To my family

for their unconditional love

and everlasting support

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Assoc. Prof. Dr. Ceylan Yozgatligil, and co-advisor, Prof. Dr. İnci Batmaz, who accepted me as a student. I am deeply indebted to them, for patience, help, suggestions and encouragements in the whole process of this study. Their guidance, support, and feedbacks have turned this study to an immeasurable learning experience.

I would like to thank Prof. Dr. Gerhard Wilhelm Weber for valuable ideas and contributions that he made to this study. I really appreciate Assoc. Prof. Dr. Özlem İlk Dağ for building a profound content of this thesis. I also want to express thanks to Assoc. Prof. Dr. Seher Nur Sülkü and Assist. Prof. Dr. Özlem Türker Bayrak for their relevant discussions, suggestions and comments. I owe my special thanks to examining committee members for their acceptance of reviewing my thesis study by spending their invaluable time.

Fatma Yerlikaya-Özkurt and Elçin Kartal-Koç, express full support, everlasting patience, motivation and kindness all the time. They always tried to answer my questions and contributed to this study. I also want to thank to them for the amusing time we spent.

I owe my special thanks to my dear friend Nazire Canver for her perfect friendship, patience, kindness, confidence and the enjoyable time we spent from the first moment I have met her. I also want to express many thanks to Neşe Bayram for friendship, kindness and support during my study.

But most of all, I am forever indebted to my parents, and my brother who have always stood by me, for their endless love, understanding, endless patience and encouragement when it was most required.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES

LIST OF TABLES

TABLES

CHAPTER 1

INTRODUCTION AND MOTIVATION

Time series is a series that consists of observations that are measured in an order. Even though this ordering is generally made through time with equally spaced intervals, continuously recorded series are also possible. Time series has many application areas including economics (Baillie et al. 1996), finance (McNeil and Frey, 2000), oceanography (Lau and Weng, 1995), and health (Zeger et al., 2006).

The main objectives of the time series are to understand the nature of the data, to make forecasts about the future values, to interpret the results and to control and simulate the system. The observations of the series are correlated and different series may be dependent on each other. Thus, independence assumption of the traditional methods is not valid for time series and other methods which considers properties of time series such as autocorrelation are needed.

Any change in the parameters of time series may influence the distribution and may cause abrupt changes or trends, sudden increase or decreases, or changes in the mean and/or variance of the series. Investigating the effects and locations of these changepoints is a branch of time series. Before conducting any statistical analyses, the quality control of the data should be considered to detect and correct the effects of changepoints, if possible.

While working with data which includes independent observations, the traditional statistical tests such as t-test or chi-square test can be easily applied, if assumptions are validated. However, these tests are not applicable if the data covers time series due to autocorrelation.

The detection of changepoints is crucial in many areas such as climate (Toreti et al., 2012) to detect climate change, health studies to detect anomalies such as an increase in heart rate (Aminikhanghahi and Cook, 2017), to detect macroeconomic fluctuations (Sobreira et al, 2014) etc. It is also needed to detect the location of the changepoint and then remove or correct the effect of it, if possible. This correction process is called as homogenization in climate studies.

One of the examples of time series is climate studies which gained much concern since it has an effect on humans and environment directly. The food, water and shelter to live are the most essential needs for living things. These are all dependent on the climate of the region. World Meteorological Organization (WMO) defines climate as the mean weather status of an area over a long period of time (WMO, 2017). However, the world is experiencing the strongest climate change since the beginning of the 20th century in its history (WMO, 2017) which may influence the next generations. This change may result in increase in temperature, floods, rise in sea-levels and melting glaciers. In order to forecast extreme climate events and take precautions, conducting statistical analyses with meteorological series is the first step. These analyses are valid with a high quality non-human effect related data.

Climate studies are also a possible area of changepoint detection because the world is continuously changing with all its components one of which is climate. Humans and living things which are adapted to the changes such as shortages on the water sources early, have a better selective advantage over those who do not in a dynamic world. Thus, perceiving changes and moving with them becomes a significant issue (Boettcher, 2011).

There are many climate related studies in the literature. To exemplify, in our project (Determination of Climate Zones and Development of Rainfall Prediction Models for Turkey by Data Mining), where we have been working for more than five years, the homogeneity problem occurred due to the nature of the data. The main concerns of the project are to determine the change in climate and to develop precipitation models motivated by changes in the climate.

It is also essential to work with data that do not have any outer effect so that reliable results can be obtained; otherwise unreliable inferences may be made. If any outer effect is detected, it should be removed if possible before conducting any kind of analyses.

Meteorological variables such as precipitation, minimum temperature, maximum temperature or air pressure can be easily influenced by outer effects and they are collected in meteorological stations by instruments. Thus, the variables are subject to instrument, location or station which may have significant effect on them. For instance, urbanization around the stations, any breakout or change in the instruments, or changes in the calculations can result in abrupt changes, sudden increase or decreases, gradual changes, artificial trends or multiple changepoints (Yozgatligil and Yazici, 2016). That occurrence of non-climatic effect on the variables is called as *homogeneity* problem in climate studies. That problem is also named as structural break, segmentation, edge detection, event detection, regime switching, breakouts and anomaly detection (Aminikhanghahi and Cook, 2017). Historic metadata support is the best solution to this problem and essential for evaluating the breaks detected. Unfortunately, most of the data sets do not have accompanying metadata to check the sources of inhomogeneity.

In homogeneity studies, changepoint detecting algorithms are classified as *absolute* and *relative* depending on whether the need for a reference series or not (Tuomenvirta, 2002). If a highly correlated series is needed to conduct the test, the method is classified as relative, otherwise it is classified as an absolute test. The methods are also classified as *online* and *offline* algorithms (Aminikhanghahi and Cook, 2017). The offline algorithms take the whole series and look back in time to detect when the change occurred, while online algorithms tries to detect the changepoint as soon as it occurs rather than investigating the past observations.

Note that, since the observations are taken from instruments in the meteorological stations, it is obvious that these variables are spatio-temporal variables. However, the methods used in the literature to detect inhomogeneities have some drawbacks. For instance, the Standard Normal Homogeneity Test (SNHT) has independent and identically distributed (i.i.d) assumption which is not realistic for time series. The nonparametric tests such as Kruskal Wallis do not indicate the exact breakpoint but gives an interval and the number of observations in each block is not determined. The relative tests need homogeneous references, but if there is no prior information, then it is almost impossible to classify the stations. Moreover, this type of data may include extreme values due to its nature. Thus, these extreme values should be determined and treated separately.

There are many methods used for changepoint analysis. Yozgatligil and Yazici (2016) compared the performances of the offline methods under different scenarios of inhomogeneity based on simulated data. The results show that relative tests work better than absolute tests especially when the changepoint is in the middle of the series, while the detection performances become worse when the location of the changepoint is close to the beginning or end of the series. Moreover, detection performances become better when the magnitude of the level shift increases. Thus, there is a need for an absolute test to detect small amounts of level shifts especially when the shift is close to the boundaries of the series.

In this study, likelihood ratio test based on the exact likelihood for autoregressive models is used. The other alternative is to use conditional likelihood which is an approximation of the exact likelihood and both have the large sample properties (Hamilton, 1994). The critical values of the test are obtained by a simulation study. Then, moving block bootstrap procedure is proposed for detection of changepoints to capture the mean shift starting at the beginning or end of the observations. This study considers the mean shift (level shift) scenario. The method is also tried to detect multiple changepoints. Even though this study is motivated by a meteorological data, the proposed method can be applicable to any time series such as economics or health data which may include a changepoint.

The current study consists of five chapters and each one is organized in the following way. Literature review for the applications of changepoint detection is given in Chapter 2. In Chapter 3, likelihood ratio tests for autoregressive models, Standard Normal Homogeneity Tests (SNHT) and bootstrap methods that are used in this study are explained in detail. Application results of the methods on the simulations are presented in Chapter 4. Conclusions and future studies, which can be developed depending on the findings of this thesis, are given in Chapter 5.

CHAPTER 2

LITERATURE REVIEW AND BACKGROUND

Changepoint detection in time series has been studied in different aspects in the literature. Chen and Gupta (2012) stated that the first aim of the changepoint analysis is to detect whether there is any change in the series or not and then try to estimate the number of possible changepoints and associated locations.

There are few studies to make inferences about time series such as the equality of means or variances. For instance, Panaretos et al. (2010) considered two-sample problem in terms of functional setting and developed inferential methodology. Horvath et al. (2009) compared linear operators in two functional regression models. Horvath et al. (2013) proposed methods to test whether the mean functions of the functional samples are equal or not. They considered the samples which exhibit temporal dependence valid for stationary data by using kernels. Degras et al. (2012) proposed a simulation-based test for parallelism among trends in nonstationary time series.

Likelihood ratio test (LRT) is generally used to test the changepoint problems in time series. For instance, Tsay (1988) tried to handle outliers by using least squares and residual variance ratios. Chang et al. (1988) studied the estimation of parameters when there are outliers in ARMA model and then Chen and Liu (1993) improved their model. Then, Battaglia and Orfei (2005) suggested a similar method of Chen and Liu (1993) to nonlinear time series models. Apart from these studies, Davis et al. (1995)

proved that if there is no change in the parameters and order of an AR model, the LRT obtained by conditional likelihood is distributed as Gumbel's extreme value distribution. Karioti and Caroni (2004) compared the powers of LRT and Normal outlier test to detect whether the means of different small time series are equal or not. Moreover, McQuarrie and Tsai (2003) investigated the effect of outliers on the parameter estimates and model selection and proposed a LRT based method to classify the outlier as an innovation or an additive one. Then, Gombay (2008) studied the large sample properties of the test statistic derived to detect the change in any parameter of the AR models. All these works mentioned are based on conditional likelihood or least squares. Moreover, Davis et al. (2008) studied the break detection in nonlinear time series models. Yau and Zhao (2015) tried to estimate the multiple changepoints by using scan statistics.

Since changepoint detection is generally studied in economics to detect structural breaks, there are many studies conducted in that area. One of the first studies is developed by Chow (1960) which needs the possible breakpoint a priori. Then, it is improved with the F-statistic for the case of unknown changepoint. It is based on testing the change of parameters of a linear model. Then, Zivot and Andrews (1992) proposed an endogenous structural break test to detect possible break associated with the minimum unit root t-test statistic. Another regression based method for structural change with cointegration is developed by Gregory and Hansen (1996). In another study, Bai and Perron (1998) developed a test statistic to multiple changes in linear models. Perron (2017) made a comprehensive literature review on unit root and structural break tests. Banerjee et al. (1998) used bootstrapping for inferential purposes such as deriving the confidence interval for parameters of the marginal and conditional model to locate the multiple break. Jiang (2009) developed a Bayesian structural break model and considered the number of breaks as random and allowed a regime coefficient to include information about the other regime coefficients.

Apart from these studies, there are regime switching methods in the literature. For instance, Azavedo et al. (2014) studied Markov-switching jump, Temocin and Weber (2014) developed an approximation for controlled autonomous stochastic hybrid systems with jumps, Yerlikaya-Ozkurt et al. (2016) proposed a robust hybrid approach for CMARS, MSOM and CQP to handle outliers and Savku et al. (2015) applied stochastic hybrid models to sudden paradigm changes.

On the other hand, there are many climate related studies. Ribeiro et al. (2016) made a comprehensive review on the detection and homogenization methods used in climate studies and Peterson et al. (1998) evaluated many methods used in the literature. The most well-known relative homogeneity test is the Standard Normal Homogeneity Test (SNHT) which was proposed by Alexandersson (1986) as a likelihood ratio test. Even though Rienzner and Gandolfi (2011) improved it to capture multiple changepoints, this test has some drawbacks. First, it has the assumption of i.i.d. series which is not realistic for time series. The test also requires homogeneous reference stations; thus, their reliability and homogeneity should be validated before the test. Moreover, it needs close relative stations so that the correlation between the series obtained in the test station and reference stations must be at least 0.80. If this assumption is not satisfied, then the test station is classified as non-testable (Gokturk et al., 2008). Finally, at the end of the test, stations are classified as homogeneous, inhomogeneous and inconsistent. Later, Alexandersson and Moberg (1997) modified the test to identify the linear trends. The SNHT captures the changepoints close to the beginning and the end of the series better (Wijngaard et al., 2003). This test detects the location of the year in which the break occurs. Buishand range and the Pettitt test have this property, null and alternative hypotheses in common. Thus, these three tests are named as *location-specific* tests (Wijngaard et al., 2003).

Since it is the most popular homogeneity test, there are many studies which involve this test. For instance, Gokturk et al. (2008) applied SNHT to Turkish monthly precipitation from 267 stations. Apart from the other studies, in this study all months are treated separately and the test is applied to each individual monthly series and at the end of the test the stations are classified as homogeneous or not if the test is applicable.

Buishand range test is an absolute test which is proposed by Buishand (1982). It assumes the series is i.i.d and captures the breakpoints in the middle of the series easily (Hawkins, 1977). The test based on the adjusted partial sums is defined as

$$S_0^* = 0,$$
$$S_k^* = \sum_{i=1}^{k} \left(X_i - \overline{X}\right) \quad (k = 1, 2, \ldots, n). \tag{1}$$

If there is a break in year *K*, then S_k^* gets the maximum (negative shift) or minimum (positive shift) close to the year *k* = *K* for the series of *X,* where *n* is the sample size.

The test statistic is defined as

$R = (\max S_k^* - \min S_k^*)/s$ (s is the sample standard deviation) and the critical values can be obtained from Buishand (1982), where the max and min denote the maximum and minimum values of *S*, respectively.

Pettitt test is a nonparametric and absolute test. The ranks ($r_1, r_2, \ldots, r_n$) of the variables $X_1, X_2, \ldots, X_n$, are used to obtain the test statistic. This test needs the ranks of the series rather than the original values; hence it does not need the normally distributed X_i values. However, this makes it less sensitive to outliers (Wijngaard et al. 2003), but more sensitive to breakpoints in the middle of a time series (Hawkins, 1977). The test statistic is calculated as

$$P_k = 2\sum_{i=1}^{k} r_i - k(n+1) \quad (k = 1, 2, \ldots, n). \tag{2}$$

If there is a break in the year *E*, then the test statistic gets its maximal or minimal near the year *k* = K.

$$P_K = \max_{k=1,\ldots,n} |P_k|. \tag{3}$$

The critical values can be obtained from the original work of Pettitt (1979).

The von Neumann ratio test is also an absolute test, but cannot detect the location of the break. This property of the test is complementary to other three tests since it is more sensitive to breaks other than strict step-wise shifts (Wijngaard et al., 2003). The Von Neumann ratio is calculated by the ratio of the mean square successive difference to the variance

$$N = \sum_{i=1}^{n-1} \left(X_i X_{i+1}\right)^2 / \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 . \tag{4}$$

When there is not any breakpoint in the sample, the value of N is 2. In case of inhomogeneity, the value of the N is smaller than this expected value. If there are rapid variations in the mean, the values are higher than 2.

The KW test (Kruskal, 1952; Kruskal and Wallis, 1952) is a popular nonparametric test. This test needs at least two independent groups of samples and then compares them. Even though it is one of the mostly used tests for detecting homogeneity, the assumptions of random observations and independent populations are not valid for time series. The test statistic H is calculated as

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{m} \frac{R_i^2}{n_i} - 3(n+1), \tag{5}$$

where *n* represents the overall sample size in the whole data, n_i and R_i represent the sample size and the rank of the ith series, respectively. In addition to them, *m* represents the number of groups.

The KW test does not consider the autocorrelation in the series; thus, the performance of the Friedman Test is studied. This test which is proposed by Friedman (1937) is the nonparametric version of the repeated Analysis of Variance test. The test statistic is obtained by the sum of ranks within each column by the following expression *F*:

$$F = \frac{12}{bc(c+1)} \sum_{i=1}^{c} R_i^2 - 3b(c+1), \tag{6}$$

where R_i represents the underlying rank for the i th series ($i = 1, 2, \ldots, t$), b denotes the number of columns and c denotes the rows of the data matrix.

The KPSS test which is proposed by Kwiatkowski et al., (1992) tries to detect the trend within the series if exists. The test may capture the positive or negative shift as a trend in case of mean shift. The first step to conduct the KPSS is to regress X_t on a constant and then to obtain the least-squares residuals $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$. In the next step, the partial sum of the residuals is calculated as $e_t = \sum_{i=1}^{t} \varepsilon_i$. The KPSS test statistic is obtained by

$$KPSS = n^{-2} \sum_{t=1}^{n} \frac{e_t}{\hat{\sigma}^2}, \tag{7}$$

where $\hat{\sigma}^2$ represents the estimate of the long-run variance of the residuals. In case of nonstationarity, i.e., inhomogeneity, the null hypothesis should be rejected. An asymptotic distribution of the test statistic uses the standard Brownian bridge.

Another test which is proposed to detect the stationarity of the time series is the ADF Test. This test also tries to detect the trend within the series. Considering the *p*-th order autoregressive process, AR(p),

$$X_t = \theta_0 + \phi_1 X_{t-1} + \ldots + \phi X_{t-p} + a_t, \tag{8}$$

where θ_0 represents the function of process mean, while ϕ are model parameters and a_t is a White Noise process with zero mean and constant variance. In order to conduct the test, the process should be represented by backshift operator form, $(1 - \phi_1 B - \ldots - \phi_p B^p) X_t = \theta_0 + a_t$. If some of roots of the polynomial $(1-\phi_1 B - \ldots - \phi_p B^p)$ are 1, then the series is not stationary. To detect whether the series is stationary or not, the Equation 9 is used as

$$\Delta X_t = \gamma X_{t-1} + \sum_{j=1}^{p-1} \varphi_j \Delta X_{t-j} + \theta_0 + a_t. \tag{9}$$

If γ =1, there is a unit root in the series, *i.e.*, the series is not stationary. The critical values of the test can be found in the study of Said and Dickey (1984).

GAHMDI which is proposed by Toreti et al. (2012) tries to find the *M* possible changepoints and the homogeneous segments. This method uses Hidden Markov models and a genetic algorithm (GA) in order to obtain the global maximum. The first step of the method is to decide on the maximum number of changepoints, *Mmax*. The GAHMDI method uses a likelihood approach and then an expectation-maximization procedure in order to complete maximization. GA helps to estimate the initial state sequence in order to get global maxima. Then, the number of states is obtained by using the minimum description length.

