AUG-Segmenter: a user-friendly tool for segmentation of long time series

Abdullah Gedikli, Hafzullah Aksoy and N. Erdem Unal

ABSTRACT

In this study, three algorithms are presented for time series segmentation. The first algorithm is based on the branch-and-bound approach, the second on the dynamic programming while the third is a modified version of the latter into which the remaining cost concept of the former is introduced. A user-friendly computer program called AUG-Segmenter is developed. Segmentation-by-constant and segmentation-by-linear-regression can be performed by the program. The program is tested on real-world time series of thousands of terms and found useful in performing segmentation satisfactorily and fast.

Key words | branch-and-bound approach, change point, dynamic programming, remaining cost, segmentation, time series

Abdullah Gedikli
Hafzullah Aksoy (corresponding author)
N. Erdem Unal
Department of Civil Engineering,
Istanbul Technical University,
34469 Maslak,
Istanbul,
Turkey
E-mail: *haksoy@itu.edu.tr*

INTRODUCTION

Due to various natural or man-made causes datasets are not necessarily homogeneous or consistent. Therefore structural characteristics (jump, trend, randomness, intermittency, probability distribution function, etc) of the data should be known prior to its use (Fanta *et al.* 2001; Adeloye & Montaseri 2002; Xiong & Guo 2004; Aksoy 2007; Dahamsheh & Aksoy 2007; Aksoy *et al.* 2008*b*). A trend in a time series can result from gradual natural and human-induced disruptive and evolutionary changes in the environment whereas a jump may result from sudden catastrophic natural events. Without looking if it is gradual or sudden, inhomogeneity and inconsistency in the time series should be identified, detected or be estimated and then corrected (Peterson *et al.* 1998; Aguilar *et al.* 2003; Wijngaard *et al.* 2003; DeGaetano 2006). Any change in the time series is most reliable if it is detected by statistical tests and also has physical and historical evidence. Projections for the future cannot be made without having this information to hand. Detection of irregularities, jumps and changes is of great importance and therefore such changes must be taken into account when the past is extrapolated into the future.

doi: 10.2166/hydro.2009.084

Change point (jump) analysis is used to determine if there is a change in the statistical properties of a time series (mean, variance, etc) and to estimate the number and position(s) of change point(s). Chen & Gupta (2000, 2001) present an extensive literature review of change point analysis. Homogenization, on the other hand, is an application of change point analysis in hydrology and climatology, which consists of detecting and correcting artificial change points only, which is often done by comparing the candidate series to neighbour series to isolate artificial changes only. Jump analysis is a change point detection problem that can also be referred to as segmentation of a time series simply aiming at dividing a given number of observations into subseries (segments) with statistical characteristics that are similar within each segment and different between segments. The simplest case is the segmentation with regression-by-constant by which subsequent segments with significantly different means are determined. The usual criterion to decide if a change point exists is based on the segmentation cost defined as the sum of the squared deviation of the data from the means of their respective segments. Various techniques, including the

classical parametric or nonparametric tests, are available to compare the average of successive segments and test if they are different from each other by checking the null hypothesis that the mean values of successive segments are equal.

The development of fast and efficient segmentation algorithms emerges as a practically significant problem. An early study on hydrological time series segmentation was made by Buishand (1982). Many segmentation methods and a very extensive bibliography are presented in Basseville & Nikiforov (1993). Hubert *et al.* (1989) worked on the segmentation of hydrometeorological time series with a continuous effort (Hubert 2000; Hubert *et al.* 2007).

For the purpose of time series segmentation, the dynamic programming (DP) algorithm has been presented in a preliminary form in Kehagias *et al.* (2006). Similarly, early versions of the newly developed branch-and-bound (BB) approach-based algorithm (denoted as AUG) appeared in Aksoy *et al.* (2007, 2008*a*) and Gedikli *et al.* (2008). All of these studies contain fairly extensive references to the segmentation literature, which was therefore only mentioned in brief here.

Motivation for this study comes from the work on *multiple segments*, *offline segmentation* by Hubert (2000). In the *offline segmentation* an entire time series $(x_1, x_2, \ldots, x_T)$ is given and it is required to divide it into segments. In the *online segmentation*, on the other hand, the data points $(x_1, x_2, \ldots, x_t, \ldots)$ arrive one at a time and, at every time step t, it is required to decide whether x_t belongs to the previous segment or is assigned to a new segment which starts at t (Dobigeon & Tourneret 2007).

In this study, three *offline* segmentation algorithms, one based on the BB approach and another on the DP approach, are used. The third algorithm (mDP) is a new version of the latter modified by the *remaining cost* concept of the former. The *remaining cost* concept was introduced for the first time by Gedikli *et al.* (2008), and later by Aksoy *et al.* (2008*a*) and Gedikli *et al.* (2010). The BB, DP and mDP algorithms have been evaluated on several real-world hydrometeorological time series (Kehagias *et al.* 2006; Aksoy *et al.* 2007, 2008*a*; Gedikli *et al.* 2008; 2010) and progress has been achieved in the segmentation algorithm of Hubert (2000) by using the *remaining cost* concept. A user-friendly computer program (AUG-Segmenter) was developed in this study that uses the above algorithms. For details on the BB algorithm see Gedikli *et al.* (2008) and Aksoy *et al.* (2008*a*); on the DP algorithm see Kehagias *et al.* (2006) and on the mDP algorithm see Gedikli *et al.* (2010). The unique perspective of this study is the AUG-Segmenter software using the above algorithms. The software is presented together with its pros and cons, and updates foreseen for the future. Currently, it is fast and capable to segment long time series of thousands of items and hence is advised for the use of researchers as well as practicing engineers in the earth-related studies.

This study is organized as follows. The three algorithms are briefly defined in the following section. Then the developed software (AUG-Segmenter) is detailed together with an example experiment of the minimum water level data in the Nile River, after which the abilities and disabilities of the software are discussed. Finally, conclusions and future developments foreseen for the computer program are given.

