杨利娟,金武,黄珊珊,等. 人工神经网络在环棱螺体质量缺失值预测中的应用[J]. 华中农业大学学报,2021,40(5): 154-159.
DOI:10.13300/j.cnki.hnlkxb.2021.05.019

# 人工神经网络在环棱螺体质量缺失值预测中的应用

杨利娟[1],金武[2],黄珊珊[1],闻海波[2],
马学艳[2],唐小林[1],王卫民[1],曹小娟[1]

1.华中农业大学水产学院/教育部长江经济带大宗水生生物产业绿色发展工程研究中心/
农业农村部淡水生物繁育重点实验室,武汉 430070;
2.中国水产科学研究院淡水渔业研究中心/中美淡水贝类种质资源保护及利用国际联合实验室,无锡 214081

**摘要**　环棱螺育种时,往往会出现部分个体体质量数据缺失的情况。为尽可能利用育种性能优异的所有个体的信息,采用人工神经网络对来自 5 个地理群体(阳澄湖、江阴、官莲湖、洪湖和仙桃)的 784 个环棱螺的 4 个形态学指标(包括壳高、壳宽、壳口高和壳口宽)和体质量数据进行训练,再使用太湖群体的 261 个环棱螺的相应数据进行人工神经网络模型测试,建立了用于环棱螺体质量缺失值预测的人工神经网络模型,利用该人工神经网络模型对微山湖群体的 201 个环棱螺缺失的体质量进行预测,并比较该方法与另外 2 种缺失值预测方法(即预测均数匹配法和随机森林预测法)的决定系数。结果显示,研究构建的人工神经网络模型对环棱螺体质量缺失值预测的决定系数为 0.96,明显高于预测均数匹配法(0.87)和随机森林预测法(0.85)的决定系数。以上结果表明,本研究建立的人工神经网络模型可以用于环棱螺体质量缺失值的预测。

**关键词**　环棱螺;人工神经网络;体质量;缺失值预测;决定系数
**中图分类号**　S 966.2　　**文献标识码**　A　　**文章编号**　1000-2421(2021)05-0154-06

环棱螺俗称螺蛳、豆田螺、石螺,隶属于腹足纲(Gastropoda)、前鳃亚纲(Prosobranchia)、田螺科(Viviparidae)、环棱螺属(*Bellamya*)。环棱螺属常见的种有铜锈环棱螺、方形环棱螺和梨形环棱螺等[1]。因有着营养价值高[2]、用途多[3]的优点,环棱螺越来越受到人们的关注和喜爱[4-5]。然而,随着长江全面禁渔推行,作为水域生态系统中重要成员的环棱螺已被纳入禁捕行列。因此,开展环棱螺繁育工作以推进其养殖业发展势在必行。

目前,环棱螺育种重点关注体质量性状的遗传改良[6],但在育种过程中常因种群保管不善、养殖水环境剧变、饵料不适口及流行性疾病暴发等因素导致环棱螺死亡。虽然环棱螺死亡个体形态学数据(如壳高、壳宽、壳口高和壳口宽等)仍能测量获得,但其体质量数据则会缺失。育种数据缺失的处理包括直接删除[7]、尝试填补[8]、不处理[9]3 种方法。在实践中,因为育种性能优异的个体来之不易,为了尽可能利用所有的信息,往往需要对缺失值进行处理。本研究基于人工神经网络的预测功能,利用测得的环棱螺 4 个形态学数据和体质量数据构建模型,继而对缺失的体质量数据进行预测并评估其效率,以期为环棱螺选择育种提供高效的数据分析工具。

## 1　材料与方法

### 1.1　环棱螺采集及数据测量

从阳澄湖、太湖、江阴、官莲湖、洪湖和仙桃共采集获得 1 045 个环棱螺,利用游标卡尺测量其形态学(包括壳高(SH)、壳宽(SW)、壳口高(AH)和壳口宽(AW))数据,同时测量体质量。此外,从微山湖采集 201 个环棱螺,测量其壳高、壳宽、壳口高和壳口宽。本研究从以上 7 个采样点(含体质量缺失的微山湖群体),共采集获得 1 246 个环棱螺,具体情况见表 1。

