

A Neural Network Approach to the Problem of Recovering Lost Data in a Network of Marine Buoys

S. Puca*, B. Tirozzi*, G. Arena**, S. Corsini**, R. Inghilesi**

* University of Rome "La Sapienza", Department of Physics

** SIMN (National Hydrological and Marine Survey – National Technical Surveys Dept. of Italy)

Abstract

Neural Network (NN) technology provides several reliable tools for analysis in many science and technology applications. In particular NN are often applied to the development of statistical models for intrinsically non-linear systems, since NN usually behave better than ARMA or GARCH models in complex conditions. A member of this class of problems is the analysis of time series of significant wave heights from a network of buoys. A project is being carried out by the Italian DSTN-SIMN (Technical Surveys Dept. – National Hydrological and Marine Survey) and the Dept. Of Physics of the University of Rome "La Sapienza", in order to reproduce the time series collected by the Italian SWaN network of buoys (Sea Wave monitoring Network). Aim of the project is the determination of the best way to fill gaps and long periods of missing data with the best accuracy by means of a reanalysis of the whole ten years' data set of the SWaN. Here a NN model is proposed for time-space analyses of the marine data. Main feature of the tool is the ability to reproduce long time series of data without any increase of the error. The method is based on a preliminary spatial analysis of the wave climates in order to classify the degree of overlapping of information from different stations. This overlapping, where possible, led to an optimal and selective training of the NN by means of data collected in different, nearby, locations. NN numerical simulations of some important historical storm are compared with the data originally observed at the stations of Crotona (Ionian Sea), Pescara (Adriatic Sea) and Monopoli (Adriatic Sea).

. Introduction

The Sea Wave monitoring Network (SWAN) is a network of 10 buoys moored all round the coasts of Italy (Arena et al., 1997). It has been working since June 1989 with the original 8 stations of Alghero, La Spezia, Ponza, Mazara, Catania, Monopoli, and Ortona. Two stations, Cetraro and Ancona, have been added since 1999. At present time six buoys provide real time data, all the network will be real time working within the current year. In more than eleven years of activity, the efficiency of the monitoring system was high on the average, Overall statistics indicate that less than 5% of data were lost for 6 out of 8 buoys. For only two stations, namely Mazara and Ponza, the percentage is higher, about 10-15% (Corsini et al., 2000). But even for the best of the circumstances no system can provide for 100% of the data for indefinitely long periods. Causes of relatively long periods of missing



data were mainly unmoorings of the buoy due to occasional collisions with ships, or prolonged radio transmission problems. The occurrence of problems in the measure or even in the transmission of data affects both the principal activities of the network. It decreases the real-time monitoring efficiency (every buoy fault produces a decrease of 16% of the overall efficiency of the system), and depletes the quality of the long term statistics. It must be observed that most of the problems at sea are associated with severe weather conditions. That is to say that more often than not lost information would have been the most valuable. An other aspect of the problem is that in the ordinary time series statistical analysis there are tasks, like the evaluation of the autocorrelation function, the Fourier analysis or the extreme wave analysis, which

requires some sort of replacement of the gaps. Methods' reliability depends usually on the length of the gap, univariate methods hardly can provide reliable estimates of entire storm episodes hidden by large (two weeks long, for instance) gaps. The aim of the present work is the assessment of a practical and reliable method of recovering lost information by means of multivariate Neural Network methods.

