Some examples of spurious correlation in the literature

Mike Kirkby *m.j.kirkby@leeds.ac.uk (corresponding author)*
Stephanie Bond
Joseph Holden

School of Geography, University of Leeds, UK

Several articles (Carollo et al, 2021; Di Stefano et al ,2018; Di Stefano et al, 2019 a & b; Di Stefano et al, 2020; Nicosia et al, 2019; Nicosia et al, 2020 a, b & c; Palmeri et al, 2019) have been published recently that claim to show a meaningful relationship between the measured hydraulic variables taken from a series of independent overland flow and rill flow measurements. These papers insert the experimental data into a relationship between Darcy-Weisbach roughness and other hydraulic variables, including the flow Reynolds and Froude numbers. The resulting exceptionally high correlation coefficient is claimed to show a meaningful general relationship.

It is argued here that the authors have inadvertently derived a relationship that is, instead, an example of spurious correlation (Brett, 1996) in that the two variables regressed are strongly interdependent, both, to a large extent, derived from the same underlying hydraulic variables. As a result, any reasonable (or unreasonable) set of variables chosen at random will generate the same excellent correlation, providing no new insight into the relationships between the measured variables

Spurious correlations can arise in a number of contexts, from purely coincidental relationships, through relationships where the two variables are both functionally linked to a third causal variable, to contexts where the two variables correlated have a strong functional inter-dependence. This paper is concerned with the last of these cases, in which the quality of inference that can properly be drawn from the correlation must take into account the underlying inter-dependence: the stronger the inter-dependence, the weaker the associated inference.

One frequently used example of a potentially spurious correlation is that between stream discharge and dissolved or suspended sediment load. Although these variables may be those of greatest hydrological relevance, the load is commonly calculated from independent measurements of discharge and concentration, and load calculated as their product. Gao and Zhang (2016) have explored this case, showing that there is a spectrum of inference from the discharge: load correlation.   Where concentration is almost independent of discharge, no valid inference can be drawn from the discharge:load relationship. Where concentration varies systematically with discharge, then the discharge:load relationship adds inferential value.

The papers cited in the opening paragraph all argue that there is an empirical relationship that can be used to estimate the Darcy-Weisbach roughness from other measurable hydraulic parameters.

The variables involved are:
Darcy Weisbach roughness, $f = 8\ g\ r\ s\ /\ V^2$
Reynolds Number, $Re = V.h/v$
Froude number, $F=V/\sqrt{g.h}$
Where g is the gravitational acceleration,
h is flow depth
r is hydraulic radius, = $hw/(2h+w)$ for a rectangular channel of width w
s is channel slope
V = mean flow velocity
v is the kinematic viscosity of water.

The final form of the relationship presented varies in detail between the papers, but a typical example (Di Stefano et al, 2017, equation 14) proposes

$$f = 8\left[\frac{(1+\delta).(2+\delta)}{2.465\ 2^{1-\delta}}\right]^{\frac{2}{1+\delta}}.\left(\frac{s^{0.477}}{F^{1.106}}\right)^{\frac{2}{1+\delta}} \qquad (1)$$

where δ =1.5/ln(Re). Supplementary Table 1 shows some of the variations on this form in the cited publications.

As a primary conclusion of their papers, the high correlation coefficient shown by this relationship (Nicosia et al, 2020b equation 31 – see Figure 1) is suggested to provide a valid inference. However, the definition of the Darcy-Weisbach roughness above can be re-expressed in a form that shows a strong functional dependence with the expression on the right hand side of equation (1). From the definitions, the roughness can be expressed in the form:

$$f=8s/[F^2(1+2h/w)]. \qquad (2)$$

The exponents in equation (2) can be seen to have a lot in common with those in equation (1), suggesting the possibility of a correlation that is spurious due to the lack of independence between the measured and calculated values of f shown in Figure 1.

To check on the severity of this effect and so illustrate the strength of the interdependence of the two sides of this equation, a spreadsheet has been constructed, choosing 100 sets of values, at random (uniformly distributed on a geometric scale) and independently, from the ranges shown in Table 1 for the necessary driving variables in the context of overland flow and rill flow. Other variables have been derived from these as follows:

Width, w = h ×. width-depth ratio
Hydraulic radius, r =h.w/(2h+w)
Velocity V = Re.ν/h

These variables then allow calculation of the various elements of equation (1).

The first term on the right hand side is $\left[\frac{(1+\delta).(2+\delta)}{2.465 \text{ x } 2^{1-\delta}}\right]^{\frac{2}{1+\delta}}$.
For the random set of variables chosen initially, this expression took values that ranged from 0.38 to 1.45. Combining it to give the complete left and right hand sides of equation (1), the left hand side (*f*) ranges from 2.6 x 10^{-7} to 3.91 x 10^{+5} and the right hand side from 7.3 x 10^{-7} to 5.08 x 10^{+4}. In other words, it is clear that almost all of the variability on the right hand side resides in the second term, $\left(\frac{s^{0.477}}{F^{1.106}}\right)^{\frac{2}{1+\delta}}$. The exponent 2/(1+δ) varies from 1.21 to 1.72, so that, using the median value of δ, this second term behaves as proportional to $\left(\frac{s^{0.75}}{F^{1.74}}\right)$.

