МЕТОДИЧЕСКИЕ РАБОТЫ И КРАТКИЕ СООБЩЕНИЯ

УДК 631.4

В. А. РОЖКОВ, Н. В. ПРОШИНА

ОПЫТ ЧИСЛЕННОЙ ТАКСОНОМИИ ПОЧВ

Развиваются положения известной работы [15] по использованию дендрограмм и отсеву неинформативных признаков. Расширение рассматриваемого подхода выполнено включением разнообразных методов построения дендрограмм и оценок информативности признаков, а также завершением исследования диагностической классификации новых объектов.

Основной задачей почвенной таксометрии (численной таксономии) является исследование заданного множества почвенных разностей с целью выявления их однородных (в определенном смысле) или компактных групп (таксонов). Часто при этом отсутствует какая-либо априорная информация о числе и составе таксонов, кроме предположения о неоднородности множества. Элементы множества задаются через признаки, т. е. свойства почв. При ограниченном числе наблюдений и большой размерности пространства признаков не представляется возможным построить критерии для проверки статистических гипотез. Поэтому для решения задач таксономии могут быть использованы методы, основанные на понятиях сходства объектов и компактности таксонов. Одним из вариантов таких методов является расчет, построение и использование дендрограмм. В почвоведении накоплен значительный опыт таких исследований [9—11, 13—15, 17—19]. Идеи и техника метода изложены в работах [1, 9, 16].

Применению дендрограмм, оценке информативности признаков и диагностической классификации почв посвящена и настоящая статья.

Обозначим множество (выборку) почвенных разностей, подлежащее исследованию, матрицей X:

$$X = x_{ij} = \begin{bmatrix} x_{11}x_{12} & \dots & x_{1j} & \dots & x_{1q} \\ x_{21}x_{22} & \dots & x_{2j} & \dots & x_{2q} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1}x_{i2} & \dots & x_{ij} & \dots & x_{iq} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1}x_{n2} & \dots & x_{nj} & \dots & x_{nq} \end{bmatrix}, \qquad (1)$$

где x_{ij} — значение (или градация) j-го признака для i-го наблюдения (почвы), $i=1, 2 \dots, n$ — число наблюдений, $j=1, 2 \dots, q$ — количество признаков. Как уже указывалось, возможен случай, когда $n \leqslant q$, что недопустимо при статистическом подходе.

Каждое наблюдение может рассматриваться как точка в многомерном пространстве, координатами которого является количество признаков. Задача таксономии ставится исходя из предположения неоднородности множества X, т. е. точки в q-мерном пространстве признаков распределены не равномерно, а образуют сгущения и разрежения. Понятно, что сгущения образуют сходные объекты. Чем более они сходны (ближе друг к другу) и более отличны (дальше) от объектов других сгущений, тем компактней группа. Расстояние между объектами определяется из выражения

$$d_{il} = \left[\frac{1}{q}\sum_{j=1}^{q}\left(\frac{x_{ij}-x_{lj}}{\omega_j}\right)^p\right]^{\frac{1}{p}}, \qquad (2)$$

где d_{il} — расстояние между i-тым и l-тым объектами, q — количество признаков, p — показатель степени, обычно принимаемый равным 1 или 2, но при необходимости усилить различия и более [8], ω_j — нормирующий множитель.

При значениях $p=2$ и $\omega_j=1$ выражение (2) называется эвклидовым расстоянием. Применимо оно лишь в случаях одинаковой размерности признаков. Более общими условиями являются:

$$p=1 \quad \text{и} \quad \omega_j = x_{ij} + x_{lj}, \qquad (4)$$

$$p=2 \quad \text{и} \quad \omega_j = x_{ij} + x_{lj}, \qquad (5)$$

$$p=1 \quad \text{и} \quad \omega_j = \max(x_{ij};\ x_{lj}), \qquad (6)$$

$$p=2 \quad \text{и} \quad \omega_j = s_j, \qquad (7)$$

где s_j — квадратическое отклонение j-го признака.

