Wedding Connectionist and Algorithmic Modelling towards Forecasting *Caulerpa taxifolia* Development in the North-Western Mediterranean Sea

Alex Aussem and **David Hill**
Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS)
ISIMA - Université Blaise Pascal, Clermont Ferrand II
Campus des Cézeaux - B.P. 125 63173 AUBIERE Cedex FRANCE
e-mails : alex@sp.isima.fr, benny@sp.isima.fr

Abstract - We discuss the use of supervised neural networks as a metamodelling technique for discrete event stochastic simulation in order to reduce significantly the computational burden involved by discrete simulations. A sophisticated computer model, coupling a Geographical Information System with a stochastic discrete event simulator, has been developed to anticipate the propagation of the green alga *Caulerpa taxifolia* in the North-Western Mediterranean sea. The simulation model provides reliable predictions, a couple of years in advance, of: i) the local expansion patterns of the alga, ii) the increase of *C. taxifolia* biomass and iii), the covered surfaces. However because the algorithmic model accounts for spatial interactions and anthropic dispersion/activities such as eradication, introduction of specific predators etc., simulations are extremely time and memory consuming. Therefore, to reduce the computational burden, a neural network was successfully trained on artificially generated data provided by the simulation runs to provide accurate forecasts 12 years in advance along with associated confidence intervals. The ability of the neural networks to capture the underlying physics of the phenomena is clearly illustrated by several preliminary experiments on a large coastal area. The neural network is able to construct, on this site, estimates of the *Caulerpa taxifolia* expansion 12 years in advance in good agreement with the simulation trajectories.

Keywords - Metamodelling, neural networks, discrete event simulation, *Caulerpa taxifolia*, invasive species.

1 Introduction

In 1984, the French coast of the Mediterranean sea near Monaco was the initial site of the development of *Caulerpa taxifolia*, a green alga of tropical origin introduced by mistake [4, 17]. Twelve years later, this species had colonized several thousand hectares of the French and Italian coasts and was detected in numerous places of the north-western Mediterranean coast, from Croatia (Adriatic sea) to the Balearic islands (Spain). This surprising development may locally induce an intense and irreversible alteration of the coastal ecosystems, such as endogenous species distributions (alga, cnidaria, sponges, echinoderms, fishes, etc.) as well as ecosystem functioning (trophic levels relationships). Indeed, the progressive elimination of the benthic flora and fauna is observed at stations heavily occupied by *C. taxifolia*. Though the social and economic repercussions of such damage on the ecosystems are considerable, the precise incidence on fisheries and tourism is difficult to estimate.

With a view to understanding the underlying rules governing the development of *C. taxifolia*, an interdisciplinary joint venture between marine ecologists, biologists and computer scientists was undertaken [12, 13]. They concurred on the observation that spatial effects like currents, spatial heterogeneity, fragments spreading etc., are major parameters influencing the colonization process and should therefore be integrated into a sophisticated discrete-event simulation model. The major goals are to provide quantitative results associated with the alga expansion several years in advance, e.g., biomass, production, contaminated surfaces and residual biomass of competing species (especially the protected seagrass *Posidonia oceanica*), given various environmental parameters, e.g., bathymetry, substrates and biocenosis on different sites [13]. To provide forecasts with sufficient accuracy, the model had to take into account spatial effects as well (heterogeneity of the sites, long and short distances interactions between organisms, etc.).

Unfortunately, one of the major drawbacks of discrete-event simulation is the amount of computational resources required to fully explore numerous possible trajectories of the ecosystem several years ahead according to distinct experimental scenarios, e.g., pattern of currents, anthropic dispersion, etc. Clearly, performing multiple replications for selecting optimal decision variables or policies becomes swiftly inhibitory when dealing with large scale ecosystems.

To supplement the computationally intensive (stochastic) discrete-event simulation software, we developed a so-called *metamodel*, based on supervised multilayer neural networks, to provide direct forecasts several years in advance in terms of covered surface levels [6]. Once trained with artificially generated data provided by the simulation runs, neural networks are shown to be reliable approximations of the underlying biological system that perform satisfactorily and are significantly more computationally efficient than the simulator itself. In addition, such techniques promise insights into the biological mechanisms that discrete simulation alone cannot provide.

This paper first presents the different modelling constraints which were kept in mind while designing the model, as well as the technical choices and the first results. Neural networks are then discussed and applied towards the prediction of the covered surface 12 years in advance.