The Bayesian method of change-point analysis is also known as the product partition model (Barry and Hartigan, 1992; 1993). This procedure can be used as an absolute and a relative test. The posterior probability of a break point for all time points of the series is presented and the posterior mean is approximated by Markov Chain Monte Carlo (MCMC). This method is also capable of detecting multiple changepoints. It is assumed that the change-points are identical and have the geometric distribution and independent priors for the parameters in order to capture them in the mean of normally distributed variables. The partition $U = (U_1, U_2, \ldots, U_n)$ where U_i = 1 indicates a changepoint at position *i*+1, is used to conduct the test. In the first step, U_i is equal to 0 for all $i < n$, with U_n = 1. In every process of the Markov chain, at each position *i*, a value of U_i is selected from the conditional distribution of U_i given all observations and the current partition. The transition probability for the conditional probability of a changepoint at the position *i*+1 may be calculated by using the ratio

$$\frac{p_i}{1-p_i} = \frac{P\left(U_i = 1 \middle| \mathrm{X}, U_j, j \neq i\right)}{P\left(U_i = 0 \middle| \mathrm{X}, U_j, j \neq i\right)}, \tag{10}$$

where *X* represents the series. This series can also be used with another series, *Y,* and include some explanatory variables, such as reference information.

RHTest is proposed by Wang (2008a, 2008b). This method can capture single or multiple mean changepoints and it is based on an empirical approach which accounts for lag-1 autocorrelation. RHTest is based on the penalized maximal t-test (PMT) or F test (PMF) suggested by (Wang et al., 2007). First, the most possible changepoint c$_0$ $\left(t \in \{n_{\min}, n_{\min}+1, \ldots, n-n_{\min}\}\right)$ is identified. Then, the next possible changepoints; *c*$_{01}$ and *c*$_{03}$ are searched in the segments $t \in \{n_{\min}, \ldots, (c_0 - n_{\min})\}$ and $t \in \{c_0 + 1 + n_{\min}, \ldots, (n - n_{\min})\}$, respectively. Then, the new estimate of c$_0$, represented by c$_{02}$ is identified. PMT test statistic, PTmax is needed to determine the most significant changepoint. The following models are conducted in order to estimate the p-value of the PTmax,

$$\begin{aligned} &R_t = X_t - \hat{X}\left(t = 1,2,\ldots,n\right), \\ &W_1 = R_1, \\ &W_t = R_t - \hat{\phi} R_{t-1}\left(t = 2,\ldots,n\right). \end{aligned} \tag{11}$$

where $\hat{X}_t$ indicates the full model fit to the series *X*$_t$, and *R*$_t$ denotes the residuals. Then, the prewhitened series, *W*$_t$ is obtained in order to get the t-statistics and p-value for *c*. The changepoint associated with the maximum PTmax is determined. If it is significant, this point is classified as a changepoint.

Caussinus and Mestre (2004) (CauMe) suggested a method to capture the changepoints and then correct them in a climate series. This test assumes the normally distributed differences between the test and reference series. One of the advantages of this method is that it uses adapted penalized log-likelihood procedure which makes it useable when there is not any reliable reference series. This method considers that each observation in a climate series represents the climate and station effect in addition to a random white noise. The following model is used to conduct the test

$$E(X_{ij}) = \mu_i + \nu_{jh(i,j)}, \tag{12}$$

where μ_i represents the climate effect at time i and ν_{jh} represents the station effect of station j for level L_{jh}. When the sample size of the series is large, the number of hypotheses increases which makes it computationally inefficient. In order to prevent this, a dynamic programming algorithm is used for pairwise comparisons.

Two Phase Regression (TPR) is also used to detect changepoints and proposed by Easterling and Peterson (1995). Similar to other relative tests, the differences between the test and reference series is calculated. A simple linear regression (SLR) is conducted where the differences are regressed on the time variable and the residual sum of squares, RSS$_1$ of the model is calculated. Then, two different SLR is conducted before and after the possible breakpoint, k, and the sum of the residual sum of squares (RSS$_2$) of these two models are obtained for each possible breakpoint k. The maximum value of RSS$_2$ is determined as the test statistic and then the significance of the test is determined by using the test statistic U:

$$U = \frac{(RSS_1 - RSS_2)/2}{RSS_2/(n-4)}. \tag{13}$$

Lund and Reeves (2002) presented the critical values. If a possible breakpoint is determined, its magnitude is estimated by the differences of the averages before and after the breakpoint.

After introducing the common methods to detect inhomogeneities, the studies comparing the performances of the methods are discussed. The comparison studies (Guijarro, 2013; Ducre-Robitaille et al., 2003) generally include simulation studies. However, they either simulated i.i.d. data or they create only one type of shift like mean shift which is not realistic or not enough for comparison. Toreti et al. (2012) proposed an approach based on a genetic algorithms and hidden markov models (GAHMDI) to detect inhomogeneities. Caussinus and Mestre (2004) detects the number of breaks and outliers by using an adapted penalized log-likelihood procedure. Wang (2008a, 2008b) proposed an approach in order to account for lag-1 autocorrelation to capture mean shifts. Toreti et al. (2012) compared their method with the SNHT, the RHtest (Wang, 2008a), and the method developed by Caussinus and Mestre (2004). In their work, GAHMDI overperforms the other three methods. Huskova and Kirch (2012) used bootstrapping regression methods to find the critical values of the sequential changepoint tests.

Buishand (1982) applied his test to annual data from 264 stations of 30 years long and compared the results with that of the von Neumann ratio test and concluded that Buishand range test performs better than von Neumann ratio test for the model of one changepoint in the mean. Wijngaard et al. (2003) applied SNHT, the Buishand range, the Pettitt test and the Von Neumann tests to the daily European series and classified the series as *useful* (if at least one test indicates inhomogeneity), *doubtful* (if two tests indicate inhomogeneity) and *suspect* (if at least three tests indicates inhomogeneity). Hanssen-Bauer and Forland (1994) proposed a four-step approach to define the reference series as homogeneous in SNHT if there is no prior information about the homogeneity of the reference series. Gonzalez-Rouco et al. (2001) used SNHT for the 95 monthly precipitation series of the Southwest Europe and added a one step to the proposed method of Gonzalez – Rouco et al. (2001) in order to classify the stations according to its homogeneity. Karabork et al. (2007) compared the performances of the SNHT and Pettitt tests for the annual precipitation totals of 212 stations of Turkey. If both tests indicate homogeneity, the series is classified as homogeneous. Thus, 43 out of 212 stations are classified as inhomogeneous and the other stations are detected as homogeneous. Sahin and Cigizoglu (2010) compared the performances of SNHT, Pettitt, von Neumann, Buishand range and the bivariate test developed by Maronna and Yohai (1978) on 250 meteorological stations of Turkey and concluded that the relative tests performs better than the absolute tests. Tayanc et al. (1998) conducted a comparative study by using Kruskal–Wallis and Wald–Wolfz methods to detect the inhomogeneity of the temperature series of Turkey. Their study revealed that 50 stations out of 82 were classified as homogeneous.

Li and Lund (2012) proposed a method based on a genetic algorithm to detect the number of changepoints and their locations. Guijarro (2013) compared *t*-test, SNHT, two-phase regression (TPR), Wilcoxon-Mann-Whitney test, Durbin-Watson (DW) test and squared relative mean difference on windows running along the series and conclude that SNHT is the best performing test Moreover, Ducre-Robitaille et al. (2003) compared the performances of the SNHT with and without trend, Multiple Linear regression (MLR), TPR, Wilcoxon rank-sum, sequential testing for equality of means, Bayesian approach with and without reference series on a simulated temperature series and concluded that SNHT and MLR have the better performances.

The studies in the literature that are mentioned tried to detect the mean shift type but not any change in the variance, sudden increase or decrease or trend cases. On the other hand, the studies that covers AR or ARMA models deal with conditional likelihood, but not exact likelihood.

In our research group, a simulation study is conducted in order to compare the performances of the homogeneity tests most commonly used. First, the temperature model is estimated from the data set which consists of 244 meteorological stations of Turkey. The results imply that the appropriate model for the monthly temperature series is a seasonal dummy model. By using similar coefficients, two reference series are created and the performances of the SNHT, Pettitt, Buishand range, Chow test, von Neumann and Kruskal Wallis are compared (Yazici et al., 2012; Yozgatlıgil, 2011) under several scenarios which represent the sources of inhomogeneity like mean shift, trend, gradual change and sudden decrease. The results imply that SNHT is the best method in terms of detecting the breakpoints.

CHAPTER 3

METHODS

The methods that are used in the mentioned study are explained in this chapter. First, the motivation of the study is explained and then the likelihood ratio test is given in general. Next, the derivation of LRT for level shifts for AR(1) and AR(*p*) are presented. In the next subsections, the best performing relative and absolute changepoint detection tests; the SNHT and F-test are explained in detail. Then, the proposed approach; the moving block bootstrap is given and the application of bootstrap to LRT is explained in the last subsection.

3.1. Motivation

Detection of inhomogeneity, if exists, is an important problem in time series data. There are many sources from which inhomogeneity can be originated such as mean shift, variance and trend change, gradual change, or sudden decrease or increase in time series. Figure 1 illustrates the examples of these cases.

There are many methods developed for changepoint detection. In Yozgatligil and Yazici (2016) the methods in the literature are compared based on simulation of temperature series. The results indicate that the SNHT has the best performance in terms of capturing the breakpoint. However, it has some drawbacks such as its performance is getting worse if the level shift occurs at the beginning or at the end of the series especially for small amounts of shifts. Moreover, it is a relative method which needs highly correlated reference series. Thus, an absolute test which detects changepoints especially in the beginning or at the end of the series is needed.

In this thesis, first likelihood ratio test (LRT) based on the exact likelihood for autoregressive models to capture mean shift is constructed. A simulation study is conducted to obtain the critical values of the test since its distribution is not known. Then, computational approach involving bootstrapping is used to improve the performance of likelihood ratio test for level shifts in autoregressive models to detect changepoints occurred at the beginning or end of the series.

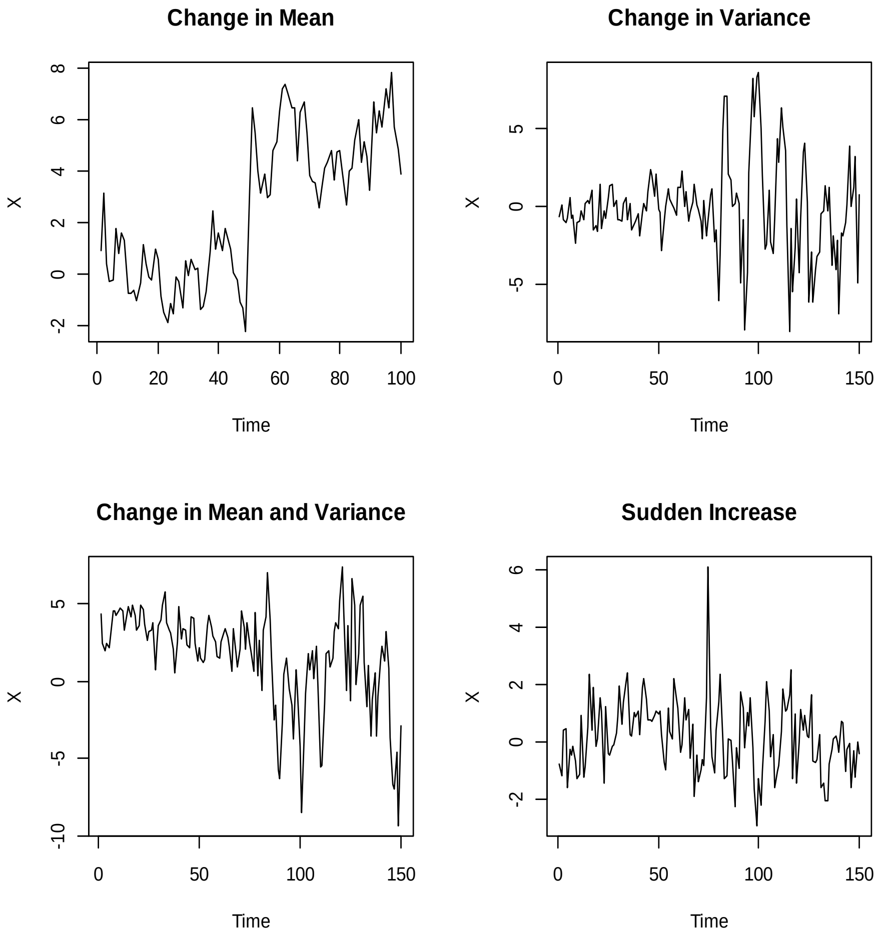


Figure 1. Examples of changepoint in time series.

3.2. Likelihood Ratio Test

The likelihood ratio test is a likelihood based test designed for testing hypotheses and related with maximum likelihood estimators. If the null and alternative hypotheses are defined as $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^C$, where θ is $b\times 1$ vector of parameters, Θ_0 and Θ_0^C are the parameter spaces specified in the null and alternative hypotheses, respectively, and X represents the i.i.d. data. Then the likelihood ratio test statistic is defined as

$$\lambda(x) = \frac{\sup_{\Theta_0} L(\theta \mid X)}{\sup_{\Theta} L(\theta \mid X)} \tag{14}$$

where Θ represents the parameter space and L represents the likelihood of the data. The rejection region of the test is $\{x : \lambda(x) \leq k\}$, where k is any number which satisfies $0 \leq k \leq 1$ (Bain and Engelhardt, 1992).

Wilks (1938) stated that if the study consists i.i.d. data for large n, $-2\ln(\lambda(x)) \sim \chi_b^2$. Thus, an approximate size α test is to reject the null hypothesis if $-2\ln(\lambda(x)) \sim \chi_\alpha^2(b)$. However, it is no longer valid for unknown changepoints (Wei, 2017; Karioti and Caroni, 2002).

3.2.1. Likelihood Ratio Test for Level Shift for Autoregressive Models

In that section, the likelihood of AR(1) and AR(p) are studied and the related test statistics are obtained.

3.2.1.1. The First-Order Autoregressive Model (AR(1))

AR(1) which is a linear model used to predict the present value of a time series. It also uses the immediately prior value in time and the model is represented as

$X_t = \delta + \phi\ X_{t-1} + \varepsilon_t$

where $\varepsilon_t \sim WN(0, \sigma^2)$. (15)

Here, δ determines the mean of the process. If $\delta = 0$, then the mean of the process is 0. In order to be stationary, the $|\phi| < 1$ condition should be satisfied. For a stationary AR(1), the process mean is $E(X_t) = \dfrac{\delta}{1-\phi}$ and the process variance is $Var(X_t) = \dfrac{\sigma^2}{1-\phi^2}$. The autocorrelation function is defined as $\rho_k = \dfrac{\gamma_k}{\gamma_0} = \phi^k$, $k \geq 1$.

3.2.1.2. Maximum Likelihood Estimators of AR(1)

In the case of identically distributed and independent random variables, the likelihood function is just the multiplication of marginal pdf of random variables. However, in time series analysis, the dependence structure of observation is specified and joint pdf is considered. To ease the calculations, conditional densities are used.

Consider the AR(1) model with Gaussian errors. For the model $X_t = \delta + \phi X_{t-1} + \varepsilon_t$, the i.i.d. errors are $\varepsilon_t \sim N(0, \sigma^2)$ and the parameter vector is $\boldsymbol{\theta} = (\delta, \phi, \sigma^2)^T$.

The distribution of the first observation is $X_1 \sim N\left(\dfrac{\delta}{1-\phi}, \dfrac{\sigma^2}{1-\phi^2}\right)$ and

$$f(x_1; \boldsymbol{\theta}) = f(x_1; \delta, \phi, \sigma^2) = \frac{1}{\sqrt{2\pi}\sqrt{\dfrac{\sigma^2}{1-\phi^2}}} \exp\left\{ -\frac{1}{2} \frac{\left[x_1 - \dfrac{\delta}{1-\phi}\right]^2}{\dfrac{\sigma^2}{1-\phi^2}} \right\}.$$

(16)

The conditional distribution of the second observation X_2 conditional on $X_1 = x_1$ is obtained from the AR(1) model

$$X_2 = \delta + \phi X_1 + \varepsilon_2 . \tag{17}$$

Hence, $\left(X_2 \mid X_1 = x_1\right) \sim N\left(\delta + \phi X_1, \sigma^2\right)$ and

$$f_{X_2|X_1}\left(x_2 \mid x_1;\boldsymbol{\theta}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{\left(x_2 - \delta - \phi x_1\right)^2}{\sigma^2}\right\}. \tag{18}$$

The joint density of X_1 and X_2 is obtained by the multiplication of the conditional and marginal densities as

$$f_{X_2,X_1}\left(x_2, x_1;\boldsymbol{\theta}\right) = f_{X_2|X_1}\left(x_2 \mid x_1;\boldsymbol{\theta}\right) f_{X_1}\left(x_1;\boldsymbol{\theta}\right). \tag{19}$$

The conditional distribution of X_3 given the first two observations can be derived similarly and obtained as

$$f_{X_3|X_2,X_1}(x_3 \mid x_2, x_1;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{\left(x_3 - \delta - \phi x_2\right)^2}{\sigma^2}\right\}. \tag{20}$$

The joint density of the first three observations can be obtained as

$$\begin{aligned} f_{X_3,X_2,X_1}(x_3, x_2, x_1;\boldsymbol{\theta}) &= f_{X_3|X_2,X_1}\left(x_3 \mid x_2, x_1;\boldsymbol{\theta}\right) f_{X_2,X_1}\left(x_2, x_1;\boldsymbol{\theta}\right) \\ &= f_{X_3|X_2,X_1}\left(x_3 \mid x_2, x_1;\boldsymbol{\theta}\right) f_{X_2|X_1}\left(x_2 \mid x_1;\boldsymbol{\theta}\right) f_{X_1}\left(x_1;\boldsymbol{\theta}\right). \end{aligned} \tag{21}$$

The value of X_1, X_2, ..., X_{t-1} has an effect on X_t only through the value of X_{t-1} and the density of observation t conditional on the preceding t-1 observations given by

$$\begin{gathered} f_{X_t|X_{t-1},X_{t-2},\ldots X_1}\left(x_t \mid x_{t-1}, x_{t-2}, \ldots, x_1;\boldsymbol{\theta}\right) = f_{X_t|X_{t-1}}\left(x_t \mid x_{t-1};\boldsymbol{\theta}\right) = \\ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{\left(x_t - \delta - \phi x_{t-1}\right)^2}{\sigma^2}\right\}. \end{gathered} \tag{22}$$

The likelihood of the complete sample is then obtained as

$$f_{X_n,X_{n-1},X_{n-2},\ldots X_1}(x_n,x_{n-1},\ldots,x_1;\boldsymbol{\theta}) = f_{X_1}(x_1;\boldsymbol{\theta})\prod_{\mathbf{t=2}}^{\mathbf{n}} f_{X_T|X_{T-1}}(x_t \mid x_{t-1};\boldsymbol{\theta}) \qquad (23)$$

The exact log-likelihood for a sample of size n from a Gaussian AR(1) process is

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\log\left(\frac{2\pi\sigma^2}{1-\phi^2}\right) - \frac{\left(x_1 - \dfrac{\delta}{1-\phi}\right)^2}{2\dfrac{\sigma^2}{1-\phi^2}} + \sum_{t=2}^{n}\left\{-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{\left(x_t - \delta - \phi x_{t-1}\right)^2}{2\sigma^2}\right\}. \qquad (24)$$

The exact log-likelihood is a non-linear function of the parameters θ. Thus, there is no closed form solution in order to obtain the maximum likelihood estimates. The exact estimates can be obtained by numerically maximizing the log-likelihood.