收稿日期: 2021-03-29
基金项目:广西柳州市财政资金项目(LZT18-201);中央高校基本科研业务费专项(2662020SCPY002);中国水产科学研究院淡水渔业研究中心基本科研业务费专项(2017JBFM11)
杨利娟,E-mail:1875997025@qq.com
通信作者:曹小娟,E-mail:caoxiaojuan@mail.hzau.edu.cn

表 1　环棱螺采样点和数目

Table 1　Sampling sites and number of *Bellamya*

| 数据测量类型 Types | 采样点 Sampling sites | 样本数 Sample number |
|---|---|---|
| 体质量数据未缺失群体 Populations with body weight data | 阳澄湖 Yangcheng Lake | 257 |
| | 江阴 Jiangyin | 171 |
| | 官莲湖 Guanlian Lake | 106 |
| | 洪湖 Hong Lake | 143 |
| | 仙桃 Xiantao | 107 |
| | 太湖 Tai Lake | 261 |
| 体质量数据缺失群体 Population without body weight data | 微山湖 Weishan Lake | 201 |
| 合计 Summary | | 1 246 |

### 1.2　数据分析

1）人工神经网络构建。体质量未缺失采样群体数据集中随机抽取 75%的数据(784 个)用于训练模型，总体数据中剩余的 25%的数据(261 个)用于测试模型。在建模过程中，经过预先多次的参数调整，人工神经网络设定为 1 个隐含层和 3 个神经元的结构。人工神经网络类似于生物神经元结构，经训练的模型利用输入的 4 个形态学数据生成 1 个输出预测的体质量值。神经元的输出都是输入的加权和加上偏差的函数。一旦接收到的信号总量超过激活阈值，则每个神经元都执行简单的操作[10]。每个典型的神经元用数学函数可以表示为式(1)：

$$y=f(x)=\sum x_i w_i \tag{1}$$

其中，$x_i$ 为输入变量，$w_i$ 为权重，$i$ 为输入变量的个数，$1\leqslant i\leqslant n$。

2)不同预测方法之间的比较。体质量数值预测分别采用 R 统计软件[11]的人工神经网络 neuralnet 包[12]和 mice 包[13]中的预测均数匹配法(predictive mean matching，PMM)[14]和随机森林预测法(random forest，RF)[15]。不同缺失值预测的方法统一以模型的决定系数 $R^2$ 来进行比较[16]。决定系数的计算方法为式(2)：

$$R^2=1-\frac{\sum_{i=1}^{n}(X_t-X_t')^2}{\sum_{i=1}^{n}(X_t')^2} \tag{2}$$

其中，$X_t$ 和 $X_t'$ 分别为第 $t$ 个真实值与第 $t$ 个预测值。

## 2　结果与分析

### 2.1　描述性统计

表 2 统计了体质量数据未缺失的 6 个地理群体环棱螺壳高、壳宽、壳口高、壳口宽和体质量数据(形态学数据精确到 0.01 mm，体质量数据精确到 0.01 g)。体质量数据未缺失群体的 4 个形态学性状数据的分布如图 1 所示。本研究构建的人工神经网络模型预测的微山湖环棱螺体质量为(3.91±1.30) g。微山湖环棱螺的形态学性状值小于其他 6 个地理群体环棱螺的形态学性状值，本研究基于人工神经网络模型预测的微山湖环棱螺的体质量也小于其他 6 个地理群体环棱螺的体质量(表 2)，这在一定程度上反映了本研究构建的人工神经网络模型对环棱螺体质量预测的准确性。

表 2　形态学数据和体质量的描述性统计(平均值±标准差)

