ПОЧВОВЕДЕНИЕ, 2011, № 12, с. 1411–1424

ГЕНЕЗИС И ГЕОГРАФИЯ ПОЧВ

УДК 631.4

ФОРМАЛЬНЫЙ АППАРАТ КЛАССИФИКАЦИИ ПОЧВ*

© 2011 г. В. А. Рожков

Почвенный институт им. В.В. Докучаева РАСХН, 119017, Москва, Пыжевский пер., 7
e-mail: rva39@mail.ru

Поступила в редакцию 20.12.2010 г.

Обсуждается математический аппарат основных разделов классификации почв: оценки информативности почвенных признаков; группировки почвенных объектов на заданное и автоматически определяемое число классов; оптимизации качества классификаций; построение моделей и решающих правил классифицирования (распознавания) почвенных объектов. Приведены алгоритмы многомерных статистических методов и кластер-анализа, реализованных в пакете прикладных программ, широко апробированных в решении конкретных задач из перечисленных разделов. Акцент сделан на разъяснении малоиспользуемых методов создания систем информативных признаков объектов и классов, оценке качества и сравнения классификаций. Приведены примеры и источники опубликованных результатов решения отдельных и комплекса задач из области классификации почв. Теоретические и практические проблемы классификации в науке не могут выйти из области субъективных личных или корпоративных построений и схем без приложения современных средств информатики и в первую очередь математических методов. Описанная структура задач численной классификации и сопровождающих их программных средств послужит каркасом для создания не только формальной, но и содержательной теории классификации почв.

ВВЕДЕНИЕ

Внедрение математических методов в биологию началось с публикации 1763 г. работы М. Адансона “Семейства растений” [7], в которой он изложил основы классификации растений по частоте совпадений характерных признаков растений или их таксонов. В классификации почв широко применяется математическая статистика, но сравнительно мало используются многомерные методы, хотя технические дисциплины могут предоставить эффективные алгоритмы распознавания образов, кластер-анализа, многомерной статистики. С 60-х годов прошлого столетия вошла в употребление численная классификация (**ЧК**). Теории и методам ЧК посвящена фундаментальная монография “Numerical Taxonomy” [35]. Позже появились обобщающие монографии по проблемам ЧК в почвоведении [21, 32, 36] и многочисленные статьи в мировой научной периодике. Следует отметить возможности, во многом уже реализованные, привлечения достижений в области ЧК из других дисциплин, в том числе биологии и геологии.

Проблемы классификации почв бурно обсуждались в 60–70-х годах, в наше время больше публикаций по почвенным корреляциям. Приводятся сводки параллельных списков названий почв и их аналогов в классификациях разных авторов и стран. Однако, как было давно сказано, главное не в списках, а в необходимости разработки научных принципов почвенных классификаций разного назначения. Классификации могут меняться, а принципы останутся.

Все более широко признается необходимость внедрения методов математики в классификации почв. В популяризации нового подхода, называемого ЧК, почвоведы ссылаются на выдвинутый Уайтхедом еще в 1925 г. тезис: “Классификация необходима. Но если вы не можете перейти от классификации к математике, ваши рассуждения не продвинут вас далеко” [7]. Уместно вспомнить и высказывание Бейли (цит. по [3]) относительно классификаций, не имеющих количественной основы: “Хотя они, несомненно, имеют известную ценность, истинное научное значение их сомнительно. Теория таких методов классификаций остается пока чисто интуитивной и является скорее искусством, чем наукой”. Содержание понятия ЧК довольно простое: “Оценка численными методами связей или сходства между таксономическими единицами и использование этих связей в построении их таксономического порядка” [35]. Под таксономическими единицами здесь понимаются объекты и их классы (таксоны). ЧК подразумевает “...разбиение некоторой совокупности многомерных объектов на классы, основанное только на том, что каждому классу соответствует обособленная группа точек в пространстве параметров” [2]. Таким образом, под ЧК следует понимать формальные (с применением математики) классификационные построения и классифицирование (распознавание) почвенных объектов, заданных их признаками.

* Выполнено при поддержке РФФИ грант № 11-04-01123а.

Существует логика классификации, формулируются концепции, требования к классификациям и их функциям, но не правила их построения. Однако можно утверждать, что накопленный опыт позволяет предложить последовательность этапов построения численной классификации почв с использованием апробированных методов. Для разработки теории классификации данная последовательность послужит каркасом, на который нужно нанести предметное (почвенное) содержание.

Интенсивно развивающиеся дистанционное зондирование Земли, цифровое картографирование и внедрение геоинформационных систем дали новый стимул применению математических методов [12, 25]. Однако высокий уровень технического оснащения и строгая унификация методов автоматизированной обработки соответствующих материалов ограничили потребность отрасли в детальном описании алгоритмов анализа данных. К тому же, многие методы оценки информативности показателей и качества цифровых картографических материалов отличаются от разработанных в почвоведении. Следовательно, перед почвоведением возникает перспектива не просто использовать достижения геоинформационных технологий, технического распознавания образов, но и, в свою очередь, предложить апробированные эффективные методы. Поэтому цель настоящей статьи – представить основные алгоритмы, на примерах показать их эффективность и наметить направление дальнейшего конструирования теории классификации почвенных объектов: агрегатов, шлифов, горизонтов, профилей, почв.

ФОРМАЛИЗАЦИЯ КЛАССИФИКАЦИИ ПОЧВ

Термин “*классификация*” имеет три значения: процесс создания таковой, описание классификации (результат разработки) и процедура ее использования при распознавании конкретной почвы. Последнее значение здесь будет заменено на “классифицирование”. Два других будут понятны из контекста.

Формально понятие классификации (ее логика), может быть представлено в теоретико-множественных терминах. Множество объектов A делится на классы: $A = A_j$, где $j = 1, 2, ..., k$ и k – число классов; $A_j \neq \emptyset$, то есть все классы непустые (содержат хоть один объект); $A_j \cap A_l = \varnothing$, где $i, j = 1, 2, ..., k$ и $j \neq l$, то есть классы не пересекаются (не имеют общих объектов); $A_j = A$ – объединение классов составляет исходное множество.

Отдельные классы называются классами эквивалентности. Отношение эквивалентности обладает свойствами рефлективности (xRx), симметричности ($xRy \longrightarrow yRx$) и транзитивности (xRy & $yRz \longrightarrow xRz$), где R – некоторое отношение – сходства, различия, подобия и др. Следовательно, классификация – есть система классов эквивалентности.

Типологическое районирование также характеризуется отношениями эквивалентности, но региональное районирование не обладает свойством транзитивности: (xRy & yRz) уже не означает (xRz), то есть “сосед моего соседа” не обязательно является моим соседом (отношение толерантности). Это своего рода каноническое определение классификации.

Начальным шагом создания любой классификации является формулировка ее цели. Цель может быть лишь конкретной и непротиворечивой, а не философской и всеобъемлющей. Для этого из многообразия почвенных показателей отбираются такие, которые имеют непосредственное отношение к поставленной цели и наилучшим образом отражают связанные с ней представления. Исходная субъективность выбора цели может быть ограничена только четкостью ее формулировки, точнее, адекватностью концептуальной модели объекта классификации структуре раскрывающих ее показателей. Исключение многообразия аспектов, ради которых строится классификация, – главная гарантия успеха в ее создании [27]. В этом *отличие классификации от базы данных*: она включает информацию, высоко релевантную узкому запросу, а база данных предназначена объять как можно более широкий круг аспектов фактических данных для удовлетворения многих запросов. Иными словами, классификация по существу является *компактной информационной системой*, которая содержит максимум информации о классах почвенных объектов (почв, горизонтов, образцов, шлифов, агрегатов и др.) в принятом пространстве признаков. Под “признаком”, атрибутом или показателем здесь понимаются свойства, состав, строение почв, факторное или смешанное описание почвы.

