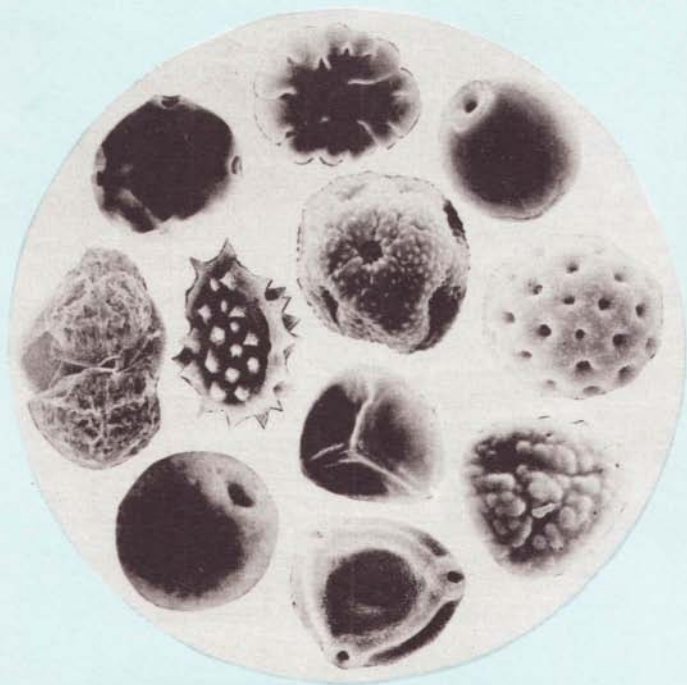


МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ПАЛИНОЛОГИИ



АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ
ИНСТИТУТ ГЕОЛОГИИ И ГЕОФИЗИКИ

МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ПАЛИНОЛОГИИ

Выпуск 2

КОЛИЧЕСТВЕННАЯ ОЦЕНКА
ЭЛЕМЕНТОВ ПАЛЕОКЛИМАТА
МЕТОДОМ ГЛАВНЫХ КОМПОНЕНТ
И ПОШАГОВЫМ РЕГРЕССИОННЫМ АНАЛИЗОМ
ПО ПАЛИНОЛОГИЧЕСКИМ ДАННЫМ

МЕТОДИЧЕСКИЕ РАЗРАБОТКИ

НОВОСИБИРСК 1989

УДК 551.583.7 + 561 + 519.272

Математические методы в палинологии. Вып.2. Количественная оценка элементов палеоклимата методом главных компонент и пошаговым регрессионным анализом (по палинологическим данным): Метод. разработ. / Г.Ф.Букреева. Новосибирск: Изд. ИГиГ, 1989. 98 с.

В работе рассмотрено применение математических методов в палинологических исследованиях. Предложены методики построения расчетных уравнений для количественной оценки элементов палеоклимата голоцена и плейстоцена по данным спорово-пыльцевому анализу методом главных компонент и пошаговым регрессионным анализом. Приведены алгоритмы программ математического обеспечения и примеры реализации.

Для палинологов и специалистов, интересующихся применением математических методов и ЭВМ для обработки многомерных данных.

Составитель

канд.геол.-мин. наук Г.Ф.Букреева

Отв.редактор д-р геол.-мин.наук А.Н.Дмитриев

Р е ц е н з е н т ы

д-р геол.-мин.наук В.С.Волкова (ИГиГ СО АН СССР),
канд.геол.-мин.наук В.В.Зуенко (Новосибирский ун-т)

Материал рассмотрен и рекомендован к публикации как методические разработки секцией стратиграфии, тектоники, литологии и осадочных полезных ископаемых Ученого совета Института геологии и геофизики СО АН СССР

© Институт геологии
и геофизики СО АН СССР,
1989

ВВЕДЕНИЕ

Данные палинологического анализа широко используются для реконструкции климата и растительности прошлого. Исходной предпосылкой в задаче количественной оценки элементов климата по результатам этого вида анализа служит предположение, что закономерности, существующие в настоящее время между климатом и растительностью, были такими же в те геологические эпохи, которые характеризовались развитием растительности современного типа. Следовательно, если смоделировать эти зависимости на современном материале, то они могут быть использованы, с известной долей вероятности, для получения количественных характеристик палеоклимата.

Соответствующую информацию о климате, при котором происходило захоронение ископаемых пыльцевых зерен, несут в себе качественный состав спорово-пыльцевого спектра и количественное содержание его компонентов. Выявление этой закономерности и есть решение задачи количественной оценки элементов палеоклимата. Этот этап называется этапом обучения. Здесь решается прямая задача – поставить в соответствие определенным климатическим условиям тип спорово-пыльцевого спектра. Следовательно, исходная информация должна включать в себя палинологические данные по поверхностным пробам и количественные оценки элементов современного климата.

На этапе распознавания решается обратная задача: среди рецентных спорово-пыльцевых спектров ищется аналог составу ископаемого спектра, и если таковой находится, то считается, что ископаемый спектр сформировался в сходной с современным спектром климатической обстановке. Этот же подход реализуется палинологами при качественной оценке палеоклиматов. Зачастую процентное соотношение между компонентами спорово-пыльцевого спектра считается адекватным соотношению между компонентами соответствующей растительности в природе. Но по разным причинам количественное соотношение растений в биоценозе не совпадает с количественным соотношением соответствующих пыльцевых зерен в спектре. Более объективную интерпретацию можно получить при математической обработке палинологического материала.

В последних публикациях /Методы ..., 1985; Веклич, 1987;

Ясаманов, 1985/ отмечается, что в настоящее время еще мало методик, использующих математические методы для восстановления количественных оценок элементов палеоклимата. Среди них следует отметить методики: 1) реализующую математическими методами ареалогический метод В.П.Гричука /Муратова и др., 1972/, 2) основанную на использовании метода информационного анализа /Климанов, 1976/, 3) использующую методы многомерного анализа /Гелета, Спиридонова, 1981/, 4) базирующуюся на выделении информативных систем компонентов спорово-пыльцевых спектров для построения регрессионных моделей с целью количественной оценки элементов палеоклимата /Букреева и др., 1984; 1986/.

Однако отсутствуют практически работы, в которых предлагаемые методики описываются с исчерпывающей полнотой, начиная от постановки задачи, алгоритма решения и кончая расчетным примером. Иными словами, нет работ, которыми может воспользоваться любой исследователь, заинтересованный в применении математических методов, не тратя времени и не повторяя уже проделанную до него работу. Опыт такого подхода реализован в монографии /Букреева и др., 1986/, где описаны все этапы применения методики, в том числе и опробование на конкретных геологических разрезах Западной Сибири. В этом же ключе выполнена и данная работа.

В обзорной работе В.Томпсона и Д.Р.Кларка /Thompson, Clark, 1977/ среди перечисленных методов многомерного анализа, применяемых для решения интересующей нас задачи, фигурируют методики, использующие метод главных компонент и пошаговый регрессионный анализ. В отечественной же литературе отсутствует подробное описание и применение перечисленных методик для количественных оценок элементов палеоклимата по палинологическим данным. Эти вопросы рассматриваются в данной публикации, которая может быть полезна еще и тем, что позволяет выявить положительные и отрицательные стороны разных методик. В работе сравниваются две указанные методики с методикой определения количественных оценок элементов палеоклимата по информативным системам компонентов спорово-пыльцевых спектров /Букреева и др., 1984; 1986/.

Для количественной оценки элементов палеоклимата в качестве объекта исследования выбраны палинологические данные, а метода решения – многомерный статистический анализ. Подтверждение правильности такого подхода находим в работе А.А.Величко /1985/.

В ней сравнительным анализом палеогеографических объектов (биогенных, биокостных и литогенных) и методов, используемых для реконструкции древних климатов, убедительно показано, что "наиболее богатую и точную палеоклиматическую информацию, особенно для теплых эпох, удастся извлечь из палеоботанических данных" /Величко, с.14/, а "успехи дальнейших исследований в области экспериментальной палеоклиматологии связаны с дальнейшим развитием методик количественных оценок, обогащением новым фактическим материалом" / Там же, с.19 /.

Одной из целей этой работы и является пополнение числа методик количественной оценки элементов палеоклимата по данным палинологического анализа, с разработкой и опробованием математического обеспечения для них на конкретном примере. Однако при этом следует отметить универсальность этих методик. Они с успехом могут быть использованы в тех же или других целях на любом современном и ископаемом фактическом материалах, структура которых отвечает требованиям, предъявляемым к исходным данным математическим аппаратом.

Работа построена с учетом возможности ее практического использования: формулируется постановка задачи, перечисляются требования к исходным данным, описываются алгоритмы методик, приводятся блок-схемы двух программ, входящих в пакет математического обеспечения обеих методик.

ПОСТАНОВКА ЗАДАЧИ И ТРЕБОВАНИЯ К ИСХОДНЫМ ДАННЫМ

Задача количественной оценки элементов палеоклимата по данным палинологического анализа в содержательном плане сводится к следующему. Имеется множество спорово-пыльцевых спектров (СПС) поверхностных проб изучаемого района, отражающих все многообразие растительных зон, представленных в нем, и данные наблюдений метеостанций по элементам климата этого же района в виде осредненных многолетних количественных показателей (среднемесячная температура, сумма годовых осадков и др.). Каждому рецентному спорово-пыльцевому спектру поставлены в соответствие те количественные показатели элементов современного климата, которые зафиксированы метеостанцией, ближайшей к месту отбора соответствующей поверхностной пробы.

Требуется найти закономерную зависимость – в виде линейной регрессионной модели – между количественной оценкой элемента климата и содержанием компонентов в рецентных спорово-пыльцевых спектрах.

Найденные зависимости можно использовать для расчета элементов палеоклимата голоцена и плейстоцена, т.е. геологического прошлого, которое, судя по составу ископаемых спорово-пыльцевых спектров, характеризуется растительностью, а следовательно, и климатом, сходным с современным.

Чтобы перейти к математической формулировке задачи, отметим, что здесь в качестве зависимого признака выступает элемент климата, а в качестве независимых признаков – компоненты СПС или их линейные комбинации (главные компоненты).

Математически задача количественной оценки элемента климата по данным спорово-пыльцевого анализа может быть сформулирована следующим образом.

Имеется n наблюдений, каждое из которых охарактеризовано одним и тем же набором m независимых признаков ($x_1, x_2, \dots, x_m$) и одним зависимым признаком x_{m+1} . Совокупность наблюдений представляет собой матрицу данных размерностью $X(n, m+1)$, которая служит обучающей выборкой.

Требуется определить параметры расчетной модели регресси-

онного вида для зависимой переменной x_{m+1} по m независимым переменным $(x_1, x_2, \dots, x_m)$:

$$x_{m+1} = a_0 + a_1x_1 + a_2x_2 + \dots + a_mx_m, \quad (I.I)$$

где $x_1, x_2, \dots, x_m$ — компоненты СПС или их линейные комбинации (главные компоненты), x_{m+1} — элемент климата.

Найденную модель можно использовать для определения неизвестного зависимого признака x_{m+1} по известным на этом объекте значениям m признаков.

К исходным данным, используемым при решении поставленной задачи, предъявляется ряд требований.

1. Исходная фактическая информация должна содержать: составы спорово-пыльцевых спектров поверхностных проб и соответствующую им количественную характеристику элементов современного климата.

2. Отбор поверхностных проб на спорово-пыльцевой анализ производится во всех растительных зонах, представленных в изучаемом регионе. Если предполагается, что в прошлом в изучаемый район происходила миграция растительности из соседних зон, то поверхностные пробы отбираются и в них. Следовательно, в обучающей выборке должны быть представлены спорово-пыльцевые спектры всех типов растительности, которые могли существовать на изучаемой территории в голоцене и плейстоцене. Иначе говоря, в обучающей выборке должны присутствовать все типы спорово-пыльцевых спектров, которые могут встретиться в ископаемом материале района исследования.

3. Охарактеризованность пробами каждой растительной зоны должна быть одинаковой, т.е. по каждой зоне отбирается приблизительно одно и то же число поверхностных проб. При этом оно должно быть достаточно представительным (30 и более). Иначе на построение расчетных моделей существенное влияние могут оказать данные по той растительной зоне, спорово-пыльцевые спектры которой имеют наибольший удельный вес в обучающей выборке.

4. Для достижения однородности обучающей выборки следует предусмотреть отбор проб с учетом фациальной характеристики отложений. Это связано с тем, что сохранность пыльцевых зерен сильно зависит от условий их захоронения. Количественная оценка элементов палеоклимата осуществляется на материале обучения по

отложениям одной какой-либо фации. Так, например, для реконструкции климата по ископаемым торфяникам предварительное обучение должно проводиться по спорово-пыльцевым спектрам поверхностных проб торфяников.

5. В обучающей выборке должны быть максимально представлены данные о современном климате изучаемого региона, т.е. по возможности со всех метеостанций, расположенных на изучаемой территории.

6. Каждой поверхностной пробе, взятой на спорово-пыльцевой анализ, а следовательно соответствующему спектру, должны быть поставлены в соответствие данные о современном климате, полученные на ближайшей к месту ее отбора метеостанции. Если это место равноудалено от нескольких метеостанций, то для этого спектра берутся усредненные данные.

7. В исследовании нужно использовать компоненты СПС, статистически обеспеченные участием в спектрах по всем или большинству представленных в выборке растительным зонам. Те компоненты СПС (даже если это такие индикаторы, как морошка, ольховник и др.), которые присутствуют лишь в отдельных спорово-пыльцевых спектрах, включать в процесс обучения не имеет смысла. Ведь для обработки палинологического материала применяются статистические методы, а они разработаны лишь для статистически обеспеченных данных. Однако находки единичных пыльцевых зерен-индикаторов можно с успехом использовать для подтверждения выводов о климатических условиях, которые будут получены по расчетным моделям с участием статистически обеспеченных компонентов СПС в обучающей выборке.

Понятно, что наличие только поверхностных проб при отсутствии множества данных со станций наблюдения за современным климатом исключает возможность использования предлагаемых методик, основанных на статистических методах обработки совместной исходной информации.

Г л а в а П

ПОСТРОЕНИЕ РЕГРЕССИОННЫХ МОДЕЛЕЙ КОЛИЧЕСТВЕННОЙ ОЦЕНКИ ЭЛЕМЕНТОВ ПАЛЕОКЛИМАТА ПОШАГОВЫМ РЕГРЕССИОННЫМ АНАЛИЗОМ

2.1. Общие сведения

Данная методика позволяет получать расчетную модель вида

$$x_0 = a_0 + a_1 x_1 + a_j x_j + \dots + a_k x_k, \quad (2.1)$$

где x_0 — элемент климата, для которого ищется зависимость, $x_1, x_j, \dots, x_k$ — компоненты спорово-пыльцевых спектров, выбранные из общего числа исследуемых и введенные в уравнение регрессии.

Суть метода пошагового регрессионного анализа состоит в том, что в уравнение регрессии включаются только те переменные, которые оказывают существенное влияние на зависимую переменную. Это влияние оценивается величиной частного коэффициента корреляции между зависимой и независимой переменными.

Поскольку в решаемой задаче в качестве зависимой переменной выступает элемент климата, а в качестве независимых — компоненты СПС, то частный коэффициент корреляции оценивает "чистую" связь между элементом климата и каждым из компонентов СПС при устранении влияния на эту связь остальных компонентов СПС. Следовательно, выбор компонентов СПС для построения регрессионной зависимости методом пошагового регрессионного анализа базируется на расчете и последующем анализе величин частных коэффициентов корреляции между исследуемым элементом климата и компонентами СПС по фактическому материалу. На каждом шаге в уравнение регрессии вводится тот компонент СПС, который составляет наибольшую величину дисперсии между ним и элементом климата, для которого ищется зависимость. Иными словами, вводится компонент СПС, имеющий наибольшую частную корреляцию с элементом климата.

На следующем шаге анализу подвергаются частные коэффициенты корреляции между элементом климата и оставшимися компонентами СПС, еще не введенными в уравнение. Выбирается компонент СПС, имеющий максимальное значение частного коэффициента корреляции в оставшемся множестве и т.д.

Коэффициент частной корреляции характеризует тесноту корре-

ляционной связи между двумя переменными при исключении влияния всех прочих рассматриваемых переменных (в отличие от коэффициента парной корреляции, который оценивает взаимосвязь между двумя величинами при наличии влияния всех прочих переменных).

Для определения частного коэффициента корреляции используется исходная корреляционная либо ковариационная матрица. Для простейшего случая, когда исследуются три переменные x_1, x_2, x_3 , определение частных коэффициентов с использованием парных коэффициентов осуществляется по следующим формулам: частный коэффициент корреляции r между двумя переменными x_1, x_2 при исключении влияния третьей переменной x_3

$$r_{12(3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}, \quad (2.2)$$

между двумя переменными x_1, x_3 при исключении переменной x_2

$$r_{13(2)} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}, \quad (2.3)$$

между переменными x_2, x_3 при исключении переменной x_1

$$r_{23(1)} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}, \quad (2.4)$$

Здесь первые два индекса при коэффициенте r указывают на переменные, связь между которыми измеряется; индекс в скобке — переменную, влияние которой исключается; r_{12}, r_{13}, r_{23} — коэффициенты парной корреляции соответственно между переменными x_1 и x_2 , x_1 и x_3 , x_2 и x_3 .

В нашей задаче, т.е. при изучении корреляционных связей между исследуемым элементом климата и m компонентами СПС, расчетную формулу для коэффициента частной корреляции между j -ым компонентом СПС и элементом климата при исключении влияния k -го компонента СПС можно записать в обобщенном виде:

$$r_{\varepsilon j(k)} = \frac{r_{\varepsilon j} - r_{\varepsilon k}r_{jk}}{\sqrt{(1-r_{\varepsilon k}^2)(1-r_{jk}^2)}}, \quad \begin{matrix} j = 1, 2, \dots, m; \\ k = 1, 2, \dots, m; \end{matrix} \quad (2.5)$$

где ε — элемент климата, для которого рассчитываются частные коэффициенты корреляции с каждым из m компонентов СПС; j — индекс компонента СПС, с которым определяется парная корреляция

исследуемого элемента климата; k - индекс компонента СПС, влияние которого на эту связь исключается.

В данной методической разработке использован алгоритм пошагового регрессионного анализа, где частные коэффициенты корреляции вычисляются по ковариационной матрице /Пакет ..., 1981/. Поскольку же в нашем случае решается задача оценки элементов палеоклимата по содержанию компонентов СПС, то и описание алгоритма приводится в терминах этой задачи: вместо термина "зависимая переменная" употреблен термин "элемент климата", а вместо "независимая переменная" - "компонент СПС".

2.2. Исходные данные

Для определения параметров уравнения регрессии необходимо сформировать обучающую выборку и количественно оценивать элемент климата в зависимости от процентного содержания компонентов СПС. В состав выборки входят данные спорово-пыльцевого анализа поверхностных проб и количественные показатели современного климата, для которых требуется построить расчетные модели. Размерность исходной таблицы получается равной $X(n, m+1)$, где n - число рецентных спорово-пыльцевых спектров; m - число всех компонентов СПС; 1 - число всех элементов климата, участвующих в исследовании. Каждая из n строк представляет собой вектор, первые m элементов которого содержат данные по компонентам СПС, а следующие 1 - количественные оценки элементов современного климата, соответствующие пробе. Структура исходной выборки, на которой выполняется пошаговый регрессионный анализ, показана в табл. I.

2.3. Алгоритм определения параметров регрессионных моделей пошаговым регрессионным анализом

Последовательность операций, реализующая построение регрессионных моделей пошаговым регрессионным анализом, отражена на блок-схеме (рис. I).

I. На вход программы STEP подаются исходные данные по ре-

Таблица I

Структура исходной таблицы для количественной оценки элементов климата по палинологическим данным методом пошагового регрессионного анализа

№ п/п	Компонент спорово-пыльцевого спектра, %							Элемент современного климата		
	Общий состав			Picea sp.	Pinus sp.	...	Brya- les	Средне- месячная температу- ра июля, °C	...	Средне- годовое количество осадков, мм
	Пыльца древесных пород и кустар- ников	Пыльца травянистых растений и кустар- ничков	Споры							
	I	2	3	4	5	...	m	I	...	1
I	18,5	70,5	11,0	86,0	4,0	...	30,6	17,8	...	242
2	9,1	89,5	1,3	71,0	23,8	...	1,3	18,8	...	119
...	...	...	...	...	...	...	...	...	...	...
n	53,5	44,7	1,8	98,9	0	...	25,6	16,9	...	386

Ввод: RAB - название объекта исследования;
 N - число строк в исходной таблице; M - число столбцов;
 NS - число моделей; PCT - величина, ограничивающая ввод
 переменной; PRI1 - признак печати исходной таблицы;
 PRI2 - признак печати корреляционной матрицы; D(N,M) - ис-
 ходная таблица

1. CORR (N,M,DAT,XBAR,STD,RX,R,B);

Вычисление средних значений XBAR(M), стандартных откло-
 нений STD(M), матрицы ковариаций RX(M,M), матрицы корреле-
 ций R(M,M).

DAT - подпрограмма читает данные (матрицу наблюдений) с
 внешнего устройства

R = RX
 (R ← матрица ковариаций)

IK = 1

RX = R
 (RX ← матрица ковариаций)

Ввод: IDX(M), SAG;

IDX(M) - вектор признаков: 0 - ввод переменной; 1 - обя-
 зательный ввод; 2 - переменную не вводить; 3 - зависимая
 переменная; SAG - задаваемое число шагов (максимальное
 число вводимых переменных)

2. STRG(M,N,RX,XBAR,IDX,PCT,NSTEP,ANS,LL,BB,STD,OUT);

Выполнение пошагового множественного линейного регресси-
 онного анализа для зависимой переменной и множества неза-
 висимых переменных.

OUT - подпрограмма печати результатов пошаговой регрессии

IK=IK+1

+

IK < NS

-

Конец

Рис.1. Блок-схема программы STEP "Пошаговый регрессионный анализ"

центным спорово-пыльцевым спектрам и элементам современного климата и необходимая сопутствующая информация.

2. По исходной таблице вычисляется матрица ковариаций. Элементы этой матрицы впоследствии используются для расчета частных коэффициентов корреляции между элементом климата и каждым компонентом СПС. На этом заканчивается подготовительный этап обработки исходной информации по палинологическим и климатическим данным.

Следующий этап – построение регрессионной модели – выполняется программой столько раз, сколько задано моделей, и контролируется переменной NS.

3. Формируется вектор состояния признаков, указывающий действие, производимое с переменной. Он состоит из $(m+1)$ ячеек, по числу столбцов в исходной таблице. Код, содержащийся в j -ом элементе вектора, указывает на действие с j -ой переменной (компонентом СПС или элементом климата): 0 – ввод переменной в уравнение, 1 – обязательный ввод, 2 – переменная не вводится, 3 – признак зависимой переменной.

Вводом вектора состояния признаков начинается цикл построения регрессионного уравнения для одного элемента климата, т.е. в одном векторе состояния можно указать только один элемент климата, для которого и будет найдена расчетная модель пошаговым вводом в нее тех компонентов СПС, которые имеют максимальные частные коэффициенты корреляции с заданным элементом климата.

Исходя из вышесказанного, в ячейку вектора, соответствующую исследуемому элементу климата, засылается цифра 3, в ячейки остальных элементов климата – 2, в ячейки компонентов СПС – 0 или 1.

