Data Validation in Italian Wave Metric Datasets

Mauro Bencivenga[1], Marco Picone[1], Francesco Lagona[2], Gabriele Nardone[1], Sara Morucci[1]

[1]ISPRA - Rome (Italy), [2]University of Roma Tre - Rome (Italy)

ISPRA Istituto Superiore per la Protezione e la Ricerca Ambientale

EMODnet European Marine Observation and Data Network

Italian Buoy Network

The Italian National Wave Measurement Network run 15 real-time directional buoys uniformly distributed along the Italian coasts. Data have been collected since 1989 at 8 measurement stations; in 1999 two other stations were added and the remaining five buoys were moored in 2001. From 2010 all stations are equipped with meteorological instruments. It provides real-time meteorological and marine parameters such as significant wave height, peak period, mean period, mean direction, sea surface temperature, wind speed, wind gust, wind direction, air temperature, dew point, relative humidity and atmospheric pressure.
Data are disseminated through the Internet, Televideo RAI and Global Telecommunication System with the co-operation of the Italian Meteorological Service (USAM). These data provide information foundamental to all studies, design and management of coastal and offshore works as well as to the relative environmental impacts.
RON data are necessary in order to characterize the sea state in the Central Mediterranean sea and to develop predictive models and statistical analysis.

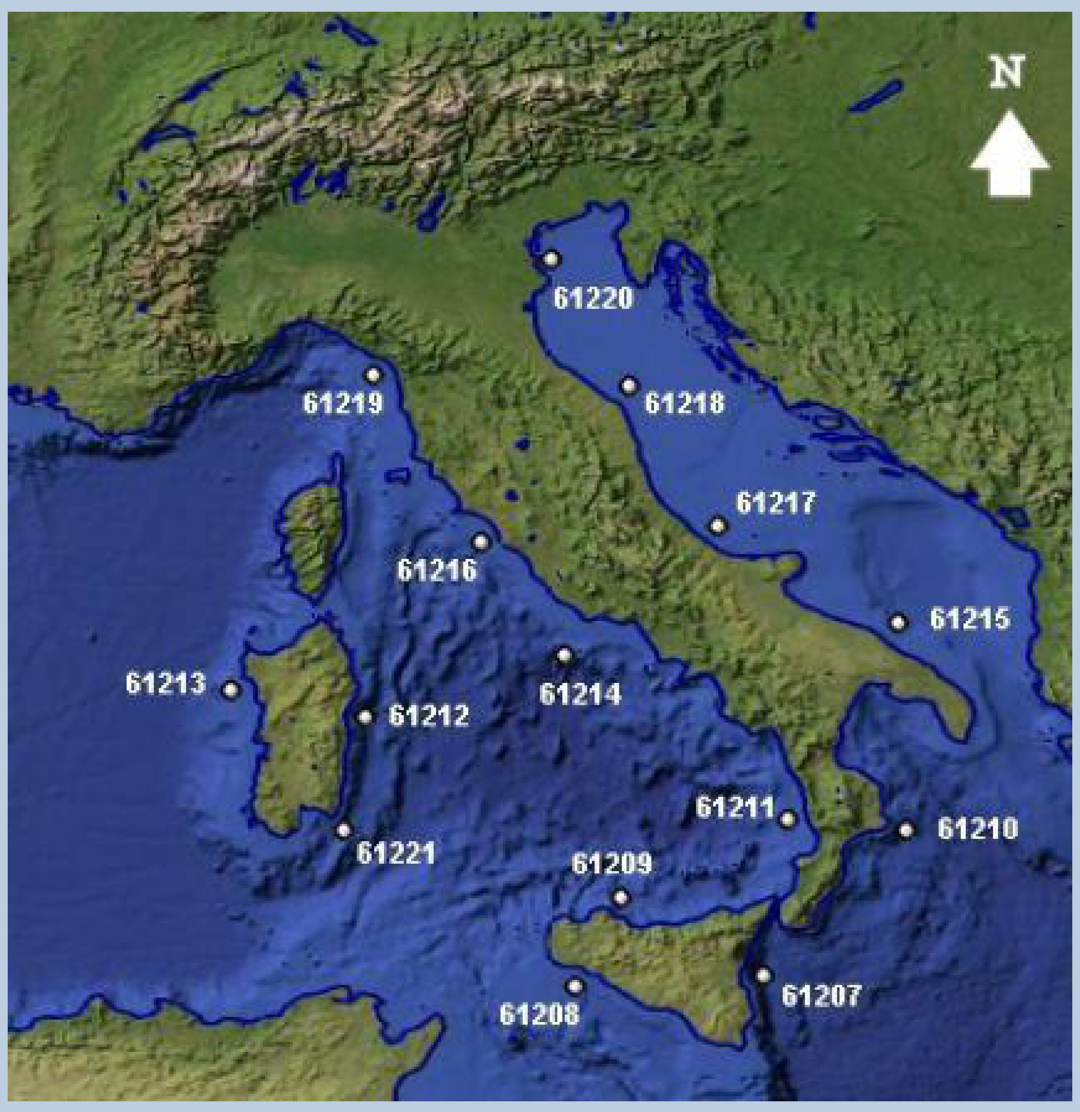

Figure 1: The Italian Data Buoy Network (RON), http://www.telemisura.it

Wind-Wave Data

Studies of air-sea interactions involve the analysis of multivariate time series of marine data, that include measurements of mixed type variables, like wind and wave direction (circular variables) and wind speed and wave height (linear variables). Marine datasets, anomalous values occour because of device transmission malfunctioning, measurement errors, unmoorings, missing information. Data quality control is essential in order to obtain validate dataset and rigorous statistical, numerical and physical results.
The exploited data include time series of semi-hourly wave and wind directions, wind speeds and wave heights, recorded in the period 2009-2012 by the buoy of Ancona, located in the Adriatic Sea, about 30 km far from the coast. Relevant wind events in the Adriatic Sea are typically generated by the southeastern Sirocco, the northern Bora and the northwestern Maestral. Bora and Sirocco episodes are usually associated with high-speed flows (especially in wintertime), while Maestral is usually related to good meteorological conditions.
Height and direction of waves are just partially influenced by wind conditions in semi-enclosed basins. The orography of the Adriatic Sea plays a key role in this case and most of the waves tend to propagate from north-northwest and southeasterly along the major axis of the basin, where they can travel freely, without being obstructed by physical obstacles, such as coastlines.
A multivariate mixture model (MMM), describing wind-wave data in terms of latent environmental regimes, i.e., specific distributions that the data take under latent environmental conditions, is presented in this work. It allows to cluster, under unsupervised conditions, any kind of data such as positive data (almost all oceanographic and meteorological measurements), circular data (typically directional data), univariate and multivariate observations.