2 The *Caulerpa taxifolia* simulation model

As discussed above, a sophisticated stochastic discrete-event simulation model was developed few years ago, in collaboration with marine ecologists, biologists and computer scientists, in order to better understand the parameters influencing the colonization of *Caulerpa taxifolia* along the French coast of the Mediterranean sea, and to explore distinct experimental scenarios [12, 13, 14]. The major goals were to provide quantitative results associated with the alga expansion several years in advance, given various environmental parameters. For completeness, we briefly detour to present the simulation model before we discuss the metamodelling approach and the experiments.

2.1 The *Caulerpa taxifolia* settlements

First, a thorough study of *C. taxifolia* settlement and development on precise locations was carried out to facilitate the validation of simulation results. Three settlement sites have been mapped yearly by marine ecologists since the beginning of 1990. All the experiments presented in this paper focus the first zone, namely Villefranche-sur-Mer. This site was mapped at a decametric scale. In spite of the absence of exhaustive mapping, enough data were available in that area to run simulations on a large scale [14].

As discussed above, spatial effects like currents, spatial heterogeneity, fragment spreading, etc., are considered major parameters influencing the colonization process and were therefore integrated in the simulation model. The modelling technique is based on discrete-event simulation used in the last decade for ecological modelling purposes. Consequently, the model is no only specified by a mathematical formalism but also by an algorithm describing the system functioning. This required the development of simulation software able to handle any littoral site by interfacing it with a Geographical Information System (GIS). Part of the simulation model input was initialized with digitized maps created in the GIS software MapGraphix. For each studied site, two maps were used : one for substrates and one for bathymetry. In the same way, the spatial simulation results are provided in GIS format in addition to traditional curves and statistics.

The outputs are composed of means along with confidence intervals obtained by bootstraping techniques [5]. Independence, common mean, variance and normal distribution of the responses were evaluated by a Kolmogorov test applied on results obtained from both 1 000 and 10 000 replications. With 1 000 replications, the Kolmogorov never rejected the null hypothesis, namely that the observed distribution is normally distributed, for a level of significance of 95% (see [13, 14] for more details).

2.2 Modelling elements

The simulation model relies on a set of holistic variables which are estimated and changed with the occurrence of discrete events. The site under study are divided into cells whose size is an important model parameter. The cell size varies from 16 cm^2 up to 370 m^2 depending on the scale of the studied site. Each cell possesses substrate and depth attributes (provided precisely by the GIS). The probability that *C. taxifolia* grows in each cell is linked to a set of evolution rules depending on (i) the depth, (ii) the kind of substrate, and (iii) the number of fragments arriving in the cell. The growth during a simulation session is based on a list of active cells (i.e., those containing *C. taxifolia* individuals). An exploratory approach allowed us to determine the best parameters for each site. Simulations are initialized from experiment files (with 98 parameters). The experiment files contain:

- simulation control parameters (duration, number of replications, GIS map specification etc.)
- model initial values (distribution rate of the fragments, current direction and strength, etc.).
- settlement probabilities as a function of bathymetry, substrate, season, etc.

- monthly parameters for stolon growth, spread of fragments, biomass, and degeneration.

2.3 Model validation

The calibration and the sensitivity analysis have demonstrated that the model is robust enough and the conceptual model is consistent. Numerous replications - up to 100 000- were done for verification purposes (the random number pseudo-period being long enough). The modifications of internal data flow never lead the model to exhibit abnormal behavior or biased statistical results. Under such conditions the model could be considered as *reliable* according to the definition in use in the simulation community.

Different techniques [6] were used for validation purposes : i) comparison of results from site to site, ii) confrontation with empirical knowledge of marine ecologists, and iii) graphic visualization and animation to make use of the human ability to comprehend spatial relationships.

For this kind of simulation study, the spatial auto-correlation is strong since *C. taxifolia* contaminated zones tend to form aggregated spots. However, from one replicate to the other, peripheral spots distribution is totally different without apparent correlation. Thus, a large number of replicates were carried out and the results combined to perform a discrete spectral analysis. This analysis is useful to point out areas which have a high probability of invasion by *C. taxifolia* although such plots are not always easily interpreted.

2.4 Computational burden

The basic objective of this modelling approach is the quest for a deeper understanding of the *C. taxifolia* expansion phenomena which manifests itself along the Mediterranean coast. Unfortunately, in order to better understand the parameters influencing the colonization process, an exploration of distinct experimental scenarios is required, which in turn translates into a great number of simulation replications. Moreover, many environmental parameters (bathymetry, substrates, biocenosis, etc.) on different sites, have to be varied according to distinct experimental scenarios, e.g., pattern of currents, anthropic dispersion, etc.

