Inmaculada López [a,*], Carmelo Rodríguez [a], Manuel Gámez [a], Zoltán Varga [b] and József Garay [c]

[a] Department of Mathematics, University of Almería, La Cañada de S. Urbano, 04120 Almería, Spain; E-Mails: milopez@ual.es (I.L.); crt@ual.es (C.R.); mgamez@ual.es (M.G.)

[b] Institute of Mathematics and Informatics, Szent István University, Páter K. u. 1., H-2103 Godollo, Hungary; E-Mail: Varga.Zoltan@gek.szie.hu

[c] MTA-ELTE Theoretical Biology and Evolutionary Ecology Research Group and Department of Plant Systematics, Ecology and Theoretical Biology, L. Eötvös University, Pázmány P. sétány 1/C H-1117 Budapest, Hungary; E-Mail: garayj@ludens.elte.hu

* Corresponding author; E-Mail: milopez@ual.es; Phone: +34 950015775

Change-Point Method Applied to the Detection of Temporal Variations in Seafloor Bacterial Mat Coverage

Abstract: The paper is aimed at a methodological development of change-point detection, applicable to identify abrupt changes in temporal or spatial data sequences of ecosystem monitoring. In earlier papers we developed a method for the detection of a change in the parameters of a *discrete* distribution, with the simultaneous estimation of the distribution parameters before and after the change.

In the present paper we not only extend this method to the case of *normal distributions*, but also provide a new algorithm for the refining of the estimation of the change-point, based on the following idea: It is intuitively clear that, the more samples are need to distinguish between the two distributions, the more sample elements should be eliminated near the estimated change-point in order to 'clean' the mixed-up samples. The appropriate size of the cut-down part of the sample is analytically calculated for the case of normal distributions. This cleaning is combined with our original change-point detection method. This new algorithm is validated, and applied to the detection of change-points and the parameter estimation of the separated distributions in the time-series data on the bacterial mat coverage of a seafloor area, collected by other authors using a multi-sensor seafloor observatory. Since the normality of the distributions involved is an import condition for the new algorithm, the application of a normality test was also necessary. Our results corroborate the abrupt changes of bacterial mat coverage of a seafloor area, obtained recently by other authors using a different method.

Keywords: change-point detection; maximum likelihood method; time-series; multi-sensor seafloor observatories; bacterial mat coverage

1. Introduction

The statistical detection of abrupt changes (change-points) in time-series data goes back to the initiative of Shewhart, 1931, concerning quality control of industrial production lines. Following the methodological article (Page, 1954), where the so-called cumulative sum (CUSUM) control chart was introduced, and a technically involved branch of mathematical statistics, the change-point analysis has been developed. Important theoretical contributions are summarized in Camarero et al., 2000; Csörgő and Horváth, 1997. For recent surveys on change-point analysis, see Chen and Gupta, 2000; Eckley et al., 2011. Since the developed methodology is appropriate to explore the possible temporal or spatial structure of local homogeneity from collected data, change-point analysis found applications in various fields of science and human activity, ranging from quality control to environmental studies, from economy to biology and medicine. For example, in earlier papers (López et al., 2010, 2012) we applied a change-point method for border or edge detection in the study of patchiness in plant ecology and forest use. We also note that in our method, the type of distributions was known and we estimate their parameters simultaneously with the change-point in an iterative way.

In López et al., 2010, for a given data system (number of individuals of the considered species in each quadrat) collected along a straight line, two areas were considered, where the data of each area came from different *discrete distributions*, with unknown parameters. A method was presented that simultaneously estimated the change-point separating the different distributions and the unknown parameters of the

latter distributions. The proposed algorithm was based on the maximum likelihood method. In addition, another algorithm was implemented to find the so-called change-interval for K, that is, a kind of transition zone where both distributions are mixed and the estimation of the change-point is included with a given probability. In López et al., 2012, this method was applied in the field of forest use, namely, to the analysis of the effect of a gap-cut on the spatial distribution of undergrowth plants and tree seedlings.

While in the above mentioned papers we developed and applied a method for the detection of a change in the parameters of a *discrete* distribution occurred in a data sequence linearly ordered in space, in the present paper we extend this method to the case of *normally* distributed data. In our change-point detection method, at the same time, we also estimate the parameters of the separated normal distributions in an iterative way.

Moreover, we propose a possible improvement of this extended method, based on the following new idea: It is intuitively clear that, the more samples are need to distinguish between the two distributions, the more sample elements should be eliminated near the already estimated change-point in order to clean the 'mixed-up' samples. The appropriate size of the cut-down part of the sample is analytically calculated for the case of normal distribution. Then, from the cleaned sample we get a finer estimate of the separated distributions, and obtain a new estimate for the change-point. We repeat this process until the change-point remains unchanged.

This new algorithm is validated and applied to the detection of change-points in the time-series data on the bacterial mat coverage of a seafloor area, described in Matabos et al., 2011a, and deposited in repository Matabos et al., 2011b. Although the theory of

change-point analysis is mathematically rather involved, we emphasize that our method uses only sophomore statistics.

The paper is organized as follows: In Section 2, the conceptual model is set up. Section 3 is dedicated to the mathematical description of the model and to the validation of the corresponding new algorithm. In Section 4 the experimental data are presented, in Section 5, the results of the application of our method are summarized. Section 6 contains the discussion of the proposed algorithms, obtained results and a short outlook. Finally, as a theoretical background of the proposed method, some mathematical details are presented in the Appendix.

2. Conceptual model

In this paper, similarly to our papers López et al., 2010, 2012, the calculation of the change-point is also based in a maximum likelihood approach. The main difference is that in López et al., 2010, 2012, discrete distributions were considered and here the distributions separated by the obtained change-point are assumed to be normal distributions. It is supposed that there exists a time moment or spatial point along a line where there is a change in the parameters of the distribution, and the question now is when or where this change is produced, in order to understand what took place in this moment of the time or space point that could have affected our data. In this sense, in nature, the detection of a change-point in a data sequence on a given object can help us to understand e.g. how the environment can affect the object in question.

To estimate the change-point K an algorithm is implemented with the help of the statistical software "R" (version 3.1.1.). In López et al., 2010, 2012, for a fixed data position K in time or space, the probability distributions on the left- and right-hand side of the original sample were estimated by the statistic sample proportion. Here, since we suppose that both sides are normally distributed, for a fixed K we estimate the unknown

parameters: mean and standard deviation of both normal distributions by the sample mean and sample standard deviation. Then for this K we calculate the product of the likelihood functions of both estimated distributions. Another difference in relation to the algorithm implemented in the above papers is that now the likelihood function is defined for continuous variables, while previously it was defined for discrete variables. Now, as the estimated change-point, we choose the value K that maximizes the product of the corresponding likelihood functions. Once K is estimated, the estimations of the parameters of both required distributions are also obtained.

