WATER RESOURCES RESEARCH, VOL. ???, XXXX, DOI:10.1029/,

A comparison of homogeneity tests for regional frequency analysis

A. Viglione

Dipartimento di Idraulica, Trasporti ed Infrastrutture Civili, Politecnico di Torino, Torino, Italy

F. Laio

Dipartimento di Idraulica, Trasporti ed Infrastrutture Civili, Politecnico di Torino, Torino, Italy

P. Claps

Dipartimento di Idraulica, Trasporti ed Infrastrutture Civili, Politecnico di Torino, Torino, Italy

A. Viglione, Dipartimento di Idraulica, Trasporti ed Infrastrutture Civili, Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy. (alberto.viglione@polito.it)

Abstract. The assessment of regional homogeneity is a critical point in regional frequency analysis. To this end many homogeneity tests have been proposed, even though a general comparison among them is still lacking. Commonly used homogeneity tests, based on L-moments ratios, are considered here in a comparison with two rank tests that do not rely on particular assumptions regarding the parent distribution. The performance of these tests is assessed in a series of Monte Carlo simulation experiments. In particular, the power and Type I error of each test are determined for different scale and shape parameters of the regional parent distributions. The tests are also evaluated by varying the number of sites belonging to the region, the series length, the type of the parent distributions and the degree of heterogeneity. We find that L-moments based tests are more powerful when the samples are slightly skewed while the rank tests have better performances in case of high skewness. Based on these findings, we propose a simple method to guide the choice of the homogeneity test to be used for the different possible cases.

1. Introduction

Estimation of the frequency of extreme events is often required in the hydrological practice. The procedures for the analysis of a single set of data are well-established, but often observations of the same variable at different measuring sites are available, and more accurate conclusions can be reached by analyzing many data samples together. This constitutes the basis for regional frequency analysis [e.g., *Hosking and Wallis*, 1997]. Critical points of the regional approach to frequency analysis are in the choice of the method to group the data samples together, and in the assessment of the plausibility of the obtained groupings. This involves testing whether the proposed regions may be considered homogeneous or not. The hypothesis of homogeneity implies that frequency distributions for different sites are the same, except for a site-specific scale factor.

Many Authors have proposed homogeneity tests in the hydrologic literature, including *Dalrymple* [1960], *Wiltshire* [1986a,b,c], *Chowdhury et al.* [1991], *Lu and Stedinger* [1992], *Fill and Stedinger* [1995], and *Hosking and Wallis* [1993; 1997]. However, few comparisons have been carried out between the tests, with the effect of leaving the user without clear ideas regarding the merits and drawbacks of each method. L-moments based statistics [*Hosking and Wallis*, 1993; 1997] are nowadays routinely used in regional analyzes, but no detailed studies are available that demonstrate their superiority towards other methods. Here we compare, in a very general setting, four homogeneity tests: the first two tests, proposed by *Hosking and Wallis* [1993], are based on L-moments statistics. The other considered tests are novel in the hydrologic field: these are the k-sample Anderson-Darling test [*Scholz and Stephens*, 1987], opportunely modified to account for the normalization by the index value, and the *Durbin and Knott* [1971] test, routinely used as a goodness-of-fit test but adopted here for the heterogeneity assessment. The following Section is devoted to the description of the considered tests. In Section 3 we describe the procedure adopted for carrying out the comparison among the tests, in Section 4 the obtained results are presented, and in Section 5 some conclusions are drawn.

2. Homogeneity tests

Suppose that k samples of observations of the same variable at different measuring sites are available, and that one wishes to verify if they can be grouped to form a statistically homogeneous region: let Y_{ij} be the j-th observation in the i-th sample, sorted in ascending order ($Y_{i1} \leq Y_{i2} \leq \ldots \leq Y_{in_i}$, where $i = 1, \ldots, k$). Following an index value procedure, the observations are first rescaled with respect to a site specific index value $\overline{Y_i}$ (details on the choice of the index value are provided in Section 4.1) obtaining $X_{ij} = \frac{Y_{ij}}{\overline{Y_i}}$. If the observations are independent and the i-th rescaled sample has distribution function F_i, the homogeneity test corresponds to verifying the hypothesis $H_0 : F_1 = ... = F_k = F$, without specifying the common distribution F. The merits and drawbacks of a test statistic are evaluated by considering its power and its Type I error. Given the null hypothesis H_0 (in our case the hypothesis of regional homogeneity), the power of the test is defined as the probability of correctly rejecting H_0 when it is not true. If instead the hypothesis is rejected when it should be accepted, one makes a Type I error. The test is unbiased when the probability of making a Type I error is equal to the selected level of significance, α, of the test.

Homogeneity tests involve finding, for each site, an estimate of a quantity, θ_i, that measures some aspects of the (at-site) frequency distributions, and verifying if the dispersion of the θ_i values around their regional counterpart, θ^R, is consistent with the hypothesis of homogeneity. This requires defining the distribution of θ under the null hypothesis H_0, $G_{H_0}(\theta)$, which in many cases implies that the common distribution F is selected *a priori*. This is a theoretical problem affecting the application of many homogeneity tests (an exception is the *Wiltshire* [1986a] CV-based test). The necessity to preselect F implies that the test actually do not allow one to verify the homogeneity hypothesis alone, but the composite (homogeneity + goodness of fit) hypothesis that the parent distribution is the same at each site, and has a pre-defined mathematical form F. As a consequence, the possible reasons why the test is not passed can be either that the region is heterogeneous, or that the adopted regional probability distribution F is inadequate. We will return to this point in Section 2.2, where the Anderson-Darling test is described.

A second problem occurs as an effect of the normalization by the index value, which in some cases can distort the distribution $G_{H_0}(\theta)$ of the test statistic under the null hypothesis: this is the case, for example, of the *Wiltshire* [1986a] rank-based test or of the k-sample Anderson-Darling test. The problem will be treated in detail in Section 2.3. We now describe the four homogeneity tests selected for the comparison. The R package HOMTEST, developed to facilitate the practical application of the tests, is available at the web page `http://www.idrologia.polito.it/~alviglio/software/Rindex.htm` .

2.1. The Hosking and Wallis heterogeneity measures

The idea underlying *Hosking and Wallis* [1993] heterogeneity statistics is to measure the sample variability of the L-moment ratios and compare it to the variation that would be expected in a homogeneous region. The latter is estimated through repeated simulations of homogeneous regions with samples drawn from a four parameter kappa distribution [see *Hosking and Wallis*, 1997, pp. 202-204]. More in detail, the steps are the following:

1. With regards to the k samples belonging to the region under analysis, find the sample L-moment ratios (see *Hosking and Wallis* [1997] for details) pertaining to the i-th site: these are the L-coefficient of variation (L-CV),

$$t^{(i)} = \frac{\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\frac{2(j-1)}{(n_i-1)} - 1\right) Y_{i,j}}{\frac{1}{n_i}\sum_{j=1}^{n_i} Y_{i,j}}, \tag{1}$$

the coefficient of L-skewness,

$$t_3^{(i)} = \frac{\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\frac{6(j-1)(j-2)}{(n_i-1)(n_i-2)} - \frac{6(j-1)}{(n_i-1)} + 1\right) Y_{i,j}}{\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\frac{2(j-1)}{(n_i-1)} - 1\right) Y_{i,j}}, \tag{2}$$

and the coefficient of L-kurtosis

$$t_4^{(i)} = \frac{\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\frac{20(j-1)(j-2)(j-3)}{(n_i-1)(n_i-2)(n_i-3)} - \frac{30(j-1)(j-2)}{(n_i-1)(n_i-2)} + \frac{12(j-1)}{(n_i-1)} - 1\right) Y_{i,j}}{\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\frac{2(j-1)}{(n_i-1)} - 1\right) Y_{i,j}}. \tag{3}$$

Note that the L-moment ratios are not affected by the normalization by the index value, i.e. it is the same to use $X_{i,j}$ or $Y_{i,j}$ in Equations (1)-(3).

