АЗЖ
http://www.bgpu.ru/azj/ *© Амурский зоологический журнал. X(3-4), 2018. 197-205*
http://elibrary.ru/title_about.asp?id=30906 *© Amurian zoological journal. X(3-4), 2018. 197-205*
УДК 574.55

ПРИМЕНЕНИЕ РЕГРЕССИОННЫХ УРАВНЕНИЙ В ЛИМНОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ: ПРЕИМУЩЕСТВА ИСПОЛЬЗОВАНИЯ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

О.П. Сосновская, В.В. Скворцов

APPLICATION OF REGRESSION EQUATIONS IN LIMNOLOGICAL RESEARCHES: ADVANTAGES OF USING ARTIFICIAL NEURAL NETWORKS

O.P. Sosnovskaia, V.V. Skvortsov

Российский государственный педагогический университет им. А.И. Герцена, наб. реки Мойки, 48, Санкт-Петербург, 191186, Россия. E-mail: olgasosnovskaia@gmail.com, vlad_skvortsov@mail.ru

Ключевые слова: *лимнология, регрессионные модели, искусственные нейронные сети, экосистема, первичная продукция, хлорофилл a, зообентос*

Резюме. В настоящей работе на основе литературных данных проведен анализ точности предсказания регрессионных моделей некоторых важных параметров биологических озерных экосистем (первичная продукция, концентрация хлорофилла *a*, биомасса зоопланктона и зообентоса). Было показано, что точность предсказания, измеряемая как средняя абсолютная ошибка в процентах (MAPE) практически во всех случаях составляет 60-100%, что не позволяет использовать эти модели для экспертных оценок экосистемных параметров озёр. Использовав литературные данные, с помощью технологии искусственных нейронных сетей были сгенерированы множественные регрессионные модели. Проверка точности этих моделей производилась на независимых данных, которые не использовались для построения конкретной модели. Нейросетевые регрессионные модели оказались более точны – их средняя абсолютная ошибка в процентах не превышала 25%. Таким образом, по нашему мнению, дальнейшее применение регрессионных нейросетевых моделей в лимнологических исследованиях представляется весьма перспективным.

Herzen State Pedagogical University of Russia, 48, Moika Emb., Saint-Petersburg, 191186, Russia. E-mail: olgasosnovskaia@gmail.com, vlad_skvortsov@mail.ru

Key words: *limnology, regression models, artificial neural networks, ecosystem, primary production, chlorophyll a, zoobenthos*

Summary. In the present paper the accuracy of regression models prediction of some important parameters of lake ecosystems (primary production, chlorophyll a concentration, zooplankton and zoobenthos biomass) is analyzed on the basis of literature data. It was shown that the prediction accuracy, measured as the mean absolute percentage error (MAPE), in almost all cases reaches 60-100%, what does not allow these models to be used for expert assessments of the ecosystem parameters of lakes. Using the literary data, multiple regression models were generated on the base of artificial neural network technology. Verification of the accuracy of these models was performed on independent data that were not used to build this model. Neural network regression models turned out to be more accurate – their mean absolute percentage error did not exceed 25%. Thus, in our opinion, the advantage of using regression neural network models in limnological studies is very perspective.

ВВЕДЕНИЕ

Начиная с 70-х годов прошлого столетия и до нашего времени лимнологи во всем мире стараются аппроксимировать регрессионными моделями зависимости между биотическими и абиотическими характеристиками озерных экосистем. На первом этапе обычно применялись простые линейные регрессионные уравнения, а в последствии в исследовательской практике стали использоваться более сложные модели –множественные, которые учитывали влияния на исследуемый

параметр нескольких факторов. С.П. Китаев в своей монографии [Китаев, 2007] приводит несколько десятков подобных уравнений регрессии (регрессионных моделей). Большая часть этих уравнений аппроксимирует зависимости характеристик фитопланктонных сообществ озер (биомасса, первичная продукция, концентрация хлорофилла *a*), биомассы зоопланктона, биомассы макрозообентоса, биомассы и вылова рыб от концентрации общего фосфора в воде. Иногда подобные модели усложняются введением дополнительных переменных (главным образом, морфометрических и гидрохимических характеристик озер). Авторы подобных моделей подтверждали тот факт, например, что величина первичной продукции прямо пропорциональна концентрации общего фосфора. Однако даже беглого взгляда на эти регрессионные модели оказывается достаточным, чтобы заметить, что все однотипные модели отличаются друг от друга своими коэффициентами и становится понятно, что регрессионные модели, рассчитанные по данным какой-либо конкретной группы озер конкретного региона, вряд ли будут способны предсказать значений изучаемого параметра в другом регионе. В этом, очевидно, кроется главный недостаток линейных регрессионных (в том числе и множественных) моделей, используемых обычно лимнологами.

