Aquat.sci.62 (2000) 350–363
1015-1621/00/040350-14 $ 1.50+0.20/0

Aquatic Sciences

Gonadosomatic index estimates of an introduced pumpkinseed (*Lepomis gibbosus*) population in a Mediterranean stream, using computational neural networks

Juan Carlos Gutiérrez-Estrada[1,*], Inmaculada Pulido-Calvo[1] and José Prenda[2]

[1] Departamento de Ciencias Agroforestales, Universidad de Huelva, Campus Universitario de La Rábida, 21819 Palos de la Frontera (Huelva), Spain, e-mail: ipulido@uhu.es
[2] Departamento de Biología Ambiental y sawd Pública, Universidad de Huelva, Campus Universitario de La Rábida, 21819 Palos de la Frontera (Huelva), Spain, e-mail: jprenda@uhu.es

Key words: Fish biology, GSI, artificial intelligence, biological prediction, Guadalquivir basin.

ABSTRACT

In this paper, we propose an alternative method to predict the Gonadosomatic Index (GSI), based on a technique known as computational neural networks (CNNs), with two main objectives: (1) to develop a quick and reliable method for the prediction of the fish reproductive period under variable environmental conditions, and thus (2) to reduce the field sampling and laboratory efforts. Three different neural architectures (5-6-6-1, 5-8-8-1 and 5-10-10-1), whose 'training' was carried out controlling three threshold determinism coefficients (R_t^2: 0.7, 0.8 and 0.9), were trained to estimate the GSI of an introduced pumpkinseed (*Lepomis gibbosus*) population inhabiting a highly fluctuating Mediterranean stream in southern Spain. This GSI estimate was made using several easily measured fish and environmental variables. The correlation (R) between the GSI observed (GSI_r) and the GSI predicted by the CNN (GSI_e) was very high (>0.8 in all cases). The optimal CNN structure was the 5-6-6-1 with $R_t^2 = 0.8$ because it produced the best generalization of the confidence limits of GSI_e with respect to GSI_r. To compare with traditional multiple regression analysis, we submitted the data to the same process as with CNN. The validation of the regression model produced a much lower correlation (R) than the CNN models. As an example of the predictive capacities of the CNN models, we predict the hypothetical pumpkinseed reproductive cycle of our population but under the environmental conditions found in the Camargue marshes (South France).

Introduction

The gonadosomatic index (GSI) refers to the relationship between the gonad weight and the fish somatic weight (Wootton, 1991). The variation in GSI throughout an annual cycle usually indicates the beginning and end of the fish reproductive

* Corresponding author, e-mail: juanc@uhu.es

period, an important component of the fish's life history. To estimate precisely the GSI it is necesary to analyse a large number of specimens, and this process is commonly quite time-consuming. In multiple spawning species, GSI needs to be calculated for each reproductive period, due to the high variability of the main environmental parameters influencing the reproductive timing, such as temperature (Kaya, 1973; Burns, 1976), especially in the Mediterranean area (Encina and Granado-Lorencio, 1997).

A group of techniques known as Computational Neural Networks (CNNs) are considered as a powerful alternative to the regression models. They are nowadays being used in highly non-linear system modelling and function fitting. They have a great capacity to fit highly scattered data, far from normality. In addition, they have the advantage that they are not being based on any mathematical expression relating the dependent variable (here the GSI) with any other independent variable. Another main property of CNNs is their capacity to produce powerful models from very few data, thus providing reliable preditions (see Govindajaru, 2000, for a review and additional strengths of this method). It follows that it is possible to quickly predict the reproductive timing of a population, such as that of the introduced pumpkinseed (*Lepomis gibbosus*) in southern Spain, under new environmental conditions and thus to foresee their dynamics and, if necessary, to program precisely any control mechanism.

CNNs models are increasingly being applied in many research fields, usually providing highly satisfactory results. Hsu et al. (1995) developed CNNs to model the superficial runoff. They obtained much better results than those produced by the statistical models of ARMAX temporal series and also than traditional models. Guegan et al. (1998) predicted the riverine fish diversity patterns on a global scale by local river conditions. They used the potential of CNNs to deal with some of the persistent fuzzy and nonlinear problems that confound classical statistical methods in ecology. Also, in complex cases, the CNNs models are more efficient than multiple regression or discriminant analysis, as has been repeatedly observed in several fish ecology studies (Lek et al., 1995, 1996; Baran et al., 1996; Mastrorillo et al., 1997).

In this paper we apply the CNN approach to predict the GSI of an introduced pumpkinseed population in a Mediterranean stream in southern Spain from easily measured field data. Also, we determine the optimal structure of the CNNs by comparing the observed GSI data with those predicted by several alternative CNN models. To compare with traditional multiple regression analysis, we submitted the data to the same process as with CNN. Finally, we used the best CNNs model obtained to predict the hypothetical pumpkinseed reproductive cycle of our population but in other environmental conditions.

