Journal of Data Science **9**(2011), 585-605

A Latent-Class Model for Clustering Incomplete Linear and Circular Data in Marine Studies

Francesco Lagona* and Marco Picone
Roma Tre University

Abstract: Identification of representative regimes of wave height and direction under different wind conditions is complicated by issues that relate to the specification of the joint distribution of variables that are defined on linear and circular supports and the occurrence of missing values. We take a latent-class approach and jointly model wave and wind data by a finite mixture of conditionally independent Gamma and von Mises distributions. Maximum-likelihood estimates of parameters are obtained by exploiting a suitable EM algorithm that allows for missing data. The proposed model is validated on hourly marine data obtained from a buoy and two tide gauges in the Adriatic Sea.

Key words: Circular data, cross-validation, EM algorithm, Gamma distribution, latent classes, marine data, missing values, Von Mises distribution.

1. Introduction

Wave regimes are specific shapes that the distribution of wave attributes (such as wave height and direction) takes under latent environmental conditions. The identification of relevant regimes in a particular area is often necessary to estimate the drift of floating objects and oil spills (Huang *et al*, 2011), in the design of offshore structures (Faltinsen, 1990) and in studies of sediment transport (Jin and Ji, 2004) and coastal erosion (Pleskachevsky *et al.*, 2009). The description of wave data in terms of regimes is also useful in the analysis of coastal areas and enclosed seas, where numerical wind-wave models, traditionally used for ocean waves, can give inaccurate results (Bertotti and Cavalieri, 2004). For these reasons, the Assembly of the International Maritime Organization has repeatedly encouraged the publication of wave data atlas that include a description of representative wave regimes in specific areas, characterized by probability of occurrence, and corresponding to dominant environmental conditions (e.g., wind conditions) over

*Corresponding author.

the area of interest. This has motivated an increasing interest in methods for clustering wave data according to a finite number of regimes.

Traditionally, techniques of wave data clustering are based on distance-based methods. Recent proposals require the use of a finite number of target distributions, defined as cluster centroids, and an optimization algorithm that associates the observed data to the closest centroid (Boukhanovsky *et al.*, 2007). Hierarchical agglomerative clustering methods (Hamilton, 2010) have been also suggested to avoid the specification of a family of target distributions.

The limitations of distance-based based methods are well known (Fraley and Raftery, 2002). The statistical properties of these methods are generally unknown, precluding the possibility of formal inference on the clustering results. This is a critical issue in marine studies, because the identification of wave regimes without a measure of the statistical uncertainty of regime-specific parameters is of little practical use. In addition, there is little systematic guidance associated with distance-based methods for solving basic questions that arise in cluster analysis, such as the choice of an optimal number of clusters and the choice of an optimal clustering algorithm.

A general framework to address these issues is provided by latent-class models (Hagenaars and McCutcheon, 2002), which cluster multivariate data according to a finite number of classes, approximating the joint distribution of the data by a mixture of parametric densities, which represent the distributional shape of the data within each cluster. From a methodological viewpoint, a latent-class approach allows to solve the clustering problem as a missing value problem, by treating the unknown cluster membership of each observation as a missing value, to be estimated from the data. From a technical viewpoint, the clustering algorithm reduces to likelihood maximization and the choice of the optimal number of clusters reduces to a model selection problem in parametric inference.

In this paper we take a latent-class approach to describe sea conditions in terms of wave regimes, by clustering multivariate environmental profiles in a finite number of classes. Specifically, we model the data by a mixture of product densities, i.e. a particular latent-class model where the observed variables are assumed conditionally independent, given a latent multinomial variable. This model is tailored to identify wave regimes in practical settings that often arises in marine studies, where (1) environmental profiles include measurements taken on linear and circular supports and (2) some of these observations are missing, due to malfunctioning of the devices that provide the data.

While there is an extensive literature on modelling multivariate continuous, categorical and mixed continuous-categorical variables by multivariate normal models, log-linear models or a combination of both, the joint modelling of variables on linear and circular supports is still an open area of research. Recent

attempts include multivariate circular distributions defined on toroidal supports (Mardia *et al.*, 2008), distributions on cylinders that are based on nonnegative trigonometric sums (Fernández-Durán, 2007) and multivariate distributions with specified marginals on cylinders, discs and tori (Kato and Shimizu, 2008). When however the goal of an analysis is the identification of typical wave regimes, the specification of the joint distribution of marine variables should aim at clustering the data according to a finite number of classes in a way that the dependence structure between the data is well approximated by this partitioning of the sample. Mixtures of product densities provide such clustering of the data and flexibly accommodate for the mixed supports on which linear and circular data are taken. Moreover, the semi-parametric nature of the model allows for a parsimonious specification of the association structure between linear and circular measurements, which is of great help in marine studies, where too little is often known about the data generating process to assume a fully parametric specification.

Wave regimes identification is additionally complicated by the occurrence of missing values. Marine databases are often incomplete because of device malfunctioning or maintenance-related reasons. For mixture-based data clustering, maximum-likelihood estimation could be carried out by discarding incomplete data profiles from the sample and using the complete cases to build up the likelihood function to be maximized (CC; complete case analysis). If the joint distribution of the variables of interest is correctly specified and the data are missing at random (MAR; i.e., the conditional probability of not observing a value, given the observed data, does not depend on the unobserved value; Rubin, 1987), CC-based maximum-likelihood estimation is known to be (asymptotically) unbiased but inefficient (Rotnitzky and Wypij, 1994). Loss of efficiency is due to the fact that incomplete profiles are informative of the parameters of the joint distribution of several variables, especially when these variables are strongly correlated. Efficient maximum-likelihood estimation from MAR multivariate data often requires data-augmentation or multiple-imputation methods (Shafer, 1997). Mixture of product densities, instead, can be efficiently estimated by including both complete and incomplete profiles into the likelihood, because likelihood contributions of incomplete profiles are available in closed form and data-augmentation/imputation methods are not necessary.

Mixtures of product densities have been already suggested in the statistical literature to cluster multivariate categorical data (Vermunt *et al.*, 2008) and mixed linear and categorical data (Hunt and Jorgensen, 2003) in the presence of missing values. From a technical viewpoint, therefore, our application extends this strand of literature to the case of linear and circular data with missing values. On the methodological side, our proposal is an alternative to the existing distance-based methods for wave regime identification, with three practical advantages. First,

it is based on an EM algorithm that is less computationally demanding than the algorithms currently in use for distance-based identification of wave regimes. Second, missing values are efficiently handled, while distance-based methods normally require complete data information. Third, while formal inference is not possible with a distance-based approach to clustering, mixture-based clustering is carried out within a parametric inferential framework and, as a result, it can be validated by using traditional methods of parametric inference.

