Vol. 140: 251–256, 1996 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published September 12

# Artificial neural network versus multiple linear regression: predicting *P/B* ratios from empirical data

**T. Brey[1,*], A. Jarre-Teichmann[2], O. Borlich[1]**

[1]Alfred Wegener Institute, Postfach 120 161, D-27515 Bremerhaven, Germany
[2]Institut für Meereskunde, Düsternbrooker Weg 20, D-24105 Kiel, Germany

ABSTRACT: Traditionally, multiple linear regression models (MLR) are used to predict the somatic production/biomass (*P/B*) ratio of animal populations from empirical data of population parameters and environmental variables. Based on data from 899 benthic invertebrate populations, we compared the prediction of *P/B* by MLR models and by Artificial Neural Networks (ANN). The latter showed a slightly (about 6%) but significantly better performance. The accuracy of both approaches was low at the population level, but both MLR and ANN may be used to estimate production and productivity of larger population assemblages such as communities.

KEY WORDS: Productivity · Benthic invertebrates · Artificial Neural Network

## INTRODUCTION

Somatic production of populations is an important component of energy flow and organic matter cycling in all ecosystems. The assessment of production, however, is time-consuming and expensive work, even at the level of a single population. Therefore, many attempts have been made to establish empirical relations between easy-to-obtain parameters and production, *P*, or the ratio of production to mean biomass, *B* (*P/B* ratio)

With respect to benthic invertebrates, Zaika (1970), Robertson (1979), Warwick (1980), Banse & Mosher (1980), Schwinghamer et al. (1986) and many others tried to predict the *P/B* ratio from 1 independent parameter. Their models depend solely on the negative exponential relation between metabolic rate and body mass in animals (see Schmidt-Nielsen 1984). Other authors, such as Plante & Downing (1989), Brey (1990), Edgar (1990) and Morin & Bourassa (1992), introduced more parameters, such as biomass, temperature and taxon, to improve the accuracy of the prediction.

All these authors used linear models to predict the *P/B* ratio, i.e. the data were transformed to achieve linear relations between the independent parameters and *P/B*. The main weakness of these multiple linear regression (MLR) models is that transformations include *a priori* assumptions about the type and consistency of the relation between 2 parameters which may not be met completely.

Artificial neural networks (ANN) are computer programs which are characterized by a massively parallel but highly interconnected architecture. They are able to learn and to generalize relations between input and output data from examples presented to the network. The strength of ANN is pattern recognition and pattern classification, but these programs can also be used for predictive purposes (Dayhoff 1990). ANN are used in various fields such as in industrial process control, speech recognition, financial market forcasts and chemical compound identification (Nelson & Illingworth 1991, Zupan & Gasteiger 1991) but are increasingly being applied to tasks in aquatic ecology as well (see e.g. Culverhouse 1992, Potter et al. 1993, French & Recknagel 1994). The main advantage in using ANN for prediction is that *a priori* assumptions about the relations between independent and dependent vari-

*E-mail: tbrey@awi-bremerhaven.de

ables are not necessary. However, those relations learned by an ANN are hidden in its neural architecture and cannot be expressed in traditional mathematical terms.

In the present paper we compare the performance of the 'classic' approach, MLR, and of ANN in estimating *P/B* ratios of benthic invertebrates from empirical data.

## METHODS

**Data.** This study is based on data from 899 unexploited benthic invertebrate populations collected by Brey (1996) from the literature (Table 1). Each data set consisted of an annual *P/B* ratio and 8 variables used to estimate *P/B* (Table 2). Three of these, mean water temperature (*T*), water depth (*D*) and mean individual body mass (*M*), were used as continuous variables, whereas the other parameters were transformed to categorical binary variables (0 or 1, see Draper & Smith 1981). To check the performance of MLR and ANN with a 'real world' example, we used 30 production data sets of Sprung (1993, 1994, M. Sprung pers. comm.; our Table 3) which were not part of our data base.