Material and methods

Field Sampling

The study area was a stream of the Guadalquivir River basin (Guadiato River, 37º N, 4º W) about 127 km long. This river runs throughout the northern part of Córdoba Province (Spain). The climate is meso-Mediterranean, with strong seasonal and interannual fluctuations.

A total of 1422 pumpkinseed (*L. gibbosus*) was collected by electrofishing in four sampling sites located in the middle reach of the stream, between March 1993 and September 1994. Specimens were obtained from weekly collections during the reproductive period (April–September) and monthly collections during the rest of the year. Samples were collected on 49 occasions. Fresh fish were transported to the laboratory (30 minutes away from sampling sites), where they were kept frozen until laboratory analysis.

Laboratory Analysis

In the laboratory, fork length (FL, mm) and fresh weight (FW, g) of each specimen was recorded. Also, several scales were removed from the left side between the beginning of the dorsal fin and the lateral line. Scales were washed and dry mounted between two slides and read with a microfilm reader. Scales were measured in arbitrary units (1 mm = 62 a.u., arbitrary units). These measures were subsequently used to determine the age of the fish following Bagenal and Tesch (1978).

Sex (male, female or juvenile) was determined by visual examination of the gonads. Female gonads were dried (24 h) to constant weight (± 0.1 mg) in an oven at 80ºC and weighed (G_d). After removal of gonads, specimens were eviscerated and similarly dried and weighed (S_d).

Temporal patterns in female gonad development were described using the gonadosomatic index (GSI_r) according to the formula (Wootton, 1991):

$$GSI_r = 100 \frac{G_d}{S_d} \quad (1)$$

GSI predictions using Computational Neural Networks (CNNs)

Computational neural networks (CNNs) are mathematical models inspired by the neural architecture of the human brain (Rumelhart et al., 1986). The model neuron or node is a simple non-linear unit. The neuron collects inputs from single or multiple sources and produces a simple output. Interconnecting many of these single neurons or nodes in a known layer configuration creates a model neural network. Each layer is made up of several nodes, and layers are interconnected by sets of weights (Fig. 1).

The nodes receive input from either outside the model (the initial inputs) or from the interconnections. Nodes operate on the input transforming it to produce an output. Given a set of inputs x_i, at the $i = 1, \ldots, n$ input nodes, the values are multiplied by the first set of interconnectios weights (W_{ji}) where W_{ji} is the connection from the *i*th node and *j*th node:

$$I_j = \sum_{i=1}^{n} x_i W_{ji} \quad (2)$$

In this work, linear output nodes are used and the transformation associated with each node is a sigmoid function in the hidden layers, where y_{pj} is the output of *j*th node from *p*th node:

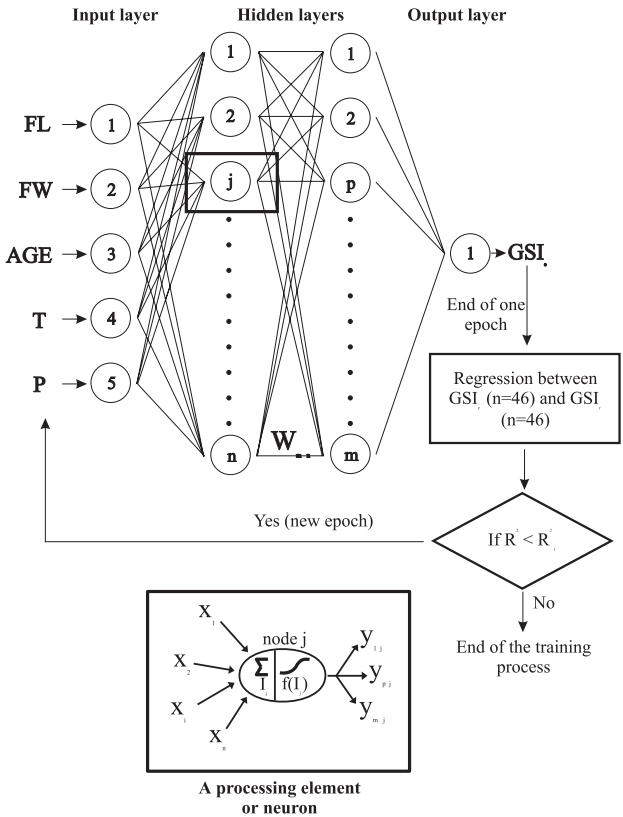

Figure 1. Four-layer feed forward computational neural network for computing GSI. Input variables: FL, fork lenght; FW, fresh weight; age; T, temperature and P, photoperiod. Output variable: GSI, gonadosomatic index. The input layer receive the input data and the nodes operate on these inputs transforming them to produce an output, i.e. the GSI_e. At the end of each epoch the GSI_r and the GSI_e are compared by linear regression. If the determination coefficient of this regression (R^2) is less than the threshold determination coefficient, a new epoch is carried out. If R^2 is higher than or equal R_t^2 then the training process will stop

