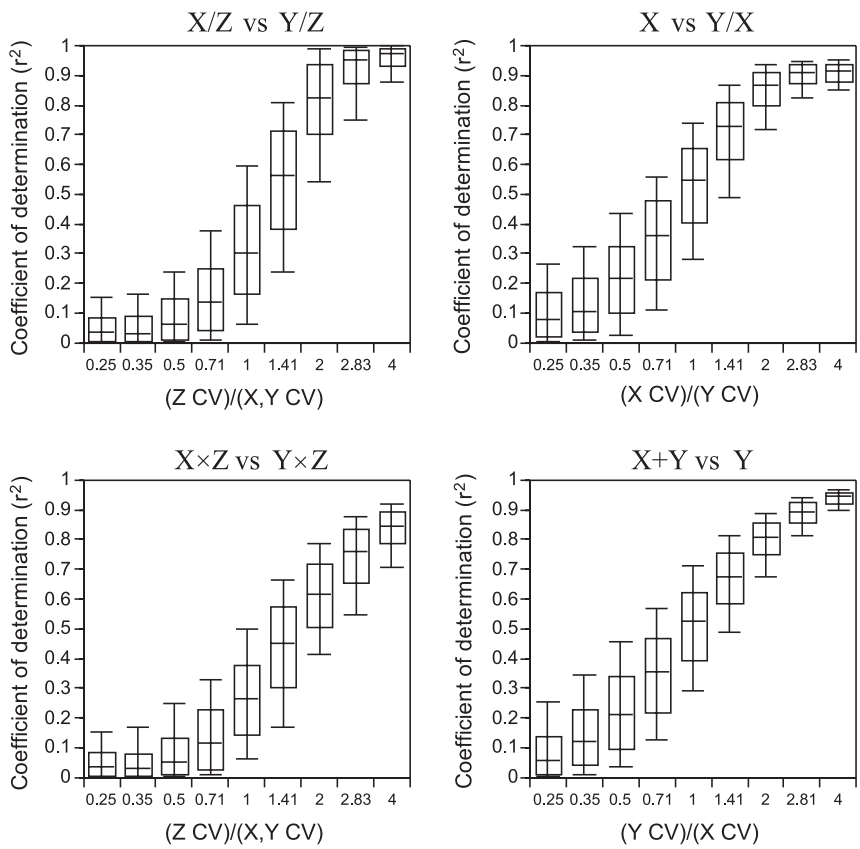

Fig. 2. The impact of disparities in variability between the shared and non-shared terms for each of the four types of regressions considered in this study. For these simulations the mean and variability of the non-shared term was held constant at 25 ± 5 (± 1 SD), and the variability of the shared term was varied in an exponential series from 1.25 to 20. For each case, 500 simulations with sample sizes of 20 were run. The results for simulations with samples sizes of 100 had virtually identical means and much less variability.