Table 1. Benthic invertebrate population data used in this analysis. Data set available on request from the first author (ASCII on 3.5" disk for Apple Macintosh)

| Taxon | No. of species[a] | No. of data sets |
|---|---|---|
| Mollusca | 138 + 3 | 297 |
| Polychaeta | 41 + 3 | 93 |
| Crustacea | 53 + 1 | 154 |
| Echinodermata | 38 + 0 | 51 |
| Insecta larvae | 124 + 92 | 304 |
| Total | 394 + 99 | 899 |

[a]No. of species + No. of populations identified to genus only

Table 2. Independent variables collected together with *P/B* ratio

| Variable group | Variables |
|---|---|
| Abiotic variables | Mean annual temperature *T* (K) |
| | Water depth *D* (m) |
| Biotic variables | Mean individual body mass *M* (kJ) |
| | Motility |
| | Vagile - Sessile |
| | Living |
| | Epifauna - Infauna |
| | Feeding |
| | Herbivorous - Omnivorous - Carnivorous |
| | Biotope |
| | Marine - River - Lake |
| | Taxon |
| | Mollusca - Crustacea - Polychaeta - Echinodermata - Insecta larvae |

**MLR.** We linearized the relations between *T*, *D* and *M* and the *P/B* ratio according to theoretical considerations and to empirical evidence. The relation between metabolic rates and body mass is exponential, as shown by many experimental and empirical studies (Schmidt-Nielsen 1984 and references therein). Therefore we used log-transformations for *P/B* and *M*. Metabolic rates show a positive nonlinear increase with temperature in aquatic poikilotherms, as expected from theoretical considerations and shown by many investigations (e.g. Ikeda 1985, Alongi 1990 or Clarke 1991). Linearisation can be approximated by transforming according to the Arrhenius equation with $\log(P/B)$ and $1/T$. Food is likely to be the limiting resource in many aquatic systems (Levinton 1982), and

Table 3. Annual production (*P*) and *P/B* ratio estimates for 30 populations (23 species) from Ria Formosa, Portugal (Sprung 1993, 1994, M. Sprung pers. comm.) computed by the increment summation method (ISM). The original ash-free dry mass data were converted to kJ using factors taken from the literature

| No. | Species | *P* (kJ $m^{-2}$ $yr^{-1}$) | *P/B* ($yr^{-1}$) |
|---|---|---|---|
| 1 | *Abra ovata* | 13.00 | 2.309 |
| 2 | *Corastoderma edule* | 40.24 | 4.810 |
| 3 | *Cerastoderma edule* | 42.42 | 5.697 |
| 4 | *Loripes lacteus* | 2.54 | 1.142 |
| 5 | *Scrobicularia plana* | 400.33 | 1.791 |
| 6 | *Scrobicularia plana* | 32.62 | 3.060 |
| 7 | *Tellina tenuis* | 8.70 | 1.875 |
| 8 | *Venerupis aureus* | 54.21 | 6.270 |
| 9 | *Amyclina corniculum* | 17.86 | 1.611 |
| 10 | *Bittium reticulatum* | 27.09 | 1.252 |
| 11 | *Bittium reticulatum* | 146.61 | 2.145 |
| 12 | *Cerithium vulgatum* | 14.08 | 0.848 |
| 13 | *Cyclope neritea* | 24.44 | 2.032 |
| 14 | *Haminea hydatis* | 30.07 | 1.702 |
| 15 | *Hydrobia ulvae* | 6.48 | 1.811 |
| 16 | *Mesalia brevialis* | 238.12 | 1.759 |
| 17 | *Cyathura carinata* | 0.46 | 1.802 |
| 18 | *Cyathura carinata* | 11.74 | 3.013 |
| 19 | *Idotea chelipes* | 0.77 | 3.778 |
| 20 | *Upogebia pusilla* | 3.98 | 5.090 |
| 21 | *Upogebia pusilla* | 57.86 | 3.121 |
| 22 | *Audouinia filigera* | 58.44 | 2.996 |
| 23 | *Audouinia filigera* | 11.21 | 2.433 |
| 24 | *Capitella* sp. | 6.34 | 1.636 |
| 25 | *Glycera convoluta* | 22.65 | 3.447 |
| 26 | *Melinna palmata* | 144.81 | 2.337 |
| 27 | *Nephtys hombergii* | 31.53 | 4.601 |
| 28 | *Nereis diversicolor* | 448.43 | 5.263 |
| 29 | *Nereis diversicolor* | 739.67 | 3 275 |
| 30 | *Terebella lapidaria* | 12.10 | 2.758 |
| | Community | 2648.74 | 2.649 |

