Опубликовано: ***ВОДНЫЕ РЕСУРСЫ.*** *2004. том 3. № 3. с. 354-364*

УДК 504.064: 574.587+ 574. 633

ОЦЕНКА КАЧЕСТВА ПОВЕРХНОСТНЫХ ВОД ПО ИНДИКАТОРНЫМ ВИДАМ МАКРОЗООБЕНТОСА

© 2004 г. В. К. Шитиков, Т. Д. Зинченко, Л. В. Головатюк
Институт экологии Волжского бассейна Российской академии наук
445003 Тольятти, ул. Комзина, 10
Поступила в редакцию 05.06.2002 г.

Рассматриваются математические аспекты прогнозирования класса качества вод по данным гидробиологического анализа с использованием теории распознавания образов. Выполнена серия расчетов индикаторных валентностей видов зообентоса с применением сапробного анализа методом Зелинки-Марвана [12] и на основе частоты встречаемости гидробионтов в водоемах разного типа. Предложены модели распознавания класса качества вод по зообентосу с помощью методов нелинейной оптимизации. Приведен подробный сравнительный анализ результатов, полученных на основе данных наблюдений на малых реках Самарской обл.

Основная задача биоиндикации - разработка методов и критериев, которые могли бы адекватно отражать уровень антропогенных воздействий с учетом комплексного характера загрязнения и диагностировать ранние нарушения в наиболее чувствительных компонентах биотических сообществ. Современная теория и практика биоиндикации характеризуется пространным изложением исследователем отмеченных им фактов поведения различных видов гидробионтов в конкретных условиях среды. Иногда эти описания сопровождаются выводами, основанными на чисто визуальных методах сравнения или использовании недостаточно достоверных индексов. Редки работы, где делаются практические попытки оценить лимитирующий уровень загрязнения (так называемый «анализ биологически значимых нагрузок») или выполняется «индикация», когда с использованием биоиндикаторных показателей прогнозируются неизвестные факторы среды и оценивается их значимость для всей экосистемы в ближайшем и отдаленном будущем.

Сейчас уже можно преодолеть математическую ущербность биоиндикации и широко использовать современные методы обработки многомерных измерений, так как:

- сформированы банки многолетних данных по наблюдениям за природными экосистемами;
- разработан и апробирован ряд методов и математических моделей интегральной оценки состояния сложных систем различного типа;
- развиты аппаратные и программные информационные компьютерные технологии, позволяющие анализировать необходимые массивы экологических данных.

МАТЕМАТИЧЕСКАЯ ПОСТАНОВКА ЗАДАЧИ ОЦЕНКИ КАЧЕСТВА ВОД

Одна из актуальных проблем биоиндикации - идентификация информативных компонентов водных экосистем и прогнозирование на этой основе категории качества водоема. Формальная математическая постановка этой задачи может быть выполнена следующим образом:

Пусть задано пространство признаков **X** размерностью $n>1$, соответствующее некоторому списку видов гидробионтов. Точки этого пространства - данные конкретных гидробиологических измерений $\boldsymbol{x} = \{x_1,\ldots, x_i,\ldots,x_n\}$, где x_i – значение обилия i-го вида в пробе. Пусть также известно число L классов $y_1, y_2, \ldots, y_L$, к которым относятся данные всевозможных измерений $\boldsymbol{x}$ из **X**. Предположим, что из объектов пространства **X** можно

скомплектовать обучающую последовательность – таблицу, разбитую на L непересекающихся подмножеств строк, где каждой строке $\{x_1,\ldots, x_i,\ldots,x_n\}$ поставлен в соответствие некоторый класс качества y_k, причем любому из L классов принадлежит не менее одного объекта.

Основная задача состоит в том, чтобы на основе обучающей выборки определить набор формальных решающих правил, позволяющих для произвольного измерения $\boldsymbol{x_p}$ из $\mathbf{X}$ указать класс качества y_k, к которому оно принадлежит.

Содержание задаваемой системы классификации $\{y_1, y_2, \ldots, y_L\}$ не имеет принципиального значения для последующего изложения и может быть вполне произвольным (например, любые градации сапробности, токсобности, классов качества вод, типов водоемов, природно-климатических зон).

Решить поставленную задачу можно с использованием алгоритмов распознавания образов [6], которые методически основаны на следующей гипотезе компактности: “в используемом пространстве признаков измерения, принадлежащие одному и тому же классу, близки между собой, а измерения, принадлежащие разным классам хорошо разделимы друг от друга” [11, стр.96]. Традиционной особенностью этих методов является наличие двух последовательных этапов: обучения и собственно распознавания. В процессе обучения формируется правило разделения множества объектов на несколько категорий, а при распознавании это классификационное правило используется для отнесения неизвестного объекта к одной из рассматриваемых категорий. Надежность процесса распознавания оценивается в ходе экзамена и зависит как от математической сущности алгоритма классификации, так и информационной репрезентативности выборки, использованной на стадии обучения.

В настоящей работе с помощью метода распознавания образов:

на этапе обучения рассчитывается матрица оценок $\mathbf{R}$, элементы которой r_{ik} являются тем больше, чем выше индикаторная ценность вида i для класса k;

на этапе экзамена подбирается такой алгоритм распознавания F, т.е. последовательность формул преобразования матрицы $\mathbf{R}$ в вектор результирующих оценок классов $\boldsymbol{t}$:

$$F(\mathbf{R}, \boldsymbol{x_p}) \Rightarrow \boldsymbol{t}\ , \qquad (1)$$

который обеспечивает минимум ошибок распознавания для широкого набора тестовых примеров $\boldsymbol{x_p}$.. Элементы матрицы $\mathbf{R}$ будем называть *индикаторными валентностями* (value of indication) вида i в классе k.