3.2.1.3. Test for a Breakpoint in AR(1) Models

The purpose of this study is to conduct a test for detecting a breakpoint of an AR(1) for mean shift. The hypothesis is that the AR(1) process does not have any changepoint against the alternative hypothesis that the sequence does have a changepoint. The changepoint is considered as the change of either the correlation coefficient ϕ, or the change of the drift δ. In this test, only a changepoint in the drift δ is considered. Otherwise, the variance of the process also varies by time.

The stationary AR(1) model can be shown as

$$X_t = \begin{cases} \delta_0 + \phi X_{t-1} + \varepsilon_t,\ 1 \le t \le k-1, \\ \delta_1 + \phi X_{t-1} + \varepsilon_t,\ k \le t \le n \end{cases} \qquad (25)$$

where $(t = 1, 2, \ldots, n)$ and the stationarity condition is satisfied if $|\phi| < 1$.

If there is not a change in the mean, δ_0 and δ_1 should be the same. Thus, the hypothesis can be represented as

$H_0 : \delta_o = \delta_1$,
$H_1 : \delta_o \neq \delta_1$.

The null hypothesis indicates no changepoint, while the alternative expresses a single changepoint at time *k*. Under *H*1, the model can be specified as in Eqn (25).

Then, the likelihood ratio test is adopted to test the hypothesis. Under *H*0, the parameter space is $\Theta_0 = \{\delta_0 = \delta_1 = \delta, \sigma > 0, -1 < \phi < 1\}$, while under *H*1, it is defined as $\Theta^C = \{\delta_0 \neq \delta_1, \sigma > 0, -1 < \phi < 1\}$ and the overall parameter space is $\Theta = \{\delta_0, \delta_1, \sigma > 0, -1 < \phi < 1\}$.

The likelihood ratio, $\lambda(x)$, defined as

$$\lambda(x) = \frac{\sup\{L(\theta; x) : \theta \in \Omega_0\}}{\sup\{L(\theta; x) : \theta \in \Omega\ \}} \tag{26}$$

and the ratio is always between 0 and 1. It is an evidence to reject the null hypothesis if the ratio is small.

Since the function -2log($\lambda(x)$) is a decreasing function of *x*, the critical region of the test can be expressed in the form

$$C = \{x : -2\log\lambda(x) \geq c\}. \tag{27}$$

Then, the likelihood ratio statistic is defined as below

$\Lambda(x) = -2\log\lambda(x) = 2\left(l(\hat{\theta}; x) - l(\theta_0; x)\right)$. Under regularity conditions (Rao and Scott, 1987),

$\Lambda(x) \xrightarrow{d} \chi^2_q$, where q = dim(Θ)-dim(Θ_0).

In order to test the hypothesis for an AR(1) model of a possible breakpoint, the denominator of the $\lambda(x)$ is obtained as follows:

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\log\left(\frac{2\pi\sigma^2}{1-\phi^2}\right) - \frac{\left(x_1 - \frac{\delta_0}{1-\phi}\right)^2}{2\frac{\sigma^2}{1-\phi^2}} + \sum_{t=2}^{k-1}\left\{-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{\left(x_t - \delta_0 - \phi\, x_{t-1}\right)^2}{2\sigma^2}\right\}$$

$$+\sum_{t=k}^{n}\left\{-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{\left(x_t - \delta_1 - \phi\, x_{t-1}\right)^2}{2\sigma^2}\right\}. \tag{28}$$

Then, the following likelihood ratio test statistics is obtained:

$$\Lambda(x) = 2\left\{\begin{array}{l} -\frac{1}{2}\log\left(\frac{2\pi\sigma^2}{1-\phi^2}\right) - \frac{\left(x_1 - \frac{\delta_0}{1-\phi}\right)^2}{2\frac{\sigma^2}{1-\phi^2}} + \sum_{t=2}^{k-1}\left\{-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{\left(x_t - \delta_0 - \phi\, x_{t-1}\right)^2}{2\sigma^2}\right\} \\ + \sum_{t=k}^{n}\left\{-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{\left(x_t - \delta_1 - \phi\, x_{t-1}\right)^2}{2\sigma^2}\right\} \end{array}\right\}$$

$$-2\left\{-\frac{1}{2}\log\left(\frac{2\pi\sigma^2}{1-\phi^2}\right) - \frac{\left(x_1 - \frac{\delta}{1-\phi}\right)^2}{2\frac{\sigma^2}{1-\phi^2}} + \sum_{t=2}^{n}\left\{-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{\left(x_t - \delta - \phi x_{t-1}\right)^2}{2\sigma^2}\right\}\right\}. \tag{29}$$

Since there is no closed form solution for the maximum likelihood estimates in exact log-likelihood functions, Newton-Raphson method is used.

Since the location of the breakpoint, *k*, is not known in advance, the test statistic is calculated as

$$\sup_{k=2,\dots,n} \left|\Lambda_k\right| = \left|\Lambda_s\right|,$$

where $s \in \{2,\dots,n\}$ for a sample size of *n* (Wei 2017, Karioti and Caroni, 2002) and the distribution of this test statistic is not known. If the supremum is greater than the critical level, *C*, then the null hypothesis is rejected. In the previous studies, several values for *C* is recommended. For conditional likelihood approach, Tsay (1988) used 3.0, 3.5, and 4.0; Chen and Tiao (1990) used 2.8, 3.0, and 3.3; Chen and Liu (1993) used 2.3 and 3.4 while Galeano et al. (2006) used values between 2.9 and 4.0. Conditioning leads to loss of information (Karioti and Caroni, 2002), thus, in this study exact likelihood is used to derive the test statistic to prevent this. Hence, suggested critical values are not valid for our case. To obtain the critical values, first, the likelihood based on exact likelihood is obtained and then a simulation study for $\phi = -0.9,\ldots,0.9$ of sample sizes of 50, 75 and 100 is conducted to obtain the critical values of the test.

3.2.2. Likelihood Ratio Test for Level Shift for AR(*p*) Models

To be able to provide a general testing procedure, we derive a test for AR(*p*) process.

The stationary AR(*p*) model can be expressed as

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t , \tag{30}$$

where $\varepsilon_t \sim N(0,\sigma^2)$.

The likelihood function is calculated by conditional pdfs. The vector of parameters for an AR(*p*) model is $\boldsymbol{\theta} = (\delta, \phi_1, \phi_2, \ldots, \phi_p, \sigma^2)^T$. Here, the joint density of the first *p* time series variables, ($x_1, x_2, \ldots, x_p$), is Multivariate Normal Distribution, i.e.,

$$f_{X_1,X_2,\ldots,X_p}(x_1,x_2,\ldots,x_p;\boldsymbol{\theta}) = (2\pi)^{-p/2}\left|\sigma^{-2}\boldsymbol{V}_p^{-1}\right|^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{x}_p - \boldsymbol{\mu}_p)^T \boldsymbol{V}_p^{-1}(\boldsymbol{x}_p - \boldsymbol{\mu}_p)\right\}, \tag{31}$$

where $\mu = \delta/(1-\phi_1-\phi_2-\ldots-\phi_p)$ and $\sigma^2\boldsymbol{V}_p$ is the variance – covariance matrix of the first *p* observations which is defined as

$$\sigma^2 \boldsymbol{V}_p = \begin{bmatrix} E(X_1-\mu)^2 & E(X_1-\mu)(X_2-\mu) & \cdots & E(X_1-\mu)(X_p-\mu) \\ E(X_2-\mu)(X_1-\mu) & E(X_2-\mu)^2 & \cdots & E(X_2-\mu)(X_p-\mu) \\ \vdots & \vdots & \cdots & \vdots \\ E(X_p-\mu)(X_1-\mu) & E(X_p-\mu)(X_2-\mu) & \cdots & E(X_p-\mu)^2 \end{bmatrix}. \tag{32}$$

For the other variables, the conditional density is used (Hamilton, 1994). The *p* most recent observations are used for the remaining observations.

$$\begin{aligned} & f_{X_t|X_{t-1},X_{t-2},\ldots,X_1}(x_t \mid x_{t-1}, x_{t-2}, \ldots, x_1; \boldsymbol{\theta}) = \\ & f_{X_t|X_{t-1},X_{t-2},\ldots,X_{t-p}}(x_t \mid x_{t-1}, x_{t-2}, \ldots, x_{t-p}; \boldsymbol{\theta}) = \\ & \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x_t - \delta - \phi_1 x_{t-1} - \phi_2 x_{t-2} - \ldots - \phi_p x_{t-p})^2}{2\sigma^2}\right]. \end{aligned} \tag{32}$$

The likelihood function for the complete sample of sample size *n* is

$$\begin{aligned} & f_{X_1,X_2,\ldots,X_T}(x_1, x_2, \ldots, x_t; \boldsymbol{\theta}) = \\ & f_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p; \boldsymbol{\theta}) \times \prod_{t=p+1}^{n} f_{X_t|X_{t-1},X_{t-2},\ldots,X_{t-p}}(x_t \mid x_{t-1}, x_{t-2}, \ldots, x_{t-p}; \boldsymbol{\theta}). \end{aligned} \tag{33}$$

Thus, the log likelihood is

$$\begin{aligned} & \ell(\boldsymbol{\theta}) = -\frac{p}{2}\log(2\pi) - \frac{p}{2}\log(\sigma^2) + \\ & \frac{1}{2}\log\left|\boldsymbol{V}_p^{-1}\right| - \frac{1}{2\sigma^2}(\boldsymbol{x}_p - \boldsymbol{\mu}_p)^T \boldsymbol{V}_p^{-1}(\boldsymbol{x}_p - \boldsymbol{\mu}_p) - \frac{n-p}{2}\log(2\pi) \\ & -\frac{n-p}{2}\log(\sigma^2) - \sum_{t=p+1}^{n} \frac{(x_t - \delta - \phi_1 x_{t-1} - \phi_2 x_{t-2} - \ldots - \phi_p x_{t-p})^2}{2\sigma^2} \end{aligned} \tag{34}$$

where $\theta \in \Theta$.

3.2.2.1. Test for a Breakpoint in AR(*p*) Models

Hypotheses for a single break are defined as follows

$$H_0 : \delta_o = \delta_1,$$
$$H_1 : \delta_o \neq \delta_1.$$

Under H_1, the model can be specified as

$$X_t = \begin{cases} \delta_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t,\ 1 \le t \le k-1, \\ \delta_1 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t,\ k \le t \le n, \end{cases}$$

where $t = 1, \ldots, n.$ (35)

where $\varepsilon_t \sim N(0, \sigma^2)$. The log likelihood under H_1 is

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= -\frac{p}{2}\log(2\pi) - \frac{p}{2}\log(\sigma^2) \\ &+ \frac{1}{2}\log\left|\boldsymbol{V}_p^{-1}\right| - \frac{1}{2\sigma^2}(\boldsymbol{x}_p - \boldsymbol{\mu}_p)^T \boldsymbol{V}_p^{-1}(\boldsymbol{x}_p - \boldsymbol{\mu}_p) - \frac{n-p}{2}\log(2\pi) \\ &- \frac{n-p}{2}\log(\sigma^2) - \sum_{t=p+1}^{k-1} \frac{(x_t - \delta_0 - \phi_1 x_{t-1} - \phi_2 x_{t-2} - \ldots - \phi_p x_{t-p})^2}{2\sigma^2} \\ &- \sum_{t=k}^{n} \frac{(x_t - \delta_1 - \phi_1 x_{t-1} - \phi_2 x_{t-2} - \ldots - \phi_p x_{t-p})^2}{2\sigma^2}. \end{aligned} \tag{36}$$

Thus, the likelihood ratio test statistic is

$$\begin{aligned} \Lambda(x) = 2&\left\{ \begin{aligned} &-\frac{p}{2}\log(2\pi) - \frac{p}{2}\log(\sigma^2) \\ &+\frac{1}{2}\log\left|\boldsymbol{V}_p^{-1}\right| - \frac{1}{2\sigma^2}(\boldsymbol{x}_p - \boldsymbol{\mu}_p)^T \boldsymbol{V}_p^{-1}(\boldsymbol{x}_p - \boldsymbol{\mu}_p) - \frac{n-p}{2}\log(2\pi) \\ &-\frac{n-p}{2}\log(\sigma^2) - \sum_{t=p+1}^{k-1} \frac{(x_t - \delta_0 - \phi_1 x_{t-1} - \phi_2 x_{t-2} - \ldots - \phi_p x_{t-p})^2}{2\sigma^2} \\ &-\sum_{t=k}^{n} \frac{(x_t - \delta_1 - \phi_1 x_{t-1} - \phi_2 x_{t-2} - \ldots - \phi_p x_{t-p})^2}{2\sigma^2} \end{aligned} \right\} \\ -2&\left\{ \begin{aligned} &-\frac{p}{2}\log(2\pi) - \frac{p}{2}\log(\sigma^2) + \frac{1}{2}\log\left|\boldsymbol{V}_p^{-1}\right| - \frac{1}{2\sigma^2}(\boldsymbol{x}_p - \boldsymbol{\mu}_p)^T \boldsymbol{V}_p^{-1}(\boldsymbol{x}_p - \boldsymbol{\mu}_p) \\ &-\frac{n-p}{2}\log(2\pi) \\ &-\frac{n-p}{2}\log(\sigma^2) - \sum_{t=p+1}^{n} \frac{(x_t - \delta - \phi_1 x_{t-1} - \phi_2 x_{t-2} - \ldots - \phi_p x_{t-p})^2}{2\sigma^2} \end{aligned} \right\}. \end{aligned} \tag{37}$$

If the underlying process is ARMA(p,q) process, the process is written in inverted form and approximated by AR(pmax) when $P_{\max} = \left[12 \left(\frac{n}{100} \right)^{1/4} \right]$ where $\lfloor x \rfloor$ denotes the integer part of x (Schwert, 1989). Then, by using similar methodology, we obtain the critical values for different p-values under various sample sizes to be able to conduct the test.

3.3. Standard Normal Homogeneity Test (SNHT)

SNHT which is the most popular homogeneity test is a likelihood based test proposed by Alexandersson (1986) to capture breaks in climate studies. It is designed to detect single mean shifts and does not concern the autocorrelation in time series. This relative test which assumes normality needs reference series that are highly correlated (at least 0.80) with the test series. The relative series should also be close to the test series in terms of location.

The calculation of the test statistic starts with obtaining Q values which is explained below.

$$Q_i = Y_i - \sum_{j=1}^{r} \rho_j^2 \left[X_{ji} - \overline{X}_j - \overline{Y} \right] / \sum_{j=1}^{r} \rho_j^2, \quad \text{and} \qquad i = 1, \ldots, n, \tag{38}$$

where X and Y (Y_{j} represents the j$^{\text{th}}$ reference series) are the test reference series, respectively and ρ represents the weight between test and reference series. This weight is generally used as the correlation between these series. This formula is used for temperature series, while it becomes

$$Q_i = \frac{X_i}{\left[\sum_{j=1}^{r} \rho_j^2 Y_{ji} \overline{X} / \overline{Y}_j \right] / \sum_{j=1}^{r} \rho_j^2} \tag{39}$$

for precipitation series.

Then, the Q values are normalized as follows:

$$Z_i = \frac{\left(Q_i - \overline{Q}\right)}{\hat{\sigma}_Q} \ , \tag{40}$$

where $\overline{Q}$ is the mean and $\hat{\sigma}_Q$ is the estimated standard deviation of the Qi values.

The normality assumption is applied here, and the null and alternative hypotheses are constructed as

$$H_0 : Z_i \in N(0,1), \ \ i \in \{1,\ldots,n\},$$

$$H_1 : \begin{cases} Z_i \in N(\mu_1,\sigma) \ \ \mathit{if}\ i \in \{1,\ldots,a\}, \\ Z_i \in \ N(\mu_2,\sigma) \ \ \mathit{if} \ \ i \in \{a+1,\ldots,n\}, \end{cases}$$

where a is the possible changepoint, μ_1 is the mean of the first a observations and μ_2 is the mean of the last (n-a) observations for a sample size n.

Hence, the test statistic is calculated as

$$T_{\max} = \max_{1 \le a \le n-1} \{T_a\} = \max_{1 \le a \le n-1} \left\{a\bar{z}_1^2 + (n-a)\bar{z}_2^2\right\} \tag{41}$$

where $\bar{z}_1$ and $\bar{z}_2$ are the mean values before and after the shift. The test statistic is compared with the critical values obtained by Alexandersson (1986), and if $T_{\max}$ is greater than the critical value which is given in Alexandersson (1986) for the selected significance level and then, the test series is classified as an inhomogeneous series.

3.4. F-Test

In economics, the Chow test is proposed for structural breaks. It is a model based test proposed by Chow (1960). To apply the test, the following model is considered.

$$Y_t = X_t^T \beta_t + \varepsilon_t \quad (t = 1, 2, \ldots, n), \tag{42}$$

where X_t is the vector of independent variables and variables ε_t are i.i.d. with $E(\varepsilon_t) = 0$ and constant variance. The test is capable of being used as an absolute test and a relative test if there are reference series. X_t represents the time series obtained in j reference series if it used as a relative series. The null hypothesis is $H_0 : \beta_t = \beta_0$, while the alternative hypothesis states that the series includes a structural break. Thus, in the alternative hypothesis, the parameters are represented as

$$\beta_t = \begin{cases} \beta_A \,, & 1 \le t \le K, \\ \beta_B \,, & K < t < n, \end{cases} \tag{43}$$

where *K* is the changepoint in the interval (*k, n* − *k*). In the original version, the breakpoint is known in advance and then it is modified for all possible changepoints in the interval (1, *n*). The related test statistic is defined as

$$F_i = \frac{\hat{u}^T \hat{u} - \hat{e}^T \hat{e}}{\hat{e}^T \hat{e} / (n - 2k)},$$

where $\hat{e} = (\hat{u}_A, \hat{u}_B)^T$ and the F test statistic is distributed as χ_k. The results are compared by SNHT and F-test.

