

ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

УДК 004.942; 519.876.5

DOI 10.22213/2410-9304-2018-2-84-96

СИСТЕМНЫЙ АНАЛИЗ ПОКАЗАТЕЛЕЙ ИСХОДНОЙ ВОДЫ ПРИ ПРОИЗВОДСТВЕ ПИТЬЕВОЙ ВОДЫ В СИСТЕМЕ ЦЕНТРАЛЬНОГО ВОДОСНАБЖЕНИЯ

Г. А. Благодатский, кандидат технических наук, доцент, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия

А. А. Бас, кандидат технических наук, доцент, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия
М. М. Горохов, доктор физико-математических наук, профессор, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия

Д. С. Пономарев, аспирант, ИжГТУ имени М. Т. Калашникова, Ижевск, Россия

Работа посвящена системному анализу данных показателей исходной воды при производстве питьевой воды в системе центрального водоснабжения крупного населенного пункта. На сегодняшний день на фоне увеличивающегося негативного антропогенного воздействия на окружающую среду наблюдается ухудшение состояния многих источников питьевого водоснабжения в широком спектре показателей, в частности, таких как органолептические свойства воды. Как следствие, возникает проблема и для питьевой воды. В работе приводится процесс подготовки данных о параметрах исходной воды, забираемой из водохранилища, которые ежемесячно (с 2002 по 2014 год) учитывались на предприятии при дезодорации воды. Приведенные параметры оказывают существенное влияние на органолептические свойства конечной воды. Подготовка данных для анализа проводится методом главных компонент К. Пирсона. Данные, полученные в пространстве R^9 , переводятся в пространство меньшей размерности R^3 . Понижение размерности позволяет снизить автокорреляцию между компонентами. Отбор компонент в пространство R^3 проводится по правилу Парето. В пространстве R^3 методом сферической кластеризации данных «Форель» с постоянным радиусом группировки проводится кластеризация. Приводится пошаговое визуальное представление алгоритма кластеризации в пространстве R^3 . В работе показано, что в данных показателях качества исходной воды имеются кластеры. Проводится корреляционно-регрессионный анализ данных, представленных в главных компонентах. Строятся регрессионные зависимости показателей органолептических свойств от главных компонент из пространства R^3 .

Ключевые слова: производство питьевой воды, метод главных компонент, кластеризация, регрессия, визуализация.

Введение

На сегодняшний день на фоне увеличивающегося негативного антропогенного воздействия на окружающую среду наблюдается ухудшение состояния многих источников питьевого водоснабжения в широком спектре показателей, в частности, таких как органолептические свойства воды. Как следствие, возникает проблема и для питьевой воды. Оптимальным решением является очистка воды порошкообразными активированными углями (ПАУ). Такой способ очистки может быть использован на любых сооружениях водоподготовки. Однако применение ПАУ является исключительно дорогостоящим процессом, требующим значительных инвестиций.

Кроме того, отсутствуют рекомендации по выбору параметров дезодорации, а именно марки активированного угля и его смесей, оптимального дозирования и выбора времени контакта в зависимости от параметров исходной воды. Таким образом, системный подход к решению данного вопроса является достаточно актуальным.

Проведение экспериментальных исследований очистки и дезодорации питьевой воды непосредственно на очистных сооружениях является довольно дорогостоящим процессом с необходимостью привлечения практически всех ресурсов предприятия и вероятностью нарушения водоснабжения города, что является недопустимым. Поэтому перспективным в данной области пред-

ставляется проведение исследований на теоретическом уровне, а именно – системный анализ объекта исследования и разработка математической модели для основных параметров очистки и дезодорации воды. Данный подход позволяет осуществить экономию ресурсов, детально изучить протека-

ние процессов с возможностью построения прогноза, а также выявить закономерности их протекания. Особый интерес представляет применение методов кластеризации и регрессии, что позволяет описать нелинейные зависимости и тем самым более точно представить рассматриваемые процессы.