Рассмотрим три возможных варианта расчета индикаторных валентностей, основанных на различных концепциях:

- использование техники сапробных валентностей М.Зелинки и П.Марвана [12];
- вероятностные оценки на основе частот встречаемости видов в водоемах разной категории;
- сбалансированные индикаторные валентности, полученные с использованием методов оптимизации.

ЭКСПЕРИМЕНТАЛЬНЫЙ МАТЕРИАЛ

Индикаторные валентности рассчитывали на основе данных гидробиологических наблюдений [10], проведенных на 34 малых реках разных типов и уровней антропогенной нагрузки, которые расположены в степной и лесостепной зонах Самарской обл. Массив данных включал в себя 542 пробы, взятые на 247 станциях с 1987 по 2001 г. Для каждой станции наблюдения по совокупности гидрохимических показателей оценивали класс качества вод в соответствии с методиками [3, 9]. В качестве объектов биологического мониторинга в отобранных точках наблюдения использовали значения численности и биомассы 546 видов зообентоса, принадлежащих к различным таксономическим группам.

РАСЧЕТЫ ПО ФОРМУЛЕ ВЫЧИСЛЕНИЯ САПРОБНЫХ ВАЛЕНТНОСТЕЙ

Оценка зон сапробности по показательным организмам методом М.Зелинки и П.Марвана основывается на следующих предпосылках [12]:

– для каждого *i*-го вида гидробионтов устанавливаются значения сапробных валентностей a_{ik}, которые теоретически совпадают с оценками распределения вероятности встречаемости вида в каждой *k*-й ступени сапробности и выражаются одной или несколькими цифрами, сумма которых равна 10;
– вводится шкала индикаторного веса J_i, выраженного в баллах (от 1 до 5) и оценивающего роль (дискриминирующую важность) *i*-го вида при оценке степени загрязнения;
– для произвольной гидробиологической пробы, в которой измерены значения обилия видов $G_{i,}$, средневзвешенная сапробная валентность сообщества t_k, рассчитанная как

$$t_k = \sum_i a_{ik} G_i J_i \,/\, \sum_i G_i J_i \,, \qquad (2)$$

- эффективная и несмещенная оценка принадлежности пробы к *k*-й зоне сапробности.

Сапробные валентности и индикаторные веса обычно оценивают с помощью гистограмм количественного распределения показательных организмов по выделенным зонам сапробности [1, 14]. Однако использование «стандартизованных» значений a_{ik}, представленных в различных публикациях, среди которых видное место занимает основополагающая работа В.Сладечека [18], приводит во многих случаях к неудовлетворительным результатам по ряду следующих очевидных причин:

- сапробные валентности принимались с учетом реакции гидробионтов на разложение органических веществ естественного происхождения, в то время как появление в воде токсических веществ (минеральных или органических ингредиентов) создает обстановку отравления организмов, не имеющую себе аналогов в ступенях сапробности;
- в различных географических регионах и в различных типах водоемов сапробиологические характеристики видов не остаются постоянными, что может сильно исказить результаты анализов. При изменении пространственно-временных условий эти показатели могут претерпеть серьезные модификации: появляются новые виды-индикаторы, либо коренным образом изменяется индикаторная сущность уже имеющихся видов;
- оценка валентностей часто связана с субъективным подходом исследователей: один и тот же вид животных разными авторами характеризуется от олиго- до полисапробного [1];
- для многих бентосных организмов сапробных валентностей нет в доступных источниках (например, из 546 видов макрозообентоса малых рек Самарской обл. в литературе представлены значения a_{ik} только по 62 видам);
- техника классификации зон сапробности изобилует чисто математическими неточностями.

В то же время, как считает П.Шреверс [17], «вместе с водой из купели не следует выплескивать и младенца»: подвергая законному сомнению надежность готовых таблиц индикаторных видов, полученных 40 лет назад на реках Средней Европы, нельзя распространять эту критику на сам метод, который остается единственным практическим способом количественной водной биоиндикации. Поэтому понятно стремление исследователей-гидробиологов, накопивших значительный массив экспедиционных данных, провести самостоятельные расчеты сапробных валентностей a_{ik} с учетом региональных особенностей бентофауны, характера загрязнений и типологии водоемов. Например, П.А.Цимдинем [5] была предложена формула расчета

$$a_{ik} = 10 N_{ik} D_{ik} / \sum_{k=p}^{o} N_{ik} D_{ik}$$

с использованием численности N_{ik} и встречаемости D_{ik} гидробионта i-го вида в основных зонах сапробности (о-р). Эту формулу легко обобщить, положив $N_{ik} = \sum_{j=1}^{m_{jk}} G_{ijk}$, где m_{ik} – встречаемость i-го вида в k-й зоне, $D_{ik} = m_{ik} / m$, m – общее число измерений, G_{ijk} - произвольный показатель обилия вида-индикатора (численность, биомасса, проективное покрытие или др.) Тогда выражение для сапробных валентностей приобретает тривиальный вид отношения групповых средних G_{ijk} к глобальной средней:

$$a_{ik} = 10 \sum_{j=1}^{m_{jk}} G_{ijk} / \sum_{k=p}^{o} \sum_{j=1}^{m_{jk}} G_{ijk} \ . \qquad (3)$$

Значения индикаторных весов J_i устанавливают, ориентируясь на характер распределения сапробных валентностей по зонам [18]. Например, индикаторный вес J = 5 присваивается «хорошим» индикаторам, если все 10 баллов a_{ik} сосредоточены в одной зоне сапробности. Если валентности равномерно распределяются по ступеням, то такие виды считаются индифферентными или «плохими» индикаторами и получают небольшой балл. Нами индикаторные веса вычислялись по уравнению регрессии, связывающему J_i с энтропией h_i, которая математически строго характеризует равномерность распределения валентностей по классам и рассчитывается по известной формуле К.Шеннона:

$$h_i = \sum_{k=2}^{6} a_{ik} \log_2 (a_{ik}) .$$

Используем формулы (2) и (3) для прогнозирования класса качества вод на основе видовых численностей зообентоса ($G_i = N_i$ экз/м^2) по аналогии с анализом зон сапробности. На этом этапе примем тождественными понятия “индикаторная валентность” и “сапробная валентность” (**R ≡ A**). Для некоторых видов рассчитанные значения J_i и a_{ik} представлены в табл. 1. По первым пяти видам приведена более подробная информация, позволяющая детально анализировать механизм расчета. Можно отметить следующие свойства оценок a_{ik}, вычисленных по формуле (2):

