ISEIS

Journal of Environmental Informatics XX(X) 1-12 (2016)

Journal of Environmental Informatics

www.iseis.org/jei

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

I. López[1,*], C. Rodríguez[1], M. Gámez[1], M. Varga[2], and J. Garay[3]

[1]*Department of Mathematics, University of Almería, La Cañada de S. Urbano, Almería 04120, Spain*
[2]*Institute of Mathematics and Informatics, Szent István University, Páter K. u. 1., Godollo H-2103, Hungary*
[3]*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, Budapest H-1117, Hungary*

Received 27 January 2015; revised 3 August 2015; accepted 12 January 2016; published online xx xx 2016

ABSTRACT. The paper is aimed at a methodological development of change-point detection, applicable in identifying abrupt changes in temporal or spatial data sequences. In earlier papers we developed a method for detecting a change in the parameters of a discrete distribution, with the simultaneous estimation of the (deterministic but unknown) distribution parameters before and after the change. In this paper we not only extend this method to the case of normal distributions, but also provide a new algorithm for the iterative refining of the estimation of the change-point, based on a "cleaning" of mixed-up parts of the samples. The appropriate size of reduced part of the sample is analytically calculated for the case of normal distributions. This "cleaning" is combined with our original change-point detection method. Our new algorithm is not only validated on artificial data, but also applied to a real environmental data set collected and analysed by other authors in a seafloor observatory. Our results detecting abrupt changes of bacterial mat coverage of a seafloor area are in harmony with the biological fluctuations and changes in the abiotic environment, analysed recently by other authors using a different method. We also provide a comparison with other existing change-point detection methods: a one-dimensional version of the gradient method widely used for edge detection, and a maximum type statistical method well-known in environmental studies. Although normality conditions of our method are rather restrictive, its application potential for environmental data sets is also demonstrated.

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 dates back to the initiative in Shewhart, 1931, concerning quality control of industrial production lines. Following the methodological article (Page, 1954), where the 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 summarised 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, and 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 of 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, 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 to analise of the effect of a gap-cut on

* Corresponding author. Tel.: +34 950015775; fax: +xxxxxxxxxxxxx.
E-mail address: milopez@ual.es (I. López).

ISSN: 1726-2135 print/1684-8799 online
 doi: 10.3808/jei.201500313

the spatial distribution of undergrowth plants and tree seedlings.

In the above mentioned papers we developed and applied a method for detecting 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. At the same time, our change-point detection method also estimates 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 organised as follows: In Section 2, the conceptual model is established. 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 summarised. 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 and 2012), discrete distributions were considered while 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 where a change in the parameters of the distribution occurs. The question becomes when or where this change is produced in order to understand what took place at this point point that could have affected our data. Thus, in nature, the detection of a change-point in a data sequence on a given object can help us to understand 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 and 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, we estimate for a fixed K 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.

Additionally, in López et al. (2010 and 2012), another algorithm was implemented to find the change-interval for K, which is a kind of transition zone containing the estimation of the change-point with a given probability where both distributions are mixed. There, this change-interval was constructed by an adaptation of the bootstrap method, generating bootstrap samples that consist of two linearly arranged "homogeneous" parts. The original sample is divided into two parts, so 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, implemented in the software "R" 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 exists 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, then the original sample is divided in two parts and the parameters of both distributions are estimated. Now we repeat this process but with a reduced sample from the original. 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, but centering this elimination interval in the estimated K. For the new sample, smaller than the original 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 reduce again

the original sample. We repeat this process until the change-point remains constant. 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 by taking into account that normal distributions are considered. We should know the necessary sample size to distinguish between two normal distributions. For example, we will establish for a general sample a hypothesis test where the null hypothesis is: this sample is extracted from a given normal distribution and the alternative hypothesis is: the sample is extracted from another normal distribution. 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.) Consider the sum of both errors (total error), in the following question: Given $\varepsilon > 0$, from what threshold sample size n_0, would it be verified that the total error is smaller than ε? We explain in the Appendix how we calculate the sample size n necessary to distinguish between two normal distributions given a total error.

We note that the above approach is new, different from the algorithm for calculating a change-interval from the papers of López et al., 2010, 2012. In theirs, a sample with the original sample size was always considered. However, 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. Another novelty when compared to our previous studies, here we also show how to deal with the case of several change points.