$$y_{pj} = f(I_j) = \frac{1}{1 + e^{-I_j}} \tag{3}$$

The determination of the set of weights is a corrective-repetitive process called "learning' or "training". This "training" forms the interconnections between neurons, and is accomplished using known inputs and outputs ("training" set), and presenting these to the CNN in some ordered manner, adjusting the interconnection weights until the desired output is reached. The strength of these interconnections is adjusted using an error convergence technique so that a desired output will be produced for a given input. Three typical "training" pattern combinations are shown in Table 1. Six data points (five inputs and one output) compose each "training" pattern. In the CNN used for this study, inputs (similarly as independent variables in regression models) are: fish fork length (FL, mm), fish weight (FW, g),

age (years), air temperature (T, ºC, Trasierra Meteorological Station) and day-length (P, hours), and the output: GSI_r (the "dependent variable") (Fig. 1). The "training" method used is a standard back-propagation variation, proposed by Rumelhart et al. (1986), known as Extended-Delta-Bar-Delta (EDBD) (Minai and Williams, 1990).

The change in the weight $W_{ji}(t)$ on the link between *i*th node and *j*th node is calculated from the expression:

$$\Delta W_{ji}(t+1) = \alpha(t) \cdot \delta_{pj}(t) \cdot y_{pj}(t) + \mu(t) \cdot \Delta W_{ji}(t) \tag{4}$$

where:

$$W_{ji}(t+1) = W_{ji}(t) + \Delta W_{ji}(t+1) \tag{5}$$

where $\alpha(t)$ is a constant of proportionality called "learning" rate which controls the speed with which the algorithm converges, $\mu(t)$, namely momentum factor, is used to add a term to the weight adjustment that is proportional to the amount of the previous weight change, and $\delta_{pj}(t)$ is the error term. $\alpha(t)$ and $\mu(t)$ are adjusted similarly, according to the rules below. First, the $\bar{\delta}(t)$ (exponential average of previous gradient components at time t) is calculated:

$$\bar{\delta}(t) = (1-\theta) \cdot (\delta(t) + \theta \cdot \delta(t-1)) \tag{6}$$

The "learning" rate change for EDBD is:

$$\Delta\alpha(t) \begin{cases} \kappa_\alpha \cdot e^{(-\gamma_\alpha \cdot |\bar{\delta}(t)|)} & \text{if} \quad \bar{\delta}(t-1) \cdot \delta(t) > 0 \\ -\varphi_\alpha \cdot \alpha(t) & \text{if} \quad \bar{\delta}(t-1) \cdot \delta(t) < 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

and the momentum rate change is similarly,

$$\Delta\mu(t) \begin{cases} \kappa_\mu \cdot e^{(-\gamma_\mu \cdot |\bar{\delta}(t)|)} & \text{if} \quad \bar{\delta}(t-1) \cdot \delta(t) > 0 \\ -\varphi_\mu \cdot \alpha(t) & \text{if} \quad \bar{\delta}(t-1) \cdot \delta(t) < 0 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where κ_α is a constant "learning" rate scale factor, κ_μ is a constant momentum rate scale factor, γ_α is a constant "learning" rate exponential factor, γ_μ is a constant momentum rate exponential factor, φ_α is a constant "learning" rate decrement factor, φ_μ is constant momentum rate decrement factor and θ is the convex weighting factor. In this work: $\kappa_\alpha = 0.095$, $\kappa_\mu = 1$, $\gamma_\alpha = 0.1$, $\gamma_\mu = 0.05$, $\varphi_\alpha = 0.1$, $\varphi_\mu = 0.01$ and $\theta = 0.7$ (Minai and Williams, 1990).

The weights are updated after the display of each pattern. Epoch is the time period that encompasses all the iterations done after the display of all the patterns. As the CNN architectural complexity increases (more neurons in the intermediate layers), the time for processing an epoch will be longer. However, the total time employed in the "learning" process can be reduced, due to a reduction in the number of epochs used.

The answer of each neuron will be in the interval [0,1] for any input $(-\infty, \infty)$. This scaling prevents the node from collapsing. Thus, the neuron will always produce the same response (0 or 1). The scaling of the experimental data of which one makes use in this work is defined as:

$$V_b^* = \frac{V_b - V_{min,b}}{1.1 \cdot V_{max,b} - V_{min,b}} \quad (9)$$

where V_b^* is the scaled value (V_b) of variable b, $V_{min,b}$ is the minimum value of the b set and $V_{max,b}$ is the maximum value of the b set.

By dividing by 1.1 in (9), during the "training" the values are scaled in a smaller range than the [0,1] interval. Thus, during the validation we prevent that large unexpected values may produce inadequate responses in the network.