- индикаторные валентности зависят только от соотношения групповых средних численности и никак не связаны с абсолютными значениями обилия (вид, представленный единичными экземплярами, будет получать такие же оценки, что и вид, встречающийся в пробах тысячами особей);
- как и любые оценки, основанные на средних, валентности сильно зависят от характера распределения данных, наличия аномальных «выбросов» данных и т.д. Так для вида Polypedilum scalaenum (Sch.) большое значение валентности четвертого класса (a_4= 8.5) во многом определяется одним измерением с N = 17 840 экз/м^2;
- большим преимуществом в формировании оценок обладают редкие виды, в том числе виды, встретившиеся только в одной пробе, например, Eukiefferiella minor (Edw.). Они автоматически приобретают максимальный вес J = 5 и максимальную валентность a = 10.

Рассчитанные индикаторные валентности a_{ik} использовались на втором этапе для прогнозирования класса водоема по комбинации из m_p видов, встретившихся в произвольно взятой пробе. Предварительно по формуле (2) для каждого k-го класса (k = 2 - 6) вычислялись средневзвешенные сапробные валентности t_k по численностям видов N_i (i = 1,…,m_p) с учетом индикаторных весов J_i. Пример расчета по данным, полученным на участке

р. Сок, относящемся ко 2 классу качества вод [3, 7, 9], приведен в табл. 2. Индикаторные валентности и индикаторные веса для всех семи видов зообентоса, найденных в этой пробе, представлены в табл. 1.

Получив значения t_k, имеющие смысл оценок принадлежности к классам качества вод, для реализации алгоритма (1) можно избрать одну из двух стратегий классифицирования. Первая стратегия (наиболее простая и традиционно рекомендуемая математической теорией распознавания образов) предлагает отнести объект к классу, набравшему максимальную оценку, что соответствует минимальному среднему риску ошибки. В случае для ст.5 на р. Сок такой выбор безошибочен и не предполагает иных толкований: $t_2 = 5.3$ и второй класс качества.

Вторая стратегия основана на аналогиях с расчетом индекса сапробности S_R, различные модификации которого предлагались Р.Пантле, Г.Букком, Дж.Ротшайном и др. И.К.Тодераш [14] предложил следующую формулу пересчета средневзвешенных сапробных валентностей $\{t_x, t_o, t_\beta, t_\alpha, t_p\}$ в индекс сапробности:

$$S_R = 0\, t_x, + \ 1\, t_o + 2\, t_\beta + 3\, t_\alpha \ + 4\, t_p.$$

Используем установленное ГОСТом [7] соответствие между классами качества вод и зонами сапробности: $\cdot t_2 = \cdot t_o$ для олигосапробной, $t_3 = t_\beta$ для β - мезосапробной, $t_4 = t_\alpha$ для α - мезосапробной и $\cdot t_5 = t_p$ для полисапробной зон. Подставив рассчитанные значения t_k в формулу [14], получим для примера, рассматриваемого в табл. 2, индекс сапробности:

$$S_R = 1\times0.53 + 2\times0.04 + 3\times0.24 + 4\times0.01 + 5\times0.19 = 2.25,$$

который также уверенно, но уже ошибочно относит тестируемый водоем к третьему классу (т.е. β - мезосапробной зоне, $S_R = 1.5 - 2.5$).

Распределение численности и биомассы организмов в пробах характеризуется определенными статистическими законами. Анализ показал [15], что для бентосных организмов наиболее характерно логнормальное распределение, а следовательно, логарифмирование численности существенно нормализует распределение значений N_i и обеспечивает корректное вычисление статистических характеристик: средней, дисперсии и т.д. В соответствии с этим предположением, был выполнен вариант расчета валентностей a_{ik} и t_k по формулам (2 – 3), использующим предварительное логарифмирование показателей обилия : $G_i = \ln(N_i + 1)$. Добавление единицы вызвано необходимостью учета единичных экземпляров, т.к. $\ln(1) = 0$.

Для сравнения работоспособности отдельных модификаций метода проводился тест в режиме скользящего контроля: из обучающей выборки поочередно удалялись описания одного объекта, на оставшемся материале вычислялись оценки a_{ij} и по ним классифицировался исключенный объект. Такая процедура повторялась (m - 1) раз. Доля ошибок классификации при скользящем контроле - несмещенная оценка вероятности ошибки на всей генеральной совокупности [2] и, следовательно, наиболее репрезентативная оценка качества алгоритма распознавания. Результаты теста на 542 пробах зообентоса тремя описанными способами классифицирования приведены в табл. 3 (алгоритмы 1.1-1.3).

По результатам расчета можно сделать следующие выводы:

- подтверждается предположение о недостаточной математической состоятельности подхода Пантле-Букка и всей техники вычисления индексов сапробности. Попытка выразить одним числом некоторую функцию распределения вероятностей принадлежности измерения к четырем зонам сапробности (или шести классам качества) неизбежно приводит к осреднению факторов видовой структуры биоценоза, а следовательно, к сильному смещению прогноза к некоторому «сапробному центру тяжести» $S_R = 2 - 3$. В частности, алгоритмом 1.1 (табл. 3) 371 наблюдение из 542 было отнесено к «среднему» четвертому классу (α - мезосапробной зоне), причем в 178 случаях это было сделано ошибочно.

