Limnol. Oceanogr., 57(2), 2012, 665–668

doi:10.4319/lo.2012.57.2.0665

Perils of correlating CUSUM-transformed variables to infer ecological relationships (Breton et al. 2006; Glibert 2010)

James E. Cloern,[a,*] Alan D. Jassby,[b] Jacob Carstensen,[c] William A. Bennett,[d] Wim Kimmerer,[e] Ralph Mac Nally,[f] David H. Schoellhamer,[g] and Monika Winder [h,i]

[a] U.S. Geological Survey, Menlo Park, California
[b] Department of Environmental Science and Policy, University of California, Davis, California
[c] National Environmental Research Institute, Aarhus University, Roskilde, Denmark
[d] Center for Watershed Sciences, and Bodega Marine Laboratory, University of California, Davis, Bodega Bay, California
[e] Romberg Tiburon Center, San Francisco State University, Tiburon, California
[f] Australian Centre for Biodiversity, School of Biological Sciences, Monash University, Victoria, Australia
[g] U.S. Geological Survey, Sacramento, California
[h] John Muir Institute of the Environment, Tahoe Environmental Research Center, Watershed Sciences Center, University of California, Davis, California
[i] Leibniz-Institute of Marine Sciences at Kiel University (IFM-GEOMAR), Kiel, Germany

We comment on a nonstandard statistical treatment of time-series data first published by Breton et al. (2006) in *Limnology and Oceanography* and, more recently, used by Glibert (2010) in *Reviews in Fisheries Science*. In both papers, the authors make strong inferences about the underlying causes of population variability based on correlations between cumulative sum (CUSUM) transformations of organism abundances and environmental variables. Breton et al. (2006) reported correlations between CUSUM-transformed values of diatom biomass in Belgian coastal waters and the North Atlantic Oscillation, and between meteorological and hydrological variables. Each correlation of CUSUM-transformed variables was judged to be statistically significant. On the basis of these correlations, Breton et al. (2006) developed "the first evidence of synergy between climate and human-induced river-based nitrate inputs with respect to their effects on the magnitude of spring *Phaeocystis* colony blooms and their dominance over diatoms."

Using the same approach, Glibert (2010) reported correlations between CUSUM-transformed abundances of organisms occupying many trophic levels and a range of environmental variables in the San Francisco Estuary, California. These correlations were reported to be statistically significant, and on this basis Glibert (2010) concluded that recent large population declines of diatoms, copepods, and several species of fish were responses to a single factor—increased ammonium inputs from a municipal wastewater treatment plant. The study by Breton et al. (2006) is consistent with a large body of research demonstrating the importance of climate and human activity on phytoplankton communities in Belgian coastal waters (Lancelot et al. 2007). However, Glibert's (2010) study piqued our curiosity about correlations between CUSUM-transformed variables because it contradicts the overwhelming weight of evidence that population collapses of native fish (Sommer et al. 2007) and their supporting food webs in the San Francisco Estuary are responses to multiple stressors, including landscape change, water diversions, introductions of exotic species, and changing turbidity (Bennett and Moyle 1996; Kimmerer et al. 2005; Cloern 2007; Jassby 2008; Mac Nally et al. 2010; Thomson et al. 2010). We ask here how CUSUM transformation leads to inferences about such cause-effect relationships when visual inspection of the data series (e.g., Fig. 1) shows no association between wastewater ammonium and fish abundance.

We emphasize an important distinction between the CUSUM chart and CUSUM transformation. The CUSUM chart is a well-established technique of quality assurance for industrial processes (Page 1954). The method involves keeping a running summation of the deviations of the quality of the quantity of interest (e.g., concentration of an industrial chemical) based on a sample of size n. If the quantity suddenly jumps, or gradually drifts from the specified tolerance, then a warning is raised and the process is stopped. The CUSUM chart has been used as a valuable off-line method in aquatic sciences to detect and resolve climatic (Breaker 2007) and ecological (Briceño and Boyer 2010) regime shifts, as well as departures of water-quality indicators from compliance conditions (Mac Nally and Hart 1997). In contrast, there appears to be no history for regression (or correlation) analyses on CUSUM-transformed variables prior to its use by Breton et al. (2006), and we have found no theoretical development or justification for the approach. We prove here that the CUSUM transformation, as used by Breton et al. (2006) and Glibert (2010), violates the assumptions underlying regression techniques. As a result, high correlations may appear where none are present in the untransformed data (e.g., Fig. 1). Regression analysis on CUSUM-transformed variables is, therefore, not a sound basis for making inferences about the drivers of ecological variability measured in monitoring programs. This issue is sufficiently important to warrant exploration of the approach, which we present here.