4. С помощью переменной SAG указывается число шагов, равное числу компонентов СПС, вводимых в уравнение. Программа закончит построение уравнения, когда будет достигнут ввод в модель указанного числа компонентов СПС.

5. Вычисляются частные коэффициенты корреляции между элементом климата и каждым компонентом СПС по формуле

$$c_j = \frac{a_{jz}^2}{a_{jj}} \quad (2.6)$$

где a_{jj} – элемент матрицы ковариаций; $j = 1, 2, \dots$; q – компоненты СПС (индекс j не равен индексам компонентов СПС, уже вве-

денных в уравнение регрессии или выведенных из него); z — индекс элемента климата.

6. Из общего числа частных коэффициентов корреляции находится наибольшее значение c_j . Тот компонент СПС, для которого оно получено, вводится в уравнение, так как именно он оказывает самое существенное влияние на данный элемент климата.

7. Вычисляются параметры уравнения регрессии для элемента климата и тех компонентов СПС, которые на данном шаге оказались в числе выбранных для ввода в модель.

8. Производится статистическая оценка полученной зависимости путем определения коэффициента множественной корреляции, стандартной ошибки и т.д.

9. Компонент СПС, введенный в уравнение, исключается из числа компонентов, рассматриваемых на следующем шаге.

10. Если в уравнение не введено указанное число компонентов СПС, т.е. оно меньше количества, заданного переменной SAG , действие возвращается на ввод следующего компонента СПС, т.е. в пункт 5.

11. После выполнения заданного числа шагов — кроме информации о параметрах расчетной модели, выдаваемой на каждом промежуточном шаге, — распечатываются фактические значения исследуемого элемента климата и расчетные значения, полученные по найденной регрессионной модели, их абсолютные и относительные расхождения.

12. Проверяется число построенных регрессионных моделей, заданное переменной NS . Если оно меньше заданного, то управление программы передается на пункт 3, выполнение которого начинается с ввода нового вектора состояния, где помечен следующий элемент климата. Если построено заданное число моделей, программа заканчивает свою работу.

Реализация описанного алгоритма осуществляется программой $STEP$, составленной на языке ПЛ/I.

В качестве расчетной модели оценки элемента палеоклимата принимается и используется та регрессионная модель, которая удовлетворяет пользователя по совокупности свойств — количеству введенных компонентов СПС, величине множественного коэффициента корреляции, оценке стандартной ошибки и величине усредненной ошибки (в %). Зачастую можно значительно уменьшить число вводи-

мых в уравнение компонентов СПС при незначительном ухудшении статистических характеристик. Это позволяет повысить устойчивость расчетного уравнения регрессии.

Глава III

ПОСТРОЕНИЕ РЕГРЕССИОННЫХ МОДЕЛЕЙ КОЛИЧЕСТВЕННОЙ ОЦЕНКИ ЭЛЕМЕНТОВ ПАЛЕОКЛИМАТА МЕТОДОМ ГЛАВНЫХ КОМПОНЕНТ*

3.1. Общие сведения

Метод главных компонент (МГК) – это один из методов многомерного анализа. Начало его разработки относится к концу тридцатых годов /Обухов, 1938/. Он неоднократно освещался в литературе и подробно описан в ряде книг /Андерсон, 1963; Крамбейн и др., 1973; Йереског и др., 1980; Белонин и др., 1982; Елисеева, Рукавишников, 1982/. Поэтому раздел 3.1 данной работы в основном носит реферативный характер, но методически приспособлен для перехода к практической реализации метода в целях количественной оценки элементов палеоклимата, излагаемой в разделах 3.2 и 3.3.

Интерпретация результатов МГК дана в исследовании статистических закономерностей метеорологических /Багров, 1959/ и гидрологических полей /Скляренко, Смирнов, 1974/, процессов осадконакопления /Деч, Кноринг, 1985/, геохимических /Смирнов и др., 1988/ и палинологических /Thompson, Clark, 1977/ материалов.

Метод эффективен при разбиении исходного множества на подмножества объектов, применим также для выяснения возможной структуры совокупности объектов, характеризуемых множеством признаков. Он обеспечивает наилучшую, в среднеквадратическом смысле, аппроксимацию числового поля фиксированным числом ряда разложения, тем самым предоставляя возможность выбрать из большого числа переменных их ограниченный набор. Метод позволяет сконцентрировать основную информацию о состоянии поля в небольшом числе слагаемых – главных компонент, – которые и подвергаются дальнейшему анализу. Он помогает выявлять интерпретируемый смысл зависимостей между переменными. Преобразование результатов мно-

* В данном разделе употребляются два термина, сходные по звучанию, но разные по содержанию: "главная компонента" – математический термин и "компонент спорово-пыльцевого спектра" – палинологический термин.

гомерных наблюдений, производимое методом, устраняет возможную нелинейность зависимостей между переменными. Выявленные главные компоненты можно использовать как для описания, так и для классификации результатов наблюдений.

Краткая характеристика метода главных компонент состоит в следующем.

Имеется множество n объектов, каждый из которых описан набором из m признаков $X(x_1, x_2, \dots, x_m)$. Требуется разложить исходную совокупность на такие составляющие $U_1(x_1, x_2, \dots, x_m)$, которые полностью описывали бы ее. При этом необходимо, чтобы составляющие обладали свойством независимости и вклады их в общую дисперсию располагались в порядке уменьшения.

Полученная случайная составляющая U представляет собой линейную комбинацию m исходных признаков x_j и называется главной компонентой

$$U = \sum_{j=1}^m \alpha_j x_j, \quad (3.1)$$

где α_j — коэффициент при j -ом признаке.

Так, если в анализе участвуют две переменные (признака) x_1 и x_2 , то первая главная компонента рассчитывается по формуле:

$$U_1 = \alpha_1 x_1 + \alpha_2 x_2, \quad (3.2)$$

а вторая главная компонента по формуле:

$$U_2 = \beta_1 x_1 + \beta_2 x_2. \quad (3.3)$$

Первая главная компонента задает направление с максимальным вкладом в общую дисперсию объектов в пространстве всех признаков. Вторая главная компонента определяется как линейная функция тех же признаков, имеющая максимальную дисперсию при условии независимости от первой главной компоненты.

Вклад каждой следующей компоненты в общую дисперсию убывает. Вообще же число главных компонент равно числу исходных признаков, поскольку производимый при анализе поворот осей координат не изменяет евклидовых расстояний между объектами. Выявлено, что зачастую несколько первых главных компонент (k) заключают в себе максимум информации о результатах наблюдений, т.е. учитывают большую часть общей изменчивости. В дальнейшем анализе имеет смысл рассматривать первые главные компоненты, поэтому на практике число k меньше m .

В МГК оценивается доля вклада каждой компоненты в описание общей изменчивости и определяются коэффициенты корреляции между каждым признаком и главной компонентой, так называемые нагрузки на главные компоненты. Предположение, что главные компоненты с наибольшей энергией (дисперсией) обладают и большей информативностью, не всегда подтверждается. Возможны, как утверждают А.В.Александров и Н.Д.Горский /1983/, и другие ситуации. Поэтому следует различать термин "энергия" (дисперсия) и "информативность" главной компоненты. Авторы этой работы предлагают решение об отбрасывании принимать исходя из каких-либо дополнительных критериев, например, путем построения оценки функции плотности распределения для проекции выборки на каждую главную компоненту.

Таким образом, метод главных компонент служит для нахождения в пространстве признаков направлений (в виде линейных комбинаций исходных признаков), максимально учитывающих исходную дисперсию. Определение главных компонент исходной многомерной совокупности сводится к оценке коэффициентов главных компонент.

Главные компоненты определяются нормированными коэффициентами линейного уравнения. Это достигается при условии, что сумма квадратов коэффициентов главной компоненты должна равняться единице. Так, в случае двух переменных $m = 2$ для коэффициентов первой главной компоненты (3.2) условие запишется в виде

$$\alpha_1^2 + \alpha_2^2 = 1. \quad (3.4)$$

Это приводит еще и к тому, что коэффициенты главных компонент можно рассматривать как косинусы углов направления главной компоненты по отношению к исходным признакам.

Независимость главных компонент достигается при условии выполнения другого ограничения, накладываемого на коэффициенты: сумма произведений j -ых ($j = 1, 2, \dots, m$) коэффициентов двух главных компонент должна быть равна 0:

$$\alpha_1 \beta_1 + \alpha_2 \beta_2 = 0. \quad (3.5)$$

При выполнении этого условия векторы двух главных компонент будут ортогональны, т.е. в пространстве направлены под прямым углом друг к другу.

Налагаемые ограничения на коэффициенты главных компонент записываются для первой главной компоненты как

$$\sum_{j=1}^m \alpha_j^2 = 1, \quad (3.6)$$

для двух главных компонент как

$$\sum_{j=1}^m \alpha_j^2 = 1; \quad \sum_{j=1}^m \beta_j^2 = 1; \quad \sum_{j=1}^m \alpha_j \beta_j = 0, \quad (3.7)$$

где m — число исходных признаков.

Существует несколько способов определения коэффициентов главных компонент. Способ, реализуемый методами корреляционной теории, рассмотрен в работе /Крамбейн и др., 1973/. Сущность его состоит в том, что определение коэффициентов α_j главной компоненты U в выражении (3.1), осуществляется путем максимизации среднего квадрата корреляции T между главной компонентой U и исходными признаками x_j :

$$T = \frac{1}{m} \sum_{j=1}^m r^2(x_j, U), \quad (3.8)$$

где $r^2(x_j, U)$ — квадраты коэффициента корреляции между признаком x_j и набором случайных величин U . Линейная комбинация, которая максимизирует это выражение, является первой главной компонентой корреляционной матрицы /Rao, 1964/. Эта комбинация отвечает наибольшему собственному значению (первому корню) $\lambda_{\max}$ корреляционной матрицы, а ее коэффициенты есть не что иное, как элементы собственного вектора этого собственного значения $\lambda_{\max}$. Для положительно определенной корреляционной матрицы существует m действительных положительных корней λ (собственных значений). Каждому характеристическому корню λ соответствует характеристический (собственный) вектор корреляционной матрицы. Подбирая собственные значения и связанные с ними собственные векторы корреляционной матрицы, можно вычислить оценки параметров всех главных компонент.

Таким образом, коэффициенты главной компоненты суть элементы собственного вектора корреляционной матрицы, а связанное с ним собственное значение λ характеризует суммарный эффект изменчивости каждой главной компоненты. Согласно теории Андерсона /1963/, вклад в общую дисперсию каждой главной компоненты U может быть определен из отношения

$$U = \frac{\lambda_i}{\sum_{i=1}^m \lambda_i}, \quad (3.9)$$

Метод главных компонент иногда называют геометрическим методом распознавания образов. Геометрическая интерпретация метода состоит в том, что главные компоненты представляют в виде удобной системы координат. Её построение начинается с проведения прямой, для которой сумма квадратов проекций всех точек на эту прямую была наибольшей или, иначе, была максимальной дисперсия исходной совокупности полей. Эта ось, максимально описывающая исходную дисперсию, и есть первая главная компонента. Вторую главную компоненту находим поворотом на 90° системы координат. Соответствующая ось в свою очередь максимально описывает первый остаток дисперсии. Процедура нахождения последующей главной компоненты сводится к повороту системы координат и максимальному описанию остаточной дисперсии.

При решении задачи классификации методом главных компонент не требуется знание априорной информации о числе классов однородных объектов и признаков исследуемой совокупности. Эту информацию получают при помощи самого метода путем проекции объектов и признаков на плоскость первых главных компонент. Причем объективность такой информации зависит от величины обобщенной дисперсии, приходящейся на эти главные компоненты.

Классификация объектов носит название Q -анализа, а признаков — R -анализа /Миллер, Кан, 1965/. В первом случае выявляются группы однородных объектов, во втором — признаков.

Методом главных компонент классифицируются строки многомерной таблицы. Поэтому в зависимости от того, что требуется классифицировать — объекты или признаки, в качестве строк представляют объекты или признаки. За меру близости двух объектов (или признаков) принимается их коэффициент ковариации.

Q -анализ проводится на матрице ковариации объектов, рассчитанной на исходной таблице $X(n, m)$, в которой n объектов зафиксированы в виде строк, а m признаков в виде столбцов.

R -анализ выполняется на матрице ковариации признаков, полученной по таблице $X(m, n)$, представляющей собой транспонированную исходную матрицу, в которой поменяли местами столбцы и строки: теперь признаки — это строки, а столбцы — объекты. Здесь классифицируемыми объектами будут сами признаки.

мата по палинологическим данным используется до нескольких десятков компонентов СПС, поэтому вышеприведенные формулы (3.11), (3.12) получаются громоздкими и практически непригодны для ручного счета. Чтобы воспользоваться ими для расчета элементов палеоклимата, необходимо составить программу для ЭВМ и внести в нее (запомнить) все параметры этого уравнения. Выходом из этой ситуации является минимизация исходного признакового пространства без потери существенной информации. Такая возможность заложена в самом методе главных компонент.

Минимизация признакового пространства здесь осуществляется исходя из того, что корреляция признаков означает их избыточность, а сведение m избыточных признаков к l признакам, свободным от избыточности, означает снижение размерности исходного признакового пространства. Реализуется это выполнением R-анализа (выявлением групп сходных признаков). В результате признаки, дублирующие друг друга, т.е. линейно взаимосвязанные, окажутся скомпонованными в одну группу. Поэтому (без потери существенной информации) из каждой выделенной группы признаков для дальнейшего рассмотрения можно оставить по одному признаку. Сформированный таким образом набор окажется состоящим из независимых признаков, число которых значительно меньше исходного.

Минимизацию удобно проводить на графике проекций признаков на плоскость первой и второй главной компоненты, представляющих собой в этом случае линейные комбинации исходных объектов. В работе В.В.Александрова и Н.Д.Горского /1983/ перечислены случаи, когда можно отбросить признаки: а) если признаки линейно зависимы (тогда они изобразятся точками, лежащими на прямой, проходящей через начало координат, см. признаки 5 и 6 на рис.2), б) если группа признаков связана сильной корреляционной связью (тогда она будет представлена компактной группой точек, см. признаки I-4 на рис.2). В обоих случаях рекомендуется оставить по одному признаку из группы, например, для первого случая - признак 6, а для второго - признак 2.

В этой же работе сказано, как оценить важность признака. Она тем больше, чем дальше от начала координат расположен признак. Это объясняется тем, что вклад признака в обобщенную дисперсию пропорционален расстоянию до начала координат. Исходя из этого утверждения признаки 7 и 8 на рис.2 не представляют интереса и их можно отбросить.

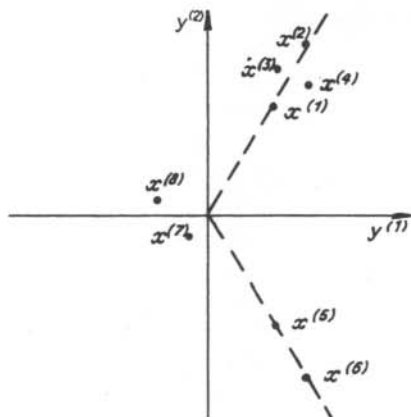


Рис.2. Анализ линейной зависимости признаков при помощи метода главных компонент /по В.В.Александрову, Н.Д.Горскому, 1983/

Кроме того, минимизацию исходного признакового пространства, очевидно, можно осуществлять исходя из сведений об относительной информативности того или иного исследуемого признака, которые отражены в величине коэффициента корреляции между главной компонентой и каждым признаком, т.е. в нагрузках на главные компоненты. Признаки, имеющие максимальную нагрузку, несут основную информацию об исходном материале, учитываемую главными компонентами. Поэтому первоначально заданное признаковое пространство можно сократить путем исключения из дальнейшего рассмотрения признаков с малыми нагрузками. В последующих вычислениях следует использовать только признаки с максимальными нагрузками на главные компоненты. Как показала практика, для выбора относительно информативных признаков достаточно воспользоваться анализом нагрузок на первые две главные компоненты, вычисленные на всем исходном признаковом пространстве.

Расчетная модель количественной оценки элементов палеоклимата по палинологическим данным, полученная методом главных компонент, после проведенной минимизации признаков запишется:

$$x_0 = a_0 + \sum_{i=1}^k a_i \sum_{j=1}^l q_{ij} x_j, \quad (3.13)$$

где x_0 - элемент климата; k - число главных компонент, введенных в уравнение; l - число компонентов СПС, оставленных для рассмотрения.

Итак, количественная оценка элемента климата по данным палинологического анализа методом главных компонент сводится:

1) к получению предварительной информации о структуре исходной матрицы спорово-пыльцевых спектров поверхностных проб (Q-анализ), 2) к минимизации исходного числа компонентов СПС (R-анализ), 3) к расчету параметров регрессионной модели на материале вновь созданной многомерной таблицы данных, включающей в себя набор главных компонент и вектор-столбец фактических данных по исследуемому элементу климата.

На первых двух этапах математической обработки применяется метод главных компонент, на третьем – метод множественного регрессионного анализа.

3.2. Исходные данные

При реализации методики количественной оценки элементов палеоклимата методом главных компонент сначала анализируется выборка, состоящая исключительно из рецентных спорово-пыльцевых спектров, и только после этого (на последнем этапе математической обработки) к анализу подключаются данные о современном климате.

Как отмечалось выше, для получения предварительных сведений о структуре исходной совокупности применяется Q-анализ. Поэтому исходной информацией на самом начальном этапе служит многомерная таблица палинологических данных $X(n, m)$, где n – число строк (по числу рецентных спорово-пыльцевых спектров в выборке), m – число компонентов СПС, описывающих каждый спектр. Строка таблицы представляет собой вектор, состоящий из m элементов и несущий информацию об одном спорово-пыльцевом спектре. Элементы вектора содержат в заданной последовательности процентное участие каждого из m компонентов в этом спектре. Отсутствие в спектре того или иного компонента отмечается занесением цифры 0 в соответствующий элемент вектора. Структура описываемой выборки приведена в табл.2.

На следующем этапе, когда классифицируются признаки в целях их минимизации (R-анализ), указанная исходная палинологическая матрица транспонируется и после этого подвергается обработке (табл.3).

Таблица 2

Структура исходной таблицы для количественной оценки
элементов климата по палинологическим данным
методом главных компонент

№ п/п	Компонент спорово-пыльцевого спектра, %						
	Общий состав			Picea Schren- kiana	Pinus sp.	...	Brya- les
	Пыльца древесных пород и кустар- ников	Пыльца травянистых растений и кустар- ничков	Споры				
	I	2	3	4	5	...	m
I	18,5	70,5	11,0	86,0	4,0	...	30,6
2	9,1	89,5	1,3	71,0	23,8	...	1,3
...	...	...	...	...	...	...	...
n	53,5	44,7	1,8	98,9	0	...	25,6

Исходной информацией на последнем этапе математической обработки служит вновь сформированная многомерная матрица $U(n, k+1)$, в которой k столбцов – это преобразованные исходные палинологические данные (главные компоненты) и только $(k+1)$ -й столбец содержит исходные фактические данные по тому климатическому показателю, для которого отыскивается расчетная модель. В каждую из n ячеек последнего $(k+1)$ -го столбца помещается соответствующее конкретному спорово-пыльцевому спектру количественное значение исследуемого элемента климата. Структура описанной матрицы приведена в табл.4.

3.3. Алгоритм определения параметров регрессионных моделей методами главных компонент и регрессионного анализа

Построение расчетной модели осуществляется в процессе трех основных этапов математической обработки (рис.3).

Первый этап – определение параметров главных компонент на

Таблица 3

Структура таблицы для выполнения R-анализа
методом главных компонент
(выявление однородных групп признаков)

№ п/п	Компонент спорово-пыльцевого спектра, %	Процентный спорово-пыльцевой спектр, № п/п					
		1	2	3	4	...	n
1	Пыльца древесных пород и кустар- ников	18,5	9,1	6,1	12,1	...	16,8
2	Пыльца травянис- тых растений и кустарничков	70,5	89,5	90,9	81,2	...	66,4
3	Споры	11,0	1,3	3,0	6,7	...	16,8
4	<i>Picea Schrenkiana</i>	86,0	71,0	80,0	88,9	...	53,4
5	<i>Pinus sp.</i>	4,0	23,4	20,0	5,6	...	0,0
...	...	...	...	...	...	...	...
n	<i>Bryales</i>	30,6	1,3	0,0	15,4	...	25,6

Таблица 4

Структура таблицы, формируемая для выполнения
множественного регрессионного анализа
на материале главных компонент
и количественных оценок элемента современного климата

№ п/п	Главная компонента				Количественная оценка элемента современного климата
	1-я	2-я	...	k-я	Температура июля, °C
	1	2	...	k	k+1
1	33,84	66,11	...	32,7	17,8
2	-10,74	11,13	...	-8,66	18,8
...	...	...	...	...	...
n	9,85	52,63	...	19,32	16,3

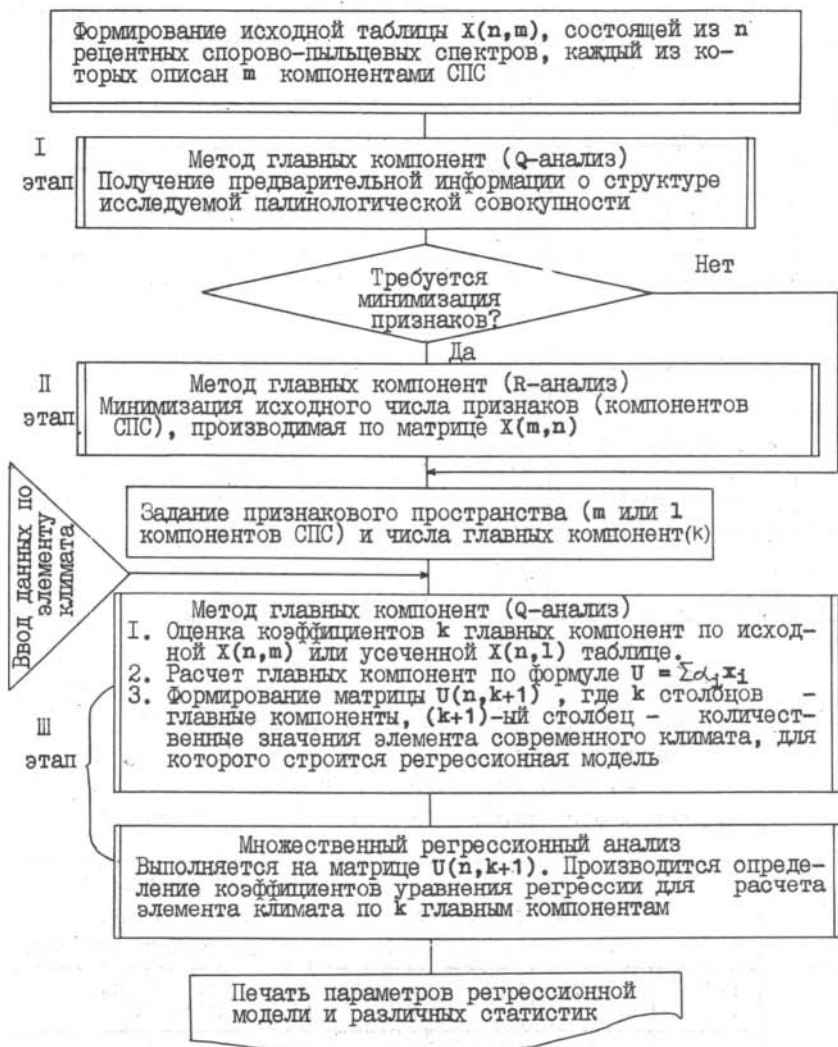


Рис.3. Блок-схема методики количественной оценки элемента палеоклимата по палинологическим данным методами главных компонент и регрессионного анализа

исходной таблице палинологических данных с использованием всего набора компонентов СПС (информация о климате пока в рассмотрении не участвует). Получаемые при этом результаты – выявление групп однородных рецентных спорово-пыльцевых спектров в исследуемой совокупности, оценка доли вклада каждой главной компоненты в описание общей изменчивости, определение нагрузок исходных признаков на первые главные компоненты – предварительно анализируются с целью установления оптимального числа главных компонент, вводимых в уравнение, и относительной информативности каждого компонента (доли вклада в первые две компоненты).