An important aspect of the CNN is the capacity to generalize, starting from examples. The generalization is the capacity of the CNN to provide a correct answer with patterns that have not been employed in their "training" (i.e. to produce a model from a subset of data and later on to make predictions from the remaining data set). In this study, "learning" was controlled by an arbitrary threshold determination coefficient (R_t^2: 0.7, 0.8 and 0.9). At the end of each epoch, the GSI_e (GSI predicted by the CNN) is regressed against the GSI_r. If the determination coefficient of this regression (R^2) is smaller than the threshold determination coefficient, then a new epoch is carried out. If R^2 is higher than or equal R_t^2 then the training process will end (Fig. 1). Thus, with this method we fit the original data to the predicted one as much as we want. On the other hand, this criterion does not overfit the network as would be the case if we employed the criteria $R^2 = 1$. This procedure has been originally developed here as an alternative to crossvalidation.

One individual female was selected at random from each sampling, thus using 46 observations (on three sampling dates no female was captured) for the "training" (see Table 1). The "training" is an iterative process by means of which the network progressively approaches the observed pattern that tries to emulate until it reaches the criterion established to stop (here the three threshold determination coefficients; R_t^2: 0.7, 0.8 and 0.9).

To check the generalization capacity, the remaining females (290 individuals corresponding to the total of captured females minus the 46 that were employed during

Table 1. Examples of several "training" patterns. Each one of them is composed of five data inputs (FL, fork length; FW, fresh weight; Age; T, temperature; P, photoperiod) and one output (GSI_r, real Gonadosomatic Index). The input data points are sequentially presented to the CNN until the desired output is reached

"training" pattern number	Input					Output
	FL (mm)	FW (g)	Age (years)	T (ºC)	P (hours)	GSI_r
1	97	20.79	2	13.8	12.29	3.49
2	77	10.98	1	24.6	15.02	8.38
3	129	58.03	4	18.5	12.21	3.42

the "training" process) were used. Using the data from these 290 individuals as input values for the trained network, 290 new predicted GSI values were obtained.

Although a CNN with a single hidden layer is a structure capable of identifying complex nonlinear relationships between input and output data sets, here only CNNs with two hidden layers have been used. This choice can be explained because *a priori* we do not know the type of relationship among the variables. And it can happen that with a single hidden layer the number of necessary intermediate nodes to reach a certain error is so high that its application is unapproachable in practice (Müller and Reinhardt, 1990). In this study, the number of hidden nodes employed in the CNN was 6, 8 and 10. The selection of these values was motivated by two factors: (1) the desire to train the CNN in a relatively short time period and (2) the need to provide enough working space within the structure of the CNN to allow adequate "learning" of the forecasting problem.

The calculation of the GSI_e has been carried out with the neural network simulation software Redgen V 1.0 (Gutiérrez-Estrada and Pulido-Calvo, unpublished) for Windows developed in MS Visual Basic® language.

Comparison with multiple regression analysis

To compare with conventional multiple regression analysis we submitted the data exactly to the same process as with CNN:

1. Model elaboration with six data from 46 females (dependent variable: GSI_r; independent variable: FL, FW, age, T, P) selected at random, one per sampling date. These data were the same as those used in the CNN model training.
2. Prediction of the dependent variable (GSI_e) from the regression model, using the remaining data set not employed in the model elaboration (n = 290).
3. Model validation after the regression of GSI_e obtained in point 2 and the GSI_r calculated from the 290 data not used in the model.

CNN model prediction example

We used the temperature and day-length regimes published by Crivelli and Mestre (1988) for the Camargue (South France), for the same months (March-February), to evaluate the capacity of the CNN selected model to predict the GSI under different environmental conditions. To evaluate the coherence of these predictions, we compared them with the real data, both from the population used in the model – the Guadiato- and from the GSI observed for the Camargue pumpkinseed population.

Results

Temporal pattern in GSI calculated from field data

Gonad development began at the end of March-early April and peaked approximately a month later in both 1993 and 1994 (mean ± 95% CL: May 1993: GSI =

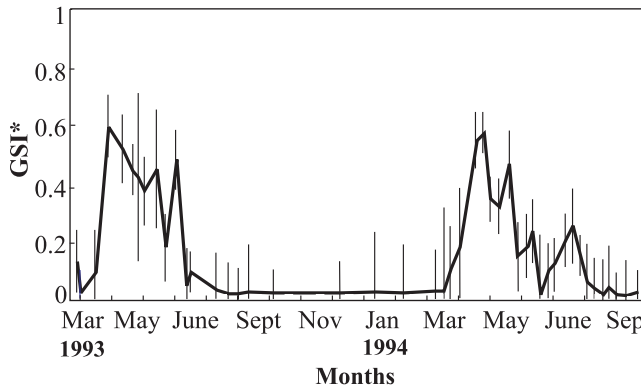

Figure 2. Seasonal changes in GSI_r (real Gonadosomatic Index; solid line) for female pumpkinseed of Guadiato River (S Spain). Solid line: mean values for samples of five or more specimens; vertical lines: 95% confidence intervals. GSI* are normalized values on a scale from 0 to 1

17.65 ± 3.19; April 1994: GSI = 16.95 ± 2.05) (Fig. 2). From then on, the GSI progressively dropped to a minimum at the end of August-early September, at which point reproduction ended and a quiescence period began that lasted almost 8 months (August-March). In July 1994, an increase in gonad development was observed (Fig. 2), probably related to a second reproductive event.