Relevant details on the data that motivated this work are presented in Section 2, while Section 3 is devoted to maximum-likelihood estimation of mixture of product densities in the case of missing observations. In Section 4 we specify the Gamma-von Mises latent-class model that was exploited to examine the data presented in Section 2. Estimation and model validation results are summarized in Section 5. Relevant points of discussion are listed in Section 6.

2. Data

The Italian Institute for Environmental Research and Protection (ISPRA; www.isprambiente.it) maintains a network of buoys to monitor wave direction and height at various points of the Italian seas. A network of ISPRA tide gauges, located along the Italian coast, additionally provide data about wind direction and speed.

The data that we have exploited in this work include hourly measurements of wave height and direction, taken in the period 11/18/2002-01/17/2003 by the buoy of Ancona, which is located in the Adriatic sea at about 30 Km from the coast (Figure 1). During the same period, hourly data on wind speed and direction were obtained from the two nearest tide gauges, respectively located at Ancona (about 30 Km from the buoy) and at Ravenna (about 120 Km from the buoy). To account for the cumulative effect that wind has on waves, wind data were smoothed by taking, for each hour, the average of wind speeds and the circular average of wind directions, observed during the last eight hours.

Table 1 reports the percentages of missing data observed during the study period. Measurements taken by buoys and tide gauges can be missing because of devices maintenance or discontinuous functioning. Occurrence of missing values on wave measurements is more frequent than the occurrence of missing wind data because buoys are more exposed to transmission errors than tide gauges. We remark that our data are in the form of hourly profiles of six observations. As a result, different patterns of missing values occur: while about the 28% of the data profiles include at least one missing value, the modal missingness pattern (15.3%) includes a missing circular and a missing linear variable. During the study period, there is a very small portion (about 0.1%) of hourly profiles with no information.

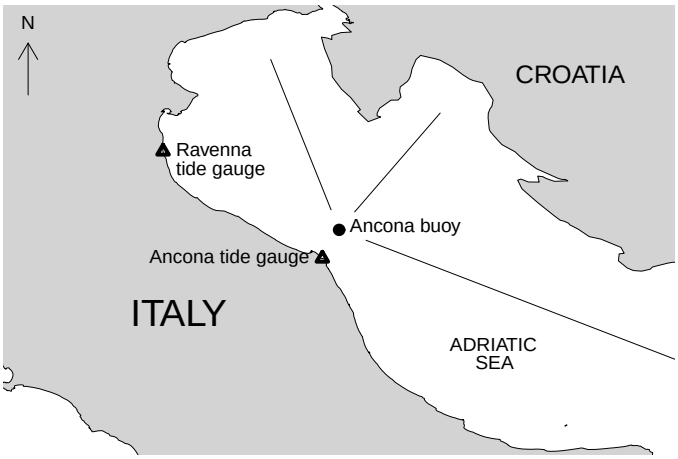


Figure 1: Locations of the buoy and the two tide gauges, from which the data displayed in Figure 2 were obtained; segments indicate the three directions of maximal fetch, i.e. the distance between the buoy and the closest coastline

Table 1: Percentages of missing values

Site	Measurement (Unit)	Percentages
Ancona (buoy)	Wave Height (Meters)	16.3%
	Wave Direction (Radians)	16.3%
Ancona (tide gauge)	Wind Speed (Meter/Sec)	1.1%
	Wind Direction (Radians)	2.2%
Ravenna (tide gauge)	Wind Speed(Meter/Sec)	10.4%
	Wind Direction (Radians)	2.3%

Univariate distributions of the available data are displayed in Figure 2. Rose diagrams indicate the distribution of directions from which the wind and the wave come from. As expected, waves mostly come from two modal directions (south-east and north-east), which relate to two of the three angles at which the distance between the buoy and the nearest coast (fetch) is maximum (Figure 1). Waves from North-West (along the third maximum-fetch direction) are rarely observed in wintertime, because winter winds do not typically blow from this direction. As displayed by the circular wind distributions at the two tide gauges of Ancona and Ravenna, two are the winds that dominate the Adriatic Sea in wintertime: bora, a typical cold wind, blowing from West/North-West, and Sirocco, blowing from South-East, and responsible for the storm surges in the northern part of the Adriatic sea, and hence for the famous floods of Venice.

The histograms on the right side of Figure 2 show the distributions of wave height, as observed at the buoy of Ancona, and wind speed, as measured at the

two tide gauges of Ancona and Ravenna. The multi-modal shape of these distributions is less apparent than that displayed by directional data. This is typical of wave and wind data that are observed in enclosed seas, such as the Adriatic, where the geometry of the coastline makes it difficult to separate components of dominant wind speeds and wave heights and is responsible for the inaccurate results provided by numerical wind-wave models that are normally used for modelling ocean waves.

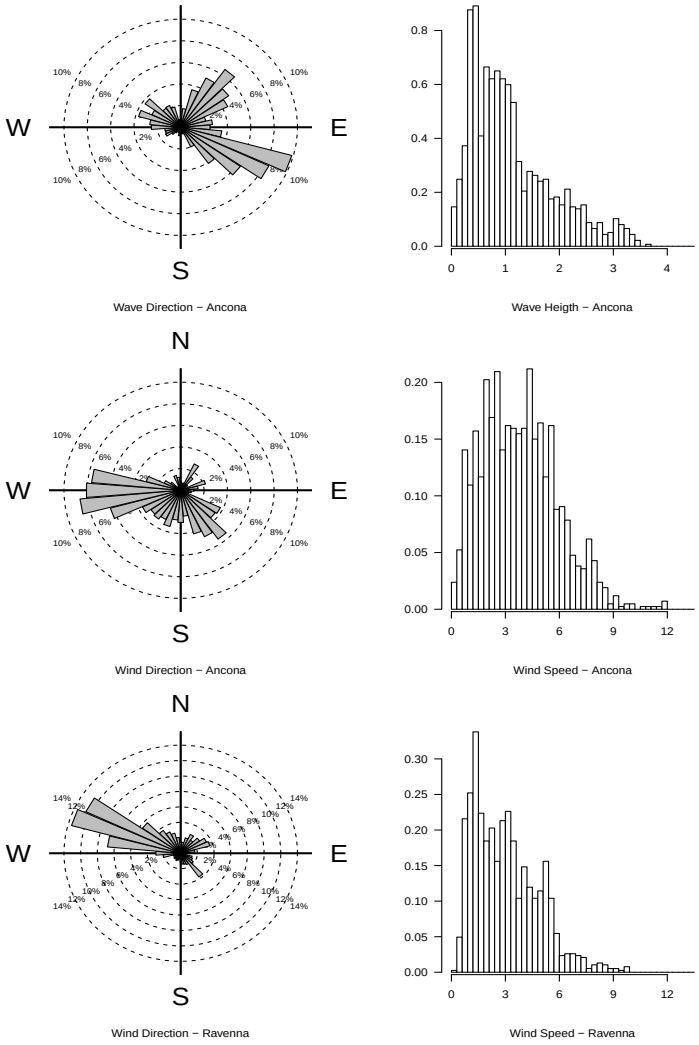


Figure 2: Distribution of the available wave metric data at the buoy of Ancona and wind data at the two nearest tide gauges (Ancona and Ravenna)

