Oecologia (1989) 81:285–288

Oecologia

Short communication

Some misconceptions about the spurious correlation problem in the ecological literature*

Yves T. Prairie and David F. Bird*****
Department of Biology, McGill University, 1205 avenue Docteur Penfield, Montréal, Québec, Canada H3A 1B1

Summary. It is a common misconception that correlations between variables that share a common term are statistically invalid. Although the idea that such relationships are wholly or partially spurious was rejected decades ago by statisticians, ecologists continue mistakenly to exclude legitimate hypotheses on this basis. Besides directing attention to the statistical literature on the subject, we briefly reconsider the problem from 3 viewpoints: first, the confusion between spurious correlation and spurious inference, second, the problem of concept familiarity and definition, with particular reference to the self-thinning rule for plants, and third, a legitimate concern with measurement error of shared variable components.

Key words: Spurious correlation – Ratios – Relationships – Statistical inference – Self-thinning

Ecology is, by definition, a discipline devoted in part to determining the relative and quantitative importance of different relationships among the biotic and abiotic constituents of the world. The power of statistical techniques has made them an indispensable tool for ecological research in this regard, but their use requires caution and familiarity. To this end, there have been a number of cautionary reports exploring and defining the range for the proper application of different methods. In particular, there is a literature extending back to the turn of the century advising us to beware of "spurious" correlations (see Kenney 1982).

A correlation between two variables is said to be spurious when its magnitude is attributable to a term common to both correlated variables. Pearson (1897), who coined the term, examined this problem in the correlation between ratios sharing the same denominator. He concluded that the correlation induced by the common denominator among a set of otherwise uncorrelated variables represented a "real danger". Pearson's intent was to distinguish correlations that are spurious or partly spurious from "genuine" correlations expressing some meaningful relationship between two variables. Many natural scientists have taken his warning to heart. There are reiterations of Pearson's arguments (or variations on them) in research papers in hydrology (Yalin and Kamphuis 1971), eutrophication studies (Kenney 1982), plant community ecology (Weller 1987), and allometry (Atchley et al. 1976). Apparently unknown to most, however, has been the intervening reconsideration of the problem by statisticians. This reconsideration has resulted in a recanting and subsequent abandonment of the term "spurious" in reference to correlations. The purpose of this paper is to draw the attention of ecologists to the facts and, particularly, to the misconceptions surrounding the notion of spurious correlation. It is hoped thereby to rehabilitate legitimate hypotheses and results that were previously discarded on grounds of spuriousness.

A more precise definition of spurious correlation

We wish to specify clearly from the outset the problem we are addressing, and those we are not. We will not consider the problem of statistical estimation bias discussed by Garsd (1984) which is usually (and more appropriately) termed model misspecification rather than spurious correlation. Similarly, spurious correlations have been equated in some cases with nonsense correlations, which refers to the correlation observed between variables that are, from a logical and causal point of view, unrelated (e.g. Bhattacharyya and Johnson 1977; Lapin 1980). Instead, we consider the case of a correlation labelled spurious when its statistical significance depends, in whole or in part, on a necessary mathematical relationship between two variables or their components and not on any true empirical informative content. This is the traditional definition of spurious correlation.

A simple example will clarify this definition. Let X, Y, and Z be normal random variables, where $Z=X+Y$. The correlation between Z and X can be predicted from the formula (Sokal and Rohlf 1981):

$$r_{zx}=(S_x+r_{yx}S_y)/\sqrt{(S_y^2+2r_{yx}S_yS_x+S_x^2)}, \quad (1)$$

where S^2 and r denote variance and correlation coefficient, respectively. In cases where X and Y are known to be independent (so that $r_{xy}=0$), Eq. 1 reduces to:

$$r_{zx}=1/\sqrt{(1+(S_y^2/S_x^2))} \quad (2)$$

demonstrating that the correlation depends on the ratio of the variances of X and Y. Is the correlation between Z and X spurious? According to the definitions given in