GSI_r neural network modelling and prediction capacity

During the CNN "training", nine models were obtained after the application of three neural structures and three threshold determination coefficients (R_t^2) (Table 2). The time occupied and the epoch number during the "training" period to obtain the CNNs were both dependent on the R_t^2 (Kruskal-Wallis statistic = 6.48, n = 9, $P < 0.05$). Both parameters significantly increased with R_t^2 (Table 2).

Table 2. Time and epoch number employed in the "training" process for three neural architectures and three threshold determination coefficients

Neural structure	Threshold determination coefficient (R_t^2)	Epoch number	Time (seconds)
5-6-6-1	0.7	341	126
	0.8	1046	378
	0.9	987	367
5-8-8-1	0.7	278	125
	0.8	770	323
	0.9	1601	709
5-10-10-1	0.7	472	261
	0.8	653	334
	0.9	1131	597

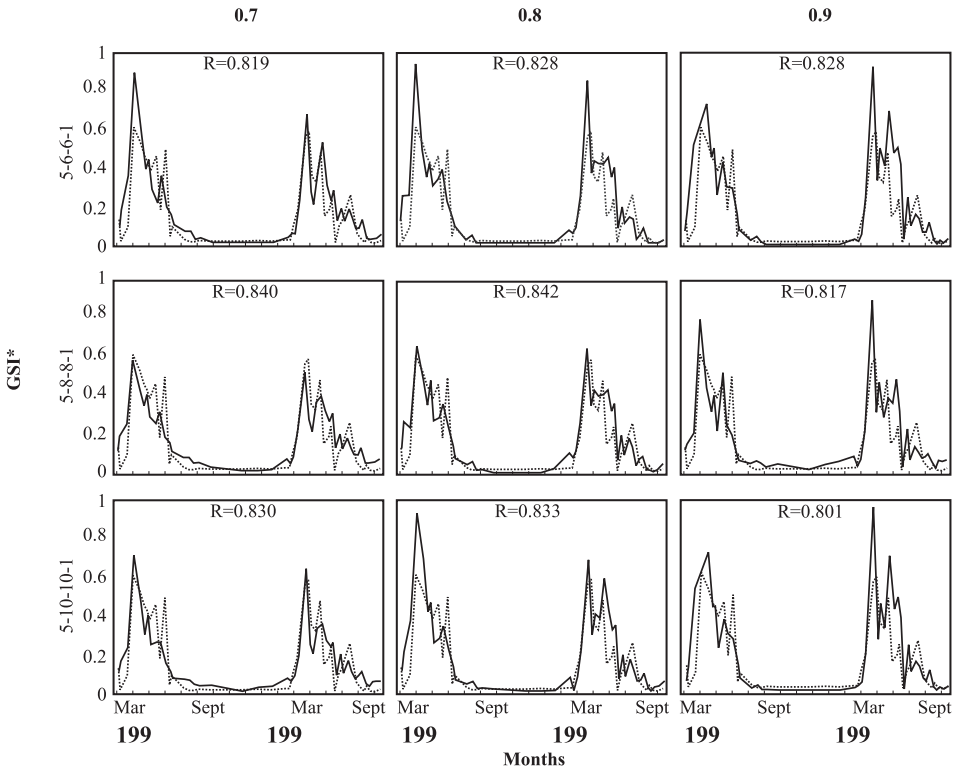

Figure 3. Generalization capacity of the nine CNNs obtained: comparison of the seasonal changes of the GSI_r (real Gonadosomatic Index; dotted line) and GSI_e (estimated Gonadosomatic Index; solid line) estimated by the different CNNs (5-6-6-1, 5-8-8-1 and 5-10-10-1 and three R_t^2). R: correlation coefficient between the GSI_r and GSI_e. GSI* are normalized values on a scale from 0 to 1

To test the generalization capacity of the CNNs, the GSI_r and the GSI_e were correlated. The correlations found between the GSI_r and the GSI_e after the application of any of the nine CNNs were always higher than 0.8. Also, there were no differences in the correlations between the three neural structures (Kruskal-Wallis statistic = 1.17, n = 9, P = 0.56) nor between the three R_t^2 (Kruskal-Wallis statistic = 3.95, n = 9, P = 0.14) (Fig. 3). Thus, based on the correlation between the observed and predicted data, the nine models were similarly good. An alternative way to identify the best model was to compare the dispersion of the predicted data around their mean values for each of the nine models. There were some statistically significant differences between the real confidence limits (CL_r) and the confidence limits predicted after the CNNs (CL_e). From the nine two-sample comparisons, the CNN 5-6-6-1, with $R_t^2 = 0.8$ was chosen due to the smallest value provided (Table 3).