3. Model description and algorithms

3.1. Model description

In the following we will use time-series terminology, but emphasize that 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 to ξ as the *left distribution* and to η as the *right distribution*. First, from a given sample vector X= (x_1, x_2,…, 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 $2 \le K \le N$-2, we estimate the parameters of both distributions in the same way.

Let

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

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

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

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3}$$

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 as follows.

$$l(\mu,\sigma \mid X) = \prod_{i=1}^{n} f(x_i;\mu,\sigma) \tag{4}$$

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 respectively from the left and right distributions, 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:

$$l_K := l(\mu_1,\sigma_1 \mid X_{lK}) \cdot l(\mu_2,\sigma_2 \mid X_{rK}). \tag{5}$$

This function l_K will be considered as the "validity" 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) and (2).

b) Calculate:

$$Log\ l_K := Log\ l(\hat{\mu}_1, \hat{\sigma}_1 \mid X_{lK}) + Log\ l(\hat{\mu}_2, \hat{\sigma}_2 \mid X_{rK}) = \\ = \sum_{i=1}^{K} Log\ f(x_i; \hat{\mu}_1, \hat{\sigma}_1) + \sum_{s=K+1}^{N} Log\ f(x_s; \hat{\mu}_2, \hat{\sigma}_2). \quad (6)$$

(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 l2, …, Log lN-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) and (2), 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, $\hat{\mu}_1$, $\hat{\sigma}_1$, $\hat{\mu}_2$, and $\hat{\sigma}_2$, according to (1) and (2) from the obtained K=K_0.

4. Introduce the error ε, see Appendix. (This error is 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 "clean" the mixed up samples. Therefore, it 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, $X_n = (x_1, \ldots, x_{K_0-n-1}, x_{K_0+n+1}, \ldots, x_N)$ we estimate again the left and right distributions:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{K_0-n-1} x_i}{K_0 - n - 1}, \quad \hat{\sigma}_1 = \sqrt{\frac{\sum_{i=1}^{K_0-n-1} (x_i - \hat{\mu}_1)^2}{K_0 - n - 2}} \quad (7)$$

$$\hat{\mu}_2 = \frac{\sum_{i=K_0+n+1}^{N} x_i}{N - K_0 - n}, \quad \hat{\sigma}_2 = \sqrt{\frac{\sum_{i=K_0+n+1}^{N} (x_i - \hat{\mu}_2)^2}{N - K_0 - n - 1}} \quad (8)$$

c) Apply again Algorithm 1 to the complete sample X, but now change the calculation of Step 2a), that is, we keep the previously calculated values of $\hat{\mu}_1$, $\hat{\sigma}_1$, $\hat{\mu}_2$, and $\hat{\sigma}_2$ according to (7) and (8), for this application of Algorithm 1. We then 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. We will calculate the change-point for the complete sample but supposing the 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 (7) and (8).

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 would have three change-points in total, 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: the last obtained change-point of a sample piece is not considered when one of the two new obtained parts of the corresponding sample piece is too small, or the field researcher decides to stop the procedure.

3.4. Validation of the Algorithms

In order to validate the presented methods, for a given change-point, we will generate several samples from given different left and right normal distributions. After applying our method, we will reconstruct the given change-point and the parameters of both distributions. 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 random sample of size 13,500, where the left-hand side of the sample (the first 7,500 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. With a small-

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

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

b) We generate a random sample of size 135, where the left-hand side of the sample (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. 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 random sample of size 4500, with 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. Algorithm 1 may work very well even when there are more mixed elements from both distributions, due to the close values of the means and variances, as evidenced from the following example.

b) The left-hand side of the sample, the first 1000 elements, are randomly generated from $N(1,2)$ and the 800 elements of the right-hand side from $N(3,4)$. 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 distinguishment 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 $N(1,2)$ and the 40 elements of the right-hand side from $N(3,4)$. Obviously K=100. The whole sample is given in Table 2. Algorithm 1 provides an estimate for K equal to 103. Algorithm 2 improves this estimate, resulting in K=99.