3.5. Bootstrap

The bootstrap is a computer-intensive computational method that provides answers to inference problems (Lahiri, 2003). It is proposed by Efron (1979) as a resampling technique which considers the data as a population and obtains samples from it with replacement. Bootstrap is widely used for estimating biases, standard errors and parameters. The approach is generalized to solve problems in independent but not identically distributed data sets, dependent data, and discrimination and regression problems. The application of bootstrap includes constructing confidence intervals, estimation of standard errors and biases and to obtain critical values of some tests.

The bootstrap methods for dependent data have provided new approaches to solve problems about inferential statistics. The bootstrap methods for dependent data is still an active research area. In this study, a bootstrap method for dependent data is considered to increase the performance of our testing methodology. The data generating process for dependent data is not fully defined, thus there is no unique way to resample from the data (Mammen and Nandi, 2012). The important point is to capture the dependence structure. The most widely used bootstrap methods for dependent data are block, sieve, the nonparametric autoregressive bootstrap, frequency domain and Markov bootstrap and subsampling. Block bootstrap methods have been studied under the assumption of stationarity (Gonçalves and Politis, 2011).

3.5.1. Moving Block Bootstrap (MBB)

The block bootstrap has a similar approach to the nonparametric i.i.d. bootstrap and proposed by Künsch (1989). However, in this method, blocks of consecutive observations are taken with replacement instead of single observations. This method is valid for stationary processes. First, a set of blocks of consecutive observations are constructed and then, the blocks are selected with replacement. When it is first proposed, nonoverlapping blocks of fixed length *l*: $\{X_j : j = 1,\ldots,l\},\{X_{l+j} : j = 1,\ldots,l\},\ldots$ was used. Then, it is suggested to use all possible (overlapping and nonoverlapping) blocks of length *l*, i.e. the *r*-th block consists of the observations $\{X_{r-1+j} : j = 1,\ldots,l\}$ which is also known as Moving Block Bootstrap (MBB). The bootstrap sampling procedure which is illustrated in Figure 2 is constructed by sampling n/l blocks randomly with replacement and combining to a time series of length *n* for different *B* bootstrap samples. The distribution of the bootstrap time series is a nonstationary (conditional) distribution by construction. When the block length, *l*, is random and generated from a geometric distribution, the resample becomes stationary and called as stationary bootstrap. Recently, Paparoditis and Politis (2001, 2002) proposed another modification to correct the effects of boundaries between consecutive blocks.

MBB has better higher-order properties which make it superior to the one which uses non-overlapping blocks. Both methods get higher order accuracy similar to the stationary bootstrap. Even though the block bootstrap does not achieve the accuracy of the bootstrap for i.i.d. data, it works better than the subsampling. Moreover, MBB performs well under weak conditions on the dependency structure. Mammen and Nandi (2012) stated that there is no specific assumption about the data generating process to apply block bootstrap.

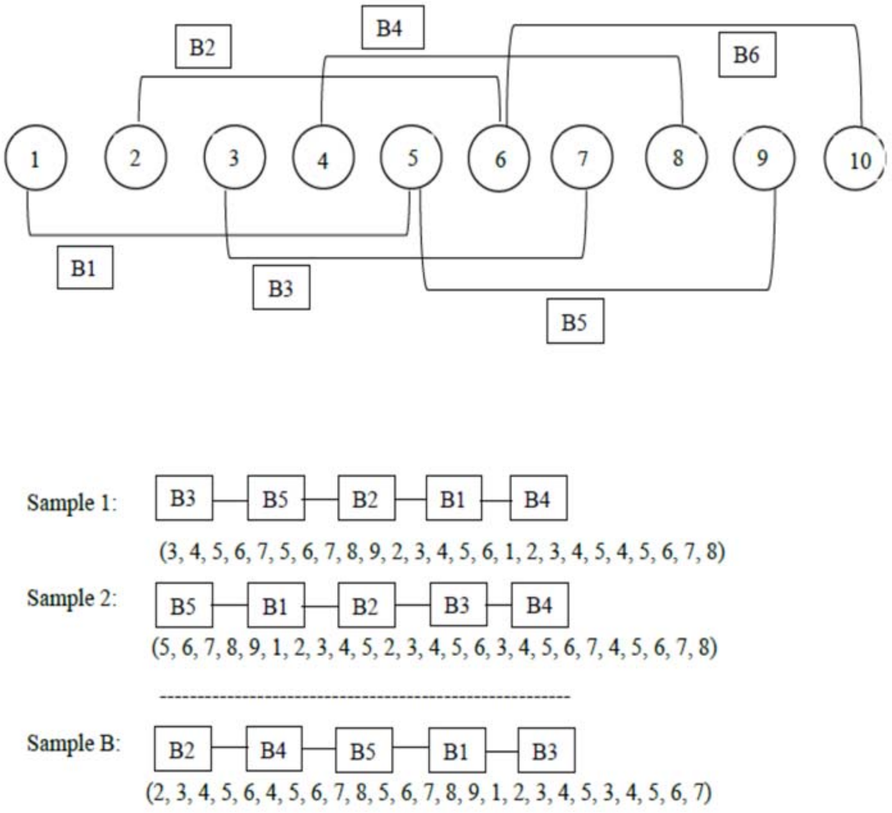


Figure 2. Moving Block Bootstrap Scheme

It is important to determine the block length in order to apply bootstrap. A common approach is to select the block lengths which minimizes the Mean Squared Error (MSE) of the bootstrap estimators. The main nonparametric methods of estimating block lengths are proposed by Hall et al. (1995) and

Lahiri et al. (2007). Hall et al. (1995) proposed to use $n^{1/3}$, $n^{1/4}$ and $n^{1/5}$, where n is the sample size for the estimation of variance or bias. In their work, an empirical block length selection is proposed with the formula $b \approx Dn^{1/k}$, where $k = 3$, 4 or 5 and the constant D is determined by the underlying process. Then, Lahiri et al. (2007) proposed another method based on the jackknife-after-bootstrap method to estimate variance and Nordman and Lahiri (2014) compared the convergence rates of the block length procedures in variance estimation and conclude that the second method has better convergence properties.

3.6. Proposed Approach

In this thesis, the use of moving block bootstrap method described in Section 3.5.1 is proposed to detect the small mean shifts close to the boundaries of the data for autoregressive models. Since the moving block bootstrap keeps the autocorrelation and has no assumption, it is applicable for time series and it can be used for autoregressive models. The algorithm to detect the changepoint and its location is explained below.

Algorithm

Step 0. Apply stationarity or unit root test to the series. If it is not stationary, convert it to a stationary series by detrending or differencing. Then, decide on order of the time series model.

Step 1. LRT based on exact likelihood for an AR model is conducted to the original stationary series and the test statistic is compared with the critical values obtained by simulation. If the test detects a changepoint, classify it as a changepoint.

Step 2. If the test cannot detect any significant changepoint, a bootstrap sample is selected by using moving block bootstrap and original locations of the observations are recorded. The original locations of the observations are the locations in stationary time series. For instance, in Figure 2, in the first bootstrap sample, starts with 3rd block, B*3*. The observations covered in that block are 3, 4, 5, 6 and 7 and their original locations are their locations in the data, that is also 3, 4, 5, 6 and 7.

Step 3. LRT is conducted to the bootstrap sample and if a changepoint is detected, its original location is recorded.

Step 4. The frequency of the locations is calculated and then the possible location of the changepoint with the highest frequencies are determined as the location of the breakpoint.

CHAPTER 4

APPLICATIONS

In this chapter, the applications and obtained results are presented. First, the break types that are covered are explained and then a combination of moving block bootstrap with SNHT is given. In the next section, the comparison of the changepoint detection methods based on a simulation study is given. Then, the application of LRT and its comparison with the best performing methods, SNHT and F-test, is given. In the next subsection, the critical values of the LRT are presented and then the moving block bootstrap is applied to LRT is applied.

4.1. The Artificial Breaks

In this study, two types of inhomogeneity is considered. These are mean shift and sudden decrease or increase which are explained as below.

1. Mean Shift

Mean shift may represent the abrupt discontinuity. Since there are 60 yearly aggregates, the 0.5, 1 and 2 °C shifts are applied to the series starting from 5th year (starting from the beginning), 27th year (starting from the middle) and 53th year (starting from the end).

2. Sudden Decrease/Increase

Sudden decreases or increases may represent the change or any breakdown in the instrument. In the application, 1°C is decreased from the 5th year (in the beginning), 27th year (in the middle) and 53th year (in the end).

The LRT is designed for mean shift. However, the first purpose is to improve the performance of SNHT. In that part of the study (4.2), the performance of the proposed method is compared under mean shift and sudden decrease cases.

4.2. A Modification of the SNHT based on Bootstrap

The comparison of the methods indicate that SNHT is the best test in terms of detecting the breakpoints. Even though Yozgatligil and Yazici (2016) show that SNHT is superior to other tests in terms of detecting inhomogeneity, this test has some drawbacks. The test requires reference stations, whose reliability and homogeneity should be validated before the test is conducted. Moreover, similar to other relative tests, it needs close relative series so that the correlation between the test series and reference series must be at least 0.80. If this assumption is not satisfied, then the test series is classified as non-testable (Gokturk et al., 2008).

In order to overcome this drawback and improve the test, an application of one of the dependent bootstrap methods is proposed if there is a reference series with correlation smaller than 0.80. Since the data is time series, one of the dependent bootstrap methods called moving block bootstrap (MBB) is applied to construct an empirical distribution of the test statistics and decide whether the test statistics obtained from the original series is insignificant or not by constructing percentile intervals. Since the assumption of SNHT which needs the high correlation between test and reference series is not validated it is not applied here, but another successful test, called F-test is applied to the series to compare the performances of the both methods.

4.2.1. Data Generation

First, a time series model is fitted to temperature variable of one of the series. The best model obtained for the temperature variable is the seasonal exponential smoothing method

$$Y_t = \mu_t + S_p(t) + \varepsilon_t$$

where μ_t is a level parameter indicates the mean of the series, the terms $S_p(t)$ indicates the seasonal parameters and ε_t is an error term $t = 1,2,\dots,720; p = \textit{seasonality period} = 12$. The ε_t term has 0 mean and constant variance. The coefficients estimated from the data are as follows;

$$S_{12}(1) = -7.4, S_{12}(2) = -7.5, S_{12}(3) = -6.1, S_{12}(4) = -3.0, S_{12}(5) = 1.4, S_{12}(6) = 6.2,$$

$$S_{12}(7) = 9.0, S_{12}(8) = 9.4, S_{12}(9) = 5.7, S_{12}(10) = 1.8,$$

$S_{12}(11) = -2.6$ and $S_{12}(12) = -5.5.$

These coefficients are the same for each month of each year, but adding an error term ε_t, produces different values each of the 732 values.

However, to conduct relative tests, reference series having high correlation with the test series are needed. To simulate the two reference series, similar coefficients are used and yearly aggregates are obtained. However, the correlations of the yearly aggregates are not satisfactory for the tests, so a random $\varepsilon_t \left(\varepsilon_t \sim N_3 \left(\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.6 & 0.8 \\ 0.6 & 1 & 0.9 \\ 0.8 & 0.9 & 1 \end{bmatrix} \right) \right)$ vector is added to all variables. When paired t-test is conducted to the simulated series and the series in the dataset that is used in NINLIL project and it is concluded that there is not statistically significant difference between the two series. Thus, a simulation setup based on real data is obtained.

4.2.2. Application

The test and reference series are simulated as explained in Section 4.2.1 and yearly aggregates are obtained. Kernel density estimation is tried to be fitted to the test series to decide on the density estimates. The lowest and highest 3 density values are taken as possible breakpoints. After excluding these observations MBB is applied on the three series. The aim of removing these values is to apply bootstrap, since bootstrap is not an appropriate method if there are extreme values in the data. Thus, the weights given to values are not affected by the possible breakpoints. However, these values are used in the test process.

Gaussian kernel and the bandwith selection of $0.786 \times IQR \times n^{-1/5}$ are used for density estimation, where IQR is the interquartile range of the series. The observations associated with the lowest and highest 3 density values are removed from the data to exclude the possible breakpoints. The length selection method to decide on the block length of MBB of Politis and White (2004) gives the block length of 1 (one) for the yearly aggregates and MBB is applied 250 times. SNHT statistic $T^s_{\max}$ is obtained for each bootstrap sample and the 5% percentile interval (PI) is calculated from the empirical distribution. If the test statistic $T^s_{\max}$ of the original data falls in the PI, then the test series is classified as homogeneous, otherwise it is classified as inhomogeneous. SNHT is not applied to the series, since its assumption is not validated. Moreover, the F-test is also applied on the series without conducting bootstrap to compare the performances. The same procedure is repeated after creating the breaks in the data. 1 ^{o}C and 2 ^{o}C mean shifts are applied. The whole analysis is repeated 250 times due to computational inefficiency of the bootstrap method.

4.2.3. Results

The simulation results are given in Table A1 and Table A2 for the mean shift and sudden decrease cases, respectively. SNHT-BS represents the SNHT test applied with bootstrapping and F-test represents the F-test applied with reference series. Y_t column represents the frequencies of inhomogeneity detection after simulating the original series i.e. when there is not any artificial break. That column also represents the Type-I error probability. $Y_{t,shift}$ column represents the frequencies of inhomogeneity detection after creating the artificial change. For instance, the first row of the table represents the output when the shift starts from the beginning of the series. For the increase of 1 ^{o}C, Type-I error probability is 0.284 for the SNHT-BS and 0.044 for the F-test which should be close to 5%. When there is a 1 ^{o}C increase which starts from the beginning of the series, SNHT-BS captures 38.8% of the breakpoints while F-test captures 5.6% of the series.

The frequencies obtained from the original series are better for F-test since they are close to Type-I error probability of 5%. However, the detection rates for the SNHT-BS are higher than the F-test. The detection frequencies of the SNHT-BS are higher if the break is located in the middle of the series in both scenarios. Moreover, the detection frequencies increase as the magnitude of the shift increases from 1 ^{o}C to 2 ^{o}C. In addition to this, both tests are better to capture the mean shift than the sudden decrease.

4.2.4. Conclusions

The homogeneity analysis is the quality control part of the meteorological studies which should be conducted carefully. The non-climatic effects should be detected and removed if possible to obtain reliable inferences. The most widely used relative homogeneity test, SNHT, needs highly correlated reference series to conduct the test. In this part of the study a computational statistics method; bootstrap for dependent data is applied, if there are reference series with correlation less than 0.80. Otherwise, SNHT cannot be applied since its assumption is not validated. Thus, the non-classified SNHT test stations are tested. The results are compared with another relative test F-test which does not make any assumptions on the reference series and applied on two inhomogeneity scenarios of mean shift and sudden change. The performance of the SNHT-BS method works well especially if the break is in the middle of the series; while it needs to be improved to capture the

Type-I probability better in both inhomogeneity scenarios. However, it is still a relative test and needs relative series. Since the Type-I error is not correctly captured, SNHT-BS is not a reliable test. The ratio that a homogenous series to be classified as inhomogeneous series is high.

The relative tests need homogeneous and highly correlated reference series, while it is not possible to obtain such series in some datasets. Even though an attempt is done to improve the application of the best performing test, SNHT, there are still problems such as high Type-I error probability. Then, the study is continued to obtain an absolute test which captures the breakpoints and application of likelihood ratio test is studied.

4.3. Likelihood Ratio Test for AR(1) for a Single Changepoint

The previous part of the study indicates the need for an absolute test in order to capture breakpoints. Here, the purpose is to propose a likelihood ratio test for time series models to detect the changepoint by resampling methods by using an absolute test. Then, the comparison of the performances of the proposed methods for mean shift is conducted and the methods are applied to real life datasets such as economics, meteorology, energy.

A likelihood ratio test for AR(1) model is conducted to detect a single changepoint. AR(1) series of length 732 which has the same size with the data is simulated and then a single changepoint is created artificially by changing the parameter, δ and three changepoint scenarios such as at the beginning, in the middle and at the end of the series and their outputs are investigated.

In this part of the study, the distribution of the test statistic is tried to be derived by using its asymptotic distribution. Since $\Lambda(x) \sim \chi_1^2$ asymptotically (Bain and Engelhardt, 1992), the distribution of the maximum order statistics is needed. Let Y_n represent the maximum order statistics and G_Y represent the cumulative distribution function of $Y_{n.}$. That is, if $X \sim \chi_k^2$, then $F(x) = P\left(\frac{k}{2}, \frac{x}{2}\right), x > 0$ where $P(a,z)$ is a regularized gamma function and $G_Y(y) = [F_X(y)]^n$ (Abramowitz, 1972). In order to obtain the 95th percentile, the solution of $G_Y(m) = [F_X(m)]^n = (0.95)$ with respect to *m* is needed. The solution

$$G_Y(m) = \left[P\left(0.5, \frac{m}{2}\right) \right]^n = (0.95)$$

gives the value of 16.99 for n = 732. Since the test statistics is the maximum of the $\Lambda(x)$ should be compared with 16.99.

