УДК 004.65:519.237.8

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА ДЛЯ ЗАДАЧ МОНИТОРИНГА НА ПРИМЕРЕ ДАННЫХ ПО ТУНГУССКОМУ МЕСТОРОЖДЕНИЮ ПОДЗЕМНЫХ ВОД

[1]Кондратьева Л.М., [2]Кожевникова Т.В., [2]Манжула И.С.

[1]Институт водных и экологических проблем Дальневосточного отделения Российской академии наук, Хабаровск, e-mail: kondratevalm@gmail.com;

[2]Вычислительный центр Дальневосточного отделения Российской академии наук, Хабаровск, e-mail: ktvsl@mail.ru

Разработан алгоритм, позволяющий обрабатывать экспериментальные данные, которые входят в базу данных внешнего мониторинга Тунгусского месторождения подземных вод. Водозаборы, эксплуатирующие месторождения подземных вод в прибрежной зоне речных долин, являются интересными объектами для исследования. Основной их характеристикой является наличие «тесной взаимосвязи» между подземными и поверхностными водами. При эпизодическом (аварийном), хроническом загрязнении и при устойчивом низком качестве речной воды существует риск загрязнения подземных вод. Проблема ухудшения качества подземных вод в результате речной фильтрации усугубляется во время сильных наводнений. В работе выполнен подбор математических методов для решения задач имитационного моделирования. Показана возможность применения методов кластерного анализа K-Means, Tree Clustering и метода Главных компонент для выделения схожих объектов из экспериментальных данных. Математические методы применяются для выявления закономерностей и оценки влияния наводнений на р. Амур на качество подземных вод Тунгусского месторождения, с позиций многофакторного анализа. В рамках исследования выполняется анализ данных по различным разделам естественных наук: гидрологии, гидрохимии и гидробиологии. Приведены результаты использования алгоритма на примере выборки из базы данных о содержании ароматических веществ в пробах подземных вод из скважин и Пемзенской протоки. Установлено, что Пемзенская протока обособлена по показателям содержания органических веществ, а скважины сгруппировались по мере удаленности от ее берега. Отмечено, что метод Главных компонент показал наилучшие результаты, по оценке специалистов предметной области.

Ключевые слова: сплавы, многомерное шкалирование, статистическая выборка, описательная статистика, предельно допустимая концентрация

INFORMATION TECHNOLOGIES AND METHODS OF STATISTICAL ANALYSIS FOR THE TASKS OF MONITORING ON THE EXAMPLE OF DATA ON THE TUNGUS DEPOSIT OF UNDERGROUND WATER

[1]Kondrateva L.M., [2]Kozhevnikova T.V., [2]Manzhula I.S.

[1]Institute of water and ecology problems Russian academy of sciences, Khabarovsk, e-mail: kondratevalm@gmail.com;

[2]Computer Center Far East Branch Russian Academy of Science, Khabarovsk, e-mail: ktvsl@mail.ru

An algorithm has been developed that allows processing experimental data that are included in the external monitoring database of the Tunguska groundwater field. Water intakes that exploit groundwater deposits in the coastal zone of river valleys are interesting sites for research. Their main characteristic is the presence of a «close relationship» between groundwater and surface water. With episodic (emergency), chronic pollution and with a steady low quality of the river water there is a risk of pollution of groundwater. The problem of deterioration of groundwater quality as a result of river filtration is aggravated during severe flooding. In the work made the selection of mathematical methods for solving problems of simulation. The possibility of applying the cluster analysis methods K-Means, Tree Clustering and the Principal Components method to isolate similar objects from experimental data is shown. Mathematical methods are used to identify patterns and assess the impact of flooding on the river. Amur on the quality of groundwater of the Tunguska field, from the standpoint of multivariate analysis. As part of the study, an analysis of data in various branches of the natural sciences: hydrology, hydrochemistry and hydrobiology. The results of using the algorithm are given on the example of a sample from a database of the content of aromatic substances in groundwater samples from wells and the Pemze channel. It was established that the Pemze duct is separated by indicators of the content of organic substances, and the wells were grouped as they were remote from its shore. It is noted that the method of the Main Components showed the best results, according to the assessment of domain specialists.