DEFINITIONS AND FORMULATION

Let us say a time series $\mathbf{x} = (x_1, x_2, \ldots, x_T)$ is given. The procedure by which the segments of the time series are determined is called *time series segmentation*. The segmentation can be described by a sequence $\mathbf{t} = (t_0, t_1, \ldots, t_K)$ to satisfy $0 = t_0 < t_1 < \ldots < t_{K-1} < t_K = T$. The times $t_0, t_1, \ldots, t_K$ where changes take place are called *segment boundaries* (or *change points*). Any interval between two subsequent change points $[t_0 + 1, t_1]$, $[t_1 + 1, \ldots, t_2]$, $\ldots$, $[t_{K-1} + 1, t_K]$ is a *segment* (of the time series), and K, the number of segments, is called the *order of the segmentation*.

The set of all segmentations of $(1, 2, \ldots, T)$ is denoted by $\mathbf{T}$ and the set of all segmentations of order K by $\mathbf{T}_K$. Clearly, $\mathbf{T} = \cup_{K=1}^{T} \mathbf{T}_K$. The number of all possible segmentations of $(1, 2, \ldots, T)$ is 2^{T-1}. This can be formulated as an optimization problem. In other words the *optimal* segmentation depends on $\mathbf{x}$. The segmentation cost $J(\mathbf{t})$ is defined by

$$J(\mathbf{t}) = \sum_{k=1}^{K} d_{t_{k-1}1, t_k} \tag{1}$$

where $d_{s,t}$ (for $0 \le s < t \le T$) is the segment error corresponding to the segment $[s,t]$. The segment error depends on the data vector $\mathbf{x} = (x_s, x_{s+1}, \ldots, x_t)$.

A variety of $d_{s,t}$ functions is available. In this study

$$d_{s,t} = \sum_{\tau=s}^{t} (x_\tau - \mu_{s,t})^2 \quad (2)$$

is used in which the segment mean is given by

$$\mu_{s,t} = \frac{\sum_{\tau=s}^{t} x_\tau}{t-s+1} \quad (3)$$

The optimal segmentation, denoted by $\hat{\mathbf{t}} = (\hat{t}_0, \hat{t}_1, \ldots, \hat{t}_K)$, is defined as $\hat{\mathbf{t}} = \underset{t \in T}{\operatorname{argmin}} J(\mathbf{t})$ and the optimal segmentation of order K, denoted by $\hat{\mathbf{t}}^{(K)} = \left(\hat{t}_0^{(K)}, \hat{t}_1^{(K)}, \ldots, \hat{t}_K^{(K)}\right)$, is defined as $\hat{\mathbf{t}}^{(K)} = \underset{\mathbf{t} \in \mathbf{T}_K}{\operatorname{argmin}} J(\mathbf{t})$. The optimal segmentation can be found by exhaustive enumeration of all possible segmentations (and computation of the corresponding $d_{s,t}$). This is a computationally infeasible way as the total number of segmentations increases exponentially with T. Hubert (2000) uses a branch-and-bound approach to search efficiently the set of all possible segmentations and states that this approach "currently" (in the year 2000) can segment time series with several tens of terms but is not able "… to tackle series of much more than a hundred terms …" because of the combinatorial increase of the computational burden. In this study, fast algorithms which can segment time series with thousands of terms in a time frame of seconds are presented. In order to obtain these algorithms, it is first required to develop a fast method for computing the cost $d_{s,t}$. For this aim, the recursive formulation of

$$d_{s,t+1} = d_{s,t} + (t-s+1)(\mu_{s,t} - \mu_{s,t+1})^2 + (x_{t+1} - \mu_{s,t+1})^2 \quad (4)$$

is easily proved where

$$\mu_{s,t+1} = \frac{(t-s+1)\mu_{s,t} + x_{t+1}}{t-s+2}. \quad (5)$$

The BB approach

As stated before, the AUG segmentation algorithm is based on the BB-type technique. The *branches* are the possible segments of the kth-order segmentation. As suggested by Hubert (2000), the upper bound, u, of the kth segment in the Kth-order segmentation can trivially be given as

$$t_k \leq u = T - K + k. \quad (6)$$

In the AUG algorithm, the *upper bound* corresponds to the highest possible value that t_k can take.

In order to determine the optimal segmentations of any order from $K = 2$ to $T - 1$, three alternative approaches were made available in detail by Aksoy *et al.* (2008*a*) who supplied the pseudo-codes of each approach. The easiest one was called the *primitive code*, the second (a more complicated one) the *intermediate code*. The most complicated one resulted in the AUG algorithm to be detailed below.

Loops in the *primitive code* are always completed from $K = 2$ to $T - 1$ and then a comparison and an update is made to minimize the cost which initially is taken equal to $d_{1,T}$. Considering Equation (2) one can easily realize that $d_{1,T}$ corresponds to the highest cost. This also means that the cost of any kth-order segmentation of the first t_k elements, $c_{t_k}^k$, where $k < T$, is not considered in the *primitive code*. When this cost is considered, a more efficient way, called *intermediate code*, is obtained (Aksoy *et al.* 2008*a*).

The basic idea of the AUG algorithm (and, more generally, of the BB-type technique) is to enumerate (branch into) the possible solutions of the segmentation problem but, at the same time, to avoid exhaustive enumeration by eliminating clearly suboptimal solutions (bounds). Hence, before presenting the AUG algorithm, it is worth discussing upper and lower bounds of the segmentation, more specifically the boundaries t_k (for $k = 1, 2, \ldots, K$).

In addition to those eliminated in the *intermediate code*, it is possible to further eliminate segmentations by reducing the upper bound of the segments as defined in Equation (6). It is also easy to check that

$$c_{t+1}^k \geq c_t^k \geq \left(c_t^{k+1} \quad \text{and} \quad c_{t+1}^{k+1}\right) \quad (7)$$

is valid for $t = 2, \ldots, T-1$ and $k = 1, \ldots, t$. Equation (7) is rather obvious; a detailed derivation of it can be found in Gedikli *et al.* (2008), where four lemmas – one with a proof – are given. In addition to Equation (7), it is also known that any k sequential segments extracted from the optimal segmentation are also optimal, i.e. if the cost of the optimal segmentation is $J(\hat{\mathbf{t}})$, then the cost $J(\hat{\mathbf{t}}_k)$ with change points $\mathbf{t}_k = \{t_0, t_1, \ldots, t_k\}$ also satisfies the optimality condition. It then becomes clear that a kth-order segmentation of $(x_1, x_2, \ldots, x_t)$ with cost $c_t^{k-1} > c_T^K$ can not be optimal (Gedikli *et al.* 2008).