Unfortunately, one of the major drawbacks of discrete-event simulation is the amount of computational resources required to fully explore system responses. Under these circumstances, performing multiple replications for the selected values of the decision becomes swiftly inhibitory when dealing with large scale ecosystems, hence the idea of using a so-called metamodel to provide direct forecasts of the relevant variables with sufficient accuracy.

3 The metamodel

To overcome the limitations of discrete-event simulation models, researchers and practitioners in the environmental, management, industrial and production sciences have developed so-called *metamodels*, which are approximations that perform satisfactorily and are significantly more computationally efficient. Metamodelling techniques can be traced back to the early 70's, although the term was developed more recently (see [15, 16] and references therein). Applications in the literature mainly cover industrial and production applications, such as production planning and control, facilities storage and design and shop floor control [15, 19]. Most of these articles concur in the observation that metamodels are easier to manage and provide more insight than simulation alone.

The main issues in metamodelling are:

- the choice of the underlying functional form,
- the choice of the inputs and their corresponding response variables to be used,
- selection of the appropriate samples from the simulation to construct the metamodel,
- validation of the model.

3.1 Metamodelling principles

To define more explicitly a stochastic simulation metamodel, let $X_j, j = 1, \ldots, n$ denote the variables influencing the response, Y, of the physical system. Assume that the system is subject to some additive random fluctuations, then the unknown relationship between Y and the inputs X_j may be written as

$$Y = f(X_1, \ldots, X_n) + \gamma$$

where γ is the additive noise, with zero mean and is independent of the inputs X_j. It is well known that the minimum mean squared error predictor is then given by

$$\bar{Y} = E[Y|X_1, \ldots, X_n] = f(X_1, \ldots, X_n)$$

A simulation model usually integrates only the input variables which lend themselves to the observation,

$$\hat{Y} = g(X_1, \ldots, X_p) + \delta$$

where $p \leq n$ and δ is some additive noise, with zero mean and is independent of the inputs X_j, and represents the random fluctuation of the simulation model, which was shown, in our case, to be normally distributed.

Now the minimum mean squared error predictor for the simulation model is given by

$$\bar{\hat{Y}} = g(X_1, \ldots, X_p)$$

Let $\bar{\hat{Y}} - \bar{Y} = \epsilon_s$ denote the error term accounting for the simulation error, owing to the excluded variables and model mispecification.

Now, a metamodel, h, is a further simplification of the previous model, adjusted to output $\bar{\hat{Y}}$, the optimal simulation response. The metamodel may be written as

$$\hat{\hat{Y}} = h(X_1, \ldots, X_m)$$

where $m \leq p \leq n$. Since the metamodel is adjusted from the simulation model, let $\bar{\hat{Y}} - \hat{\hat{Y}} = \epsilon_m$ be some additional error accounting for the metamodelling error, owing to the further excluded variables and the metamodelling error of fitting the metamodel to the simulation model.

The identification of the optimal metamodel is the identification of the parameters of $h()$. In other words, the metamodel can be viewed as a simplified model of the simulation model acting as a surrogate for the study of the physical system. The metamodelling error is therefore the sum of three terms, namely $Y - \hat{\hat{Y}} = \epsilon_m + \epsilon_s + \gamma$.

Now, several metamodelling techniques exist, e.g. kernel-based regression models, neural networks, genetic algorithms, etc. As far as we are concerned, they are distinguished by their tradeoffs, e.g., between accuracy and computational expense, between local and global fitting techniques, that must be dealt with when developing a metamodel. The next section aims at introducing the reader to the very basic knowledge of neural networks, making it suitable to readers with interests in non-technical areas.

3.2 Neural Network Metamodels

It is difficult to present the connectionist network, especially when the audience consists of a mixture of meteorologists, and ecologists, some of whom will know a great deal about connectionist approaches to modelling and prediction, while the knowledge of others is limited in scope. However, we have tried to highlight the specificities in a brief review in which neural networks are analyzed as regression models.

In this paper, we experiment with a multi-layer neural network (also called preceptron or feed-forward network): a simple, well-known and in-depth studied input-output (i.e., static) model [20], with non-linear transfer functions, offering universal function approximation capability [7]. Such a model can accommodate a combination of continuous (usually interval-scaled) and discrete numeric variables, as will be the case in our experiments - even if, in practice, they have severely heterogeneous certainties [18]. Additionally, classical neural paradigms are global models, that is, a single neural model is trained to model the entire simulation response surface. This differs from polynomial regression metamodelling, where the regression surface is fitted locally. Finally, the parallel architecture provides robustness to incomplete or erroneous data sets and offers fault tolerant, real-time performance.