3. Estimation of Mixtures of Product Densities from Incomplete Mixed Data

The multivariate data described in Section 2 can be represented as n vectors $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iJ})$, $i = 1, \cdots, n$, drawn from the multivariate distribution of J variables $Y_j, j = 1, \cdots, J$, measured on different supports (e.g., linear or circular). We assume that these vectors can be clustered into K groups (or classes) and that the association structure between the variables Y_j is well approximated by this partitioning of the sample. Formally, we introduce a latent (unobserved) multinomial random vector $\boldsymbol{Z} = (Z_1, \cdots, Z_K)$ with one trial and cell probabilities $(\pi_1, \cdots, \pi_K)$, and assume that the J variables Y_j are conditionally independent given Z. Within this conditional independence assumption, we specify $K \times J$ distributions $f_k(y|\boldsymbol{\beta}_{kj})$, each known up to a parameter vector $\boldsymbol{\beta}_{kj}$, and model the multivariate distribution of vector $\boldsymbol{y}_i$ as a finite mixture of J-dimensional product densities, say

$$f(\boldsymbol{y}_i) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_{ij}|\boldsymbol{\beta}_{kj}), \tag{3.1}$$

where $f_k(y|\boldsymbol{\beta}_{kj})$ denotes the conditional distribution of Y_j within the kth latent class. We observe that (3.1) specifies a multivariate distribution without imposing consistency constraints on the conditional densities $f_k(y|\boldsymbol{\beta}_{kj})$, which, hence, do not necessarily need to be member of the same parametric family. This flexibility is of great help in the modelling of mixed linear and circular data. Given the number K of classes, mixtures of product densities are furthermore strictly identifiable, provided that the densities $f_k(\boldsymbol{y}) = \prod_{j=1}^{J} f_k(y_j|\boldsymbol{\beta}_{kj})$ are linearly independent (Teicher, 1967; Yakowitz and Spragins, 1968).

Mixture (3.1) is a particular latent-class model and is often presented in the literature as a model-based alternative to the traditional cluster-analysis methods that are based on distance-based procedures, such as hierarchical agglomerative clustering or iterative relocation procedures. Typically exploited in social science studies and marketing research, mixtures of product densities such as (3.1) have been successfully implemented in the classification of mixed profiles that include quantitative (continuous or discrete) and categorical (nominal or ordinal) observations.

Maximum-likelihood estimation of a mixture model is normally based on an EM algorithm. Hunt and Jorgensen (2003) developed an EM algorithm for estimating latent-class models from MAR data, in the case of mixed multi-normal and categorical data. In the case of mixtures of product densities, such as (3.1), their algorithm can be greatly simplified as follows.

We account for the occurrence of missing values by splitting the complete

data vector $\boldsymbol{y}_i = (\boldsymbol{y}_{O(i)}, \boldsymbol{y}_{M(i)})$ into a vector $\boldsymbol{y}_{O(i)}$ of observed data and a vector $\boldsymbol{y}_{M(i)}$ of missing values, $O(i) \cup M(i) = \{1, \cdots, J\}$. We furthermore introduce a $n \times J$ matrix R, whose generic component $r_{ij} = 1$ if y_{ij} is missing and 0 otherwise. Accordingly, the row-sums of R, say $r_{i\cdot} = \sum_{j=1}^{J} r_{ij}$, indicate the number of missing values within each ith profile.

If the data are MAR, i.e. the probability of a missing value does not depend on the value that is missing, maximum likelihood estimates of model (3.1) can be found by maximizing the marginal log-likelihood function

$$\begin{aligned} l(\boldsymbol{\beta}, \pi) &= \sum_{i=1}^{n} \log \int_{\boldsymbol{y}_{M(i)}} \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_{ij}|\boldsymbol{\beta}_{kj}) d\boldsymbol{y}_{M(i)} \\ &= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \left(f_k(y_{ij}|\boldsymbol{\beta}_{kj})\right)^{1-r_{ij}} \\ &= \sum_{i:r_{i\cdot}=0} \log \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_{ij}|\boldsymbol{\beta}_{kj}) + \sum_{i:r_{i\cdot}>0} \log \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \left(f_k(y_{ij}|\boldsymbol{\beta}_{kj})\right)^{1-r_{ij}} \\ &= l_{CC}(\boldsymbol{\beta}, \pi) + l_{IC}(\boldsymbol{\beta}, \pi), \end{aligned} \tag{3.2}$$

which is the sum of the log-likelihood contributions of the complete (CC) and incomplete cases (IC). We observe that the log-likelihood contribution of a completely missing profile, i.e. such that $r_{i\cdot} = J$, is given by $\log \sum_k \pi_k = 0$. Under a CC strategy, the log-likelihood contribution l_{IC} is ignored, leading to inefficient estimates.

Local maximum points of the log-likelihood (3.2) can be found by an EM algorithm (Dempster, Laird and Rubin, 1977) that iteratively maximizes the expectation of the complete data log-likelihood function. In the case of MAR data drawn from a mixture of product densities, the complete log-likelihood can be written as follows

$$l_{\text{comp}}(\boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left(\log \pi_k + \sum_{j=1}^{J} (1 - r_{ij}) \log f_k(y_{ij}|\boldsymbol{\beta}_{jk}) \right), \tag{3.3}$$

where $(z_{i1}, \cdots, z_{ik})$ is the ith realization of the multinomial random variable Z. At the hth step of the algorithm, the expectation of $l_{\text{comp}}(\boldsymbol{\beta}, \boldsymbol{\pi})$ with respect to the conditional distribution $p(Z|\boldsymbol{y})$ is computed on the basis of the estimates