2. Define the regional averaged L-CV, L-skewness and L-kurtosis coefficients,

$$t^R = \frac{\sum_{i=1}^k n_i t^{(i)}}{\sum_{i=1}^k n_i} \qquad t_3^R = \frac{\sum_{i=1}^k n_i t_3^{(i)}}{\sum_{i=1}^k n_i} \qquad t_4^R = \frac{\sum_{i=1}^k n_i t_4^{(i)}}{\sum_{i=1}^k n_i} \tag{4}$$

and compute the statistic

$$V = \left\{ \sum_{i=1}^{k} n_i (t^{(i)} - t^R)^2 / \sum_{i=1}^{k} n_i \right\}^{1/2} . \tag{5}$$

3. Fit the parameters of a four-parameters kappa distribution to the regional averaged L-moment ratios t^R, t_3^R and t_4^R, and then generate a large number N_{sim} of realizations of sets of k samples. The i-th site sample in each set has a kappa distribution as its parent and record length equal to n_i. For each simulated homogeneous set, calculate the statistic

in Equation 5, obtaining N_{sim} values. On this vector of V values determine the mean μ_V and standard deviation σ_V that relate to the hypothesis of homogeneity (actually, under the composite hypothesis of homogeneity and kappa parent distribution).

4. An heterogeneity measure, which is called here HW_1, is finally found as

$$\theta_{HW_1} = \frac{V - \mu_V}{\sigma_V}. \tag{6}$$

θ_{HW_1} can be approximated by a normal distributed with zero mean and unit variance: following *Hosking and Wallis* [1997], the region under analysis can therefore be regarded as "acceptably homogeneous" if $\theta_{HW_1} < 1$, "possibly heterogeneous" if $1 \leq \theta_{HW_1} < 2$, and "definitely heterogeneous" if $\theta_{HW_1} \geq 2$. *Hosking and Wallis* [1997] suggest that these limits should be treated as useful guidelines. Even if the θ_{HW_1} statistic is constructed like a significance test, significance levels obtained from such a test would in fact be accurate only under special assumptions: to have independent data both serially and between sites, and the true regional distribution being kappa.

The θ_{HW_1} statistic measures heterogeneity only in the dispersion of the samples, since it is based solely on the differences between the sample L-CV's in the region. As such, it is insensitive to heterogeneity that arises between sites having equal L-CV but different L-skewness. *Hosking and Wallis* [1993] also give an alternative heterogeneity measure (that we call HW_2), in which V in Equation (5) is replaced by:

$$V_2 = \sum_{i=1}^{k} n_i \left\{ (t^{(i)} - t^R)^2 + (t_3^{(i)} - t_3^R)^2 \right\}^{1/2} / \sum_{i=1}^{k} n_i \ , \tag{7}$$

The test statistic in this case becomes

$$\theta_{HW_2} = \frac{V_2 - \mu_{V_2}}{\sigma_{V_2}} \ , \tag{8}$$

with similar acceptability limits as the HW_1 statistic. *Hosking and Wallis* [1997] judge θ_{HW_2} to be inferior to θ_{HW_1} and say that it rarely yields values larger than 2 even for grossly heterogeneous regions. Moreover they stress that in practice it is uncommon to have sites with equal L-CV and different L-skewness (sites with high L-skewness tend to have high L-CV too). Anyway we decided to consider also this statistic in the present paper because it is used in the most systematic and documented regional flood study available [*Robson and Reed*, 1999].

2.2. The *k*-sample Anderson-Darling test

As mentioned, the HW_1 and HW_2 heterogeneity measures suffer from the limitation that they take a kappa parent distribution, thus reverting the homogeneity test into a goodness-of-fit + homogeneity test. The kappa distribution is probably flexible enough to limit the consequences of this assumption [*Hosking and Wallis*, 1997], but the theoretical inconsistency remains. We therefore decided to propose in the comparison also tests that do not have this problem. A possible candidate could be the *Wiltshire* [1986a] CV-based test, unless it was shown by the same Author to be unreliable. Another test that does not make any assumption on the parent distribution is the Anderson-Darling (AD) rank test [*Scholz and Stephens*, 1987]. The AD test is the generalization of the classical Anderson-Darling goodness of fit test [e.g., *D'Agostino and Stephens*, 1986], and it is used to test the hypothesis that k independent samples belong to the same population without specifying their common distribution function.

The test is based on the comparison between local and regional empirical distribution functions. The empirical distribution function, or sample distribution function, is defined by $F(x) = \frac{j}{\eta}, x_{(j)} \leq x < x_{(j+1)}$, where η is the size of the sample and $x_{(j)}$ are the order statistics, i.e. the observations arranged in ascending order. Denote the empirical distribution function of the i-th sample (local) by $\hat{F}_i(x)$, and that of the pooled sample of all $N = n_1 + ... + n_k$ observations (regional) by $H_N(x)$. The k-sample Anderson-Darling test statistic is then defined as

$$\theta_{AD} = \sum_{i=1}^{k} n_i \int_{\text{all } x} \frac{[\hat{F}_i(x) - H_N(x)]^2}{H_N(x)[1 - H_N(x)]} dH_N(x) \ . \tag{9}$$

If the pooled ordered sample is $Z_1 < ... < Z_N$, the computational formula to evaluate Equation (9) is:

$$\theta_{AD} = \frac{1}{N} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{N-1} \frac{(NM_{ij} - jn_i)^2}{j(N-j)} \ , \tag{10}$$

where M_{ij} is the number of observations in the i-th sample that are not greater than Z_j. The homogeneity test can be carried out by comparing the obtained θ_{AD} value to the tabulated percentage points reported by *Scholz and Stephens* [1987] for different significance levels.

