Critical values improvement for the Standard Normal Homogeneity Test by combining Monte Carlo and regression approaches

Author: Michele Rienzner[1], Francesca Ieva[2]
Affiliation: 1) Università degli Studi di Milano, DISAA
          2) Università degli Studi di Milano, Department of Mathematics
Address: 1) Via Celoria, 2, 20133, Milano, Italy
       2) Via Saldini 50, 20133 Milano, Italy
E-mail: michele.rienzner@unimi.it
Phone: +39 (0)2 503 16 901
Fax: +39 (0)2 503 16 911

Abstract

The distribution of the test statistics of homogeneity tests is often unknown, requiring the estimation of the critical values through Monte Carlo simulations. The computation of the critical values at low α, especially when the distribution of the statistics changes with the series length (sample cardinality), requires a considerable number of simulations to achieve a reasonable precision of the estimates (i.e., 10^6 simulations or more for each series length). If, in addition, the test requires a noteworthy computational effort, the estimation of the critical values may need unacceptably long runtimes.

To overcome the problem, the paper proposes a regression-based refinement of an initial Monte Carlo estimate of the critical values, also allowing an approximation of the achieved improvement.

Moreover, the paper presents an application of the method to two tests: SNHT (standard normal homogeneity test, widely used in climatology), and SNH2T (a version of SNHT showing a squared numerical complexity). For both, the paper reports the critical values for α ranging between 0.1 and 0.0001 (useful for the p-value estimation), and the series length ranging from 10 (widely adopted size in climatological change-point detection literature) to 70,000 elements (nearly the length of a daily data time series 200 years long), estimated with coefficients of variation within 0.22%. For SNHT, a comparison of our results with approximated, theoretically derived, critical values is also performed; we suggest adopting those values for the series exceeding 70,000 elements.

Keywords

Critical value; Monte Carlo simulation; Changepoint Detection; Time Series; Regression

AMS classification

62J02 Statistics - general nonlinear regression; 65C04 Numerical analysis - probabilistic methods-Monte Carlo methods; 65G99 Numerical analysis - error analysis and interval analysis

Abbreviations

SNHT – Standard normal homogeneity test;
SNH2T – Standard normal homogeneity test for couples of change-points;
n - series length (i.e. sample size);
P - number of considered values of n;
M - number of Monte Carlo iterations;
$α$ - significance of the test;
$C_{n,α}^{est}$ - Monte Carlo estimation of the critical value;
$C_{(n)}$ - unknown actual critical value;
$R_{(n)}$ - estimation given by the regression curve;

$e_{n,\alpha}$ - error (unknown) affecting the Monte Carlo estimations;

$s_{n,\alpha}^2$ - estimated variance of the Monte Carlo error;

$r_{n,\alpha}$ - regression residual;

$\varepsilon_{n,\alpha}$ - error affecting the estimation given by the regression curve.

1. Introduction

In climatology it is common to deal with long time-series. It is also common to incur, during the time span covered by the series, in changes in the measurement technology or practice, instrument substitution, changes in the environment surrounding the station, installation type or site. Those changes usually affect the measurement artificially altering the signal, then adding some bias in the estimation of the main quantities of interest that consequently reflect on most of the analysis performed on the time-series itself (e.g. trend evaluation, monthly mean values and distributions, etc.). A number of researchers, and the World Meteorological Organization (WMO), suggest applying methods to identify and remove the artificial changes in the mean value of the time-series, obtaining homogenized series that are suitable for the analysis purposes (e.g. Peterson et al., 1998; Aguilar et al., 2003; Rust et al., 2008; Venema et al., 2012).

A variety of different homogenization techniques have been developed and usually involve different phases, often customized for the application of interest (e.g. Ducré-Robitaille et al., 2003; Beaulieu et al., 2008; Venema et al., 2012; Domonkos, 2013). Most of the customization concerns the method for obtaining of the work series starting from the available series; the adopted method for the definition of the work series is suited to reduce the non-interesting behaviours enhancing the change-points (see Menne and Williams, 2005).

The core of the homogenization methods is typically the application of a homogeneity test to the work series (derived in various ways from the original one). Such tests aim at distinguishing if the series is homogeneous or includes one or more change-points (or breaks). The homogeneity tests are generally suited to apply to series that are the sum of an uncorrelated, zero mean, normal noise and a series containing the shifts (that is supposed to keep a constant value between two subsequent shifts). However, since the features of the series are often far from complying the conditions the tests are based on, the test must be robust to departures from the assumptions.

The most adopted homogeneity tests are based on the hypothesis that 'at the most one change-point' exists (i.e., the series should be homogeneous or include a single step change), and look for a single change-point a time.

The Standard Normal Homogeneity Test (SNHT – Alexandersson 1986), is a single change-point detection test widely adopted in climate time-series analysis for the homogeneity assessment and the change-point detection (see Khaliq and Ouarda 2007; Venema et al. 2013). SNHT is known for its power, simplicity and robustness (Aguilar et al. 2003; Ducrè-Robitaille et al. 2003; Beaulieu et al. 2008; Domonkos 2011; Rienzner and Gandolfi 2011; Venema et al. 2012).

For identifying further change-points, repeated applications are performed on pieces of the series, which is cut according to the identified change-points. As a matter of fact, in most cases 'single change-point' tests applications work fine also in case multiple change-points are included in the series (which is the common case), provided that such steps form a so-called *hierarchic pattern* (Hawkings 2001). On the other hand, the single change-point tests lose in reliability (Hawkings 2001; Rienzner and Gandolfi 2011) in case the steps follow a non-hierarchic pattern, as the platform-like inhomogeneity, i.e. two subsequent changes having equal amplitude and opposite sign. Therefore, multiple change-point detection methods should be preferred since, in the real cases, number and pattern of the change-points are usually unknown.

Some multiple change-point detection methods were proposed in literature. Among them Caussinus and Mestre (2004) minimizes a penalized residual sum of squares, whereas Hannart and Naveau

(2009) adopt a Bayesian approach. On the other hand, multiple change-point detection methods are computationally cumbersome and their application to time series with many thousands of elements is often unfeasible.