Table 2　Descriptive statistics of morphological data and body weights (Mean±SD)

| 群体 Population | 壳高/mm Shell height | 壳宽/mm Shell width | 壳口高/mm Aperture height | 壳口宽/mm Aperture width | 体质量/g Body weight |
|---|---|---|---|---|---|
| 体质量数据未缺失群体 Populations with body weight data | 27.81±4.45 | 17.59±3.01 | 13.14±1.81 | 10.6±1.59 | 4.75±1.81 |
| 体质量数据缺失群体(微山湖) Population without body weight data | 25.14±2.75 | 16.37±1.71 | 8.67±1.18 | 9.88±1.43 | 3.91±1.30$^{\$}$ |

注 Note：$：人工神经网络预测值 Predicted value made by the artificial neural network.

### 2.2　人工神经网络模型建模

人工神经网络模型的准确度经多次参数调整后，经过 622 810 次迭代后收敛(图 2)。连接实线的数值为该连接的权重，连接虚线上的数值为每一步计算添加的权重。广义权重散点图显示，壳高、壳宽、壳口高、壳口宽这 4 个性状对体质量的线性相关关系很强(图 3)。壳宽和壳口宽的广义权重多数分布于 0 附近，说明这 2 个性状对体质量的作用相对较弱。壳高和壳口高这 2 个性状对体质量的作用较强，这 2 个性状和体质量存在一定的非线性相关性。

人工神经网络模型对环棱螺体质量预测的决定系数为 0.96，说明该模型具有较高的准确性。预测均数匹配法和随机森林预测法的决定系数分别为 0.87和 0.85，这说明人工神经网络和其他 2 种体质量缺失值预测方法相比，具有明显优势。

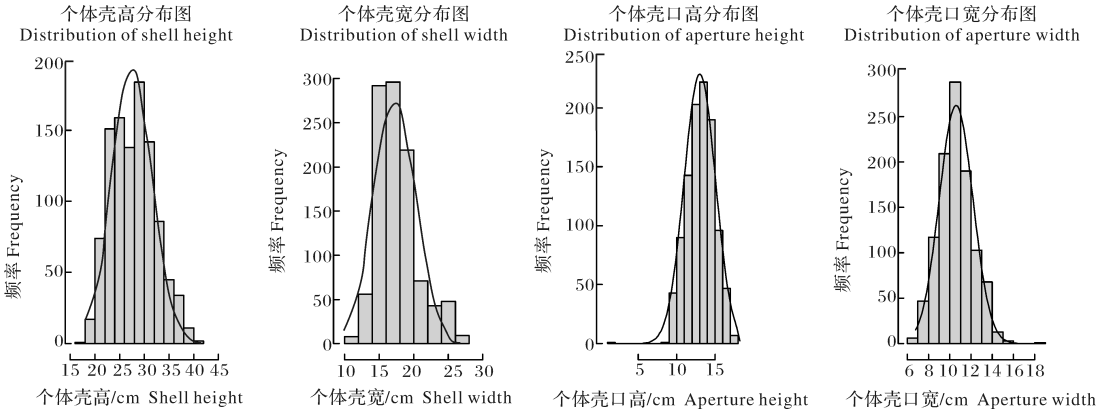


**图 1 微山湖环棱螺 4 个形态学性状数据的分布图**

Fig.1 Distribution of four morphological traits in *Bellamya* sampled from Weishan Lake