Keywords: multidimensional scaling, statistical sampling, descriptive statistics, maximum permissible concentration

В последнее десятилетие было показано, что количество и качество поверхностных вод в значительной степени определяется глобальным изменением климата и усилением антропогенной нагрузки. Катастрофическое наводнение на р. Амур в 2013 г. широко обсуждалось в научных публикациях с различных позиций: климатических, метеорологических и гидрологических. Интенсивное весеннее половодье из-за снежной зимы и летние дождевые паводки формировались практически на всех

притоках р. Амур. Смещающийся паводок с западной части бассейна принимал на своем максимуме паводки рек восточной части бассейна, обуславливая «каскадное» развитие наводнения [1].

Согласно ранее проведенным расчетам гидрогеологами было высказано мнение, что в период наводнения 2013 г. активного влияния речных вод на гидродинамические условия подземной гидросферы не было отмечено за счет наличия покровных суглинков. В то же время высказываются рекомендации «о разумном удалении подземных водозаборов от контура реки» [2].

Несмотря на то, что предпринимаются многочисленные попытки моделирования поведения загрязняющих веществ в подземных водах в зоне речной фильтрации, существует много нерешенных проблем. Прежде всего, это связано с многокомпонентным загрязнением водоносного горизонта и сложной динамикой биогеохимических процессов, происходящих при взаимодействии воды с горными породами и вновь поступившими с поверхностными водами органическими веществами.

В работе предлагается алгоритм исследования вод (на основе имитационного моделирования) с целью классификации объектов (кустов, группы скважин) по наличию схожих биохимических показателей, в предположении, что можно выделить класс объектов по расстоянию от берега Пемзенской протоки. Это докажет, что на содержание органических веществ в пробах подземных вод, оказывают влияние поверхностные воды р. Амур, распространяющиеся в Пемзенской протоке.

Для понимания природных процессов необходимо проанализировать данные мониторинга за природными объектами. В процессе анализа требуется выделить схожие объекты, наблюдая за которыми можно лучше понять законы, по которым происходят изменения в этих объектах. Отметим, что классификация является одним из фундаментальных процессов в науке. Достаточно часто возникает необходимость проведения классификации множества объектов по нескольким факторам. Для проведения такой многомерной классификации используются методы кластерного анализа. Кластеризацию можно считать процедурой, которая, начиная работать с тем или иным типом данных, преобразует их в данные о кластерах.

Наибольшее распространение получили иерархические агломеративные методы и итерационные методы группировки. При использовании методов кластерного анализа достаточно сложно дать однозначные рекомендации по предпочтению применения тех или иных методов. Необходимо понимать, что получаемые результаты классификации не являются единственными. Предпочтительность выбранного метода и полученных результатов следует тщательно обосновать.

В условиях необходимости многофакторного анализа исследуемых показателей подземных вод и низкой структурированности данных, эффективным методом выявления схожих признаков (количество органических веществ в пробах воды) выступает кластерный анализ – метод множественной количественной классификации. При этом элементы и их сочетания важны не сами по себе, а как индикаторы наличия биохимических показателей воды, которые зависят от удаленности скважин, глубины установки фильтров для отбора проб подземных вод и сезонности наблюдений.

Предполагается, что анализ данных выполняется в среде программирования R. R – язык программирования для статистической обработки данных и работы с графикой, а также свободная программная среда вычислений с открытым исходным кодом в рамках проекта GNU. R поддерживает широкий спектр статистических и численных методов и обладает хорошей расширяемостью с помощью пакетов. Пакеты представляют собой библиотеки для работы специфических функций или специальных областей применения. Ещё одна особенность R – возможность создания качественной графики, которая может включать математические символы [3].

На основании всего вышеперечисленного сформулирована цель исследований: на основе имитационного моделирования разработать алгоритм для выявления пространственно-временных факторов, способных оказывать влияние на качество подземных вод в зоне речной фильтрации как в процессе многолетнего мониторинга, так и во время катастрофических наводнений.

