СТАТИСТИЧЕСКИЙ АНАЛИЗ РЕЗУЛЬТАТОВ МОНИТОРИНГА ХИМИЧЕСКОГО СОСТАВА ПОВЕРХНОСТНЫХ ВОД ПУРОВСКОГО РАЙОНА ЯМАЛО-НЕНЕЦКОГО АВТОНОМНОГО ОКРУГА

Научный отчет TR-07-2003

Воронеж, 2003

Лаборатория статистического
контроля качества, прикладной
статистики и хемометрики

394004, Г. Воронеж, ул. Щорса,
д.107, к. 28
тел. (0732) 49-04-09
e-mail: **igor@anch.vsu.ru**

Содержание

1. Резюме

Показаны возможности применения контрольных карт при решении задач экологического мониторинге объектов окружающей среды. Проведен обзор методов и алгоритмов принятия решения по контрольным картам.

Показаны сезонные изменения суммарного содержания органических веществ в р. Пякупур. Повышено содержание нефтепродуктов в период 1998-2000 г.г. по сравнению с периодом 2002-2003 г.г. Снижены величины перманганатной окисляемости в осенние периоды по сравнению с летними.

В работе проведен многомерный статистический анализ хромато-масс-спектроскопических результатов количественного определения содержания органических веществ в пробах воды рек Пякупур, Айваседопур и Пур Пуровского района Ямало-Ненецкого автономного округа. Определены фоновые концентрации 43 органических веществ в этих реках.

2. Исследуемые объекты

Места (точки) отбора проб были пронумерованы. Нумерация точек отбора проб приведена в табл. 1. В табл. 2 и 3 приведены кодировки измеряемых показателей.

Таблица 1. Нумерация мест отбора проб

Номер точки отбора проб	Место отбора пробы
1	Р. Пякупур, выше г. Муравленко
2	Р. Пякупур, выше г. Губкинский
3	Р. Пякупур, ниже г. Губкинский
4	Р. Пякупур, вблизи г. Пуровска
5	Р. Пякупур, вблизи г. Тарко-Сале
6	Р. Айваседопур, вблизи г. Карнада
7	Р. Айваседопур, устье реки
8	Р. Пур, вблизи г. Каратчаево
9	Р. Пур, вблизи г. Самбург

Таблица 2. Кодировка исследуемых физико-химических показателей качества воды

Показатель	Кодировка показателя
pH	P1
Перманганатная окисляемость, мг O_2/л	P2
Общая жесткость, ммоль/л	P3
Кальций, мг/л	P4
Магний, мг/л	P5
Натрий, мг/л	P6

Показатель	Кодировка показателя
Калий, мг/л	P7
СПАВ, мг/л	P8
Нитриты, мг/л	P9
Нитраты, мг/л	P10
Аммонийный азот, мг/л	P11
Сульфаты, мг/л	P12
Хлориды, мг/л	P13
Фосфаты, мг/л	P14
Железо, мг/л	P15
Марганец, мг/л	P16
Кремний, мг/л	P17
Медь, мг/л	P18
Цинк, мг/л	P19
Свинец, мг/л	P20
Никель, мг/л	P21
Кадмий, мг/л	P22
Ртуть, мг/л	P23
Мышьяк, мг/л	P24
Алюминий, мг/л	P25
Фенольный индекс, мг/л	P26
Нефтепродукты	P27
Гидрокарбонаты, мг/л	P28
Общая минерализация, мг/л	P29
Сухой остаток, мг/л	P30
Хлорорганические пестициды, мг/л	P31
Цезий-137, к/л	P32

Таблица 3. Кодировка органических веществ, определяемых методом хромато-масс-спектроскопии

Код вещества	Органическое вещество
Org 1	Толуол
Org 2	Масляная кислота
Org 3	Хлоруксусная кислота, этиловый эфир
Org 4	4-гидрокси-4-метилпентанон-2
Org 5	Сложный эфир
Org 6	2,6-Ди-t-бутил-4-гидрокси-4-метил-2,5-циклогексадиен –1-он
Org 7	Ионол
Org 8	Карбоновая кислота
Org 9	Σ Нафтенов
Org 10	ΣРазветвл. Алканов
Org 11	Пентановая кислота
Org 12	Хлорэтилхлорме-тиловый эфир
Org 13	Алкилдиамин
Org 14	Дихлорпропанол
Org 15	Дихлоруксусная кислота, этиловый эфир
Org 16	1,1 – Дихлор – 2 – этоксиэтан
Org 17	Серосодержа-щие соединения
Org 18	Гексановая кислота
Org 19	Этиленгликоль диацетат
Org 20	Гептановая кислота
Org 21	Октановая кислота
Org 22	Нонановая кислота
Org 23	Амид кислоты
Org 24	Декановая кислота
Org 25	Ароматич. Амин
Org 26	Додекановая кислота (лауриновая)
Org 27	Пентилтиазол
Org 28	Тетрадекановая кислота (миристиновая)
Org 29	Акриловая кислота, эфир
Org 30	Пальмитиновая кислота
Org 31	4-(1,5-Диметил-3-оксогексил) -1-циклогексен -1-карбоновая кислота (Javabione)
Org 32	Метоксикоричная кислота
Org 33	Производное фенилизоцианата
Org 34	Эфир бензойной кислоты
Org 35	Сквален
Org 36	Диэтилфталат
Org 37	Диизобутилфталат
Org 38	Бутилизобутилфталат
Org 39	Дибутилфталат
Org 40	Бис(2-этилгексил)Фталат

Код вещества	Органическое вещество
Org 41	Нафталин
Org 42	Фенантрен
Org 43	Диметилсульфид

Электронная база данных была создана средствами управления данными (Data Management) пакета прикладных статистических программ Statistica 5.1. База данных представляет двумерную электронную таблицу 20x73.

3. Теоретические основы используемых статистических методов

3.1. X-R – статистические контрольные карты количественного признака [1-14]

При гауссовском законе распределения в пределах трехсигмовых границ лежит 99,73% всех значений контролируемого параметра качества. Отсюда следует, что почти все средние, вычисленные по результатам выборок из генеральной совокупности с математическим ожиданием MX и стандартным отклонением σ, приходятся на участок с границами $MX \pm 3\sigma/\sqrt{n}$. Эти две границы называются границами регулирования контрольной карты для средних значений количественного признака $\overline{X}$ (X-карты).

На контрольную карту наносятся обычно три линии: средняя и две крайние, представляющие собой верхнюю и нижнюю границы регулирования. По оси ординат откладываются значения контролируемого параметра, а по оси абсцисс – номера выборок.

Если неизвестно стандартное отклонение σ генеральной совокупности, то его можно оценить с помощью среднего выборочного отклонения s или с помощью средней величины размаха R.

Чтобы получить более наглядные представления об изменениях случайной величины, наряду с X-картой ведут либо s-карту, с помощью которой непрерывно контролируют стандартное отклонение, либо R-карту – для контроля размахов выборок.

Распределение размахов R выборок одинакового объема асимметрично, так как размах по определению является положительной величиной и теоретически может принимать какое угодно значение. Несмотря на асимметричность распределений R, при небольших объемах выборок в большинстве случаев на

практике пользуются формулами, выведенными для гауссовского распределения, поскольку точные формулы сложны для расчетов. Хотя в таких случаях и неизвестна вероятность того, что контрольная точка попадет за границы регулирования, но очевидно, что для статистически управляемого процесса эта вероятность очень мала.

Преимущество двойных карт заключается в наглядности изображения процесса, простоте принятия решения, достоверности вывода относительно величины рассеяния значений контролируемого параметра. По двойной карте можно непрерывно следить за составляющими общей дисперсии признака – рассеянием внутри выборок (внутригрупповая дисперсия) с помощью R-карты, рассеянием между средними значениями $\bar{X}$ различных выборок (межгрупповая дисперсия) с помощью X-карты. Процесс считается лишь тогда статистически управляемым, когда об этом свидетельствуют оба типа карт. Вывод, сделанный по X- карте, до тех пор не будет иметь значения, пока по R-карте процесс не станет статистически управляемым.

При осуществлении контроля характеристик с помощью контрольных карт проверяют, попадают ли все точки графика в диапазон между двумя контрольными границами. Этот диапазон характеризует контрольные нормативы, в пределах которых разброс показателей качества считается допустимым. Такой разброс вызван случайными отклонениями и называется неизбежным разбросом (рассеянием) показателей качества и не требует вмешательства в ход процесса.

Если же на графике часть точек выходит за пределы верхней или нижней контрольной границы, это значит, что показатели качества испытывают разброс, выходящий за пределы контрольных нормативов. Как только на контрольной карте появляется одна или несколько точек на графике, выходящих за пределы контрольных границ необходимо немедленно принять все меры для выявления и устранения причины отклонения. В том случае, когда на графике X-карты какая-то точка выходит за контрольную границу, это означает, что возникает

отклонение от среднего для групп. В случае, когда за контрольные границы выходит точка на графике размахов, это означает, что значительно меняется разброс в группах.

3.2. Критерии нарушения процесса

Контрольная карта – характерный инструмент статистического контроля технологического процесса. Если результаты последовательных выборок вышли за границы регулирования, то процесс отклонился от нормы. Контрольные карты позволяют заблаговременно обнаружить эти нарушения.

Предположим, что среднее значение параметра, например, тридцатой выборки, вышло за контрольные пределы, но еще задолго до этой выборки исследователь по возрастающему характеру расположения соседних точек мог сделать вывод о начале нарушения нормального режима работы. Другим типом предупреждения может служить слишком длительный выброс точек от средней линии. Любое необычное расположение точек на контрольной карте позволяет предположить, что нарушен установленный режим работы.

Вот некоторые из предупреждающих сигналов нарушения процесса:

- точка вне контрольных пределов;
- расположение двух последовательных (или двух из трех последовательных) точек за двухсигмовым пределом;
- выброс последовательности точек по одну сторону от средней линии или наличие группы точек, значительная часть которых находится по одну сторону от среднего значения.

3.3. Критерий серий

Как уже было отмечено ранее, когда точка на контрольной карте, соответствующая выборочному значению контролируемой характеристики (например, среднему значению в X-карте) оказывается вне ограниченной контрольными переделами области, это дает основания предполагать, что производственный процесс разладился. При этом необходимо отслеживать появление систематической тенденции в расположении точек (например, выборочных средних) на контрольной карте, так как наличие такой тенденции может служить свидетельством тренда среднего значения контролируемого процесса. Эти критерии иногда называют критериями серий типа AT&T или критериями против альтернатив специального вида. Термин «специальные альтернативы», как альтернатива случайным или общим причинам, был использован в работе Шуэрта (Shewhart) для того, чтобы сделать разграничение между нормальным производственным процессом, вариации в котором появляются только в силу действия случайных причин, и вышедшим из-под контроля процессом, в котором вариации характеристик обусловлены некоторыми неслучайными, то есть специальными факторами.

Как и обсуждавшиеся ранее контрольные пределы, выраженные в единицах сигмы, критерии серий имеют в своей основе "статистическое" обоснование. Так, например, вероятность того, что любое выборочное среднее значение для X-карты окажется выше центральной линии, равна 0.5 при следующих условиях:

производственный процесс находится в нормальном состоянии (т.е. центральная линия проведена через значение, равное среднему контролируемой характеристики генеральной совокупности изделий),

средние значения следующих друг за другом выборок независимы (т.е. отсутствует автокорреляция) и выборочные средние значения контролируемой характеристики распределены по нормальному закону. Проще говоря, при таких

условиях для выборочного среднего значения шансы попасть выше или ниже центральной линии составляют 50 на 50. Поэтому вероятность того, что два следующих друг за другом выборочных средних окажутся выше центральной линии, будет равна 0.5, умноженному на 0.5 , т.е. 0.25.