* A contribution to the Limnology Research Centre of McGill University

** *Current address and address for offprint requests:* Département des Sciences Biologiques. Université du Québec à Montréal, Case Postale 8888, succursale "A", Montréal, Québec, Canada H3C 3P8
*** *Current address:* Hawaii Institute of Geophysics, University of Hawaii, 1000 Pope Road, Honolulu, HI 96822, USA

Kenney (1982), the correlation between Z and X is entirely spurious if $r_{xy}=0$ since r_{zx} is then completely dependent on the presence of X in the definition of Z. Similarly, when $|r_{xy}|>0$, the relationship is only partially spurious. Equations analogous to Eqs. 1 and 2 have been developed for situations where $Z=X-Y$, $Z=X\cdot Y$, or $Z=X/Y$, and also for more complex cases where, for example, only a component of the correlated variables is common to both.

The so-called 'spurious correlation' problem

The thrust of the present criticism has been touched on repeatedly in the statistical literature. For example, in a brief passage by Kendall and Stuart (1973) about Pearson's original article (Pearson 1897), they pointed out that, even in a case where the correlation between two ratios is due entirely to their sharing a common denominator, the term 'spurious' is "inapt if one is fundamentally interested in the ratios" (Kendall and Stuart 1973, p. 327–328). Similarly, Sokal and Rohlf (1981, p. 578) suggested that correlations between parts and wholes are "not really" spurious, but are logical consequences of particular variable formulations. They suggested that there is no theoretical reason for avoiding such calculations, as long as the formulation is deliberate and well-considered. Kuh and Meyer (1955) made the point that questions of spuriousness in correlations "quite obviously [do] not arise" when the hypothesis under examination has been formulated in terms of ratios. They consider the phrase 'spurious correlation' to be of "historical interest only". Long (1980), a sociological methodologist, offers probably the most thorough analysis on the subject and argues that "this belief [in the potential spuriousness of correlation among ratios sharing a common denominator], despite its intuitive appeal, is groundless." Many ecologists, however, have been less careful and have argued that in a case of spurious correlation, the correlation exists but has little meaning because of its mathematical necessity (Kenney 1982): one might have predicted the existence of the relationship based on a knowledge of the relationship between the components.

The equation of mathematical necessity with meaninglessness, then, is the first aspect of the problem we would like to address, starting with an example. Replace variables in the model described above with liver weight (X) and total human body weight (Z). Y is then the weight of the body free of liver. It is impossible, short of sheer luck, to predict a priori what the empirical relationship (and its correlation) between liver weight and total body weight will be. It could be positive, negative, or zero. Note that knowing, for example, that liver weight (X) and liver-free body weight (Y) are uncorrelated would not help matters much. It would imply only that the overall correlation, if different from zero, is non-negative. From Eqs. 1 and 2, both the correlation between X and Y and their variances are required to derive the needed correlation.

For the sake of argument, however, imagine a situation wherein these values are known. Then the correlation between liver weight and total body weight in humans would be completely predictable. However, this prediction would be merely an indirect way of deriving the relationship of interest. It does not in any way threaten the relationship's validity since the fundamental interest here lies in the allometry of the liver. It would be very odd indeed to examine the relationship between liver weight and the weight of the body minus its liver in order to find out about the correlation between liver and total body weight. Yet, this is precisely the prescription offered to avoid the pitfalls of spurious correlation (Kenney 1982). As common sense would suggest, there is nothing wrong with examining the relationship between liver and body weight directly, since these are the variables of interest. Aside from the usual assumptions inherent to correlation analysis (i.e. bivariate normality, homoscedasticity, linearity, random sampling) the only important questions in interpreting a correlation are to decide (A) whether the variables represent intelligible concepts and (B) whether they are the concepts of interest. Legitimate arguments can arise over the appropriateness of the concepts, i.e. the variable formulations, for the problem at hand. Once these are agreed upon, however, correlation analysis is always, in some sense, informative.

