ARTICLE IN PRESS

CENG-02511; No of Pages 10

Coastal Engineering xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Coastal Engineering

journal homepage: www.elsevier.com/locate/coastaleng

Review

Analysis of clustering and selection algorithms for the study of multivariate wave climate

Paula Camus [a,*], Fernando J. Mendez [a], Raul Medina [a], Antonio S. Cofiño [b]

[a] *Environmental Hydraulics Institute "IH Cantabria", Universidad de Cantabria, Spain*

[b] *Santander Meteorology Group, Dep. of Applied Mathematics and Computer Sciences, Universidad de Cantabria, Spain*

ARTICLE INFO

Article history:
Received 12 April 2010
Received in revised form 8 February 2011
Accepted 14 February 2011
Available online xxxx

Keywords:
Data mining
~~Hindcast~~
K-means
~~Long-term distribution~~
Maximum dissimilarity algorithm
Probability density function
~~Propagation~~
Reanalysis database
Self-organizing maps

ABSTRACT

Recent wave reanalysis databases require the application of techniques capable of managing huge amounts of information. In this paper, several clustering and selection algorithms: K-Means (KMA), self-organizing maps (SOM) and Maximum Dissimilarity (MDA) have been applied to analyze trivariate hourly time series of met-ocean parameters (significant wave height, mean period, and mean wave direction). A methodology has been developed to apply the aforementioned techniques to wave climate analysis, which implies data pre-processing and slight modifications in the algorithms. Results show that: a) the SOM classifies the wave climate in the relevant "wave types" projected in a bidimensional lattice, providing an easy visualization and probabilistic multidimensional analysis; b) the KMA technique correctly represents the average wave climate and can be used in several coastal applications such as longshore drift or harbor agitation; c) the MDA algorithm allows selecting a representative subset of the wave climate diversity quite suitable to be implemented in a nearshore propagation methodology.

Contents

1. Introduction

In the last decade, long-term wave databases from numerical models have been developed improving the knowledge of deep water wave climate, especially at locations where instrumental data is not

* Corresponding author at: Environmental Hydraulics Institute, IH Cantabria, Universidad de Cantabria, E.T.S.I. Caminos Canales y Puertos, Avda de los Castros s/n, 39005, Santander, Spain. Tel.: +34 942 201810; fax: +34 942 201860.
E-mail address: camusp@unican.es (P. Camus).

0378-3839/$ – see front matter © 2011 Published by Elsevier B.V.
doi:10.1016/j.coastaleng.2011.02.003

available (see, for instance, Dodet et al., 2010; Pilar et al., 2008; Ratsimandresy et al., 2008; Weisse et al., 2002). These reanalysis (or hindcast) databases present the advantage of having an adequate spatial and temporal resolution, not presenting the problems of instrumental buoys such as missing data or sparse locations. This increase of information requires different data mining techniques, in particular clustering and selection techniques, to deal with such amounts of information and to provide an easier analysis and description of the multidimensional wave climate. An example of an application of a classification process to obtain representative sea Q3 states can be found in Abadie et al. (2006).

The reanalysis database provides long-term hourly time series (say, >300,000 data) of several sea state met-ocean variables (such as significant wave height—H_s, mean period—T_m, mean wave direction—θ_m, wind velocity, wind direction, swell significant wave height, or even, the directional spectra), which can be used for the statistical characterization of wave climate. Usually, the long-term distribution of wave climate is limited to the analysis of significant wave height by means of parametric probabilistic models. The multivariate analysis of wave climate (e.g. of H_s, T_m and θ_m) is usually carried out defining the empirical joint probability density function $p(H_s, T_m,$ and $\theta_m)$, sorting the observed values in classes and visualizing the results using two-dimensional histograms of H_s and T_m for a given directional sector $\Delta\theta$ (Holthuijsen, 2007). The development of an analytical parametric multivariate model is not an easy task due to the complicated form of the corresponding probability density functions (Athanassoulis and Belibassakis, 2002). The availability of an analytical expression for the probability density function (pdf) is very useful for several applications, e.g. the extrapolation to calculate extreme values or the integration to obtain different return value quantiles. ~~These models allow extracting useful information~~, the joint analysis of all the variables is difficult and the visualization is limited to 2D marginal pdfs. Therefore, a statistical tool able of representing graphically multivariate data is highly demanded.

On the other hand, the characterization of nearshore wave climate requires long-term time series of wave parameters at a particular location. The available information is usually located in deep water and must be transferred to shallow water using a state-of-the-art wave propagation model capable of simulating the most important wave transformation processes. The huge number of sea states to propagate leads to different strategies which aim to reduce the computational effort. The more common methodologies consist of replacing all available data with a small number of representative sea states, which are later propagated to shallow water areas. A transfer function is defined allowing the propagation of all the sea states of the long-term series of wave parameters in deep waters by means of an Q4 interpolation algorithm (Groeneweg et al., 2007; Stansby et al., ~~2006~~). The success of the interpolation scheme depends totally on the correct selection of the most representative sea states, requiring new algorithms that synthesize the huge amount of information.

Several clustering methods have been developed in the field of data mining to efficiently deal with huge amounts of information. These techniques extract features from the original N data, giving a more compact and manageable representation of some important properties contained in the data. Standard methods in data mining include clustering techniques (to obtain a set of reference vectors representing the data), dependency graphs (to represent dependencies among the variables), association rules, etc. The K-means algorithm (KMA) and the self-organizing maps (SOM) are some of the most popular clustering techniques in this field. The KMA computes a set of M prototypes or centroids, each of them characterizing a group of data, formed by the vectors in the database for which the corresponding centroid is the nearest one (Hastie et al., 2001). A SOM algorithm is a version of the KMA that preserves the topology of the data in the original space in a low-dimensional lattice. The cluster centroids are forced with a neighborhood adaptation mechanism to a space with smaller dimension (usually a two-dimensional regular lattice) and which is spatially organized. A number of applications of SOM for different geophysical parameters have been presented over the last decade (Cavazos, 1997; Gutiérrez et al., 2004, 2005; Lin and Chen, 2005; Liu and Weisberg, 2005; Solidoro et al., 2007).

Regarding the selection algorithms, the requirements of high-throughout screening and combinatorial synthesis in pharmaceutical discovery programs have led to much interest in the development of computer-based methods for selecting sets of structurally diverse compounds from chemical databases. Dissimilarity-based compound selection has been suggested as an effective method, as it involves the identification of a subset comprising the M most dissimilar molecules in a database containing N molecules (Snarey et al., 1997). One Q5 subclass of these selection algorithms, referred to as maximum-dissimilarity algorithm (MDA), has been considered. The subset selected by this algorithm is distributed fairly evenly across the space with some points selected in the outline of the data space.

The objectives of this work are to develop numerical tools for: a) describing graphically multivariate wave climate; b) describing statistically multivariate wave climate; c) defining a propagation strategy consisting of a selection of a reduced number of multidimensional sea states representative of the wave climate in deep waters to be propagated to shallow water. For this reason, we adapted the above-mentioned algorithms to analyze the trivariate (H_s, T_m, and θ_m) time series at a specific location and compare their performance in the proposed objectives.

In Section 2, the KMA, the SOM and the MDA are described and the differences between them are established. Section 3 gives a brief description of the data used to define wave climate at a particular area in Galicia (Spain). The proposed methodology to analyze the trivariate wave climate is presented in Section 4. Some results are described in detail in Section 5. Finally, conclusions are given in Section 6.

2. Clustering and selection algorithms

The initial database is composed of N three-dimensional vectors, defined as $X=\{x_1,x_2,\ldots,x_N\}$ where $x_i=\{H_{s,i},T_{m,i},\theta_{m,i}\}$. In order to generalize the algorithms to be valid for different met-ocean parameters, in this section we used a notation for ~~n~~-dimensional data ($n=3$ in this work) and x_k is defined as $x_{1k}=H_{s,k}$, $x_{2k}=T_{m,k}$ and $x_{3k}=\theta_{m,k}$.

2.1. K-means algorithm (KMA)

The KMA clustering technique divides the high-dimensional data space into a number of clusters, each one defined by a prototype and formed by the data for which the prototype is the nearest.