Соответственно, вероятность того, что выборочные средние девяти последующих выборок (или серия из 9 точек контрольной карты) окажется с одной стороны от центральной линии, составит 0,95. Заметим, что это значение приблизительно равно вероятности того, что отдельное выборочное среднее значение не попадет в интервал, ограниченный контрольными пределами в 3 сигма (при условии нормального распределения выборочных средних и нормальности производственного процесса). Поэтому, в качестве еще одного индикатора разладки процесса можно рассматривать ситуацию, когда девять последовательных выборочных средних находятся с одной стороны от центральной линии.

Зоны A, B, C. Обычно для задания критериев поиска серий область контрольной карты над центральной линией и под ней делится на три "зоны".

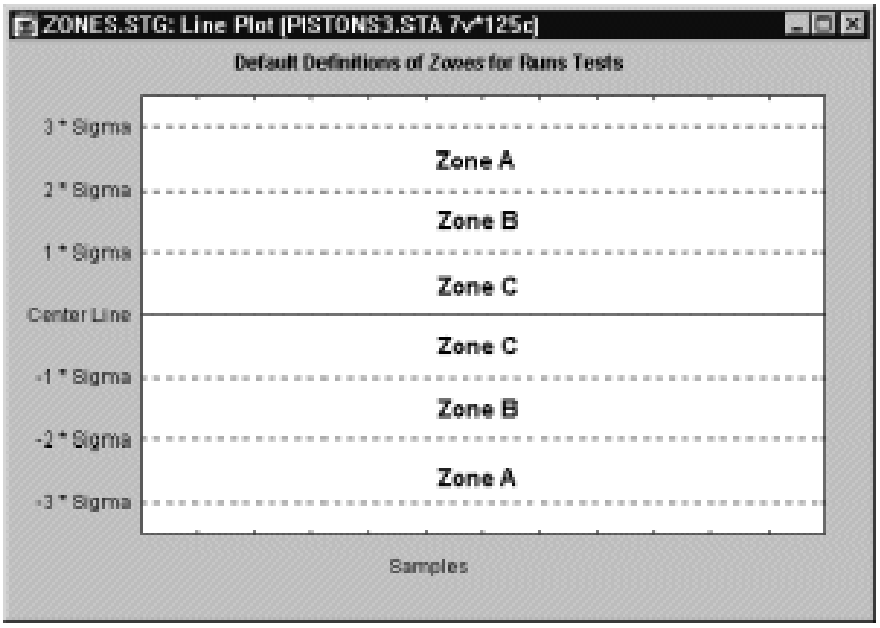

Рисунок 1. Зоны для критерия серий

По умолчанию, зона A определяется как область, расположенная на расстоянии от 2 до 3 сигма по обе стороны от центральной линии. Зона B определяется как область, отстоящая от центральной линии на расстояние от 1 до 2 сигма, а зона C - как область, расположенная между центральной линией по обе ее стороны и ограниченная прямой, проведенной на расстоянии одной сигма от центральной линии.

9 точек в зоне C или за ее пределами (с одной стороны от центральной линии). Если этот критерий выполняется (т.е. если на контрольной карте обнаружено такое расположение точек), то делается вывод о возможном изменении среднего значения процесса в целом. Заметим, что здесь делается предположение о симметричности распределения исследуемых характеристик качества вокруг среднего значения процесса на графике. Но это условие не выполняется, например, для R-карт. Тем не менее, данный критерий полезен для того, чтобы указать занимающемуся контролем качества на присутствие потенциальных трендов процесса. Например, здесь стоит обратить внимание на последовательные выборочные значения с изменчивостью ниже среднего, так как с их помощью можно догадаться, каким образом снижается вариация.

6 точек монотонного роста или снижения, расположенные подряд. Выполнение этого критерия сигнализирует о сдвиге среднего значения процесса.

14 точек подряд в "шахматном" порядке (через одну над и под центральной линией). Если этот критерий выполняется, то это указывает на действие двух систематически изменяющихся причин, которое приводит к получению различных результатов.

2 из 3-х расположенных подряд точек попадают в зону A или выходят за ее пределы. Этот критерий служит "ранним предупреждением" о начинающейся разладке процесса. Заметим, что для данного критерия вероятность получения ошибочного решения (критерий выполняется, однако процесс находится в нормальном режиме) в случае X-карт составляет приблизительно 2 %.

4 из 5-ти расположенных подряд точек попадают в зону В или за ее пределы. Как и предыдущий, этот критерий может рассматриваться в качестве индикатора - "раннего предупреждения" о возможной разладке процесса. Процент принятия ошибочного решения о наличии разладки процесса для этого критерия также находится на уровне около 2%.

15 точек подряд попадают в зону С (по обе стороны от центральной линии). Выполнение этого критерия указывает на более низкую изменчивость по сравнению с ожидаемой (на основании выбранных контрольных пределов).

8 точек подряд попадают в зоны В, А или выходят за контрольные пределы, по обе стороны от центральной линии (без попадания в зону С). Выполнение этого критерия служит свидетельством того, что различные выборки подвержены влиянию различных факторов, в результате чего выборочные средние значения оказываются распределенными по бимодальному закону.

3.4. Многомерный кластерный анализ [15-52]

Метод контрольных карт очень информативен для небольшого количества контролируемых параметров. Если же количество измеряемых признаков велико, как это обычно бывает при анализе объектов окружающей среды, то использование отдельных контрольных карт для каждого измеряемого признака приводит к большому объему графической информации и в общем случае, ее потере.

Экологическая система – это сложная, многомерная система, которая характеризуется, как правило, десятками, сотнями, а то и тысячами показателей, и один человек не может одновременно отслеживать пространственно-временные изменения каждого показателя. Возможны два решения этой проблемы. Во-первых, компьютер может сообщать оператору только о тех показателях, которые

«выходят» за критические значения. Во-вторых, можно использовать многомерные статистические методы для составления обобщенных показателей.

Первое решение является неадекватным для задач экологического мониторинга, так как в нем отсутствуют механизмы предупреждения «внештатных» ситуаций. Во втором случае, реальный экологический объект представляется как многомерный вектор параметров, количественно определяемый матрицей средних значений компонентов и матрицей дисперсий-ковариаций (корреляций). Мониторинг заключается в отслеживании всех изменений многомерного вектора во времени и принятии статистически обоснованных решений на ранних этапах изменений, что позволяет своевременно проводить необходимые корректирующие мероприятия.

В ходе данного исследования была поставлена следующая задача: “Разбить имеющуюся совокупность объектов (образцов речной воды) на отдельные группы - кластеры, таким образом, чтобы объекты входящие в один кластер имели сходство между собой больше, чем с объектами другого кластера”. При этом схожесть объектов должна наблюдаться по многим признакам одновременно. Такая группировка позволит в дальнейшем содержательно интерпретировать различие в химическом составе воды.

Общий вопрос состоит в том, как организовать наблюдаемые данные в наглядные структуры. Классификация основана на интуитивном понятии “близости” или “сходства”. Удачным считается такое рабиение выборочной совокупности, при котором похожие объекты объединены в один класс. Для того, чтобы провести многомерную классификацию необходимо:

- выбрать набор признаков классификации (признаковое пространство);
- определить меру сходства;
- провести расчеты;
- оценить результаты.

"Кластерный анализ - совокупность математических методов, предназначенных для формирования относительно "отдаленных" друг от друга групп "близких" между собой объектов по информации о расстояниях или связях (мерах близости) между ними. Фактически "кластерный анализ" - это обобщенное название достаточно большого набора алгоритмов, используемых при создании классификации. Кластерный анализ широко используется в науке как средство типологического анализа. В любой научной деятельности классификация является одной из фундаментальных составляющих, без которой невозможны построение и проверка научных гипотез и теорий. Анализ отечественных и зарубежных публикаций показывает, что кластерный анализ находит применение в самых разнообразных научных направлениях: биология, медицина, археология, история, география, экономика, филология и т.д.

Термин "кластерный анализ" впервые был предложен Трионом. Слово "cluster" переводится с английского языка как "гроздь, кисть, пучок, группа". По этой причине первоначальное время этот вид анализа называли "гроздевым анализом". В начале 50-х годов появились публикации Р. Люиса, Е.Фикса и Дж. Ходжеса по иерархическим алгоритмам кластерного анализа. Заметный толчок развитие работ по кластерному анализу дали работы Р.Розенблатта по распознающему устройству (персептрону), положившие начало развитию теории "распознавания образов без учителя".

Толчком к разработке методов кластеризации явилась книга "Принципы численной таксономии", опубликованная в 1963 г. двумя биологами - Робертом Сокэлом и Питером Снитом. Авторы этой книги исходили из того, что для создания эффективных биологических классификаций процедура кластеризации должна обеспечивать использование всевозможных показателей характеризующих исследуемые организмы, производить оценку степени сходства между этими организмами и обеспечивать размещение схожих организмов в одну и ту же группу. При этом сформированные группы должны быть достаточно

"локальны", т.е. сходство объектов (организмов) внутри групп должно превосходить сходство групп между собой. Последующий анализ выделенных группировок, по мнению авторов, может выяснить, отвечают ли эти группы разным биологическим видам. Иными словами, Сокэл и Снит предполагали, что выявление структуры распределения объектов в группы, помогает установить процесс образования этих структур. А различие и сходство организмов разных кластеров (групп) могут служить базой для осмысления происходившего эволюционного процесса и выяснения его механизма.

В эти же годы было предложено множество алгоритмов таких авторов, как Дж. Мак-Кин, Г. Болл и Д. Холл по методам k-средних; Г. Ланса и У. Уильямса, Н. Джардайна и др. - по иерархическим методам. Заметный вклад в развитие методов кластерного анализа внесли и отечественные ученые - Э.М. Браверман, А.А.Дорофеюк, И.Б.Мучник, Л.А,Растригин, Ю.И.Журавлев, И.И.Елисеева и др.

В том или ином объеме методы кластерного анализа имеются в большинстве наиболее известных отечественных и зарубежных статистических пакетах: SIGAMD, DataScope, STADIA, СОМИ, ПНП-БИМ, СОРРА-2, СИТО, SAS, SPSS, STATISTICA, BMDP, STATGRAPHICS, GENSTAT, S-PLUS и т.д.

3.5. Расстояние между объектами в признаковом пространстве (метрика)

Интуитивно многие понимают, что понятие “расстояния между объектами” должно отражать меру сходства, близости объектов между собой по всей совокупности используемых признаков. Иными словами служит интегральной мерой сходства объектов между собой. В многочисленных изданиях посвященных кластерному анализу описано более 50 различных способов вычисления расстояния между объектами. Наиболее доступно для восприятия и понимания в случае количественных признаков является так называемое

“евклидово расстояние” или “евклидова метрика”. Оно попросту является геометрическим расстоянием в многомерном пространстве и вычисляется следующим образом:

$$D(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Кроме эвклидовой метрики часто используются и такие метрики, как расстояние Минковского, Хемминга, Махаланобиса, коэффициенты Рао, Роджерса-Танимото, Жаккара, Миркина, метрики Брея-Кертиса, Канберрова, Манхэттен и многие другие.

3.6. Правила объединения объектов в кластеры (связи)

Многократные попытки классификации методов кластерного анализа приводят к десяткам, а то и сотням разнообразных классов. Такое многообразие порождается большим количеством возможных способов вычисления расстояния между отдельными наблюдениями, не меньшим количеством методов вычисления расстояния между отдельными кластерами в процессе кластеризации и многообразием оценок оптимальности конечной кластерной структуры (под структурой кластеров подразумевается состав отдельных кластеров и их взаимное расположение в многомерном пространстве).