To this aim, Rienzner and Gandolfi (2011) proposed the SNHCM (standard normal homogeneity composite method), a composite multiple change-point detection method including SNHT and SNH2T (this test is peculiar to identify the platform-like inhomogeneity). Therefore, SNHCM is an intermediate point between the complete multiple change-point detection methods and the single change-point tests. In case of multiple change-points, SNHCM performance are close to the Caussinus and Mestre (2004) ones, while the simplified approach allows its application to long series.

SNHCM was developed for the application in a change point detection tool, the MAC-D procedure (Multiple Abrupt Changes in Daily time series - Rienzner and Gandolfi 2013), in order to make it able to deal with strong seasonal fluctuations, autocorrelation, non-normality and other departures from the most common requirements of the homogeneity tests. To do this, MAC-D works in a recursive fashion with three filtering phases, eventually giving to SNHCM a work series complying with most of its theoretical requirements. However, unpredictable minor features of the work series cannot be efficiently removed (e.g. seasonal fluctuation of autocorrelation parameters, variance non-stationarity, and chaotic behaviour) producing a relevant increase in the false positive rate. MAC-D was then re-parameterized in order to get a reasonable false positive rate (0.01) by further reducing the test's significance. Through an extensive Monte Carlo analysis, Rienzner and Gandolfi (2013) demonstrated the good performance of the whole approach. On the other hand, since the significance of both SNHT and SNH2T tests is a calibration parameter inside the MAC-D procedure, their critical values had to be obtained up to very low significance (0.0001) for different series length (n), say from 10 up to 70,000 elements. The lower limit for n is widely adopted in climatological change-point detection (see for example Ducré-Robitaille et al, 2003; Della Marta and Wanner, 2006; Reeves et al., 2007) while the higher is nearly the length of a daily data series 200 years long (which is the span of the longer meteorological series apart from the Antarctic ones).

As it is common for many widely applied homogeneity tests, the critical values of the tests follow unknown distributions changing with the series length (i.e. sample size). For this reason, the Monte Carlo approach is adopted for their estimation (e.g. Hawkins, 1977; Alexandersson, 1986; Csorgo and Horvath, 1997; Reeves et al., 2007). In order to perform such estimates for any series length, a large amount of series complying H0 are generated (i.e., uncorrelated random normal numbers free from any change-point). Then the tests are performed and the corresponding test statistics are collected. The estimation of the critical values is then carried out through the frequency approach, i.e., the estimation of the critical value $C_{n,\alpha}$ is obtained as $100 \cdot (1-\alpha)$ percentile of the set of statistics obtained with the series of length n. Obviously, the precision in the estimation of $C_{n,\alpha}$ is strictly related to the number of series processed and to the significance α.

For SNHT, an estimation of the critical values (thorough an asymptotic and an approximated approach) can be also obtained by the results presented in Csorgo and Horvath (1997). However, a comparison between those results and an intensive Monte Carlo estimation is still missing.

In order to reach a precise estimation of the critical values using Monte Carlo methods, intensive simulations are needed. As an example, Khaliq and Ouarda (2007) used 30×10^6 iterations, for each value of n (n being the length of the time series), reaching good estimations for SNHT with alpha up to 0.01. However, SNH2T requires at each step the computation of an $n \times n$ matrix, bringing to a considerable runtime in case of large n, e.g. 1.2 seconds per iteration with n=70,000. If 1.2 seconds is usually considered a short lapse of time, using the same iterations of Khaliq and Ouarda (2007) brings to a runtime of 416 days that is clearly unaffordable for the lifespan of any middle term research project.

Moreover, further decreasing alpha brings the focus on the thin part of the distribution's tail, whose area is described by a little number of Monte Carlo estimations, i.e. with a higher uncertainty with respect to the case considered in the cited paper for SNHT, and a p-fold reduction of the standard error of a Monte Carlo simulation needs an increase of p^2 times in the number of iterations (Kiviet, 2011). In other words, it is difficult to assess a priori if a suitable precision can be reached for SNH2T with alpha = 0.0001 after 416 days of computations. As a matter of fact, to reach a good estimation of the critical values with low significance, long series, and a time consuming test, we need a computationally efficient approach that is alternative to a mere increase of the Monte Carlo iterations.

Therefore, after an initial Monte Carlo estimation with a limited number of iterations, we performed a refinement of the obtained critical values was performed through a regression-based approach, also providing a method for the estimation of the precision of the regression based critical values.

This paper presents, after a brief description of the tests (Section 2), the method adopted for the critical values estimation and refinement (Section 3), the application of the method to the tests and the related optimality conditions assessment (Section 4), also comparing the results with the findings of other authors concerning SNHT critical values (i.e. Csorgo and Horvath, 1997; Khaliq and Ouarda, 2007). Finally, in Section 5 results are discussed. The parameters of the optimal regression curves obtained for the tests adopted as case studies, useful for the application of the tests, are included in Appendix B.

All the analyses were carried out in MATLAB® (R2011b, version 7.13), along with the MATLAB Optimization Toolbox® (R2011b, version 6.1), the MATLAB Statistics Toolbox® (R2011b, version 7.6), and the routines implementing SNHT and SNH2T as described in Rienzner and Gandolfi (2011).

2. SNHT and SNH2T

Since a complete description of the tests and their application details goes beyond the aims of this paper, only a brief description is provided here. For further reading concerning SNHT, see Hawkins (1977), Alexandersson (1986), Alexandersson and Moberg (1997), Khaliq and Ouarda (2007); for those concerned with SNH2T, refer to Rienzner and Gandolfi, (2011).

The single change-point homogeneity test SNHT is suited to distinguish between a series of uncorrelated normal variables with a constant unitary variance and constant mean value $\bar{z}_0$, and two different mean values ($\bar{z}_1 \neq \bar{z}_2$), one for the first part of the series and one for the second. This implies the following hypothesis test: H0: no change-point is present vs H1: one change-point is present.

The test is performed by:

i) computing for each index (*a*) of the series the statistics described in equation (1)(Alexandersson, 1986),

ii) picking the index *a** where the statistics *T* reaches its maximum (denoted by *T**),

iii) comparing *T** with the critical value that depends upon the series length (*n*) and the significance (*α*).

$$ T = \max_{1 \leq a < n-1} \left[a\bar{z}_1^2 + (n-a)\bar{z}_2^2 \right] \qquad (1) $$

where z is the standardized work series of length *n*, $\bar{z}_1$ is the mean of the first *a* elements and $\bar{z}_2$ is the mean of the remaining *n-a* ones. If *T** exceeds the critical value then the series is not homogeneous to the test and a step change occurs, with probability 1-*α*, between the indices *a** and *a**+1. In the case the series is not homogeneous, it is usually split according to the change-point position (i.e. between *a** and *a**+1) and the test is separately applied to the resulting two series.