Территория Приамурья входит в провинцию железосодержащих, марганецсодержащих и кремнийсодержащих пресных подземных вод. В междуречье р. Амур и Тунгуска разведано Тунгусское месторождение подземных вод для водоснабжения г. Хабаровска. По гидрохимическому составу это гидрокарбонатно-натриевые, маломинерализованные (до 200 мг/дм^3) воды с повышенным содержанием железа и марганца [4].

На территории Тунгусского месторождения подземных вод сооружена наблюдательная сеть мониторинга подземных вод [5], состоящая из нескольких кустов

скважин, расположенных на разном расстоянии от основного русла р. Амур и левобережной Пензенской протоки (табл. 1). Ярусные кусты состоят из трех компактно расположенных скважин, оборудованы фильтрами длиной 2 м на разной глубине водоносного горизонта. Куст 1 расположен на расстоянии 50 м от уреза воды, куст 2 – 300 м от берега, куст 3 более 1000 м от берега.

Таблица 1

Общая характеристика скважин по отбору проб подземных вод

Кусты скважин	Расстояние от берега, м	Номер скважины	Глубина установки фильтра, м
Куст 1	50 м	К 1-1	14,7
		К 1-2	24,7
		К 1-3	34,7
Куст 2	300	К 2-1	13,7
		К 2-2	26,7
		К 2-3	37,7
Куст 3	1000	К 3-1	20,0
		К 3-2	39,40
		К 3-3	53,8

Для разработки алгоритма, специалистами Института водных и экологических проблем Дальневосточного отделения Российской академии наук были предложены данные о содержании ароматических веществ в пробах подземных вод из 9 скважин, а также из Пензенской протоки (табл. 2).

Специалистами предметной области предложена рабочая гипотеза: содержание органических веществ (ОВ) в подземных водах изменяется в зависимости от удаления скважин от береговой линии и глубины отбора проб. Динамика поднятия уровня воды во время наводнения оказывает влияние на изменение содержания ароматических соединений в подземных водах в зоне речной фильтрации. С целью исследования достоверности этой гипотезы предлагается следующий алгоритм исследования на основе имитационного моделирования, в процессе реализации которого возникает необходимость анализа многомерных данных, полученных при проведении имитационных экспериментов, в частности – задачи разделения множеств данных на непересекающиеся подмножества. Для решения данной задачи используются методы кластерного анализа, в частности – задачи разделения множества данных на непересекающиеся подмножества.

Пусть в ходе имитационных экспериментов получено множество наблюдений, которое необходимо разбить на непересекающиеся подмножества (кластеры) [6].

Материалы и методы исследования

В качестве исходных данных для имитационного моделирования используются данные о содержании ароматических веществ (табл. 1), которые необходимо разбить на непересекающиеся подмножества (кластеры). Объектами для кластеризации выступают скважины, данные о которых содержат биохимические показатели и сезонность.

Для исследования выбраны методы кластеризации, являющиеся представителями основных методологических подходов к разделению исходного множества объектов на кластеры: K-Means, Tree Clustering, метод Главных компонент.

Алгоритм имитационного моделирования на основе методов кластерного анализа для визуализации шагов обработки экспериментальных данных, представленных в табл. 2, представлен на рис. 1.

Таблица 2

Пространственно-временная динамика ароматических органических соединений по спектральной характеристике (275 нм) в подземных водах Тунгусского месторождения в 2013–2014 гг.

Место отбора проб, № скважины	2013 г.				2014 г.			
	Апрель	Август	Сентябрь	Ноябрь	Март	Июнь	Август	Ноябрь
Пензенская протока	0,264	0,714	0,398	0,385	0,582	0,558	0,354	0,324
К1-1	0,137	0,540	0,386	0,357	0,265	0,269	0,201	0,156
К1-2	0,203	0,511	0,349	0, 298	0,243	0,257	0,153	0,162
К1-3	0,188	0,422	0,333	0,187	0,187	0,197	0,106	0,124
К2-1	0,261	0,420	0,256	0,200	0,204	0,225	0,110	0,161
К2-2	0,169	0,386	0,197	0,194	0,165	0,154	0,087	0,092
К2-3	0,213	0,547	0,298	0,198	0,182	0,228	0,135	0,104
К3-1	0,034	0,158	0,231	0,147	0,025	0,205	0,303	0,017
К3-2	0,118	0,238	0,230	0,163	0,065	0,132	0,074	0,034
К3-3	0,142	0,411	0,281	0,205	0,084	0,102	0,126	0,085

