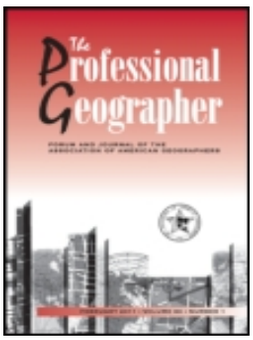

The Professional Geographer

ISSN: 0033-0124 (Print) 1467-9272 (Online) Journal homepage: http://www.tandfonline.com/loi/rtpg20

Determining Spurious Correlation between Two Variables with Common Elements: Event Area-Weighted Suspended Sediment Yield and Event Mean Runoff Depth

Peng Gao & Lianjun Zhang

To cite this article: Peng Gao & Lianjun Zhang (2016) Determining Spurious Correlation between Two Variables with Common Elements: Event Area-Weighted Suspended Sediment Yield and Event Mean Runoff Depth, The Professional Geographer, 68:2, 261-270, DOI: 10.1080/00330124.2015.1065548

To link to this article: http://dx.doi.org/10.1080/00330124.2015.1065548

Published online: 20 Aug 2015.

Determining Spurious Correlation between Two Variables with Common Elements: Event Area-Weighted Suspended Sediment Yield and Event Mean Runoff Depth

Peng Gao
Syracuse University

Lianjun Zhang
State University of New York, Syracuse

Spurious correlation is a classic statistical pitfall pervasive to many disciplines including geography. Although methods of calculating the spurious correlation between two variables possessing a common element in the form of sum, ratio, or product have been developed for a long time, controversial assertions on whether the spurious correlation should be treated or ignored are still prevalent. In this study, we examined this well-known but intriguing issue using the data representing two nonindependent variables, event area-weighted suspended sediment yield (SSY_e) and event mean runoff depth (h). By transferring the correlation between SSY_e and h to that between suspended sediment transport rate (Q_s) and water discharge (Q), we developed a new method of determining whether Q_s is truly correlated to Q. The method involves calculating coefficients of spurious correlation (r_{AB}) and the associated "pure" spurious correlation (r^0_{AB}), a hypothesis test, and regression between r_{AB} and r^0_{AB}. Our analysis showed that (1) there exists a true correlation between SSY_e and h and (2) the spurious correlation is strongly related to the variability of the variables. We then proposed a general rule stating that the apparent spurious correlation between two variables could be ignored if the two have a true causal relation. At last, we distinguished the difference between spurious correlation and spurious reference. **Key Words: sediment transport, spurious correlation, spurious inference.**

虚假相关是经典的统计陷阱，在包含地理学等诸多领域中相当普遍。儘管计算具有总和、比例或乘积形式的共同元素的两个变项之间的虚假相关之方法已建立了一段时间，但有关虚假相关应进行处理或忽略的争议主张仍然相当盛行。我们于本研究中，运用呈现事件面积权重悬浮沉积产出（SSY_e）和事件平均迳流深度（h）两个非相关变项的数据，检视此般众所週知但却引人好奇的议题。透过将 SSY_e 和 h 之间的相关性转移至悬浮沉积运输率（Q_s）和水排放（Q）之间，我们建立了一个决定 Q_s 是否真正与 Q 相关的新方法。该方法涉及计算虚假相关的系数（r_{AB}）以及有关的"纯粹"虚假相关（r^0_{AB}）、假说检定、以及 r_{AB} 和 r^0_{AB} 之间的迴归。我们的分析显示：(1) SSY_e 和 h 之间存在真正的相关性，以及 (2) 虚假相关强烈地与变项的变异性有关。我们接着提出一个普遍的法则，声明两个变项若具有真实的因果关係，那麼两者之间的明显虚假相关则可被忽略。最后，我们区辨虚假相关和虚假参考之间的差异。**关键词：沉积物运送，虚假相关，虚假推论**。

La correlación espuria es un inconveniente estadístico clásico que ocurre en muchas disciplinas, la geografía incluida. Aunque desde hace tiempo se han desarrollado métodos para calcular la correlación espuria entre dos variables que posean un elemento común en forma de suma, razón o producto, todavía son prevalentes las aseveraciones polémicas sobre si la correlación espuria deba ser tratada o ignorada. En este estudio, examinamos este asunto, tan bien conocido como intrigante, usando los datos que representan dos variables no independientes, el evento producto de sedimento en suspensión por área ponderada (SSY_e) y el evento profundidad media de la escorrentía (h). Transfiriendo la correlación entre SSY_e y h a aquella entre la tasa del transporte de sedimento suspendido (Q_s) y la descarga de agua (Q), desarrollamos un nuevo método para determinar si (Q_s) está verdaderamente correlacionado con Q. El método implica calcular los coeficientes de correlación espuria (r_{AB}) y la asociada correlación espuria "pura" (r^0_{AB}), un test de hipótesis y la regresión entre r_{AB} y r^0_{AB}. Nuestro análisis mostró que (1) existe una correlación verdadera entre SSY_e y h, y (2) la correlación espuria está fuertemente relacionada con la variabilidad de las variables. Entonces, propusimos luego una regla general declarando que la relación espuria aparente entre dos variables podría ser ignorada si las dos tienen una relación causal verdadera. Finalmente, distinguimos la diferencia entre la correlación espuria y la referencia espuria. **Palabras clave: transporte de sedimentos, correlación espuria, inferencia espuria.**

Spurious correlation refers to the correlation between two variables sharing a common element. It has been debated and discussed for more than a century in many disciplines such as ecology, anthropology, economics, earth science, and geography (Irvine and Drake 1987; Schlager et al. 1988; Kronmal 1993; Salles, Poesen, and Sempere-Torres 2002; Gani, Gani, and Abdelsalam 2007; Akkoyunlu et al. 2010). The common element typically appears in the form of sum or ratio with the prevalence of the latter in practice. Surprisingly, this old and apparently simple issue has received great controversy. The term *spurious correlation*, originally coined by Pearson (1897), was meant to alert the users that the nonzero correlation

The Professional Geographer, 68(2) 2016, pages 261–270 © Copyright 2016 by American Association of Geographers.
Initial submission, January 2015; revised submission, March 2015; final acceptance, April 2015.
Published by Taylor & Francis Group, LLC.

coefficient between two variables containing a common element (i.e., nonindependent variables) is artificial and misleading with no inference of any possible causation between the two. The alert was reiterated in a number of examples in ecology and public health to remind of the danger of using ratio variables in regression (Kronmal 1993; Jasienski and Bazzaz 1999). Brett (2004), however, concluded by analyzing the statistical relationships between nonindependent variables with sum, multiple, or ratio structures that true correlation actually exists between some nonindependent variables. Others (Prairie and Bird 1989) further contended that spurious correlation should not be considered as long as the interpretation is based on the ratio rather than part of it.

