OPTIMIZING NEURAL NETWORK TRAINING DATA SELECTION FOR NEARSHORE SEA STATE FORECASTING THROUGH CLUSTERING ALGORITHMS

Elisa Castro, University of Catania, elisa.castro@phd.unict.it
Luca Cavallaro, University of Catania, luca.cavallaro@unict.it
Claudio Iuppa, University of Messina, claudio.iuppa@unime.it
David Lucio Fernández, IH Cantabria, david.lucio@unican.es
Javier López Lara, IH Cantabria, jav.lopez@unican.es
Enrico Foti, University of Catania, enrico.foti@unict.it
Rosaria Ester Musumeci, University of Catania, rosaria.musumeci@unict.it

The growth of maritime traffic led to increased maritime accidents and multiple factors, including weather conditions, can contribute to these accidents. To mitigate the accidents in port areas, the scientific community is dedicated to recognizing risks and formulating effective strategies for prevention (Marino et al., 2023).
Therefore, understanding local weather and sea conditions well in advance can significantly reduce the likelihood of accidents in port areas. Nevertheless, the ability to forecast sea conditions in deep water, where wave-seabed interactions are irrelevant, is generally available. Using this information, numerical models assess the sea state in the nearby coastal areas.
However, these models require extensive computational resources and are unsuitable for immediate short-term predictions and forecasting systems (Salah et al., 2016).

Artificial Neural Networks (ANNs) offer a promising solution to address this challenge by efficiently acquiring the necessary metocean data (Cavallaro et al., 2023). However, the current ANNs developed to date have been trained using a multitude of scenarios provided by physically based models, demanding substantial memory and computational resources.

In a study conducted by Iuppa et al. (2022), the sensitivity of these ANNs was assessed across three datasets. The initial dataset comprised 146,875 scenarios, while the subsequent two datasets were reduced by 95% and 99% from the first. These datasets were generated by altering wind characteristics (velocity and direction) and wave attributes (significant wave height, mean direction, and peak period) without following specific criteria. The results of this study showed that a reduction in the input data had relevant consequences, especially in the estimation of the wave direction.

Numerous clustering methods have emerged in this context to manage vast data volumes effectively. These methods extract characteristics from the initial set of N data, creating a more condensed and feasible presentation of significant data properties (Camus et al., 2011). In this sense, the primary objective of this research is to contribute to developing a novel approach for determining the number of simulations to use as input data in the ANNs. The initial step involves evaluating clustering methods to select a representative dataset of wave climate. Specifically, the study focuses on the area surrounding the port of Augusta (SR), one of Italy's most significant ports.

The Maximum Dissimilarity Algorithm (MDA), which aims to select a representative subset of reduced size M from an initial database of size N, was applied to the offshore wave data acquired from the CMENS Copernicus repository and to the wind data acquired from ERA5 to reduce the number of simulations to be carried out on a spectral wave propagation model (SWAN).
The MDA was applied to wind and sea characteristics, which include: the significant wave heights *Hs*, the peak period *Tp*, the direction of the wave Dir_{wave}, the velocity *V* and the wind direction Dir_{wind}. Therefore, different possible sea states were examined by varying the number of the selected cases from 250 to 5000. An example of the application of MDA for 500 cases is represented in Figure 1 by the red points.

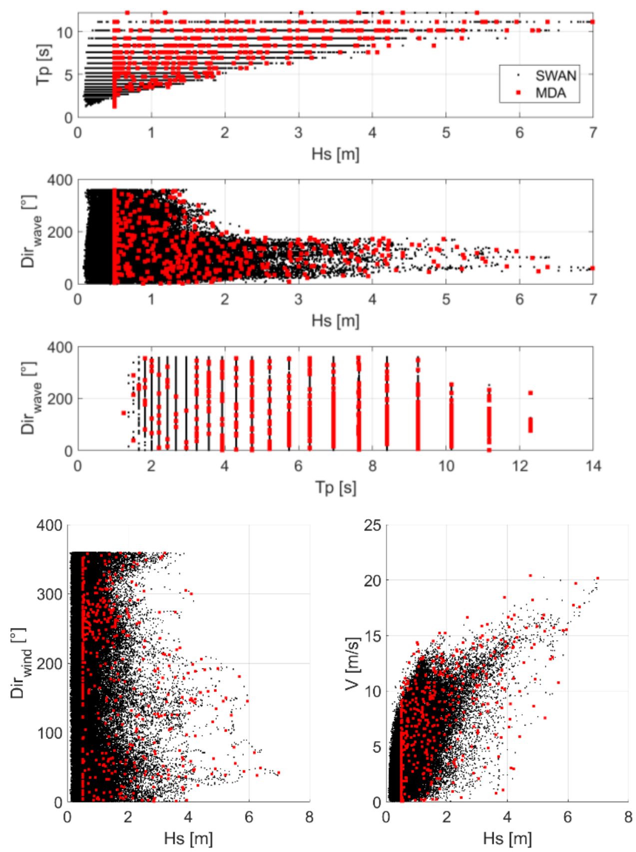

Figure 1 - MDA results for 500 cases

A set of feed-forwards MLP Multilayer Perceptron neural networks with one hidden layer was developed and