$\hat{\boldsymbol{\beta}}^{(h-1)}$ and $\hat{\boldsymbol{\pi}}^{(h-1)}$, obtained at the previous iteration, by evaluating (E-step)

$$\begin{aligned}
&Q(\boldsymbol{\pi}, \boldsymbol{\beta}|\hat{\boldsymbol{\pi}}^{(h-1)}, \hat{\boldsymbol{\beta}}^{(h-1)}) = \mathbb{E}(l_{\text{comp}}(\boldsymbol{\beta}, \boldsymbol{\pi})) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\left(\log \pi_k + \sum_{j=1}^{J}(1-r_{ij})\log f_k(y_{ij}|\boldsymbol{\beta}_{jk})\right)\mathbb{E}(z_{ik}|\hat{\boldsymbol{\beta}}^{(h-1)}, \hat{\boldsymbol{\pi}}^{(h-1)}) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{\pi}_{ik}^{(h-1)}\log \pi_k + \sum_{j=1}^{J}\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}^{(h-1)}(1-r_{ij})\log f_k\left(y_{ij}|\boldsymbol{\beta}_{jk}\right) \\
&= Q\big(\boldsymbol{\pi}|\hat{\beta}^{(h-1)}, \hat{\pi}^{(h-1)}\big) + \sum_{j=1}^{J} Q_j\big(\boldsymbol{\beta}_{kj}|\hat{\beta}^{(h-1)}, \hat{\pi}^{(h-1)}\big),
\end{aligned} \tag{3.4}$$

where

$$\begin{aligned}
\hat{\pi}_{ik}^{(h-1)} &= \mathbb{E}(z_{ik}|\hat{\boldsymbol{\beta}}^{(h-1)}, \hat{\boldsymbol{\pi}}^{(h-1)}) \\
&= \frac{\int_{\boldsymbol{y}_{M(i)}} \hat{\pi}_k^{(h-1)} \prod_{j=1}^{J} f_k(y_{ij}|\hat{\boldsymbol{\beta}}_{kj}^{(h-1)}) d\boldsymbol{y}_{M(i)}}{\int_{\boldsymbol{y}_{M(i)}} \sum_{k=1}^{K} \hat{\pi}_k^{(h-1)} \prod_{j=1}^{J} f_k(y_{ij}|\hat{\boldsymbol{\beta}}_{kj}^{(h-1)}) d\boldsymbol{y}_{M(i)}} \\
&= \frac{\hat{\pi}_k^{(h-1)} \prod_{j=1}^{J} \left(f_k(y_{ij}|\hat{\boldsymbol{\beta}}_{kj}^{(h-1)})\right)^{1-r_{ij}}}{\sum_{k=1}^{K} \hat{\pi}_k^{(h-1)} \prod_{j=1}^{J} \left(f_k(y_{ij}|\hat{\boldsymbol{\beta}}_{kj}^{(h-1)})\right)^{1-r_{ij}}}
\end{aligned} \tag{3.5}$$

indicates the conditional probability of vector $\boldsymbol{y}_{O(i)}$ to belong to the kth latent class. The previous E-step is followed by an M-step where vector $(\hat{\boldsymbol{\beta}}^{(h-1)}, \hat{\boldsymbol{\pi}}^{(h-1)})$ is updated by a new vector $(\hat{\boldsymbol{\beta}}^{(h)}, \hat{\boldsymbol{\pi}}^{(h)})$ that maximizes the expected log-likelihood (3.4). We observe that (3.4) is the sum of $J+1$ functions, which depend on independent sets of parameters, and, as a result, the M-step can be carried out by separately solving $J+1$ maximization problems. In particular, the maximum point of $Q(\boldsymbol{\pi}|\hat{\beta}^{(h-1)}, \hat{\pi}^{(h-1)})$ is available in closed form and it is equal to

$$\hat{\pi}_k^{(h)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_{ik}^{(h-1)}.$$

The form of the updating equations for parameters $\boldsymbol{\beta}$ that maximize the remaining J functions $Q_j(\boldsymbol{\beta}_{kj}|\hat{\beta}^{(h-1)}, \hat{\pi}^{(h-1)})$ depend on the form of the densities $f_k(y_j|\boldsymbol{\beta}_{kj})$. In Section 4 we derive these updates under Gamma and von Mises densities.

The algorithm alternates the E-step and the M-step up to convergence of the estimates, whose limit (Wu, 1983) is a local maximum point of the likelihood function (3.2).

4. A Gamma-Von Mises Latent-Class Model

The $J = 6$ variables of our case study can be clustered in two groups according to the scale on which they are measured. A first group includes three circular variables, say Y_1 (wave direction at the Ancona buoy), Y_2 (wind direction at the Ancona tide gauge) and Y_3 (wind direction at the Ravenna tide gauge). A second group includes three variables on a linear support, say Y_4 (wave height at the Ancona buoy), Y_5 (wind speed at the Ancona tide gauge) and Y_6 (wind speed at the Ravenna tide gauge).

The mixture model presented in Section 3 allows for a flexible choice of the univariate distributions that can be placed within each latent class.

We have decided to model wave and wind directions by exploiting three von Mises distributions, i.e.

$$f_k(y|\boldsymbol{\beta}_{kj}) = \mathrm{VM}(\beta_{kj0}, \beta_{kj1}) = \frac{\exp(\beta_{kj1}\cos(y - \beta_{kj0}))}{2\pi I_0(\beta_{kj1})}, \quad j = 1, 2, 3, \qquad (4.1)$$

where the parameters β_{kj0} and β_{kj1}, $j = 1, 2, 3$, respectively indicate the mean (or modal) direction and the concentration of each conditional circular distribution, given the kth latent class, and I_0 is the modified Bessel function of order 0.

Wave height at the buoy and wind speeds at the two tide gauges have been instead modeled by three Gamma distributions, i.e.

$$f_k(y|\boldsymbol{\beta}_{kj}) = \mathrm{Gam}(\beta_{kj0}, \beta_{kj1}) = \frac{\beta_{kj0}^{\beta_{kj1}} y^{\beta_{kj1}-1} \exp -(y/\beta_{kj0})}{\Gamma(\beta_{kj1})}, \quad j = 4, 5, 6, \quad (4.2)$$

where parameters β_{kj0} and β_{kj1}, $j = 4, 5, 6$, respectively indicate the scale and shape of the conditional distributions, given the latent class.