In order to reduce the upper bound, u, in this way, the *remaining cost* concept is defined as

$$R_{T,t}^{K,k} = c_T^K - c_t^k \tag{8}$$

where $k \le K$ and $t \le T$. Considering Equation (7), the reduced upper bound of the kth segment, e, can be obtained as the largest integer satisfying

$$s \le e \le T - K + k \tag{9}$$

and

$$d_{s,e} \le R_{T,s-1}^{K,k-1} \tag{10}$$

where s is the starting point of the kth segment. Based upon Equation (10), it is seen that the cost of the kth segment must be less than or equal to the remaining cost. When Equation (8) and Equation (10) are combined, it is noted, for $k = 1$, that Equation (10) takes the form of

$$d_{1,e} \le c_T^K \tag{11}$$

since it is already known that

$$R_{T,0}^{K,0} = c_T^K \tag{12}$$

is valid. Considering the kth-order segmentation of the sub-series made of the first r items, and using Equation (10)

$$d_{s,r} \le R_{e,s-1}^{k,k-1} \tag{13}$$

can be written and hence a new upper bound, r, satisfying

$$s \le r \le e \tag{14}$$

can be obtained.

The DP algorithm

Consider the optimal segmentation of $(x_1, x_2, \ldots, x_t)$ which contains k segments and suppose that its last segment is $[s+1, t]$. Then the first $k-1$ segments form an optimal segmentation of $(x_1, x_2, \ldots, x_s)$. More specifically, if c_t^k is the minimum segmentation cost of $(x_1, x_2, \ldots, x_t)$ into k segments then

$$c_t^k = c_s^{k-1} + d_{s+1,t} \tag{15}$$

is satisfied. Equation (15) allows the use of a typical DP approach to efficiently compute the optimal costs and the corresponding optimal segmentations.

On termination, the DP algorithm computes the optimal segmentation cost $c_T^K = J(K) = \min_{\mathbf{t} \in T_K} J(\mathbf{t})$ and, by backtracking, the optimal segmentation $\hat{\mathbf{t}}^{(K)} = \left(\hat{t}_0^{(K)}, \hat{t}_1^{(K)}, \ldots, \hat{t}_K^{(K)}\right)$; these quantities are computed for $K = 1, 2, \ldots, K_{\max}$: in other words, a *sequence* of minimization problems is solved recursively.

The mDP algorithm

The remaining cost concept defined through Equations (7)–(14) was coupled to the DP algorithm to obtain the mDP. By the remaining cost concept, the upper bound of the segments is reduced. From Equation (15), it is seen that computation is made from $t = s$ to T in the DP algorithm while in mDP it is ended when $t = e$, $e < T$, as given in Equation (9).

Optimal segmentation

The algorithms compute a sequence of optimal segmentations $\hat{\mathbf{t}}_1, \hat{\mathbf{t}}_2, \ldots, \hat{\mathbf{t}}_K$; where $\hat{\mathbf{t}}_k$ is the kth-order optimal segmentation. Determining the optimal order of segmentation, i.e. selecting the number of segments, is, however, a subsequent step in the segmentation procedure for which the Scheffe (1959) test is employed in this study. The test is based on the following idea.

For a given segmentation ($\hat{\mathbf{t}}_k$ for instance), the null hypothesis that the means of consecutive segments are significantly different is tested. This is done using the Scheffe (1959) test which, in short, is a very general *multiple means comparison* test. The test is run on the optimal segmentations $\hat{\mathbf{t}}^{(1)}, \hat{\mathbf{t}}^{(2)}, \ldots, \hat{\mathbf{t}}^{(K)}$. Hubert (2000) accepts $\hat{t}^{(k)}$as the optimal segmentation when $\hat{t}^{(k+1)}$ is the *first lowest order* segmentation which is rejected by the Scheffe (1959) test (i.e. the first segmentation for which at least two consecutive segments do not show a statistically significant difference in their means). In the BB, DP and mDP algorithms in this study, the *highest order* segmentation accepted by the test is considered optimal instead of the *first lowest*.

Bayesian information criterion (BIC) is a criterion for model selection among a class of parametric models with different numbers of parameters such as the order and the degree of regression used in the segmentation. Developed by Schwarz (1978) the BIC is defined as

$$BIC = -2\ln L + k\ln(n) \tag{16}$$

where n is the number of observations (sample size = length of time series in this study), k is the number of free parameters (for example, the regressors plus the constant in the linear regression used in this study) and L is the maximized value of the likelihood function. Among the alternatives, the model with the lower value of BIC is the one to be preferred.

THE AUG-SEGMENTER VERSION - 1.0.0

The AUG-Segmenter Version - 1.0.0 is a software that uses the three segmentation algorithms above. Based upon a user-friendly interface it is able to segment long time series efficiently and fast. The software is briefly introduced below by stressing then its abilities and disadvantages.