Мы полагаем, что основным назначением регрессионных моделей является предсказание с достаточной точностью важнейших характеристик озерных экосистем, ранее не исследованных. Поэтому мы согласны с мнением P. J. Dillon и F. H. Rigler, которые высказали следующее мнение: «Мы считаем, что следует уделить некоторое внимание разработке моделей, ценность которых будет определяться *только их предсказательной способностью, а не их причинно-следственными или эвристическими свойствами*» [Dillon & Rigler, 1974].

Исходя из вышеизложенного, мы задались целью определить точность предсказания линейных регрессионных моделей с предсказательной способностью моделей, построенных по технологии искусственных нейронных сетей. В качестве исходного положения было принято, что первый класс моделей (линейные модели) не может служить инструментом для предсказания, поскольку существуют ограничения: линейность связей между переменными и соответствие нормальному распределению исходных данных. Для нейронных же сетей таких ограничений не существует, то есть регрессионные модели этого типа являются более гибкими.

МАТЕРИАЛЫ И МЕТОДЫ

Для оценки прогностической способности регрессионных моделей, мы взяли из опубликованного перечня [Китаев, 2007] наиболее интересные, и проверили их «работоспособность» на независимых выборках, взятых из литературных источников.

Для определения точности прогноза моделей использовался показатель средней абсолютной ошибки в процентах (the mean percentage absolute error, MAPE), который вычисляется по формуле:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|Oi - Mi|}{Oi} * 100\%$$

где *N – число измерений, Oi – измеренные значения, Mi – предсказанные значения*

Статистическая обработка материалов производилась в MS Excel и Statistica 12 [Нейронные сети…, 2008].

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Простые и множественные линейные регрессионные модели

Первичная продукция. Тестировалась модель связи первичной продукции и концентрации общего фосфора [Иконников и др., 2003] на данных, полученных Институтом озероведения РАН на озерах Карельского перешейка в 70-80-х годах прошлого столетия [Трифонова, 1989]. Точность прогноза оказалась крайне низкой – MAPE=95%;

Концентрация хлорофилла *a*. Тестировалась модель связи концентрации хлорофилла *a* от концентрации общего фосфора [Dillon and Rigler, 1974]) на данных, полученных Институтом озероведения РАН на озерах

Таблица 1

Тестирование моделей биомассы профундального зообентоса по [Hanson & Peters, 1984]

Table 1

Testing models of biomass profundal zoobenthos according to [Hanson & Peters, 1984]

Величины биомассы (г/м^2)	Модель 1	Модель 6	Модель 7	Модель 8	Исходные данные
Среднее	6,08	0,45	0,02	0,0002	6,08
Минимальное	2,50	0,18	0,01	0,0001	2,50
Максимальное	36,00	2,91	0,12	0,0008	36,00

Карельского перешейка в 70-80-х годах прошлого столетия [Трифонова, 1989]. Точность прогноза оказалась исключительно низкой – MAPE=114%;

Биомасса зоопланктона. Тестировались две модели [Hanson & Peters, 1984]) на данных, из литературных источников, опубликованных авторами в статье. В первой из них в качестве предиктора выступала концентрация общего фарфора, во второй – концентрация хлорофилла *a*. Точность первой модели оказалась несколько ниже (MAPE=25,5%, чем второй (MAPE=19,5%);

Биомасса профундального зообентоса. Рассматривались модели тех же авторов [Hanson & Peters, 1984], из которых наиболее интересны 1, 6, 7 и 8 (нумерация моделей из указанной статьи). Модель 1 в качестве предиктора содержала концентрацию общего фосфора, модель 6 содержала два предиктора – концентрацию общего фосфора и площадь озера, модель 7 также содержала два предиктора – концентрацию общего фосфора и среднюю глубину озера, модель 8 содержала два предиктора – концентрацию общего фосфора и максимальную глубину. К сожалению, авторы не привели исходных данных, по которым они строили модели, и поэтому мы приводим расчётные результаты для минимальных, средних и максимальных величин параметров исследованных озёр (табл. 1). Как видно из таблицы 1 о точности предсказания биомассы зообентоса по рассмотренным моделям говорить не приходится.