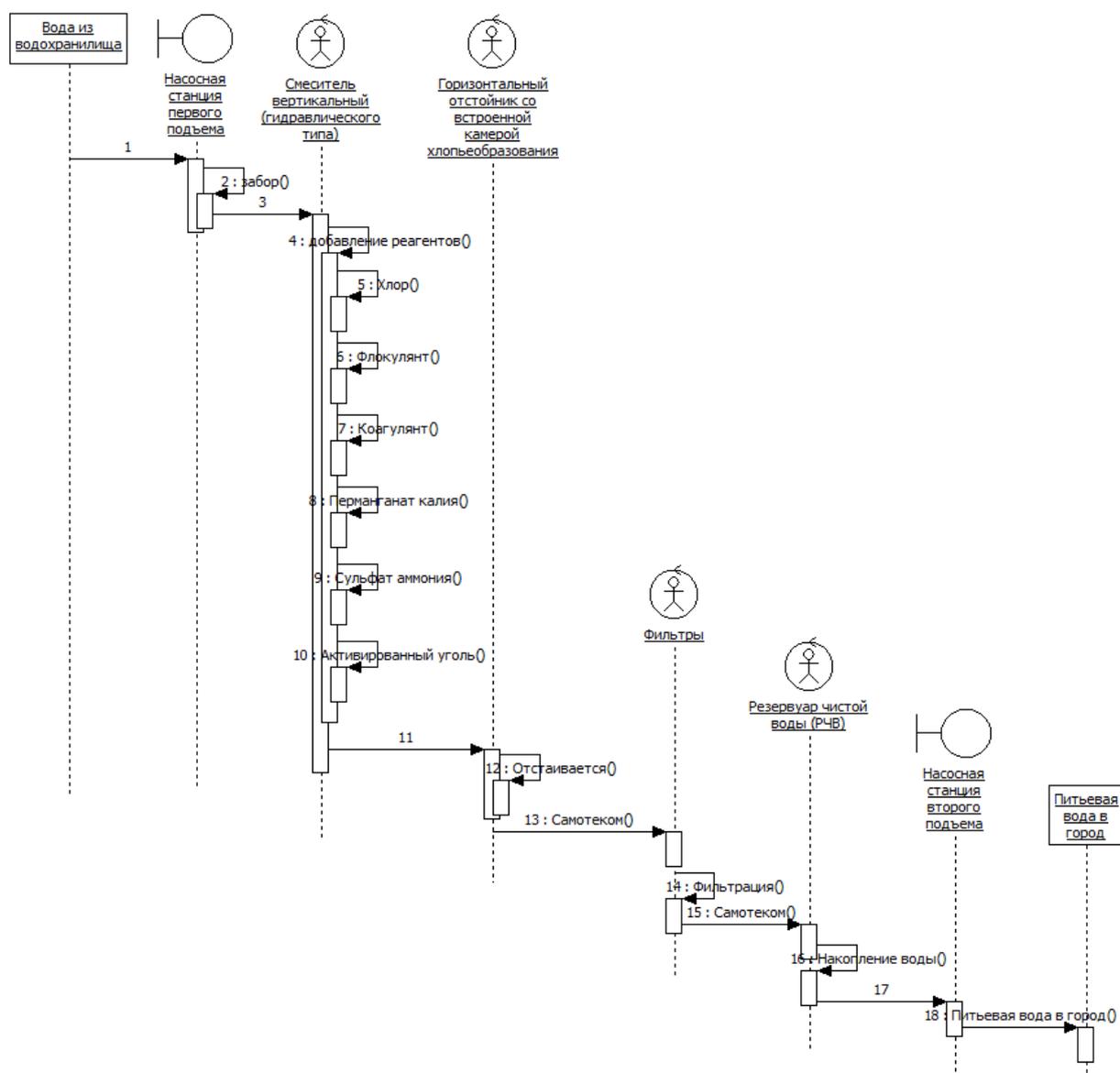


Рис. 1. Процесс производства питьевой воды в системе центрального водоснабжения

На рис. 1 изображены: 1 – насосная станция первого подъема. Применяется для забора воды из водохранилища; 2 – смеситель (вертикальный, гидравлического типа). Представляет собой резервуар, в котором перемешиваются реагенты, применяется для равномерного их распределения. Добавление реагентов: хлор (средний годовой расход со-

ставил 80 тонн); флокулянт («Праестол 650TR» расход 5 тонн/год затраты), коагулянт (расход 10 тонн), перманганат калия (расход 6 тонн), сульфат аммония (расход 40 тонн), активированный уголь (расход 50 тонн). 3 – горизонтальный отстойник со встроенной камерой хлопьеобразования (представляет собой резервуар, в котором

отстаивается вода. Вода движется самотеком); 4 – фильтры (фильтрация воды путем ее пропускания через слой загрузки (т. е. через слой гальки, песка). Вода движется самотеком); 5 – резервуар чистой воды (РЧВ) (представляет собой резервуар для накопления воды для стабильной, постоянной подачи в город к потребителям); 6 – насосная станция второго подъема (расход электроэнергии на работу насосов 600 кВт·ч на 1000 м³ для двух станций первого и второго подъемов, при обработке воды от 30 000 до 67 000 м³); 7 – питьевая вода в город.

Подготовка данных для анализа проводится методом главных компонент К. Пирсона (метод разработан К. Пирсоном в 1901 году [1]). Данные, полученные в пространстве R₉, переводятся в пространство меньшей размерности R₃.

Цель состоит в том, чтобы найти как можно меньшее число m главных компонент, объясняющих как можно большую долю общей дисперсии. Таким образом, будет снижено количество главных компо-

нент, нужных для объяснения большей части дисперсии данных.

Применим метод главных компонент для основных параметров исходной воды, которые ежемесячно (с 2002 по 2014 год) учитывались на предприятии при дезодорации воды и которые оказывают существенное влияние на ее органолептические свойства.

Основная часть

Для расчета используем исходные данные значений x_{ij} стохастических переменных X_i . Проведем приведение значений случайных переменных к безразмерному виду, так как метод главных компонент очень чувствителен к абсолютным значениям переменных.

Применяя (1), получим значения случайных переменных в интервале [0; 1]:

$$\tilde{x}_{ij} = \frac{x_{ij} - \min_i x_{ij}}{\max_i x_{ij} - \min_i x_{ij}}. \quad (1)$$

Средние значения величин X_i приведены в табл. 1.

Таблица 1. Средние значения показателей качества воды, приведенных к безразмерному виду

Переменная	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Среднее значение	0,328	0,602	0,560	0,098	0,469	0,417	0,195	0,466	0,465

Найдем выборочную ковариационную матрицу

$$S = \frac{1}{n-1} (X - \bar{X})(X - \bar{X})^T = \\ = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_i)(X_i - \bar{X}_i)^T$$

для матрицы $X = [\tilde{x}_{ij}], i = \overline{1, m}, j = \overline{1, n}$, где X_i – вектор значений случайных переменных, \bar{X}_i – его среднее значение (табл. 2).

Таблица 2. Выборочная ковариационная матрица для показателей качества воды

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
X_1	0,067	0,035	0,021	0,001	-0,008	0,024	0,026	0,032	0,025
X_2	0,035	0,094	0,073	0,000	-0,004	0,010	0,000	0,015	0,012
X_3	0,021	0,073	0,065	-0,004	-0,006	0,003	-0,006	0,007	0,006
X_4	0,001	0,000	-0,004	0,089	-0,011	-0,008	0,008	0,002	0,018
X_5	-0,008	-0,004	-0,006	-0,011	0,050	-0,002	-0,020	-0,021	-0,030
X_6	0,024	0,010	0,003	-0,008	-0,002	0,034	0,021	0,021	0,007
X_7	0,026	0,000	-0,006	0,008	-0,020	0,021	0,055	0,041	0,034
X_8	0,032	0,015	0,007	0,002	-0,021	0,021	0,041	0,063	0,033
X_9	0,025	0,012	0,006	0,018	-0,030	0,007	0,034	0,033	0,074

В практических исследованиях предпочитают использовать вместо матрицы S корреляционную матрицу для выравнивания весов критериев (табл. 3).

Пусть $R = [r_{ij}] = \frac{\text{cov}(x_i, x_j)}{S_i S_j}$ – матрица коэффициентов корреляции показателей качества воды.

Таблица 3. Матрица коэффициентов корреляции

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
X_1	1	0,436411	0,320324	0,007422	-0,1374	0,510804	0,432515	0,49054	0,359066
X_2	0,436411	1	0,942217	0,000639	-0,05927	0,184814	-0,00308	0,193069	0,137772
X_3	0,320324	0,942217	1	-0,05907	-0,10022	0,06793	-0,10273	0,109868	0,092018
X_4	0,007422	0,000639	-0,05907	1	-0,16911	-0,15346	0,114741	0,03265	0,222297
X_5	-0,1374	-0,05927	-0,10022	-0,16911	1	-0,04798	-0,37921	-0,37283	-0,49106
X_6	0,510804	0,184814	0,06793	-0,15346	-0,04798	1	0,478461	0,446703	0,148481
X_7	0,432515	-0,00308	-0,10273	0,114741	-0,37921	0,478461	1	0,700632	0,529043
X_8	0,49054	0,193069	0,109868	0,03265	-0,37283	0,446703	0,700632	1	0,486906
X_9	0,359066	0,137772	0,092018	0,222297	-0,49106	0,148481	0,529043	0,486906	1

Определим собственные векторы и собственные значения матрицы R (табл. 4).

Так как матрица R симметрическая, то можно воспользоваться методом вращений

Якоби для нахождения собственных векторов и собственных значений [2].