Назначение весового множителя ω_j — нормировка признаков в безразмерные величины [2, 12]. Существуют и некоторые другие его виды [4, 5], однако нет каких-либо правил выбора в конкретной задаче. И хотя в целом все они дают сходные результаты, возможны и различия, о чем будет ниже указано. Наилучшим основанием для выбора p и ω_j является соответствие результатов счета фактическому соотношению исследуемых почв, но такая практическая проверка не всегда возможна. В значительной мере роль этих параметров зависит от характера признаков размерности и вариабельности и меняется вместе с ними. Другим подходом, также не всегда выполнимым, является дублирование разделения множества с использованием выражений (2)—(7), например, методом главных компонент [6]. Совпадение результатов помогает выбрать необходимые параметры. Однако такое внимание к ним оправдано лишь в очень тонких и ответственных исследованиях. В массовых исследованиях этот выбор определяется наличием соответствующей программы для ЭВМ и временем счета.

Заметим, что нормировка не имеет отношения к весу или диагностической нагрузке признаков. Для введения последних необходимо к ω_j ввести дополнительные коэффициенты, но это не является предметом рассмотрения в настоящей работе. Веса определяют целевую ориентацию анализа, но здесь она ставится через состав учтенных признаков, взятых с равным весом.

Для двух одинаковых объектов расстояние будет равно 0, для максимально отличных 1 при условиях (3)—(6) и может быть больше 1 при (7).

Расстояние между объектами может быть преобразовано в показатель сходства

$$r_{il} = 100\,(1 - d_{il}),\ \%. \qquad (8)$$

Для выражения (2) при условиях (3)—(6) значение показателя сходства изменяется от 100% для пары одинаковых объектов до 0 для максимально отличных. При условии (7) выражение (8) может дать отрицательный результат.

В качестве показателя сходства объектов при большом числе признаков может быть использован коэффициент корреляции (объектов)

$$R_{il} = \frac{\sum_{j=1}^{q} x_{ij}x_{lj} - \frac{1}{n}\sum_{j} x_{ij}\sum x_{lj}}{\left\{\left[\sum x_{ij}^2 - \frac{1}{n}\left(\sum x_{ij}\right)^2\right]\left[\sum x_{lj}^2 - \frac{1}{n}\left(\sum x_{lj}\right)^2\right]\right\}^{0,5}}. \qquad (9)$$

Для исключения отрицательных значений R_{il} в случае обратной сопряженности можно использовать выражение $\arccos R_{ij}$ или просто прибавить к полученной величине 1. В последнем случае максимальному различию будет соответствовать 0 показателя сходства, а полному сходству — 2 [16, 19].

Расчет расстояний (сходства) составляет первую часть анализа. Использование их для построения дендрограммы может быть проведено разными методами [1, 16], из которых наиболее распространен так называемый взвешенный парногрупповой. Для его реализации вначале находят два наиболее близких (сходных) объекта. Эта пара объединяется (усредняется) и в дальнейшем выступает как один объект. Далее снова находят два самых близких объекта, усредняют их и т. д. При этом такими объектами могут оказаться уже вычисленные, и их объединяют с учетом численности ранее объединенных объектов.

Таким образом, все объекты исходного множества объединяются в единую иерархическую систему. Находимые минимальные расстояния служат исходными данными для построения дендрограммы. Последняя является графическим отображением соотношений объектов множества по комплексу признаков, структурой этого множества. Это дает наглядное представление о множестве и позволяет выдвигать гипотезы о числе и составе компактных групп.

Следующим этапом анализа является оценка информативности признаков. В данной задаче могут быть сделаны лишь относительные оценки. Естественно предположить, что чем больше варьирование признака, тем выше его информативность в смысле различий объектов. Основание для такого предположения чисто логическое, которое используется при отсутствии информации о числе групп почвенных объектов. В реальной ситуации варьирование признаков может и не иметь столь простой связи с разделением групп, поэтому подобная оценка информативности требует подтверждения, которое и дано ниже.

Мерой варьирования может быть принят показатель

$$t = \frac{\bar{X}_j}{m_j}, \qquad (10)$$

где $\bar{X}_j$ — среднее арифметическое j-го признака, $m_j = \left[\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_{ij} - \bar{X}_j)^2\right]^{0,5}$ — ошибка среднего.