Latent Class Model for validation porpose

Multivariate Latent Class Models, or Mixture Models (MMM), are particularly usefull under complex coastal/offshore conditions, such as closed basins or coastal areas, where the correlation structure of the data can be decomposed according to a finite number of easily interpretable distributions. We approximate the joint distribution of the data by a mixture of multivariate densities, each one specified as the product of a bivariate von Mises and a bivariate skew normal density. In this setting, wind and wave directions are segmented by toroidal clusters, while bivariate observations of wind speed and wave height are clustered within skewed ellipses. This allows to cluster mixed linear and circular data separately, avoiding the definition of hyper-cylindrical clusters that could be hardly understood. The time series $\boldsymbol{z} = (\boldsymbol{z}_i, i = 0, \ldots, T)$, can be split in bivariate circular and linear components, say $\boldsymbol{z}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i)$, $\boldsymbol{x}_i = (x_{1i}, x_{2i}) \in (-\pi, \pi]^2$ and $\boldsymbol{y}_t = (y_{1i}, y_{2i}) \in \mathbb{R}^2$. Formally, the multivariate distribution of $\boldsymbol{z}_i$ is given by

$$f(\boldsymbol{z}|\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{k=1}^{K} \pi_k f_c(\boldsymbol{x}|\boldsymbol{\beta}_k) f_l(\boldsymbol{y}|\gamma_k),$$

where β_k e γ_k are set of parametrs of the mixture model, π_k is the probability that the observation i belongs to cluster k and K is the number of clusters. A maximum likelihood approach is used to estimate parameters and to evaluate the class membership, through the implementation of an EM algorithm that solve the equation

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f_c(\boldsymbol{x}_i|\boldsymbol{\beta}_k) f_l(\boldsymbol{y}_i|\gamma_k)$$

and that iteratively converges to a solution. This procedure allow to cluster the observations in a finite number of relevant sea regimes. According to the $f(\boldsymbol{z}_i)$, conditional features of the model can be examined by computing predictive intervals from the conditional distributions of each variable given the values of the remaining variables, as estimated by the model. As a result, it is possible to define a 95% confident interval for each time, so the i-th observation could be pointed out as good or anomalous value.

Application of Validation Procedure

Given a semi-hourly quadrivariate profiles of wind speed, wave height, wind direction and wave direction measured in Ancona, various mixture models have been estimated, using a number of components K ranging from 2 to 5. K is fitted according to the Bayesian Information Criterion (BIC) and the Integrated Complete Likelihood (ICL) statistic, that suggest a model with $K = 3$ components.
The first component of the model (Fig.2 - red) is related to periods of calm sea. As expected, wind and wave directions are poorly synchronized in case of good sea conditions.
The second component (blue) is associated with bora and maestral episodes. Winds drive high waves that come from the north. Wind and wave directions appear strongly correlated.
The third component (green) is related to episodes of sirocco. Wind and waves come from the major axis of the basin and generate rough sea.