The statistic θ_{AD} depends on the sample values only through their ranks. This guarantees that the test statistic remains unchanged when the samples undergo monotonic transformations, an important stability property not possessed by HW heterogeneity measures. However, problems arise in applying this test in a common index value procedure. In fact, the index value procedure corresponds to dividing each site sample by a different value, thus modifying the ranks in the pooled sample. In particular, this has the effect of making the local empirical distribution functions much more similar to the other, providing an impression of homogeneity even when the samples are highly heterogeneous. The effect is analogous to that encountered when applying goodness-of-fit tests to distributions whose parameters are estimated from the same sample used for the test [e.g., *D'Agostino and Stephens*, 1986; *Laio*, 2004]. In both cases, the percentage points for the test should be opportunely redetermined. This can be done with a nonparametric bootstrap approach presenting the following steps:

1. Build up the pooled sample $\mathcal{S}$ of the observed non-dimensional data.

2. Sample with replacement from $\mathcal{S}$ and generate k artificial local samples, of size $n_1, \ldots, n_k$.

3. Divide each sample for its index value, and calculate $\theta_{AD}^{(1)}$.

4. Repeat the procedure for N_{sim} times and obtain a sample of $\theta_{AD}^{(j)}$, $j = 1, \ldots, N_{sim}$ values, whose empirical distribution function can be used as an approximation of $G_{H_0}(\theta_{AD})$, the distribution of θ_{AD} under the null hypothesis of homogeneity.

5. The acceptance limits for the test, corresponding to any significance level α, are then easily determined as the quantiles of $G_{H_0}(\theta_{AD})$ corresponding to a probability $(1 - \alpha)$.

We will call the test obtained with the above procedure the bootstrap Anderson-Darling test, hereafter referred to as AD.

2.3. Durbin and Knott test

The last considered homogeneity test derives from a goodness-of-fit statistic originally proposed by *Durbin and Knott* [1971]. The test is formulated to measure discrepancies in the dispersion of the samples, without accounting for the possible presence of discrepancies in the mean or skewness of the data. Under this aspect, the test is similar to the HW_1 test, while it is analogous to the AD test for the fact that it is a rank test. The original goodness-of-fit test is very simple: suppose to have a sample X_i, $i = 1, ..., n$, with hypothetical distribution $F(x)$; under the null hypothesis the random variable $F(X_i)$ has a uniform distribution in the $(0, 1)$ interval, and the statistic $D = \sum_{i=1}^{n} \cos[2\pi F(X_i)]$ is approximately normally distributed with mean 0 and variance 1 [*Durbin and Knott*, 1971]. D serves the purpose of detecting discrepancy in data dispersion: if the variance of X_i is greater than that of the hypothetical distribution $F(x)$, D is significantly greater than 0, while D is significantly below 0 in the reverse case. Differences between the mean (or the median) of X_i and $F(x)$ are instead not detected by D, which guarantees that the normalization by the index value does not affect the test.

The extension to homogeneity testing of the *Durbin and Knott* (DK) statistic is straightforward: we substitute the empirical distribution function obtained with the pooled observed data, $H_N(x)$, for $F(x)$ in D, obtaining at each site a statistic

$$D_i = \sum_{j=1}^{n_i} \cos[2\pi H_N(X_j)], \tag{11}$$

which is normal under the hypothesis of homogeneity. The statistic $\theta_{DK} = \sum_{i=1}^{k} D_i^2$ has then a chi-squared distribution with $k - 1$ degrees of freedom, which allows one to determine the acceptability limits for the test, corresponding to any significance level α. Note that the implementation of the DK test is much simpler compared to the other considered statistics.

3. Basis for test comparison

The main issue of this work is to analyze, through Monte Carlo simulations, which of the tests described in Section 2 works better, i.e. is less biased (Type I error close to the adopted significance level) and more powerful. The Monte Carlo simulation experiment requires that:

1. an artificial region is defined by providing the number of samples k, their length n (which is kept constant for all sites), the (3-parameter) parent distribution $\mathcal{P}$ used for the generation of the samples, and the regional average L-moment ratios τ^R and τ_3^R;

2. the artificial region has a known heterogeneity, with the local L-moment ratios, $\tau^{(i)}$ and/or $\tau_3^{(i)}$ varying linearly from site 1 through site k, with an overall range of variation $\Delta\tau$ and $\Delta\tau_3$ (when $\Delta\tau$ and $\Delta\tau_3$ are both equal to zero, the region is homogeneous);

3. for each site in the region, the three parameters of the parent distribution $\mathcal{P}$ are estimated from the local L-moments, and a sample of size n is generated from $\mathcal{P}$ and normalized by the index value;

4. the four homogeneity tests are applied to the obtained artificial region, after having selected a significance level α for the AD and DK tests, or an almost equivalent acceptability limit for the HW_1 and HW_2 heterogeneity measures;

5. 1000 replications of the artificial regions are generated, and each replication is separately tested for homogeneity with the four tests; the power of each test (or its Type I error) is estimated as the percentage of the 1000 replicates recognized as heterogeneous.

The comparison among the tests should be as general as possible; different values of k, n, $\mathcal{P}$, τ, τ_3, $\Delta\tau$, $\Delta\tau_3$, and α need then to be considered, which complicates the numerical simulation. In particular, the average dispersion and skewness of the samples, τ^R and τ_3^R, are very likely to relevantly affect the performances of the test. The same is true for the other parameters, but the effects on the tests of a change of, say, n is much easier to predict and therefore less interesting. For this reason we decided to consider several τ^R and τ_3^R values, i.e. to explore in our simulation experiment a large portion of the τ-τ_3 diagram. Numerical constraints to the τ and τ_3 values are given by *Hosking and Wallis*

[1997]: these are $0 \leq \tau < 1$, $-1 < \tau_3 < 1$, and $2\tau - 1 < \tau_3$ (valid for variables that can take only positive values). However, the portion of the $\tau - \tau_3$ space bounded by these constraints remains still too big in an operational perspective.

To choose tighter bounds in the $\tau - \tau_3$ space we refer to a hydrological perspective considering *Vogel and Wilson* [1996] work, who use L-moment diagrams to select a regional distribution for annual minimum, average and maximum streamflows. *Vogel and Wilson* [1996] build these diagrams for more than 1400 river basins in the continental United States. All the observed $\tau - \tau_3$ values, independently of the type of flow, occupy a bisector band of the graphic with $\tau_3 - 0.2 < \tau < \tau_3 + 0.4$ (see Figure 1) and very few points have a τ_3 larger than 0.5 or smaller than -0.1. We therefore choose to limit our investigations to the region with the following bounds (Figure 1):

$$\begin{cases} 0.1 & < \tau < & 0.6 \ , \\ -0.1 & \leq \tau_3 < & 0.5 \ , \\ \tau_3 - 0.2 & < \tau < & \tau_3 + 0.4 \ , \end{cases} \tag{12}$$

We consider all τ^R and τ_3^R pairs inside that region on a grid with a 0.1 spacing (gray points in Figure 1).