Given a database of n-dimensional vectors $X=\{x_1,x_2,\ldots,x_N\}$, where N is the total amount of data and n is the dimension of each data $x_k=\{x_{1k},\ldots,x_{nk}\}$, KMA is applied to obtain M groups defined by a prototype or centroid $v_k=\{v_{1k},\ldots,v_{nk}\}$ of the same dimension of the original data, being $k=1,\ldots,M$. The classification procedure starts with a random initialization of the centroids $\{v_1^0,v_2^0,\ldots,v_M^0\}$. On each iteration r, the nearest data to each centroid are identified and the centroid is redefined as the mean of the corresponding data. For example, on the $(r+1)$ step, each data vector x_i is assigned to the jth group, where $j=\min\{\|x_i-v_j^r\|, j=1,..,M\}$, $\|\|$ defines the Euclidean distance and v_j^r are the centroids on the r step. The centroid is updated as:

$$v_j^{r+1} = \sum_{x_i \in C_j} \frac{x_i}{n_j} \quad (1)$$

where n_j is the number of elements in the jth group and C_j is the subset of data included in group j. The KMA iteratively moves the centroids minimizing the overall within-cluster distance until it converges and data belonging to every group are stabilized (more details in Hastie Q6 et al., 2001).

Please cite this article as: Camus, P., et al., Analysis of clustering and selection algorithms for the study of multivariate wave climate, Coast. Eng. (2011), doi:10.1016/j.coastaleng.2011.02.003

The K-means algorithm has been applied to a sample of $N=1000$ two-dimensional data to obtain a number of $M=16$ clusters. In Fig. 1, the initialization of centroids $\{v_1^0,\ldots,v_{16}^0\}$, the updating (represented by its tracks) and the final prototypes $\{v_1,\ldots,v_{16}\}$ are shown. The data corresponding to each cluster is represented in the same color as its prototype. The separation lines between different clusters correspond to the Voronoi diagram associated with the centroid.

2.2. Self-organizing maps (SOM)

The SOM automatically extract patterns or clusters of high-dimensional data and project them into a bidimensional organized space, allowing an intuitive visualization of the classification and the transformation of the distributions from the high-dimensional space into Probability Density Functions (PDF) on the lattice (Kohonen, 2000).

The algorithm is similar to the KMA, starting from an initialization of the reference vectors $\{v_1^0,\ldots,v_M^0\}$ and the prototypes are adjusted iteratively to data trying to minimize an overall within-cluster distance from the data vectors v_j to the corresponding centroid vector x_i for each cluster j.

The training proceeds in cycles: during each training cycle, each of the data vectors x_i is considered, and the 'winning' centroid vector $v_{w(i)}$ is found to be the one closest to the data vector:

$$\|v_{w(i)}-x_i\| = \min_j \left\{\|v_j-x_i\|, j=1,\ldots,M\right\} \tag{2}$$

where $1 \le w(i) \le M$ is the index of the winning reference vector.

The training procedure includes a neighborhood adaptation mechanism in the lattice of projection, so not only the winning centroid moves toward the data vector but also the neighboring centroids in the lattice are adapted towards the sample vector:

$$v_j = v_j + \alpha h(w(i),j)\left(x_i-v_j\right), j=1,\ldots,M \tag{3}$$

where $0 \le \alpha \le 1$ is the learning rate and controls the velocity of the adaptation process. The function ~~$h(j,k)$~~ is a neighborhood kernel on the SOM lattice, which determines the rate of change around the winning centroid and which projects the topological relationships in the data space onto the lattice. This means that similar clusters in the multidimensional space are located together in the lattice of projection. The self-organizing maps (bidimensional projections with spatial organization) can be rectangular or hexagonal, the number of neighbors being 4 or 6 respectively. Each cluster of a SOM is defined by two vectors: one in the data space v_j (prototype) and the other one (m_j, n_j) describing the position on the lattice (Fig. 2). For a given SOM of size $M=A\cdot B$, the jth index of a cluster is related with the lattice dimensions and its position in the lattice by the expression: $j=B\cdot(m-1)+n$.

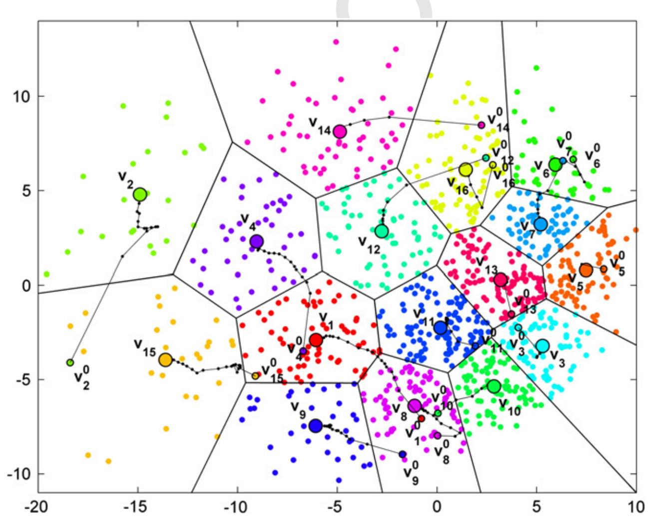

Fig. 1. KMA clustering: initialization $\{v_1^0,\ldots,v_{16}^0\}$, updating tracks and final centroids $\{v_1,\ldots,v_{16}\}$ with their corresponding clusters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

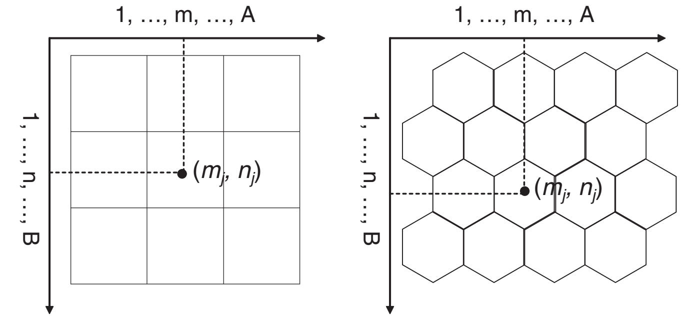

Fig. 2. SOM lattice of projection: rectangular (left) and hexagonal (right).

In Fig. 3, the $M=16$ SOM centroids have been randomly initiated over the bidimensional sample considered previously in the description of KMA. The initial centroids and their updating tracks are represented in the same color as the corresponding final centroid. As a consequence of the neighborhood kernel, the SOM behaves like a flexible lattice folding onto the cloud formed by the data in the original n dimensional space. The final centroids and lattice are also shown in Fig. 3.

2.3. Maximum dissimilarity algorithm (MDA)

The aim of MDA is to select a representative subset of size M from a database of size N. Therefore, given a data sample $X=\{x_1,x_2,\ldots,x_N\}$ consisting of N n-dimensional vectors, a subset of M vectors $\{v_1,\ldots,v_M\}$ representing the diversity of the data is obtained by applying this algorithm. The selection starts initializing the subset by transferring one vector from the data sample$\{v_1\}$. The rest of the M-1 elements are selected iteratively, calculating the dissimilarity between each remaining data in the database and the elements of the subset and transferring the most dissimilar one to the subset. The process finishes when the algorithm reaches M iterations. This algorithm was first described by Kennard and Stone (1969). Many variants, depending upon the precise implementation of the initialization and the definition of the most dissimilar vector, are available (Willet, 1996). In this work, the initial data of the subset is considered to be the vector with the largest sum of dissimilarities relative to the others within the data sample. In the selection process, the dissimilarity between each remaining vector in the database and each vector in the subset is calculated, and a unique dissimilarity between each vector in the database and the subset is established to define the most dissimilar one. In this work, the MaxMin version of the algorithm has been considered.

For example, if the subset is formed by R ($R \le M$) vectors, the dissimilarity between the vector i of the data sample N–R and the j vectors belonging to the R subset is calculated:

$$d_{ij} = \|x_i-v_j\|; i=1,\ldots,N-R; j=1,\ldots,R. \tag{4}$$

Subsequently, the dissimilarity $d_{i,subset}$ between the vector i and the subset R, is calculated as:

$$d_{i,subset} = \min\left\{\|x_i-v_j\|\right\}; i=1,\ldots,N-R; j=1,\ldots,R. \tag{5}$$

Once the N–R dissimilarities are calculated, the next selected data is the one with the largest value of $d_{i,subset}$.