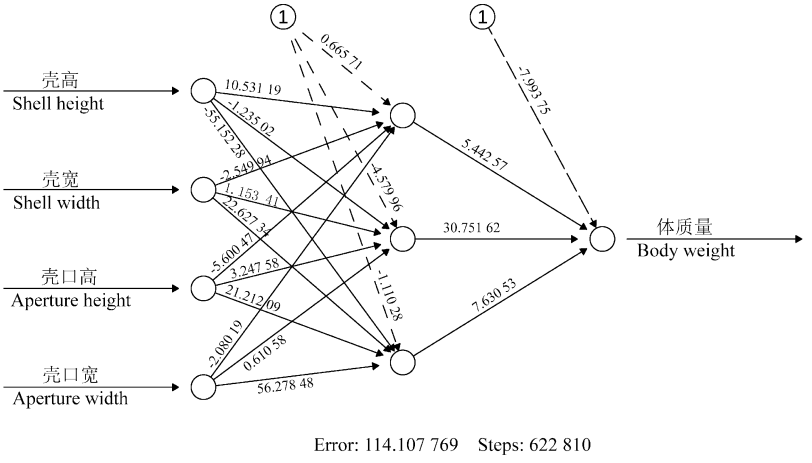


**图 2 人工神经网络结构图**

Fig.2 Neural network structure diagram

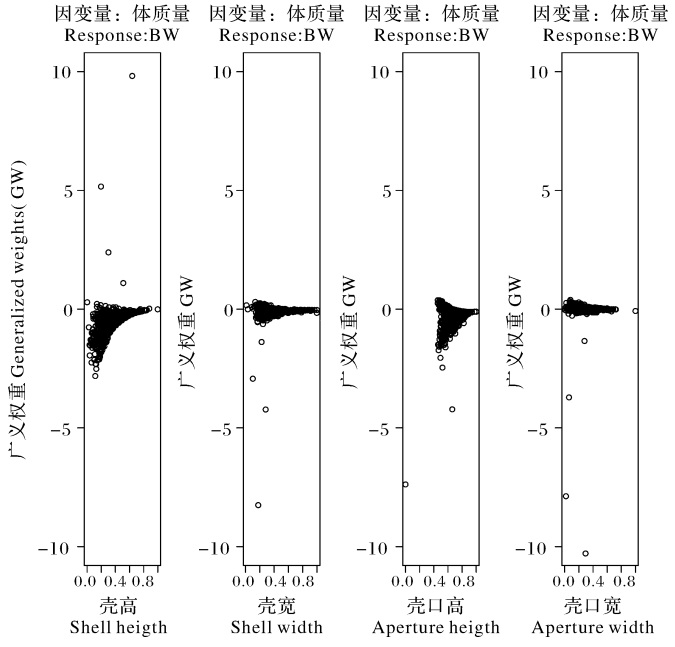


**图 3 广义权重的散点图**

Fig.3 Scatter plot of generalized weights

# 3 讨　论

## 3.1 人工神经网络在缺失值预测中的应用

人工神经网络作为一种并行的计算模型，不需要对研究对象的数据规律有大致的了解，只需要通过网络本身的学习功能就可以得到网络输入与输出的关系[9]。与传统建模方法相比，人工神经网络对非线性相关的数据的学习能力更强。基于神经网络进行缺失数据估计的基本步骤是：利用该系统中的已知数据训练网络，在网络满足要求后，把其他参数的数据（不含缺失值）输入网络，网络输出值即为缺失数据的估计值[9]。人工神经网络在一些复杂系统如飞机发动机[9]、农业气象[17]、原子反应堆[18]、农田生态系统[19]、湖泊水体[20]中数据处理中已取得了一定进展。对活立木茎干水分缺失数据的研究表明，人工神经网络较传统的插值方法优势明显，且神经网络方法预测精度受数据缺失量增多的影响较小[21]。对农业生产资料数据库中缺失数据的神经网络预测结果也显著好于传统的线性插补和加权分析[22]。与农学研究相似，生态学监测中也较易出现缺失值。基于神经网络的参数学习方法也取得了比其他算法更高的精度[23]。本研究率先探索建立了在水产育种领域较易出现的缺失值预测方法，并得到了比传统缺失值处理方法更高的决定系数。针对环棱螺育种过程中常涉及的体质量缺失问题，本研究提前进行了技术储备（即构建相应的高效人工神经网络模型），但同时也存在实验数据量偏少的不足之处，我们将在日后的研究中，加大数据量的采集。