На первом шаге иерархической классификации, каждый объект представляет собой отдельный кластер (монокластер), расстояние между которыми определяются выбранной мерой - евклидовой. Однако когда связываются вместе несколько объектов, возникает вопрос, как следует определить расстояние между кластерами? Другими словами, необходимо правило объединения или связи для двух кластеров. После такого объединения “объектом” является не отдельная точка в многомерном признаковом пространстве, а отдельные группы таких точек. Отметим, что в этом случае

разнообразных возможностей еще больше, нежели в случае вычисления расстояния между двумя наблюдениями в многомерном пространстве. Эта процедура осложняется тем, что в отличие от точек, кластеры занимают определенный объем многомерного пространства и состоят из многих точек.

Кроме объединяющих методов иерархической кластеризации существуют и противоположные методы - дивизимные, в которых на начальном этапе вся выборка рассматривается как единый кластер, а затем уже начинается процесс его деления на составляющие части. Процесс деления продолжается до тех пор, пока каждое наблюдение не превратится в отдельный кластер. В свою очередь дивизимные алгоритмы делятся на монотетические и политетические. В монотетической классификации деление производится на основании единственного признака, имеющего максимальную информативность. В политетических же алгоритмах учитываются все признаки. Поскольку данные алгоритмы оперируют расстояниями между наблюдениями, то в некоторых программах предусмотрена возможность работы не с исходной матрицей "объект - признак", а с симметричной матрицей расстояний между наблюдениями.

Среди итерационных методов наиболее популярным методом является метод k-средних Мак-Кина. В отличие от иерархических методов в большинстве реализаций этого метода сам пользователь должен задать искомое число конечных кластеров, которое обычно обозначается как "k". Как и в иерархических методах кластеризации, пользователь при этом может выбрать тот или иной тип метрики. Разные алгоритмы метода k-средних отличаются и способом выбора начальных центров задаваемых кластеров. В некоторых вариантах метода сам пользователь может (или должен) задать такие начальные точки, либо выбрав их из реальных наблюдений, либо задав координаты этих точек по каждой из переменных. В других реализациях этого метода выбор заданного числа k начальных точек производится случайным образом, причем эти начальные точки

(зерна кластеров) могут в последующем уточняться в несколько этапов. Можно выделить 4 основных этапа таких методов:

- выбираются или назначаются k наблюдений, которые будут первичными центрами кластеров;
- при необходимости формируются промежуточные кластеры приписыванием каждого наблюдения к ближайшим заданным кластерным центрам;
- после назначения всех наблюдений отдельным кластерам производится замена первичных кластерных центров на кластерные средние;
- предыдущая итерация повторяется до тех пор, пока изменения координат кластерных центров не станут минимальными.

3.7. Плотность и локальность кластеров

Одно из важных свойств кластера - это плотность распределения точек, наблюдений внутри кластера. Это свойство дает нам возможность определить кластер в виде скопления точек в многомерном пространстве, относительно плотное по сравнению с иными областями этого пространства, которые либо вообще не содержат точек, либо содержат малое количество наблюдений. Несмотря на достаточную очевидность этого свойства, однозначного способа вычисления такого показателя (плотности) не существует. Наиболее удачным показателем, характеризующим компактность, плотность "упаковки" многомерных наблюдений в данном кластере, является дисперсия расстояния от центра кластера до отдельных точек кластера. Чем меньше дисперсия этого расстояния, тем ближе к центру кластера находятся наблюдения, тем больше плотность кластера. И наоборот, чем больше дисперсия расстояния, тем более

разрежен данный кластер, и, следовательно, есть точки находящиеся как вблизи центра кластера, так и достаточно удаленные от центра кластера.

Другое важное свойство кластера - их локальность, отделимость. Оно характеризует степень перекрытия и взаимной удаленности кластеров друг от друга в многомерном пространстве.

3.8. Расстояние между кластерами

В кластерном анализе широко используются межкластерные расстояния, вычисляемые по принципу ближайшего соседа, центра тяжести, дальнего соседа, медиан.

Наиболее широко используются четыре метода: одиночной связи, полной связи, средней связи и метод Варда. В методе одиночной связи объект будет присоединен к уже существующему кластеру, если хотя бы один из элементов кластера имеет тот же уровень сходства, что и присоединяемый объект. Отсюда и название метода - одиночная или единственная связь. Для метода полных связей присоединение объекта к кластеру производится лишь в том случае, когда сходство между кандидатом на включение и любым из элементов кластера не меньше некоторого порога. Для метода средней связи имеется несколько модификаций, которые являются некоторым компромиссом между одиночной и полной связью. В них вычисляется среднее значение сходства кандидата на включение со всеми объектами существующего кластера. Присоединение производится в том случае, когда найденное среднее значение сходства достигает или превышает некоторый порог. Наиболее часто используют среднее арифметическое сходство между объектами кластера и кандидата на включение в кластер.

Многие из методов кластеризации отличаются между собой тем, что их алгоритмы на каждом шаге вычисляют разнообразные функционалы качества

разбиения. Такие экстремальные задачи позволяют определить тот количественный критерий, следуя которому можно было бы предпочесть одно разбиение другому. Под наилучшим разбиением понимает такое разбиение, на котором достигается экстремум (минимум или максимум) выбранного функционала качества. Выбор такого количественного показателя качества разбиения опирается подчас на эмпирические соображения. В качестве таких функционалов качества часто берутся "взвешенная" сумма внутриклассовых дисперсий расстояний, сумма попарных внутриклассовых расстояний между внутрикластерными элементами и т.д. Популярный метод Варда построен таким образом, чтобы оптимизировать минимальную дисперсию внутрикластерных расстояний. На первом шаге каждый кластер состоит из одного объекта, в силу чего внутрикластерная дисперсия расстояний равна 0. Объединяются по этому методу те объекты, которые дают минимальное приращение дисперсии, вследствие чего данный метод имеет тенденцию к порождению гиперсферических кластеров.

4. Результаты и их обсуждение

4.1. Дескриптивная статистика результатов

На первом этапе проводили описательный (дескриптивный) статистический анализ результатов. Теоретической предпосылкой такого усреднения являлось положение, что вся вода принадлежит одной генеральной совокупности, а разброс показателей обусловлен случайными отклонениями, которые могут быть оценены с использованием выборочных характеристик разброса (например, стандартных отклонений). Результаты такого исследования приведены в табл. 4 и табл. 5.

Таблица 4. Дескриптивная статистика физико-химических показателей воды

Показатель	Среднее значение	95%-ный доверительный интервал	ПДК
P1	6,25	6,01 – 6,48	6-9
P2	16,56	12,26 – 20,87	5*
P3	0,574	0,033 – 0,814	7
P4	6,134	4,118 – 8,150	100
P5	3,252	1,272 – 5,232	50
P6	1,995	1,442 – 2,468	200
P7	0,394	0,232 – 0,555	12
P8	0,040	0,020 – 0,060	0,5
P9	0,032	0,007 – 0,058	3,0
P10	0,494	0,232 – 0,755	45
P11	0,596	0,445 – 0,747	2,0
P12	8,730	5,904 – 11,535	500
P13	5,876	4,789 – 6,964	350
P14	0,173	0,086 – 0,261	3,5
P15	2,376	1,853 – 2,898	0,3*

Показатель	Среднее значение	95%-ный доверительный интервал	ПДК
P16	0,104	0,068 – 0,141	0,1*
P17	1,943	0,461 – 3,424	10
P18	0,007	0,004 – 0,009	1,0
P19	0,072	0,027 – 0,116	5,0
P20	0,011	0,004 – 0,017	0,03
P21	0,034	0,013 – 0,056	0,1*
P22	0,0005	0,0003 – 0,0007	0,0001
P23	0,00004	0,00001 - 0,0001	0,0005
P24	0,0058	0,0041 – 0,0073	0,05
P25	0,241	-0,091 – 0,574	0,5*
P26	0,009	-0,004 – 0,023	0,25
P27	0,415	0,142 – 0,692	0,1*
P29	30,07	19,40 – 40,73	-
P30	74,16	33,44 – 114,87	1000
P31	49,08	32,47 – 65,68	-

*Показатели, не отвечающие стандартам ПДК

В дальнейшем исследовании будем рассматривать только те показатели, которые находятся на уровне ПДК или его превышают. Таких показателей пять – Перманганатная окисляемость (P2), Железо (P15), Марганец (P16), Никель (P21), Нефтепродукты (P27)

Таблица 5. Дескриптивная статистика количественного анализа содержания органических веществ

Код вещества	Вещество	Средняя концентрация	Стандартное отклонение	Границы 95%-ного доверительного интервала		Размах
				Мин.	Макс.	
Org 1	Толуол	1,296	0,404	0,793	1,799	1,070

Код вещества	Вещество	Средняя концентрация	Стандартное отклонение	Границы 95%-ного доверительного интервала		Размах
				Мин.	Макс.	
Org 2	Масляная кислота	0,186	0,415	-0,330	0,702	0,930
Org 3	Хлоруксусная кислота, этиловый эфир	0,454	0,450	-0,106	1,014	1,010
Org 4	4-гидрокси-4-метилпентанон-2	1,558	3,483	-2,768	5,884	7,790
Org 5	Сложный эфир	1,678	1,184	0,207	3,149	3,170
Org 6	2,6-Ди-t-бутил-4-гидрокси-4-метил-2,5-циклогексадиен –1-он	2,744	0,941	1,575	1,740	2,070
Org 7	Ионол	0,100	0,161	-0,100	0,300	0,370
Org 8	Карбоновая кислота	0,890	0,818	-0,126	1,906	2,090
Org 9	Σ Нафтенов	18,368	16,215	-1,767	38,503	37,810
Org 10	ΣРазветвл. Алканов	2,850	1,767	0,656	5,044	4,180
Org 11	Пентановая кислота	0,090	0,201	-0,160	0,340	0,450
Org 12	Хлорэтилхлорметиловый эфир	3,462	4,802	-2,501	9,425	9,740
Org 13	Алкилдиамин	0,476	0,685	-0,375	1,327	1,490
Org 14	Дихлорпропанол	0,590	0,621	-0,181	1,361	1,480
Org 15	Дихлоруксусная кислота, этиловый эфир	0,912	1,610	0,176	1,647	0,592
Org 16	1,1 – Дихлор – 2 - этоксиэтан	0,408	1,280	-0,250	1,066	0,530
Org 17	Серосодержа-щие соединения	0,060	0,134	-0,106	0,226	0,300
Org 18	Гексановая кислота	14,312	3,897	9,472	19,151	10,640
Org 19	Этиленгликоль диацетат	0,252	0,563	-0,447	0,951	1,260
Org 20	Гептановая кислота	2,756	0,999	1,514	3,997	2,380
Org 21	Октановая кислота	10,582	2,038	8,050	13,113	4,700
Org 22	Нонановая кислота	17,468	2,006	14,976	19,959	4,310
Org 23	Амид кислоты	1,758	0,951	0,576	2,939	2,060
Org 24	Декановая кислота	1,680	0,680	0,835	2,524	1,730
Org 25	Ароматич. амин	0,222	0,325	-0,182	0,626	0,720

Код вещества	Вещество	Средняя концентрация	Стандартное отклонение	Границы 95%-ного доверительного интервала		Размах
				Мин.	Макс.	
Org 26	Додекановая кислота (лауриновая)	0,770	0,781	-0,200	1,740	1,670
Org 27	Пентилтиазол	0,246	0,163	0,042	0,449	0,400
Org 28	Тетрадекановая кислота (миристиновая)	0,398	0,422	-0,126	0,992	1,000
Org 29	Акриловая кислота, эфир	0,134	0,299	-0,238	0,506	0,670
Org 30	Пальмитиновая кислота	6,652	4,593	0,948	12,356	10,130
Org 31	4-(1,5-Диметил-3-оксогексил) -1-циклогексен -1-карбоновая кислота (Javabione)	0,268	0,262	-0,058	0,594	0,590
Org 32	Метоксикоричная кислота	1,056	0,822	0,034	2,077	2,160
Org 33	Производное фенилизоцианата	0,286	0,316	-0,107	0,679	0,670
Org 34	Эфир бензойной кислоты	0,174	0,209	-0,086	0,434	0,510
Org 35	Сквален	0,442	0,988	-0,785	1,669	2,210
Org 36	Диэтилфталат	0,124	0,046	0,066	0,182	0,120
Org 37	Диизобутилфталат	1,280	0,331	0,868	1,691	0,830
Org 38	Бутилизобутил-фталат	0,268	0,140	0,093	0,442	0,350
Org 39	Дибутилфталат	63,880	17,566	42,068	85,691	35,230
Org 40	Бис(2-этилгексил) Фталат	1,850	0,309	1,465	2,234	0,780
Org 41	Нафталин	0,002	0,004	-0,003	0,007	0,010
Org 42	Фенантрен	0,056	0,042	0,003	0,108	0,100
Org 43	Диметилсульфид	0,046	0,011	0,031	0,060	0,030

По результатам содержания органических веществ в речной воде были рассчитаны парные коэффициенты корреляции. Всего было рассчитано *(43*43 – 43)/2 = 903* коэффициента парной корреляции. В таблице 3 представлены статистически значимые (при уровне значимости p=0,05) коэффициенты корреляции.