Биомасса литорального зообентоса. J.B. Rasmussen в своей статье [Rasmussen, 1988] на основании литературных источников и собственных неопубликованных данных проанализировал связь биомассы литорального зообентоса (LZB) озера Мемфремагог (Канада, Квебек – Онтарио) с такими параметрами, как уклон литорали в месте отбора проб, экспозиция (площадь озера, видимая с точки отбора проб) и концентрация хлорофилла *a*. Полученная мультирегрессионная модель смогла объяснить 81% дисперсии биомассы литорального зообентоса. Автор попытался протестировать предсказательную способность полученной модели на множестве *других озер* (термин автора). Другие озера – 20 озер (42 точки взятия проб) расположены преимущественно на территории Канады, несколько озер – на территории США, одно в Голландии и три на территории Российской Федерации (озера Зеленецкое, Кривое и Круглое). К сожалению, автор не дал никаких количественных оценок сходства модельных (расчётных) данных с натурными, а лишь ограничился замечанием, что совпадение *довольно слабое*. Затем, автор объединил данные со всех пунктов на всех озёрах (n=71) и на основании объединенного множества построил новую регрессионную модель, в которую вошли в качестве предикторов: концентрация хлорофилла *a*, уклон литорали, экспозиция, концентрации ионов кальция и хлора. Новая модель смогла объяснить 80% дисперсии биомассы литорального зообентоса.

Воспользовавшись опубликованной моделью, мы сравнили расчетные данные биомассы литорального зообентоса (в потенцированном виде) с измеренными, которые были использования при расчете модели (рис. 1). Прежде всего обращает внимание, что коэффициент дисперсии (R^2) оказался заметно

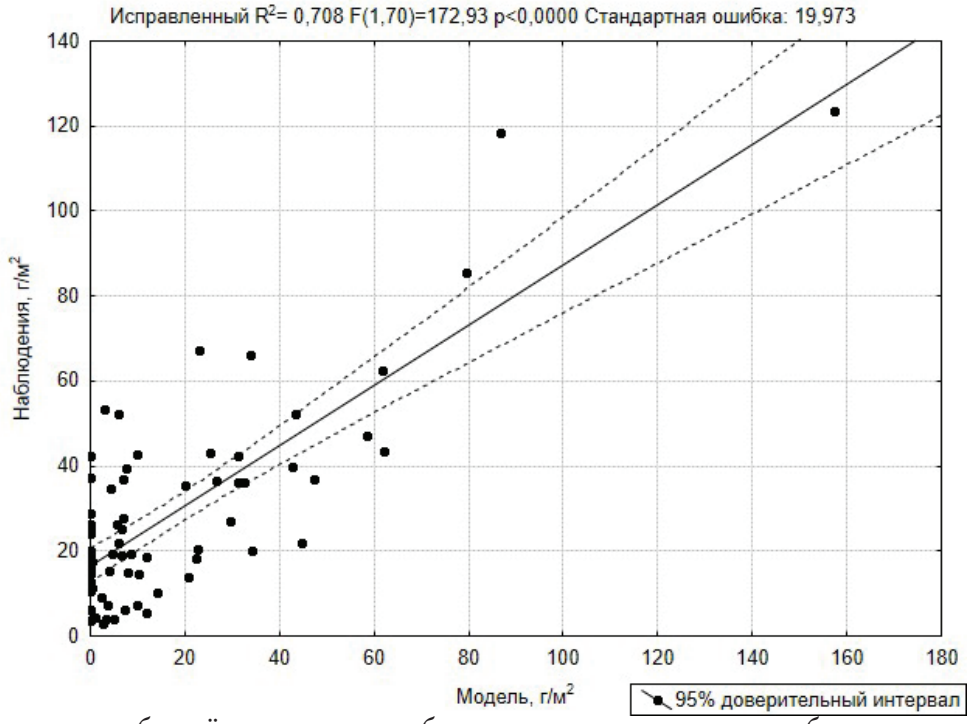

Рис. 1. Сравнение наблюдённых величин биомассы литорального зообентоса с модельными по [Rasmussen, 1988]

Fig. 1. Comparison of the observed values of littoral zoobenthos biomass with the model data by [Rasmussen, 1988]