Under the above distributional assumptions, the mixture of product densities

$$f(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_j|\boldsymbol{\beta}_{kj})$$

is a multivariate distribution on a six-dimensional hyper-cylinder. According to the sufficient conditions stated by Teicher (1967) and Yakowitz and Spragins (1968), identifiability of this mixture follows by the linear independence of the families of the Gamma and the von Mises densities. Moreover, the marginal distribution of each variable on a linear support is approximated by a mixture of K Gamma densities and the marginal distribution of each circular variable is approximated by a mixture of K von Mises densities. As a result, the J-dimensional profiles of wave and wind data ($J = 6$ in our application) are clustered according to K wind-wave regimes. Because von Mises and Gamma densities are known

up to 2 parameters, each regime is defined on the basis of $2J$ parameters, which indicate not only class-specific modal directions of waves and winds and their average heights and speeds, but also the amount of variation of the circular and linear measurements around these means. In particular, the association between each variable and the remaining variables is semi-parametrically described by conditional densities that take the following mixture form

$$f(y_j|y_l, l \neq j; \boldsymbol{\beta}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \frac{\pi_k \prod_{h \neq j} f_k(y_h|\boldsymbol{\beta}_{hk})}{\sum_{k=1}^{K} \pi_k \prod_{h \neq j} f_k(y_h|\boldsymbol{\beta}_{hk})} f_k(y_j|\boldsymbol{\beta}_k j). \tag{4.3}$$

Class-specific parameters of the above Gamma-von Mises mixture model can be separately updated by the EM algorithm within the M-step. In particular, standard derivative computations show that contributions to the expected log-likelihood function given by the circular data, $Q_j(\boldsymbol{\beta}_j|\hat{\boldsymbol{\pi}}^{(h-1)}, \hat{\boldsymbol{\beta}}^{(h-1)}), j = 1, 2, 3$ are separately maximized by

- an update of the modal directions, given by

$$\hat{\beta}_{kj0}^{(h)} = \operatorname{arctg} \frac{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)} \sin y_{ij}}{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)} \cos y_{ij}}, \tag{4.4}$$

- and by the roots $\hat{\beta}_{kj1}^{(h)}$ of the three equations

$$\frac{I_0(\beta_{kj1})}{I_0'(\beta_{kj1})} = \frac{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)} \cos(y_{ij} - \hat{\beta}_{kj0}^{(h)})}{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)}}, \tag{4.5}$$

  which are the updated concentrations of wave and wind directions on the circle.

Analogous derivative computations show that the remaining three functions $Q_j\big(\boldsymbol{\beta}_j|\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\beta}}\big), j = 4, 5, 6$, i.e. the contributions of the linear data to the expected log-likelihood function, are separately maximized by

- an update of the shape parameters, given by

$$\hat{\beta}_{kj0}^{(h)} = \frac{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)} y_{ij}}{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)}},$$

- and by the roots $\hat{\beta}_{kj1}^{(h)}$ of the three equations

$$\begin{aligned} \log(\beta_{kj1}) - \psi(\beta_{kj1}) = & \log \left(\frac{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)} y_{ij}}{\sum_{i=1}^{n} \hat{\pi}_{ik}^{(h-1)}} \right) \\ & - \left(\frac{\sum_{i=1}^{n} (1 - r_{ij}) \hat{\pi}_{ik}^{(h-1)} \log y_{ij}}{\sum_{i=1}^{n} \hat{\pi}_{ik}^{(h-1)}} \right), \end{aligned}$$

where $\psi(\beta_{kj1})$ is the Digamma function.

5. Results and Model Validation

The proposed model was estimated from the data illustrated in Section 2, by considering $K = 4, \cdots, 10$ classes. According to the BIC criterion, a model with $K = 7$ classes is needed to adequately describe the data (results not reported here and available upon request to the corresponding author).

Table 2 displays the maximum likelihood estimates and the standard errors of the $7 \times 12 + 7 = 91$ parameters of the model with minimum BIC. The last row of the table indicate the estimated class probabilities $\hat{\boldsymbol{\pi}}$. While point estimates were computed by exploiting the EM algorithm of Section 3, standard errors were computed by taking the square root of the diagonal elements of the inverse observed information matrix, obtained by extracting the observed information from the complete log-likelihood (Louis, 1982). These estimates (all significant at a 95% significance level) can be directly exploited for a variety of applications that include for example the computation of the expected wave load to ships and off-shore structures. In addition, these estimates have an immediate physical interpretation, which can be summarized with the help of Figure 3, which displays the 6×7 densities that have been estimated under model (3.1). To draw this picture, we have used seven different colors (listed in Table 2) to show the grouping of the conditional densities according to the seven latent classes. Latent classes can be interpreted with the help of the map in Figure 1. Components 1, 2 and 7 cluster S-E waves of high (comp. 1), medium (comp. 2) and low (comp. 7) average heights, respectively. As expected, components 1, 2 and 7 are respectively associated with Sirocco winds of high, medium and low speed at both the tide gauges considered for analysis. Components 3 and 5 cluster N-W waves of medium (comp. 3) and low (comp. 5) height, associated with Bora winds of medium (comp. 3) and low (comp. 5) speed, blowing from west and north-west at the two tide gauges. Components 4 and 6 cluster waves with a direction that is perpendicular to the coast (coastal waves) and, as expected, are of moderate/medium heights. However, while waves within latent class 4 are associated with winds blowing along the same direction as waves, waves within latent class 6 are associated with winds coming from north. We note that the occurrence of coastal waves of moderate heights, regardless of wind and speed direction, is responsible for numerical wind-wave models giving inaccurate results in coastal areas. Our mixture model correctly separates coastal waves and wind-generated waves moving along maximal fetch directions. The results additionally suggest that regimes that generate coastal waves cannot be ignored in the analysis of sea conditions, because the probability of occurrence of classes 4 and 6 is about 0.22. We also remark that the model seems able to separate regimes that drive

severe and moderate conditions of the sea. Component 1 detects the distributional shape of wave height and direction of sea storms and identifies the wind conditions under which this event occur.

Table 2: Parameter estimates and standard errors (within brackets)

	Parameters	Component 1 (red)	2 (blue)	3 (green)	4 (yellow)	5 (cyan)	6 (magenta)	7 (orange)
Wave Dir [a] (radians)	mean	1.935	2.102	5.927	0.820	5.097	0.821	2.409
		(0.013)	(0.020)	(0.076)	(0.035)	(0.051)	(0.036)	(0.100)
	concentration	99.278	16.778	1.215	9.377	6.951	6.269	0.898
		(17.476)	(1.862)	(0.108)	(1.343)	(1.341)	(0.784)	(0.116)
Wind Dir [b] (radians)	mean	2.339	2.864	4.632	1.195	4.534	5.701	3.411
		(0.022)	(0.049)	(0.018)	(0.058)	(0.052)	(0.103)	(0.063)
	concentration	30.995	3.122	11.111	3.181	7.057	1.240	1.656
		(5.223)	(0.335)	(0.964)	(0.447)	(1.626)	(0.139)	(0.122)
Wind Dir [c] (radians)	mean	2.305	2.697	5.103	1.065	5.322	5.942	5.319
		(0.022)	(0.169)	(0.012)	(0.020)	(0.034)	(0.064)	(0.082)
	concentration	33.105	0.632	24.259	24.716	13.778	2.467	1.064
		(6.116)	(0.130)	(2.357)	(3.908)	(2.543)	(0.241)	(0.110)
Wave Height [a] (meters)	shape	99.226	5.782	12.556	35.778	26.081	10.147	3.031
		(20.640)	(0.701)	(1.148)	(5.125)	(5.116)	(1.360)	(0.247)
	scale	0.029	0.174	0.078	0.055	0.014	0.179	0.146
		(0.006)	(0.021)	(0.007)	(0.008)	(0.003)	(0.024)	(0.014)
Wind Speed [b] (meters/sec)	shape	10.372	5.140	11.779	9.358	9.768	7.685	2.276
		(1.756)	(0.592)	(0.915)	(1.289)	(1.630)	(1.013)	(0.177)
	scale	0.517	0.628	0.420	0.386	0.192	0.771	0.896
		(0.089)	(0.070)	(0.033)	(0.056)	(0.033)	(0.099)	(0.084)
Wind Speed [c] (meters/sec)	shape	22.656	4.497	10.517	12.732	13.004	4.159	6.232
		(4.712)	(0.586)	(0.817)	(1.683)	(2.354)	(0.452)	(0.577)
	scale	0.223	0.626	0.296	0.448	0.083	0.969	0.210
		(0.045)	(0.077)	(0.023)	(0.060)	(0.015)	(0.110)	(0.021)
	probability	0.053	0.166	0.250	0.087	0.060	0.135	0.249
		(0.006)	(0.013)	(0.012)	(0.008)	(0.007)	(0.010)	(0.015)