Таблица 6. Коэффициенты парной корреляции признаков, статистически значимые на уровне значимости p=0,05

Org 1-Org 23 0,92	Org 6-Org 25 0,92	Org 9-Org 13 0,99	Org 18-Org 21 0,92	Org 29-Org 41 1,00
Org 2-Org 4 1,00	Org 6-Org 37 0,91	Org 9-Org 14 0,95	Org 20-Org 21 0,90	Org 32-Org 43 -0,95
Org 2-Org 11 1,00	Org 6-Org 38 0,96	Org 10-Org 31 0.92	Org 20-Org 22 0,93	Org 33-Org 40 0,94
Org 2-Org 17 11,00	Org 6-Org 39 0,95	Org 10-Org 40 0,92	Org 21-Org 22 0,99	Org 35-Org 41 1,00
Org 2-Org 19 1,00	Org 7-Org 8 0,93	Org 10-Org 42 0,92	Org 23-Org 25 0,89	Org 36-Org 37 0,96
Org 3-Org 20 0,99	Org 7-Org 25 0,98	Org 11 -Org 17 1,00	Org 23-Org 37 0,89	Org 36-Org 38 0,89
Org 3-Org 22 0,91	Org 7-Org 29 0,94	Org 11-Org 19 1,00	Org 23-Org 38 0.92	Org 37-Org 38 0,97
Org 4-Org 11 1,00	Org 7-Org 35 0,94	Org 12-Org 13 0,89	Org 23-Org 39 0,97	Org 37-Org 39 0,93
Org 4-Org 17 1,00	Org 7-Org 41 0,94	Org 13-Org 14 0,90	Org 25-Org 39 0,90	Org 38-Org 39 0,94
Org 4-Org 19 1,00	Org 8-Org 25 0,94	Org 15-Org 40 0,95	Org 28-Org 30 0,98	
Org 6-Org 23 0,94	Org 8-Org 31 0,96	Org 16-Org 38 0,90	Org 29-Org 35 1,00	

По результатам корреляционного анализа можно выделить 5 групп признаков («корреляционные плеяды»), которые имеют высокие коэффициенты взаимной парной корреляции (табл. 4).

На рис. 1 приведена дендрограмма иерархической классификации всех 43 признаков по методу Варда.

Таблица 7. Группы наиболее коррелированных признаков

Группа	Признаки
1	Масляная кислота
	4-гидрокси-4-метилпентанон-2
	Пентановая кислота

Группа	Признаки
	Серосодержащие соединения
	Этиленгликоль диацетат
2	2,6-Ди-t-бутил-4-гидрокси-4-метил-2,5-циклогексадиен –1-он
	Амид кислоты
	Ароматический амин
	Диэтилфталат
	Диизобутилфталат
	Бутилизобутилфталат
	Дибутилфталат
3	Ионол
	Карбоновая кислота
	Ароматический амин
	Акриловая кислота, эфир
	Сквален
	Нафталин
4	Σ Нафтенов
	Алкилдиамин
	Дихлорпропанол
5	Хлоруксусная кислота, этиловый эфир
	Гексановая кислота
	Гептановая кислота
	Октановая кислота
	Нонановая кислота

Как видно из дендрограммы, достаточно четко выделяются три кластера. Дибутилфталат (Org 39) образует монокластер, который присоединяется ко всем остальным на последнем шаге. Во второй кластер входят пять веществ: пальмитиновая кислота (Org 30), октановая кислота (Org 21), нонановая кислота (Org 22), гексановая кислота (Org 18), нафтены (Org 9). В третий кластер можно объединить все остальные вещества.

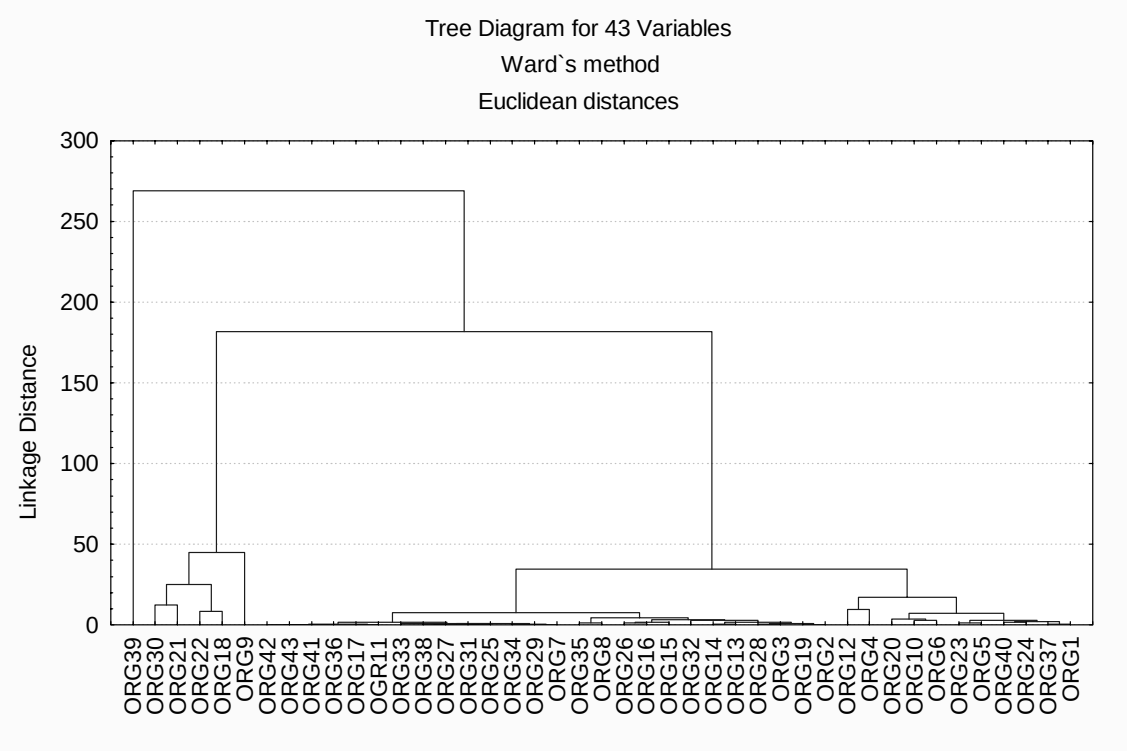

Рисунок 2. Дендрограмма иерархической классификации органических веществ в пробах речной воды рек Пякупур, Пур, Айваседопур.

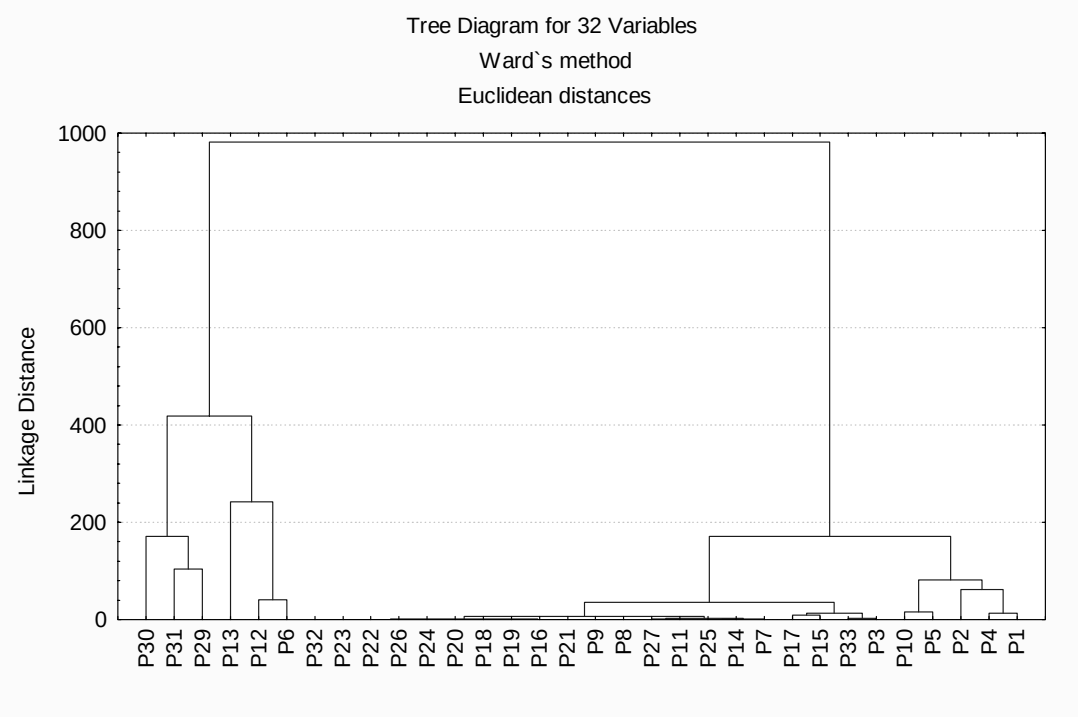

Рисунок 3. Дендрограмма иерархической классификации физико-химических показателей качества речной воды по методу Варда.

По данным рис. 3 достаточно четко выделяются четыре кластера физико-химических показателей (табл. 8).

Таблица 8. Состав кластеров физико-химических показателей

Номер кластера	Показатели, входящие в кластер
1	Минерализация (P29); Сухой остаток (P30); Гидрокарбонаты (P28)
2	Натрий (P6), Сульфаты (P12), Хлориды (P13)
3	pH (P1), Окисляемость (P2), Кальций (P4), Магний (P5), Нитраты (P10)
4	Все остальные показатели

Выделенные кластеры физико-химических показателей достаточно четко отражают их естественные взаимосвязи в поверхностных водах.

4.2. Статистические контрольные карты физико-химических показателей

На рис. 4 -9 приведены X-R контрольные карты для исследуемых физико-химических показателей (нулевая точка на всех картах отвечает значению ПДК для питьевой воды). Эти карты показывают, что средние значения всех показателей, кроме алюминия (P25, рис.8) превышают значения ПДК при очень высокой статистической достоверности ($p<0.01$).

Обращает на себя внимание достаточно высокий уровень колебаний средних значений показателей по сезонам. Например, содержание нефтепродуктов (рис. 9) в период 1998-2000 г.г. намного больше, чем в 2002-2003 г.г., что может быть связано с интенсивностью добычи полезных ископаемых в эти периоды.