As for the other involved variables (k, n, $\mathcal{P}$, $\Delta\tau$, $\Delta\tau_3$, and α), the adopted simulation strategy involves building up a main case study, with reasonable parameter values, and then carrying out a sort of sensitivity analysis. The parameters selected for the main case study are the following: $k = 11$; $n = 30$; $\mathcal{P} \equiv$ generalized extreme value (GEV) distribution; $\alpha = 5\%$ (or, equivalently, $\theta_{HW} \leq 2$); $\Delta\tau = 0$ and $\Delta\tau_3 = 0$ for verifying the Type I error, or $\Delta\tau = 0.5\tau$ and $\Delta\tau_3 = 0$ for verifying the power of the tests (see Section 4.2). The type and degree of heterogeneity, the sample size, the number of sites in the region, the significance level, and the parent distribution are then varied once at a time (see Section 4.3), and the results are analyzed for 4 points in the central part of the $\tau - \tau_3$ diagram (points A, B, C and D in Figure 1).

4. Results

This section is divided into three parts: in the first one the choice of the index value is discussed, in the second one the main case study is described and in the third part the effects of the variation of k, n, $\mathcal{P}$, $\Delta\tau$, $\Delta\tau_3$, or α is analyzed.

4.1. Choice of the index-value

A relevant issue in regional frequency analysis, which is related to the main subject of this paper, is the choice of the index-value, i.e. of the parameter used to normalize the samples. We decided to include a specific section regarding this topic both because the choice of the index value can affect the performances of the homogeneity tests, and because we wish to raise some discussion on this important, but often neglected, topic. In the original formulation of the index-value method by *Dalrymple* [1960], the index value was intended to be the population mean. However, the passage from theory to practice involved replacing the population mean by the sample mean. As clearly pointed out by *Sveinsson et al.* [2001], this change is not trouble-free, since replacing the population mean by its sampling counterpart can produce relevant distortions in the regional frequency analysis. The induced distortions can be expected to be rather large when the sample mean is not a "good" estimator of the population mean, i.e. when it is either biased or has a large estimation variance. In those cases a possible alternative would be to use the sample median as the index value, as proposed for example by *Robson and Reed* [1999]. The advantages of this choice are described hereafter.

A numerical investigation is conducted for each simulation point in Figure 1. 100000 samples of length 30 are generated from a GEV distribution with known mean and median. The distortion of the sample estimates of the mean and median are estimated by the normalized root mean square error,

$$RMSE_{\%} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{x}_i - \mu)^2}}{\mu} \cdot 100 \ , \tag{13}$$

where μ and $\bar{x}_i$ are, respectively, the population and sample mean (or median) of each sample. The difference between the $RMSE_{\%}$ for the mean and for the median is shown in Figure 2. Where the differences are negative, the estimation of the mean by its sample counterpart is less biased than the corresponding median estimation, and the mean can therefore be regarded as a more reliable index value. It is clear from Figure 2 that the differences are almost negligible, except that in the very right part of the graph, corresponding to highly skewed samples, where the sample median performs considerably better than the sample mean. In fact, the sample median is known to be less sensitive than the sample mean to the presence of outliers, and the latter are more likely found in samples from highly skewed distributions [*Hampel*, 1974]. Overall, we believe that Figure 2 demonstrates the advantages of using the sample median as the index value when skewed parent distributions are suspected, as in flood frequency analysis studies. Similar results are obtained with distributions other than the GEV. We therefore use the sample median as the index value in the following of the paper.

4.2. Main case study

The main case study corresponds to a full analysis of the performances of the tests for all points in the τ-τ_3 diagram, with $k = 11$, $n = 30$, $\mathcal{P} \equiv$ GEV distribution and $\alpha = 5\%$

(or $\theta_{HW} \leq 2$). The Type I error of the tests is considered first, through simulation from homogeneous regions, with $\Delta\tau = 0$ and $\Delta\tau_3 = 0$. Figure 3 reports on the background (gray numbers) the percentage of regions considered heterogeneous by each test, and in the foreground (black lines) a fitted "trend-surface" whose isolines show how the Type I error varies in the $\tau - \tau_3$ space. It can be noticed that the average sample values $< t^R >$ and $< t_3^R >$ (i.e., the averages of t^R and t_3^R over the 1000 replications) can be different from their theoretical counterparts τ^R and τ_3^R, i.e. the gray numbers in Figure 3 do not precisely lie on the grid defined in Figure 1. This is due to the fact that in small samples t and t_3 are not unbiased estimators of τ and τ_3 [*Hosking and Wallis*, 1997].

None of the tests has the expected Type I error everywhere in the $\tau - \tau_3$ space. In a large part of the $\tau - \tau_3$ space the percentage of regions stated as non-homogeneous by the heterogeneity measures of Hosking and Wallis is $2 \div 4\%$; this percentage rises to $8 \div 10\%$ for high L-skewness coefficients ($t_3^R > 0.4$, Figure 3). The rank tests have a correct Type I error in the central-diagonal part of the L-moments space, while the percentage of regions mistakenly assumed as heterogeneous increases towards the borders (especially for the DK test).

Figure 4 reports the results of the tests for simulated regions whose heterogeneity is due to the different dispersion of the frequency distributions at different sites. The range of variation of the L-CV's ($\Delta\tau$) inside the region is 0.5 times the regional average L-CV (τ^R). Being $k = 11$ as before, in a region with $\tau^R = 0.2$ the samples are generated from distributions characterized by τ values respectively equal to $0.15, 0.16, 0.17, ..., 0.25$. The gray points and trend lines in Figure 4 show the power of the tests, i.e. the percentage of times when the test succeed in detecting the heterogeneity. The lack of power of the measure HW_2, as anticipated by *Hosking and Wallis* [1997], is evident. For all other tests, the power tends to be greater in the diagonal line of the $\tau - \tau_3$ space and to grow towards the upper-right corner of the investigated space. HW_1, if compared to the DK and AD tests, has a higher power in the bottom-left part of the L-moments space. In contrast, for highly skewed regions it has considerably lower power than the non-parametric tests, among which the AD test is the most powerful.

4.3. Sensitivity analysis

As mentioned in Section 3, the effect of a variation of k, n, $\mathcal{P}$, $\Delta\tau$, $\Delta\tau_3$, and α is considered in four points (A, B, C and D) located in the central part of the $\tau - \tau_3$ diagram (Figure 1), rather than through the whole diagram. As an example, we report in Figure 5 the behavior of the tests for regions whose heterogeneity is only due to the shape parameter ($\Delta\tau = 0$, $\Delta\tau_3 \neq 0$). In this case the non-parametric tests, in particular the AD test, and the Hosking and Wallis heterogeneity measure HW_2 are (obviously) more powerful than HW_1. This is particularly evident when the average shape parameter is rather large ($\tau_3^R \geq 0.2$) since for low values of τ_3^R (point A) all tests fail to detect the heterogeneity. As expected, the power of the tests increases with increasing heterogeneity, i.e. with increasing $\Delta\tau_3$.

As a second example, we show in Figure 6 the power of the tests for regions generated from different parent distributions, when the heterogeneity is only due to differences in the L-CV's ($\Delta\tau = 0.5\tau^R$). In addition to the GEV distribution, which is considered in the main case study, the other adopted 3-parameter distributions are the Generalized Logistic distribution (GL), the three-parameter Lognormal distribution (LN), the Pearson Type III distribution (P3) and the Generalized Pareto distribution (GP). The reader is referred to *Hosking and Wallis* [1997, pp. 191-208] for a description of the parametrization of these distributions and of the relations between their parameters and the L-moments. The four tests behave in a very similar manner with varying parent distribution: in point A (low skewness) the Hosking and Wallis heterogeneity measure HW_1 outperforms the non-parametric tests, while in point D (high skewness) the reverse is true. Points B and C reflect the transition between the two cases, and are characterized by a substantial equivalence of the different testing techniques. In all cases HW_2 lacks power to discriminate between homogeneous and heterogeneous regions.