This "check and split" procedure is repeated until no further change-point is found. A table of high quality critical values for SNHT are reported in Khaliq and Ouarda (2007).

SNH2T statistics is derived from equation (1) for the case of two change-points of equal amplitude and different sign (i.e. a platform-like inhomogeneity). The double change-point homogeneity test SNH2T is suited to distinguish, between a series of uncorrelated normal variables with a constant unitary variance with mean $\bar{z}_0$, and the same series with two different mean values ($\bar{z}_1 \neq \bar{z}_2$) due to two change-points: the first change-point introduces the new mean value (a change from $\bar{z}_1$ to $\bar{z}_2$) and the second restores the mean to the previous value (from $\bar{z}_2$ to $\bar{z}_1$). This implies the following hypothesis test: H0: no change-points are present vs H1: two change-points are present. SNH2T has 2 running indices, a and b, with $1 \leq a \leq b$ and $a \leq b \leq n$. Without loss of generality, we will exclude the meaningless couple ($a=1$, $b=n$). The statistics of SNH2T is given by:

$$T = \max_{1 \leq a \leq b \leq n} (t_{a,b}) \quad \text{with} \quad t_{a,b} = \left| \sum_{t=a}^{b} z_t \right| \sqrt{\frac{n}{(b-a+1)(n+a-b-1)}} \tag{2}$$

Denote (a^*, b^*) the couple of indices where the maximum is obtained. If T exceeds the critical value (depending on n and α), a platform-like inhomogeneity is included in the series from a^* to b^*, at a given significance α. As for SNHT, a tree-splitting recursive application is adopted for SNH2T, but with a three-segments split (i.e. one segment with indices from 1 to a-1, one from a to b, and the third from b+1 to n; special cases with two segments occur with $a=1$ or $b=n$). A list of the SNH2T critical values, computed with the method described hereafter, is reported in Rienzner and Gandolfi (2011).

3. Estimation of the critical values

Hereafter, the method for the estimation of the critical values is described, pointing out the differences in the application to SNHT and SNH2T according to their features. This Section is composed by three parts: the first one concerns the Monte Carlo estimation of the critical values and their uncertainty; the second explains the refinement of the estimations via regression approach; the third provides tools for estimating the distance between the calibrated curve and the unknown true values.

3.1 Monte Carlo estimations

As usual in the case of unknown distribution of the test statistic, the critical values for SNHT and SNH2T are firstly estimated through Monte Carlo approach (Kiviet, 2011). A number (P) of different series lengths, equally spaced on the logarithmic scale, were considered, with length n going from 10 to 70,000. Then, for each n, a number (M) of Monte Carlo iterations was performed in order to carry out the corresponding critical value of the T statistic.

Since the SNH2T runtime is proportional to n^2 (Rienzner and Gandolfi, 2011), whereas SNHT complexity is proportional to n (Alexandersson, 1986), different values of M and P were chosen for the tests as a compromise between the need of a large dataset and the need of maintaining feasible runtimes (e.g. a couple of weeks).

In particular: $P_{SNHT} = 109$; $M_{SNHT} = 30 \times 10^6$; $P_{SNH2T} = 50$; $M_{SNH2T} = 200,000$. The choice of 109 values of n for SNHT was made to match the 108 values used in Khaliq and Ouarda (2007), allowing a comparison with their results; while the 109th value (70,000) extends further the maximum inspected series length.

As usual, the Monte Carlo approach was performed under H0 (i.e. no change-points included in the series) by using uncorrelated standard normal series (using MATLAB[®] R2011b normal random

number generator). The estimation of the base critical values ($C_{n,\alpha}^{est}$) and their uncertainties ($s_{n,\alpha}^2$) were performed for the two tests as follows:

- For SNHT, $C_{n,\alpha}^{est}$ was estimated as the mean of the (1-α)·100 percentiles of M_{SNHT} = 30 independent Monte Carlo applications with 10^6 iterations each (like Khaliq and Ouarda, 2007). The 6 values of α adopted in Khaliq and Ouarda (2007) are included in the 19 values (ranging from 0.1 to 0.0001) considered here. Since the Monte Carlo percentiles have an asymptotical normal distribution (Kiviet, 2011), the variance of each estimation $C_{n,\alpha}^{est}$ (i.e., $s_{n,\alpha}^2$) was computed accordingly. This is also supported by a normality check performed on the 109x19=2071 collected samples. The Lilliefors test (Lilliefors 1967), with α = 0.05, rejected 5% of the samples as expected under H0 (i.e., the samples are normally distributed).
- For SNH2T, the estimation of the critical values was obtained by extracting the corresponding percentiles. A set of 13 values of α were considered, ranging from 0.1 to 0.0001, matching a corresponding number of SNHT α values. The variance of the Monte Carlo thresholds ($s_{n,\alpha}^2$) was computed through the bootstrap approach (Efron and Tibshirani 1993) with B=500 resamplings.

3.2 Refinement of the critical values

Without loss of generality, hereafter α is considered as fixed, if not otherwise specified, and the values are considered along the *n* index only. In fact, the same procedure was separately applied for each selected value of α.

Moreover, in the following, in order to distinguish between realizations of random variables and values obtained by known deterministic functions, or deterministic values, we included the subscript of the latter within brackets.

Denote $C_{(n,\alpha)}$ the unknown true critical value for a given *n*, and $e_{n,\alpha}$ the error affecting the Monte Carlo estimation procedure implemented in order to obtain the corresponding estimates $C_{n,\alpha}^{est}$, i.e.,

$$C_{n,\alpha}^{est} = C_{(n,\alpha)} + e_{n,\alpha} \qquad (3)$$

According to Kiviet (2011), we assume that

$$e_{n,\alpha} \sim N\left(0, \sigma_{(n,\alpha)}^2\right) \qquad n = 1, \dots, P$$

Where $\sigma^2_{(n,\alpha)}$ is the Monte Carlo error, estimated through $s_{n,\alpha}^2$.