It is important to stress that neural networks are *deterministic* models, meaning that the optimal mean-squared predictor is nothing else than the *conditional mean* of the desired output, given the input. This intuitive result represents the closed-form optimal solution for a neural network trained using least squares. It should be understood, however, that a neural network of fixed size can only approximate the optimal function thus introducing an effective *bias* in the solution. Furthermore, training on limited data results in an effective *variance* over the possible converged solutions. A larger network has smaller *bias* but requires more samples to train and so has an effectively larger *variance*. This trade-off is typical in regression theory [8]. What makes neural networks better at finding solutions than splines for instance is a far more difficult problem to answer and requires thorough statistical understanding of the notions of *consistency* and *generalization*. But it is well beyond the scope and purpose of this paper to delve deeper into statistical theory of optimal predictors. Readers interested in more theoretical material are encouraged to consult the literature.

3.3 Measure of fit & confidence intervals

The training set and the test set of a supervised neural network are made up from a number of input-output patterns. The test set is used to assess the performance of the neural network. The target value is the estimate from the simulation of the conditional mean. We implicitly assume that the average system output is the correct answer, and thus any deviation of the simulation from the actual system is disregarded. On the other hand, for the research presented in this paper, two target outputs were used: the average simulation response and its estimated variance calculated over $n = 10$ simulation replications,

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2$$

Using a single network for simultaneous prediction of both targets can be regarded as detrimental rather than beneficial, because it creates added complexity, and thus increases the risk of overfitting, given the limited number of training patterns. Therefore, two distinct neural networks were used separately in this paper for the prediction of the mean $\hat{\hat{Y}}$ and the variance $\hat{\hat{\sigma}}^2$.

A standard measure of fit [1, 21] for the neural network is given by the *normalized mean-squared error*:

$$\text{NMSE} = \frac{1}{\sigma^2 N}\sum_{k=1}^{N}(\epsilon_m^k)^2 \tag{1}$$

where ϵ_m^k is the modelling error, i.e., the difference between the average simulation value, at iteration k, and the neural network prediction, and σ^2 is the empirical variance of the simulation values calculated over all the N training patterns. A value of the NMSE = 1 thus corresponds to predicting the unconditional mean.

In addition, to gauge the practical relevance of the neural network metamodels, the neural network forecasts were combined to form confidence intervals. Intervals took the form of:

$$Interval = \hat{\hat{Y}} \pm \frac{\hat{\hat{\sigma}}}{\sqrt{n}} t_{n-1,\alpha/2}$$

where $\hat{\hat{Y}}$ stands for the neural network forecast of the average simulation response; $t_{n-1,\alpha/2}$ is the Student t-value for a level of significance of $\alpha = 95\%$ and $n = 10$, the number of replications of the simulation [9]. Note again that the normality of the simulation responses was statistically tested on results obtained from more than 1 000 replications. Therefore, we implicitly considered that the 10 values above were drawn from a normal distribution. Simulation runs were independent so that the necessary statistical assumptions held. The confidence intervals constructed from the neural network predictions were compared to those constructed from the simulation responses, using the same formula.

3.4 Training set-up

A traditional multilayer perceptron trained with a standard back-propagation algorithm was used. Before we turn to the experiments, we detour to review two (very basic) techniques we found to be useful for networks to run optimally.

In an attempt to optimize the learning rate η automatically, we allow the η to vary in the course of training since the best values at the outset of training may not be so good later on. Our approach is to check whether the weight updates at each iteration actually decrease the cost function. If not then the process is overshooting and η should be reduced. On the other hand, if several steps in a row have actually decreased the cost function, then η is increased. It appears best in the literature [1, 3, 10] to increase η by a constant and to decrease it geometrically to allow rapid decay when necessary. This gives the overall scheme

$$\eta = \begin{cases} +\alpha & \text{if } \Delta E \leq 0 \\ -\beta\eta & \text{if } \Delta E > 0 \end{cases} \tag{2}$$

where α and β are appropriate time constants. In this paper, we took $\alpha = 10^{-3}$ and $\beta = 0.9$.

In order to preempt data overfitting, a topological reduction approach was adopted in these studies. The approach consists in pruning appropriately non-useful connections during training. This is achieved by giving each connection a tendency to decay to zero, such that

$$w_{ij}^{new} = (1 - \epsilon) w_{ij}^{old} \tag{3}$$

so that the connections disappear unless reinforced [10, 21]. For small w_{ij}'s to decay more rapidly than larger ones, we made ϵ dependent on w_{ij} by

$$\epsilon_{ij} = \frac{\gamma\eta}{(1 + w_{ij}^2)^2} \tag{4}$$

where γ was varied in the course of training. We found out that to it is useful to begin with $\gamma = 10^{-3}$ until performance reaches a minimum and then to decrease the value up to 10^{-4} so the error continues to decrease at a slow rate.

