Тема 12. КЛАСТЕРНЫЙ АНАЛИЗ В ЭКОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Объяснение. *Кластерный анализ* – совокупность алгоритмов обработки данных, предназначенных для распределения исследуемых объектов на относительно однородные группы (кластеры).

В области экологии широко применяется для выделения пространственно однородных групп организмов, сообществ и т. п. Реже методы кластерного анализа применяются для исследования сообществ во времени.

Иногда биологи ставят цель разбить животных на различные виды, чтобы содержательно описать различия между ними. В соответствии с современной системой, принятой в биологии человек принадлежит к приматам, млекопитающим, амниотам, позвоночным и животным. В данной классификации, чем выше уровень агрегации, тем меньше сходства между членами в соответствующей группе. Человек имеет больше сходства с другими приматами (т.е. с обезьянами), чем с "отдаленными" членами класса млекопитающих (например, собаками) и т.д.

Проведение кластерного анализа может быть реализовано в программе Statistica несколькими методами: Иерархическая классификация, Двухвходовое объединение и Метод *K* средних.

Последний метод используется, если вы уже имеете гипотезы относительно числа кластеров по наблюдениям или по переменным. Вы можете "сказать" системе образовать ровно три кластера так, чтобы они были настолько различны, насколько это возможно. Это именно тот тип задач, которые решает алгоритм метода *K* средних. В общем случае метод *K* средних строит ровно *k* различных кластеров, расположенных на возможно больших расстояниях друг от друга.

Исходными данными для анализа могут быть первичные данные (объекты и их параметры) или матрица расстояний между объектами. В основе процедур кластерного анализа лежит группировка сходных между собой объектов в некоторые группы (или кластеры). Именно поэтому понятие сходства имеет для него первостепенное значение. Несмотря на кажущуюся простоту, понятие сходства и особенно процедуры, используемые при измерении сходства, не так просты. Количественное оценивание сходства отталкивается от понятия метрики или расстояния

(*distance*) между объектами. Интуитивно понятно, что чем меньше расстояние между объектами, тем больше сходство между ними.

Одной из мер сходства является Евклидово расстояние. Например, если объект описывается двумя параметрами, то он может быть изображен точкой на плоскости, а расстояние между объектами – это расстояние между точками, вычисленное по теореме Пифагора. Нужно возвести в квадрат расстояние по каждой координате, суммировать их и из полученной суммы извлечь квадратный корень.

Расстояние d $(x, y) = \sqrt{(x_i - y_i)}$

Рассмотрим применение кластерного анализа в экологических исследованиях. Метод позволяет сгруппировать водные объекты в кластеры с разным уровнем загрязнения. Исходные данные приведены в таблице 13.

Таблица 13 – Среднегодовая концентрация вещества в реке или на значительном ее протяжении в 2009 г. (в долях ПДК)

Речной бассейн	Фосфаты (по Р)	Азот аммонийный	Азот нитритный
р. Западная Двина	1,2	1,4	1,7
р. Неман	1,2	1,6	0,9
р. Западный Буг	3,5	1,9	1,9
р. Днепр	4,4	2,8	3,3
р. Березина	1,3	2,3	1,2
р. Свислочь (ниже Минска)	11	6,1	5,5
р. Сож	1,1	1,1	0,5
р. Припять	1,2	1,5	1,2

В результате обработки данных методом полной связи *(Complete linkage)* получена дендрограмма (рис. 3), на которой объекты исследования разделены на несколько групп. Наиболее тесным сходством по уровню загрязнения характеризуются Западная Двина, Неман, Припять, Березина, Сож. Во вторую группу объединены Западный Буг и Днепр. Концентрации загрязнителей в Свислочи наиболее сильно отличаются от других объектов.