Please cite this article as: Camus, P., et al., Analysis of clustering and selection algorithms for the study of multivariate wave climate, Coast. Eng. (2011), doi:10.1016/j.coastaleng.2011.02.003

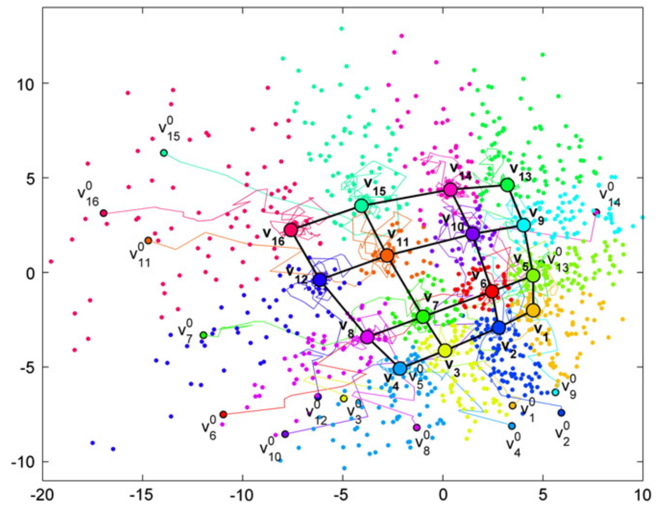

Fig. 3. SOM technique: initialization $\{v_1^0,\ldots,v_{16}^0\}$, updating tracks, final centroids $\{v_1,\ldots,v_{16}\}$ with its corresponding clusters and the final projection lattice. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

MDA has an expected time complexity of $O(M^2N)$ for ~~m-~~member subsets from an N-member database. The more efficient algorithm $O(MN)$ developed by Polinsky et al. (1996) has been considered. In this case, the definition of the distance $d_{i,subset}$ does not imply the calculation of the distance between the different vectors d_{ij}. For example, in the selection of the ~~r th~~ vector, the distance $d_{i,subset}$ is defined as the minimum distance between the vector i of the data sample (consisting of N-~~$(R-1)$~~ vectors at this cycle) and the last vector transferred to the subset R, and the minimum distance between the vector i and the R-1 vectors of the subset determined in the previous cycle:

$$d_{i,subset}^{min} = min\left[d_{i,R}, d_{i,subset(R-1)}^{min}\right]. \quad (6)$$

The subset of size $M = 16$ obtained by the maximum dissimilarity algorithm applied to the same sample used with the classification techniques is shown in Fig. 4. The subset vectors are represented by the larger dots and have been numbered in the order of selection. The first selected vector $\{v_1\}$ is the one that is most dissimilar to the rest of the data, representing one of the points located on the edge of the data space. Then the point $\{v_2\}$is selected, representing the one which is most dissimilar from the first one, located on the opposite corner; it continues selecting points $\{v_3, v_4,....\}$ not only from the periphery but also from all domain of the data sample, the final subset being quite uniformly distributed. Although, this algorithm is not a clustering technique, each data has been considered to be represented by the closest vector of the selected subset and therefore they are shown in the same color.

2.4. Graphical comparison between algorithms

The three algorithms considered have been applied to a data sample located in the space defined by a circle with a diameter equal to one. ~~The~~ distribution of the KMA centroids (left panel), the SOM centroids (middle panel) and the MDA subset (right panel) are represented (blue points) over the data sample (red points). The effect of the topology preserving projection in the SOM algorithm can be observed in the distribution of the SOM centroids. KMA distributes the clusters over the data covering a large area, but there are none on the edge of the data domain. MDA begins by selecting one data on the edge of the data space and continues extending over the data domain until M vectors belong to the subset ~~(Fig. 5)~~.

The different density of information in the data space determines the random initialization of the KMA and SOM classifications. This initial distribution has a great influence on the final KMA centroids. In the SOM algorithm, the flexible lattice folds with more resolution onto the data areas with more density of information. The MDA subset is not influenced by a higher density in some regions of the data space. Another difference between the clustering and selection techniques is that the classification centroids are not vectors from the database. For the clustering algorithms, the KMA and SOM centroids are defined as an average of the corresponding data; however, in the selection algorithm, the MDA subset is formed by vectors from the database.

3. Data

In order to apply the considered algorithms to analyze trivariate wave climate at a specific location, the data used to define a typical wave climate is described. A wave reanalysis time series located in Galicia (NW Spain), see left panel of Fig. 6, is extracted from the SIMAR-44 database, developed by Puertos del Estado (Spain) using the WAM model and forced by 10-m winds from REMO model (Jacob and Podzun, 1997). The temporal coverage spans 44 years (1958–2001) with an hourly resolution and a spatial resolution of 1/12 degree. In this paper, the three main parameters: significant wave height (H_s), mean period T_{02} (T_m) and mean direction (θ_m) are used in the definition of each sea state. Therefore, the multivariate database is defined as: $\{H_{s,i}, T_{m,i}$, and $\theta_{m,i}\}$; $i = 1,\ldots,N$, where N is almost 400,000 sea states. In the right panel of Fig. 6, the empirical bivariate distribution of significant wave height and mean direction is shown. This directional distribution provides information about the direction of the most frequent sea states as well as the largest significant wave heights. Wave climate at this particular location is influenced by waves from sectors SW to NE, with the most energetic sea states from sectors W to NW.

4. Methodology to analyze the multidimensional wave climate

The three above-mentioned algorithms have been considered to analyze wave climate. The purpose of this section is to establish which technique is the most suitable to describe the multidimensional wave climate or to select the most representative subset of sea states. Sea states can be defined by different spectral scalar and directional parameters which imply data pre-process and transformations of the clustering and selection algorithms.

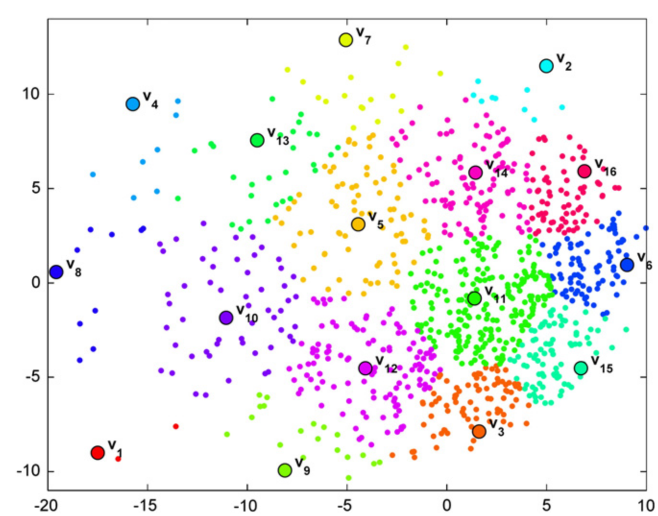

Fig. 4. Maximum dissimilarity selection. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article as: Camus, P., et al., Analysis of clustering and selection algorithms for the study of multivariate wave climate, Coast. Eng. (2011), doi:10.1016/j.coastaleng.2011.02.003

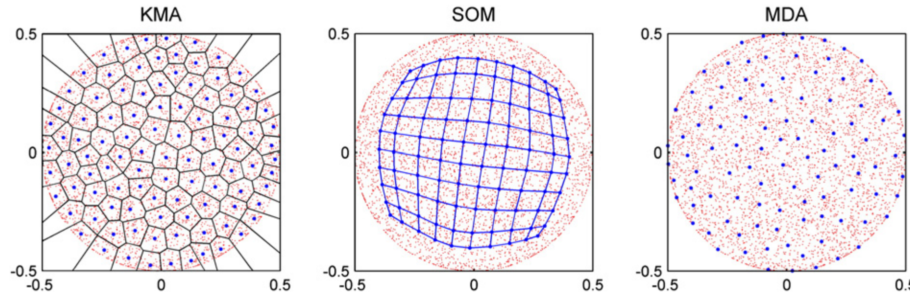

Fig. 5. Distribution of the classified or selected data in the circle domain. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The conditioning factors imposed by the wave data and the steps of the proposed methodology for the application of these techniques to analyze multidimensional wave climate are described below.