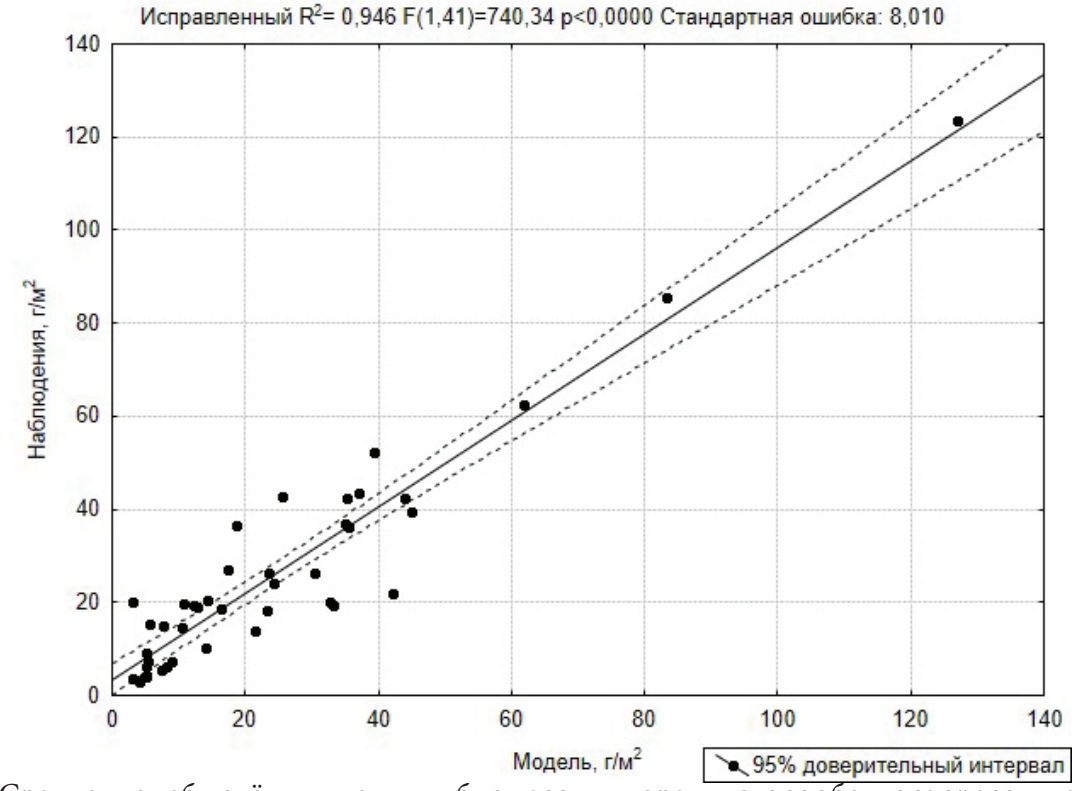

Рис. 2. Сравнение наблюдённых величин биомассы литорального зообентоса с рассчитанными по модели «Другие озера»

Fig. 2. Comparison of the observed values of littoral zoobenthos biomass with those calculated using the “Other Lakes” model