The CUSUM function

The CUSUM function is a mathematical discrete operator that transforms an input time series (x_t) to an output time series (y_t) representing the running total of the input.

* Corresponding author: jecloern@usgs.gov

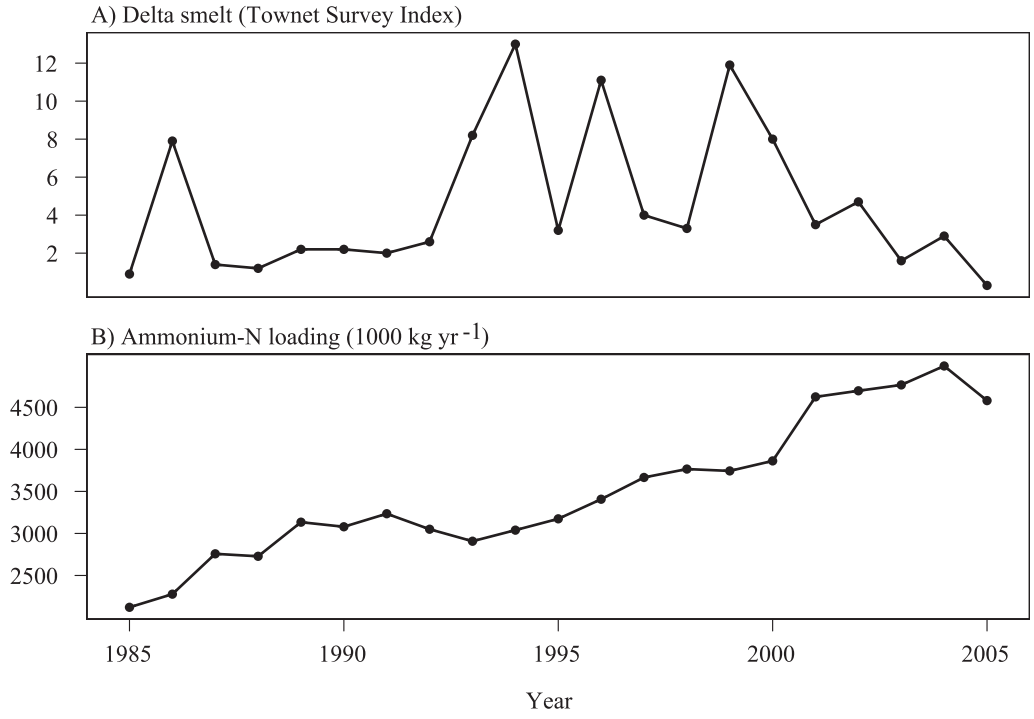

Fig. 1. Annual (A) abundance index of delta smelt (*Hypomesus transpacificus*) in the San Francisco Estuary and (B) wastewater loadings of ammonium to the Sacramento River, 1985–2005. Treatment plant data were obtained from the Sacramento Regional County Sanitation District (S. Nebozuk pers. comm., 28 July 2006). Monthly loading was calculated from discharge-weighted ammonium concentrations using the methods described by Jassby and Van Nieuwenhuyse (2005). Delta-smelt abundance data were obtained from the California Department of Fish and Game (http://www.dfg.ca.gov/delta/data/townet/indices.asp?species=3).

$$y_t = \sum_{i=1}^{t} x_i \tag{1}$$

The CUSUM function often is applied to time series of standardized residuals to detect changes in the mean of the time series (Zeileis et al. 2003; Breaker 2007). The CUSUM function changes the statistical properties of the input time series. If the standardized input time series consists of independent observations with zero mean ($E[x_t] = 0$) and variance σ^2 ($V[x_t] = \sigma^2$) then

$$E[y_t] = \sum_{i=1}^{t} E[x_i] = 0 \tag{2}$$

$$V[y_t] = \sum_{i=1}^{t} V[x_i] = t \times \sigma^2 \tag{3}$$

$$\mathrm{Cov}[y_t, y_{t-1}] = \mathrm{Cov}\left[\sum_{i=1}^{t} x_i, \sum_{i=1}^{t-1} x_i\right] = (t-1) \times \sigma^2 \tag{4}$$

$$\begin{aligned} \mathrm{Corr}[y_t, y_{t-1}] &= \frac{\mathrm{Cov}[y_t, y_{t-1}]}{\sqrt{V[y_t] \times V[y_{t-1}]}} \\ &= \frac{(t-1) \times \sigma^2}{\sqrt{t \times \sigma^2 \times (t-1) \times \sigma^2}} = \frac{t-1}{\sqrt{t \times (t-1)}} \end{aligned} \tag{5}$$