There is, clearly, a strong similarity to the exact relationship $f = \frac{2grs}{V^2} = 2\frac{r}{h}\frac{s}{F^2}$, suggesting an inappropriate level of inter-dependence between the two sides of equation (1), and this is borne out by the graphical comparison (Figure 2) below, based on the 100 randomly assigned sets of variables. Repetition of choosing the random points makes only minor differences to the form of this relationship. Figure 3 illustrates the lack of any relationship between any two (here Reynolds Number and mean depth) of the four independently assigned input variables.

It is concluded that the correlations presented are spurious because there is a high degree of dependence between the expressions on either side of equation (1), so that it is not justifiable to make valid inferences from the relationship or imply any deeper understanding of the underlying processes. In different papers by this group of authors the exact exponents in equation (1) differ slightly, but the conclusions drawn here remain valid. The spreadsheet provided in the Supplementary Information allows the ranges of variables and exponents in equation (1) to be varied to confirm these conclusions.

The inference to be drawn is that expressions similar to equation (1) have no scientific merit, due to the high degree of dependence between the variables. Progress in understanding the mechanisms of overland flow and other hydrological processes should instead focus on relationships between the independently measured variables.

References cited

Brett, M. T. (2004). "When is a correlation between non-independent variables “spurious”?" Oikos **105**(3): 647-656.

Carollo, F. G., C. Di Stefano, A. Nicosia, V. Palmeri, V. Pampalone and V. Ferro (2021). "Flow resistance in mobile bed rills shaped in soils with different texture." European Journal of Soil Science.

di Stefano, C., V. Ferro, V. Palmeri and V. Pampalone (2017). "Flow resistance equation for rills." Hydrological Processes **31**: 2793-2801.

Di Stefano, C., V. Ferro, V. Palmeri and V. Pampalone (2018a). "Assessing dye-tracer technique for rillflow velocity measurements." Catena **171**: 523-532.

Di Stefano, C., V. Ferro, V. Palmeri and V. Pampalone (2018b). "Testing slope effect on flow resistance equation for mobile bed rills." Hydrological Processes **32**(5): 664-671.

Di Stefano, C., A. Nicosia, V. Palmeri, V. Pampalone and V. Ferro (2019a). "Comparing flow resistance law for fixed and mobile bed rills." Hydrological Processes **33**(26): 3330-3348.

Di Stefano, C., A. Nicosia, V. Palmeri, V. Pampalone and V. Ferro (2020). "Flow resistance law under suspended sediment laden conditions." Flow Measurement and Instrumentation **74**.

Di Stefano, C., A. Nicosia, V. Pampalone, V. Palmeri and V. Ferro (2019b). "Rill flow resistance law under equilibrium bed-load transport conditions." Hydrological Processes **33**(9): 1317-1323.

Gao, P. and L. 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.

Nicosia, A., C. Di Stefano, V. Palmeri, V. Pampalone and V. Ferro (2020a). "Flow resistance of overland flow on a smooth bed under simulated rainfall." Catena **187**.

Nicosia, A., C. Di Stefano, V. Pampalone, V. Palmeri, V. Ferro and M. A. Nearing (2019). "Testing a new rill flow resistance approach using the Water Erosion Prediction Project experimental database." Hydrological Processes **33**(4): 616-626.

Nicosia, A., C. Di Stefano, V. Pampalone, V. Palmeri, V. Ferro and M. A. Nearing (2020b). "Testing a theoretical resistance law for overland flow on a stony hillslope." Hydrological Processes **34**(9): 2048-2056.

Nicosia, A., C. Di Stefano, V. Pampalone, V. Palmeri, V. Ferro, V. Polyakov and M. A. Nearing (2020c). "Testing a theoretical resistance law for overland flow under simulated rainfall with different types of vegetation." Catena **189**.

Palmeri, V., V. Pampalone, C. Di Stefano, A. Nicosia and V. Ferro (2018). "Experiments for testing soil texture effects on flow resistance in mobile bed rills." Catena **171**: 176-184.

Table 1. Ranges of values from which independent variable values have been selected. Assignment is uniform random over the log/geometric range.

Variable	Lower limit	Upper limit
Slope, s	0.001	1.0
Mean depth (m), h	0.0001	0.1
Width-depth ratio, w/h	1	100
Reynolds Number	10	10000

Figure Captions

Figure 1. A reproduction of Figure 4 from Nicosia et al 2020b showing a comparison between Darcy-Weisbach friction factor values and those calculated by equation 31 from Nicosia et al.

Figure 2. Example of the relationship between the left and right hand sides of equation(1) for 100 randomly selected variable sets, obtained as in Table 1.

Figure 3. Example of the 100 independently assigned random variable values that have been used to calculate the terms in equation (1) and the relationship graphed in Figure 1. It can be seen that there is no meaningful relationship between the variables.