3.5. Comparing with other change-point methods

For the comparison with other methods, we will use the randomly generated data used for validation in 3.4.

3.5.1. Gradient method

One of the most popular change point detection methods for continuous variables is the gradient method (GM). This method (the one-dimensional version of a planar or spatial method used for edge detection) consists of searching the maximum module of the derivative of the considered function (the maximum module of the difference sequence in our discrete-time case). Applying it to the data of 3.4, we can compare the performance of the considered methods. As we can judge from the above comparisons of both methods, our algorithms performs better than the GM, especially for large size samples.

Table 3. Comparison with the Gradient Method (GM)

Based on samples of 3.4, obtained from normal distributions with equal variances			
Sample used	Theoretical change-point	From Algorithms 1 and 2	From GM
section a)	K=7500	K=7500	K=2808
section b) (Table 1)	K=75	K=76	K=84
Based on samples of 3.4, obtained from normal distributions with different variances			
section a)	K=2500	K=2500	K=3815
section b)	K=1000	K=1000	K=1695
section c) (Table 2)	K=100	K=99	K=104

Table 4. Comparison with the Maximum Type Statistical Method (MTSM)

Based on samples of 3.4, obtained from normal distributions with equal variances			
Sample used	Theoretical change-point	From Algorithms 1 and 2	From MTSM
section a)	K=7500	K=7500	K=7320
section b) (Table 1)	K=75	K=76	K=76
Based on samples of 3.4, obtained from normal distributions with different variances			
section a)	K=2500	K=2500	K=2043
section b)	K=1000	K=1000	K=1057
section c) (Table 2)	K=100	K=99	K=85

Advantages and disadvantages of GM and Algorithms 1 and 2

The advantages of the GM are: It is computationally very fast, performs rather well for small size samples and requires no assumption on piecewise normality of the distribution of the observed data. Our methods advantage over the GM is that ours seems to perform better for both small and large size samples and also estimates the parameters of the observed data distribution.

The disadvantages of the GM are its uselessness for large data series, and it fails to estimate the parameters of the observed data distribution. The disadvantage of our method is it requires an assumption of piecewise normality for the observed data and we sometimes have to transform the data to obtain the normality. Furthermore, for large size samples, the GM (taking only a few seconds) is faster than our method (requireing several minutes).

3.5.2. Statistical test method applying "maximum type" statistics

We have considered this method to make a comparison because it is frequently used in meteorological and hydrological data series to discover systematic changes in the mean of the measured quantities, such as precipitation, air pressure or temperature. The idea is to test the null hypothesis that claims no change in the parameters of the normal distribution of the series against the alternative hypothesis claiming that there exists a time k when the distribution of the series changed its mean. For testing the null hypothesis the used test statistic is the maximum of the absolute values of the corresponding test statistics, which are calculated for each possible time point. In other words, the maximum is taken over all possible time points where the change might occur. The null hypothesis is rejected if the test statistics is larger than a corresponding critical value (which depends on the sample size n), and the time point where the previous maximum is attained will be the change point k to be found. For more details, see Jaruskova, 1997, section 4, and also Jaruskova, 1996.

Applying this maximum type statistical method (MTSM) to the data of 3.4, we can compare the performance of MSTM and our method (Table 4). Again, we can judge from the above comparisons of both that our method works better than the MTSM, especially for large size samples.

Advantages and disadvantages of MTSM and Algorithms 1 and 2

The advantage of the MTSM is its performance for small size samples, and in most cases performs much better than the GM. The advantages of our method over the MTSM are that our method seems to perform better for both small and large size samples, and it also estimates the parameters of the observed data distribution. Moreover, in our method it is not neessary calculate critical values according to the sample size.

A disadvantage of both methods is the assumption that piecewise normality of observed data is required, and that sometimes a data transformation is necessary to obtain it. The disadvantages of the MTSM are that for large data series, it gives a worse change-point estimation than ours and it doesn't estimate the parameters of the observed data distribution. Furthermore, the calculation of exact critical values for testing is complicated and an approximation of the critical values is necessary when $n > 10$.