Furthermore, denote $R_{(n,\alpha)}$ a regression function calibrated on the $C_{n,\alpha}^{est}$ (according to Baltagi, 2011), *n*= 1,…,*P*. Moreover, denote $r_{n,\alpha}$ the corresponding regression errors, i.e.,

$$C_{n,\alpha}^{est} = R_{(n,\alpha)} + r_{n,\alpha}. \qquad (4)$$

The following assumptions for the residuals ($r_{n,\alpha}$) are supposed to hold:

(i)   sequentially independent (i.e. $r_{n,\alpha}$ values should be negligibly correlated at increasing *n*),
(ii)   with zero mean,
(iii)   normally distributed,
(iv)   homoscedastic.

The last assumption may be relaxed if algorithms with proper tools for dealing with heteroscedasticity are employed. Since $s_{n,\alpha}^2$ increases with increasing series lengths (see fig. 1), i.e.,

$e_{n,\alpha}$ is heteroscedastic, standardization of residuals and proper calibration algorithms were adopted (see Baltagi 2011 among others, details are available in Appendix A).

[Figure 1(a) and 1(b) near here]

In particular, the variance of $r_{n,\alpha}$ was reduced by calibrating the regression functions with MATLAB® algorithms (Appendix A); root mean square error (*RMSE*) and R^2 statistics of the calibrated curves are reported in the tables resuming the application results (Sections 4.1 and 4.2).

To check the sequential dependence, it is common to carry out a visual inspection of the residuals (plotted versus the independent variable) and/or to apply a non-parametric test. We applied the runs test (non-parametric - Bradley, 1968) that gave *p*-values higher than 0.05 in all the cases. Moreover, since we have theoretical reasons supporting a quasi-normality of the residuals (as discussed at the end of this Section), and they are nearly evenly spaced along the independent variable, a parametric test for autocorrelation can be applied. In the case of interest, we firstly assessed the independence of the standardized residuals ($r_{n,\alpha}/s_{n,\alpha}$ – see below), then we run the Ljung-Box autocorrelation test (Ljung and Box, 1978), with respect to 3 lags, were applied with a 5% significance level. The normality assumption of the standardized residual distribution was checked using the Lilliefors test (Lilliefors 1967).
The application of this parametric test, that is somehow redundant, is aimed to further stress the absence of a significant sequential correlation of the residuals. On the other hand, we acknowledge that, even if a parametric test can be used, small (undetected) incompliances with the null hypothesis of the test may disturb the results, mainly, by increasing the type 1 error (erroneously reject H0). On the other hand, despite the non-parametric tests are usually less powerful, they are always reliable and may be better in performances under specific circumstances (see Gibbons and Chakraborti, 2011 for details).

If independence of the standardized residuals does not hold, this implies that there are ranges of the independent variable where the regression curve stays systematically above or below the data points instead of laying in the core of the data. In these cases, even if the curve partially improves the base estimations in some ranges of the independent variable, the opposite may occur in the areas where the curve does not follow the data; therefore, the curve is not optimal and further effort must be spent in the choice of the regression function.

Finally, the zero mean condition was matched by simply including an additive parameter into the regression function. In the application of interest, the resulting bias was always negligible.

From (3) and (4) we may define $\varepsilon_{n,\alpha}$ to be the distance between the estimation given by the regression curve and the unknown true critical value ($\varepsilon_{n,\alpha} = C_{(n,\alpha)} - R_{(n,\alpha)}$), i.e.,

$$r_{n,\alpha} = C_{n,\alpha}^{est} - R_{(n,\alpha)} = C_{(n,\alpha)} - R_{(n,\alpha)} + e_{n,\alpha} = \varepsilon_{n,\alpha} + e_{n,\alpha} \qquad (5)$$

In so doing, a useful partition of the regression error (as the sum of the Monte Carlo error and the difference between regression and the true critical values) is pointed out. In fact, equation (4) may be rewritten as

$$C_{n,\alpha}^{est} = R_{(n,\alpha)} + \varepsilon_{n,\alpha} + e_{n,\alpha} \qquad (6)$$

This is useful because highlights the role of an optimal calibration. In fact, the interest lies in allowing the term $e_{n,\alpha}$ to be the greatest part of the global error term $r_{n,\alpha}$. To this aim, the more optimal is the regression function, the smaller $\varepsilon_{n,\alpha}$ will be obtained. Indeed, if the $r_{n,\alpha}$s match the

optimality conditions and the function $R_{(n,\alpha)}$ is calibrated accounting for the different variances of the estimations $C_{n,\alpha}^{est}$, the distribution of each $r_{n,\alpha}$s approaches the corresponding distributions of the $e_{n,\alpha}$s. In other words, due to the procedure adopted for the estimation of the critical values $C_{n,\alpha}^{est}$, the $e_{n,\alpha}$s are sequentially uncorrelated, and normally distributed with zero mean and variances $s_{n,\alpha}^2$. Finally, since the curve was fitted on the actual values of $C_{n,\alpha}^{est}$ affected by the actual value of the errors $e_{n,\alpha}$s, the calibration errors $r_{n,\alpha}$s are expected to approach asymptotically the corresponding values of the $e_{n,\alpha}$s (by increasing the number of data points - P), reducing the distances $\varepsilon_{n,\alpha}$s between the regression curve and the unknown dotted curve of the true critical values $C_{(n,\alpha)}$.

3.3 Estimation of the error reduction

Assuming the optimality of the calibration procedure described above, and given the estimate of the MC variance, now we would like to quantify the improvement in estimation we can achieve using the critical points obtained by the regression approach with respect to the MC simulations.

The estimation proposed below is valid up to some simplifications (discussed and detailed in Appendix C); nevertheless, it provides a reasonable insight in terms of Percentage Variance Reduction of the improvements gained by the proposed method in estimating the critical values of the T statistics. The achievable Percentage of error Variance Reduction (hereafter PVR) can be written as (proof in Appendix C)

$$PVR = 100\left\{1+\left(1-\frac{1}{P}\right)\left[1-\mathrm{var}\left(\frac{r_{n,\alpha}}{s_{n,\alpha}}\right)\right]\right\} \qquad (7)$$

where, $r_{n,\alpha}$ is the regression error in equation (4), $s_{n,\alpha}$ is the standard deviation of the MC error included in the corresponding calibration data-point, and P is the number of simulated data-points at each given α.