Чем меньше t, тем более информативен признак. Коэффициент вариации также может быть такой мерой:

$$V = 100\frac{s_j}{\bar{X}_j}, \qquad (11)$$

где $s_j = m\sqrt{n}$ — квадратическое отклонение j-го признака. Чем больше V, тем выше информативность.

Показателем варьирования может служить и выражение

$$I_j = \left[\frac{1}{n}\sum_{i=1}^{n-1}\sum_{k=i+1}^{n}\left(\frac{|x_{ij}-x_{i<j|}}{\omega_j}\right)^p\right]^{\frac{1}{p}}, \qquad (12)$$

где i, $k=1, 2 \ldots, n$ и $i \neq k$, p и ω_j — аналогичны (3)—(6). Чем больше I_j, тем выше информативность j-го признака. Обращает на себя внимание общность выражений (12) и (2), что позволяет использовать сходные по параметрам выражения для расчета дендрограммы и оценки информативности.

Перечисленные показатели основаны на исследовании каждого отдельного признака без учета их взаимосвязи. В противоположность им корреляционный метод, апробированный в почвоведении [7, 10, 11, 15], основан только на учете их взаимосвязи. Для этого находится корреляционная матрица признаков

$$r_{jk} = \frac{\sum_{i=1}^{n} x_{ij}x_{ik} - \frac{1}{n}\sum_{i}^{n} x_{ij}\sum x_{ik}}{\left\{\left[\sum x_{ij}^2 - \frac{1}{n}\left(\sum x_{ij}\right)^2\right]\left[\sum x_{ik} - \frac{1}{n}\left(\sum x_{ik}\right)^2\right]\right\}^{0,5}}. \qquad (13)$$

Наиболее информативным считается признак, который не имеет сопряженности с другими, и, наоборот, чем выше корреляция, тем менее информативен признак. При равных показателях корреляций менее информативен тот признак, который имеет больше связей (с бо́льшим числом других признаков). Логика подхода заключается в том, что признак, тесно связанный с другими, добавляет мало информации о системе, поскольку она уже имеется в тех других, а потому может быть исключен. Однако при этом нельзя указать нижних пределов ни по уровню, ни по количеству связей, а поэтому, так же как и в предыдущем подходе, отсев каждого признака должен строго контролироваться повторением группировки.

Ниже будет показана возможность применять комбинацию корреляционного метода и метода, основанного на учете варьирования признаков, как дополняющих друг друга.

Итогом любых классификационных построений, не исключая и излагаемого, является проверка их на «экзаменационной» выборке, т. е. на объектах с априорно известной принадлежностью. В данном случае с использованием дендрограмм это можно сделать или по среднему сходству диагностируемых объектов с какой-либо группой объектов, выделенных на дендрограмме, или просто по максимальному сходству с отдельными объектами. Очевидно, для четкой диагностики уровень сходства при этом не может быть меньше минимального в дендрограмме и иметь единственный максимум.

Для иллюстрации описанного анализа использовано по три повторности из слоя 0—20 *см* шести почв Звенигородского района Московской обл. (всего 18 образцов). Каждый образец описан 30 признаками, значения которых приведены в табл. 1. Задача имела узкую цель разделить почвы по заданному описанию и минимизировать последнее, поэтому состав признаков и интерпретация результатов не вполне соответствуют принятым в классификации почв традиционным методам. Необходимые расчеты выполнены на ЭВМ «Мир-2» по нашим программам, реализующим все описанные методы.

Для выбора метрики расстояний между объектами была использована информация о принадлежности образцов к почвенным разностям.