The results of the likelihood ratio test are compared with that of SNHT applied to the same series. The test (*X*) and two reference series (*Y*, *Z*) are simulated with arbitrarily chosen parameter values so that the correlation of X and Y, and X and Z are at least 0.80 both for the monthly series and yearly aggregates.

$$X_t = 0.4 + 0.5X_{t-1} + \varepsilon_t,$$

$$Y_t = 0.25 + 0.4X_{t-1} + 0.55Y_{t-1} + \varepsilon_t$$

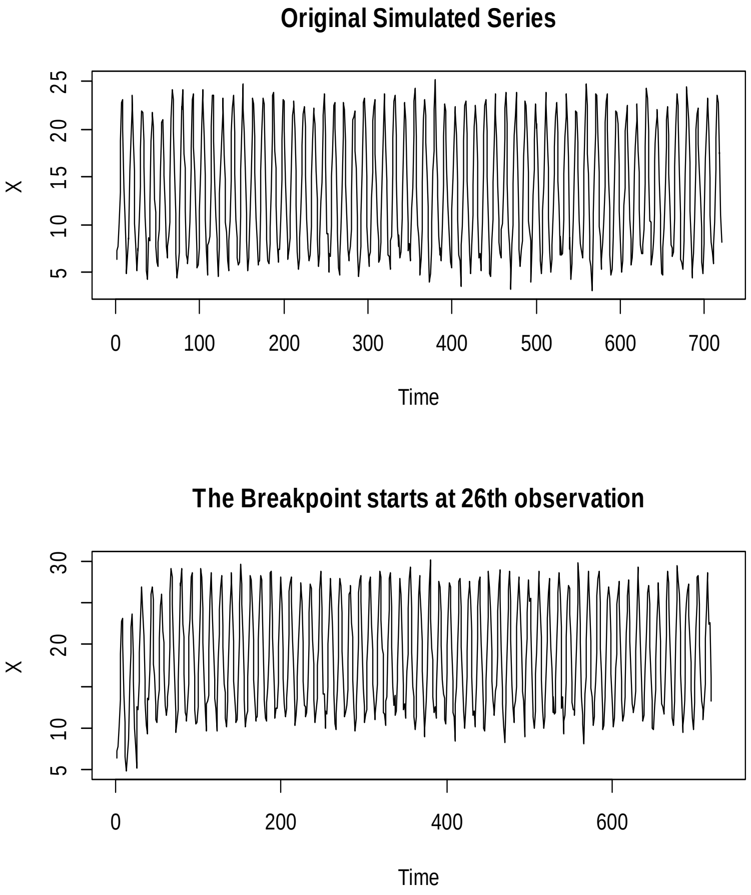
$$Z_t = 0.31 - 0.3X_{t-1} + 0.41Y_{t-1} + 0.45Z_{t-1} + \varepsilon_t$$

where the error terms are generated from $\varepsilon_t \sim N_3(\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & .9 & .9 \\ .9 & 1 & .9 \\ .9 & .9 & 1 \end{bmatrix})$.

The simulation is conducted to obtain 732 monthly (61 years) series.

The model is simulated 1000 times and each time the detected breakpoints are calculated. Note that, the SNHT-BS is not used in the comparisons due to its high Type-I error probability and also the correlations between the test and reference series are high enough (greater than 0.80) to conduct SNHT. The SNHT is applied to monthly and annual data while both tests (LR and SNHT) are applied before creating the artificial change and after the change.

“In the beginning” case, the observations starting from 26 (2th year) to the end of the series (732), “In the middle” case, the observations starting from 355 (29th year) to the end of the series (732) and “at the end” case, the observations starting from 700 (58th year) to the end of the series (732) are increased by 1, 2, 3, 5 and 10, respectively. The original series and the series that have breaks are illustrated in Figure 3 and Figure 4.



Original Simulated Series

The Breakpoint starts at 26th observation

Figure 3. Original Series and a Single Change that Starts at the Beginning (26th observation)

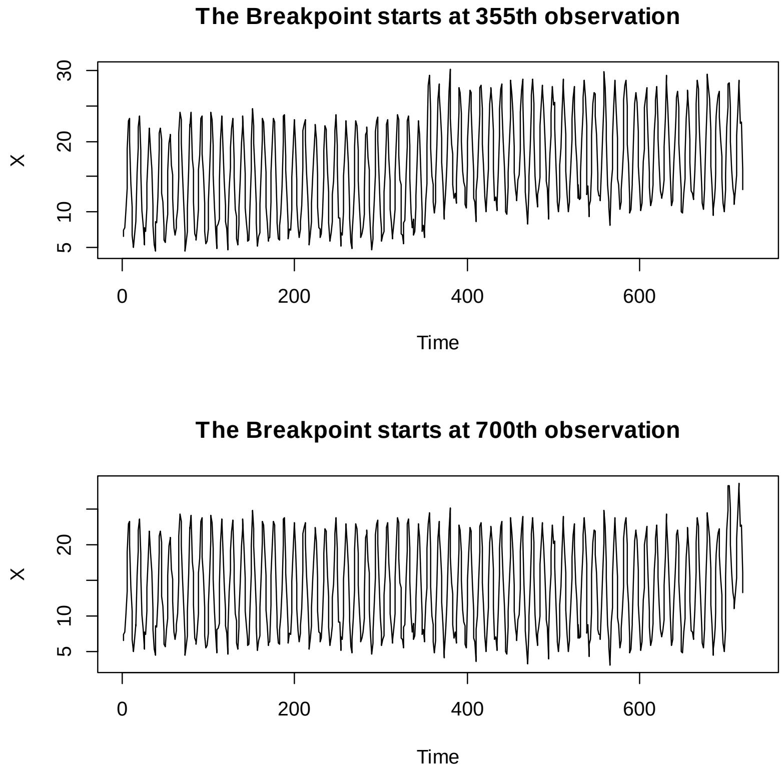


Figure 4. Single Change that Starts in the Middle (355th observation) and at the end of the series (700th observation)

Tables A3-A7 present the breakpoint detection frequencies and location detections. In these tables, the "Methods" column represents the applied methods, namely LRT, $\mathrm{SNHT}_{\mathrm{monthly}}$ when the test is applied to monthly series and $\mathrm{SNHT}_{\mathrm{annual}}$ when the test is applied to yearly aggregates. Similar to previous tables, Y_t column represents the detection frequencies before creating the change. It also indicates the Type-I error probability. $Y_{t,\mathrm{shift}}$ column represents the detection frequencies after the change. "Location detection" column presents the detected frequencies of the true locations of breakpoints among the detected shifts. All analyses are conducted under $\alpha = 0.05$. For instance, in the Table A3, when there is no breakpoint in the beginning case, LRT captures 1%, $\mathrm{SNHT}_{\mathrm{monthly}}$ captures 92%, $\mathrm{SNHT}_{\mathrm{annual}}$ captures 9% of inhomogeneous cases. However, when 1 °C of mean shift is applied starting from the beginning case, LRT detects 12%, $\mathrm{SNHT}_{\mathrm{monthly}}$

detects 11% and SNHT$_{\text{annual}}$ detects 2% of the true shifts. On the other hand, LRT captures the true breakpoint (26th month) 1.1% of the detected series.

It can be easily observed that the detection frequencies are higher in SNHT$_{\text{monthly}}$, close to zero in LR and close to 0.1 in SNHT$_{\text{annual}}$ when there is no change. Even though the detection frequencies and location detection in SNHT$_{\text{monthly}}$ are higher, it has worse performance in terms of Type-I error. The detection frequencies are lower, if the changepoint is started in the beginning of the series and they increase as the unit of mean shift increase. SNHT$_{\text{annual}}$ produces close results with LR, however the location detected is a year not a month, while the LR can be applied on monthly series. Moreover, LR can be applied to a series without any reference series and that makes it to be a powerful absolute test.

Thus, the LRT can be proposed to be used as an absolute test to detect changepoints. However, its Type-I error probability should be improved.

4.4. Empirical Null Distributions

The simulations done in the previous part of the study indicate that LRT can be used as an absolute test to detect mean shifts if it is improved to capture Type-I better. In order to achieve this, an attempt is done to derive the critical values of the test statistic.

In the previous part of the study, the distribution of the LRT statistic is derived by using the property that the asymptotic distribution of the test statistic is chi-square. Karioti and Caroni (2002) states that the asymptotic distribution of the test statistics is distributed as chi-square, thus they should be obtained by a simulation study. In this section, the derivation of the critical values by simulation is explained.

The exact distribution of the LRT statistic for mean shift is not known, a simulation study is conducted to obtain the empirical distribution of the test statistic. First, 1,000 different series of AR(1) models with the ϕ = -0.9, …, 0.9 with σ^2 = 1 are simulated for sample sizes, n = 50, 75 and 100.

Then, the original test statistic $\Lambda(x)$ is calculated for each series and the supremum is calculated to obtain the empirical distribution. The corresponding values of 95% percentile is calculated for each case and the critical values are presented in Table 1. For instance, if a LRT is applied for a sample size of 50 the test statistic should be compared with 8.33 if ϕ of the series is obtained as 0.9. The value of the test statistic is obtained as 8.37 for a sample size of 75 and 9.40 for a sample size of 100. If the sample size of the series is greater than 100, the corresponding value for the sample size of 100 is suggested to be used. For other sample sizes, interpolation can be used to decide on the critical value.

Table 1. Critical Values for the LRT

	ϕ					
Sample Size (n)	ϕ = -0.9	ϕ = -0.8	ϕ = -0.7	ϕ = -0.6	ϕ = -0.5	ϕ = -0.4
50	8.33	8.34	8.05	8.69	8.47	9.00
75	8.37	8.69	8.73	8.68	9.05	8.76
> 100	9.40	8.91	8.89	8.97	8.94	8.52
	ϕ					
	ϕ = -0.3	ϕ = -0.2	ϕ = -0.1	ϕ = 0.1	ϕ = 0.2	ϕ = 0.3
50	8.72	9.68	8.78	8.32	8.46	8.01
75	9.05	9.19	9.40	8.49	8.70	8.29
> 100	8.96	9.34	8.98	8.73	9.21	8.90
	ϕ					
	ϕ = 0.4	ϕ = 0.5	ϕ = 0.6	ϕ = 0.7	ϕ = 0.8	ϕ = 0.9
50	7.77	7.83	7.98	8.15	8.10	9.09
75	8.40	8.32	8.32	8.46	8.81	9.10
> 100	8.87	8.59	8.55	8.32	8.70	9.07

Table 1 indicates that the critical values of the LRT differs when the sample size changes. Moreover, the values are also different from the values obtained from the chi-square distribution. For instance, the corresponding value of that distribution, $\chi^2_{0.05}$ is 3.84, while it depends on the ϕ of the fitted AR(1) model and the sample size in the proposed method. Moreover, the values and the method differs from the values obtained from the previous part of the study. The main reason for that is the assumption of the approximate distribution of the test statistic which is not valid when the exact location of the breakpoint is not known.

4.5. Moving Block Based Likelihood Ratio Test (LRT-BS) for AR(1) Model for a Single Change

The critical values of the LRT statistic are obtained by simulation, thus moving block bootstrap application to LRT can be suggested in order to improve the mean shift detection. The detection performance of the LRT-BS method which is the name of the application of MBB to LRT is investigated with a simulation study.

The data used in this part of the study is generated as explained in Section 4.2.1. The moving block bootstrap with different block lengths (*l*) is also applied 1,000 times to detect the changepoints for different sample sizes. In that bootstrap model consecutive observations are selected. After constructing each bootstrap sample, LRT is applied to that sample. The frequencies of the detected locations among the detected inhomogeneities are calculated. The results are given in Table A8 and A9 when the change starts at the beginning and at the end, respectively. In these tables, sample sizes from 50 to 90 are studied. "Location of the breakpoint" column presents the exact location of the true breakpoint. Different block sizes for each sample size is tried. For instance, block length of 15, 20 and 30 are tried when the sample size is 50. Y_t column represents the results obtained when there is no changepoint in the series and $Y_{t,\text{shift}}$ column represents the output obtained after creating the artificial changepoint. $\text{SNHT}_{\text{monthly}}$, $\text{SNHT}_{\text{annual}}$ and LRT columns represent the results obtained from the application of these tests, while the LRT-BS column represents the results obtained from the application of MBB to LRT.

The highest frequency is presented in the last column. For instance, the first row of the Table A8 presents the output of a sample size 50. In that case, the breakpoint is at the 10th observation. In the application of bootstrap, block size of 15 is used. None of the tests detected a change before the artificial break (10th observation) and only $\text{SNHT}_{\text{monthly}}$ detected changepoint at the 25th observation which is not the true location. However, the LRT-BS detected a change 41% of the times. Among these detected ones, the frequencies of the detected locations are calculated. In that case, 25% classifies the 10th observation as the breakpoint which is the highest frequency.

The output in the Table A8 and Table A9 show that the Type-I error rates are low (almost 0) even for the small samples. While the tests' performance decrease for small samples, LRT-BS is capable of detecting the true location of the breakpoint. However, $\text{SNHT}_{\text{monthly}}$ usually detects a changepoint close to the exact location among the detected changepoints. On the other hand, it has large Type-I error probability. In the "Beginning" case, the LRT-BS performs better when n = 50 and n = 90. However, the best performance of the detection of true location is obtained when the block size is 20 or 30. In the "End" case, the LRT-BS test has the highest percent of capturing the inhomogeneity when n = 50. If the size of the blocks is 30 or 40, the capability of detecting the true location is generally higher.

4.6. Comparison of LRT-BS with SNHT and F-test

In the comparison study of all Yozgatligil and Yazici (2016), the best performing tests are SNHT as the best relative test, F-test as the best absolute test and RHTest as the best one to capture multiple changepoints. Thus, another simulation study is performed to compare performance of LRT-BS with these tests.

First, 100 series of AR(1) model with the parameters ϕ = -0.9, -0.6, -0.3, 0.3, 0.6, 0.9 and σ = 1 are simulated for sample sizes of n = 75 and n = 100. Since SNHT is a relative test, two reference series are obtained. First, error terms are simulated from multivariate normal distribution $\left(\varepsilon_t \sim N_3\left(\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{bmatrix}\right)\right)$ and then the first reference series is obtained by adding the first column of the error terms to the test series and the second reference series is obtained by adding the second column of the error terms to the test series. Thus, highly correlated test and reference series are obtained and SNHT is performed on these series.

Before creating artificial change, the four tests are conducted on the original series. Single and multiple changes are applied to series. When single level shift is considered, 5-unit increase is applied after 20th observation for n = 100 and the same amount of increase is applied after 10th observation for n = 75. In the case of multiple change, two level shifts are considered. The first shift is applied as explained at the 10th and 20th observations for n = 75 and n = 100 respectively. Then, a second level shift of 5-unit is applied at the 70th observation and 80th observation for n = 75 and n = 100, respectively. Figures 5 and 6 exemplify the level shifts for sample size of 75 and 100, respectively for the model with ϕ = -0.3.

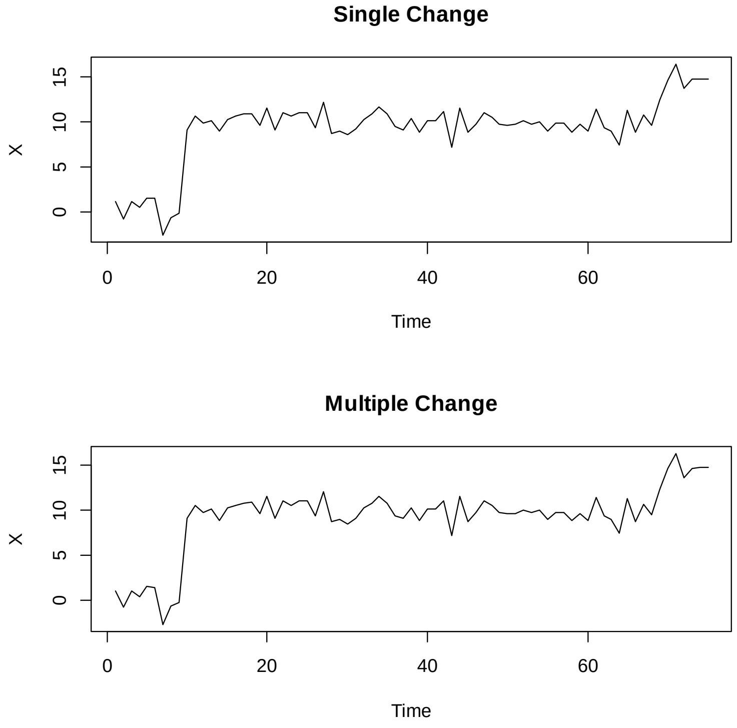


Figure 5. Single and Multiple Changes for n = 75 and ϕ = -0.3

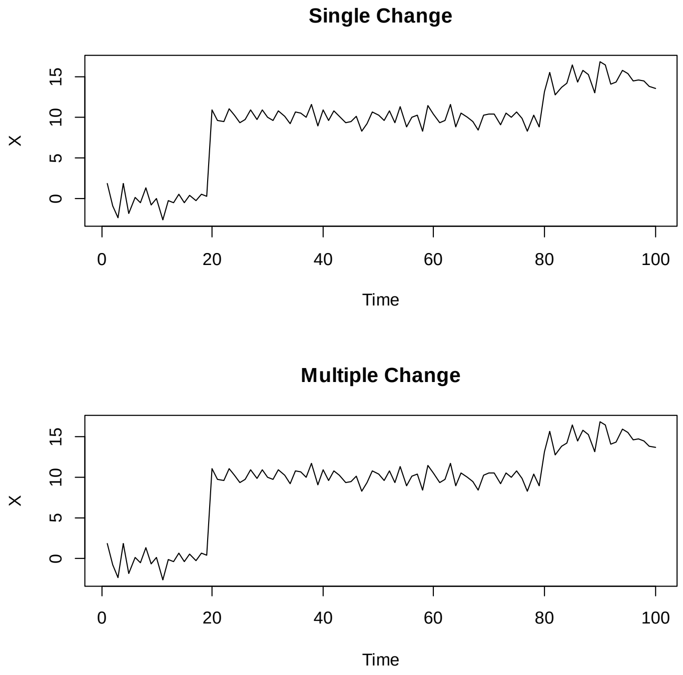


Figure 6. Single and Multiple Changes for n = 100 and ϕ = -0.3

The detection rates for single change are presented in Table 2 and Table 3 for n = 75 and n = 100, respectively. The detection rates in the original samples and the rates after creating shifts are presented in the Y_t and $Y_{t,shift}$ columns, respectively. Y_t column also presents the Type-I error and the "Breakpoint detection" column represents the frequency of the detection of true locations among the detected series.

When Table 2 is investigated for the performance of the tests when there is a single change for the sample size of 75, it can be seen that Type-I error probabilities of the SNHT and LRT are close to 5%, while it increases as the ϕ increases for the F-test and it is always greater than 5% for the RHTest. Moreover, the F-test and LRT captures the breakpoint 99% of the series. SNHT is the worst test in terms of detecting the location of the true breakpoint. This test only captures the breakpoint when ϕ = 0.6. On the other hand, the true breakpoint detection frequencies are higher for the F-test and LRT. For instance, when the series is generated from the ϕ = -0.6, F-test detects 11th observation 91% of the detected series and 12th observation 8% of the detected series. However, the detection frequency and the detection of the breakpoint is the smallest when $\phi = 0.9$ for the LRT when only this tests' output is investigated.