Conditions for the application of equation (7) are:
  (i)   a large number (P) of calibration points well distributed along the abscissa according to the function curvature (e.g., the greater the curvature of the function, the higher the density of data needed in order to guarantee a reliable estimation);
  (ii)   a precise estimate of uncertainty of the data-points;
  (iii)   the regression function must fulfil the optimality conditions.

Once assessed the fulfilment of the conditions above, the estimation of the average coefficient of variation of the regression estimates, CV_{reg}, can be obtained as:

$$CV_{reg} = \underset{n}{mean}\left(\hat{\varepsilon}_{n,\alpha} / R_{(n,\alpha)}\right); \qquad (8)$$

where

$$\hat{\varepsilon}_{n,\alpha} = s_{n,\alpha} \cdot \sqrt{1-0.01\cdot PVR} . \qquad (9)$$

Since (8) focuses on the changes in the CV index, it might be sensitive to the values of different CVs obtained for the different values of n. In general, the more CV values are uniform along n, the more CVreg is expected to be a reliable measure of the uncertainty affecting the critical values estimated through the regression curve.

Figure 2 shows the CV of the initial Monte Carlo estimations at changing n for some values of α (0.1, 0.01, 0.001, 0.0001) for SNHT (panel a) and SNH2T (panel b). Those values of α where selected for avoiding a too crowded graph. As it is evident from the graph, after $n = 20$, the behaviour of the CVs is pretty flat, allowing PVR to become a good indicator of the enhancement of the critical values for (nearly) all the values of n.

[Figure 2(a) and 2(b) near here]

4. Application

4.1 SNHT

The 19 values of α considered for the application are listed in Table B2, Appendix B. As we said before, they include the six values considered by Khaliq and Ouarda (2007). For n, 109 values were used according to the following criterion: for n ranging from 10 to 100, one value every 2 elements was chosen; for n ranging from 100 to 200, one value every 5; for n ranging from 200 to 600, one value over 25; for n ranging from 600 to 1000, one value every 50; for n ranging from 1000 to 1600, one value every 100; for n ranging from 2000 to 5000, one every 500; and finally for the concluding values of $n = 7500, 10000, 15000, 20000, 50000$ and 70000.

All the reported values of n match with those adopted in Khaliq and Ouarda (2007), apart from the last one which is not included in their study.

As long as n increases, $s_{n,\alpha}$ ranges from 0.00066 to 0.0017 with α = 0.1, and from 0.003 to 0.046 with α = 0.0001. The coefficient of variation (CV_α) of the Monte Carlo estimations range from a minimum of 1% to the unacceptable precision of 14%.

We did not find a function family able to follow the Monte Carlo estimations for each α, therefore, a two-phase approach was used.

Firstly, an optimal regression function (i.e. obeying at the optimality conditions defined in Section 3.2 − optimality checks are reported in Table 1) was found, by trial and error (among different functional forms), for the critical values of SNHT with α = 0.1:

$$R_{n,0.1} = a \cdot \{\arctan[\ln(n) + b]\}^c + d \cdot \ln(n)^e + f \tag{10}$$

where: $R_{n,0.1}$ is the regression-based estimation of the critical values for α = 0.1, ln is the natural logarithm, $arctan$ is the inverse tangent function, and a, b, c, d, e and f are the function parameters (calibrated with $lsqcurvefit$) reported in Table B1 (Appendix B).

Secondly, a polynomial function of $R_{n,0.1}$ was found to give optimal regression functions for the other values of α (i.e. $R_{(n,\alpha)} = f(R_{n,0.1}(n), \boldsymbol{q}_\alpha)$ where $\boldsymbol{q}_\alpha$ is a set of regression parameters changing with α):

$$R_{(n,\alpha)} = q_{1,\alpha} R_{n,0.1}^4 + q_{2,\alpha} R_{n,0.1}^3 + q_{3,\alpha} R_{n,0.1}^2 + q_{4,\alpha} R_{n,0.1} + q_{5,\alpha} \tag{11}$$

where $R_{(n,\alpha)}$ is the regression estimation of the critical value for SNHT, according to the notation introduced in Section 3, and $q_{1,\alpha}$, $q_{2,\alpha}$, $q_{3,\alpha}$, $q_{4,\alpha}$, $q_{5,\alpha}$ are parameters (calibrated with the $lscov$) reported in Table B2 (Appendix B).

Fig. 3 shows a clear agreement between $C_{n,\alpha}^{est}$ points and the regression curves, as confirmed by the R^2 statistics reported in Table 1 together with the statistics concerning the goodness of fit and the analysis of the residuals. As expected in a case of optimal regression, the statistical independence and the normality assumption were not rejected for the standardized residuals ($r_{n,\alpha}/s_{n,\alpha}$). Moreover,

the bias is negligible, and the RMSE has a value between the minimum and the maximum $s_{n,\alpha}$ for every α. Since all the suitability conditions are fulfilled, the proposed regression curves are optimal in resuming all the information provided by the Monte Carlo estimations.

The estimated PVRs (equation (7)) have a mean value close to 50%, i.e., a reduction around 50% of the initial error variance can be expected for most of the curves leading the CV_{reg} values to range from 0.01% to 0.1%, which are fully reliable even for extremely low significances.

The calibrated regression curves and Monte Carlo results of Khaliq and Ouarda (2007) show a high agreement, since their difference with the calibrated curve have a negligible bias and a standard deviation of the same order of the corresponding uncertainties that Khaliq and Ouarda (2007) give for their Monte Carlo estimations.

As an example, fig. 4 illustrates the plot of the normalized residuals of the regression curve with α = 0.01. In the figure, the squares represent the standardized residuals ($r_{n,\alpha}/s_{n,\alpha}$) and the circles are the normalized differences with respect to the values of Khaliq and Ouarda (2007) (our computed standard deviations were used to have a normalization of their estimations).