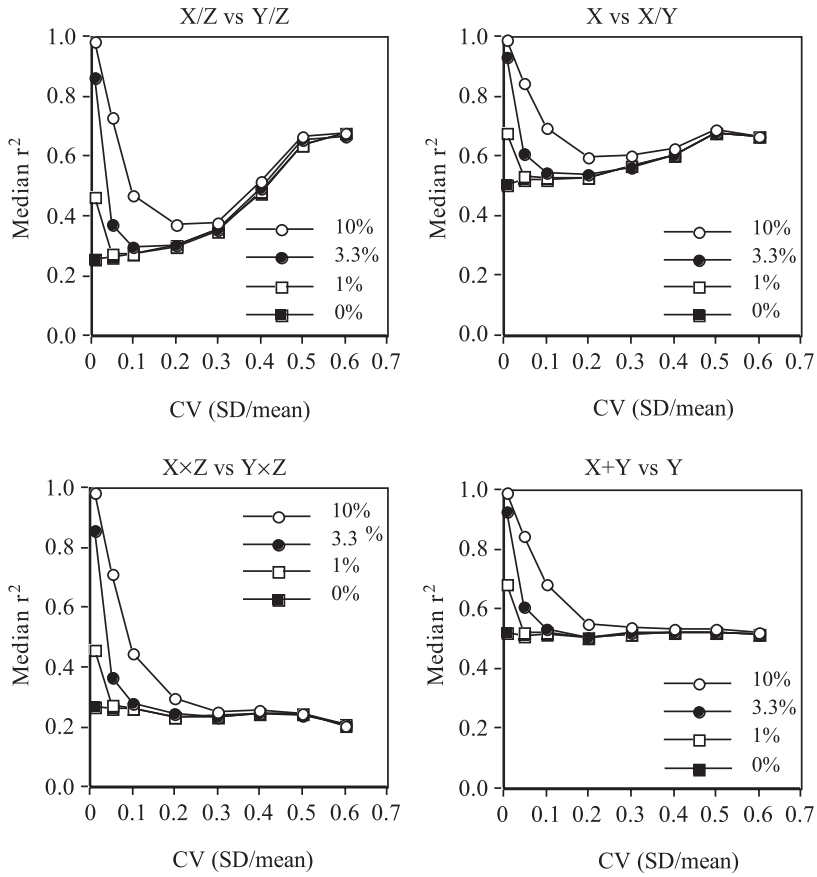

Fig. 3. The impact of measurement error for the shared term on the magnitude of spurious correlations for each of the four types of regressions considered in this study. These simulations were run across a gradient of sample variability and measurement error. These simulations were conducted by first assigning a common variability to the shared and non-shared terms and then adding a random error term equivalent to some percentage of the sample mean to the shared term. For each case, 500 simulations with sample sizes of 20 were run. The medians of these distributions were plotted.

correlation when the CV for Z was greater than that for X and Y (Fig. 4). In contrast, Eq. 1 did a very good job of predicting the expected spurious correlation for the case of $X \times Z$ vs $Y \times Z$ across a wide range of shared CVs (Fig. 1), as well as when the CV for Z differed from that for X and Y (Fig. 4). Equation 2 did a good job of predicting the expected spurious correlation for the case of X vs Y/X when the variable CV was ≤ 0.20 (Fig. 1), but tended to strongly underestimate the spurious correlation at higher sample CVs (Fig. 4). Furthermore, Eq. 2 only provided accurate predictions when the data were first log(1+X) transformed. Equation 3 predicted the spurious correlation perfectly for the case of X+Y vs Y when X = Y (Fig. 4), but failed to accurately predict the expected spurious correlation in all other cases (Fig. 1).

Vitousek (1982) reconsidered

The litterfall nitrogen and dry weight in Vitousek's (1982) Fig. 4 had means $\pm$1 SD of 65 ± 50 and 5553 ± 2782 (kg ha^{-1} yr^{-1}), respectively. When the data from Vitousek's Fig. 4 was statistically analyzed, after log(1+X) transformation, a coefficient of determination of 0.65 was obtained. The mean $\pm$1 SD r^2 values of the 1000 bootstrap simulations, after randomly sorting these data, was 0.69 ± 0.07. According to these results the relation reported by Vitousek (1982) in his Fig. 4 had a P-value of 0.7265. To test whether a significant positive intercept when regressing litterfall dry weight (y) against litterfall nitrogen content (x) is positive support for Vitousek's "nutrient-use efficiency" hypothesis, we randomly sorted Vitousek's litterfall dry weight and nitrogen content data and calculated 100 regression equations. The mean intercept of these regressions was 5566 ± 455. The t-value for these intercepts averaged 12.56 ± 1.36 and was significant in each case at the 0.0001 level.

