Proceedings of the Seventeenth (2007) International Offshore and Polar Engineering Conference
Lisbon, Portugal, July 1-6, 2007

ISBN 978-1-880653-68-5; ISBN 1-880653-68-0(Set); ISSN 1098-6189 (Set)

A Comparison of Segmentation Procedures and Analysis of the Evolution of Spectral Parameters

José B. Hernández Ch.
Universidad Central de Venezuela
Caracas, Venezuela

Joaquín Ortega S.
CIMAT, A.C.
Guanajuato, Gto, Mexico

ABSTRACT

In this work we consider the evolution of power spectra of waves during a period of one year. Soukissian and Samalekos (2005) have proposed a segmentation method for significant wave height based on determining periods of stability, increase and decrease using time-series techniques. The second segmentation method is based on the mean value over a moving window, and uses a fixed-width band to determine the change-points in the register. We compare both segmentation methods for several spectral characteristics and give a statistical analysis of duration and intensity of sea states in each case.

KEY WORDS: Spectral analysis; stationary periods; time series; segmentation procedure.

INTRODUCTION

In this work we consider the evolution of power spectra of waves during a period of one year with data from one recording station situated at Waimea Bay, Hawaii. Using the wave-height record we calculate the spectra every 15 minutes in order to capture the short term evolution of some wave characteristics that can be obtained from the spectra. WAFO was used for obtaining the spectra and the spectral characteristics.

Soukisissian and Samalekos (2005) have proposed a segmentation method for significant wave height based on determining periods of stability, increase and decrease using time-series techniques. Their method is based on local linear regression and the initial and end points of the intervals are extreme points (local maxima and minima) of the time series. They use a cost function to determine the best configuration of intervals. We apply this method to some spectral characteristics and compare the results obtained with another segmentation method which will be described next.

The second segmentation method is based on calculating mean values over moving windows, and using a fixed-width band to determine change points in the wave-height data. Those intervals in which the values remain within a fixed-width interval around the mean are considered to be stationary, those in which the values go above (or below) will be considered increasing (or decreasing). In this way the stationary, increasing and decreasing intervals are determined. Both methods were implemented in MATLAB.

We will consider the following spectral characteristics: Significant wave height ($H_m = 4\sqrt{Var(X)}$), spectral moments of order zero ($m_0 = \int_0^\infty \hat{s}(\omega)d\omega$), and two ($m_2 = \int_0^\infty \omega^2 \hat{s}(\omega)d\omega$) and up-crossing peak periods ($Tp = 2\pi\sqrt{m_0/m_2}$).

After calculating the spectral characteristics the results were smoothed using a finite moving average filter of order 5, to get rid of the local noise, see Brockwell and Davis (1996) for details.

The time series we considered are from Station 10601 in Waimea Bay, Hawaii, with the following characteristics. Deployment latitude: 21°40.364' N, longitude: 158°06.949' W, water depth (m): 198.00. The time series has a sampling rate of 1.280 Hz.

The rest of the paper is organized as follows. In the next section we describe the Soukissian and Samalekos algorithm, and apply this method to the time series. Next we describe the band method and its application to the time series. In the following two sections we make an analysis of the results for both methods and give our conclusions.

SOUKISSIAN'S ALGORITHM

Consider a time series of significant wave height observations $H_m = h_1, h_2, \ldots, h_n$ with n terms; the goal is to find a k-segmentation of H_m, i.e. $H_m = H_{m1}, H_{m2}, \ldots, H_{mk}$ with H_{mi} disjoint and non-overlapping intervals. The first step in time series segmentation is to define a representation model that approximates the data in each segment. Once we find the representation model, the quality of the approximation is evaluated by a cost function to minimize the representation error. For this a linear regression model is used (Charbonnier, 2005), and the representation error is defined based on the sum of squares of distances between the actual values of the time series and the values of the representation model fitted.

The total cost of a k-segmentation is

$$COST = \sum_{i=1}^{k} cost(i,k) \quad (1)$$

where $cost(i,k), 1 \le i \le k$ is the cost of i-th segment of a k-segmentation.

The linear regression model is employed because H_m data exhibit

alternating occurrences of monotonically increasing and decreasing trends and the model is well suited to detect these features.

