Locating Multiple Change-Points Using a Combination of Methods

Johan Andersson

Master of Science Thesis

Stockholm, Sweden 2014

Lokalisering av multipla brytpunkter med en kombination av metoder

Johan Andersson

Examensarbete

Stockholm, Sverige 2014

Abstract

The aim of this study is to find a method that is able to locate multiple change-points in a time series with unknown properties. The methods that are investigated are the CUSUM and CUSUM of squares test, the CUSUM test with OLS residuals, the Mann-Whitney test and Quandt's log likelihood ratio. Since all methods are detecting single change-points, the binary segmentation technique is used to find multiple change-points. The study shows that the CUSUM test with OLS residuals, Mann-Whitney test and Quandt's log likelihood ratio work well on most samples while the CUSUM and CUSUM of squares are not able to detect the location of the change-points. Furthermore the study shows that the binary segmentation technique works well with all methods and is able to detect multiple change-points in most circumstances. The study also shows that the results can, most of the time, be improved by using a combination of the methods.

Sammanfattning

Syftet med studien är att hitta en metod som identifierar tidpunkterna för strukturella brott i en tidsserie med okända egenskaper. De metoder som undersöks är CUSUM och CUSUM av kvadrater, CUSUM test med OLS-residualer, Mann-Whitney-test samt Quandts log likelihood ratio. Eftersom alla metoder identifierar enbart en brytpunkt används binära uppdelningstekniken för att hitta multipla brytpunkter. Studien visar att CUSUM-test med OLS-residualer, Mann-Whitney-test och Quandt's log likelihood ratio fungerar bra för de flesta stickproven medan CUSUM och CUSUM av kvadrater inte hittar tidpunkten för brytpunkterna. Vidare så visar studien att binära uppdelningstekniken fungerar bra med alla metoder och kan identifiera multipla brytpunkter i de flesta fallen. Studien visar också att resultaten för det mesta kan förbättras genom att använda en kombination av metoderna.

Acknowledgements

I would like to thank my supervisors, Camilla Landén at KTH and Jovan Zamac at Handelsbanken. Camilla, thank you for your support and advice throughout the process. Jovan, thank you for your suggestions, feedback and constant support.

I would also like to thank Erik Svensson at Handelsbanken. Thank you for your support.

Contents

List of figures

List of tables

1 Introduction

1.1 Background

Time series are frequently used for analyzing the past and making predictions about the future. Generally it is assumed that a longer time series contains more information and that the prediction of the future becomes better the longer the time series is. It has however been observed that many time series contain structural breaks. A structural break is a change in the underlying distribution of the data and the point that the change occurs at is called a change-point. This means that the assumption of homogenous distribution throughout the whole time series is not correct. To be able to use all data available for analysis, the breaks need to be located. The theory surrounding change-point detection aims to discover if and where such shifts are present in the data.

The detection of change-points does not only help to avoid faulty assumptions about the data. The location and size of the shifts can be used to extract further information about the underlying properties of the data. When the change-points have been identified the analysis can be continued to explain why the break occurred.

Change-point detection has applications in a wide variety of fields e.g. ecology (Beckage, et al., 2007), economics (Talwar, 1983) and medicine (Barros & Nunes, 2010) and is therefore of interest to many, both practitioners and theoreticians. But even though the theory has a wide range of applications, the properties of data are often similar from case to case. It consists of longitudinal data with a number of dependent and independent variables. This means that results based on a study in one field can often be applied in another field.

However, even though the change-point detection finds many uses in practice, most of the methods are developed in a theoretical framework. This is often necessary to derive the correct properties of the tests but it presents an issue when applying the methods to real life data. Real life data do seldom fit the narrow assumptions made in the theoretical framework. A consequence of this is that there is uncertainty as to how the performance of the methods is affected when they are applied to real life data.

The complicated mathematical models behind the investigated methods often lead to other simplifications. One such simplification is that most methods are only able to detect one change-point in the data. This is a restriction that is not present in real life data and in reality it can often be observed that longer time series have multiple change-points. To be able to perform an appropriate analysis all change-points need to be located.

As mentioned the interest for change-points is vast and ranges between many different subjects of which one is the financial sector. The financial sector deals with a massive data flow from which it often needs to identify if there is a structural break or not in different data ranging from market data to client accounts. Identifying change-points in individual accounts might be a tool to find irregular activities that would motivate further action from the bank. For this reason this thesis is carried out at Svenska Handelsbanken AB (publ), hereafter known as ‘the Bank’. The Bank provides data that represents the pattern of the daily balances for the Bank’s Swedish accounts.

1.2 Purpose

This study aims to find a method that is able to detect the location of multiple change-points in a time series. Since the analyzed data does not always fit the assumptions of the methods, this study aims to investigate how the performance of the methods is affected when the assumptions are not satisfied.

1.3 Outline

The rest of this paper is structured as follows. Section 2 provides an overview of the literature in the field of change-point analysis and the methods used in this study are described. Section 3 explains the methodology for this study and details of the calculations. Section 4 presents the results from the analysis performed on both artificially generated data and real life data. Section 5 contains interpretations and discussions of the results. Section 6 consists of concluding remarks and suggestions for further research.

2 Literature

This chapter covers the theory surrounding change-point detection. It begins with defining what a change-point is and continues with presenting the different types of change-point problem. The final section of the chapter describes the different methods that are used to detect change-points.

2.1 The change-point problem

The change-point theory is usually based on the regression model

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta}_t + \varepsilon_t \ (\mathrm{t} = 1,2, \ldots, \mathrm{T}) \tag{2.1}$$

Where at time t, y_t is the observation on the dependent variable and $\boldsymbol{x}_t$ is the column vector of observations on k regressors. The first regressor is equal to one for all values of t if the model contains a constant. $\boldsymbol{\beta}_t$ is a $k \times 1$ vector containing the model parameters and ε_t is the error term which is usually assumed to be normally independently distributed.

The null hypothesis of no structural breaks is

$$\boldsymbol{\beta}_i = \boldsymbol{\beta}_0 \ \forall i \tag{2.2}$$

The alternative hypothesis is that there exist points $1 < k_1 < k_2 < \cdots < k_n < T$ such that

$$\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_{k_1} \neq \boldsymbol{\beta}_{k_1+1} = \cdots = \boldsymbol{\beta}_{k_2} \neq \boldsymbol{\beta}_{k_2+1} = \cdots = \boldsymbol{\beta}_{k_n} \neq \boldsymbol{\beta}_{k_n+1} = \cdots = \boldsymbol{\beta}_T \tag{2.3}$$

If the null hypothesis is rejected the time series $(\boldsymbol{x}_i, y_i)$ is said to have n structural breaks with the change-points $k_1, k_2, \ldots, k_n$.

The presence of structural breaks in time series has been known to exist for a long time. One method to detect the breaks is to employ a moving average instead of a regression on the whole time period. However, this approach runs into problems when deciding the length of the moving window. As Page (1954) points out, a small window will be able to detect a large change rapidly but small changes are detected slowly. A large window is better for detecting small changes but large changes will be dampened by the smoothing effect of the moving average. The size of the window hence depends on the properties of the breaks, which are often not known in advance.

Instead Page chooses another approach and develops the foundation for change-point detection theory (Page, 1954). The first application is quality control in factory productions and hence the initial theory is developed for so called online data. For online data new data is sequentially added to the time series and then tested for structural breaks. A key concept when analyzing online data is Average Run Length (ARL) which is defined as the expected number of articles sampled before action is taken. In a satisfactory setting ARL is measuring the Type I error[1], since it measures the rate of false alarms. When the quality is poor it is measuring Type II error[2], since it measures the time it takes to react to errors in the production.

There is another branch of the theory that is concerning offline data, which is a fixed sample of historical data. The methods are based on the same principles as those for online data but they have somewhat different focuses. Tests for offline data are concerned with the significance of the

[1] A type I error is the incorrect rejection of a true null hypothesis

[2] A type II error is the failure to reject a false null hypothesis

detected change-points and search for change-points in the whole data set, while online tests only investigate changes at the end of the sample. This study will focus on tests on offline data.

Depending on the nature of the problem and the prior knowledge about the data, a change-point problem can be sorted into three different categories. The method is then chosen appropriately to fit the category. The main categories of change-point problems are:

1) A known number of change-points at known times
2) A known number of change-points at unknown times
3) An unknown number of change-points at unknown times

The method for detecting change-points in category 1) is usually a Chow test (Chow, 1960). This test is appropriate when there is a specific point in time that is suspected to be a change-point, e.g. crime statistics before and after a law is passed. In many applications this is not the case, and in particular it cannot be applied to this study.

Category 2) provides a simple model since the null and alternative hypotheses are easily constructed. In this category all parameters including the change-points can be estimated. Their statistical properties are investigated by Bai and Perron (1998). It is however unusual that the number of change-points are known in advance.

This study is concerned with category 3) since neither the number nor the locations of the change-points are known in advance. Several methods have been suggested on how to find unknown change-points but there is no universally correct method that applies to all circumstances. Instead a method is chosen depending on the data the analysis is performed on and what the applications of the results should be.

One of the difficulties of an unknown number of change-points is the construction of the null and alternative hypothesis. Since the number of change-points is unknown, it is impossible to in advance construct an accurate alternative hypothesis. Many articles are employing a test for one change-point (versus none). In longer time series there is often multiple structural breaks and disregarding the possibility for multiple breaks will result in flawed results.

One way to utilize single change-point methods when investigating the possibility of multiple change-points is to employ the binary segmentation technique described by e.g. Chen and Gupta (2012). The technique starts with testing the whole period for a single change-point (versus none). If one change-point is discovered, the time period is divided into two parts (divided by the change-point). The separate parts are then tested individually for the occurrence of a change-point. The procedure is repeated until no new change-points are discovered. This method allows the complex problem with an unknown number of change-points to be reduced to a test for one change-point versus none performed sequentially. Bai and Perron (1998) suggest a similar procedure to test for l versus $l + 1$ breaks.

2.2 Change-point detection methods

The literature contains a plethora of change-point detection methods. A selection of methods is tested in this study. The methods are chosen because they are easy to implement, fast to calculate and commonly occurring in the literature. They are also chosen to work on different kinds of data, creating a heterogeneous set of methods.

CUSUM (cumulative sum) is a method that is well described in the literature. Page (1954) is the first to describe the method and introduces it as a test for online data. The method is further developed and made to fit offline data by e. g. Hinkley (1971) and Brown, Durbin and Evans (1975). The CUSUM method is based on the cumulative sum of the sequential residuals from an Ordinary Least Squares (OLS) regression. The null hypothesis is that there is no change-point and the alternative is that there is one change-point at an unknown time. The model rests on the assumption that the error terms are normally independently distributed. This assumption is then used to derive a limit for the rejection of the null hypothesis.

The CUSUM of Squares (CUSUMSQ) test is a similar method that is also described by Brown et al. (1975). The method is instead using the sum of squared sequential residuals. This version of the test fits better to find haphazard changes rather than systematic changes and works well as a complement to the CUSUM test (Brown, et al., 1975).

One disadvantage with the standard method of CUSUM is that the power for late structural breaks is rather low, meaning that structural breaks that occur late in the time series risk being undetected by the method. Ploberger and Krämer (1992) examine the relative performance of a CUSUM test on the OLS residuals (hereafter referenced as OLSCUSUM). Instead of sequential residuals, the method is using OLS residuals over the whole time period. This means that it performs better at detecting late structural breaks. The method is based on the assumption that the residuals are independent and identically distributed (i.i.d). The null distribution for this test is harder to derive since the OLS residuals are correlated and heteroscedastic even under the null hypothesis (Ploberger & Krämer, 1992). Also, since the residuals sum to zero, the cumulative sum does not tend to drift off after a structural change. Despite those problems Ploberger and Krämer (1992) derive the null distribution and construct a test on the OLS residuals. They conclude that the OLS based CUSUM method reacts better on late structural shifts but that no version of the test is uniformly superior to the other.

Another test for a single change-point is Quandt's log-likelihood test (hereafter referenced as 'Quandt method' or 'Quandt') which is first introduced by Quandt (1958). The method is based on the likelihood ratio λ defined as

$$\lambda = \frac{L(\widehat{\omega})}{L(\widehat{\Omega})} \tag{2.4}$$

where $L(\widehat{\Omega})$ is the unrestricted maximum of the likelihood function over the entire parameter space Ω and $L(\widehat{\omega})$ is the maximum of the likelihood function over the subspace $\omega \subset \Omega$ to which one is restricted by the hypothesis. In this context Ω corresponds to no breaks while ω corresponds to one break in the time series. The simple computations of this measurement make it preferred over more complicated methods. However, a severe limitation of the model is that the distribution of the ratio is unknown and hence the results are only indicative. This issue is resolved by Deutsch (1992) who calculates the distribution empirically, making the test more viable outside of an indicative nature. Since the method is based on the likelihood-function it is assumed that the underlying distribution of the data is known.

Most of the described tests are constructed under the assumption that the data is normally distributed. This is not always true in practice and a more robust test seems appropriate. Talwar (1983) provides a comparison between some methods and their robustness. One such method is the homogeneity test discussed in the article by Brown et al. (1975) which turns out well against heavy tailed distributions. The method is dividing the time series in different parts and performing a piecewise regression and then analyzing the differences in variances. It is however difficult to use the method to identify the locations of the change-points, and it rather works as an indicative method.

Another robust method is the method based on the Mann-Whitney two sample test (MW) described by Pettitt (1979). The Mann-Whitney method is non-parametric which means that is not based on any assumptions of the underlying distribution of the residuals. Since it is based on ranks it is also insensitive to outliers.

Some of the methods that are described are based on an assumption of i.i.d. residuals. This could become an issue since there is often a time-dependence in a time series. Alippi et al. (2013) try to resolve this problem by using an ensemble method where random subsamples of the data are drawn, which removes the time dependence from the sample. The analysis is then performed on the subsample. Then a new random sample is drawn and the analysis is performed on this sample. This procedure is repeated for a fixed number of times. The results are then combined by a weighted average. Alippi, et al. (2013) use the Lepage statistic which is based on the Mann-Whitney test statistic combined with the Mood test statistic. The Mann-Whitney statistic locates changes in the mean while the Mood statistic locates changes in the variance. Alippi et al. (2013) show that the Ensemble method improves the change-point estimates when the residuals are not i.i.d.

3 Methodology

In this chapter the methods used in the study are described in detail. First the binary segmentation technique is presented. Then the change-point detection methods CUSUM, CUSUMSQ, OLSCUSUM, Quandt and MW are described. Thereafter the ensemble and combined method are explained. The chapter is concluded with a description of the data and how the methods are evaluated.

3.1 Binary segmentation technique

The binary segmentation technique is a technique that makes it possible to use a single change-point method to detect multiple change-points sequentially. The technique is described by e. g. Chen & Gupta (2012) and performed as follows.

Let $I_i = \{1, k_1, \dots, k_n, T\}$ denote the partition of the interval $[1, T]$ into subintervals $[1, k_1], [k_1, k_2], \dots, [k_{n-1}, k_n], [k_n, T]$ where $1 < k_1 < \cdots < k_n < T$

1) The initial partition is $I_0 = \{1, T\}$ i.e. the whole sample
2) Test each of the subintervals given by I_i for change-points k_j
3) Add the l found change-points k_j to the partition: $I_{i+1} = \{1, k_1, \dots, k_{n_i}, T\}$, $n_i = n_{i-1} + l$
4) Repeat from 2)

The algorithm is iterated until no more change-points are found or with a set amount of iterations. One possible variation is to only add the most significant change-point to the partition in step 3. This is however more computationally demanding and most of the time does not make any difference in the final result.

The binary segmentation technique is illustrated in Figure 1 on a fictional time series symbolized by rectangles.

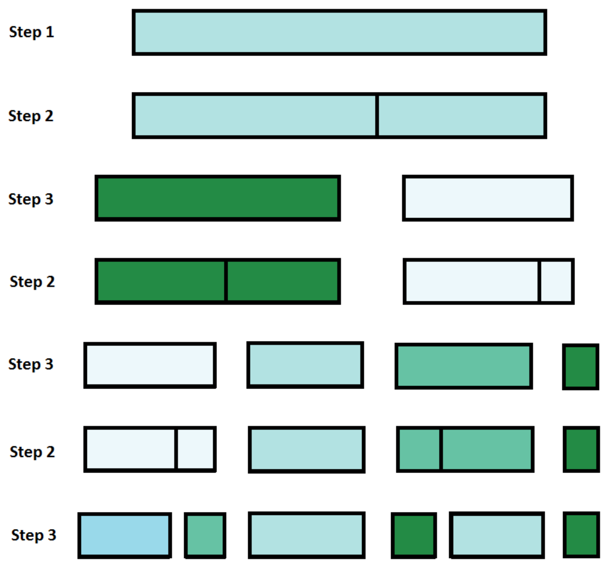


Figure 1 – Illustration of the binary segmentation technique

Figure 1 demonstrates that the binary segmentation technique can locate multiple change-points using single change-point methods sequentially. It also shows that the time intervals become smaller and smaller. It should be noted that if no change-points are found in the subinterval, the subinterval is kept intact.

3.2 CUSUM and CUSUMSQ

The CUSUM- and CUSUMSQ-test are based on the recursive residuals from a regression. The tests are described in detail by Brown et al. (1975) but the calculations are outlined in this section. The recursive residuals w_t from the observations $(\boldsymbol{x_t}, y_t)$, $1 \leq t \leq T$ are calculated as follows

$$w_t = \frac{e_t}{\sqrt{v_t}} \tag{3.1}$$

Where

$$e_t = y_t - \boldsymbol{x}_t'\boldsymbol{\beta}^{(t)} \tag{3.2}$$

$$\boldsymbol{\beta}^{(t)} = \left[\sum_{i=1}^{t-1} \boldsymbol{x}_i\boldsymbol{x}_i'\right]^{-1} \left(\sum_{i=1}^{t-1} \boldsymbol{x}_i y_i\right) \tag{3.3}$$

$$v_t = 1 + \boldsymbol{x}_t'\left[\sum_{i=1}^{t-1} \boldsymbol{x}_i\boldsymbol{x}_i'\right]^{-1} \boldsymbol{x_t} \tag{3.4}$$

The CUSUM and CUSUMSQ statistics are then calculated as

$$CUSUM_t = \sum_{i=k+1}^{t} \frac{w_i}{\sigma_w} \tag{3.5}$$

$$CUSUMSQ_t = \frac{\sum_{i=k+1}^{t} w_i^2}{\sum_{i=k+1}^{T} w_i^2} \tag{3.6}$$

Where

$$\sigma_w = \sqrt{\frac{\sum_{i=k+1}^{T}(w_i - \widehat{w})^2}{(T-k-1)}} \tag{3.7}$$

$$\widehat{w} = \frac{1}{T-k}\sum_{i=k+1}^{T} w_i \tag{3.8}$$

And k is the number of regressors.

The upper and lower critical values for $CUSUM_t$ are

$$\pm a\left[\sqrt{T-k} + 2\frac{(t-k)}{(T-k)^{\frac{1}{2}}}\right] \tag{3.9}$$

Where a in equation (3.9) is given by Brown et al. (1975) to be 1.143 for significance level 0.01, 0.948 for 0.05 and 0.850 for 0.10.

The critical values for $CUSUMSQ_t$ are

$$\pm a + \frac{(t-k)}{T-k} \tag{3.10}$$

Where the value of a in equation (3.10) is obtained from a table by Durbin (1969) if $\frac{1}{2}(T-k) - 1 \leq 60$. Edgerton and Wells (1994) provide a method of obtaining the value of a for larger samples.

If the statistic moves outside of the critical value-boundaries the conclusion is that there is a structural break and the change-point is set to be the point where the statistic first crosses the boundary.

The CUSUM and CUSUMSQ statistics and their boundaries are illustrated with an example in Figure 2 and Figure 3.

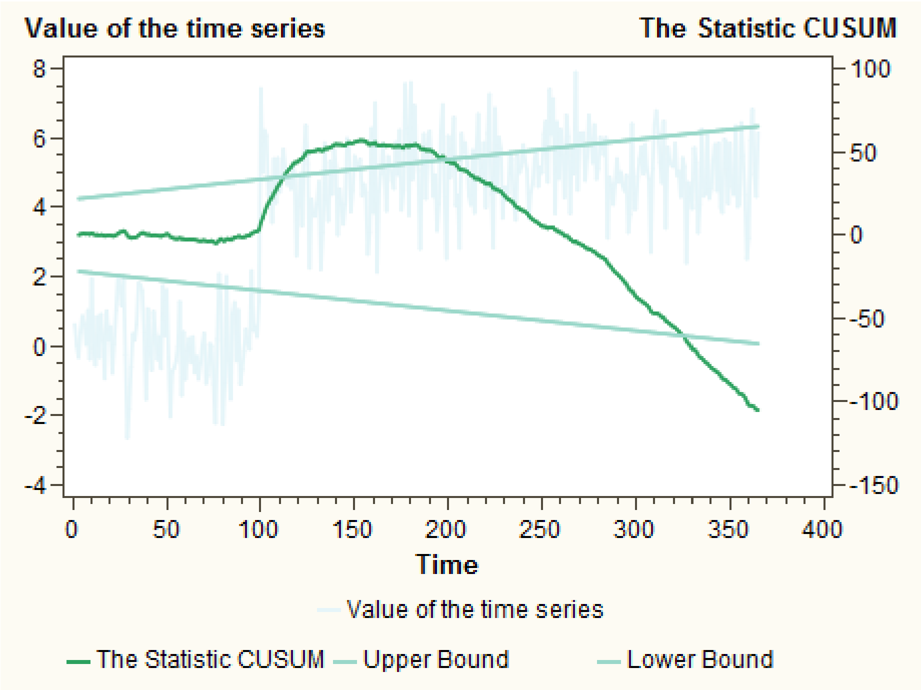


Figure 2 – A time series with a break and the statistic $CUSUM_t$ with upper- and lower confidence bounds

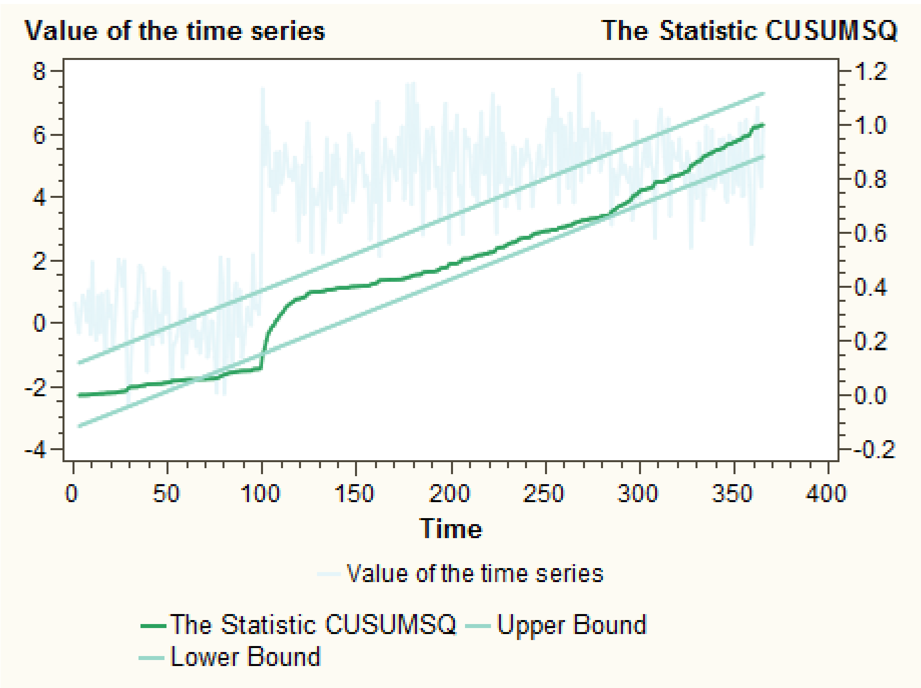


Figure 3 – A time series with a break and the statistic $CUSUMSQ_t$ with upper- and lower confidence bounds

3.3 OLSCUSUM

The calculations of the OLSCUSUM test are similar to those of the CUSUM and CUSUMSQ test, but they are based on an OLS on the entire sample. The details and the proof of the method can be found in the article by Ploberger and Krämer (1992). In this section the calculation of the method is outlined.

