Лекция № 12

Тема: Методы исследования компонентного анализа геосистем

Цель: рассмотреть основные методы оценки степени антропогенных воздействий на геосистемы

1. Общая характеристика
2. Этапы компонентного анализа геосистем

Компоненты геосистем – это результат взаимопроникновения и взаимодействия качественно разных тел, мы имеем основание рассматривать их как первую ступень географической интеграции; вторая ступень – собственно геосистемы как наиболее сложная форма организации природных тел на Земле. По отношению к геосистемам географические компоненты служат структурными частями первого порядка, точнее – частями их вертикальной (радиальной, ярусной) структуры, поскольку им присуще упорядоченное ярусное расположение внутри геосистемы. Особенность географических компонентов состоит в том, что в каждом из них присутствует вещество всех остальных компонентов, и это придает им новые свойства, которыми не могло бы обладать химически чистое и физически однородное вещество.

Компонентный анализ – один из методов многомерной статистики, в его основу положена гипотеза: наблюдаемые или измеряемые параметры являются лишь косвенными характеристиками изучаемого объекта или явления. Существуют внутренние (скрытые, не измеряемые) параметры, называемые главными компонентами, как предполагается, сохраняют всю информацию, содержащуюся во множестве наблюдаемых переменных. И хотя такие компоненты заранее нам не известны, компонентный анализ ставит задачу представить наблюдаемые параметры в виде линейных комбинаций главных компонент и определить их, т.е. для каждого объекта указать значение каждой главной компоненты. В таком случае модель компонентного анализа может быть записана в виде, формула (1):

$$Y\,[n \times m] = F\,[n \times m] \cdot A\,[m \times m] \qquad (1)$$

где $Y\,[n \times m]$ представляет собой совокупность всех *n* наблюдаемых значений всех *m* параметров;

$F\,[n \times m]$ - матрица, включающая совокупность всех *n* получаемых значений всех *m* главных компонент, это искомая матрица значений новых переменных в каждой точке опробования;

$A\,[m \times m]$ - так называемая матрица компонентных нагрузок, или весовая матрица, она является связующим звеном между старыми и новыми переменными.

В данном матричном уравнении две неизвестные матрицы – $F\,[n \times m]$ и $A\,[m \times m]$, а так как из одного уравнения можно найти только одно неизвестное, для определения второго требуется дополнительное условие. Таким дополнительным условием, исходной предпосылкой анализа, является наличие взаимосвязи между несколькими одновременно наблюдаемыми переменными. В качестве количественной меры связи между двумя переменными используется коэффициент корреляции. Он может принимать значения от -1 до +1. При этом если он приближается к 0, это свидетельствует об отсутствии линейной связи, и чем более он близок к +1 или -1, тем более тесная линейная связь существует между переменными.

Все вычисленные коэффициенты корреляции между каждой парой переменных располагаются соответствующим образом в корреляционной матрице. В ней содержится важная информация о взаимоотношениях переменных с учетом влияния помех, причиной которых может явиться, например, неоднородность материала. При анализе такой корреляционной матрицы получают структуру искомых гипотетических величин (матрицу $A\,[m \times m]$), которые находятся в определенных взаимоотношениях с переменными.

Эта структура находится в результате математических преобразований корреляционной матрицы, вычисленной по исходным результатам наблюдений. Преобразования основаны на теореме, использующей симметричность матрицы коэффициентов корреляции и некоторых несложных операциях над матрицами, в результате которых получают, формула (2) и (3):

$$R\,[m \times m] = U'\,[m \times m] \cdot \Lambda\,[\mathrm{m} \times \mathrm{m}] \cdot \mathrm{U}\,[\mathrm{m} \times \mathrm{m}]$$
(2)

$$A\,[m \times m] = U'\,[m \times m] \cdot \Lambda^{1/2}[\mathrm{m} \times \mathrm{m}]$$
(3)

В последних уравнениях $\mathrm{U}\,[\mathrm{m} \times \mathrm{m}]$ - ортогональная матрица, столбцами которой являются собственные векторы матрицы $R\,[m \times m]$, а в $\Lambda\,[\mathrm{m} \times \mathrm{m}]$ диагональная матрица, составленная из собственных чисел матрицы $R\,[m \times m]$, соответствующих собственным векторам, причем элементы в матрице $\Lambda\,[\mathrm{m} \times \mathrm{m}]$ расположены в порядке убывания: $\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0$.

Таким образом, матрицу $A\,[m \times m]$ можно считать определенной, если известны собственные вектора и собственные числа матрицы $R\,[m \times m]$. Этот процесс легко осуществим на ЭВМ, так как существует много стандартных алгоритмов для вычисления собственных значений и собственных векторов симметричных матриц. Так решается основная проблема компонентного анализа - определение матрицы весовых коэффициентов, учитывающих тесноту связи между признаками и главными компонентами.