4.1. Conditioning factors imposed by the wave data

The input data is defined by the multivariate time series of the sea states defined in Section 3. The first two parameters (significant wave height, H_s, and mean period, T_m) are scalar variables, and the third one (mean direction, θ_m) is a circular variable.

The criterion of similarity implemented in the three considered algorithms is defined by the Euclidian distance. The wave direction θ_m is recorded on a continuous scale with 360° being identical to 0° while the Euclidian distance is adapted to an open linear scale. Note, that the circular variables entail a problem for the application of these techniques. For example, the directions N1°W (1° respect to the North) and N1°E (359° respect to the North) are supposed to be completed differently (differences of 358° with the Euclidian distance when the real distance is 2°). The problem is solved by implementing the distance in the circle for the directional variables. Therefore, a Euclidian-circular distance has been introduced into the clustering and selection algorithms, namely EC distance ('E' for the Euclidian distance in scalar parameters and 'C' for the circular distance in directional parameters). Besides, the vector components are normalized in order to be similarly weighted in the EC distance calculation.

Another conditioning factor is the redundancy of the average wave climate conditions defined in the reanalysis database. The clustering centroids depend on the distribution of the data to be classified, with more groups in those areas with higher density of information. In the SOM case, the neighborhood function produces a higher effect. A representative sample of all the sea states of reanalysis data base must be selected, trying to cover the range of the variable values without repeated data. In the case of KMA, a pre-selection avoids a conditioned initialization of the clusters in the data area with an excessive density of information.

The pre-selection is not necessary in the MDA application because the subset is selected independently to that of the different density of information in the data space. Besides, the version developed by Polinsky et al. (1996) is capable of working with high amounts of data without an excessive computational effort.

Therefore, the methodology has been divided into several steps. In the case of KMA and SOM, these are as follows: a) preselection of the input data; b) normalization of the variables which define the sea states; c) application of the clustering algorithm with the EC distance implemented; and d) denormalization of the clusters obtained. In the case of MDA the steps are: a) normalization; b) application of the algorithm with EC distance implemented; and c) denormalization of the subset. An explanatory sketch of the methodology is shown in Fig. 7 and is explained below.

4.2. Steps of the methodology

The pre-selection step consists of a "cube sampling" scheme: from the empirical 3-D histogram (composed of small cubic classes), we select only one data per class. The resolution of the equispaced division in all dimensions of data space has to assure that the centroids with its corresponding probably enable reproduce the mean values and the extreme values of different sea states parameters (e.g. H_s, and θ_{FE}). In the example, the H_s, T_m and θ_m dimensions are divided in 50 segments, obtaining a sample of 10,000 data. The input data,

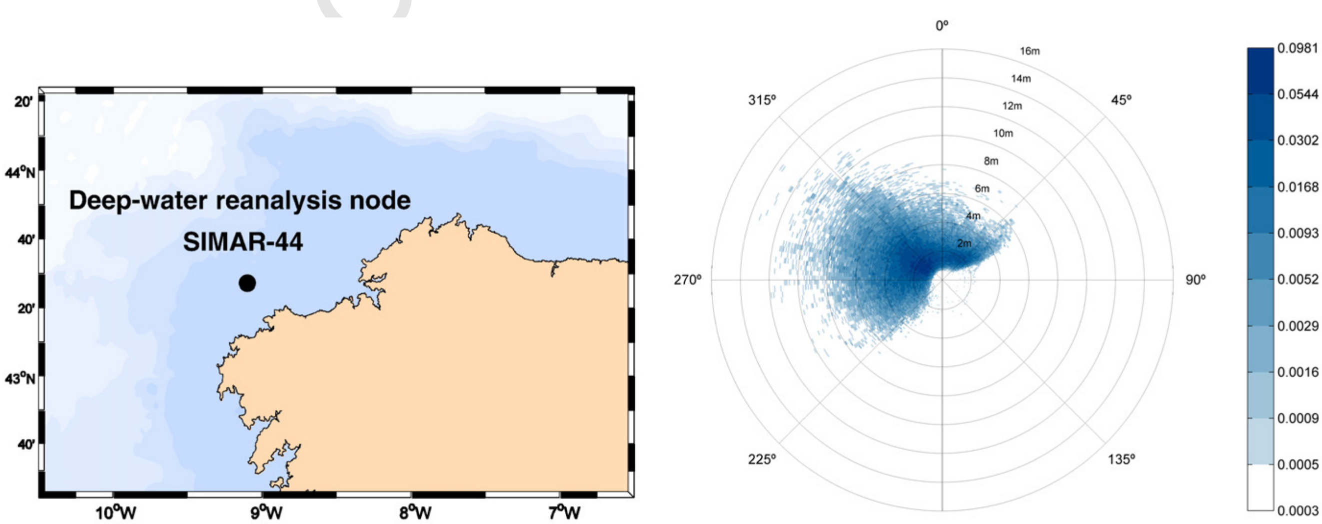

Fig. 6. Localization, near Villano deep-water buoy, Galicia, NW Spain (left panel). Empirical joint distribution of H_s and θ_m (right panel).

Please cite this article as: Camus, P., et al., Analysis of clustering and selection algorithms for the study of multivariate wave climate, Coast. Eng. (2011), doi:10.1016/j.coastaleng.2011.02.003