Значения почвенных признаков могут представляться в различных шкалах, теория и приложения которых рассмотрены в теории измерений [14] и ряде конкретных дисциплин [8]. В основу выделения шкал положен принцип допустимости преобразований значений признаков, с которым связано определение допустимых арифметических операций на признаках и методов обработки. Понятие допустимости преобразования тесно связано с прогнозированием проявления одних признаков по другим: если оно не нарушает такого прогноза, то оно допустимо. Для отдельного признака допустимость означает, что преобразование не изменяет его основных свойств, в частности результаты некоторых арифметических

Таблица 1. Шкалы значений признаков

Шкала	Допустимые в данной шкале							Примеры
	преобразования*	операции**					статистическая обработка	
		1	2	3	4	5		
Номинальная (наименований, классификационная)	Взаимнооднозначные	+	–	–	–	–	1) распределение частот, 2) определение модального класса	Цвет, структура, названия почв и горизонтов, форма границ
Порядка (ординальная)	Монотонные непрерывные	+	+	–	–	–	1, 2, 3) оценка медианы, 4) центилей, 5) ранговая корреляция	Степень оподзоленности, окультуренности, влажность, плотность
Интервалов	$y(x) = ax + b$ $a > 0$	+	+	+	–	–	1–5, 6) оценка математического ожидания, 7) дисперсия, 8) ассиметрия, 9) моменты	Температура, абсолютный возраст
Разностей	$y(x) = ax + b$ $a = 1$	+	+	+	+	–	1–9	Определяемые по разности в сумме показатели
Отношений	$y(x) = ax \;\; a > 0$	+	+	+	–	+	Все возможные	Глубины, мощности
Абсолютная	$y(x) = x$ $a = 1$	+	+	+	+	+	Все возможные	Количество образцов, горизонтов

* Допустимые изменения значений признаков и их использование в качестве аргументов уравнений.

** 1 – равно (=) или неравно (≠); 2 –больше (>) – меньше (<); 3 – $(x_1 - x_3)/(x_2 - x_3)$; 4 – $(x_1 - x_2)$; 5 – x_1/x_2, где x_1 – значения признака.

операций и отношения значений признака сохраняются.

В табл. 1 приведены типы шкал, исчерпывающие разнообразие значений почвенных признаков, даны их общая характеристика и примеры. Типы почвенных признаков достаточно широко обсуждалась в печати [9, 17, 21], что исключает необходимость останавливаться на таблице детально. Важно лишь подчеркнуть, что учет шкал необходим при выборе правильного метода математической обработки почвенных данных.

Информативность признаков в классификации означает способность отделять один объект и/или класс от других. Оценка информативности может проводиться в ситуации, когда классы объектов неизвестны или когда заданы. В первом случае для многомерной выборки $X = x_{ij}$ (где $i = 1, ..., n$ и n – число объектов; $j = 1, ..., m$ и m – число признаков) эффективным средством служит метод главных компонент [17, 24]. Собственные числа (λ_{lj}) и собственные векторы (υ_{lj}, $l = 1, ..., k$ – число найденных чисел и векторов) корреляционной матрицы такой выборки. Метод иллюстрируется следующим примером.

Описание почвенного профиля подзолистой почвы включило по три образца из гор. А пах, А2, А2В и В, описанных семью признаками: pH, содержания углерода, кислотность, содержание ила и глины, вынос ила и вынос Ca + Mg. Фрагмент результата расчетов показан на рис. 1.

Первая главная компонента описывает 67% варьирования признаков в многомерной выборке. Наибольшие значения в первом собственном векторе соответствуют четырем последним признакам почвы, то есть гранулометрического состава горизонтов. Следовательно, именно они определяют разброс горизонтов по оси ординат (проекции горизонтов на первую главную компоненту). Наибольший вес по второй компоненте (0.73) имеет значение содержания углерода – оно характеризует разброс горизонтов по оси абсцисс.

Для дальнейшего анализа можно оставить только наиболее весомые показатели, что неоднократно проверено в экономических задачах [1], хотя при этом может происходить потеря информации. Критерием качества может служить визуально обнаруживаемое качество разделения объектов на рисунке. Здесь оно вполне адекватно: четко группируются гор. А пах (образцы с номерами 1–3) и А2 (4–6), и смешиваются гор. А2В (1–9) и В (10–12). Это более наглядно видно на дендрограмме этих объектов (рис. 2).

Представленная на рисунке группировка горизонтов по сходству аналогична таковой в координатах главных компонент. Рис. 1 и 2 иллюстрируют ординатную и иерархическую классификации. Дендрограмма является одним из наглядных средств визуализации результатов классификации. При ее вычислении особенно важно учитывать шкалу значений признаков, которая определяет выбор меры сходства [21].

Корреляционная матрица свойств дерново-подзолистй почвы:

1	:	(	2)	0.50	(	3)	−0.47	(	4)	−0.68	(	5)	−0.49	(	6)	−0.70	(	7)	−0.48		
2	:	(	3)	0.26	(	4)	−0.50	(	5)	−0.41	(	6)	−0.50	(	7)	−0.34					
3	:	(	4)	0.55	(	5)	0.49	(	6)	0.52	(	7)	0.41								
4	:	(	5)	0.93	(	6)	0.98	(	7)	0.84											
5	:	(	6)	0.93	(	7)	0.85														
6	:	(	7)	0.85																	

i	Собственные числа матрицы	Нагрузки главных компонент, %
1	4.69	67.0
2	1.26	85.0

Собственные векторы матрицы:

ГК1–	1	:	−0.34	−0.24	0.26	0.45	0.43	0.45	0.40
ГК2–	2	:	−0.09	−0.73	−0.67	−0.01	−0.04	−0.02	−0.04

Положение образцов в координатах главных компонент:

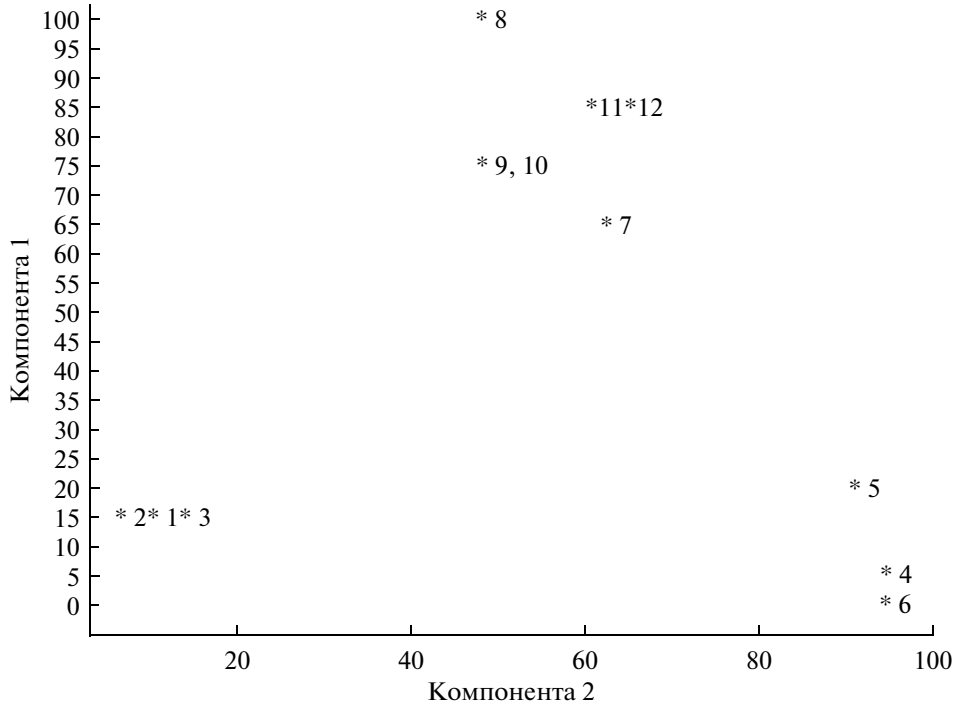

Рис. 1. Результаты обработки 12 горизонтов дерново-подзолистой почвы по 7 признакам методом главных компонент: 1–3 образцы гор. A1; 4–6 – гор. A2; 7–9 – гор. A2B; 10–12 – гор. B.

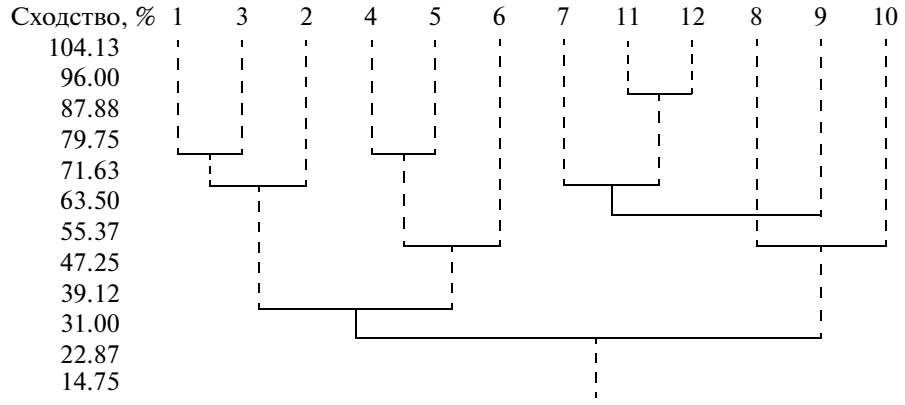

Рис. 2. Дендрограмма описанных выше 12 горизонтов по 7 признакам (номера горизонтов на рис. 1).

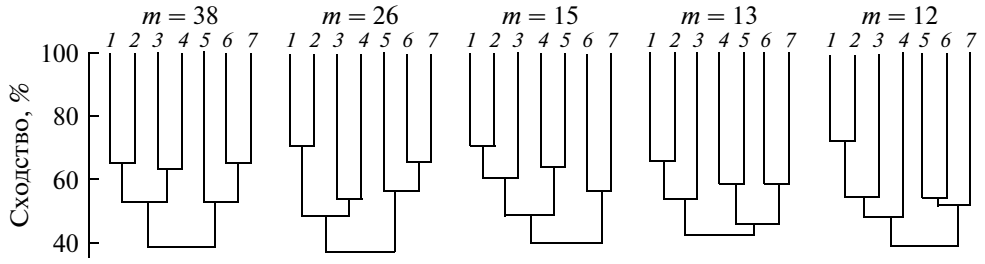

Рис. 3. Оценка информативности признаков по дендрограмме (семь произвольных объектов).

Дендрограмма также может использоваться для визуальной оценки информативности признаков, как это показано на рис. 3. Дендрограмма, построенная для произвольных объектов по 38 признакам, не изменяется при исключении малоинформативных до тех пор, пока не останется всего 12 из них. Однако для больших выборок такой подход может быть заменен более строгим сравнением иерархических структур.

Определение и исключение мало информативных признаков, когда нет информации о классах объектов, производится не очень строгими методами. Например, вариация значений признаков: низкая вариация означает слабую относительную информативность и наоборот. Довольно распространенная оценка – корреляция. Из двух сильно коррелирующих признаков можно оставить один, так как другой не добавляет информации. Контролем допустимого исключения признаков служат дендрограммы, подобные вышеприведенным. Дендрограмма непосредственно самих признаков (используется транспонированная матрица исходной выборки наблюдений) также может дать информацию об их относительной информативности. При обнаружении высокого сходства каких-либо признаков, можно исключить некоторые из них аналогично корреляции. Преимуществом метода дендрограмм является возможность вычислять и строить их в любой шкале признаков, не только в арифметической.

Более строгие критерии информативности можно применить при заданных (экспертно или формально) классах объектов. При достаточных объемах выборок оценка информативности признаков осуществляется методами многомерной статистики [ttp://lem.edu.mhost.ru/doc/presentations/Rozhkov.pdf; 21]. Для этого число объектов в классах должно быть больше числа описывающих признаков. Потеря информации при их исключении осуществляется сопоставлением сходства классов объектов по полному (p) и сокращенному (q) набору показателей:

$$\chi_f^2 = -n(p+k)/2\ln(\lambda q/\lambda p),$$

где $f = p(k-1)$, $n = n_1 + n_2$ – число объектов в классах, k – число классов; $\lambda q = |W|/|T|$ – отношение определителей матриц внутри- и межклассового варьирования (сходства) с p и q признаков.

Сходные результаты дает сравнение расстояний Махаланобиса между классами по полному и сокращенному набору признаков:

$$F = \left[(n_1 n_2 - 1)\, n_1 n_2 \left(D_p^2 - D_q^2\right)\right] \Big/ \left[(q-p)(n_1+n_2) \times \right.$$
$$\left. \times (n_1 + n_2 - 2) + n_1 n_2 D_p^2\right],$$

где F – распределение с $f_1 = q - p$ и $f_2 = n_1 + n_2 - q - 1$ степенями свободы. Здесь n_1 и n_2 число объектов в сопоставляемых классах, q и p – соответственно исходное и сокращенное число признаков ($q > p$); D_q^2 и D_p^2 – расстояния Махаланобиса. Если $F \le F_{\alpha f1, f2}$, это значит, что исключение признака не привело к потере к потере информации.

Расчеты выполняются путем циклического перебора всех признаков и выявления наиболее мало информативного. Он исключается из выборки и повторяется процедура поиска наименее информативного до тех пор, пока потеря информации (по критерию χ^2) не станет существенной. Исключение мало информативных показателей имеет и материальную основу – количество анализов можно сократить более чем вдвое.

Кроме исключения признаков сокращение пространства признаков и их значений производится *сверткой* данных до обозримых структур путем аппроксимации полиномами[1] распределений значений показателей почв по профилю [20, 21]:

$$P_m(x) = \sum_{k=0}^{m} a_k x^k, \quad \text{где } m - \text{степень полинома.}$$

Например, полином

$$P_2(w) = 32.0 - 0.123w + 0.000361w^2$$

[1] Степенной полином служит базисной функцией в различных формальных процедурах аппроксимации структур данных, регрессии, представлении линейных дискриминантных функций, линейных преобразований векторов признаков в главные компоненты, в канонической корреляции служит моделью в проектировании сложных технических систем и их самоорганизации [4, 12].