Furthermore, in López et al., 2010, 2012, another algorithm was implemented to find the so called change-interval for K, that is, a kind of transition zone where both distributions are mixed, a "change-zone" containing the estimation of the change-point with a given probability. There, this change-interval was built up by an adaptation of the bootstrap method, generating bootstrap samples with the particularity that it consists of two linearly arranged "homogeneous" parts, the original sample is divided into two parts, such that the elements of the original sample are mixed only within these parts. Finally a distribution for the estimates of K is obtained, and the algorithm calculates the required change-interval.

In this paper, we do not construct the analogous algorithm for normal distributions because our purpose is to refine the change-point estimation and not to find a change-zone containing the change-point with a certain probability. Therefore, apart from the algorithm to estimate the change-point for normal distributions, we present another one, implemented in the software "R" in order to improve this estimation. This algorithm is based on the iteration of the change-point estimation obtained from the first algorithm. At first, it is supposed that there is a change-point in the normal distribution parameters, which are unknown. Applying the first algorithm the change-point K is obtained by a

maximum likelihood approach, the original sample is divided in two parts and the parameters of both distributions are estimated. Now we repeat this process but cutting the original sample. We eliminate n elements from the left and right- hand side of the calculated change-point K with the objective of eliminating the elements where we doubt if they come from the first distribution or from the second one, but centering this elimination interval in the estimated K. For the new sample, smaller than the original one and separated in two clearly defined parts, we estimate again the parameters of the left and right distributions from the left- and right-hand sides of the smaller sample, respectively. Then, we apply again the first algorithm to the original sample to estimate the change-point but considering known the parameters of both distributions from these last estimations, and from the new K obtained, we cut again the original sample. We repeat this process until the change-point remains unchanged. However, the question is what sample size n we should eliminate from both sides of the change-point? How should we calculate n? This question can be answered taking into account that normal distributions are considered. We should know what sample size we need to distinguish between two normal distributions. For example, for a general sample, we will establish a hypothesis test where the null hypothesis is that this sample is extracted from a given normal distribution and the alternative hypothesis is that the sample is extracted from another normal distribution. Therefore, two types of errors can be made: type I error is made when we reject the null hypothesis when it is true, and type II error is made when we accept the null hypothesis when it is not true. (Terms type I error and type II error are also used for their probabilities.) If we consider the sum of both errors (total error), the question is the following: Given $\varepsilon > 0$ from what threshold sample size n_0, would it be verified that the total error is smaller than ε? In this way in the Appendix we

calculate the sample size n necessary to distinguish between two normal distributions given a total error.

We note that this is a new approach concerning the algorithm for the calculation of a change-interval of papers López et al., 2010, 2012. There, a sample with the original sample size was always considered, but in the present method we remove the uncertain parts from the original sample to estimate the distribution parameters and consider them as known, and after that we can estimate the change-point again.

A further novelty compared to our previous studies is that here we also show how to deal with the case of several change points.

3. Model description and algorithms

3.1. Model description

In what follows, we will use the time-series terminology, but we emphasize that the construction is also valid for spatially structured data sequences. We consider N sampling times and fix $0<<K<<N$. Suppose that the values of the considered characteristic (observed quantity) collected at sampling times 1, 2, 3,…, K are independent random variables with the same continuous probability distribution $\xi \in N(\mu_1, \sigma_1)$, that is, a normal distribution with mean μ_1 and standard deviation σ_1 ; whereas the characteristic at sampling times K+1, K+2, K+3, …, N are independent random variables with the same continuous probability distribution $\eta \in N(\mu_2, \sigma_2)$.

1	2	…	K- 1	K	K +1	K +2	…	N
ξ	ξ	…	ξ	ξ	η	η	…	η

We also refer us to ξ as the *left distribution* and to η as the *right distribution*.

First, from a given sample vector $X:=(x_1, x_2,\ldots, x_N)$, for each possible K, we estimate distributions of ξ and η , and the likelihood of "realization" of the given sample. Then,

from the possible values of K we obtain the required estimate for K, applying the maximum likelihood approach.

3.2. Estimation of distributions ξ and η

For given $2 \le K \le N\text{-}2$, we estimate the parameters of both distributions in the same way. Let

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{K} x_i}{K}, \quad \hat{\sigma}_1 = \sqrt{\frac{\sum_{i=1}^{K} (x_i - \hat{\mu}_1)^2}{K - 1}}; \qquad \hat{\mu}_2 = \frac{\sum_{i=K+1}^{N} x_i}{N - K}, \quad \hat{\sigma}_2 = \sqrt{\frac{\sum_{i=K+1}^{N} (x_i - \hat{\mu}_2)^2}{N - K - 1}}. \tag{1}$$

be the corresponding sample means and standard deviations. Then, we estimate the left normal distribution by a $N(\hat{\mu}_1, \hat{\sigma}_1)$, and the right normal distribution by a $N(\hat{\mu}_2, \hat{\sigma}_2)$.

Let

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

be the probability density function of a normal distribution $N(\mu, \sigma)$. Then, given a sample $X = (x_1, x_2, \ldots, x_n)$ obtained from a population with normal distribution $N(\mu, \sigma)$, the likelihood function is

$$l(\mu, \sigma \mid X) = \prod_{i=1}^{n} f(x_i; \mu, \sigma).$$

Since our sample X consists of two parts, the left part, $X_{lK} = (x_1, \ldots, x_K)$, and the right part $X_{rK} = (x_{K+1}, \ldots, x_N)$, extracted from the left and right distributions, respectively, and both distributions have different parameters, let us consider the

likelihood of "realization" of the sample *X*, calculated as the product of the corresponding left and right likelihood functions

$$l_K := l(\mu_1, \sigma_1 \mid X_{lK}) \cdot l(\mu_2, \sigma_2 \mid X_{rK}).$$

This function l_K will be considered as the "goodness" of *K*. Based on the given sample *X*, our purpose is to find a *K* which maximizes l_K, providing the "best" (i.e. the "most likely") value of *K*. We will deal with this in the next subsection.

3.3. Algorithms

Algorithm 1 (Estimation of the change-point *K*):

1. Introduce sample *X*. *N*:= Size (*X*).
2. FOR *K*=2 until *N*-2:
   a) Calculate: $\hat{\mu}_1, \hat{\sigma}_1, \hat{\mu}_2, \hat{\sigma}_2$, according to (1).
   b) Calculate:

$$Log\ l_K := Log\ l(\hat{\mu}_1, \hat{\sigma}_1 \mid X_{lK}) + Log\ l(\hat{\mu}_2, \hat{\sigma}_2 \mid X_{rK}) =$$
$$= \sum_{i=1}^{K} Log\ f(x_i; \hat{\mu}_1, \hat{\sigma}_1) + \sum_{s=K+1}^{N} Log\ f(x_s; \hat{\mu}_2, \hat{\sigma}_2).$$

(It is supposed that the left part of the sample is obtained from a normal distribution $N(\hat{\mu}_1, \hat{\sigma}_1)$ and the right part of the sample is extracted from a normal distribution $N(\hat{\mu}_2, \hat{\sigma}_2)$.)