Table 3. Two sample comparisons of the real and estimated mean 95 % confidence limits for three neural architectures and three threshold determination coefficients (R_t^2)

Neural structure	R_t^2	t value
5-6-6-1	0.7	6.7**
	0.8	– 0.2
	0.9	0.9
5-8-8-1	0.7	9.5**
	0.8	3.8**
	0.9	0.4
5-10-10-1	0.7	5.8**
	0.8	2.5*
	0.9	– 1.4

* $0.001 < P \leq 0.05$; ** $P < 0.001$.

Comparison with multiple regression models

The model obtained after the multiple regression had an $R = 0.77$, $P < 0.001$, $n = 46$. The coefficients for only two independent variables, T and P, were both statistically significant ($P < 0.05$) (Table 4). It must be remembered that in the CNN model we predetermined the fitting level (0.7, 0.8 or 0.9).

The model validation (GSI_e and GSI_r regression) produced a correlation coefficient $R = 0.63$ (Table 4), lower than the one obtained for the CNN model ($R = 0.83$).

Table 4. (A) Multiple regression model between several independent variables [fish fork length (FL, mm), fish weight (FW, g), age (years), air temperature (T, ºC) and day-length (P, hours)] and the GSI_r, as the dependent variable. (B) Validation of the model, after the regression between GSI_r and GSI_e

(A)	*Regression summary*			
	$R = 0.77$, $F = 12.02$, $P < 0.001$, $n = 46$			
	dependent variable	independent variable	b	P
	GSI_e	Intercept	– 34.17	0.015
		FL	– 0.07	0.966
		FW	– 0.10	0.547
		Age	2.31	0.109
		T	– 0.67	0.000
		P	3.89	0.000
(B)	*Regression summary*			
	$R = 0.63$, $F = 187.04$, $P < 0.001$, $n = 290$			
	dependent variable	independent variable	b	P
	GSI_e	Intercept	0.28	0.590
		GSI_r	0.99	0.000

Application of the selected CNN to new situations to predict the GSI

As an example of the potential predictive capacity of the CNN, we used the best CNNs model to predict the hypothetical pumpkinseed reproductive cycle from the Guadiato River population but under the temperature an day-length conditions observed in The Camargue (South France) (Fig. 4). As we know the real GSI, both from the population used in the model – the Guadiato- and from The Camargue pumpkinseed population (Crivelli and Mestre, 1988), the coherence of the predictions can thus be easily evaluated (Fig. 4).

Comparing the three GSI patterns shown in Fig. 4, three significant aspects can be observed: (1) a displacement of the GSI maxima from April (Guadiato population) to May-June (Guadiato population under the Camargue environmental regime). In the Camargue, the GSI reaches a maximum at the end of the reproductive period. In the Guadiato, in contrast, such a maximum is reached at the beginning of the reproductive period (mid April). (2) The way the maximum is reached varies for the three cases: in southern Spain it occurs in just a few days; in the Camargue this process is very progressive and lasts almost six months; in our prediction for the Guadiato population but under new environmental conditions, the situation is intermediate. (3) The way the quiescent period is reached also varies in the three cases: in the Camargue the quiescence is suddenly reached after the September GSI maximum. In the Guadiato population GSI progressively diminishes from the middle of April with a partial interruption during June-July, after which there is a gonad reactivation, probably related to a second reproductive event. In the predicted situation, similarly as in the Camargue, there is no gonad reactivation, and consequently, no new reproductive period.

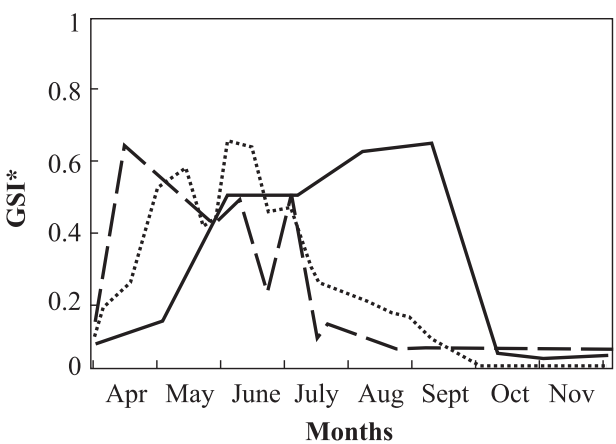

Figure 4. Prediction of GSI values under different environmental conditions. Solid line: seasonal changes in GSI_r for female pumpkinseed in the Camargue (Crivelli and Mestre 1988); discontinuous line: GSI_r for female pumpkinseed in Guadiato River (South Spain); dotted line: GSI_e for the Guadiato pumpkinseed population under the temperature and day-length regime of the Camargue. GSI* are normalized values on a scale from 0 to 1

Discussion

Pumpkinseed Reproductive Period

Fish gonad development is controlled by environmental factors (temperature, day length, etc.) (Kaya, 1973; Burns, 1976) in such a way that the spawning period coincides with the time in which the larval survival is maximum (Stacey, 1984).