Таблица 2. Аппроксимация таблицы валового анализа иллювиально-гумусового подзола

Генетический горизонт	Глубина (x), см	Валовые, %	Анализ	Расчет по формуле	Разность	Разность, % от анализа
AoA1	5	SiO_2	46.0	56.0	−10.0	21
		Fe_2O_3	14.0	13.0	1.0	7
		Al_2O_3	6.5	5.0	1.5	23
A2	10	SiO_2	85.0	69.0	16.0	14
		Fe_2O_3	9.0	11.0	−2.0	22
		Al_2O_3	0.8	0.3	0.5	62
Bf	15	SiO_2	83.0	89.0	−6.0	7
		Fe_2O_3	9.4	8.6	0.8	8
		Al_2O_3	1.8	0.8	1.0	56

Коэффициенты полинома: $P_{2,2}(x, y) = 1.20 - 86.0x + 16.0x^2 + 11.0 - 9.5y + 1.9y^2$, где x – глубина, y – порядковый номер окисла.

довольно точно аппроксимирует наименьшую влагоемкость (w, %) по профилю дерново-среднеподзолистой почвы [20].

Таблицы типа "состав–свойства" $f(x, y)$ также аппроксимируются полиномами вида [15].

$$\mathbf{P}n,m(x,y) = \sum_{r=1}^{m+1} y^{r-1} \sum_{s=1}^{n+1} A_{(r-1)(n+1)} + x_s^{s-1},$$

где n – степень аппроксимирующего полинома по переменной x; $\underline{s}$ – число узлов аппроксимации по переменной x (число столбцов таблицы); m – степень полинома по переменной y; r – число узлов аппроксимации по переменной y; A – коэффициенты полинома.

В табл. 2 показана принципиальная возможность двумерной аппроксимации таблицы почвенных показателей. Можно было найти более точное приближение данных, применяя другие начальные показатели или другой алгоритм, но в данном случае иллюстрируется сама возможность экономной свертки исходных данных. Подход удобен при необходимости привести данные к однородному представлению (распределений по профилю, по катене) и получить возможность интерполяции. Параметры полиномов могут служить исходными данными для последующего анализа, что неоднократно использовалось в почвоведении [5, 13, 15, 17, 29, 30].

ЧК располагает средствами оценки качества и сравнения классификаций между собой. На рис. 4 приведен пример и формулы некоторых критериев качества разделения множества объектов на классы. Собранные из разных источников данные по содержанию гумуса и обменного Ca в гор. $A1$ семи почв представлены группировкой в верхней части рисунка. По специальной программе производилось разделение 27 образцов последовательно на 2, 3, ..., 12 классов (нижняя часть рисунка) с вычислением на каждом этапе 6 критериев качества разделения (KK1–KK6 – средняя часть рисунка). KK1 – отношение числа правильно установленных классов к общему их числу (то есть здесь исходные классы известны). Более интересны ситуации, когда такая информация отсутствует. Это все другие критерии: KK2 – разность среднего внутри- и межгруппового сходства образцов; KK3 – среднее внутригрупповое сходство; KK4 – отношение определителей матрицы внутри- и общего варьирования; KK5 – след матрицы среднего внутригруппового сходства; KK6 – определитель матрицы среднего внутригруппового сходства.

Первые три критерия дают сходные и более четкие результаты. Лучшему разделению, более всего приближающемуся к заданному разделению, соответствуют пять и близко к нему восемь классов. Но остальные критерии почти дают точный ответ, указывая именно семь классов. Тем не менее, наиболее распространен второй критерий качества, вычисляемый непосредственно в процессе разделения объектов.

На основании вычисляемого критерия качества можно сравнивать разные классификации. Однако имеются и специальные подходы. Для разных ординатных классификаций одних и тех же почв критерием сравнения служат коэффициент ассоциации и полихорический показатель связи [21, 31].

Алгоритм расчета критерия ассоциации ординатных классификаций. Исходные данные: k – число сопоставляемых классификаций $q(k)$ – число классов в них; $n(k, q_{max})$ – число объектов в классах; $m(k, n)$ – номера объектов по классам.

1. Строится таблица парной сопряженности классификаций (n_{ij}).

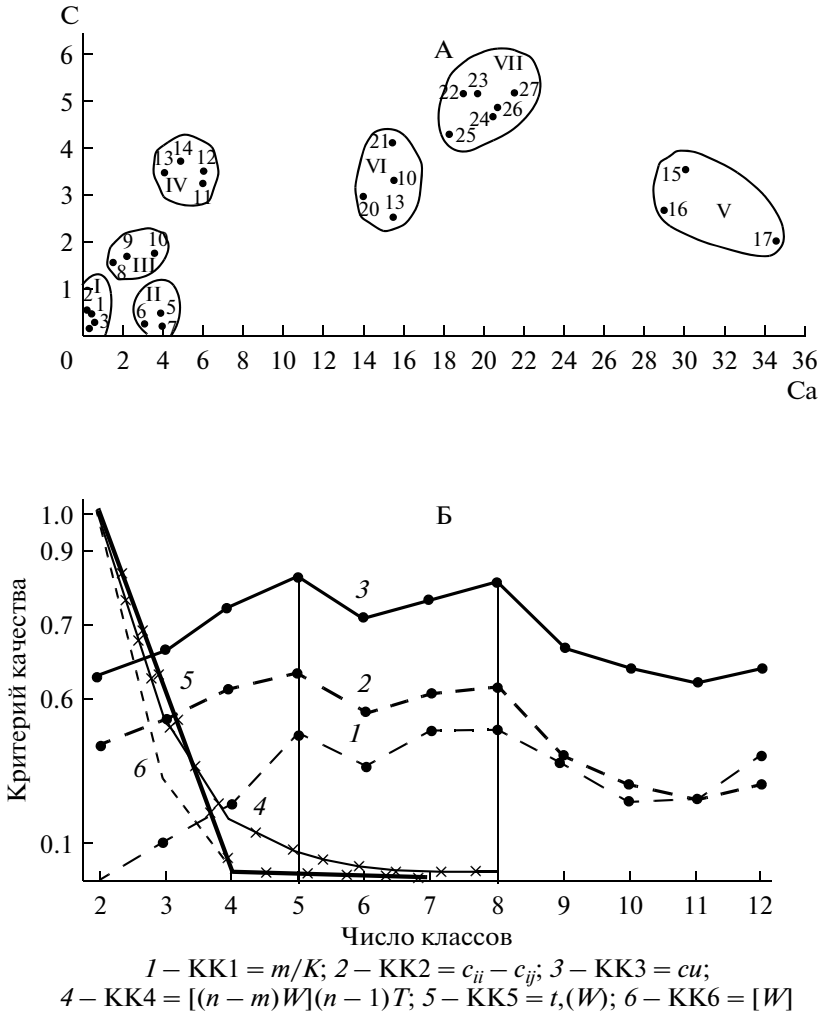

Рис. 4. Критерии качества классификаций (пояснения в тексте). А – ординатное представление групп почв: I – бурые пустынные; II – серо-бурые; III – дерново-среднеподзолистые; IV – дерново-сильноподзолистые; V – каштановые; VI – лесные светло-серые; VII – серые лесные. 1–17 номера профилей. Б – изменение качества классификации с увеличением числа групп.