3. *LogLikelihood:= (Log* l_2*,…,Log* l_{N-2}*).*
4. *EstimateK:=* [Position with maximum value among the coordinates of *LogLikelihood*] + 1
5. Return *EstimateK.*

If we are also interested in the estimation of the left and right distributions, we can calculate the corresponding estimated parameters $\hat{\mu}_1, \hat{\sigma}_1, \hat{\mu}_2, \hat{\sigma}_2$, according to (1), for *K*:= *EstimateK*.

Algorithm 2 (Refining the estimation of the change-point *K*):

1. Introduce sample *X*. *N* := Size (*X*).
2. We apply Algorithm 1 to the sample *X*, to obtain an estimate K_0 for the change-point.
3. We estimate the parameters of the left and right distributions, $\hat{\mu}_1, \hat{\sigma}_1, \hat{\mu}_2, \hat{\sigma}_2$, according to (1) from the obtained $K=K_0$.
4. Introduce the error ε, see Appendix. (This error is a bound for the sum of the probabilities of both type I and II errors).
5. a) Calculate *n* from $\hat{\mu}_1, \hat{\sigma}_1, \hat{\mu}_2, \hat{\sigma}_2$ and ε, see Appendix for this calculation.

b) It is intuitively clear that, the more samples are need to distinguish between the two distributions, the more sample elements should be eliminated near K_0 in order to clear the mixed up samples. Therefore is at hand to eliminate *n* sample elements from both the left and the right hand sides of change-point K_0, and from the remaining part of the sample, $X_n = (x_1, \ldots, x_{K_0-n-1}, x_{K_0+n+1}, \ldots, x_N)$ estimate again the left and right distributions:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{K_0-n-1} x_i}{K_0 - n - 1}, \quad \hat{\sigma}_1 = \sqrt{\frac{\sum_{i=1}^{K_0-n-1} (x_i - \hat{\mu}_1)^2}{K_0 - n - 2}};$$
$$\hat{\mu}_2 = \frac{\sum_{i=K_0+n+1}^{N} x_i}{N - K_0 - n}, \quad \hat{\sigma}_2 = \sqrt{\frac{\sum_{i=K_0+n+1}^{N} (x_i - \hat{\mu}_2)^2}{N - K_0 - n - 1}}. \qquad (2)$$

c) Apply again Algorithm 1 to the complete sample *X*, but now changing the calculation of Step 2a), that is, we keep the previously calculated values of $\hat{\mu}_1, \hat{\sigma}_1, \hat{\mu}_2, \hat{\sigma}_2$ according to (2), for this application of Algorithm 1. Therefore we obtain the change-point *K* supposing that the left and right distributions are $N(\hat{\mu}_1, \hat{\sigma}_1)$, $N(\hat{\mu}_2, \hat{\sigma}_2)$, respectively. That is, we will calculate the change-point for the complete sample but supposing known parameters for both distributions, what we have previously estimated from the original sample without the elements $(x_{K_0-n}, \ldots, x_{K_0+n})$, according to (2).

d) IF $K \neq K_0$

$K_0 := K$

REPEAT Step 5

ELSE

RETURN *K*.

Search for more than one change-point

If we want to find more than one change-point, once we have obtained the change-point *K* from the previous algorithms, we would apply them again to the left and right samples independently, obtaining two new change-points K_l and K_r. Then we can have, in total, three change-points and four new parts of the complete sample. In principle, we can repeat this process for each sample piece independently until the following stop criterion: we do not consider the last obtained change-point of a sample piece when the size of one of the two new obtained parts of the corresponding sample piece is too small or the field researcher thinks that it would be appropriate to stop the procedure for a particular reason, according to the kind of collected data or establish an own stop criterion for the given data set.

3.4. Validation of the Algorithms

In order to validate the presented methods, we will generate several samples from different left and right normal distributions such that we know previously, which is the theoretical change-point value. After that, we will consider that we do not know, which are the left and right distribution parameters, from which the sample has been obtained; and the change-point is also unknown. We will also suppose that there is an only change-point and calculate it applying Algorithms 1 or 2.

Samples obtained from normal distributions with equal variances

a) If we generate a sample of size 13500 in a random way, where the left-hand side of the sample, concretely the first 7500 elements, are obtained from a normal distribution $N(1,1)$ and the rest of elements (the right-hand side) are obtained from a normal distribution $N(3,1)$, obviously the theoretical change-point is 7500. Applying only Algorithm 1 we obtain K=7500.

If the sample size is not so large, the means of the distributions are closer and the variances are large enough as to not distinguish so easily the change-point, it may be necessary to improve Algorithm 1, as we have done it obtaining Algorithm 2. We will show this in the following example.

b) We generate a sample of size 135 randomly, where the left-hand side of the sample, concretely the first 75 elements, are obtained from a normal distribution $N(1,1)$ and the rest of elements (the right-hand side) from a normal distribution $N(2,1)$. Therefore, the theoretical change-point is 75. The whole sample is given in Table 1.

TABLE 1

Applying only Algorithm 1 we obtain K=83. If we apply Algorithm 2 the estimate of change-point is much better, K=76.

Samples obtained from normal distributions with different variances

a) We generate a sample of size 4500 in a random way, the first 2500 elements from a distribution $N(1,4)$ and the rest from a distribution $N(7,6)$. Then K=2500. If we apply Algorithm 1, we obtain K=2500.

Even sometimes when apparently we could have more mixed elements from both distributions, due to the values of the means and variances, Algorithm 1 works very well, as we can judge from the following example.

b) The left-hand side of the sample, that is, the first 1000 elements are randomly generated from a $N(1,2)$ and the 800 elements of the right-hand side from a $N(3,4)$. The theoretical change-point is 1000 and applying Algorithm 1 to the whole sample, the estimate K is 1000.

When the size of the sample is not so large and means and variances do not allow distinguish well between both distributions, sometimes Algorithm 1 needs an improvement, carried out in Algorithm 2.

c) In this case the first 100 elements are randomly generated from a $N(1,2)$ and the 40 elements of the right-hand side from a $N(3,4)$. Obviously K=100. The whole sample is given in Table 2.

TABLE 2

Algorithm 1 provides an estimate for K equal to 103. Algorithm 2 improves this estimate, resulting in K=99.

4. Experimental data

The developed change-point methodology can be applied in the analysis of temporal or spatial data sequences in a wide range of fields, for the monitoring of agro-ecological and forest systems, aquatic ecosystems, etc. In the present paper we illustrate the efficiency of our method applying it for the detection of change-points in the “ready-made” time-series data on the bacterial mat coverage of a seafloor area. The data we

will use have been collected by the authors of Matabos et al., 2011a, and made available at Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.db2gd, see Matabos et al., 2011b. Since we use these data for the illustration of our methodology, below we only shortly summarize the circumstances of data collection, for a complete description of the experiments we refer the reader to Matabos et al., 2011a,b.

For the study of biological cycles in benthic ecosystems, the VENUS multi-sensor cabled seafloor observatory had been established in deep-water environment in Saanich Inlet, British Columbia, Canada. Three species were observed by a remotely operated digital camera, providing abundances of shrimp (Spirontocaris spp.) and squat lobster (Munida quadrispina) and bacterial mat coverage (Beggiatoa spp.).