Таблица 1. Сапробные показатели классов качества водоемов, рассчитанные для некоторых видов зообентоса (ΣN_{ik} , m_{ik} и a_{ik} - сумма численностей особей, встречаемость и сапробная валентность i-го вида в пробах k-го класса соответственно; m_k – общее число проб k-го класса; J_i - индикаторный вес i-го вида, прочерк здесь и в табл. 2, 4 соответствует нулевому значению)

Виды (таксоны)		Классы качества					Итого	J_i
		2	3	4	5	6		
Parachironomus varus	ΣN_{ik}	63	107	490	603	1080	2343	1.4
	m_{ik}	2	2	5	4	4	17	
	m_{ik}/m_k, %	3.6	1.5	2.6	4.3	5.8	3.14	
	a_{ik}	0.3	0.5	2.1	2.6	4.6	10	
Parametriocnemus lundbecki	ΣN_{ik}	628	316	40	-	-	984	3.2
	m_{ik}	8	4	3	-	-	15	
	m_{ik}/m_k, %	14.3	3.1	1.6	-	-	2.76	
	a_{ik}	6.4	3.2	0.4	-	-	10	
Polypedilum scalaenum	ΣN_{ik}	60	3239	21697	290	320	25606	4.2
	m_{ik}	3	15	22	4	2	46	
	m_{ik}/m_k, %	5.4	11.5	11.4	4.3	2.9	8.5	
	a_{ik}		1.3	8.5	0.1	0.1	10	
Eukiefferiella minor	ΣN_{ik}	10	-	-	-	-	10	5.0
	m_{ik}	1	-	-	-	-	1	
	m_{ik}/m_k, %	1.8	-	-	-	-	0.2	
	a_{ik}	10.0	-	-	-	-	10	
Eukiefferiella gr. claripennis		3.0	0.3	4.2	2.6	-	10	1.7
Cricotopus bicinctus		3.3	4.9	1.6	0.1	-	10	2.0
Cricotopus gr. sylvestris		2.7	0.0	2.9	0.0	4.4	10	2.0
Baetis rhodani		8.0	0.4	1.6	-	-	10	3.9
Ephemeroptera (прочие)		6.4	1.1	2.2	0.3	-	10	2.4
Simulium sp.		1.0	1.7	7.2	0.1	-	10	3.1
m_k		56	131	193	93	69	542	

Таблица 2. Расчет средневзвешенных сапробных валентностей t_k на примере данных экспедиционных наблюдений на ст. 5 р. Сок от 30.07.1999 (N_i численность особей i-го вида, J_i и a_{ik} - индикаторные веса и сапробные валентности k-го класса из табл. 1)

Виды (таксоны)	N_i	$N_i\, a_{ik} J_i$, $k = 2 - 6$					$N_i J_i$
		2	3	4	5	6	
Eukiefferiella minor	10	500	-	-	-	-	50
Cricotopus gr.sylvestris	150	829	-	884	-	1360	307
Cricotopus bicinctus	10	66	97	33	2	1	20
Eukiefferiella gr.claripennis	10	50	4	70	44	-	17
Simulium sp.	10	31	53	220	3	-	31
Ephemeroptera (прочие)	10	155	28	53	7	-	24
Baetis rhodani	70	2199	103	454	-	-	276
Сумма по всем видам		3830	285	1714	56	1361	725
$t_k = \Sigma\, N_i\, a_{ik} J_i / \Sigma\, N_i J_i$		5.3	0.4	2.4	0.1	1.9	

Таблица 3. Сравнительный анализ адекватности различных алгоритмов вычисления оценок индикаторных валентностей и техник распознавания (числитель – число, знаменатель - %)

Алгоритм вычисления оценок принадлежности к классам	На всей выборке			Без учета проб из 6 класса	Критерий оптимизации D^2
	Правильно распознано	Ошибочно распознано	Ошибка на 2 класса и более		
1.1. По натуральным численностям и по индексам сапробности	260 48.0	282 52.0	90 16.6	259 54.8	691
1.2. По натуральным численностям и по максимумам оценок	315 58.1	227 41.9	88 16.2	307 64.9	607
1.3. По логарифмам численностям и по максимумам оценок	341 62.9	201 37.1	84 15.5	338 71.5	558
2.1. Простое усреднение оценок вероятности	319 58.9	223 41.1	86 15.9	317 67.0	571
2.2. Усреднение вероятностей с угловым преобразованием Фишера	366 67.5	176 32.5	81 14.9	350 74.0	580
3.1. Полная черехпараметрическая модель оптимальных оценок	391 74.9	131 25.1	41 7.9	359 78.6	274
3.2. Трехпараметрическая модель ($\alpha = 0$, $\beta = 0.41$, $\gamma = 1.21$, $\lambda = 0.71$)	370 70.9	152 29.1	54 10.3	345 75.5	345
3.3. Трехпараметрическая модель ($\lambda = 0$, $\alpha = -0.53$, $\beta = 0.39$, $\gamma = 1.189$)	354 67.8	168 32.2	77 14.8	343 75.1	499
3.4. Двухпараметрическая модель ($\lambda = 0$, $\alpha = 0$, $\beta = 0.35$, $\gamma = 1.225$)	345 66.1	177 33.9	76 14.6	339 74.2	500
3.5.Двухпараметрическая модель ($\lambda = 0$, $\beta = 0$, $\alpha = -0.47$, $\gamma = 1.185$)	345 66.1	177 33.9	79 15.1	340 74.4	514
3.6.Однопараметрическая модель ($\lambda = 0$, $\alpha = 0$, $\beta = 0$, $\gamma = 1.2$)	329 63.0	193 37.0	82 15.7	325 71.1	534