Второй этап – минимизация исходного признакового пространства (компонентов СПС). На транспонированной исходной матрице (см. табл.3) вычисляются две первые главные компоненты. На плоскость, образованную первой и второй главными компонентами, наносятся проекции исходных признаков. Из каждой группы однородных или линейно связанных между собой признаков (на графике проекции соответствующих признаков располагаются на прямой, проведенной из начала координат) оставляется по одному, а остальные отбрасываются. Наиболее информативными являются те компоненты СПС, которые максимально удалены от начала координат. В результате проведенного анализа формируется система информативных компонентов СПС, которая участвует в построении расчетного уравнения.

Третий этап – определение параметров расчетного регрессионного уравнения по заданному числу компонентов СПС и главных компонент. На этом этапе результаты, полученные методом главных компонент, используются как исходные данные для применения метода множественного регрессионного анализа, который и позволяет вывести зависимость, существующую между количественным выражением данного элемента климата и составом спорово-пыльцевого спектра.

Реализация описанной методики осуществляется программой MGKMPK. Функционально программа делится на две части. Первая реализует метод главных компонент, вторая – множественный регрессионный анализ.

Реализация метода главных компонент (I-я часть программы)

Параметры главных компонент определяются на материале корреляционной матрицы, получаемой по исходной палинологической таблице. Коэффициенты главных компонент вычисляются путем последовательного подбора собственных значений и связанных с ними собственных векторов корреляционной матрицы. Собственный вектор, соответствующий максимальному собственному значению матрицы, представляет собой набор коэффициентов для первой главной компоненты, а соответствующий второму наибольшему значению характеристического корня матрицы, есть набор коэффициентов для второй главной компоненты и т.д.

Вычисление параметров главных компонент осуществляется в следующем порядке.

1. По исходной таблице рецентных спорово-пыльцевых спектров $X(n,m)$ рассчитывается матрица корреляций (рис.4, блок I).

2. Вычисляются собственные значения λ_j и собственные векторы v_j действительной симметричной матрицы корреляций по методу диагонализации, улучшенному К.Г.Я.Якоби и переработанному Дж. Нейманом /Пакет ..., 1981/. Собственные значения располагаются по диагонали матрицы в убывающем порядке (см. рис.4, блок 2).

3. Находится число k собственных значений, которые больше или равны переменной CON . Это делается для того, чтобы сохранить наименьшее возможное число независимых переменных. В дальнейший анализ вводятся только те главные компоненты, собственные значения которых больше или равны CON .

При первом обращении к программе, т.е. на первом этапе обработки, рекомендуется задавать $CON = 0$, чтобы вычислить параметры всех главных компонент.

4. Определяются накопленные отношения собственных значений, представляющие собой оценку суммарной доли вклада главных компонент в описание исходной изменчивости (см. рис.4, блок 3):

$$d_j = \sum_{i=1}^k = \frac{\lambda_i}{m}, \quad j = 1, 2, \dots, k,$$

где k — число собственных значений, $k \leq m$.

5. Вычисление нагрузок главных компонент путем умножения

Ввод: RAB - название объекта исследования; N - число строк в исходной таблице; M - число столбцов (признаков); NS - число выборок (моделей); PRI1, PRI2, PRI3 - признаки печати соответственно: исходной таблицы, усеченной таблицы и корреляционной матрицы; WEK(M) - вектор состояния признаков; XIS(N,M) - исходная таблица

NS $\neq$ 0?
модель стро-
ить?

Нет

Да

Ввод: BLK(N) - целевой столбец (количественные значения признака для которого строится модель); K(NS) - вектор, содержащий информацию о числе главных компонент, вводимых в каждую из моделей

1. CORRO(N,M,X,XBAR,STD,RX,R,V);

Вычисление средних значений XBAR(M), стандартных отклонений STD(M), ковариаций RX(M,M) и корреляций R(M,M)

2. MSDU(R,V,M,MV);

Вычисление собственных значений R(M,M) и собственных векторов симметричной матрицы V(M,M)

3. TRAC(M,R,CON,K2,DX);

Вычисление DX(K2) накопленных отношений собственных значений $\geq$ CON

4. LOAD(M,K2,R,V);

Вычисление матрицы факторных нагрузок по собственным значениям и собственным векторам. V(M,M) - заданная матрица собственных векторов, расположенных по столбцам, и вычисляемая матрица факторных нагрузок V(M,K2)

А

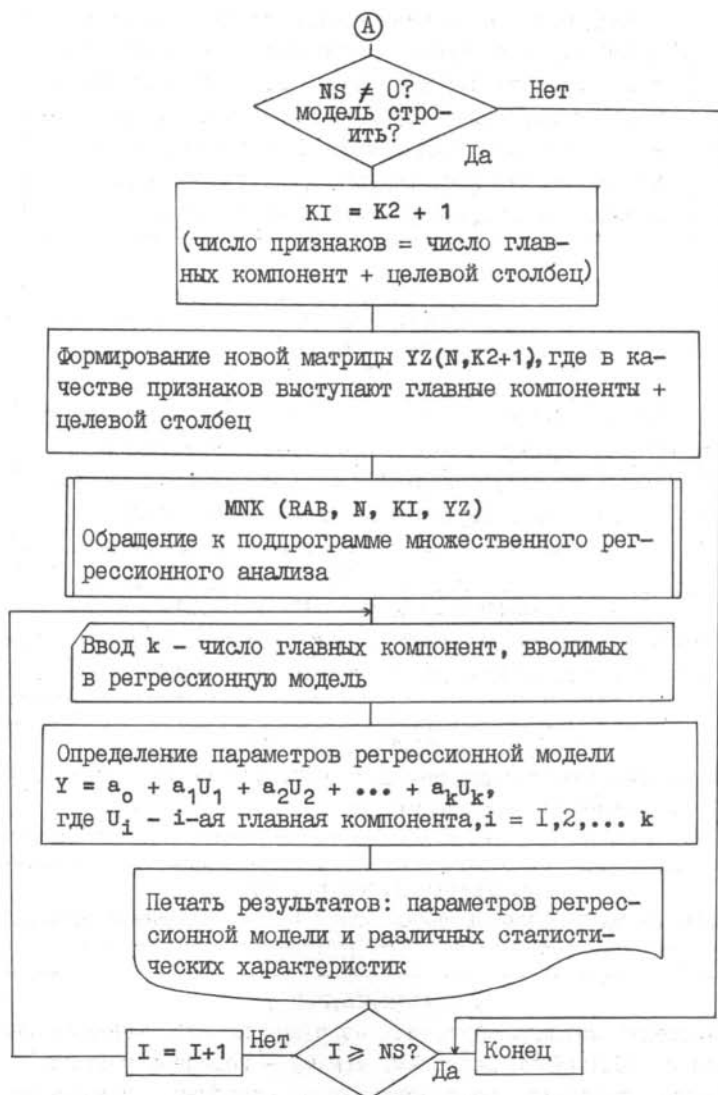


Рис.4. Блок-схема программы MGKMNK "Построение модели методами главных компонент и регрессионного анализа"

элементов каждого нормализованного собственного вектора на квадратные корни соответствующих собственных значений (см. рис.4, блок 4):

$$r_{ij} = v_{ij} \sqrt{\lambda_j}, \quad (3.14)$$

где $i = 1, 2, \dots, m$ — индексы переменных (компонентов СПС), $j = 1, 2, \dots, k$ — оставленные для рассмотрения собственные значения, $k \leq m$.

6. Расчет главных компонент. По определению главная компонента представляет собой линейную комбинацию исходных признаков. Набор коэффициентов главной компоненты есть соответствующий собственный вектор (см. выше пункт 2). Главная компонента рассчитывается по формуле:

$$i = 1, 2, \dots, k;$$

$$U_i = \gamma_{i1}x_1 + \gamma_{i2}x_2 + \dots + \gamma_{ij}x_j; \quad j = 1, 2, \dots, m; \quad (3.15)$$

где U_i — i -я главная компонента; $\gamma_{i1}, \gamma_{i2}, \dots, \gamma_{ij}$ — коэффициенты i -ой главной компоненты; $x_1, x_2, \dots, x_j$ — компоненты СПС.

7. Формируется матрица $U(n, k)$ новых независимых переменных, т.е. главных компонент. Завершена первая часть программы.

Реализация метода множественного регрессионного анализа (2-я часть программы MGKMNK)

8. Вводится вектор-столбец данных по исследуемому элементу климата.

9. Матрица главных компонент $U(n, k)$ (см. пункт 7) достраивается вектором-столбцом фактических значений по элементу климата. Полученная многомерная таблица $U(n, k+1)$ служит материалом для выполнения множественного регрессионного анализа.

10. Задается общее число моделей (переменной NS), которое требуется построить для исследуемого элемента климата.

11. Указывается число главных компонент, вводимых в данную регрессионную модель.

12. Выполняется множественный регрессионный анализ. Результатом его является определение параметров регрессии, связывающей между собой элемент климата с заданным числом главных компонент:

$$\text{ЭК} = a_0 + a_1U_1 + a_2U_1 + a_2U_2 + \dots + a_kU_k, \quad (3.16)$$

где ЭК - исследуемый элемент климата; $a_1, a_2, \dots, a_k$ - коэффициенты при главных компонентах; $u_1, u_2, \dots, u_k$ - главные компоненты (линейные комбинации исходных признаков).

13. Статистическая оценка полученной зависимости путем расчета коэффициента множественной корреляции, стандартной ошибки и др.

14. Производится контроль числа построенных моделей. Если их меньше, чем указано переменной NS, то действие программы передается на пункт II, иначе она завершает свою работу.

Регрессионные уравнения, позволяющие получать хорошую сходимость с фактическими климатическими данными, в дальнейшем могут быть использованы для количественной оценки элементов климата по составу ископаемых спорово-пыльцевых спектров.

Г л а в а I V

ПРИМЕРЫ ПОСТРОЕНИЯ УРАВНЕНИЯ РЕГРЕССИИ ДЛЯ КОЛИЧЕСТВЕННОЙ ОЦЕНКИ ЭЛЕМЕНТОВ ПАЛЕОКЛИМАТА ПО ПАЛИНОЛОГИЧЕСКИМ ДАННЫМ

Применение пошагового регрессионного анализа и метода главных компонент для определения параметров расчетного уравнения, описывающего зависимость количественной оценки элемента климата от состава спорово-пыльцевого спектра, иллюстрируется на одном и том же исходном материале – результатах палинологического анализа поверхностных проб и количественных показателей современного климата территории Иссык-Кульской котловины. Исходный фактический материал был любезно предоставлен Г.Н.Бердовской (Институт озероведения АН СССР). Обучающая выборка состояла из 30 спорово-пыльцевых спектров поверхностных проб, каждый из которых был охарактеризован 28 компонентами СПС (списочный состав приведен в распечатках 1, 2 и на рис.8). Пробы отбирались в зоне распространения лугово-степной растительности (7 проб), сухих ступей (10 проб) и полупустынь (8 проб), т.е. в исходной выборке представлены рецентные спорово-пыльцевые спектры основных типов растительности, распространенной в районе Иссык-Кульской котловины. Принадлежность остальных пяти спектров к одному из перечисленных классов однозначно не определена. Таким образом, исходная совокупность палинологических данных представлена таблицей размерностью $x(30,28)$.

В качестве элемента климата, для которого требуется найти расчетную регрессионную модель, выбрана среднемесячная температура июля. Поэтому в состав исходного фактического материала, кроме палинологического, обязательно входят метеоданные о среднемесячных температурах июля. Каждой поверхностной пробе, а значит и рецентному спорово-пыльцевому спектру, ставится в соответствие количественное значение этого элемента климата. Исходные температурные данные представляют собой вектор-столбец размерностью в 30 элементов (по числу проб). При этом использованы данные наблюдений шести метеостанций, расположенных в Иссык-Кульской котловине.

Решаемая задача состоит в поиске зависимости между среднемесячной температурой июля и составом спорово-пыльцевых спектров

поверхностных проб разных растительных зон Иссык-Кульской котловины.

Допускается, что модель, построенная на современном фактическом материале, в дальнейшем может быть применена для восстановления палеотемператур июля по составу ископаемых спорово-пыльцевых спектров из отложений плейстоцен-голоценового возраста Иссык-Кульской котловины.

Ниже описывается последовательность определения параметров регрессионной модели для оценки температуры июля методами пошагового регрессионного анализа и главных компонент. Для сравнения результатов дополнительно рассматривается модель, построенная согласно методике количественной оценки элемента палеоклимата по информативным системам компонентов спорово-пыльцевых спектров /Букреева и др., 1984, 1986/.

4.1. Определение параметров расчетной модели для оценки температуры июля пошаговым регрессионным анализом

Построение моделей пошаговым регрессионным анализом осуществляется программой STEP. За одно выполнение можно определить параметры уравнений для нескольких элементов климата.

На вход программы подается исходная матрица данных, которая одновременно содержит информацию о рецентных спорово-пыльцевых спектрах и элементах современного климата (см. табл.1). При этом необходимо соблюдать следующее условие: число наблюдений (спорово-пыльцевых спектров) должно быть больше, чем число независимых переменных (компонентов СПС и элементов климата) плюс 1. Так, в случае оценки расчетного уравнения для определения июльских температур по составу спорово-пыльцевых спектров поверхностных проб Иссык-Кульской котловины исходная матрица имела размерность $x(30;28+1)$. В ней соотношение числа строк и столбцов не отвечает требованию, так как при наличии 30 наблюдений минимальное число независимых переменных должно быть 28, а в нашем случае - 29. Поэтому приходится выводить из рассмотрения один из компонентов СПС, а именно пыльцу неопределенной травянистой растительности. Теперь остается 27 компонентов СПС. Размерность исходной таблицы становится $x(30;27+1)$.

Кроме этого, программе нужно сообщить, сколько моделей следует построить. По числу заданных моделей должно быть подготовлено и число векторов состояния признаков. Вектор несет информацию о зависимом признаке и множестве независимых. Иначе говоря, в этом векторе отмечается тот элемент климата, для которого строится модель, и те компоненты СПС, из числа которых могут выбираться аргументы в эту модель.

Составим пакет в соответствии с требуемой последовательностью ввода данных (см. рис.1).

Пакет 1

ИССЫК-КУЛЬСКАЯ КОТЛОВИНА. ПОВЕРХНОСТНЫЕ ПРОБЫ

```

30  28      I      0 0 0
18.5 70.5 II 86 4 0 6 4 0 31.4 5.2 0 I.I 0 26.2 ... 22 17.8
...
16.8 66.4 I6.8 53.4 0 0 20 26.6 0 0 I6.9 I.7 0 0 ... 0 I6.3
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3
4

```

Данная последовательность цифр передаст в программу, что:

- объект исследования ИССЫК-КУЛЬСКАЯ КОТЛОВИНА, ПОВЕРХНОСТНЫЕ ПРОБЫ;

- в исходной таблице 30 строк и 28 столбцов;
- требуется построить одно регрессионное уравнение ($NS = I$);
- величина, ограничивающая ввод переменных (компонентов СПС), равна 0, т.е. ввод не ограничен ($PST = 0$);
- не нужно печатать исходную таблицу данных ($PR11 = 0$) и матрицу коэффициентов корреляций ($PR12 = 0$);
- исходная таблица $X(30;27+I)$ вводится построчно. Каждая строка содержит данные о процентном содержании в спектре 27 компонентов СПС и температуре июля;
- вектор состояния признаков (во всех 27-ми позициях, принадлежащих компонентам СПС, содержит цифру 0, указывающую, что данные признаки суть независимые переменные и из них следует выбирать компоненты СПС для ввода в модель). В последней 28-ой позиции вектора находится цифра 3, этот код свидетельствует, что соответствующая переменная (а в нашем случае в 28-ом столбце нахо-

дятся данные о температуре июля) есть зависимая переменная и именно для нее следует строить модель;
 - число шагов равно 4, т.е. в модель будет введено 4 компонента СПС.

Распечатка I

Результаты определения параметров регрессии
 для количественной оценки температуры июля,
 полученные пошаговым регрессионным анализом

ИССЫК-КУЛЬСКАЯ КОТЛОВИНА. ПОВЕРХНОСТНЫЕ ПРОБЫ

N - ЧИСЛО НАБЛЮДЕНИЙ 30
 M - ЧИСЛО ПЕРЕМЕННЫХ 28
 NS - КОЛИЧЕСТВО ВЫБОРОК I
 PCT - ВЕЛИЧИНА, ОГРАНИЧИВАЮЩАЯ ВВОД ПЕРЕМЕННОЙ В УРАВНЕНИЕ, 0
 PRI1 = 0 PRI2 = 0

D:

*** CORR ***

ERROR = 0

ПЕРЕМЕННАЯ	СРЕДНЕЕ ЗНАЧЕНИЕ	СТАНДАРТНОЕ ОТКЛОНЕНИЕ	Пыльца
I)	14.177	13.857	← древесных пород
2)	72.390	16.149	← травянистых
3)	13.390	12.410	← споры
4)	54.067	41.675	← ель Шренка
5)	5.847	10.604	← сосна
6)	0.980	2.884	← береза
7)	2.727	6.100	← облепиха
8)	5.947	18.969	← селитрянка
9)	0.410	1.659	← неопр. кустарн.
10)	9.253	9.557	← эфедра
11)	8.137	6.924	← злаковые
12)	0.140	0.438	← злаковые (культурн.)
13)	1.227	3.127	← осоковые
14)	0.513	1.208	← гречишные
15)	31.263	13.681	← маревые
16)	0.790	1.620	← гвоздичные

17)	0.333	1.061	←	лютиковые
18)	0.540	1.254	←	крестоцветные
19)	0.967	1.574	←	розоцветные
20)	0.333	0.758	←	бобовые
21)	0.153	0.514	←	мезофиты
22)	0.247	0.600	←	зонтичные
23)	0.080	0.292	←	Dipsacaceae
24)	1.763	1.825	←	губоцветные
25)	13.233	9.479	←	сложноцветные
26)	0.340	0.890	←	Centaurea sp.
27)	29.080	12.760	←	полыни
28)	17.773	0.764	←	Т ⁰ июля

ПОШАГОВЫЙ МНОЖЕСТВЕННЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

ВЫБОРКА I

SAG - КОЛИЧЕСТВО ЗАДАНЫХ ШАГОВ 4

ВЕКТОР idx(M):00000000000000000000000000000003

*** STRG ***

ЗАВИСИМАЯ ПЕРЕМЕННАЯ 28

КОЛИЧЕСТВО ПЕРЕМЕННЫХ, УЧАСТВУЮЩИХ В ПОСТРОЕНИИ РЕГРЕССИИ, 27
КОЛИЧЕСТВО ПЕРЕМЕННЫХ, НЕ ВКЛЮЧАЕМЫХ В УРАВНЕНИЕ РЕГРЕССИИ, 0
ШАГ I

ВВОДИМАЯ ПЕРЕМЕННАЯ 4 ← пыльца ели Шренка

ОБЩЕЕ ЧИСЛО ВВЕДЕННЫХ ПЕРЕМЕННЫХ I

ПЕРЕМЕННАЯ ЧАСТНЫЙ СТАНДАРТНОЕ

КОЭФФИЦИЕНТ ОТКЛОНЕНИЕ

РЕГРЕССИИ

4 -0.00962 0.00295 ← пыльца ели Шренка

СВОБОДНЫЙ ЧЛЕН

18.29362

КОЭФФИЦИЕНТ МНОЖЕСТВЕННОЙ КОРРЕЛЯЦИИ 0.52475

ОЦЕНКА СТАНДАРТНОЙ ОШИБКИ 0.66209

ОШИБКА УСРЕДНЕННАЯ,% 3.02

ШАГ 2

ВВОДИМАЯ ПЕРЕМЕННАЯ I2 ← пыльца злаковых

ОБЩЕЕ ЧИСЛО ВВЕДЕННЫХ ПЕРЕМЕННЫХ 2 гр.Cerealia

ПЕРЕМЕННАЯ ЧАСТНЫЙ СТАНДАРТНОЕ

КОЭФФИЦИЕНТ ОТКЛОНЕНИЕ

РЕГРЕССИИ

4	-0.01263	0.00216	← пыльца ели Шренка
12	-1.10929	0.20526	← злаковые гр. Cerealia

СВОБОДНЫЙ ЧЛЕН

18.61121

КОЭФФИЦИЕНТ МНОЖЕСТВЕННОЙ КОРРЕЛЯЦИИ 0.80740

ОЦЕНКА СТАНДАРТНОЙ ОШИБКИ 0.46731

ОШИБКА УСРЕДНЕННАЯ, % 1.99

ШАГ 3

ВВОДИМАЯ ПЕРЕМЕННАЯ 3 ← содержание спор

ОБЩЕЕ ЧИСЛО ВВЕДЕННЫХ ПЕРЕМЕННЫХ 3

ПЕРЕМЕННАЯ	ЧАСТНЫЙ КОЭФФИЦИЕНТ РЕГРЕССИИ	СТАНДАРТНОЕ ОТКЛОНЕНИЕ
------------	-------------------------------------	---------------------------

4	-0.01375	0.00196	← пыльца ели Шренка
---	----------	---------	---------------------

12	-1.21305	0.18648	← злаковые гр.
----	----------	---------	----------------

3	-0.01822	0.00643	← споры
---	----------	---------	---------

СВОБОДНЫЙ ЧЛЕН

18.93072

КОЭФФИЦИЕНТ МНОЖЕСТВЕННОЙ КОРРЕЛЯЦИИ 0.85674

ОЦЕНКА СТАНДАРТНОЙ ОШИБКИ 0.41628

ОШИБКА УСРЕДНЕННАЯ, % 1.74

ШАГ 4

ВВОДИМАЯ ПЕРЕМЕННАЯ 13 ← пыльца осоковых

ОБЩЕЕ ЧИСЛО ВВЕДЕННЫХ ПЕРЕМЕННЫХ 4

ПЕРЕМЕННАЯ	ЧАСТНЫЙ КОЭФФИЦИЕНТ РЕГРЕССИИ	СТАНДАРТНОЕ ОТКЛОНЕНИЕ
------------	-------------------------------------	---------------------------

4	-0.01196	0.00181	← пыльца ели Шренка
---	----------	---------	---------------------

12	-1.13668	0.16463	← злаковые гр. Cerealia
----	----------	---------	-------------------------