Таблица 4. Собственные векторы и собственные значения матрицы R

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9
Собственное число	3,24	1,96	1,35	0,82	0,51	0,45	0,37	0,25	0,04
Собственный вектор	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9
	0,41	0,14	-0,20	0,27	0,29	0,48	0,61	-0,06	0,05
	0,26	0,60	0,14	0,04	-0,05	-0,13	-0,12	-0,09	-0,71
	0,21	0,62	0,19	-0,11	-0,08	-0,13	-0,11	-0,09	0,69
	0,06	-0,16	0,56	0,76	-0,25	-0,03	-0,02	0,08	0,06
	-0,28	0,19	-0,42	0,52	0,52	-0,35	-0,21	0,00	0,07
	0,33	-0,01	-0,53	0,19	-0,45	0,30	-0,49	0,22	0,05
	0,42	-0,33	-0,11	0,04	-0,03	-0,32	-0,08	-0,77	0,03
	0,45	-0,16	-0,10	-0,07	-0,01	-0,61	0,28	0,55	0,01
	0,38	-0,20	0,32	-0,13	0,62	0,21	-0,49	0,18	0,01

Определим сумму долей накопленной выборочной дисперсии, определяемую главными компонентами (табл. 5).

По правилу Парето [3] были выделены четыре главные компоненты, формирующие 81,9 % дисперсии данных ($P_1 - 36,3$ %, $P_2 - 21,82$ %, $P_3 - 14,98$ %, $P_4 - 9,2$ %).

В результате, согласно уравнению перехода к главным компонентам:

$$y_i = \sum_{j=1}^n b_{ij} x_j, \quad (2)$$

определим матрицу перехода к выбранным главным компонентам (табл. 6).

Таблица 5. Доля общей дисперсии, определяемая главными компонентами

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
i	1	2	3	4	5	6	7	8	9
$\frac{\lambda_i}{\sum_{i=1}^k \lambda_i}$	0,360	0,218	0,149	0,092	0,056	0,050	0,041	0,028	0,005
$\sum_{i=0}^i \lambda_i$	0,360	0,578	0,728	0,819	0,876	0,926	0,967	0,995	1,000

Таблица 6. Коэффициенты главных компонент

Параметры исходной воды	Коэффициенты уравнения главных компонент			
	P_1 (36,03 %)	P_2 (21,82 %)	P_3 (14,94 %)	P_4 (9,2 %)
X_1	0,41	0,14	-0,20	0,27
X_2	0,26	0,60	0,14	0,04
X_3	0,21	0,62	0,19	-0,11
X_4	0,06	-0,16	0,56	0,76
X_5	-0,28	0,19	-0,42	0,52
X_6	0,33	-0,01	-0,53	0,19
X_7	0,42	-0,33	-0,11	0,04
X_8	0,45	-0,16	-0,10	-0,07
X_9	0,38	-0,20	0,32	-0,13

Перейдем к представлению данных через полученные главные компоненты.

Анализ результатов

Проведем анализ рассеивания данных по осям нового базиса (рис. 1–6). Визуально

выделим кластеры показателей воды в новом базисе.