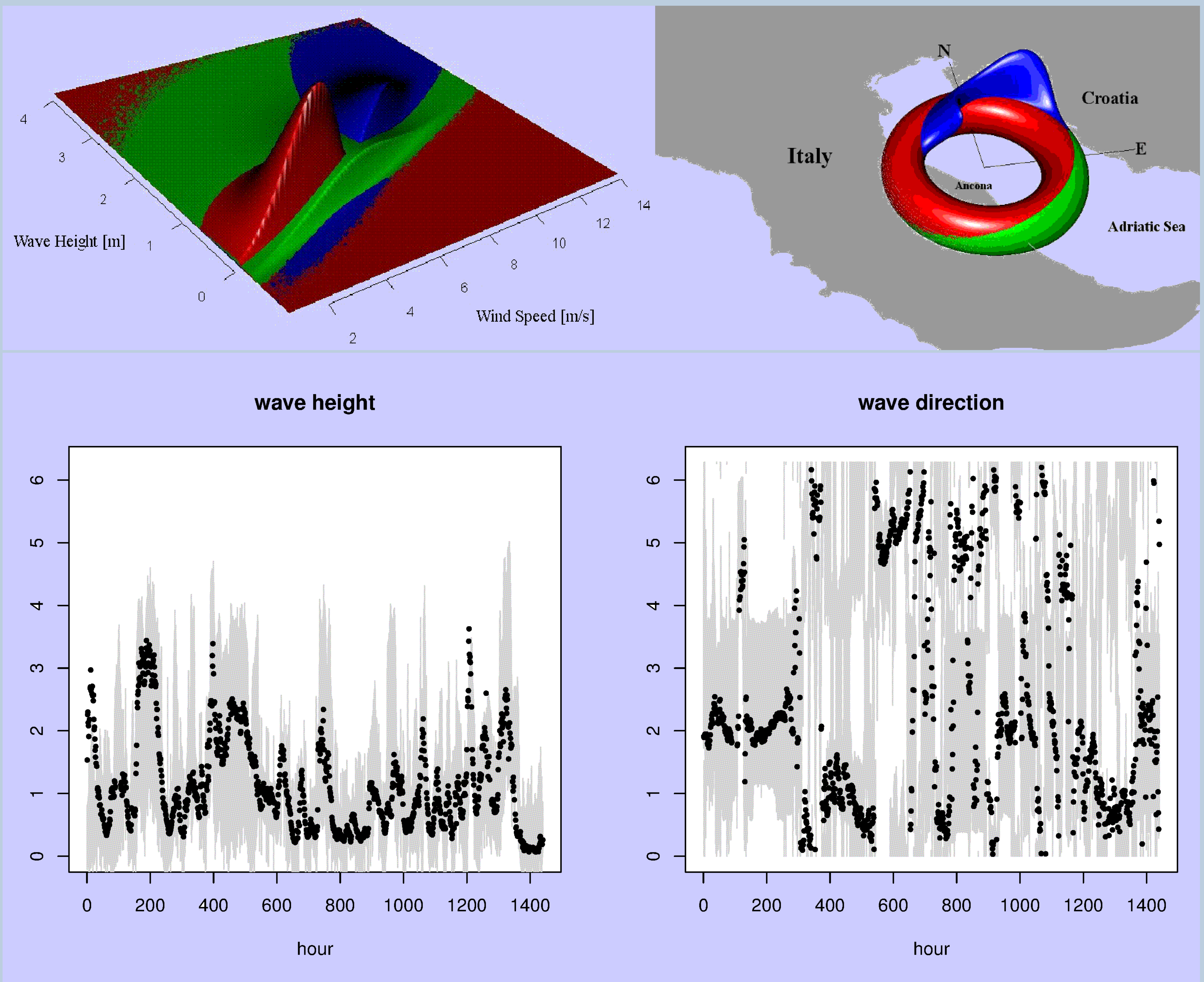

Figure 2: Joint distribution of wind-wave data (above) and confidence interval for each observation (below) according to the mixture model.

References

[1] F. Lagona, M. Picone: *Model-based clustering of multivariate skew data with circular components and missing values*, Journal of Applied Statistics, 39 (5), pp.927-945 (2012).

[2] M. Bencivenga, G. Nardone, F. Ruggiero, D. Calore: *The Italian Data Buoy Network (RON)*, Advances in Fluid Mechanics IX (2012).