NN approach

There are several methods currently available to fill the gaps of a time series, or, at least, to evaluate their possible influence on the statistical analysis. Most of them are univariate techniques, i.e. methods that operate on the same time series. There can be statistical or empirical methods, which just help to fill the gaps in a way that preserves some feature (the overall expected value, as an example). More sophisticated methods have the aim of simulating the actual time series by means of the past history (ARMA or Neural Networks). The intrinsic error in the case of ARMA model comes from the fact that the prediction (or the estimate) of missing data is obtained by conditional expectation with respect to the known data. In this algorithms each estimate of the variate at a certain time implies a prediction error, which adds up at every time step. Same situation occurs with neural networks applied in the same way, that is, using a NN which captures the relationship among the data at time $t+1$ and those at previous times $t, t-1, t-m$. Ordinary NN approach would result in errors in the estimates growing really fast with the size of the gap, giving meaningless estimates after a few iterations. The problem of numerically evaluating the time series of observations collected at a single station has little chances to be successfully solved. Dealing with networks of evenly positioned stations, on the other hand, allows the application of different strategies, in particular it suggest the use of adaptive methods with multivariate data. The method which led to the best results in simulating the behaviour of a long time series for the SWAN was found to be the use of a superposition of the direction and significant wave height (H_{m0}) information provided from the nearby stations with non-linear coefficients (weights) estimated by means of a SWAN-tailored NN. The directional information was responsible for the weights (correlations) attributed by the NN to H_{m0} data collected at different locations.

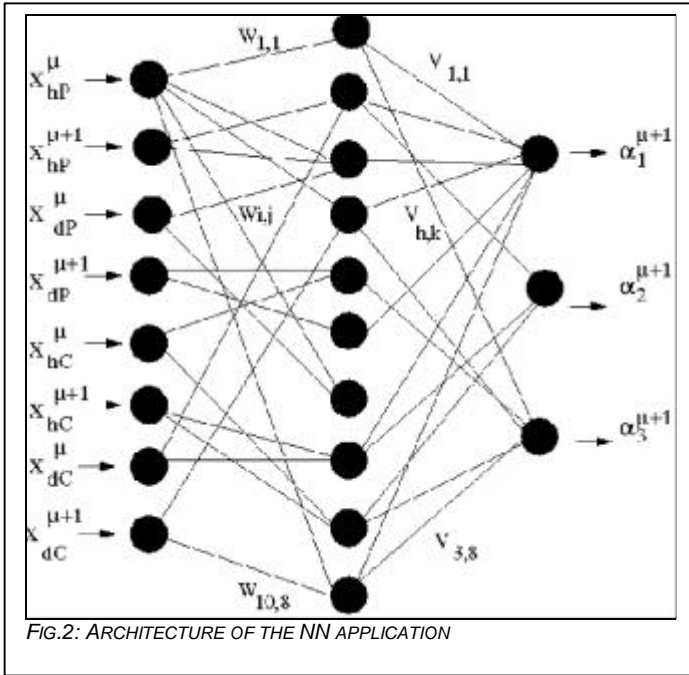


FIG.2: ARCHITECTURE OF THE NN APPLICATION

Neural Network Architecture

Neural Networks are a class of algorithms particularly suitable for the modelling of time series. Data can be chaotic or stochastic, usually governed by deterministic evolution equations with stochastic terms (Herz et al., 1993). These algorithms can detect any form of non linear relation which generates a sequence of input-output pairs (patterns)

$\{\bar{x}^m, \bar{y}^m\}_{m=1}^P$, where \bar{x}^m is a n -dimensional vector and \bar{y}^m is a m -dimensional vector, μ is a temporal index. Once it is assumed the existence of a function $f: R^n \rightarrow R^m$ of the following form: $\bar{y}^m = f(\bar{x}^m)$, the aim of the algorithm is then to determine the best approximation of the function $f(\bar{x}^m)$. If the output of the neural

network is called \bar{z}^m , a learning process is introduced in order to minimise the learning error E_L

$$1) \quad E_L = \frac{1}{P} \sum_{m=1}^P \left(\sum_{i=1}^m |z_i^m - y_i^m| \right)$$

on a sequence of P known patterns by means of the optimisation of the synaptic weights \bar{w}, \bar{v} . The complete data set is divided in two different subsets. One, the Learning Set (LS), coincides with the patterns collection on which the weights are estimated by means of the minimisation of E_L . The generalisation skill of the NN is then evaluated applying the error E_T on the second subset of data, called the Testing Set (TS).