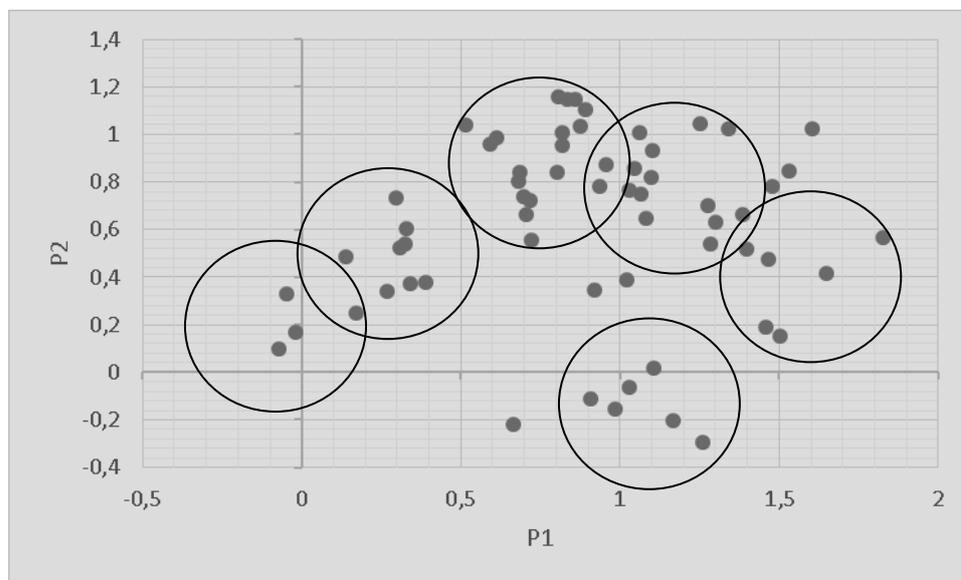


Рис. 1. Рассеивание данных в осях $P_1 P_2$ главных компонент

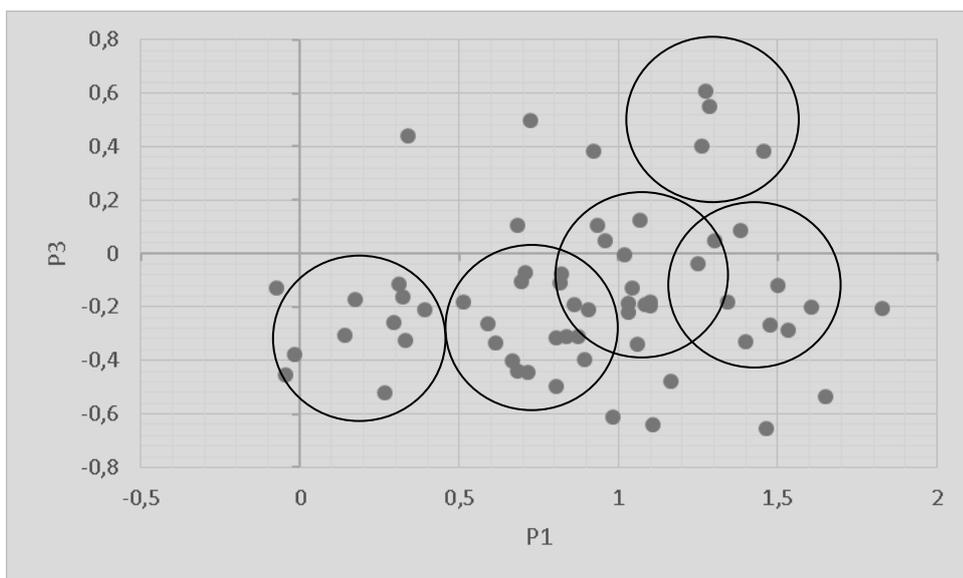


Рис. 2. Рассеивание данных в осях $P_1 P_3$ главных компонент

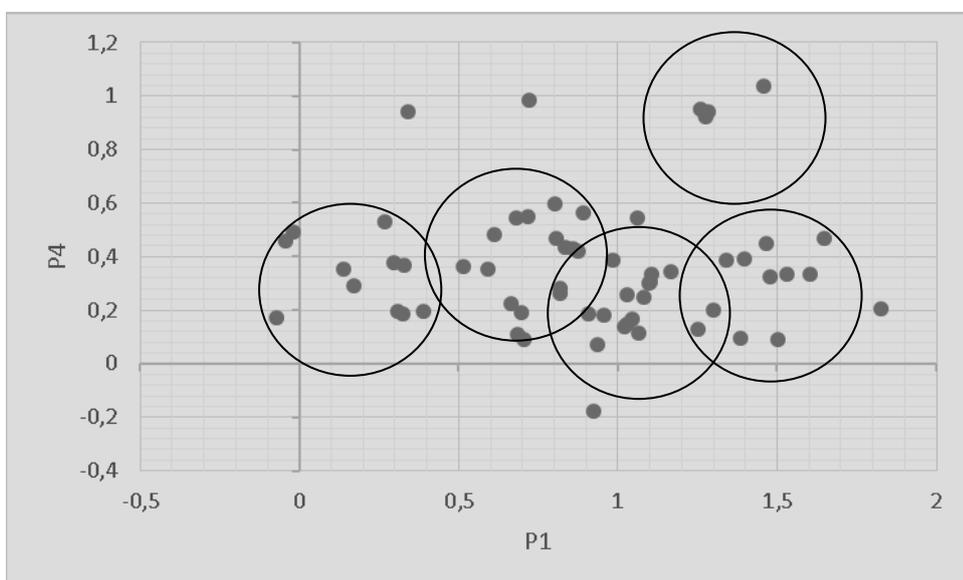


Рис. 3. Рассеивание данных в осях $P_1 P_4$ главных компонент