The algorithm initially creates a fine segmentation of n-1 segments and n breakpoints $[t_1,t_2],[t_2,t_3],\ldots,[t_{n-1},t_n]$ based on the raw data. This first partition is based on local extreme (local maxima and minima). Then a linear regression model is fitted for each segment of partition and the representation error is calculated.

Assuming $[s_i,e_i]$ where s and e denote the start and end points for i-th segment, the fitted values of the data $h_{s_i},h_{s_{i+1}},\ldots,h_{e_{i-1}},h_{e_i}$ is calculated with the linear regression model described as follows

$$h_i = \alpha_i + \beta_i t + \varepsilon_i$$
or
$$h_i = a_i + b_i t \quad (2)$$

for $s_i < t < e_i$ and $h_{s_i} < h_i < h_{ei}$, where t is time, h_t the significant wave height at time t, ε_t is the random error and a_i, b_i are the estimates of α_i and β_i respectively. The parameter a_i represents the intercept and b_i the slope (regression coefficients) of the regression lines. Next the algorithm merges the lowest cost pair of segments until the representation error is less than the maximum error defined by the user. This process gives all the increasing and decreasing intervals. For extracting the stationary sea states one proceeds by applying a criterion which for the increasing intervals is of the form $h_{e_i} \le h_{s_i}(100+p)\%$ and for decreasing intervals is of the form $h_{s_i} \le h_{e_i}(100+p)\%$. (see Soukissian and Samalekos, 2006; Labeyrie, 1990 and Athanassoulis et al., 1992). Next we present the results obtained for station 106.

Table 1. Statistics for significant wave height, H_m, max-error = 0.015, p = 4% (min)

	Increase	Decrease	Stationary	Total
Num. inter	337	368	446	1151
Minimum	15	15	15	15
1st Quartile	60	75	15	45
Median	120	150	90	120
Mean	159.21	180.82	114.99	148.98
3rd Quartile	220	240	165	210
Maximum	720	900	675	900
Variance	17383.83	19822.63	14082.11	17658.53

Table 2. Statistics for spectral moment or order zero, m_0, max-error = 0.001, p = 6%. (min)

	Increase	Decrease	Stationary	Total
Num. inter	384	428	373	1185
Minimum	15	15	15	15
1st Quartile	45	71.25	15	45
Median	105	135	45	105
Mean	153.20	175.20	100.98	144.71
3rd Quartile	195	225	135	195
Maximum	1020	1260	675	1260
Variance	21877.57	26517.93	13546.69	21858.55

Table 3. Statistics for spectral moment or order two, m_2, max-error = 0.0025, p = 7%. (min)

	Increase	Decrease	Stationary	Total
Num. inter	402	415	372	1189
Minimum	15	15	15	15
1st Quartile	60	60	15	45
Median	120	135	45	105
Mean	157.09	174.90	96.09	144.22
3rd Quartile	180	210	135	180
Maximum	960	1275	810	1275
Variance	22892.88	33110.38	13716.94	24659.77

Table 4. Statistics for Up-crossing peak periods, Tp, max-error = 0.07, p = 3% (min)

	Increase	Decrease	Stationary	Total
Num. inter	375	366	417	1158
Minimum	15	15	15	15
1st Quartile	60	75	15	45
Median	120	135	105	120
Mean	158.68	166.52	122.37	148.08
3rd Quartile	210	225	195	210
Maximum	855	855	705	855
Variance	16583.48	15690.37	14097.96	15761.49

Notice that for each spectral characteristic we choose a different value for the max-error and for p, because the scale of each spectral characteristic is different. The range for H_m is [0.8298, 5.3873], for m_0 it is [0.0431, 1.82], for m_2 it is [0.041, 2.1926] and for Tp it is [3.4958, 12.2582].

The slope for each sea state and each spectral characteristic was also calculated. The mean values were as follows: significant wave height: increasing slope 0.035, decreasing slope -0.030; spectral moment of order zero: increasing slope 0.013, decreasing slope -0.012; spectral moment of order two: increasing slope 0.017, decreasing slope -0.016; up-crossing peak periods: increasing slope 0.090, decreasing slope -0.079. To give an idea of the corresponding distributions we show the corresponding boxplots in figures 9 – 12.

ALGORITHM OF BANDS

We now describe the band algorithm. This segmentation procedure is based on calculating the mean values over a moving window with fixed bandwidth. We start by calculating the mean of the first two data points and then we add successively new points and recalculate the mean.