In the pumpkinseed population of the Guadiato River, the length of the reproductive period was similar to other European populations of the same species (Crivelli and Mestre, 1988; Neophitou and Giapis, 1992). However, in southern Spain the GSI peaks in a few days, while in the other populations it takes between three and four months. This strong difference may be a consequence of both the sudden temperature increase usually observed in southern Spain at the end of the winter and the mild winter conditions (i.e. more food available) found in southern Spain with respect to colder situations observed in more northerly locations.

The application of Neural Networks to estimate and predict the GSI

The CNNs are very good predictors of GSI, as demonstrated by correlations higher than 0.8 between the observed and predicted data. It should be possible to reach still higher correlations during the modelling process, by the use of more neurons in the hidden layers and by the use of a larger number of "training" patterns. But this probably would greatly lengthen the "learning" process and the total time devoted to the model elaboration. The correlations obtained here can be considered as statistically very satisfactory and are higher than in other CNNs applications (Yang et al., 1997), and also higher than conventional multiple regression models.

The time and epoch number employed during the "training" process was clearly lower than those obtained by other authors (Ranjithan et al., 1993; Rizzo and Dougherty, 1994; Alvarez and Bolado, 1996). These differences can be explained by the type of parameter estimated, but also by the use in this work of a modification of the standard backpropagation model EDBD as the "learning" algorithm. This accelerates the effective "learning" process in certain directions.

A disadvantage of the standard model is that it can overfit the examples during the "training" process. However, the "training" method proposed here, based on threshold coefficients, was very efficient during the generalization, as the CNNs provided correct answers with data not used during the "learning".

The CNN 5-6-6-1 with $R_t^2 = 0.8$ was considered the optimal model because it produced the best fit to the observed confidence limits. The poorest estimates of the 95 % confidence limits were produced with the CNNs trained with $R_t^2 = 0.7$. This can be explained as an incomplete "learning" process, and implies that the estimated values are very similar to the pattern used during the "training" process. Similarly, it can be observed that all CNNs overestimate the maximum values, a probable consequence of the few "training" patterns available for these extreme data. Thus, a good CNN model should include a "training" data set with all common values, besides the most extreme elements that can be found.

The CNNs have important advantages relative to classical methods. The CNNs do not need any empirical relationship between the GSI and any other variable,

such as those included in this work (fish length, weight and age, temperature and day-length). This permits an adaptability of the model to a much larger data range, which in turn implies that the results will always be comparable to or better than those obtained with the best fitted classic model.

Another advantage of this technique (Rumelhart et al., 1986) is the possibility of CNNs to be used in real time control of different systems (McClendon et al., 1996; Seginer et al., 1996; Yang et al., 1997; Thirumalaiah and Deo, 1998). In the case presented here, once GSI is obtained from field data and the CNN trained accordingly, it is possible to predict the GSI in future years from different environmental conditions (e. g. temperature or day-length). This can also be a useful tool in the management of non-desired introduced species, such as the pumpkinseed in the Guadiato River.

Apart from the GSI estimates under variable environmental conditions, the CNNs can also be employed to estimate and predict other fish variables such as fecundity, egg-laying density, biomass, migratory timing, etc. But this alternative method is only advisable when the relationship between the variables is highly non-linear and the field or laboratory methods and more traditional statistical techniques prove to be laborious or do not provide satisfactory results.

One of the main problems associated to the CNNs is the identification of their parameters (Ranjithan et al., 1993; Xu et al., 1994) and a longer computing time than other alternative applications (Ranjithan et al., 1993; Rizzo and Dougherty, 1994).

Predicting GSI values under different environmental conditions

The result of the example in which the CNN model was used to predict the GSI for the Guadiato population under other environmental conditions was coherent. This points out that this sort of models can be good GSI estimators, although this has to be properly tested.

However, in our prediction two factors may have influenced the fitting of the GSI temporal evolution between the Camargue and the new scenario imposed on the Guadiato River population (the simulation). The one is the population structure, as both the age structure and the length-weight relationship were very different for both populations. The other is the habitat which varied greatly between the Camargue and the Guadiato River: a Mediterranean stream vs. a freshwater marsh. In the Guadiato River, the flow is highly variable, fluctuating between large autumn-winter floods and a summer dry period, when flow practically ceases.

ACKNOWLEDGEMENTS

Pieter Jelle de Visser van Bloemen, Dr. Emili García-Berthou and J. Palazón made valuable suggestions and comments on an earlier version of the manuscript.