Discussion

The most fundamental premise of statistics is that the strength of any association should be judged against the likelihood that that association could have occurred purely due to random chance. The present study shows regressing non-independent variables against each other frequently results in regressions which appear to be strong based on their r^2 values, but which are not

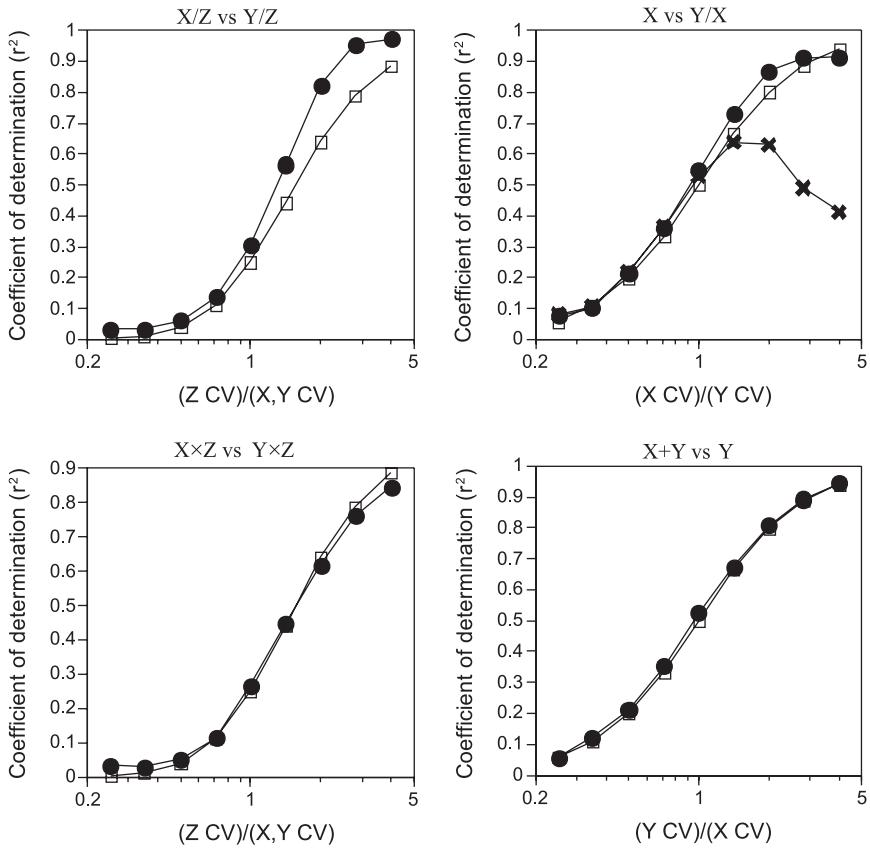

Fig. 4. An analysis of the predicted spurious correlation from Eq. 1, 2 and 3 (open squares) and the spurious correlations obtained from the simulations (closed circles) when variability between the shared and non-shared terms differed. The cross-hatches (×) represent cases for the X + Y vs Y simulations where the values were not log transformed. For each point, 500 simulations with sample sizes of 20 were run. Median values are plotted.

significant (at $P = 0.05$) as determined by Monte Carlo simulations. These results indicate researchers should be cautious when drawing inferences from analyses of non-independent terms. In many cases such analyses will not produce meaningful results, e.g. regressions of the form X/Z vs Y/Z and X vs Y/X when the sample CVs = 0.4, or regressions of the form X + Y vs Y when the Y term is equal to or greater than the X term. In addition, all non-independent analyses assessed in this study produced very large spurious correlations when the shared variable CV was greater than that for the non-shared terms. In the preceding cases regressing non-independent variables against each other produced a range of results encompassing virtually all large and interesting coefficients of determination. Counter-intuitively, sample size did not have a substantial impact on the average magnitude of spurious correlations, although large sample sizes did result in less variance in expected outcomes. It should be emphasized that these results do not mean that all observed correlations between non-independent variables are solely due to spurious correlations. The results of the simulations summarized in this study simply suggest that in many cases the expected spurious correlation may be so large that it will be very difficult to discern the actual magnitude of any true correlations.