Модель компонентного анализа предполагает точное определение, как компонентных нагрузок, так и значений главных компонент для каждой точки опробования (для каждого объекта). На практике же оставляют обычно небольшое число компонент, если на их долю приходится достаточно большой процент суммарной дисперсии параметров, которая равна следующей матрицы $\Lambda\,[\mathrm{m} \times \mathrm{m}]$, или размерности матрицы $R\,[m \times m]$, то есть числу параметров. Следовательно, можно выбрать незначительное число главных компонент *(m < n)* и добиться значительного облегчения описания признаков. Поскольку $\sum_{i=1}^{m} \lambda_i = m$, обычно ${}^{q}/_{m}$ = 0,8÷0,9, то есть чаще всего ограничиваются 80-90% суммарной дисперсии, хотя эту величину можно устанавливать в зависимости от целей исследования и меньше этой величины, и больше.

После нахождения матрицы компонентных (иногда говорят о весовых) нагрузок при наличии ЭВМ не составит большого труда определить второе неизвестное - матрицу значений главных компонент в каждой точке опробования (для каждого объекта) - по уравнению записи модели. Довольно часто анализ ограничивают вычислением матрицы $A\,[m \times m]$, которая позволяет установить структуру новых (гипотетических) переменных - главных компонент, а иногда и интерпретировать их с профессиональной точки зрения.

В качестве главной компоненты выбирается ассоциация исходных признаков, вошедших в компоненту (или фактор) со статистически значимыми величинами нагрузок на соответствующие признаки. И хотя на первом этапе анализа (определение матрицы $A\ [m \times m]$) по выделенным главным компонентам можно сделать некоторые выводы об условиях, определяющих ход исследуемых процессов, матрица значений главных компонент $F\ [n \times m]$ представляет собой важный и полезный результат, особенно для картируемых величин. Первый этап компонентного анализа можно считать завершающим в том случае, когда его применяют для подтверждения выдвигаемых гипотез, то есть когда мы заранее предполагаем определенную иерархию признаков или их сочетаний, и выделенные ассоциации признаков (компоненты) подтверждают это предположение.

Далее в [1], отмечается, что если исследования компонентным анализом проводятся с целью сокращения числа параметров (при условии, что исходные переменные являются картируемыми величинами), то картирование значений одного фактора по точкам наблюдения будет равносильно изображению информации, представленной ранее на таком количестве карт, сколько исходных параметров входит в главную компоненту. Последнее замечание справедливо лишь в том случае, когда каждый из признаков входит только в одну компоненту, чего практически никогда не наблюдается при интерпретации главных компонент.

Главные компоненты отражают не простую сумму параметров, описывающих систему - они являются результатом системного взаимодействия этих параметров, тем новым свойством, которое появляется при построении системы. И процесс интерпретации главных компонент заключается в выявлении общих причин, вызывающих «параллельное» или «антипараллельное» изменение измеряемых параметров. Свертывание информации заключается в том, что число действующих на систему факторов всегда меньше числа параметров, являющихся их проявлением, примером чему может служить проявление климатического фактора через температурный режим и т.д.

Таким образом, второй этап компонентного анализа, или результат решения обратной компонентной задачи, представляет значительный интерес. Это решение находится по вычисленной предварительно матрице компонентных нагрузок и нормированной матрице исходных данных по следующему уравнению, формула (4):

$$F[m \times n] = A^{-1}\ [m \times m] \cdot Y\ [m \times n] \qquad (4)$$

Это уравнение справедливо лишь для случая, когда вычисляется полная матрица компонентных нагрузок $A\ [m \times m]$, и обратную матрицу можно найти. Если же ищутся только первые q компонент, удовлетворяющие заданной точности, то алгоритмы решения несколько усложняются, поскольку матрицу $A[m \times q]$ нужно вначале дополнить до квадратно, формула (5):

$$F[qm] = (A'[q \times m] \cdot A[m \times q]) \cdot A'[q \times m] \cdot Y[m \times n] \qquad (5)$$

Полученное решение - есть окончательный результат компонентного анализа в смысле математических построений, далее предстоит содержательная, предметная интерпретация выделенных компонент, которая, как отмечается во всех теоретических работах по компонентному анализу, представляет главную трудность в применении этого метода.

С точки зрения геометрического представления результатов компонентного анализа, его задачей является поиск такой ортогональной системы, которая при наименьшем числе координатных декартовых осей описывает наибольшее число исходных векторов-признаков. Положение новой системы координатных осей в компонентном анализе определяется таким образом, чтоб на каждую последующую ось (главную компоненту) приходился бы максимум оставшейся суммарной дисперсии после учета предыдущих осей.