The expression for the E_T error is:

$$2) \quad E_T = \frac{1}{P_T} \sum_{m=1}^{P_T} \left(\sum_{i=1}^m |z_i^m - y_i^m| \right),$$

where P_T is the number of patterns present in the TS.

The method described in the present paper was based on a two-layered neural network (NN), in which the input layer was made of eight neurons, the hidden of ten, and the output of three neurons, as shown in fig.2

The input vector \bar{x}^{m+1} was defined as:

$$3) \quad \bar{x}^{m+1} = (x_{H1}^m, x_{H1}^{m+1}, x_{d1}^m, x_{d1}^{m+1}, x_{H2}^m, x_{H2}^{m+1}, x_{d2}^m, x_{d2}^{m+1})$$

Where μ is a time index, $x_{H1,2}^m$ and $x_{d1,2}^m$ are respectively the significant wave heights and mean wave direction at the time μ of two stations near the buoy whose series is to be simulated. Each of the 8 input neuron was connected to all 10 neurons of the first layer by the synaptic interaction (weights) $w_{(i,j)}$ with $i=1, \dots, 10$ and $j=1, \dots, 8$. In the same way each neuron of the hidden layer is connected with the three neurons of the output by the synaptic weights $v_{(h,k)}$ with $h=1, \dots, 3$ and $k=1, 2, \dots, 10$.

The values of the output vector $\bar{a}^{m+1} = (a_0^{m+1}, a_1^{m+1}, a_2^{m+1})$, are:

$$4) \quad \bar{a}_h^{m+1} = \mathbf{s}_2 \left(\sum_{k=1}^{10} v_{h,k} \mathbf{s}_1 \left(\sum_{j=1}^8 w_{k,j} x_j^{m+1} \right) \right)$$

with $h=1, 2, 3$ and $\sigma(x)$ is the non linear input-output function of the neurons, defined as follows:

$$5) \quad \mathbf{s}_i(x) = \frac{1}{1 + e^{-I_{i,x}}} \text{ with } i=1,2.$$

The components of the output vector \bar{a}^{m+1} are the coefficients of the linear combination of the vector $\bar{I}^{m+1} = (D_p, x_{H1}^{m+1}, x_{H2}^{m+1})$, where D

is an average value of the significant heights at the site 0 where the prediction is to be made:

$$6) \quad D_P = (\max(x_{H0}^m) - \min(x_{H0}^m))$$

The final output of the model z^{m+1} is the following scalar value:

$$7) \quad z^{m+1} = (\mathbf{a}_0^{m+1} D_P + \mathbf{a}_1^{m+1} x_{H1}^{m+1} + \mathbf{a}_2^{m+1} x_{H2}^{m+1})$$

which represents the best estimate for the unknown x_{H0}^{m+1} .

The NN learns how to evaluate the different H_{m0} contributions of the correlated stations at every time step m .

The E_L actually adopted in the application of the learning algorithm differs from (1), in fact $E_L(\bar{x}, \bar{y}, \bar{w}, \bar{v}, \bar{I})$ was defined as:

$$8) \quad E_L = \frac{1}{P} \sum_{m=1}^P (x_{H0}^m \| x_{H0}^m - z^m)$$

This particular form was chosen in order to give more emphasis to the events of greater magnitude, i.e. to calibrate the NN on storms rather than calm periods. To get the global minimum in the learning phase, the Monte Carlo Method was used in order to explore all the possible values of the free variables $\{w_{i,j}\}_{i=8,j=1}^{10}$, $\{v_{h,k}\}_{h=10,k=1}^{3}$, λ_1 and λ_2

belonging to a discrete bounded set. The Simulated Annealing (SA) algorithm was adopted as the more effective Monte Carlo method in the present conditions. The SA skill in determining the global minimum result was theoretically stated in the Geman and Geman theorem (Geman et Al. 1984) and checked numerically many times (Aarts et al, 1990.). A problem with the SA method, sometimes cited in literature (for example Mhaskar 1996), deals with the algorithm convergence velocity. Nevertheless, in the present work the convergence velocity was found to be high enough to operate efficiently. One of the principal reasons for the use of the SA is the stability with respect to the choice of random initial conditions. This aspect was carefully tested during the application of the method.