Table 5. 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

*Matabos et al., 2011a,b

4. Experimental data

The developed change-point methodology can be applied to the analysis of temporal or spatial data sequences in a wide range of fields, and for monitoring agro-ecological and forest systems, aquatic ecosystems, etc. In the present paper we illustrate the efficiency of our method in applying it to detect 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 illustrating our methodology, we only shortly summarize the circumstances of data collection below. 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 was established in the deep-water environment of Saanich Inlet, British Columbia, Canada. Three species were observed by a remotely operated digital camera, providing an abundance of shrimp (*Spirontocaris spp.*), squat lobster (*Munida quadrispina*) and bacterial mat coverage (*Beggiatoa spp.*).

We will only discuss 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 5. The first observation corresponds to November 2, 16:00 hrs, the next to 17:00 hrs and the rest of the observations were taken hourly 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 (applying Algorithm 2 provides no improvement of this value). This change-point corresponds to November 3, 19:00 hrs.

If we want to search for another change of distribution, and we apply again Algorithm 1 only for the right-hand side of the sample, we find that there must be another change-point for $Kr = 77$, the second change-point for the complete sample would be at $K_2 = 105$, which is on November 7 at 0:00 hrs.

In many statistical procedures normal distribution of the involved samples is required. It is very important therefore to check for this normality assumption because, if 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). According to Nornadiah and Yap (2011), Shapiro-Wilk is the most powerful normality test among them. 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, there are substantial deviations from a straight line in the resulting plot, which means that the complete sample does not proceed from a normal distribution, as the formal normality tests will confirm. In Figure 2 we see what happens 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.

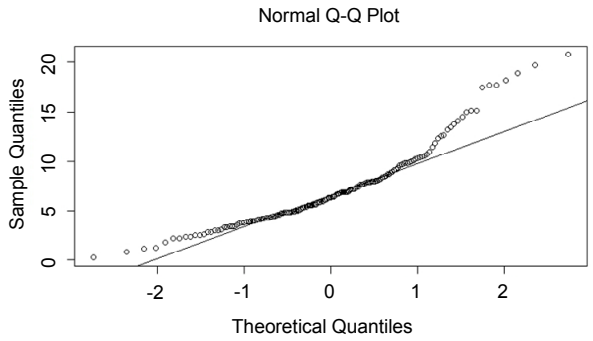

Figure 1. Normal quantile-quantile plot for the complete sample.

If we apply the Shapiro-Wilk normality test to the whole sample with a significance level $\alpha = 0.05$, the p-value obtained is 3.562·10-8, (applying Lilliefors test for normality, p-value = 3.569·10-5), which for both normality tests means that there is enough evidence to reject the normality of the whole data set. However if we use the information obtained previously and consider two samples, one between the first element and 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 0.27 71). This indicates that both tests to accept that both data sets proceed from normal distributions. If we divide the se- cond sample in two parts, according to the obtained Kr =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 we applied the Anderson-Darling normality test to all the considered samples in a similar way. The estimate of these three bacterial mat coverage distributions by the sample means and sample standard deviations are N(12.36534,4.83452), N(7.051384, 2.6 93788), and N(4.63 1949, 1.834058).

In summary, we have accepted that the data corresponding to the percentage of bacterial mat coverage during the period November 2 ~ 9 do not proceed from only a normal distribution. Normality tests have proven 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.