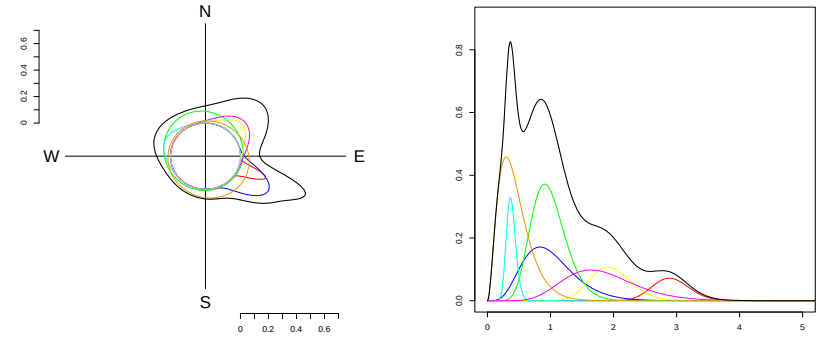
[a] Ancona buoy - [b] Ancona tide gauge - [c] Ravenna tide gauge

Figures 4 and 5 display the classification of the multivariate profiles, as obtained by modal allocation, i.e. assigning each profile i to the latent class k with the highest probability $\hat{\pi}_{ik}$. Fiducial intervals for each single observation y_{ij} were obtained on the basis of the estimated conditional distribution (5.1) whose expectation

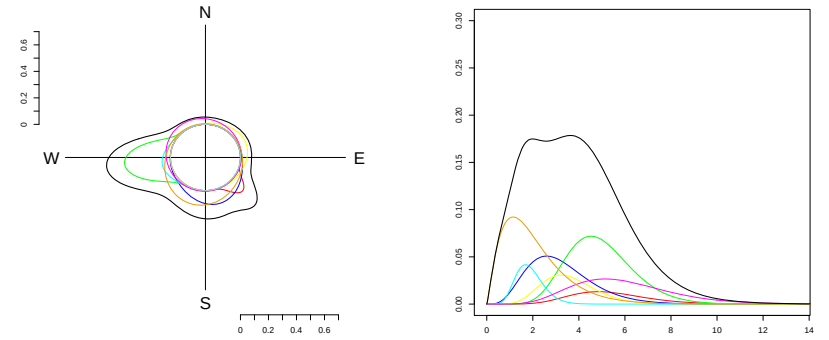
$$\mathbb{E}(y_{ij}|y_{il}, l \neq j; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}) = \sum_{k=1}^{K} \frac{\hat{\pi}_k \prod_{h \neq j} f_k(y_{ih}|\hat{\boldsymbol{\beta}}_{hk})}{\sum_{k=1}^{K} \hat{\pi}_k \prod_{h \neq j} f_k(y_{ih}|\hat{\boldsymbol{\beta}}_{hk})} \mathbb{E}_k(y_{ij}|\hat{\boldsymbol{\beta}}_k j) \tag{5.1}$$

was exploited to impute missing values (the black dots in Figures 4 and 5). The model gives an adequate fit of the observed data (right-hand histograms in Figures

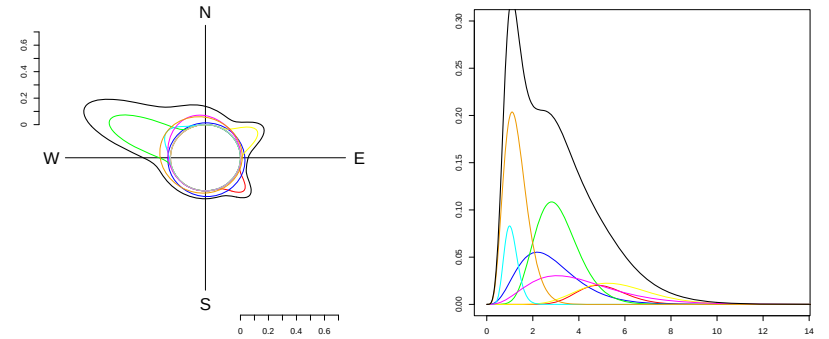
4 and 5 and, simultaneously, operates an intuitively appealing classification of complete and incomplete profiles of wind and wave measurements.



wave direction - Ancona buoy wave height - Ancona buoy



wind direction - Ancona tide gauge wind speed - Ancona tide gauge



wind direction - Ravenna tide gauge wind speed - Ravenna tide gauge

Figure 3: Densities of wave direction and height at the Ancona buoy (top) and wind direction and speed (middle: Ancona tide gauge; bottom: Ravenna tide gauge), as estimated by a 7-components LC cluster model; coloured lines indicate conditional densities and black lines indicate mixture densities