We just stress that more sophisticated and powerful topological reduction approaches exist. They consist in appropriately pruning (or penalizing) non-useful connections as training proceeds, preventing the network from overfitting the data. For further details, the reader is directed to the references.

4 Experiments

This section discusses some preliminary experiments performed with neural network meta-models. Given concerns about the rapid, intense and irreversible alteration of the coastal system, and the computational burden of discrete-event simulation over periods spanning several months, we mainly focused on predictions of *C. taxifolia* contamination on the long-term (12 years in advance). In this regard, our objective was to forecast the *C. taxifolia* expansion in terms of undersea contaminated surface over a 12-year period according to most of the available biological parameters.

4.1 Contaminated surface analysis

Visual inspection of the simulation runs is an important first step towards intuitively understanding how neural networks could help in modelling and forecasting the *C. taxifolia* expansion. One way to assess this effect of the initialization parameters is by visually inspecting different plots obtained after a simulation run. Therefore, we have plotted the most informative curve, namely the contaminated surface versus time, with plausible parameters, Fig.1. Visual inspection of the same curve but in log-scale, Fig.2, clearly reveals yearly growing cycles with rapid increases during summer and stagnation during winter seasons. Indeed, the *C. taxifolia* expansion requires light and heat. A second and maybe more subtle point, is that the alga expansion exhibits three distinct regimes: i) during the first 50 months, a regular growth is observed, ii) then, a temporary stagnation occurs, corresponding to the bathymetric limit and the coastal edge, and iii) a new expansion phase takes place, not as strong as the first though, corresponding to the expansion along the coast and in deep water. An easily interpreted plot of *C. taxifolia* expansion shows the variation rate $(x_t - x_{t-1})/x_{t-1}$ versus time (Fig.3). The three regimes are more apparent. As expected, the biomass increase is more or less proportional to the covered surface, therefore, we limit ourselves to the analysis and prediction of the contaminated surface.

To gauge the effects of the *C. taxifolia* microscopic description parameters on the subsequent evolution of the contaminated surface, some trial replications were run. The simulation initialization parameters were all arbitrarily varied within biological plausible ranges. Interestingly, no sensible change on the global shape of the curve, representing the contaminated surface coverage versus time, was observed.

Consequently, instead of trying to reconstruct the overall curve by performing iterated predictions several steps ahead, we decided first to forecast the final value and then fit a prototype curve between the first and the final points.

4.2 Input parameter selection

The objective being set, we now proceed to the selection of the relevant parameters influencing the contamination process.

A relevant representation can make useful information explicit and strip away obscuring clutter. Different representations can be equivalent in terms of expressive power but differ dramatically in the efficiency to solve problems [1]. Also, the more parameters to be considered, the more parameters are required by the regression model, and thus the higher the risk of overfitting. Therefore, it is best for a small number of input variables to be inserted at the input of the neural network.

According to the experimental knowledge of the ecologists we are collaborating with, a very restricted subset of parameters was selected from the overall simulation initialization parameters for the prediction of the final contaminated surface level, namely

- the ground projection surface of the cuttings,
- the ground projection surface of the stolons,
- the yearly maximum growing rate of *C. taxifolia*,
- the number of new cuttings per year.

Table. 1 shows their respective plausible range of variation. The key parameters being identified, we now proceed to the training set construction.

4.3 12-year contaminated surface forecasts

Coming up with good data sets addressing parameters of interest is tricky. The practical experience of the ecologists has limitations, which points to the inherent difficulty of selecting the best training set. Therefore, a uniform discretized lattice over the allowable ranges was used for the selection of the neural network training tuples. Afterwards, these values were linearly range-normalized onto the interval $[0, 1]$.

Insufficient time and resources were the major stumbling blocks. For research reported in this paper, the number of replications per training pair was kept constant. Due to the expensive computational effort incurred in conducting each simulation run for obtaining 12-years contaminated surface values (from a few minutes up to a few hours for a single simulation run on a bi-processsor IBM work station), only 10 replications per input tuple were carried out. A total of 6 000 simulation replications were run.

The contaminated surface 12-year forecats obtained by simulation were not found to vary drastically over the training tuples. Also instructive is the following observation: although each simulation run - for the same input parameters - yields a somewhat distinct trajectory, they all converged to the an almost identical contaminated surface value (the variance is negligible, below 10^{-2}).