Table 2. Detection Rates for Single Changes for 100 simulated series for $n = 75$

SINGLE CHANGE									
ϕ	**Method**	Y_t	Y_t, s	**Breakpoint Detection (%)**	ϕ	**Method**	Y_t	Y_t, s	**Breakpoint Detection (%)**
-0.9	SNHT	0.02	0.04	-	**0.3**	SNHT	0.06	0.05	-
	F-test	0.00	0.99	11→0.01, 12→0.99		F-test	0.26	0.99	11→0.98
	RHTest	0.23	0.26	11→0.08		RHTest	0.32	0.36	11→0.08
	LRT	0.05	0.99	8→0.59, 9→0.40		LRT	0.03	0.99	8→0.90
-0.6	SNHT	0.02	0.02	-	**0.6**	SNHT	0.02	0.06	9→0.16
	F-test	0.00	0.99	11→0.91, 12→0.08		F-test	0.49	0.99	11→0.92
	RHTest	0.35	0.31	11→ 0.13		RHTest	0.26	0.32	11→0.12
	LRT	0.10	0.99	8→0.81, 9→0.18		LRT	0.07	0.99	8→0.87
-0.3	SNHT	0.04	0.04	-	**0.9**	SNHT	0.04	0.03	-
	F-test	0.00	0.99	11→0.95		F-test	0.82	0.99	11→0.70
	RHTest	0.25	0.38	11→0.15		RHTest	0.26	0.25	11→0.08
	LRT	0.06	0.99	8→0.88, 9→0.10		LRT	0.05	0.74	8→0.62

The detection performance of the methods when a single shift is applied to a sample size of 100 is investigated and the related output is presented in Table 3. Type-I error probabilities present similar results with the sample size of 75 except for LRT in the case of $\phi = 0.6$. The F-test and LRT captures the breakpoint almost all of the series except for $\phi = 0.9$. In addition to this, the frequency of breakpoint detection is also higher for these two tests.

Table 3. Detection Rates for Single Changes for 100 simulated series for $n = 100$

ϕ	Method	Y_t	Y_t, s	Breakpoint Detection (%)	ϕ	Method	Y_t	Y_t, s	Breakpoint Detection (%)
-0.9	SNHT	0.03	0.04	-	**0.3**	SNHT	0.07	0.07	-
	F-test	0.00	1.00	19→0.77		F-test	0.23	1.00	19→0.97
	RHTest	0.24	0.30	11→0.06		RHTest	0.22	0.22	11→0.18
	LRT	0.06	1.00	19→0.54, 20→0.46		LRT	0.08	1.00	19→0.95
-0.6	SNHT	0.07	0.01	19→0.97	**0.6**	SNHT	0.06	0.06	-
	F-test	0.00	1.00	19→0.97		F-test	0.66	1.00	19→0.98
	RHTest	0.24	0.26	11→0.15		RHTest	0.25	0.25	11→0.16
	LRT	0.07	1.00	19→0.85		LRT	0.13	1.00	18→0.98
-0.3	SNHT	0.05	0.04	-	**0.9**	SNHT	0.04	0.02	-
	F-test	0.01	1.00	19→0.98		F-test	1.00	0.99	19→0.67
	RHTest	0.20	0.25	11→0.08		RHTest	0.25	0.28	11→0.07
	LRT	0.07	1.00	19→0.94		LRT	0.02	0.73	17→0.20, 18→0.42

The output obtained when the multiple change applied to the sample size of 75 is presented in Table 4. According to that table, SNHT and LRT produce similar Type-I error probabilities, while F-tests' Type-I error probability

Table 4. Detection Rates for Multiple Changes for 100 simulated series for $n = 75$

MULTIPLE									
ϕ	**Method**	Y_t	Y_t, s	**Breakpoint Detection (%)**	ϕ	**Method**	Y_t	Y_t, s	**Breakpoint Detection (%)**
-0.9	SNHT	0.07	0.04	69→0.25	**0.3**	SNHT	0.05	1.00	-
	F-test	0.00	0.99	11→0.77, 12→0.22		F-test	0.33	0.99	11→0.85
	LRT	0.06	0.99	8-→0.76, 9→0.22		LRT	0.08	0.99	8→0.64, 68→ 0.23
-0.6	SNHT	0.02	0.07	74→0.14	**0.6**	SNHT	0.07	0.03	-
	F-test	0.00	0.99	11→0.90		F-test	0.54	0.99	11→0.72
	LRT	0.06	0.99	7→0.10, 8→0.81		LRT	0.02	0.97	8→0.44, 68→0.31
-0.3	SNHT	0.04	0.06	-	**0.9**	SNHT	0.05	0.04	-
	F-test	0.00	0.99	11→0.92		F-test	0.80	0.99	11→0.48, 64→0.12
	LRT	0.05	0.99	11→0.86		LRT	0.06	0.77	10→0.37, 70→0.18

becomes worse when the ϕ value increases. The F-test and LRT captures the breakpoint almost all of the cases. However, the detection frequency of LRT becomes worse when $\phi = 0.9$. Even though F-test and the LRT detect the true breakpoint in most of the cases, the detection performance of the LRT again worsens for the $\phi = 0.9$.

When the multiple change is applied to sample size of 100, its output is presented in Table 5. SNHT is still the best performing test in terms of Type-I error probabilities. Even though the LRT performs better when compared with F-test in terms of Type-I error probability, its performance also becomes worse when ϕ is close to 0.9. Similarly, the F-test has zero Type-I probability in most of the cases, while it increases for large values of ϕ. When the detection of changepoints are investigated, it can be easily stated that the F-test and LRT again have the best performances. However, when ϕ = 0.9, the detection of LRT decreases. These two tests again have the best performance in terms of capturing the exact location of the breakpoint.

Table 5. Detection Rates for Multiple Changes for 100 simulated series for $n = 100$

MULTIPLE									
ϕ	**Method**	Y_t	Y_t, s	**Breakpoint Detection (%)**	ϕ	**Method**	Y_t	Y_t, s	**Breakpoint Detection (%)**
-0.9	SNHT	0.08	0.03	20→0.33	**0.3**	SNHT	0.06	0.02	-
	F-test	0.00	1.00	19→0.23, 79→0.51		F-test	0.00	1.00	19→0.29, 79→0.63
	LRT	0.02	1.00	19→0.18, 79→0.55		LRT	0.06	1.00	19→0.39, 79→0.45
-0.6	SNHT	0.05	0.04	17→0.24	**0.6**	SNHT	0.08	0.07	78→0.14
	F-test	0.00	1.00	19→0.29, 79→0.63		F-test	0.58	1.00	19→0.33, 79→0.56
	LRT	0.06	1.00	19→0.39, 79→0.45		LRT	0.14	0.99	19→0.36, 79→0.30
-0.3	SNHT	0.03	0.04	20→0.25	**0.9**	SNHT	0.06	0.03	79→0.33
	F-test	0.00	1.00	19→0.24, 79→0.70		F-test	0.95	1.00	19→0.35, 79→0.26
	LRT	0.02	1.00	19→0.38, 79→0.36		LRT	0.16	0.78	19→0.32, 79→0.20

When there is no changepoint in the series, the Type-I error rates are close to 0.05 for the SNHT and LRT especially for the sample size of 75. However, when n = 100 is considered, SNHT still captures the 0.05 while LRT performs worse when $\phi = 0.6, 0.9$. On the other side, the Type-I

probabilities of F-test and RHTest are not close to 0.05. Even though, in the simulation study of Yozgatligil and Yazici (2016), F-test performs well in terms of Type-I, that is probably because of the small sample size. It is also the same for the multiple changes which are presented in Table 4 and Table 5, respectively for n = 75 and n = 100.

When the detection rates are investigated, F-test and LRT are the best performing tests. Since the Type-I rates are higher in F-test, LRT is preferable in all cases. The reason why the SNHT does not perform well can be the small sample sizes.

After creating single or multiple changes, the stationarity of the model usually is not valid. Thus, ADF test is applied on each series and if the test concludes that the series is nonstationary, the series are made stationary by differencing. Table 6 and Table 7 presents the orders of the models for single and multiple shifts respectively. When the series are simulated with positive ϕ values after creating shifts and differencing, almost all models become White Noise. However, for the cases with negative ϕ values, the series still follow the AR(1) type. Moreover, the MA1 model frequencies are also high especially when ϕ is 0.3 or -0.3. The estimation is done with AR(1) model. Note that, even if the wrong model is used for estimation, the performance of the test is not changing.

LRT-BS is applied to each series 100 times with different block length, l. Since the purpose is to detect a changepoint, the best length is determined by simulation. For this purpose, the length of 0.1×n, 0.2×n and 0.3×n are considered where n is the sample size of the series.

Table 6. The frequency of models after creating a single shift

Sample Size	ϕ	SINGLE												
75	0.9	WN					**100**	0.9	AR1	WN				
		1.00							0.10	0.90				
	0.6	AR1	MA1	WN				0.6	AR1	ARMA11	MA1	MA2	WN	
		0.01	0.10	0.89					0.01	0.05	0.12	0.01	0.81	
	0.3	AR1	MA1	MA2	WN			0.3	AR1	ARMA11	MA1	MA2	WN	
		0.06	0.30	0.01	0.63				0.03	0.08	0.59	0.08	0.22	
	-0.3	AR1	AR2	ARMA11	MA1	MA2		-0.3	AR1	AR2	ARMA11	ARMA23	MA1	MA2
		0.30	0.12	0.34	0.22	0.02			0.05	0.05	0.31	0.02	0.57	0.03
	-0.6	AR1	AR2	ARMA11	ARMA21	MA2		-0.6	AR1	AR2	ARMA11	ARMA12	MA2	
		0.37	0.46	0.15	0.01	0.01			0.08	0.40	0.49	0.01	0.02	
	-0.9	AR1	AR2	ARMA11				-0.9	AR1	AR2	AR3	ARMA11	ARMA12	
		0.43	0.49	0.08					0.07	0.47	0.02	0.43	0.01	

Table 7. The frequency of models after creating multiple shifts

Sample Size	ϕ	MULTIPLE												
75	0.9	MA1	WN					100	0.9	AR1	WN			
		0.02	0.98							0.02	0.98			
	0.6	MA1	WN						0.6	AR1	ARMA11	ARMA21	MA1	WN
		0.04	0.96							0.01	0.01	0.01	0.10	0.87
	0.3	AR1	MA1	WN					0.3	AR1	ARMA11	MA1	MA2	WN
		0.06	0.20	0.74						0.05	0.03	0.63	0.01	0.28
	-0.3	AR1	AR2	ARMA11	ARMA21	MA1	MA2		-0.3	AR1	AR2	ARMA11	MA1	MA2
		0.58	0.10	0.01	0.01	0.28	0.02			0.26	0.07	0.04	0.56	0.07
	-0.6	AR1	AR2	ARMA11	ARMA21	MA1	MA2		-0.6	AR1	AR2	ARMA11	ARMA12	MA1
		0.88	0.06	0.01	0.02	0.02	0.01			0.55	0.28	0.14	0.01	0.02
	-0.9	AR1	AR2						-0.9	AR1	AR2	ARMA11		
		0.98	0.02							0.68	0.31	0.01		

Table 8. The frequency of true breakpoint detection for single shift

			Single					
Sample Size	**Block Length**	**Location**	**-0.9**	**-0.6**	**-0.3**	**0.3**	**0.6**	**0.9**
***n* = 75**	***l* = 7**	**9**	**0.33**	**0.75**	**0.65**	**0.63**	**0.62**	**0.72**
		10	0.05	0.12	0.15	0.27	0.32	0.50
		11	0.26	0.74	0.46	0.43	0.33	0.35
	***l* = 15**	**9**	0.00	0.01	0.03	0.08	0.08	0.17
		10	0.05	0.14	0.11	0.22	0.31	0.47
		11	**0.65**	**0.65**	**0.60**	**0.45**	**0.39**	**0.40**
	***l* = 22**	**9**	**0.75**	**0.77**	**0.67**	**0.69**	**0.70**	**0.62**
		10	0.06	0.10	0.16	0.27	0.36	0.53
		11	**0.70**	**0.72**	**0.64**	**0.48**	**0.42**	**0.44**
***n* = 100**	***l* = 10**	**19**	**0.40**	**0.71**	**0.77**	**0.74**	**0.75**	**0.67**
		20	0.20	0.08	0.15	0.13	0.29	0.39
		21	0.39	0.61	0.61	0.60	0.38	0.32
	***l* = 20**	**19**	**0.79**	**0.74**	**0.78**	**0.84**	**0.78**	**0.70**
		20	0.08	0.09	0.18	0.24	0.34	0.29
		21	**0.75**	**0.68**	**0.63**	**0.61**	**0.43**	**0.37**
	***l* = 30**	**19**	**0.73**	**0.84**	**0.80**	**0.76**	**0.88**	**0.73**
		20	0.08	0.08	0.16	0.31	0.35	0.44
		21	**0.81**	**0.77**	**0.70**	**0.51**	**0.50**	**0.33**

The highest frequencies of the detected breakpoints are kept. Note that the exact breakpoints are 10 and 20 for *n* = 75 and *n* = 100 respectively in the single change case. Multiple change is applied at 10th and 70th observations for *n* = 75 and 20th and 80th observations are *n* = 100. The detection of the true breakpoints is presented in Table 8 and Table 9 for single and multiple changes respectively. According to Table 8, the highest detection rates belong to 9th and 11th observations for *n* = 75 and 19th and 21st observations for *n* = 100, while the exact location is 10th and 20th observations. This is probably losing one observation by taking a differencing. On the other hand, even though the models becomes White Noise for positive ϕ values, the detection rates are still high. When the detection rates are compared, *l* = 22 and *l* = 30 produce close detection rates for two different sample sizes.

When multiple change detections are compared in Table 9, it can be seen that for n = 75, l = 15 generally produces better detection rates for the locations of 11 and 68. On the other hand, if n = 100 is considered, the block length of l = 30 is not as good as the other block lengths for capturing the second changepoint at the end of the series. The best detection rate is obtained for l = 15 and l = 20 for the sample size of 75 and 100, respectively.

The LRT-BS method is applicable to capture multiple changepoints. Table 10 presents the performance of the LRT-BS method for capturing the true breakpoints at the same time. For instance, when the exact breakpoint is 10 and 70 for the sample size of 75, if l = 15 and the model is simulated from ϕ = -0.9, LRT-BS detects both changepoint 43% of the series. According to the results, the highest detection rate is obtained when the block length, l is 15 and 20 for the sample size of 75 and 100.

4.7. Application of LRT-BS to a Real Data

The LRT-BS method is applied to two different datasets. The first one is the Nile data set that is used in many studies (Figure 7). The data include the annually flow volume Nile River at Aswan between 1871 and 1970. According to Cobb (1978), river flow levels in 1877 and 1913 are possible additive outliers and also there was a mean shift in the flow levels starting from 1899. This is connected partly to the climate changes and partly to the beginning of construction of a new dam at Aswan.

When LRT-BS is applied with the block length, l = 20 which is the $0.2\times n$, the detected years are 7th (1877), 27 (1897), 44 (1913) and 82 (1952), while F-test captures 1898 as the only breakpoint. On the other hand, after simulating highly correlated reference series, SNHT cannot detect any changepoint for that data when it is applied to monthly and yearly aggregates. Thus, LRT-BS is capable of detecting the true changepoint of the series while the other tests cannot detect any changepoint. Moreover, the

Table 9. The frequency of true breakpoint detection for multiple shift

			Multiple					
Sample Size	**Block Length**	**Location**	**-0.9**	**-0.6**	**-0.3**	**0.3**	**0.6**	**0.9**
n **= 75**	***l*** **= 7**	**9**	0.43	0.56	0.56	0.68	0.69	0.72
		10	0.04	0.07	0.07	0.19	0.29	0.39
		11	0.27	0.46	0.46	0.41	0.30	0.29
		67	0.04	0.04	0.04	0.03	0.07	0.11
		68	0.12	0.08	0.08	0.07	0.03	0.14
		69	0.31	0.62	0.62	0.72	0.72	0.59
	l **= 15**	**9**	0.00	0.01	0.01	0.08	0.06	0.15
		10	0.02	0.11	0.11	0.21	0.30	0.43
		11	0.64	0.68	0.68	0.51	0.42	0.26
		67	0.01	0.01	0.01	0.03	0.05	0.06
		68	0.07	0.10	0.10	0.06	0.05	0.12
		69	0.64	0.78	0.78	0.78	0.82	0.62
	l **= 22**	**9**	0.11	0.65	0.80	0.82	0.78	0.67
		10	0.00	0.11	0.09	0.22	0.26	0.48
		11	0.09	0.51	0.78	0.54	0.41	0.30
		67	0.00	0.02	0.05	0.03	0.10	0.06
		68	0.01	0.05	0.09	0.03	0.00	0.04
		69	0.06	0.60	0.19	0.11	0.01	0.09
n **= 100**	***l*** **= 10**	**19**	0.26	0.55	0.70	0.64	0.73	0.65
		20	0.05	0.07	0.16	0.16	0.16	0.30
		21	0.37	0.47	0.39	0.38	0.37	0.29
		79	0.28	0.62	0.67	0.69	0.67	0.64
		80	0.02	0.08	0.15	0.25	0.15	0.22
		81	0.29	0.45	0.40	0.37	0.26	0.28
	l **= 20**	**19**	0.72	0.65	0.72	0.65	0.75	0.68
		20	0.10	0.11	0.15	0.21	0.23	0.31
		21	0.54	0.52	0.54	0.38	0.32	0.31
		79	0.75	0.70	0.72	0.67	0.69	0.73
		80	0.15	0.13	0.18	0.22	0.23	0.24
		81	0.54	0.55	0.49	0.38	0.27	0.28
	l **= 30**	**19**	0.58	0.75	0.74	0.80	0.77	0.94
		20	0.10	0.12	0.18	0.20	0.23	0.93
		21	0.64	0.72	0.67	0.52	0.45	0.68
		79	0.25	0.13	0.04	0.06	0.04	0.01
		80	0.00	0.00	0.00	0.00	0.00	0.00
		81	0.00	0.00	0.00	0.00	0.00	0.00

Table 10. The frequency of true breakpoint detection for multiple shift

n **= 75**							
Location	**Block Length**	**-0.9**	**-0.6**	**-0.3**	**0.3**	**0.6**	**0.9**
[9,67]	***l*** = 7	0.16	0.33	0.37	0.35	0.24	0.21
[9,67]	***l*** = 15	0.43	0.53	0.59	0.42	0.34	0.18
[9,67]	***l*** = 22	0.04	0.16	0.59	0.10	0.05	0.03
n **= 100**							
Location	**Block Length**	**-0.9**	**-0.6**	**-0.3**	**0.3**	**0.6**	**0.9**
[19,79]	***l*** = 10	0.10	0.32	0.44	0.44	0.52	0.46
[19,79]	***l*** = 20	0.54	0.45	0.51	0.38	0.54	0.54
[19,79]	***l*** = 30	0.16	0.08	0.02	0.04	0.03	0.46

year “1899” is an example of inhomogeneity in climate studies since it is an effect of non-climatic effect.