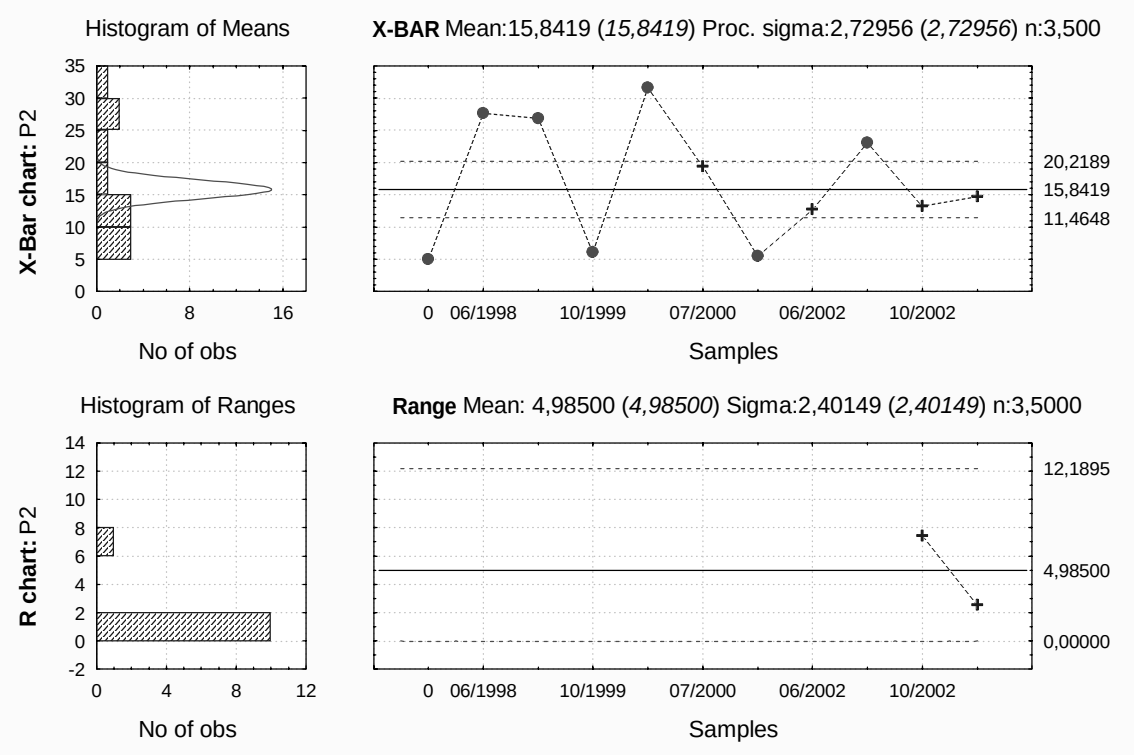

Рисунок 4. Контрольная карта перманганатной окисляемости

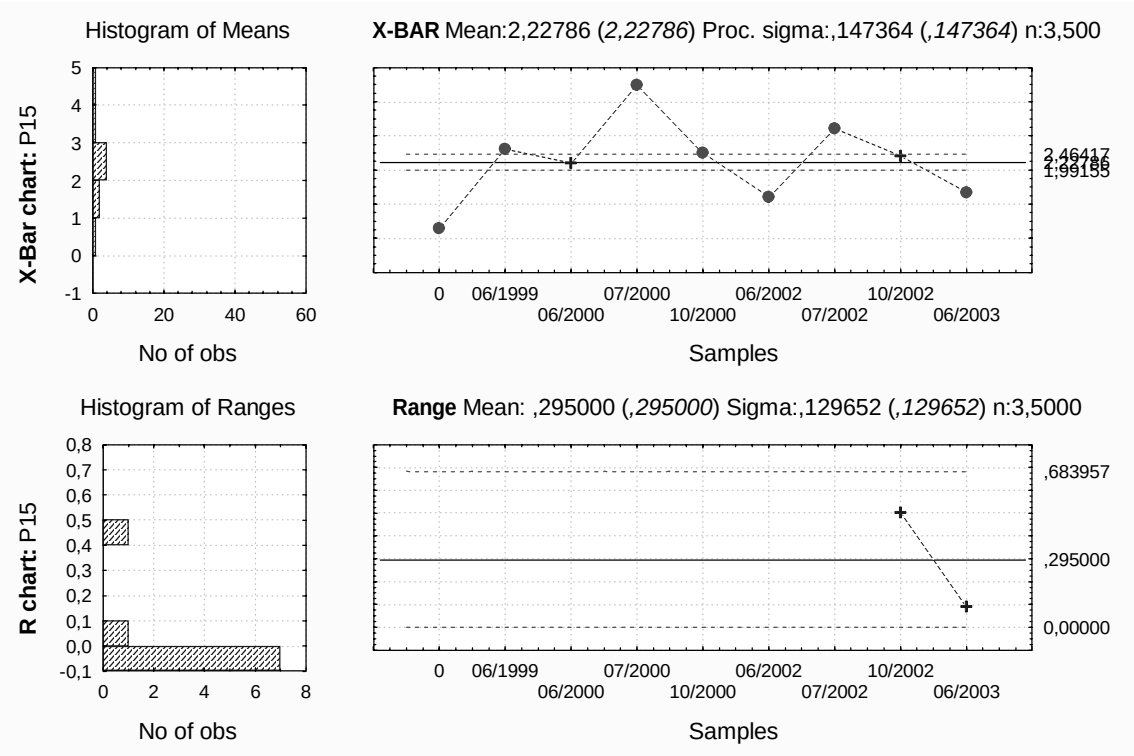

Рисунок 5. Контрольная карта содержания железа

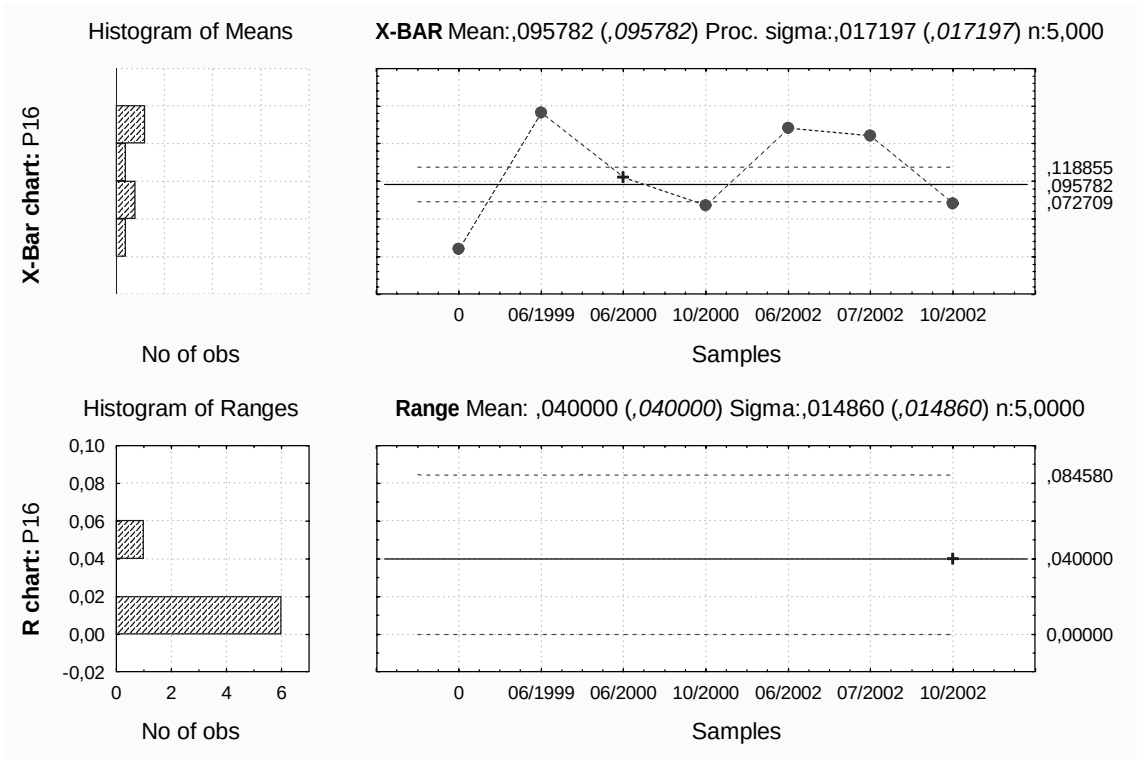

Рисунок 6. Контрольная карта содержания марганца

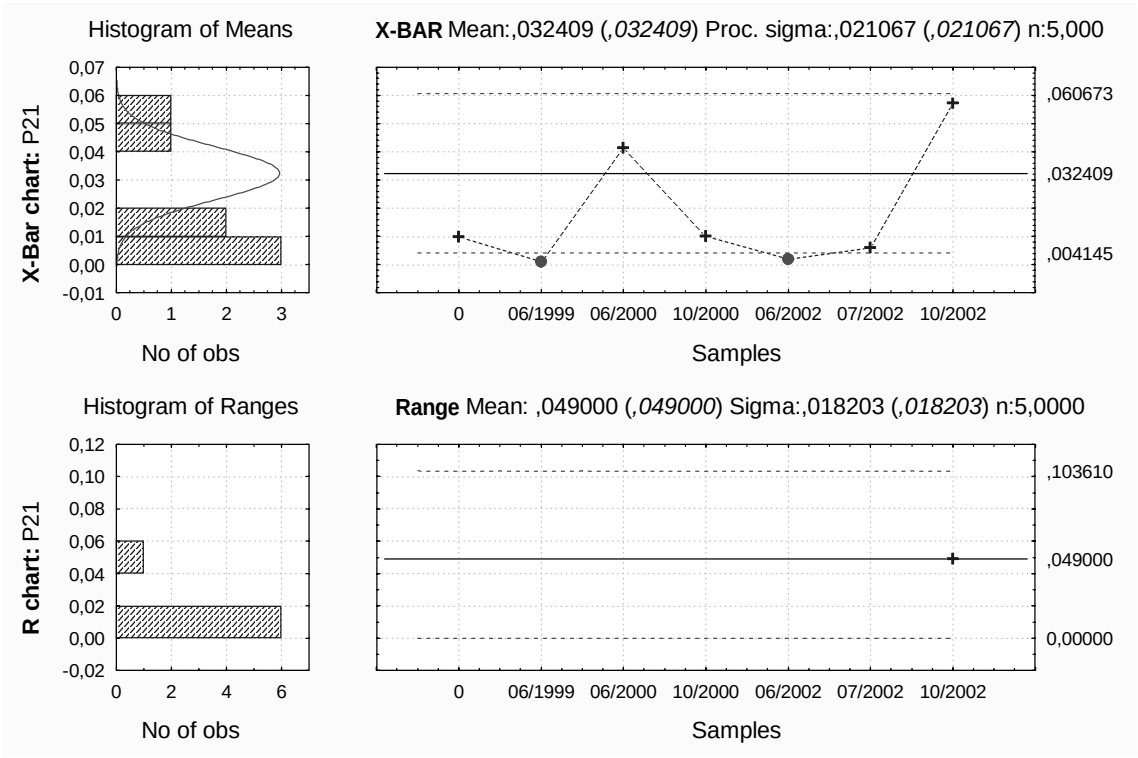

Рисунок 7. Контрольная карта содержания никеля

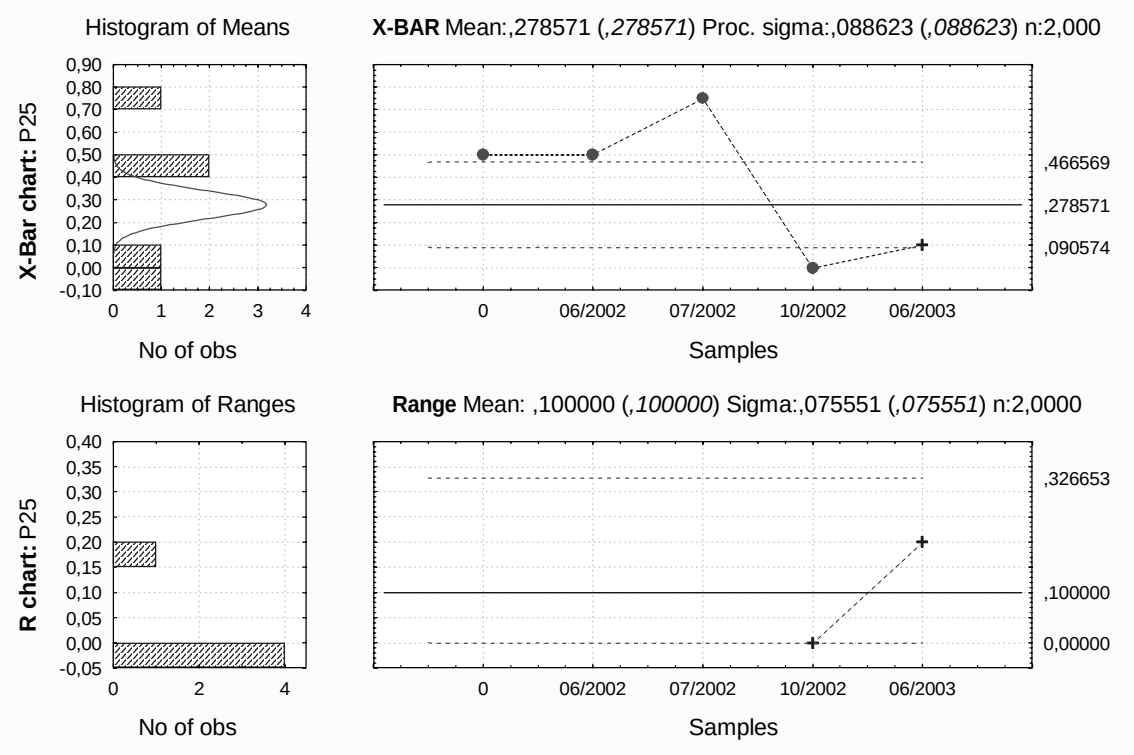

Рисунок 8. Контрольная карта содержания алюминия

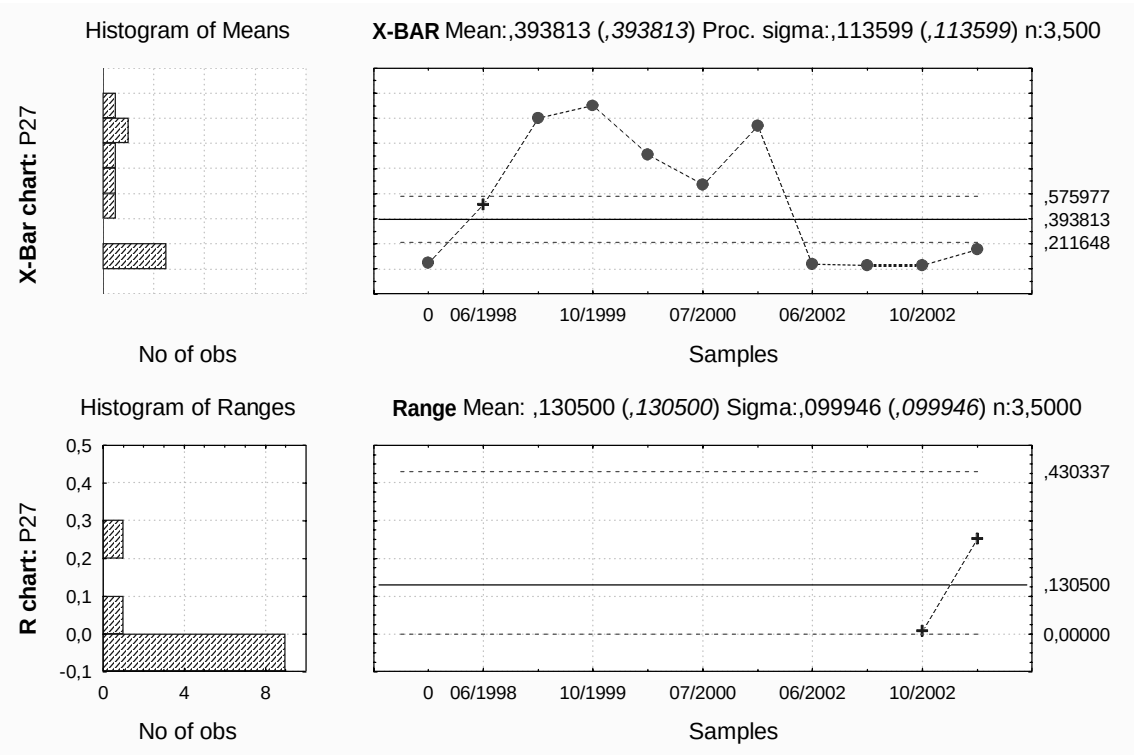

Рисунок 9. Контрольная карта содержания нефтепродуктов

4.3. Кластер-анализ образцов в пространстве физико-химических показателей

В данном разделе приведены результаты кластерного анализа образцов речной воды Пуровского района по методу k-средних с использованием пяти показателей качества: окисляемости (P2), содержания железа (P15), марганца (P16), никеля (P21), нефтепродуктов (P27). Эти показатели превышают в среднем значения соответствующих ПДК.

На рис. 10 представлена соответствующая иерархическая дендрограмма Варда. Данные дендрограммы показывают возможность разделения всех исследуемых образцов воды на четыре кластера.

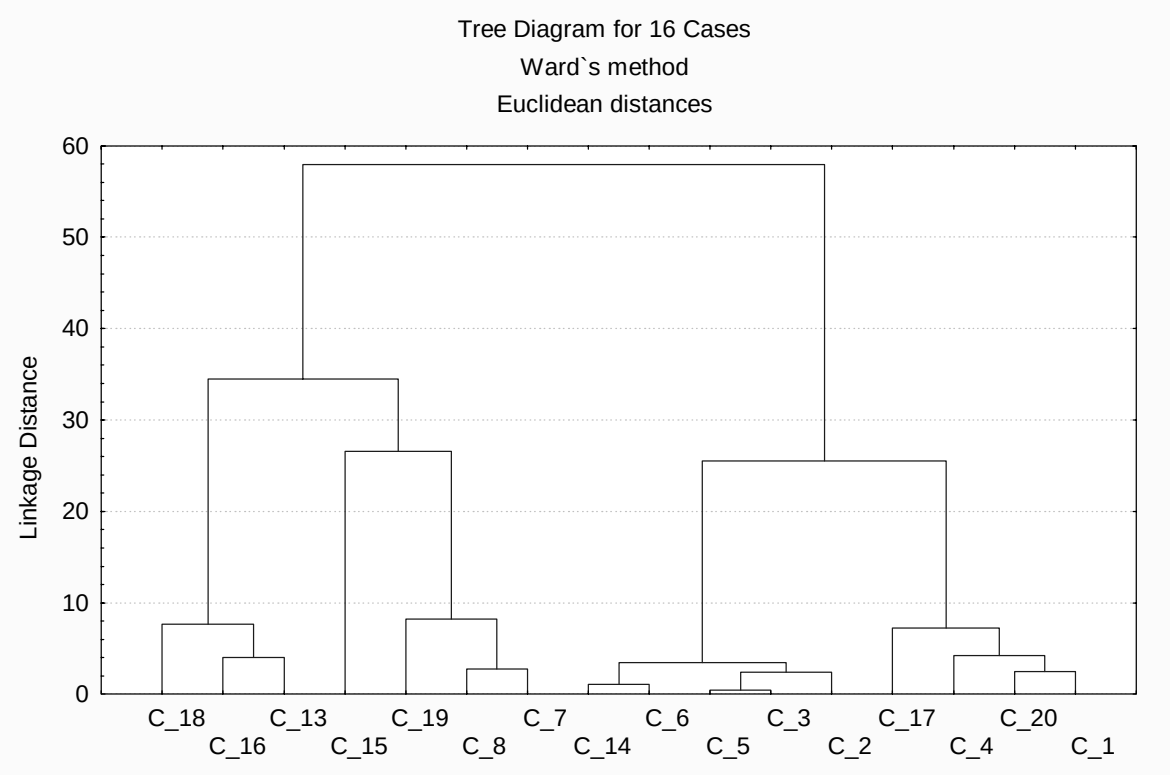