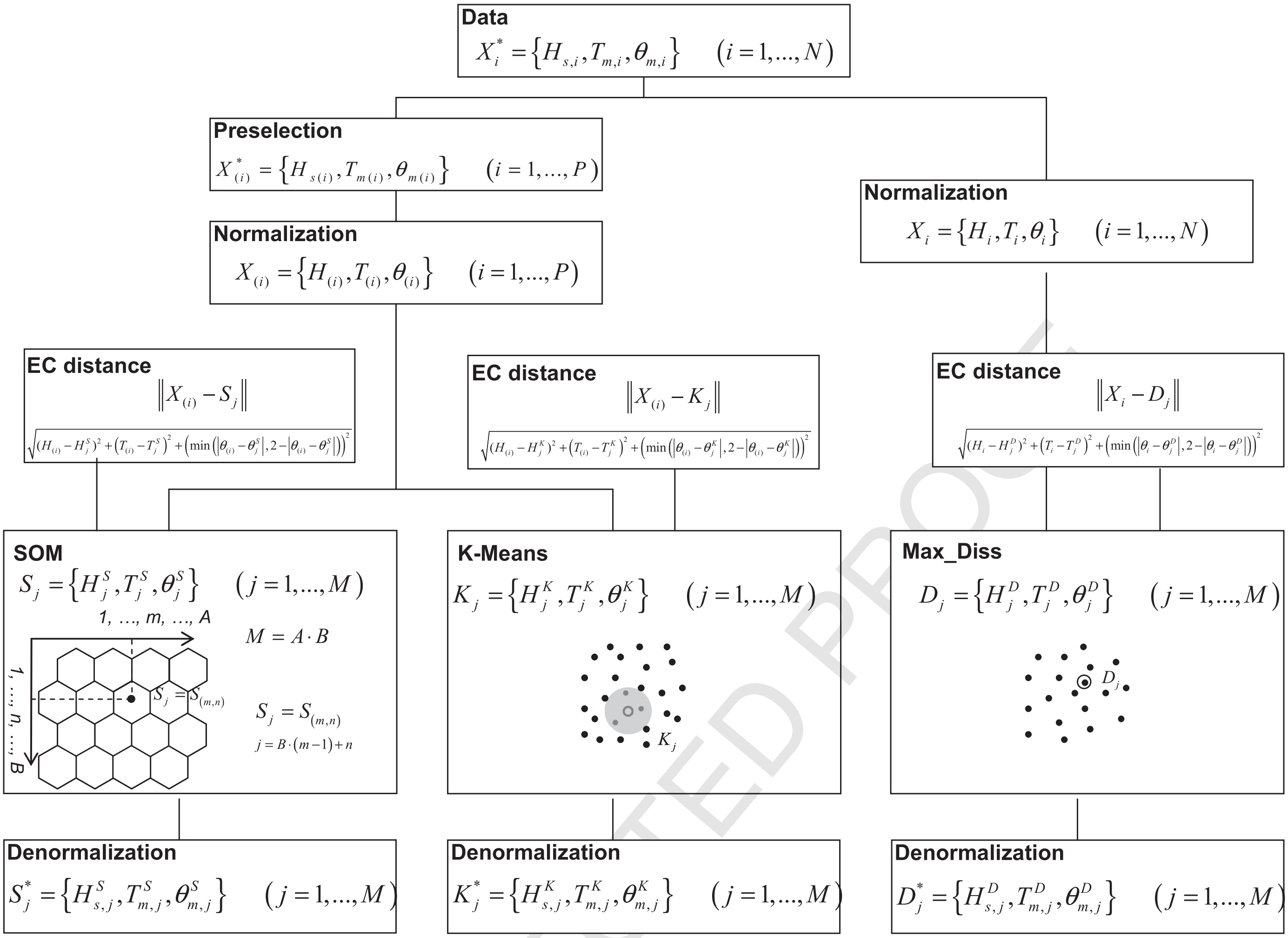

Fig. 7. Methodology to analyze the multidimensional wave climate.

composed of N tridimensional-vectors, $X_i^* = \{H_{s,i}, T_{m,i}, \theta_{m,i}\}; i = 1,..,N$, is reduced to a set of P vectors $X_{(i)}^* = \{H_{s(i)}, T_{m(i)}, \theta_{m(i)}\}; i = 1, \ldots, P$.

The scalar variables are normalized by scaling the variables values between [0,1] with a simple linear transformation, which requires two parameters, the minimum and maximum value of the two scalar variables.

$$\begin{aligned} H_s^{min} &= min(H_s); \quad H_s^{max} = max(H_s) \\ T_m^{min} &= min(T_m); \quad T_m^{max} = max(T_m). \end{aligned} \tag{7}$$

For the circular variables (defined in radians or in sexagesimal degrees using the scaling factor π/180), taking into account that the maximum difference between two directions over the circle is equal to π and the minimum difference is equal to 0, this variable has been normalized by dividing the direction values between π, therefore rescaling the circular distance between [0,1].

After these transformations, the dimensionless input data $X = \{H, T, \theta\}$ are defined as:

$$H = \frac{H_s - H_s^{min}}{H_s^{max} - H_s^{min}}; T = \frac{T_m - T_m^{min}}{T_m^{max} - T_m^{min}}; \theta = \frac{\theta_m}{\pi}. \tag{8}$$

The clusters obtained by the KMA technique are defined as $K_j = \{H_j^K, T_j^K, \theta_j^K\}; j = 1, \ldots, M$, the centroids obtained by SOM $S_j = \{H_j^S, T_j^S, \theta_j^S\}; j = 1, \ldots, M$, while the subset obtained by the MDA are $D_j = \{H_j^D, T_j^D, \theta_j^D\}; j = 1, \ldots, M$, where M is the number of centroids.

The EC distance in the KMA, SOM and MDA, presents the following expressions:

$$||X_{(i)} - K_j|| = \sqrt{\left(H_{(i)} - H_j^K\right)^2 + \left(T_{(i)} - T_j^K\right)^2 + \left(\min\left(|\theta_{(i)} - \theta_j^K|, 2 - |\theta_{(i)} - \theta_j^K|\right)\right)^2} \tag{9}$$

$$||X_{(i)} - S_j|| = \sqrt{\left(H_{(i)} - H_j^S\right)^2 + \left(T_{(i)} - T_j^S\right)^2 + \left(\min\left(|\theta_{(i)} - \theta_j^S|, 2 - |\theta_{(i)} - \theta_j^S|\right)\right)^2} \tag{10}$$

$$||X_i - D_j|| = \sqrt{\left(H_i - H_j^D\right)^2 + \left(T_i - T_j^D\right)^2 + \left(\min\left(|\theta_i - \theta_j^D|, 2 - |\theta_i - \theta_j^D|\right)\right)^2}. \tag{11}$$

Finally, the last step is the denormalization of clusters, applying the opposite transformation of the normalization step:

$$H_{s,j}^S = H_j^S \cdot \left(H_s^{max} - H_s^{min}\right) + H_s^{min}; T_{m,j}^S = T_j^S \cdot \left(T_m^{max} - T_m^{min}\right) + T_m^{min}; \theta_{m,j}^S = \theta_j^S \cdot \pi \tag{12}$$

$$H_{s,j}^K = H_j^K \cdot \left(H_s^{max} - H_s^{min}\right) + H_s^{min}; T_{m,j}^K = T_j^K \cdot \left(T_m^{max} - T_m^{min}\right) + T_m^{min}; \theta_{m,j}^K = \theta_j^K \cdot \pi \tag{13}$$

$$H_{s,j}^D = H_j^D \cdot \left(H_s^{max} - H_s^{min}\right) + H_s^{min}; T_{m,j}^D = T_j^D \cdot \left(T_m^{max} - T_m^{min}\right) + T_m^{min}; \theta_{m,j}^D = \theta_j^D \cdot \pi. \tag{14}$$

5. Results

The proposed methodology has been applied to analyze the multidimensional wave climate at the location in Galicia, in NW Spain (shown in Fig. 6). In this section, we describe the centroids obtained by KMA, SOM and MDA and we analyze the cluster variance within and the representativeness of centroids.

5.1. Description of classifications and selection

The original data and the results of the three algorithms are shown in Fig. 8 with a 3D representation in the upper panel and different 2D projections in the rest of the panels. In the upper panel, the pre-selected data (gray points), the $M = 529$ centroids (black points), six selected centroids (black circles) and the corresponding data which define the clusters (in different colors) are shown. The KMA centroids (in the left upper panel) are expanded over the input data space, with some centroids in areas with little information. These are areas with the largest significant wave heights or southern sea states. In the case of the SOM algorithm, most of the centroids (in the middle upper panel) are located in the area with more density of information, and no clusters are found around the data edges due to the topological restrictions. The MDA subset (in the right upper panel) is distributed over the data space, even at the edges.

The 2D projections of the six selected clusters (cyan, magenta, green, red, yellow, and blue) allow us to analyze the differences