Geography shares the similar controversy, which could be highlighted by the dispute over the correlation between suspended sediment transport rate (Q_s), defined as the product of sediment concentration (*SSC*) and the associated water discharge (Q), and Q. McBean and Al-Nassri (1988) argued that the common element Q causes spurious correlation between the two and thus their correlation should be replaced by that between *SSC* and Q. By contrast, Annandale et al. (1990) supported the legitimacy of establishing the Q_s–Q relationship given that (1) there is a true *SSC*–Q correlation and (2) the established relationship is for predicting unknown Q_s values. The debate remains open (McBean and Al-Nassri 1990) and either relationship has been widely used since (Crawford 1991; Jansson 1996; Crowder, Demissie, and Markus 2007; Toor et al. 2008; Vanmaercke et al. 2010).

Because sediment transport has shown high variability both spatially and temporally, the established Q_s–Q or *SSC*–Q relationship often fails to characterize the processes of sediment transport (Gao 2008). A logical and common alternative has focused on the sum of sediment loads over a certain time period (an event or a year) to represent the lumped effect of sediment transport and relate it to various variables that imght explain the processes of sediment transport (Hicks 1994; Alexandrov et al. 2009; Zheng et al. 2012). Many such attempts involve common element(s) in both variables and thus run into the issue of spurious correlation. A well-known instance is the relationship between area-specific annual suspended sediment yield, *SSY*, defined as the ratio of annual sediment yield (*SY*) to the associated watershed area A_r and A_r (Walling 1983; de Vente et al. 2007), where A_r is a common element. Although the suggestion of replacing *SSY* by *SY* was proposed to avoid the apparent spurious correlation (Waythomas and Williams 1988; De Boer and Crosby 1996), use of the *SSY*–A_r relationship is still common (de Vente et al. 2006; de Vente et al. 2007). Another eminent case ties to a suite of studies linking event-based and area-specific suspended sediment yield (SSY_e, t/km^2) to either event peak discharge, Q_p (m^3/s), or event mean runoff depth, h, (mm; Hicks 1994; Zheng, Cai, and Cheng 2008; Duvert et al. 2012; Gao and Josefson 2012; Gao, Nearing, and Commons 2013). SSY_e and h are defined as

$$SSY_e = \frac{Q_s}{A} = \frac{a\sum_{i=1}^{n} Q_{si} \cdot t_i}{A} = \frac{a\sum_{i=1}^{n} Q_i \cdot SSC_i \cdot t_i}{A} \tag{1a}$$

$$h = \frac{V}{A} = b\frac{\sum_{i=1}^{n} Q_i \cdot t_i}{A}, \tag{1b}$$

where Q_i and SSC_i are the water discharge and sediment concentration measured at time i, respectively, t_i is the time interval between two consecutive data points, and a and b are the unit conversion factors. The common elements Q, A_r, and t in both variables suggest that SSY_e might be spuriously correlated to h.

The mathematical forms of these common elements in Equations 1a and 1b are neither sum nor ratio, however, such that the potential spurious correlation might be directly determined using the existing methods (Benson 1965; Kenney 1982; Kim 1999; Brett 2004). In this study, we developed a new method that can determine whether SSY_e is truly correlated to h. Furthermore, we showed that the degree of spurious correlation varies with the degree of variability the variables involved in SSY_e and h have and proposed a general rule of judging whether spurious correlation between two nonindependent variables should be considered. Finally, we clarified four different types of correlation and the difference between spurious correlation and spurious inference.

Methods

We compiled event-based suspended sediment and water discharge data available in twenty-two watersheds with a wide range of sizes and physical conditions: one in central New York with a humid continental climate (Gao and Josefson 2012), one in Idaho with a mixture of locally dry and humid climates (Pierson, Slaughter, and Cram 2001), four in Puerto Rico (Murphy and Stallard 2012) with a humid climate, and sixteen in Walnut Gulch, Arizona, with a semiarid climate (Gao, Nearing, and Commons 2013). For each watershed, a series of field-measured instantaneous water discharge (Q, m^3/s) and suspended sediment concentration (*SSC*, mg/1) with uneven time intervals between the data points (normally varying between and to eleven depending on durations of rainfall events and sampling frequencies) were collected for multiple storm events (Table 1). Details of data collection and compilation can be found in Gao and Josefson (2012) and Gao, Nearing, and Commons (2013). By definition, suspended sediment discharge, Q_s, was subsequently calculated using $Q_s = SSC \cdot Q$, which was followed by determining values of SSY_e and h based on Equations 1a and 1b.

Table 1 *Basic information and statistical results of the selected data sets from twenty-two watersheds*

Region	Watershed	No. of events	No. of samples	r_{XY}	r_{AB}	r^0_{AB}	R^a (%)	CV_p	CV_q	CV_q/CV_p	Hypothesis test
Puerto Rico	Cayaguas	30	411	0.9584	0.7410	0.6123	83	0.9809	1.2665	1.291161	Reject
Puerto Rico	Icacos	19	283	0.7421	0.4769	0.2700	57	0.6686	2.3843	3.566108	Reject
Puerto Rico	Canovanas	29	512	0.8880	0.6360	0.4859	76	0.9852	1.7720	1.798620	Reject
Puerto Rico	Mameyes	40	524	0.8239	0.6939	0.5089	73	0.9516	1.6098	1.691677	Reject
Idaho	Reynolds	65	14,178	0.5354	0.6790	0.4144	61	0.6723	1.4763	2.195895	Reject
New York	Oneida Creek	120	7,475	0.6028	0.7064	0.5352	76	1.0585	1.6704	1.578082	Reject
Walnut Gulch	Flume 1	75	997	0.9468	0.9526	0.9409	99	1.7390	0.6258	0.359862	Reject
Walnut Gulch	Flume 2	22	116	0.9005	0.8939	0.8162	91	1.0015	0.7090	0.707938	Reject
Walnut Gulch	Flume 3	7	36	0.9892	0.5281	0.3968	75	0.9945	2.3003	2.313022	Reject
Walnut Gulch	Flume 4	8	64	0.9700	0.9205	0.8744	95	1.6637	0.9234	0.555028	Reject
Walnut Gulch	Flume 6	102	1,702	0.9294	0.8302	0.8117	98	1.8716	1.3467	0.719545	Reject
Walnut Gulch	Flume 7	7	63	0.9915	0.9739	0.9610	99	2.0017	0.5761	0.287805	Reject
Walnut Gulch	Flume 8	8	95	0.6954	0.9599	0.9543	99	1.4235	0.4460	0.313312	Reject
Walnut Gulch	Flume 11	8	44	0.9571	0.9478	0.9436	100	1.0992	0.3857	0.350892	Not reject
Walnut Gulch	Flume 102	36	194	0.9614	0.9362	0.9395	100	0.9350	0.3408	0.364492	Not reject
Walnut Gulch	Flume 103	60	418	0.8938	0.8717	0.8726	100	1.1876	0.6647	0.559700	Not reject
Walnut Gulch	Flume 104	51	280	0.8118	0.5351	0.4878	91	1.2155	2.1752	1.789552	Not reject
Walnut Gulch	Flume 105	45	230	0.7402	0.9933	0.9932	100	1.5430	0.9739	0.631173	Reject
Walnut Gulch	Flume 106	59	368	0.8935	0.8062	0.6937	86	1.0480	1.0880	1.038168	Reject
Walnut Gulch	Flume 112	39	262	0.5555	0.8726	0.8367	96	1.4617	0.9567	0.654512	Reject
Walnut Gulch	Flume 121	57	403	0.9176	0.8442	0.7535	89	1.2762	1.1134	0.872434	Reject
Walnut Gulch	Flume 125	40	281	0.8333	0.7286	0.652	89	1.3105	1.5240	1.162915	Reject