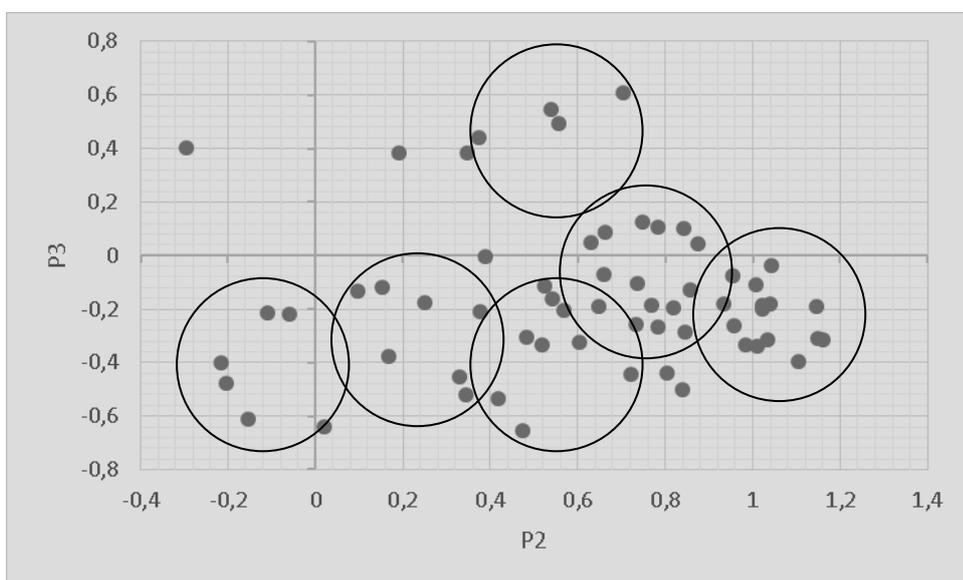


Рис. 4. Рассеивание данных в осях $P_2 P_3$ главных компонент

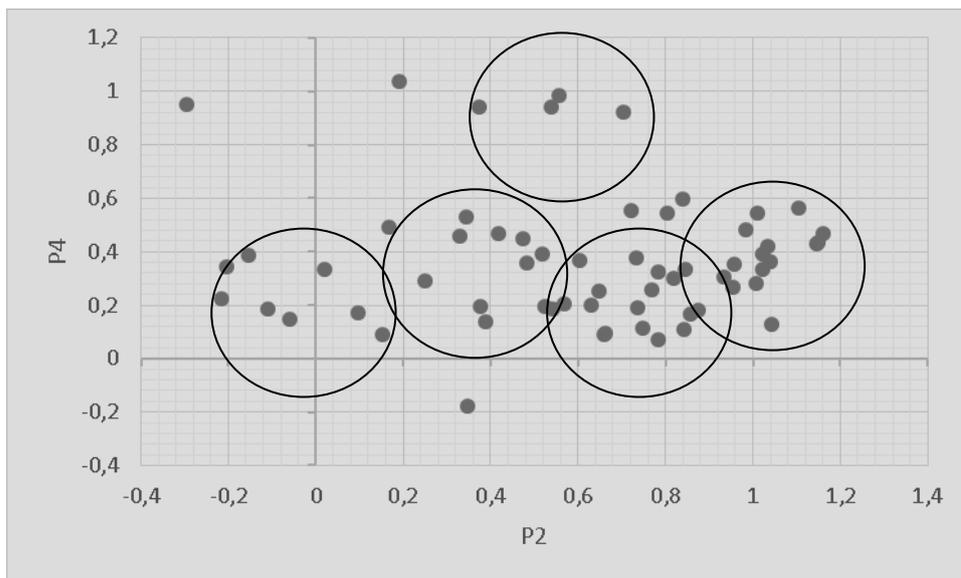


Рис. 5. Рассеивание данных в осях P_2 P_4 главных компонент

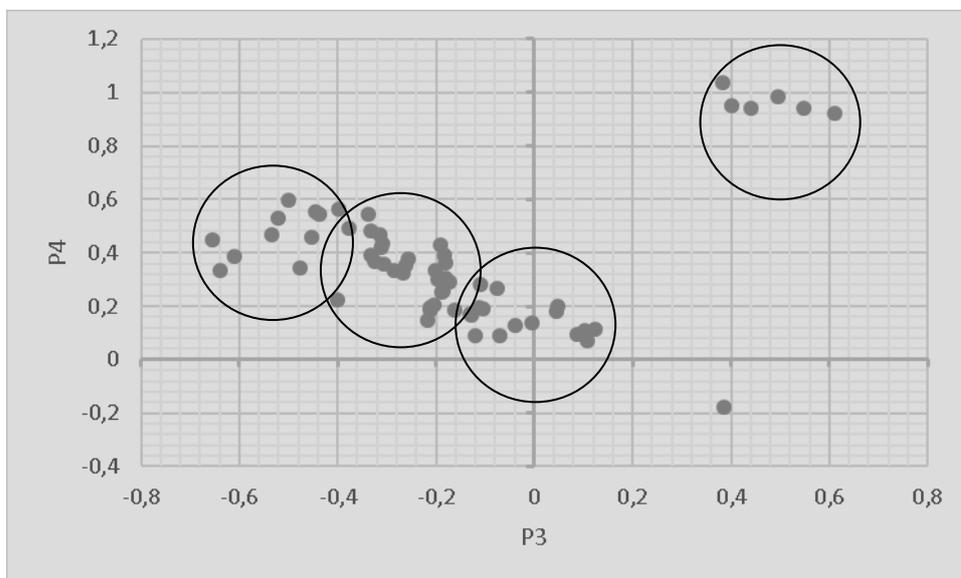


Рис. 6. Рассеивание данных в осях P_3 P_4 главных компонент

Перейдем к размерности R^3 (рис. 7–11).