Let m_i be the mean value of $X_1,\ldots, X_i$ and let $2h$ be the chosen bandwidth. If the next point X_{i+1} belongs to the interval $[m_i-h, m_i+h]$ the mean is recalculated adding the new point X_{i+1}, to obtain m_{i+1}, and the new interval is $[m_{i+1}-h, m_{i+1}+h]$. This process continues until the new point does not belong to the interval, in which case it is marked as a breakpoint. The previous points form a stationary interval. The process starts again. If successive points fall above (or below) the corresponding fixed-width band, they determine an increasing (decreasing) interval.

The algorithm is as follows

1. Read data file of spectral characteristic
2. Set j = 1, n = length of data file
3. While $j<n$-2,
m_i = mean (p_i,…,p_f), where p_i is starting point and p_f is endpoint
If $p_{f+1} \in [m_i - h, m_i + h]$ then $p_f = p_{f+1}$
Else
Set p_f as breakpoint
Set $p_i = p_f$ and $p_f = p_f + 1$
End
4. Go to 3.

Next we present the results of applying this algorithm to the data of station 106.

Table 5. Statistics for significant wave height, H_m, bandwidth = 0.125 (min)

	Increase	Decrease	Stationary	Total
Num. inter	792	848	1179	2819
Minimum	15	15	15	15
1st Quartile	30	45	30	30
Median	45	45	45	45
Mean	55.63	56.76	67.26	60.84
3rd Quartile	60	75	90	75
Maximum	360	240	600	600
Variance	1253.18	866.40	4358.36	2464.00

Table 6. Statistics for the spectral moment of order zero, m_0, bandwidth = 0.035 (min)

	Increase	Decrease	Stationary	Total
Num. inter	748	798	1054	2600
Minimum	15	15	15	15
1st Quartile	30	45	30	30
Median	45	45	45	45
Mean	55.99	57.31	79.58	65.96
3rd Quartile	63.75	75	90	75
Maximum	375	240	1110	1110
Variance	1350.76	911.12	10484.11	5042.16

Table 7. Statistics for the spectral moment of order two, m_2, bandwidth = 0.06 (min)

	Increase	Decrease	Stationary	Total
Num. inter	740	773	1009	2522
Minimum	15	15	15	15
1st Quartile	30	45	30	30
Median	45	45	45	45
Mean	59.23	61.07	79.74	68.00
3rd Quartile	75	75	90	75
Maximum	300	300	1020	1020
Variance	1542.44	1686.65	13102.26	6299.97

Table 8. Statistics for Up-crossing peak periods, Tp, bandwidth = 0.28 (min)

	Increase	Decrease	Stationary	Total
Num. inter	809	751	1114	2674
Minimum	15	15	15	15
1st Quartile	30	45	30	30
Median	45	45	45	45
Mean	56.68	57.82	73.80	64.13
3rd Quartile	75	75	90	75
Maximum	210	255	705	705
Variance	1012.89	1067.75	5344.99	2898.31

In this case, as with Soukissian's algorithm and for the same reason, the value of the bandwidth parameter is different for each spectral characteristic. The value of bandwidth h for each spectral characteristics were chosen after a many proof with several values of h and analyzing the number of intervals, duration of intervals, starting and ending point of it, etc.

Again, we calculated the slope for each sea state and each spectral characteristic, obtaining the following results for the mean values: significant wave height: increasing slope 0.054, decreasing slope -0.051; spectral moment of order zero: increasing slope 0.018, decreasing slope -0.017; spectral moment of order two: increasing slope 0.028, decreasing slope -0.026; up-crossing peak periods: increasing slope 0.121, decreasing slope -0.119. The corresponding boxplots are given in figures 9 – 12.

ANALYSIS OF RESULTS

In the analysis of wave data with both algorithms one can see that the number of breakpoints as well as the distribution of interval length are different. The band algorithm shows more breakpoints than Soukissian's algorithm. Moreover, the breakpoints do not always match one another. In the next table (Table 9) we give the total number of breakpoints and matched breakpoints. The fourth column of table 9 shows the percentage of breakpoints obtained with Soukissian's algorithm that matched with those obtained the band algorithm. As can be seen the best agreement was obtained using the significant wave height time series. And in Figure 1 we show a segment of the time series with both breakpoints for significant wave height.