The cumulated sum of the OLS residuals is given by

$$B_t^{(T)} = \frac{1}{\hat{\sigma}\sqrt{T}} \sum_{i=1}^{t} e_i^{(T)} \tag{3.11}$$

Where

$$e_i^{(T)} = y_i - \boldsymbol{x}_i' \beta^{(T)} \tag{3.12}$$

$$\beta^{(T)} = \left[\sum_{i=1}^{T-1} \boldsymbol{x}_i \boldsymbol{x}_i'\right]^{-1} \left(\sum_{i=1}^{T-1} \boldsymbol{x}_i y_i\right) \tag{3.13}$$

$$\hat{\sigma} = \sqrt{\frac{1}{T} \sum_{t=k+1}^{T} \left(e_i^{(T)}\right)^2} \tag{3.14}$$

And finally the test statistic is

$$OLSCUSUM = \sup_{1 \le t \le T} \left|B_t^{(T)}\right| \tag{3.15}$$

Ploberger and Krämer (1992) show that

$$P(OLSCUSUM > a) \to 2 \sum_{j=1}^{\infty} (-1)^{j+1} \exp\left(-2j^2 a^2\right) \tag{3.16}$$

As $T \to \infty$

They also provide the critical values of a for different significance levels, 1.22 ($\alpha = 10\%$), 1.36 ($\alpha = 5\%$) and 1.63 ($\alpha = 1\%$). Furthermore they show that the asymptotic approximation works well for moderate sample sizes and that the test is almost always conservative.

The test statistic for an example time series is illustrated in Figure 4.

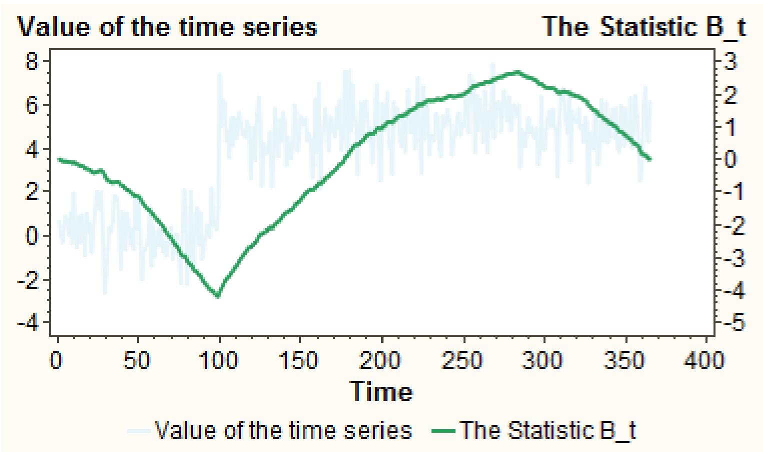


Figure 4 – A time series with a break and the statistic $B_t^{(T)}$ from the OLSCUSUM method

3.4 Quandt's log likelihood ratio

The Quandt's log likelihood ratio is based on the likelihood ratio of one break versus no breaks in the time series. Assumptions of the distribution of the data have to be made to calculate the likelihood function. Because of its simple implementation and its frequent occurrence a normal distribution is assumed. The calculation of the statistic is outlined next, but described in more detail by Quandt (1958) and Deutsch (1992).

The likelihood ratio statistic is given by

$$\lambda = \frac{(\hat{\sigma}_1^2)^{\frac{t_0}{2}}(\hat{\sigma}_2^2)^{\frac{T-t_0}{2}}}{(\hat{\sigma}^2)^{\frac{T}{2}}} \tag{3.17}$$

Where

$$\hat{\sigma}_1^2 = \frac{\sum_{t=1}^{t_0}\left(y_t - x_t'\hat{\beta}_1\right)^2}{t_0} \tag{3.18}$$

$$\hat{\sigma}_2^2 = \frac{\sum_{t=t_0+1}^{T}\left(y_t - x_t'\hat{\beta}_2\right)^2}{T - t_0} \tag{3.19}$$

and

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{T}\left(y_t - x_t'\hat{\beta}\right)^2}{T} \tag{3.20}$$

Where $\hat{\beta}_1$ is the estimated parameters from an OLS regression on $(\boldsymbol{x_t}, y_t), t \leq t_0$, $\hat{\beta}_2$ is the estimated parameters from an OLS regression on $(\boldsymbol{x_t}, y_t), t > t_0$ and $\hat{\beta}$ is the estimated parameters from an OLS regression on the whole sample.

The test statistic is then calculated as

$$Quandt = \max_{k<t_0<T-k} -log\lambda \tag{3.21}$$

The critical values are given by Deutsch (1992) empirically. He also shows that the tail is practically not affected by the sample size when it is larger than 100. The critical values for a sample size of 200 are given in Table 1.

P	0.50	0.75	0.90	0.95	0.99
Critical value	6.2	8.1	10.5	12.6	17.2

Table 1 - Cumulative distribution of Quandt's statistic with 2 explanatory variables and a sample size of 200

The Quandt test statistic is illustrated in Figure 5.

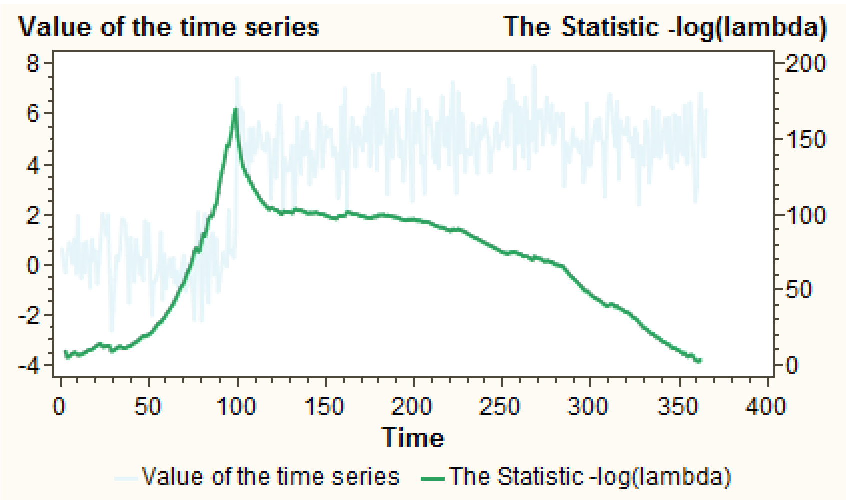


Figure 5 – A time series with a break and the statistic -log(λ) from the Quandt method

3.5 Mann-Whitney

The Mann-Whitney test is a non-parametric test which means that it can be used on data regardless of the underlying distribution. It is described in detail by Pettitt (1979) but the calculations of the test are outlined in this section. The test is based on the residuals $e_i^{(T)}$ from an OLS on the whole sample (calculated as in equation (3.12)).

Let

$$D_{ij} = sgn(e_i^{(T)} - e_j^{(T)}) \quad \textbf{(3.22)}$$

Where

$$sgn(x) = \begin{cases} 1 \; if \; x > 0 \\ 0 \; if \; x = 0 \\ -1 \; if \; x < 0 \end{cases} \quad \textbf{(3.23)}$$

Then

$$U_{t,T} = \sum_{i=1}^{t} \sum_{j=t+1}^{T} D_{ij} \quad \textbf{(3.24)}$$

$U_{t,T}$ in equation (3.24) is sometimes easier to calculate using the equivalent formula

$$U_{t,T} = 2W_t - t(T+1) \quad \textbf{(3.25)}$$

Where

$$W_t = \sum_{j=1}^{t} R_j \quad \textbf{(3.26)}$$

And R_j is the rank of the residual $e_j^{(T)}$.

Then the test statistic is calculated as

$$MW = \max_{1 \le t \le T} |U_{t,T}| \quad \textbf{(3.27)}$$

The significance level is given by Pettitt (1979) to be

$$p_{OA} = 2 \sum_{r=1}^{\infty} (-1)^{r+1} \exp\{-\frac{6kr^2}{T^3 + T^2}\} \cong 2\exp\{-\frac{6k^2}{T^3 + T^2}\} \quad \textbf{(3.28)}$$

Where $k = MW$and the approximation works well, accurate for two decimal places, for $p_{OA} \le 0.5$

The statistic is illustrated with an example time series in Figure 6.

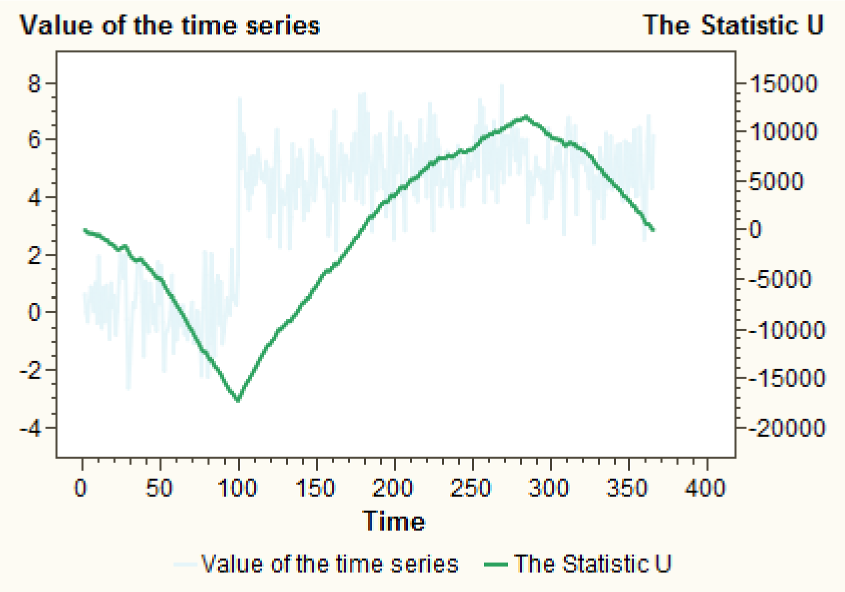


Figure 6 – A time series with a break and the statistic $U_{t,T}$ from the MW method

3.6 Ensemble method

The purpose of the ensemble method is to remove time dependency in observations by drawing random subsamples iteratively and then combining the results. The algorithm used in the method is described in detail by Alippi et al. (2013). In this study a simplified version is tested. It is based on the sequence of residuals $R = \{e_t^{(T)}, t = 1, \ldots, t\}$ from an OLS on the whole sample (calculated as in equation (3.12)).

1) Draw (without replacement) n random observations from R, denoted by $I_n^{(i)}$
2) Apply a method for finding a change-point in $I_n^{(i)}$
3) If a change-point is found, add it to the list of found change-points $\{M_i\}$
4) Repeat from 1) d times
5) Apply the method for finding a change-point in R
6) If a change-point is found, add it to the list of found change-points $\{M_i\}$
7) The change-point is the (weighted) average of all discovered change-points in $\{M_i\}$

In this study the change-point detection method in the ensemble method is the MW method. The random sampling parameter $n = \frac{T}{2}$ and the number of individual estimates $d = 100$. In step 7 the average is taken to be the arithmetic mean with all change-points having equal weights.

3.7 Combined method

In an attempt to improve the performance of the individual methods a combined method is examined. The main reason for the construction of this method is to reduce the number of false positives. By combining several methods the probability of multiple false positives coinciding is significantly reduced.

The three best performing methods are used to generate change-points. Then the combined method identifies change-points only if multiple methods have detected the same point. Depending on the preferences of the analysis, the critical value is for two methods to identify the same change-point or for all three methods to identify the same change-point.

That the combined method reduces the number of false positives can be explained by calculating the probability of random subsamples overlapping.

Let the sizes of the subsamples be n_1, n_2, n_3 and the entire sample size be N

The probability of k triple overlapping points when drawing 3 subsamples is

$$P(X = k) = \sum_{i=k}^{\min(n_1,n_2)} \frac{\binom{n_1}{i}\binom{N-n_1}{n_2-i}}{\binom{N}{n_2}} * \frac{\binom{i}{k}\binom{N-i}{n_3-k}}{\binom{N}{n_3}} \tag{3.29}$$

Which gives

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \sum_{i=0}^{\min(n_1,n_2)} \frac{\binom{n_1}{i}\binom{N-n_1}{n_2-i}}{\binom{N}{n_2}} * \frac{\binom{N-i}{n_3}}{\binom{N}{n_3}} \tag{3.30}$$

Since N usually is much larger than any of n_1, n_2, n_3 the probability of points identified as change-points using the triple change-point criterion is very low. And this is assuming that all points are chosen randomly, in reality the individual methods are constructed to result in very few false positives (1 % error risk) which means that the probability of false positives is reduced further.

The requirement for triple overlapping points can be relaxed to double overlapping points. Then the probability of k (at least) double overlapping points when drawing 3 subsamples is

$$P(X = k) = \sum_{i=\max(0,k-n_3)}^{\min(n_1,n_2,k)} \frac{\binom{n_1}{i}\binom{N-n_1}{n_2-i}}{\binom{N}{n_2}} * \frac{\binom{n_1+n_2-2i}{k-i}\binom{N-(n_1+n_2-2i)}{n_3-(k-i)}}{\binom{N}{n_3}} \tag{3.31}$$

Which gives

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{\binom{N-n_1}{n_2}}{\binom{N}{n_2}} * \frac{\binom{N-(n_1+n_2)}{n_3}}{\binom{N}{n_3}} \tag{3.32}$$

The details of the calculations of the probabilities and plots for some examples of sample sizes can be seen in Appendix A.

3.8 Data

3.8.1 Real life data

Real life data is provided by the Bank and consists of anonymized and scaled daily balances for all Swedish accounts at the Bank. The maximum length of the time series is between 26 November 2012 and 13 Jan 2014 (414 days), but accounts opened and closed in that time period are also included. The data is mainly used to extract properties of real life data in order to generate artificial data with similar properties.

3.8.2 Generated data

To evaluate the performance of the methods, artificial data are generated. This data have known properties that make it easy to judge how well the methods are performing. For each set of different distributed data 1000 time series are generated with 365 observations (which correspond to daily observations over a year).

First data from a normal distribution is generated. This is in order to test the performance of the methods under ideal conditions and provide a baseline for further tests. The first tests are performed on data with a single structural break with varying size. This will hence illustrate the performance of the methods on a single change-point and how the size of the shift affects the performance.

Secondly normal distributed data with two structural breaks is tested. This will illustrate how well the methods perform on data with multiple breaks. In this stage the location of the change-points will be varied to extract the performance of the methods on differently located change-points.

The next stage of the study is performed on data with different distribution than a normal distribution. Some heavy tail distributions are investigated, including the student's t-distribution and the Cauchy distribution. Data from a uniform distribution is also tested. The results show how well the methods perform on data that is not following the assumptions of normal distribution.

Lastly data generated from an autoregressive model with lag 1, AR(1), is used to test how well the methods are performing on data that is not i.i.d.

3.9 Evaluation of the methods

The analysis on the data, both real life and generated, are performed using the program SAS (Statistical Analysis System).

All methods are first tested on the artificially generated data. Since the true locations of the change-points are known it is easy to evaluate the performance of the methods both regarding the number of correctly identified change-points and the spread of falsely identified change-points. The ensemble method is also tested to see if it is able to improve the performance of the regular MW method.

The performance of the methods is evaluated according to two criteria, the number of correctly identified change-points (true positives) and the number of incorrectly identified change-points (false positives). A good method produces a large amount of true positives and a low amount of false positives. A low amount of false positives is the most important property in this study. This is because wrongly identified change-points create misinformation about the properties of the time series while missed true positives only limits the amount of information available. The sooner is more severe than the latter, at least in this study.

The binary segmentation technique is used for each method, even on data with a single change-point. This is because in reality the number of change-points is unknown and it is of interest to know how the binary segmentation technique performs on data where it should not be needed.

When the individual methods have been evaluated the combined method is tested in a similar way. The different criteria are evaluated to see if they are able to improve the performance of the individual methods.

As a final evaluation, the analysis is performed on the real life data provided by the Bank with unknown properties and locations of the change-points. This will give an indication on how well the methods perform on data with the complex properties and the unpredictability that real life data has. However, since the true properties of the data are unknown the results from these tests can only be of an indicative nature.

4 Results

This chapter begins with an investigation on how the data from the Bank is distributed. It is followed by the results of how the methods are performing on generated data with different (known) properties. Furthermore the performance of the combined method is evaluated. The chapter ends with an evaluation on how well the methods perform on real life data.

4.1 The distributions of the residuals

In this section the properties of the data from the Bank are investigated. It should be noted that the data contains structural breaks while the analysis assumes homogenous data. That is because it is impossible to in advance locate the change-points and make appropriate corrections. The results from this section should hence only be used indicatively and not as a confirmation about the true properties of the data.

First the distribution of the residuals is examined by producing Q-Q plots. Figure 7 shows a Q-Q plot of residuals from an OLS regression (in relation to the value of the time series) on 1,000,000 accounts versus a normal distribution. The aggregated result is resting on the assumption that the residuals from all accounts have the same distribution, which is a somewhat questionable assumption. But it gives an overview of how the distribution of the residuals could look.

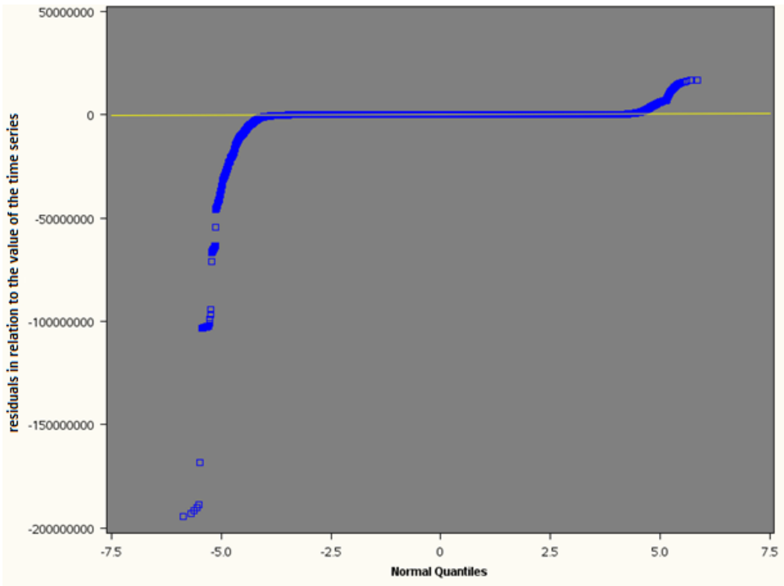


Figure 7 – Q-Q plot of the residuals versus a normal distribution

As Figure 7 shows, the data does not seem to come from a normal distribution. The most obvious reason for this is that the tails are much heavier for the data than for the normal distribution. However, those heavy tails could also be seen as outlier that does not fit the general distribution of the data. Hence a Q-Q plot is produced where residuals larger than two times the value of the time series are removed. This is presented in Figure 8.

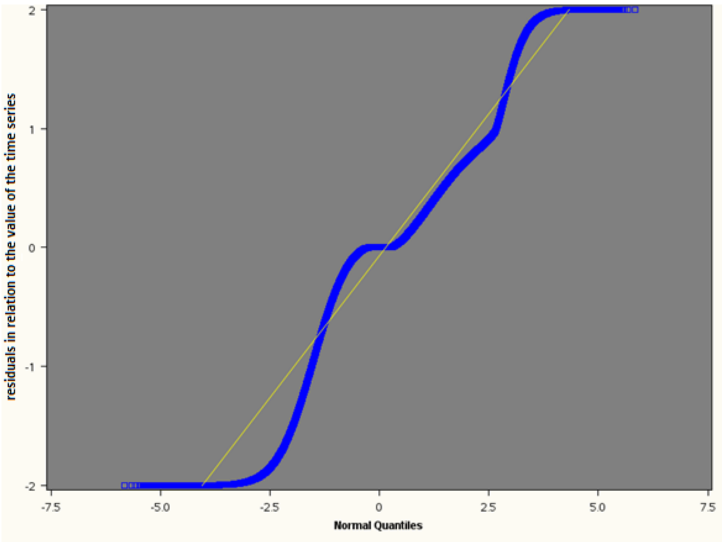


Figure 8 – Q-Q plot of the cropped residuals versus a normal distribution

In Figure 8 the behavior of the data in the center is more clearly illustrated. As can be seen, the tails can still be considered heavy, but it can also be observed that the data does not follow the normal distribution in the middle either.

The autocorrelation of the data is investigated next to see how well motivated the assumption of i.i.d. seems to be. An initial inspection shows that the most accounts have time dependence. Some examples are presented in Appendix B . To investigate the AR lag for all real life time series the partial autocorrelation function is used on roughly 1,000,000 accounts. The histogram over the identified lags is presented in Figure 9.

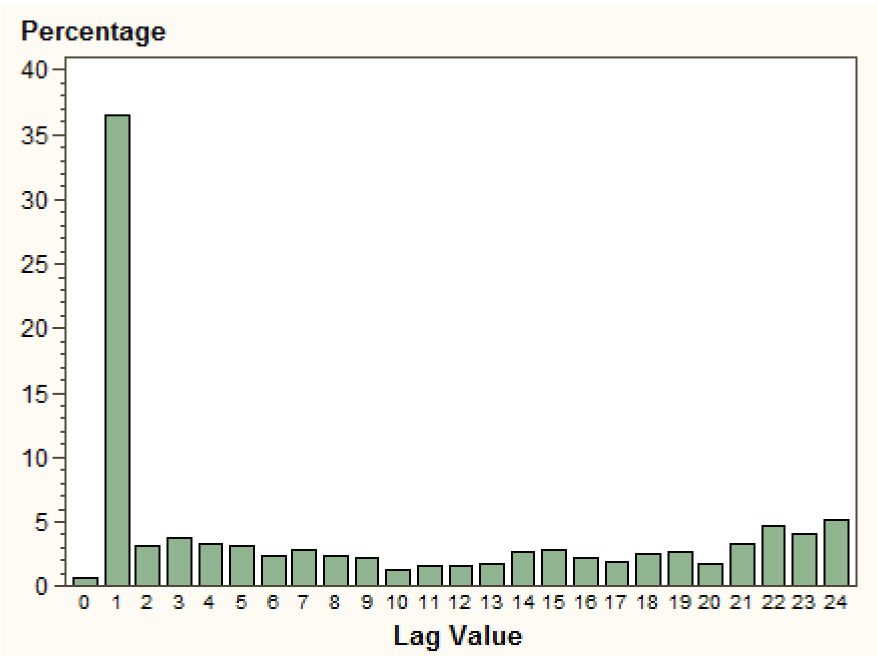


Figure 9 – Identified lags of the AR-process using the partial autocorrelation function on roughly 1,000,000 accounts

Figure 9 shows that for most of the accounts the data seem to come from an AR(1)-process. It can also be noted that lag 0 has the lowest frequency meaning that the i.i.d. assumption does not seem to hold.