Рисунок 10. Дендрограмма проб речной воды Пуровского района, построенная по методу Варда с использованием пяти показателей.

Состав кластеров представлен в табл.9.

Таблица 9. Состав полученных кластеров

Номер кластера	Состав соответствующего кластера
1	П.1 (октябрь 2002) П.3 (июнь 2003) П.5 (июль 2000)
2	П.2 (октябрь 2002) П.4 (октябрь 2002) П.5 (октябрь 2002) П.5 (июнь 2002) П.5 (июнь 2003)
3	ПДК П.3 (октябрь 2002) П.5 (октябрь 1999) П.5 (октябрь 2000)
4	П.5 (июнь 1998) П.5 (июнь 1999) П.5 (июнь 2000) П.5 (июль 2002)

В табл. 10 и на рис.11 приведены средние значения исследуемых показателей по кластерам 1-4.

Таблица 10. Средние значения показателей в соответствующих кластерах

Показатель	Кластер			
	1	2	3	4
Окисляемость	17,19	13,52	6,32	27,25
Железо	2,66	2,01	1,83	2,55
Марганец	0,08	0,09	0,06	0,13
Никель	0,04	0,03	0,02	0,02
Нефтепродукты	0,24	0,08	0,62	0,66

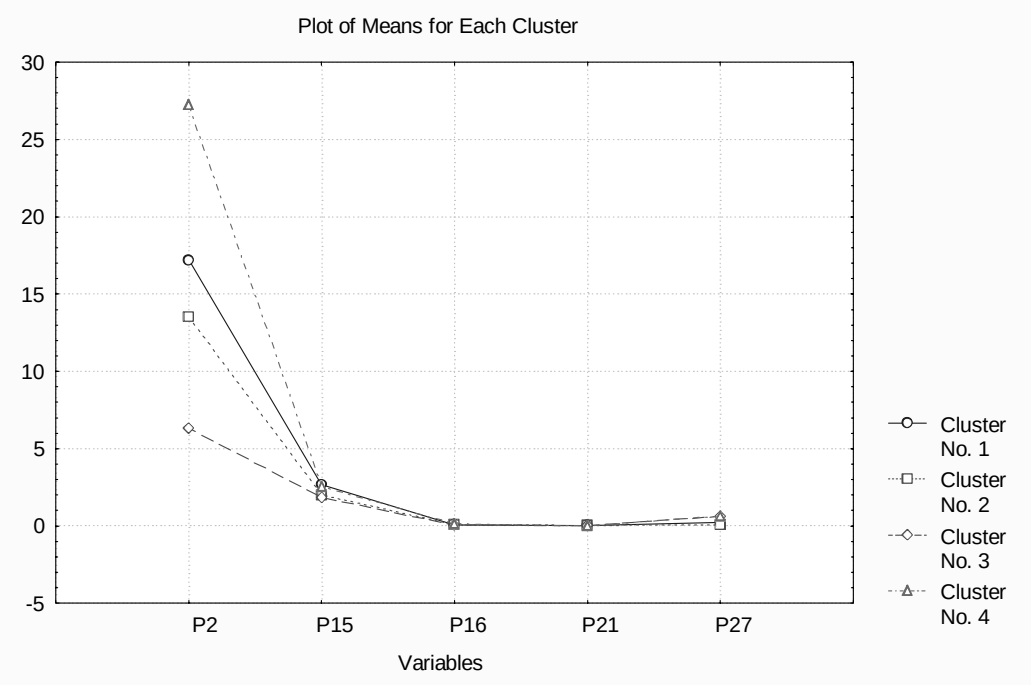

Рисунок 11. Средние значения показателей качества воды в соответствующих кластерах

Данные табл. 10 и рис. 11 показывают основное различие кластеров по перманганатной окисляемости, которая зависит от содержания органических веществ. Наиболее низкое значение окисляемости приходится на третий кластер, объединяющий пробы воды, взятые из р. Пякупур в октябре 1999-2002.