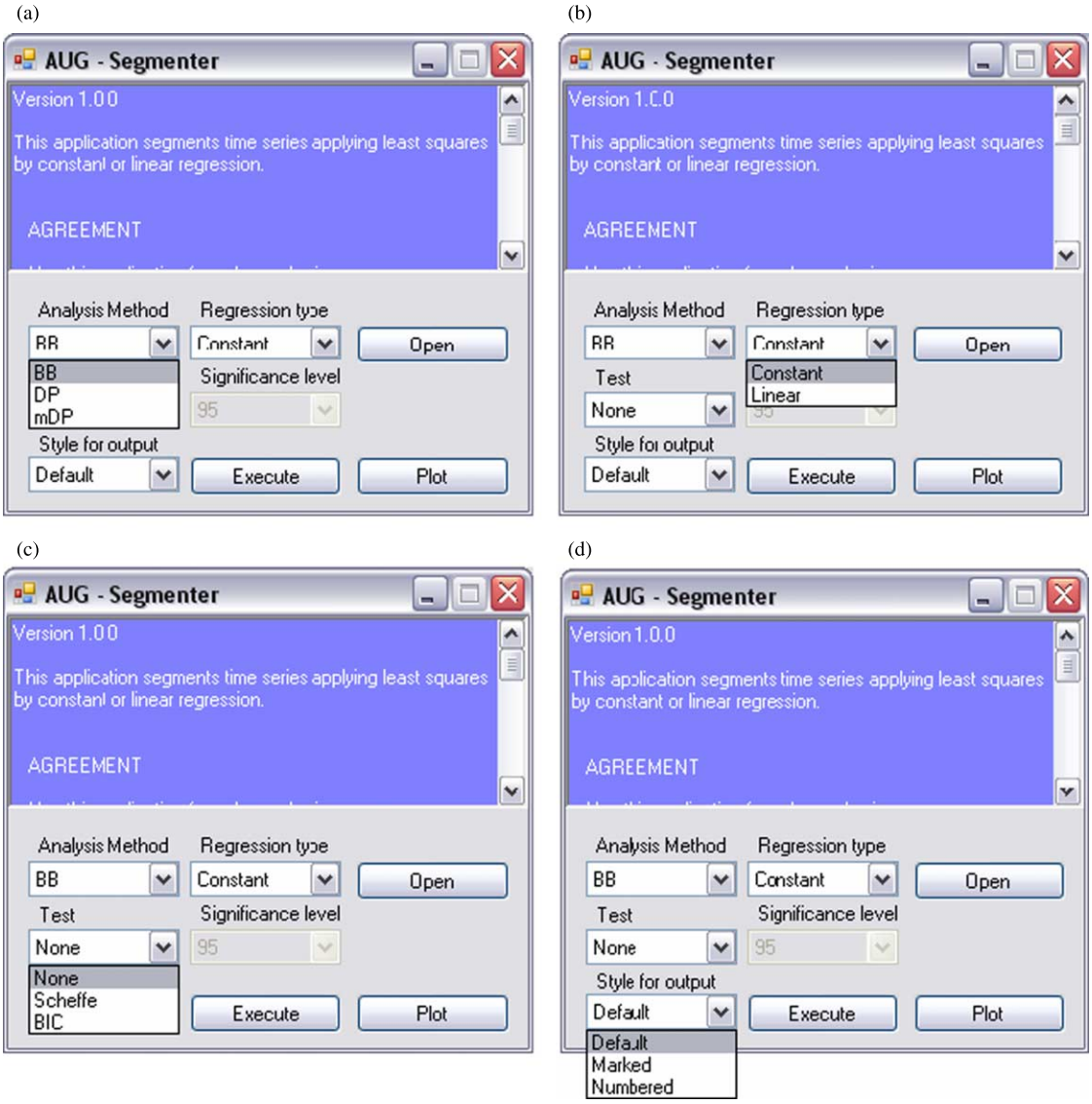

Figure 1 | Main window of the AUG-Segmenter and options for (a) methods of analysis, (b) types of regression, (c) test used and significance level, and (d) style of the output file.

The AUG-Segmenter Version - 1.0.0 has a main window (Figure 1) with an information box, command buttons and scroll-down menus from which the method of analysis, type of regression, test to be applied, significance level to be selected and style for output can be selected. The methods used for segmentation contain the BB, DP and mDP algorithms (Figure 1(a)). The BB-approach-based AUG segmentation algorithm of Gedikli *et al.* (2008) was supplied together with the DP of Kehagias *et al.* (2006) and its modified version, mDP, as briefly described above. Two types of regressions are included in the tool, namely the constant and linear (Figure 1(b)). The constant type regression is selected for segmentation-by-constant. The segmentation-by-regression is obtained when linear regression is selected. If the optimal order of segmentation is desired then either the Scheffe test or BIC should be selected (Figure 1(c)). If none is selected then the time series is segmented into as many pieces as it can be divided into. This option is useful in cases when the user requires the change points for any order of segmentation although it might not be optimal. The highest order for the constant segmentation is the same with the number of data in the time series whereas it goes naturally down to half in the case of linear regression. Style for output is to detail the output file (Figure 1(d)).

The "Open" button in the main window allows one to browse in the computer and retrieve the data file to be segmented. Information on the dataset appears in the information box once the dataset is retrieved, which includes the name of the file, total items in the dataset and its statistical characteristics (Figure 2).

In this study, minimum water level in the River Nile (in m) will be used for demonstration purposes. The length of the data is 1297 years for the period 622–1918. Statistical characteristics of the data used are given in Table 1. This dataset has previously been used and well defined by Kehagias (2004), Kehagias & Fortin (2006), Kehagias *et al.* (2006, 2007), Aksoy *et al.* (2007, 2008*a*) and Gedikli *et al.* (2008).

Constant-mean segmentation

When the AUG-Segmenter is executed for the constant-mean segmentation (constant regression type) as displayed in Figure 2, an output file is generated and the report is seen on the computer screen (Figure 3). In this report, a list of characteristics of the time series is given first after which change points are marked with numbers as given in Figure 3. Each number corresponds to instances where change points are located. In the meantime, the same list of characteristics in the report is seen in the information box of the main window once the execution of the program is completed (Figure 4). The "Plot" button creates the "Plot window" for the resulting optimal segmentation, which is a graph as seen in Figure 5. The 16th-order segmentation is provided together with the input data. The total cost and the order of the segmentation are given. A scroll-down menu for the segmentation order, two tick boxes and six command buttons are also supplied on the "Plot" window. By scrolling down or up, the segmentation order can be changed. The total cost as well as the graph change accordingly.

Figure 2 | Main window after the dataset was retrieved.