3	-0.02937	0.00671	← споры
---	----------	---------	---------

13	0.08507	0.02811	← осоковые
----	---------	---------	------------

СВОБОДНЫЙ ЧЛЕН

18.86811

КОЭФФИЦИЕНТ МНОЖЕСТВЕННОЙ КОРРЕЛЯЦИИ 0.89741

ОЦЕНКА СТАНДАРТНОЙ ОШИБКИ 0.36317

ТАБЛИЦА ФАКТИЧЕСКИХ И ВЫЧИСЛЕННЫХ ЗНАЧЕНИЙ ЗАВИСИМОЙ ПЕРЕМЕННОЙ

НОМЕР ПРОБЫ	ФАКТИЧ.	ВЫЧИСЛ.	РАЗНОСТЬ	ОШИБКА, %
I	I7.79999	I7.60987	0.19012	I.07
2	I7.79999	I7.98060	-0.18061	I.01
3	I7.79999	I7.82301	-0.02303	0.13
4	I7.79999	I7.67594	0.12404	0.70
5	I7.79999	I7.49022	0.30977	I.74
6	I7.79999	I7.62923	0.17076	0.96
7	I7.79999	I7.63217	0.16782	0.94
8	I7.79999	I7.73334	0.06665	0.37
9	I7.79999	I7.70607	0.09392	0.53
10	I7.79999	I7.90348	-0.10349	0.58
11	I7.79999	I7.37212	0.42787	2.40
12	I7.79999	I7.90151	-0.10152	0.57
13	I7.00000	I7.16309	-0.16309	0.96
14	I7.00000	I7.59927	-0.59927	3.53
15	I8.79999	I8.33949	0.46050	2.45
16	I8.79999	I8.52407	0.27591	I.47
17	I8.79999	I8.68016	0.11983	0.64
18	I8.79999	I8.66254	0.13745	0.73
19	I8.79999	I8.86811	-0.06812	0.36
20	I8.79999	I9.08387	-0.28389	I.51
21	I8.79999	I8.27429	0.52570	2.80
22	I8.79999	I8.86811	-0.06812	0.36
23	I7.29999	I7.44778	-0.14779	0.85
24	I7.29999	I7.16628	0.13370	0.77
25	I7.29999	I6.95028	0.34971	2.02
26	I6.89999	I7.48705	-0.58705	3.47
27	I6.89999	I7.39252	-0.49253	2.91
28	I6.89999	I7.43082	-0.53083	3.14
29	I6.29999	I7.00080	-0.70081	4.30
30	I6.29999	I5.80357	0.49641	3.05
ОШИБКА УСРЕДНЕННАЯ, %				I.54

Анализ распечатки показывает, что на первом шаге в расчетное уравнение для температуры июля программа выбрала пыльцу ели Шренка, определившую 52 % исходной изменчивости фактических температурных данных, на втором - пыльцу злаков гр.Cerealia, при

Таблица 5

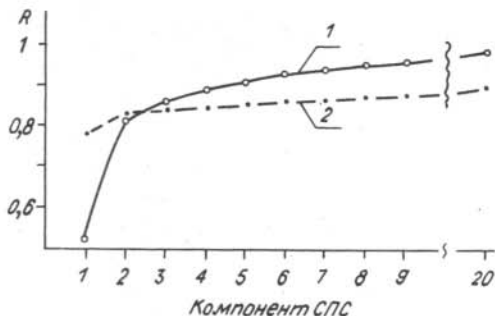
Зависимость множественного коэффициента корреляции
и стандартной ошибки
от числа введенных в уравнение компонентов СПС

Шаг	Вводимая переменная	Коэффициенты при переменных на соответствующем шаге					
1	<i>Picea Schrenkiana</i>	-0,01	-0,013	-0,01	-0,01	-0,01	-0,01
2	<i>Gramineae</i> <i>gr.Cerealia</i>		-1,11	-1,21	-1,14	-1,18	-1,18
3	Количество спор в СПС, %			-0,02	-0,03	-0,04	-0,04
4	<i>Cyperaceae</i>				0,09	0,1	0,11
5	<i>Rosaceae</i>					0,1	0,1
6	<i>Nitraria sibirica</i>						0,01
Коэффициент множественной корреляции		0,52	0,81	0,86	0,89	0,91	0,94
Стандартная ошибка, °C		0,66	0,46	0,42	0,36	0,34	0,32
Ошибка усредненная, %		3,0	2,0	1,7	1,5	1,5	1,3

этом суммарное описание дисперсии составило 81 %. На третьем шаге в уравнение вошел такой показатель спектра, как общее содержание спор. Его ввод повысил описание изменчивости лишь на 5 %, что в сумме составило 86 %. Четвертой по значимости частного коэффициента корреляции с температурой июля оказалась пыльца осокровых. Она добавила всего 3 % в учет общей изменчивости и совместно с остальными это составило 89 %. Если бы мы продолжили вводить и дальше в модель компоненты СПС, то лишь незначительно улучшили бы этот показатель (табл.5, рис.5). Этим объясняется тот факт, что мы решили ограничиться вводом всего четырех компонентов СПС в расчетное уравнение температуры июля. Пошаговым регрессионным анализом была построена следующая модель (см. распечатку I):

$T_{\text{июля}}^0 = 18,87 - 0,03SP - 0,01P_{sch} - 1,14GrC + 0,09C_{cyr}$. (4.1)
Для нее коэффициент множественной корреляции $R = 0,9$, оценка стандартной ошибки $\sigma = 0,36$ °C, а расхождение между фактическими

Рис.5. Графики изменения величины коэффициента множественной корреляции для модели $T_{\text{июля}} = f(\text{СПС})$ от ввода в уравнение нового компонента СПС. Данные по Иссык-Кульской (1) и Минусинской (2) котловинам



и расчетными температурными данными в среднем составило 1,54 %.

В ранее опубликованных работах /Букреева и др., 1964, 1986; Букреева, 1987, 1988/ выделялась в качестве главной при решении задачи количественной оценки элементов палеоклимата по данным палинологического анализа проблема выбора оптимального набора эффективных компонентов СПС, вводимых в расчетное уравнение. Указывалось, что возрастание числа компонентов СПС приводит к смешению причинно-следственных связей. Чтобы уменьшить вероятность такой ошибки, как принятие ложной зависимости за истинную, необходимо представить начальную информацию в виде комбинации значений сравнительно небольшого числа компонентов СПС. Поэтому требование достаточной полноты системы компонентов СПС неотделимо от требования минимизации их начального числа.

Воспользуемся тем свойством пошагового регрессионного анализа, которое позволяет оценивать на каждом шаге, насколько учитывается общая изменчивость от ввода очередного компонента СПС в уравнение. Результаты такого анализа лишний раз свидетельствуют о том, что основная дисперсия описывается уже при введении первых 4–5 компонентов СПС. Ввод следующих изменяет ее столь незначительно, что ими можно пренебречь (см. рис.5). Аналогичный результат получен на материалах Минусинской котловины. Следовательно, ввод в уравнение более 10 компонентов СПС может только нарушить устойчивость найденной закономерности. Материалы, позволившие прийти к такому выводу, приведены (частично) в табл.5 и графически представлены на рис.5. Они подтверждают полученные ранее В.Томпсоном и Д.Р.Кларком /Thompson, Clark, 1977/ выводы о

зависимости общей изменчивости информации от числа компонентов СПС, вводимых в прогнозную регрессионную модель.

4.2. Определение параметров расчетной модели для оценки температуры июля методами главных компонент и регрессионного анализа

Решение поставленной задачи методом главных компонент осуществляется, как отмечалось, на материале тех же палинологических и температурных данных, что и ее решение методом пошагового регрессионного анализа (см. раздел 4.1). Поэтому и здесь исходные данные представлены, с одной стороны, спорово-пыльцевой характеристикой поверхностных проб, а с другой – среднемесячными температурами июля, зафиксированными метеостанциями за последние десятилетия на территории Иссык-Кульской котловины.

Однако в соответствии со спецификой методик, структуры исходных таблиц для математической обработки несколько отличаются между собой (сравн. табл.1, 2 и 4). В случае применения метода главных компонент прежде всего анализируется фактический материал, состоящий исключительно из палинологических данных. По существу, он представляет собой ту же исходную таблицу, используемую пошаговым регрессионным анализом (см. табл.1), но с удаленной из ее состава информацией о климате (см. табл.2), т.е. на вход программы, реализующей метод главных компонент, подается палинологическая матрица размерностью $X(30,28)$. Данные о климате вводятся позже – на этапе подключения метода множественного регрессионного анализа. Эта информация в виде вектора-столбца присоединяется к матрице главных компонент (см. табл.4).

Поиск параметров регрессионного уравнения для расчета температуры июля по содержанию компонентов в спорово-пыльцевом спектре методом главных компонент выполнен в соответствии с последовательностью, приведенной на обобщенной блок-схеме алгоритма методики (см. рис.3). Построение моделей показано на трех примерах: с использованием всего признакового пространства (28-ми компонентов СПС) и минимизированного. Во втором случае рассмотрен вариант выбора информативной системы компонентов СПС при помощи R-анализа и вариант выбора относительно информативных ком-

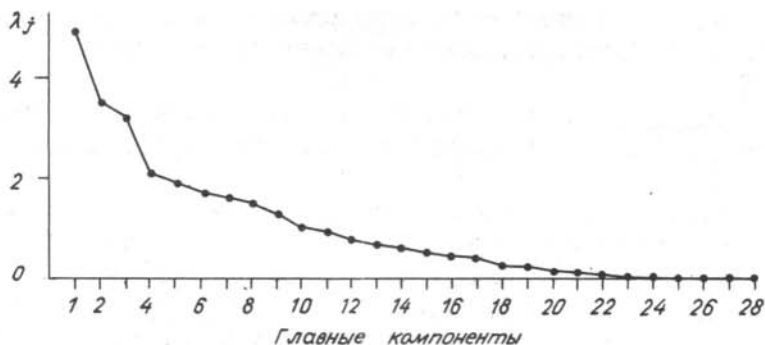


Рис. 6. График собственных значений λ_j исходной корреляционной матрицы рецентных спорово-пыльцевых спектров территории Иссык-Кульской котловины

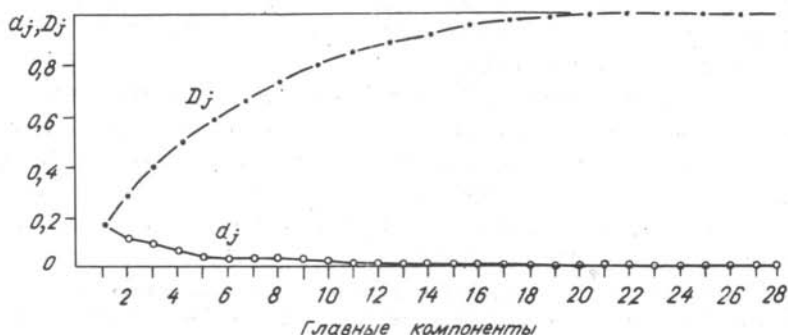


Рис. 7. Графики доли вклада d_j каждой главной компоненты и суммарной доли вклада D_j в дисперсию исходного описания рецентных спорово-пыльцевых спектров территории Иссык-Кульской котловины. Здесь главные компоненты — линейные комбинации 28 исходных компонентов СПС

понентов СПС (по их доле вклада в первую и вторую главные компоненты).

Пример I. Требуется построить расчетную модель для оценки температуры июля с участием главных компонент, представляющих собой линейную комбинацию всех 28 исходных компонентов СПС. Так

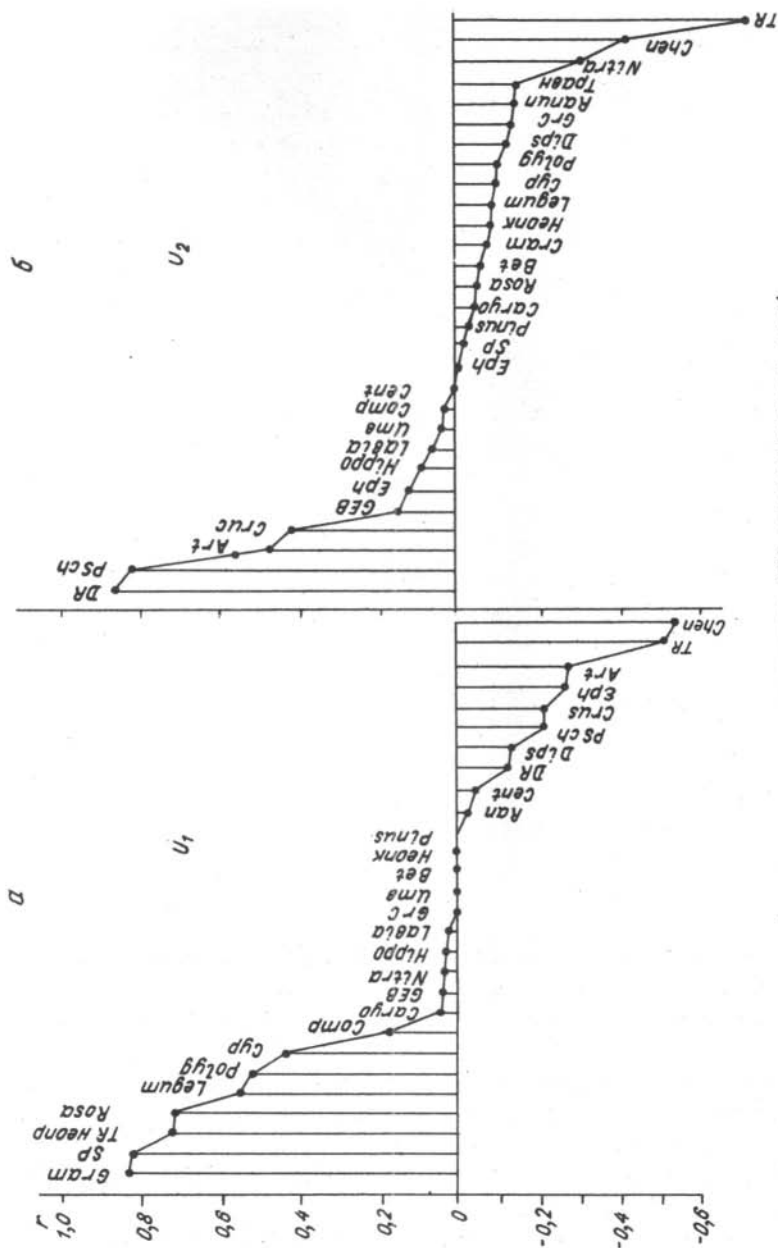


Рис. 8. Относительная информативность компонентов СИС поверхностных проб Иссик-Кульской котловины (вклады в первые две главные компоненты, а и б)

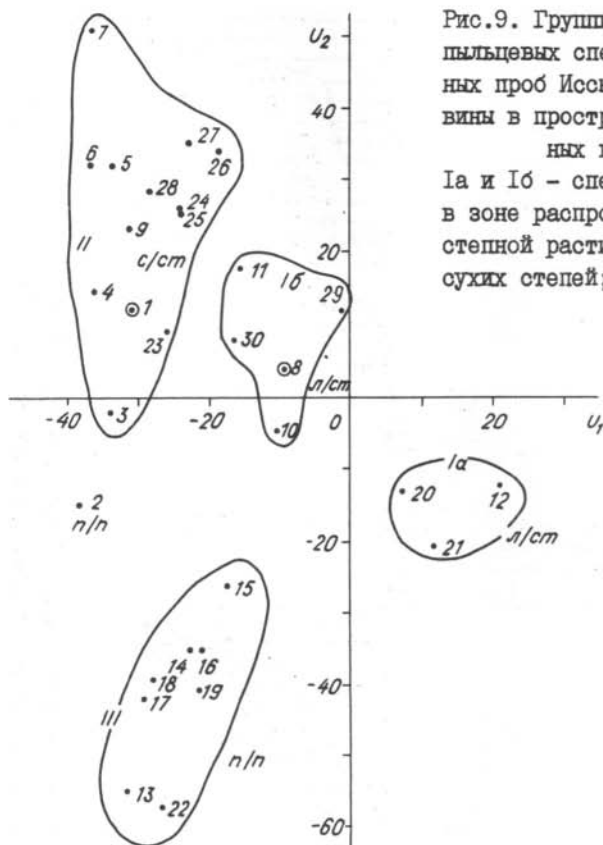


Рис.9. Группирование спорово-пыльцевых спектров поверхностных проб Иссык-Кульской котловины в пространстве двух главных компонент:

Ia и Ib – спектры, отобранные в зоне распространения луговой степной растительности; II – сухих степей; III – полупустынь

- исследуется ИССЫК-КУЛЬСКАЯ КОТЛОВИНА. ПОВЕРХНОСТНЫЕ ПРОБЫ;
- в исходной таблице 30 строк, 28 столбцов;
- ввод собственных значений для рассмотрения не ограничен (CON = 0);
- строить регрессионные модели не требуется (NS = 0);
- требуется распечатать исходную таблицу (PRI1 = 1) и вычисленную корреляционную матрицу (PRI2 = 1), но не печатать усеченную исходную таблицу (PRI3 = 0), ведь анализируется признаковое пространство в полном составе;
- 28 элементов вектора состояния признаков заполнены цифрой 1.

Это означает, что параметры главных компонент следует рассчитывать с учетом всех 28 компонентов СПС:

Распечатка 2

ИССЫК-КУЛЬСКАЯ КОТЛОВИНА. ПОВЕРХНОСТНЫЕ ПРОБЫ

9)	0.410	1.659	←	неопр. кустарн.
10)	9.253	9.557	←	эфедра
11)	8.137	6.924	←	злаковые
12)	0.140	0.438	←	злаковые (культурн.)
13)	1.227	3.127	←	осоковые
14)	0.513	1.208	←	гречишные
15)	31.263	13.681	←	маревые
16)	0.790	1.620	←	гвоздичные
17)	0.333	1.061	←	лютиковые
18)	0.540	1.254	←	крестоцветные
19)	0.967	1.574	←	розоцветные
20)	0.333	0.758	←	бобовые
21)	0.153	0.514	←	мезофиты
22)	0.247	0.600	←	зонтичные
23)	0.080	0.292	←	Dipsacaceae
24)	1.763	1.825	←	губоцветные
25)	13.233	9.479	←	сложноцветные
26)	0.340	0.890	←	Centaurea sp.
27)	29.080	12.760	←	полюны
28)	1.177	2.544	←	неопр. травян.

***** MSDU ***** ERROR = 0

***** TRAC ***** ERROR = 0

K = 28

СОБСТВЕННЫЕ ЗНАЧЕНИЯ, R(M,M):

1) 4.931 2) 3.517 3) 3.217 4) 2.108 5) 1.934 6) 1.718
7) 1.681 8) 1.536 9) 1.311 10) 1.008 11) 0.955 12) 0.717
13) 0.664 14) 0.632 15) 0.523 16) 0.425 17) 0.374 18) 0.220
19) 0.189 20) 0.115 21) 0.091 22) 0.054 23) 0.030 24) 0.026
25) 0.006 26) 0.004 27) 0.002 28) 0.0001

ДОЛЯ ВКЛАДА КАЖДОЙ КОМПОНЕНТЫ:

1) 0.176 2) 0.126 3) 0.115 4) 0.075 5) 0.061 6) 0.061
7) 0.060 8) 0.055 9) 0.047 10) 0.036 11) 0.034 12) 0.026
13) 0.024 14) 0.023 15) 0.019 16) 0.015 17) 0.013 18) 0.008
19) 0.007 20) 0.004 21) 0.003 22) 0.002 23) 0.001 24) 0.001
25) 0.0002 26) 0.0001 27) 0.0001 28) 0.000

НАКОПЛЕННЫЕ ОТНОШЕНИЯ СОБСТВЕННЫХ ЗНАЧЕНИЙ, DX(M):

0.176 0.301 0.417 0.492 0.561 0.622 0.682 0.737

0.784	0.820	0.854	0.880	0.903	0.926	0.945	0.960
0.973	0.981	0.988	0.992	0.995	0.997	0.998	0.999
0.999	0.999	0.999	0.99996				

ВЕКТОРЫ СОБСТВЕННЫХ ЗНАЧЕНИЙ, $v(m, m)$:

ВЕКТОР I

-0.112	-0.163	0.339	-0.109	0.089	0.166	-0.076
-0.011	-0.001	-0.236	0.312	-0.076	0.246	0.264
-0.226	0.063	0.105	-0.081	0.303	0.372	0.071
-0.034	-0.143	0.086	0.215	0.178	-0.145	0.256

ВЕКТОР 2

0.379	-0.408	0.106	0.404	-0.124	-0.141	0.103
-0.037	-0.051	-0.082	-0.018	-0.189	-0.038	-0.128
-0.309	0.065	-0.097	0.247	-0.035	-0.047	0.182
0.077	-0.262	0.070	0.140	-0.046	0.308	-0.021

ВЕКТОР 28

0.561	0.651	0.510	0.010	-0.001	-0.011	-0.004
0.002	0.006	-0.003	-0.008	0.002	-0.001	-0.000
0.004	-0.001	0.007	0.002	0.008	-0.001	-0.001
-0.005	-0.003	0.013	-0.011	-0.011	0.024	0.004

***** LOAD *****

ЧИСЛО ПЕРЕМЕННЫХ, $m = 28$

ЧИСЛО ЗНАЧИМЫХ ГЛАВНЫХ КОМПОНЕНТ, $k = 28$

НАГРУЗКИ НА ГЛАВНЫЕ КОМПОНЕНТЫ, $v2(m, k)$

ПЕРЕМЕННАЯ	НАГРУЗКИ НА ГЛАВНЫЕ КОМПОНЕНТЫ			Пыльца
1)	-0.250	0.711	0.185	← древесных пород
2)	-0.363	-0.766	0.138	← травянистых
3)	0.754	0.200	-0.386	← спор
4)	-0.243	0.758	0.330	← ель Шренка
5)	0.200	-0.233	0.597	← сосна
6)	0.370	-0.264	0.723	← береза
7)	-0.171	0.194	-0.306	← облепиха
8)	-0.025	-0.070	-0.169	← селитрянки
9)	-0.003	-0.095	0.305	← неопр. кустарн.
10)	-0.525	-0.153	-0.092	← эфедра
11)	0.693	-0.035	-0.447	← злаковые

12)	-0.169	-0.354	-0.368	злаковые (культ.)
13)	0.546	-0.072	-0.322	осоковые
14)	0.588	-0.240	0.056	гречишные
15)	-0.503	-0.581	0.346	маревые
16)	0.140	0.123	-0.364	гвоздичные
17)	0.233	-0.183	-0.166	лютиковые
18)	-0.181	0.465	0.257	крестоцветные
19)	0.673	-0.066	-0.199	розоцветные
20)	0.826	-0.088	0.299	бобовые
21)	0.158	0.342	0.161	мезофиты
22)	-0.076	0.144	-0.063	зонтичные
23)	-0.318	-0.491	-0.111	Dipsacaceae
24)	0.190	0.132	0.327	губоцветные
25)	0.478	0.263	0.499	сложноцветные
26)	0.395	-0.086	0.658	Centaurea sp.
27)	-0.322	0.578	-0.135	полыни
28)	0.568	-0.040	-0.187	неопр. травян.