Характеристика объектов исследования

Таблица 1

№ признака	Наименование признака	Индекс	I			II			III			IV			V			VI		
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	Содержание, %	H	1,84	2,86	1,87	0,67	0,52	0,55	0,34	0,23	0,25	0,40	0,34	0,30	0,47	0,45	0,39	0,29	0,26	0,23
2		C	11,25	16,86	11,37	2,49	2,07	2,31	0,97	0,68	0,70	0,77	0,76	0,75	1,34	1,39	1,21	0,69	0,73	0,53
3		N	0,54	0,52	0,46	0,20	0,14	0,26	0,06	0,06	0,06	0,10	0,08	0,10	0,14	0,16	0,12	0,08	0,07	0,05
4	Коэффициенты отражения при 680 *нм*: $\Delta R = R_{680} + R_{510}$;	R_{680}	12,4	10,1	16,0	23,6	24,0	24,2	40,8	39,5	37,5	31,3	27,5	27,5	35,1	34,9	34	36,3	36,0	37,0
5		ΔR	3,0	2,8	3,1	7,2	6,8	6,7	13,4	11,7	11,1	9,7	9,5	9,5	8,8	9,0	9,7	10,4	11,2	12,0
6	$\Sigma R = \frac{R_{440}+R_{490}+R_{540}+R_{590}+R_{640}+R_{950}}{6}$	ΣR	8,8	7,8	10,8	16,9	17,1	17,4	29,1	29,2	27,9	20,5	18,8	18,6	23,8	25,3	24,8	26,5	26,0	26,4
7	Оптическая плотность	D_{382}	1,3	1,33	1,31	0,16	0,12	0,12	0,20	0,18	0,10	0,08	0,10	0,07	0,15	0,21	0,16	0,17	0,15	0,08
8	Амплитуды линий в спектре ЭПР	$Fe_{пр}$	35	32	16	480	420	490	300	370	450	410	500	380	60	60	160	250	360	200
9	(ферромагнитных примесей железа, органических радикалов, Mn	R_p	200	200	350	115	120	130	22	30	40	120	115	105	18	25	37	81	85	65
10	карбонатов)	$Mn_к$	0	0	1	33	36	40	0	0	0	25	18	10	0	0	0	0	0	0
11	Содержание гигроскопической воды, %	Гигр.	4,92	4,02	4,67	1,72	1,42	1,48	0,89	0,72	0,71	0,99	1,01	1,16	0,92	1,05	0,98	0,76	0,69	0,59
12		$pH_{сол}$	4,4	4,4	4,4	6,5	7,1	7,1	4,6	4,5	4,5	7,4	7,5	7,0	4,3	4,3	4,3	4,4	4,6	4,7
13	Содержание элементов, *мг/кг*/100 *г*	B	15	13	11	23	20	20	18	15	19	19	27	25	24	23	21	17	15	18
14	почвы	Mn	44,7	81,3	64,6	850	740	700	351	331	490	1000	785	851	1050	1313	1460	950	1000	881
15		Cr	9	8	15	20	19	22	18	18	17	27	28	21	17	17	18	18	18	18
16		Cu	8	38	6,7	7,8	7,2	6,4	4	4	4	6	7	6	3	3	3	3,9	4,2	4,5
17		V	8	9	11	28,7	18	25	14	16	18	26	35	21	15	17	18	16	14	17
18		Ti	750	660	684	1990	1440	1730	1410	1450	1600	3090	3080	2190	1840	1990	1990	1590	1360	1600
19		Ni	15	10	15	22	22	27	15	15	16	23	23	22	18	15	16	17	18	18
20		Y	8	7,2	6,5	15	23	29	14	16	13	22	24	18	18	17	16	17	13	18
21		Zr	340	340	230	470	470	550	400	550	440	470	430	400	500	470	470	440	440	470
22		Yb	1,2	1,0	0,9	1,5	1,8	2,1	1,7	1,8	1,4	2,3	2,7	2,2	1,9	1,9	1,8	1,8	1,8	1,7
23		Sc	4,5	3,8	54	4,0	5,2	5,5	3,8	3,9	3,7	4,4	6,4	5,3	4,3	4,6	4,3	4,7	4,8	4,3
24		Co	2,4	2,3	2,2	3,8	4,6	4,9	3,2	3,3	3,4	5,4	5,6	4,3	3,7	7,4	8,0	5,2	5,3	5,1
25		Sr	72	90	68	78	84	82	90	81	72	76	105	84	80	76	90	86	94	72
26	Интенсивность почернения линий	Ba	50	48	44	48	54	45	49	45	54	52	51	51	50	46	52	51	53	47
27		Ca	30	33	36	66	65	63	45	54	52	60	59	57	46	42	51	51	50	49
28		Fe	56	97	48	100	131	86	95	82	108	100	126	101	92	88	88	107	106	126
29	Активность ферментов (инвертазы,	Инв.	38,5	34,3	44,4	31,7	29,8	29,4	4,8	4,2	5,1	5,5	5,4	8,1	2,1	5,8	3,8	1,3	1,4	2,8
30	*мг* глюкозы на 1 *г* почвы за сутки; каталазы, *см*^{3} O_2 на 1 *г* почвы за 5 мин)	Кат.	25	31	27	11	14	13	4,0	3,5	4,0	10	10	8	4	4,5	5,5	5,0	4,5	3,5