After the learning phase, in the following testing phase, a different set of data was used to evaluate the error E_T .

Test and preliminary results

The NN method was applied to simulate the time series observed at Monopoli during the first ten years of activity. The learning phase was set on by means of the first 3000 available patterns (which correspond to the first year of data, since the 1st of June 1989, of the buoys of Crotona, Monopoli and Pescara). In the learning phase the weights in the expression (3) were evaluated, then in the test phase the whole set of data of Crotona and Pescara were used to simulate the time series at Monopoli until December 1999. Overall statistics indicate as average learning error to be $E_L=0.021m$, while the average testing error was $E_T=0.023 m$. The closeness of the E_T to the E_L strongly suggest that the learning phase was successful in determining the behaviour of the system, so any further increase of dimension of the learning set would slow down the convergence process without any significant enhancement in the result accuracy. In order to assess the reliability of the results, i.e. to ascertain that simulated time series can trustfully be used to recover large gaps in the original time series of observations in real-time monitoring and statistical analysis, the three major storm activity periods were selected for a direct comparison with observations at Monopoli. A further comparison was introduced with the numerical simulation given by the deterministic 'physical' Wave Model (Komen et al, 1994) operationally run at the European Centre for Medium Range Forecasts (ECMWF). For the comparison, the nearest grid point to Monopoli was singled out from the Mediterranean 0.25/0.25 degree high resolution ECMWF WAM analysis. The obtained 6-hourly numerical time series (H_{WAM}) is consistent (in the sense that it is meaningfully comparable) with the 3 hourly observations averaged over

30 minutes, as can be verified by the comparison of the averaging period of the measurement with the time scale T_s which characterises the dynamics of the model. The T_s can be roughly estimated by the time it takes a physical signal travelling with typical speed of 10 m/s to cover the model grid length at 40° of latitude. T_s estimates for high-resolution models give times not greater than 80 minutes, being of the same order of magnitude of the averaging periods of measurement.

1st comparison period: 28.12.94-10.02.95

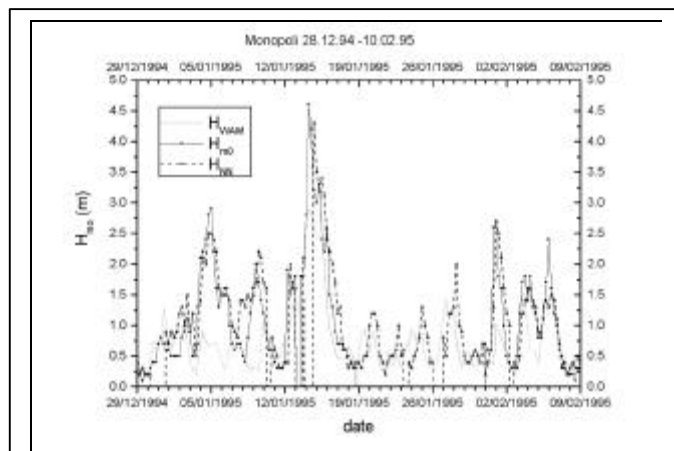


FIG.3: 1ST COMPARISON PERIOD BETWEEN OBSERVED SIGNIFICANT HEIGHT (H_{m0}) AND NUMERICAL EVALUATION BASED ON NN (H_{NN}) AND WAVE MODEL (H_{WAM}) AT MONOPOLI