2. Вычисляется коэффициент ассоциации (КА):

$$\text{КА} = (\omega_1 + \omega_2 - t/(2n - t),$$

где $\omega_1 = \sum_{j=1}^{k_1} m_{j\max}$; $m_{j\max} = \max(n_{ij})$;

$j = 1, k_1$ – число классов в первой классификации;

$$\omega_2 = \sum_{i=1}^{k_2} m_{i\max}; \quad m_{i\max} = \max(n_{ij});$$

$i = 1, k_2$ – число классов во второй классификации.

$$t = \max(m_j) + \max(m_i); \quad m_i = \sum_{j=1}^{k_1} n_{ij}; \quad m_j = \sum_{i=1}^{k_2} n_{ij},$$

Расчет полихорического показателя связи ординатных классификаций (исходные данные и п. 1 те же, что и в предыдущем алгоритме):

1. Вычисляется полихорический показатель (ПП) связи Чупрова:

$$\text{ПП} = \varphi k^{-0.25}, \quad \text{где } k = (k_1 - 1)(k_2 - 1),$$

где k_1 и k_2 – числа классов в двух классификациях соответственно.

$$\varphi^2 = \sum_{i=1}^{k_2} \left(1 \Big/ m_i \sum_{j=1}^{k_1} n_{ij}^2 \Big/ m_j \right) - 1 - k/n;$$

$$n = \sum_{j}^{i=1} \sum_{i} n_{ij} = \sum_{i}^{j=1} m_i = \sum_{j} m_j;$$

$$\chi^2_{\alpha,f} = n\varphi^2; \quad f = k.$$

Например, пусть для $n = 107$ объектов получено две классификации: в первой (I) все объекты раз-

Таблица 3. Число объектов, совпадающих в классификациях I и II, по классам*

Классификация II	Классификация I					
	1	2	3	$K_{1=4}$	m_i	$m_{i\max}$
1	20	–	–	–	20	20
2	8	–	–	–	8	28
3	–	13	4	–	17	13
4	–	1	4	–	5	4
5	–	–	14	28	42	28
6	–	–	–	–	2	2
7	–	–	5	2	7	5
8	1	–	1	–	2	1
9	2	–	–	–	2	2
$K_2 = 10$	–	–	1	1	2	1
m_i	31	14	31	31	$n = 107$	$\omega_2 = 84$
$m_{j\max}$	20	13	14	28	$\omega_1 = 75$	

* Объяснения показателей – в тексте.

делились на четыре класса, а во второй (II) – на 10. В табл. 3 приведены числа совпадающих объектов по классам обоих разбиений и выполнены необходимые суммирования.

В результате получаем значение КА = 75 + 84 – – (31 + 42)/2.107 – (31 + 42) = 0.61, то есть классификации I и II согласуются на уровне 61%.

Для полихорической связи имеем: $k = 27$; $\varphi^2 = 2.59$; ПП = 0.71; $\chi^2 = 277$, что указывает на высокую сопряженность этих классификаций ($\chi^2_{\text{табл}} = 55.5$ с вероятностью 0.999).

Сравнение иерархических классификаций проводится методом, предложенным Сокалом и Рольфом [34] и выполняется автоматически специальной программой сравнения дендрограмм На рис. 5 и в табл. 4–5 показан алгоритм сравнения двух дендрограмм вышеописанных 12 горизонтов, построенных по разному набору признаков (или с использованием различных показателей сходства).

Размах значений сходства на дендрограммах делится на одинаковое число интервалов, которое рекомендуется принимать равным 4 при числе объектов $n \le 10$ и более 10 – при $n \ge 100$.

Далее строятся матрицы сходства $cV_а$ и $cV_б$ размером $[n, n]$ (табл. 4).

Они указывают номера интервалов, в которые попадают значения сходства образцов по совокупности 7 свойств между собой. Так, для первого объекта дендрограммы (А) сходство с объектами 2–3 и 4–6 попадает соответственно в интервалы II и III,

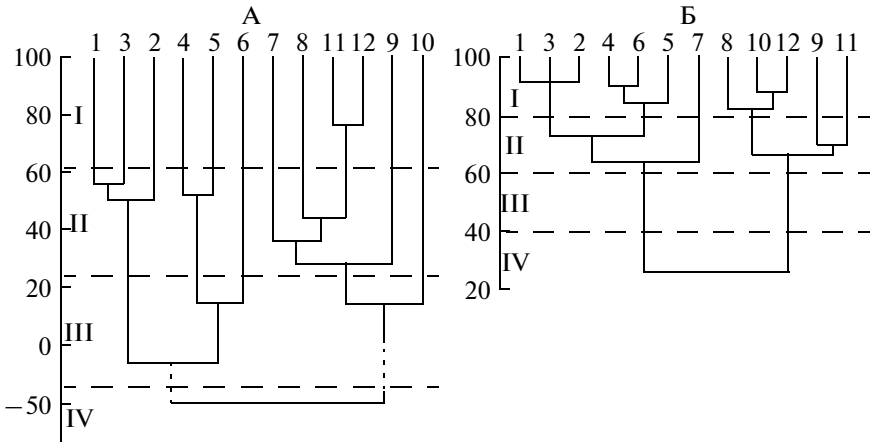

Рис. 5. Сравнение иерархических классификаций.

Таблица 4. Матрицы *cV*: нижний треугольник для дендрограммы *А*, верхний – для *Б* (см. рис. 5)

А/*Б*	1	2	3	4	5	6	7	8	9	10	11	12
1	–	1	1	2	2	2	2	4	4	4	4	4
2	2	–	1	2	2	2	2	4	4	4	4	4
3	2	2	–	2	2	2	2	4	4	4	4	4
4	3	3	3	–	1	1	2	4	4	4	4	4
5	3	3	3	2	–	1	2	4	4	4	4	4
6	3	3	3	3	3	–	2	4	4	4	4	4
7	4	4	4	4	4	4	–	4	4	4	4	4
8	4	4	4	4	4	4	2	–	2	1	2	1
9	4	4	4	4	4	4	2	2	–	2	2	2
10	4	4	4	4	4	4	3	3	3	–	2	1
11	4	4	4	4	4	4	2	2	2	3	–	2
12	4	4	4	4	4	4	2	2	2	3	1	–

$R = 0.5$ при $R_{0.99} = 0.35$.

а с объектами 7–12 – в IV. Это записано в первом столбце матрицы *cV*. В качестве меры сходства этих дендрограмм использовался коэффициент корреляции *R*. Для его расчета матрицы развертываются в два ряда соответствующих значений или строится корреляционная таблица. Когда число интервалов менее 4, вместо *R* используются непараметрические показатели. В данном случае $R = 0.5$, что достоверно с вероятностью 0.99.

Рассмотренный подход до недавнего времени считался единственным для сравнения дендрограмм [24]. Однако возможна его модификация с помощью вышеописанных показателей ПП и КА для иерархических разделений (табл. 5).

Соответствие $cV_а$ и $cV_б$ определяется по числу совпадений элементов этих матриц. ПП здесь равен 0.36, то есть достоверен с вероятностью 0.99. КА также составил 0.36, однако оценки его достоверности авторами не предложено, поэтому он имеет только иллюстративное значение.

Сравнение дендрограмм таким методом не является строгим, поскольку число интервалов назначается довольно произвольно. Более объективный подход состоит в выделении на дендрограммах классов объектов по критерию качества и проведении сравнения по показателям КА и ПП.