- вектор значений валентностей Зелинки–Марвана $\boldsymbol{t}$ более полно отражает индикаторное значение видов в сообществе. Принадлежность пробы x_p к водоему k-й категории целесообразнее определять по max t_k – наибольшей средневзвешенной валентности из $\{t_x, t_o, t_\beta, t_\alpha, t_p\}$.
- выбросы обучающей выборки (аномально высокие значения численностей по некоторым видам в отдельных пробах) могут сильно сказываться на устойчивости расчета индикаторных валентностей. Этим можно объяснить относительно большое для алгоритма 1.2 число «грубых» ошибок прогноза (на два класса и более). При использовании предварительно логарифмированных значений N_i результаты можно назвать вполне стабилизировавшимися (см. алгоритм 1.3 в табл.3). Для сравнения приведем значения индикаторных валентностей, рассчитанных на основе $G_i = \ln(N_i + 1)$ для вида Polypedilum scalaenum (Sch.) , которые сильно отличаются от представленных в табл. 1: $\boldsymbol{A} = \{0.5,\ 3.2,\ 5.1,\ 0.8,\ 0.5\}$ при $J = 1.6$.
- весьма неудовлетворительными оказались результаты экзамена для проб, взятых в загрязненных водоемах пятого и шестого классов, на которые преимущественно пришлась доля ошибок. С одной стороны, сказывается некоторая жесткость санитарно-гигиенического подхода к оценке класса качества по гидрохимическим показателям, когда по некоторому лимитирующему фактору, например концентрации поверхностно-активных веществ, водоем относится к шестому классу, что не мешает развиваться структурно деформированному, но количественно полноценному сообществу гидробионтов. С другой стороны, ошибки определяются и чисто статистическими эффектами: того небольшого числа видов-эврибионтов, обычно характерного для грязных водоемов, недостаточно для точной идентификации, поскольку эти виды имеют невысокие индикаторные веса J и валентности, вносящие равномерный вклад в средневзвешенные оценки t_k.

ВЕРОЯТНОСТНЫЕ ОЦЕНКИ, ПОЛУЧЕННЫЕ С ИСПОЛЬЗОВАНИЕМ ЧАСТОТ ВСТРЕЧАЕМОСТИ

Для гидробиологии традиционен анализ видовой встречаемости, когда для исследователя имеет значение только факт наличия вида в пробе. Такой подход, например, широко используется в кластерном анализе для оценки общности видового состава с использованием коэффициента сходства Т.Съёренсена [12, 13]. Поэтому целесообразно рассмотреть вариант расчета индикаторных валентностей на основе частот встречаемости.

Сформируем матрицу гидробиологических наблюдений **X** в альтернативной шкале измерений, положив все ее элементы равными единице, если значение численности i-го вида в j-й пробе не меньше некоторого заданного порога $N_{пор}$, и нулю в противном случае. Эта операция позволяет абстрагироваться от абсолютных значений обилия видов, нестационарных по своей природе, и использовать в качестве байесовских оценок условных вероятностей более устойчивые и унифицированные значения частот встречаемости.

На первом этапе с использованием примеров обучающей выборки, для которых известны значения классов качества вод $y_1, y_2, \ldots, y_L$, сформируем:

- матрицу **P** оценок условных вероятностей класса k для вида i ($p_{ik} = m_{ik} / m_i$);
- вектор-столбец оценок априорной вероятности вида i ($p_i = m_i / m$);
- вектор-строку оценок априорной вероятности класса k ($p_k = m_k / m$),

где m - общее число примеров обучающей выборки; m_i - число измерений, где встретился вид i; m_k - число объектов, принадлежащих к классу k; m_{ik} - число измерений класса k, где встретился вид i. Значения p_i и p_{ik} для первых четырех видов приведены в табл. 1.

Как и в предыдущем случае, на втором этапе для прогнозирования класса водоема по комбинации видов, встретившихся в произвольно взятой пробе x_p , и найденным значениям p_{ik} априорных вероятностей вычисляется вектор $\boldsymbol{t}$ результативных оценок (1).

Вывод о принадлежности пробы к водоему k-й категории делается, если t_k – наибольшая оценка из $t_1, t_2, \ldots, t_L$.

Простейший вариант расчета оценок t_k - осреднение значений условных вероятностей для всех m_p видов, встретившихся в тестируемом измерении:

$$t_k = \frac{1}{m_p} \sum_{i=1}^{m_p} p_{ik} \ ,$$

т.е. вероятность принадлежности пробы к k-му классу есть средняя вероятность класса k для всех видов, найденных в пробе.

Формула «простых средних вероятностей», традиционная для многих работ в этом направлении, часто дает вполне удовлетворительную точность. Однако ряд теоретико-вероятностных предположений заставляет усомниться в конечной оптимальности аппроксимации первого порядка. Поэтому другой вариант осреднения вероятностей p_{ik} - использование известного углового преобразования Р.Фишера [8], при котором частотные оценки вероятностей имеют ошибку, почти не зависящую от самих вероятностей. Кроме того, функция $\arcsin(2P - 1)$ ведет себя почти так же, как используемая в байесовских процессах функция $\ln(P/(1-P))$, но в то же время при P, близких к нулю или единице, не вырождается, устремляясь в бесконечность, а ограничена $\pm\pi/2$. На этих принципах основана работа компьютерной системы PASS [16], прогнозирующей спектр биологической активности химических соединений по их структурным формулам. При угловом преобразовании результирующие оценки принадлежности к классам тестируемого измерения $\boldsymbol{x_p}$ выражаются через условные вероятности p_{ik} и априорные вероятности классов p_k следующим образом:

$$t_k = \frac{1}{m_p} \sum_{i=1}^{m_p} J_i [\arcsin(2p_{ik} - 1) - \arcsin(2p_k - 1)] \ ,$$

где J_i – индикаторный вес вида i, который интерпретировался как величина, обратная к «шенноновской» энтропии распределения вероятностей по классам:

$$J_i = 1/[1 - \sum_{k=2}^{6} p_{ik} \log_2(p_{ik})] \ .$$

Результаты теста скользящего контроля на 542 пробах зообентоса по формуле «простых средних вероятностей» (алгоритм 2.1) и с использованием углового преобразования Р.Фишера (алгоритм 2.2) представлены в табл. 3. Расчеты показывают, что оценки априорных вероятностей, полученные на основе частот встречаемости видов и не учитывающие абсолютные значения численностей организмов, оказались значительно лучше оценок индикаторных валентностей, рассчитанных с использованием формулы П.А.Цимдиня (табл.3, п. 1.1, 1.2). Это еще одно подтверждение известного тезиса о некорректности сопоставления средних значений без учета закона статистического распределения выборок.