[a] R is the ratio of r^0_{AB} to r_{AB}.

A common statistical analysis quantifying the relationship between SSY_e and h is to compute the Pearson correlation coefficient between the two. Let $Y = SSY_e$ and $X = h$; the Pearson correlation coefficient is

$$r_{XY} = \frac{\sum_{i=1}^{n} (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X - \overline{X})^2 \cdot \sum_{i=1}^{n} (Y - \overline{Y})^2}}. \quad (2)$$

Unfortunately, the value of r_{XY} does not reflect their true correlation, because SSY_e and h share the common elements of A_r, Q_i, and t_i. One of them, A_r, appeared in the denominator of both variables and so can be canceled out in Equation 2 and hence has no impact on the magnitude of r_{XY}. Pearson (1897) provided an approximate formula for computing the spurious correlation coefficient between two variables A and B, each of which can be expressed as a ratio or product of two elements $A = p/q$ and $B = t/s$, where p, q, s, and t are four elements, as follows (i.e., equation 8 in Kenney [1982]):

$$r_{AB} = \frac{r_{pr}C_pC_r - r_{ps}C_pC_s - r_{qr}C_qC_r + r_{qs}C_qC_s}{\sqrt{C_p^2 + C_q^2 - 2r_{pq}C_pC_q} \cdot \sqrt{C_r^2 + C_s^2 - 2r_{rs}C_rC_s}}, \quad (3)$$

where r is the Pearson correlation coefficient between any two elements, and $C = S/M$ is the coefficient of variation of an element where S and M are the standard deviation and mean of the element, respectively.

By definition, SSY_e is a function of sediment load Q_{si}, and Q_{si} is a function of water discharge Q_i and suspended sediment concentration SSC_i; that is, $SSY_e = f(Q_{si}) = f(Q_i, SSC_i)$. Similarly, the runoff depth h is a function of event total water volume V_i, which is a function of water discharge Q_i, $h = f(V_i) = f(Q_i)$, Unfortunately, Equation 3 cannot be directly applied to the two variables, SSY_e and h, because even after the area A_r is removed, they still hold the common elements Q_i and t_i in a form much more complex than a simple sum, ratio, or product, such that the elements p, q, t, and s cannot be mathematically defined. Alternatively, Equation 3 can be used to calculate the spurious correlation between Q_{si} and Q_i. The complex format of Q_i and t_i in SSY_e and h by definition suggests that the degree of spurious correlation between SSY_e and h is less than that between Q_{si} and Q_i. Thus, we propose that if the correlation coefficient between SSY_e and h is statistically related to that of Q_{si} and Q_i, the determination of the spurious correlation between Q_{si} and Q_i can be transferred to that between SSY_e and h. This proposal can be verified by examining whether there is a statistically significant linear relationship between correlation coefficients of the former and those of the latter.

For Q_{si} and Q_i, the corresponding items in Equation 3 are $A = p/q = Q_{si} = Q_i \cdot SSC_i = Q_i/(1/SSC_i)$, indicating $p = Q_i$ and $q = 1/SSC_i$ and $B = t/s = Q_s = Q_i/1$, indicating $t = Q_i$ and $s = 1$. Hence, $r_{pt} = 1$, $r_{ps} = r_{qs} = r_{rs} = 0$, $r_{pq} = r_{qt}$, $C_p = C_t$, $C_s = 0$. It follows that Equation 3 could be reduced to

$$r_{AB} = \frac{C_p - r_{pq}C_q}{\sqrt{C_p^2 + C_q^2 - 2r_{pq}C_pC_q}}. \quad (4)$$

Equation 4 is essentially the same as the one developed by Jawitz and Mitchell (2011). When all elements p, q, t, and s are uncorrelated (i.e.,

$r_{pq} = r_{pt} = r_{ps} = r_{qt} = r_{qs} = r_{ts} = 0$), Equation 4 can be further simplified to

$$r_{AB}^0 = \frac{C_p^2}{\sqrt{C_p^4 + C_q^2 C_p^2}} = \frac{1}{\sqrt{1 + \left(C_q / C_p\right)^2}}, \quad (5)$$

where C_p is the coefficient of variation of Q_i and C_q is the coefficient of variation of ($1/SSC_i$).

Statistically, r_{AB}^0 is the "pure" spurious correlation when the correlation between p (i.e., Q_i) and q (i.e., $1/SSC_i$) is zero, whereas r_{AB} is the spurious correlation between Q_{si} and Q_i that includes the possible correlation between p and q. Comparison between r_{AB} and r_{AB}^0 might show the impact of spurious correlation on the correlation between Q_{si} and Q_i. The specific magnitude of the "pure" spurious correlation (i.e., r_{AB}^0), however, varies with the form of ratios, the coefficients of variation of the elements (i.e., p, q, t, and s), and the original correlation between the elements (e.g., r_{pq}; Kenney 1982; Jackson and Somers 1991). Therefore, it is still not clear what the threshold value of the difference between r_{AB} and r_{AB}^0 is below which one can claim statistically that there is no true correlation between Q_{si} and Q_i. By investigating the spurious correlation between food price per kilocalorie (i.e., price of food per gram divided by energy density) and energy density (i.e., kilocalories per gram), Davis and Carlson (2012) provided an approach of using a one-sided t test to reveal the statistical significance of the spurious correlation. Based on it, we developed a statistical method of determining whether a correlation between two variables is completely spurious or not as follows.