Table 9. Number of matches breakpoints

	Bandwidth algorithm	Soukissian algorithm	Number or matchpoint	Percent
H_m	2819	1151	938	91.49%
m_0	2600	1185	957	82.64%
m_2	2522	1189	974	82.19%
Tp	2674	1158	956	80.40%

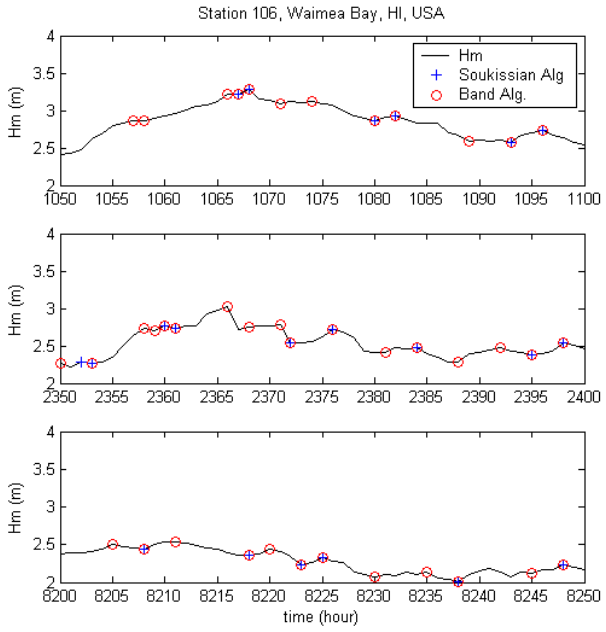

Figure 1. Three segments of significant wave height calculated from the data of Station 106.

Next are two tables that show number of breakpoint that match between two different spectral characteristics with both algorithms

Table 10. A comparison of segmentation results for different spectral characteristics with the band algorithm.

Spectral characteristics	Number of breakpoints	Number of matchpoints	Percentage
H_m m_0	2819 2600	2402	85.21%
m_0 Tp	2600 2674	799	29.88%
m_2 Tp	2522 2674	894	33.43%
H_m m_2	2819 2522	1112	39.45%
H_m Tp	2819 2674	882	31.29%
m_0 m_2	2600 2522	1607	61.81%

Table 11. A comparison between different spectral characteristics with Soukissian's algorithm.

Spectral characteristics	Number of breakpoints	Number of matchpoints	Percentage
H_m m_0	1151 1185	930	78.48%
m_0 Tp	1185 1158	198	16.71%
m_2 Tp	1189 1158	246	20.69%
H_m m_2	1151 1189	282	23.72%
H_m Tp	1151 1158	192	16.58%
m_0 m_2	1185 1189	283	23.80%

In the boxplots (Figs. 2 ~ 5) the segmentation obtained with both algorithms can be compared For the band algorithm the duration of stationary intervals is larger than the duration of periods of increase and decrease for all spectral characteristics considered, while for Soukissian's algorithm it is the other way round. For the band algorithm increase and decrease periods have similar distributions while for Soukissian's algorithm decrease periods tend to longer than periods of increase. The distributions, however, vary with the spectral characteristic being considered. 0On the other hand, for both algorithms the percentages of number of different kinds of intervals are similar for all spectral characteristics as can be seen in table 12.

Table 12. Percentage and number of different types of intervals obtained with both algorithms for station 106

Significant wave height				
	Soukissian algorithm		Band algorithm	
	Num. interv	Percent	Num. interv	Percent
Increase	337	29.28%	792	28.10%
Decrease	368	31.97%	848	30.08%
Stationary	446	38.75%	1179	41.82%

Spectral moment of order zero				
	Soukissian algorithm		Band algorithm	
	Num. interv	Percent	Num. interv	Percent
Increase	384	32.41%	748	28.77%
Decrease	428	36.12%	798	30.69%
Stationary	373	31.48%	1054	40.54%
Spectral moment of order two				
	Soukissian algorithm		Band algorithm	
	Num. interv	Percent	Num. interv	Percent
Increase	402	33.81%	740	29.34%
Decrease	415	34.90%	773	30.65%
Stationary	372	31.29%	1009	40.01%
Up-crossing peak periods				
	Soukissian algorithm		Band algorithm	
	Num. interv	Percent	Num. interv	Percent
Increase	375	32.38%	809	30.25%
Decrease	366	31.61%	751	28.09%
Stationary	417	36.01%	1114	41.66%