Finally, the data set is made up from 600 pairs consisting of the 4 selected input parameters and the two target values, namely the average contaminated surface level 12 years in advance, provided by the simulation software, and its variance. Both targets were afterwards range-normalized in the unit interval.

To assess the generalization ability of the metamodels, 5 distinct training + test sets were constructed. The training set (strictly speaking) were made up from 80% randomly selected patterns and the test set made up from the 20% remaining patterns for validation purposes.

4.4 Intermediate results

Preliminary tests have indicated diminishing returns beyond five nonlinear units for both tasks. The two networks used to learn the series have a total of 26 links. The networks were stopped when the minimum of the mean-squared error in the training set was reached and their performance were afterwards tested on the test set.

Predictions of the average surface on the test set are shown in Fig. 4 for a particular test set. Results were far above our expectations. As can be seen, the task was perfectly learned by a relatively small neural network. The 5 independent trainings performed with the 5 data sets yielded the same result, in terms of NMSE over the test set, here $10^{-2.7}$ The predictions did not degrade from one test set to the other, indicating very good generalization ability. Predictions of the variance (not shown) yielded comparable performance with a NMSE=$10^{-2.4}$.

4.5 Reconstructing the overall curve

With a view towards reconstructing with reasonable assuredness the overall curve - and not only the final value -, all 6 000 surface curves obtained by intensive simulation runs were normalized and afterwards averaged to obtain a *prototype* curve. Unfortunately, our attempts to reconstruct the overall curve by stretching out the prototype curve up to the predicted final surface point, were not as successful. Indeed, as discussed previously, three distinct regimes are observed: existence of these regimes is closely related to the bathymetry.

For the same input parameters, the 3 regimes occur at the same time. Also, we decided to forecast the month, the magnitude and its variance of the two breakpoints (i.e. the X-coordinate, the Y-coordinate and $Var(Y)$) using the same 4 input parameters. With this in mind, 6 training + test sets were constructed from 600 patterns each. Six distinct neural networks were trained separately in order to infer the exact location of these two breakpoints.

The resulting networks had between 20 and 50 links including bias. No more than five nonlinear units were used. Selection of these dimensions were mostly by trial and error, along with various heuristics. In performing these preliminary simulations, we stress that we have made no effort to obtain an optimal architecture. Instead, we focused on these configurations and tried to obtain the lowest mean-squared error on the training set, and assessed the performance on the test set. Note that in general, selection of dimensions for neural networks remains a difficult problem in need of further research.

Final results are encouraging. The neural networks have learned to infer the month and the average surface with very good accuracy: NMSE varied between $10^{-1.6}$ and 10^{-3}. Predictions of the surface variance were not as good though: $NMSE = 0.2$ for the first point and 0.3 for the second. We believe that predicting the variance of the surface at these points depends on further parameters that were not taken into account in our approach.

Therefore, we constructed basic confidence intervals for each testing tuple, given the estimate of the average surface at the first point, at the second point, and at the last point 12 years ahead, and their variance. Again, simulations runs were independent and the normality of the responses were verified so that the necessary assumptions held.

To gauge the practical relevance of our metamodels, we ran the neural networks and the simulation model with randomly selected input parameters within their respective range (see Table.1) and compared the curves. Comparisons were done between the 95% confidence intervals provided by the simulator and the neural network metamodel. As seen in Fig.5, very encouraging results were obtained. The 95% confidence interval of the covered surface reconstructed from the neural network overlaps the confidence interval given by the simulation model. Significant reduction in computational time can thus be achieved without a sacrifice in prediction accuracy.

4.6 Discussion

Our approach may be criticized on two accounts. First, few replications were run for each input tuple: the number of replications (10) was set by balancing the computation cost with the desired accuracy of the estimate. This raises the question of whether our confidence intervals of the covered surface are statistically reliable.

Second, the results presented in this paper only relate to a single site. Though not discussed here, a topic of considerable interest would be to analyze the differences in terms of performance between several metamodels trained at different sites to find out whether the contaminated surface 12 year forecasts is also deterministically related to the same limited set of parameters. This needs further substantiation through more experiments and analysis. We are also interested in assessing the usefulness of the neural network metamodelling approach for other variables and *C. taxifolia* local expansion patterns.