Table 1 | Summary statistics of minimum water level in the River Nile

Mean (m)	11.60
Standard deviation (m)	1.10
Coefficient of variation	0.094
Coefficient of skewness	0.011
Correlation coefficient	0.761

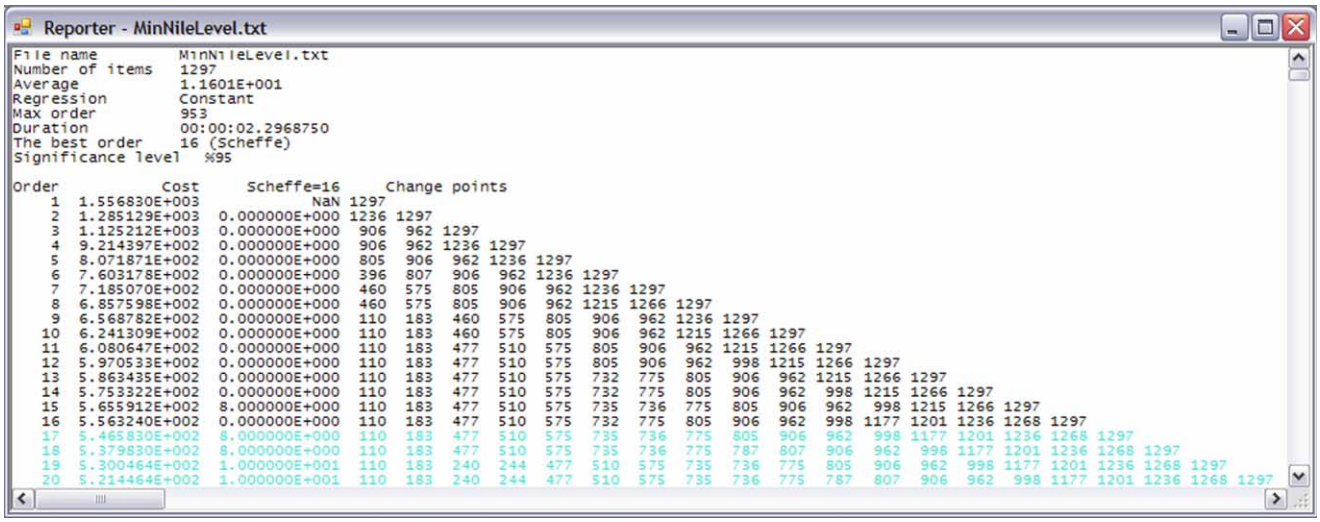

Figure 3 | Report generated after the computer program was executed (change points are numbered and segmentation order higher than the optimal is colored).

The AUG-Segmenter plots the optimal segmentation order in red, different from others.

Vertical lines in the graph show where the *dominant change points* are located. The dominant change point is a new concept that has first been introduced in Gedikli *et al.* (2008). If a change point in any order of segmentation satisfies the change point condition (i.e. becomes a change point) for all possible higher-order segmentations, then it is called a *dominant change point*. Any graph can be stamped by pushing the button "Stamp". As an example, the 16th-order segmentation was stamped in Figure 6 in which the sixth-order segmentation was also seen together with the "Legend" box ticked. If it is desired to go to the stamped segmentation, the button "Go Stamped" is pushed after which only the stamped segmentation remains in the window; the other segmentation disappears. When the stamped segmentation is not needed anymore in the window then it can simply be removed by clicking on "Clear Stamp". Information previously seen in the information box can also be retrieved by pushing the "Info" button. The "Close" button simply closes the "Plot" window.

Linear segmentation

The AUG-Segmenter is not only able to do segmentation-by-constant but also segmentation-by-linear-regression, again using BB, DP or mDP segmentation algorithms, and it can decide on what order of segmentation is optimal with respect to BIC. When the linear segmentation is concerned, a time series can be segmented into as many linear pieces as half the number of items in the time series by selecting BB, DP or mDP as the regression type and selecting "None" as the test. This information might be useful in many cases. For instance, one might be interested in dividing a time series into six segments and fitting linear equations to each segment. Such information is valuable in particular when

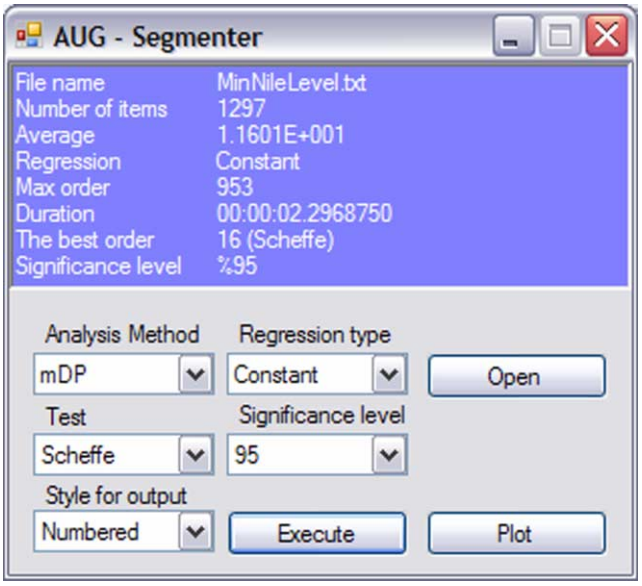

Figure 4 | Summary information after the program was executed.