Содержательным примером является сравнение трех классификаций почв поймы средней Оби [26].

Почвы в поле классифицировали по системе Добровольского [10]. Затем они были переопределены в терминах классификации Шрага [28]. Каждый разрез описывался девятью признаками, отражающими степень гидроморфизма, гранулометрический состав гор. Ад, А1, В, Д (С) и уровень грунтовых вод (табл. 6).

На основе указанных морфологических показателей проведена автоматическая классификация, для чего описания разрезов кодировались. Коды являются не ранжировкой, а только цифровым обозначением, то есть номинальной шкалой. На рис. 6 представлены результаты автоматической группировки почв. Условными знаками показаны наименования почв по классификации Г.В. Добровольского.

Получено пять компактных групп, которые объединяют почвы преимущественно одного подтипа. Но так как почвы при переходе от одного типа к другому изменяются постепенно, то естественно, что к группе дерновых почв (группа 1) присоединяются дерново-луговые почвы, которые отличаются от дерновых слабыми признаками гидроморфизма. Большая группа луговых почв (II), содержащая 35 разрезов, включает часть дерново-луговых почв, которые по условиям гидроморфности близко примыкают к луговым поч-

Таблица 5. Соответствие дендрограмм (число классов в классификациях K_1 и K_2 =4)

K_1/K_2	1	2	3	4	m_i	$m_{i\max}$
1	–	1	–	–	1	1
2	5	4	–	4	13	5
3	4	11	–	1	15	11
4	–	6	–	30	36	30
m_i	9	22	0	35	$N = 66$	$S_2 = 47$
$m_{j\max}$	5	11	0	30	$S_1 = 46$	

1) ПП = 0.36, $\varphi^2 = 0.38$.

$\chi^2 = 25.1, f = 9$, $\chi^2_{0.99} = 21.7$.

2) КА = 0.36.

Таблица 6. Коды морфологических описаний разрезов

Код	Признаки проявления гидроморфизма	Гранулометрический состав	Глубина грунтовых вод, см
0	Признаков гидроморфизма нет	Песок	Более 200
1	Бурые, охристые, ржавые пятна, примазки, прожилки	Слоистость: песок (супесь) – супесь (суглинок)	150–200
2	Сизоватый оттенок, охристые (ржавые) пятна на буром (сером) фоне	Супесь	100–150
3	Сизые пятна на буром (сером) фоне, иногда с охристыми (ржавыми) пятнами	Легкий суглинок	50–100
4	Охристый (ржавый) горизонт с железистыми (марганцевыми) образованиями	Средний суглинок	0–50
5	На буро (охристо)-сизом фоне железистые (марганцевистые) образования (ортштейны, бобовины, конкреции)	Тяжелый суглинок	
6	Сизый (грязно-сизый) горизонт, иногда с охристыми (ржавыми, бурыми) пятнами	Глина	
7	Оторфованный горизонт (слабооторфованная дернина)	Заиленный песок, торф	
8	Торф (заиленный, полу- и полностью разложившийся)	Торф	

вам. Точно так же к группе лугово-болотных почв (IV) тесно примыкает часть луговых почв с повышенной гидроморфностью. В табл. 7 показано, как распределяются 126 профилей пойменных почв рассматриваемого района по таксонам трех классификаций.

На основании этих данных в табл. 8 приведены критерии сравнения – коэффициенты ассоциации и полихорические показатели связи.

Достоверное сходство можно отметить лишь между классификацией автоматической и Г.В. Добровольского. Остальные связи слабые.

Такие соотношения в значительной мере объясняются тем, что для обработки на ЭВМ использованы показатели, более отвечающие основаниям классификации Г.В. Добровольского, чем В.И. Шрага. Однако главное здесь в самой возможности количественно сопоставить различные классификации. Проведение специально организованных сопоставлений почвенных классификаций на выделенных в разных регионах полигонах позволит решить многие спорные вопросы.

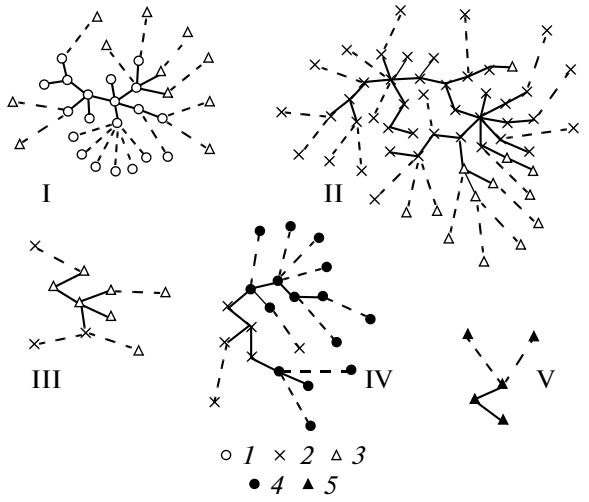

Рис. 6. Результат автоматической группировки почв. Почвы: *1* – дерновые; *2* – луговые; *3* – дерново-луговые; *4* – лугово-болотные; *5* – болотные.

В задачах, когда классы (таксоны) почв определены, дискриминантный анализ позволяет исследовать отношения между ними, в том числе включения или пересечения таксонов (классов) почв, описанных в арифметической шкале (рис. 7).

Этот вид многомерного статистического анализа реализует формализацию результатов классификации и предоставляет возможность классифицирования (распознавания) новых почв относительно установленных классов. Метод применим на всех уровнях организации почвенной системы [23].

Для примера взяты три класса объектов с 8, 14 и 12 объектами, описанными 4 признаками. На рис. 8 приведены результаты счета по программе линейного дискриминантного анализа.

В результирующей таблице приведены расстояния Махаланобиса между классами и соответствующие F критерии с числами свободы $f_1 = m -$ число признаков и $f_2 = n_i + n_l - m - 1$.

Линейная дискриминантная функция (**ЛДФ**), максимизирующая различия между классами 1 и 2 по признакам $x_1 - x_4$, имеет вид:

$$L_{1,2,j} = 492.59 + 1.46x_1 + 2.41x_2 + 1.04x_3 + 11.17x_4.$$

Различия классов по критерию Фишера F_{f1f2} вполне достоверны, что подтверждается и тем,