The stronger sea storm (H_{m0} peak =5 m) of the 10 years of Monopoli buoy's activity was observed in the period considered. Less intense episodes were observed respectively one week before and one month later the peak. It can be seen in fig (3) that there is a gap in the observed time series (H_{m0}) ranging from immediately after the extreme peak to the next (smaller) one. It may lead to wander whether significant episodes were hidden in the analysis by the gap. The comparison between numerical WAM simulation (H_{WAM}) and the NN simulation (H_{NN}) showed that NN performs much better than WAM in all the period considered. This is not really a surprise, as it is well known that WAM outputs (especially before 1996, when significant improvements were made at the ECMWF) in the Mediterranean Area, and in particular in the Adriatic Sea, tends to underestimate H_{m0} . It is perhaps worth to mention that all three episodes were correctly reproduced by NN, but during the strongest peak there was a significant loss of data at the nearby station of Crotona. This caused the relative failure of the NN in predicting the exact maximum at the real time. The H_{NN} maximum was then shifted 3 hours forward ($H_{NN}=4 m$), when the actual measurement H_{m0} was significantly smaller ($H_{m0}=3 m$). Nevertheless, taken as a whole the episode was the best possible simulation of the event (considered the simulated hypothetical operational conditions: nor information from Monopoli, and nor from Crotona during the storm rise. What really happened was that at least one of the two was always working in the period considered).

2nd comparison period: 15.03.95-8.04.95

Two distinct storm episodes were observed in the period: the first one ($H_{m0}=3m$) occurred around the 20th of March, and the second, more severe ($H_{m0}=4m$), at the end of the month. Both episodes, shown in Fig (4), were very well reproduced by NN, differences from the H_{m0} being within the order of centimetres at the peaks. The H_{WAM} was found again to underestimate systematically the event. As in the previous comparison period it can be seen that a part of the series could not be reproduced by NN due to a minor loss of data in the leading edge of the

higher peak curve. Nevertheless, in the present case it didn't affect significantly the results.

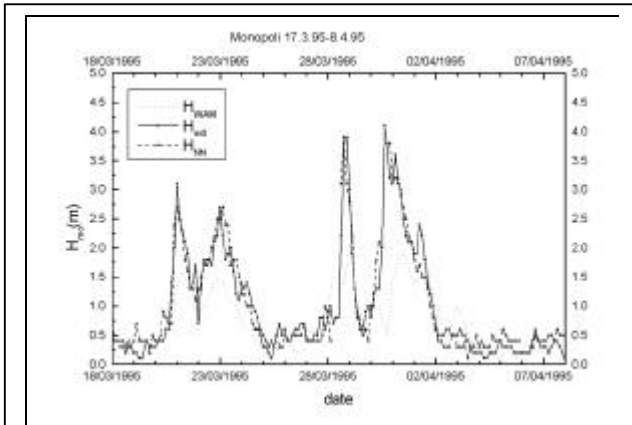


FIG.4: 2ND COMPARISON PERIOD BETWEEN OBSERVED SIGNIFICANT HEIGHT (H_{MO}) AND NUMERICAL EVALUATION BASED ON NN (H_{NN}) AND WAVE MODEL(H_{WAM}) AT MONOPOLI

3rd comparison period: 15.02.97-15.03.97

Three episodes were observed in the period, (fig 5) the first occurring in the mid-January, the second (the extreme one) at the beginning of March, and the last one just after two weeks. The major feature to be observed is a neat overestimation (difference between peaks of about 1 m) of the H_{mo} by H_{NN} in the January episode. Further investigations will be carried out in order to assess whether the overestimate can be related to the gap in the time series of the nearby buoy which can be seen along the leading (rising) edge of the peak line, or just to an inadequate representation of the directional information in the learning phase. The higher peak is well reproduced by NN, giving a reliable representation for the descending tail of the curve, which was lost in the observations.

