

Discordant Sites Detection in the Regional Frequency Analysis by Means of Robust Distances

P. Neytchev¹ {joint work with N. Neykov, P.H.A.J.M. Van Gelder and V. Todorov}

¹ National Institute of Meteorology and Hydrology, BAS, 66 Tsarigradsko chaussee, 1784 Sofia, Bulgaria

Keywords: Regional frequency distribution analysis, minimum covariance determinant estimator, robust distances, Detection of discordant sites, Mahalanobis distance, L-moments.

Abstract

The estimation of the interval between rare events such as extreme floods, precipitation, rainstorms, droughts, high winds, or extreme pollution for a site or a group of sites is a challenging problem because the data record is often short. According to Hosking and Wallis (1997): "Regional Frequency Analysis (RFA) resolves this problem by 'trading space for time'; data from several sites are used in estimating event frequency at any one site". The approach to the RFA, developed by these authors, involves objective and subjective techniques for defining homogeneous regions, assigning of sites to regions, identifying and fitting regional probability distribution to data, and testing hypotheses about distributions using the theory of L-moments (a modification of the probability weighted moments) described by Hosking (1990).

In RFA data are assumed to come from homogeneous regions. Let Q_{ij} , $j = 1, \dots, n_i$ be observed data at N sites of a region, with sample size n_i at site i , and let $Q_i(F)$, $0 \leq F \leq 1$, be the quantile function of the distribution at site i . A region of N sites is called homogeneous if $Q_i(F) = \mu_i q(F)$, $i = 1, \dots, N$, where μ_i is the site dependent scale factor and $q(F)$ is the quantile function of the regional frequency distribution, the common distribution of the rescaled data $q_{ij} = Q_{ij}/\hat{\mu}_i$. Here $\hat{\mu}_i$ is an estimate of μ_i , for example the mean of the at site frequency distribution, but other location estimates could be considered. It is assumed that $q(F)$ is a known function that depends on p unknown parameters. The quantile estimates are given by $\hat{Q}_i(F) = \hat{\mu}_i \hat{q}(F)$, where $\hat{q}(F)$ denotes the estimated regional quantile function.

The cornerstone in RFA is the assumption that the data come from homogeneous regions. Usually the process of formation of homogeneous regions is based on the at site characteristics using objective (clustering) and subjective techniques. Once the regions of sites are formed they are subsequently evaluated by a measure of regional heterogeneity, H , developed by Hosking and Wallis (1993). A region can be regarded as homogeneous if $H < 1$, possibly heterogeneous if $1 \leq H \leq 2$, and definitely heterogeneous if $H > 2$. The H statistics, however, do not tell us anything about those sites responsible for the heterogeneity. Thus if regions of sites are identified as heterogeneous some redefinition of these regions must be made. In the screening process of atypical sites the standard discordancy measure based on Mahalanobis distance, in terms of the sample L-moment ratios (the L-CV, L-skewness and L-kurtosis) of the site's data, is recommended although as a guideline rather than a formal test by Hosking and Wallis (1997). It is well known that the Mahalanobis distance is not robust against discordant sites as it is based on the sample mean and covariance matrix of the sample L-moment ratios. Alternatives to it based on robust M-estimates or Minimum Covariance Determinant (MCD) estimator of multivariate location and scatter are superior and look promising in this particular case. Thus the robust distances based on the MCD and their one-step improvement, developed by Rousseeuw and Van Zomeren (1990), were recommended by Neykov (1998) instead off.

In this talk we will, provide an overview of the RFA based on L-moments approach, and present our results and findings concerning the performance of the classical Mahalanobis distance and robust distances based on adjusted MCD estimates as Pison et al. (2003) in heterogeneous regions using an extended Monte Carlo simulation study within the framework of the RFA.

References

- J.R.M. Hosking. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *J. R. Statist. Soc. B*, 52, 105–124, (1990).
- J.R.M. Hosking and J.R. Wallis. Some statistics useful in Regional Frequency Analysis, *Water Resour. Res.*, 29, 271–281, 1993.
- J.R.M. Hosking and J.R. Wallis. *Regional Frequency Analysis: An Approach Based on L-moments*, Cambridge University Press, 1997.
- G. Pison, S. Van Aelst and G. Willems. Small sample corrections for LTS and MCD. In: R. Dutter et al. editors, *Developments in Robust Statistics*, Physica-Verlag, Heidelberg, 330–343, 2003.
- P.J. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212–223, 1999.
- P.J. Rousseeuw and B.C. Van Zomeren. Unmasking Multivariate Outliers and Leverage Points (with discussion), *J. Amer. Statist. Assoc.*, 85, 633–651, 1990.
- N.M. Neykov. Review of "Regional Frequency Analysis: An Approach Based on L-moments" by Hosking and Wallis, *Roy. Statist. Soc. ser D*, 47, 718–719, 1998.
- P.H.A.J.M. Van Gelder, N.M. Neykov, P.N. Neytchev, J.K. Vrijling and H. Chbab. Probability Distributions of Annual Maximum River Discharges in North-western and Central Europe. In M.P. Cottam et al. editors, *Foresight and Precaution*, A.A Balkema publishing, 899–903, 2000.