Наиболее высокое содержание нефтепродуктов в р. Пякупур приходится на период 1998-2000 г.г.

4.4. Кластер-анализ образцов в пространстве органических веществ

Матрица эвклидовых расстояний между образцами воды в 43-мерном пространстве органических веществ представлена в табл. 11.

Таблица 11. Матрица эвклидовых расстояний между объектами

Точка отбора проб	5	6	7	8	9
5	0,0	50,6	40,7	26,2	11,2
6	50,6	0,0	23,4	42,7	53,1
7	40,7	23,4	0,0	25,2	41,7
8	26,2	42,7	25,2	0,0	26,5
9	11,2	53,1	41,7	26,5	0,0

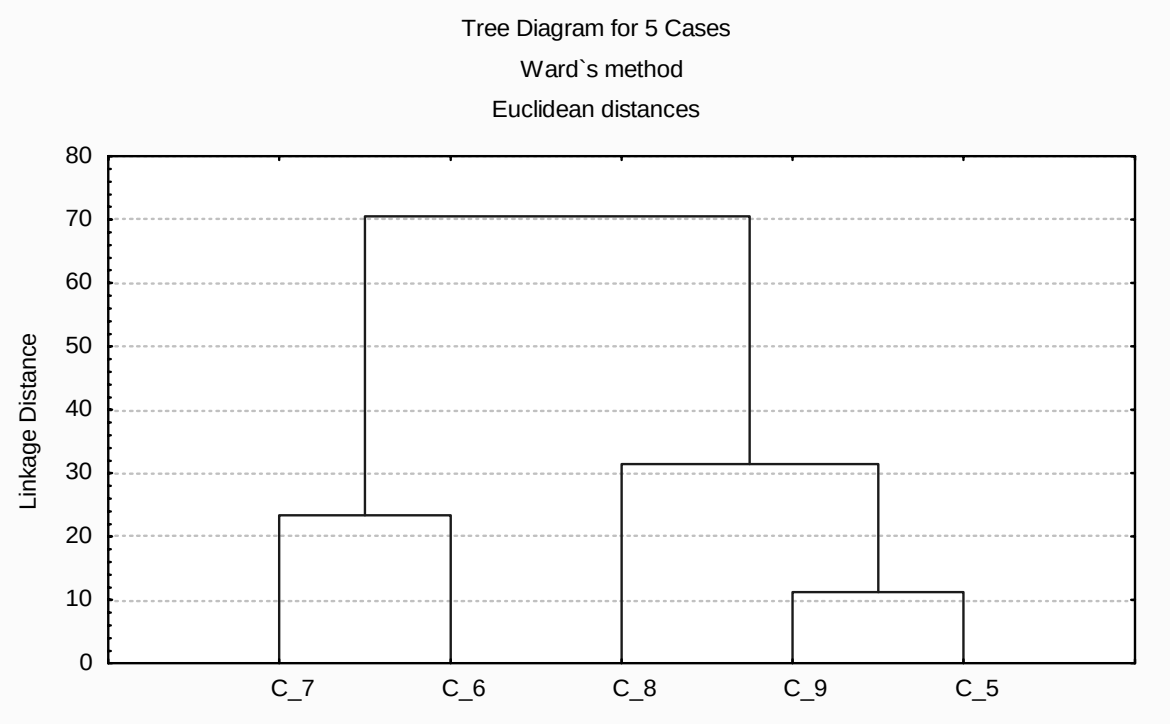

Рисунок 12. Дендрограмма образцов речной воды Пуровского района ЯНАО по содержанию органических веществ.

На рис. 12. показана дендрограмма многомерной классификации образцов речной воды Пуровского района. Согласно этой дендрограмме можно выделить два кластера. В первый кластер входят три образца, взятые из рек Пякупур и Пур, во второй – образцы, взятых из реки Айваседопур.

Для количественного анализа составляющих обоих кластеров применим к этим данным метод k-средних.

Результаты многомерного дисперсионного анализа для двух кластеров приведены в табл. 12.

Таблица 12. Результаты многомерного дисперсионного анализа (MANOVA) для двух кластеров

Вещество	Среднее для кластера		MANOVA			
	1	2	SS между кластерами	SS внутри кластера	F (1; 2)	p
Org 1	1,476	1,025	0,24	0,41	1,78	0,27
Org 2	0,000	0,465	0,25	0,43	1,80	0,27
Org 3	0,250	0,760	0,31	0,50	1,87	0,26
Org 4	0,000	3,895	18,20	30,34	1,80	0,27
Org 5	1,813	1,475	0,13	5,47	0,07	0,80
Org 6	3,283	1,935	2,18	1,36	4,80	0,12
Org 7	0,166	0,000	0,03	0,07	1,42	0,32
Org 8	1,113	0,555	0,37	2,30	0,48	0,54
Org 9	7,633	34.470	864,24	187,56	13,82	0,03**
Org 10	2,906	2,765	0,02	12,47	0,01	0,94
Org 11	0,000	0,225	0,06	0,10	1,80	0,27
Org 12	0,000	8,655	89,89	2,35	114,53	0,01**
Org 13	0,000	1,190	1,69	0,18	28,32	0,01**
Org 14	0,216	1,150	1,04	0,49	6,27	0,08**
Org 15	0,900	0,930	0,001	1,40	0,002	0,96
Org 16	0,593	0,130	0,26	0,86	0,89	0,41
Org 17	0,000	0,150	0,02	0,04	1,80	0,27
Org 18	12,173	17,520	34,03	26,46	3,88	0,14
Org 19	0,000	0,630	0,47	0,79	1,80	0,27

Вещество	Среднее для кластера		MANOVA			
	1	2	SS между кластерами	SS внутри кластера	F (1; 2)	p
Org 20	2,220	3,560	2,15	1,84	3,50	0,16
Org 21	9,476	12,240	9,16	7,46	3,68	0,15
Org 22	16,483	18,945	7,27	8,83	2,47	0,21
Org 23	2,243	1,030	1,76	1,85	2,85	0,19
Org 24	1,560	1,860	0,11	1,74	0,18	0,69
Org 25	0,370	0,000	0,16	0,25	1,89	0,26
Org 26	0,803	0,720	0,01	2,43	0,01	0,92
Org 27	0,300	0,165	0,02	0,09	0,76	0,44
Org 28	0,330	0,500	0,03	0,67	0,15	0,72
Org 29	0,223	0,000	0,05	0,29	0,60	0,49
Org 30	6,330	7,135	0,77	83,62	0,02	0,87
Org 31	0,340	0,160	0,03	0,23	0,49	0,53
Org 32	0,836	1,385	0,36	2,34	0,46	0,54
Org 33	0,223	0,380	0,02	0,37	0,23	0,65
Org 34	0,290	0,000	0,10	0,07	4,06	0,13
Org 35	0,736	0,000	0,65	3,25	0,6	0,49
Org 36	0,153	0,080	0,01	0,002	8,54	0,06**
Org 37	1,483	0,975	0,31	0,12	7,21	0,07**
Org 38	0,346	0,150	0,04	0,03	4,31	0,12
Org 39	74,490	47,965	844,29	389,99	6,49	0,08**
Org 40	1,796	1,930	0,02	0,36	0,18	0,70
Org 41	0,003	0,000	0,00	0,00	0,60	0,49
Org 42	0,050	0,065	0,00	0,01	0,11	0,75
Org 43	0,050	0,040	0,00	0,00	0,90	0,41

Таблица содержит результаты однофакторного дисперсионного анализа, в котором в качестве группирующего фактора выступает номер кластера. В первом

столбце - список 43 переменных, далее идут средние значения для кластеров, суммы квадратов (SS), затем F-критерий Фишера и в последнем столбце - достигнутый уровень значимости "p".

Как видно из этой таблицы, нулевая гипотеза о равенстве средних значений отвергается при p=0,10 для семи соединений (Org 9, Org 12, Org 13, Org 14, Org 36, Org 37, Org 39). Ниже на рис. 13 приведен график средних значений всех переменных по отдельным кластерам.

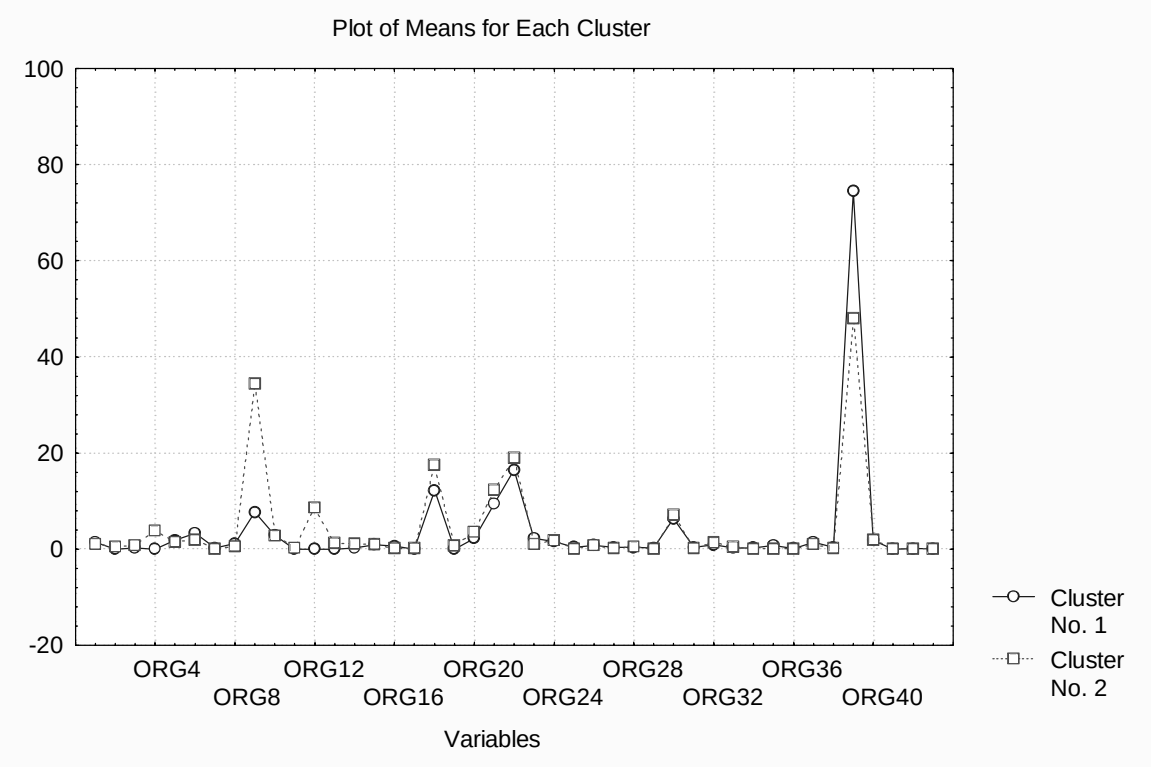

Рисунок 13. Средние значения концентраций органических веществ в кластерах 1 и 2.

По результатам проведенного кластерного анализа хромато-масс-спектроскопических данных мы можем составить таблицу (табл. 13) фонового содержания органических веществ в реках Пуровского района ЯНАО.