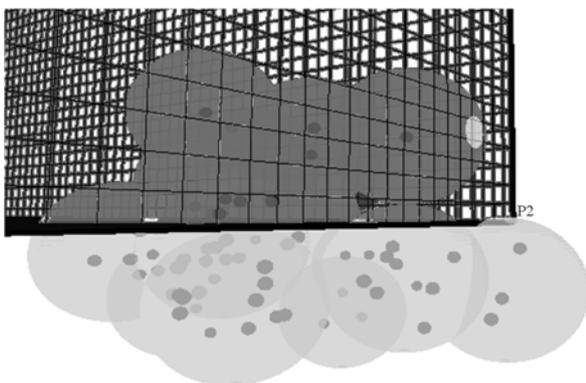


Рис. 7. Пример расположения кластеров

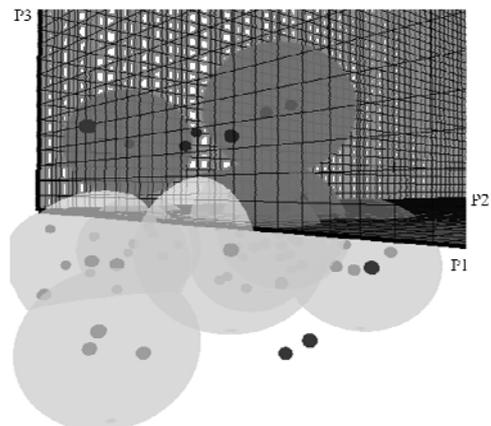


Рис. 8. Пример расположения кластеров

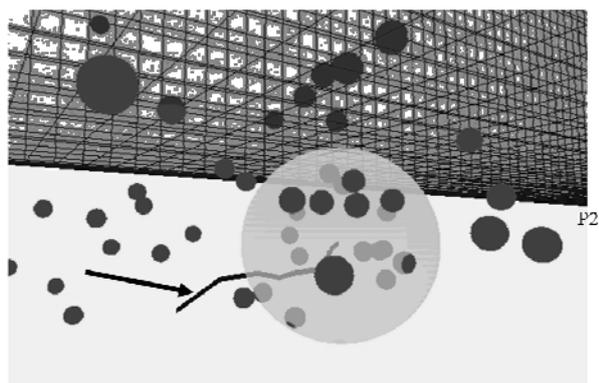


Рис. 9. Поиск оптимального расположения центра кластера. Стрелкой отмечен путь центра сферы в точку локального сгущения

Алгоритм поиска оптимального центра кластера следующий. Вначале по методу Монте-Карло [4] (в границах изменения показателей) выбирается начальное приближение центра сферы. Определяются все точки, попавшие в начальную сферу. Находится центр тяжести этих точек как среднее их координат, и центр сферы перемещается в эту точку. Процесс повторяется до тех пор, пока центр сферы не перестанет изменяться – кластер найден. Итеративное изменение положения центра сферы показано на рис. 9 (зигзагообразный след центра сферы).

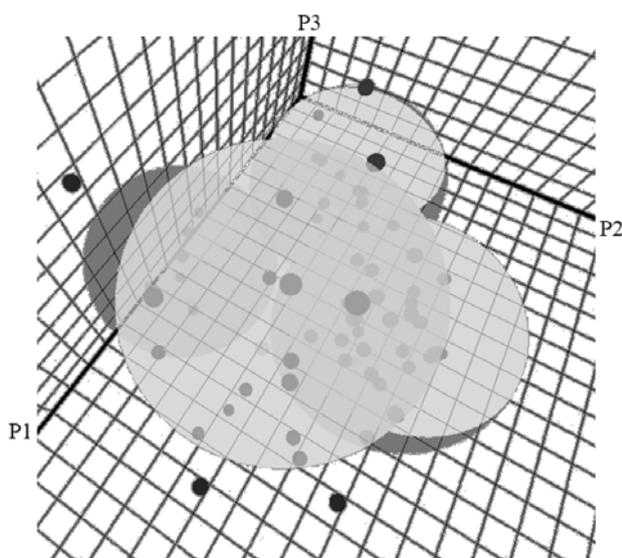


Рис. 10. Кластеры с большим радиусом ($R = 0,5$)

При выборе большого радиуса кластеров (например, 0,5 на рис. 4) получается небольшое число кластеров, в данном случае 4.

В интерфейсе (рис. 11) реализованы возможности поворота трехмерного изображения в градусах (параметры a , b , c), чтения координат точек из файла, однократного расчета координат центров кластеров, указания радиуса кластера, задания максимального количества точек (в процентах от общего числа), которое при расчете может не попасть ни в один кластер, отключения визуализации для ускорения расчетов, множественного расчета расположения кластеров, визуализации лучшего варианта расположения кластеров после расчета, сохранения изображения в файл.

На рис. 12 показана зависимость минимального критерия качества расположения кластеров от итераций. Видно, что критерий не уменьшается уже после тысячи итераций, откуда можно сделать вывод, что достигнуто хорошее (близкое к оптимальному) решение.

Для выявления параметрических зависимостей показателя качества питьевой воды проведен корреляционный анализ данных для трех главных компонент (P) и параметров питьевой воды (Z). Корреляционная связь была установлена для концентрации геосмина в питьевой воде, мг/дм^3 (Z_2); концентрации хлороформа, мг/дм^3 (Z_3); концентрации хлоридов, мг/дм^3 (Z_4). Результаты представлены в табл. 7.

Концентрация геосмина (Z_2) имеет корреляционную зависимость со второй главной компонентой (P_2) (коэффициент корреляции 0,6).

Построим зависимости параметров чистой воды от главных компонент.

Упорядочим данные согласно возрастанию главной компоненты P_2 и построим зависимость концентрации геосмина от P_2 (рис. 7).