[Figure 3 near here]

[Figure 4 near here]

[Table 1 near here]

An interesting theoretical work concerning the change-point detection problems is presented in Csorgo and Horvath (1997). Their results also include two methods for the estimation of the critical values of the ratio-changepoint detection test, written as the square root of T (equation (1)). The first method is based on an asymptotic result and provides an asymptotic estimation of the critical points given by

$$u^* = \frac{1}{\sqrt{2\ln(\ln(n))}} \left[-\ln\left(\frac{1}{2}\ln\left(\frac{1}{1-\alpha}\right)\right) + 2\ln(\ln(n)) + \frac{d}{2}\ln\left(\ln(\ln(n))\right) - ln\left(\Gamma\left(\frac{d}{2}\right)\right) \right] \quad (12)$$

where ln is the natural logarithm, d is the number of changing statistics (here $d = 1$ since the test is looking only for a change in the mean), and Γ is the gamma function and α is the significance.

According to the authors the estimator u^* can suffer from an excessive overestimation of the actual value of the statistics, also if n is not that small. Therefore, they provided an approximated result leading to a more useful estimation of the critical value. In this case, the resulting equation do not allow an analytical forward solution. Indeed, the value of the critical value u, for the given value α, have to be computed backward using the following equation:

$$\alpha = \frac{u^d e^{-x^2/2}}{2^{d/2}\Gamma(d/2)}\left[\ln\left(\frac{(1-h)(1-l)}{hl}\right) - \frac{d}{u^2}\ln\left(\frac{(1-h)(1-l)}{hl}\right) + \frac{4}{u^2}\right], \text{ with } h = l = \frac{1}{n}\ln(n)^{3/2} \quad (13)$$

Figure 5 shows, for a selection α values (0.1, 0.01 and 0.001), the squares of the estimators u^* and u and the corresponding values of R. In the same graph are also displayed the squares of the values given in table 1.3.1 of Csorgo and Horvath (1997), i.e., estimations for some couples n and α arising from the theorems and from a Monte Carlo simulations using series of random normal variables having the same mean and equal and unknown variance.

Looking at fig. 5, the asymptotic estimations provided by theorem 1.3.1 (squared to be comparable with R) are much higher than the ones provided by equation (11) (and also by our Monte Carlo evidences) even for $n = 70000$, confirming the need of the approximated approach.

The results u obtained from equation (13), once squared, are higher than our results for low n and become lower, but really close to them, for n large.

[Figure 5 near here]

4.2 SNH2T

The standard deviation of the SNH2T estimations shows the same behaviour as described for SNHT and, as long as n increases, $s_{(n,\alpha)}$ goes from 0.001 to 0.002 with $\alpha = 0.1$ and from 0.004 to 0.063 with $\alpha = 0.0001$. Due to the limited amount of iterations (200,000), the coefficients of variation of the Monte Carlo estimations are quite large, varying from 4% to 62%, and most of them are too high for a reliable application of the corresponding critical values.

In case of SNH2T, a suitable, general regression function was identified:

$$R_{(n,\alpha)} = a \cdot \left\{ \arctan\left[\ln(n) + b\right] \right\}^2 + c \cdot \ln(n)^d + e \tag{14}$$

where $R_{(n,\alpha)}$ is the critical value estimated by the regression function, ln is the natural logarithm, $arctan$ is the inverse tangent function, and a, b, c, d, and e are the function parameters (calibrated with *lsqcurvefit*) reported in Table B3 (Appendix B).

Table 2 reports the performance of the procedure: RMSEs are close to the standard deviation of the Monte Carlo estimations; the values of R^2 are close to one; the *p*-values of the statistical independence and normality tests are always higher than 5%. Moreover, in this case the Ljung-Box test is applied also complying the evenly spaced data condition. According to the results, the selected curves are optimal regressions of the data points.

The *PVR* index, is nearly normally distributed (*p*-value of the Lilliefors test = 0.5) and shows a mean value close to 90%. Due to the statistical error affecting the PVR estimations, some values of PVR are higher than 100%. The regressions' CV values of Table 2 are computed with the same procedure adopted for SNHT, except for the four values higher than 99% that were lowered to that value. The final CVs range from 0.02% to 0.22%, i.e., notwithstanding the large error included in the original Monte Carlo estimations, the refined estimations are fully reliable even for extremely low significances.

Figure 6 shows the Monte Carlo estimations and the calibrated regression curves. Looking at the figure, it can be seen how the agreement between points and curves can be considered high and satisfying.

[Table 2 near here]

[Figure 6 near here]

5. Conclusions

The paper proposes a method for improving the estimation of the critical values of the statistical tests adopted for detecting change points in time series analysis. In particular, the methodology is focused on those cases where the distribution of the test statistics is both unknown and varying with the sample size or the series length. According to the proposed method, critical values are firstly obtained by Monte Carlo percentiles, and then refined by calibrating suitable regression models, obtaining an estimation of the average precision of the curves in terms of average coefficient of variation (CV).

Two applications, concerning the change-point detection tests, SNHT and SNH2T, are discussed. Since the tests have a different computational complexity (SNHT is linear, while SNH2T is quadratic in the series length), in order to obtain a reasonable runtime (two weeks for SNHT and three for SNH2T on a standard personal computer), different values of Monte Carlo iterations and

the series lengths were considered. According to the results, the proposed regression approach reduced the standard deviation of the error of the initial Monte Carlo estimations by 1.5 times for SNHT, and 3.4 times for SNH2T, achieving average coefficients of variation within 0.2%. Therefore, the estimated critical values are fully reliable also at the very low significance of 0.0001, allowing their application in the MAC-D package.

Focusing on SNHT, for a series length (n) higher than 70,000, the estimations given by equation (11) lose quickly their precision, as it is expected from any extrapolation on polynomial functions. In that cases, we can suggest to adopt the squares of the estimations provided by equation (13) (u^2, theorem 1.3.3 of Csorgo and Horvath, 1997) that behave fairly well for very large n.

Since a reduction of p times in the Monte Carlo standard error implies an increase of p^2 times in the number of iterations, the same improvement obtained as described here could be achieved by increasing 2.25 times the iterations of SNHT (4.5 weeks runtime), and 12 times the iterations of SNH2T (i.e. 8.4 months). While the first is rather a long runtime, the second is clearly unacceptable and/or unaffordable, showing that increasing the Monte Carlo iterations is not an appropriate solution for all the cases.