This means that the variance of the CUSUM-transformed variables and the autocovariance between two consecutive observations of the CUSUM-transformed variables both grow linearly with time and, consequently, the autocorrelation of the CUSUM-transformed variables quickly approaches 1.

Two key assumptions behind tests derived from standard regression analyses are that the observations comprising the sample are independently and identically distributed (IID). As shown above, both assumptions are violated when a random input variable is CUSUM-transformed because: the variance is not constant, so the transformed observations are not identically distributed; and the transformed observations are autocorrelated and therefore not independent of one another. Thus, applying statistical regression techniques to CUSUM-transformed time series violates the two most crucial assumptions for these tests.

CUSUM transformation inflates correlation

The CUSUM of a purely random process is a pure random walk, an example of a difference-stationary variable (because its first difference is stationary). Pfaff (2006) described the difficulty of using difference-stationary variables in regression and correlation: "In this case, the error term is often highly correlated and the t and F statistics are distorted such that the null hypothesis is rejected too often for a given critical value; hence the risk of a 'spurious regression' or 'nonsense regression' exists. Furthermore, such regressions are characterized by a high

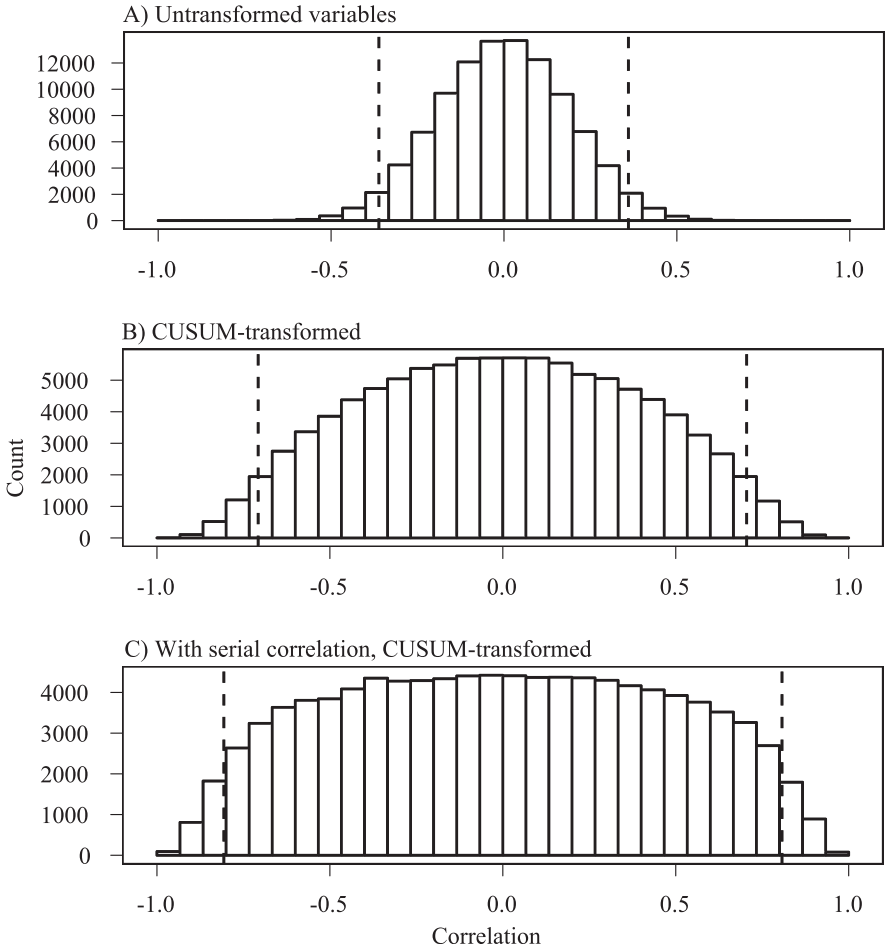

Fig. 2. (A) Frequency distribution of correlation coefficients for two independent random normal series of length 30 (n = 100,000). (B) Same as A after the samples are CUSUM-transformed. (C) Same as B, but with first-order serial correlation of 0.5 introduced into the otherwise random normal processes. Vertical dashed lines, 95% CI.