A second point contributed to the confusion over the spurious correlation question, particularly about correlations involving ratios. The confusion results from the failure to distinguish between spurious correlation and what might be appropriately termed 'spurious inference'. Ratio variables are derived often in an attempt to standardize measurements, or deflate the influence of an extraneous variable on two absolute measurements. The procedure is legitimate but the ratios cannot, mathematically or conceptually, be equated to the original measurements. For example, to consider the correlation between ratios to be representative of the relationship between numerators alone is simply faulty reasoning and this represents a case of spurious inference. An example of spurious inference in ecology is the interpretation of the relationship between bacterial abundance and organic matter content in lake sediments (e.g. Rublee 1982). Both bacterial numbers and organic matter are expressed per unit dry weight of sediment and therefore produce ratio variables sharing a common denominator. The difficulty here is that the relationship was thought to express an association between bacteria and organic matter alone without regard to other sediment characteristics. However, the hypothesis had been mathematically formulated in terms involving the amount of sediments. The correlation is not spurious, only its interpretation was. In the preceding terminology, the flaw here was that these ratios were not the variables of interest, only the numerators were. Unfortunately, concern appropriate to instances of mistaken inference has led too often to the inappropriate and overzealous rejection of legitimate correlations as spurious.

Concepts, priority of concepts, and spurious correlation: the self-thinning rule

The roots of the notion of spurious correlation go deeper than a simple confusion between spurious correlation and faulty inference. A proper account of the spurious correlation problem must consider the more fundamental question of how we choose the variables we measure. That this is so may be most clearly demonstrated with examples from the ecological literature. A useful model case from which to describe the interplay among concept definition, familiarity and spurious correlations is the self-thinning rule for plants. The rule describes the broad trend existing between the weight of individual plants in crowded monospecific stands and their density. Also called the $-3/2$ power law, it is one of the few widely recognized, quantitative generalities of plant ecology. However, the mathematical form in

which the self-thinning rule has commonly been expressed was recently claimed to be potentially spurious and "statistically invalid" (Weller 1987). In a thorough review of the literature pertaining to self-thinning, Weller (1987) raised a number of legitimate questions about the solidity of the existing evidence supporting the rule. His arguments regarding the statistical validity of the rule, however, are questionable in the light of the comments made above, and it is those arguments that we are interested in.

The self-thinning rule usually takes the form:

$$\log W = \alpha + \beta \log D \tag{3}$$

where W is the mean weight of individual plants and D, the density of the stand (ind./m^2). For practical reasons, W is often determined by harvesting whole quadrats of monospecific stands, weighing the total biomass (B) and dividing it by the number of plants in the quadrat (which is the density D). Thus $W = B/D$, and the self-thinning rule is the relationship (on a double log scale) between B/D and D. Not surprisingly, the relationship is considered potentially spurious (Weller 1987) since D appears on both sides of the equation, i.e. D is common to both correlated variables. The usual remedy in such cases is to restrict the analysis to an examination of the relationship between, in this particular case, B (the biomass) and D (the density) in the hope of removing the shared component. The difficulty here is that the new relationship is clearly equally "spurious" since, by definition, $B = W \cdot D$ and thus the relationship between B and D is equivalent to that between $W \cdot D$ and D. Note that the question of whether W is determined from B/D or measured directly (by weighing each individual plant) is irrelevant since the two estimates, ignoring weighing and counting errors (considered below), must be numerically the same. In this case, the two relationships ($B-D$ and $W-D$) are derivable from one another because of the definitional equation linking the three concepts (B, W, and D). Note also that this derivability does not mean that the two relationships are equivalent. They are distinct, although related, empirical patterns both of which are appropriately described by conventional correlation and regression analysis. When applied to the self-thinning rule, this means that one does not have to find significant correlations for both the weight-density and biomass-density relationships to infer competition in crowded stands, especially if the competition hypothesis is stated in terms of the effect of crowding on individual plant weight (Gorham 1979).