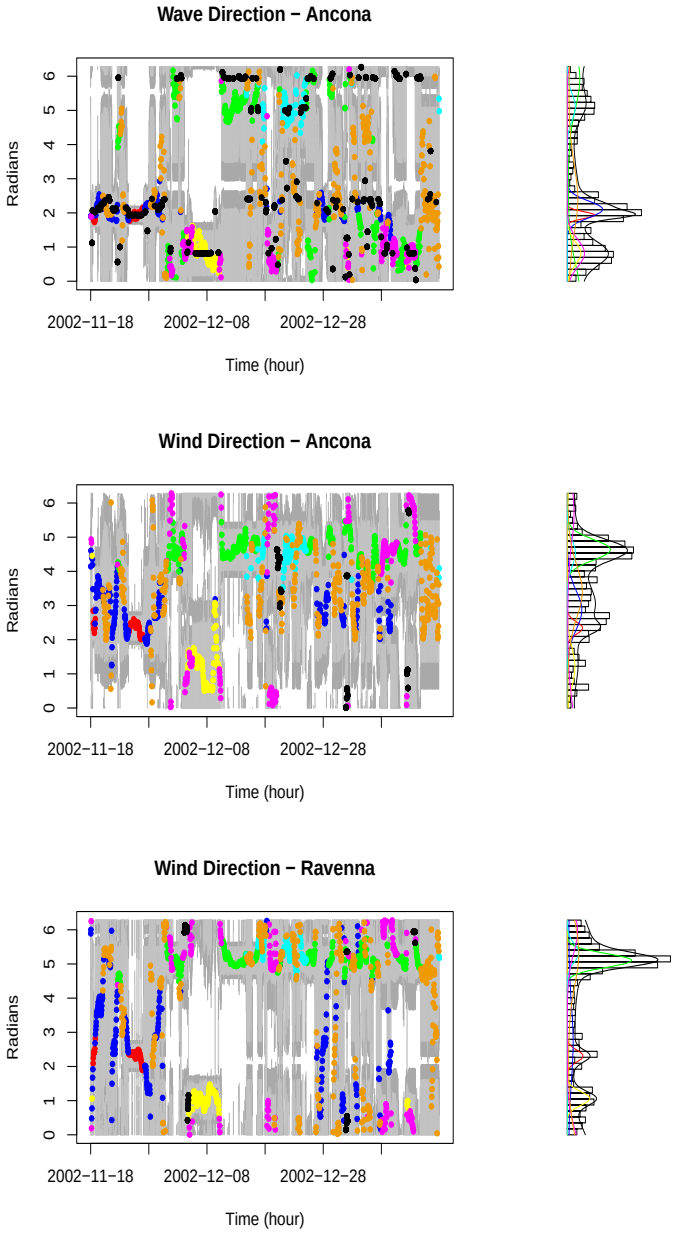


Figure 4: Left: directional data, clustered into seven latent classes and 95% (grey) and 99% (dark grey) fiducial intervals, as estimated by a 7-components mixture model. Black dots indicate missing values, imputed by the expectation of the conditional distribution of the missing values given the observed data, as estimated by the model. Right: histograms of complete data fitted by the model

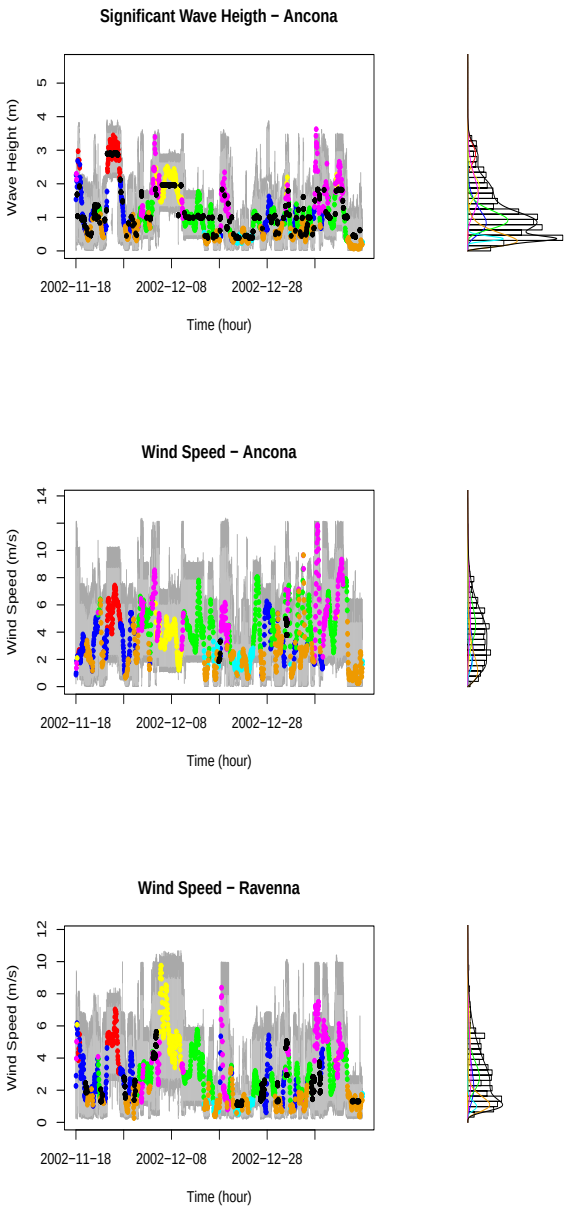


Figure 5: Left: linear data, clustered into seven latent classes and 95% (grey) and 99% (dark grey) fiducial intervals, as estimated by a 7-components mixture model. Black dots indicate missing values, imputed by the expectation of the conditional distribution of the missing values given the observed data, as estimated by the model. Right: histograms of complete data fitted by the model

Goodness of fit was also evaluated by comparing the squared cross-correlations between the observed data and those expected by the mixture model. To compute the empirical correlation between intensity observations (wind speed and wave height), we have used the standard Pearson correlation. The empirical correlation between circular data (wind and wave direction) was computed by exploiting the Fisher-Lee correlation index (Fisher and Lee, 1983). Finally, we computed the cross-correlation between linear and circular data (e.g., between wind direction and wave height) by exploiting the Mardia's linear-circular correlation index (Mardia, 1976). Table 3 displays a reasonable matching between the empirical correlations against their expected counterparts, under the estimated mixture model, showing that the conditional independence assumption of model (3.1) (coupled with the choice of 7 latent classes) explains a significant part of data variability.

Table 3: Observed and expected squared correlations

	Wave H.	Wind S. [a]	Wind S. [b]	Wave D.	Wind D. [a]	Wind D. [b]
Wave Heigth	1					
(expected)	(1)					
Wind Speed [a]	0.191	1				
(expected)	(0.320)	(1)				
Wind Speed [b]	0.385	0.142	1			
(expected)	(0.517)	(0.199)	(1)			
Wave Direction	0.199	0.111	0.168	1		
(expected)	(0.193)	(0.117)	(0.165)	(1)		
Wind Direction [a]	0.233	0.108	0.193	0.002	1	
(expected)	(0.222)	(0.165)	(0.111)	(0.005)	(1)	
Wind Direction [b]	0.184	0.007	0.119	0.008	0.017	1
(expected)	(0.194)	(0.003)	(0.156)	(0.011)	(0.027)	(1)

[a] Tide gauge: Ancona - [b] Tide gauge: Ravenna

We also evaluated the predictive accuracy of the model by non-parametric cross-validation (Gelman *et al.*, 1998). More precisely, we randomly split the sample in 10 subsamples. From each subsample, we discarded the 10% of the observations and (1) use the remaining portion of the subsample to fit a new model and (2) draw 5 imputations for each discarded vector of data, from the estimated conditional distribution of the discarded values given the observed data. If multiple imputations were of good quality, then we would expect than the actual outcome and the multiple imputations to have the same distributions, so that if one ranked the actual response along to the 5 imputations, then all 6 possible

orderings (actual outcome lowest, second lowest,$\cdots$, highest) would be equally likely. Figure 6 displays the cumulative distribution functions of the 6 ranks of circular and linear outcomes (overlapped to that of the uniform distribution), showing the good predictive accuracy of the model.