R^2." Regressions involving cumulative variables such as those produced by CUSUM transformation are classic examples of spurious regression and a well-known problem in econometrics (Hendry 1980).

To illustrate the problem more concretely, we conducted the following Monte Carlo experiment. We first generated two independent, standardized (mean 0, standard deviation 1), normal random processes of length 30, about the length of many annualized time series available from monitoring data (e.g., those analyzed by Glibert 2010). We then calculated the Pearson correlation between these two series and also between their CUSUM-transformed values. We repeated the process 100,000 times, yielding two distributions of correlation coefficients from which we generated 95% confidence intervals (CIs). The distribution of CUSUM correlations is very different from the distribution of correlations of the untransformed variables (Fig. 2). The 95% CI is (−0.36, 0.36) for the original variables (Fig. 2A), but (−0.71, 0.71) for the CUSUM-transformed variables (Fig. 2B). Thus, correlations must exceed 0.71 (instead of 0.36) for CUSUM-transformed variables to be considered significant at the $p < 0.05$ levels. This implies that the CUSUM transformation increases the probability of making a Type I error (incorrectly rejecting a null hypothesis of no correlation) from 5% to 42% when Pearson's statistics are applied. Therefore, on this basis alone, the p-values for correlations of CUSUM-transformed variables reported by Breton et al. (2006) and Glibert (2010) are incorrect.

The above experiment was based on independent random processes. Water resources data, however, commonly exhibit serial correlation (Helsel and Hirsch 2002). The introduction of serial correlation accentuates the problem by broadening the distribution of correlation coefficients even further than in the example above. To measure this effect, we repeated the simulations after introducing varying amounts of first-order serial correlation (r_1, r_2) into the paired series that otherwise represented random normal processes (using the *arima.sim* function of R; R Development Core Team 2010). This second experiment shows how the 95% CIs for the correlations broaden in proportion to the strength of serial correlation (Table 1; Fig. 2C). The presence of serial correlation thus increases the probability of making a Type I error further (53% when $r_1 = r_2 = 0.5$), making any conclusions from such correlations correspondingly less reliable. Even if a significance level of $p < 0.0001$ were used, the probability

Table 1. Upper limits of the 95% CIs for correlation between two untransformed and CUSUM-transformed random variables with different combinations of serial correlation coefficients, r_1 and r_2.

r_1	r_2	Untransformed	CUSUM-transformed
0.0	0.0	0.36	0.71
0.1	0.1	0.36	0.73
0.1	0.5	0.38	0.77
0.1	0.9	0.39	0.82
0.5	0.5	0.44	0.81
0.5	0.9	0.51	0.86
0.9	0.9	0.71	0.92

of making a Type I error (19% when $r_1 = r_2 = 0.5$) would still be much greater than 5%.

We showed that two CUSUM-transformed variables often have an apparent statistically significant correlation even if none exists between the original untransformed series. Moreover, even if a statistically significant relationship could be established between CUSUM-transformed variables, there is no proven basis for inferring relationships between the original variables. Given these difficulties, we wonder what purpose is served by CUSUM transformation for exploring relationships between two variables. As a real example, Glibert (2010) inferred a strong negative association between delta smelt abundance and wastewater ammonium from regression of CUSUM-transformed time series. However, the Pearson correlation ($r = -0.096$) between the time series (Fig. 1) is not significant, even under the naive IID assumptions ($p = 0.68$). In short, correlations between CUSUM-transformed variables should not be used as a substitute for analysis of the original untransformed variables.

References

Bennett, W. A., and P. B. Moyle. 1996. Where have all the fishes gone? Interactive factors producing fish declines in the Sacramento–San Joaquin Estuary, p. 519–542. *In* p. 519–542, J. T. Hollibaugh [ed.], San Francisco Bay: The ecosystem. American Association for the Advancement of Science.