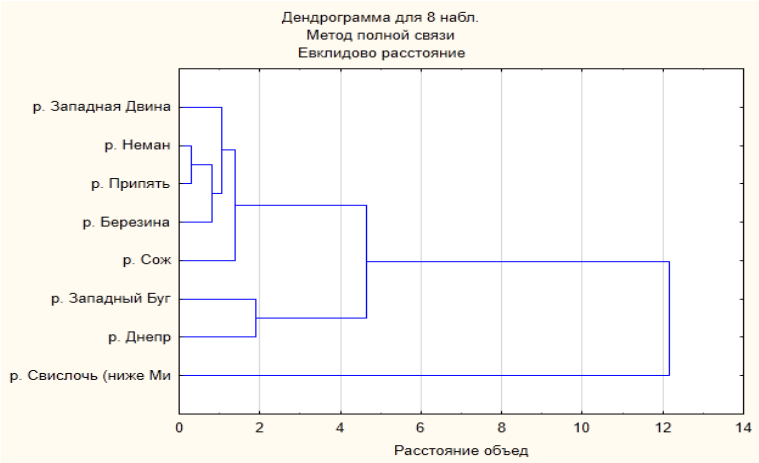

Рисунок 3 – Дендрограмма результатов кластерного анализа

Так как в имеющейся структуре связей возможно выделение трех групп, то продолжение обработки данных можно провести при помощи метода *K* средних (рис. 4). Это итеративный метод, который работает непосредственно с объектами, а не с матрицей сходства. Он отличается тем, что позволяет заранее задать число кластеров. Это число определяет сам пользователь, исходя из имеющейся задачи и предсказаний теории. Метод *K* средних разобьет все объекты на заданное количество кластеров, которые будут максимально различаться между собой.

В этом методе объект относится к тому классу, расстояние до которого минимально. Расстояние понимается как евклидово расстояние, то есть объекты рассматриваются как точки Евклидова пространства. Вначале задается некоторое разбиение данных на кластеры (число кластеров определяется пользователем) и вычисляются центры тяжести кластеров. Затем происходит перемещение каждой точки в ближайшей к ней кластер. Затем снова вычисляются центры тяжести новых кластеров и процесс повторяется, пока не будет найдена стабильная конфигурация (то есть кластеры перестанут изменяться) или число итераций не превысит заданное пользователем.

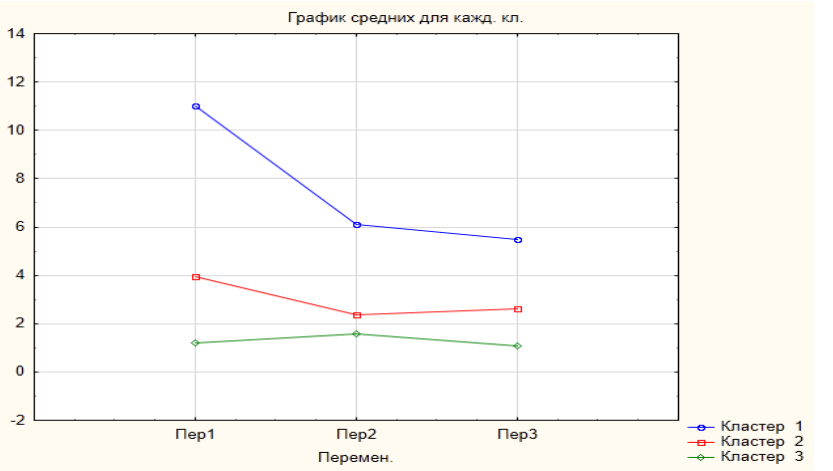

Рисунок 4 – Результаты кластеризации методом К средних

В кластер 1 входит река Свислочь (№6), в кластер 2 входят реки Западный Буг и Днепр (№3 и 4), к третьему кластеру относятся водные объекты с номерами 1, 2, 5, 7 и 8. От первого к третьему кластеру уменьшается уровень загрязнения рек.

З**адание**. Выполнить кластерный анализ данных (приложение 8) в программе Statistica.