Because many studies have shown that the relationship between SSC_i and Q_i is commonly nonlinear in nature (Gao 2008; Lopez-Tarazon et al. 2009; Navratil et al. 2011), it is reasonable to take a double logarithm transformation to form a log-linear regression model, such that

$$\ln(SSC_i) = \alpha + \beta \cdot \ln(Q_i) + \varepsilon, \quad (6)$$

where ln is the natural logarithm, α is the intercept coefficient, β is the slope coefficient, and ε is the error term. If $\beta = 0$, there is no relationship (zero correlation) between SSC_i and Q_i. Given that $Q_{si} = Q_i \cdot SSC_i$, we have by the rule of logarithms

$$\ln(Q_{si}) = \ln(Q_i) + \ln(SSC_i). \quad (7)$$

Substituting Equation 6 into Equation 7 yields

$$\ln(Q_{si}) = \ln(Q_i) + \alpha + \beta \cdot \ln(Q_i) + \varepsilon = \alpha + (\beta + 1) \cdot \ln(Q_i) + \varepsilon = \alpha + \lambda \cdot \ln(Q_i) + \varepsilon, \quad (8)$$

where $\lambda = (\beta + 1)$. It is clear that if there is no relationship between $\ln(SSC_i)$ and $\ln(Q_i)$ (i.e., $\beta = 0$), then a positive relationship between $\ln(Q_{si})$ and $\ln(Q_i)$ exists because $\lambda = (\beta + 1) = 1$. This is the case where the correlation between SSC_i and Q_i is completely spurious (Davis and Carlson 2012). If $0 < \beta \leq 1$, there is still a positive relationship between $\ln(Q_{si})$ and $\ln(Q_i)$ because $1 < \lambda \leq 2$. This is the case where the spurious correlation dominates the relationship between the two variables (Davis and Carlson 2012). Further, if $\beta > 1$, but $1 < \beta < \lambda$, then the increase of $\ln(Q_{si})$ is greater than that of $\ln(Q_i)$. This is the case where there exists a true correlation between the two variables. Therefore, for the range of $0 < \lambda \leq 1$, the positive relationship indicates the complete spurious correlation, which leads to a simple hypothesis test for a completely spurious relationship between $\ln(Q_{si})$ and $\ln(Q_i)$: H_0: $\lambda \leq 1$ and H_a: $\lambda > 1$. If the null hypothesis is not rejected, the relationship is completely spurious. Otherwise, the relationship is statistically genuine.

Nonetheless, for all twenty-two watersheds, if the previously mentioned method determined that the Q_s–Q correlation is statistically spurious for some but genuine for others, then we cannot conclude that Q_s is truly correlated to Q in general. For making a general conclusion, we developed using the data from all twenty-two watersheds a linear regression model between r_{AB} and r_{AB}^0 and between r_{AB} and r_{XY}. Then, we used these regression models to infer the general nature of the Q_s–Q correlation and hence the SSY_e–b correlation.

Results and Analysis

Calculated values of r_{AB} and r_{AB}^0 showed (Table 1) that the former tended to be greater than the latter. In twelve out of twenty-two selected watersheds, though, numerical values of r_{AB}^0 took more than 90 percent of those of r_{AB}. Some of them were even close or slightly greater than those of r_{AB}. It appeared that in these watersheds, spurious correlation between Q_{si} and Q_i is large enough to lead to the conclusion that Q_{si} is not actually correlated to Q_i. This, nonetheless, is misleading because the correlation coefficient between Q_{si} and Q_i is controlled by not only the true correlation between the two but also the ratio of coefficient of variance between $1/SSC$ (i.e., q) and Q (i.e., p). Examining the mathematical structure of the definitions of r_{AB} (i.e., Equation 4) and r_{AB}^0 (i.e., Equation 5; see Figure 1) provided insight into the interconnection among these variables. When $C_q/C_p < 1$ (e.g., $C_q/C_p = 0.6$ in Figure 1), values of r_{AB} do not change much (e.g., $r_{AB} = 0.8441$, 0.8102, and 0.8023 for $r_{qp} = 0.1$, 0.4, and 0.7, respectively) and are very close to that of r_{AB}^0, which is 0.8575. When $C_q/C_p > 1$ (e.g., $C_q/C_p = 1.4$ in Figure 1), values of r_{AB} become less and less as r_{qp} increases (e.g., $r_{AB} = 0.5253$, 0.3244, and 0.020 for $r_{qp} = 0.1$, 0.4, and 0.7, respectively). Therefore, comparing the numerical value of r_{AB}^0 with that of r_{AB} is not sufficient to infer whether two nonindependent variables are truly correlated with each other or not.

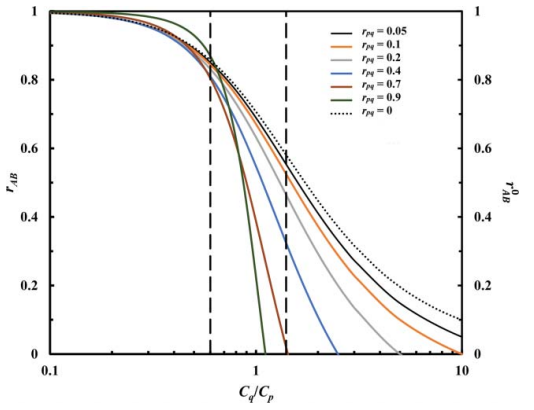

Figure 1 *The relationship based on Equation 4 among r_{AB}, r^0_{AB}, C_Q / C_P, and r_{qp}. The curve with $r_{qp} = 0$ represents Equation 5. The two vertical dashed lines represent two scenarios of different C_Q / C_P values discussed in the text. (Color figure available online.)*

The previously mentioned hypothesis testing method, which serves as a more robust tool of verifying the true correlation between Q_{si} and Q_i, signified that 82 percent of the selected watersheds have statistically significant correlation between *SSC* and *Q*, regardless of how close the value of r^0_{AB} is to that of r_{AB} (i.e., those marked as reject in Table 1). In the four watersheds where the null hypothesis was not rejected, the Q_{si}–Q_i correlation is completely spurious in the statistical sense. For three of them with C_q/C_p less than 1, the ratio of r^0_{AB} to r_{AB} reached 100 percent, whereas for the remaining one with C_q/C_p greater than 1, the ratio was 91 percent. The specific values of the r^0_{AB}/r_{AB} ratio are essentially controlled by the values of C_q/C_p and r_{qp} (Figure 1), which further signifies that the ratio cannot be used alone to judge in general whether a correlation between two nonindependent variables is true or not.

