Journal of Hydrology 530 (2015) 328–335

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

The problem of self-correlation in fluvial flux data – The case of nitrate flux from UK rivers

Fred Worrall [a,*], Tim P. Burt [b], Nicholas J.K. Howden [c]

[a] Dept. of Earth Sciences, Science Laboratories, South Road, Durham DH1 3LE, UK
[b] Dept. of Geography, Science Laboratories, South Road, Durham DH1 3LE, UK
[c] Dept. of Civil Engineering, University of Bristol, Queens Building, Bristol, UK

ARTICLE INFO

Article history:
Received 6 August 2015
Received in revised form 24 September 2015
Accepted 25 September 2015
Available online 3 October 2015
This manuscript was handled by Geoff Syme, Editor-in-Chief

Keywords:
Spurious correlation
Induced correlation
Biogeochemical stationarity

SUMMARY

This study proposes a general method for testing for self-correlation (also known as spurious or induced correlation) in comparisons where there is a common variable, e.g. the comparison of the fluvial flux of a component with water yield. We considered the case of the fluvial flux of nitrate from 153 catchments from across the UK for which there were at least 10 years of data. The results show that 66% of records (102 catchments) could be rejected as significantly self-correlated ($P < 95\%$). Amongst the 51 catchments, which proved to be significantly different from the spurious, or self-correlated result, the response was variable with linear, convex, *s*-curve and mixed results proving the best description. There was no spatial pattern across the UK for the results that were and were not rejected as spurious; the most important predictor of not being self-correlated was the length of record rather than any catchment characteristic. The study shows that biogeochemical stationarity cannot be assumed and that caution should be applied when examining fluvial flux data.

1. Introduction

Significant correlations can occur in data that are entirely spurious and not related to any causal or physical relationships between the variables being compared (Kenney, 1982). Such spurious correlations (also referred to as induced or self-correlation) will occur where there is common variable in the comparison, e.g. A vs. A/B, A ∗ B vs. B or A/B vs. C/B). The strength of the self-correlation will increase where the variance in the common variable is large in comparison to that of the other or unique variables. Kenney (1982) pointed out that the reverse can also be the case: self-correlation can weaken a strong relationship when the variance of the common variable is equal to or less than that of the other or unique variable. Furthermore, spurious correlation is enhanced whenever log-transformation has been used. Interpretations based upon plots or relationships with common variables are frequently applied and their occurrence has been discussed relative to atmospheric sciences (e.g. Baas et al., 2006); is raised as an issue in geomorphological data (Gani et al., 2007); and the correlations of parts and wholes in ecology (e.g. stand biomass and tree measurements – Dean and Cao, 2003).

Several studies have proposed methods for identifying spurious correlation. The strength of relationship between two variables can be tested by standard statistical tests of the correlation coefficient. Pearson (1897), in the original study of spurious correlation, gives an approximation of the correlation coefficient that would be expected for the correlation of two variables both ratioed to a third. However, Pearson's formula is only an approximation and so tests based upon randomisation have become more common. McCuen and Surbeck (2008) observed that many environmental models (e.g. Michaelis–Menten kinetics) are calibrated by linearisation based upon plots that include a common variable; they recommend avoiding such a linearisation step and used non-linear fitting methods but provided no test of spurious correlation. Lenahan et al. (2011) discussed "induced correlations" with respect to hydrochemical data and especially the comparison of ratio data and summed data, and used randomisation to view self-correlation but, they provided no formal account of how randomisation was performed nor how to compare between the observed and the randomised data. Jackson and Somers (1991) placed randomisation tests in the same context as the null hypothesis and therefore stated that, when comparing ratioed variables, the null hypothesis used not that the regression coefficient was zero but rather the hull hypothesis was that the regression coefficient was the expected value (usually the arithmetic mean) of the distribution of regression coefficients resulting from randomisation of

∗ Corresponding author. Tel.: +44 (0)191 334 2295; fax: +44 (0)191 334 2301.
E-mail address: Fred.Worrall@durham.ac.uk (F. Worrall).

http://dx.doi.org/10.1016/j.jhydrol.2015.09.068
0022-1694/

the data, however, Jackson and Somers (1991) provided no suggestion for making this comparison. Kenney (1982) formulated the self-correlated regression coefficient for normally distributed data in the comparison A + B vs. B and Vickers et al. (2009a) used this approach to consider surface exchange of CO_2; they considered the test to be that the regression coefficient was greater than the value predicted by the formulation of Kenney (1982).

There are several problems with these approaches used to assess self-correlation; firstly, some of the above provide no formal test method at all; secondly, the formulations so far provided are either approximations or for specific comparisons; thirdly, many methods provide no formal test of the difference between the randomised results and the observed. Alternatively, we propose a different method and test for the detection of induced correlation. For the randomisation we propose that we do not assume a normal distribution and that other distributions or no distribution at all would be more appropriate for some datasets. The difference between the randomly generated and the observed data will be statistically tested and, furthermore the nature of the best-fit line will be considered relative to the randomised data.