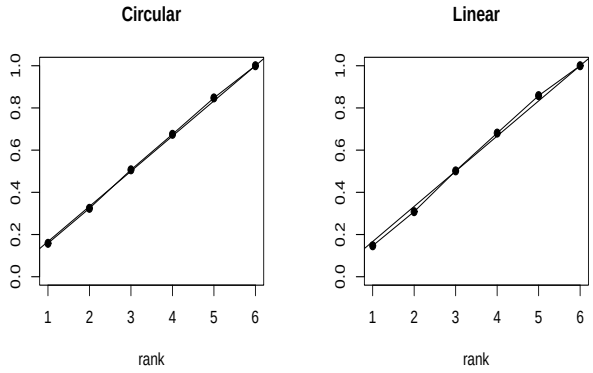


Figure 6: Rank cumulative distribution of the actual outcome with respect to 5 multiple imputations in a cross-validation experiment and cumulative distribution function of a uniform distribution

6. Discussion

We propose a latent-class approach to identify wave regimes under various wind conditions and estimate regime-specific wave parameters, such as modal wave directions and average wave heights, in the case of incomplete data, observed at different locations.

Using mixtures of product densities to model multivariate data allows for a simple specification of the dependence structure between variables that are measured on different supports (e.g. linear and circular) and, simultaneously, provides a flexible framework within which a variety of different parametric families can be exploited to model the univariate distribution of each single variable, given the latent class. We exploited von Mises and Gamma distributions, but the estimation procedure of Section 3 can be implemented by choosing different parametric families that can be more suitable in different case studies. By assuming a mixture densities, moreover, missing values are efficiently handled in a maximum-likelihood framework.

Modelling flexibility and computational efficiency in the case of incomplete data information come at the price of a simplifying constraint on the dependence structure among variables, given by the conditional independence assumption. In marine studies, this assumption can be often motivated by empirical evidence of a number of latent sea regimes and by the need of clustering the data in a

way that the association structure between the observed variables is well approximated by this partitioning of the sample. Nevertheless, issues of goodness of fit should be carefully addressed. Rigorous goodness-of-fit methods are however problematic with missing values. We have obtained reassuring results by computing case-wise fiducial intervals, overlaying the estimated marginal densities of the variables on the observed histograms (Figures 4 and 5) and comparing expected and empirical squared correlations between the variables (Table 3). These results should be interpreted with care, because empirical histograms and correlations are computed after discarding the missing values and because having most of the observed values within fiducial intervals says little about their ability to include missing values. These issues motivated our cross-validation experiment, whose results indicate that the proposed model was capable to explain most of the data variability and to re-impute artificially-removed values with a reasonable accuracy.

References

Bertotti, L. and Cavalieri, L. (2009). Wind and wave predictions in the Adriatic Sea. *Journal of Marine Systems* **78**, S227-S234.

Boukhanovsky, A. V., Lopatouhkin, L. J. and Guedes Soares, C. (2007). Spectral wave climate of the North Sea. *Applied Ocean Research* **29**, 146-154.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1-38.

Faltinsen, O. M. (1990). *Sea Loads on Ships and Offshore Structures*. Cambridge University Press, Cambridge.

Fernández-Durán, J. J. (2007). Models for circular-linear and circular-circular data constructed from circular distributions based on nonnegative trigonometric sums. *Biometrics* **63**, 579-585.

Fisher, N. I. and Lee, A. J. (1983). A correlation coefficient for circular data. *Biometrika* **70**, 327-332.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of American Statistical Association* **97**, 611-631.

Gelman, A., King, G. and Liu, C. (1998). Not asked or not answered: multiple imputation for multiple surveys. *Journal of the American Statistical Association* **93**, 846-874.

Hagenaars, J. A. and McCutcheon, A. L. (2002). *Applied Latent Class Analysis.* Cambridge University Press, Cambridge.

Hamilton, L. J. (2010). Characterising spectral sea wave conditions with statistical clustering of actual spectra. *Applied Ocean Research* **32**, 332-342.

Huang, G., Wing-Keung Law, A. and Huang, Z. (2011). Wave-induced drift of small floating objects in regular waves. *Ocean Engineering* **38**, 712-718.

Hunt, L. and Jorgensen, M. (2003). Mixture model clustering for mixed data with missing information. *Computational Statistics and Data Analysis* **41**, 429-440.

Jin, K. R. and Ji, Z. G. (2004). Case study: modeling of sediment transport and wind-wave impact in Lake Okeechobee. *Journal of Hydraulic Engineering* **130**, 1055-1067.

Kato, S. and Shimizu, K. (2008). Dependent models for observations which include angular ones. *Journal of Statistical Planning and Inference* **138**, 3538-3549.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44**, 226-233.

Mardia, K. V. (1976). Linear-circular correlation coefficients and rhythmometry. *Biometrika* **63**, 403-405.

Mardia, K. V., Hughes, G., Taylor, C. C. and Singh, H. (2008). A multivariate von Mises distribution with applications to bioinformatics. *Canadian Journal of Statistics* **36**, 99-109.

Pleskachevsky, A., Eppel, D. P. and Kapitza, H. (2009). Interaction of waves, currents and tides, and wave-energy impact on the beach area of Sylt Island. *Ocean Dynamics* **59**, 451-461.

Rotnitzky, A. and Wypij, D. (1994). A note on the bias of estimators with missing data. *Biometrics* **50**, 1163-1170.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* Chapmanan and Hall, London.

Teicher, H. (1967). Identifiability of mixtures of product measures. *Annals of Mathematical Statistics* **38**, 1300-1302.

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A. and Sijtsma, K. (2008). Multiple imputation of categorical data using latent class analysis. *Sociological Methodology* **33**, 369-297.

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *Annals of Mathematical Statistics* **39**, 209-214.

Wu, C. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11**, 95-103.

Received November 18, 2010; accepted April 21, 2011.

Francesco Lagona
DIPES and GRASPA Unit of Rome
Roma Tre University
Chiabrera 199, 00145 Rome, Italy
lagona@uniroma3.it

Marco Picone
Department of Economics and GRASPA Unit of Rome
Roma Tre University
Chiabrera 199, 00145 Rome, Italy
marco.picone@uniroma3.it