Table 4. Binary variables used in multiple linear regression analysis. Pilot studies showed that the combination of 'Motility' and 'Living' as well as the re-grouping of 'Feeding' improved performance

| Group | No. | Binary variable | | | | |
|---|---|---|---|---|---|---|
| Motility & living | 1 | 1<br>Vagile & epifauna | 0<br>Others | | | |
| Feeding | 2 | 1<br>Carniv | 0<br>Omniv & Herbiv | | | |
| Biotope | 3 | 1 | 0 | 0 | | |
| | 4 | 0<br>River | 1<br>Lake | 0<br>Marine | | |
| Taxon | 5 | 1 | 0 | 0 | 0 | 0 |
| | 6 | 0 | 1 | 0 | 0 | 0 |
| | 7 | 0 | 0 | 1 | 0 | 0 |
| | 8 | 0 | 0 | 0 | 1 | 0 |
| | 9 | 0<br>Moll | 0<br>Crus | 0<br>Poly | 0<br>Echi | 1<br>Inse-La |

shortage of food seems to reduce metabolic rates (Parry 1983, Steen et al. 1991). In aquatic systems, the input of food into the benthic system decreases exponentially with water depth, therefore we used $\log(P/B)$ and $\log(D+1)$. Binary variables representing the remaining parameters were constructed according to Table 4.

**ANN.** We pre-transformed the continuous variables *T*, *D* and *M* and *P/B* ratio to achieve distributions as even as possible over the whole range of each variable using the Box-Cox algorithm (Sokal & Rohlf 1995). Pilot studies showed that continued 'flattening' of the distributions, by e.g. transformations based on standardised cumulative frequency dustributions, did not significantly improve the results.

We used 1 binary variable to represent both the groups 'Motility' and 'Living', because there are only 2 alternatives in either group (Table 2). All other variables were represented by 1 binary variable each, which resulted in a total of 13 binary variables. We used 'NeuralWorks Predict' by NEURALWARE to train multilayer backpropagation networks to predict *P/B* from the 16 input variables mentioned above. This software performs semi-automated data analysis, variable selection and network construction, using elements of fuzzy logic and genetic algorithms.

**Comparison of performance.** In order to compare the performance of MLR and ANN we divided the data sets randomly into 750 training data sets and 149 test data sets. Both MLR and ANN were applied to the training data. The resulting models were used to estimate the *P/B* ratio of each test data set from the independent variables. Then we computed the correlation between calculated *P/B* ratios and *P/B* ratios estimated by either method using log-transformed data. The correlation coefficients were interpreted as a measure of prediction accuracy. For statistical comparison of the 2 methods we applied a straightforward bootstrap resampling approach (Efron & Gong 1983, Efron & Tibshirani 1993): the whole procedure from data selection to correlation computation—was repeated 10 times which resulted in 10 pairs of correlation coefficients for MLR and ANN, respectively. These data were tested for significant differences between MLR and ANN by ANOVA with randomized-complete-block design (Sokal & Rohlf 1995).

Based on the data of Sprung (shown in our Table 3), estimates of a MLR model based on all 899 data sets were compared with average estimates of the 10 ANN established previously.

## RESULTS

Pilot trials showed the MLR models including the variables *T*, *D*, and *M* and the binary variables Nos. 1, 2, & 4 to 8 (Table 4) to work best; hence, this model was used for all 10 experimental trials. In all trials, *M* and *T* explained the highest proportion of variance in *P/B*.

The 10 ANN constructed by 'NeuralWorks Predict' differed in number and type of input variables selected. In all trials, the variable selection algorithm decided in favor of the parallel use of several different transformations of at least one of the continuous variables (Table 5).

The $r^2$ coefficients of the correlation between calculated *P/B* ratios and *P/B* ratios estimated by MLR and ANN, respectively, are shown in Table 6. $r^2$ coefficients of ANN (mean = 0.799) were significantly ($p < 0.001$) higher than those of MLR (mean = 0.751), the average difference was about 6%.

The performance of the MLR in predicting annual production based on all 899 data sets compared to the average of the 10 ANN is shown in Fig 1. The average absolute deviation of estimated from calculated population production was 85% for MLR and 69% for ANN, respectively. These values did not differ significantly ($p > 0.10$). MLR overestimated total production (= sum of the 30 populations in Table 3) by 8.4%, whereas the ANN estimate was 7.0% below Sprung's figure.