We will only deal with the bacterial mat coverage. The latter was registered at hourly intervals during three periods: November 2–9, 20–23 and November 30 to December 4, in 2009, related to the changes in the abiotic environmental data.

5. Results

In the experimental situation shortly described in the previous section, we apply our change-point estimation method using the time-series data of Table 3.

TABLE 3

The first observation corresponds to November 2, 16:00 hrs, the next observation to 17:00 and the rest of the observations were taken hourly during the same day and consecutive days until the last considered observation taken on November 9, 8:00 hrs.

If we apply Algorithm 1 to these data, we obtain that there is a change-point at K=28, and applying Algorithm 2 no improvement of this value is obtained. This change-point corresponds to November 3, 19:00 hrs.

If we want to see where there is another change of distribution, and we apply again Algorithm 1 only for the right-hand side of the sample, we obtain that there must be

another change-point for K_r =77, that is, the second change-point for the complete sample would be at K_2 =105, that is, on November 7, at 0:00.

In many statistical procedures normal distribution of the involved samples is required. Therefore, it is very important to check for this normality assumption because if it is violated, interpretation and inference may not be reliable or valid. For this reason, we have checked normality applying three of the most common normality tests (Shapiro-Wilk, Lilliefors (Kolmogorov-Smirnov) and Anderson-Darling), being Shapiro-Wilk test the most powerful normality test of them, see Nornadiah and Yap, 2011. These formal normality tests support graphical methods as the normal quantile-quantile plot (QQ-plot) that we present next. As we can see in Figure 1, in the resulting plot there are substantial deviations from a straight line, what means that the complete sample does not proceed from a normal distribution, as the formal normality tests will confirm. In Figure 2 we can observe how if we divide this original sample in three subsamples according to the two obtained change points, the corresponding resulting plots are approximately linear, which means that these three subsamples proceed from normal distributions as the previous normality tests will confirm.

FIGURE 1

FIGURE 2

If we apply the Shapiro-Wilk normality test to the whole sample with a significance level $\alpha = 0.05$, the p-value obtained is $3.562 \cdot 10^{-8}$, (applying Lilliefors test for normality, p-value = $3.569 \cdot 10^{-5}$), which for both normality tests means that there is a sample evidence to reject the normality of the whole data set. However if we use the information obtained previously and consider two samples, one from the first element until position K=28 and the other one the rest of the sample, the Shapiro-Wilk test for

normality applied to both samples separately provides the following p-values, 0.4234 and 0.1364, for the first and second samples, respectively, (for Lilliefors test the corresponding p-values are 0.623 and 0.2771), indicating both tests to accept that both data sets proceed from normal distributions. If we divide the second sample in two parts, according to the obtained K_r =77, the Shapiro-Wilk test applied to these two last samples separately provides p-values equal to 0.9507 and 0.5213, respectively (0.8555 and 0.2328, respectively, for Lilliefors test). This means that we can also accept that considering these three samples, the three data sets proceed from three normal distributions. The same conclusions were obtained when in a similar way we applied the Anderson-Darling normality test to all the considered samples.

The estimate of these three bacterial mat coverage distributions by the sample means and sample standard deviations are the following:

$$N(12.36534,4.83452),\ N(7.051384,2.693788),\ N(4.631949,1.834058).$$

Summarizing, we have accepted that the data corresponding to the percentage of bacterial mat coverage during the period November 2-9 do not proceed from an only normal distribution. Normality tests have proved that the data could proceed from the previous three normal distributions. At this moment it seems interesting to check through hypotheses tests and confidence intervals the values of their means.

From November 2, 16:00 hrs until November 3, 19:00 hrs, the data proceed from a normal distribution with mean 12.36534. A hypothesis test to check if the mean is this value or not provide a p-value equal to 1, that is, there is no sample evidence to reject that the mean is this value and the 95% confidence interval for the mean of the normal distribution is [10.49071, 14.23997].

From November 3, 20:00 hrs until November 7, at 0:00, the data proceed from a normal distribution with mean 7.051384, providing the same conclusion for the

corresponding hypothesis test (p-value = 1) and the 95% confidence interval for the mean of the normal distribution is [6.439969, 7.662799].

From November 8, 1:00 hrs until the end of the period, November 9, 8:00 hrs, the data proceed from a normal distribution with mean 4.631949, providing the same conclusion for the corresponding hypothesis test (p-value = 1) and the 95% confidence interval for the mean of the normal distribution is [4.140785, 5.123113].

We can observe how the mean of the normal distributions has gone decreasing.

For a comparison with other approaches we recall that to deal with the uncertainty of the change point, either a confidence interval for the change-point estimate was calculated (e.g. in Wang and Wang,1994), or the change-interval is constructed (see López et al., 2010, 2012). In our present approach the uncertainty of the change-point was taken into account in the cleaning procedure of our Algorithm 2. Of course, as we have shown in the Validation section 3.4, the cleaning may improve the estimate of the change-point (especially in case of relatively small samples), or just leave it unchanged, depending on the size of the concrete data set, and on the closeness of the parameters of the involved normal distributions. A disadvantage of our method might be that, at the present stage, it is developed only for normal distributions. Nevertheless, in environmental monitoring, samples from continuous variables often give positive answer to normality tests, as it was the case in our application to seafloor bacterial mat coverage data.

6. Conclusion

Change-point method is a powerful tool for detecting changes in space or time. In particular, our proposed change-point estimation method turned out to be efficient, not

only in previous cases of spatially structured data (see edge detections carried out in López et al., 2010, 2012), but also in the case of time-series data.

The extension of our change-point detection method to normal distributions, developed in the present paper (Algorithm 1) opens the way to a large scale of applications, in particular in environmental studies where normal distribution often occurs.

Under the normality assumption on the distributions separated by the change-point, a further novelty is a new additional method (Algorithm 2) that may improve the estimation of the change point K_0 already estimated by Algorithm 1. In fact, using this K_0 and Algorithm 2, we can clean the original sample eliminating a certain number n of sample elements near K_0, and from this cleaned sample we estimate again the left and right distributions and then calculate the change-point from the original sample by Algorithm 1. In fact, Algorithm 2 consists in the iterative combination of Algorithm 1 and the cleaning procedure. Examples used for the validation of Algorithm 2 show that the latter really improves the estimate of the change-point. It is also seen that this does not happen always, but anyway it is worth it to try.

For a comparison with other methods used to detect of abrupt changes in time-series, first of all we remind that originally, in industrial production lines, control charts have been introduced to detect changes. As the overview by Taylor (2000) pointed out, control charting and the more recent change-point method should be considered as complementary tools, since the first one has the advantage to work online, the latter one, although needs data about the whole process, offers a deeper insight to the process in question.

For a comparison with other change-point detection methods, we call the attention to the fact that our method needs an *a priori* knowledge on the type of the distribution. For

an overview of nonparametric methods, where no such knowledge is supposed, see Brodsky and Darkhovsky, 1993, and Cheng, 2012, 2013.