## 3.2 常见缺失值预测方法的比较

在本研究中，尽管环棱螺形态学性状和体质量测量数据有限（1 045 个个体），但构建好的人工神经网络模型对 201 个体质量缺失的样本预测仍取得了较高的准确度，在缺失数据增加若干数量级是否能取得类似效果仍待深入研究[24]。由于预测均数匹配法只有在某些特定的缺失数据类型时才能取得较好的效果[25]，本研究中缺失的体质量数据与环棱螺自身形态学数据相关，可能也会造成该方法预测缺失值的决定系数偏低。此外，随机森林对缺失数据和非平衡的数据的结果分析比较稳健，能够在高维数据中有效地分析具有交互作用和非线性关系的数据[26]，但对多元共线性不敏感[27]。在本研究建立模型过程中，可能由于训练集样本量偏小导致随机森林模型的决定系数低于人工神经网络，随机森林预测缺失值的优势未得到完全显示。后期可以通过增加训练样本量，进一步挖掘随机森林预测法的优势。尽管缺失值预测的方法有很多，但在实际分析中仍需谨慎对待预测结果，并进行多种方法的比较[28]。

## 参考文献 References


[1] 刘月英，张文珍，王跃先，等.中国经济动物志：淡水软体动物[M].北京：科学出版社，1979：14. LIU Y Y，ZHANG W Z，WANG Y X，et al. Economic fauna of China：freshwater mollusks[M]. Beijing：Science Press，1979：14(in Chinese).

[2] 夏树华，王璋.螺蛳腹足肌的酶解工艺[J].食品与生物技术学报，2006，25(5)：91-97. XIA S H，WANG Z. Hydrolysis of *Bellamya purificat*a foot muscle[J]. Journal of food science and biotechnology，2006，25(5)：91-97(in Chinese with English abstract).

[3] CHEN X，SHEN Q Y，GU X M，et al. Effects of different live food extracts on fish attraction activities[J]. Agricultural science & technology，2014，15(6)：942-946，963.

[4] 金武，马学艳，闻海波，等.梨形环棱螺 3 个群体形态性状与体质量的相关及通径分析[J].中国农学通报，2017，33(32)：135-139. JIN W，MA X Y，WEN H B，et al. Correlation and path analysis between morphology traits and body weight of *Bellamya purificata* from three populations[J]. Chinese agricultural science bulletin，2017，33(32)：135-139(in Chinese with English abstract).

[5] HUANG S Q，JIANG H J，ZHANG L，et al. Integrated proteomic and transcriptomic analysis reveals that polymorphic shell colors vary with melanin synthesis in *Bellamya purificata* snail[J/OL]. Journal of proteomics，2021，230：103950[2021-03-29]. https：//doi.org/10.1016/j.jprot.2020.103950.

[6] 颜元杰，金武，闻海波，等.梨形环棱螺 60 日龄 6 个生长性状遗传参数估计[J].淡水渔业，2018，48(6)：108-111. YAN Y J，JIN W，WEN H B，et al. Estimation of genetic parameters for growth traits of *Bellamya purificata* in 60 days[J]. Freshwater fisheries，2018，48(6)：108-111(in Chinese with English abstract).

[7] 叶健，胡晓湘，边成，等.大白猪主要生长性状的遗传参数估计及育种中存在问题的探讨[J].华南农业大学学报，2017，38(1)：1-4. YE J，HU X X，BIAN C，et al. Estimation of genetic parameters of major growth traits and existing problems in breeding of Large White pigs[J]. Journal of South China Agricultural University，2017，38(1)：1-4(in Chinese with English abstract).

[8] 岳勇，田考聪.数据缺失及其填补方法综述[J].预防医学情报杂志，2005，21(6)：683-685. YUE Y，TIAN K C. A review of data gaps and their filling methods[J]. Journal of preventive

medicine information,2005,21(6):683-685(in Chinese).