The results from the initial analysis is hence that the residuals of the real data comes from a heavier tailed distribution than a normal distribution and that the assumption of i.i.d. is questionable, with an AR(1)-process being more likely. These results could come from the fact that the time series contains structural breaks but are still used when generating the data that the methods are tested on.

4.2 Evaluating the individual methods

The analysis is performed using the methods CUSUM, CUSUMSQ, OLSCUSUM, Quandt's log likelihood ratio and MW on generated test data. The binary segmentation technique is used to be able to identify multiple change-points. The iteration limit is set to two which mean that the maximal identified change-points are three. This is because the number of change-points in these tests is known to be at most two. Since the number of false positives is an important factor the significance level is set to 1 %.

4.2.1 Normal distribution

4.2.1.1 Single change-point

The first test is constructed to investigate the performance of the individual methods and how the size of the structural break affects the ability of locating the change-point. The test data used have only one change-point to keep the results clear and easy to interpret.

A set of 1000 time series consisting of 365 observations (with a change-point at 99) are generated as follows:

$$\begin{cases} X_i \in N(0,1), & 0 \le i \le 99 \\ X_i \in N(j,1), & 99 < i \le 365 \end{cases}, j = 0,1,2,\dots,8 \tag{4.1}$$

Where $N(\mu, \sigma)$ denotes a normal distribution with expected value μ and standard deviation σ.[3]

Each of the tested methods is then used respectively on the generated data. To see how the identified change-points are distributed the figures present three levels of identified change-points:

1) All identified change-points (false positives and true positives)
2) Identified change-points within 3 observations from the true value (almost true positives)
3) Correctly identified change-points (true positives)

This will hence demonstrate how well the methods perform regarding both true positives and false positives. Level 2) is included to get an overview of the distribution of the false positives and whether they lie close to the true value or not.

A good performance is for the bars in each category to be of equal height (meaning that all identified change-points are at the true location) and of height one (meaning all change-points are identified). The exception is for the shift of size 0 where no change-point should be identified.

[3] The probability density function of $X \in N(\mu, \sigma)$ is $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for $-\infty < x < \infty$

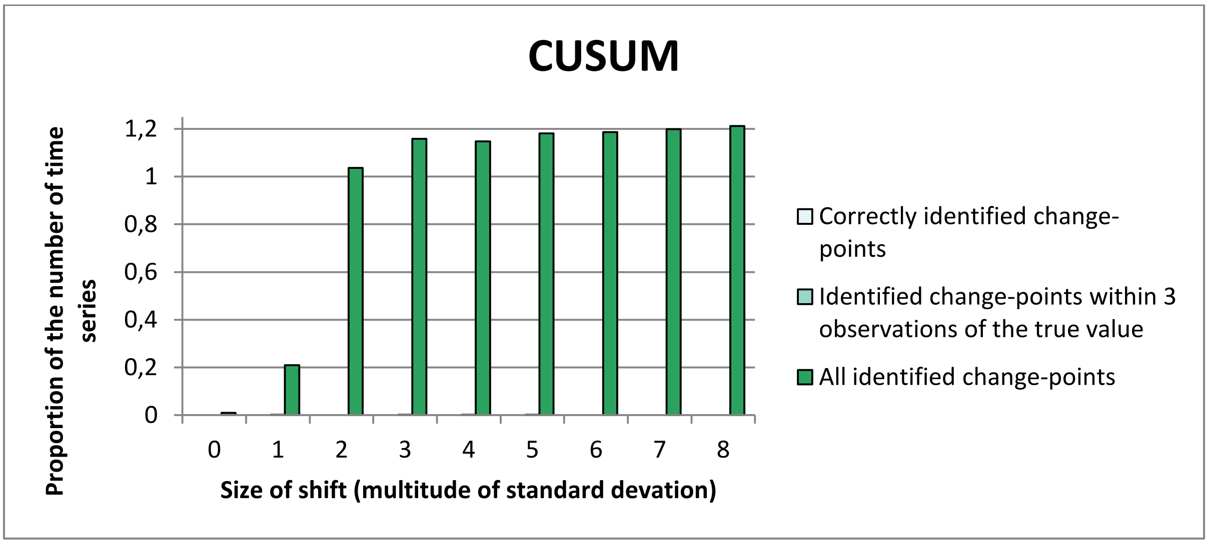


Figure 10 – Identified change-points using the CUSUM method on normally distributed data with varying shifts

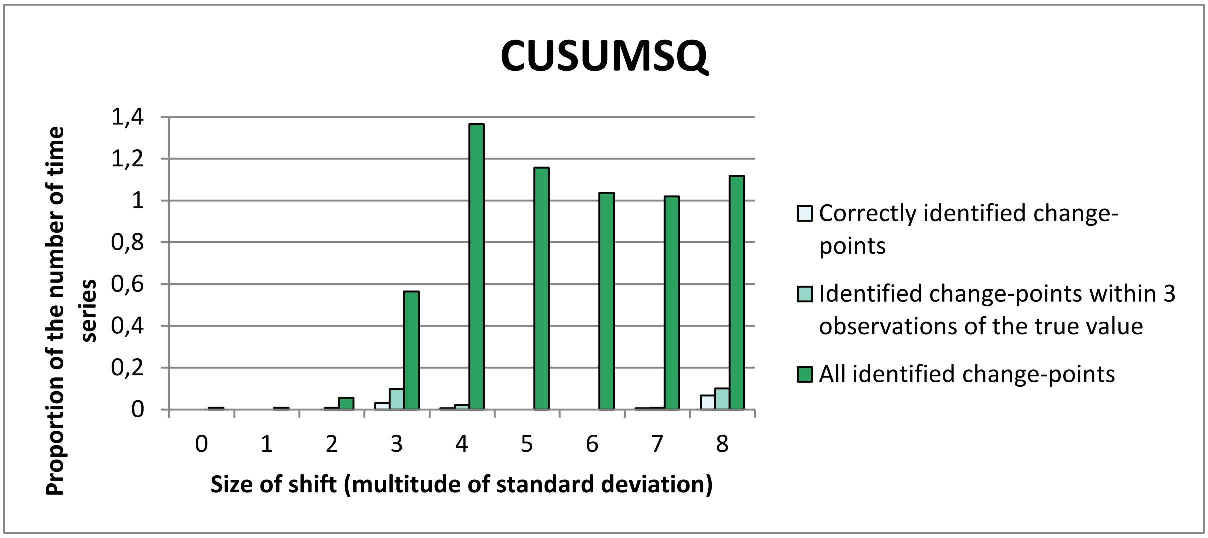


Figure 11 – Identified change-points using the CUSUMSQ method on normally distributed data with varying shifts

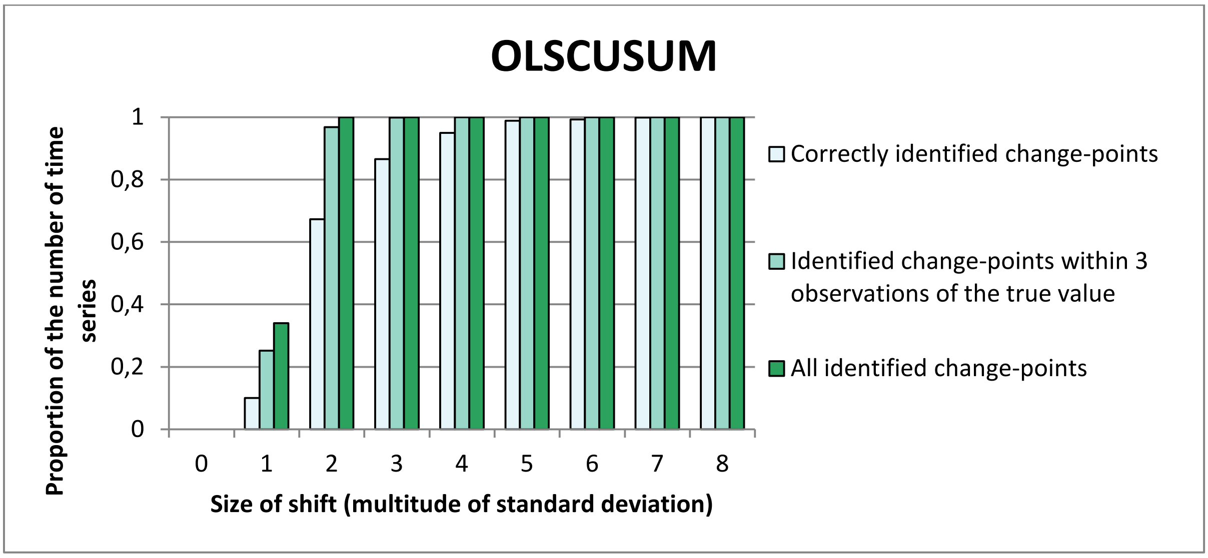


Figure 12 – Identified change-points using the OLSCUSUM method on normally distributed data with varying shifts

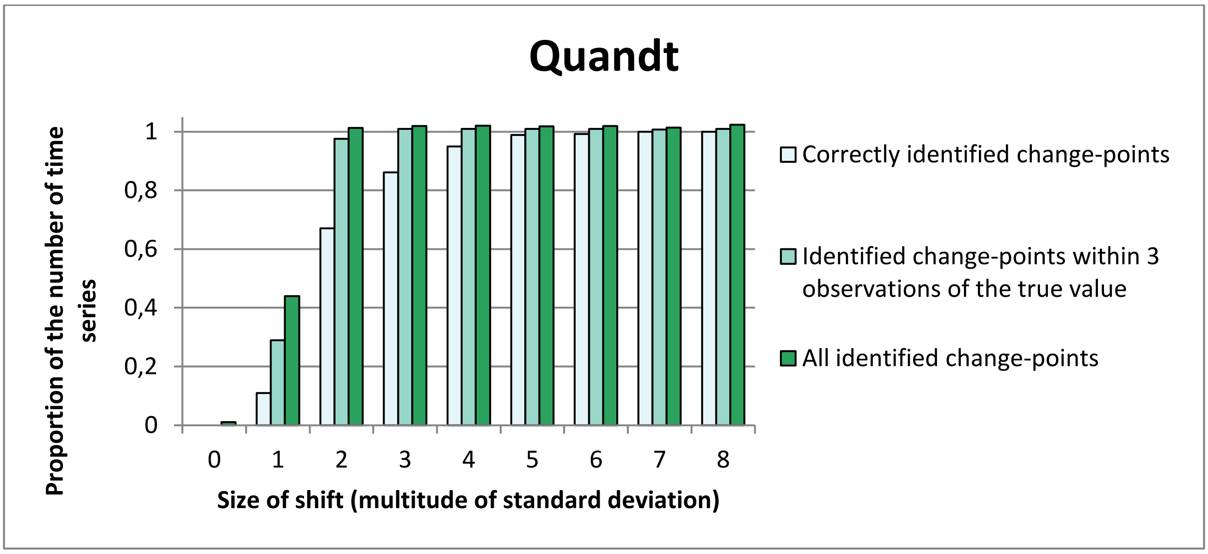


Figure 13 – Identified change-points using the Quandt method on normally distributed data with varying shifts

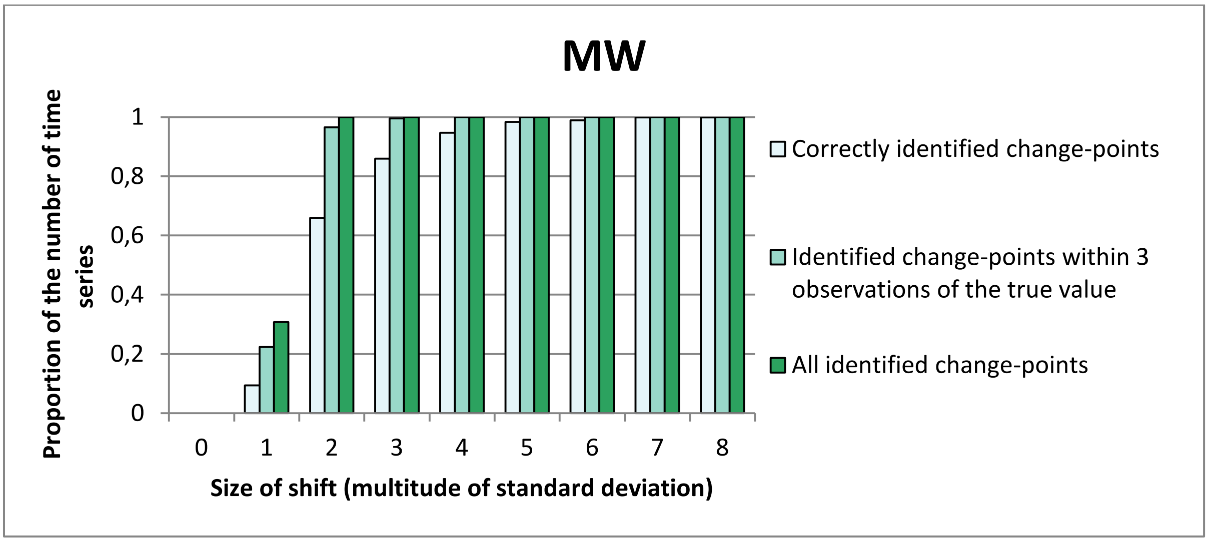


Figure 14 – Identified change-points using the MW method on normally distributed data with varying shifts

As can be seen in Figure 10-Figure 14 the methods OLSCUSUM, Quandt and MW are outperforming the CUSUM and CUSUMSQ test using both performance criteria (number of true positives and number of false positives).

All methods are producing a small amount of false positives when there is no change-point in the data (where the size of the shift is 0). This means that all methods perform as expected when used on data with no structural break and no method is superior when the null hypothesis of no shifts is true.

For small breaks (1 standard deviation) no method is able to detect a large amount of the true change-points. As the shift increases the number of identified change-points is increasing and the differences between the methods become more apparent. The OLSCUSUM, Quandt and MW methods have a large percentage of true positives compared to false positives and for shifts larger than 8 standard deviations the methods are able to detect all change-points without any false positives. The CUSUM and CUSUMSQ are not able to detect the location of the change-points.

4.2.1.2 Multiple change-points

Next, the performance of the methods on data with two change-points is investigated. The data is generated as follows.

$$\begin{cases} X_i \in N(0,1), & 0 \le i \le t_0 \\ X_i \in N(2,1), & t_0 < i \le t_1 \\ X_i \in N(-1,1), & t_1 < i \le 365 \end{cases} \tag{4.2}$$

Where t_0 and t_1 are the true change-points. The time series are generated in sets with different values for t_0 and t_1 to examine how well the methods are performing depending on the location of the change-points. 1000 time series are produced for each configuration of change-points.

$t_0 = 99, t_1 = 199$

First two change-points that are relatively evenly spaced are investigated. This aims to isolate the multiple change-points from interference effects of the change-points and the edges. All methods are used on the test data and the identified change-points are plotted in histograms. The histograms have a discrete x-axis which means that only points that are observed are on the axis. The axis is hence not linear. The reason for this is to improve the visibility of the plots. The true locations of the change-points are marked with a lighter colour (note that Figure 15 – Identified change-points using the CUSUM method on normally distributed data with shifts at 99 and 199Figure 15 does not have any light coloured bar). A method is considered to perform well if the lighter coloured bars are close to 1.0 (which means that all true change-points are located) and the rest of the bars are close to zero (which means that no false change-points are identified).

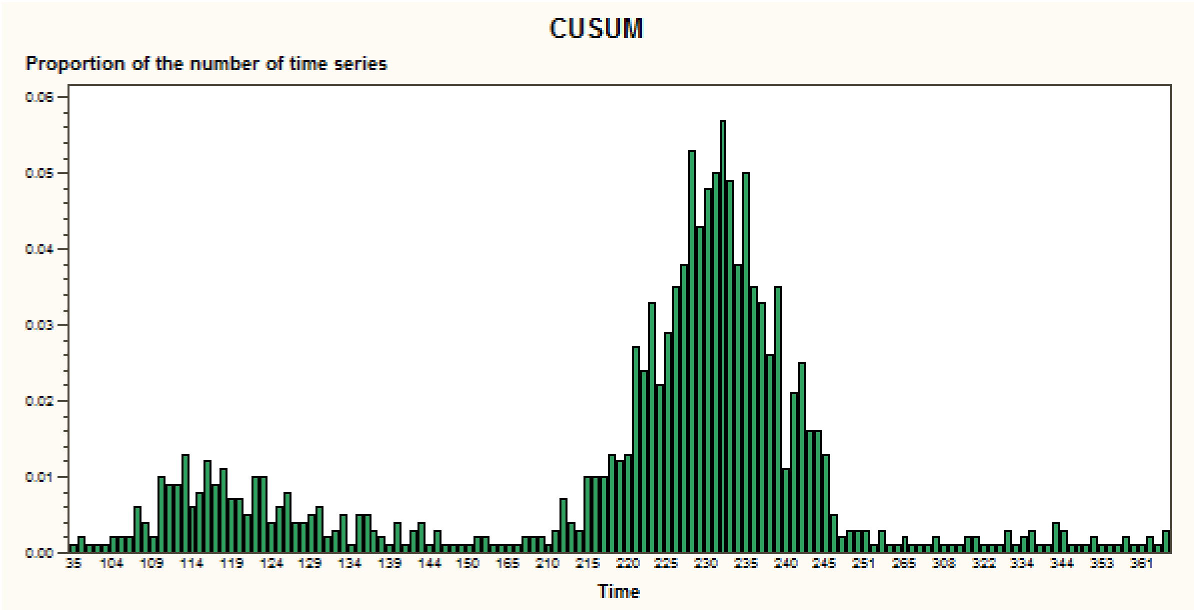


Figure 15 – Identified change-points using the CUSUM method on normally distributed data with shifts at 99 and 199

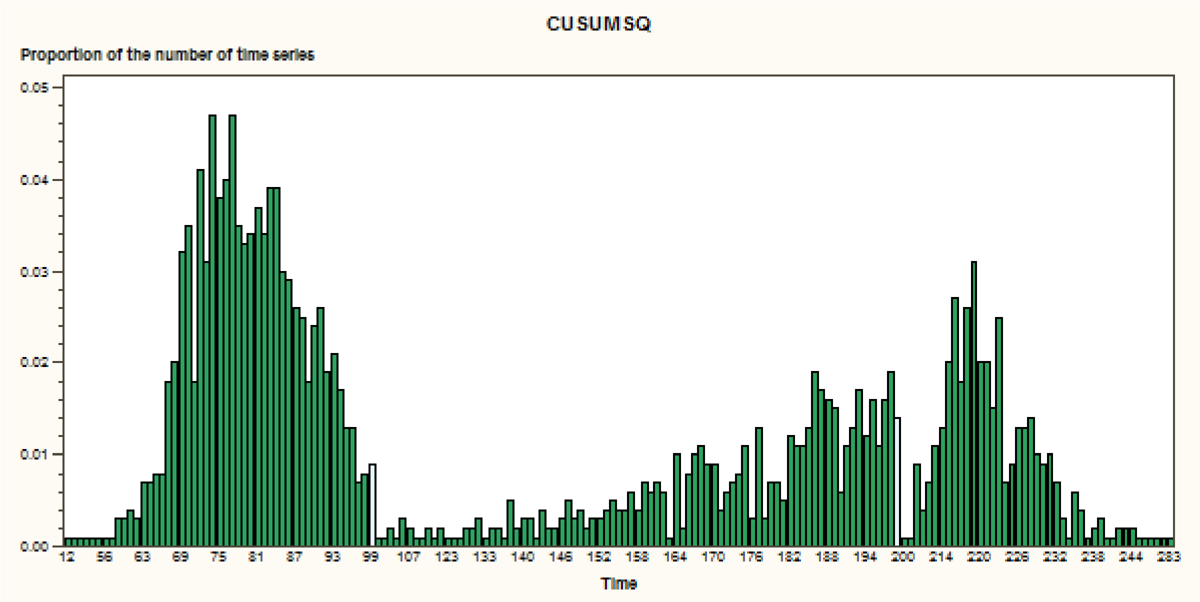


Figure 16 – Identified change-points using the CUSUMSQ method on normally distributed data with shifts at 99 and 199

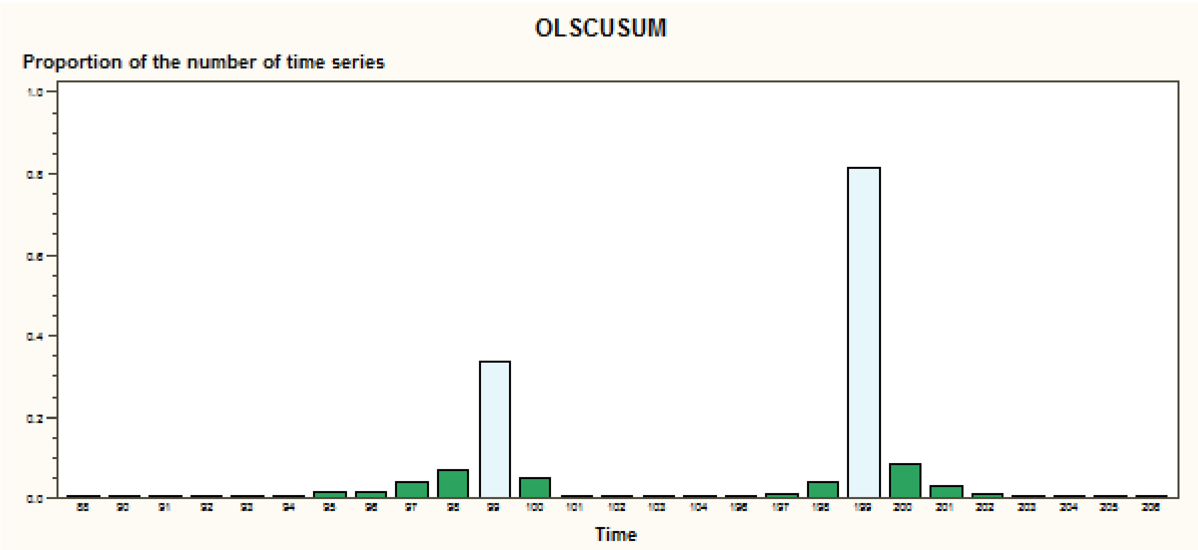


Figure 17 – Identified change-points using the OLSCUSUM method on normally distributed data with shifts at 99 and 199