Рис. 1. Схема алгоритма обработки данных

Таблица 3

Принятые обозначения данных

Период наблюдения		Пензенская протока	К1-1	К1-2	К1-3	К2-1	К2-2	К2-3	К3-1	К3-2	К3-3
Обозначения	2013 г.	1	2	3	4	5	6	7	8	9	10
	2014 г.	11	12	13	14	15	16	17	18	19	20

Результаты исследования и их обсуждение

При выполнении каждого из представленных видов кластерного анализа, данные для проведения исследования были обозначены следующим образом (табл. 3).

После предварительного исследования выборки данных (проверки однородности, вычисления описательных статистик) была проведена кластеризация методом K-Means, результаты которой приведены на рис. 2.

Метод K-Means. Задается 5 кластеров, данные по которым распределились по мере удаленности скважины от Пензенской протоки, с учетом года взятия проб. Объекты (скважины) распределились на 5 кластеров. Кластеризующий признак – удаление скважин от Пензенской протоки, что согласуется с выдвинутой гипотезой. Отметим, что однозначного суждения о влиянии расстояния на содержание ОВ в пробах только по этим результатам выявить не представляется возможным.

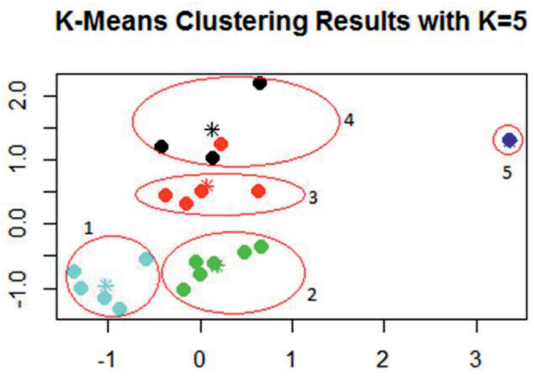

*Рис. 2. Результаты анализа данных методом K-Means (символом * обозначены центры выделенных кластеров, цифрами от 1 до 5 обозначены кластеры разбиения данных)*

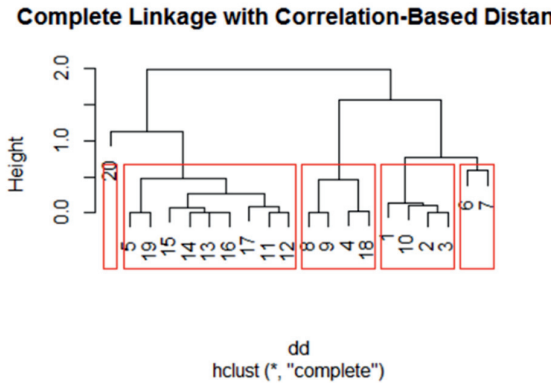

Рис. 3. Результаты анализа данных методом Tree Clustering

Метод Tree Clustering. Результаты приведены на рис. 3. По данным исследования можно выделить следующие кластеры:

1. Данные скважины К3-3 за 2014 г.
2. Данные преимущественно за 2014 г. по всем скважинам 1 и 2 кустов, и нижнего слоя водоносного горизонта 3 куста.
3. Данные преимущественно 3 куста за 2013 г.
4. Данные преимущественно за 2013 г. по Пемзенской протоке и первому кусту.
5. Данные по 2 кусту за 2013 г.

Результаты исследования данных по методу Tree Clustering вполне соответствуют представлениям о гидрологических и биогеохимических процессах, которые могли происходить в зоне речной фильтрации во время наводнения и после этого события.

Метод главных компонент. По результатам обработки данных Пемзенская протока выделяется как обособленный объект со специфическими показателями, характерными для поверхностных вод, в отличие от подземных вод.