We also note that the intuitive and elementary way we deal with the case of several change-points, turned out to be efficient in the considered environmental application. For a theoretically elaborated approach to the multiple change-point case see e.g. Hawkins, 2001.

Although the application to time-series data on bacterial mat coverage was intended to illustrate the extension of this method from discrete to normally distributed variables, it also corroborates certain observations of Matabos et al., 2011a. In fact, applying cross-correlation analysis, they showed that the bacterial mat coverage was significantly correlated with oxygen concentration in the water. Depending on the time lag considered after a change in dissolved oxygen concentration a weak but significant correlation is obtained, for instance, $r = -0.27$, for 6 hour lag. Namely, following a major oxygen intrusion, they found a rapid disappearance of bacterial mats. This disappearance coincided with a rapid increase in shrimp abundance in the highly oxic environment, which might be a feeding impact on the bacterial mats. Another option to explain the disappearance of Beggiatoa spp. mats is that they migrate downward (and out of sight) to avoid high oxygen levels. In any case, the question remains open, which one is the real (or the dominant) cause of the observed phenomenon. The results of our change-point analysis, to some extent, also contributes to the study of this problem: Before the observed major oxygen intrusion, our method also provided two change-points (each of them follows a local maximum of the oxygen concentration, see Figure 2 of Matabos et al., 2011a), separating normal distributions and at each change-point the mean value changes to a smaller one, giving an insight to the effect of minor peaks in oxygen concentration. For a complex automated image analysis for the detection

bacterial mat coverage, based on the data collected in the VENUS Undersee Cabled Observatory, see Aguzzi et al., 2011.

It should be remarked that we continued searching for further change-points inside these three samples. However, going on with the procedure, we obtained too small subsamples, so we stopped the search, keeping the previously obtained two change-points as final results.

Finally, for an outlook we note that the developed change-point methodology can be also applied to temporal or spatial data sequences for the monitoring of epibenthic marine ecosystems, or similarly, for the detection of heterogeneities in certain terrestric ecosystems, see e.g. Healey et al., 2014; Boluwade and Madramootoo, 2015.

Acknowledgements

The research was supported by the Regional Government of Andalusia (Spain), Programme of Excellence Projects (ref: P11-TIC-7821) of the Junta de Andalusia, Consejería de Economía, Innovación y Ciencia, with joint financing from FEDER Funds, and by the Spanish Ministry of Economy and Competitiveness, under project No. MTM2010-20774-C03-03 and by EFRD (FEDER) funds. The authors also thank the Editor and the anonymous Referees for their helpful suggestions to improve the manuscript.

References

Aguzzi, J., Costa, C., Robert, K., Matabos, M., Antonucci, F., Juniper, S.K. and Menesatti, P. (2011). Automated Image Analysis for the Detection of Benthic

Crustaceans and Bacterial Mat Coverage Using the VENUS Undersea Cabled Network. Sensors 11(11), 10534-10556.

Boluwade, A. and Madramootoo, C. (2015). Determining the Influence of Land Use Change and Soil Heterogeneities on Discharge, Sediment and Phosphorus. Journal of Environmental Informatics, 25(2), 126-135.

Brodsky, E. and Darkhovsky, B. S. (1993). Nonparametric Methods in Change Point Problems. Springer, New York

Camarero, J.J., Gutiérrez, E. and Fortin, M.J. (2000). Boundary detection in altitudinal treeline ecotones in the Spanish Central Pyrenees. Arct. Antarct. Alp. Res. 32, 117–126.

Chen, J. and Gupta, A.K. (2000). Parametric Statistical Change Point Analysis. Birkhauser.

Cheng, Z. (2012). Using LS-SVM Pattern Recognizer to Detect Change-Point in ARMA Process, Applied Mechanics and Materials, 271-272, 1731.

Cheng, Z. (2013). An intelligent method of change-point detection based on LS-SVM algorithm. HKIE Transactions, 20.3: 141-147.

Csörgő, M. and Horváth, L. (1997). Limit Theorems in Change-Point Analysis. Wiley, Chichester.

Eckley, I.A., Fearnhead, P. and Killick, R. (2011). Analysis of Change point Models, in: Barber, D., Cemgil, A.T., Chiappa, S. (Eds.), Bayesian Time Series Models. Cambridge University Press.

Hawkins, D. M. (2001), Fitting multiple change-point models to data. Computational Statistics & Data Analysis Volume 37, Issue 3, 28 September, 323–341.

Healey, N.C., Oberbauer, S.F., Ahrends, H.E., Dierick, D., Welker, J.M., Leffler, A.J., Hollister, R.D., Vargas, S.A. and Tweedie, C.E. (2014). A Mobile Instrumented Sensor

Platform for Long-Term Terrestrial Ecosystem Analysis: An Example Application in an Arctic Tundra Ecosystem. Journal of Environmental Informatics, 24(1), 1-10.

López, I., Gámez, M., Garay, J., Standovár, T. and Varga, Z. (2010). Application of change-point problem to the detection of plant patches. Acta Biotheor 58, 51-63. DOI 10.1007/s10441-009-9093-x

López, I., Standovár, T., Garay, J., Varga, Z. and Gámez, M. (2012). Statistical detection of spatial plant patterns under the effect of forest use. International Journal of Biomathematics 5 (6), 1250054 (15 pages).

Matabos, M., Aguzzi, J., Robert, K., Costa, C., Menesatti, P., Company, J.B. and Juniper, S.K. (2011a). Multi-parametric study of behavioural modulation in demersal decapods at the VENUS cabled observatory in Saanich Inlet, British Columbia, Canada. J. Exp. Mar. Biol. Ecol. 401(1-2): 89-96. http://dx.doi.org/10.1016/j.jembe.2011.02.041

Matabos, M., Aguzzi, J., Robert, K., Costa, C., Menesatti, P., Company, J.B. and Juniper, S.K. (2011b). Data from: Multi-parametric study of behavioural modulation in demersal decapods at the VENUS cabled observatory in Saanich Inlet, British Columbia, Canada. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.db2gd

Nornadiah Mohd Razali and Yap Bee Wah (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. Journal of Statistical Modeling and Analytics, 2 (1), 21-33.

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

Shewhart, W.A. (1931). Economic Control of Quality of Manufactured Products. Van Nostrand, New York and MacMillan, London.

Taylor, W. (2000). Change-Point Analysis: A Powerful New Tool for Detecting Changes. Deerfield, IL: Baxter Healthcare Corporation.

http://www.variation.com/cpa/tech/changepoint. html

Wang, Jing-Long and Wang, Jin (1994). The test and confidence interval for a change-point in mean vector of multivariate normal distribution. Multivariate analysis and its applications, Institute of Mathematical Statistics, Hayward, CA, 397—411.

Appendix

In this Appendix we explain how we calculate the sample size *n* used in Algorithm 2, Step 5a).