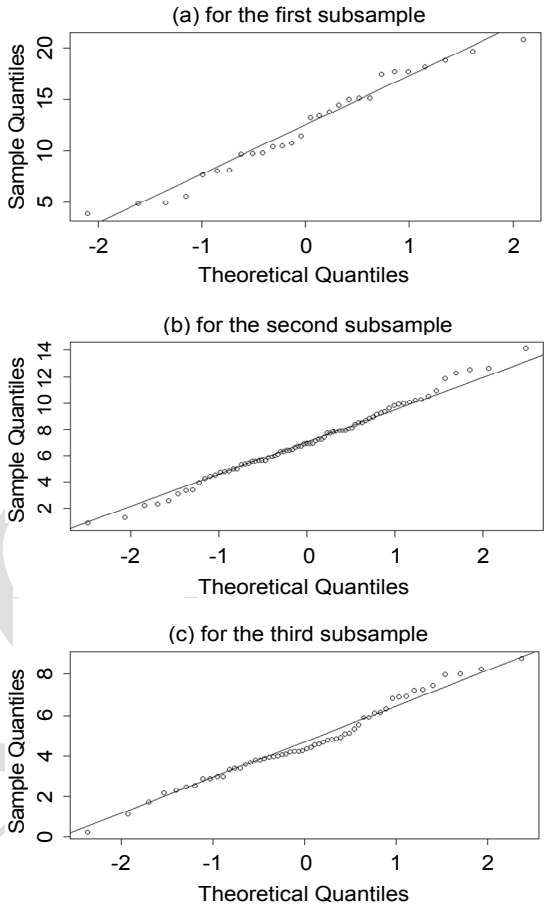

Figure 2. Normal quantile-quantile plots: (a) for the first subsample; (b) for the second subsample; (c) for the third subsample.

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, which means there is no sample evidence to reject the mean at this value. 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 decreased over time.

For a comparison with other approaches we recall that to deal with the uncertainty of the change point, we can either calculate a confidence interval for the change-point estimate (e.g. in Wang and Wang,1994), or construct the change-interval (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 leave it unchanged, depending on the size of the concrete data set and the closeness of the parameters of the involved normal distributions. A disadvantage of our method might be that, in its present stage, it is developed only for normal distributions. Nevertheless, samples from continuous variables often give positive answer to normality tests in environmental monitoring, as 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, particularly in environmental studies where normal distribution often occurs.

Under the normality assumption on the distributions separated by the change-point, Algorithm 2 is a new additional method 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 by 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 is essentially 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 is worth it to try.

For a comparison of our method with others used to detect of abrupt changes in time-series, we go back to the control charts originally used to detect changes in industrial production lines. As the overview by Taylor (2000) points 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, while the latter requires data about the whole process but offers a deeper insight to the process in question.

For a comparison with other change-point detection methods, we note that our method needs an *a priori* knowledge on the type of the distribution. For an overview of non-parametric 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 agrees with certain observations of Matabos et al., 2011a. In fact, when applying cross-correlation analysis, the bacterial mat coverage showed significant correlation to 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, $r = -0.27$, for a 6 hour lag. In other words, following a major oxygen intrusion, they found a rapid disappearance of bacterial mats. This disappearrance coincided with a rapid increase in shrimp abundance in the highly oxic environment, which might impact feeding on the bacterial mats. Another option to explain the disappearrance of Beggiatoa spp. mats is to consider 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). This separated normal distributions and at each change-point the mean value changed to a smaller one, giving an insight to the effect of minor peaks in oxygen concentration. For a complex automated image analysis of detected bacterial mat coverage based on the data collected in the VENUS Undersea Cabled Observatory, see Aguzzi et al., 2011.

It should be noted that we continued searching for further change-points inside these three samples. However, as we proceeded, the subsamples obtained were too small, and so we stopped the search and kept the previously obtained two change-points as final results.

Finally, as we look forward we note the developed change-point methodology might also be applied to temporal or spatial data sequences for monitoring epibenthic marine ecosystems, or similarly, for detecting heterogeneities in certain terrestric ecosystems, see e.g. Healey et al., 2014; Bo luwade and Madramootoo, 2015. Of course, as we have discussed in the Results section, normality tests should be applied for a correct application of our method, lest normal distribution of environmental data be taken for granted, see e.g. Rong, 2000. As already emphasized in general terms, our method can be applied to detect abrupt changes, which gives it potential *environmental and ecological* applications under the normality condition of the distributions involved. We will shortly discuss this issue below for an outlook. The well-

known observation that many variables turn out to be approximately normally distributed, is theoretically justified by the central limit theorem of probability theory, see e.g. Durrett, 2010. In environmental context important examples of time series are confirmed to be normally distributed. E.g. in Jones and Hulme,1996 it is reported that in regions where boundary conditions do not change dramatically, monthly mean temperature (as well as monthly mean maximum or minimum temperature) have the tendency to be *normally distributed*. (At this point we also note that most precipitation time series are not normally distributed, see Legates, 1991). In time series of environmental data, the random variable in question often takes only positive values, and hence has an asymmetric distribution which, of course, cannot be normal. It often occurs, however, that the variable *log-normally distributed*, i.e., its log-transformed is normally distributed. Therefore, in these cases our change-point detection method can be also applied. Here we only recall e.g. Bell, 2001 where it is shown that log-normal distribution is appropriate for particulate data and the majority of the nitric oxide, oxides of nitrogen and sulphur dioxide data sets. Also, Holland and FitzSimons, 1982 argue that log-normal distributions can represent aerometric data of positive skewness, like ozone data. To finish, for other transformations to achieve normal distributions in environmental data we refer to Mateu, 1997, where such transformations have been applied to atmospheric parameters and particle concentrations.