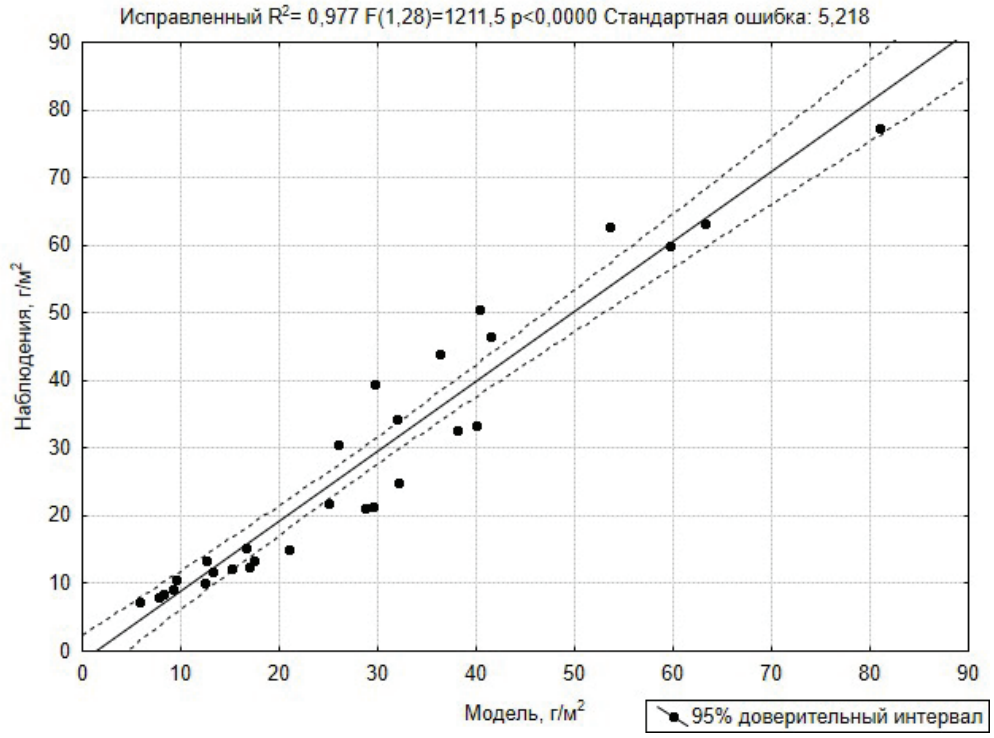

Рис. 3. Сравнение наблюдённых величин биомассы литорального зообентоса в озере Мамфремагог с рассчитанными по модели «Другие озера»

Fig. 3. Comparison of the observed values of littoral zoobenthos biomass in Lake Memphremagog with those calculated using the "Other Lakes" model

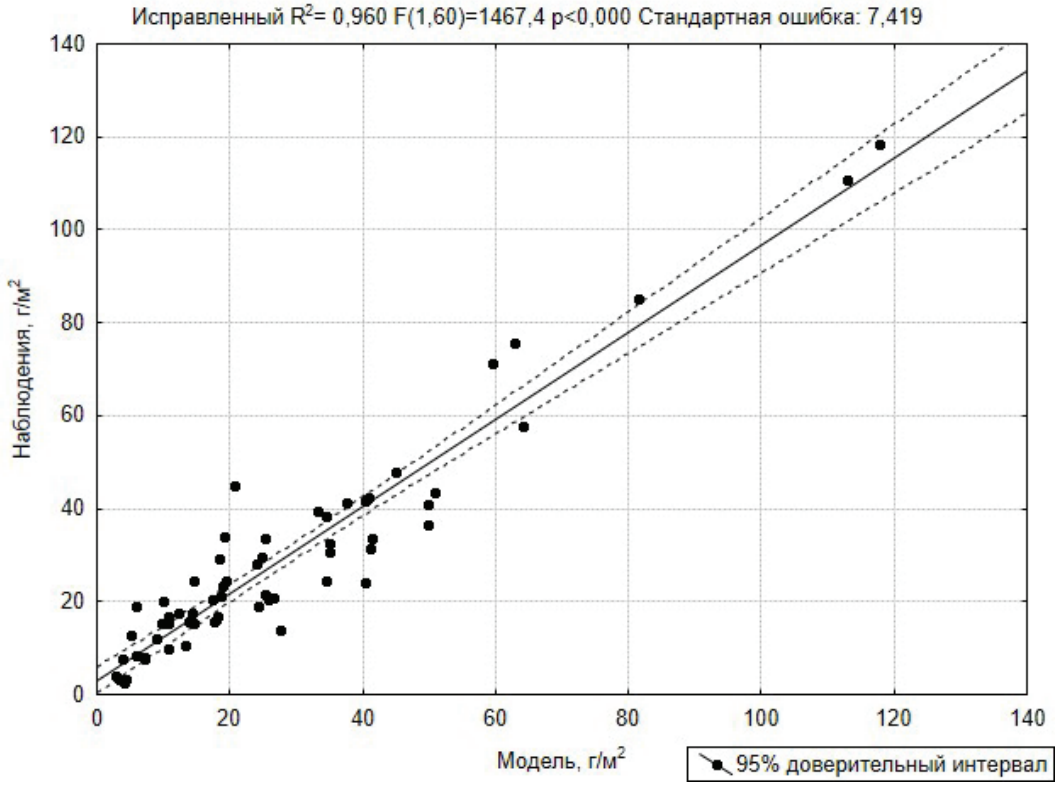

Рис. 4. Сравнение наблюдённых величин биомассы литорального зообентоса с предсказанием обобщенной модели

Fig. 4. Comparison of the observed values of littoral zoobenthos biomass of with the prediction of the generalized model