Хотя математически это решение единственно, оно в действительности является одним из бесконечного множества других систем, так же хорошо описывающих конфигурацию исходных данных в пространстве объектов, как и система осей, выбранная по дисперсиям, то есть возникает проблема выбора оптимального положения осей – проблема вращения. Геометрическое вращение компонентных осей рассматривается как задача ортогонального вращения матрицы компонентных нагрузок. Однако всегда следует помнить, что процедура вращения осуществляется исключительно для облегчения процесса интерпретации выделенных компонент, которые и без вращения могут быть вполне интерпретируемыми.

Главным результатом компонентного анализа, с точки зрения практических приложений, можно считать матрицу собственных векторов матрицы парных коэффициентов корреляции исходных признаков, представляющей фактически ее внутреннюю структуру. Количество собственных векторов окажется меньше числа исходных параметров, это будет сигналом наличия линейной взаимосвязи между частью исходных признаков. Поскольку матрица компонентных нагрузок представляет собой матрицу собственных векторов, умноженную на диагональную матрицу из квадратных корней собственных чисел, вполне естественно перенести все рассуждения о свертывании пространства признаков на матрицу компонентных нагрузок. Это как раз и есть реализация гипотезы компонентного анализа о существовании внутренних свойств. Для геосистем такими свойствами являются факторы ее формирования, поэтому с полным основанием можно сказать, что компонентный анализ исследует, выявляет скрытые взаимосвязи исходными признаками геосистем. Полученные в результате данные можно рассматривать как формальную запись процессов через ту или иную ассоциацию параметров, составляющих компоненту. Эти ассоциацию выделяются по признаку совпадения знаков параметров, причем в одну компоненту могут входить две ассоциации, если они имеют противоположные знаки, однако и в этом случае компонента описывает один процесс. Противоположность знаков нагрузок на исходные переменные, выделенные моделью в одну компоненту, свидетельствует о разнонаправленности воздействия изучаемого процесса геосистем на ассоциации признаков с одинаковым знаком.

Вторым результатом компонентного анализа являются собственные числа матрицы коэффициентов парных корреляций. Данный результат позволяет судить о степени иерархичности процессов, выделяемых по матрице компонентных нагрузок.

Третий результат - матрица значений компонент показывает интенсивность проявления факторов на территории исследования. Значения главных компонент имеют положительные и отрицательные значения, однако, здесь различие в знаках означает не разную направленность процесса, а разную интенсивность его проявления – положительные значения характеризуют области с более интенсивным его проявлением, чем отрицательные. Положительные значение демонстрирует различную интенсивность его проявления к среднему значению.

Анализируя вышесказанное, можно сделать вывод, что компонентный анализ совмещает возможности статистического моделирования и системного анализа конкретных объектов (геосистем). Системный анализ при построении компонентного анализа позволяет выделять функции геосистемы и ранжировать их по вкладу в суммарную дисперсию системы, учитывает характер взаимосвязи и самоорганизацию системы. Основным результатом модели является не только выявление функции системы (системообразующих факторов) на основе интерпретации системы взаимосвязей исходных признаков, но и территории по интенсивности проявления этой функции. Таким образом, компонентный анализ позволяет выявлять взаимосвязи в системе, исследовать структуру взаимосвязей, ранжировать процессы по степени влияния на формирование состояния геосистемы.

Вопросы:

1. Матрица компонентных нагрузок?
2. «Параллельное» или «антипараллельное» изменение измеряемых параметров?
3. Матрица собственных векторов матрицы парных коэффициентов корреляции исходных признаков?

Литература:

1 Йёреског К.Г., Клован Д.И., Реймент Р.А. Геологический факторный анализ. – Л.: Недра, 1980. – 223 с.

2 Андерсон Т. Введение в многомерный статистический анализ. – М.: ГИФМЛ, 1963. – 500 с.

3 Павличенко Л.М. Многомерные статистические модели в геоэкологии. – Алматы: ProService LTD, 2007. – 173 с.

4 Айвазян С.А. Модельно- и методоориентированные интеллектуализированные программные комплексы по статистическому анализу данных // В кн.: Многомерный статистический анализ и вероятностное моделирование реальных процессов. – М.: Наука, 1990. – С. 6-30.

5 Павличенко Л.М. Системное моделирование природно-технических геосистем // В кн.: Новые подходы и методы в изучении природных и природно-хозяйственных систем. – Алматы: Қазақ университеті, 2000. – С. 132-135.

6 Павличенко Л.М. К технологии построения моделей прогноза изменений экогеосистем // Гидрометеорология и экология. – 2000. – №1. – С. 37-60.