Таблица 7. Классификация пойменных почв р. Обь

Класс	Аллювиальные почвы	Номера классифицируемых объектов
	Классификация Г.В. Добровольского	
1	Дерновая	1–3, 11–13, 31, 32, 34, 37,40, 59, 66, 88, 89, 99
II	Дерново-луговая	5–7, 20, 29, 35, 38, 54, 60, 61, 63, 67, 68, 70, 72, 73, 79, 84, 93, 94, 97, 101–103, 105,107–110
III	Луговая	4, 9, 10, 15, 17–19, 21–23, 25–28, 30, 33, 36, 39, 41–50, 52, 53, 55–58, 62, 65, 71, 74–78, 80–83, 85–87, 95, 96, 98, 100, 104, 106, 113, 116,118, 125, 126
IV	Лугово-болотная	14, 16, 24, 51, 64, 69, 90, 92, 112, 114, 117, 119, 121–124
V	Болотная	8, 91, 111, 115, 120
	Классификация В.И. Шрага	
I	Слоистые	2, 59, 99, 103
II	Зернисто-слоистые	1, 11–13, 29, 31, 40, 42, 61, 66, 67, 89, 96
III	Зернистые	3–7, 17–23, 26–28, 32, 34, 35, 37, 39, 41, 47, 50, 52–55, 57, 58, 60, 62, 63, 65, 68, 70, 73, 76, 78–80, 82, 84–86, 88, 93–95, 102, 105, 107–109, 113, 124–126
IV	Слоистые и зернисто-слоистые заболоченные	15, 33, 38, 43–45, 72, 74, 75, 97, 101, 106, 110
V	Зернистые заболоченные	9, 10, 16, 24, 25, 30, 36, 46, 48, 49, 51, 56, 64, 69, 71, 77, 81, 83, 87, 90, 92, 98, 100, 104, 116–118, 121–123
VI	Иловато-глеевые	14, 120
VII	Иловато-болотные	8, 91, 111, 112, 115, 119
	Классификация на ЭВМ	
I	–	1, 3, 11–13, 31, 32, 34, 37, 40, 59, 60, 66, 72, 79, 84, 89, 97, 99, 101, 103, 105, 107, 109, 110
II	–	4, 10, 15, 17–23, 25–28, 35, 38, 41–43, 48–50, 52–54, 56–58, 61–63, 65, 70, 71, 73–78, 80–83, 85–87, 93–96, 98, 100, 102, 104, 106
III	–	5–7, 29, 30, 33, 39, 44, 55, 67, 68
IV	–	9, 14, 16, 24, 36, 45–47, 51, 64, 69, 90, 92, 112, 114,117–119, 121–124, 8, 91, 111, 115, 120

что проекции объектов на цифровые оси (значения L_{ilj}) не пересекаются.

При наличии обучающих выборок, подобных приведенным в данном примере, классифицирование новых объектов осуществляется расчетом значения L_{ijl} и соотнесением их с интервалами проекций объектов соответствующих классов. Для рассмотренных первых двух классов это соответственно 34–42 и 0–7. В случае промежуточного результата объект относят по близости к границам того или иного интервала, хотя возможно и возникновение неопределенности в классифицировании, особенно при пересечении проекций.

Классифицирование новых объектов по системе ЛДФ осуществляется вычислением проекций по каждой из них и отнесением объекта к тому из классов, в который он попадает чаще других (метод голосования).

Существуют варианты дискриминантного анализа с учетом неравенства ковариационных матриц, с ЛДФ на три и более классов и другими решающими правилами классифицирования, в том числе в бинарной шкале признаков. Классифицирование объектов может производиться по мере сходства с объектами обучающих классов: по близости с ближайшим соседом, средним сходством со всеми объектами классов, по интервалам сходства.

Приведенные алгоритмы и методы реализуются в рамках общей структуры задач численной классификации (рис. 9). Возможны два варианта постановки задачи – диагностическая классификация предполагает распознавание (классифицирование) новых почв относительно заданных классов (обучающих образов); автоматическая классификация – группировка почв по автомати-

Таблица 8. Критерии ассоциации и связи классификаций КА (ПП)

Классификация	В.И. Шрага	ЭВМ
Г.В. Добровольского	0.27 (0.30)	0.53 (0.56)
В.И. Шрага	–	0.25 (0.12)

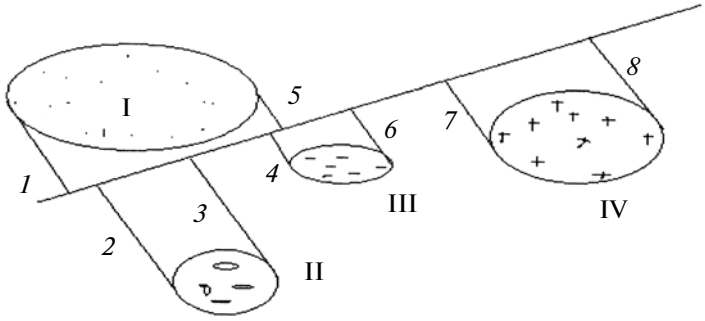

Рис. 7. Линейная дискриминантная функция. $L = MS^1 = L_0 + L_1 X_1 + L_2 X_2 + ... + L_M X_M$.

Номера классов		Расстояние Махаланобиса D2	F	f2
i	l			
1	2	223.43	241.7	17
1	3	371.00	371.0	15
2	3	18.35	25.9	21

Коэффициенты линейных дискриминантных функций:

.....Между классами 1 и 2:
B[0] = 492.59 B[1] = 1.46 B[2] = 2.41 B[3] = 1.04 B[4] = 11.17
.....Между классами 1 и 3:
B[0] = 508.53 B[1] = 0.73 B[2] = 2.58 B[3] = −0.35 B[4] = 19.17

.....Между классами 2 и 3:
B[0] = 108.46 B[1] = 0.54 B[2] = 0.78 B[3] = −0.25 B[4] = 3.91

Проекции объектов на разделяющую плоскость:

Между классами 1 и 2
класс 1
41 41 42 36 40 36 34 41
класс 2
4 3 1 2 0 7 0 2 2 5 5 5 2 6

Между классами 1 и 3
класс 1
57 56 58 52 57 50 49 57
класс 3
3 1 3 5 3 5 4 1 3 7 0 3

Между классами 2 и 3
класс 2
23 22 17 18 17 26 17 20 19 24 24 24 19 25
класс 3
6 4 7 9 7 9 8 1 4 13 0 8

Рис. 8. Многомерный статистический анализ (линейные дискриминантные функции) ($f1 = 4$).

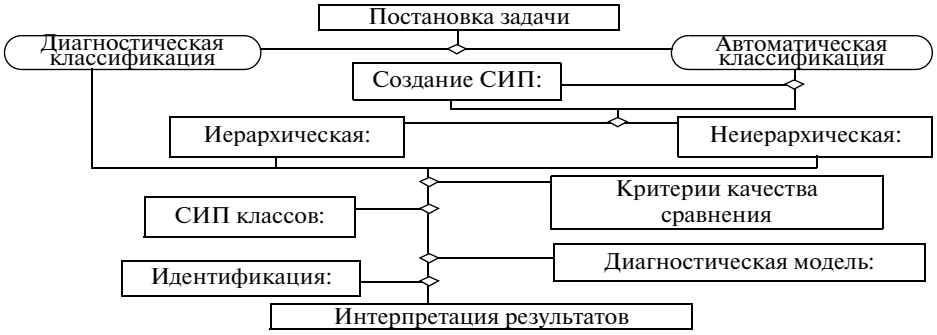

Рис. 9. Структура задач численной классификации почв.

чески определяемым классам в иерархической или другой структуре. Критерии качества и/или сравнения могут быть определены для заданных или построенных классификаций. Для выбранных из них можно определить более компактную систему информативных признаков (СИП).

Различия классов фиксируются системой ЛДФ, образуя диагностические модели для классифицирования новых объектов. Имеются и другие методы классифицирования (распознавания, идентификации) объектов, то есть определения класса, к которому их следует отнести.

Описанные методы реализованы в пакете прикладных программ MERON [23], с помощью которого решаются рассмотренные выше задачи численной классификации почв.