Regression analysis between r_{AB} and r^0_{AB} indicated (Figure 2) that the two types of correlation coefficients are very well correlated with each other. As $r^0_{AB} = 0$,

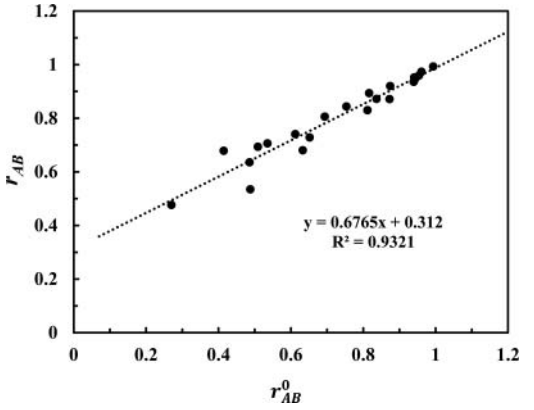

Figure 2 *The linear relationship between r_{AB} and r^0_{AB}.*

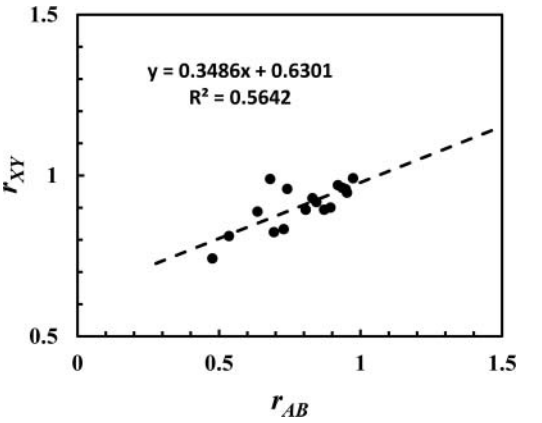

Figure 3 *The linear relationship between r_{AB} and r_{XY}.*

$r_{AB} = 0.312$, suggesting that statistically, when the spurious correlation between Q_{si} and Q_i is completely eliminated, there still exists certain genuine correlation between them, which essentially means that *SSC* is generally correlated with *Q*. This statistical implication is consistent with the physical causality between the two: Suspended sediment is predominantly transported by surface runoff. The more surface runoff (i.e., water discharge, *Q*), the more suspended sediment tends to be carried out, although the actual amount of carried suspended sediment also depends on other environmental factors, such as degrees of sediment supply, land use/cover types of uplands, and watershed topography. Therefore, Q_{si} and Q_i generally have some degree of real correlation, although the magnitude of correlation coefficient is inflated by the spurious correlation between the two due to the common element of *Q*.

Examining the relationship between r_{AB} and r_{XY} showed that the former for the data from five watersheds was even greater than the latter (Table 1), suggesting that the impact of spurious correlation on the SSY_e–*h* relationship is even less, if it does exist. For the remaining watersheds that have the values of r_{AB} less than those of r_{XY}, values of both correlation coefficients were significantly and linearly correlated to each other (Figure 3), which suggested that SSY_e values for these watersheds in general are genuinely correlated to *h*, although their r_{XY} values are affected by the spurious correlation.

The statistical characteristics of the four watersheds with complete spurious correlations (i.e., the four with not reject results in Table 1) could be further examined using examples of scatter plots of *SSC* versus *Q* (Figure 4A and 4B). When C_q/C_p was less than 1 (e.g., Flume 102 in Walnut Gulch), the trend was almost parallel to the horizontal axis, whereas when C_q/C_p was greater than 1 (i.e., Flume 104 in Walnut Gulch), no trend was discernible. Either pattern indicated that *SSC* is not statistically related to *Q*, which was consistent with the results of the hypothesis testing. Examples of two counterparts (Figures 4C and 4D),

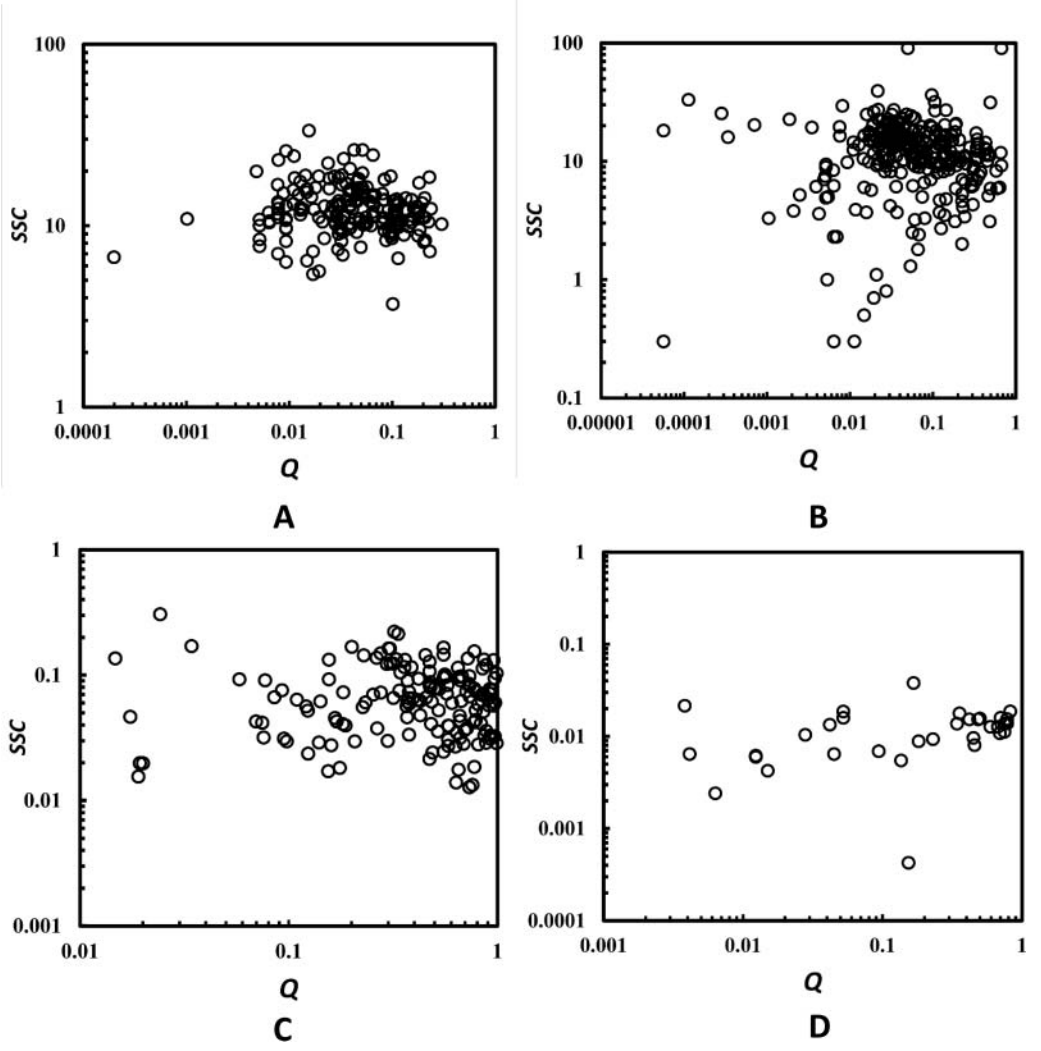

Figure 4 *Scatter plots of SSC versus Q for four watersheds: (A) Flume 102, (B) Flume 104, (C) Flume 1, and (D) Flume 3.*