Breaker, L. C. 2007. A closer look at regime shifts based on coastal observations along the eastern boundary of the North Pacific. Cont. Shelf Res. **27:** 2250–2277, doi:10.1016/j.csr.2007.05.018

Breton, E., V. Rousseau, J. Parent, J. Ozer, and C. Lancelot. 2006. Hydroclimatic modulation of diatom/*Phaeocystis* blooms in nutrient-enriched Belgian coastal waters (North Sea). Limnol. Oceanogr. **51:** 1401–1409, doi:10.4319/lo.2006.51.3.1401

Briceño, H. O., and J. N. Boyer. 2010. Climatic controls on phytoplankton biomass in a sub-tropical estuary, Florida Bay, USA. Estuar. Coasts **33:** 541–553, doi:10.1007/s12237-009-9189-1

Cloern, J. E. 2007. Habitat connectivity and ecosystem productivity: Implications from a simple model. Am. Nat. **169:** E21–E33, doi:10.1086/510258

Glibert, P. 2010. Long-term changes in nutrient loading and stoichiometry and their relationships with changes in the food web and dominant pelagic fish species in the San Francisco Estuary, California. Rev. Fish. Sci. **18:** 211–232, doi:10.1080/10641262.2010.492059

Helsel, D., and R. Hirsch. 2002. Statistical methods in water resources, chapter A3. *In* Techniques of water-resources investigations of the U.S. Geological Survey, Book 4: Hydrologic analysis and interpretation. U.S. Geological Survey.

Hendry, D. 1980. Econometrics: Alchemy or science? Economica **47:** 387–406, doi:10.2307/2553385

Jassby, A. D. 2008. Phytoplankton in the upper San Francisco Estuary: Recent biomass trends, their causes and their trophic significance. San Francisco Estuary and Watershed Science [accessed 2011 December 30]. Available from: http://escholarship.org/uc/item/71h077r1

———, and E. E. Van Nieuwenhuyse. 2005. Low dissolved oxygen in an estuarine channel (San Joaquin River, California): Mechanisms and models based on long-term time series. San Francisco Estuary and Watershed Science [accessed 2011 December 30]. Available from: http://escholarship.org/uc/item/0tb0f19p

Kimmerer, W., D. D. Murphy, and P. L. Angermeier. 2005. A landscape-level model for ecosystem restoration in the San Francisco Estuary and its watershed. San Francisco Estuary and Watershed Science [accessed 2011 December 30]. Available from: http://escholarship.org/uc/item/5846s8qg

Lancelot, C., N. Gypens, G. Billen, J. Garnier, and V. Roubeix. 2007. Testing an integrated river-ocean mathematical tool for linking marine eutrophication to land use: The *Phaeocystis*-dominated Belgian coastal zone (southern North Sea) over the past 50 years. J. Mar. Syst. **64:** 216–228, doi:10.1016/j.jmarsys.2006.03.010

Mac Nally, R., and B. T. Hart. 1997. Use of CUSUM methods for water-quality monitoring in storages. Environ. Sci. Technol. **31:** 2114–2119, doi:10.1021/es9609516

———, and others. 2010. Analysis of pelagic species decline in the upper San Francisco Estuary using multivariate autoregressive modeling (MAR). Ecol. Appl. **20:** 1417–1430, doi:10.1890/09-1724.1

Page, E. S. 1954. Continuous inspection schemes. Biometrika **41:** 100–115.

Pfaff, B. 2006. Analysis of integrated and cointegrated time series with R. Springer-Verlag.

R Development Core Team. 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 [accessed 2011 December 30]. Available from: http://www.r-project.org/

Sommer, T., and others. 2007. The collapse of pelagic fishes in the upper San Francisco Estuary. Fisheries **32:** 270–277, doi:10.1577/1548-8446(2007)32[270:TCOPFI]2.0.CO;2

Thomson, J. R., and others. 2010. Bayesian change point analysis of abundance trends for pelagic fishes in the upper San Francisco Estuary. Ecol. Appl. **20:** 1431–1448, doi:10.1890/09-0998.1

Zeileis, A., C. Kleiber, W. Krämer, and K. Hornik. 2003. Testing and dating of structural changes in practice. Comput. Stat. Data Anal. **44:** 109–123, doi:10.1016/S0167-9473(03)00030-6

Associate editor: Hans W. Paerl

Received: 07 March 2011
Accepted: 25 April 2011
Amended: 27 April 2011