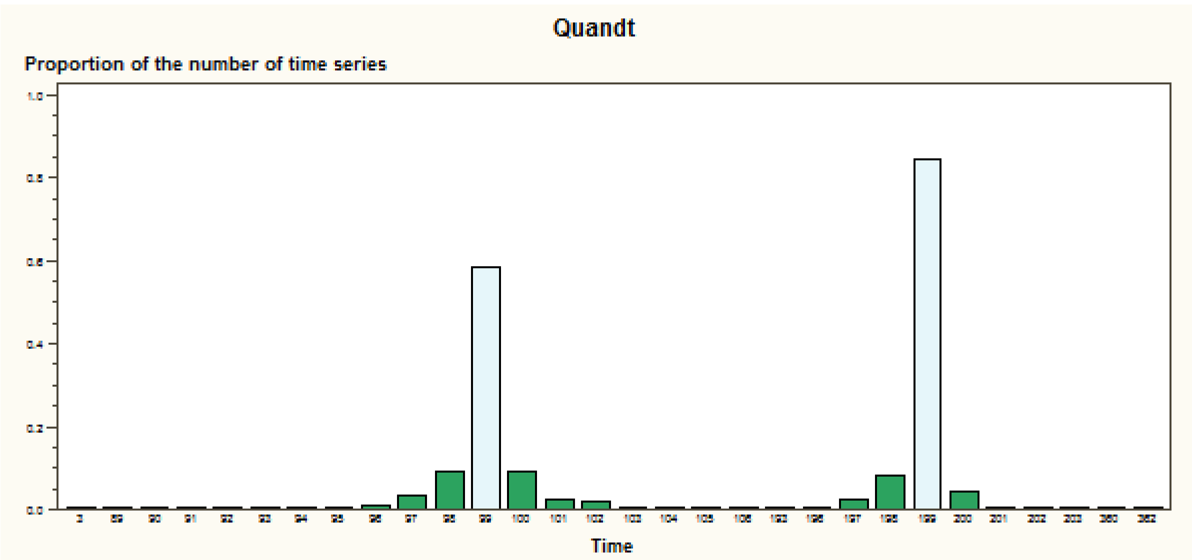


Figure 18 – Identified change-points using the Quandt method on normally distributed data with shifts at 99 and 199

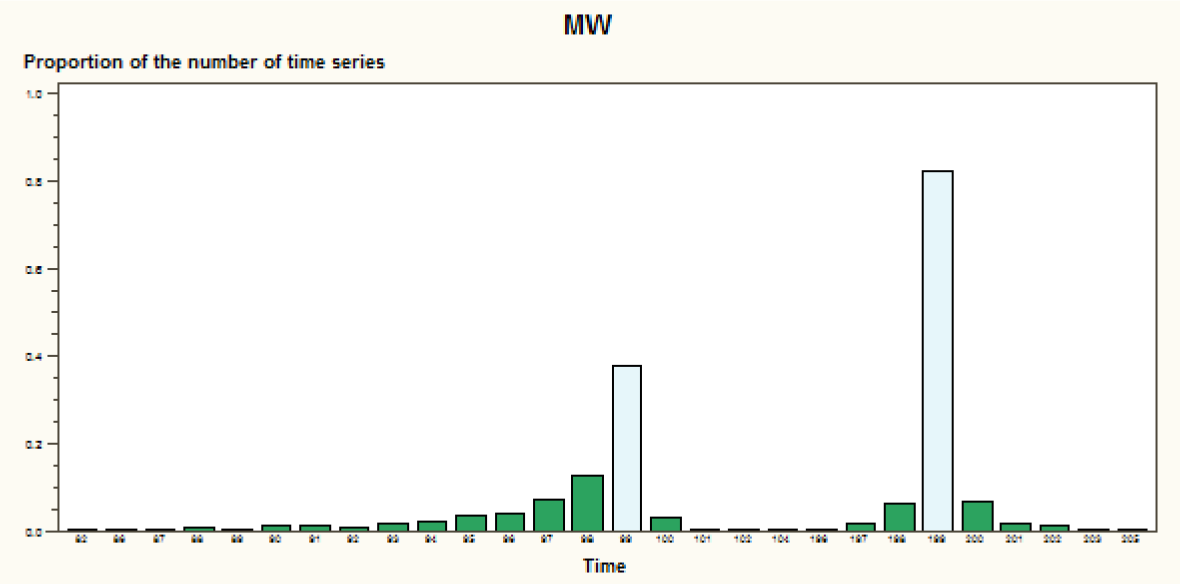


Figure 19 – Identified change-points using the MW method on normally distributed data with shifts at 99 and 199

As can be seen in Figure 15-Figure 19 the OLSCUSUM, Quandt and MW methods are producing the best results according to both performance criteria. There are a high percentage of true positives with a smaller portion of false positives which are closely spread around the true change-point. The CUSUM and CUSUMSQ are not able to locate the change-points and instead are much more spread out. From the histograms it can be observed that the change-point locations identified by the CUSUM method are lagging the true value. Because of the large variance and inability to produce true positives in close to ideal circumstances the CUSUM and CUSUMSQ are deemed to not perform well at locating change-points and are not investigated further.

The ensemble method is a method that aims to be used on time dependent time series to improve the performance of the methods. To see how well it performs on time series with i.i.d. data and with multiple change-points it is tested on the same data as the individual methods. Since the ensemble method is used with the MW method it should be compared with the regular MW method. The performance of the ensemble method is presented in Figure 20.

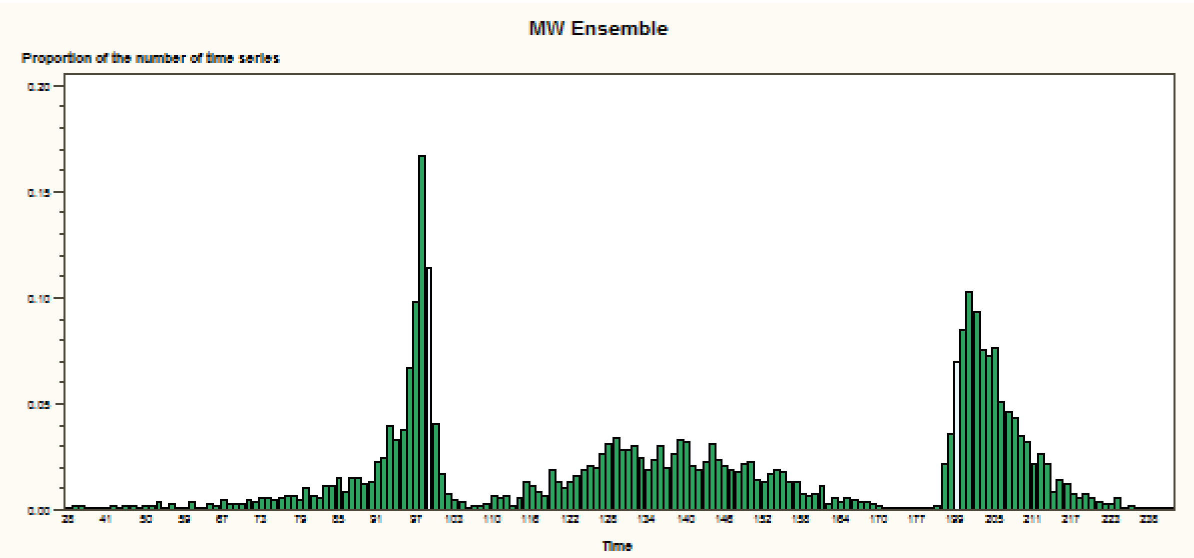


Figure 20 – Identified change-points using the MW Ensemble method on normally distributed data with shifts at 99 and 199

In Figure 20 it can be seen that the MW Ensemble method is not performing as well as the regular MW method (Figure 19). In Figure 20 the largest spikes are not at the true locations of the change-points. Also, the identified change-points are more spread out which means that the uncertainty of the results from the method is larger. It can furthermore be observed that there is an increased probability of detecting change-points in the middle of the true change-points. This result together with the inaccuracy leads to the conclusion that the ensemble method is not working well when multiple change-points are present in the data and the method is not investigated further in this study.

$t_0 = 9, t_1 = 357$

1000 new time series are generated as in (4.2) with true change-points close to the beginning and end of the time series. This is then used to see how well the methods perform on data with change-points on the edges of the data and if it matters whether the change-point is in the beginning or the end. Neither the OLSCUSUM nor the MW method manages to locate a single change-point. The Quandt method works better and the results are presented in a histogram (with a discrete x-axis).

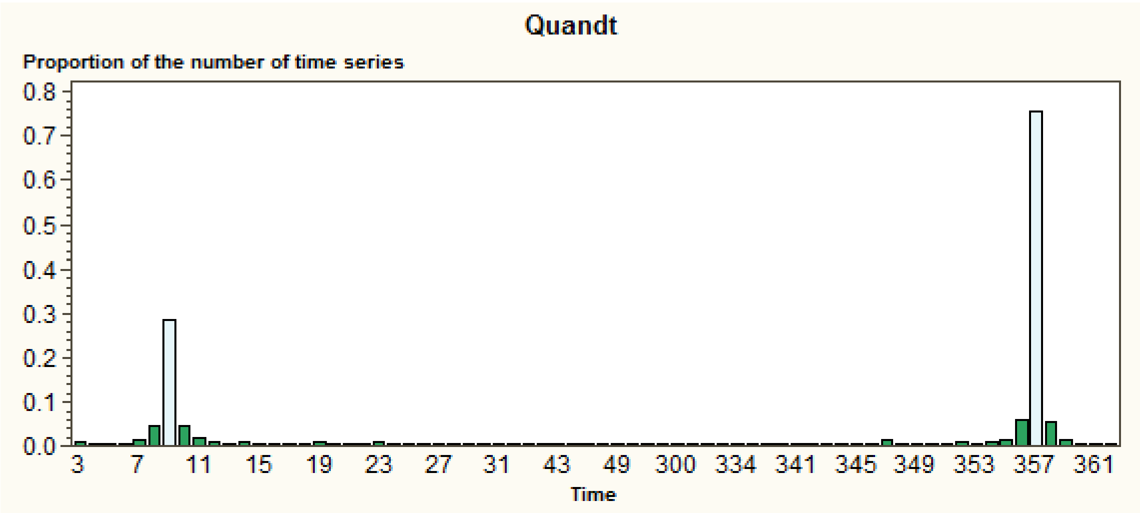


Figure 21 – Identified change-points using the Quandt method on normally distributed data with shifts at 9 and 357

As these results demonstrate, the OLSCUSUM and MW method are performing poorly when the true change-points lie close to the edges of the time series. Only the Quandt method is able to produce a similar histogram as for the previous data (Figure 21) and manages to identify both change-points with sufficiently large accuracy.

$t_0=99, t_1=109$

To investigate if there is an interference effect of change-points that lie close to each other 1000 time series with change-points closely located are generated as in (4.2). The results are presented in Figure 22-Figure 24.

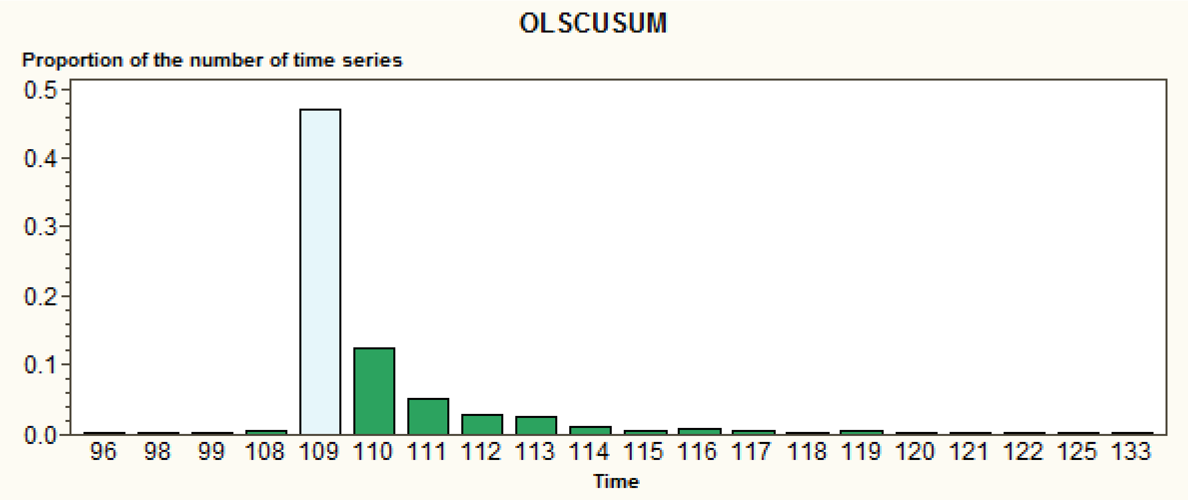


Figure 22 – Identified change-points using the OLSCUSUM method on normally distributed data with shifts at 99 and 109

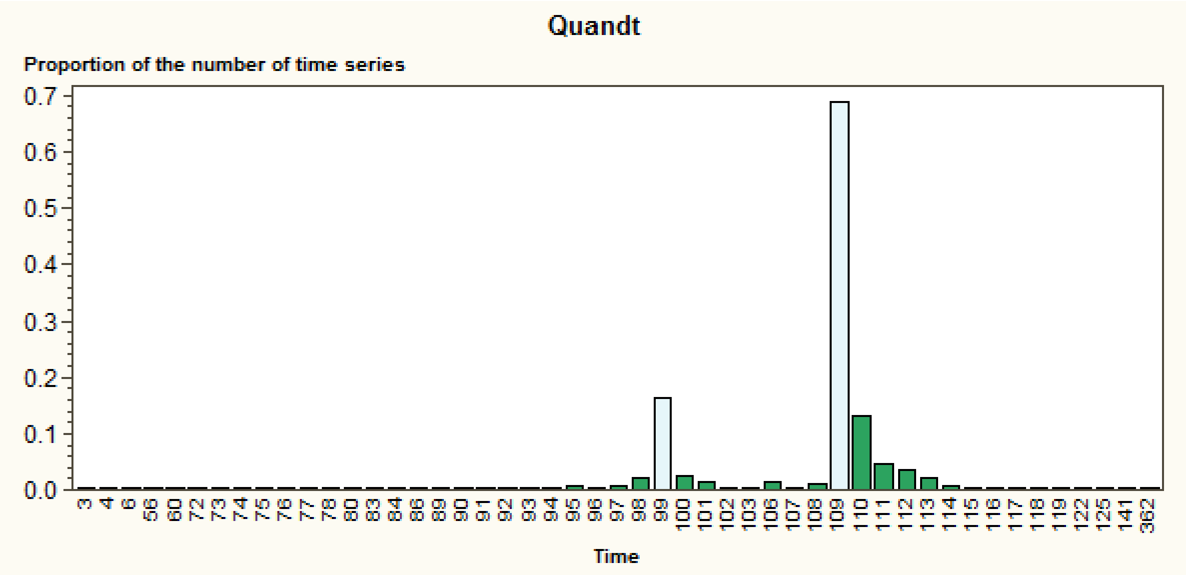


Figure 23 – Identified change-points using the Quandt method on normally distributed data with shifts at 99 and 109

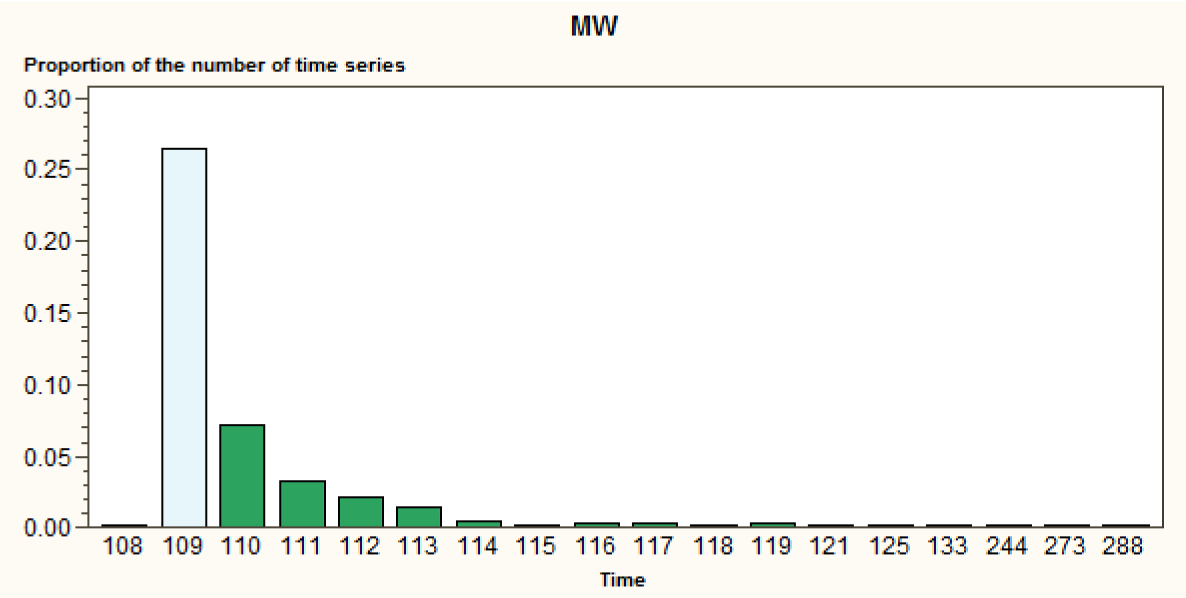


Figure 24 – Identified change-points using the MW method on normally distributed data with shifts at 99 and 109

As can be seen in the results, the MW method is only able to detect the last (and largest) of the two change-points (Figure 24). The OLSCUSUM performs similarly but with a larger amount of true positives (Figure 22). The only method that reliably identifies both change-points is the Quandt method (Figure 23).

4.2.2 Student's t-distribution

Data generated from a Student's t-distribution are used next to evaluate the methods. The sizes of the shifts are chosen to correspond to the magnitude of the shift for the normal distribution. The true change-points are put in the middle of the time series and relatively evenly spaced at 99 and 199. 1000 time series are generated as follows.

$$\begin{cases} X_i \in t(3), & 0 \le i \le 99 \\ X_i \in t(3) + 2\sqrt{3}, & 99 < i \le 199 \\ X_i \in t(3) - \sqrt{3}, & 199 < i \le 365 \end{cases} \qquad \textbf{(4.3)}$$

Where $t(\nu)$ denotes a student's t-distribution with ν degrees of freedom[4].

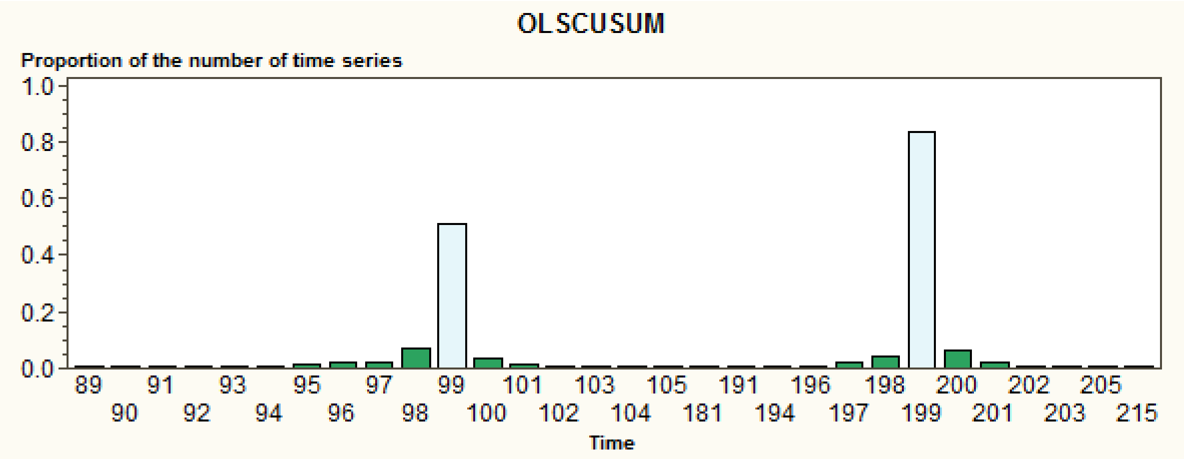


Figure 25 – Identified change-points using the OLSCUSUM method on Student's t-distributed data with shifts at 99 and 199

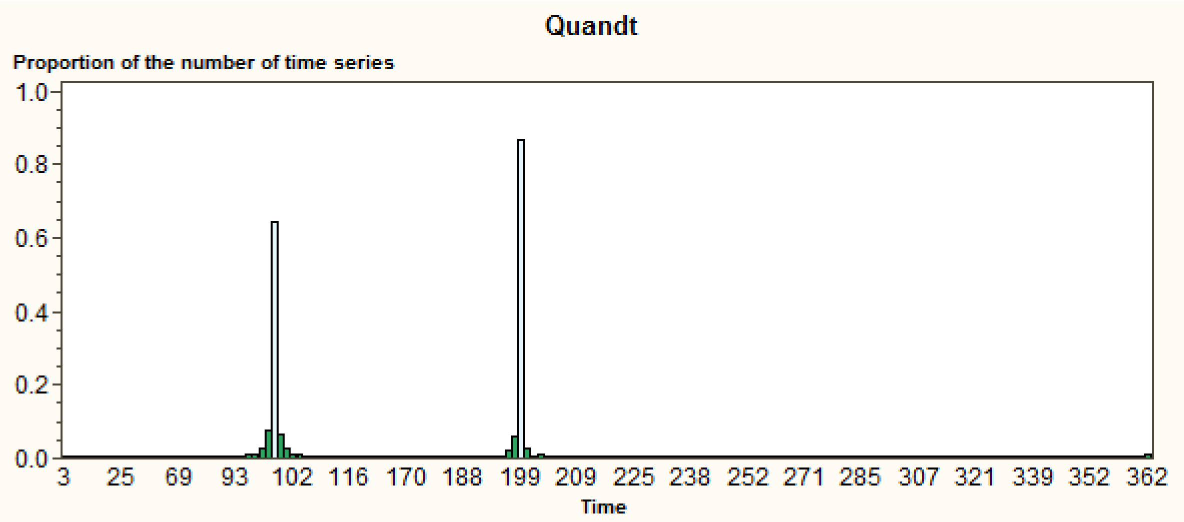


Figure 26 – Identified change-points using the Quandt method on Student's t-distributed data with shifts at 99 and 199

[4] The probability density function of $X \in t(\nu)$ is $f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$ for $-\infty < x < \infty$, where $\Gamma(.)$ is the gamma function

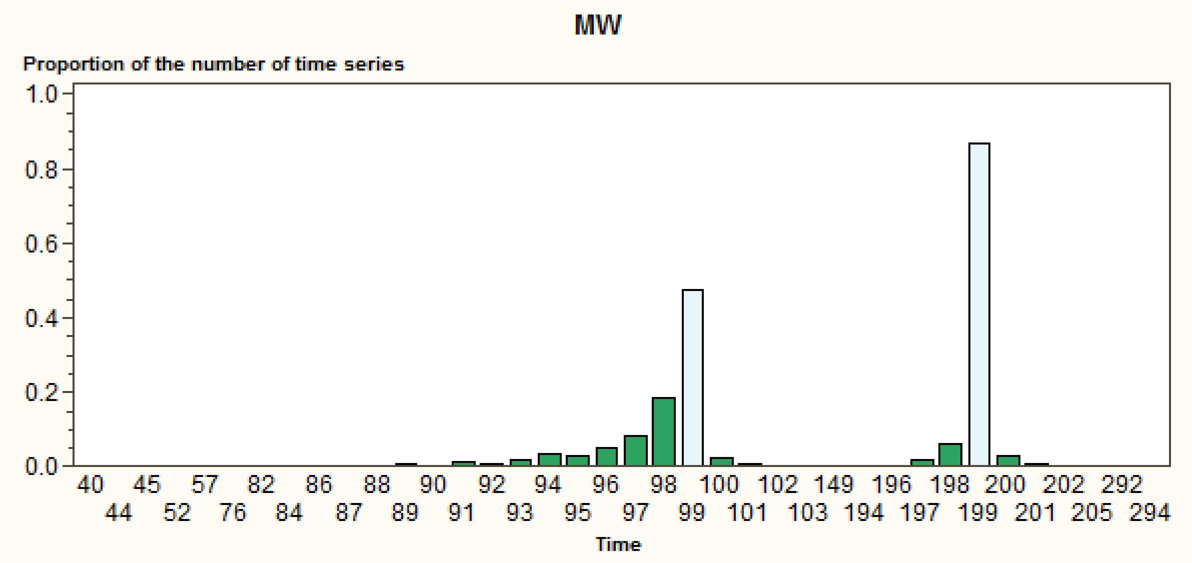


Figure 27 – Identified change-points using the MW method on Student's t-distributed data with shifts at 99 and 199

As can be seen in Figure 25-Figure 27 the methods seems to perform roughly the same as for the normally distributed data. Most of the detected change-points are at the correct location. The false positive change-points are spread around the true change-points.