[9] 张宏亭,李学仁,孔韬.BP神经网络在缺失数据估计中的应用[J].计算机工程与设计,2007,28(14):3457-3459.ZHANG H T,LI X R,KONG T.Application of BP neural network in predicting absent data[J].Computer engineering and design,2007,28(14):3457-3459(in Chinese with English abstract).

[10] CIABURRO G,VENKATESWARAN B.神经网络:R语言实现[M].李洪成,译.北京:机械工业出版社,2018.CIABURRO G,VENKATESWARAN B.Neural networks with R [M].LI H C,translator.Beijing:China Machine Press,2018(in Chinese).

[11] TEAM C R.R:a language and environment for statistical computing[CP/OL].R foundation for statistical computing,2010.http://www.r-project.org.

[12] GUNTHER F,FRITSCH S.Neuralnet:training of neural networks[J].The R journal,2010,2(1):30-38.

[13] VAN BUUREN S,GROOTHUIS-OUDSHOORN K.Mice:multivariate imputation by chained equations in R[J].Journal of statistical software,2011,45(3):1-67.

[14] 梁永厚,张文广,王瑞军,等.绒山羊育种过程中对缺失数据处理的新方法[J].中国草食动物科学,2007,27(1):8-11.LIANG Y H,ZHANG W G,WANG R J,et al.The new method of losing data treatment in Cashmere Goat breeding[J].China herbivores,2007,27(1):8-11(in Chinese).

[15] 邹永潘,王儒敬,李伟.随机森林算法在小麦育种辅助评价中的应用[J].计算机系统应用,2017,26(12):181-185.ZOU Y P,WANG R J,LI W.Application of the random forest algorithm in wheat breeding evaluation[J].Computer systems & applications,2017,26(12):181-185(in Chinese with English abstract).

[16] MOHAMED Z E.Using the artificial neural networks for prediction and validating solar radiation[J].Journal of the Egyptian mathematical society,2019,27(1):1-13.

[17] 赵兰兰,王恺,赵兵.农业气象资料中连续性数据缺失插补方法研究[J].水电能源科学,2010,28(5):4-6,172.ZHAO L L,WANG K,ZHAO B.Interpolation method of continuous missing data in agro-meteorology[J].Water resources and power,2010,28(5):4-6,172(in Chinese with English abstract).

[18] 宋梅村,蔡琦.基于BP神经网络的反应堆功率预测[J].原子能科学技术,2011,45(10):1242-1246.SONG M C,CAI Q.Reactor power prediction based on BP neural network[J].Atomic energy science and technology,2011,45(10):1242-1246(in Chinese with English abstract).

[19] 米湘成,马克平,邹应斌.人工神经网络模型及其在农业和生态学研究中的应用[J].植物生态学报,2005,29(5):863-870.MI X C,MA K P,ZOU Y B.Artificial neural network and its application in agricultural and ecological research[J].Acta phytoecologica sinica,2005,29(5):863-870(in Chinese with English abstract).

[20] 刘恒.BP神经网络在千岛湖水体富营养化变化预测中的应用[D].杭州:浙江大学,2007.LIU H.Back-propagation network model for predicting the change of eutrophication of Qiandao Lake[D].Hangzhou:Zhejiang University,2007(in Chinese with English abstract).

[21] 宋维,高超,赵玥,等.基于LSTM的活立木茎干水分缺失数据填补方法[J].林业科学,2020,56(2):134-141.SONG W,GAO C,ZHAO Y,et al.Method of filling the missing water loss data of living plant stem by sequence based on LSTM[J].Scientia silvae sinicae,2020,56(2):134-141(in Chinese with English abstract).

[22] 蒋丽丽,姜大庆.基于BP神经网络的农资库存数据插补技术[J].江苏农业科学,2018,46(20):268-271.JIANG L L,JIANG D Q.Interpolation technology of agricultural assets inventory data based on BP neural network[J].Jiangsu agricultural sciences,2018,46(20):268-271(in Chinese).

[23] 邵佳.生态站中时间序列缺失值填补研究[D].北京:北京林业大学,2017.SHAO J.Research of time series missing values imputation method in ecological monitoring stations[D].Beijing:Beijing Forestry University,2017(in Chinese with English abstract).