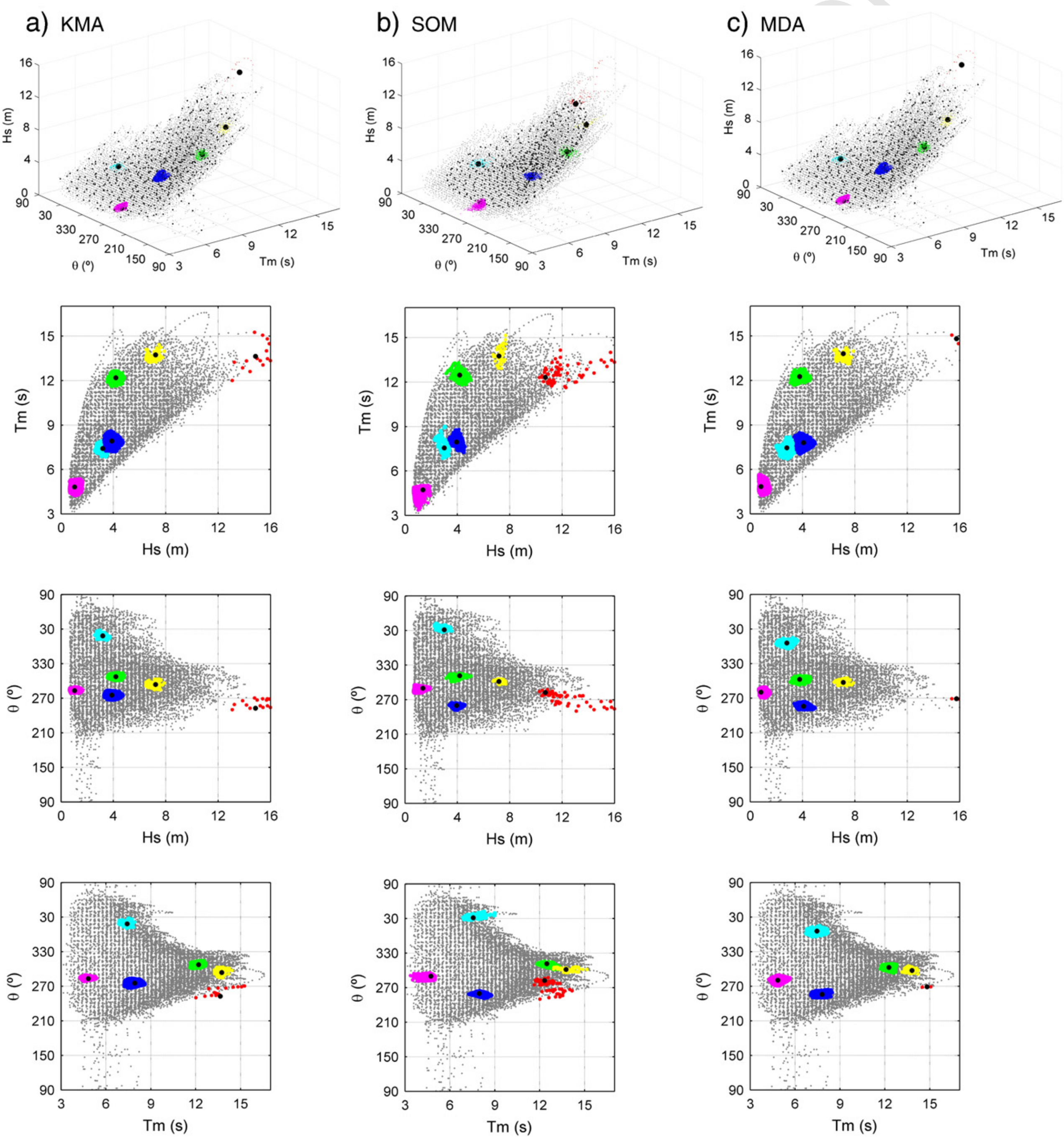

Fig. 8. Pre-selected wave climate data and centroids obtained by KMA (a), SOM (b) and MDA (c). Distribution of the six selected groups obtained by three algorithms. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

between the three algorithms in more detail. The centroids (in cyan, green, yellow and blue), which represent data located in the area with higher density of information, are similarly classified by the three techniques. However, SOM does not classify as well as it does the others the cluster in red, which represents the wave data with the largest significant wave height. The SOM centroids are not able to expand over the whole data space. The SOM clusters located on the edges are made up of a larger range of data variables. In the case of the red MDA centroid, the amount of data represented by this vector is smaller than that of the rest of the algorithms, and the variance of the variable values are smaller than the corresponding KMA centroid.

An important property of the SOM algorithm is that it projects the topological relationships of the high-dimensional data space onto a lattice, providing an easy visualization of the classification. A hexagonal SOM of 23 × 23 $\{H_s, T_m, \text{and } \theta_m\}$ clusters is shown in Fig. 9. The significant wave height H_s, the wave period T_m and the mean wave direction θ_m are represented by the size, the gray color intensity and the direction of the arrow, respectively. The smaller hexagon, in a light yellow-dark red scale, defines the H_s magnitude. The background of each hexagon has been filled in shades of blue, showing the relative frequency. The input data has been projected into a toroidal lattice which means that the centroids located on the upper, lower and in lateral sides of the sheet are joined in the toroidal projection, being similar in the data space.

As seen, this technique is capable of detecting all the possible sea states, similar clusters are located together in the projection space, and the magnitudes of the parameters which define the centroids vary smoothly from one cell to another. The value of the H_s varies from 1.22 m to 10.8 m, T_m has a minimum value of 4.66 s and a maximum value of 13.8 s, and θ_m varies from 220° (SSW) to 45° (NE).

The clusters with the largest significant wave heights, with a range of values between 9.01 m and 10.83 m, centered around the cluster $S^*_{(18,15)} = S^*_{406}$ (= 10.83 m), show high periods (values between 11.07 s and 13.26 s) and western directions (273.6°–310.6°).

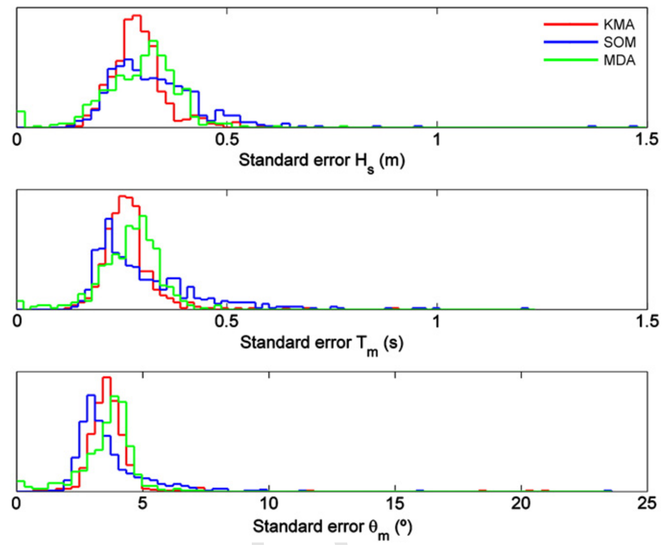

Fig. 10. Standard errors of H_s, T_m and θ_m of the corresponding data to each cluster obtained by the KMA, SOM and MDA.

The centroids with the largest period values, centered at the clusters $S^*_{(21,12)} = S^*_{472}$, $S^*_{(21,13)} = S^*_{473}$, $S^*_{(22,12)} = S^*_{495}$ and $S^*_{(22,13)} = S^*_{496}$, with periods around 13.7 s, present wave heights between 5.83 m and 9.14 m with corresponding directions around W-NW (293.15°–315.9°).

The clusters with directions from the first quadrant are located in the corners of the SOM map. These clusters present low-average significant wave heights and periods (range values between 1.35 m–6.19 m and 4.77 s–9.27 s), with a predominance of low energetic sea states.

Regarding the frequency (represented in a log-scale), we can distinguish areas with very frequent sea states (around $S^*_{(8,3)}$, and