4.2.3 Cauchy distribution

In this section the performance of the methods on Cauchy distributed data is investigated. Since the Cauchy distribution does not have any defined mean or variance, it is not possible to produce a shift of the same magnitude (in relation to the variance) as for the previous distributions, instead the same size of shift as for the student's t-distribution is chosen. The change-points are again set to be roughly evenly spaced at 99 and 199. 1000 time series are generated as follows.

$$\begin{cases} X_i \in Cauchy(0,0.5), & 0 \le i \le 99 \\ X_i \in Cauchy(2\sqrt{3},0.5), & 99 < i \le 199 \\ X_i \in Cauchy(-\sqrt{3},0.5), & 199 < i \le 365 \end{cases} \tag{4.4}$$

Where $Cauchy(x_0,\gamma)$ denotes a Cauchy distribution with location parameter x_0 and scale parameter γ.[5] The results are plotted in histograms.

[5] The probability density function of $X \in Cauchy(x_0,\gamma)$ is $f_X(x) = \frac{1}{\pi}\left[\frac{\gamma}{(x-x_0)^2+\gamma^2}\right]$ for $-\infty < x < \infty$

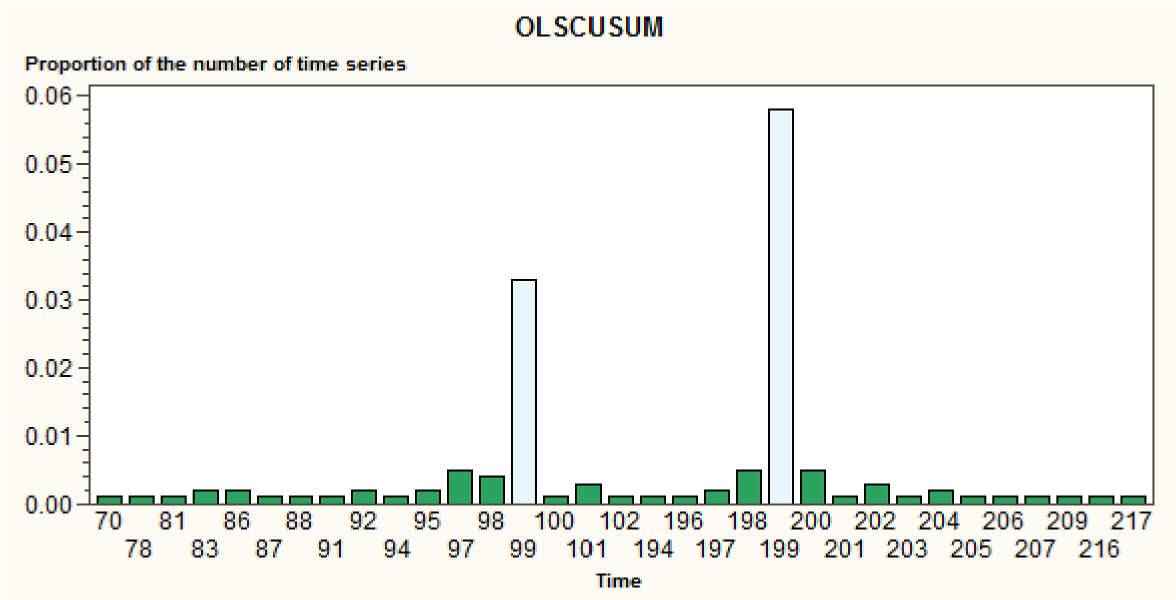


Figure 28 – Identified change-points using the OLSCUSUM method on Cauchy distributed data with shifts at 99 and 199

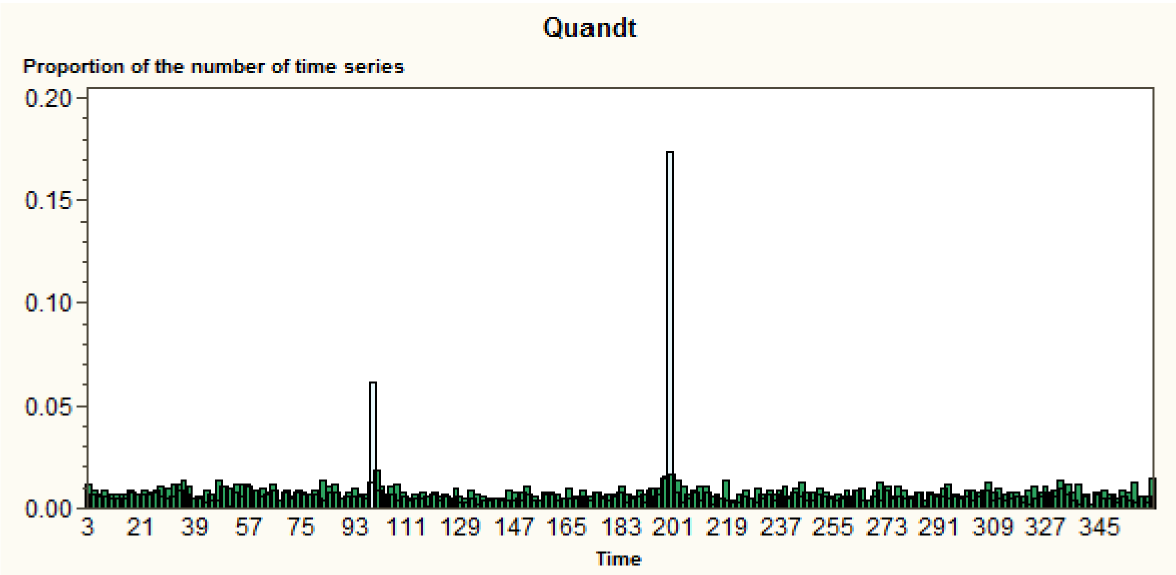


Figure 29 – Identified change-points using the Quandt method on Cauchy distributed data with shifts at 99 and 199

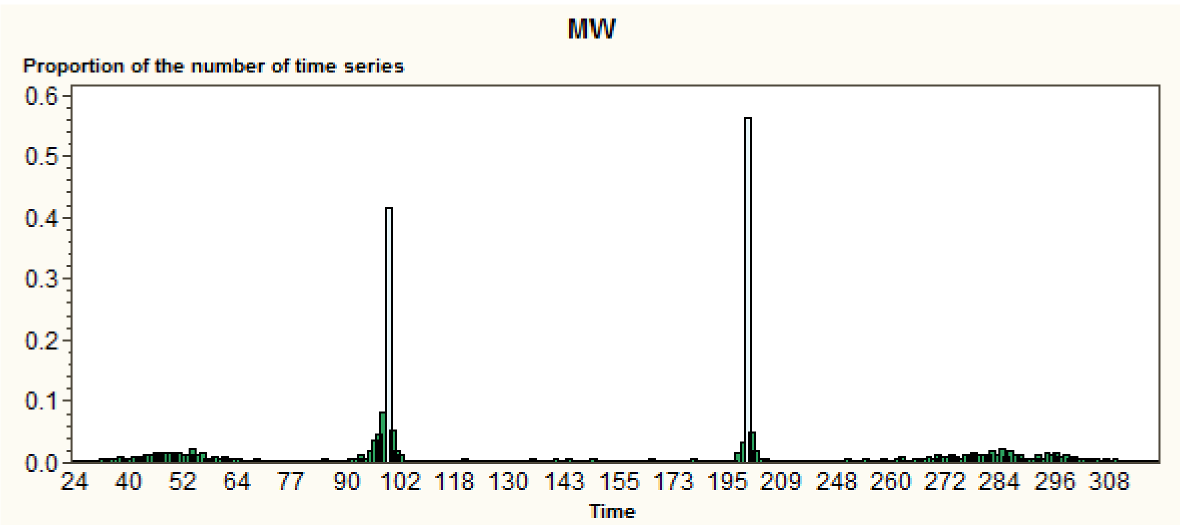


Figure 30 – Identified change-points using the MW method on Cauchy distributed data with shifts at 99 and 199

The efficiency of the methods is somewhat reduced compared to the distributions with lighter tails as can be seen in Figure 28-Figure 30. But for the OLSCUSUM and the MW method, the majority of the identified change-points are on the correct location or nearby. The MW method does however produce a larger amount of true positives. The method that is most affected by the heavy tailed distribution is the Quandt method (Figure 29) with a large spread of false positives over most of the interval.

4.2.4 Uniform distribution

Next, the methods are tested on data from a uniform distribution. As before the true change-points are at 99 and 199.

$$\begin{cases} X_i \in U(-4,4), & 0 \le i \le 99 \\ X_i \in U(-2,6), & 99 < i \le 199 \\ X_i \in U(-5,3), & 199 < i \le 365 \end{cases} \tag{4.5}$$

Where $U(a,b)$ denotes a (continuous) uniform distribution with minimum a and maximum b.[6]

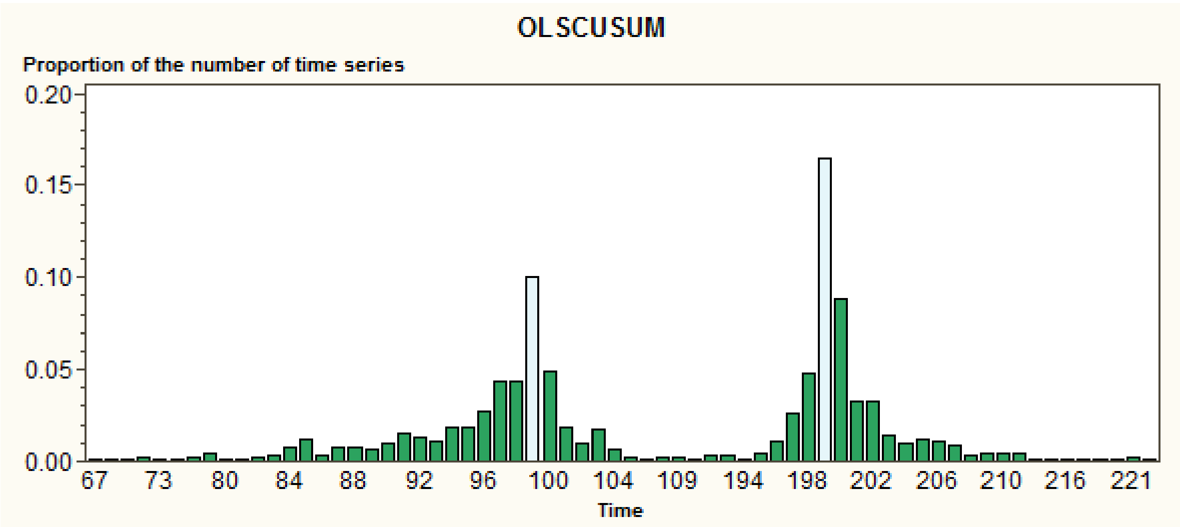


Figure 31 – Identified change-points using the OLSCUSUM method on uniformly distributed data with shifts at 99 and 199

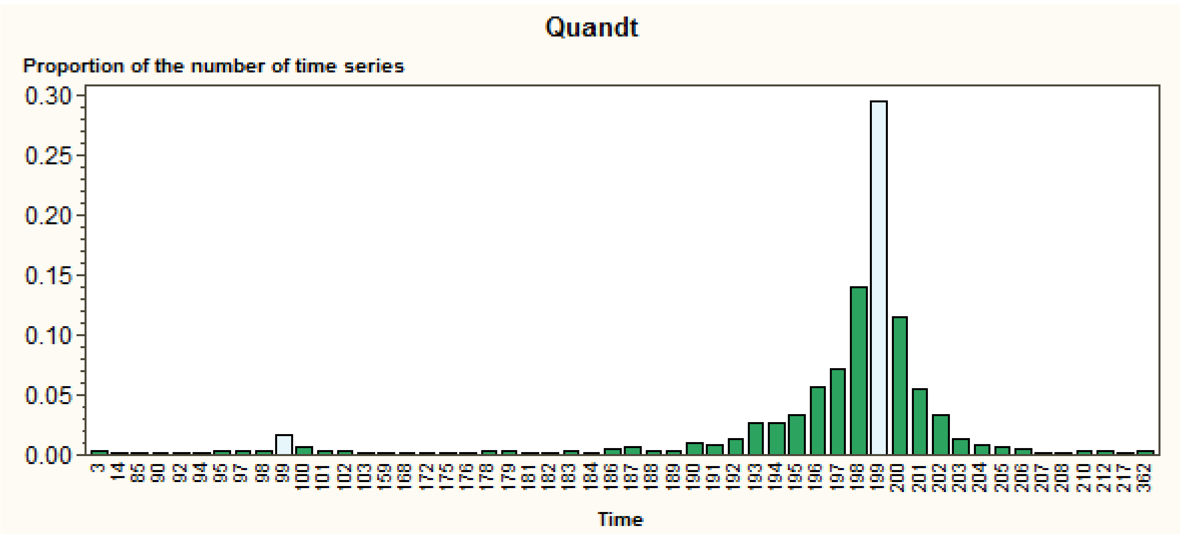


Figure 32 – Identified change-points using the Quandt method on uniformly distributed data with shifts at 99 and 199

[6] The probability density function of $X \in U(a,b)$ is $f_X(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & otherwise \end{cases}$

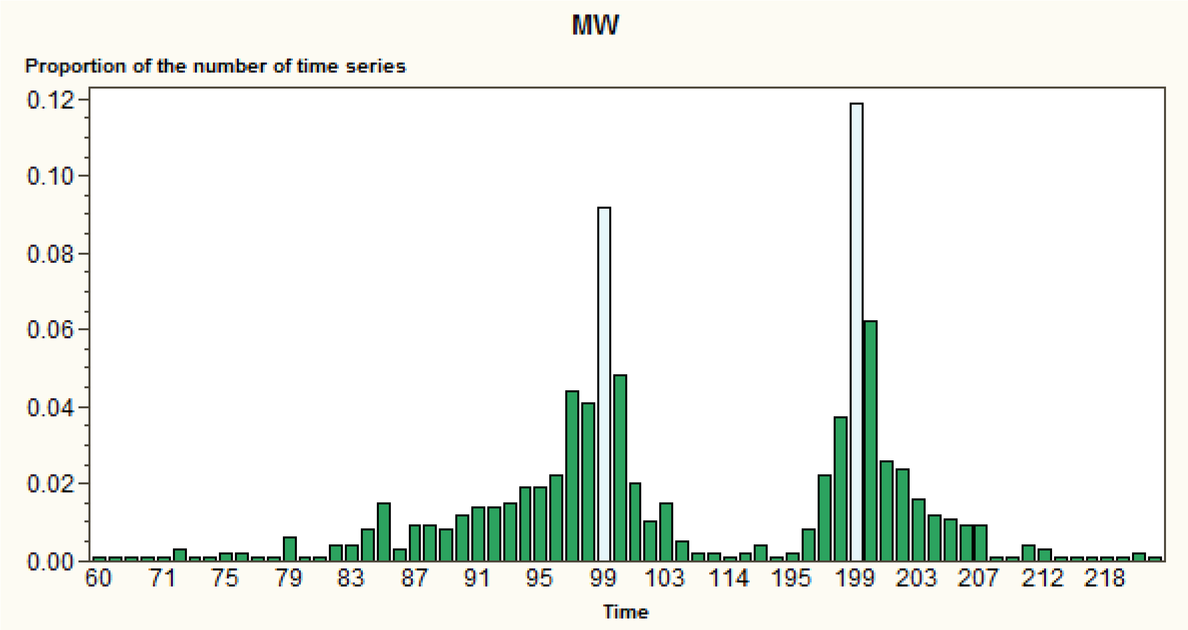


Figure 33 – Identified change-points using the MW method on uniformly distributed data with shifts at 99 and 199

For the uniform distribution the methods are also able to detect the change-points as can be seen in Figure 31-Figure 33. However, the discovered change-points are more spread out than for the normal distribution. The Quandt method is only managing to find one of the change-points, the other change-point produces a local maximum but it is too small to be noted globally (Figure 32).

4.2.5 AR(1)-process

To test the effect of time dependence on the methods a time series with AR(1)-process is constructed. The coefficient $\varphi_1 = 0.8$ is chosen to be relatively large to get a clearer result of the effect of the time dependence. As before the true change-point locations are at $t_0 = 99$ and $t_1 = 199$ and 1000 time series are generated as follows.

$$X_i = 0.8 * X_{i-1} + \varepsilon_i \tag{4.6}$$

$$\begin{cases} \varepsilon_i \in N(0,1), & 0 \le i \le t_0 \\ \varepsilon_i \in N(2,1), & t_0 < i \le t_1 \\ \varepsilon_i \in N(-1,1), & t_1 < i \le 365 \end{cases} \tag{4.7}$$

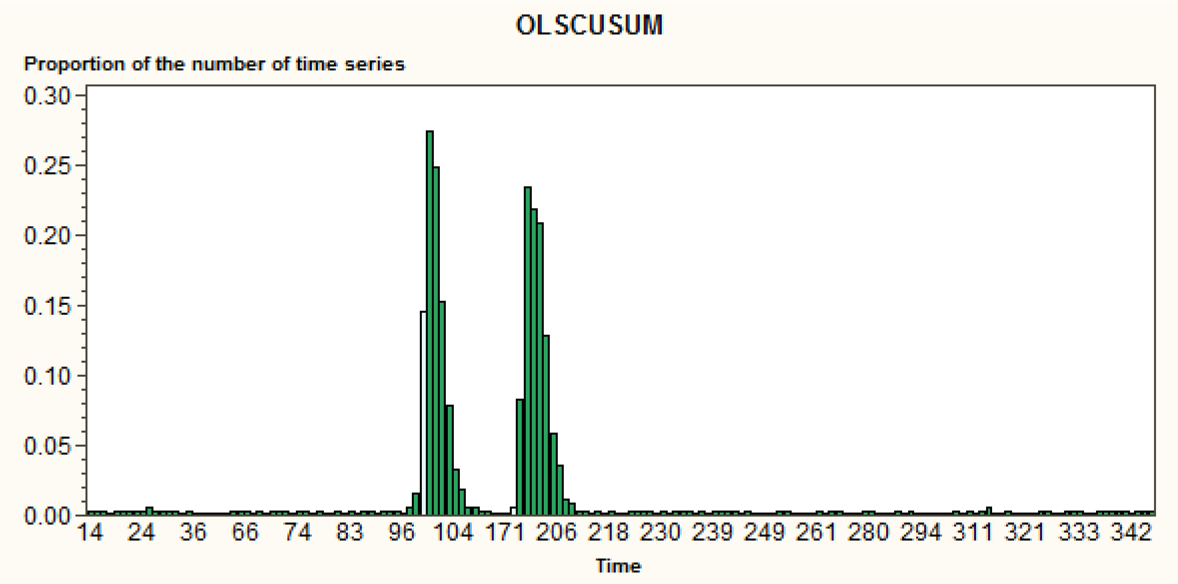


Figure 34 – Identified change-points using the OLSCUSUM method on data from an AR(1)-process with shifts at 99 and 199

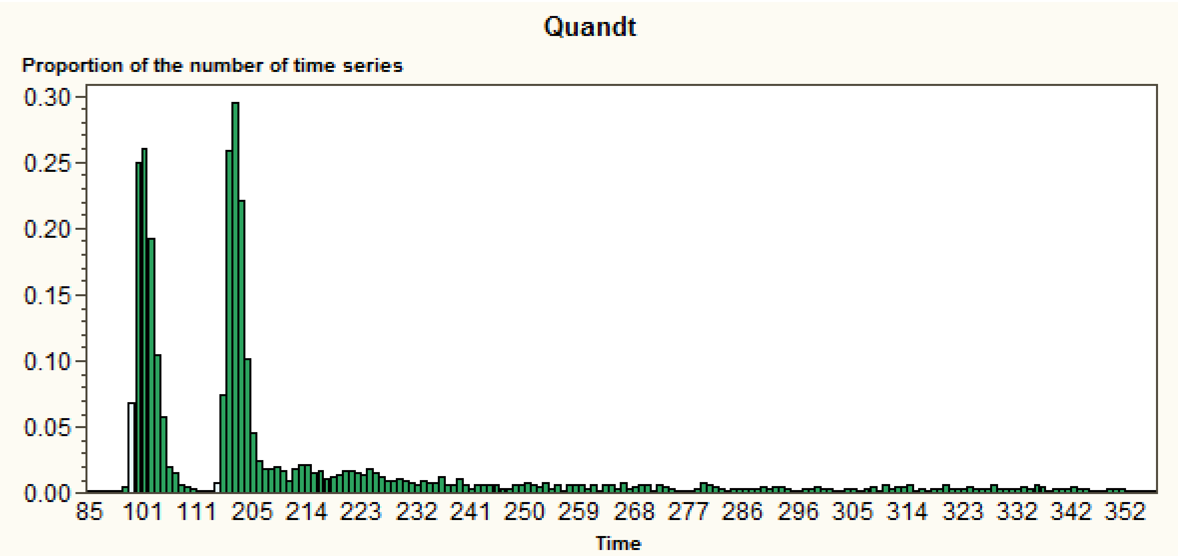


Figure 35 – Identified change-points using the Quandt method on data from an AR(1)-process with shifts at 99 and 199

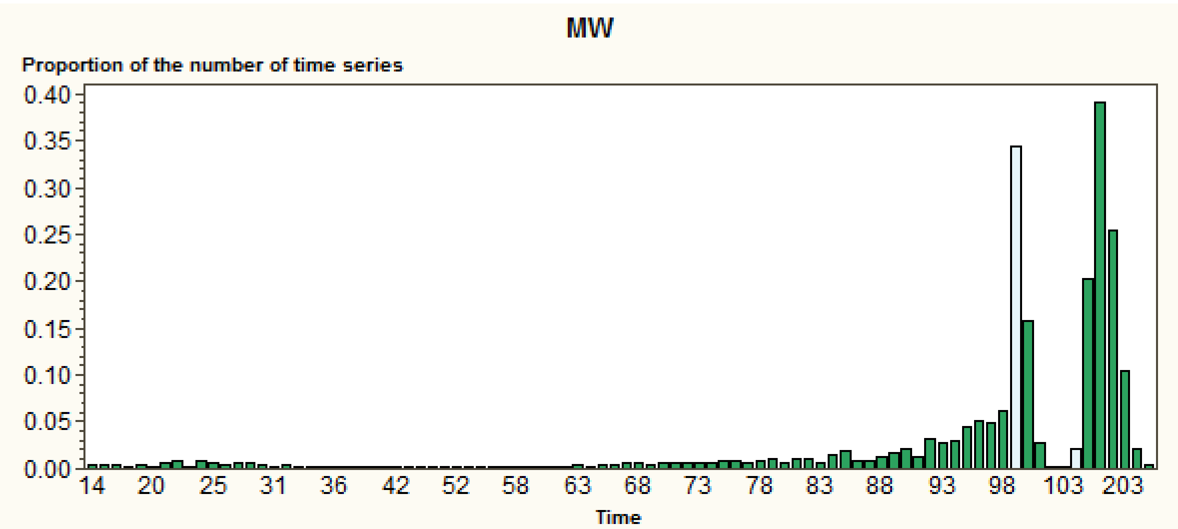


Figure 36 – Identified change-points using the MW method on data from an AR(1)-process with shifts at 99 and 199

Figure 34-Figure 36 shows that even with strong time dependence the methods are able to locate the (approximate) location of the change-points. However, there is a larger variance than for a purely normal distributed process. The peaks of the located change-points are also slightly lagged in comparison to the true change-points.