## DISCUSSION

Ten re-sampling trials are rather few replicates for a proper bootstrap estimate of a correlation coefficient. However, we did not aim for an exact estimate, but for

Table 5. Architecture of the 10 artificial neural networks constructed. –: Parameter not included; +: 1; ++: 2; +++: 3; ++++: 4 transformations of this parameter included

| | Trial | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **Parameter** | | | | | | | | | | |
| Vagile - Sessile | + | – | + | + | + | + | + | + | – | – |
| Epifauna - Infauna | – | – | + | – | – | + | + | + | – | – |
| Carnivorous | + | – | – | + | + | – | – | + | – | – |
| Omnivorous | + | – | + | – | + | + | – | + | – | + |
| Herbivorous | – | – | – | – | – | – | – | – | – | – |
| Lake | – | – | – | – | – | – | – | + | + | – |
| River | – | + | + | – | + | – | – | – | – | + |
| Marine | + | – | + | – | + | – | – | – | – | – |
| Mollusca | + | + | – | – | – | – | – | – | – | – |
| Crustacea | + | – | – | + | – | + | + | – | – | + |
| Polychaeta | – | – | – | – | – | + | + | + | + | – |
| Echinodermata | – | + | – | – | – | – | + | + | – | – |
| Insecta larvae | – | + | – | + | + | + | + | – | – | + |
| Water depth *D* | + | +++ | + | + | + | + | ++ | +++ | +++ | +++ |
| Temperature *T* | ++ | +++ | ++ | + | +++ | ++ | +++ | ++++ | + | +++ |
| Mean body mass *M* | ++ | + | +++ | ++ | +++ | ++ | ++ | ++ | +++ | +++ |
| **Network structure** | | | | | | | | | | |
| Input nodes | 11 | 11 | 11 | 8 | 13 | 11 | 13 | 16 | 9 | 13 |
| Hidden nodes | 17 | 6 | 20 | 13 | 17 | 18 | 14 | 11 | 4 | 12 |
| Output nodes | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

a comparison of MLR and ANN performance. Because of the parallel evaluation of both methods, we could use a randomized-complete-block-test design which counteracts high sample variance. Our results showed ANN to perform slightly but significantly better in predicting population *P/B* ratio than MLR (Table 6). The final test with Sprung's data revealed only slight and insignificant differences in prediction accuracy (Fig. 1), but this single example provides only limited evidence.

Two conclusions can be drawn from this small difference in accuracy between the 2 models: On the one hand, the *a priori* assumptions concerning the relations between *M*, *T*, *D* and *P/B* ratio used for the MLR are not far from reality. However, the simultaneous use of several different transformations of *M*, *T* and *D* in the ANNs (Table 5) may point towards somewhat more complex relations present in our data. On the other hand, all relevant relations between input variables and *P/B* seem to be covered by the MLR. The remaining 20 to 25% variance in either trial are 'true noise' which cannot be explained by the input parameters used here. This may be due to important parameters missing in our data set, e.g. primary production (Suess 1980) hydrodynamic parameters such as turbulence or current regime which affect

Table 6. Performance of multiple linear regression (MLR) and artificial neural network (ANN) in predicting log(*P/B*) of 149 test data sets not used to construct the respective model. The table shows the squared correlation coefficient $r^2$ of the relation between calculated and estimated log(*P/B*). MLR and ANN performance are significantly different ($p < 0.001$)

| Trial | MLR | ANN | Difference |
|---|---|---|---|
| 1 | 0.758 | 0.803 | 0.045 |
| 2 | 0.758 | 0.808 | 0.050 |
| 3 | 0.714 | 0.766 | 0.052 |
| 4 | 0.769 | 0.812 | 0.043 |
| 5 | 0.786 | 0.835 | 0.049 |
| 6 | 0.724 | 0.769 | 0.045 |
| 7 | 0.769 | 0.796 | 0.027 |
| 8 | 0.732 | 0.808 | 0.076 |
| 9 | 0.775 | 0.787 | 0.012 |
| 10 | 0.724 | 0.808 | 0.084 |
| Mean | 0.751 | 0.799 | 0.048 |