Third, the training and generation of confidence intervals requires several replication of the same process. Therefore, because the actual data about the colonization process only reflect a single trajectory of the process (moreover over a period of time less than 12 years), actual data cannot be used for training purposes. As far as validation is concerned, the simulation model runs were already shown to be in good agreement with the observations (not only in terms of the contaminated surface but also the 2-dimensional maps). Actual data collected on different sites will be used when the neural network will take in to account bathymetric and substrate information in order to be able to extrapolate to new sites. Results will be reported in due course.

5 Conclusion

In this article we discussed work carried out with supervised neural networks, as a metamodelling technique for discrete event stochastic simulation, with a view to dramatically reducing the computational burden involved by the simulations. The studies currently being undertaken are aimed to be incorporated later into a prediction package for exploring numerous possible trajectories of *Caulerpa taxifolia*, a green algaprovoking intense and irreversible alteration of the coastal ecosystems, including endogenous species distribution.

The preliminary work presented in this paper addressed the prediction of the covered surface 12 years in advance. A neural network was successfully trained on artificially generated data provided by the simulation runs to provide reliable forecasts. The overall expansion curve was also reconstructed with reasonable accuracy. Future work will address other variables, such as expansion patterns at different sites.

6 Acknowledgments

We thank Cédrick Laballery, Jérôme Bart, Yannick Tranchier and Damien Azambourg of ISIMA/University of Blaise Pascal, Clermont-Ferrand, France, for their contribution to some ideas reported in this paper. We are greatful to James Paul Hoffman for careful reading of this manuscript. We also thank the anonymous referees for numerous suggestions that greatly improved the clarity of the manuscript. This work was supported by the European LIFE DGXI 95/F/A31/EPT/782 program and is now funded by the French Ministry of Environment (ref. MATE/98154).

Captions :

Fig.1: Contaminated (normalized) surface vs. time expressed in months.

Fig.2: Contaminated (normalized) surface in log scale vs. time expressed in months.

Fig.3: Contaminated surface growing rate, $(x_t - x_{t-1})/x_{t-1}$, vs. time expressed in months.

Fig.4: 12-year forecasts of the normalized surface contaminated by *C. taxifolia* on the test set.

Fig.5: *C. Taxifolia* expansion curves in log scale vs. time in months; in plain line, the 95% confidence interval output by the simulator, in grey line, the 95% confidence interval reconstructed from the neural network forecasts.

References

[1] A. Aussem, F. Murtagh and M. Sarazin, "Dynamical recurrent neural networks - towards environmental time series prediction," *International Journal on Neural Systems*, Vol. 6, No. 2, pp. 145-170, 1995.

[2] A. Aussem, "Nonlinear modelling of chaotic processes with dynamical recurrent neural networks," *Proceedings Neural Networks and Their Applications NEURAP'98*, Marseille, France, pp. 425-433, March 1998.

[3] A. Aussem, "Dynamical dynamical recurrent neural networks towards prediction and modeling of dynamical processes," to appear in *Neurocomputing*, 1999.

[4] T. Belsher, A. Meinesz, "Deep-water dispersal of the alga Caulerpa taxifolia introduced in the Mediterranean," *Aquatic Bontany*, Vol. 51, pp. 163-169, 1995.

[5] E. Bradley, "The jackknife, the bootstrap and other resampling plans", *SIAM*, no.38, 1982.

[6] P. Coquillard, D. Hill, "Modélisation et Simulation des Ecosystèmes", Masson, Paris, 1997

[7] G. Cybenko, "Continuous value neural networks with two hidden layers are sufficient," *Math. Control Signals and Systems*, Vol. 2, pp. 303-314, 1989.

[8] S. Geman, E. Bienenstock et R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, Vol. 4, no. 1, pp. 1-58, 1992.

[9] G. Gordon, "System Simulation", Prentice Hall, 1978.

[10] J. Hertz, A. Krogh and R. Palmer, *An Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1991).

[11] D. Hill, "Object-Oriented Analysis and Simulation", Addison-Wesley Longman, UK, 1996.

[12] D. Hill, P. Coquillard, J. De Vaugelas, A. Meinesz, "An algorithmic model for invasive species application to caulerpa taxifolia (Vahl) C. Agardh development in the north-western Mediterranean sea," *Ecological Modelling*, Vol. 109, p p. 251-265.

[13] D. Hill, P. Coquillard, J. De Vaugelas, "Discrete-event simulation of alga expansion", *Simulation*, Vol. 68, no.5, pp. 269-277, 1997.

[14] D. Hill, "Modélisation du processus d'expansion : Application à *Caulerpa taxifolia*", *French Science Academy*, ed. Tec & Doc, pp. 219-230, 1997.

[15] R.A. Kilmer, "Artificial Neural Network Metamodels of Stochastic Computer Simulations", Ph.D. Dissertation, University of Pittsburgh, 1994.