4.3 Evaluation of the combined method

Even though most methods seem to perform sufficiently well on their own it can be noted that the performance of the methods varies with the distribution of the data. In some cases the amount of false positives is too large to make the method viable and even in the optimal circumstances the methods produce a spread around the change-point. To remedy the problem with too many false positives, a combined method is investigated. This method uses the results from OLSCUSUM, Quandt and MW since they were the best performing individual methods and combines them. Depending on the preferred number of true positives versus false positives, three criteria are presented:

1) The single criterion – all change-points identified with OLSCUSUM, Quandt and MW method are classified as change-points
2) The double criterion – Only change-points identified by at least two of the OLSCUSUM, Quandt and MW methods are classified as change-points
3) The triple criterion – Only change-points identified by all of OLSCUSUM, Quandt and MW methods are classified as change-points

4.3.1 Normal distribution

As for the individual methods a set of 1000 time series is generated as follows:

$$\begin{cases} X_i \in N(0,1), & 0 \leq i \leq 99 \\ X_i \in N(j,1), & 99 < i \leq 365 \end{cases}, j = 0,1,2, \dots, 8 \tag{4.8}$$

The combination method is then tested by counting how many change-points are identified, how many change-points are identified within 3 observations of the correct location and how many are on the correct location.

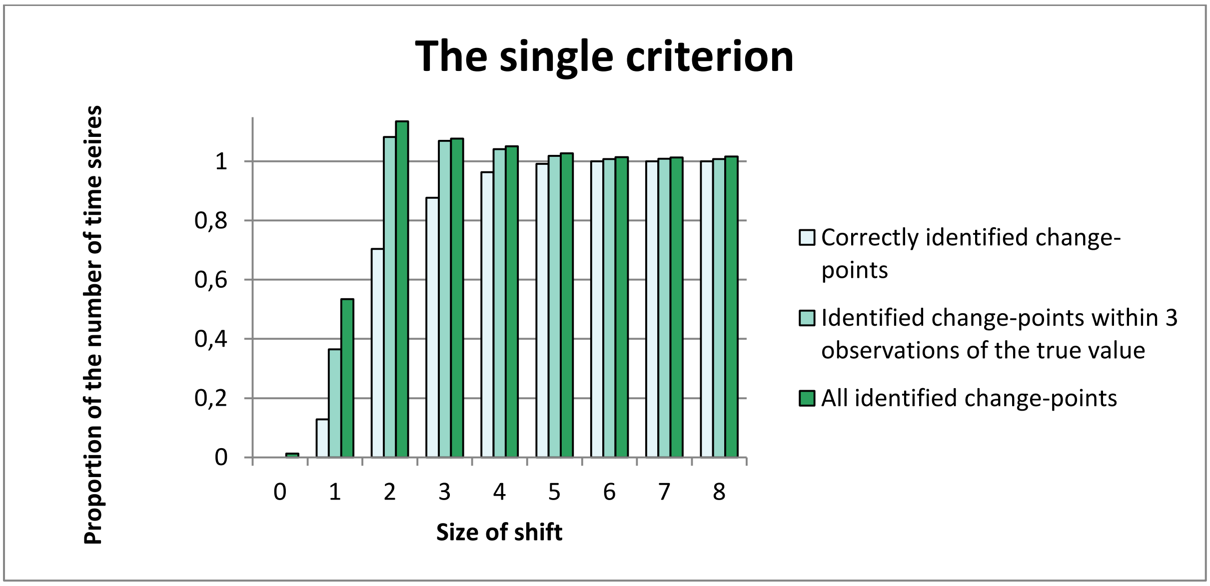


Figure 37 – Identified change-points using the single criterion on normally distributed data with a varying shift

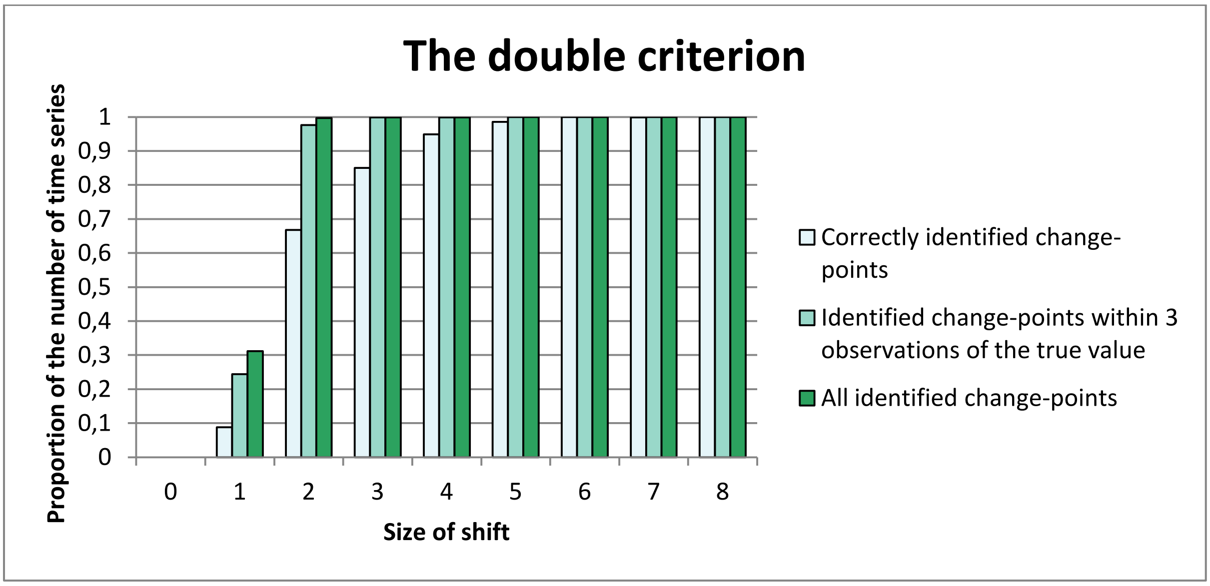


Figure 38 – Identified change-points using the double criterion on normally distributed data with a varying shift

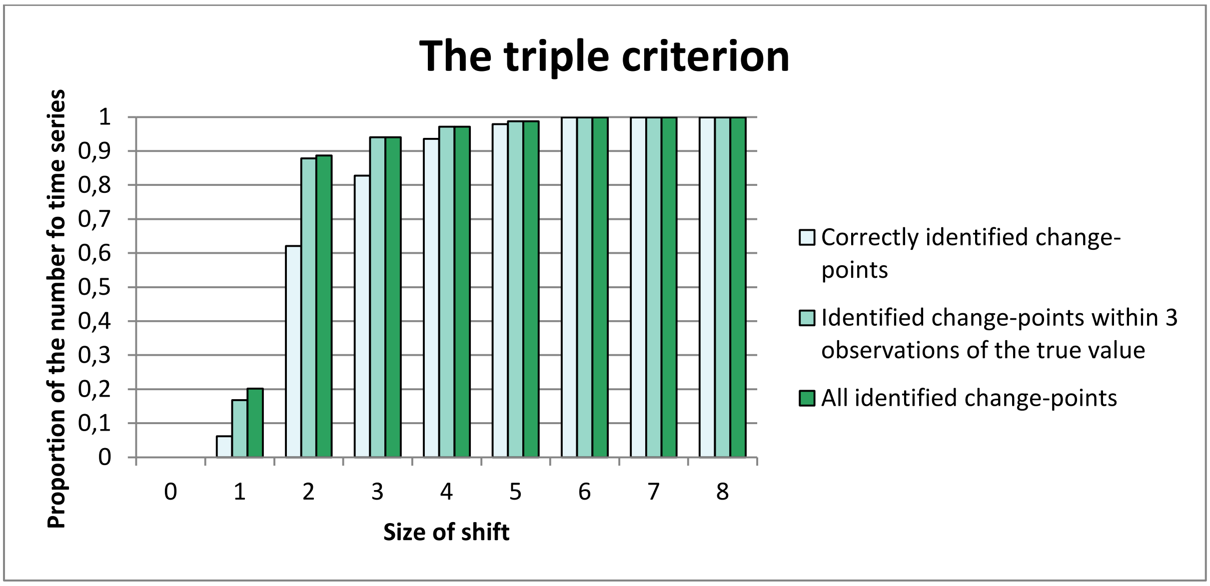


Figure 39 – Identified change-points using the triple criterion on normally distributed data with a varying shift

As can be seen by Figure 37-Figure 39 the double and triple criterion produces less amount of false positives while the single criterion produces a larger amount of true positives. The difference is almost negligible though and as the shift increases the methods perform almost equally well.

4.3.2 Cauchy distribution

The data is not necessarily normally distributed and therefore the combination method is evaluated on data which is performing weaker when applying the individual methods. One such case is the Cauchy distributed data where the individual methods are able to locate the change-points but are also producing plenty of false positives. 1000 time series are generated as follows.

$$\begin{cases} X_i \in Cauchy(0,0.5), & 0 \le i \le 99 \\ X_i \in Cauchy(j,0.5), & 99 < i \le 365 \end{cases}, j = 0,1,2,\dots,10 \tag{4.9}$$

A similar procedure as for the normal distribution is employed. The results are presented in the following figures.

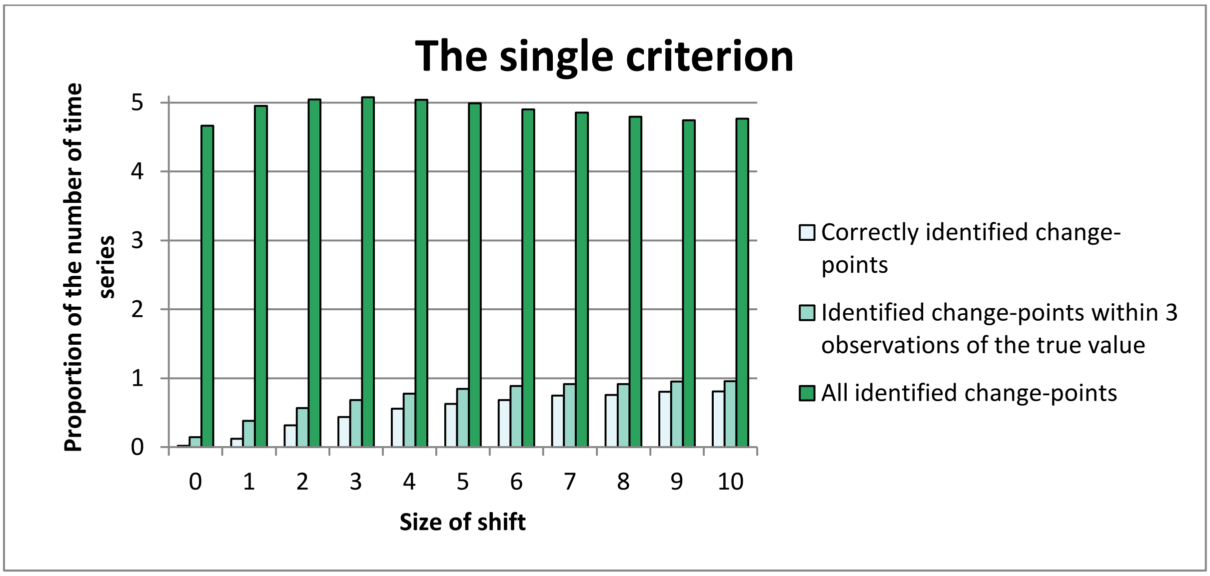


Figure 40 – Identified change-points using the single criterion on Cauchy distributed data with a varying shift

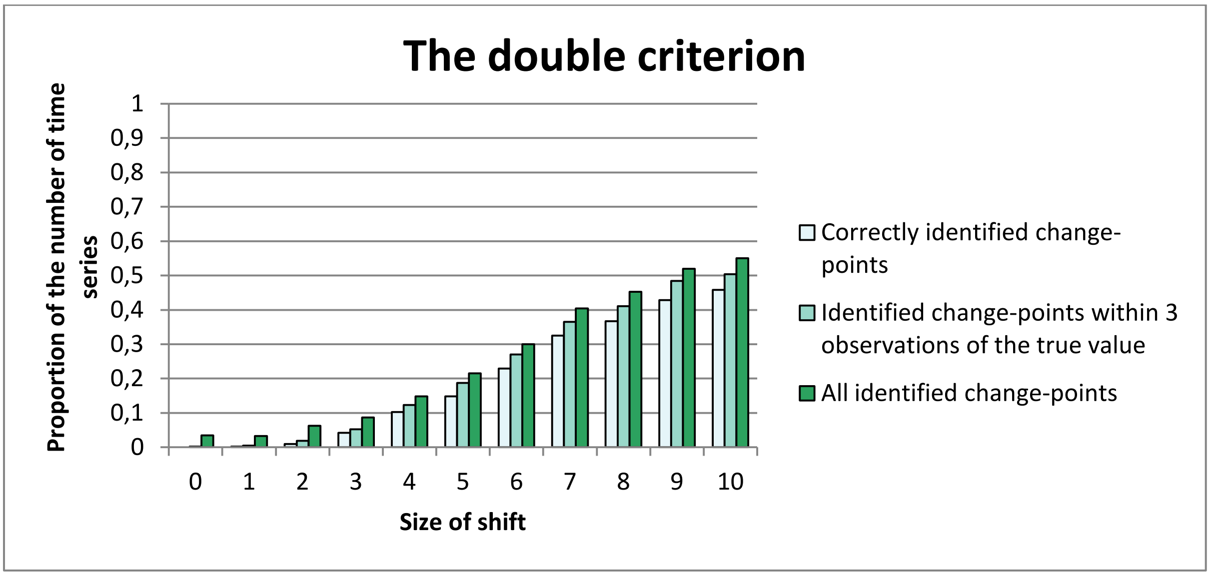


Figure 41 – Identified change-points using the double criterion on Cauchy distributed data with a varying shift

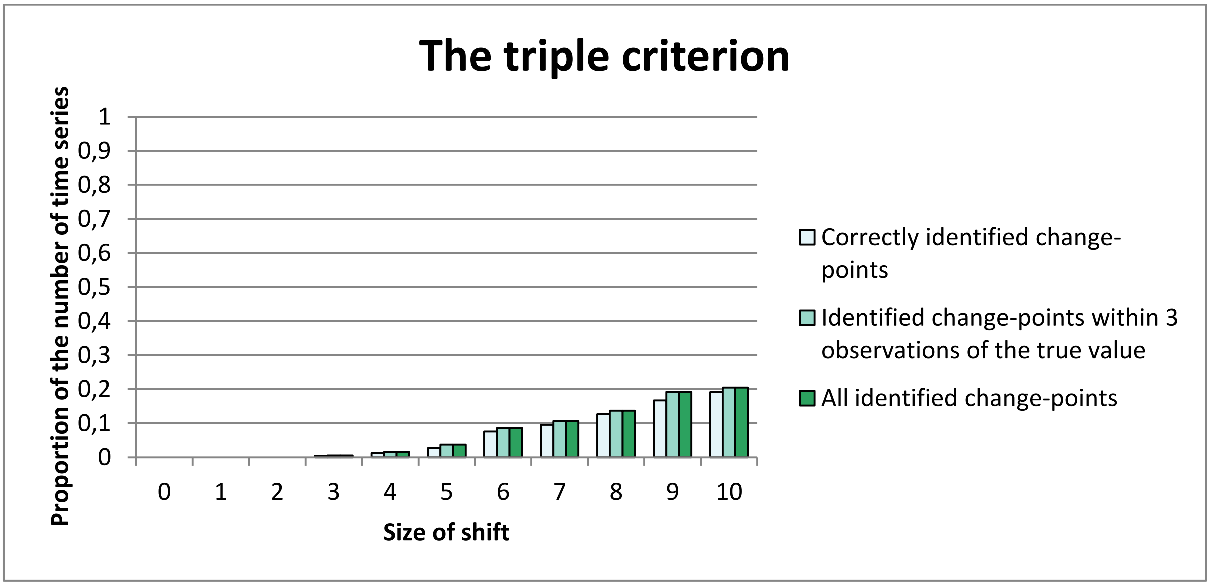


Figure 42 – Identified change-points using the triple criterion on Cauchy distributed data with a varying shift

In Figure 40-Figure 42 the differences between the three criteria are clearly illustrated. The single criterion is able to identify most of the true change-points. However it produces a massive amount of false positives (note the different scale of the y-axis in Figure 40). The triple criterion produces almost only true positives but the amount is smaller than for the single criterion. The double criterion performs somewhere in between with more true positives than the triple criterion but also slightly more false positives.

4.3.3 AR(1)-process

The individual methods are not performing well on data generated from an AR(1)-process. The spread is large and the peaks are not centered at the true change-points. The combined method is tested on data from an AR(1)-process to see if the method is able to resolve the issues from the use of the individual methods. 1000 time series are generated as follows with a varying shift.

$$\mathrm{X_i} = 0.8 * \mathrm{X_{i-1}} + \varepsilon_\mathrm{i} \tag{4.10}$$

$$\begin{cases} \varepsilon_i \in N(0,1), & 0 \leq i \leq 99 \\ \varepsilon_i \in N(j,1), & 99 < i \leq 365 \end{cases}, j = 0,1,2,\dots,15 \tag{4.11}$$

The results are presented as before where the spread of the identified change-points can be seen.

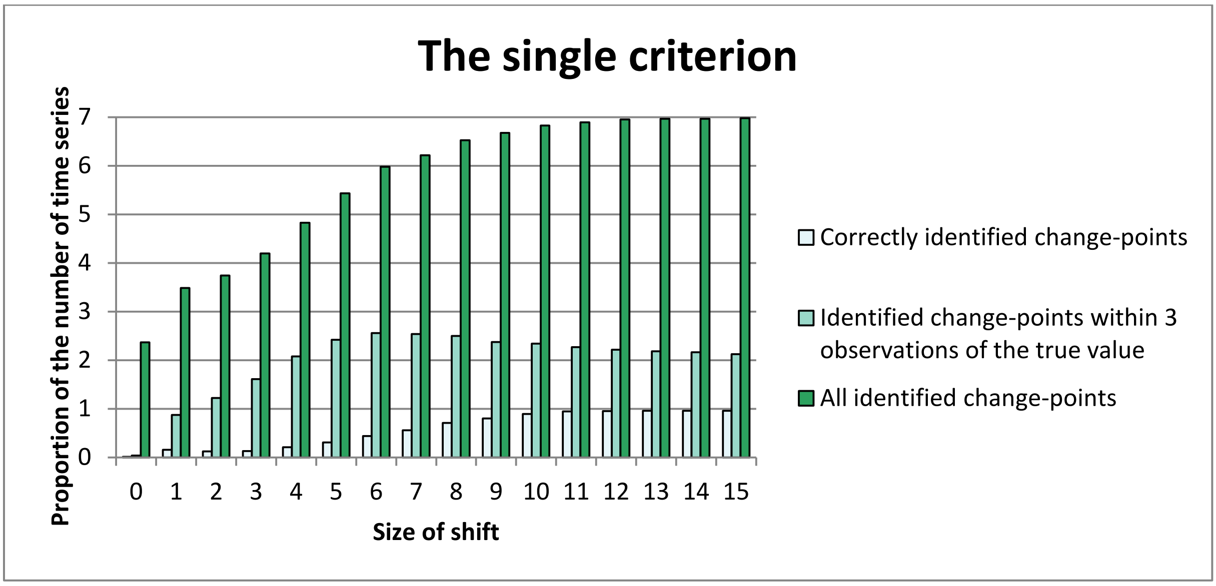


Figure 43 – Identified change-points using the single criterion on data from an AR(1)-process with a varying shift

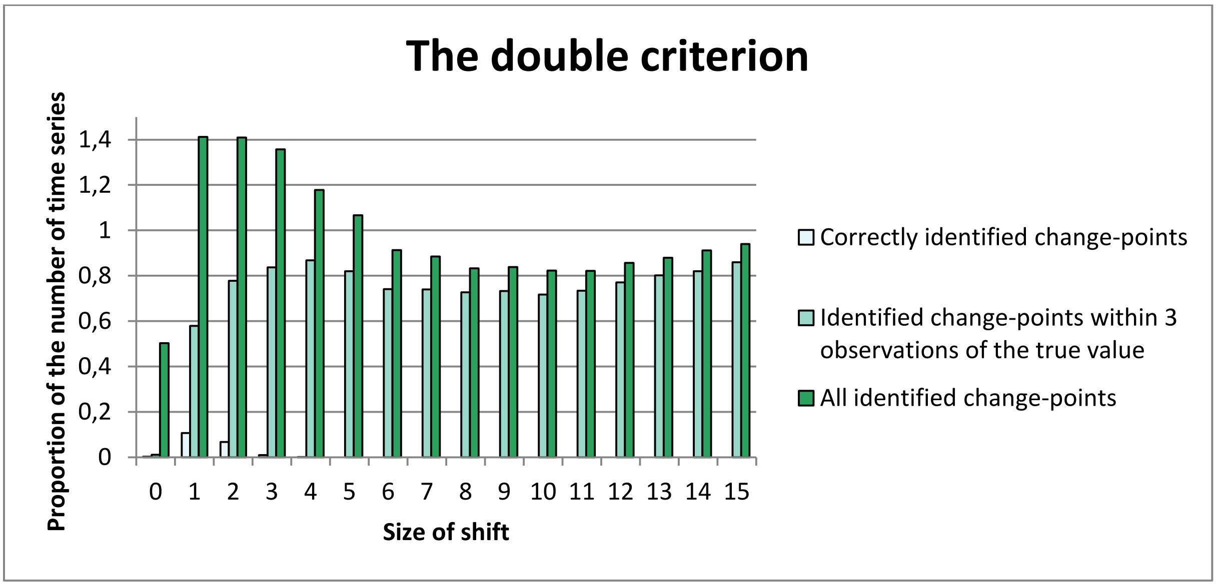


Figure 44 – Identified change-points using the double criterion on data from an AR(1)-process with a varying shift