The effects of a variation of the other parameters are more trivial, and the corresponding diagrams are not shown for reasons of space: the power of the tests increases with increasing number of sites k in a region and with increasing series length n. The tests are much more affected by the length of the series (n values from 10 to 100 are considered) than by the number of sites k (values from 3 to 21 have been considered). As for an increase of the degree of heterogeneity in the dispersion parameter ($\Delta\tau/\tau^R$), its effect is obviously to increase the power of the tests. The power reaches a 100% value when $\Delta\tau/\tau^R = 1$ (except that for HW_2). In all of the considered cases the HW_1 test is more powerful in points A and B, while the DK and AD tests are more powerful in points C and D. The differences in power can be relevant, under a practical viewpoint, especially for intermediate degrees of heterogeneity.

5. Discussion and conclusions

A practical problem in regional frequency analysis is the choice of a test for regional homogeneity assessment. In this paper, the Hosking and Wallis heterogeneity measures (based on L-moment ratios) are compared with the bootstrap Anderson-Darling test and with the Durbin and Knott rank test. This comparison shows that the Hosking and Wallis heterogeneity measure HW_1 (only based on L-CV) is preferable when skewness is low, while the bootstrap Anderson-Darling test should be used for more skewed regions. As for HW_2, the Hosking and Wallis heterogeneity measure based on L-CV and L-CA, it is shown once more how much it lacks power.

Our suggestion is to guide the choice of the test according to Figure 7, that we have obtained as a compromise between power and Type I error of the HW_1 and AD tests. The L-moment space is divided into two regions: if the t_3^R coefficient for the region under analysis is lower than 0.23, we propose to use the Hosking and Wallis heterogeneity measure HW_1; if $t_3^R > 0.23$, the bootstrap Anderson-Darling test is preferable. Further comments arise from the observation of Figure 7 that displays some (t^R, t_3^R) points. Each of these points is representative of a homogeneous region, considered in three flood frequency studies: *Hosking and Wallis* [1997], that directly report the t^R and t_3^R values for several regions in the Apalachian area; *De Michele and Rosso* [2002] and *Farquharson et al.* [1987], that give the three parameters of the GEV distribution (estimated using L-moments) for many regions in Italy [*De Michele and Rosso*, 2002] and around the world [*Farquharson*, 1987]. Note that, as expected, these empirical regions lay in the part of the parameter space that was considered in our simulations. Also note that the majority of the points belong to the upper-right region of $\tau - \tau_3$ space, where the bootstrap Anderson-Darling test is more powerful.

The good performances of the Hosking and Wallis heterogeneity measure HW_1, largely used in hydrology, deserve further comments. The HW_1 test is based solely on the L-CV coefficient (see Equations (5) and (6)), and the fact that it performs well suggests that the heterogeneity among the series is mainly due to variations in the sample variance of the samples. In contrast, the variations in skewness and kurtosis are in many cases masked by the sample variability of higher order moments and L-moments. As a consequence, other tests of constancy of the variance in different samples can be used as alternatives to the HW_1 test. Possible examples are the "classical" Levene and Barlett tests [*Conover et al.*, 1981], that, however, resulted to be weaker than the HW_1 test in a preliminary case study.

Acknowledgments. This work was carried out under a C.I.P.E. (Comitato Interministeriale per la Programmazione Economica) grant, and a COFIN (2005) project of M.I.U.R. (Ministero dell'Istruzione, dell'Università e della Ricerca), Italy.

References

Chowdhury, J.U., Stedinger, J.R. and Lu, L.H., Goodness-of-fit tests for regional generalized extreme value flood distributions. *Water Resources Research*, **27**, 1765-76, 1991.

Conover, W.J., Johnson, M.E. and Johnson, M.M., A comparative study for homogeneity of variances, with applications to the outer cantinental shelf bidding data. *Technometrics*, **23**(4), 351-361, 1981.

D'Agostino, R.B. and Stephens, M.A., *Goodness-of-fit techniques*, Department of Statistics, Southern Methodist University, Dallas, Texas, 1986.

Dalrymple, T., Flood frequency analyzes. *Water Supply Paper 1543-A*, U.S. Geological Survey, Reston, Va, 1960.

De Michele, C. and Rosso R., A multi-level approach to flood frequency regionalization. *Hydrology and Earth System Sciences*, **6**(2), 185-194, 2002.

Durbin, J. and Knott M., Components of Cramér-von Mises Statistics. *London School of Economics and Political Science*, 290-307, 1971.

Farquharson, F.A.K., Green, C.S., Meigh, J.R. and Sutcliffe, J.V., Comparison of flood frequency curves for many different regions of the world. V.P. Singh (ed.), *Regional Flood Frequency Analysis*, 223-256, 1987. In: Proceedings of the International Symposium on Flood Frequency and Risk Analyses, 14-17 May 1986, Louisiana State University, Baton Rouge, U.S.A.

Fill, H.D. and Stedinger, J.R., Homogeneity tests based upon Gumbel distribution and a critical appraisal of Darlymple's test. *Journal of Hydrology*, **166**, 81-105, 1995.

Hampel, F.R., The influence curve and its role in robust estimation, *J. Am. Stat. Ass.*, **69**(346), 383-393, 1974.

Hosking, J.R.M. and Wallis, J.R., Some statistics useful in regional frequency analysis, *Water Resour. Res.*, **29**(2), 271-281, 1993.

Hosking, J.R.M. and Wallis, J.R., *Regional Frequency Analysis: an approach based on L-moments*, Cambridge University Press, Cambridge, UK, 1997.

Laio, F., Cramer-von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters, *Water Resour. Res.*, **40**, W09308, doi:10.1029/2004WR003204.

Lu, L. and Stedinger, J. R., Sampling variance of normalized GEV/PWM quantile estimators and a Regional Homogeneity Test, *Journal of Hydrology*, **138**(1/2), 223-245, 1992.

Robson, A. and Reed, D., *Flood Estimation Handbook Volume 3: Statistical procedures for flood frequency estimation*, Istitute of Hydrology Crowmarsh Gifford, Wallingford, Oxfordshire, 1999.

Scholz, F.W. and Stephens M.A., K-Sample Anderson-Darling Tests, *Journal of American Statistical Association*, **82**(399), 918-924, 1987.

Sveinsson, G.B.O., Boes, D.C. and Salas, J.D., Population index flood method for regional frequency analysis. *Water Resour. Res.*, **37**(11), 2733-2748, 2001.

Vogel, R.M. and Wilson, I., Probability distribution of annual maximum, mean, and minimum streamflows in the United States, *Journal of Hydrologic Engineering*, 69-76, 1996.