The LRT-BS is also capable of detecting inhomogeneity in meteorological studies. The first real life application of the method shows that the proposed method works better than the other methods.

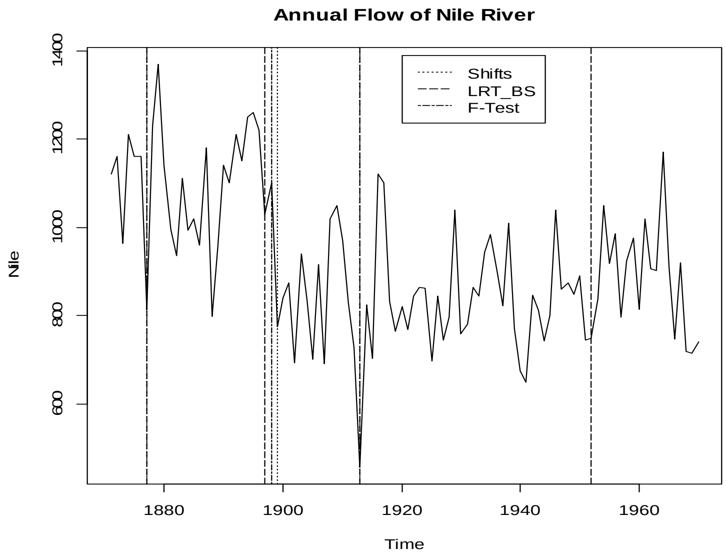


Figure 7. Time Series Plot of Annual River Flow of Nile River

The second data is temperature and precipitation of Fethiye station which belongs to our project data. According to metadata, the station is moved to somewhere else at 1962, but still represents the same area. This data is an example of a sample size greater than 100. Figure 8 and 9 represent the time series plots of precipitation and temperature series of the data, respectively with the year of inhomogeneity.

The three methods, LRT-BS, SNHT and F-test are applied to both series of length 732 monthly data separately for both series. SNHT is applied monthly and yearly aggregates. However, SNHT and F-test cannot detect any changepoint in both of the series, while the proposed method, LRT-BS detects breakpoints.

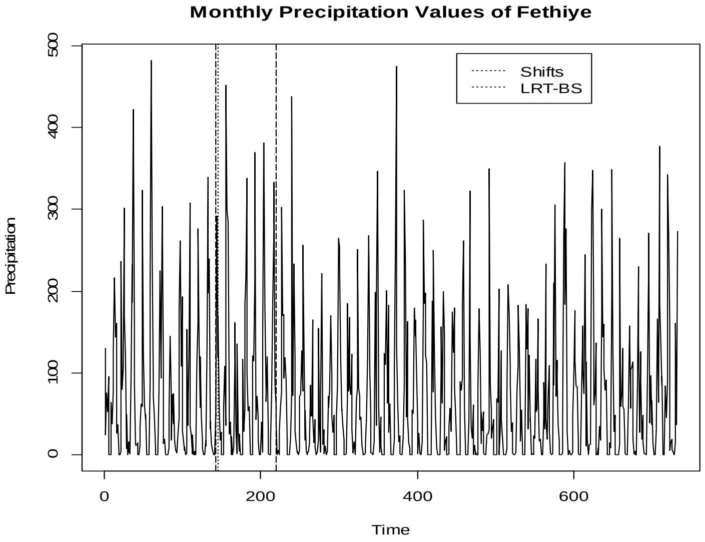


Figure 8. Time Series Plot of Precipitation of Fethiye

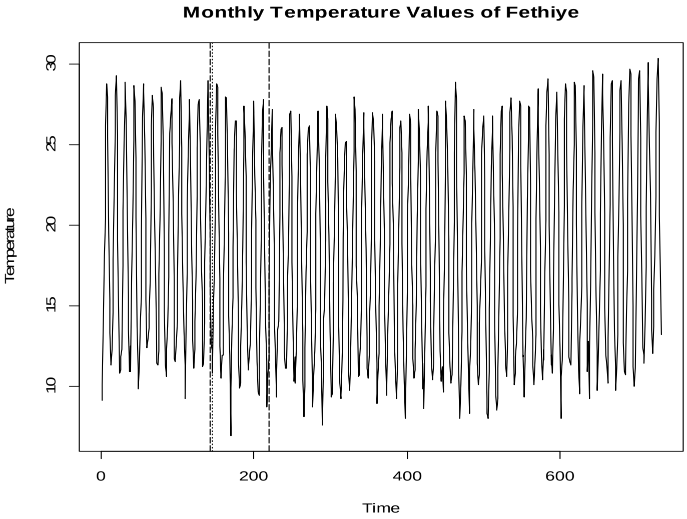


Figure 9. Time Series Plot of Temperature of Fethiye

LRT-BS indicate that there is a breakpoint at October of 1961. Another changepoint is detected as April of 1968. This data is also an example of an inhomogeneity since there is an effect which is not related with climate. Similar to the first real life application, this one also shows that LRT-BS is superior to SNHT and F-test when the changepoint detection performances are compared.

CHAPTER 5

CONCLUSION AND FURTHER RESEARCH

The changepoint is an important issue in time series analysis. The effect of a changepoint can be change in mean, variance, abrupt or sudden changes, gradual increases or multiple changes. The methods in the literature have some drawbacks that may lead to unreliable inferences. For instance, these include i.i.d. or normality assumption of observations whose validation may not be possible for dependent data. In this thesis, a computational approach involving bootstrapping is used to improve existing methods to detect whether data is homogeneous or not.

First, an attempt is done to improve the best performing method, SNHT based on the computation method, moving block bootstrap. However, the application shows that the proposed SNHT-BS method, has high Type-I error probabilities. Moreover, it is still a relative method which needs homogeneous reference series similar to SNHT. Then, the study is continued to propose an absolute test which captures the breakpoints.

Then, the use of likelihood ratio test is considered and the applications show that this method can be used to detect changepoints in the series. However, the moving block bootstrap method is applied on the LRT to capture the breakpoints close to the beginning or end of the series in addition to detect multiple changepoints.

The study covers LRT based on the exact likelihood whose distribution is not known. Thus, a simulation study is conducted to obtain the critical values of the test statistic. Then, moving block bootstrap is applied to LRT capture single or multiple changepoints.

The proposed approach, LRT-BS consists of selecting block length of $0.2 \times n$ to capture the breakpoints especially multiple changepoints starting at the beginning or end of the series. These points are starting values of the mean shifts and when the performance of the proposed method is compared with the other methods such as SNHT and F-test, it is concluded as the best one in terms of detecting the changepoint. Moreover, LRT-BS is also capable of detecting the multiple changes.

The simulations show that F-test and LRT are the best performing tests. Since the Type-I error probabilities are higher in F-test, LRT is preferable in all cases. The performance of SNHT is not as good as the other methods, for instance this test cannot capture the true breakpoint most of the time. This can be due to the small sample size of the series.

Since the results imply better detection rates, the study is applied to two real datasets whose breakpoints are known. These data sets are also examples of the inhomogeneous series in climate studies. The result of the proposed approach show that the method is capable of detecting the changepoint while the SNHT and F-test cannot detect any changepoint especially in the long series. Thus, this real life example also shows that the proposed method is capable of detecting the true single or multiple changepoints.

On the other hand, the frequencies of the model types show that after creating changepoints, the models do not always keep the AR(1) model type. The series simulated by ϕ values produce MA(1) or ARMA(1,1) models. Even though the model type is different from AR(1), the LRT based on that model still detects the true changepoints.

Moreover, in that study, AR(p) models is considered since the approximations can be applied to ARMA models to represent them with AR models. When the series are simulated with positive ϕ values after creating shifts and differencing, almost all models become White Noise. However, for the cases with negative ϕ values, the series protect the AR(1) type. Moreover, the MA(1) model frequencies are also high especially when ϕ is 0.3 or -0.3.

In conclusion, the proposed method, LRT-BS is capable of detecting the single or multiple changes close especially to the beginning or end of time series data when mean shift type of breakpoint is considered. In addition to this, the block length of the proposed method is studied and appropriate length is suggested to capture mean shifts in the series while other tests in the literature such as the F-test and SNHT cannot detect them.

Even though the applications are done based on climate studies, the method can be applicable to any time series including economics or health studies. Moreover, the method can be used in the classification studies.

The method is going to be applied to variance changes and changes in ϕ. The LRT test is then going to be applied to AR(p) models for single and multiple changepoints under different simulation cases. Since the bootstrap is an inefficient method in terms of computation, another study is going to be conducted to decrease this inefficiency. Moreover, the application of the method to spatial data is also considered. On the other hand, the combination of the proposed method with stochastic differential equations and Markov switching techniques can be considered for the other types of shifts.

REFERENCES

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th printing. New York: Dover, pp. 940-943.

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

Alexandersson, H., & Moberg, A. (1997). Homogenization of Swedish temperature data. Part I: Homogeneity test for linear trends. *International Journal of Climatology*, *17*(1), 25-34.

Aminikhanghahi, S., & Cook, D. J. (2017). A survey of methods for time series changepoint detection. *Knowledge and Information Systems*, *51*(2), 339-367.

Azevedo, N., Pinheiro, D., & Weber, G. W. (2014). Dynamic programming for a Markov-switching jump–diffusion. *Journal of Computational and Applied Mathematics*, *267*, 1-19.

Bai, J., & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 47-78.

Baillie, R. T., Bollerslev, T., & Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *74*(1), 3-30.

Bain, L. J., & Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics*. Brooks/Cole.

Banerjee, A., Lazarova, S., & Urga, G. (1998). Bootstrapping sequential tests for multiple structural breaks. *European University Institute*.

Barry D, Hartigan J.A. (1992). Product partition models for changepoint problem. *Ann. Stat*. 20: 260–279.

Barry D, Hartigan JA. (1993). A Bayesian analysis for changepoint problems. *Journal of American Statistical Association.* 88: 309–319.

Battaglia, F., & Orfei, L. (2005). Outlier detection and estimation in nonlinear time series. *Journal of Time Series Analysis*, *26*(1), 107-121.

Bauwens, L., Dufays, A., & Rombouts, J. V. (2014). Marginal likelihood for Markov-switching and change-point GARCH models. *Journal of Econometrics*,178, 508-522.

Boettcher, M. (2011). Contrast and change mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(3), 215-230.

Buishand T.A. (1982). Some methods for testing the homogeneity of rainfall records*. J Hydrol* 58: 11–27.

Caussinus H & Mestre O. (2004). Detection and correction of artificial shifts in climate series, *J. Roy. Stat. Soc. Series C53:* 405–425.

Chang, I., Tiao, G. C., & Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics*, *30*(2), 193-204.

Chen, C., & Liu, L. M. (1993). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, *88*(421), 284-297.

Chen, C., & Tiao, G. C. (1990). Random level-shift time series models, ARIMA approximations, and level-shift detection. *Journal of Business & Economic Statistics*, *8*(1), 83-97.

Chen, J., & Gupta, A. K. (2012). *Parametric statistical changepoint analysis: with applications to genetics, medicine, and finance*. Springer Science & Business Media.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society*, 591-605.

Cuenod, C. A., Favetto, B., Genon-Catalot, V., Rozenholc, Y., & Samson, A. (2011). Parameter estimation and change-point detection from Dynamic Contrast Enhanced MRI data using stochastic differential equations. *Mathematical Biosciences*, 233(1), 68-76.

Davis, R. A., Huang, D., & Yao, Y. C. (1995). Testing for a change in the parameter values and order of an autoregressive model. *The Annals of Statistics*, 282-304.

Davis, R. A., Lee, T., & Rodriguez□Yam, G. A. (2008). Break detection for a class of nonlinear time series models. *Journal of Time Series Analysis*, *29*(5), 834-867.

Degras, D., Xu, Z., Zhang, T., & Wu, W. B. (2012). Testing for parallelism among trends in multiple time series. *IEEE Transactions on Signal Processing*, *60*(3), 1087-1097.

Ducre-Robitaille, J.F, Vincent, L.A., & Boulet, G. (2003). Comparison of techniques for detection of discontinuities in temperature series, *Int. J. Climatol.* 23: 1087-1101.

Easterling, D.R., Peterson, T.C. (1995). A new method for detecting undocumented discontinuities in climatological time series. *Int. J. Climatol*. 15: 369–377.

Efron, B. (1979). Computers and the theory of statistics: thinking the unthinkable. *SIAM review*, *21*(4), 460-480.

Friedman M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc* 32(200): 675–701.

Galeano, P., Peña, D., & Tsay, R. S. (2006). Outlier detection in multivariate time series by projection pursuit. *Journal of the American Statistical Association*, *101*(474), 654-669.

Gombay, E. (2008). Change detection in autoregressive time series. *Journal of Multivariate Analysis*, *99*(3), 451-464.

Gonçalves, S., & Politis, D. N. (2011). Discussion on the review by J.-P. Kreiss and E. Paparoditis. *Journal of the Korean Statistical Society*, *40*, 357395.

Gonzales-Rouco, J.F, Jimenez, J.L, Quesada, V., & Valero, F. (2001). Quality control and homogeneity of precipitation data in the Southwest of Europe. *Journal of Climate*, 14: 964–978.

Göktürk, O. M., Bozkurt, D., Şen, Ö. L., & Karaca, M. (2008). Quality control and homogeneity of Turkish precipitation data. *Hydrological processes*, *22*(16), 3210-3218.

Gregory, A. W., & Hansen, B. E. (1996). Residual-based tests for cointegration in models with regime shifts. *Journal of Econometrics*, *70*(1), 99-126.

Guijarro JA. (2013). Climatological series shift test comparison on running windows. *Quarterly Journal of the Hungarian Meteorological Service*. 117(1): 35-45.

Hall P, Horowitz JL, Jing BY. (1995). On blocking rules for the bootstrap withe dependent data. *Biometrika*, 82:561–574.

Hamilton, J. D. (1994). *Time series analysis* (Vol. 2). Princeton: Princeton university press.

Hanssen-Bauer I, Førland E. (1994). Homogenizing long Norwegian precipitation series. *J. Clim*. 7: 1001–1013.

Hawkins P M. (1977). Testing a sequence of observations for a shift in random location, *J. Am. Statist. Assoc.* 73: 180–185.

Horváth, L., Kokoszka, P., & Reeder, R. (2013). Estimation of the mean of functional time series and a two□sample problem. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(1), 103-122.

Horváth, L., Kokoszka, P., & Reimherr, M. (2009). Two sample inference in functional linear models. *Canadian Journal of Statistics*, *37*(4), 571-591.

Hušková, M., & Kirch, C. (2012). Bootstrapping sequential change-point tests for linear regression. *Metrika*, *75*(5), 673-708.

Jiang, Y. (2009). Inference and prediction in a multiple structural break model of economic time series. The University of Iowa.

Karabork, M.C, Kahya, E., & Komuscu, A.U. (2007). Analysis of Turkish precipitation data: homogeneity and the Southern Oscillation forcings on frequency distributions. *Hydrological Processes* 21: 3203–3210.

Karioti, V., & Caroni, C. (2004). Simple detection of outlying short time series. *Statistical Papers*, *45*(2), 267-278.

Kruskal, W.H. (1952). A nonparametric test for the several sample problem. *Ann. Math. Stat.* 23: 525–540.

Kruskal, W.H., & Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* 47: 583– 621.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 1217-1241.

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics* 4: 159-178.

Lahiri, S. N. (2013). *Resampling methods for dependent data*. Springer Science & Business Media.

Lahiri, S. N., Furukawa, K., & Lee, Y. D. (2007). A nonparametric plug-in rule for selecting optimal block lengths for block bootstrap methods. *Statistical Methodology*, 4(3), 292-321.

Lau, K. M., & Weng, H. (1995). Climate signal detection using wavelet transform: How to make a time series sing. *Bulletin of the American Meteorological Society*, *76*(12), 2391-2402.

Li S, Lund R. (2012). Multiple changepoint detection via genetic algorithms. *Journal of Climate*. 25 (2), 674-686.

Lund R, Reeves, J. (2002). Detection of undocumented changepoints: A revision of the two-phase regression model. *Notes and Correspondence*. 15: 2547-2554.

Mammen, E., & Nandi, S. (2012). Bootstrap and resampling. In *Handbook of Computational Statistics* (pp. 499-527). Springer Berlin Heidelberg.

Maronna R, & Yohai VJ. (1978). A bivariate test for detection of systematic change in mean. *J. Amer. Statis. Assoc.* 73: 640-645.

McNeil, A. J., & Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, *7*(3), 271-300.

McQuarrie, A. D., & Tsai, C. L. (2003). Outlier detections in autoregressive models. *Journal of Computational and Graphical Statistics*, *12*(2), 450-471.

Nordman, D. J., & Lahiri, S. N. (2014). Convergence rates of empirical block length selectors for block bootstrap. *Bernoulli*, 20(2), 958-978.

Panaretos, V. M., Kraus, D., & Maddocks, J. H. (2010). Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *Journal of the American Statistical Association*, *105*(490), 670-682.

Paparoditis, E., & Politis, D. N. (2001). Tapered block bootstrap. *Biometrika*, *88*(4), 1105-1119.

Paparoditis, E., & Politis, D. N. (2002). Local block bootstrap. *Comptes Rendus Mathematique*, *335*(11), 959-962.

Perron, P. (2017). Unit Roots and Structural Breaks. *Econometrics*, 5(2), 22.

Peterson, T.C., Easterling, D.R., Karl, T.R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullet, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Førland, E.J., Hanssen-Bauer, I., Alexandersson, H., Jones, P., & Parker, D. (1998). Homogeneity adjustments of *in situ* atmospheric climate data: A review. *Int. J. Climatol.* 18, 1493-1517.

Pettitt, A.N. (1979). A Non-Parametric Approach to the Change-Point Problem, *Journal of Applied Statistics* 28: 126-135.

Politis, D. N., & White, H. (2004). Automatic block-length selection for the dependent bootstrap. Econometric Reviews, 23(1), 53-70.

R Development Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rao, J. N., & Scott, A. J. (1987). On simple adjustments to chi-square tests with sample survey data. *The Annals of Statistics*, 385-397.

Ribeiro, S., Caineta, J., & Costa, A. C. (2016). Review and discussion of homogenisation methods for climate data. *Physics and Chemistry of the Earth, Parts A/B/C*, *94*, 167-179.

Rienzner, M., & Gandolfi, C. (2011). A composite statistical method for the detection of multiple undocumented abrupt changes in the mean value within a time series. *International Journal of Climatology*, *31*(5), 742-755.