РАСЧЕТ ИНДИКАТОРНЫХ ВАЛЕНТНОСТЕЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ОПТИМИЗАЦИИ

Сделаем основные исходные предположения относительно природы индикаторных валентностей R_{ik}, оценивающих сродство (т.е. резистентность или экологическую толерантность) i-го вида гидробионтов к воде k-го класса качества. Будем считать, что их значения является некоторой сложной математической функцией F от следующих факторов:

$$R_{ik} = F\ (\ m_i\ , p_k, J_i, k, a_{ik}\ ),$$

где:

- m_i - встречаемость вида, для которого находится R_{ik}. С точки зрения классической статистики, чем больше m_i, тем выше надежность и устойчивость рассчитываемых оценок. Однако гидробиологический опыт свидетельствует о том, что целостная картина наиболее характерных черт биоценоза определяется в первую очередь редкими видами с m_i , близким к единице, в то время, как массовые виды-эврибионты составляют размытый фон. Встречаемость различных видов зообентоса на данной обучающей выборке из 542 измерений составляла от 1 до 226.
- p_k - относительная вероятность оцениваемого класса. Поскольку частота наблюдений в водоемах различных классов неодинакова, то разумно предположить, что оценки R_{ik} для «редких» классов должны обладать некоторыми преимуществами при их сопоставлении. В качестве конкретной дефиниции этой вероятности приняли величину m/m_k, которая в эксперименте менялась от 2.8 для наиболее многочисленного четвертого класса до 9.7 для второго класса.
- J_i - индикаторная значимость вида i, которую, как было показано выше, можно интерпретировать как величину, обратную к энтропии h_i распределения оценок по классам. Поскольку для пяти классов $\max(h_i) = 2.33$, приняли $J_i = 3.33/(h_i + 1) = 1 \div 3.33$.
- k - порядковый номер класса. Поскольку биоразнообразие и обилие гидробионтов существенно уменьшаются с увеличением градации класса качества воды, разумно предположить, что некоторым приоритетом должны обладать оценки пятого и шестого классов.
- a_{ik} - оценки относительного вклада вида i в развитие биоценозов, характерных для каждого k-го класса качества вод. По аналогии с сапробными валентностями a_{ik} рассчитывали по формуле (3), в которую подставляли показатель обилия $G_i = \ln(\sqrt{B_i N_i})$, учитывающий как численность N_i, так и биомассу видов B_i мг/м^2 в логарифмической шкале.

В расчете была принята мультипликативная модель обобщенных оценок, в которой были учтены все пять перечисленных факторов:

$$R_{ik} = a_{ik}\,(m_i)^{\alpha}\,(m/m_k)^{\beta}\,(3.33/(h_i+1))^{\gamma}\,k^{\lambda} \quad , \qquad (4)$$

где α, β, γ и λ - параметры модели, т.е. некоторые специально подобранные коэффициенты.

Для определения качества водоема по комбинации видов, встретившихся в произвольно взятой пробе, как и в предыдущих версиях, оценки t_k принадлежности к классам рассчитывали по формуле:

$$t_k = \sum_{i=1}^{m_p} G_i R_{ik} \Big/ \sum_{i=1}^{m_p} G_i \text{ , где } G_i = \ln(\sqrt{B_i N_i}).$$

Тестируемый вектор наблюдений относили к классу, которому соответствует максимальная оценка t_k .

Оптимальность значений индикаторных валентностей, а следовательно, и достоверность процесса распознавания, зависит от подбора настроечных коэффициентов α, β, γ и λ, которые регулируют долю участия каждого из пяти перечисленных факторов в формировании оценок R_{ik}. Поэтому ставилась следующая задача поиска экстремума: необходимо получить такие значения α, β, γ и λ модели (4), которые сводят к минимуму число ошибок классификации всех m примеров обучающей выборки в режиме скользящего контроля:

$$D^2 = \sum_{j=1}^{m} (Y_j^{\phi} - Y_j^{p})^2 \Rightarrow \min ,$$

где Y_j^{ϕ} и Y_j^{p} – расчетные и фактические значения класса качества вод.

Решение этой задачи не может быть представлено в аналитическом виде или по крайней мере сведено к системе конечных линейных уравнений. Поэтому минимизация функционала D^2 может быть выполнена только специальными численными методами нелинейного программирования. Для нахождения оптимальных значений коэффициентов α, β, γ и λ воспользуемся симплексным методом, обеспечивающим достаточно быструю сходимость к экстремуму с использованием выпуклого многогранника. Модификация Нелдера-Мида [4] предполагает автоматическое изменение размеров ребер симплекса, что обеспечивает эффективное преодоление «оврагов» и быстрое движение по пологим спускам.