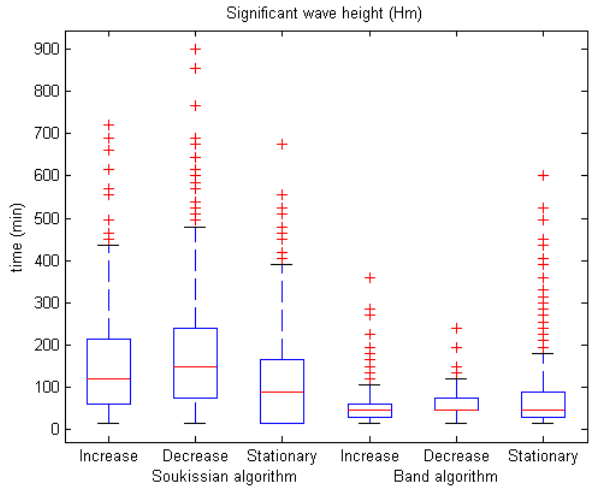

Figure 2. Boxplot for segmentation of H_m obtained with Soukissian's algorithm (left) and band algorithm (right)

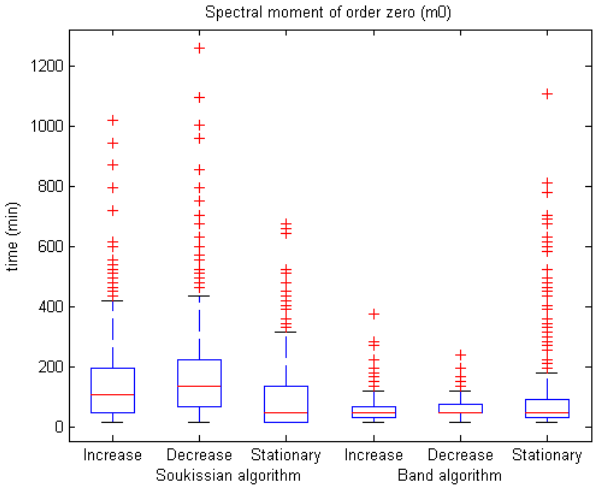

Figure 3. Boxplot for segmentation of m_0 obtained with Soukissian's algorithm (left) and band algorithm (right)

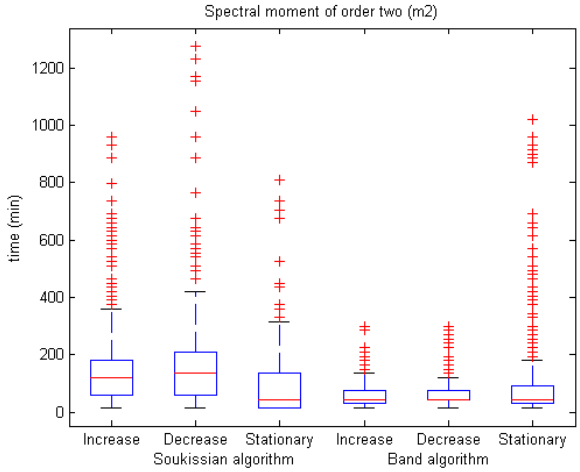

Figure 4. Boxplot for segmentation of m_2 obtained with Soukissian's algorithm (left) and band algorithm (right).

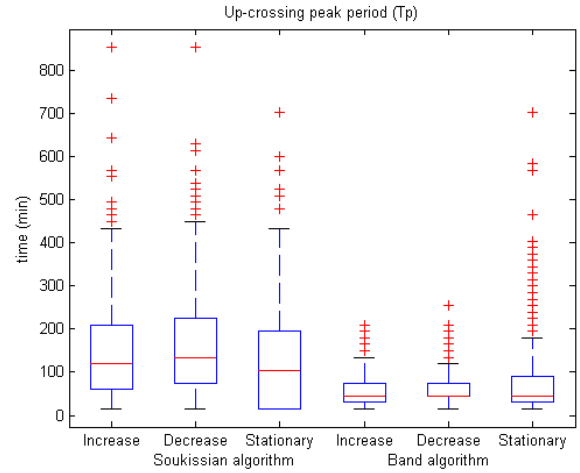

Figure 5. Boxplot for segmentation of *Tp* obtained with Soukissian's algorithm (left) and band algorithm (right).

In figures 6 – 8, one for each sea state, we show boxplots of all spectral characteristics for both algorithms.