Let us assume we have an *n*-sample $(x_1,...x_n)$, which is homogeneous. We know that this sample is taken either from a normal distribution $\xi \in N(\mu_1,\sigma_1)$ or from another normal distribution $\eta \in N(\mu_2,\sigma_2)$. Suppose that $\mu_1, \mu_2 \in R$ with $\mu_1 < \mu_2$ and $\sigma_1, \sigma_2 > 0$.

We have to find out, whether our *n*-sample is taken either from ξ or η. Let us suppose firstly that σ_1, σ_2 are equal, but keep the distinctive notation. We consider two hypotheses:

H_0: the sample is taken from ξ, that is, the population mean is μ_1;

H_1: the sample comes from η, that is, the population mean is μ_2.

We use a statistic $S_1 : R^n \to R$ and let us denote by $Q_r(\alpha)$ the rejection region and $Q_a(\alpha)$ the acceptance region.

Type I error is

$$P\left[S_1(x_1,...x_n) \in Q_r(\alpha) \middle| H_0 \text{ is true}\right] = \alpha .$$

Type II error is

$$P\left[S_1(x_1,...x_n) \in Q_a(\alpha) \middle| H_1 \text{ is true}\right] = \beta(\alpha) .$$

For each fixed sample size *n* and significance level α we have a total error: $E : n \times R \to R$,

$$E(n,\alpha)=\alpha+\beta(\alpha).$$

The question is, for a fixed n, where is the minimum of $E(n,\alpha)$ attained?

If we consider that we have a sample of size 1, it has sense that the rejection region was of the form $Q_r(\alpha)=[X>y]$. Therefore:

Type I error is $\int_y^{+\infty} f(x)dx$, where f is the probability density function of a $N(\mu_1,\sigma_1)$,

Type II error is $\int_{-\infty}^{y} g(x)dx$, where g is the probability density function of a $N(\mu_2,\sigma_2)$.

We try to find out which y would minimize the sum of both errors. Let us denote

$$\Lambda(y)=\int_y^{+\infty} f(x)dx+\int_{-\infty}^{y} g(x)dx=\int_y^{+\infty}\frac{1}{\sigma_1\sqrt{2\Pi}}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}dx+\int_{-\infty}^{y}\frac{1}{\sigma_2\sqrt{2\Pi}}e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}dx.$$

That is, we want to find the value of y such that $\Lambda'(y)=0$ and y is a minimum:

$$\Lambda'(y)=-f(y)+g(y)=\frac{-1}{\sigma_1\sqrt{2\Pi}}e^{-\frac{(y-\mu_1)^2}{2\sigma_1^2}}+\frac{1}{\sigma_2\sqrt{2\Pi}}e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}=0$$

Case 1: *Suppose equal variances,* $\sigma_1=\sigma_2=\sigma$.

It is easy to prove that $y=\frac{\mu_2+\mu_1}{2}$ verifies $\Lambda'(y)=0$ and $\Lambda''(y)>0$:

Therefore $y=\frac{\mu_2+\mu_1}{2}$ is a minimum point of $\Lambda(y)$.

Case 2: *Suppose different variances,* $\sigma_1\neq\sigma_2$.

It is not difficult to prove that in this case

$$y=\frac{\mu_1\sigma_2^2-\mu_2\sigma_1^2+\sigma_1\sigma_2\sqrt{(\mu_2-\mu_1)^2+\ln\left(\frac{\sigma_1}{\sigma_2}\right)^{2(\sigma_1^2-\sigma_2^2)}}}{\sigma_2^2-\sigma_1^2}$$

1 is a minimum point of $\Lambda(y)$.

2 Now our aim is the following:

3 Given these two distributions

4 $$\xi(n) := \sum_{i=1}^{n} \xi_i \in N\left(n\mu_1, \sigma_1\sqrt{n}\right),\ \ \eta(n) := \sum_{i=1}^{n} \eta_i \in N\left(n\mu_2, \sigma_2\sqrt{n}\right),$$

5 where we suppose $\mu_1 < \mu_2$ and $\sigma_1, \sigma_2 > 0$, see how the error depends on μ_1, μ_2,

6 σ_1, σ_2 :

7 $$\Lambda(y) = \int_y^{+\infty} f(x)dx + \int_{-\infty}^{y} g(x)dx = \int_y^{+\infty} \frac{1}{\sigma_1\sqrt{2\Pi n}} e^{-\frac{(x-n\mu_1)^2}{2\sigma_1^2 n}} dx + \int_{-\infty}^{y} \frac{1}{\sigma_2\sqrt{2\Pi n}} e^{-\frac{(x-n\mu_2)^2}{2\sigma_2^2 n}} dx$$

8 <u>Case 1:</u> $\sigma_1 = \sigma_2 = \sigma$ and $y = \frac{n(\mu_2+\mu_1)}{2}$.

9 We have that the total error in function of n (we will denote it by $E(n)$) would be:

10 $$E(n) = \int_{\frac{n(\mu_2+\mu_1)}{2}}^{+\infty} f(x)dx + \int_{-\infty}^{\frac{n(\mu_2+\mu_1)}{2}} g(x)dx =$$

11 $$= \int_{\frac{n(\mu_2+\mu_1)}{2}}^{+\infty} \frac{1}{\sigma\sqrt{2\Pi n}} e^{-\frac{(x-n\mu_1)^2}{2\sigma^2 n}} dx + \int_{-\infty}^{\frac{n(\mu_2+\mu_1)}{2}} \frac{1}{\sigma\sqrt{2\Pi n}} e^{-\frac{(x-n\mu_2)^2}{2\sigma^2 n}} dx =$$

12 $$= 1 - P\left(\xi(n) \le \frac{n(\mu_2+\mu_1)}{2}\right) + P\left(\eta(n) \le \frac{n(\mu_2+\mu_1)}{2}\right) = 1 - P\left(Z \le \frac{\frac{n(\mu_2+\mu_1)}{2} - n\mu_1}{\sigma\sqrt{n}}\right)$$
$$+ P\left(Z \le \frac{\frac{n(\mu_2+\mu_1)}{2} - n\mu_2}{\sigma\sqrt{n}}\right) = 1 - P\left(Z \le \frac{n(\mu_2-\mu_1)}{2\sigma\sqrt{n}}\right) + P\left(Z \le \frac{n(\mu_1-\mu_2)}{2\sigma\sqrt{n}}\right) =$$

13 $$= P\left(Z \le \frac{n(\mu_1-\mu_2)}{2\sigma\sqrt{n}}\right) + P\left(Z \le \frac{n(\mu_1-\mu_2)}{2\sigma\sqrt{n}}\right) = 2P\left(Z \le \frac{n(\mu_1-\mu_2)}{2\sigma\sqrt{n}}\right).$$

14 where Z follows a distribution N(0,1).