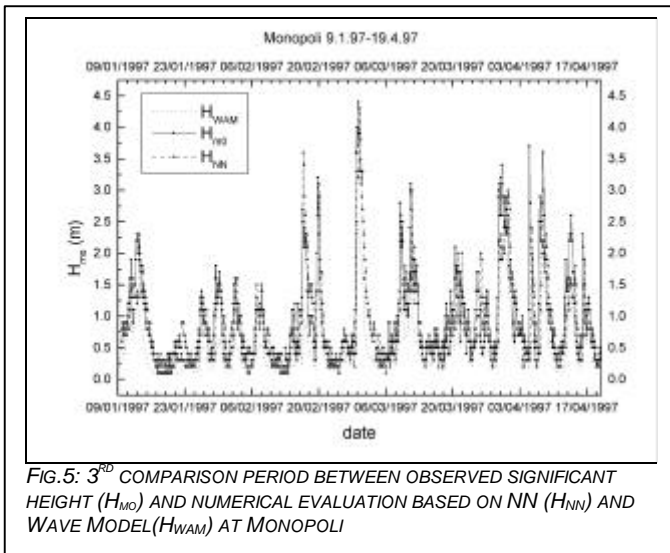


FIG.5: 3RD COMPARISON PERIOD BETWEEN OBSERVED SIGNIFICANT HEIGHT (H_{MO}) AND NUMERICAL EVALUATION BASED ON NN (H_{NN}) AND WAVE MODEL(H_{WAM}) AT MONOPOLI

Conclusions

The preliminary analysis carried out on the test cases proposed, showed that the numerical simulation of a very long time series, as the Monopoli time series, by means of the information collected in the nearby locations (Pescara and Crotone) processed by optimal NN algorithms, is at least promising in order to provide a method to recover

the effects of the loss of data of whatever long period. In the entire test periods considered NN were found to be far more effective in the simulation of the physical process than the numerical high resolution ECMWF Wave Model. Comparison with the observations showed that NN could reproduce most of the sea storm almost exactly in terms of the H_{mo} time series. Only one of the episodes considered in the three periods was 30% overestimated by NN, the failure being possibly related to the loss of nearby location data in the development stage of the storm. Minor weak points of NN were found to be the impossibility of data recovery when more than one nearby stations fails (situation that fortunately was seldom seen to occur) and the uncertainty about the mean wave directional information. It must be said that, despite the poorer ability of H_{WAM} simulation, WAM was found to be quite effective in the determination of the mean wave direction.

Further analysis is being carried out in order to assure the optimisation of the learning sets for all the SWAN time series. More complete results will indicate the effects of the recovery of the gaps currently present in time series in standard statistical analyses (i.e. wave climate) as well as in the analysis of extreme waves.

References

- E. Aarts and J. Korst, "Simulated Annealing and Boltzmann Machine", John Wiley & Sons New York, 1990.
- Arena G, Corsini S., "Activities of the National Hydrographic And Oceanographic Service in the maritime field", PIANC-PIC Congress, Venice, 1997.
- Corsini S., F. Guiducci, R. Inghilesi "Statistical Extreme Wave Analysis of the Italian Sea Wave Measurement Network Data in the period 1989-1999" ISOPE 2000 proc., Seattle.
- J. Herz, A. Krog, R. G. Palmer, "Introduction to the theory of neural computation.", Addison-Wesley, 1993.
- Geman S. and D. Geman, "**** IEEE Trans. on Pattern Analysis and Machine Intelligence, 6, p.721-741, 1984.
- Komen G. J., L. Cavaleri, M. Donelan, K. Hasselmann, S. Hasselmann, P.A.E.M. Janssen, "Dynamics and Modelling of Ocean Waves", Cambridge Univ. Press, 1994.
- Mhaskar H. N., "Neural Networks for optimal approximation of smooth and analytic functions", Neural Computation, 8, 164-177, 1996.
- Tirozzi B., "Modelli Matematici di Reti Neurali", CEDAM Milano, 1995.