Rodionov, S. (2015). A sequential method of detecting abrupt changes in the correlation coefficient and its application to Bering Sea climate. *Climate*, 3(3), 474-491.

Sahin, S., & Cigizoglu, H.K. (2010). Homogeneity analysis of Turkish meteorological data set. *Hydrol. Process.* 24: 981–992.

Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3), 599-607.

Savku, E., D. Pinheiro, Azevedo, N. and G.-W. Weber. (2015) Optimal Control of Stochastic Hybrid Models in the Framework of Regime Switches in Finance and Economics. 55th Meeting of EWGCFM, EURO Working Group "Commodities and Financial Modelling, 55, (2015), p.7.

Schwert, G. W. (2002). Tests for unit roots: A Monte Carlo investigation. *Journal of Business & Economic Statistics*, 20(1), 5-17.

Sobreira, N., Nunes, L. C., & Rodrigues, P. M. (2014). Characterizing economic growth paths based on new structural change tests. *Economic Inquiry*, *52*(2), 845-861.

Tayanc M, Dalfes N, Karaca M, & Yenigun O. (1998). A comparative assessment of different methods for detecting inhomogeneties in Turkish temperature data set. *Int. J. Climatol.* 18: 561–578.

Temoçin, B. Z., & Weber, G. W. (2014). Optimal control of stochastic hybrid system with jumps: a numerical approximation. *Journal of Computational and Applied Mathematics*, *259*, 443-451.

Toreti, A., Kuglitsch, F. G., Xoplaki, E., & Luterbacher, J. (2012). A novel approach for the detection of inhomogeneities affecting climate time series. *Journal of Applied Meteorology and Climatology*, *51*(2), 317-326.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *Journal of Forecasting*, *7*(1), 1-20.

Tuomenvirta H. (2002). Homogeneity testing and adjustment of climatic time series in Finland. *Geophysica* 38: 15-41.

Wang, X.L. (2008a). Accounting for autocorrelation in detecting mean shifts in climate data series using the penalized maximal t or F test. *J. Appl. Meteor. Climatol*. 47: 2423– 2444.

Wang, X.L. (2008b). Penalized maximal F test for detecting undocumented mean shift without trend change. *J. Atmos. Oceanic Technol*. 25: 368–384.

Wang, X.L., Wen, Q.H., & Wu, Y. (2007). Penalized maximal t test for detecting undocumented mean change in climate data series. *J. Appl. Meteor. Climatol*. 46: 916–931.

Wei, William Wu-Shyong. (1994). *Time series analysis*. Reading: Addison-Wesley publ.

Wijngaard, J. B., Klein Tank, A. M. G., & Können, G. P. (2003). Homogeneity of 20th century European daily temperature and precipitation series. *International Journal of Climatology*, 23(6), 679-692.

World Meteorological Organization. (2017). World Meteorological Organization. [ONLINE] Available at: https://public.wmo.int/en/our-mandate/climate. [Accessed 20 August 2017].

Yau, C. Y., & Zhao, Z. (2016). Inference for multiple changepoints in time series via likelihood ratio scan statistics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(4), 895-916.

Yazici C, Yozgatligil C,& Batmaz I. (2012). A simulation study on the performances of homogeneity tests applied in meteorological studies, ICACM: International Conference on Applied and Computational Mathematics Ankara, Turkey. Book of Abstracts, 93.

Yerlikaya-Özkurt, F., Askan, A., & Weber, G. W. (2016). A Hybrid Computational Method Based on Convex Optimization for Outlier

Problems: Application to Earthquake Ground Motion Prediction. *Informatica*, *27*(4), 893-910.

Yozgatligil C, Purutcuoglu V, Yazici C, Batmaz I. (2011). Validity of homogeneity tests for meteorological time series data: A simulation study, *Proceedings of the 58th World Statistics Congress (ISI2011)*.

Yozgatligil, C., & Yazici, C. (2016). Comparison of homogeneity tests for temperature using a simulation study. *International Journal of Climatology*, *36*(1), 62-81.

Zeger, S. L., Irizarry, R., & Peng, R. D. (2006). On time series analysis of public health and biomedical data. *Annu. Rev. Public Health*, *27*, 57-79.

Zivot, E., & Andrews, D. W. K. (1992). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business & Economic Statistics*, *20*(1), 25-44.

Table A1. The detection frequencies for mean shift

Detection rates	1°C				2 °C			
	Y_t		$Y_{t,shift}$		Y_t		$Y_{t,shift}$	
	SNHT-BS	**F-Test**	**SNHT-BS**	**F-Test**	**SNHT-BS**	**F-Test**	**SNHT-BS**	**F-Test**
At the beginning	0.284	0.044	0.388	0.056	0.328	0.064	0.532	0.116
In the middle	0.32	0.092	0.628	0.264	0.336	0.04	0.912	0.848
At the end	0.296	0.052	0.444	0.108	0.3	0.06	0.696	0.34

Table A2. The detection frequencies for sudden decrease

Detection rates	**Sudden Decrease (1^{o}C)**			
	Y_t		$Y_{t,shift}$	
	SNHT-BS	**F-Test**	**SNHT-BS**	**F-Test**
At the beginning	0.264	0.044	0.272	0.052
In the middle	0.324	0.056	0.324	0.056
At the end	0.288	0.048	0.280	0.052

Table A3. Frequencies of inhomogeneity when there is a mean shift of 1-unit increase

	At the beginning			In the middle			At the end		
Methods	Y_t	$Y_{t,shift}$	**Location Detection**	Y_t	$Y_{t,shift}$	**Location Detection**	Y_t	$Y_{t,shift}$	**Location Detection**
$LR_{monthly}$	0.01	0.12	26→0.011	0.01	1.00	355→0.13	0.00	0.19	700→0.26
$SNHT_{monthly}$	0.92	0.11	25→0.270	0.91	1.00	355→0.02	0.94	0.18	699→0.11
$SNHT_{annual}$	0.09	0.02	-	0.10	1.00	29th→0.84	0.10	0.66	58th→0.06 59th→0.06

Table A4. Frequencies of inhomogeneity when there is a mean shift of 2-unit increase

	At the beginning			**In the middle**			**At the end**		
Methods	Y_t	$Y_{t,shift}$	**Location Detection**	Y_t	$Y_{t,shift}$	**Location Detection**	Y_t	$Y_{t,shift}$	**Location Detection**
$LR_{monthly}$	0.00	0.86	26→0.37	0.01	0.99	355→0.34	0.01	0.97	700→0.30
$SNHT_{monthly}$	0.93	0.99	25→0.77	0.92	0.95	355→0.94	0.95	0.97	699→0.65
$SNHT_{annual}$	0.17	0.85	2th→0.91	0.11	0.99	29th→0.36	0.11	0.96	58th→0.83 59th→0.09

Table A5. Frequencies of inhomogeneity when there is a mean shift of 3-unit increase

	At the beginning			In the middle			At the end		
Methods	Y_t	$Y_{t,shift}$	**Location Detection**	Y_t	$Y_{t,shift}$	**Location Detection**	Y_t	$Y_{t,shift}$	**Location Detection**
$LR_{monthly}$	0.00	0.88	26→0.75	0.00	1.00	355→0.69	0.00	1.00	700→0.44
$SNHT_{monthly}$	0.89	0.98	25→0.85	0.93	1.00	354→0.90	0.96	0.96	699→0.62
$SNHT_{annual}$	0.10	0.98	2th →0.97	0.06	1.00	29th→0.5 30th→0.5	0.13	0.96	58th→0.88

Table A6. Frequencies of inhomogeneity when there is a mean shift of 5-unit increase

	At the beginning			**In the middle**			**At the end**		
Methods	Y_t	$Y_{t,shift}$	**Location Detection**	Y_t	$Y_{t,shift}$	**Location Detection**	Y_t	$Y_{t,shift}$	**Location Detection**
LR$_{monthly}$	0.00	0.97	26→0.91	0.01	1.00	355→0.84	0.02	1.00	700→0.85
SNHT$_{monthly}$	0.96	1.00	25→0.97	0.96	1.00	354→1.00	0.90	1.00	699→1.00
SNHT$_{annual}$	0.08	0.97	2th →0.97	0.08	1.00	29th→0.48 30th→0.52	0.10	1.00	58th→1.00

Table A7. Frequencies of inhomogeneity when there is a mean shift of 10-unit increase

	At the beginning			**In the middle**			**At the end**		
Methods	Y_t	$Y_{t,shift}$	**Location Detection**	Y_t	$Y_{t,shift}$	**Location Detection**	Y_t	$Y_{t,shift}$	**Location Detection**
$\text{LR}_{\text{monthly}}$	0.03	0.98	26→0.98	0.00	1.00	355→1.00	0.00	1.00	700→1.00
$\text{SNHT}_{\text{monthly}}$	0.97	0.98	25→1.00	0.93	1.00	354→1.00	0.94	1.00	699→1.00
$\text{SNHT}_{\text{annual}}$	0.14	0.98	2th→1.00	0.07	1.00	29th→0.51 30th→0.49	0.10	1.00	58th→1.00

Table A8. Frequencies of inhomogeneous results when the mean shift starts at the beginning of the series

Sample Size		**BEGINNING**							
	Location of the Breakpoint		Yt			Yt,shift			
		Block Size	$\text{SNHT}_{\text{monthly}}$	$\text{SNHT}_{\text{annual}}$	$\text{LR}_{\text{monthly}}$	$\text{SNHT}_{\text{monthly}}$	$\text{SNHT}_{\text{annual}}$	$\text{LR}_{\text{monthly}}$	LRT-BS
$n = 50$	10	$l = 15$	-	-	-	~0 (25)	-	-	0.41 (10→0.25)
		$l = 20$	-	-	-	~0 (25)	-	-	0.28 (10→0.90)
		$l = 30$	-	-	-	~0 (25)	-	-	0.36 (9→0.30, 10→0.23)
$n = 60$	10		$\text{SNHT}_{\text{monthly}}$	$\text{SNHT}_{\text{annual}}$	$\text{LR}_{\text{monthly}}$	$\text{SNHT}_{\text{monthly}}$	$\text{SNHT}_{\text{annual}}$	$\text{LR}_{\text{monthly}}$	LRT-BS
		$l = 10$	~0 (49)	-	-	~0 (9)	-	-	0.12 (10→0.84)
		$l = 20$	~0 (49)	-	-	~0 (9)	-	-	0.18 (10→0.87)
		$l = 30$	~0 (49)	-	-	~0 (9)	-	-	0.10 (10→0.87)
		$l = 40$	~0 (49)	-	-	~0 (9)	-	-	0.13 (10→0.85)

Table A8 (contd'). Frequencies of inhomogeneous results when the mean shift starts at the beginning of the series

Sample Size		BEGINNING							
	Location of the Breakpoint		Yt			Yt,shift			
		Block Size	SNHT$_{\text{monthly}}$	SNHT$_{\text{annual}}$	LR$_{\text{monthly}}$	SNHT$_{\text{monthly}}$	SNHT$_{\text{annual}}$	LR$_{\text{monthly}}$	LRT-BS
n = 70	15		SNHT$_{\text{monthly}}$	SNHT$_{\text{annual}}$	LR$_{\text{monthly}}$	SNHT$_{\text{monthly}}$	SNHT$_{\text{annual}}$	LR$_{\text{monthly}}$	LRT-BS
		l = 10	~0 (5)	-	-	~0 (14)	-	-	0.30 (15→0.72)
		l = 20	~0 (5)	-	-	~0 (14)	-	-	0.15 (15→0.80)
		l = 30	~0 (5)	-	-	~0 (14)	-	-	0.10 (15→0.82)
		l = 40							0.3 (15→0.83)
n = 80	20		SNHT$_{\text{monthly}}$	SNHT$_{\text{annual}}$	LR$_{\text{monthly}}$	SNHT$_{\text{monthly}}$	SNHT$_{\text{annual}}$	LR$_{\text{monthly}}$	LRT-BS
		l = 10	-	-	-	-	-	-	0.7 (20→0.50)
		l = 20	-	-	-	~0 (17)	-	-	0.11 (20→0.66)
		l = 30	-	-	-	~0 (17)	-	-	0.14 (20→0.74)
		l = 40	-	-	-	~0 (17)	-	-	0.14 (20→0.75)

Table A8 (contd’). Frequencies of inhomogeneous results when the mean shift starts at the beginning of the series

Sample Size		**BEGINNING**							
	Location of the Breakpoint		Yt			Yt,shift			
		Block Size	SNHT$_{monthly}$	SNHT$_{annual}$	LR$_{monthly}$	SNHT$_{monthly}$	SNHT$_{annual}$	LR$_{monthly}$	LRT-BS
n = 90	20		SNHT$_{monthly}$	SNHT$_{annual}$	LR$_{monthly}$	SNHT$_{monthly}$	SNHT$_{annual}$	LR$_{monthly}$	LRT-BS
		l = 10	~0 (13)	-	-	~0 (16)	-	-	0.34 (19→0.16)
		l = 20	~0 (13)	-	-	~0 (16)	-	-	0.22 (19→0.23)
		l = 30	~0 (13)	-	-	~0 (16)	-	-	0.40 (19→0.18)
		l = 40	1~0 (13)	-	-	~0 (16)	-	-	0.46 (19→0.15)

Table A9. Frequencies of inhomogeneous results when the mean shift starts at the end of the series

Sample Size		END							
	Location of the Breakpoint		**Yt**			**Yt,shift**			
		Block Size	SNHT$_{monthly}$	SNHT$_{annual}$	LR$_{monthly}$	SNHT$_{monthly}$	SNHT$_{annual}$	LR$_{monthly}$	LRT-BS
n = 50	40	*l* = 10	~0 (53)	-	-	~0 (43)	-	-	0.8 (40→0.77)
		l = 20	~0 (53)	-	-	~0 (43)	-	-	0.23 (40→0.90)
		l = 30	~0 (53)	-	-	~0 (43)	-	-	0.38 (40→0.92)
		l = 40	~0 (53)	-	-	~0 (43)	-	-	0.38 (40→0.92)
			SNHT$_{monthly}$	SNHT$_{annual}$	LR$_{monthly}$	SNHT$_{monthly}$	SNHT$_{annual}$	LR$_{monthly}$	LRT-BS
n = 60	45	*l* = 10	-	-	-	~0 (44)	-	-	0.2 (45→0.87)
		l = 20	-	-	-	~0 (44)	-	-	0.4 (45→0.92)
		l = 30	-	-	-	~0 (44)	-	-	0.6 (45→0.94)
		l = 40	-	-	-	~0 (44)	-	-	0.6 (45→0.94)

Table A9(contd'). Frequencies of inhomogeneous results when the mean shift starts at the end of the series

Sample Size		**END**							
	Location of the Breakpoint		**Yt**			**Yt,shift**			
		Block Size	SNHT$_{monthly}$	SNHT$_{annual}$	LR$_{monthly}$	SNHT$_{monthly}$	SNHT$_{annual}$	LR$_{monthly}$	LRT-BS
n = 70	55		SNHT$_{monthly}$	SNHT$_{annual}$	LR$_{monthly}$	SNHT$_{monthly}$	SNHT$_{annual}$	LR$_{monthly}$	LRT-BS
		l = 10	~0 (29)	-	-	~0 (54)	-	-	0.16 (55→0.33)
		l = 20	~0 (29)	-	-	~0 (54)	-	-	0.22 (55→0.32)
		l = 30	~0 (29)	-	-	~0 (54)	-	-	0.24 (55→0.35)
		l = 40	~0 (29)	-	-	~0 (54)	-	-	0.30 (55→0.30)
n = 80	65		SNHT$_{monthly}$	SNHT$_{annual}$	LR$_{monthly}$	SNHT$_{monthly}$	SNHT$_{annual}$	LR$_{monthly}$	LRT-BS
		l = 10	~0 (40)	-	-	~0 (57)	-	-	0.2 (64→0.35)
		l = 20	~0 (40)	-	-	~0 (57)	-	-	0.12 (64→0.44)
		l = 30	~0 (40)	-	-	~0 (57)	-	-	0.10 (64→0.35)
		l = 40	~0 (40)	-	-	~0 (57)	-	-	0.15 (64→0.57)

Table A9(contd'). Frequencies of inhomogeneous results when the mean shift starts at the end of the series

Sample Size		**END**							
	Location of the Breakpoint		**Yt**			**Yt,shift**			
		Block Size	$\text{SNHT}_{\text{monthly}}$	$\text{SNHT}_{\text{annual}}$	$\text{LR}_{\text{monthly}}$	$\text{SNHT}_{\text{monthly}}$	$\text{SNHT}_{\text{annual}}$	$\text{LR}_{\text{monthly}}$	LRT-BS
n = 90	70		$\text{SNHT}_{\text{monthly}}$	$\text{SNHT}_{\text{annual}}$	$\text{LR}_{\text{monthly}}$	$\text{SNHT}_{\text{monthly}}$	$\text{SNHT}_{\text{annual}}$	$\text{LR}_{\text{monthly}}$	LRT-BS
		l = 10	~0 (6)	-	-	~0 (65)	-	-	0.16 (70→0.55)
		l = 20	~0 (6)	-	-	~0 (65)	-	-	0.25 (70→0.46)
		l = 30	~0 (6)	-	-	~0 (65)	-	-	0.35 (70→0.52)
		l = 40	~0 (6)	-	-	~0 (65)	-	-	0.35 (70→0.52)

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Yazıcı, Ceyda
Nationality: Turkish (TC)
Date and Place of Birth: 11 January 1986, Ankara
Marital Status: Single
Phone: +90 312 210 53 11
Fax: +90 312 210 29 59
email: ceydayazici86@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
MS	METU Department of Statistics	2011
BS	METU Department of Statistics	2009
High School	Milli Piyango Anadolu High School, Ankara	2004

WORK EXPERIENCE

Year	Place	Enrollment
2009 – Present	METU Department of Statistics	Research Assistant

PUBLICATIONS

1. Batmaz, I., Danisoglu, S., Yazici, C., and Kartal-Koc, E. “A data mining application to deposit pricing”, Applied Soft Computing, Doi: 10.1016/j.asoc.2017.07.047 (2017).

2. Yozgatligil, C. and Yazici, C. “Comparison of homogeneity tests for temperature using a simulation study”. International Journal of Climatology, 36(1), 62-81 (2016).

3. Yazici, C., Yerlikaya-Ozkurt, F., and Batmaz, I. “A computational approach to nonparametric regression: bootstrapping CMARS method”. Machine Learning 101 (1-2) 211-230, Doi: 10.1007/s10994-015-5502-3 (2015).