Hydrol. Earth Syst. Sci. Discuss., 9, C2143–C2147, 2012
www.hydrol-earth-syst-sci-discuss.net/9/C2143/2012/

Hydrology and Earth System Sciences Discussions

Interactive comment on "Record extension for short-gauged water quality parameters using a newly proposed robust version of the line of organic correlation technique" *by* B. Khalil and J. Adamowski

F. Serinaldi (Referee)

francesco.serinaldi@ncl.ac.uk

Received and published: 10 June 2012

I enjoyed reading this manuscript. The Authors suggest a modified version of the LOC regression (I am more familiar with the acronym MOVE) as a simple analytical tool to perform data record extension preserving the variance of the extended records. I like such a type of simple but effective engineering solutions, especially when they are supported by well designed Monte Carlo experiments. As the editing details have been already recognized by the other Reviewer, I would like to add some remarks about the

methodology.

Specific comments

The Monte Carlo experiments reveal that the proposed RLOC provides estimates of the extreme quantiles less accurate than LOC for small datasets even though the data follow a bivariate Gaussian distribution. In my opinion, this shortcoming is due to the non-optimal use of the information contained in the data. In more detail, while the LOC slope is the ratio of standard deviations computed by the standard estimator based on n_1 and n_2 values of X and Y, respectively, the RLOC relies on only four values, that is, the first and third quartiles of X and Y. Therefore, even though RLOC is robust against outliers, it is also unavoidably more imprecise under uncertain estimates of the quartiles in small datasets. These remarks and a closer look at the KTRL estimator can suggest a straightforward modification of the RLOC that can be explored and can possibly improve the precision preserving the robust nature of the RLOC. From Eq. 10, it is rather clear that a similar relationship exists not only between the interquartile range and variance but also between every interquantile range and the variance for symmetric distributions. Therefore, instead of computing the RLOC slope by using only the interquartile range (four values), the median of the slopes resulting from a set of suitable interquantile ranges (say, $(y_{c(90)} - y_{c(10)})/(x_{c(90)} - x_{c(10)})$, $(y_{c(80)} - y_{c(20)})$ / $(x_{c(80)} - x_{c(20)})$,..., $(y_{c(70)} - y_{c(30)})$ / $(x_{c(70)} - x_{c(30)})$,...) can be used. In this way, the information stored in the data can be better used by applying the same rationale of the KTRL method, and simultaneously the estimation procedure is kept robust by discarding the outliers from the computations. Of course, the effectiveness of this approach should be checked, but in principle, it seems to be a viable trade-off between unbiasedness and accuracy.

Referring to the above comments, I think that the Authors should distinguish the hypotheses of Gaussianity and symmetry throughout the paper. While the parametric

methods require that the data follows a biviariate Gaussian distribution, the symmetry of the marginal distribution is enough for the non-parametric techniques, such as RLOC (otherwise, these methods should be referred to as "parametric"). A clear distinction is also fundamental to define an appropriate transformation of the original marginal distribution of the data. Namely, while we need that a logarithmic transformation (or whatever else) returns bivariate Gaussian data in order to apply the parametric regressions, on the other hand, we only need that the transformations simply adjust the symmetry of the marginal distributions when nonparametric RLOC is applied. Perhaps, the Authors can be interested to the work by Serinaldi et al. (2012), for a discussion on the effects of different marginal transformations and the role of the structure of dependence on the regression outcomes for skewed data. That work also points out the good performance (in terms of point and interval estimates) of a very simple weighted regression, which easily accounts for the heteroskedasticity of the errors of skewed data without applying preliminary marginal transformations. In this context, it is also worth mentioning that more refined techniques based on e.g. Generalized Linear Models (GLMs), Generalized Additive Models (GAMs) and their extensions are readily available and widely used in the industry and research. These techniques are not as simple as the closed form formulas provided in the paper under review, but I think that they must be taken into account when one requires a more refined augmentation of data that exhibit complex temporal and cross correlation patterns.

A final remark concerns the leave-one year-out cross validation used in the real world data analysis. As mentioned at P4669L12-15, water quality data show special characteristics such as seasonal patterns and autocorrelation. However, every regression technique relies on the basic hypothesis that data are time independent. Thus, I wonder if the Author can better clarify if the monthly data used in the case study show an evident seasonality (as it could be expected) and how they accounted for that. In my understanding, simple regression techniques, such as MOVE, work well for data augmentation when the signal-to-noise ratio between the seasonal pattern and the local

fluctuations is high (as for monthly streamflow data). In this context, small and high values of the variables often correspond to particular seasons or months (summer, winter, etc.). In other words, an approximate linear relationship in the scatter plots results from clusters of data corresponding to particular months or seasons. In more detail, it can be worth specifying how the paper fits e.g. into the framework of cyclic and noncyclic procedures suggested by Alley and Burns (1983) and mentioned by Hirsh (1982). A figure showing some time series as well as the corresponding scatter plots of X versus Y can help the visual understanding of the data on hand and their statistical properties along with the regression outcomes.

Editing notes

In a previous work, the first Author mentions the Sen's slope. Actually, in my understanding the Sen's regression is a generalization of the Theil's method when there are ties in the covariate. Perhaps, it is fair to cite Sen (1968).

I suggest avoiding the terms "independent" and "dependent" variables. Even though they are commonly used, perhaps, "response variable" and "explanatory variable" or "covariate" are more appropriate as X are rarely statistically independent.

References

Alley, W. M. and Burns, A. W.: Mixed-station extension of monthly streamflow records, J. Hydraul. Eng. 109, 1272-1284 (1983), http://dx.doi.org/10.1061/(ASCE)0733-9429(1983)109:10(1272).

Hirsch, R. M: A comparison of four streamflow record extension techniques, Water Resources Research, 18 (4), 1081-1088, 1982.

Serinaldi, F., Grimaldi, S., Abdolhosseini, M., Corona, P., Cimini, D.: Testing copula regression against benchmark models for point and interval estimation of tree

wood volume in beech stands, European Journal of Forest Research, 2012, DOI: 10.1007/s10342-012-0600-2.

Sen, P. K.: Estimates of the regression coefficient based on Kendall's tau, Journal of the American Statistical Association 63, 1379-1389, 1968.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 9, 4667, 2012.