СФОРМИРОВАННАЯ МАТРИЦА ГЛАВНЫХ КОМПОНЕНТ, DD (N, KG)

	U_1	U_2	U_3	
1)	-31.174	12.007	24.778	Пробы по порядку ↓
2)	-38.618	-14.699	32.219	
3)	-33.728	-2.218	34.443	
4)	-35.697	14.441	26.714	
5)	-33.618	31.917	34.387	
6)	-36.371	31.722	27.141	
7)	-36.371	31.722	27.141	
8)	-9.692	-3.783	40.769	
9)	-30.605	22.847	24.654	
10)	-10.588	-4.466	43.743	
11)	-15.526	17.842	30.419	
12)	20.802	-12.053	-3.679	
13)	-31.798	-55.048	14.402	
14)	-23.238	-35.346	10.880	
15)	-17.420	-26.437	-3.263	
16)	-21.383	-34.955	28.999	
17)	-29.671	-41.873	16.249	
18)	-28.331	-39.392	16.072	
19)	-21.479	-40.665	8.634	

20)	7.062	-13.236	-7.172
21)	11.628	-20.461	-8.801
22)	-27.151	-56.915	25.596
23)	-26.293	8.742	13.716
24)	-23.701	26.221	23.353
25)	-23.691	25.651	22.184
26)	-17.577	33.954	29.181
27)	-17.577	33.964	26.838
28)	-28.475	28.639	26.058
29)	-0.743	12.214	21.759
30)	-16.957	6.730	2.855

Примечание к распечатке 2. Для экономии места приведены содержания только первых двух и последнего собственных векторов (1, 2 и 28), а также нагрузки на первые три главные компоненты и сами три главные компоненты (из общего числа - 28).

Анализ и интерпретацию распечатки 2 начнем с собственных значений, минуя информацию о средних значениях, стандартных отклонениях и парных коэффициентах корреляции компонентов СПС, которая должна рассматриваться раньше - на этапе формирования исследуемой палинологической совокупности.

1. Собственные значения корреляционной матрицы (см. распечатку 2 и рис.6). Здесь следует обратить внимание на их численную величину. Это пригодится на втором этапе обработки, при ограничении числа рассматриваемых главных компонент, которое реализуется заданием порога с помощью переменной *CON*. При первом анализе исходной палинологической таблицы рекомендуется задавать *CON* = 0, чтобы иметь информацию о всех главных компонентах.

2. Доля вклада каждой главной компоненты в описание исходной дисперсии и накопленные отношения собственных значений. Последний параметр представляет собой суммарную долю вклада главных компонент в общую изменчивость. Анализ перечисленных показателей помогает правильно выбрать число главных компонент, вводимых в расчетную модель. Оптимальным считается такой вариант, когда только первые две-три главные компоненты участвуют в построении регрессии. Использование в качестве аргументов уравнения большо-

го числа главных компонент при значительном количестве исходных признаков делает его громоздким и непригодным для ручного счета.

В литературе нередко утверждается, что первые две-три главные компоненты обуславливают до 85-90 % общей исходной дисперсии. Однако оказалось, что палинологический материал данной выборки таким свойством не обладает. Доли вклада даже первых главных компонент сравнительно невелики: 0,18; 0,13; т.е. 18 и 13 %. Так что здесь первые две главные компоненты учитывают 31 % общей исходной дисперсии, четыре - 49 %, десять - 82 % и только 20 главных компонент - почти всю исходную изменчивость - 99 % (см. распечатку 2 и рис.7).

На основе анализа доли их вклада приходим к выводу, что нельзя ограничиваться участием только двух-трех главных компонент. Их число следует существенно увеличить. Поэтому принимаем решение построить регрессионные модели с участием в них двух, четырех и десяти главных компонент, чтобы оценить результаты сходности расчетных данных с фактическими температурными при использовании минимального, разумного максимального и промежуточного числа главных компонент.

Исходя из того, что в регрессионные модели мы вводим не более десяти главных компонент, выбираем порог CON для собственных значений, которым воспользуемся непосредственно при построении моделей. Для этого вновь обратимся к собственным значениям. Из распечатки 2 видно, что 10-е собственное значение $\lambda_{10} = 1,008$, а 11-е (которое не нужно вводить) $\lambda_{11} = 0,956$. Следовательно, задав $\text{CON} = 1$ на втором этапе обработки, мы тем самым ограничиваем математическую обработку десятью главными компонентами.

Значение переменной CON запомним для себя и используем при определении параметров расчетной модели.

3. Векторы собственных значений. Каждому собственному значению соответствует собственный вектор, представляющий собой набор коэффициентов соответствующей главной компоненты. Каждый вектор состоит из 28 коэффициентов (по числу исходных компонентов СПС).

Таким образом, эта часть распечатки содержит необходимые данные для расчета самих главных компонент по исходной палинологической таблице. Ею же пользуются для аналитической записи главной компоненты, используя элементы соответствующего собственного вектора.

4. Нагрузки на главные компоненты. Они представляют собой коэффициенты парной корреляции между главной компонентой и исходными признаками. Первый столбец (см. распечатку 2) содержит нагрузки на первую главную компоненту каждого из 28 компонентов СПС. Второй столбец соответственно указывает на корреляцию второй главной компоненты с каждым из компонентов СПС. Как известно, коэффициент парной корреляции изменяется в пределах от -1 до $+1$. Линеальной функциональной зависимости между величинами соответствует абсолютное значение $r = |I|$, полному отсутствию корреляционной связи отвечает $r = 0$.

Анализ нагрузок на главные компоненты позволяет выявить, какой из исходных признаков (компонентов СПС) максимально описывается данной главной компонентой. Его можно проводить по значениям начальных нагрузок (именно они представлены в распечатке 2) и окончательных, получаемых в результате дополнительно проводимого варимаксного вращения. На рис.8 вынесены результаты второго варианта. На графиках нагрузок на первую и вторую главные компоненты видно, что максимальная информация в описании исходного палинологического материала заключена в содержании пыльцы злаковых ($r = 0,83$), общего количества спор ($r = 0,82$), пыльцы неопределенной травянистой растительности ($r = 0,72$) и розоцветных ($r = 0,72$) (см. рис.8,а). Эта группа определена первой главной компонентой. Таким образом, основная информация в исходной палинологической таблице принадлежит пыльце травянистых растений и содержанию спор в спектре.

Следующая по значимости информация – содержание пыльцы древесных пород и кустарников в спектре ($r = 0,87$), а также ели Шренка ($r = 0,81$). Эта группа компонентов СПС определена второй главной компонентой. Значит, следующая по важности информация касается участия пыльцы древесных пород в спектре.

Анализ нагрузок на главные компоненты может быть полезен для проведения минимизации признакового пространства, т.е. сокращения числа членов для описания главных компонент, а следовательно числа членов расчетного уравнения регрессии. Поэтому отмечаем для себя, что имеет смысл построить расчетные модели на усеченном признаковом пространстве, используя только те компоненты СПС (они перечислены выше), которые выявили максимальную корреляцию с главными компонентами. Построение расчетной модели на

усеченной исходной палинологической матрице показано в примере 2 (см. с.65).

5. Матрица главных компонент. Она представляет собой набор новых переменных, независимых между собой. Каждая главная компонента — это линейная комбинация исходных признаков. Поскольку коэффициенты главных компонент уже известны (см. собственные векторы), а значения исходных признаков представлены в палинологической таблице, то программа, вычислив первую главную компоненту по формуле (3.1), помещает ее в виде первого столбца матрицы. Второй столбец представляет собой вторую главную компоненту и т.д.

Распределив в пространстве двух первых главных компонент объекты изучения (в нашем случае — рецентные спорово-пыльцевые спектры), можно выявить реально существующие группы однородных объектов в исследуемой выборке (см. рис.9). На графике четко выделяются три компактные группы.

В группу Ia объединились три рецентных лугово-степных спектра, остальные из них (точки II, 29 и 30 на графике) составили группу Ib. По своему составу они оказались ближе к спектрам сухих степей (группа II).

Группа II объединила в себе все спектры сухих степей. Кроме того, сюда попали три из тех пяти спектров, для которых не был однозначно определен тип растительности, и два спектра, относимые к растительности полупустынь (точки I и 2). И, как отмечалось выше, близкими по содержанию к спектрам этой группы оказались четыре лугово-степных спектра (точки IO, II, 29 и 30).

Группа II самая многочисленная.

Группа III состоит в основном из рецентных спектров поверхностных проб, отобранных в зоне распространения растительности полупустынь. Кроме того, сюда попали два спектра из тех пяти, принадлежность которых к одной из растительных зон однозначно не была определена (точки I3 и I4).

Чтобы получить обобщенную характеристику спорово-пыльцевых спектров, вошедших в каждую из выделенных групп, составим таблицу, в которой для представителей каждой группы запишем содержание компонентов СПС, имеющих максимальные нагрузки на первую и вторую главные компоненты (табл.6).

Оказалось, что в группу Ia (лугостепь) объединились спектры,

Таблица 6

Характеристика спорово-пыльцевых спектров,
сгруппировавшихся в пространстве двух главных компонент,
по признакам с максимальными нагрузками

№ п/п	Тип растительности в месте отбора по- верхностной пробы	Группа	Компонент спорово-пыльцевого спектра, %						
			Общее содержание			Picea Schren- kiana	Grami- neae	Rosa- ceae	Cheno- podia- ceae
			пыльцы древесных пород	пыльцы травянистых растений	спор				
12	Лугостепь	Ia	0	55,5	44,5	0	24,0	4,0	12,0
20	— " —		5,5	55,5	39,0	0	16,0	0	14,4
21	— " —		3,5	58,9	37,6	0	24,0	6,0	16,0
29	— " —	Iб	8,0	57,5	34,5	71,4	24,0	0	20,0
30	— " —		16,8	66,4	16,8	53,4	16,9	3,4	20,0
5	Сухая степь	II	46,0	47,0	7,0	98,0	3,0	0	44,0
6	— " —		35,7	62,4	1,9	98,0	6,0	1,3	25,2
17	Полупустыня	III	6,4	87,2	6,4	0	0	0	60,0
18	— " —		5,3	87,7	7,0	0	0	0	52,0

для которых характерно минимальное участие пыльцы древесных пород (0-5,5 %) и маревых (12-16 %), а злаковых (16-24 %) и розоцветных (4-6 %) - самое максимальное по сравнению с остальными двумя группами. Содержание пыльцы травянистых растений составляет 55-59 %, а спор 38-45 %. Пыльца ели Шренка отсутствует.

Спектры группы Iб (лугостепь) отличаются от спектров группы Iа небольшим увеличением количества пыльцы древесных пород (8-17 %) и присутствием пыльцы ели Шренка (52-60 %).

Группу II (сухая степь) составляют спектры, для которых отмечено почти равное содержание в спектре пыльцы древесных (36-46%) и травянистых растений (46-62 %), с обязательным участием пыльцы ели Шренка (25-44 %), с заниженным количеством пыльцы злаковых (3-6 %) и розоцветных (0-1,3 %), но повышенным (в 2-3 раза) содержанием маревых (25-44 %) по сравнению с лугово-степными спектрами.

Группа III (полупустыня) состоит из спектров, в которых преобладает пыльца травянистых растений (87-88 %) и максимально участвуют маревые (52-60 %). В них отсутствует пыльца ели Шренка, злаковых и розоцветных.

Таким образом, анализ показал, что рецентные спорово-пыльцевые спектры дифференцируются по своему составу в соответствии с растительной зоной, где отбирались поверхностные пробы. В то же время существуют различия в среднемесячных температурах (амплитуда колебаний составляет от 16,3 до 18,8 градусов). Поэтому делаем вывод, что исходная выборка рецентных спорово-пыльцевых спектров и имеющиеся температурные данные пригодны для решения поставленной задачи.

Итак, по результатам обработки методом главных компонент исходной палинологической информации выяснилась возможная структура искомым регрессионных моделей. Их нужно построить, во-первых, на исходном признаковом пространстве $X(30,28)$, т.е. при участии всех 28 компонентов СПС, во-вторых, на усеченном пространстве $X(30,5)$, т.е. оставив для рассмотрения те пять компонентов СПС, для которых выявилась максимальная корреляция с первой и второй главными компонентами (см. пункт 4). Для первого случая уже выяснено (см. пункт 2), что следует построить модели с участием двух, четырех и десяти главных компонент. Для второго это число еще предстоит определить, исходя из анализа доли вклада главных

компонент, полученных на усеченной палинологической выборке $\chi(30,5)$ (см. пример 2).

Перейдем к рассмотрению процесса построения регрессионных моделей, т.е. ко второму (завершающему) этапу математической обработки.

Второй этап математической обработки – определение параметров регрессионных моделей, осуществляемое с помощью множественного регрессионного анализа. Выполняется той же программой MGKMNK с подключением второй ее части.

Рассмотрим пример I – построение моделей с учетом всего признакового пространства (28 компонентов СПС) и участием в качестве аргументов 2, 4 и 10 главных компонент.

Исходя из того, что на этом этапе обработки к методу главных компонент подключается метод множественного регрессионного анализа, должна быть дополнена и информация, которая использовалась на первом этапе (см. пакет 2). Здесь необходимо указать число моделей, которые нужно построить, число главных компонент, которое при этом должно быть использовано, и, наконец, данные о том элементе климата, для которого рассчитывается модель.

Выполняя предусмотренную программой последовательность ввода (см. рис.4), исходная информация должна быть представлена следующим образом.

Пакет 3

ИССЫК-КУЛЬСКАЯ КОТЛОВИНА. ПОВЕРХНОСТНЫЕ ПРОБЫ

30	28	I	3	0	0	0																							
I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
18.5	70.5	II	86	4	0	6	4	0	3I.4	5.2	0	I.I	0	26.2	...	22	0												
.....																													
16.8	66.4	I6.8	53.4	0	0	20	26.6	0	0	I6.9	I.7	0	0	...	0	30.6													
I7.8	I7.8	I7.8	I7.8	I7.8	I7.8	I7.8	I7.8	I7.8	I7.8	I7.8	I7.8	I7.8	...	I6.3															
2	4	I0																											

Эти данные передадут программе, что:

- исследуется ИССЫК-КУЛЬСКАЯ КОТЛОВИНА. ПОВЕРХНОСТНЫЕ ПРОБЫ;
- в исходной таблице 30 строк, 28 столбцов;
- ввод собственных значений ограничивается переменной $CCN = I$;

- требуется построить три регрессионные модели ($NS = 3$);
- не нужно распечатывать исходную таблицу ($PR11 = 0$), корреляционную матрицу ($PR12 = 0$) и усеченную исходную таблицу ($PR13 = 0$);
- вектор состояния признаков содержит в каждой своей позиции цифру I (это означает, что для определения параметров главных компонент следует учитывать все 28 компонентов СПС);
- исходная палинологическая таблица $X(30,28)$. Здесь построчно представлено содержание всех 30 спектров, каждый из которых описан 28 компонентами СПС;
- количественные оценки среднемесячной температуры июля, расположены в соответствии с пробами, к которым они приписаны, т.е. указывается 30 значений температуры июля (по числу проб);
- 2, 4, 10 - это число главных компонент соответственно в первой, второй и третьей модели.

Программа последовательно реализует метод главных компонент с определением параметров заданных 10 главных компонент, а потом перейдет к построению моделей множественным регрессионным анализом. Информация о главных компонентах была рассмотрена выше. Здесь же ее содержание остается тем же самым, с той лишь разницей, что на печать выдаются данные только о десяти главных компонентах. Поэтому, чтобы избежать ненужного повторения, рассмотрим распечатку, начиная с того места, где начался вывод результатов множественного регрессионного анализа.

Распечатка 3

Результаты определения параметров регрессионных моделей на материале палинологической матрицы $X(30,28)$

ЦЕЛЕВОЙ СТОЛБЕЦ, W_K

17.8	17.8	17.8	17.8	17.8	17.8	17.8	17.8	17.8	17.8	17.8	17.8
17.0	17.0	18.8	18.8	18.8	18.8	18.8	18.8	18.8	18.8	17.3	17.3
17.3	16.9	16.9	16.9	16.3	16.3						

МАТРИЦА ГЛАВНЫХ КОМПОНЕНТ + ЦЕЛЕВОЙ СТОЛБЕЦ, $T^0_{\text{июля}}$

	U_1	U_2	U_3	...	$T^0_{\text{июля}}$
1)	-31.174	12.007	24.778	...	17.8

2)	-38.618	-14.699	32.219	...	I7.8
3)	-33.728	-2.218	34.443	...	I7.8
4)	-35.697	I4.44I	26.714	...	I7.8
5)	-33.618	3I.9I7	34.387	...	I7.8
6)	-36.37I	3I.722	27.14I	...	I7.8
7)	-36.044	50.984	28.669	...	I7.8
8)	-9.692	-3.783	40.769	...	I7.8
9)	-30.605	22.847	24.654	...	I7.8
10)	-10.588	-4.466	43.743	...	I7.8
II)	-15.526	I7.842	30.4I9	...	I7.8
I2)	20.802	-12.053	-3.679	...	I7.8
I3)	-3I.798	-55.048	I4.402	...	I7.0
I4)	-23.238	-35.346	I0.880	...	I7.0
I5)	-17.420	-26.437	-3.263	...	I8.8
I6)	-2I.383	-34.955	28.999	...	I8.8
I7)	-29.67I	-4I.873	I6.249	...	I8.8
I8)	-28.33I	-39.392	I6.072	...	I8.8
I9)	-2I.479	-40.665	8.634	...	I8.8
20)	7.062	-13.236	-7.I72	...	I8.8
2I)	II.628	-20.46I	-8.80I	...	I8.8
22)	-27.I5I	-56.9I5	25.596	...	I8.8
23)	-26.293	8.742	I3.7I6	...	I7.3
24)	-23.70I	26.22I	23.353	...	I7.3
25)	-23.69I	25.65I	22.I84	...	I7.3
26)	-17.577	33.954	29.I8I	...	I6.9
27)	-22.430	35.I54	26.838	...	I6.9
28)	-28.475	28.639	26.058	...	I6.9
29)	-0.743	I2.2I4	2I.759	...	I6.3
30)	-16.957	6.730	2.855	...	I6.3

МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

ЧИСЛО НАБЛЮДЕНИЙ - 30

ЧИСЛО ПЕРЕМЕННЫХ - 28

ВЫБОРКА - I

NDER - ЗАДАННЫЙ ИНДЕКС ЗАВИСИМОЙ ПЕРЕМЕННОЙ = II ← T⁰_{ИЮЛЯ}

ЧИСЛО ЗАВИСИМЫХ ПЕРЕМЕННЫХ, K = IO

*** DRDR ***

ВЕКТОР, СОДЕРЖАЩИЙ В ВОЗРАСТАЮЩЕМ ПОРЯДКЕ ИНДЕКСЫ НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ, В K+I - ЗАВИСИМОЙ: I 2 II

ПЕРЕМЕННАЯ	СРЕДНЕЕ	СТАНД.	КОРРЕЛЯЦИЯ	КОЭФ.	СТАН.
НОМ	ЗНАЧЕНИЕ	ОТКЛОНЕНИЕ	Х С У	РЕГРЕССИИ	ОШИБКА
					КОЭФ.
					РЕГР.

$U_1 \rightarrow I$	-21.084	14.635	0.078	-0.00004	0.00864
$U_2 \rightarrow 2$	-1.416	29.928	-0.527	-0.01346	0.00423

ЗАВИСИМАЯ ПЕРЕМЕННАЯ

$T_{\text{июля}}^0$ II 17.773 0.764

СВОБОДНЫЙ ЧЛЕН УРАВНЕНИЯ 17.75339

КОЭФФИЦИЕНТ МНОЖЕСТВЕННОЙ КОРРЕЛЯЦИИ 0.52682

ОЦЕНКА СТАНДАРТНОЙ ОШИБКИ 0.67323

ТАБЛИЦА ФАКТИЧЕСКИХ И ВЫЧИСЛЕННЫХ ЗНАЧЕНИЙ ЗАВИСИМОЙ ПЕРЕМЕННОЙ

НОМЕР	ФАКТИЧ.	ВЫЧИСЛ.	РАЗНОСТЬ	ОШИБКА, %
I	17.79999	17.59296	0.20703	1.16
2	17.79999	17.95259	-0.15260	0.86
3	17.79999	17.78445	0.01553	0.09
4	17.79999	17.56038	0.23961	1.35
5	17.79999	17.32515	0.47484	2.67
6	17.79999	17.32788	0.47211	2.65
7	17.79999	17.06868	0.73131	4.11
8	17.79999	17.80463	-0.00464	0.03
9	17.79999	17.44710	0.35289	1.98
10	17.79999	17.81386	-0.01387	0.08
11	17.79999	17.51387	0.28612	1.61
12	17.79999	17.91478	-0.11479	0.64
13	17.00000	18.49527	-1.49527	8.80
14	17.00000	18.22984	-1.22984	7.23
15	18.79999	18.10974	0.69025	3.67
16	18.79999	18.22450	0.57549	3.06
17	18.79999	18.31790	0.48209	2.56
18	18.79999	18.28447	0.51552	2.74
19	18.79999	18.30135	0.49864	2.65
20	18.79999	17.93120	0.86879	4.62
21	18.79999	18.02826	0.77173	4.10
22	18.79999	18.52020	0.27979	1.49
23	17.29999	17.63670	-0.33672	1.95
24	17.29999	17.40143	-0.10144	0.59
25	17.29999	17.40909	-0.10910	0.63

26	I6.89999	I7.297I3	-0.397I4	2.35
27	I6.89999	I7.28II9	-0.38I20	2.26
28	I6.89999	I7.36908	-0.46909	2.78
29	I6.29999	I7.58905	-I.28906	7.9I
30	I6.29999	I7.66345	-I.36346	8.36

ОШИБКА УСРЕДНЕННАЯ, % 2.83

Представленная часть распечатки 3 содержит:

1. Фактические данные по температуре июля (целевой столбец).
2. Сформированную матрицу главных компонент, к которой в виде последнего столбца присоединен столбец температурных данных. Поскольку в рассмотрении участвуют 10 главных компонент, то размерность таблицы - $U(30, 10+1)$. На ее материале выполняется множественный регрессионный анализ.

Для экономии места в распечатке приведены только первые три столбца (три первые главные компоненты) и последний столбец (целевой).

3. Далее следуют непосредственные результаты регрессионных построений. Для каждой модели указывается, сколько независимых переменных (главных компонент) вводится в уравнение, распечатываются свободный член и коэффициенты уравнения, коэффициент множественной корреляции, стандартная ошибка, расхождения между фактическими и расчетными температурными данными (абсолютные и относительные). Для экономии места в распечатке приводится только одна модель: с участием двух главных компонент.

Сравним качество трех полученных моделей по величине множественного коэффициента корреляции R , стандартной σ и усредненной Q (в %) ошибок. Оказалось, что для модели с участием:

двух главных компонент	$R = 0,53$, $\sigma = 0,67^0$, $Q = 2,83 \%$,
четырех	"- $R = 0,53$, $\sigma = 0,7^0$, $Q = 2,85 \%$,
десяти	"- $R = 0,77$, $\sigma = 0,6^0$, $Q = 2,02 \%$.

На рис.10 вынесены фактические данные о температуре июля и расчетные данные с участием двух и десяти главных компонент. Оценки по модели с четырьмя главными компонентами близки к оценкам по модели с двумя главными компонентами. Выносить их на график не имело смысла. Как и следовало ожидать, лучшее приближение

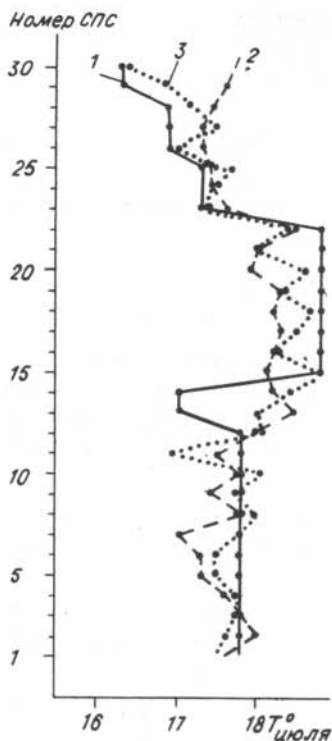


Рис.10. Сопоставление фактических (1) данных о температуре июля Исык-Кульской котловины с расчетными, полученными по двум (2) и десяти (3) главным компонентам. Здесь главные компоненты – линейные комбинации исходных 28 компонентов СПС

к исходным данным показала модель с десятью главными компонентами. Таким образом, исходная информация описывается тем точнее, чем больше главных компонент задействовано в расчетном уравнении.