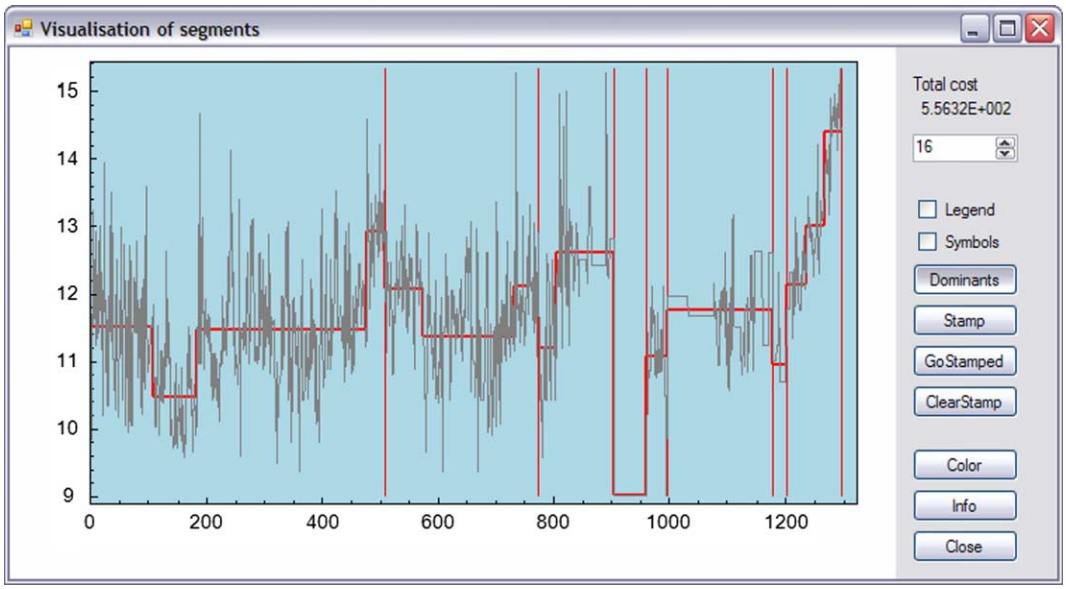

Figure 5 | The "Plot" window showing the segments and the dominant change points (horizontal axis refers to the year of the observation and vertical axis is the minimum water level in the Nile River in m).

the inner trends might behave differently than the trend taken over the whole dataset, as seen in Figure 7. For the minimum water level data of the Nile River, the positive general trend has, in fact, a very early negative and two positive trends as well as two most recent positive components plus a constant period in the middle. It should be noticed that the most recent trend is severe and highly significant.

On the other hand, the best-fit lines can be determined for each constant-mean segments which can again be useful. A graph showing the best-fit lines to these segments is shown in Figure 8 for which the sixth-order

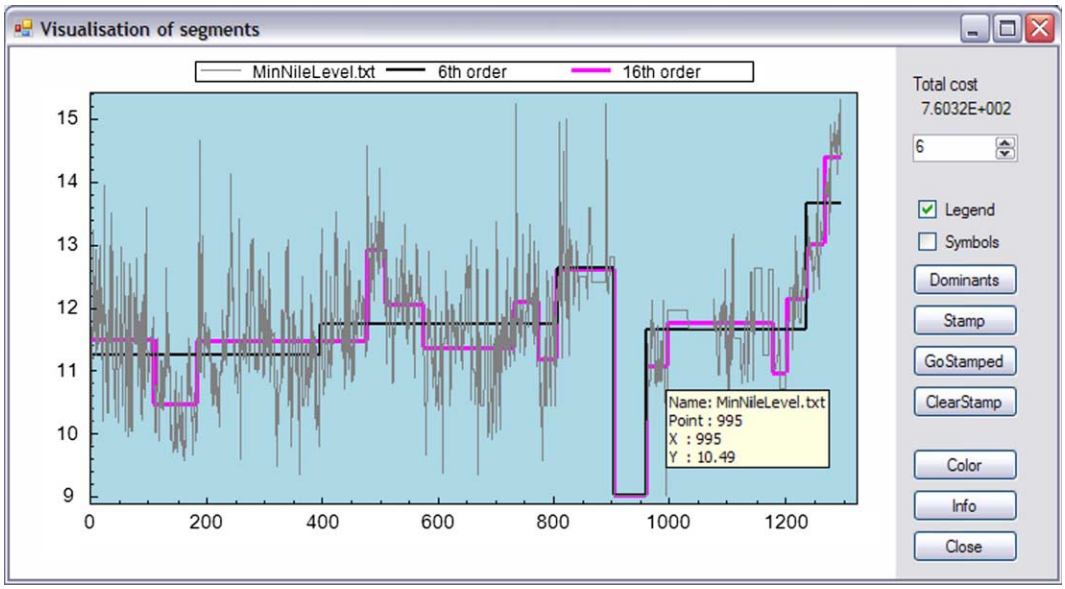

Figure 6 | The "Plot" window with the 16th-order segmentation stamped and the sixth-order segmentation (horizontal axis refers to the year of the observation and vertical axis is the minimum water level in the Nile River in m).

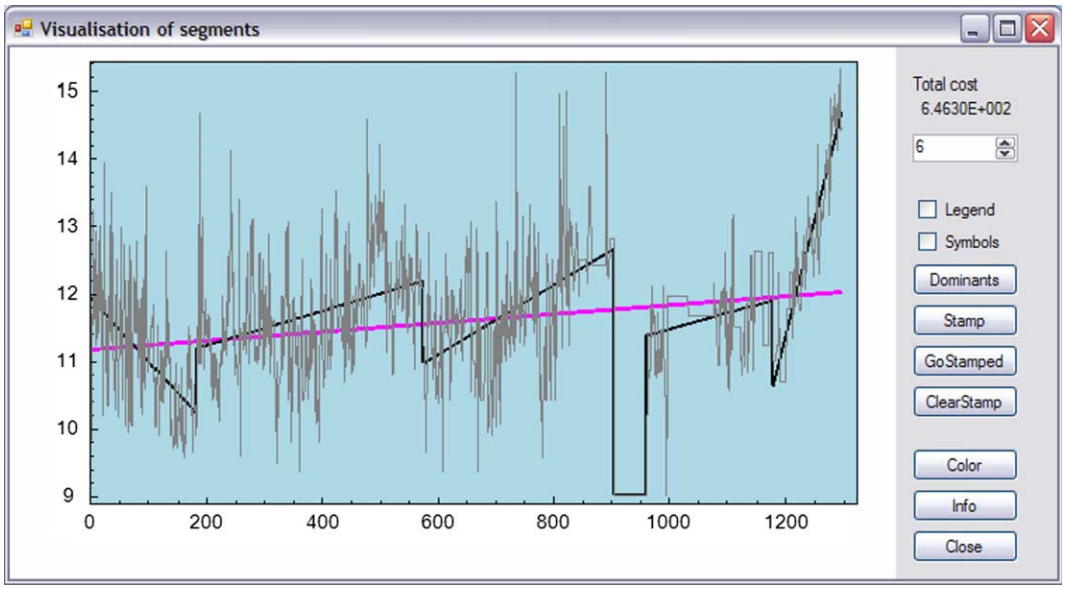

Figure 7 | Sixth-order linear segmentation together with first-order linear segmentation stamped (horizontal axis refers to the year of the observation and vertical axis is the minimum water level in the Nile River in m).