[24] 任云志,贺跃光,吴弘,等.基于BP神经网络的不完全测量数据处理方法研究[J].现代测绘,2013,36(1):9-11,15.REN Y Z,HE Y G,WU H,et al.The approach of incomplete surveying data based on BP neural net[J].Modern surveying and mapping,2013,36(1):9-11,15(in Chinese with English abstract).

[25] 鲍晓蕾,高辉,胡良平.多种填补方法在纵向缺失数据中的比较研究[J].中国卫生统计,2016,33(1):45-48.BAO X L,GAO H,HU L P.Comparative study of various imputation methods in dealing with longitudinal missing data[J].Chinese journal of health statistics,2016,33(1):45-48(in Chinese with English abstract).

[26] 李贞子,张涛,武晓岩,等.随机森林回归分析及在代谢调控关系研究中的应用[J].中国卫生统计,2012,29(2):158-160,163.LI Z Z,ZHANG T,WU X Y,et al.Methodology of regression by random forest and its application on metabolomics[J].Chinese journal of health statistics,2012,29(2):158-160,163(in Chinese with English abstract).

[27] 李欣海.随机森林模型在分类与回归分析中的应用[J].应用昆虫学报,2013,50(4):1190-1197.LI X H.Using "random forest" for classification and regression[J].Chinese journal of applied entomology,2013,50(4):1190-1197(in Chinese with English abstract).

[28] 胡玄子,陈小雪,钱叶亮,等.数据处理中缺失数据填充方法的研究[J].湖北工业大学学报,2013,28(5):82-84.HU X Z,CHEN X X,QIAN Y L,et al.Research on the method of filling missing data in data processing[J].Journal of Hubei University of Technology,2013,28(5):82-84(in Chinese with English abstract).

# Application of an artificial neural network in prediction of missing body weights data of *Bellamya*

YANG Lijuan[1], JIN Wu[2], HUANG Shanshan[1], WEN Haibo[2],
MA Xueyan[2], TANG Xiaolin[1], WANG Weimin[1], CAO Xiaojuan[1]

1. *College of Fisheries/Engineering Research Center of Green Development for Conventional Aquatic Biological Industry in Yangtze River Economic Belt, Ministry of Education/Key Lab of Freshwater Animal Breeding, Ministry of Agriculture and Rural Affairs, Huazhong Agricultural University, Wuhan* 430070, *China*;
2. *Sino-US Cooperative Laboratory for Germplasm Conservation and Utilization of Freshwater Mollusks/Freshwater Fisheries Research Center, Chinese Academy of Fishery Sciences, Wuxi* 214081, *China*

**Abstract** In the breeding of *Bellamya*, weight data of some individuals are often missing. To make best use of information on all individuals with excellent breeding performance, an artificial neural network was trained on four morphological traits (including shell height, shell width, aperture height and aperture width) and body weight data of 784 individuals from five geographical populations including Yangcheng Lake, Jiangyin, Guanlian Lake, Hong Lake and Xiantao. After this, data of 261 individuals from Tai Lake were used to test the artificial neural network model. In the end, an artificial neural network model for predicting missing body weights of *Bellamya* was successfully established. In addition, the artificial neural network model was used to predict the missing body weights of 201 *Bellamya* from Weishan Lake, and the determination coefficient of this method was compared with those of two other prediction methods (i.e., the predicted mean matching method and the random forest prediction method). The results showed that the determination coefficient of the artificial neural network model constructed in this study was 0.96 for predicting the missing body weight, which was obviously higher than those of the predictive mean matching method (0.87) and the random forest prediction method (0.85). This study could provide an efficient method for the prediction of missing values of body weight involved in the breeding process of the *Bellamya*.

**Keywords** *Bellamya*; artificial neural network; body weight; prediction of missing values; determination coefficient

（责任编辑：边书京）