Запишем расчетную модель оценки среднемесячной температуры июля, построенную по двум главным компонентам, представляющим собой линейную комбинацию 28 компонентов СПС. Для этого воспользуемся данными о свободном члене уравнения и коэффициентах при двух

главных компонентах, рассчитанными множественным регрессионным анализом (см. распечатку 3), а также коэффициентами линейных уравнений самих главных компонент (см. собственные векторы в распечатке 2):

$$\begin{aligned}
 T_{\text{июля}}^0 &= 17,75 - 0,00004 U_1 - 0,01346 U_2 = \\
 &= 17,75 - 0,00004(-0,11263\text{DR} - 0,16342\text{TR} + 0,33959\text{SP} - \\
 &- 0,1092\text{Psch} + 0,08999\text{Pinus} + 0,16645\text{Bet} - 0,07699\text{hippo} - \\
 &- 0,01146\text{Nitra} - 0,00128\text{Neopk} - 0,23654\text{Eph} + 0,31199\text{Gram} - \\
 &- 0,07609\text{Grc} + 0,24604\text{Cyp} + 0,26485\text{Polyg} - 0,22629\text{Cheno} + \\
 &+ 0,06292\text{Caryo} + 0,10509\text{Ranun} - 0,08147\text{Cruc} + 0,30319\text{Rosa} + \\
 &+ 0,37213\text{Legu} + 0,0711\text{GEB} - 0,03445\text{Umb} - 0,14331\text{Dips} + \\
 &+ 0,08561\text{Labi} + 0,21503\text{Comp} + 0,17809\text{Cent} - 0,14511\text{Art} + \\
 &+ 0,2556\text{Neopt}) - 0,01346(0,37882\text{DR} - 0,40832\text{TR} + \\
 &+ 0,10686\text{SP} + 0,040428\text{Psch} - 0,12438\text{Pinus} - 0,14079\text{Bet} +
 \end{aligned}$$

Эта последовательность данных передаст программе, что:

- объект исследования ИССЫК-КУЛЬСКАЯ КОТЛОВИНА. ПОВЕРХНОСТНЫЕ ПРОБЫ;
- в исходной таблице 30 СТРОК И 28 СТОЛБЦОВ;
- ввод собственных значений не ограничен ($CON = 0$);
- не требуется строить регрессионные модели ($NS = 0$);
- не требуется распечатывать исходную таблицу ($PRI1 = 0$) и корреляционную матрицу ($PRI2 = 0$), но нужно распечатать усеченную исходную матрицу ($PRI3 = 1$);
- вектор состояния признаков содержит цифру 1 в позициях 1, 3, 4, 11 и 19, соответствующих номерам столбцов, отведенных для интересующих нас компонентов СПС в исходной таблице $X(30,28)$. В остальных ячейках проставлены нули, и эти признаки не допускаются к участию в расчетных уравнениях;
- исходная палинологическая таблица $X(30,28)$, из которой будут выбраны требуемые столбцы и из них будет сформирована усеченная таблица, которая будет рассматриваться как исходная.

Распечатка 4

Результаты определения параметров главных компонент
на материале палинологической матрицы $X(30,5)$

ИССЫК-КУЛЬСКАЯ КОТЛОВИНА. ПОВЕРХНОСТНЫЕ ПРОБЫ

N - ЧИСЛО ОБЪЕКТОВ 30

M - ЧИСЛО ПРИЗНАКОВ 28

ОГРАНИЧЕНИЕ ЧИСЛА СОБСТВЕННЫХ ЗНАЧЕНИЙ, $CON = 0$

NS - ЧИСЛО ВЫБОРОК 5

PRI1 - ПРИЗНАК ПЕЧАТИ ИСХОДНОЙ ТАБЛИЦЫ 0

PRI2 = 0 PRI3 = 1

ВЕКТОР ПРИЗНАКОВ: 1011000000100000001000000000

ЧИСЛО ИССЛЕДУЕМЫХ ПРИЗНАКОВ M = 5

РАСПЕЧАТКА ПЕРВЫХ ЭЛЕМЕНТОВ СТРОК ИСХОДНОЙ ТАБЛИЦЫ

1) 18.5	2) 9.1	3) 6.1	4) 12.1	5) 46.0	6) 35.7
7) 53.5	8) 14.2	9) 15.0	10) 14.2	11) 4.8	12) 0.0
13) 2.8	14) 6.2	15) 4.4	16) 6.4	17) 6.4	18) 5.3
19) 0.0	20) 5.5	21) 3.5	22) 0.0	23) 7.7	24) 20.5
25) 11.8	26) 40.0	27) 27.8	28) 23.0	29) 8.0	30) 16.8

***** CORRO ***** ERROR= 0

СРЕДНЕЕ ЗНАЧЕНИЕ		СТАНДАРТНОЕ ОТКЛОНЕНИЕ		
I)	I4.I77	I3.857	←	пыльца древесных пород
3)	I3.390	I2.4I0	←	споры
4)	54.067	4I.675	←	ель Шренка
II)	8.I37	6.974	←	злаковые
I9)	0.967	I.574	←	розоцветные

***** MSDU ***** ERROR = 0

***** TRAC ***** ERROR = 0

K = 5

СОБСТВЕННЫЕ ЗНАЧЕНИЯ, R (M,M):

I) 2.4I4 2) I.324 3) 0.597 4) 0.428 5) 0.235

ДОЛЯ ВКЛАДА КАЖДОЙ ГЛАВНОЙ КОМПОНЕНТЫ:

I) 0.483 2) 0.265 3) 0.II9 4) 0.086 5) 0.047

НАКОПЛЕННЫЕ ОТНОШЕНИЯ СОБСТВЕННЫХ ЗНАЧЕНИЙ, DX(M):

0.483 0.748 0.867 0.953 0.999

ВЕКТОРЫ СОБСТВЕННЫХ ЗНАЧЕНИЙ, v(M,M)

ВЕКТОР I

0.399 -0.499 0.402 -0.502 -0.42I

ВЕКТОР 2

0.578 0.363 0.58I 0.270 0.350

ВЕКТОР 3

-0.075 -0.239 -0.023 -0.523 0.8I4

ВЕКТОР 4

0.5I9 -0.539 -0.476 0.436 0.I56

ВЕКТОР 5

-0.482 -0.520 0.523 0.460 0.II3

***** LOAD ***** ERROR = 0

ЧИСЛО ПЕРЕМЕННЫХ, M = 28

ЧИСЛО ЗНАЧИМЫХ ГЛАВНЫХ КОМПОНЕНТ, K = 5

НАГРУЗКИ НА ГЛАВНЫЕ КОМПОНЕНТЫ, v2(M,K)

ПЕРЕМЕННАЯ НАГРУЗКИ НА ГЛАВНЫЕ КОМПОНЕНТЫ

I) 0.620 0.666 -0.058 ← пыльца древесных

3) -0.777 0.4I8 -0.I85 ← споры

4) 0.625 0.669 -0.0I8 ← ель Шренка

II) -0.780 0.3I0 -0.404 ← злаковые

I9) -0.654 0.403 0.629 ← розоцветные

СФОРМИРОВАННАЯ МАТРИЦА ГЛАВНЫХ КОМПОНЕНТ, DD(N,K)

	U_1	U_2	U_3
1)	33.847	66.118	-8.731
2)	30.121	47.786	-4.099
3)	31.443	52.039	-4.749
4)	35.936	62.090	-3.274
5)	52.740	86.954	-8.975
6)	49.340	81.013	-7.666
7)	58.239	90.169	-8.788
8)	17.993	51.287	-5.734
9)	38.426	66.219	-7.201
10)	19.840	50.047	-6.316
11)	30.110	68.909	-10.680
12)	-35.977	24.037	-19.943
13)	-3.153	3.912	-4.657
14)	-4.800	7.918	-6.775
15)	-10.744	11.137	-6.532
16)	6.506	28.862	-9.209
17)	-0.647	6.026	-2.014
18)	-1.385	5.607	-2.075
19)	-6.278	3.372	-6.538
20)	-25.333	21.662	-18.114
21)	-31.976	24.255	-16.927
22)	-1.808	0.971	-1.569
23)	16.435	53.176	-7.012
24)	31.031	74.440	-10.683
25)	28.706	72.987	-10.941
26)	40.469	83.412	-14.830
27)	36.359	74.056	-10.606
28)	38.307	75.024	-9.063
29)	2.604	65.160	-23.057
30)	9.854	52.626	-12.589

Распечатку 4 будем анализировать в том же порядке, что и распечатку 2 (см. пример I).

I. Собственные значения корреляционной исходной матрицы имеют оценки от 2,8 до 0,3 (см. распечатку 4 и рис. II, а). Эту информацию используем для определения порогового значения CON . Отмечаем для себя, что на этапе построения регрессионных моделей следует задать $CON = 0$, если захотим учесть все пять главных компонент.

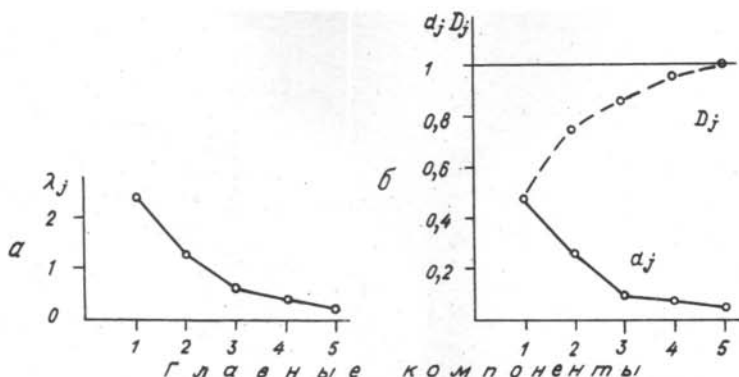


Рис. II. Графики собственных значений λ_j , доли вклада d_j каждой главной компоненты и суммарной доли вклада D_j в дисперсию исходного описания рецентных спорово-пыльцевых спектров Иссык-Кульской котловины, охарактеризованных пятью компонентами

2. Доля вклада главных компонент в описание исходной изменчивости существенно возросла (см. распечатку 4 и рис. II, б). Теперь первая главная компонента описывает 48 %, вторая – 26 % этой изменчивости (сравните с 18 и 12 %, полученными на полном признаковом пространстве в примере I).

3. Накопленная доля вклада двух главных компонент составляет 75 %, а трех – 86 % (см. распечатку 4 и рис. II, б). Пять главных компонент описывают исходную изменчивость полностью. Поэтому принимаем решение построить модели с участием одной, двух, трех, четырех и пяти главных компонент.

4. Матрица главных компонент. Обратимся к первой и второй главным компонентам (это соответственно первый и второй столбцы в таблице). Построим с их помощью распределение исходных спорово-пыльцевых спектров (объектов изучения) в пространстве первой и второй главных компонент (см. рис. I2). Убеждаемся, что по пяти компонентам СПС выделились те же группы спектров, что и по всем 28 (сравните с рис. 9). Это служит подтверждением того, что выбранные пять компонентов СПС заключают в себе достаточную информацию для дифференцированного разделения спектров на классы и

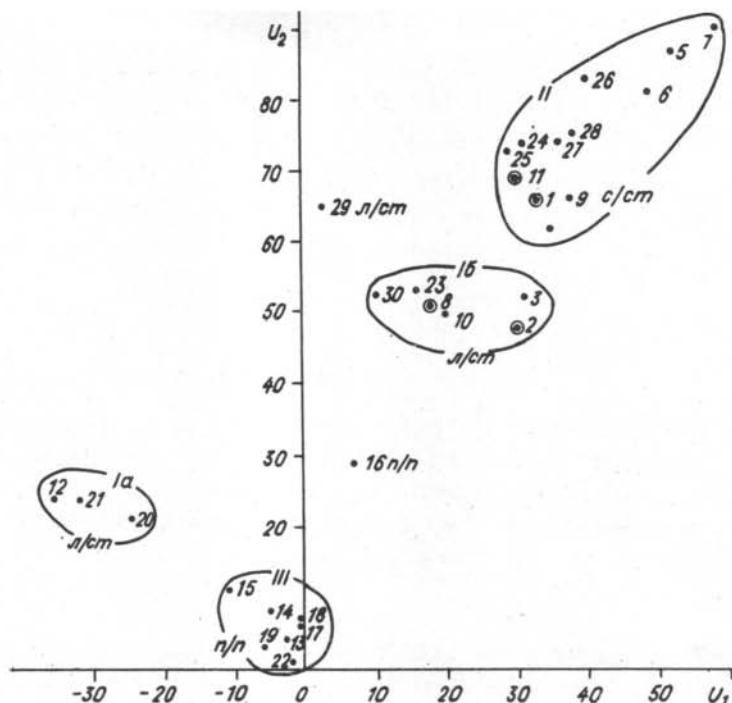


Рис.12. Группирование 30 спорово-пыльцевых спектров поверхностных проб Иссык-Кульской котловины в пространстве двух главных компонент:

Iа и Iб – спектры, отобранные в зоне распространения лугово-степной растительности; II – сухих степей; III – полупустынь. Здесь главные компоненты – линейные комбинации пяти компонентов СПС

вполне могут быть использованы для построения регрессионных моделей.

Второй этап математической обработки – определение параметров регрессионных моделей.

Вновь обращаемся к программе MGKMNK, предварительно дополнив данные пакета 4 информацией, задающей: а) количество требуемых (построить) регрессионных уравнений; б) число главных компо-

нент, вводимых в каждую модель; в) фактические данные о температуре июля, т.е. готовим пакет, аналогичный пакету 3.

ИССЫК-КУЛЬСКАЯ КОТЛОВИНА. ПОВЕРХНОСТНЫЕ ПРОБЫ

Приведенные данные передадут в программу, что:

В распечатке 5 приводятся данные по построению регрессионной модели с участием двух главных компонент, для остальных моделей - сравнительные оценки.

Распечатка 5

Результаты определения параметров регрессионных моделей
на материале палинологической матрицы X(30,5)

ЦЕЛЕВОЙ СТОЛБЕЦ, ВЛК - $T^{\circ}_{\text{июля}}$

I7.8 I7.8 I7.8 I7.8 I7.8 I7.8 I7.8 I7.8 I7.8 I7.8 I7.8 I7.8 I7.8
I7.0 I7.0 I8.8 I8.8 I8.8 I8.8 I8.8 I8.8 I8.8 I8.8 I7.3 I7.3
I7.3 I6.9 I6.9 I6.9 I6.3 I6.3

МАТРИЦА ГЛАВНЫХ КОМПОНЕНТ + ЦЕЛЕВОЙ СТОЛБЕЦ

	U_1	U_2	U_3	U_4	U_5	$T^{\circ}_{\text{июля}}$
I)	33.847	66.118	-8.731	-34.967	32.714	I7.8
2)	30.121	47.786	-4.099	-28.529	33.344	I7.8
3)	31.443	52.039	-4.749	-35.064	38.838	I7.8
4)	35.936	62.090	-3.274	-38.939	37.763	I7.8
5)	52.740	86.954	-8.975	-25.214	26.815	I7.8
6)	49.340	81.014	-7.666	-26.595	36.563	I7.8
7)	58.239	90.169	-8.788	-18.558	26.795	I7.8
8)	I7.993	51.287	-5.734	-25.581	20.345	I7.8
9)	38.426	66.219	-7.201	-37.070	41.390	I7.8
10)	I9.840	50.047	-6.316	-26.765	I9.201	I7.8
11)	30.110	68.909	-10.680	-50.205	45.314	I7.8
12)	-35.977	24.037	-19.943	-12.886	-11.676	I7.8
13)	-3.153	3.912	-4.657	5.158	2.560	I7.0
14)	-4.800	7.918	-6.775	5.151	-0.730	I7.0
15)	-10.744	11.137	-6.532	-4.668	-8.662	I8.8
16)	6.506	28.862	-9.209	-10.955	16.402	I8.8
17)	-0.647	6.026	-2.014	-0.128	-6.412	I8.8
18)	-1.385	5.607	-2.075	-1.022	-6.195	I8.8
19)	-6.278	3.372	-6.538	5.449	5.747	I8.8
20)	-25.333	21.662	-18.114	-11.183	-15.590	I8.8
21)	-31.976	24.255	-16.927	-7.041	-9.545	I8.8
22)	-1.808	0.971	-1.883	1.569	1.655	I8.8
23)	16.435	53.176	-7.012	-39.015	20.553	I7.3
24)	31.031	74.440	-10.683	-41.233	29.779	I7.3
25)	28.706	72.987	-10.941	-52.346	32.245	I7.3
26)	40.469	83.412	-14.830	-21.770	29.489	I6.9
27)	36.359	74.056	-10.606	-32.664	27.787	I6.9

28)	38.307	75.024	-9.063	-35.482	37.255	16.9
29)	2.604	65.160	-23.057	-37.934	26.550	16.3
30)	9.854	52.626	-12.589	-17.836	19.234	16.3

МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

ЧИСЛО НАБЛЮДЕНИЙ - 30

ЧИСЛО ПЕРЕМЕННЫХ - 5

ВЫБОРКА - 2

NDER - ЗАДАННЫЙ ИНДЕКС ЗАВИСИМОЙ ПЕРЕМЕННОЙ = 6 ← $T_{июля}^0$

ЧИСЛО ЗАВИСИМЫХ ПЕРЕМЕННЫХ, K = 2

*** ORDR *** ERROR = 0

ВЕКТОР, СОДЕРЖАЩИЙ В ВОЗРАСТАЮЩЕМ ПОРЯДКЕ ИНДЕКСЫ НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ, В K+1 - ЗАВИСИМОЙ: I. 2 6

***** MINV ***** ERROR = 0

***** MLTR ***** ERROR = 0

ПЕРЕМЕННАЯ СРЕДНЕЕ СТАНД. КОРРЕЛЯЦИЯ КОЭФ. СТАНД.ОШИБКА

НОМ ЗНАЧЕНИЕ ОТКЛОНЕНИЕ X C Y РЕГРЕССИИ КОЭФ. РЕГР.

$U_1 \rightarrow I$ 16.207 24.875 -0.385 0.00780 0.00898

$U_2 \rightarrow 2$ 47.042 29.293 -0.549 -0.01988 0.00763

ЗАВИСИМАЯ ПЕРЕМЕННАЯ

$T_{июля}^0 \rightarrow 6$ 17.773 0.764

СВОБОДНЫЙ ЧЛЕН УРАВНЕНИЯ 18.58179

КОЭФФИЦИЕНТ МНОЖЕСТВЕННОЙ КОРРЕЛЯЦИИ 0.56564

ОЦЕНКА СТАНДАРТНОЙ ОШИБКИ 0.65317

ТАБЛИЦА ФАКТИЧЕСКИХ И ВЫЧИСЛЕННЫХ ЗНАЧЕНИЙ ЗАВИСИМОЙ ПЕРЕМЕННОЙ

НОМЕР ПРОБЫ	ФАКТИЧ.	ВЫЧИСЛ.	РАЗНОСТЬ	ОШИБКА, %
I	17.79999	17.53166	0.26833	1.51
2	17.79999	17.86694	-0.06696	0.38
3	17.79999	17.79274	0.00725	0.04
4	17.79999	17.62802	0.17197	0.97
5	17.79999	17.26491	0.53508	3.01
6	17.79999	17.35643	0.44356	2.49
7	17.79999	17.24390	0.55609	3.12
8	17.79999	17.70277	0.09721	0.55
9	17.79999	17.56538	0.23460	1.32
10	17.79999	17.74181	0.05818	0.33
11	17.79999	17.44704	0.35295	1.98
12	17.79999	17.82336	-0.02338	0.13

I3	I7.00000	I8.47942	-I.47942	8.70
I4	I7.00000	I8.38696	-I.38696	8.16
I5	I8.79999	I8.2766I	0.52338	2.78
I6	I8.79999	I8.05888	0.74110	3.94
I7	I8.79999	I8.45697	0.34302	I.82
I8	I8.79999	I8.45950	0.34048	I.8I
I9	I8.79999	I8.46577	0.3342I	I.78
20	I8.79999	I7.95360	0.84639	4.50
2I	I8.79999	I7.85023	0.94975	5.05
22	I8.79999	I8.54837	0.25I62	I.34
23	I7.29999	I7.65306	-0.35307	2.04
24	I7.29999	I7.34430	-0.0443I	0.26
25	I7.29999	I7.35504	-0.05505	0.32
26	I6.89999	I7.23959	-0.33960	2.0I
27	I6.89999	I7.39348	-0.49348	2.92
28	I6.89999	I7.38943	-0.48944	2.90
29	I6.29999	I7.30698	-I.00699	6.18
30	I6.29999	I7.6I266	-I.3I267	8.05
ОШИБКА УСРЕДНЕННАЯ, %				2.68

Для сравнительного анализа полученных моделей, как и раньше (см. пример I), воспользуемся коэффициентом множественной корреляции R , оценками стандартной σ и усредненной Q ошибок. Выясняется, что при участии в регрессионной модели

одной главной компоненты	$R = 0,39$	$\sigma = 0,71$	$^{\circ}\text{C}$	$q = 3,31$	%
двух "- "-	$R = 0,57$	$\sigma = 0,65$	$^{\circ}\text{C}$	$q = 2,68$	%
трех "- "-	$R = 0,59$	$\sigma = 0,65$	$^{\circ}\text{C}$	$q = 2,68$	%
четырех "- "-	$R = 0,59$	$\sigma = 0,66$	$^{\circ}\text{C}$	$q = 2,64$	%
пяти "- "-	$R = 0,63$	$\sigma = 0,65$	$^{\circ}\text{C}$	$q = 2,58$	%

Как видим, здесь введение в модель более двух главных компонент не привело к резкому улучшению показателей. Приведенные характеристики дают основание отдать предпочтение модели с участием в ней двух первых главных компонент. Отметим, что она качественно не уступает модели (4.2), построенной по двум главным компонентам с учетом всего признакового пространства (для последней получено $R = 0,53$, $\sigma = 0,67$ $^{\circ}\text{C}$, $q = 2,83$ %). Кроме того, эта модель выигрывает в компактности за счет значительного сок-

Рис.13. Сопоставление фактических (1) данных о температуре июля Исык-Кульской котловины с расчетными, полученными по двум главным компонентам; 2 - линейные комбинации 28 исходных компонентов СПС; 3 - то же по пяти компонентам СПС, имеющих максимальные нагрузки на первые главные компоненты

ращения числа членов, что делает ее удобной и при ручном счете: $T_{\text{июля}}^0 = 18,58 + 0,0078U_1 - 0,01988U_2 = 18,58 + 0,0078(0,39868DR - 0,49975SP + 0,4021PSch - 0,50225Gram - 0,42114Rosa) - 0,01988(0,5783IDR + 0,3632SP + 0,58164PSch + 0,26977Gram + 0,35009Rosa)$. (4.3)

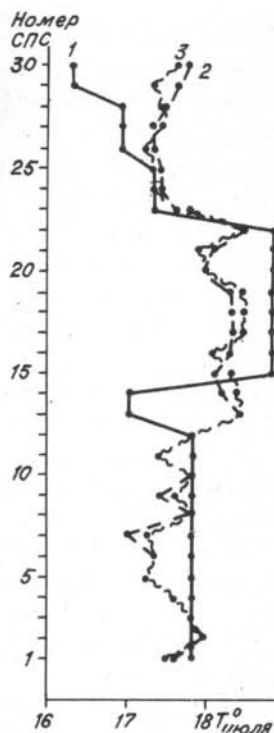
Таким образом, первую главную компоненту можно интерпретировать как зависящую в большей мере от содержания спор и пыльцы злаковых в спорово-пыльцевом спектре, а вторую - от пыльцы древесных пород и ели Шренка.