Acknowledgments. 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 (Regional Ministry of Economy, Innovation and Science) 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 would also like to thank the Editor and the anonymous Referees for their helpful suggestions to improve this 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), 10 534-10556. http://dx.doi.org/10. 3390/s111110534

Bell, G. (2001). Neutral macroecology. *Science,* 293(5539), 2413-2418. http://dx.doi.org/10.1126/science.293.5539.2413

Boluwade, A., and Madramootoo, C. (2015). Determining the Influence of Land Use Change and Soil Heterogeneities on Discharge, Sediment and Phosphorus. *J .Environ. Inf.,* 25(2), 126-135. http://dx.doi.org/10.3808/jei.201500290

Brodsky, E., and Darkhovsky, B.S. (1993). *Nonparametric Methods in Change Point Problems,* Springer, New York. http://dx.doi.org/10.1007/978-94-015-8163-9

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(2), 117-126. http://dx. doi.org/10.2307/1552443

Chen, J., and Gupta, A.K. (2012). *Parametric Statistical Change Point Analysis,* Birkhauser. http://dx.doi.org/ 10.1007/978-0-8176-4801-5

Cheng, Z. (2012). Using LS-SVM Pattern Recognizer to Detect Change-Point in ARMA Process. *Appl. Mech. Mater.,* 271-272, 17 31-1735. http://dx.doi.org/10.4028/www.scientific.net/AM-M.271-272.1731

Cheng, Z. (2013). An intelligent method of change-point detection based on LS-SVM algorithm. *HKIE Trans.,* 20(3), 141- 147. http://dx.doi.org/10.1080/1023697X.2013.813657

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

Durrett, R. (2010), *Probability: theory and examples,* Cambridge University Press. http://dx.doi.org/10.1017/CBO97805117793 98

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. http://dx.doi.org/10.1017/CBO9780511984679.011

Holland, D.M., and Fitz-Simons, T. (1982). Fitting statistical distributions to air quality data by the maximum likelihood method. *Atmos. Environ. (1967),* 16(5), 1071-1076. http://dx.doi.org/10. 1016/0004-6981(82)90196-2

Hawkins, D.M. (2001), Fitting multiple change-point models to data. *Comput. Stat. Data Anal.,* 37(3), 323-341. http://dx.doi.org/10.1016/S0167-9473(00)00068-2

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. *J. Environ. Inf.,* 24(1), 1-10. http://dx.doi.org/10.3808/jei.201400278

Jaruskova, D. (1996). Change-point detection in meteorological measurement. *Mon. Weather Rev.,* 124, 1535-1543. http://dx.doi.org/10.1175/1520-0493(1996)124<1535:CPDIMM>2.0.CO;2

Jaruskova, D. (1997). Some problems with application of change-point detection methods to environmental data. *Environmetrics,* 8(5), 469-483. http://dx.doi.org/10.1002/(SICI)1099-095X(199709/10)8:5<469::AID-ENV265>3.0.CO;2-J

Jones, P.D., and Hulme, M. (1996). Calculating regional climatic time series for temperature and precipitation: methods and illustrations. *Int. J. Climatol.,* 16(4), 361-377. http://dx.doi.org/10.1002/(SICI)1097-0088(199604)16:4<361::AID-JOC53>3.0.CO;2-F

Legates, D.R. (1991). An evaluation of procedures to estimate monthly precipitation probabilities. *J. Hydrol.,* 122(1), 129-140. http://dx.doi.org/10.1016/0022-1694(91)90176-I

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(1), 51-63. http://dx.doi.org/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. *Int. J. Biomath.,* 5(6), 1250054. http://dx.doi.org/10.1142/S1793524512500544

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

Mateu, J. (1997). Methods of assessing and achieving normality applied to environmental data. *Environ. Manage.,* 21(5), 767-777. http://dx.doi.org/10.1007/s002679900066

Razali, N.M., and Wah, Y.B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *J. Stat. Model. Analytics,* 2(1), 21-33.