Wiltshire, S.E., Regional flood frequency analysis I: Homogeneity statistics. *Hydrological Sciences Journal*, **31**, 321-333, 1986a.

Wiltshire, S.E., Regional flood frequency analysis II: Multivariate classification of drainage basins in Britain. *Hydrological Sciences Journal*, **31**, 335-346, 1986b.

Wiltshire, S.E., Identification of homogeneous regions for flood frequency analysis. *Journal of Hydrology*, **84**, 287-302, 1986c.

6. figures*

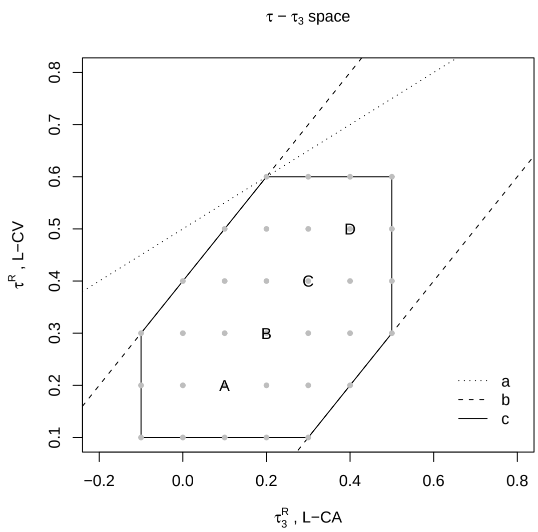

Figure 1. $\tau - \tau_3$ diagram (see Section 3). Lines are : (a) numerical constraint given by *Hosking and Wallis* [1997]; (b) bisector band identified using *Vogel and Wilson* [1996] samples; (c) the region we consider. Gray points are the τ^R and τ_3^R values considered in the main case study (Section 4.2); points A, B, C and D are considered in the sensitivity analysis of Section 4.3.

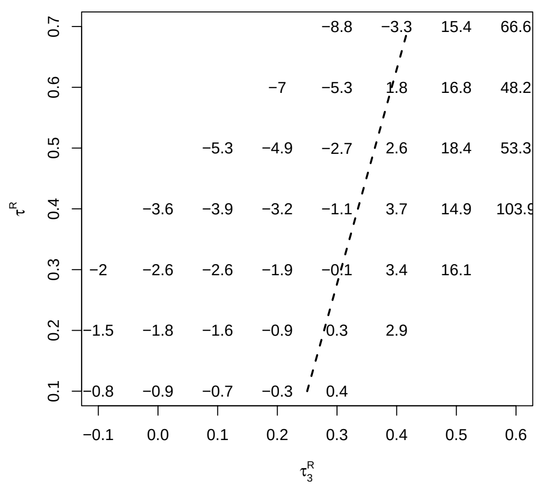

Figure 2. Difference between the $RMSE_{\%}$ of the sample mean and the $RMSE_{\%}$ of the sample median in the $\tau - \tau_3$ space (see Section 4.1). The dashed line indicates where the sample mean and sample median have, approximately, the same $RMSE_{\%}$; to the right of this line the sample median is a less distorted estimator of its population counterpart, to the left of it the sample mean performs (slightly) better.

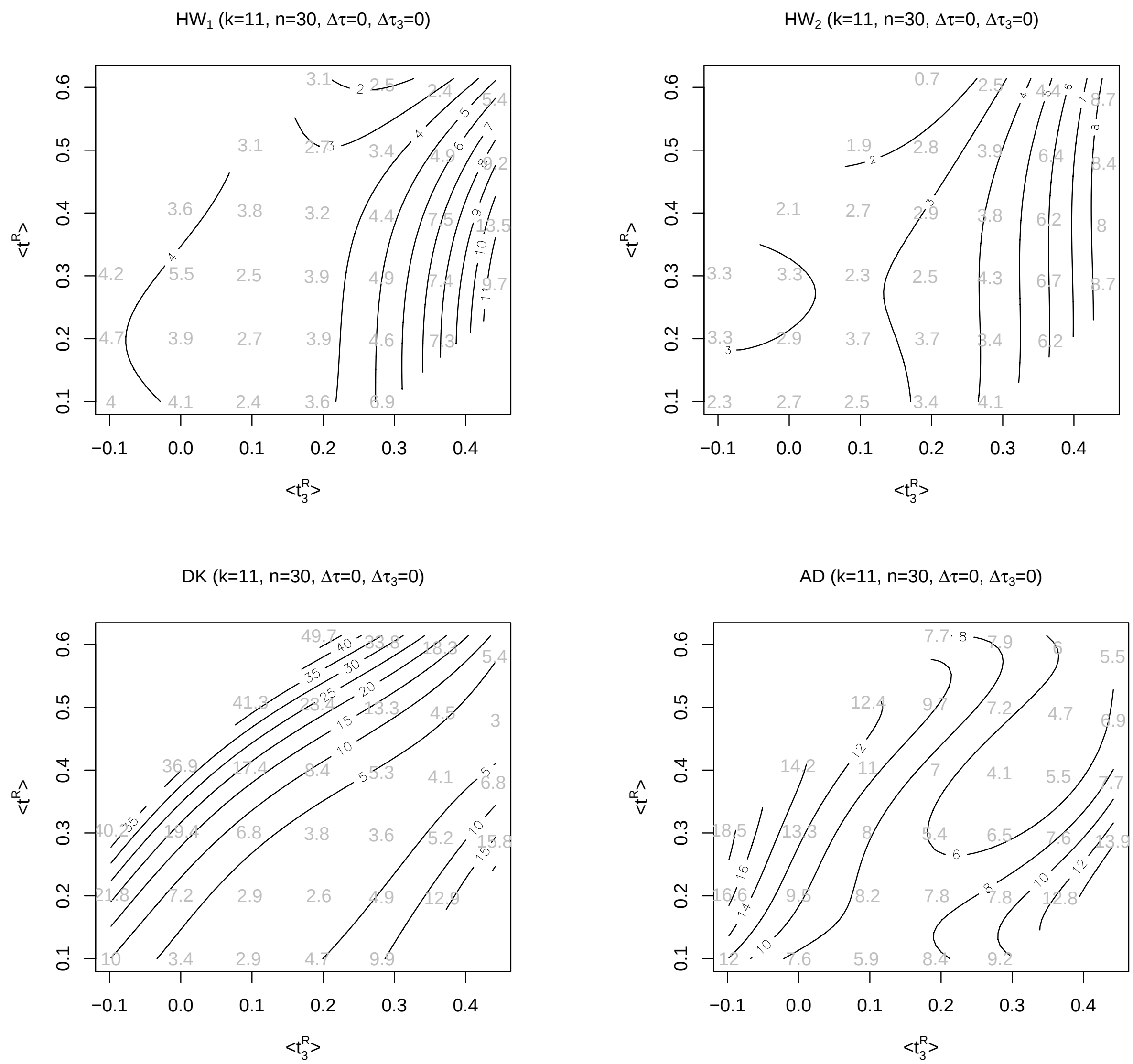

Figure 3. Percentage of regions erroneously stated as non-homogeneous in the $\tau - \tau_3$ space by the tests (Type I error). The homogeneous regions are generated using the Generalized Extreme Value distribution as the parent distribution; the other parameter values are reported in the title of each subplot.