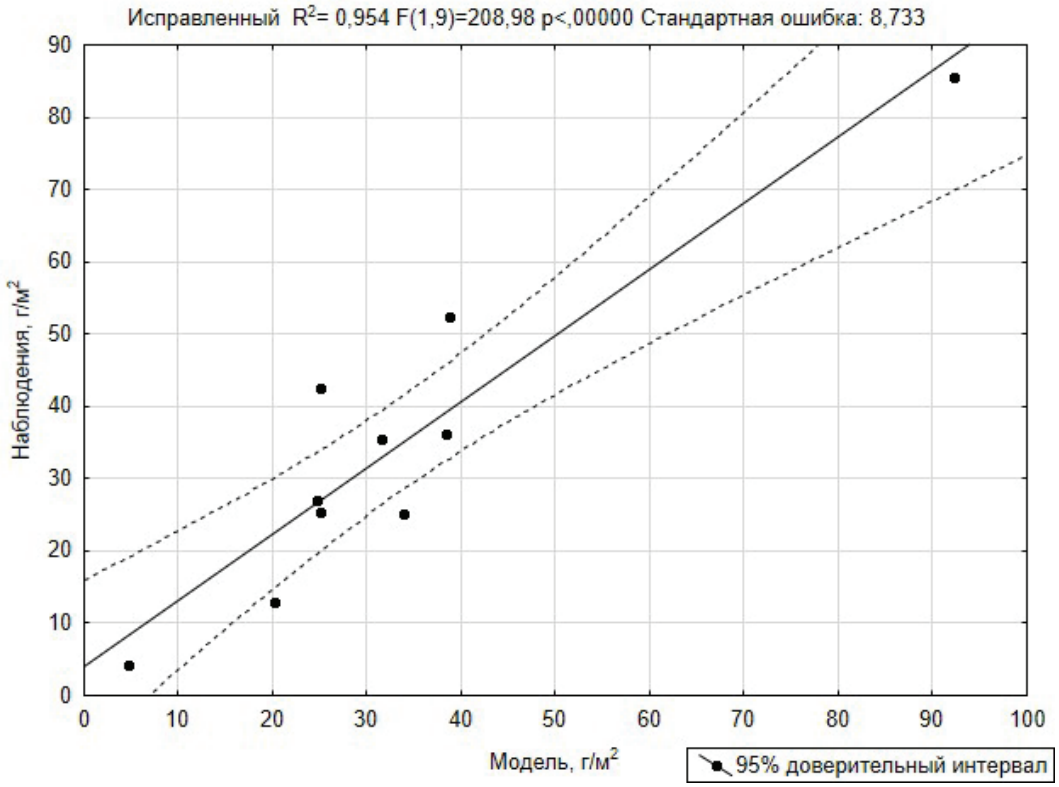

Рис. 5. Сравнение независимых наблюдённых величин биомассы литорального зообентоса с предсказанием обобщенной модели

Fig. 5. Comparison of the independent observed values of littoral zoobenthos biomass with prediction of the generalized model

меньше – всего 0,708, а вычисленная средняя абсолютная ошибка предсказания модели в процентах (MAPE) составила 64%.

Множественные регрессионные модели на основе технологии искусственных нейронных сетей

Материалом для построения нейронных регрессионных моделей послужили данные из упомянутой статьи [Rasmussen, 1988]. Данные из множества «другие озера» (n=42) были использованы для построения модели, предсказывающей биомассу литорального зообентоса, а данные озера Мемфремагог (n=29) стали множеством для проверки «работоспособности» построенной модели.

Множество «другие озера» было разделено на обучающую последовательность (80% данных) и тестовою (20%). Типом нейронной сети был выбран многослойный персептрон (MLP). Сеть автоматически подбирала функцию активации искусственных нейронов из следующих возможных: логистической, гиперболического тангенса и экспоненциальной. Из множества сгенерированных искусственной нейронной сетью моделей отбиралась наилучшая. Критерием отбора служили величины коэффициента корреляции и ошибки обучения и тестирования.

В качестве входных параметров модели (предикторов) были взяты те же, что вошли в конечное регрессионное управление (концентрация хлорофилла *a*, уклон литорали, экспозиция, концентрации ионов кальция и хлора). Из множества рассчитанных моделей выбрана наилучшая, условно названная «Другие озера», параметры которой приведены в таблице 2. Из таблицы видно, что выбраны пять входных параметров, состоит из восьми скрытых нейронов, коэффициенты корреляции (R) обучающего и тестового множеств высоки, активационные функции экспоненциальные. На рис. 2 представлен результат сравнения наблюдённых данных с модельными. Средняя абсолютная ошибка в процентах – MAPE = 29,6%.