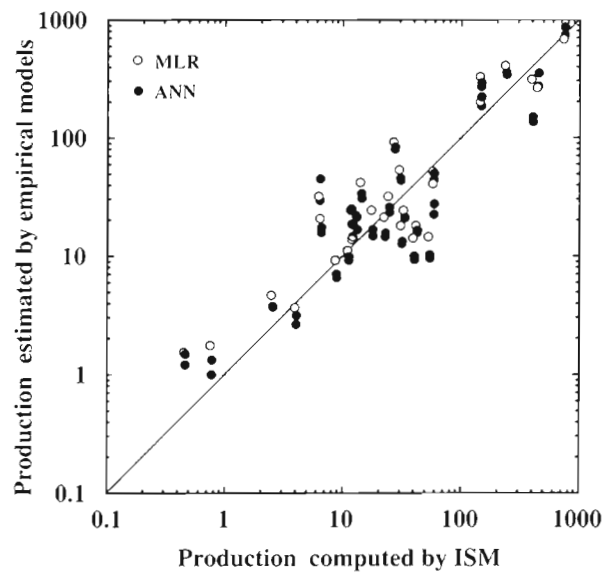


Fig. 1 Production (kJ $m^{-2}$ $yr^{-1}$) computed from *P/B* estimates using multiple linear regression (MLR based on 899 data sets, $R^2 = 0.733$) or artificial neural networks (ANN, average of the 10 nets in Tables 5 & 6) compared to estimates based on ISM (Table 3). The diagonal line represents the expected 1:1 relationship. Total production is estimated to be 2871.4 kJ $m^{-2}$ $yr^{-1}$ (MLR) and 2462.7 kJ $m^{-2}$ $yr^{-1}$ (ANN), respectively, while ISM yielded 2648.7 kJ $m^{-2}$ $yr^{-1}$

the structure and dynamics of macrozoobenthos (e.g Boesch et al. 1976, Daly & Mathieson 1977, Grant 1981, Blaricom 1982, Eckman 1983). Additionally, data accuracy is likely to be affected by methodological shortcomings. Field sampling of population data is a source of large variability (Eleftheriou & Holme 1984 and references therein). Misinterpretation of data, e.g. of age classes, can lead to distinct errors in production computations, too. This might be the case with *Capitella* sp. in Sprung's data (our Table 3), the *P/B* ratio of which is well below estimates of other authors for capitellid polychaetes (e.g. Oyenekan 1983).

It is obvious from our results that even advanced models applied to extensive data sets do not result in empirical relations which are able to estimate population *P/B* ratios or production with high accuracy. Average absolute deviations of 85% (MLR) and 69% (ANN) of the estimate from the measured value may be above acceptable limits (compare Fig 1). However, provided that deviations of estimates from true values are randomly distributed among the populations of a community, empirical models should be able to estimate community production fairly accurately, as already stated by Brey (1990). The error in estimating the total production of the 30 populations included in Sprung's data (our Table 3) is only +8.4% (MLR) and −7.0% (ANN). Although more test data are required for valid statements, we believe that these models may be extremely helpful in investigations on community energetics.

Our results indicate that one potential application of ANN in ecology may be evaluating the quality of MLR prediction models. Because of their inherent flexibility, ANN may perform better than MLR in many cases, as shown in this study. Therefore they may be a more appropriate tool when emphasis is put on the prediction itself and not on the underlying relations between independent and dependent variables.

*Acknowledgements.* We thank Dr Martin Sprung (Universidade do Algarve, Portugal) for providing unpublished data on productivity of the Ria Formosa benthic community. Alfred Wegener Institute Publication No. 1053

## LITERATURE CITED

Alongi DM (1990) The ecology of tropical soft-bottom ecosystems. Oceanogr Mar Biol Annu Rev 28:381–496

Banse K, Mosher S (1980) Adult body mass and annual production/biomass relationships of field populations. Ecol Monogr 50:355–379

Blaricom, GR van (1982) Experimental analysis of structural regulation in a marine sand community exposed to oceanic swell. Ecology 52:283–305

Boesch DF, Diaz RJ, Virnstein RW (1976) Effect of tropical storm Agnes on soft-bottom macrobenthic communities of the James and York Estuaries at the lower Chesapeake Bay. Chesapeake Sci 17:246–259

Brey T (1990) Estimating productivity of macrobenthic invertebrates from biomass and mean individual weight. Meeresforsch 32:329–343

Brey T (1996) Empirische Untersuchungen zur Populationsdynamik makrobenthischer Evertebraten. Habilitation thesis, Universität Bremen, Germany

Clarke A (1991) What is cold adaption and how should we measure it? Am Zool 31:81–92

Culverhouse PF, Ellis R, Simpson RG, Williams R, Pierce RW, Turner JT (1992) Automatic categorisation of five species of Cymatocylis (Protozoa, Tintinnida) by artificial neural network. Mar Ecol Prog Ser 107:273–280