A further point needs to be addressed with regard to the self-thinning rule. What are the effects of measurement errors on the relationships? The answer depends on the particular variable that is measured with error. In cases where biomass (and consequently weight if weight is calculated as B/D) is measured with considerable error, the correlation and its statistical significance of both the weight-density and biomass-density relationships will be reduced. This is because the total variance in biomass (or mean weight) will be the sum of the population and measurement error variances. Although this can weaken the power of statistical tests to detect significant correlations for both relationships, this problem has no bearing on the spurious correlation question *per se*. On the other hand, in cases where large *measurement* errors are made in estimating plant density, and where weight is derived from biomass and density, the correlation between weight and density can be artificially increased. It must be emphasized that the problem here is not that the same term appears on both sides of the equation but rather that both sides share a common measurement error term. Although this is a problem we must be aware of, it is in practice likely to be negligible in the self-thinning rule case as it depends on the relative magnitude of the population and error variance terms. Measurement errors in the density of individual stands are usually small compared to the total range of densities sampled. Incidentally, there appears to be an unspoken tendency to treat measurement error variance and population variance (so-called natural variability) in the same way. It must be understood that only measurement error variance is potentially problematic. The reader is referred to Long (1980) for a more detailed discussion of the effects of measurement errors on correlations.

The self-thinning rule is not a peculiar or rare example. Another familiar case involves the allometry of metabolism in animals. Respiratory metabolic rate (oxygen consumed per individual per unit time) scales as the 0.75 power of body mass (e.g. Peters 1983). It is standard procedure to calculate a related relationship between weight-specific metabolic rate (oxygen consumed per unit body mass per unit time) and body mass, by dividing each individual's total metabolism per time by its body mass. This nominally "spurious" relationship, which scales as the -0.25 power of body mass, is predictable from the former relationship, and the dependent variables are related by definition. Because we are comfortable with both the concept of total and of weight-specific metabolism, however, we should have no trouble accepting these as related, but distinct and equally viable patterns.

In summary, the claim that the correlation between variables sharing a common term is spurious is a pervasive and unfortunate misconception within the ecological literature. The correlation between such composite variables is always legitimate provided: 1) they satisfy the assumptions of correlation analysis, 2) the variables are meaningful, that is, they represent the concepts of interest and not just a component of them, and 3) the variables do not share a large measurement error term. We hope these criteria will help rehabilitate ecological conclusions and hypotheses which had been discarded too hastily.

References

Atchley WR, Gaskins CT, Anderson D (1976) Statistical properties of ratios. I. Empirical results. Syst Zool 25:137–148

Bhattacharyya GK, Johnson RA (1977) Statistical concepts and methods. John Wiley and Sons, New York

Garsd A (1984) Spurious correlation in ecological modelling. Ecol Model 23:191–201

Gorham E (1979) Shoot height, weight, and standing crop in relation to density in monospecific plant stands. Nature (London) 279:148–150

Kendall MG, Stuart A (1973) The advanced theory of statistics. Vol. 2: Inference and relationships. 3rd edition. Charles Griffin and Co., New York

Kenney BC (1982) Beware of spurious self-correlations. Water Res Res 18:1041–1048

Kuh E, Meyer JR (1955) Correlation and regression estimates when the data are ratios. Econometrica 23:400–416

Lapin LL (1980) Statistics: meaning and method. 2nd edition. Harcourt Brace Jovanovich, New York

Long SB (1980) The continuing debate over the use of ratio variables:

facts and fiction. In: Schuessler KF (ed) Sociological Methodology. Jossey-Bass Publ

Pearson K (1897) Mathematical contributions to the theory of evolution. – On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc Roy Soc (London) 60:489–498

Peters RH (1983) The ecological implications of body size. Cambridge University Press, London

Rublee PA (1982) Bacterial and microbial distribution in estuarine sediments. In: Kennedy V (ed) Estuarine comparisons. Academic Press, New York, pp 159–182

Sokal RR, Rohlf FJ (1981) Biometry. 2nd edition. Freeman and Co. San Francisco

Weller DE (1987) A reevaluation of the −3/2 power rule of plant self-thinning. Ecol Monogr 578:23–42

Yalin MS, Kamphuis JW (1971) Theory of dimensions and spurious correlation. J Hyd 9:249–265

Submitted February 13, 1989 / Accepted July 7, 1989