Следующим шагом была проверка предсказательной способности модели «Другие озера» для биомассы литорального зообенто-

Таблица 2

Параметры регрессионной модели «Другие озера»

Table 2

Parameters of the "Other Lakes" regression model

Имя модели	R Обучающее множество	R Тестовое множество	Ошибка обучения	Ошибка тестирования	Функция активации скрытых нейронов	Функция активации выходного нейрона
MLP 5-8-1	0,923	0,973	0,003512	0,00722	Экспоненциальная	Экспоненциальная

Таблица 3

Параметры регрессионной обобщенной модели

Table 3

Parameters of the regression generalized model

Имя модели	R Обучающее множество	R Тестовое множество	Ошибка обучения	Ошибка тестирования	Функция активации скрытых нейронов	Функция активации выходного нейрона
MLP 5-24-1	0,956	0,943	23,5	50,2	Гиперболический тангенс	Гиперболический тангенс

са на независимой выборке, которая не принимала участия в расчёте модели, то есть озере Мемфремагог (рис. 3). Ошибка прогноза (MAPE) составила всего 16,4%.

После того, как была протестирована модель «Другие озера» на независимых данных, мы объединили оба массива данных (озеро Мемфремагог и «другие озера») в единое множество и рассчитали так называемую *обобщенную модель*[1]. Из всего множества точек отбора проб случайным образом были удалены 10 из них, которые в дальнейшем анализе послужили для проверки обобщённой модели и определения точности предсказания (рис. 4). Параметры обобщенной модели помещены в таблице 3. Средняя абсолютная ошибка в процентах обобщенной модели MAPE=26,6%.

Результаты проверки предсказательной силы обобщенной модели на независимых данных представлены на рис. 5. Наибольшая точность предсказания оценивается в 19,4%. В общей сложности верификацию обобщенной модели подобным образом мы провели пять раз каждый раз по случайным выборкам десяти пунктов наблюдений. Все регрессии оказались значимыми, а усредненное значение MAPE составило 25%

Таким образом, наш анализ прогностических качеств линейных множественных регрессионных моделей, применяемых в лимнологии, показал, что точность последних крайне низка, и значения средней абсолютной ошибки в процентах (MAPE) иногда доходят до 100% и более. Разумеется, подобные модели невозможно использовать даже для приближенных экспертных оценок параметров озерных экосистем. Вместе с тем нами было продемонстрировано, что использование технологии искусственных нейронных сетей для построения регрессионных моделей позволяет резко снизить ошибки предсказаний до уровня, не превышающего 25%. Представляется, однако, что достигнутый уровень точности прогноза в лимнологии не сможет быть улучшен, по той причине, что многие методы сбора и обработки материалов, например, методы определения обилия и биомассы фито- и зоопланктона, зообентоса и рыбного населения сами по себе недостаточно точны и, кроме того, в рассмотренных примерах невозможно учесть межгодовую изменчивость экосистемных параметров.

ЗАКЛЮЧЕНИЕ

Лимнология – наука мультидисциплинарная. Целые коллективы учёных разных специальностей и направлений (гидрологи, ги-

[1]Программный пакет Statistica предусматривает сохранение нейронных моделей только в виде файла в формате XML на языке разметки для прогнозного моделирования (Predictive Model Markup Language – PMML) и не записывается в виде уравнения. Сгенерированная *обобщенная модель* доступна по персональному запросу по электронной почте авторов.

дрофизики, гидрохимики, гидробиологи) в течение многих десятилетий по всему миру исследуют состав, структуру, продуктивность и законы функционирования озерных экосистем. Большинство подобных исследования проводится в течение довольно длительного времени, что вызвано необходимостью учета сезонной и межгодовой изменчивости параметров озерных экосистем, а также геологического, геохимического и метеорологического разнообразия территорий. Совершенно очевидно, что лимнологические исследования являются исключительно дорогостоящими и трудоемкими. Учитывая то обстоятельство, что на Земле озер насчитывается огромное множество, (117 миллионов – по данным ученых университета Упсалы [The world's lakes …, 2018] все они исследованы быть не могут физически.

Вместе с тем, многие из озер представляют определенную ценность для людей – рекреационную, рыбопромысловую, экономическую. Многие озера нуждаются в восстановлении, и часто необходимо спрогнозировать последствия деятельности человека на озерном водосборе, что требует использования надёжных и точных эмпирических моделей. Как было показано, традиционные в лимнологической практике линейные простые и множественные регрессионные модели в большинстве не справляются с этой задачей, поскольку точность их прогнозов крайне низка.