Примечание. Почва: I — торфянисто-подзолистая, II — пойменная луговая, III — подзолистая лесная, IV — пойменная пахотная, V — дерново-подзолистая лесная, VI — дерново-подзолистая пахотная.

Таблица 2

Показатели сходства объектов по полному набору признаков ($Q = 30$, верхний треугольник) и по их минимальной комбинации ($Q = 5$, нижний треугольник), %

№ образца	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1		69	65	39	38	37	36	35	35	32	30	34	37	37	35	35	36	34
2	58		60	37	36	35	37	35	35	31	29	32	36	36	35	35	35	33
3	44	41		36	35	35	34	33	32	29	28	31	33	34	32	33	33	31
4	38	34	24		71	71	48	47	50	60	57	60	51	50	51	51	50	48
5	36	35	24	77		75	49	48	51	62	60	62	51	50	52	53	52	51
6	35	36	24	76	80		46	47	48	61	59	59	49	49	49	49	49	47
7	25	33	16	40	44	43		75	71	52	50	54	63	62	64	65	66	63
8	25	32	16	42	45	45	77		74	52	49	54	60	60	63	67	67	67
9	25	29	15	47	48	50	68	70		56	53	57	60	59	63	67	68	69
10	26	25	16	55	55	54	55	57	64		73	71	53	53	57	57	57	56
11	25	25	16	55	56	57	54	53	60	69		69	49	49	52	55	54	53
12	27	28	17	58	60	59	54	55	58	73	66		54	54	57	59	58	57
13	28	33	15	33	33	33	40	41	40	43	38	40		74	69	64	62	61
14	30	35	16	34	34	34	44	42	46	50	43	45	65		74	62	59	58
15	22	29	12	35	36	36	51	49	47	50	45	47	55	60		67	64	63
16	22	28	12	42	43	42	51	50	49	51	46	50	54	43	53		80	71
17	21	26	12	44	46	44	51	52	51	56	51	54	51	41	50	74		70
18	25	27	14	43	43	43	53	53	54	57	50	54	54	48	59	63	58	

Оказалось, что, используя выражение (2) при условиях (4)—(6), получаем результаты, полностью соответствующие полевым определениям.

В верхнем треугольнике табл. 2 приведены полученные по первой метрике показатели сходства (8) в процентах. Пунктиром выделены показатели сходства объектов внутри однородных групп. Видно, что они резко превосходят таковые для объектов разных групп. Если бы объекты исходной выборки не были упорядочены по группам, их можно было бы выделить, переставив соответствующие строки и столбцы матрицы. В этом заключается простейший подход к выделению компактных скоплений объектов.

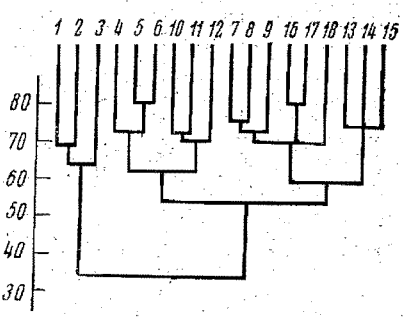

Рис. 1. Дендрограмма по полному набору признаков (q=30)

На рис. 1 представлена дендрограмма рассматриваемой совокупности, построенная по данным табл. 2. Из анализа ее следует, что по 30 признакам почвы могут быть сгруппированы по тройкам соответственно их принадлежности. Кроме того, можно видеть, что по выбранной метрике наиболее сходными (70%) оказались подзолистая лесная (III) и дерново-подзолистая пахотная (VI) почвы; на уровне 62% пойменные луговая (II) и пахотная; на уровне 59% к почвам III и VI присоединяется дерново-подзолистая лесная (V). Торфянисто-подзолистые почвы (I) отличаются от всех остальных очень резко, и уровень сходства их с другими равен 34%. Внутри трех повторностей этих почв сходства немного больше (64%), чем между разностями II и IV, что указывает на большое варьирование признаков торфянистых почв, их невысокую однородность. Иначе говоря, почва I образует наименее компактную группу.