ЗАКЛЮЧЕНИЕ

Формализация понятий и процедур является наиболее эффективным средством достижения объективности классификации почв, сравнительного анализа и достижения взаимопонимания разных научных школ. Это необходимый процесс для совершенствования теоретических представлений и практических приложений почвоведения. Классификация почв требует первоочередного внимания, поэтому с ее аксиом должна начаться такая формализация. Приведенные примеры применения формальных методов имеют определенный смысл для почвоведения. Естественно, это не всегда просто визуализируется и интерпретируется, зато инициирует новые идеи, а, возможно, и гипотезы.

Общая схема организации процедур ЧК может рассматриваться, как руководство к последовательности этапов построения и/или исследования почвенных классификаций. После формулировки цели классификации для заданных или автоматически определенных таксонов почвенных объектов, представленных ординатной или иерархической структурой, определяется оптимальная система информативных признаков, определяются критерии качества полученной классификации и формализуются решающие правила (диагностирующие модели) распознавания новых объектов. Таким образом, создается количественно обоснованная и визуализируемая классификационная система, лишенная скрытого субъективизма разработчиков.

Отдельный вопрос – применение обсуждаемых методов в цифровой картографии. Они в значительной мере пересекаются и в перспективе, очевидно, найдут свое приложение. Наряду с исследованием информативности дешифровочных признаков и качества распознавания образов, предстоит приложение мер включения, формализации рисунков структуры почвенного покрова [23], а также адаптация аппарата анализа нечетких множеств.

СПИСОК ЛИТЕРАТУРЫ.

1. *Айвазян С.А., Бежаева З.И., Староверов О.В.* Классификация многомерных наблюдений. М.: Статистика, 1974. 240 с.
2. *Аркадьев А.Г., Браверманн Э.М.* Обучение машины классификации объектов. М.: Наука, 1971. 192 с.
3. *Бейли Н.* Математика в биологии и медицине. М.: Мир, 1970. 268 с.
4. *Балык В.М., Калуцкий Н.С., Кулакова Р.Д.* Структурно-параметрическая самоорганизация сложных технических систем // Деловая Россия. 2007. С. 58–63.
5. *Бобров В.А.* Аппроксимация некоторых почвенных функций и интегральный способ расчета количества гумуса в почве // Математические методы в биолиологии и почвоведении. Алма-Ата: Наука, 1976. С. 103–111.
6. Большая Российская энциклопедия. М.: Изд-во БСЭ, 2005. Т. 1. С. 209.
7. *Боул С., Хоул Ф., Мак-Крекен Р.* Генезис и классификация почв. М.: Прогресс, 1977. 416 с.
8. *Воронин Ю.А.* Начала теории сходства. Новосибирск: Наука, 1991. 128 с.
9. *Высокос Г.Н., Рожков В.А.* Шкалы почвенных признаков и выбор мер сходства объектов // Почвенные и агрохимические исследования с применени-

ем ЭВМ // Тр. Почв. ин-та им. В.В. Докучаева. 1981. С. 30–39.

10. *Добровольский Г.В.* Почвы речных пойм центра Русской равнины. М.: Изд-во Моск. ун-та, 1968. 295 с.
11. *Князева Е.Н., Курдюмов С.П.* Основания синергетики. Человек, конструирующий себя и свое будущее. М.: КомКнига, 2007. 232 с.
12. *Лурье И.К.* Геоинформационное картографирование. М.: КДУ, 2010. 424 с.
13. *Махлин Т.Б.* Аппрохимация кривыми Джонсона функций распределения элементов вещественного состава почвы // Почвоведение. 1973. № 6. С. 123–130.
14. *Пфанцагль И.* Теория измерений. М.: Мир, 1976. 248 с.
15. *Рейнтам Л.Ю.* Корреляция и регрессия между свойствами почв буроземного, псевдоподзолистого и дерново-подзолистого типов // Почвоведение. 1971. № 1. С. 116–132.
16. *Решетуха И.В.* Аппроксимация методом наименьших квадратов таблично заданной функции $f(x,y)$ полиномом $P_{n,m}(x,y)$ // Программное обеспечение ЭВМ МИР-1 и МИР-2. Т. 2. Киев: Наукова думка, 1976. С. 126–129.
17. *Рожков В.А.* Алгебра WRB (формализация концепции) // Тр. Всеросс. конф. “Экспериментальная информация в почвоведении: теория и пути стандартизации. М.: Изд-во Моск. ун-та, 2005. С. 73–82.
18. *Рожков В.А.* Алгоритмы и программы объективной классификации почв на ЭВМ Мир-2. М.: ВАСХНИЛ, 1976. 172 с.
19. *Рожков В.А.* Метод главных компонент и его применение в почвоведении // Почвоведение. 1975. № 10. С. 141–151.
20. *Рожков В.А.* О математической формализации распределения веществ по профилю почв // Бюл. Почв. ин-та им. В.В. Докучаева. Вып. IV. 1972. С. 66–73.
21. *Рожков В.А.* Почвенная информатика. М.: Агропромиздат, 1989. 222 с.
22. *Рожков В.А.* Экономный код почвенных данных // Бюл. Почв. ин-та им. В.В. Докучаева. 1972. Вып. 53. С. 32–35.
23. *Рожков В.А., Скворцова Е.Б.* Тектология почвенной мегасистемы (общность организации и анализа данных) // Почвоведение. 2009. № 10. С. 1155–1164.
24. *Сокал Р.Р.* Кластер-анализ и классификация: предпосылки основные направления // Классификация и кластер. М.: Мир, 1980. С. 7–19.
25. *Чандра А.М., Гош С.К.* Дистанционное зондирование и географические информационные системы. М.: Техносфера, 2008. 312 с.
26. *Шеремет Б.В., Рожков В.А.* Опыт численной таксономии пойменных почв р. Оби по морфологическим свойствам // Вест. Моск. ун-та. Сер. 17, почвоведение. 1977. № 3. С. 41–50.
27. *Шишов Л.Л., Рожков В.А., Столбовой В.С.* Информационная база классификации почв // Почвоведение. 1985. № 9. С. 9–20.
28. *Шраг В.И.* Пойменные почвы, их мелиорация и сельскохозяйственное использование. М.: Россельхозиздат, 1969. 270 с.
29. *Colwell J.D.* A statistical-chemical characterization of four great soil groups in southern New South Wales based on orthogonal polynomials // Austral. J. Soil Sci. Res. 1970. V. 8. № 3. P. 221–238.
30. *Erle K.T.* Application of the spline function on soil science // Soil Sci. 1972. V. 114. № 5. P. 333–338.
31. *Goodman L.A., Kruskal W.K.* Measure of association for cross classification // J. Amer. Statist. Assoc. 1954. № 49. P. 732–764.
32. *Gruijter J.J. de.* Numerical Classification of Soils and Its Application in Survey // Soil Survey Papers. 1977. № 12. 117 p.
33. *Protz R., Arnold R.W., Presant E.W. The approximation* of the true modal profile with use of the high speed computer and landscape control // Trans. 9th. Internat. Congr. Soil Sci. Adelaide. 1968. V. 4. P. 193–204.
34. *Sokal R.R., Rohlf F.I.* The comparison of dendrogramms by the objective methods // TAXON. 1962. № 11. P. 33–40.
35. *Sokal R.R., Sneath P.H.A.* Principles of Numerical Taxonomy. San-Francisco: W.H. Freeman and Co., 1963. 359 p.
36. *Webster R.* Quantitative and numerical methods in soil classification and survey. Oxford: Clarendon Press, 1979. 269 p.