In some cases, however, the spurious correlation is sufficiently small (mean $r^2 < 0.3$) so that it is possible to conduct a careful statistical analysis of the data provided the correct null model is specified (see later). For example regressions of the form $X \times Z$ vs $Y \times Z$, regressions of the form X + Y vs Y when the Y term is much smaller than the X term, or all analyses assessed in this study when the shared term variability was substantially less than that for the non-shared terms. However, it should be noted that the statistical significance of any such regressions can only be determined through Monte Carlo (Crowley 1992) or Bootstrap simulations (Efron and Tibshirani 1993). Significance values generated from least squares or similar parametric analyses will not be accurate.

It is also worth noting that the relevance of whether a specific correlation is spurious depends on whether that statistical association is being used to predict or to explain, sensu Pedhazur (1997). When using a statistical association to merely make a prediction the fact that the statistical association might be a mathematical artifact could be irrelevant. For example, if X is being used to predict X + Y when filling in missing data it could be advantageous that X + Y is dependent on X out of mathematical necessity. However, when a correlation is

being used for explanatory purposes, i.e. when making the inference that one variable appears to cause variation in the other in a mechanistic sense, whether or not a correlation is spurious is of critical importance. For example, in a study of the biogeochemical composition of lake seston in 115 northwest Ontario lakes (Fee et al. 1989), lake water particulate phosphorus (PP) content was very highly correlated with lake water total phosphorus (TP) content ($r^2 = 0.91$). For this data-set, knowing a lake's PP concentration makes it possible to predict the corresponding TP concentration with a high degree of accuracy because in this survey PP constituted on average $45 \pm 9\%$ (± 1 SD) of TP. However, knowing the PP concentrations in these lakes provides no insight into why the 25th percentile of this sample had a TP concentration of 15.3 µg/l and the 75th percentile had a TP concentration of 22.9 µg/l. Interestingly, the mean spurious correlation between PP and TP in the Fee et al. (1989) data-set was found to be $r^2 = 0.80 \pm 0.02$ (± 1 SD), by randomly disaggregating the data and correlating PP against TP (where TP equals PP plus total dissolved phosphorus).

Several authors have suggested that measurement error for the shared term is an important (Atchley et al. 1976, Raubenheimer 1995) or the most important (Prairie and Bird 1989, Sterner et al. 1997) factor likely to generate spurious correlations. Measurement error is the variability introduced by unrepresentative sampling and/or instrument error during sample processing (i.e. the error normally assessed using field duplicates), and should be clearly distinguished from natural variability which normally occurs within a treatment or a population of interest and is the primary focus of this study. However, in direct contrast to these predictions, the present study shows measurement error will usually not generate strong spurious correlations in most statistical analyses. True to hypothetical examples cited in past studies (Kenney 1982), the impact of measurement error on spurious correlations is quite large when the measurement error is larger than the natural background variability for the shared variable, i.e. when measurement error is the main source of variation for that variable. However, large spurious correlations due to measurement error only occurred when measurement error for the shared variable was larger than the CV for that variable. The special case of measurement error exceeding natural variability for a variable is, hopefully, rare in statistical analyses of ecological data. Assessment of the impact measurement error on spurious correlation generation strongly suggests that it is the actual shared terms, and not measurement error, which is the main cause of spurious correlations.