constant-mean segmentation was selected for demonstration purposes. It might, for instance, be interesting to note that the second segment has a negative trend while the first and the middle two have almost no trend and can be considered constant segments. It is worth noticing that the positive trends in the last two segments (that in the last segment in particular) are significant and severe. This type of segment can be called *pseudo-linear segments*, which are different from the linear segments previously mentioned in Figure 7. Here, the time series is segmented by constant-mean and linear trends are then fitted to each segment.

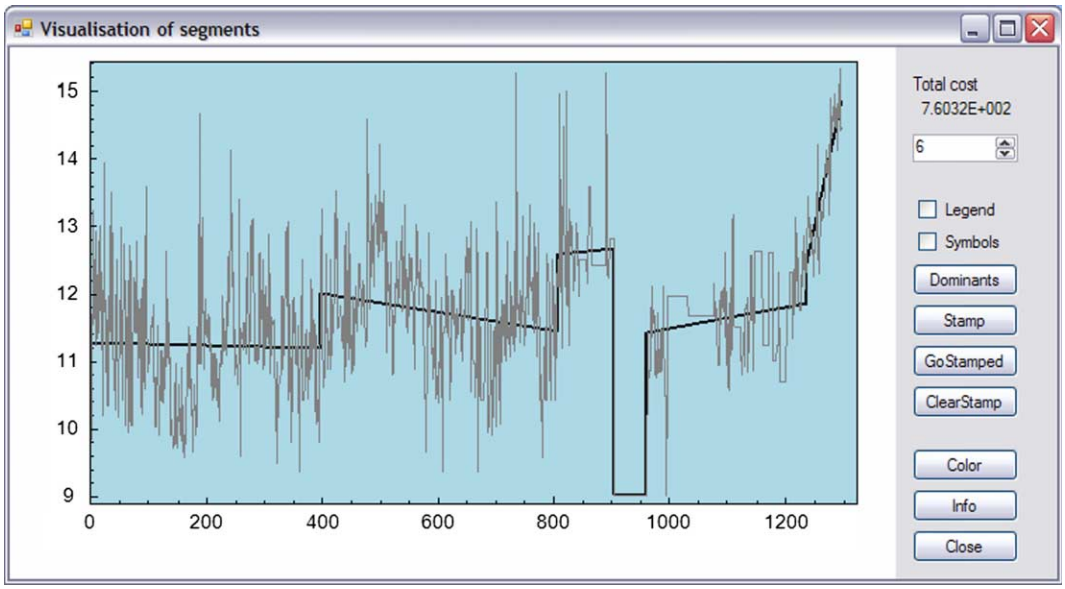

Figure 8 | Linear best fits to the sixth-order segmentation-by-constant (horizontal axis refers to the year of the observation and vertical axis is the minimum water level in the Nile River in m).

Table 2 | Execution time (hour:minute:second) for the BB, DP and mDP algorithms. The SEGMENTER was run on a computer with a dual processor 1.66 GHz (CPU) and 512 MB memory (RAM)

BB	DP	mDP
00:02:01.656	00:00:07.828	00:00:02.031

Comparison

The algorithms minimize the segmentation cost defined by Equations (1)–(3). As the algorithms work in a deterministic way, the minimization is exact and therefore the algorithms give the same change points. Hence the only comparison possible between the algorithms is in terms of the computer execution time, which can be seen in Table 2. It is observed that the BB algorithm is much slower than the other two. However, it is worth mentioning that the mDP approach, which is DP modified by the "*remaining cost*" concept of BB, is faster than the DP algorithm.

CONCLUSIONS AND FUTURE DEVELOPMENT

AUG-Segmenter, a user-friendly computer program developed for the segmentation of long time series, is presented. The software has the ability to segment time series into as many as about 8,000 items in minutes with no restriction on the segmentation order. When a time series of tens or hundreds of terms is considered, which is the usual case that hydrologists, meteorologists, climatologists and earth-related scientists or practitioners face in practice, the run time required for the program reduces to a few seconds only. The program uses three different algorithms, the BB, DP and mDP, resulting in the same change points as they are all deterministic. The difference between the three is only the execution time required for running the computer program, for which the mDP needs the shortest time. This has been achieved by the newly developed *remaining cost* concept of Gedikli *et al.* (2008) used to modify the original version of DP as given by Kehagias *et al.* (2006).

The AUG-Segmenter is able to segment time series by constant (constant-mean segmentation), i.e. subsequent segments with significantly different means are determined. The mean value of each segment as well as the best-fit line (pseudo-linear segments) can be obtained by the program. It can also divide the time series into linear segments. Other segmentation costs based on quadratic, exponential or logarithmic segmentation, when feasible, can be performed in the future. Even a segmentation algorithm can be formulated combining all these approximations. The linear segmentation among these approximations can particularly serve as an exploratory tool for trend analysis, a recent popular topic in environmetrics and climatology to detect climate change or climate variability.

Certain points in the time series always appear as change points of the optimal segmentations after a critical segmentation order is reached. Such change points can be referred to as *dominant change points* (see Figure 5). The dominant change points can help in the determination of the optimal segmentation with considerably reduced execution time provided that the time series is divided into subseries using the dominant change points. Therefore, their properties merit further analysis in the future.

Segmentation of long real-world as well as artificial (synthetic) data is performed in another study (Gedikli *et al.* 2010). However, more detailed studies have been planned using long artificial data for segmentation by different algorithms.