[16] J.P.C. Kleijnen, "Statistical Tools for Simulation Practitioners", Marcel Dekker, New York.

[17] A. Meinesz and B. Hesse, “Introduction et invasion de l’algue *Caulerpa taxifolia* en Méditerranée nord occidentale”, *Oceanologica Acta*, Vol.14, no.4, pp. 415-426, 1991.

[18] F. Murtagh, J. Cambpell, G. Zheng, A. Aussem, M. Ouberdous, E. Demirov, W. Eifler and M. Crepon, “’Data imputation and nowcasting using clustering and connectionist modelling”, *IFCS-98, International Federation of Classification Societies Biennial Conference*, Springer-Verlag, Rome, July 1998 and *Compstat’98, International Conference on Computational Statistics*, Bristol, August 1998, Springer-Verlag, in press.

[19] H. Pierreval and R.C. Huntsinger, “An investigation of neural network capabilities as simulation metamodels” *Proceedings of the 1992 Summer Simulation Conference*, pp. 155-162, 1992.

[20] D.E. Rumelhart, G.E. Hinton and R.J. Williams, “Learning internal representations by error propagation,” in *Parallel Distributed Processing Explorations in the Microstructure of Cognition*, D.E. Rumelhart and J.L. McClelland, Eds., Cambridge, MA: MIT Press, Bradfords Books, vol. 1, pp. 318-362, 1986.

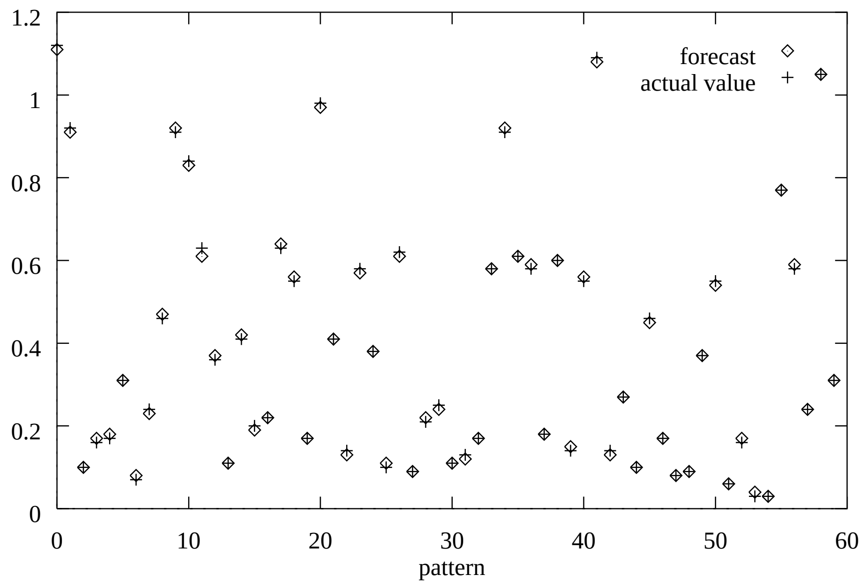
[21] A.S. Weigend, D.E. Rumelhart et B.A. Huberman, “Predicting the future: a connectionist approach,” *International Journal of Neural Systems*, Vol. 1, pp. 195-220, 1990.

[22] P. Werbos, *Beyond Regression: New Tools for Prediction and Analysis in Behavioral Sciences.* Ph.D. Dissertation, Harvard University, Cambridge, MA, 1974.

[23] G. Zheng, S. Rouxel, A. Aussem, J. Campbell, F. Murtagh, M. Ouberdous, E. Demirov, W. Eifler and M. Crepon, “Forecasting of ocean state using satellite-sensor data”, in F. Murtagh, J.G. Campbell and P. Mc Kevitt, Eds., *Eighth Ireland Conference on Artificial Intelligence (AI-97)*, Vol. 1, University of Ulster, pp. 234-240, 1997.

parameters \ range	default	min	max
cutting ground projection	10 cm^2	3 cm^2	18 cm^2
stolon ground projection	O.5 cm^2	0.2 cm^2	0.8 cm^2
max. annual growing rate	300 %	200%	500%
Num. new cuttings per year	2	1	3

Table 1: selected parameters and their corresponding range.



1.2
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
pattern
forecast
actual value

Month

Month

Month

Contaminated surface coverage

Normalized surface

Log-normalized surface

$(x_t - x_{t-1})/x_{t-1}$

Pattern

Predictions on the test set

Month

Log-normalized surface

Confidence intervals