На рис.13 представлены фактические данные о среднемесячных температурах июля и расчетные данные, полученные на моделях (4.2) и (4.3) с двумя главными компонентами. Модель (4.3) лучше описывает исходный температурный ряд и более компактна, чем модель (4.2).

Пример 3. Требуется построить расчетную модель оценки температуры июля по системе информативных компонентов СПС, определенной методом главных компонент.

Пропуская этап получения предварительной информации о структуре исследуемой палинологической совокупности, который подробно был рассмотрен в примере I, перейдем непосредственно к этапу минимизации (см. рис.3). Выполним его в соответствии с описанием, приведенным в разделе 3.1 и 3.3.

I. Определение групп однородных признаков (компонентов СПС)



6)	10.668	23.058
7)	10.714	23.877
8)	10.714	18.455
9)	10.714	23.794
10)	10.929	18.704
11)	10.714	23.821
12)	7.500	14.669
13)	7.089	20.790
14)	7.143	18.335
15)	10.714	24.420
16)	10.714	19.035
17)	7.161	19.867
18)	7.071	19.195
19)	6.411	19.700
20)	7.286	13.223
21)	6.571	13.507
22)	7.925	22.246
23)	10.707	19.297
24)	10.686	20.851
25)	10.689	22.092
26)	10.714	20.486
27)	10.714	20.641
28)	10.996	22.779
29)	10.643	18.544
30)	10.021	16.993

***** MSDU ***** ERROR = 0

***** TRAC ***** ERROR = 0

K = 29

СОБСТВЕННЫЕ ЗНАЧЕНИЯ, $\lambda(M, M)$:

.....

ДОЛЯ ВКЛАДА КАЖДОЙ ГЛАВНОЙ КОМПОНЕНТЫ:

1) 0.731 2) 0.153 3) 0.038 4) 0.025 29) 0.0

НАКОПЛЕННЫЕ ОТНОШЕНИЯ СОБСТВЕННЫХ ЗНАЧЕНИЙ, $\Delta x (M)$:

0.731 0.884 0.921 0.947 0.962 0.972 0.980 ... 1.0

ВЕКТОРЫ СОБСТВЕННЫХ ЗНАЧЕНИЙ, $V(M, M)$:

ВЕКТОР 1

....

....

ВЕКТОР 29

...

***** LOAD *****

ЧИСЛО ПЕРЕМЕННЫХ, M = 30

ЧИСЛО ЗНАЧИМЫХ ГЛАВНЫХ КОМПОНЕНТ, K = 29

НАГРУЗКИ НА ГЛАВНЫЕ КОМПОНЕНТЫ, V2(M,K):

ПЕРЕМЕННАЯ	НАГРУЗКИ	ГЛАВНЫХ КОМПОНЕНТ			
I)	0.940	-0.255	0.051	-0.029	-0.008
...	...	...	...	...	...
30)	0.919	-0.031	-0.070	0.312	-0.026

СФОРМИРОВАННАЯ МАТРИЦА ГЛАВНЫХ КОМПОНЕНТ, DD (N,KG)

	U ₁	U ₂		
I)	79.939	-47.899	←	пыльца древесных пород
2)	392.914	85.698	←	травянистых
3)	69.760	14.270	←	споры
4)	316.554	-192.159	←	ель Шренка
5)	34.741	-3.100	←	сосна
6)	5.759	0.518	←	береза
7)	16.271	-6.143	←	облепиха
8)	26.013	18.728	←	селитрянка
9)	2.339	0.515	←	неопр. кустарник
10)	51.508	-1.487	←	эфедра
11)	42.299	14.056	←	злаковые
12)	0.746	0.640	←	злаковые (культури.)
13)	5.815	6.344	←	осоковые
14)	2.564	1.970	←	гречишные
15)	169.192	42.096	←	маревые
16)	4.346	0.599	←	гвоздичные
17)	1.651	2.010	←	лютиковые
18)	3.043	-1.767	←	крестоцветные
19)	5.039	1.698	←	розоцветные
20)	1.743	0.679	←	бобовые
21)	0.884	-0.738	←	мезофиты
22)	1.347	-0.300	←	зонтичные
23)	0.424	0.359	←	Dipsacaceae
24)	9.723	0.440	←	губоцветные
25)	72.280	-4.171	←	сложноцветные
26)	2.011	-0.253	←	Centaurea sp.

27)	I60.I05	-I6.5I5	←	полыни
28)	5.476	4.332	←	неопр. травян.

Примечание. В данной распечатке наиболее ценной информацией является сформированная матрица главных компонент, поэтому промежуточные расчеты большей частью были пропущены.

для минимизации признаков (компонентов СПС) используются значения первой и второй главных компонент. Даже визуально выделяются из общей совокупности компонентов СПС: содержание в спектре пыльцы травянистой растительности (порядковый номер 2, расстояние от начала координат по оси первой главной компоненты - 392,9), пыльцы ели Шренка (порядковый номер 4, расстояние - 3I6,6), маревых (номер I5, расстояние - I69,2) и полыней (номер 27, расстояние - I60,I) (рис.I4).

Для большей наглядности изобразим проекции всех компонентов СПС в плоскости первой и второй главных компонент (см. сформированную матрицу главных компонент в распечатке 6). Из полученного распределения (рис.I5) следует, что за информативную систему можно принять перечисленные выше четыре компонента СПС. Далее по информативности следуют: содержание в спектре пыльцы древесных и спор, сложноцветных, эфедры, злаковых, сосны; остальные компоненты СПС скомпоновались в начале координат и с точки зрения информативности не представляют никакого интереса.

Итак, методом главных компонент из исходной совокупности 28 признаков выделилась информативная система, включившая в себя всего четыре компонента СПС.

2. Определение параметров расчетного уравнения регрессии методом главных компонент и множественным регрессионным анализом.

Обратимся вновь к программе MGKMNK, предварительно составив пакет данных, аналогичный пакету 5, с той лишь разницей, что укажем число моделей - четыре, а в векторе состояния признаков проставим цифру I на позициях 2, 4, I5 и 27, а в остальных - цифру 0. Тогда программа построит четыре модели с участием одной, двух, трех и четырех главных компонент. При этом в рассмотрении будут участвовать только четыре компонента СПС: содержание

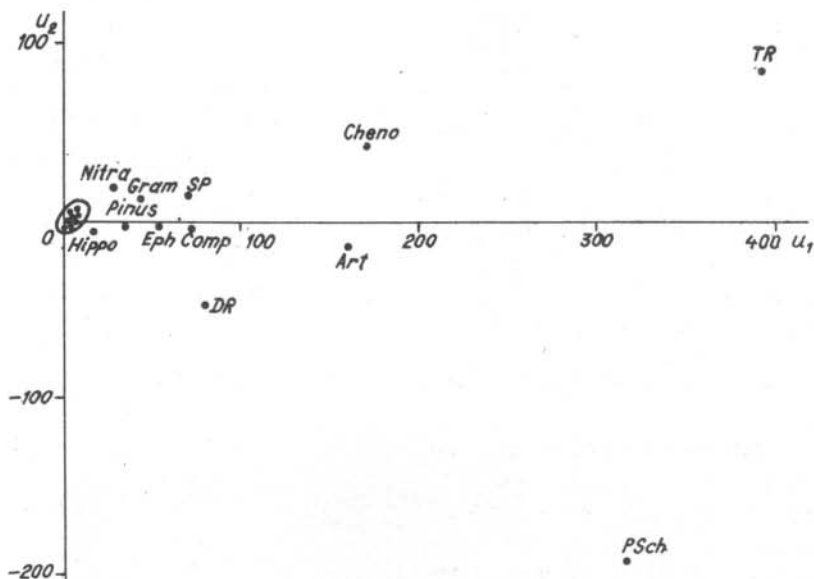
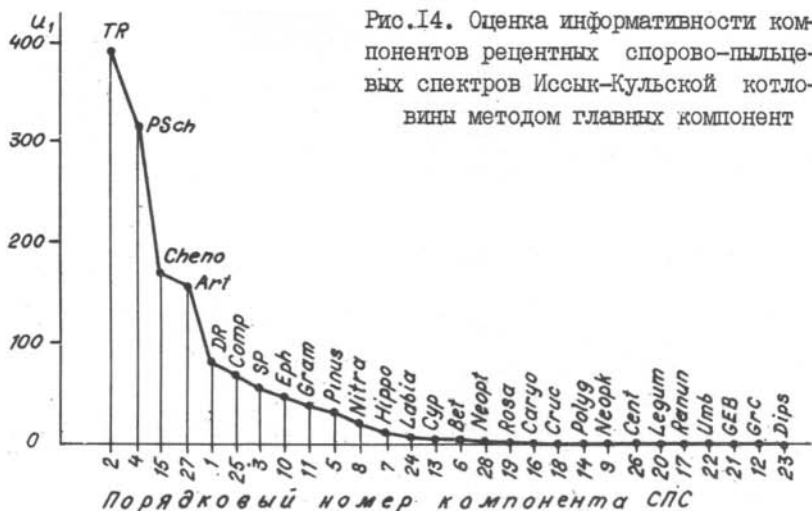


Рис.15. Анализ линейной зависимости компонентов рецентных спорово-пыльцевых спектров Иссык-Кульской котловины методом главных компонент

в спектре пылицы трав, ели Шренка, маревых и полыней. Результаты выполнения, не приводя распечатку, изобразим графически.

График собственных значений представлен на рис.16, а доли вклада каждой главной компоненты и суммарная их доля вклада на рис.17. Из него видно, что первая главная компонента описывает 55 % общей исходной дисперсии, две первых - 77 %, три - 92 %, а четыре - 100 %.

Группирование рецентных спорово-пыльцевых спектров в пространстве первой и второй главных компонент, когда в рассмотрении участвуют только четыре компонента СПС, показало, что оно осталось таким же, что и по всем исходным 28-ми компонентам СПС, а также по пяти, имеющим максимальную нагрузку на первые две компоненты (сравните рис.18 с рис.9 и 12).

Сравнительный анализ построенных четырех моделей проведем по величине коэффициента множественной корреляции R , оценкам стандартной σ и усредненной q (в %) ошибок. Оказалось, что при участии в регрессионной модели:

одной главной компоненты	$R = 0,53$,	$\sigma = 0,66$ °C,	$q = 2,83$ %,
двух " "	$R = 0,53$,	$\sigma = 0,67$ °C,	$q = 2,82$ %,
трех " "	$R = 0,54$,	$\sigma = 0,68$ °C,	$q = 2,81$ %,
четырех " "	$R = 0,56$,	$\sigma = 0,68$ °C,	$q = 2,81$ %.

Как видим, введение в модель двух и более главных компонент не привело к улучшению показателей. Поэтому принимаем в качестве расчетной - модель с участием одной первой главной компоненты:

$$\begin{aligned}
 T_{\text{июля}}^{\text{O}} &= 17,6 + 0,01216 T_1 = \\
 &= 17,6 + 0,01216(0,52236 T_R - 0,50731 p_{\text{sch}} + \\
 &\quad + 0,52408 C_{\text{cheno}} - 0,44172 A_{\text{art}}). \quad (4.4)
 \end{aligned}$$

Исходя из величин коэффициентов первой главной компоненты, можно ее интерпретировать как зависящую практически от всех четырех компонентов СПС.

На рис.19 представлены фактические данные о среднемесячной температуре июля и расчетные, полученные на моделях с минимизированным пространством признаков (4.3) и (4.4). Как видим, они показывают практически одно и то же приближение к фактическим данным, имея при этом сходные статистические показатели: коэффициент множественной корреляции ($R = 0,53$), стандартную ($\sigma = 0,67$ и $0,66$ °C) и усредненную ($q = 2,83$ %) ошибки.

Но модель (4.4) отличается компактностью (число ее членов

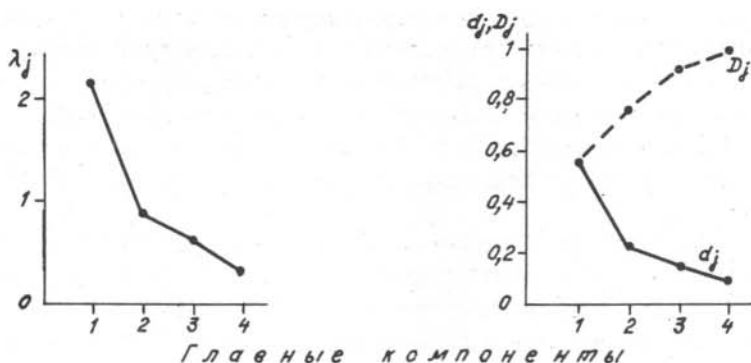


Рис.16. График собственных значений λ_j корреляционной матрицы рецентных спорово-пыльцевых спектров Иссык-Кульской котловины, охарактеризованных четырьмя информативными компонентами СПС

Рис.17. Графики доли вклада d_j каждой главной компоненты и суммарной доли вклада D_j в дисперсию исходного описания рецентных спорово-пыльцевых спектров территории Иссык-Кульской котловины. Здесь главные компоненты – линейные комбинации четырех информативных компонентов СПС

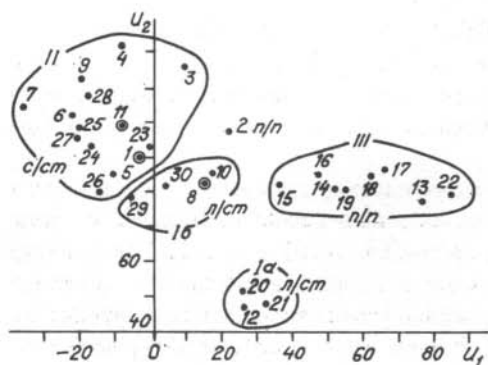


Рис.18. Группирование 30 спорово-пыльцевых спектров поверхностных проб Иссык-Кульской котловины в пространстве двух главных компонент:

Ia и Ib – спектры, отобранные в зоне распространения лугово-степной растительности; II – сухих степей; III – полупустынь. Здесь главные компоненты – линейные комбинации четырех информативных компонентов СПС

Рис.19. Сопоставление фактических (I) данных о температуре июля Иссyk-Куль-ской котловины с расчетными, полученными: 2 – по двум главным компонентам (линейным комбинациям пяти компонентов СПС с максимальными нагрузками); 3 – по одной главной компоненте (линейной комбинации четырех информативных компонентов СПС)

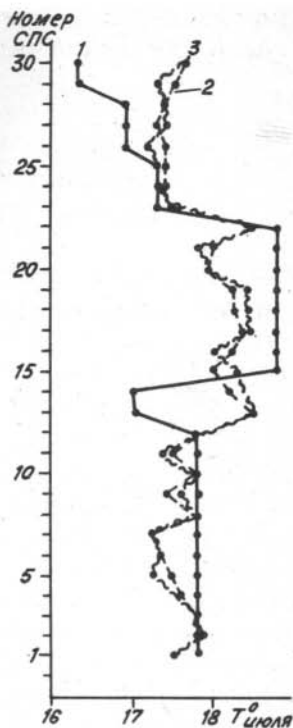
– 5, а не II) и качественным составом. В нее вошли именно те компоненты СПС, которые лучше всего отражают господствующую на изучаемой территории растительность пустынных и полупустынных сообществ, где доминируют маревые и полыни. Кроме того, в нее вошел такой индикатор климата, как пыльца ели Шренка, которым широко пользуются при реконструкциях палеоклиматов голоцена и плейстоцена Иссyk-Куль-ской котловины как показателем наступившего в прошлом увлажнения и продвижения в долины ели Шренка /Озеро Иссyk-Куль ..., 1986/.

Материал, изложенный в данном разделе, позволяет сделать следующие выводы.

1. Минимизацию исходного признакового пространства лучше производить не по величине их нагрузок на главные компоненты, а путем выявления и отбрасывания дублирующих признаков.

2. В качестве расчетной модели оценки температуры июля, полученной методом главных компонент, следует принять модель (4.4).

Таким образом, и на примере метода главных компонент показано, что для количественной оценки элементов палеоклимата по палинологическим данным нет необходимости в участии всего многочисленного состава компонентов СПС. Достаточно для этой цели выбрать систему информативных компонентов СПС. Однако при этом выбор следует производить не принятием волевых решений, отбрасывая



бездоказательно те или иные компоненты СПС, а с использованием объективных методов математической статистики.

4.3. Регрессионная модель оценки температуры июля, построенная по методике количественной оценки элемента климата по информативной системе компонентов СПС

Для сравнительной оценки расчетных моделей, полученных пошаговым регрессионным анализом и методом главных компонент, приведем модель, построенную по апробированной методике /Букреева и др., 1984, 1986/.

Согласно ее алгоритму требуется прежде всего выбрать информативную систему компонентов СПС, а затем по ней строить регрессионное уравнение. Совокупность компонентов СПС, входящих в ее состав, выбирается строго в соответствии с заданной целью, т.е. теми значениями элемента климата, для которого строится расчетная модель. Система определяется итерационным путем, т.е. повторяющимся набором операций, использующих результаты предыдущего цикла. На каждой итерации происходит взвешивание каждого из компонентов СПС в соответствии с целевым признаком (элементом климата). Слабые компоненты СПС "наказываются", а достигая определенного порога, минимизируются, сильные — поощряются. В информативную систему попадают компоненты СПС, имеющие наибольший информационный вес в отношении целевого признака и, кроме того, отличающиеся свойством дополнителности друг к другу, т.е. когда совместное их присутствие усиливает информативность системы в целом. Методика позволяет оценить информационный вес каждого из признаков, входящего в систему (в сумме он составляет 1), а также информативность системы в целом (она измеряется от 0 до 1).

Используя тот же фактический материал по рецентным спорово-пыльцевым спектрам территории Иссык-Кульской котловины и данные о среднемесячных температурах июля, находим информативную систему признаков. В ее состав вошли: 1) общее содержание в спектре пыльцы травянистых растений (информационный вес $p = 0,43$); 2) полыней ($p = 0,2$); 3) спор ($p = 0,19$) и 4) маревых ($p = 0,17$). Информативность системы составила $\bar{I} = 0,72$.

Применив к ней множественный регрессионный анализ, находим параметры расчетной модели:

$T_{\text{июля}}^0 = 15,67 + 0,016TR + 0,02SP + 0,2Chen + 0,003Art$. (4.5)
 Для нее $R = 0,44$, $b = 0,74$ °C, $q = 3,1$ % (см. распечатку 7).

Распечатка 7

Результаты определения параметров регрессионной модели
 по информативной системе компонентов СПС

МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

ЧИСЛО НАБЛЮДЕНИЙ - 30

ЧИСЛО ПЕРЕМЕННЫХ - 28

ВЫБОРКА - I

NDEF - ЗАДАННЫЙ ИНДЕКС ЗАВИСИМОЙ ПЕРЕМЕННОЙ = 28 $\leftarrow T_{\text{июля}}^0$

ЧИСЛО ЗАВИСИМЫХ ПЕРЕМЕННЫХ, K = 4

*** ORDR *** ERROR = 0

ВЕКТОР, СОДЕРЖАЩИЙ В ВОЗРАСТАЮЩЕМ ПОРЯДКЕ ИНДЕКСЫ НЕЗАВИСИМЫХ
 ПЕРЕМЕННЫХ, В K+I - ЗАВИСИМОЙ: 2 3 I5 27 28

***** MINV ***** ERROR = 0

***** MLTR ***** ERROR = 0

ПЕРЕМЕННАЯ	СРЕДНЕЕ	СТАНД.	КОРРЕЛЯЦИЯ	КОЭФ.	СТАНД.ОШИБ-
НОМ	ЗНАЧЕНИЕ	ОТКЛОНЕНИЕ	X C Y	РЕГРЕССИИ	КА. КОЭФ.
				РЕГР.	

TR $\rightarrow$ 2	72.390	I6.I48	0.348	0.0I589	0.0IIIO
SP $\rightarrow$ 3	I3.390	I2.409	-0.077	0.020I7	0.0I698
Chen $\rightarrow$ I5	3I.263	I3.680	0.320	0.0I920	0.0I60I
Art $\rightarrow$ 27	29.080	I2.760	-0.I88	0.00269	0.0I327

ЗАВИСИМАЯ ПЕРЕМЕННАЯ

$T_{\text{июля}}^0 \rightarrow$ 28 I7.773 0.764

СВОБОДНЫЙ ЧЛЕН УРАВНЕНИЯ 15.67466

КОЭФФИЦИЕНТ МНОЖЕСТВЕННОЙ КОРРЕЛЯЦИИ 0.43950

ОЦЕНКА СТАНДАРТНОЙ ОШИБКИ 0.73937

ТАБЛИЦА ФАКТИЧЕСКИХ И ВЫЧИСЛЕННЫХ ЗНАЧЕНИЙ ЗАВИСИМОЙ ПЕРЕМЕННОЙ

НОМЕР ПРОБЫ	ФАКТИЧ.	ВЫЧИСЛ.	РАЗНОСТЬ	ОШИБКА, %
I	I7.79999	I7.5788I	0.22II7	I.24
2	I7.79999	I7.88679	-0.08680	0.49
3	I7.79999	I7.99344	-0.I9345	I.09
4	I7.79999	I7.895I3	-0.095I4	0.53

5	I7.79999	I7.43998	0.3600I	2.02
6	I7.79999	I7.29278	0.5072I	2.85
7	I7.79999	I7.I269I	0.67308	3.78
8	I7.79999	I7.76685	0.033I4	0.I9
9	I7.79999	I7.46888	0.33III	I.86
IO	I7.79999	I7.84367	-0.04368	0.25
II	I7.79999	I7.7042I	0.09578	0.54
I2	I7.79999	I7.73830	0.06I69	0.35
I3	I7.00000	I8.3I530	-I.3I530	7.74
I4	I7.00000	I7.93683	-0.93683	5.5I
I5	I8.79999	I7.94466	0.85533	4.55
I6	I8.79999	I8.I6298	0.6370I	3.39
I7	I8.79999	I8.4I464	0.38535	2.05
I8	I8.79999	I8.28294	0.5I705	2.75
I9	I8.79999	I7.835I3	0.96486	5.I3
20	I8.79999	I7.67650	I.I2348	5.98
2I	I8.79999	I7.7I9I7	I.08082	5.75
22	I8.79999	I8.5096I	0.29038	I.54
23	I7.29999	I7.83684	-0.53685	3.I0
24	I7.29999	I7.6I042	-0.3I043	I.79
25	I7.29999	I7.72586	-0.42587	2.46
26	I6.89999	I7.22376	-0.32377	I.92
27	I6.89999	I7.397I8	-0.497I9	2.94
28	I6.89999	I7.60466	-0.70466	4.I7
29	I6.29999	I7.7327I	-I.43272	8.79
30	I6.29999	I7.53475	-I.23476	7.58
ОШИБКА УСРЕДНЕННАЯ, %				3.08

Применим эту же методику - получение расчетных моделей множественным регрессионным анализом по информативным системам компонентов СПС - к двум наборам, выделенным методом главных компонент, и сравним результаты с полученными ранее соответствующими моделями (4.3) и (4.4), где использовались в качестве аргументов линейные комбинации исходных признаков (главные компоненты).