Вместе с тем, применение современных методов прогнозирования, основанное на искусственных нейросетевых регрессионных моделях, значительно повышает точность предсказаний интересующих пользователя параметров озерных экосистем. Это возможно благодаря гибкости вычислительных алгоритмов и отсутствию ограничения на форму связей между переменными, включаемых в модели. К сожалению, названный метод требует очень большого количества исходной информации, желательно собранной и обработанной по стандартным методикам и схемам. Из этого следует, что необходимо создавать банки данных, содержащие сотни записей характеристик разнотипных озер. Только анализ подобных баз данных позволит генерировать надёжные прогностические модели для оценки параметров озерных экосистем в экспертных целях, не прибегая к крупным финансовым и трудовым затратам.

ЛИТЕРАТУРА

Иконников В.Б., Кузей Л.М., Суворов Д.В., 2003. Пространственные особенности трансформации лимнических систем Белоруссии // Озерные экосистемы: биологические процессы, антропогенная трансформация, качество воды. Минск. С. 25-28.

Китаев С.П., 2007. Основы лимнологии для гидробиологов и ихтиологов. Петрозаводск: Карельский научный центр РАН. 395 с.

Нейронные сети, 2008. STATISTICA Neural Networks: методология и технология современного анализа данных/ Под редакцией В.П. Боровикова. М.: Горячая линия – Телеком. 392 с.

Трифонова И.С., 1989. Содержание хлорофилла и скорость продуцирования органического вещества в озерах с разным уровнем концентрации биогенных элементов //Трансформация органического вещества при антропогенном эвтрофировании озер. Л.: Наука. С. 78-93.

Dillon P.J., Rigler F.H., 1974. The phosphorus-chlorophyll relationship in lakes // Limnology and Oceanography, Volume 19, Issue 5. P. 767-773. https://doi.org/10.4319/lo.1974.19.5.0767

Hanson J. M., Peters R.H., 1984. Empirical prediction of crustacean zooplankton Biomass and Profundal Macrobenthos Biomass in Lakes // Can. J. Fish. Aquat. Sci., Vol. 41. P.439-445.

Rasmussen J.B., 1988. Littoral zoobenthic biomass in lakes, and its relationship to physical, chemical, and trophic factors. // Can. J. Fish. Aquat. Sci. 45. P. 1436-1447.

The world's lakes have finally been counted, 2018 www.uu.se/en/media/news/ article/?id=3637&area=2,5,10,16&typ=artikel&na=&lang=en

REFERENCES

Dillon P.J., Rigler F.H., 1974. The phosphorus-chlorophyll relationship in lakes. *Limnology and Oceanography.* Volume 19. Issue 5. P. 767-773. https://doi.org/10.4319/lo.1974.19.5.0767

Hanson J. M., Peters R.H., 1984. Empirical prediction of crustacean zooplankton Biomass and Profundal Macrobenthos Biomass in Lakes. *Can. J. Fish. Aquat. Sci.,* Vol. 41. P.439-445.

Ikonnikov V.B., Kuzey L.M., Suvorov D.V., 2003. Spatial features of the transformation of limnic systems of Belarus. *Lake ecosystems: biological processes, anthropogenic transformation, water quality*. Minsk. P. 25-28. *In Russian*.

Kitaev S.P., 2007. *Basic general limnology for hydrobiologists and ichtiologists*. Petrozavodsk, Karelian Scientific Center of RAS. 395 p. *In Russian*.

Neural networks, 2008. *STATISTICA Neural Networks: Methodology and Technology of Modern Data Analysis* / Ed. V.P. Borovikov. M.: Goriachaia linia – Telecom Publ. 392 p. *In Russian*.

Rasmussen J.B., 1988. Littoral zoobenthic biomass in lakes, and its relationship to physical, chemical, and trophic factors. *Can. J. Fish. Aquat. Sci.* 45. P. 1436-1447.

The world's lakes have finally been counted, 2018. www.uu.se/en/media/news/ article/?id=3637&area=2,5,10,16&typ=artikel&na=&lang=en

Trifonova I.S., 1989. Chlorophyll content and the rate of production of organic matter in lakes with different levels of concentration of nutrients. *Transformation of organic matter during anthropogenic eutrophication of lakes*. L.: Nauka. P. 78-93. *In Russian*.

Accepted: 18.11. 2018 ***Поступила в редакцию: 18.11. 2018***

Published: 30.12. 2018 ***Дата публикации: 30.12. 2018***