REFERENCES

Alvarez, J. and S. Bolado, 1996. Descripción de los procesos de infiltración mediante redes neuronales artificiales. Ingeniería del Agua 3: 39–46.

Bagenal, T.B. and F.W. Tesch, 1978. Age and growth. In: T.B. Bagenal (ed.), IBP Handbook No. 3. Blackwell Scientific Publications, Oxford: Methods for assessment of fish production in fresh waters, pp. 101–136.
Baran, P., S. Lek, M. Delacoste and A. Belaud, 1996. Stochastic-models that predict trout population-density or biomass on a mesohabitat scale. Hydrobiologia 337: 1–9.
Burns, J.R. 1976. The reproductive cycle and its environmental control in the pumpkinseed, *Lepomis gibbosus* (Pisces Centrarchidae). Copeia 3: 449–455.
Crivelli, A.J. and D. Mestre, 1988. Life history traits of pumpkinseed, *Lepomis gibbosus*, introduced into the Camargue, a mediterranean wetland. Arch. Hydrobiol. 111: 449–466.
Encina, L. and C. Granado-Lorencio, 1997. Seasonal changes in condition, nutrition, gonad maturation and energy content in barbel, *Barbus sclateri*, inhabiting a fluctuating river. Env. Biol. Fish. 50: 75–84.
Guegan, J.F., S. Lek and T. Oberdorff, 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. Nature 391: 382–384.
Govindajaru, R.S., 2000. Artificial neural networks in hydrology. I: Preliminary concepts. J. Hydrol. Engrg. 5: 115–123.
Hsu, K., H.V. Gupta and S. Sorooshian, 1995. Artificial neural network modeling of the rainfall-runoff process. Water Resour. Res. 31: 2517–2530.
Kaya, C.M. 1973. Effects of temperature and photoperiod on seasonal regression of gonads of green pumpkinseed, *Lepomis cyanellus*. Copeia 2: 369–373.
Lek, S., A. Belaud, I. Dimopoulos, J. Lauga and J. Moreau, 1995. Improved estimation, using networks, of the food-consumption of fish populations. Mar. Freswat. Res. 46: 1229–1236.
Lek, S., M. Delacoste, P. Baran, I. Dimopoulos, J. Lauga and S. Aulagnier, 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Model. 90: 39–52.
Mastrorillo, S., S. Lek, F. Dauba and A. Belaud, 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river. Freshwat. Biol. 38: 237–246.
McClendon, R.W., G. Hoogenboom and I. Seginer, 1996. Optimal control and neural networks applied to peanut irrigation management. Trans. ASAE 39: 275–279.
Minai, A.A. and R.D. Williams, 1990. Acceleration of back-propagation through "learning" rate and momentum adaptation. Int. Joint Conf. Neural Networks I: 676–679.
Müller, B. and J. Reinhardt, 1990. Artificial Neural Networks: electronic implementations. Springer-Verlag, Berlin.
Neophitou, C. and A.J. Giapis, 1992. A study of the biology of pumpkinseed, *Lepomis gibbosus* L. in Lake Kerkini. Geotechnic Scientific Issue 3: 12–20.
Ranjithan, S.J., W. Eheart and J.H. Garrett, 1993. Neural network-based screening for groundwater reclamation under uncertainty. Water Res. Research 29: 563–574.
Rizzo, D.M. and D.E. Dougherty, 1994. Characterization of aquifer properties using artificial neural networks: Neural Kriging. Water Resour. Res. 30: 483–497.
Rumelhart, D.E., G.E. Hinton and R.J. Willians, 1986. "learning" representations by backpropagation errors. Nature 323: 533–536.
Seginer, I., Y. Hwang, T. Boulard and J.W. Jones, 1996. Mimicking an expert greenhouse grower with a neural-net policy. Trans. ASAE 39: 299–306.
Stacey, N.E. 1984. Control of the timing of ovulation by exogenous and endogenous factors. In: G.W. Potts and R.J. Wootton (eds.): Fish Reproduction: Strategies and tactics, Chapman and Hall, London, pp. 207–222.
Thirumalaiah, K. and M.C. Deo, 1998. River stage forecasting using artificial neural networks. J. Hydrol. Engrg. 3: 26–32.
Wootton, R.J. 1991. Ecology of teleost fishes. Chapman and Hall, London, 404 pp.
Xu, L., J.W. Ball, S.L. Dixon and P.C. Jurs, 1994. Quantitative structure-activity relationships for toxicity of phenols using regression analysis and computational neural networks. Environ. Toxicol. Chem. 13: 841–851.
Yang, C.C., S.O. Prasher, R. Lacroix, S. Sreekanth, N.K. Patni and Masse, L. 1997. Artificial neural network model for subsurface-drained farmlands. J. Irrig. Drain. Engrg. 123: 285–292.

Received 21 February 2000;
revised manuscript accepted 12 September 2000.