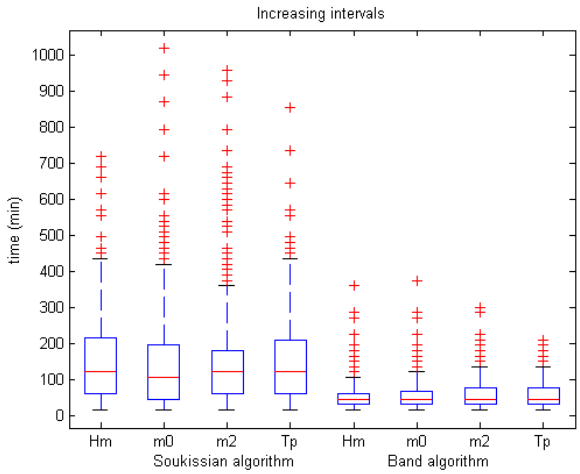

Figure 6. Boxplot for increasing intervals for all spectral characteristics obtained with both algorithms

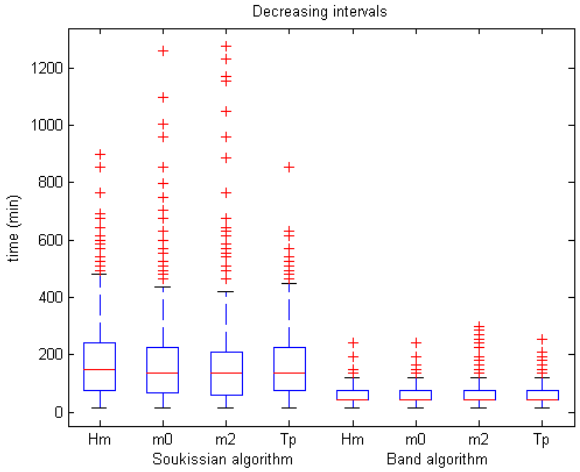

Figure 7. Boxplot for decreasing intervals for all spectral characteristics obtained with both algorithms.

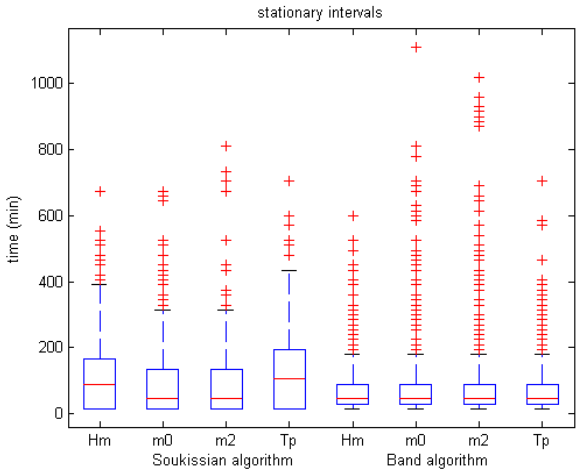

Figure 8. Boxplot for stationary intervals for all spectral characteristics obtained with both algorithms.

In figures 9-12, one for each spectral parameter, we show boxplots for the absolute value of the slope for increasing and decreasing intervals.

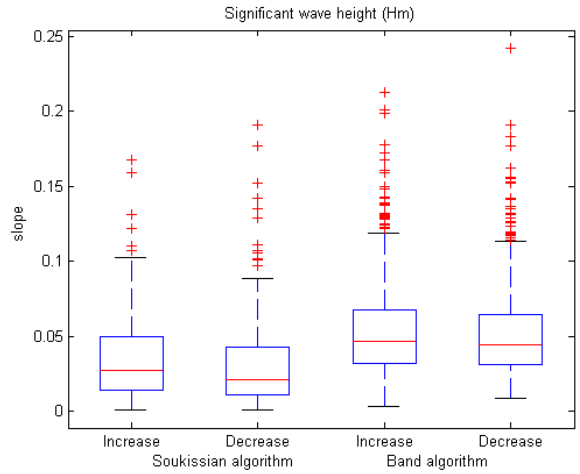

Figure 9. Boxplot for the absolute value of the slope for intervals of increase and decrease for significant wave height.