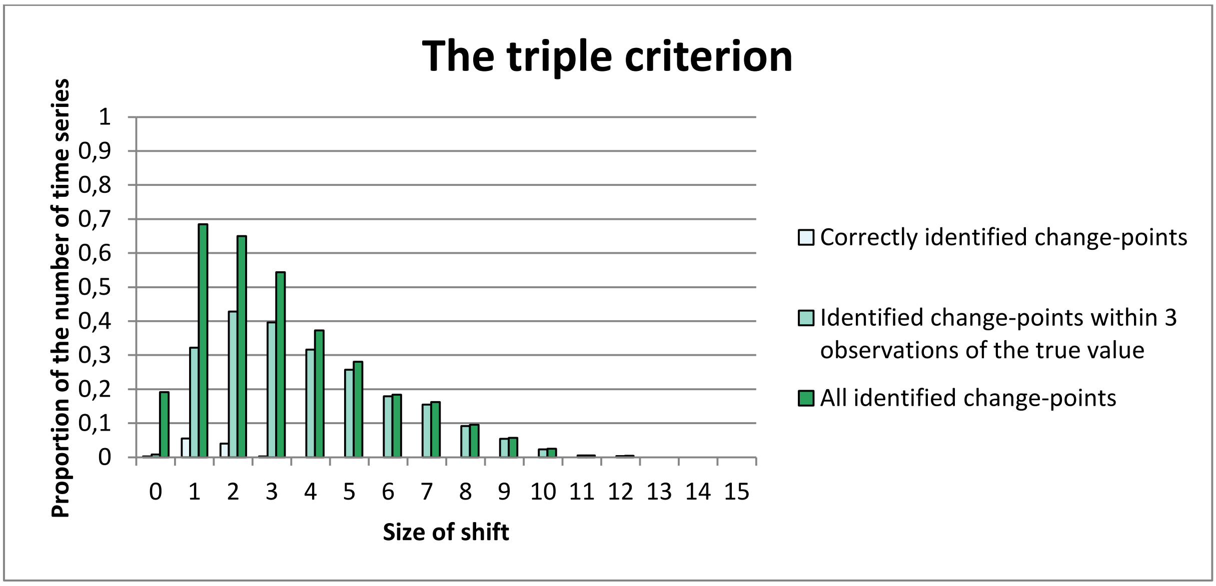


Figure 45 – Identified change-points using the triple criterion on data from an AR(1)-process with a varying shift

In Figure 43-Figure 45 it can be seen that the results are different from the earlier plots. The amount of false positives is larger for all criteria. In Figure 45, the triple criterion, it can be observed that the number of identified change-points is reduced when the size of the shift is increasing which might seem counterintuitive. The double and triple criteria are not able to find any significant amount of true change-points. The double criterion does however find plenty of change-points within three observations of the true value. The single criterion is able to identify most of the true change-points, but as for the Cauchy distribution it comes with the disadvantage of a large amount of false positives.

The strange behavior of the plots makes it interesting to investigate how well the individual methods perform with varying shifts. Hence similar plots are produced for the individual methods.

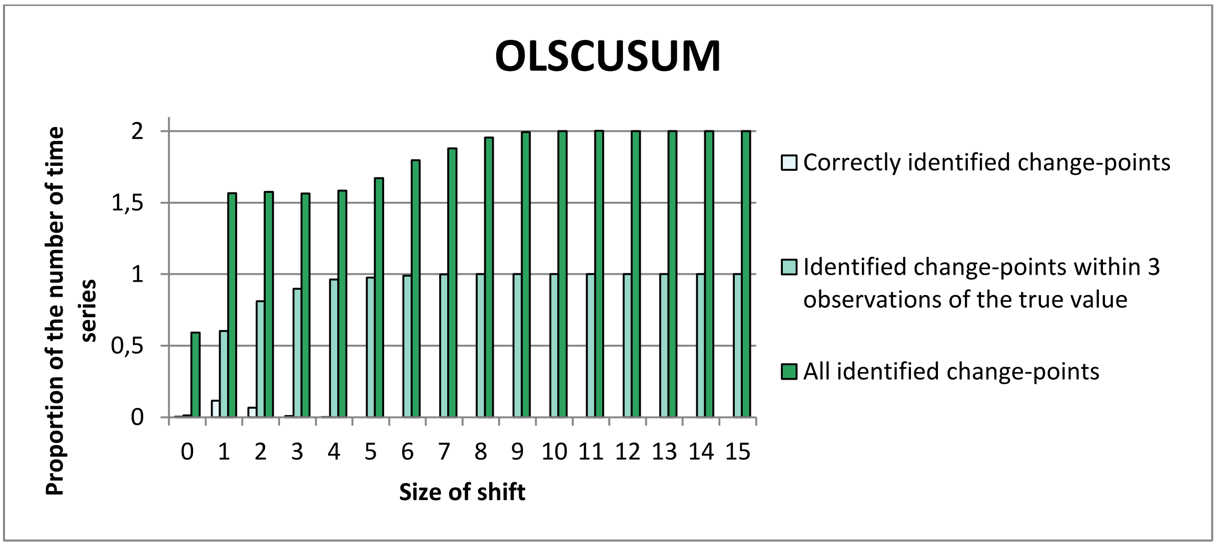


Figure 46 – Identified change-points using the OLSCUSUM method on data from an AR(1)-process with a varying shift

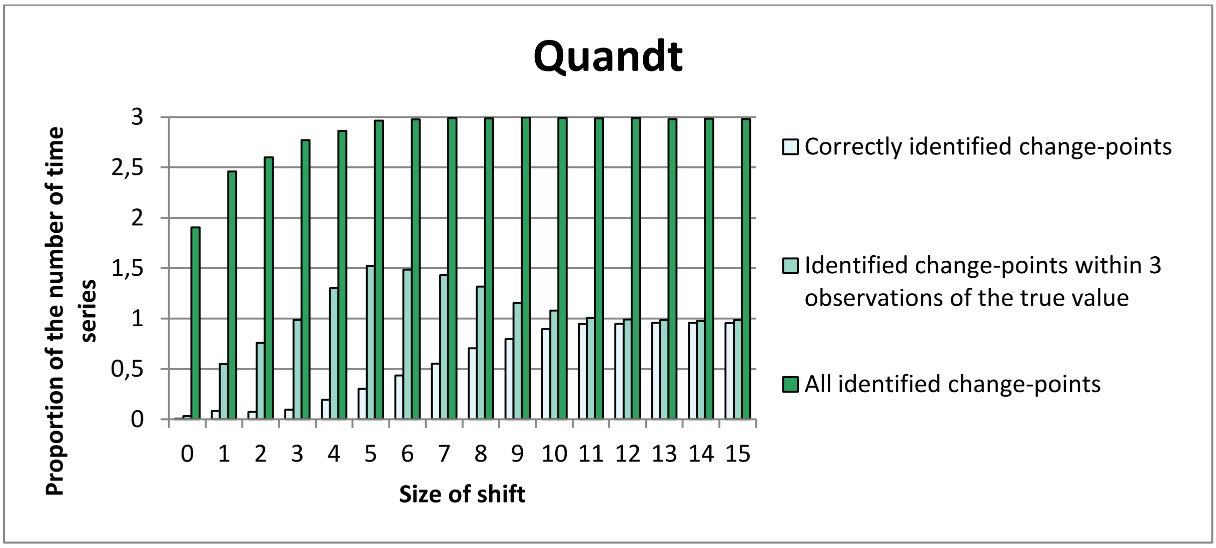


Figure 47 – Identified change-points using the Quandt method on data from an AR(1)-process with a varying shift

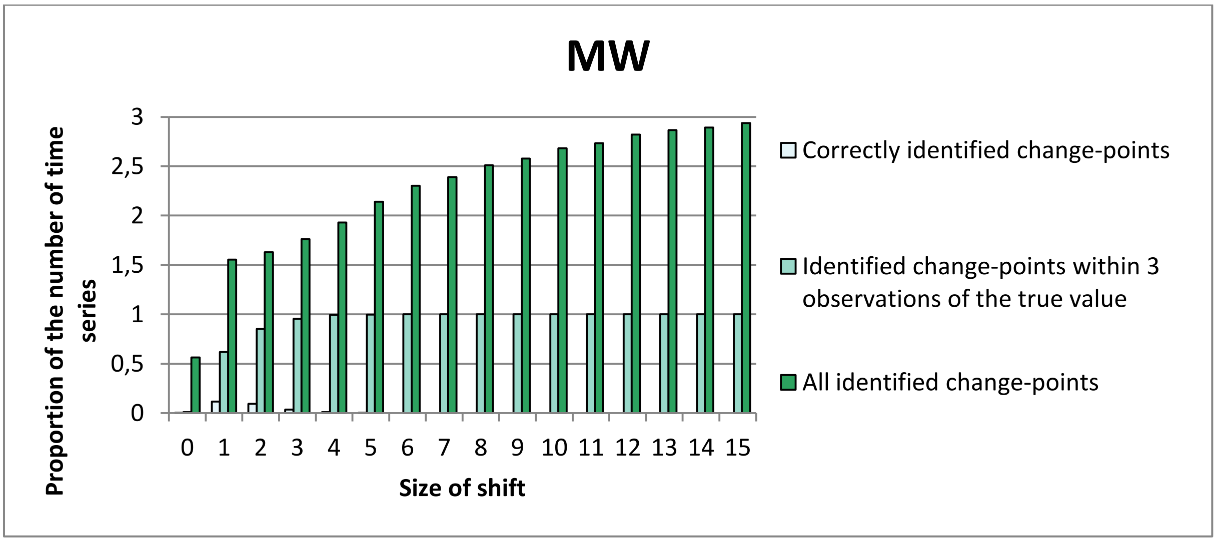


Figure 48 – Identified change-points using the MW method on data from an AR(1)-process with a varying shift

In Figure 46-Figure 48 it can be seen that the only method that is able to detect a significant amount of true locations of the change-points is the Quandt method. However, it only detects the true location for large shifts. All methods are producing large amount of false positives and even though some are lying close to the true location, many are lying further than 3 observations from the true value. Furthermore, all methods are getting around 1000 change-points within 3 observations of the true value for relatively small shifts. This indicates that the methods are detecting the shift but it is somewhat lagged from the true location. This is an observation that is confirmed by the earlier histograms in Section 4.2.5. Hence the bad performance of the individual methods is explaining why the combined method is performing much worse than for i.i.d. data. The conclusion is that the combined method does not significantly improve the results for systematic errors caused by strong time dependence.

4.4 Methods applied to real data

To further investigate how the methods perform in a real life setting they are tested on the data provided by the Bank (daily balances for individual accounts). Since the positions of the true change-points are unknown, the judgment of the performance is based on ocular inspection and relies on the observer's ability to judge if a change has occurred or not.

The binary segmentation technique is used with two iterations (meaning that a maximum of three change-points can be found). All methods are run on the data and the identified change-points are then plotted together with the underlying time series to investigate the performance of the methods. A sample of figures is presented below. The scale on the y-axis is removed for confidentiality reasons.

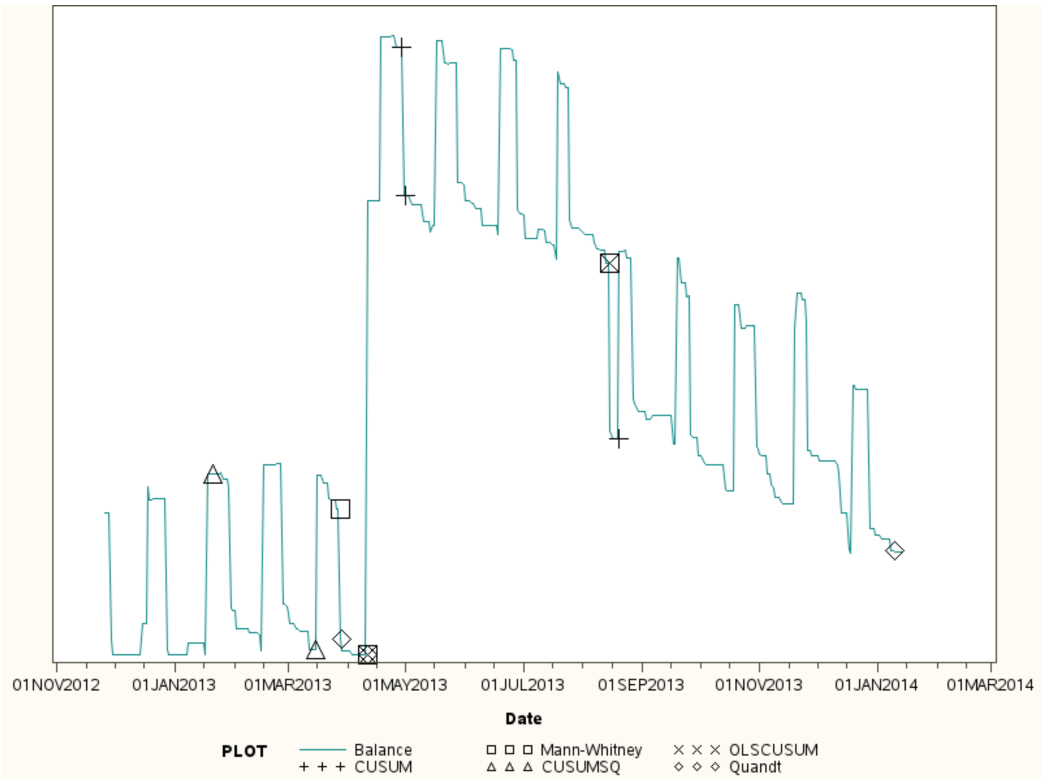


Figure 49 – Identified change-points using all methods on example time series 1

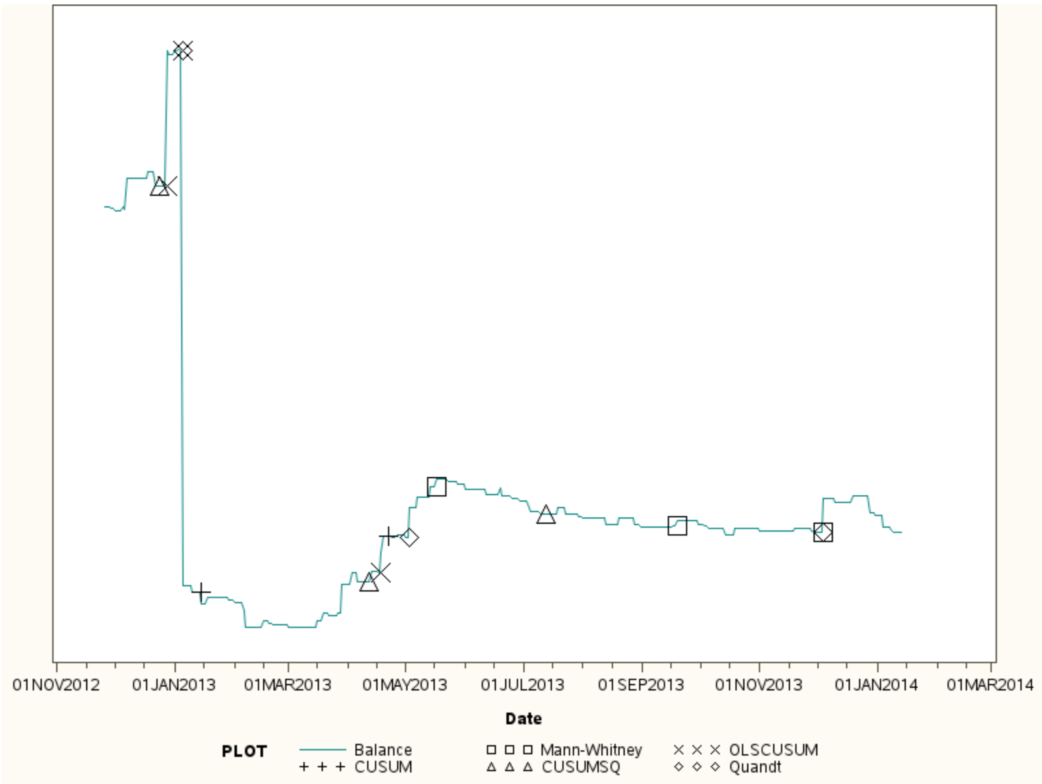


Figure 50 – Identified change-points using all methods on example time series 2

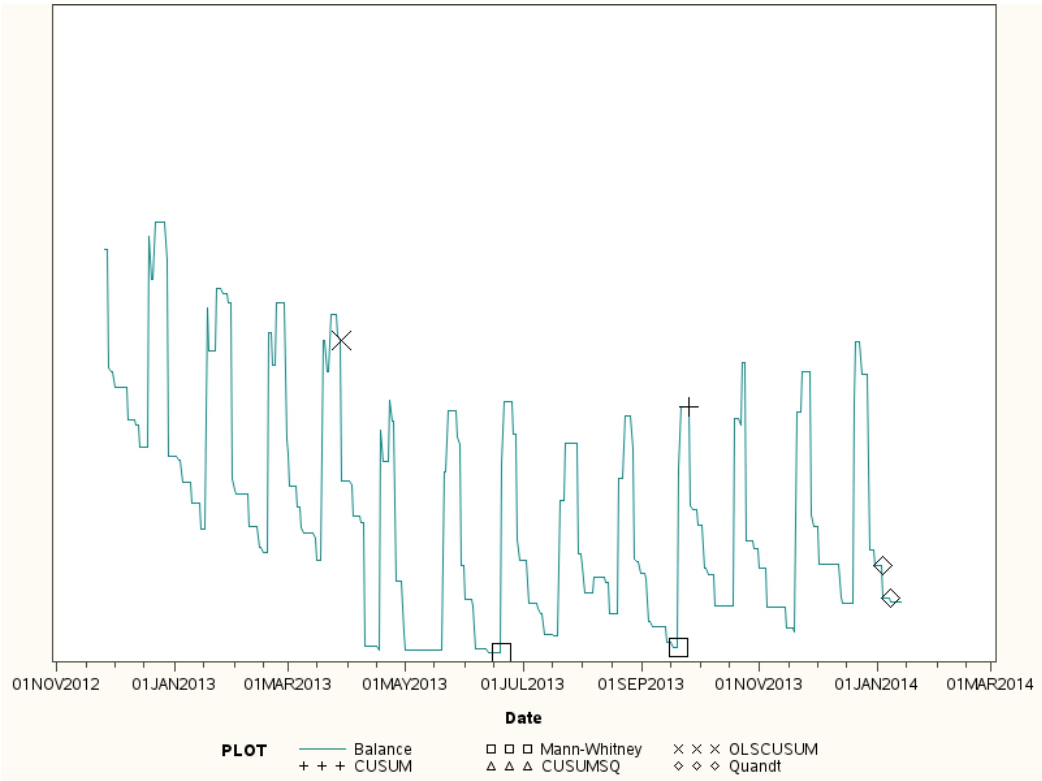


Figure 51 – Identified change-points using all methods on example time series 3

The following analysis is based on the assumption that Figure 49 has two change-points, one at 10 April and the other at 14 August, that Figure 50 has one change-point at 4 January and that Figure 51 has one change-point at 9 April. It is not certain that these points are the true change-points or that those are the only ones, but they are the points that the author identifies as an obvious shift in the underlying distribution.

The result from an ocular inspection of Figure 49-Figure 51 is that most of the results from earlier sections seem to hold. It can be observed that the CUSUM and CUSUMSQ methods are not able to find the location of the change-points in any of the plots. It can also be observed that the other three methods, OLSCUSUM, Quandt and MW are able to detect the correct location of the change-point most of the time.

There is however also differences in the performance between generated data and real life data. The number of false positives is larger than for the analysis on the generated data which can be most clearly seen in Figure 50 where the change-points are seemingly scattered all over the time series. It can also be observed that even though the shifts are large in Figure 49 and Figure 50, only OLSCUSUM manages to identify all shifts. Figure 50 seems to confirm that the MW method has problems identifying change-points close to the beginning of the time series.

Furthermore it can be seen that the OLSCUSUM method is the only method being close to identify the shift in Figure 51. It can also be seen that the Quandt method has a tendency of detecting false change-points at the end of the time series as can be seen in Figure 49 and Figure 51. Lastly it is observed that the combination method seems to perform well, since the points where the change-points coincide are the points with the most obvious shifts.

5 Discussion

In this chapter the results from the previous chapter are discussed. Firstly the performance of the binary segmentation technique is evaluated. Then the results from each of the methods are discussed. Thereafter the ensemble and combined methods are evaluated. Lastly the performance of the methods on real life data is analyzed.

5.1 Binary segmentation

The binary segmentation technique is shown to be useful when testing for an unknown number of change-points using methods that detect single change-points. The simplicity of the implementation makes it preferred over more complicated models. However, there are some disadvantages with the technique. One disadvantage is that it is computationally expensive since the analysis is run on the same data multiple times. This means that the calculations are generally slower than a technique that is able to find multiple change-points simultaneously. However, since the tested methods are quickly calculated this is not a serious issue unless one has very large data samples.

Another issue with the technique is that the time intervals are sequentially reduced. Some of the methods are based on large samples. The reduction of the size of the samples might hence affect the performance of some of the methods. OLSCUSUM is based on limit theorems which assume infinite sized samples which means that small samples will not correspond to the assumptions behind the model. Ploberger and Krämer (1992) do however show that the results are valid for moderately sized samples and that the estimation is almost always conservative for smaller sample sizes. For the Quandt method, the limits for rejecting the null hypothesis are based on a sample of 200. If the sample size is reduced below that size, the performance is affected. However, it can be seen in the article by Deutsch (1992) that the limits are conservative for smaller sample sizes. Since this study is mostly concerned with the number of false positives and the tests are conservative for small samples, the issue with small samples is not that severe.

Another disadvantage with the binary segmentation technique is that it does not combine well with methods that perform badly at the edges of time series. Since the time series is sequentially divided into subintervals it means that the number of edges is increasing for every iteration. A consequence of this is that the binary segmentation technique is reducing the ability to identify change-points that are close to each other with methods that do not work well on the edges (which are shown to be OLSCUSUM and MW). Methods that detect multiple change-points collectively are more difficult to implement but do only have two edges in the time series and hence do not have the same risk of missing change-points within the sample.

The study shows that the binary segmentation technique works well on samples with only a single change-point. It does not produce significantly more false positives even though it actively looks for change-points even after the first change-point is identified. This shows that there is no significant disadvantage of using the technique on data with few change-points and hence it could most of the time be safely used on data with an unknown amount of change-points.