Which would be the inverse function of $E(n)$? Our purpose is the following: given an error $\varepsilon > 0$ we want to obtain the corresponding n_0 such that $E(n) < \varepsilon$ for all $n > n_0$. Then we have that

$$P\left(Z \le \frac{n(\mu_1 - \mu_2)}{2\sigma\sqrt{n}} \right) = \frac{\varepsilon}{2}$$

Using the *qnorm* function of statistic software "R" we can obtain the corresponding quantile for a distribution N(0,1), then we have

$$qnorm\left(\frac{\varepsilon}{2}\right) = \frac{n(\mu_1 - \mu_2)}{2\sigma\sqrt{n}}.$$

Therefore, given the values of ε, μ_1, μ_2, σ,

$$n_0 = \left(\frac{2 \cdot \sigma \cdot qnorm\left(\frac{\varepsilon}{2}\right)}{\mu_1 - \mu_2} \right)^2.$$

Case 2: $\sigma_1 \neq \sigma_2$ and

$$y = \frac{n(\mu_1\sigma_2^2 - \mu_2\sigma_1^2) + \sigma_1\sigma_2\sqrt{n^2(\mu_2 - \mu_1)^2 + \ln\left(\frac{\sigma_1}{\sigma_2}\right)^{2n(\sigma_1^2 - \sigma_2^2)}}}{\sigma_2^2 - \sigma_1^2}.$$

In this case the total error is

$$E(n) = 1 - P(\xi(n) \le y) + P(\eta(n) \le y) = 1 - P\left(Z \le \frac{y - n\mu_1}{\sigma_1\sqrt{n}} \right) + P\left(Z \le \frac{y - n\mu_2}{\sigma_2\sqrt{n}} \right)$$

Again our objective is to obtain a value of n_0 that assure that given the error ε, $E(n) < \varepsilon$ holds for all $n > n_0$. Since we do not know which probability is greater, if $1 - P\left(Z \le \frac{y - n\mu_1}{\sigma_1\sqrt{n}} \right)$ or $P\left(Z \le \frac{y - n\mu_2}{\sigma_2\sqrt{n}} \right)$, it is guaranteed that

$$E(n) \le 2\max\left(1 - P\left(Z \le \frac{y - n\mu_1}{\sigma_1\sqrt{n}}\right), P\left(Z \le \frac{y - n\mu_2}{\sigma_2\sqrt{n}}\right)\right).$$

Then, given an error ε, we want to obtain for each one of the above probabilities a value of n, choosing finally the largest one. Let us search them in this way:

$$P\left(Z \le \frac{y - n\mu_1}{\sigma_1\sqrt{n}}\right) = 1 - \frac{\varepsilon}{2}$$
$$P\left(Z \le \frac{y - n\mu_2}{\sigma_2\sqrt{n}}\right) = \frac{\varepsilon}{2}.$$

Then with the *qnorm* function of statistic software "R" we have that

$$qnorm\left(1 - \frac{\varepsilon}{2}\right) = \frac{y - n\mu_1}{\sigma_1\sqrt{n}}$$
$$qnorm\left(\frac{\varepsilon}{2}\right) = \frac{y - n\mu_2}{\sigma_2\sqrt{n}}.$$

And taking into consideration the value of y in function of n, we solve these two previous equations with the help of the software "R", obtaining two values of n and choosing the greater one, denoted note by n_0.

Then, for both cases (equal or different variances), given an error ε, we can calculate a sample size n_0 such that, $E(n) < \varepsilon$ for all $n > n_0$. In Algorithm 2 Step 5a) we will choose $n = round(n_0) + 1$.

Table 1. Randomly generated samples with equal variances

Time	Sample
1-14	-0.63, 1.55, 2.87, 0.39, 0.23, 0.33, 0.26, 1.35, 0.57, 0.53, 2.88, 1.19, 1.35, 0.29,
15-28	-0.92, -0.26, 0.25, 0.99, 0.28, -0.02, 1.71, 2.10, 0.71, -0.20, 1.28, 0.67, -1.25, 1.67,
29-42	1.15, -0.45, 1.13, 2.04, 3.07, 1.29, 0.78, 0.78, -0.14, 1.75, 1.66, 0.92, 0.44, 1.54,
43-56	0.10, 0.67, 1.04, 1.46, 1.57, 1.15, 1.05, -0.03, 0.12, -1.39, 1.27, 1.34, 0.42, 2.21,
57-70	2.05, 0.97, -0.09, 0.45, 1.33, 1.97, -0.79, 1.51, 0.91, -0.04, 0.69, 1.86, 2.07, 1.23,
71-84	1.43, 0.48, 2.80, 0.94, -1.56, 0.98, 2.79, 2.34, 0.55, 0.59, 1.84, 0.60, 0.65, 3.83,
85-98	0.24, 1.29, 1.64, 2.33, 3.38, 1.77, 1.74, 2.53, 1.71, 3.52, 0.11, 1.27, 2.22, 4.00,
99-112	2.77, 2.32, 1.78, 2.50, 1.58, 2.57, 1.46, 0.51, 1.04, 1.43, 1.62, 2.89, 2.17, 1.80,
113-126	1.96, 1.21, 1.59, 2.22, 2.06, 1.07, 0.88, 2.79, 2.24, 0.50, 1.92, 1.11, 0.03, 0.23,
127-135	0.66, 2.29, 1.92, 1.48, 1.42, 0.40, 2.94, 2.95, 4.35

Table 2. Randomly generated samples with different variances

Time	Sample
1-14	0.50, -2.29, 0.88, 1.47, 3.14, -3.07, 1.91, 1.01, 0.62, 0.66, 0.61, 0.35, 3.22, -3.22,
15-28	1.63, 2.87, -0.22, 1.97, 3.12, 1.13, 4.33, 3.79, -1.57, 2.03, 4.90, -1.34, 1.62, -1.68,
29-42	2.97, 1.28, 1.98, 0.51, 0.83, 1.07, 4.43, 1.46, -1.30, 1.01, 4.21, 2.69, -0.06, 2.57,
43-56	1.11, 1.39, 2.29, -1.06, 1.61, 0.07, -0.50, -1.34, -1.74, 1.62, 1.54, -1.63, 0.97, -2.30,
57-70	-1.65, 0.08, 0.49, -0.78, 2.96, -0.19, -1.17, 2.08, -2.51, -1.37, -0.49, -1.11, 1.79, 1.19,
71-84	3.00, -1.07, 0.73, 2.03, -1.76, 0.65, 1.44, -0.02, 0.01, 3.63, 0.62, -0.11, -0.12, -0.14,
85-98	-2.98, 3.42, -0.28, 4.02, -1.32, -0.45, -0.13, -0.79, -0.72, -0.94, 0.32, 1.83, 3.21, -1.88,
99-112	0.79, 4.03, -2.80, 2.72, 2.09, 10.34, -1.30, 9.41, 8.62, 5.24, 3.34, 0.73, 3.60, 3.72,
113-126	4.17, 7.60, 7.84, 7.52, 6.38, -0.10, -0.63, 3.17, 6.95, -2.01, 4.60, 6.57, 6.36, 5.06,
127-140	3.90, 5.08, 2.07, 3.28, 0.71, 6.50, -4.70, 0.70, 0.46, 1.68, 9.80, -0.33, 3.77, -1.32

Table 3. (Matabos et al., 2011a,b) The data obtained in Matabos et al., 2011a,b, on the percentage of bacterial mat coverage.