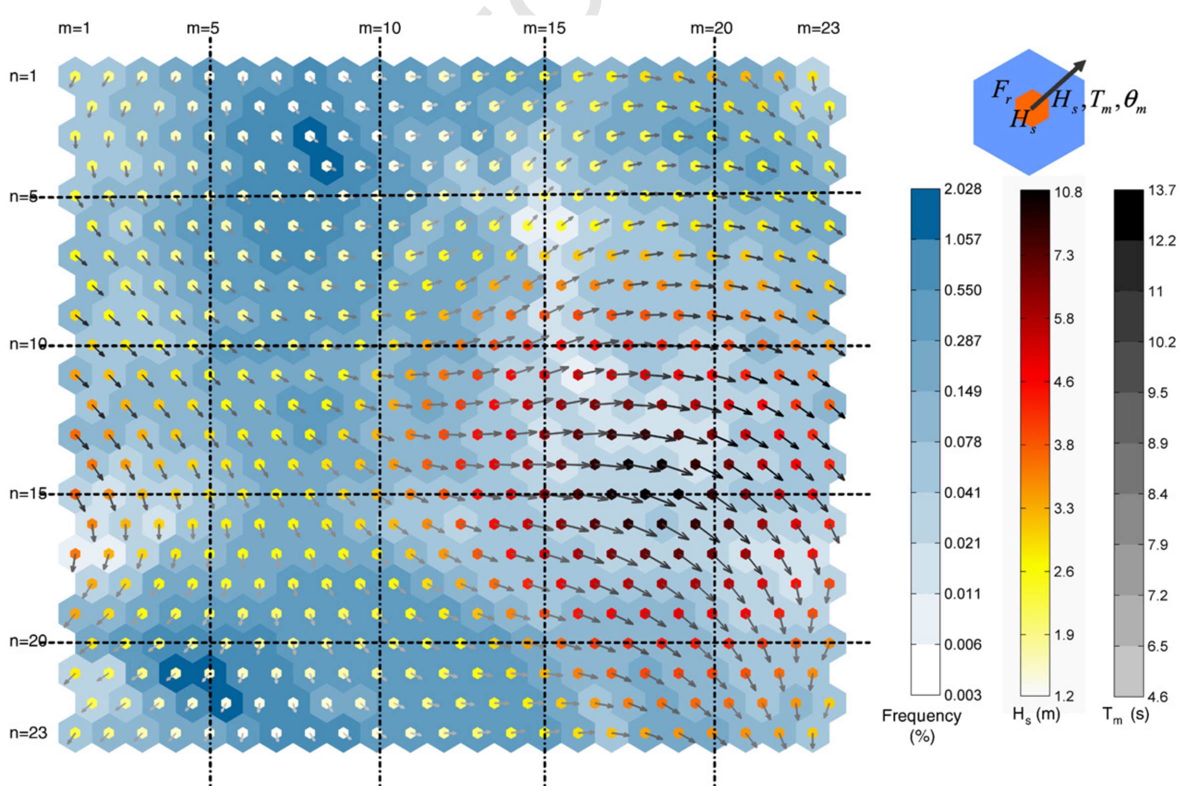

Fig. 9. SOM of size 23 × 23, corresponding to the $\{H_s, T_m, \text{and } \theta_m\}$ time series of a reanalysis database in Galicia (NW Spain). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article as: Camus, P., et al., Analysis of clustering and selection algorithms for the study of multivariate wave climate, Coast. Eng. (2011), doi:10.1016/j.coastaleng.2011.02.003

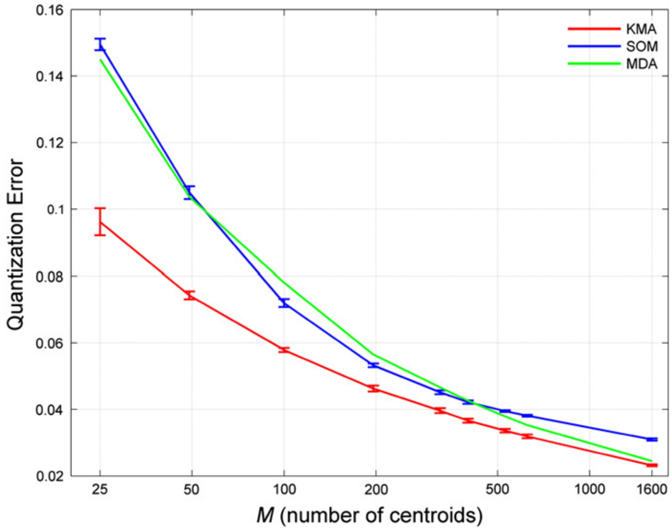

Fig. 11. Mean quantization errors of every algorithm for a different number of centroids. Standard errors for the 10 KMA and SOM trainings are also presented.

$S^*_{(4,21)}$) but also very rare sea states ($S^*_{(15,6)}$ and $S^*_{(16,11)}$) that help us to fully visualize all the possible 3D sea states at a particular location. Besides, the probability density function on the lattice allows us to consider the SOM as a multidimensional histogram, providing an interesting option to aggregate coastal engineering parameters such as mean energy flux, littoral sediment transport, port operability, etc.

5.2. Performance of the algorithms

We analyze how these techniques are able to describe wave climate through a reduced number of sea states. Nine different classification sizes have been considered (25, 49, 100, 196, 324, 400, 529, 625, and 1600) with 10 random initializations in the case of the KMA and SOM techniques, and only 1 for the MDA deterministic algorithm.

The standard errors between the corresponding data of each cluster and its centroid, for the three variables considered in the sea state definition of the KMA and SOM classifications and the MDA selection of size $M = 529$, are represented in Fig. 10. Although the KMA and SOM algorithms are applied to the pre-selected reanalysis data, the centroid corresponding to each reanalysis data is calculated and the variance and the frequency of each cluster are obtained considering the complete data time series. In the case of the SOM classification, the mean standard errors are 0.33 m, 0.31 s, and 3.7° for the variables H_s, T_m, and θ_m, respectively. In the case of the KMA classification, these mean values are 0.29 m, 0.27 s and 3.74°. For the MDA subset, the mean standard errors are 0.29 m, 0.27 s and 3.56°.

The quantization error is defined as the average distance between each vector and its corresponding centroid, and represents a measure of the SOM resolution (data far away in the high-dimensional space are close in the projection lattice). In Fig. 11, the quantization error for KMA, SOM and MDA algorithms are shown. The random initialization has no influence on the results. The best results are always obtained with KMA; for a number of centroids lower than 200 centroids, the differences in the errors between the algorithms are greater; while for sizes higher than 200 centroids, these differences are reduced, and in the case of MDA, the results tend to be similar to KMA errors.

The 90 percentile (H_{s90}) and the 99 percentile (H_{s99}) of the significant wave height statistical distribution and the mean energy flux direction (θ_{FE}) are considered to analyze the representativeness of the clusters or subset obtained to describe wave climate. We have determined the error between the real value, calculated by the complete reanalysis time series (Eq. (15)), and the estimated value, calculated using the clustering centroids or selection centroids and their frequency of occurrence (Eq. (16)). In Fig. 12, the errors (ΔH_{s90}, ΔH_{s99}, $\varepsilon_{FE} = \theta_{FE} - \theta^*_{FE}$) are shown for each size of the classification and selection considered. The exact, θ_{FE}, and the approximate, θ^*_{FE}, definitions of the mean energy flux direction are defined as:

$$\theta_{FE} = \tan^{-1}\left(\frac{\sum_{i=1}^{N} H_{s,i}^2 T_{m,i} \sin\theta_{m,i}}{\sum_{i=1}^{N} H_{s,i}^2 T_{m,i} \cos\theta_{m,i}}\right) \quad (15)$$

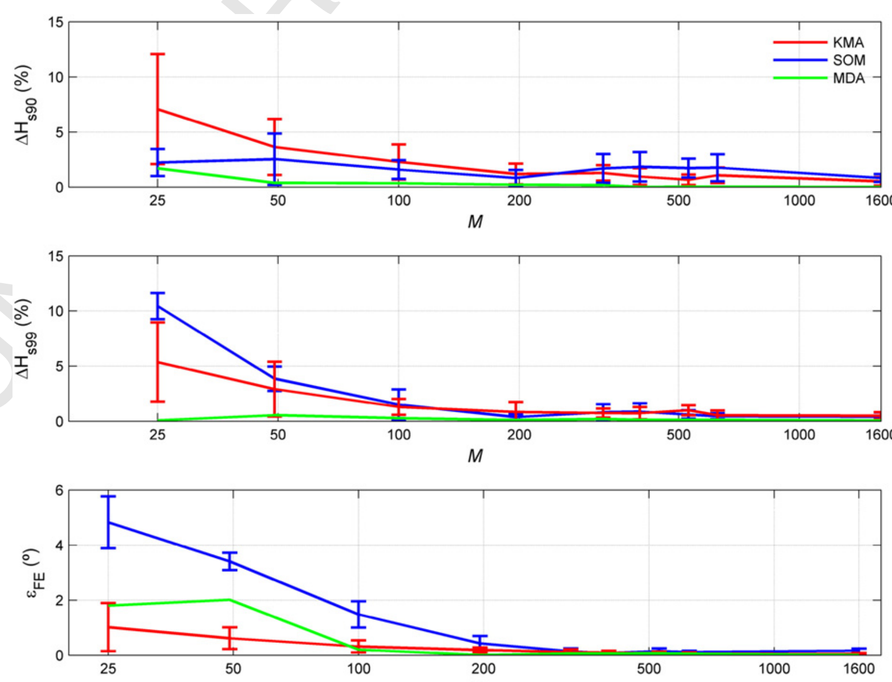

Fig. 12. Errors of H_{s90} (%), H_{s99} (%) and θ_{FE} (°).

Please cite this article as: Camus, P., et al., Analysis of clustering and selection algorithms for the study of multivariate wave climate, Coast. Eng. (2011), doi:10.1016/j.coastaleng.2011.02.003

Table 1
Goodness of the studied algorithms for different proposes.