5.2 CUSUM and CUSUMSQ

CUSUM and CUSUMSQ are methods primarily constructed to test for existence of structural breaks rather than finding the exact location of the change-points. This is confirmed by the results of this study where the methods are shown to only identify the correct location of the change-points on rare occasions.

Both the CUSUM and CUSUMSQ tests have limits in the derivation of the methods. The variance under the null hypothesis is varying but the rejection limits are straight lines. This means that the probability of crossing the line is not equal over the whole sample. This is not optimal but it is a necessary simplification since calculating the crossing probability for a non-straight line is very difficult. The implication of this simplification is that the power of the test is varying over the sample. This is shown empirically by e.g. Deutsch (1992) who points out that the power of the CUSUM test is high for structural breaks in the beginning of the time period and that the power of the CUSUMSQ test is high for structural breaks at the end of the time period.

5.3 OLSCUSUM

Similarly to the CUSUM and CUSUMSQ methods the OLSCUSUM method produces a straight critical line when in reality there is a non-constant variance in the null hypothesis. This, again, means that the probability of crossing the line is not constant in the whole sample, which affects the power function.

Another limitation is that the critical levels are calculated asymptotically with infinite sample sizes. In reality the samples are not as large and are sometimes very small. As discussed earlier the binary segmentation technique makes the sample sizes sequentially even smaller. Ploberger and Krämer (1992) do however show via Monte Carlo simulations that the OLSCUSUM method works well for moderate sample sizes and that the test is almost always conservative. Since this study is mostly concerned with false positives small sample sizes do not severely impact the results.

The OLSCUSUM method performs well on multiple change-points that lie evenly spaced in time. It does however, not perform well on data with change-points close to the edges, where it does not manage to locate either the change-point in the beginning or at the end. On data with change-points close to each other the method finds one of the change-points but not the other. This can be explained by the inability of detecting change-points close to the edges. Because the binary segmentation technique is used it means that when the first change-point is found, the other ends up on the edge of one of the subintervals and hence is not detected.

5.4 Quandt's log likelihood ratio

Quandt's log likelihood ratio is based on the distribution function of the residuals. For a simpler implementation it is assumed that the data is normally distributed in this study. As seen when analyzing the data provided by the Bank, the distribution of the residuals are heavier tailed than a normal distribution which means that the normality assumption is questionable. However this study shows that the method works sufficiently well for distributions that are similar to the normal distribution and that it performs very well on normally distributed data. Unlike the MW and OLSCUSUM method it is able to detect change-points that lie close to the edges and change-points that lie close to each other.

For data that is not similar to the normal distribution the Quandt method is performing worse than the MW and OLSCUSUM methods. This means that as long as the data is sufficiently close to normally distributed the Quandt method performs well. If the data is very heavy tailed or has a distribution that is far from normally distributed, the performance of the Quandt method is worsened. In a setting with data of unknown distribution, careful distribution analysis should be performed or a combination method that limits the effect of a single method should be used.

5.5 Mann-Whitney

One of the limitations of the Mann-Whitney test is that the significance is based on an approximation. The approximation does however hold well for significance levels under 50% (Pettitt, 1979). The Mann-Whitney test is more robust against outliers compared to other parametric methods. This is a good property since it means that random spikes in the data do not affect the detection of change-points and only permanent shifts are detected.

The Mann-Whitney test is a non-parametric test which means that it performs well on data with unknown distribution. When the data is known to be close to normally distributed it is however performing worse than methods based on the normality assumption. The benefits of the method are hence limited in that setting. This study has shown that the method works well for normally distributed data and the difference between other methods are negligible. Also, in reality it is not certain which distribution the data belongs to which makes the MW test preferred over tests based on normality assumptions.

The Mann-Whitney method performs well on evenly spaced change-points not too close to the edges. If the change-points lie too close to the edges it does not manage to find the true location of the change-points. Similarly to the OLSCUSUM method it only manages to identify one of two change-points if they lie close to each other. The reason for this is similarly that the binary segmentation technique makes one of the change-points ends up on the edge of the subinterval when the first change-point has been identified.

5.6 Ensemble method

This study shows that the ensemble method on normally independently distributed data with multiple change-points is performing worse than the regular MW method. This might be because it was developed and tested on data with a single change-point. With the presence of multiple change-points it is possible that the method detects different change-points in different iterations which can skew the results. This can be observed in Figure 20 in Section 4.2.1.2 where there is an increased probability of detecting a change-point in between the two true change-points which is not seen for the regular MW method.

Another disadvantage of the method is that it is computationally expensive and takes longer to run than a single run of the whole sample. In analysis of a few time series, this difference is negligible but for larger sets of time series this becomes an important issue that is to the method's disadvantage.

5.7 Combined method

This study has shown that the combined method can be a useful tool when the distribution of the data is unknown. The usefulness does however depend on the underlying data. The combined method applied to normally distributed data does not provide any advantage and performs similarly to the individual methods. When applying the method to the heavier tailed Cauchy distribution, the advantages of the method becomes more apparent. This study shows that when trying to reduce the number of false positives, the combined method produces much better results for the Cauchy distribution than the individual methods. When the issue of false positives is not that severe, and the number of true positives is more important, the single criterion does however perform better.

This finding shows that if the underlying distribution of the data is known to be normal, the combination method does not provide a significantly improved performance and hence could be avoided. If the underlying distribution of data is different from normally distributed the combination method can produce significantly improved results, at least for heavy tailed distributions. Since the method does not perform significantly worse for normally distributed data it is clear that the combination method is superior when used on data with an unknown underlying distribution and when the number of false positives is the important factor of the test.

The interpretation of the effect of the combined method on data from an AR(1)-process is more difficult. The performance is considerably worse than for the other two examples of distributions with plenty of false positives. The observation that the triple method identifies fewer change-points for larger shifts also seems counterintuitive. The reason for this is that there is a systematic lag in the individual methods. This means that the identified change-points are centered on an incorrect location. Since the combined method mostly reduces the spread and not systematic errors the method is not able to improve the results of systematic lags. As can be seen when the spread for individual methods are plotted the Quandt method seems to locate the correct location (for large shifts) but the other two methods are identifying positions with a lag from the true position. Since all three methods are combined in the triple criterion they are not able to center on the true position. In fact it is sufficient for one method producing systematic errors to affect the performance of the triple criterion negatively. There is hence a potential need to modify the combined method when applying it to data with (strong) time dependence.

The choice of using the double or triple criterion depends on the preferences of the number of false positives versus the number of true positives. In most cases the double criterion is performing sufficiently well since it manages to locate most change-points while keeping the number of false positives low. But if false positives have a very large impact on the analysis the triple criterion should be considered.

5.8 Methods applied to real life data

The methods are also tested on data that comes from real life situations with unknown properties. This means that it has the complexity of real life data that is difficult to replicate. It does however also mean that the true properties of the data are unknown and it is therefore impossible to base the conclusions solely on this analysis. The results can however be used indicatively and as a comparison with the results based on generated data.

The results show that most of the results from the generated data seem to hold for real life data. The best performing methods are MW, OLSCUSUM and Quandt, while CUSUM and CUSUMSQ are not able to locate the assumed positions of the change-points. It also shows that the amount of false positives is larger than what is observed for the generated data. This means that the issue with incorrectly identified change-points seems more severe in a real life application. However, since the number of change-points is unknown it is also possible that the data contains that many change-points and that the methods are able to locate them.

A common property for most of the plots is that the biggest shifts in the data are identified by multiple methods, while the other identified change-points are spread out. This means that the combination method works very well on the real life data. Since it only counts change-points if multiple methods identify them, it is able to identify the largest shift and avoid most of the false positives. As for the generated data it can be observed that the double criterion is able to identify a larger amount of true change-points than the triple method but the triple method has potentially fewer false positives.

6 Conclusions

This study has demonstrated that the binary segmentation technique is a good technique to find multiple change-points and that it mostly works well with all tested methods. There is however a disadvantage of using the technique in combination with methods that do not perform well on edges (OLSCUSUM and MW). This is because the technique is sequentially adding edges to the data. The CUSUM and CUSUMSQ methods are shown to work indicatively, but are not able to locate the change-points. The OLSCUSUM, Quandt's log likelihood ratio and MW test are individually performing well and are able to locate most of the change-points. No method is universally better than the other; instead the performance is based on the properties of the data.

The individual methods are shown to produce plenty of false positives, especially for heavier tailed distributions, which means that the usefulness of the results is reduced. This study tries to resolve this problem by using a combination of the methods where only change-points located by several methods are considered true change-points. The results from the tests show that this combined method is able to lower the number of false positives, even in a non-ideal setting. However, on data that creates systematic errors, such as for time dependant data, the combined method will not be able to improve the performance and will at most minimize the number of false positives. This study also shows that the ensemble method does not perform well in a setting with multiple change-points.

Hence, it is possible to use simple methods to locate multiple change-points even though the properties of the data are unknown in advance and the assumptions behind the models are not necessarily fulfilled.

6.1 Suggestions for further research

This study is evaluating the performance of several methods on real life data. To be able to draw any conclusions, artificial data has been created in an attempt to replicate the properties of real life data. The complexity of those properties has made it difficult to replicate all properties and how they interact with each other. Further studies on the behavior of real life data and a more accurate replication of the data will shed further light on the performance of the methods.

This study does not manage to get an improved performance by using the Ensemble method described by Alippi et al. (2013). This might be because this study is mainly performed on data with multiple shifts which is not investigated by Alippi et al. The method could possibly be modified to fit a wider range of data and if so, it could improve the performance of the methods on time dependent data.

The binary segmentation technique has proven to be a, for most cases, successful method for adapting single change-point methods to detect multiple change-points. One limitation with the technique is that it does not work well with methods that do not perform well on the edges. Neither is it optimal for methods based on limit theorems and large sample sizes. A method constructed to detect multiple change-points simultaneously could hence be investigated to see how well it performs in comparison.

There exist more change-point detection methods than those that have been investigated in this study. They could be tested to see if the performance could be enhanced. One such method is the

MOSUM method that is based on moving sums. Also more advanced methods that are more difficult to implement but works for a wider range of data could be investigated.

The combined method could possibly be extended to provide even better results. Instead of exactly coinciding change-points it could allow for a window where if multiple change-points are inside that window, is classified as a change-point. This could possibly improve the results further when there is an unsystematic spread around the true change-point.

Bibliography

Alippi, C., Boracchi, G. & Roveri, M., 2013. Ensembles of change-point methods to estimate the change point in residual sequences. *Soft Computing,* 17(11), pp. 1971-1981.

Bai, J. & Perron, P., 1998. Estimating and Testing Linear Models with Multiple Structural Changes. *Econometrica,* 66(1), pp. 47-78.

Barros, P. P. & Nunes, L. C., 2010. The impact of pharmaceutical policy measures: An endogenous structural-break approach. *Social Science & Medicine,* 71(3), pp. 440-450.

Beckage, B. et al., 2007. Bayesian change-point analyses in ecology. *New Phytologist,* 174(2), pp. 456-467.

Brown, R. L., Durbin, J. & Evans, J. M., 1975. Techniques for Testing the Constancy of Regression Relationships over Time. *Journal of the Royal Statistical Society. Series B (Methodological),* 37(2), pp. 149-192.

Chen, J. & Gupta, A. K., 2012. *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance.* 2nd ed. s.l.:Birkhäuser Boston.

Chow, G. C., 1960. Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica,* 28(3), pp. 591-605.

Deutsch, J., 1992. Linear regression under two separate regimes: an empirical distribution for Quandt's log likelihood ratio. *Applied Economics,* 24(1), pp. 123-127.

Durbin, J., 1969. Tests for Serial Correlation in Regression Analysis Based on the Periodogram of Least-Squares Residuals. *Biometrika,* 56(1), pp. 1-15.

Edgerton, D. & Wells, C., 1994. CRITICAL VALUES FOR THE CUSUMSQ STATISTIC IN MEDIUM AND LARGE SIZED SAMPLES. *Oxford Bulletin of Economics and Statistics,* 56(3), pp. 355-365.

Hinkley, D. V., 1971. Inference about the Change-Point from Cumulative Sum Tests. *Biometrika,* 58(3), pp. 509-523.

Page, E. S., 1954. Continuous Inspection Schemes. *Biometrika,* 41(1), pp. 100-115.

Pettitt, A. N., 1979. A Non-Parametric Approach to the Change-Point Problem. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 28(2), pp. 126-135.

Ploberger, W. & Krämer, W., 1992. The CUSUM test with OLS residuals. (ordinary least squares). *Econometrica,* 60(2), pp. 271-285.

Quandt, R., 1958. The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes. *Journal of the American Statistical Association,* 53(284), pp. 873-880.

Talwar, P. P., 1983. Detecting a shift in location: Some robust tests. *Journal of Econometrics,* 23(3), pp. 353-367.

Appendix A – Probabilities of overlapping subsamples

Triple overlapping subsamples

In this section the probability of three subsamples overlapping is calculated.

Let S be the original sample with cardinality N and X_1, X_2, X_3 be the sets of subsamples with cardinality n_1, n_2, n_3 such that $X_1 \subseteq S$, $X_2 \subseteq S$, $X_3 \subseteq S$

Let $Y_{1,2} = |X_1 \cap X_2|$ be the number of overlapping elements in X_1 and X_2

Let $Y_{1,2,3} = |X_1 \cap X_2 \cap X_3|$ be the number of overlapping elements in X_1, X_2 and X_3

The probability of k triple overlapping points when drawing 3 subsamples is

$$P(Y_{1,2,3} = k) = \sum_{i=k}^{\min(n_1,n_2)} P(Y_{1,2,3} = k \mid Y_{1,2} = i) * P(Y_{1,2} = i) \tag{0.1}$$

where the law of total probability is used in equation (0.1).

Furthermore, since the subsamples are drawn without replacement, the distributions are

$$Y_{1,2,3} | Y_{1,2} = i \in Hyp\left(N, n_3, \frac{i}{N}\right) \tag{0.2}$$

$$Y_{1,2} \in Hyp\left(N, n_2, \frac{n_1}{N}\right) \tag{0.3}$$

which gives the probability density function[7]

$$P(Y_{1,2,3} = k) = \sum_{i=k}^{\min(n_1,n_2)} \frac{\binom{i}{k}\binom{N-i}{n_3-k}}{\binom{N}{n_3}} * \frac{\binom{n_1}{i}\binom{N-n_1}{n_2-i}}{\binom{N}{n_2}} \tag{0.4}$$

Double overlapping subsamples

Similar calculations are performed for the (at least) double overlapping points.

Let $Z_{1,2} = |X_1 \cap X_2|$ be the number of overlapping elements in X_1 and X_2

And $Z_3 = |(X_1 \cap X_3) \cup (X_2 \cap X_3) \setminus (X_1 \cap X_2 \cap X_3)|$ be the number of overlapping elements between X_1 and X_3 and X_2 and X_3 (not counting the triple overlapping elements in X_1, X_2 and X_3).

And $Z_{1,2,3} = |(X_1 \cap X_3) \cup (X_2 \cap X_3) \cup (X_1 \cap X_2)|$ be the number of (at least) double overlapping points. It can be noted that $Z_{1,2,3} = Z_{1,2} + Z_3$

The probability of k (at least) double overlapping points when drawing 3 subsamples is

[7] The probability density function for a hypergeometric distributed stochastic variable $X \in Hyp(N, n, p)$ is $P(X = k) = \frac{\binom{Np}{k}\binom{N(1-p)}{n-k}}{\binom{N}{n}}$

$$P\left(Z_{1,2,3}=k\right)=P\left(Z_{1,2}+Z_3=k\right)=\sum_{i=\max(0,k-n_3)}^{\min(n_1,n_2,k)} P\left(Z_{1,2}+Z_3=k \middle| Z_{1,2}=i\right)P(Z_{1,2}=i) \tag{0.5}$$

Or equivalently

$$P\left(Z_{1,2,3}=k\right)=\sum_{i=\max(0,k-n_3)}^{\min(n_1,n_2,k)} P\left(Z_3=k-i \middle| Z_{1,2}=i\right)*P\left(Z_{1,2}=i\right) \tag{0.6}$$

Furthermore the distributions are

$$Z_3|Z_{1,2}=i \in Hyp\left(N,n_3,\frac{n_1+n_2-2i}{N}\right) \tag{0.7}$$

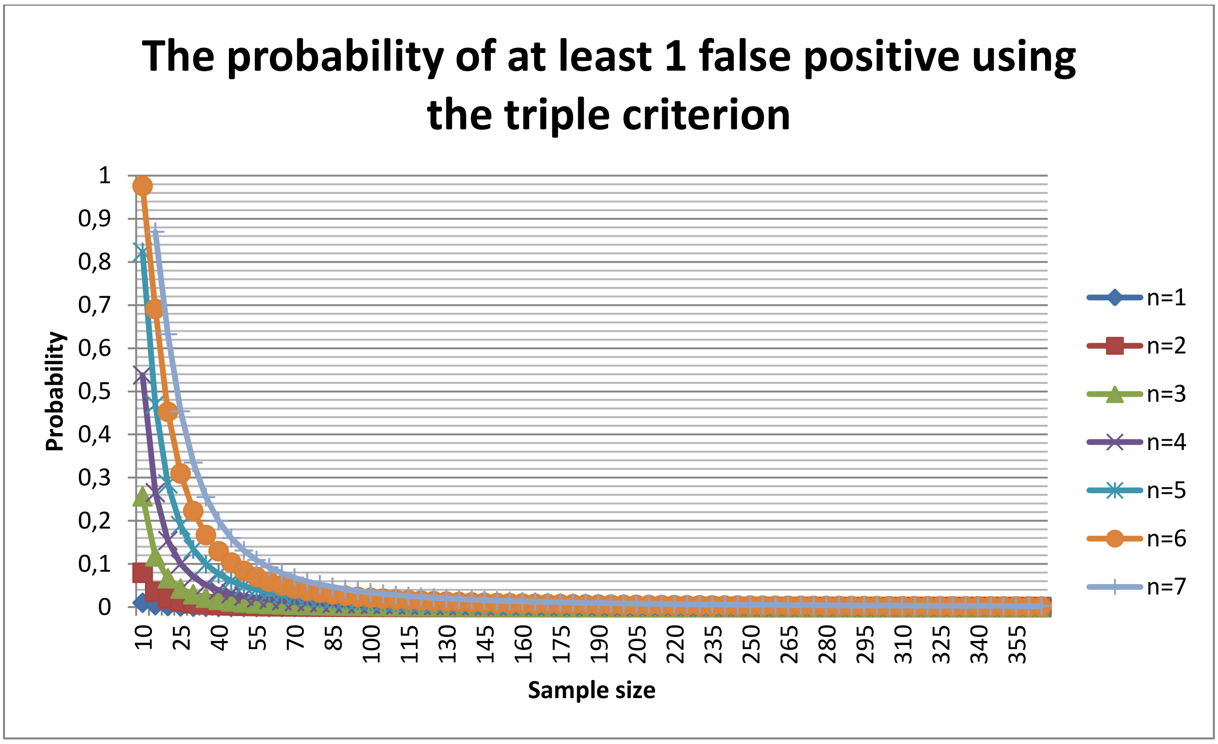
$$Z_{1,2} \in Hyp\left(N,n_2,\frac{n_1}{N}\right) \tag{0.8}$$

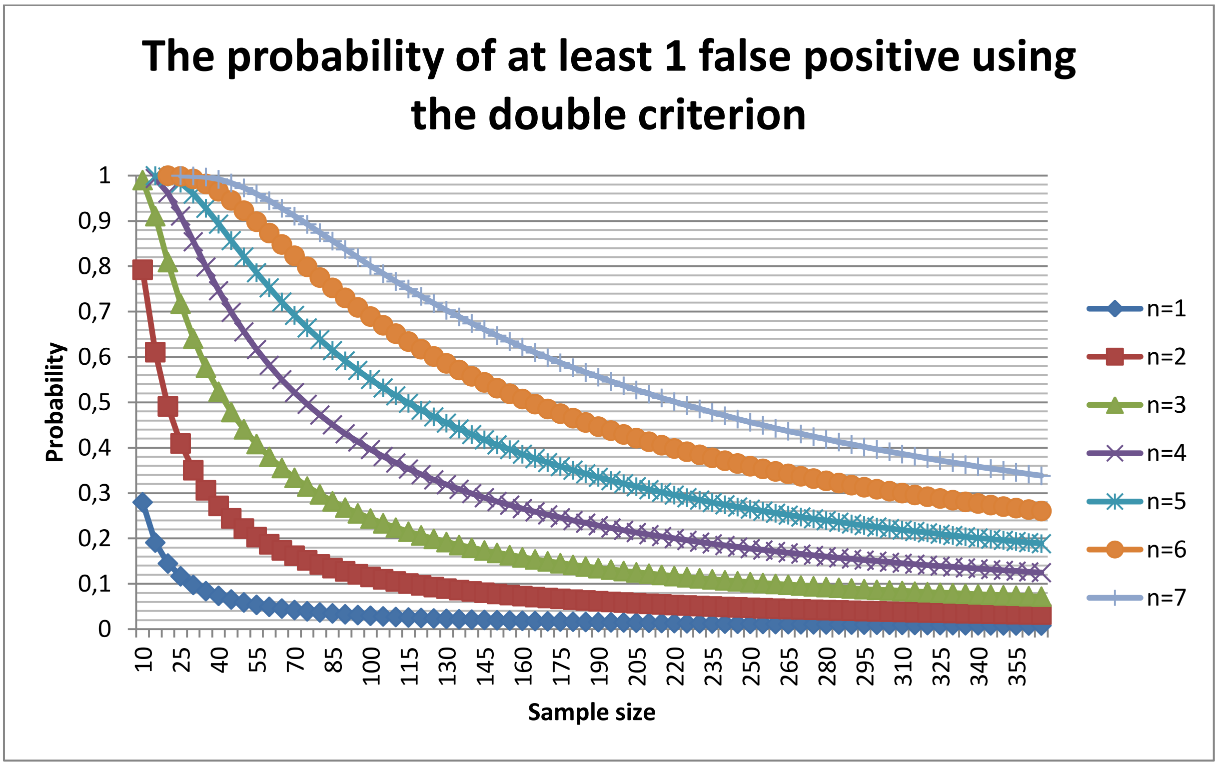
Which gives the probability density function

$$P\left(Z_{1,2,3}=k\right)=\sum_{i=\max(0,k-n_3)}^{\min(n_1,n_2,k)} \frac{\binom{n_1+n_2-2i}{k-i}\binom{N-(n_1+n_2-2i)}{n_3-(k-i)}}{\binom{N}{n_3}}*\frac{\binom{n_1}{i}\binom{N-n_1}{n_2-i}}{\binom{N}{n_2}} \tag{0.9}$$

Probabilities for different sizes of subsamples

The probabilities for different sample sizes are presented below. For a clearer presentation it is assumed that $n_1 = n_2 = n_3 = n$.




The probability of at least 1 false positive using the double criterion
Probability
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
10
25
40
55
70
85
100
115
130
145
160
175
190
205
220
235
250
265
280
295
310
325
340
355
Sample size
n=1
n=2
n=3
n=4
n=5
n=6
n=7

Appendix B – Autocorrelation plots for real life data

In this section plots of autocorrelation and partial autocorrelation for 4 time series are displayed.