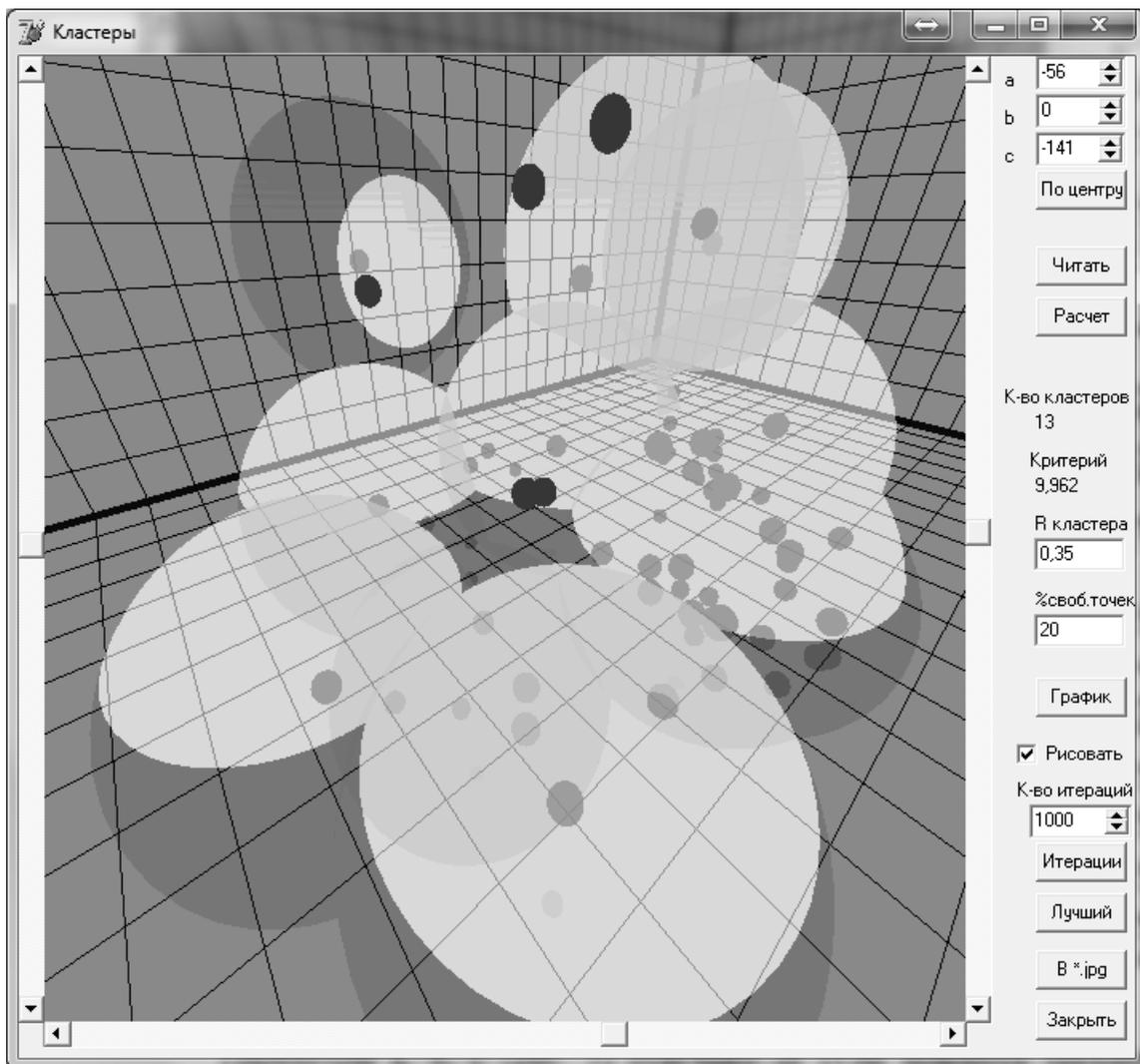


Рис. 11. Главное окно программы

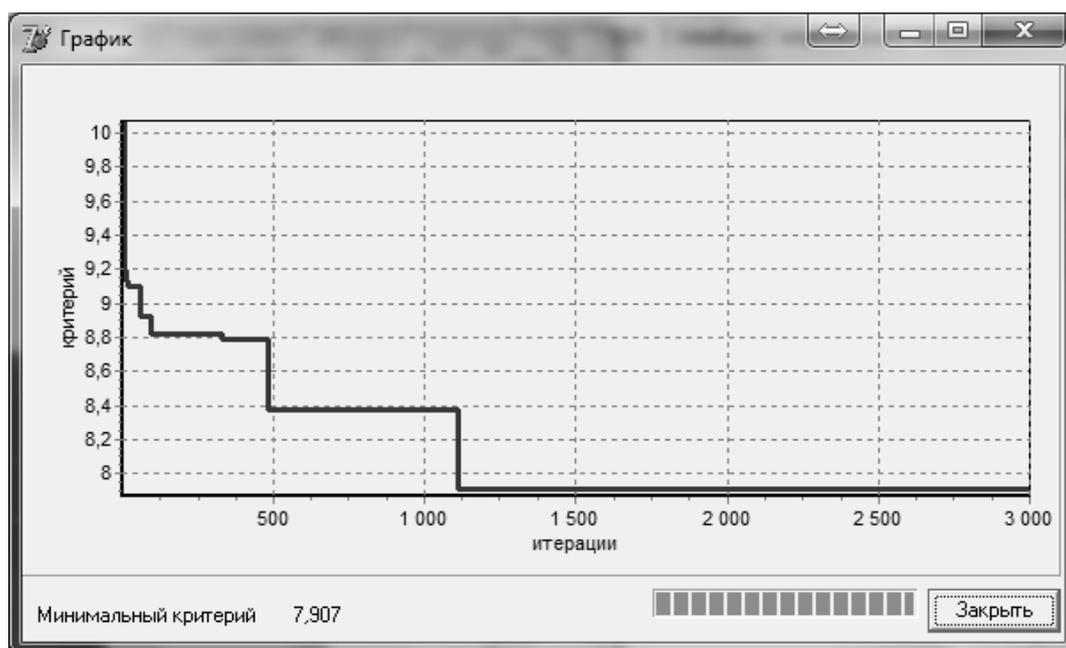


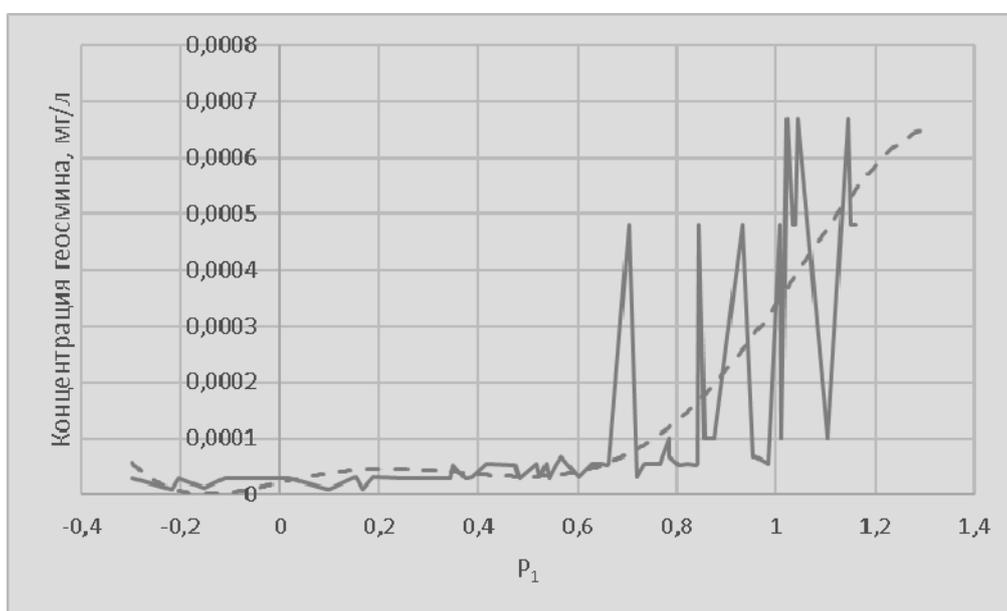
Рис. 12. График изменения критерия качества от итераций

Таблица 7. Результаты корреляционного анализа для трех главных компонент и параметров питьевой воды

Параметры	P_1	P_2	P_3	Z_2	Z_3	Z_4
P_1	1,00	-0,05	-0,47	-0,06	-0,48	0,51
P_2	-0,05	1,00	-0,13	0,62	0,12	0,16
P_3	-0,47	-0,13	1,00	-0,22	0,05	-0,97
Z_2	-0,06	0,62	-0,22	1,00	0,09	0,24
Z_3	-0,48	0,12	0,05	0,09	1,00	0,02
Z_4	0,51	0,16	-0,97	0,24	0,02	1,00

Применяя метод наименьших квадратов [5], определяем параметры линии регрессии (10) (рис. 13, пунктирная линия). Достоверность 0,57.

$$y = 10^{-3} \times (-1,8x^5 + 4,2x^4 - 2,3x^3 - 0,03x^2 + 0,2x + 0,02). \quad (3)$$

Рис. 13. Зависимость концентрации геосмина от P_2

Концентрация хлороформа (Z_3) имеет корреляционную зависимость с первой главной компонентой (P_1) (коэффициент корреляции $-0,6$). Аналогично упорядочим данные согласно возрастанию главной компоненты P_1 и построим зависимость концентрации хлороформа от P_1 (рис. 8).

Применяя метод наименьших квадратов, определяем параметры линии регрессии (11) (рис. 14, пунктирная линия). Коэффициент достоверности 0,49.

$$y = 0,36x^6 - 2,05x^5 + 4,36x^4 - 4,2x^3 + 1,7x^2 - 0,2x + 0,06. \quad (4)$$

Для концентрации хлоридов (Z_4) выявлена корреляционная зависимость с первой (P_1) и третьей (P_3) главными компонентами (коэффициенты корреляции 0,5 и $-0,9$). Упорядочим данные по P_1 .

В данном случае метод [5] будем применять:

$$F(\bar{X}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - A^T X_i + b_i)^2 \rightarrow \min, \quad (5)$$

где y_i – зафиксированное значение концентрации хлоридов; $\hat{y}_i = A^T X_i + b_i$ – модельное

значение; X_i – i -я строка матрицы главных компонент $P_1 P_3$. Минимум (5) необходимо найти относительно неизвестных A и $B = \{b_i\}, i = \overline{1, n}$.

Применяя метод наименьших квадратов, определяем параметры линии регрессии (6) (рис. 15, пунктирная линия). Коэффициент достоверности 0,65.