При расчете полной модели (4) в качестве начальных приближений были заданы наиболее ожидаемые значения ($\alpha = 0.2$, $\beta = 0$, $\gamma = 1$ и $\lambda = 0.5$). В ходе симплекс-процедуры просчитано 160 вариантов матрицы **R** размерностью 546×5 и для каждого варианта было найдено число ошибок классификации D^2, пока, наконец, не получено оптимальное решение: ($\alpha = -0.52$, $\beta = 0.41$, $\gamma = 1.21$, $\lambda = 0.68$) и формула (4) приобрела вид:

$$R_{ik} = \{a_{ik}\,(m/m_k)^{0.41}\,[3.33/(h_i+1)]^{1.21}\,k^{0.68}\} \,/\, m_i^{0.52}$$

Поскольку переусложнение расчетной модели так же вредно, как и ее недоусложнение, была оценена степень влияния каждого из составляющих факторов на конечный результат классификации. Для этого выполнялась серия расчетов матрицы оптимальных оценок **R** по моделям разной степени сложности, т.е. в уравнении (4) поочередно исключался один или несколько факторов, а значения соответствующих параметров α, β или λ априори принимались равным нулю.

Полученные результаты представлены в табл. 3 (алгоритмы 3.1-3.6). Значения рассчитанных коэффициентов λ, β и γ полностью подтверждают исходные предпосылки, в то время как отрицательное значение коэффициента α приводит к нетрадиционному для статистики, но фундаментальному для гидробиологии выводу: предпочтение в оценках следует отдавать редким видам. Серия расчетов с исключением из общей формулы (4) отдельных факторов дает уверенное основание полагать, что все компоненты модели являются информативными для прогноза класса качества вод, поскольку снижение эффективности распознавания при элиминации любого из коэффициентов весьма существенно. Однопараметрическая модель оценок 3.6 методически полностью соответствует расчетам сапробных валентностей по алгоритму 1.3 и отличается лишь использованием комбинированного показателя обилия $(NB)^{0.5}$ вместо численности экземпляров зообентоса N.

В традиционном анализе сапробности методом М.Зелинки и П.Марвана вводится жесткое условие нормировки – сумма сапробных валентностей должна быть равна 10. По-видимому, смысл введения этого условия связан со стремлением непременно пересчитывать средневзвешенные валентности во вторичные по идеологии и нерезультативные на практике индексы сапробности Пантле-Букка. В предлагаемом методе авторы отказались от этого условия, чтобы получить другую, гораздо более привлекательную возможность – сравнивать между собой отдельные виды по их индикаторной значимости в данной зоне сапробности или водоемах определенного класса качества. Если сопоставить для отдельных видов характер распределения R_{ik} по классам и сапробных валентностей, взятых из литературных источников, то можно отметить неплохое соответствие оценок для чистых водоемов, в то время как виды, объявленные классическими полисапробами, в данных региональных условиях оказались эврибионтами и не набрали высокой индикаторной значимости.

Не имея технической возможности представить полный список всех 546 видов зообентоса, в табл. 4 приводятся индикаторные валентности для некоторых видов хирономид (Chironomidae, Diptera), сгруппированные по их индикаторной значимости в отдельных классах качества вод.

Таблица 4. Индикаторные валентности некоторых видов хирономид, характерные для отдельных классов качества малых и средних рек Самарской обл.

Виды хирономид	Индикаторные валентности классов					Встречаемость
	2	3	4	5	6	
Индикаторы шестого класса						
Sergentia gr. longiventris	-	-	-	-	9.3	1
Stenochironomus sp.	-	-	-	-	9.3	1
Glyptotendipes paripes	-	-	0.9	-	1.5	2
Cricotopus gr. intersectus	0.6	-	0.4	-	0.6	3
Индикаторы пятого класса						
Lipiniella agrayloides	-	-	-	8.3	-	1
Endochironomus sp.	-	-	-	5.8	-	2
Endochironomus donatoris	-	-	0.9	1.4	-	2
Ablabesmyia longistyla	-	-	0.9	1.3	-	2
Индикаторы четвертого класса						
Lymnophyes sp.	-	-	6.3	-	-	1
Brillia sp.	-	-	4.3	-	-	2
Monopelopia sp.	-	-	4.3	-	-	2
Polypedilum scalaenum	-	-	4.3	-	-	2
Rheopelopia sp.	-	-	4.3	-	-	2
Pagastia sp.	-	-	4.3	-	-	2
Glyptotendipes barbipes	-	-	3.5	-	-	3
Microtendipes gr. rydalensis	1.1	-	1.3	-	-	2
Индикаторы третьего класса						
Psectrocladius gr. dilatatus	-	7.3	-	-	-	1
Harnischia sp.	-	5.0	-	-	-	2
Hydrobaenus distylus	-	5.0	-	-	-	2
Metriocnemus gr. hydropetricus	-	5.0	-	-	-	2
Psectrocladius simulans	-	5.0	-	-	-	2
Conchapelopia melanops	-	5.0	-	-	-	2
Chironomus anthracinus	-	1.6	0.1	-	-	9
Diamesa heterodentata	1.1	1.5	-	-	-	2
Tanytarsus usmaensis	-	1.5	0.7	-	-	2
Индикаторы второго класса						
Paratrichocladius rufiventris	10.0	-	-	-	-	1
Parorthocladius sp.	7.0	-	-	-	-	2
Paramerina sp.	7.0	-	-	-	-	2
Cricotopus albiforceps	7.0	-	-	-	-	2
Orthocladius oliveri	7.0	-	-	-	-	2
Mesocricotopus sp.	7.0	-	-	-	-	2
Tvetenia discoloripes	7.0	-	-	-	-	2
Cricotopus gr. cylindraceus	5.7	-	-	-	-	3
Rheocricotopus effusus	5.7	-	-	-	-	3
Pseudodiamesa nivosa	4.9	-	-	-	-	4
Paratanytarsus austriacus	1.8	0.9	-	-	-	2
Cryptotendipes sp.	1.7	1.0	-	-	-	2
Telopelopia sp.	1.7	0.2	-	-	-	7