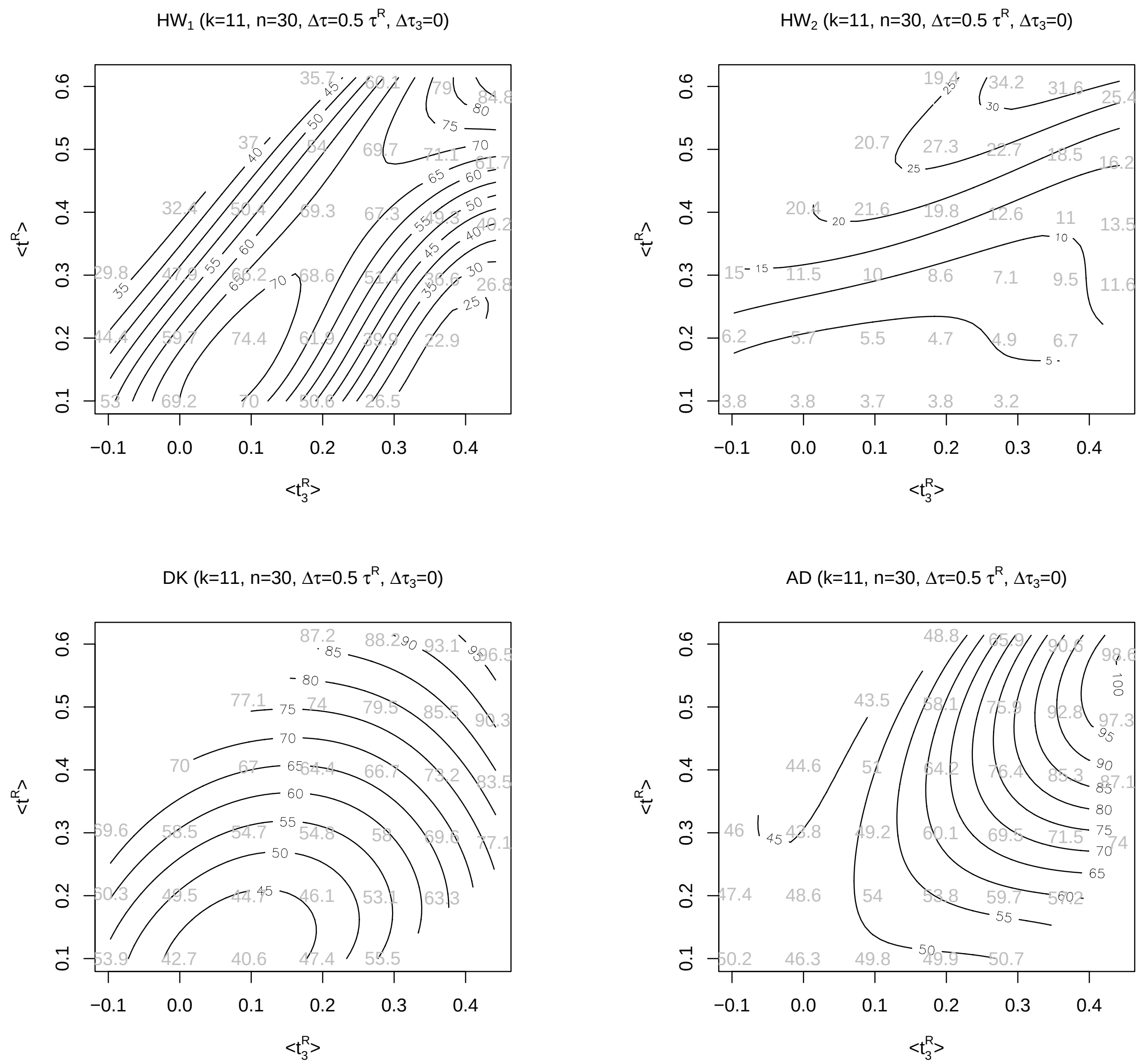

Figure 4. Power of the tests in the $\tau - \tau_3$ space with heterogeneous regions generated using the Generalized Extreme Value distribution as the parent distribution. Heterogeneity is due to the varying dispersion of the frequency distributions at different sites: the range of variation of the L-CV ($\Delta\tau$) in the region is 0.5 times the regional average L-CV (τ^R); the other parameter values are reported in the title of each subplot.