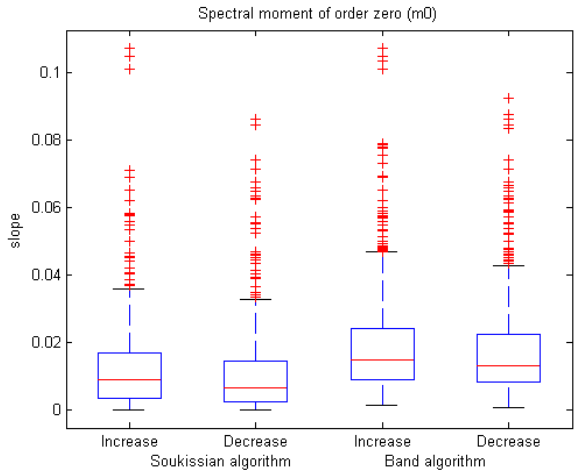

Figure 10. Boxplot for the absolute value of the slope for intervals of increase and decrease for the spectral moment of order zero.

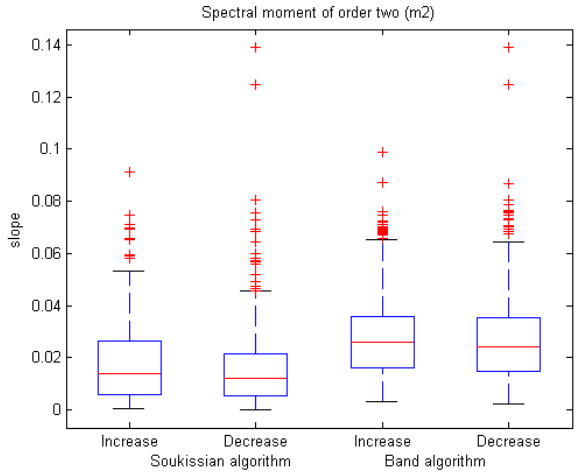

Figure 11. Boxplot for the absolute value of the slope for intervals of increase and decrease for the spectral moment of order two.

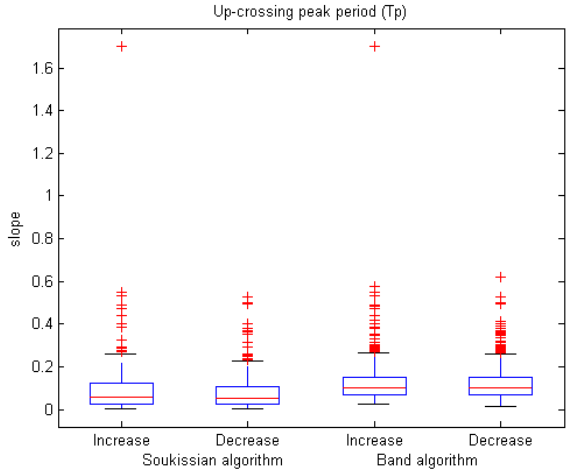

Figure 12. Boxplot for the absolute value of the slope for intervals of increase and decrease for the up-crossing peak period.

As can be seen from these previous four figures, the slopes differ from one algorithm to the other, but the increase and decrease slopes are very similar for any given algorithm. In all cases the distribution for the decreasing slope seems to have less dispersion than the distribution for the increasing case.

As can be seen from tables 1 and 5 and figure 8 the duration of intervals for the band algorithm is shorter than the duration for Soukissian's algorithm for the corresponding spectral parameters. As an example from Tables 1 and 5 one can see that the mean duration of stationary intervals for significant wave height for the band algorithm is 67.26 min while with Soukissian's algorithm is 114.99 min. In fact, the mean duration for all intervals is 60.84 min for the band algorithm and for Soukissian's algorithm is 148.98 min for significant wave height.

In figure 13, it can be seen an interval where the segmentations for significant wave height is similar in number of breakpoints, there are 12 for Soukissian's algorithm and 18 for the band algorithm. In this interval you can see that both algorithms detect some breakpoint in those points where data reach local maximum or minimum, but not at the same points. Whereas Soukissian's algorithms detect breakpoints only in local maxima and minima band algorithm can detect breakpoint in any point (local maxima, minima or any other point)