«Плохие» индикаторы						
Polypedilum nubeculosum	0.01	0.03	0.04	0.03	0.02	187
Chironomus plumosus	-	0.02	0.05	0.03	0.03	195
Procladius ferrugineus	0.01	0.04	0.05	0.01	0.02	177
Tanytarsus sp.	0.03	0.04	0.04	0.01	0.01	184
Cryptochironomus gr. defectus	0.01	0.03	0.06	0.03	0.02	143
Cladotanytarsus mancus	0.02	0.05	0.05	0.01	0.02	139
Cricotopus sylvestris	0.02	0.03	0.06	0.03	0.03	100
Cricotopus bicinctus	0.05	0.06	0.05	0.01	-	124
Dicrotendipes nervosus	-	0.06	0.04	0.04	0.07	75
Paratanytarsus confusus	0.04	0.08	0.07	0.03	0.01	64
Microchironomus tener	0.01	0.06	0.09	0.05	0.03	58
Cladopelma gr. lateralis	0.01	0.05	0.11	0.04	0.03	59
Procladius choreus	0.05	0.06	0.11	0.03	0.02	50
Prodiamesa olivacea	0.06	0.07	0.11	0.01	-	61
Paracladius conversus	0.05	0.11	0.08	0.02	-	55

ВЫВОДЫ

Предложенная техника прогноза качества вод далеко не единственная и может быть улучшена чисто математическими приемами, например аппроксимацией функции распределения видового обилия в разных классах качества вод и т.д. Другой резерв повышения эффективности распознавания - предварительно проведенная ручная или автоматизированная выбраковка гидробиологических измерений, включенных в обучающую выборку. Расчет индикаторных валентностей предпочтительнее проводить не на всем массиве наблюдений, а на некотором «опорном» подмножестве надежных примеров, где сведены до минимума случайные ошибки измерений и влияние таких посторонних факторов, как сезонность, неудачный выбор места отбора пробы и т.д. В данной работе это не делалось принципиально, чтобы оценить общий уровень и степень влияния «шума», неизбежно сопровождающего гидробиологические измерения. Зато тщательный анализ ошибок экзамена показал, что не менее 100 проб из 542 вообще не могут быть правильно классифицированы ни компьютером, ни человеком.

Любая естественно-научная теория должна выполнять как минимум две функции: объяснения и прогнозирования наблюдаемых феноменов, причем при исследовании сложных систем объединение этих функций в одной математической модели нецелесообразно [13]. Индикаторные валентности, выборочно приведенные в табл. 4, рассчитаны для выполнения конкретной задачи (обеспечить минимум ошибок прогноза класса качества воды) и изначально не предназначены для «объяснения», т.е. формирования каких-либо научных гипотез об экологии видов и их роли в формировании сообществ гидробионтов.

Подробно с результатами исследований и монографией [15] читатель может ознакомиться на авторском сайте Интернет http://www.tolcom.ru/kiril.

СПИСОК ЛИТЕРАТУРЫ

1. *Баканов А.И.* Использование зообентоса для мониторинга пресноводных водоемов // Биол. внутр. вод. 2000. № 1. С. 68-82
2. *Вапник В.Н., Червоненкис А.Я.* Теория распознавания образов. М.: Наука, 1974. 416 с.
3. Временные методические указания по комплексной оценке качества поверхностных и морских вод. М.: Госкомитет по гидрометеорологии и контролю природной среды, 1986. № 250-1163. 5 с.
4. *Гайдышев И.П.* Анализ и обработка данных: специальный справочник. СПб.: Питер, 2001. 752 с.
5. Гидробиологический режим малых рек в условиях антропогенного воздействия / Под ред. Андрушайтиса Г.П., Качаловой О.Л. Рига.: Зинатне, 1981. 166 с.
6. *Горелик А.Л., Скрипкин В.А.* Методы распознавания. М.: Высш. шк., 1984. 219 с.
7. ГОСТ 17.1.3. 07-82. Охрана природы. Гидросфера. Правила контроля качества воды водоемов и водотоков. М. : Государственный комитет СССР по стандартам, 1982. 12 с.
8. *Гублер Е.В.* Вычислительные методы анализа и распознавания патологических последствий. Л.: Медицина, 1978. 387 с.
9. *Драчев С.М.* Борьба с загрязнением рек, озер и водохранилищ промышленными и бытовыми стоками. М.-Л.: Наука, 1964. 274 с.
10. *Зинченко Т.Д., Шитиков В.К.* Гидробиологический мониторинг как основа типологии малых рек Самарской области // Изв. Самар. науч. центра РАН. 1999. Т. 1. № 1. С. 118-127
11. *Кольцов П.П.* Математические модели теории распознавания образов // Компьютер и задачи выбора. – М.: Наука, 1989. С. 89-119
12. *Макрушин А.В.* Биологический анализ качества вод. Л.: Зоол. ин-т АН СССР, 1974. 60 с.
13. *Розенберг Г.С., Мозговой Д.П., Гелашвили Д.Б. Экология.* Элементы теоретических конструкций современной экологии. Самара.: Самар. науч. центр РАН, 1999. 396 с.
14. *Тодераш И.К.* Функциональное значение хирономид в экосистемах водоемов Молдавии. Кишинев.: Штинница, 1984. 172 с.
15. *Шитиков В.К., Розенберг Г.С., Зинченко Т.Д.* Количественная гидроэкология: методы системной идентификации. Тольятти: Ин-т экологии Волж. бассейна РАН, 2003. 463 с.
16. *Poroikov V.V., Filimonov D.A., Borodina Yu.V. et al.* Robustness of Biological Activity Spectra Predicting by Computer Programm PASS for Noncongenetic Sets of Chemical Compounds // J. Chem. Inform. Comput. Sci. 2000. V. 40. № 6, P. 1349-1355
17. *Schroevers P.J.* The baby and the bath-water. Thoughts about “saprobity” // Hidrobiological Bulletin. 1988. 22(1). P. 79-80
18. *Sladeček V.* System of water quality from the biological point of view // Arch. Hydrobiol., Beihefz., Ergebnisse der Limnol. 1973. B. 7. S. 1-218