however, displayed similar patterns, although statistical analysis indicated that there existed imperceptible positive trends. Geomorphologically, the four patterns shown in Figure 4 share two fundamental characteristics of suspended sediment transport in the semiarid area of Walnut Gulch. First, sediment transport is highly variable (Nearing et al. 2007). Storms of similar intensity and amount that happened in different seasons might result in a different amount of suspended sediment loads. Second, sediment transport is prone to being more limited during high flow rates due to the relatively high degree of heterogeneity in sizes of particles on the surface (Gao, Nearing, and Commons 2013). These characteristics suggest that for two similar *Q* values, the associated *SSC* values could be either similar or greatly different, as shown in Figure 4. How an *SSC–Q* pattern would be depends on the number of samples taken during one event, rainfall properties, and other associated geomorphological conditions. Therefore, these four patterns represented commonly existing types of *SSC–Q* relationships due to the complexity of the processes controlling suspended sediment transport. In other words, the four *SSC–Q* patterns were caused by the same sets of physical processes controlling suspended sediment transport and their statistical difference does not reflect the actual causation between *SSC* and *Q*. Indeed, aggregating instantaneous *SSC* and *Q* values into event-based SSY_e and *b* reflects an effort to reduce the complexity, such that suspended sediment load might be better related to its driven force, water discharge. Correlation between SSY_e and *b* is more than a statistical issue, therefore. Pragmatically, the increased correlation between the two due to the spurious correlation helps to establish better SSY_e–*b* relationships for understanding the complex transport processes.

Discussion

In general, there exist four different types of correlations (Haig 2007). The first refers to the high correlation between the two variables that have no logic causation, which is the so-called nonsense correlation (Prairie and Bird 1989). A famous example is the high correlation between the number of stork nests and child birth rates (Didelez 2007). Although the statistical result is significant, the interpretation is unreasonable. The second is spurious correlation between two variables that are independent but have common elements. This is the artificial correlation warned by Kenney (1982). The third is indirect true correlation between two variables but might or might not reflect the true causation. An example is aiming errors made during World War II bomber flights in Europe (Mosteller and Turkey 1977). Bombing accuracy had a high positive correlation with the amount of fighter opposition. The reason is that lack of fighter opposition meant lots of cloud cover obscuring bombers from the fighters and the target from the bombers and hence low accuracy. The correlation between the two is controlled by the third variable, the degree of cloud cover. Therefore, their correlation is indirect and genuine but does not indicate causation between the two variables. By contrast, the strong correlation between suspended sediment and particulate phosphorus concentrations, although it is indirect as both are controlled by water discharges, denotes the true causation between the two variables. This is because the high correlation is caused by the fact that particulate phosphorus is easy to adhere to fine soil particles and travel with them (Buck, Niyogi, and Townsend 2004; Kronvang et al. 2012). The fourth is the direct correlation between the two variables that are truly correlated even after taking out the common element—that is, the two variables are directly related to each other, although the magnitude of the correlation coefficient is inflated by the spurious correlation due to the common element in the two variables. The correlation between Q_s and Q belongs to this type.

Confusion about the spurious correlation in geomorphology and earth science mainly lies between the second and fourth types of correlation. Theoretically, whether the correlation between the two variables is true or spurious could be tested using the previously proposed approach that contains Equations 4 and 5 and the subsequent hypothesis test. Pragmatically, however, the intrinsic features of processes controlling the variables or the errors of measurement give rise to relatively high variances of the variables, such that both r_{AB} and r^0_{AB} calculated using Equations 4 and 5 are not accurate (Kenney 1982). Figure 5 further indicates that large variations of both variables (i.e., large values of C_q and C_p) tend to lead to small pure spurious correlation between the two nonindependent variables, suggesting that the correlation between two nonindependent variables is more affected by large variability of variables than the potential spurious

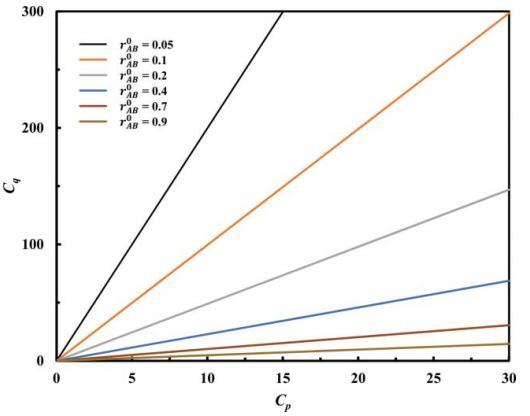

Figure 5 *The relationship among C_q, C_p, and r^0_{AB} based on Equation 5. (Color figure available online.)*

correlation between the two. Therefore, spurious correlation is not only a pure statistical pitfall (Simon 1954) but also an issue relevant to the nature of data, methods of data collection, and purposes of data analysis. In practice, we propose a general rule: If there exists physically intelligible causation between two nonindependent variables, then the spurious correlation might be negligible. In the case of the Q_s–Q or SSY_e–h correlation, it is well known that suspended sediment is transported by surface runoff and stream flows. So, there exists a cause-and-effect relationship between transported suspended sediment and surface runoff. Given that Q_s and SSY_e are two different ways of representing sediment load, and Q and h represent surface runoff, Q_s and Q or SSY_e and h are physically correlated to each other. Thus, the spurious correlation between them does not disqualify the regression analysis between the two. In the case of estimating nutrient (i.e., nitrogen and phosphorus) loads (Shivers and Moglen 2008), estimation based on the developed concentration–discharge relationship was similar to that based on the load–discharge relationship, although the latter is affected by spurious correlation. This again suggests that spurious correlation did not undermine the predictive ability of the load–discharge relationship, because nutrient movement is mainly controlled by water discharge, although other biochemical processes might also have influence.