The analytical equations developed by Pearson (1897), Reed (1921), Chayes (1949) and Bensen (1965) vary greatly in their ability to predict the expected spurious correlation. Equation 2 provided in many cases accurate predictions of the expected spurious correlation for analyses of the form $X \times Z$ vs $Y \times Z$. Equation 2 provided accurate predictions for spurious correlations of the form X vs Y/X provided the variable CV was small and the data were $\log(1+X)$ transformed. It is not surprising that log transforming the data would improve the accuracy of Eq. 2, but it is surprising that Eq. 2 provided very inaccurate predictions under some conditions without log transformation. Equation 1 did a poor job of predicting the expected spurious correlation for analyses of the form X/Z vs Y/Z under a wide range of conditions. This result is noteworthy because previous investigators (Reed 1921, Chayes 1949, Bensen 1965) pointed out that Eq. 1 should work for both analyses of the form $X \times Z$ vs $Y \times Z$ and X/Z vs Y/Z. The failure of Eq. 1 and 2 to accurately predict spurious correlations when variable CVs were ≥ 0.30 is in all likelihood due to the dependence of these types of spurious correlations on the overall variable variability and the fact that Pearson's original derivation is strictly speaking only valid at small CVs. Equation 3 only predicted the expected spurious correlation under very special circumstances, i.e. $X = Y$, which will rarely be satisfied. If these equations actually provided accurate estimates of the expected magnitude of spurious correlations for a wide range of conditions, they would be very useful tools when conducting analyses of non-independent variables. A disadvantage with all these analytical solutions, even when they provide accurate predictions, is that they do not allow investigators to establish a significance level for a calculated correlation. To do this it is still necessary to carry out appropriate bootstrap or Monte Carlo simulations. The fact that most of the above mentioned equations failed to accurately predict the true magnitude of spurious correlations in most scenarios shows there is an opportunity for statisticians to develop a new series of equations that can accurately predict spurious correlations.

The null model mis-specification problem

Some authors have argued that the problem with analyzing non-independent variables using regression approaches is not that spurious correlations are generated, but instead that virtually all authors mis-specify their null model (Prairie and Bird 1989). The intuitive and most commonly used, but incorrect, null model in cases where non-independent variables are analyzeed is $r \approx 0$ (Chayes 1949, Bensen 1965, Atchley et al. 1976, Kenney 1982, Jackson et al. 1990, Raubenheimer 1995). However, as this study has already shown, randomly generated regressions of non-independent variables almost always produce correlation coefficients larger than this. Logically, one can conduct regression analyses of non-independent variables as long as one is clear that the

null model is the average r or r^2 generated from an appropriate Monte Carlo or Bootstrap simulation, and statistical significance is judged based on the results of these simulations. Gotelli and Graves (1996) provided useful guidance on this problem "A null model is a pattern-generating model that is based on randomization of ecological data... [with] the randomization designed to produce a pattern that would be expected in the absence of a particular ecological mechanism." Some authors have misinterpreted theoretical arguments in favor of conducting regressions of non-independent variables to mean that one can also use a null model of $r \approx 0$. To quote Reed (1921) "the value of the spurious correlation involved should always be considered when drawing conclusions from the coefficient of correlation of any two index numbers regardless of their functional form".

For the sake of simplicity, this study has based all of its conclusions on analyses of distributions of simulated r^2 values. This was done because most ecologists prefer to report r^2 values, and because the coefficient of determination describes the variability explained by a statistical association. However, as a practical matter, it will usually make the most sense to base these simulations on analyses of r value distributions. This is because regressing non-independent ratios against each other produces distributions of r values which are strongly skewed towards one sign. For example, spurious correlations of the form X/Z vs Y/Z, $X \times Z$ vs $Y \times Z$, and $X+Y$ vs Y are positive, while spurious correlations of the form X vs Y/X are negative.

Vitousek's (1982) nutrient-use efficiency

In some cases regressions analyzing ratios which are not entirely independent may be the only approach that makes sense given the hypothesis being tested. Several authors have recommended randomization or Bootstrapping procedures to calculate the true significance of non-independent regression analyses (Buonaccorsi and Liebhold 1988, Jackson and Somers 1991). To demonstrate the ease with which these simulations can be preformed, the results of a reanalysis of data from a classic study will be discussed.