trained with the back-propagation algorithm for various points of the area. The ANNs were trained with the offshore condition in the input layer and the wave climate evaluated by the model SWAN in several points in the output layer. The whole dataset was divided into three sub-datasets: 70% for the training of the network, 15% for the validation and 15% for the test. In this application, the calibration was carried out using the Bayesian regularisation method, with the hyperbolic tangent as activation function in the input layer and the sigmoid one for the output layer. The number of neurons in the hidden layer was set equal to 100, according to Cavallaro et al. (2023).
The goodness of fitting was evaluated for each reduced subset obtained by the MDA using the Root Mean Squared Error (RMSE) to find the optimal configuration. For example, a point of Porto Xifonio is considered (see Figure 2), and ANNs were trained with a representative subset of 250, 500, 1000, 3000 and 5000 scenarios of possible sea states.

Figure 2 - Case study and location of the considered point of Porto Xifonio

Figure 3 shows the results of the ANNs' performance evaluation using RMSE. Regarding significant wave height, the RMSE values are 0.04 m for 250 and 500 scenarios, 0.019 m for 1000 scenarios, 0.016 m for 3000 scenarios, and 0.019 m for 5000 scenarios.
To verify the reliability of the neural networks developed for the selected point, an entire storm that recently affected the port of Augusta was simulated through SWAN, considering the characteristics of storm waves as input data. The simulations cover the period between 22 February 2019 and 25 February 2019 with a step of 1 hour. The comparison between the model SWAN and the ANNs trained with different subsets is shown in Figure 4. The comparison showed that for the point of Porto Xifonio, the reduction of the size of the dataset does not influence the estimation of all wave characteristics.

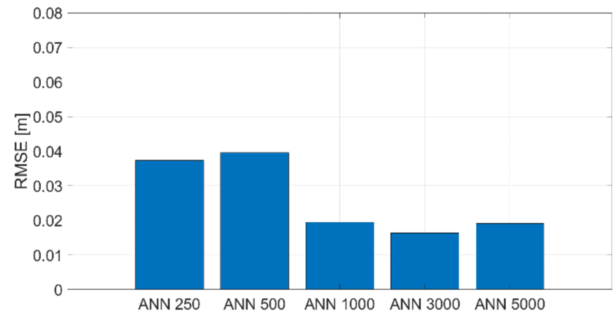

Figure 3 - Performance of the ANNs trained with different subsets provided by MDA.

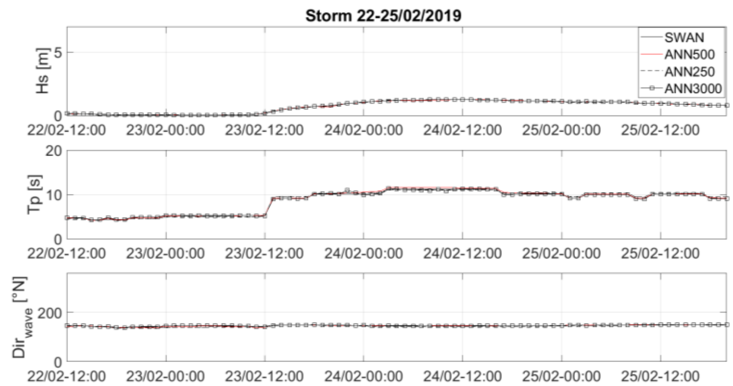

Figure 4 - Comparison between the wave characteristics estimated with SWAN and those estimated by the ANN with 100 neurons in the hidden layer. The ANNs were calibrated with the MDA's representative subset of 250, 500 and 3000 scenarios.

REFERENCES

Camus, P., Mendez, F. J., Medina, R., & Cofiño, A. S. (2011). Analysis of clustering and selection algorithms for the study of multivariate wave climate. *Coastal Engineering*, *58*(6), 453–462.
Cavallaro, L., Iuppa, C., Castro, E., Faraci, C., Musumeci, R. E., & Foti, E. (2023). WIND AND WAVE TRAINED ARTIFICIAL NEURAL NETWORKS FOR THE FORECASTING OF WAVE CLIMATE IN HARBOUR AREA. *Coastal Engineering Proceedings*, *37*(37), 58.
Iuppa, C., Cavallaro, L., Castro, E., Musumeci, R. E., Faraci, C., Carlo, L., Passalacqua, G., & Foti, E. (2022). Wave forecasts in the proximity of a harbour area based on artificial neural networks. *2022 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters - Proceedings*, 39–43.
Marino, M., Cavallaro, L., Castro, E., Musumeci, R. E., Martignoni, M., Roman, F., & Foti, E. (2023). New frontiers in the risk assessment of ship collision. *Ocean Engineering*, *274*(January), 113999.
Salah, P., Reisi-Dehkordi, A., & Kamranzad, B. (2016). A hybrid approach to estimate the nearshore wave characteristics in the Persian Gulf. *Applied Ocean Research*, *57*, 1-7.