$$y = 3,917x_1 - 5,397x_2 + 8,551. \quad (6)$$

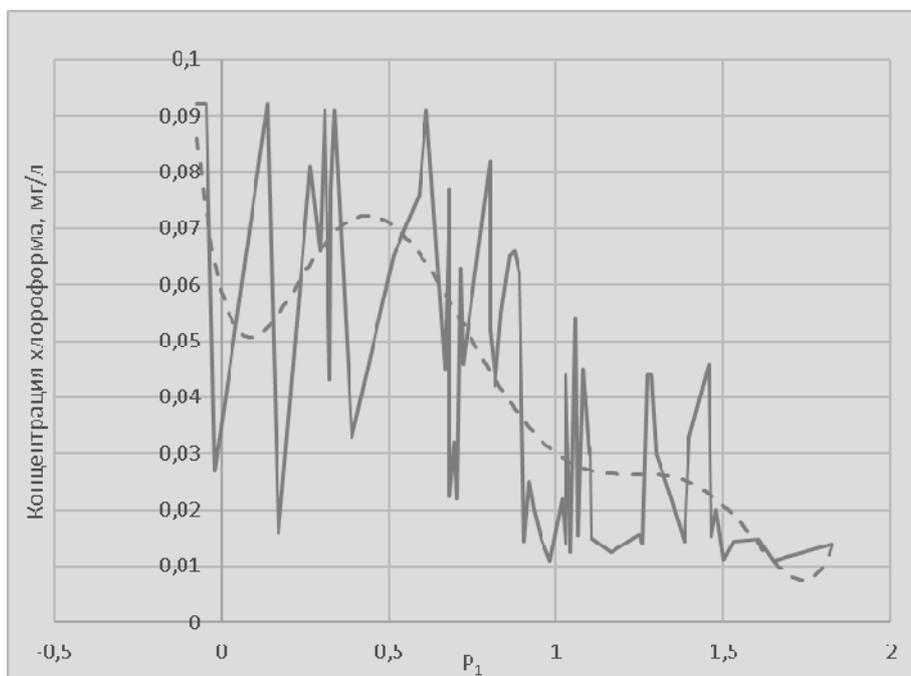


Рис. 14. Зависимость концентрации хлороформа от P_1

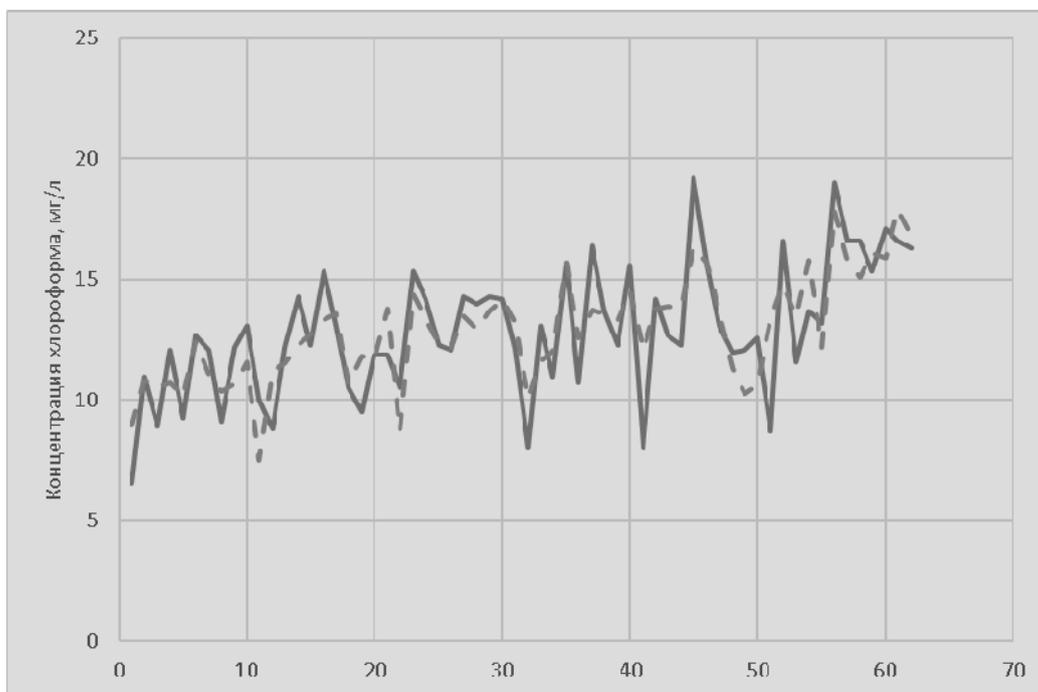


Рис. 15. Зависимость концентрации хлороформа от P_1 и P_3
(данные упорядочены по P_1 в порядке возрастания)

Выводы

Выполняя вышеизложенные расчеты и сопоставляя диаграммы, построенные по экспериментальным точкам с моделями, следует учитывать расхождение экспериментальных и теоретически рассчитанных данных, как следствие из этого, нужно оценить погрешность расчетов, достоверность результатов и отклонение теоретических данных от практических. Хорошие результаты при решении подобных задач дают интеллектуальные системы с применением генетических алгоритмов, нейронных сетей, систем нечеткого логического вывода и анализа иерархий [6–8].

Библиографические ссылки

1. Pearson K. On Lines and Planes of Closest Fit to Systems of Points in Space // *Philosophical Magazine* (London, UK). 1901. № 2 (11). P. 559–572.
2. Калиткин Н. Н. Численные методы. М. : Наука, 1978. С. 177.
3. Koch R. The 80/20 Principle : The Secret of Achieving More with Less. New York: Doubleday. 1998. 288 p.
4. Соболев И. М. Метод Монте-Карло. М. : Наука, 1968. 64 с.
5. Калиткин Н. Н. Указ. соч. С. 59.
6. Благодатский Г. А., Тенев В. А., Шаура А. С. Численная реализация алгоритма управления запасами при длительных сроках поставки комплектующих в условиях «узких» мест производственного цикла // *Интеллектуальные системы в производстве*. 2016. № 4 (31). С. 40–44. DOI: 10.22213/2410-9304-2016-4-40-44.

7. Благодатский Г. А., Горохов М. М., Переведенцев Д. А. Моделирование системы нечеткого логического вывода оценки наукоемких проектов // *Автоматизация процессов управления*. 2017. № 2 (48). С. 82–89.

8. Благодатский Г. А., Горохов М. М., Чухланцев Е. С. Программно-инструментальное средство оценки тренированности спортсменов высших квалификаций // *Вестник ИЖГТУ имени М. Т. Калашникова*. 2016. № 2 (70). С. 87–90.

References

1. Pearson K. (1901). *Philosophical Magazine* (London, UK). 1901. № 2 (11), pp. 559-572 (in Eng.).
2. Kalitkin N. N. (1978). *Chislennyye metody* [Numerical methods]. Moscow: Nauka, p. 177 (in Russ.).
3. Koch R. (1998). *The 80/20 Principle: The Secret of Achieving More with Less*. New York: Doubleday. p. 288 (in Eng.).
4. Sobol' I. M. (1968). *Metod Monte-Karlo* [Monte Carlo method]. Moscow: Nauka, 64 p.
5. Kalitkin N. N. (1978). *Op. cit.*, p. 59 (in Russ.).
6. Blagodatsky G. A., Tenenev V. A., Shaura A. S. (2016). *Intellektual'nye sistemy v proizvodstve* [Intelligent Systems in Manufacturing], no. 4 (31), pp. 40-44 (in Russ.). DOI: 10.22213/2410-9304-2016-4-40-44.
7. Blagodatsky G. A., Gorokhov M. M., Perevedentsev D. A. (2017). *Avtomatizacija processov upravlenija* [Automation Control Process], no. 2 (48), pp. 82-89 (in Russ.).
8. Blagodatsky G. A., Gorokhov M. M., Chukhlantsev E. S. (2016). *Vestnik IzhGTU imeni M. T. Kalashnikova* [Bulletin of Kalashnikov ISTU], no. 2 (70), pp. 87-89 (in Russ.).

System Analysis of Source Water Indicators at Drinking Water Production in the Central Water Supply System

G. A. Blagodatsky, PhD in Engineering, Associate Professor, Kalashnikov ISTU, Izhevsk, Russia
 A. A. Bass, PhD in Engineering, Associate Professor, Kalashnikov ISTU, Izhevsk, Russia
 M. M. Gorokhov, DSc (Physics and Mathematics), Professor, Kalashnikov ISTU, Izhevsk, Russia
 D. S. Ponomarev, Post-graduate, Kalashnikov ISTU, Izhevsk, Russia

The work is devoted to the system analysis of the given indicators of source water at the production of drinking water in the central water supply system of a large settlement. Today, within increasing negative anthropogenic impact on the environment, there is a deterioration in the state of many sources of drinking water supply in a wide range of indicators, in particular, such as organoleptic properties of water. As a consequence, there is a problem for drinking water. The paper presents the process of preparing data on the parameters of the source water taken from the reservoir, which were taken into account monthly at the enterprise (from 2002 to 2014) at the water deodorization plant. The above parameters have a significant effect on the organoleptic properties of the final water. Preparation of data for analysis is carried out by K. Pearson's principal components analysis. The data obtained in the space R9 is trans-

ferred to a space of lower dimension R^3 . Dimension reduction allows to reduce autocorrelation between components. Selection of components in the space R^3 is carried out according to the Pareto rule. Clustering is carried out in the space R^3 by the method of spherical clustering of the "Trout" data with a constant clustering radius. A step-by-step visual representation of the clustering algorithm in the space R^3 is presented. It is shown that there are clusters in these indicators of the quality of the initial water. Correlation-regression analysis of the data presented in the main components is carried out. Regression dependences of the indicators of organoleptic properties on the main components from the space R^3 are shown.

Keywords: production of drinking water, principal component analysis, clustering, regression, visualization.

Получено: 07.05.18