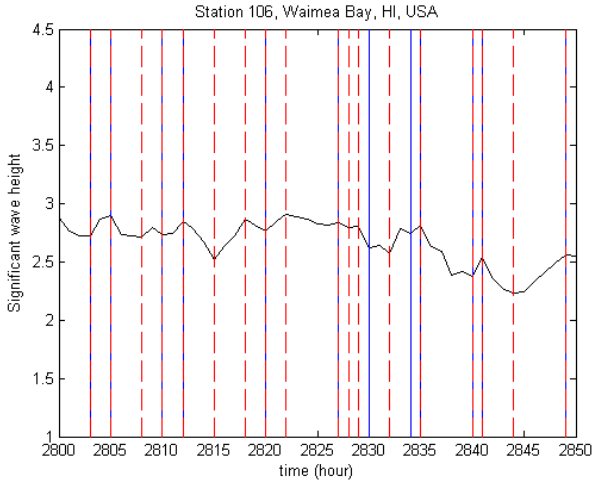

Figure 13. Segmentation for significant wave height, Station 106. The Soukissian segmentation is shown in blue (solid line), the band segmentation in red (dashed line).

CONCLUSIONS

We have considered two segmentation procedures for detecting change-point in a time series: Soukissian's algorithm and the band algorithm. These algorithms were used on the set of data coming from Station 106 for four different spectral characteristics.

The results were different with regard to the number of intervals or change-points and in the distribution of duration of intervals. The intervals obtained with Soukissian's algorithm have longer duration than the intervals obtained with the band algorithm for all spectral characteristics, but the distribution of duration is regular for both algorithms.

One disadvantage of Soukissian's algorithm is that took a long time to run with a large set of data. Despite that, the algorithm works fine to detect the change-points of a large data set. On the other hand, the band algorithm is fast and easy to use, although it tends to find much more intervals than Soukissian's algorithm. Nevertheless, for the band algorithm we are looking for ways of fixing bandwidth parameters automatically in order to avoid subjectivity of users in the choice.

For different spectral characteristics but with the same algorithm,

similar results for the number of breakpoints were obtained, and also the mean duration time for each kind of interval was similar. Significant wave height is the most commonly used spectral characteristic to determine segmentation of wave height time series. One advantage of using significant wave height is that it can be related to the evolution of sea before calculating the breakpoints, and the same holds for up-crossing peak periods. But in view of the fact that there are not large differences in the results from one spectral characteristic to another, one can use any of them.

For both algorithms parameters are fixed by the user but once this is done, the calculation of intervals is automatic, which avoids the subjective selection of intervals. We are looking at automatic methods of parameter selection, but so far results are not satisfactory.

In our view both algorithms work fine to detect change-point of a time series when the sea conditions are 'normal', it necessary to study the sea in presence of extreme conditions, for example during a hurricane, in order to establish if they work well or not.

AKNOWLEDGEMENTS

The software WAFO developed by the Wafo group at Lund University of Technology, Sweden was used for the calculation of all spectra and spectral characteristics. This software is available at http://www.maths.lth.se/matstat/wafo.

The data for station 106 were furnished by the Coastal Data Information Program (CDIP), Integrative Oceanographic Division, operated by the Scripps Institution of Oceanography, under the sponsorship of the U.S. Army Corps of Engineers and the California Department of Boating and Waterways (http://cdip.ucsd.edu/).

This work was partly done during a visit of the first author to the Centro de Investigación en Matemáticas, Guanajuato, México, which was partially supported by the Universidad Central de Venezuela and CIMAT. Their support is gratefully acknowledged.

This work was partially supported by CONACYT Mexico, grant 52554.

REFERENCES

Athanassoulis, GA, Vranas, PB and Soukissian, TH, (1992). "A new model for leng-term stochastic analysis and prediction. I: Theoretical Background", *Journal of Ship Research*, 36(1), 1-16.

Brockwell, PJ, and Davis, RA, (1996). "*Introduction to Time Series and Forecasting*". Springer-Verlag, New York Inc, New York, p 24.

Charbonnier, S. (2005). "On line extraction of temporal episodes from ICU high-frecuency data: A visual support for signal interpretation", *Computer Methods and Programs in Biomedicine*, pp 78, 115-132.

Keogh, E, Chu, S, Hart, D and Pazzani, M (2001). "An online algorithm for segmenting time series", *The IEEE International Conference on Data Mining (ICDM)*

Labeyrie, J. (1990). "Stationary and transient states fo random seas". *Marine Structures*, 3-1, 43-58.

Soukissian, TH, and Samalekos, PE, (2006). "Analysis of the Duration and Intensity of Sea States Using Segmentation of Significant Wave Height Time Series", *Proc. ISOPE 2006*, Vol. 3, 107-113.