Тогда модель с участием пяти компонентов СПС, имеющих максимальные нагрузки на первые главные компоненты, запишется в виде:

$$T_{июля}^0 = 18,7 + 0,003DR + 0,007SP - 0,012Psch - 0,15GrC + 0,015Rosa. \quad (4.6)$$

По сравнению с моделью (4.3), где фигурирует тот же набор, эта модель содержит в два раза меньше членов и при этом показывает несколько лучшее приближение к фактическим температурным данным (сравните $R = 0,63$ и $0,57$; $q = 2,58$ и $2,68$ %; стандартная ошибка одинакова и составляет $\sigma = 0,65$ °C).

Модель с участием системы из четырех информативных компонентов СПС получила выражение

$$T_{июля}^0 = 17,63 + 0,002TR - 0,009Psch + 0,01Chen + 0,007Art. \quad (4.7)$$

Здесь $R = 0,56$, $\sigma = 0,68$ °C, $q = 2,81$ % (сравните с $R = 0,53$, $\sigma = 0,66$ °C и $q = 2,83$ % для (4.4)). Видим, что использование вместо линейных комбинаций исходных компонентов СПС (главных компонент) непосредственно самих компонентов СПС не ухудшило статистические показатели моделей, а даже несколько улучшило. Поэтому делаем вывод, что к информативным системам компонентов СПС, выделенным методом главных компонент, можно непосредственно применять множественный регрессионный анализ и получать удобные для расчета уравнения с хорошими статистическими показателями.

4.4. Сравнительный анализ полученных расчетных моделей оценки температуры июля по палинологическим данным

В результате построений, проведенных по каждой из трех методик, оказалось, что в модели вошло одно и то же число компонентов СПС — четыре, представляющих собой информативные системы. Обращает на себя внимание небольшой списочный состав. Он включает всего семь компонентов СПС (из 28-ми рассмотренных). Причем только пыльца злаковых гр. *Cerealia* и осоковых принимает участие в одной модели, а остальные (общее содержание в спектре пыльцы травянистых и спор, пыльцы ели Шренка, маревых и полыней) входят в две модели из трех (табл.7).

В числе выбранных оказались такие компоненты СПС, которые отражают господствующий в районе исследования тип растительности (полупустынный) и его качественный состав (маревые, полыни, зла-

Таблица 7

Статистика компонентов СПС, выбранных разными методами
для их участия в расчетных моделях

№ п/п	Компонент спорово-пыльцевого спектра, %	Математический метод			Всего
		Пошаговый регрес- сионный анализ	Метод главных компонент + регрессионный анализ	Целевая итерационная классификация + рег- рессионный анализ	
1	Пыльца травянистых рас- тений и кустарничков		+	+	2
2	Споры	+		+	2
3	Ель Шренка	+	+		2
4	Злаковые гр. <i>Cerealia</i>	+			1
5	Осоковые	+			1
6	Маревые		+	+	2
7	Полыни		+	+	2
Всего		4	4	4	

ковые). Примечательно, что при этом двумя разными методами выбрана и пыльца ели Шренка (индикатор изменения увлажненности климата), которая в настоящее время произрастает в горах. Следовательно, применяя перечисленные методики, можно рассчитывать на объективное выявление информативных компонентов СПС, а выбор в этом качестве одних и тех же компонентов СПС разными методами рассматривать как верификацию проводимых математических построений.

Прежде чем отдать предпочтение одной из трех полученных моделей, рассмотрим качественный состав соответствующих информативных систем компонентов СПС и отметим особенности процесса их формирования каждым методом. Для выявления сравнительного веса компонентов СПС в этих системах построены графики (рис.20). Так как определение информационного веса внутри системы предусмотрено только в одной из рассмотренных методик – основанной на выборе информативной системы компонентов СПС методом целевой итерационной классификации, то для аналогичной оценки воспользуемся в случае пошагового регрессионного анализа долей вклада компонента СПС в изменение коэффициента множественной корреляции, а в случае метода главных компонент – нормированными коэффициентами при компонентах СПС в соответствующей линейной комбинации.

Пошаговым регрессионным анализом выявлено, что среднемесячная температура имеет максимальный частный коэффициент корреляции с пылью ели Шренка, а затем – со злаковыми гр. *Cerealia*, с остальными (споры и пыльца осоковых) – довольно незначительный (см. рис.20,а).

Состав информативных систем, выбранных методом главных компонент и целевой итерационной классификацией, совпал по трем из четырех позиций – это содержание пылицы травянистых, полынй и маревых. Единственное отличие состоит в том, что в первом случае дополнительно оказалась выбранной пыльца ели Шренка, а во втором – содержание спор в спектре (в ряду информативности в обоих случаях они занимают одно и то же 3-е место). В последовательности распределения весов существуют различия. Информационные веса компонентов СПС в системе, построенной методом главных компонент, оказались почти равноценными (при некотором преимуществе у пылицы маревых и общего содержания пылицы трав); в системе, выбранной целевой итерационной классификацией, они дифференцированы:

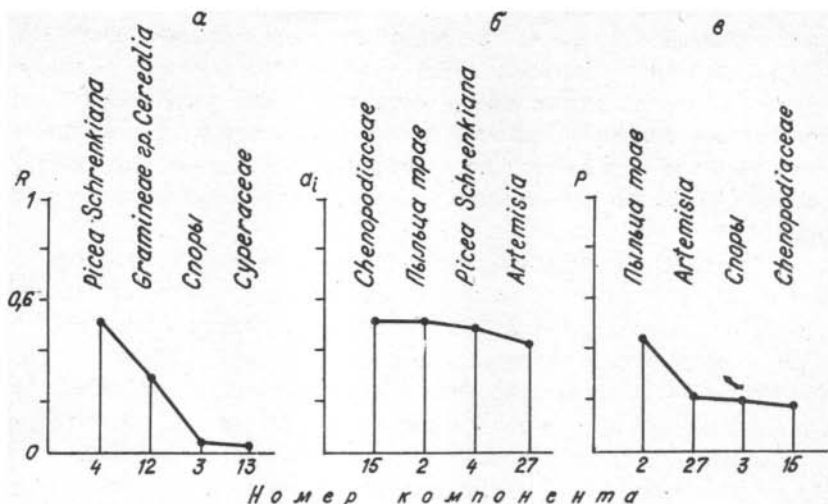


Рис.20. Характеристика информативности компонентов СПС, входящих в состав систем, полученных пошаговым регрессионным анализом (а), методом главных компонент (б) и целевой итерационной классификацией (в)

52 % приходится на содержание пылицы трав в спектре, а остальные 48 % распределились между тремя компонентами СПС – пылью полыни и маревых, содержанием спор (см. рис.20,б,в).

Для наглядности сведем в табл. 8 модели, построенные тремя методиками, и их статистические оценки, а на графике изобразим приближение расчетных значений к фактическим температурам июля (рис.21). Проведенное сравнение материалов показало, что наилучшее приближение к фактическим данным имеет модель, построенная пошаговым регрессионным анализом: коэффициент множественной корреляции максимален, а средняя квадратическая ошибка минимальна. Результаты, полученные методом главных компонент и по методике, основанной на выборе информативных компонентов СПС /Букреева и др., 1984,1986/, во многом совпали. Однако следует отдать предпочтение модели, полученной по системе информативных компонентов СПС, выделенной целевой итерационной классификацией. Дело в том, что этот метод распознавания образов, привлекаемый в данной ме-

Уравнения регрессии
для расчета среднемесячной температуры июля,
построенные по разным методикам

Модель	Уравнение регрессии	R	$\sigma_{\text{ср}}^2$	Q, %
(4.1)	1. Пошаговый регрессионный анализ $T_{\text{июля}}^{\circ} = 18,87 - 0,03SP - 0,01PSch - 1,14GrC + 0,09Cup$	0,89	0,36	1,5
(4.4)	2. Метод главных компонент + регрессионный анализ $T_{\text{июля}}^{\circ} = 17,6 + 0,01216U_1 = 17,6 + 0,01216(0,52236TR - 0,50731PSch + 0,52408Chen - 0,044172Art).$	0,53	0,66	2,83
(4.5)	3. Целевая итерационная классификация + регрессионный анализ $T_{\text{июля}}^{\circ} = 15,67 + 0,016TR + 0,02SP + 0,2Chen + 0,003Art.$	0,44	0,74	3,1

Примечание. Здесь R – коэффициент множественной корреляции; σ – средняя квадратическая ошибка; Q – усредненная ошибка, %. Район исследования – Иссык-Кульская котловина.

тодике для нахождения информативной системы признаков, не предназначен для получения наибольшей точности решения (хотя это и служит одним из оптимизирующих критериев), главное – получить наиболее устойчивое решение. При выборе информативных компонентов СПС учитывается как индивидуальная целевая значимость компонентов СПС, так и их дополнительность в сочетаниях относительно получения по спорово-пыльцевым спектрам обучающей выборки прогнозных оценок, близких или совпадающих с оценками по исследуемому элементу климата. Второй критерий зачастую не позволяет обеспечить наибольшую точность решения, зато совместно с первым помогает находить информативную систему компонентов СПС, обеспечи-

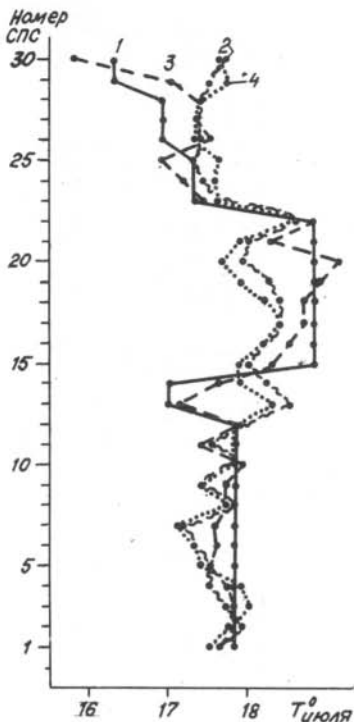


Рис.2I. Сопоставление фактических (I) данных о температуре июля в Иссык-Кульской котловине с расчетными, полученными: по методу главных компонент – рассматривается модель с участием первой главной компоненты – линейной комбинации 4-х компонентов СПС (2); по методу пошагового регрессионного анализа (3) и по методике, основанной на выявлении информативных систем компонентов СПС методом целевой итерационной классификации (4)

вающую более надежное (устойчивое) решение прогнозной задачи. Регрессионная модель, построенная с учетом компонентов СПС, входящих в эту систему, обеспечивает достаточно устойчивую взаимосвязь между ними и значениями элемента климата.

Способ формирования существенного набора компонентов СПС методом пошагового регрессионного анализа имеет тот недостаток, что выбор компонента СПС определяется исключительно величиной частного коэффициента корреляции между исследуемым элементом климата и конкретным компонентом СПС. При этом не учитывается влияние на эту связь других компонентов СПС (а она существует), а самое главное – не учитывается до ввода в уравнение регрессии взаимосвязь (свойство дополненности) выбранного компонента СПС с другими компонентами, уже включенными в число существенных.

Этот вывод подтверждается и анализом набора компонентов СПС, вошедшего в уравнение. В одном случае – это пыльца ели Шренка, культурных злаков, осоковых и содержание спор, в другом – содержание пылицы травянистых растений и спор, полыней и маревых. Все компоненты СПС, входящие во вторую систему, статистически обес-

печены материалом обучающей выборки. Так, содержание пыльцы травянистых растений и спор является обязательной характеристикой спектра и служит одним из самых важных и контролирующих его показателей. Несомненно, что и пыльца таких растений, как полыни и маревые, наиболее показательна для района полупустынь и сухих степей, откуда в основном и были отобраны поверхностные пробы в Иссык-Кульской котловине.

В то же время статистическая обеспеченность компонентов СПС, выбранных пошаговым регрессионным анализом, оказалась недостаточной. Так, пыльца ели Шренка встречается далеко не во всех спектрах и зачастую присутствует в виде единичных зерен при незначительном содержании пыльцы древесных пород в спектре. Но так как, исходя из требований к представлению исходных данных, каждый из спектров должен быть охарактеризован одним и тем же набором компонентов СПС, приходилось рассчитывать процентное содержание этого вида пыльцы по малому числу зерен. Пыльца культурных злаков и осоковых также не присутствует во всех пробах. Скорее всего, это результат локальных захоронений пыльцы указанных растений, произраставших на заливных лугах при распространении основного полупустынного типа растительности.

Таким образом, несмотря на то, что модель пошагового регрессионного анализа показала наилучшую сходимость, нет уверенности в том, что она является устойчивой закономерностью. Ее применение в целях прогноза может дать существенные ошибки.

Метод главных компонент позволяет производить существенную минимизацию исходного признакового пространства, оставляя для построения модели сильные признаки, которые несут основную информацию об исследуемой совокупности. Однако при построении модели методом главных компонент анализ набора всех спорово-пыльцевых спектров замкнут на себя, а не на элемент климата. Главные компоненты, полученные по исходной выборке, каждый раз будут одними и теми же для разных элементов климата. Поэтому, если захотим по данной исходной палинологической совокупности построить регрессионные модели оценки других климатических показателей, например среднегодовых осадков, увлажненности и т.д., то в них войдут те же четыре компонента СПС, что и для температуры июля. А в случае выбора информативной системы методом целевой итерационной классификации состав систем в каждом случае окажется ка-

чественно иным в зависимости от элемента климата, для которого ищется модель. Поэтому не исключено, что при построении модели методом главных компонент может быть не учтена дополнительная существенная информация.

* * *

Модели, получаемые пошаговым регрессионным анализом, показывают хорошее приближение к фактическим данным. Однако отсутствует уверенность, что они обеспечивают нахождение устойчивой закономерности.

Модели, построенные методом главных компонент, имеют большую устойчивость по сравнению с моделями пошаговой регрессии, так как главные компоненты вбирают в себя максимум изменчивости, существующей в материале обучения. Однако, чтобы не получать громоздких по своему составу расчетных уравнений, необходимо предварительно производить этап минимизации исходной совокупности способом, заложенным в самом методе. Метод главных компонент служит эффективным средством сокращения исходного признакового пространства. К полученной информативной системе компонентов СПС можно непосредственно, минуя линейные комбинации — главные компоненты, применять множественный регрессионный анализ и получать расчетные модели, мало уступающие по своим статистическим показателям.

Модели, формируемые по системе информативных компонентов СПС, выбранной при помощи целевой итерационной классификации, отличаются устойчивой закономерностью, обусловленной принципом их выбора путем поощрения и наказания относительно исследуемого элемента климата.

Наличие нескольких методик количественной оценки элементов палеоклимата по палинологическим данным расширяет возможности построений, позволяя получать различные варианты моделей и выбирать лучшую из них. Особенно это необходимо при поиске регрессионных зависимостей для таких климатических показателей, как среднегодовые осадки и среднемесячная температура января. В этом случае, как правило, возникают трудности. На этом фоне легко и

быстро строятся расчетные модели оценок среднемесячной температуры июля и среднегодовой температуры. Объяснение этому, видимо, содержится в том, что продуцируемый качественный и количественный состав пылицы и спор находится в прямой зависимости от летних температур. Поэтому в спорово-пыльцевых спектрах и заложен максимум информации об этих температурных показателях климата.

В заключение следует добавить, что все три перечисленные методики универсальны. Они могут быть использованы при построении расчетных уравнений для количественной оценки элементов палеоклимата с использованием данных по любым другим микрофоссилиям, а также в других областях знаний, когда для решения задачи имеется многомерная таблица данных и требуется найти регрессионную модель для зависимой переменной по множеству независимых переменных.

ЛИТЕРАТУРА

АЛЕКСАНДРОВ В.В., ГОРСКИЙ Н.Д. Алгоритмы и программы структурного метода обработки данных. Л.: Наука, 1983. 208 с.

АНДЕРСОН Т. Введение в многомерный статистический анализ. М.: Физматгиз. 1963. 374 с.

БАГРОВ Н.А. Аналитическое представление последовательности метеорологических полей посредством естественных ортогональных составляющих // Тр. ЦИП. 1959. Вып.74. С.3-24.

БЕЛОНИН М.Д., ГОЛУБЕВА В.А., СКУБЛОВ Г.Т. Факторный анализ в геологии. М.: Недра, 1982. 269 с.

БУКРЕЕВА Г.Ф. Применение математических методов в палинологических исследованиях // Изв. ВГО. 1987. Т.119. Вып.5. С.427-433.

БУКРЕЕВА Г.Ф. Решение некоторых задач палинологии математическими методами // Палинология в СССР. Новосибирск: Наука. 1988. С.157-160.

БУКРЕЕВА Г.Ф., ВОТАХ М.Р., БИШАЕВ А.А. Методика определения палеоклимата по информативным системам признаков спорово-пыльцевых спектров // Геол. и геофиз. 1984. № 6. С.16-28.

БУКРЕЕВА Г.Ф., ВОТАХ М.Р., БИШАЕВ А.А. Определение палеоклиматов по палинологическим данным методами целевой итерационной классификации и регрессионного анализа. Новосибирск: Изд. ИГиГ. 1986. 190 с.

ВЕЛИЧКО А.А. Эмпирическая палеоклиматология (принципы и степень точности) // Методы реконструкции палеоклиматов. М.: Наука. 1985. С.7-20.

ВЕКЛИЧ М.Ф. Проблемы палеоклиматологии. Киев: Наук. думка. 1987. 192 с.

ГЕЛЕТА И.Ф., СПИРИДОНОВА Е.А. Опыт восстановления климата голоцена по данным палинологии методами многомерного статистического анализа // Палинология плейстоцена и голоцена. Л., 1981. С.52-70.

ДЕЧ В.Н., КНОРИНГ Л.Д. Методы изучения периодических явлений в геологии. Л.: Недра. 1985. 255 с.

ЕЛИСЕЕВА И.И., РУКАВИШНИКОВ В.О. Логика прикладного статистического анализа. М.: Финансы и статистика. 1982. 192 с.

ЙЕРЕСКОГ К.Г., КЛОВАН Д.И., РЕЙМЕНТ Р.А. Геологический факторный анализ. Л.: Недра. 1980. 223 с.

КЛИМАНОВ В.А. К методике восстановления количественных характеристик климата прошлого // Вестн. МГУ. Сер. географ. 1976. № 2. С.92-98.

КРАМБЕЙН У., КАУФМАН М., МАК-КЕММОН Р. Модели геологических процессов. М.: Мир. 1973. 143 с.

МЕТОДЫ РЕКОНСТРУКЦИИ ПАЛЕОКЛИМАТОВ. М.: Наука. 1985. 198 с.

МИЛЛЕР Р.Л., КАН Дж.С. Статистический анализ в геологических науках. М.: Мир. 1965. 481 с.

МУРАТОВА М.В., БОЯРСКАЯ Т.Д., ЛИБЕРМАН А.А. Применение теории вероятностей для восстановления палеоклиматических условий по данным палинологического анализа // Новейшая тектоника, новейшие отложения и человек. Сб. № 3. М., 1972. С.239-246.

ОБУХОВ А.М. Нормальная корреляция векторов // Изв. АН СССР. Отд. матем. и естеств. наук. 1938. № 3. С.339-370.

ОЗЕРО Иссык-Куль и тенденции его природного развития. Л.: Наука, 1986. 256 с.

ПАКЕТ научных подпрограмм на языке ПЛ/I. Математическое обеспечение ЕС ЭВМ. Вып.2. Статистика. Таллин: 1981. 144 с.

СМИРНОВ Б.И., ЩЕРБАК А.А., АГЕЕВ В.А. Геохимический метод обнаружения седиментационной цикличности // Прикладные вопросы седиментационной цикличности и нефтегазоносности. Новосибирск: Наука. 1987. С.200-208.

СКЛЯРЕНКО В.Л., СМОРНОВ К.П. О применении многомерного анализа в гидрологии // Факторы формирования водных масс и районирование внутренних водоемов. Л.: Наука. 1974. С.180-205.

ЯСАМАНОВ Н.А. Древние климаты Земли. Л.: Гидрометеиздат, 1985. 292 с.

THOMPSON W., CLARK D.R. Calibrating micropaleontological date in climatic terms: a critical review // Annals of the New York Academy of Sciences. 1977. Vol. 288. P.93-118.

RAO C.R. The use and interpretation of principal component analysis in applied research // Sankhya. 1964. Vol.26. P.329-358.

ОГЛАВЛЕНИЕ

В В Е Д Е Н И Е	3
Глава I. ПОСТАНОВКА ЗАДАЧИ И ТРЕБОВАНИЯ К ИСХОДНЫМ ДАННЫМ	6
Глава II. ПОСТРОЕНИЕ РЕГРЕССИОННЫХ МОДЕЛЕЙ КОЛИЧЕСТВЕННОЙ ОЦЕНКИ ЭЛЕМЕНТОВ ПАЛЕОКЛИМАТА ПОШАГОВЫМ РЕГРЕССИОН- НЫМ АНАЛИЗОМ	9
2.1. Общие сведения	9
2.2. Исходные данные	II
2.3. Алгоритм определения параметров регрессионных моделей пошаговым регрессионным анализом	II
Глава III. ПОСТРОЕНИЕ РЕГРЕССИОННЫХ МОДЕЛЕЙ КОЛИЧЕСТВЕННОЙ ОЦЕНКИ ЭЛЕМЕНТОВ ПАЛЕОКЛИМАТА МЕТОДОМ ГЛАВНЫХ КОМ- ПОНЕНТ	I7
3.1. Общие сведения	I7
3.2. Исходные данные	25
3.3. Алгоритм определения параметров регрессионных моделей методами главных компонент и регресси- онного анализа	26
Глава IV. ПРИМЕРЫ ПОСТРОЕНИЯ УРАВНЕНИЯ РЕГРЕССИИ ДЛЯ КОЛИЧЕСТВЕННОЙ ОЦЕНКИ ЭЛЕМЕНТОВ ПАЛЕОКЛИМАТА ПО ПА- ЛИНОЛОГИЧЕСКИМ ДАННЫМ	35
4.1. Определение параметров расчетной модели для оценки температуры июля пошаговым регрессионным анализом	36
4.2. Определение параметров расчетной модели для оценки температуры июля методами главных ком- понент и регрессионного анализа	44
4.3. Регрессионная модель оценки температуры июля, построенная по методике количественной оценки элемента климата по информативной системе ком- понентов спорово-пыльцевых спектров	84
4.4. Сравнительный анализ полученных расчетных моде- лей оценки температуры июля по палинологическим данным	87
ЛИТЕРАТУРА	96

МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ПАЛИНОЛОГИИ

Вып. 2

Количественная оценка элементов палеоклимата
методом главных компонент и пошаговым регрессионным анализом
(по палинологическим данным)

Методические рекомендации

Составитель

Галина Фёдоровна Букреева

Редактор Р.Н.Ильина

Технический редактор Н.Н.Александрова

Подписано к печати 21.03.89. МН III36
Бумага 60x84/16. Печ.л.6,25. Уч.-изд.л.5,5.
Тираж 400. Заказ 173. Цена 40 коп.

Институт геологии и геофизики СО АН СССР
Новосибирск, 90. Ротапринт.