Таблица 13. Содержание органических веществ в реках Пякупур, Айваседопур и Пур

Код вещества	Вещество	Средняя концентрация		Стандартное отклонение	
		р. Пякупур, р. Пур	р. Айваседопур	р. Пякупур, р. Пур	р. Айваседопур
Org 1	Толуол	1,296		0,404	
Org 2	Масляная кислота	0,186		0,415	
Org 3	Хлоруксусная кислота, этиловый эфир	0,454		0,450	
Org 4	4-гидрокси-4-метилпентанон-2	1,558		3,483	
Org 5	Сложный эфир	1,678		1,184	
Org 6	2,6-Ди-t-бутил-4-гидрокси-4-метил-2,5-циклогексадиен –1-он	2,744		0,941	
Org 7	Ионол	0,100		0,161	
Org 8	Карбоновая кислота	0,890		0,818	
Org 9	Σ Нафтенов	7,633	34,470	2,091	13,067
Org 10	ΣРазветвл. Алканов	2,850		1,767	
Org 11	Пентановая кислота	0,090		0,201	
Org 12	Хлорэтилхлорметиловый эфир	0,000	8,700	0,000	1,534
Org 13	Алкилдиамин	0,000	1,190	0,000	0,424
Org 14	Дихлорпропанол	0,216	1,150	0,375	0,467

Код вещества	Вещество	Средняя концентрация		Стандартное отклонение	
		р. Пякупур, р. Пур	р. Айваседопур	р. Пякупур, р. Пур	р. Айваседопур
Org 15	Дихлоруксусная кислота, этиловый эфир	0,912		1,610	
Org 16	1,1 – Дихлор – 2 – этоксиэтан	0,408		1,280	
Org 17	Серосодержащие соединения	0,060		0,134	
Org 18	Гексановая кислота	14,312		3,897	
Org 19	Этиленгликоль диацетат	0,252		0,563	
Org 20	Гептановая кислота	2,756		0,999	
Org 21	Октановая кислота	10,582		2,038	
Org 22	Нонановая кислота	17,468		2,006	
Org 23	Амид кислоты	1,758		0,951	
Org 24	Декановая кислота	1,680		0,680	
Org 25	Ароматич. амин	0,222		0,325	
Org 26	Додекановая кислота (лауриновая)	0,770		0,781	
Org 27	Пентилтиазол	0,246		0,163	
Org 28	Тетрадекановая кислота (миристиновая)	0,398		0,422	
Org 29	Акриловая кислота, эфир	0,134		0,299	

Код вещества	Вещество	Средняя концентрация		Стандартное отклонение	
		р. Пякупур, р. Пур	р. Айваседопур	р. Пякупур, р. Пур	р. Айваседопур
Org 30	Пальмитиновая кислота	6,652		4,593	
Org 31	4-(1,5-Диметил-3-оксогексил) -1-циклогексен -1-карбоновая кислота (Javabione)	0,268		0,262	
Org 32	Метоксикоричная кислота	1,056		0,822	
Org 33	Производное фенилизоцианата	0,286		0,316	
Org 34	Эфир бензойной кислоты	0,174		0,209	
Org 35	Сквален	0,442		0,988	
Org 36	Диэтилфталат	0,153	0,080	0,032	0,014
Org 37	Диизобутилфталат	1,483	0,975	0,248	0,078
Org 38	Бутилизобутил-фталат	0,268		0,140	
Org 39	Дибутилфталат	74,490	47,965	13,952	0,827
Org 40	Бис(2-этилгексил) Фталат	1,850		0,309	
Org 41	Нафталин	0,002		0,004	
Org 42	Фенантрен	0,056		0,042	
Org 43	Диметилсульфид	0,046		0,011	

По данным представленным в табл. 7 можно утверждать, что в реке Айваседопур повышено содержание нафтенов, дихлорпропанола, алкилдиамина и

хлорметилхлорэтилового эфира по сравнению с реками Пур и Пякупур. Тогда как содержание диэтилфталата, диизобутилфталата и дибутилфталата выше в реках Пякупур и Пур, чем в Айваседопуре.

5. Заключение

1. Разработан алгоритм принятия решений по данным мониторинга химического состава речной воды, включающий набор статистических методов – дескриптивной статистики, корреляционного анализа, многомерного кластерного анализа и дисперсионного анализа. Этот алгоритм позволяет количественно охарактеризовать химический состав речной воды в виде многомерного вектора средних значений концентраций и отслеживать его изменение в географическом пространстве и во времени.

2. Методом статистических контрольных карт определены изменения во времени перманганатной окисляемости, содержания железа, никеля, марганца и нефтепродуктов в р. Пякупур. Показана существенная сезонная и временная зависимость суммарного содержания органических веществ и нефтепродуктов в поверхностной воде.

2. Определены фоновые концентрации органических веществ в реках Айваседопур, Пякупур и Пур по состоянию на июнь 2003 года.

6. Список литературы

1. Глудкин О.П., Горбунов Н.М., Гуров А.И., Зорин Ю.В. Всеобщее управление качеством. М.: ЛБЗГТ. 2001. 600 с.
2. Управление качеством/ Под ред. С.Д. Ильенковой. М.: ЮНИТИ. 2000. 199 с.
3. «Семь инструментов качества» в японской экономике. М.: Изд. стандартов. 1990. 88 с.
4. Макино Т., Охаси М., Докэ Х., Макино К. Контроль качества с помощью персонального компьютера. М.: Машиностроение. 1991. 224 с.
5. AT&T (1956). Statistical quality control handbook, Select code 700-444. Indianapolis, AT&T Technologies.
6. Nelson, L. (1984). The Shewhart control chart - tests for special causes.// Journal of Quality Technology, 15, 237-239.
7. Nelson, L. (1985). Interpreting Shewhart X-bar control charts// Journal of Quality Technology, 17, 114-116.
8. Grant, E. L., & Leavenworth, R. S. (1980). Statistical quality control (5th ed.). New York: McGraw-Hill
9. Shirland, L. E. (1993). Statistical quality control with microcomputer applications. New York: Wiley.
10. Montgomery, D. C. (1991) Design and analysis of experiments (3rd ed.). New York: Wiley.
11. Bhote, K. R. (1988). World class quality. New York: AMA Membership Publications.
12. Duncan, A. J. (1974). Quality control and industrial statistics. Homewood, IL: Richard D. Irwin.
13. Montgomery, D. C. (1985). Statistical quality control. New York: Wiley.

14. ASQC/AIAG (1991). Fundamental statistical process control reference manual. Troy, MI: AIAG.
15. Сокал З. Кластер-анализ и классификациф: предпосылки и основные направления// Классификация и кластер. - М.: Мир, 1980. - С. 7-19.
16. Sneath P.H., Sokal R.R. Numerical Taxonomy. W.H. Freeman, San Francisco. 1973. 573 p.
17. Гуд И.Д. Ботриология ботриологии// Классификация и кластер. - М.: Мир, 1980. - С. 66-82.
18. Good I.J. Categorization of classification in Mathematics and Computer Science in Biology and Medicine. - London, 1965. - P. 115-125.
19. Соломон Г. Зависящие от данных методы кластер-анализа// Классификация и кластер - М.: Мир, 1980. - С. 129-147.
20. Pearson K. On lines and planes of closest fit to systems of point in space// The London, Edinburgh and Dublin Philosophical Magazaine and Journal of Science. - 1901. - V.2. - P. 559-572.
21. Гейссер С. Распознавание: отнесение и разделение. Линейные аспекты/ / Классификация и кластер. - М.: Мир, 1980. - С. 248-274.
22. Fisher R.A. The use of multiple measurements in taxonomic probltms// Annals. Eugenics. - 1936. - V.7. - P. 179-188.
23. Beckner M. The Biological Way of Thougt. - Columdia Univ. Press, New York, 1959. - 200 p.
24. Chernoff H. The use of faces to represent points in k-dimensional space graphically// J. Amer. Stat. Assoc. - 1973. - V.68. - P. 361-368.
25. Закс Л. Статистическое оценивание. - М.: Мир, 1976. - 599 с.
26. Мандель И.Д. Кластерный анализ. - М.: Финансы и статистика, 1988. - 176 с.
27. Ward J. Hierarchical grouping to optimize an objective function// J. Amer. Stat. Assoc. - 1963. V.58. - P. 236.

28. Hotteling H. Relations between two sets of variates// Biometrika. – 1936. – V.28. – P. 321-377.
29. Андерсон Т. Введение в многомерный статистический анализ. – М.: Физматгиз, 1963. – 500 с.
30. Боровиков В.П., Боровиков И.П. Statistica. Статистический анализ и обработка данных в среде Windows. - М.: Филинъ, 1997. - 608 с.
31. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. – М.: Наука, 1976.- 736 с.
32. Уилкс С. Математическая статистика. – М.: Наука, 1967. – 632 с.
33. Сошников Л.А., Тамашевич В.Н., Уеба Г., Шефер М. Многомерный статистический анализ в экономике. – М.: ЮНИТИ, 1999. – 598 с.
34. Tryon R.C. Cluster Analysis// Ann. Arb., Edw. Brathers. - 1939
35. Tryon R.C. Cluster Analysis. New York: McGraw-Hill. - 1939.
36. Sokal R. And P.Sneat (1963) Principles of Numerical Taxonomy. San Francisco: W.H.Freeman
37. Мандель И.Д. Кластерный анализ. - М.: Финансы и статистика. 1988. - 176с.
38. Б.Болч, К.Дж. Хуань. Многомерные статистические методы для экономики/Пер. с англ. - М.: Статистика, 1979. - 317с.
39. Айвазян С.А., Бежаева З.И., Староверов О.В. Классификация многомерных наблюдений. - М.: Статистика, - 1974, - 240 с.
40. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. - М.: Финансы и статистика, 1989. - 607с.
41. Александров В.В., Горский Н.Д. Алгоритмы и программы структурного метода обработки данных. - Л.: Наука, - 1983, - 208с.
42. Дюран Б., Оделл П. Кластерный анализ. - М.: Статистика, - 1977, - 128 с.

43. Факторный, дискриминантный и кластерный анализ: Пер с англ./Дж. - О.Ким, Ч.У. Мюллер, У.Р. Клекка и др.; Под ред. И.С. Енюкова. - М.: Финансы и статистика, 1989. - 215с.
44. Хемометрика/ Шараф М.А, Иллмэн Д.Л., Ковальски Б.Р. Пер. С англ. - Л., Химия, 1989. - 272с.
45. Браверман Э.М., Мучник И.Б. Структурные методы в обработке эмпирических данных. М.: Наука, - 1983.
46. Миркин Б.Г. Группировки в социально-экономических исследованиях. М.: Финансы и статистика.
47. Жамбю М. Иерархический кластер-анализ и соответствия: Пер. с фр. М.: Финансы и статистика, 1988. - 342с.
48. Статистические методы для ЭВМ/ Под ред. К.Энслейна, Э.Рэлстона, Г.С.Уилфа: Пер с англ./Под ред. М.Б.Малютова. - М.: Наука, 1986. - 464с.
49. Айвазян С.А., Бухштабер В.М. Анализ данных, прикладная статистика и построение общей теории автоматической классификации// Методы анализа данных/ Пер. с фр. - М.: Финансы и статистика, 1985. - Вступ. ст. - с. 5-22.
50. Айвазян С.А., Бежаева З.И., Староверов О.В. Классификация многомерных наблюдений. М.: Статистика, 1974. – 240 с.
51. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. - М.: Наука, 1973. – 416 с.
52. Елисеева И.И., Рукавишников В.О. Группировка, корреляция, распознавание образов: Статистические методы классификации и измерения связей. - М.: Статистика, 1977. – 143 с.