Time (hours)	Percentage of bacterial mat coverage
1-7	10.4840000, 10.3785333, 19.6990000, 13.4586000, 18.1868667, 9.6732000, 14.9852000;
8-14	13.2225667, 11.4599333, 8.1870667, 4.9142000, 4.9316667, 3.8830667, 8.0873333;
15-21	5.6105333, 7.6965333, 9.7825333, 9.8237333, 10.7416000, 13.7971333, 20.7617333;
22-28	18.8464667, 14.4726667, 17.4436667, 15.1624000, 15.1121333, 17.7154000, 17.7116667;
29-35	0.8414667, 2.2057333, 5.2884000, 6.4312000, 9.1613000, 11.8914000, 9.9539667;
36-42	8.0165333, 10.0968000, 4.4086667, 1.2552667, 3.0557333, 9.9876000, 9.8244000;
43-49	3.3898000, 7.7288000, 7.2358000, 6.7428000, 5.5994000, 7.7983333, 5.9444000;
50-56	8.4119333, 8.0767333, 6.9683333, 4.8029333, 4.9704000, 7.2590667, 6.9236667;
57-63	12.3139333, 10.9673333, 5.9108667, 9.3456667, 8.5384667, 8.7076667, 8.8768667;
64-70	9.6138667, 12.5473333, 7.9389333, 6.4124000, 7.4238667, 6.7345333, 8.9609333;
71-77	7.9157333, 10.5557333, 5.5783333, 10.2988667, 3.3476667, 5.5553333, 5.3493333;
78-84	5.8724000, 5.3806000, 6.8749333, 4.2702000, 10.2589333, 5.5500667, 3.9351667;
85-91	2.3202667, 2.5566000, 4.5210000, 6.4854000, 4.9810667, 6.9393333, 4.7274667;
92-98	7.8811333, 14.0878667, 6.6545333, 9.2467333, 7.9180667, 7.1427333, 7.7186667;
99-105	6.1379333, 8.5431333, 5.6254667, 6.3112000, 4.8482667, 6.3447333, 12.6581333;
106-112	6.1377333, 0.2495333, 1.1498000, 3.6782000, 4.3822333, 5.0862667, 4.1902000;
113-119	2.5320667, 4.8067333, 8.2410667, 7.4472667, 8.0230667, 4.8510667, 5.9036667;
120-126	6.1734667, 6.3130667, 6.9166000, 6.8148000, 4.8423333, 2.8698667, 3.9376000;
127-133	3.8878000, 3.3624000, 2.8688000, 2.3149333, 1.7610667, 2.9851333, 3.3612000;
134-140	4.1124000, 3.9806667, 4.0778667, 2.1974667, 4.2291333, 3.8029333, 4.4337333;
141-147	7.2634000, 2.9838667, 4.9395333, 4.7098000, 8.7615333, 7.9837000, 7.2058667;
148-154	3.7946000, 5.3313333, 4.2742667, 3.9970667, 5.1236000, 6.8789333, 4.6097333;
155-161	5.8856000, 4.2232000, 5.5406000, 4.5637000, 3.5868000, 3.3167333, 2.4931333

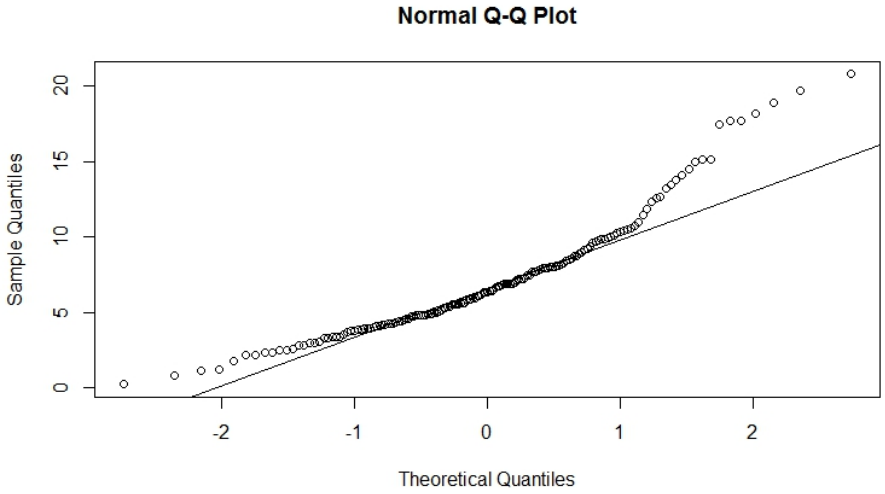
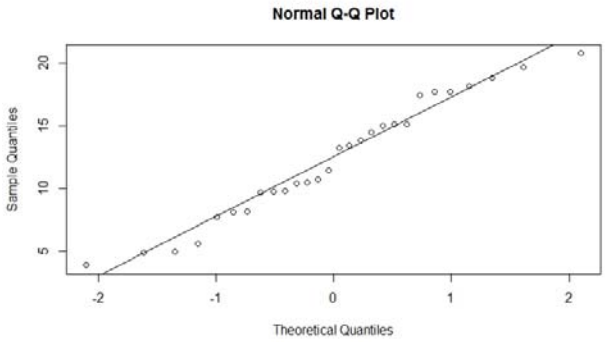
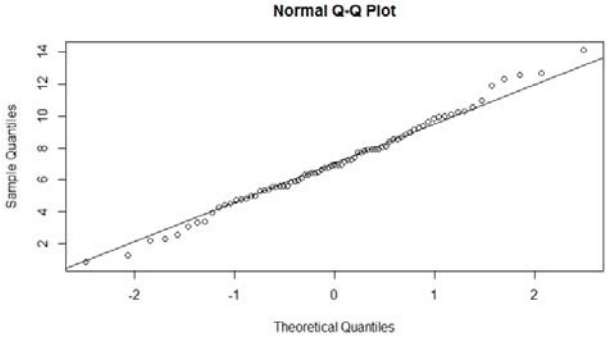


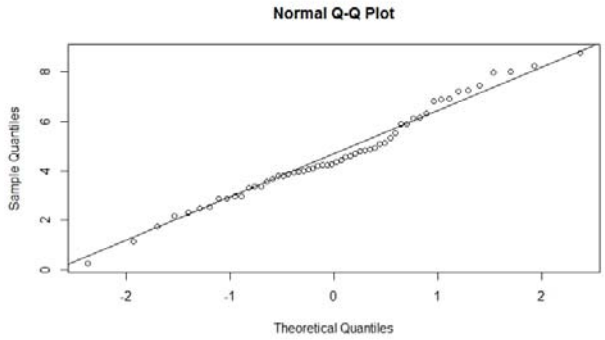
Figure 1. Normal quantile-quantile plot for the complete sample



a)



b)



c)

Figure 2. Normal quantile-quantile plots: a) For the first subsample, b) for the second subsample, c) for the third subsample.