Заключение

Разработан алгоритм, позволяющий обрабатывать экспериментальные данные, которые входят в базу данных внешнего мониторинга Тунгусского месторождения подземных вод, проводимого Институтом водных и экологических проблем ДВО РАН по заданию МУП «Водоканал» г. Хабаровска (Кулаков, Андреева, 2016).

Для решения основной задачи имитационного моделирования применялись следующие математические методы кластерного анализа: K-Means, Tree Clustering и метод Главных компонент. Отмечено, что для выборки проб вод по содержанию ароматических веществ лучшие результаты получены при использовании метода Главных компонент.

Приведены результаты использования алгоритма с использованием всех перечисленных методов кластерного анализа.

Анализ результатов имитационного моделирования позволяет сделать вывод, что для предложенной выборки данных наилучший результат получен с помощью метода Главных компонент.

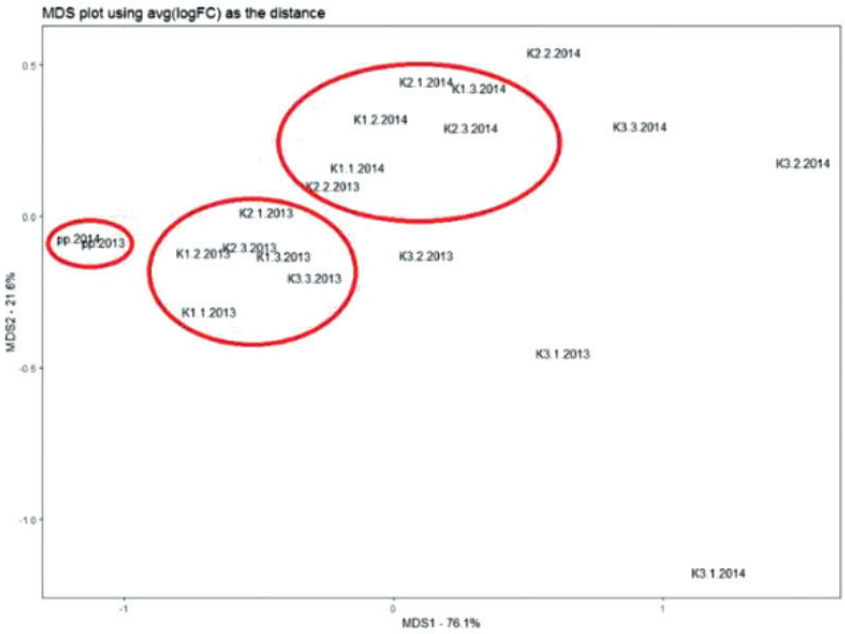

Рис. 4. Результаты анализа данных методом главных компонент

Данный алгоритм рекомендован для использования специалистами: гидрологами, гидрохимиками и гидробиологами для обработки экспериментальных данных мониторинга качества вод.

Список литературы

1. Данилов-Данильян В.И., Гельфан А.Н., Мотовилов Ю.Г., Калугин А.С. Катастрафическое наводнение 2013 года в бассейне реки Амур: условия формирования, оценка повторяемости, результаты моделирования // Водные ресурсы. 2014. Т. 41. № 2. С. 111–122. DOI: 10.7868/S0321059614020059.

2. Кулаков В.В. Загрязнение подземных вод в Среднеамурском артезианском бассейне // Известия РГО. 2017. Т. 149. Вып. 5. С. 36–47.

3. Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R / Пер. с англ. Полины А. Волковой. М.: ДМК Пресс, 2014. 588 с.

4. Кулаков В.В. Геохимия подземных вод Приамурья. Хабаровск: ИВЭП ДВО РАН, 2011. 254 с.

5. Кулаков В.В., Андреева Д.В. Растворенные газы подземных вод Амуро-Тунгусского междуречья // Тихоокеанская геология. 2016. Т. 35. № 2. С. 83–93.

6. Магнус Я.Р., Катышев П.К., Пересецкий Эконометрика. Начальный курс: учебник для вузов. 6-е изд., перераб. и доп. М.: Дело, 2004. 576 с.