Daly MA, Mathieson AC (1977) The effect of sand movement on intertidal seaweeds and selected invertebrates at Bound Rock, New Hampshire, USA. Mar Biol 43:45–55

Dayhoff JE (1990) Neural network architectures. Van Nostrand Reinhold, New York

Draper NR, Smith H (1981) Applied regression analysis. Wiley, New York

Eckman JE (1983) Hydrodynamic processes affecting benthic recruitment. Limnol Oceanogr 28:241–257

Edgar G (1990) The use of size structure of benthic macrofaunal communities to estimate faunal biomass and secondary production. J Exp Mar Biol Ecol 137:195–214

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York

Efron B, Gong G (1983) A leisurely look at the bootstrap, the jackknife and cross-validation. Am Stat 37:36–48

Eleftheriou A, Holme NA (1984) Macrofauna techniques. In: Holme NA, McIntyre AD (eds) Methods for the study of marine benthos. Blackwell Scientific, Oxford, p 140–216

French M, Recknagel F (1994) Modelling of algal blooms in freshwaters using artificial neural networks. In: Zanetti P (ed) Computer techniques in environmental studies 5, Vol. 2. Environmental systems. Computational Mechanics Inc, Billerica, MA, p 87–94

Grant J (1981) Sediment transport and disturbance on an intertidal sandflat: infaunal distribution and recolonization. Mar Ecol Prog Ser 6:249–255

Ikeda T (1985) Metabolic rate of epipelagic marine zooplankton as a function of body mass and temperature. Mar Biol 85:1–11

Levinton JS (1982) Marine ecology. Prentice-Hall, Englewood Cliffs, NJ

Morin A, Bourassa N (1992) Modèles empiriques de la production annuelle et du rapport *P/B* d'invertébrés benthiques d'eau courante. Can J Fish Aquat Sci 49: 532–539

Nelson MM, Illingworth WT (1991) A practical guide to neural nets. Addison-Wesley, Reading, MA

Oyenekan JA (1983) Production and population dynamics of *Capitella capitata*. Arch Hydrobiol 98:115–126

Parry GD (1983) The influence of the cost of growth on ectotherm metabolism. J Theor Biol 101:453–477

Plante C, Downing JA (1989) Production of freshwater invertebrate populations in lakes. Can J Fish Aquat Sci 46: 1489–1498

Potter ECE, Kell L, Reddin DG (1993) The discrimination of North American and European salmon using a genetic algorithm and by neural network. ICES (Int Counc Explor Sea) CM 1993/M:18

Robertson AL (1979) The relationship between annual production: biomass ratios and lifespans for marine macrobenthos. Oecologia 38:193–202

Schmidt-Nielsen K (1984) Scaling — why is animal size so important. Cambridge Univ Press, Cambridge

Schwinghamer P, Hargrave B, Peer D, Hawkins CM (1986)

Partitioning of production and respiration among size groups of organisms in an intertidal benthic community. Mar Ecol Prog Ser 31:131–142

Sokal RR, Rohlf FJ (1995) Biometry — the principles and practice of statistics in biological research. Freeman & Co, New York

Sprung M (1993) Estimating macrobenthic secondary production from body weight and biomass: a field test in a non-boreal intertidal habitat. Mar Ecol Prog Ser 100:103–109

Sprung M (1994) Macrobenthic secondary production in the intertidal zone of the Ria Formosa – a lagoon in southern Portugal. Estuar Coast Shelf Sci 38:539–558

Steen JD, Steen H, Stenseth NC (1991) Population dynamics of poikilotherm and homeotherm vertebrates: effects of food shortage. Oikos 60:269–272

Suess E (1980) Particulate organic carbon flux in the oceans — surface productivity and oxygen utilization. Nature 288:260–263

Warwick RM (1980) Population dynamics and secondary production of benthos. In: Tenore KR, Coull BC (eds) Marine benthic dynamics. Univ South Carolina Press, Columbia, p 1–24

Zaika VY (1970) Relationship between the productivity of marine molluscs and their life-span. Oceanology (USSR) 10:547–552

Zupan J, Gasteiger J (1991) Neural networks: a new method for solving chemical problems or just a passing phase? Anal Chim Acta 248:1–30

*This article was submitted to the editor*

*Manuscript first received: January 5, 1996*
*Revised version accepted: May 2, 1996*