Spurious correlation is often confused with spurious inference, which refers to interpreting the relationship of the two variables using the correlation established in terms of these two variables standardized by a common element (Prairie and Bird 1989). Spurious inference is particularly far-flung in analyzing geochemical data where two interested variables are normalized by the same element (Engle and Rowan 2013). Using trace metal data in various phases of sediment, van der Weijden (2002) demonstrated clearly that apparently high correlation between ratio variables normalized by the same element cannot be used to interpret the relationship between the original variables without normalization. Spurious inference is also prevalent when

comparing model results with the measured ones. In a study using a physically based erosion model to predict gully erosion, when modeled gully erosion volumes (V_g) were compared with the measured counterparts, the former were highly correlated to the latter, but when the modeled cross-section areas, defined as the ratio of V_g to the mean gully length (L_g), were compared with the measured ones, they were poorly correlated with each other (Nachtergaele et al. 2001). The good prediction of V_g is caused by the spurious correlation because of the common L_g in both modeled and measured V_g, which is one of the input parameters of the model. So, the model does not perform well as indicated by the authors. In bedload transport studies, the predicted bedload transport rates are often compared with the measured ones to show the performance of the adopted bedload equation. This comparison, however, is typically based on the dimensionless bedload transport rate, ϕ, defined as a quotient of volumetric bedload transport rate, q_b, by the median size of bedload grains (Reid, Powell, and Laronne 1996; Hayes, Montgomery, and Newhall 2002; Gao and Abrahams 2004). Consequently, the comparison is essentially based on the two nonindependent ratio variables, whereas the interpretation is based on the original transport rates. The statistically correct approach should be comparing the predicted bedload transport rates with the measured ones. Nonetheless, practically, variations caused by measurement errors and hydraulic processes such as limited sediment supply and heterogeneous grains dominate the predictive accuracy, such that using either ϕ or q_b does not affect model evaluation. In this case, spurious reference is not a concern.

Conclusions

Spurious correlation could affect the relationship between two nonindependent variables. Although a couple of mathematical methods have been available for more than a century, they are incapable of calculating the spurious correlation of two nonindependent variables with more complex forms than sum or ratio. We developed a new method of determining whether the two such variables are truly correlated with each other using data for event sediment yield, SSY_e, and event mean runoff depth, h. We concluded that there exists a genuine correlation between SSY_e and h. Our analysis also showed that the spurious correlation is significantly affected by the variability of the variables, making determining the spurious correlation more than a statistical issue. In many disciplines, however, such as geomorphology, geology, and hydrology, data often possess high variability because of complex natural processes reflected by the data and the inevitable sampling and monitoring errors in practice. Therefore, whether the two nonindependent variables are actually correlated with each other should be determined by a general rule: whether there exists a true causation between the two.

The correlation relationship between two independent variables sometimes is represented by the two ratio variables that are ratios of the original variables divided by a common element. Regardless of the nature of the spurious correlation between the two ratio variables, using the correlation of the latter to explain that of the former is misleading, causing the problem of spurious inference. ■

Acknowledgments

We thank Mark Nearing for providing us the original data collected from Reynolds watershed in Idaho.

Literature Cited

Akkoyunlu, S., F. R. Lichtenberg, B. B. Siliverstovs, and P. Zweifel. 2010. Spurious correlation in estimation of the health production function: A note. *Economics Bulletin* 30 (3): 2505–14.

Alexandrov, Y., H. Cohen, J. B. Laronne, and I. Reid. 2009. Suspended sediment load, bed load, and dissolved load yields from a semiarid drainage basin: A 15-year study. *Water Resources Research* 45:W08408.

Annandale, G. W., M. Demissie, W. P. Fitzpatrick, E. J. Gilroy, W. H. Kirby, T. A. Cohn, G. D. Glysson, C. F. Nordin, and K. L. Wahl. 1990. Comments on McBean, E. A., and S. Al-Nassri (1988), Uncertainty in suspended sediment transport curves. *Journal of Hydraulic Engineering* 116 (1): 140–50.

Benson, M. A. 1965. Spurious correlation in hydraulics and hydrology. *Journal of Hydraulics Division, ASCE* 91 (4): 35–42.

Brett, M. T. 2004. When is a correlation between nonindependent variables "spurious"? *Oikos* 105 (3): 647–56.

Buck, O., D. K. Niyogi, and C. R. Townsend. 2004. Scale-dependence of land use effects on water quality of streams in agricultural catchments. *Environmental Pollution* 130:287–99.

Crawford, C. G. 1991. Estimation of suspended-sediment rating curves and mean suspended-sediment loads. *Journal of Hydrology* 129:331–48.

Crowder, D. W., M. Demissie, and M. Markus. 2007. The accuracy of sediment loads when log-transformation produces nonlinear sediment load–discharge relationships. *Journal of Hydrology* 336:250–68.

Davis, G. C., and A. Carlson. 2012. How spurious is the relationship between food price and energy density? A simple procedure and statistical test. In *Agricultural and Applied Economics Association 2012 Annual Meeting, August 12–14, 2012*, ed. AAEA, 124716. Seattle, WA: AEAA.

De Boer, D. H., and G. Crosby. 1996. Specific sediment yield and discharge basin scale. *IAHS* 236:333–38.

de Vente, J., J. Poesen, M. Arabkhedri, and G. Verstraeten. 2007. The sediment delivery problem revisited. *Progress in Physical Geography* 31 (2): 155–78.

de Vente, J., J. Poesen, P. Bazzoffi, A. Van Rompaey, and G. Verstraeten. 2006. Predicting catchment sediment yield in Mediterranean environments: The importance of sediment sources and connectivity in Italian drainage basins. *Earth Surface Processes and Landforms* 31:1017–34.

Didelez, V. 2007. Statistical causality. In *Consilience interdisplinary communications 2005/2996*, ed. W. Ostreng, 114–20. Oslo, Norway: Center for Advanced Study.

Duvert, C., G. Nord, N. Gratiot, O. Navratil, E. Nadal-Romero, N. Mathys, J. Némery, et al. 2012. Towards prediction of suspended sediment yield from peak discharge in small erodible mountainous catchments (0.45–22 km^2) of France, Mexico and Spain. *Journal of Hydrology* 454–455:42–55.

Engle, M. A., and E. L. Rowan. 2013. Interpretation of Na–Cl–Br systematics in sedimentary basin brines: Comparison of concentration, element ratio, and isometric log-ratio approaches. *Mathematical Geosciences* 45 (1): 87–101.

Gani, N. D., M. R. Gani, and M. G. Abdelsalam. 2007. Blue Nile incision on the Ethiopian plateau: Pulsed plateau growth, Pliocene uplift, and hominin evolution. *GSA Today* 17 (9): 4–11.

Gao, P. 2008. Understanding watershed suspended sediment transport. *Progress in Physical Geography* 32:243–63.

Gao, P., and A. D. Abrahams. 2004. Bedload transport resistance in rough open-channel flows. *Earth Surface Processes and Landforms* 29:423–35.

Gao, P., and M. Josefson. 2012. Suspended sediment dynamics during hydrological events in a central New York watershed. *Geomorphology* 139–140:425–37.

Gao, P., M. A. Nearing, and M. Commons. 2013. Suspended sediment transport at the instantaneous and event time scales in semi-arid watersheds of southern Arizona, USA. *Water Resources Research* 49:1–14.

Haig, B. 2007. Spurious correlation. In *Encyclopedia of measurement and statistics*, ed. N. J. Salkind, 937–40. Thousand Oaks, CA: Sage.

Hayes, S. K., D. R. Montgomery, and C. G. Newhall. 2002. Fluvial sediment transport and deposition following the 1991 eruption of Mount Pinatubo. *Geomorphology* 45:211–24.