The choice of a suitable regression function is the main difficulty of the method in the real-case application, since positive outcomes are expected only in case the optimality conditions are fulfilled. Alternative ways, like the two phases adopted for SNHT, or the reduction of the range of n in partially overlying parts, adopting different curves for each of them, can help in satisfying all the conditions.

In conclusion, the results confirm the proposed approach to be an interesting alternative to a mere increase of the Monte Carlo iterations, and leads to a sharp improvement of the critical values in terms of reduction of the error on the estimates. Moreover, through the PVR indicator, it is possible to measure the average uncertainty (in terms of coefficient of variation) affecting a regression curve.

Supplemental material (Appendices)

Appendices A, B, and C are provided as supplemental material: Appendix A gives detailed information about the calibration process; Appendix B provides the tables concerning all the sets of parameters we calibrated for equations (10), (11), and (14); Appendix C reports the complete proof of equation (7).

References

Aguilar, E., Auer, I., Brunet, M., Peterson, T.C., Wieringa, J.: Guidelines on climate metadata and homogenization. World Meteorological Organization, WMO/TD No. 1186 (2003).

Alexandersson, H.: A homogeneity test applied to precipitation data. Int. J. Climatol. 6, 661-675 (1986).

Alexandersson, H., Moberg, A.: Homogenization of Swedish temperature data. Part I: homogeneity test for linear trends. Int. J. Climatol. 17: 25–34 (1997).

Baltagi, B., H.: Econometrics (5th ed.) - Springer Texts in Business and Econometrics. Springer, New York, 2011. ISBN: 978-3642200588

Beaulieu, C., Seidou, O., Ouarda, T.B.M.J., Zhang, X., Boulet, G.: Abderrahmane Yagouti, Intercomparison of homogenization techniques for precipitation data. Water Resour. Res. 44, 2 (2008).

Bradley, J.V., 1968. Distribution-Free Statistical Tests. Prentice Hall, Englewood Cliffs, New York.

Caussinus H., Mestre, O.: Detection and correction of artificial shifts in climate series. Applied Statistics 53(3): 405–425 (2004).

Coleman, T.F., Li, Y.: On the convergence of reflective Newton methods for large-scale nonlinear minimization subject to bounds. Math. Program. 67(2), 189-224 (1994).

Coleman, T.F., Li, Y.: An interior, trust region approach for nonlinear minimization subject to bounds. SIAM J. Optimiz. 6, 418-445 (1996).

Csorgo, M., Horvath, L.: Limit theorems in Change-point detection. J.Wiley & Sons, Chichester, 1997. ISBN: 978-0-471-95522-1

Della Marta, P.M., Wanner, H. A method of homogenization the extremes and mean of daily temperature measurements. Journal of Climate 19 (17): 4179–4197 (2006).

Domonkos, P.: Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods. Theor. Appl. Climatol. 105, 455-467 (2011) doi 10.1007/s00704-011-0399-7

Domonkos, P.: Measuring performances of homogenization methods. Idöjárás, 117, 91-112 (2013).

Draper, N.R., Smith, H.: Applied regression analysis. Jhon Wiley & Sons, inc. New York. ISBN 0 471 22170 8 (1966).

Ducrè-Robitaille, J.F., Vincent, L.A., Boulet, G.: Comparison of techniques for detection of discontinuities in temperature series. Int. J. Climatol. 23, 1087-1101 (2003). DOI: 10.1002/joc.924.

Efron B, Tibshirani R. 1993. An Introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall/CRC, London. ISBN 0412042312

Gibbons, J.D., Chakraborti, S: Nonparametric Statistical Inference, Fifth Edition. Statistics: textbooks & monographs. Chapman and Hall / CRC press – Taylor & Francis Group, Boca Raton – London – NewYork. (2011) ISBN-13: 978-1-4200-7762-9 (Ebook-PDF)

Hannart A., Naveau P.: Bayesian multiple change points and segmentation: application to homogenization of climatic series. Water Resources Research 45: W10444 (2009), DOI: 10.1029/2008WR007689

Hawkins, D.M.: Testing a sequence of observations for a shift in location. J AM STAT ASSOC, 72: 180–186 (1977).

Hawkings, D.M.: Fitting multiple change-points to data. Comput. Stat. Data An. 37, 323-341 (2001).

Khaliq, M.N., Ouarda, T.B.M.J.: Short communication on the critical values of the standard normal homogeneity test (SNHT). Int. J. Climatol. 27, 681-687 (2007).

Kiviet, F.: Monte Carlo simulation for econometricians. Foundations and Trends in econometrics. 5(1-2), 1-181 (2011) doi: 10.1561/0800000011. (2012) ISBN-10: 160198538X, ISBN-13: 978-1601985385

Menne MJ, Williams CN Jr. Detection of undocumented changepoints using multiple test statistics and composite reference series. Journal of Climate 18: 4271–4286 (2005). DOI: 10.1175/JCLI3524.1

Motulsky, H.J., Christopoulos, A.: Fitting models to biological data using linear and nonlinear regression. A pratical guide to curve fitting. GraphPad Software Inc., San Diego CA (2003).

Ledermann, W.: Handbook of applicable mathematics. Vol. IV, Statistics. John Wiley & Sons, New York (1984).

Lilliefors, H.W.: On the Komogorov-Smirnov test for normality with mean and variance unknown. J. Am. Stat. Assoc. 62, 399-402 (1967).

Ljung, G.M., Box, G.E.P.: On a Measure of a Lack of Fit in Time Series Models. Biometrika 65(2), 297-303 (1978) doi:10.1093/biomet/65.2.297

Peterson, T.C., Easterling, D.R., Thomas, R.K., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Heikki, T., Mestre, O., Szentimrey, T., Salinger, J., Førland, E.J., Hanssen-Bauer, I., Alexandersson, H., Jones, P., Parker, D.: Homogeneity adjustments of in situ atmospheric climate data: a review. International Journal of Climatology 18: 1493–1517 (1998), DOI: 10.1002/(SICI)1097-0088(19981115)18: 13<1493::AID-JOC329>3.0.CO;2-T

Reeves, J., Chen, J., Wang, X.L., Lund, R., Lu, Q.: A review and comparison of changepoint detection techniques for climate data. Journal of applied meteorology and climatology, 46: 900-915 (2007).