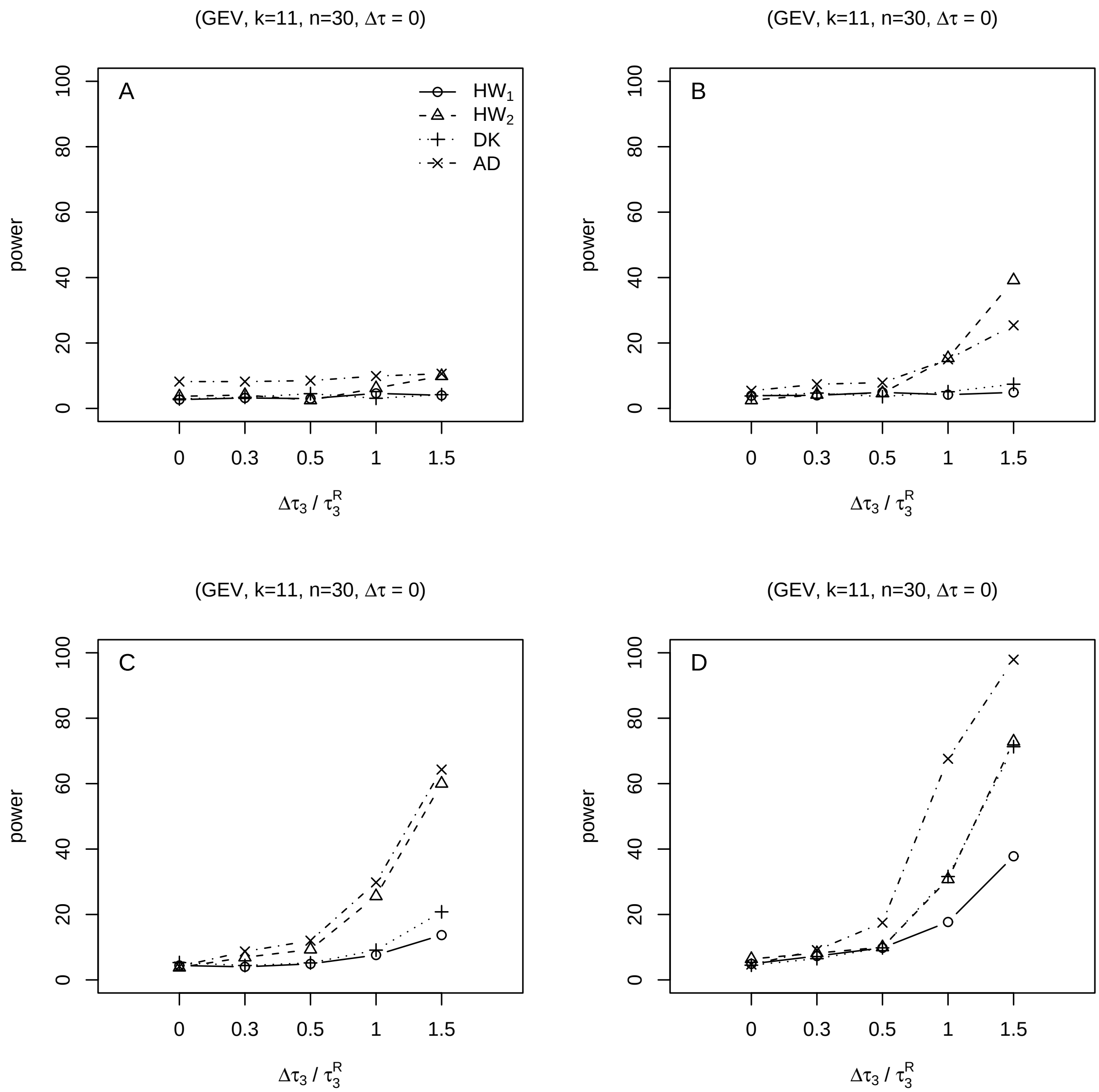

Figure 5. Power of the tests in points A, B, C and D (Figure 1) when the heterogeneity is due to the shape parameter τ_3 (see Section 4.3); parameter values are reported in the title of each subplot.

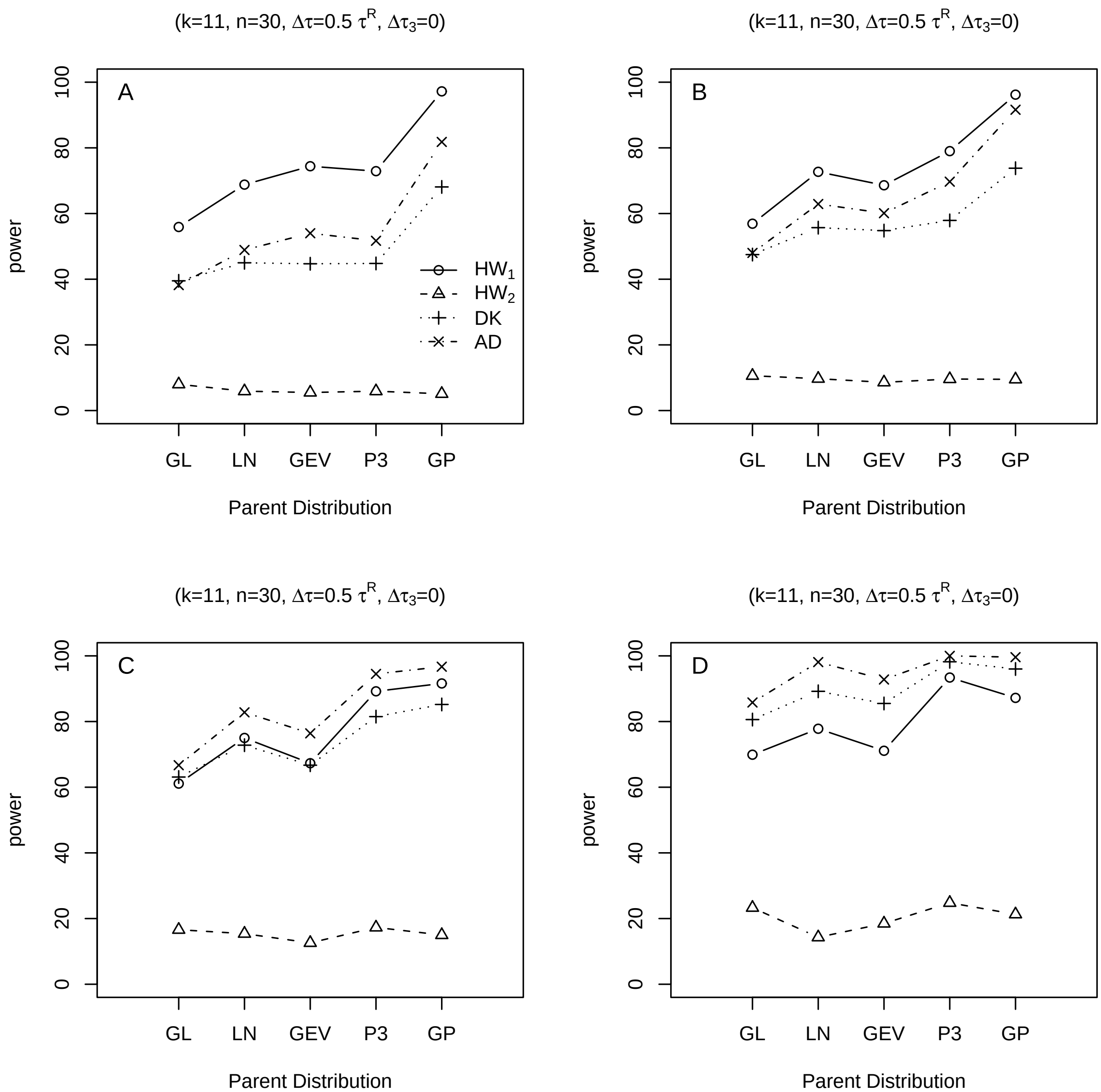

Figure 6. Power of the tests in points A, B, C and D (Figure 1) when changing the parent distribution (see Section 4.3); parameter values are reported in the title of each subplot.

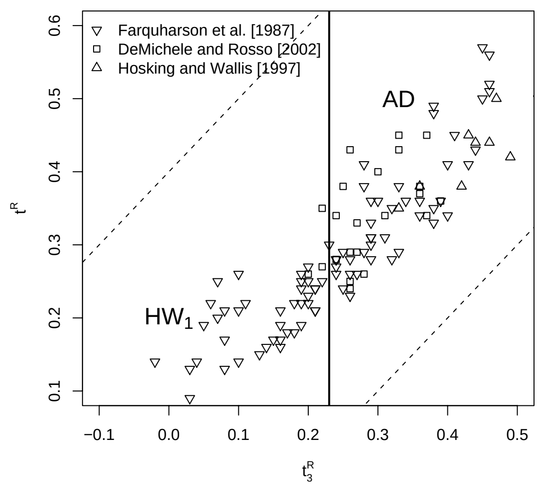

Figure 7. Regions of the $\tau - \tau_3$ space where the considered tests should be used (see Section 5); to the left of the black line ($t_3^R = 0.23$) the Hosking and Wallis heterogeneity measure HW_1 is the best test (considering both power and Type I error), to the right the bootstrap Anderson-Darling test AD should be used. Some real-world regional values are reported as points: *Farquharson et al.* [1987] computed these values considering many stations worldwide, *De Michele and Rosso* [2002] considering Italy and *Hosking and Wallis* [1997] the Apalachian region.