ACKNOWLEDGEMENTS

The authors thank Dr. Athanasios Kehagias of the Faculty of Engineering, Aristotle University of Thessaloniki, Greece, for sharing his personal experience on time series segmentation and Dr. Pierre Hubert of Université P. & M. Curie, Paris, France for sharing his software on automatic segmentation algorithms online. The AUG-Segmenter is available for supply to interested readers after a request is made to the authors at gedikliab@itu.edu.tr, haksoy@itu.edu.tr or neu@itu.edu.tr. Revision of this paper was done when the second author (HA) was working at Umwelt und Technik, Fakultät III, Leuphana Universität Lüneburg at Campus Suderburg as an experienced researcher invited by the German Alexander von Humboldt Foundation.

REFERENCES

Adeloye, A. J. & Montaseri, M. 2002 Preliminary streamflow data analyses prior to water resources planning study. *Hydrol. Sci. J.* **47** (5), 679–692.

Aguilar, E., Auer, I., Brunet, M., Peterson, T. C. & Wieringa, J. 2003 *Guidelines on Climate Metadata and Homogenization*. WMO-TD no. 1186. World Meteorological Organization, Geneva, Switzerland.

Aksoy, H. 2007 Hydrological variability of the European part of Turkey. *Iran. J. Sci. Technol., Trans. B* **31** (B2), 225–236.

Aksoy, H., Gedikli, A., Unal, N. E. & Kehagias, A. 2008*a* Fast segmentation algorithms for long hydrometeorological time series. *Hydrol. Process.* **22** (23), 4600–4608.

Aksoy, H., Unal, N. E., Alexandrov, V., Dakova, S. & Yoon, J. 2008*b* Hydrometeorological analysis of northwestern Turkey with links to climate change. *Int. J. Climatol.* **28** (8), 1047–1060.

Aksoy, H., Unal, N. E. & Gedikli, A. 2007 Letter to the editor. *Stoch. Environ. Res. Risk Assess.* **21**, 447–449.

Baseville, M. & Nikiforov, I. V. 1993 *Detection of Abrupt Changes: Theory and Application*. PRT Prentice Hall, New York.

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

Chen, J. & Gupta, A. K. 2000 *Parametric Statistical Change-point Analysis*. Birkhauser, Boston.

Chen, J. & Gupta, A. K. 2001 On change-point detection and estimation. *Commun. Statist.-Simul. Comput.* **30** (3), 665–697.

Dahamsheh, A. & Aksoy, H. 2007 Structural characteristics of annual precipitation data in Jordan. *Theor. Appl. Climatol.* **88**, 201–212.

DeGaetano, A. T. 2006 Attributes of several methods for detecting discontinuities in mean temperature series. *J. Clim.* **19** (5), 838–853.

Dobigeon, N. & Tourneret, J. Y. 2007 Joint segmentation of wind speed and direction using a hierarchical model. *Comput. Statist. Data Anal.* **51**, 5603–5621.

Fanta, B., Zaake, B. T. & Kachroo, R. K. 2001 A study of variability of annual river flow of the southern African region. *Hydrol. Sci. J.* **46** (4), 513–524.

Gedikli, A., Aksoy, H. & Unal, N. E. 2008 Segmentation algorithm for long time series analysis. *Stochast. Environ. Res. Risk Assess.* **22**, 291–302.

Gedikli, A., Aksoy, H., Unal, N. E. & Kehagias, A. 2010 Modified dynamic programming approach for offline segmentation of long hydrometeorological time series. *Stochast. Environ. Res. Risk Assess.* doi:10.1007/s00477-009-0335-x.

Hubert, P. 2000 The segmentation procedure as a tool for discrete modeling of hydrometeorological regimes. *Stochast. Environ. Res. Risk Assess.* **14**, 297–304.

Hubert, P., Bader, J.-C. & Bendjoudi, H. 2007 Un siècle de débits annuels du fleuve Sénégal. *Hydrol. Sci. J.* **52** (1), 68–73.

Hubert, P., Carbonnel, J. P. & Chaouche, A. 1989 Segmentation des series hydrométéorologiques–application à des séries de précipitations et de débits de l'afrique de l'ouest. *J. Hydrol.* **110** (3–4), 349–367.

Kehagias, A. 2004 A hidden Markov model segmentation procedure for hydrological and environmental time series. *Stochast. Environ. Res. Risk Assess.* **18**, 117–130.

Kehagias, A. & Fortin, V. 2006 Time series segmentation with shifting means hidden Markov models. *Nonlin. Process. Geophys.* **13**, 339–352.

Kehagias, A., Nidelkou, E. & Petridis, V. 2006 A dynamic programming segmentation procedure for hydrological and environmental time series. *Stochast. Environ. Res. Risk Assess.* **20**, 77–94.

Kehagias, A., Petridis, V., Nidelkou, E. *et al.* 2007 Reply by the authors to the letter by Aksoy.Stochast. *Environ. Res. Risk Assess.* **21**, 451–455.

Peterson, T. C., Easterling, D. R., Karl, T. R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Forland, E. J., Hanssen-Bauer, I., Alexandersson, H., Jones, P. & Parker, D. 1998 Homogeneity adjustments of in situ atmospheric climate data: a review. *Int. J. Climatol.* **18** (13), 1493–1517.

Scheffe, M. 1959 *The Analysis of Variance*. Wiley, New York.

Schwarz, G. E. 1978 Estimating the dimension of a model. *Annal. Statist.* **6** (2), 461–464.

Wijngaard, J. B., Klein Tank, A. M. G. & Können, G. P. 2003 Homogeneity of 20th century European daily temperature and precipitation series. *Int. J. Climatol.* **23**, 679–692.

Xiong, L. & Guo, S. 2004 Trend test and change-point detection for the annual discharge series of the Yangtze River at the Yichang hydrological station. *Hydrol. Sci. J.* **49** (1), 99–112.

First received 29 November 2008; accepted in revised form 11 June 2009. Available online 12 January 2010