Vitousek's 1982 paper on nutrient-use efficiency by terrestrial plants is a citation classic, with this and a very similar paper (Vitousek 1984) receiving over 600 citations since their publication. Based on a comparison of the nitrogen content of litterfall against the ratio of litterfall dry weight to litterfall nitrogen content (his Fig. 4), Vitousek concluded plants from nutrient poor habitats to produce more organic matter per unit nitrogen than plants from nutrient-rich habitats. As Knops et al. (1997) previously pointed out, and Vitousek (1997) acknowledged, this X vs Y/X type analysis is quite prone to spurious correlations. In fact, my reanalysis of the data in Vitousek's (1982) Fig. 4 shows it could simply be a spurious correlation. This is an important point because 13 of the 16 statements regarding the nutrient-use efficiency hypothesis in the discussion section of Vitousek (1982) referred to Fig. 4. After acknowledging that the regression in his Fig. 4 might indeed be spurious, Vitousek (1997) went on to argue that the true test of the nutrient use efficiency hypothesis is whether the "y-intercept of litterfall dry mass regressed against litterfall N" is positive. According to Vitousek (1997), "positive y-intercepts establish that there is systematically more dry mass per unit of nutrient in litterfall in the low-nutrient sites". In contrast to Vitousek's claim, the present study found a positive intercept, which can simply be caused by lack of fit in the original data. Thus, the evidence Vitousek originally (Vitousek 1982) and subsequently (Vitousek 1997) used to support his formulation of the nutrient-use efficiency hypothesis can be attributed to various forms of spurious correlations.

Conclusions

The risk of producing a spurious correlation when analyzing non-independent variables should not be surprising because one of the first things most students learn about regression analysis is the X and Y variables should always be independent. In fact, the literature on this topic is very old! There have been several excellent and clearly written papers on this topic (Chayes 1949, Atchley et al. 1976, Kenney 1982), but their warnings have often been ignored. For some reason, the rationale behind and the risks associated with this statistical problem have failed to resonate with the ecological community in the same way as Hurlbert's (1984) warning against pseudo-replication. This is true although the danger with pseudo-replication and spurious correlations is exactly the same, i.e. drawing incorrect inferences from improperly designed analyses which inflate the strength of statistical associations. It may also be true that spurious correlations are more widespread and persistent in the literature than pseudo-replication ever was.

A number of authors have argued that regressions of non-independent ratios should be avoided at all costs, and researchers should instead use analysis of covariance (ANCOVA) of the relevant variables to analyze data in these cases (Atchley et al. 1976, Packard and Boardman 1988, Jasienski and Bazzaz 1999). The results of the present study clearly show that in some cases bivariate regression analyses will provide completely meaningless results. In other cases, however, it is possible to conduct meaningful bivariate analyses of non-independent ratio variables provided the appropriate null model is speci-

fied. The statistical significance of any analysis of non-independent variables can be easily determined using bootstrap or Monte Carlo simulations. The results of the present study show researchers should exercise caution when conducting classic parametric statistical analyses of non-independent variables. As Kenney (1982) and Jackson and Somers (1991) aptly warned, "beware the spectre of 'spurious' correlations".

Acknowledgements – I would like to thank Steve Burges, Giorgos Arhonditsis, and Philip Crowley for thoughtful advice during preparation of this manuscript.

References

Atchley, W. R., Gaskins, C. T. and Anderson, D. 1976. Statistical properties of ratios. I. Empirical results. – Syst. Zool. 25: 137–148.

Beaupre, S. J. and Dunham, A. E. 1995. A comparison of ratio-based and covariance analyses of a nutritional data set. – Funct. Ecol. 9: 876–880.

Bensen, M. A. 1965. Spurious correlations in hydraulics and hydrology. – J. Hydraulic. Div., Proc. Am. Soc. Civil Eng. 91: 35–42.

Berges, J. A. 1997. Ratios, regression statistics, and "spurious" correlations. – Limnol. Oceanogr. 42: 1006–1007.