Rienzner, M., Gandolfi, C.: A composite statistical method for the detection of multiple undocumented abrupt changes in the mean value within a time series. Int. J. Climatol. 31, 742-755 (2011) doi: 10.1002/joc.2113

Rienzner, M., Gandolfi, C.: A procedure for the detection of undocumented multiple abrupt changes in the mean value of daily temperature time series of a regional network. Int. J. Climatol. 33, 1107-1120 (2013) doi: 10.1002/joc.3496

Rust, H.W., Mestre, O., Venema, V.K.C.: Fewer jumps, less memory: homogenized temperature records and long memory. J GEOPHYS RES-ATMOS. 113, d19 (2008) doi: 10.1029/2008JD009919

Venema, V.K.C., Mestre, O., Aguilar, E., Auer, I., Guijarro, J.A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos M., Williams, C.N., Menne, M.J., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, C., Prohom-Duran, M., Likso, T., Esteban, P., Brandsma, T.: Benchmarking homogenization algorithms for monthly data. Clim. Past 8, 89-115 (2012).

Wald, A., Wolfowitz, J.: On a test whether two samples are from the same population. Ann. Math Statist. 11, 147-162 (1940).

Table 1 Goodness of fit statistics, *p*-values of the tests and *PVR* indexes for the regression functions and the maximum value of the resulting CVs for Monte Carlo (MC) and regression curves (r) for each value of alpha

Alpha	RMSE	R^2	Runs test	Ljung-Box test (3)	Lilliefors test	*PVR*	CV_{MC} %	CV_r %
0.1000	0.00129	0.999992	0.51	0.17	0.43	53.3	1.4	0.01
0.0800	0.00143	0.999995	0.56	0.07	0.50	60.0	1.5	0.01
0.0750	0.00148	1.000000	0.56	0.11	0.50	54.1	1.5	0.01
0.0600	0.00183	1.000000	1.00	0.09	0.50	20.7	1.5	0.01
0.0500	0.00199	0.999997	0.81	0.09	0.22	28.6	1.6	0.01
0.0250	0.00269	0.999979	0.44	0.22	0.28	56.4	2.0	0.01
0.0100	0.00419	0.999992	0.56	0.74	0.44	65.0	2.7	0.02
0.0080	0.00503	0.999965	0.73	0.35	0.50	43.5	2.8	0.02
0.0075	0.00530	0.999963	1.00	0.29	0.50	42.0	2.9	0.02
0.0060	0.00540	0.999967	1.00	0.19	0.32	59.5	3.1	0.02
0.0050	0.00588	0.999956	0.88	0.10	0.30	60.8	3.3	0.02
0.0025	0.00806	1.000000	0.34	0.25	0.40	60.8	4.2	0.03
0.0010	0.01304	0.999993	0.43	0.53	0.50	42.8	5.9	0.04
0.0008	0.01400	0.999961	0.34	0.18	0.45	50.0	6.3	0.04
0.00075	0.01351	0.999989	0.36	0.24	0.50	63.8	6.4	0.04
0.0006	0.01579	0.999947	0.12	0.69	0.06	60.8	7.1	0.04
0.0005	0.01704	0.999890	0.78	0.95	0.05	56.6	7.5	0.05
0.00025	0.02201	1.000000	0.15	0.08	0.13	81.3	9.8	0.04
0.0001	0.03849	1.000000	0.08	0.42	0.50	36.9	14.1	0.11

Table 2 Goodness of fit statistics, *p*-values of the tests and *PVR* indexes for the regression functions and the maximum value of the resulting CVs for each value of alpha

Alpha	RMSE	R^2	Runs test	Ljung-Box test (3)	Lilliefors test	*PVR*	CV_{MC} %	CV_r %
0.1000	0.00170	0.999996	0.36	0.19	0.09	84.6	4.05	0.02
0.0750	0.00205	0.999989	0.88	0.31	0.50	74.2	4.51	0.02
0.0500	0.00221	0.999997	0.80	0.09	0.50	90.8	5.12	0.02
0.0250	0.00317	0.999974	0.92	0.69	0.50	66.9	6.61	0.04
0.0100	0.00405	0.999929	0.85	0.77	0.50	90.5	9.20	0.03
0.0075	0.00434	0.999800	0.47	0.45	0.50	99.0	10.4	0.01
0.0050	0.00494	0.999912	0.44	0.64	0.50	105.4	12.4	0.01
0.0025	0.00646	0.999771	0.20	0.44	0.50	106.3	16.5	0.02
0.0010	0.00856	0.999766	0.89	0.31	0.50	104.4	23.6	0.02
0.00075	0.00952	0.999595	0.20	0.38	0.50	87.9	26.6	0.09
0.0005	0.01083	0.998969	0.47	0.74	0.10	103.5	31.9	0.03
0.00025	0.01443	0.998471	0.47	0.50	0.50	77.8	38.8	0.19
0.0001	0.01872	0.997534	0.88	0.75	0.50	87.8	62.0	0.22

Figure captions:

Fig. 1 Standard deviation estimated for the Monte Carlo critical values ($s_{n,\alpha}^2$) for a set of α values and SNHT (*a* panel) and SNH2T (*b* panel)

Fig. 2 Behavior of the coefficient of variation (CV) of the initial Monte Carlo estimations along *n* and for some values of α: SNHT (*a* panel), SNH2T (*b* panel)

Fig. 3 Critical values of SNHT: Monte Carlo estimations (dots) and calibrated regression curves (gray lines) for the considered values of α (numbers at the right edge of the curves)

Fig. 4 Regression residuals: normalized differences (residuals) between the estimations of the regression curve (α = 0.01) and the calibration points (squares); the same with the estimations of Khaliq and Ouarda (2007) (circles)

Fig. 5 Comparison between the results of *R* (equation (11), black lines), u^{*2} (equation 12, light grey lines), u^2 (equation (13), dark grey lines), and the squares of the values reported in table 1.3.1 of Csorgo and Horvath (1997): the light grey squares refer to estimates obtained with critical values like in equation (12), the dark grey triangles to those obtained like in equation (13), and the black circles to those estimated trough Monte Carlo simulations

Fig. 6 Critical values of SNH2T: Monte Carlo estimations (dots) and calibrated regression curves (gray lines) for the considered values of α (numbers at the right edge of the curves)