	Visualization	Statistical description	Propagation
SOM	***	**	*
KMA	–	***	*
MDA	–	**	***
Achieved objectives	Multivariate histogram (SOM)	Correct definition of average wave climate (KMA, SOM, and MDA)	Ability of finding uncommon sea states (MDA)
	Visualization in the 2D lattice of parameters derived from $\{H_s, T_m, \theta_m\}$ (SOM)	Useful for defining port operability, longshore drift, (KMA, SOM, and MDA)	Good performance defining the boundaries of the data space (MDA)
			Best option for a propagation strategy including an interpolation scheme (MDA)

$$\theta^*_{FE} = tan^{-1}\left(\frac{\sum_{j=1}^{M} p_j H^2_{s,j} T_{m,j}\, sin\theta_{m,j}}{\sum_{j=1}^{M} p_j H^2_{s,j} T_{m,j}\, cos\theta_{m,j}}\right) \quad (16)$$

where p_j is the probability associated to the jth centroid.

In the case of the MDA selections, the errors ΔH_{s90} and ΔH_{s99} are almost zero for every size considered. In case of the SOM and KMA classifications, the error decreases when the number of clusters increases, with values close to zero for $M>200$. The smallest errors ε_{FE} ($\leq 1°$) are obtained by the KMA algorithm for sizes $M<100$; while for a number of clusters $M \geq 200$, the errors are closer to zero when using the KMA and MDA. For the SOM, the errors are around 5°–6° for a size of $M=25$; they decrease to values close to zero for M>200.

Summing up, these algorithms are able to extract the main features of the population data, each one showing different abilities for solving several coastal engineering problems: the SOM is the best algorithm to visualize multivariate data, the KMA is adequate to synthesize the most representative sea states to define the average wave climate, and the MDA is the algorithm that is able to explore the boundaries of the data space, suggesting that it the best option to define a wave propagation strategy.

6. Conclusions

The KMA and the SOM clustering techniques and the MDA selection algorithm have been applied to analyze the multivariate wave climate. The conditioning factors imposed by the wave database characteristics imply several modifications and processes thereby determining a methodology to analyze the multidimensional wave climate. This methodology has been applied to describe the wave climate defined by three spectral parameters (significant wave height, mean period and mean direction).

The projection of the SOM classification of multidimensional data on a lattice provides an excellent support to analyze the wave climate and to visualize a multidimensional histogram on the lattice. The SOM is the best technique to graphically characterize the multidimensional wave climate. The projection of the classification in a two-dimensional space with spatial organization allows the visualization of patterns with high dimensionality and simplifies the analysis of the multidimensional information.

The quantization error has proved that the best representation of the average wave conditions is obtained by the KMA classification. This algorithm can be adequate to study, for instance, port operability or longshore drift which require the most representative catalog of wave conditions without being interesting in the extreme situations.

The MDA algorithm is suitable for an automatic selection of a subset of sea states representative of wave climate in deep water in a methodology to transfer the wave climate to coastal areas (Camus et al, 2010).

Regarding the initial objectives of this work, the conclusions about the analysis of trivariate wave climate using the KMA, SOM and MDA algorithms are summarized in Table 1 (the number of asterisks indicates the goodness of the algorithm).

This work focuses on three parameters (H_s, T_m, and θ_m) and further research is needed to apply the algorithms to more complex problems taking into account, for instance, wind velocity and direction, sea and swell components of the sea states, storm surge level, or even the spatial variability of the met-ocean parameters.

Acknowledgments

The work was partially funded by projects “GRACCIE” (CSD2007-00067, CONSOLIDER-INGENIO 2010) from the Spanish Ministry MICIN, “MARUCA” from the Spanish Ministry MF and “C3E” from the Spanish Ministry MAMRM. The authors thank ~~the~~ Puertos del Estado (Spanish Ministry MF) for the use of the reanalysis data base.

References

Abadie, S., Butel, R., Mauriet, S., Morichon, D., Dupuis, H., 2006. Wave climate and longshore drift on the South Aquitaine coast. Continental Shelf Research 26, 1924–1939.
Athanassoulis, G.A., Belibassakis, K.A., 2002. Probabilistic description of metocean parameters by means of kernel density models 1. Theoretical background and first results. Applied Ocean Research 24, 1–20.
Camus, P., Mendez, F.J., Izaguirre, C., Reguero, B.G., Medina, R., 2010. Statistical and dynamical downscaling to transfer wave climate to coastal areas. Geophysical Research Abstracts 12, 12590 EGU.
Cavazos, T., 1997. Downscaling large-scale circulation to local winter rainfall in north-eastern Mexico. Int. J. Climatol. 17, 1069–1082.
Dodet, G., Bertin, X., Taborda, R., 2010. Wave climate variability in the North-East Atlantic Ocean over the last six decades. Ocean modelling 31, 120–131.
Groeneweg, J., van Ledden, M., Zijlema, M., 2007. Wave transformation in front of the Dutch Coast. In: Jane McKee, Smith (Ed.), Proc. of International Conference Coastal Engineering: ASCE, pp. 552–564.
Gutiérrez, J.M., Cofiño, A.S., Cano, R., Rodríguez, M.A., 2004. Clustering methods for statistical downscaling in short-range weather forecast. Monthly Weather Review 132, 2169–2183.
Gutiérrez, J.M., Cano, R., Cofiño, A.S., Sordo, C., 2005. Analysis and downscaling multi-model seasonal forecasts in Peru using self-organizing maps. Tellus 57A, 435–447.
Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, New York.
Holthuijsen, L.H., 2007. Waves in Oceanic and Coastal Waters. Cambridge University Press.
Jacob, D., Podzun, R., 1997. (1997) Sensitivity studies with the regional climate model REMO. Meteorology and Atmospheric Physics 63, 119–129.
Kennard, R.W., Stone, L.A., 1969. Computer aided design experiments. Technometrics 11, 137–148.
Kohonen, T., 2000. Self-organizing Maps, 3rd ed. Springer-Verlag, Berlin.
Lin, G.-F., Chen, L.-H., 2005. Identification of homogeneous regions for regional frequency analysis using the self-organizing map. Journal of Hydrology 324, 1–9.
Liu, Y., Weisberg, R.H., 2005. Patterns of ocean current variability on the West Florida Shelf using the self-organizing map. Journal of Geophysical Research 110, C06003.
Pilar, P., Guedes Soares, C., Carretero, J.C., 2008. 44-year wave hindcast for the North East Atlantic European coast. Coastal Engineering ~~doi:10.1016/j.coastaleng.2008.02.027~~.
Polinsky, A., Feinstein, R.D., Shi, S., Kuki, A., 1996. Librain: software for automated design of exploratory and targeted combinatorial libraries. In: Chaiken, I.M., Janda, K.D. (Eds.), Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery. American Chemical Society, Washington, D.C., pp. 219–232.
Ratsimandresy, A.W., Sotillo, M.G., Carretero Albiach, J.C., Álvarez Fanjul, E., Hajji, H., 2008. A 44-year high-resolution ocean and atmospheric hindcast for the Mediterranean Basin developed within the HIPOCAS Project. Coastal Engineering 55, 827–842.
Solidoro, C., Bandelj, V., Barbieri, P., Cossarini, G., Umani, S.F., 2007. Understanding dynamic of biogeochemical properties in the northern Adriatic Sea by using self-organizing maps and K-means clustering. Journal of Geophysical Research 112, C07S90.
Stansby, P., Zhou, J., Kuang, C., Walkden, M., Hall, J., Dickson, M., ~~2006~~. Long-term prediction of nearshore wave climate with an application to cliff erosion. In: Jane McKee, Smith (Ed.), Proc. of International Conference Coastal Engineering: ASCE, pp. 616–627.
Weisse, R., Feser, F., Günther, H., 2002. A 40-year high-resolution wind and wave hindcast for the Southern North Sea. Proceedings of the 7th International Workshop on Wave Hindcasting and Forecasting. Banff, Alberta, Canada, pp. 97–104.
Willet, P., 1996. Molecular diversity techniques for chemical databases. Information Research 2 (No. 3).

Please cite this article as: Camus, P., et al., Analysis of clustering and selection algorithms for the study of multivariate wave climate, Coast. Eng. (2011), doi:10.1016/j.coastaleng.2011.02.003