Buonaccorsi, J. P. and Liebhold, A. M. 1988. Statistical methods for estimating ratios and products in ecological studies. – Environ. Entomol. 17: 572–580.

Chayes, F. 1949. On ratio correlation in petrography. – J. Geol. 57: 239–254.

Cohn, T. A., Caulder, D. L., Gilroy, E. J. et al. 1992. The validity of a simple statistical model for estimating fluvial constituent loads – an empirical study involving nutrient loads entering Chesapeake Bay. – Wat. Resour. Res. 28: 2353–2363.

Crowley, P. H. 1992. Resampling methods for computation-intensive data analysis in ecology and evolution. – Annu. Rev. Ecol. Syst. 23: 405–447.

Efron, B. and Tibshirani, R. 1993. An Introduction to the Bootstrap. – Chapman & Hall.

Fee, E. J., Hecky, R. E., Stainton, M. P. et al. 1989. Lake variability and climate research in northwestern Ontario: study design and 1985–1986 data from the Red Lake District. – Can. Tech. Rep. Fish. Aquat. Sci. 1662.

Gotelli, N. J. and Graves, G. R. 1996. Null models in ecology. – Smithsonian Institution Press

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. – Ecol. Monogr. 54: 187–212.

Jackson, D. A. and Somers, K. M. 1991. The spectre of 'spurious' correlations. – Oecologia 86: 147–151.

Jackson, D. A., Harvey, H. H. and Somers, K. M. 1990. Ratios in aquatic sciences: statistical shortcomings with mean depth and the morphoedaphic index. – Can. J. Fish. Aquat. Sci. 47: 1788–1795.

Jasienski, M. and Bazzaz, F. A. 1999. The fallacy of ratios and the testability of models in biology. – Oikos 84: 321–326.

Kenney, B. C. 1982. Beware of spurious self-correlations! – Wat. Resour. Res. 18: 1041–1048.

Knops, J. M. H., Koenig, W. D. and Nash, T. H. 1997. On the relationship between nutrient use efficiency and fertility in forest ecosystems. – Oecologia 110: 550–556.

McCullough, B. D. and Wilson, B. 1999. On the accuracy of statistical procedures in Microsoft Excel 97. – Computational Statistics Data Analysis 31: 27–37.

Packard, G. C. and Boardman, T. J. 1988. The misuse of ratios, indices and percentages in ecophysiological research. – Physiol. Zool 61: 1–9.

Pearson, K. 1897. On a form of spurious correlation which may arise when indices are used in the measurement of organs. – Proc. R. Soc. Lond. 60: 489–498.

Pedhazur, E. J. 1997. Multiple-regression in behavioral research: explanation and prediction, 3rd ed. – Harcourt Brace College Publishers.

Peters, R. H. 1991. A critique for ecology. – Cambridge Univ. Press.

Prairie, Y. T. and Bird, D. F. 1989. Some misconceptions about the spurious correlation problem in the ecological literature. – Oecologia 81: 285–288.

Raubenheimer, D. 1995. Problems with ratio analysis in nutritional studies. – Funct. Ecol. 9: 21–29.

Reed, J. L. 1921. On the correlation between any two functions and its application to the general case of spurious correlation. – J. Wash. Acad. Sci. 11: 449–455.

Sterner, R. W., Elser, J. J., Fee, E. J. et al. 1997. The light:nutrient ratio in lakes: the balance of energy and materials affects ecosystem structure and process. – Am. Nat. 150: 663–684.

Vitousek, P. 1982. Nutrient cycling and nutrient use efficiency. – Am. Nat. 119: 553–572.

Vitousek, P. M. 1984. Litterfall, nutrient cycling, and nutrient limitation in tropical forests. – Ecology 65: 285–298.

Vitousek, P. M. 1997. On regression and residuals: response to Knops et al. – Oecologia 110: 557–559.

Weller, D. E. 1987. A reevaluation of the −3/2 power rule of plant self-thinning. – Ecol. Monogr. 57: 23–44.