Page, E.S. (1954). Continuous Inspection Schemes. *Biometrika,* 41(1), 100-115. http://dx.doi.org/10.1093/biomet/41.1-2.100

Rong, Y. (2000). Statistical methods and pitfalls in environmental data analysis. *Environ. Forensics,* 1(4), 213-220. http://dx.doi.org/10.1006/enfo.2000.0022

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. http://dx.doi.org/10.1214/lnms/1215463811

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:

$$P\left[S_1(x_1,...x_n) \in Q_r(\alpha) \middle| H_0 \text{ is true}\right] = \alpha \tag{1}$$

Type II error is:

$$P\left[S_1(x_1,...x_n) \in Q_a(\alpha) \middle| H_1 \text{ is true}\right] = \beta(\alpha) \tag{2}$$

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 makes 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

$$\Lambda(y) = \int_y^{+\infty} f(x)dx + \int_{-\infty}^{y} g(x)dx = \int_y^{+\infty} \frac{1}{\sigma_1\sqrt{2\Pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} dx + \int_{-\infty}^{y} \frac{1}{\sigma_2\sqrt{2\Pi}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} dx \tag{3}$$

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

$$\Lambda'(y) = -f(y) + g(y) = \frac{-1}{\sigma_1\sqrt{2\Pi}} e^{-\frac{(y-\mu_1)^2}{2\sigma_1^2}} + \frac{1}{\sigma_2\sqrt{2\Pi}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}} = 0 \tag{4}$$

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

It is easy to prove that $y = (\mu_2 + \mu_1)/2$ verifies $\Lambda'(y) = 0$ and $\Lambda''(y) > 0$. Therefore, $y = (\mu_2 + \mu_1)/2$ 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,

$$y = \frac{\mu_1\sigma_2^2 - \mu_2\sigma_1^2 + \sigma_1\sigma_2\sqrt{(\mu_2-\mu_1)^2 + \ln\left(\frac{\sigma_1}{\sigma_2}\right)^{2(\sigma_1^2-\sigma_2^2)}}}{\sigma_2^2 - \sigma_1^2} \tag{5}$$

is a minimum point of $\Lambda(y)$. Now our aim is the following:

Given these two distributions:

$$\xi(n) := \sum_{i=1}^{n} \xi_i \in N\left(n\mu_1, \sigma_1\sqrt{n}\right) \tag{6}$$

$$\eta(n) := \sum_{i=1}^{n} \eta_i \in N\left(n\mu_2, \sigma_2\sqrt{n}\right) \tag{7}$$

where we suppose $\mu_1 < \mu_2$ and $\sigma_1, \sigma_2 > 0$, see how the error depends on μ_1, μ_2, σ_1, σ_2 :

$$\Lambda(y) = \int_y^{+\infty} f(x)dx + \int_{-\infty}^{y} g(x)dx = \int_y^{+\infty} \frac{1}{\sigma_1\sqrt{2\Pi n}} e^{-\frac{(x-n\mu_1)^2}{2\sigma_1^2 n}} dx + \int_{-\infty}^{y} \frac{1}{\sigma_2\sqrt{2\Pi n}} e^{-\frac{(x-n\mu_2)^2}{2\sigma_2^2 n}} dx \tag{8}$$