Hicks, D. M. 1994. Land-use effects on magnitude-frequency characteristics of storm sediment yields: Some New Zealand examples. *IAHS* 224:395–402.

Irvine, K. N., and J. J. Drake. 1987. Process-oriented estimation of suspended sediment concentration. *Water Resources Bulletin* 23:1017–25.

Jackson, D. A., and K. M. Somers. 1991. The spectre of "spurious" correlations. *Oecologia* 86:147–51.

Jansson, M. B. 1996. Estimating a sediment rating curve of the Reventazon river at Palomo using logged mean loads within discharge classes. *Journal of Hydrology* 183:227–41.

Jasienski, M., and F. A. Bazzaz. 1999. The fallacy of ratios and the testability of models in biology. *Oikos* 84:321–26.

Jawitz, J. W., and J. Mitchell. 2011. Temporal inequality in catchment discharge and solute export. *Water Resources Research* 47:W00J14.

Kenney, B. C. 1982. Beware of spurious self-correlations! *Water Resources Research* 18 (4): 1041–48.

Kim, J. 1999. Spurious correlation between ratios with a common diviser. *Statistics & Probability Letters* 44:383–86.

Kronmal, R. A. 1993. Sprious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society Series A, Part 3*:379–92.

Kronvang, B., J. Audet, A. Baattrup-Pedersen, H. S. Jensen, and S. E. Larsen. 2012. Phosphorus load to surface water from bank erosion in a Danish lowland river basin. *Journal of Environmental Quality* 41:304–13.

Lopez-Tarazon, J. A., R. J. Batalla, D. Vericat, and T. Francke. 2009. Suspended sediment transport in a highly erodible catchment: The River Isábena (southern Pyrenees). *Geomorphology* 109:210–21.

McBean, E. A., and S. Al-Nassri. 1988. Uncertainty in suspended sediment transport curves. *Journal of Hydraulic Engineering* 114 (1): 63–74.

———. 1990. Closure on the comments on McBean, E. A., and S. Al-Nassri (1988), Uncertainty in suspended sediment transport curves. *Journal of Hydraulic Engineering* 116 (1): 150–51.

Mosteller, F., and J. W. Turkey. 1977. *Data analysis and regression*. Reading, MA: Addison-Wesley.

Murphy, S. F., and R. F. Stallard, eds. 2012. Water quality and landscape processes of four watersheds in eastern Puerto Rico. U.S. Geological Survey Professional Paper 1789, U.S. Geological Survey, Reston, VA.

Nachtergaele, J., J. Poesen, A. Steegen, I. Takken, L. Beuselinck, L. Vandekerckhove, and G. Govers. 2001. The value of a physically based model versus an empirical approach in the prediction of ephemeral gully erosion for loess-derived soils. *Geomorphology* 40 (3–4): 237–52.

Navratil, O., M. Esteves, C. Legout, N. Gratiot, J. Nemery, S. Willmore, and T. Grangeon. 2011. Global uncertainty analysis of suspended sediment monitoring using turbidimeter in a small mountainous river catchment. *Journal of Hydrology* 398:246–59.

Nearing, M. A., M. H. Nichols, J. J. Stone, K. G. Renard, and J. R. Simanton. 2007. Sediment yields from unit-source semiarid watersheds at Walnut Gulch. *Water Resources Research* 43:W06426.

Pearson, K. 1897. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of Royal Society, London, Series A* 60:489–502.

Pierson, F. B., C. W. Slaughter, and Z. K. Cram. 2001. Long-term stream discharge and suspended-sediment database: Reynolds Creek Experimental Watershed, Idaho, United States. *Water Resources Research* 37 (11): 2857–61.

Prairie, Y. T., and D. F. Bird. 1989. Some misconceptions about the spurious correlation problem in the ecological literature. *Oecologia* 81:285–88.

Reid, I., D. M. Powell, and J. B. Laronne. 1996. Prediction of bed-load transport by desert flash floods. *Journal of Hydraulic Engineering* 122 (3): 170–73.

Salles, C., J. Poesen, and D. Sempere-Torres. 2002. Kinetic energy of rain and its functional relationship with intensity. *Journal of Hydrology* 257:256–70.

Schlager, W., D. Marsal, P. A. G. van der Geest, and A. Springer. 1988. Sedimentation rates, observation span, and the problem of spurious correlation. *Mathematical Geology* 30 (5): 547–56.

Shivers, D. E., and G. E. Moglen. 2008. Spurious correlation in the USEPA rating curve method for estimating pollutant loads. *Journal of Environmental Engineering* 134:610–18.

Simon, H. A. 1954. Spurious correlation: A causal interpretation. *Journal of the American Statistical Association* 49 (267): 467–79.

Toor, G. S., R. D. Harmel, B. E. Haggard, and G. Schmidt. 2008. Evaluation of regression methodology with low-frequency water quality sampling to estimate constituent loads for ephemeral watersheds in Texas. *Journal of Environmental Quality* 37:1847–54.

van der Weijden, C. H. 2002. Pitfalls of normalization of marine geochemical data using a common divisor. *Marine Geology* 184:167–87.

Vanmaercke, M., A. Zenebe, J. Poesen, J. Nyssen, G. Verstraeten, and J. Deckers. 2010. Sediment dynamics and the role of flash floods in sediment export from medium-sized catchments: A case study from the semi-arid tropical highlands in northern Ethiopia. *Journal of Soils and Sediments* 10:611–27.

Walling, D. E. 1983. The sediment delivery problem. *Journal of Hydrology* 65:209–37.

Waythomas, C. F., and G. P. Williams. 1988. Sediment yield and spurious correlation—Toward a better portrayal of the annual suspended-sediment load of rivers. *Geomorphology* 1:309–16.

Zheng, M. G., Q. G. Cai, and Q. J. Cheng. 2008. Modelling the runoff–sediment yield relationship using a proportional function in hilly areas of the Loess Plateau, North China. *Geomorphology* 93:288–301.

Zheng, M. G., J. S. Yang, D. L. Qi, L. Y. Sun, and Q. G. Cai. 2012. Flow–sediment relationship as functions of spatial and temporal scales in hilly areas of the Chinese loess plateau. *Catena* 98:29–40.

PENG GAO is an Associate Professor in the Department of Geography at Syracuse University, Syracuse, NY 13244. E-mail: pegao@maxwell.syr.edu. His research interests include fluvial morphology and sediment transport.

LIANJUN ZHANG is Professor in the Department of Forest and Natural Resources Management, College of Environmental Science and Forestry at the State University of New York, Syracuse, NY 13210. E-mail: lizhang@esf.edu. His research interests include forest growth and yield modeling, modeling spatial distribution of mixed-species stands, and application of statistical methods and techniques in forest growth and yield modeling.