Within hydrology, the self-correlation is evident in a number of common approaches to analysing and interpreting data. The authors of this paper have themselves previously used comparisons that we might consider upon reflection to be vulnerable to self-correlation. Worrall et al. (2008) and Worrall and Burt (2008) compared changes in annual DOC flux in a range of catchments over periods of severe droughts to the changes in annual discharge as a means of testing whether there was a biogeochemical response to drought. Worrall et al. (2012) modelled 125 years of Ca flux data by comparing it to annual discharge and, perhaps not surprisingly, annual discharge was the most important factor in explaining the Ca flux. The USEPA (2005) actually recommend using the correlation between pollutant flux and discharge in order to improve flux estimation and Shivers and Moglen (2008) provide an alternative approach on this basis. However, a strong linear relationship between component flux (e.g. nitrate) and annual discharge has been used to suggest biogeochemical stationarity (sometimes also referred to as chemiostasis, eg. Stackpoole et al., 2014) as an emergent property of catchments (Basu et al., 2010). However, Godsey et al. (2009) argued for chemiostasis on the basis of concentration discharge relationships. Basu et al. (2010) use the term biogeochemical stationarity to a low variation in concentration of a component relative to hydrological variation. This biogeochemical stationarity is taken to arise from a legacy of available material present in the catchment that means no matter which pathways water takes through a catchment the results is very similar, i.e. stationary. A strong linear relationship between component flux and discharge (annual water yield) suggests a single concentration of the component exists across a wide range of flows. This test of stationarity has been used for dissolved carbon (Giesler et al., 2014; Jantze et al., 2013).

The sediment delivery ratio (SDR) is a common approach used to explain changes in sediment flux through a catchment (Roehl, 1962; Burt and Allison, 2010). The SDR approach is vulnerable to self-correlation yet correlations based upon the variation of sediment yield with catchment area are commonly used, interpreted and discussed (e.g. Tetzlaff et al., 2013). Worrall et al. (2014) have shown that such approaches do suffer from self-correlation. In the SDR context, a negative relationship between SDR and catchment area would be predicted and so positive relationships might be thought to be free of self-correlation. Such relationships between SDR and catchment area have been observed in a number of studies (e.g. Church and Slaymaker, 1989).

Prairie and Bird (1989) in their defence of part versus whole analysis in biology warn of "not throwing the baby out with the bath water", an attitude also taken by Francey et al. (2010, 2011) in their analysis of pollution loads. Therefore, in the hope of making the most of the available information this study considers how self-correlated, or spurious correlation, can tested for and how to consider relationships in comparisons vulnerable to spurious correlation.

2. Methods

2.1. Data set

The Harmonised Monitoring Scheme (HMS) was established in 1974 to measure important hydrochemical fluxes to the North Atlantic and to allow their trends to be monitored (Simpson, 1980). These measurements met the UK's commitment to a series of international agreements and treaties (Bellamy and Wilkinson, 2001): standards and consistency of measurement over time and space are defined within the HMS programme. There are 56 HMS sites in Scotland and 214 sites in England and Wales. Monitoring sites were placed at the tidal limits of all rivers with an average annual discharge of over 2 $m^3\ s^{-1}$, with additional sites placed on major tributaries. These criteria means that there is good spatial coverage of the coast of England and Wales but in Scotland many of the west coast rivers are too small to warrant inclusion in the HMS. A range of water quality parameters are measured at these sites: pertinent to this study, the HMS measures nitrogen as nitrate and river discharge (instantaneous discharge and daily average discharge).

Monitoring is the responsibility of regional offices of the Environment Agency in England and Wales and the Scottish Environment Protection Agency in Scotland. As a result, sampling frequencies vary ranging from sub-weekly to monthly (or even less frequently in some cases). Data from any year at any site where fewer than 12 samples were collected in that year were excluded from the analysis. Consequently, although there are 270 HMS sites across Great Britain, the number of sites which could be included in any one year was variable: the distribution of sites from which data were used is shown in Fig. 1. Furthermore, because self-correlation relies on examining a correlation and so only sites with at least 10 years of annual flux data were considered.

In addition to the use of data from the HMS sites, this study considered the World's longest water quality record, the Thames at Teddington (Howden et al., 2011). Howden et al. (2011) have demonstrated the consistency and coherence of this record over the 126 years. The record at Teddington consists of monthly average nitrate concentrations since 1867 but river discharge records were only available for complete calendar years from 1883 to 2008 – 126 years of data. Therefore, the correlation between annual nitrate flux from the Thames at Teddington and the annual water yield was analysed for self-correlation in the same way as data from the HMS.

2.2. Flux calculation

For cases where data are relatively sparse, such as in much of the HMS, Littlewood et al. (1998) suggested that the product of flow weighted concentration and the annual discharge was most appropriate. However, HMS sampling is generally aperiodic and the following method (Rodda and Jones, 1983) is more appropriate:

$$F_{jy} = KA_y \sum_{1}^{N} n_y C_i Q_i \qquad (1)$$

$$n_y = \frac{A_y}{N_y} \qquad (2)$$