Case 3: $\sigma_1 = \sigma_2 = \sigma$ and $y = n(\mu_2 + \mu_1)/2$

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

$$
\begin{aligned}
E(n) &= \int_{\frac{n(\mu_2+\mu_1)}{2}}^{+\infty} f(x)dx + \int_{-\infty}^{\frac{n(\mu_2+\mu_1)}{2}} g(x)dx \\
&= \int_{\frac{n(\mu_2+\mu_1)}{2}}^{+\infty} \frac{1}{\sigma\sqrt{2\Pi n}} e^{-\frac{(x-n\mu_1)^2}{2\sigma^2 n}} dx + \\
&\quad \int_{-\infty}^{\frac{n(\mu_2+\mu_1)}{2}} \frac{1}{\sigma\sqrt{2\Pi n}} e^{-\frac{(x-n\mu_2)^2}{2\sigma^2 n}} dx \\
&= 1 - P\left(\xi(n) \le \frac{n(\mu_2+\mu_1)}{2}\right) + P\left(\eta(n) \le \frac{n(\mu_2+\mu_1)}{2}\right) \\
&= 1 - P\left(Z \le \frac{\frac{n(\mu_2+\mu_1)}{2} - n\mu_1}{\sigma\sqrt{n}}\right) + P\left(Z \le \frac{\frac{n(\mu_2+\mu_1)}{2} - n\mu_2}{\sigma\sqrt{n}}\right) \\
&= 1 - P\left(Z \le \frac{n(\mu_2-\mu_1)}{2\sigma\sqrt{n}}\right) + P\left(Z \le \frac{n(\mu_1-\mu_2)}{2\sigma\sqrt{n}}\right) \\
&= P\left(Z \le \frac{n(\mu_1-\mu_2)}{2\sigma\sqrt{n}}\right) + P\left(Z \le \frac{n(\mu_1-\mu_2)}{2\sigma\sqrt{n}}\right) \\
&= 2P\left(Z \le \frac{n(\mu_1-\mu_2)}{2\sigma\sqrt{n}}\right)
\end{aligned}
\tag{9}
$$

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 so that $E(n) < \varepsilon$ for all $n > n_0$. Then we have that:

$$
P\left(Z \le \frac{n(\mu_1-\mu_2)}{2\sigma\sqrt{n}}\right) = \frac{\varepsilon}{2} \tag{10}
$$

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

$$
qnorm\left(\frac{\varepsilon}{2}\right) = \frac{n(\mu_1-\mu_2)}{2\sigma\sqrt{n}} \tag{11}
$$

Therefore, given the values of $\varepsilon, \mu_1, \mu_2, \sigma$, we have:

$$
n_0 = \left(\frac{2 \cdot \sigma \cdot qnorm\left(\frac{\varepsilon}{2}\right)}{\mu_1 - \mu_2}\right)^2 \tag{12}
$$

Case 4: $\sigma_1 \neq \sigma_2$ and y is given as:

$$
y = \frac{n(\mu_1\sigma_2^2 - \mu_2\sigma_1^2) + \sigma_1\sigma_2\sqrt{n^2(\mu_2-\mu_1)^2 + \ln\left(\frac{\sigma_1}{\sigma_2}\right)^{2n(\sigma_1^2-\sigma_2^2)}}}{\sigma_2^2 - \sigma_1^2} \tag{13}
$$

In this case the total error is $E(n) = 1 - P\left(\xi(n) \le y\right) + P\left(\eta(n) \le y\right)$, where

$$
P\left(\xi(n) \le y\right) = P\left(Z \le \frac{y - n\mu_1}{\sigma_1\sqrt{n}}\right) \tag{14}
$$

$$
P\left(\eta(n) \le y\right) = P\left(Z \le \frac{y - n\mu_2}{\sigma_2\sqrt{n}}\right) \tag{15}
$$

Again, our objective is to obtain a value of n_0 that assure that given the error ε, $E(n) < \varepsilon$ holds for all $n > n_0$.

It is guaranteed that:

$$
E(n) \le 2\max\left(1 - P\left(Z \le \frac{y - n\mu_1}{\sigma_1\sqrt{n}}\right), P\left(Z \le \frac{y - n\mu_2}{\sigma_2\sqrt{n}}\right)\right) \tag{16}
$$

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 for n that makes Equations (13) and (14) equals to 1- ε /2 and ε /2, respectively.

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

$$
qnorm\left(1 - \frac{\varepsilon}{2}\right) = \frac{y - n\mu_1}{\sigma_1\sqrt{n}} \tag{17}
$$

$$
qnorm\left(\frac{\varepsilon}{2}\right) = \frac{y - n\mu_2}{\sigma_2\sqrt{n}} \tag{18}
$$

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 so that, $E(n) < \varepsilon$ for all $n > n_0$. In Algorithm 2 Step 5a) we will choose $n = round(n_0) + 1$.