При использовании метрики расстояния (2) при соблюдении условия (7) объекты почв II и IV, а также III и VI перемешались друг с другом. Это может быть связано с действительно нечеткими различиями почв по свойствам, но не исключена и слабая чувствительность метрики с использованием дисперсии в условиях значительного варьирования при-

знаков в выборке. При отсутствии данных о нормальности распределений признаков использование этой метрики проблематично.

Дендрограмма, построенная по выражению (9), также не соответствовала полевому описанию почв. Были смешаны почвы V и VI. И хотя смешение пахотных и лесных дерново-подзолистых почв имеет реальный смысл, тот факт, что оно произошло по повторностям, а не группами, обусловил выбор для дальнейших иллюстраций метрики (2) при условии (4), тем более что это выражение наиболее экономично в счете.

В качестве рекомендаций для выбора вида условий для метрики (2) при отсутствии информации о возможной группировке объектов можно отметить следующее.

При дискретном описании признаков (есть — 1, нет — 0) берется условие (3), при значительных различиях признаков по объектам — (4)

Таблица 3

Ранжировка признаков по относительной информативности
(1 — наибольшая информативность, 30 — наименьшая)

Ранг	Номер признака согласно табл. 1 по формуле				
	(10)	(11)	(12) при условии (4)	при условии (5)	при условии (6)
1	2	2	29	10	29
2	10	10	10	29	8
3	7	7	8	8	9
4	1	1	2	2	2
5	29	29	9	14	14
6	3	3	14	7	3
7	11	11	3	9	30
8	30	30	7	3	10
9	9	9	30	30	7
10	8	8	1	1	1
11	14	14	11	11	11
12	18	18	18	5	18
13	17	17	5	18	5
14	5	5	20	25	20
15	24	24	6	20	24
16	20	20	24	6	17
17	16	16	17	4	6
18	6	6	4	17	4
19	4	4	16	24	16
20	15	15	15	16	22
21	22	22	22	15	15
22	25	25	25	22	13
23	12	12	13	28	19
24	19	19	28	13	28
25	13	13	19	19	27
26	28	28	27	12	25
27	27	27	12	27	12
28	21	21	21	21	21
29	23	23	23	23	23
30	26	26	26	26	26

или (5), при слабых — (6). Если число наблюдений достаточно для проверки нормальности распределения по признакам, то принимается нулевая гипотеза — (7); при большом количестве признаков ($Q \geqslant 25$) и резких различиях их по объектам — выражение (9).

Выбор выражения может определяться также объемом исходных данных и наличием времени ЭВМ (перечислены они в порядке возрастания времени счета), а также необходимостью сравнить полученные ранее и вновь проводимые группировки. Условия (4)—(6) удобны еще и тем, что аналогичные применяются и при оценке информативности признаков.

Результаты оценки и ранжировки 30 признаков по информативности в смысле рассмотренных выше выражений (10)—(12) сведены в табл. 3.

Естественно, что по t (10) и V (11) получены одинаковые результаты. По вариантам метрики (12) имеются некоторые отличия, но поскольку за основу группировки принято условие (4), то и анализ информативности ведется с ним, а остальные оценки даны для сравнения. Графическое представление об изменении информативности признаков дает рис. 2.

Последовательно исключая признаки начиная с последнего (26-го) и повторяя каждый раз расчет дендрограммы, можно найти такую их минимальную комбинацию [3], которая не меняет группировки объектов, а дальнейшее исключение хотя бы одного признака эти изменения вызывает. За истинную принимается дендрограмма по исходным 30 признакам.

Оказалось, что половина признаков может быть таким образом отсеяна и минимальная их комбинация [3] включает (согласно номерам табл. 1) 29, 10, 8, 2, 9, 14, 3, 7, 30, 1, 11, 18, 5, 20, 6. Исключение 6-го признака уже ведет к смешению образцов почв III и VI на дендрограмме.

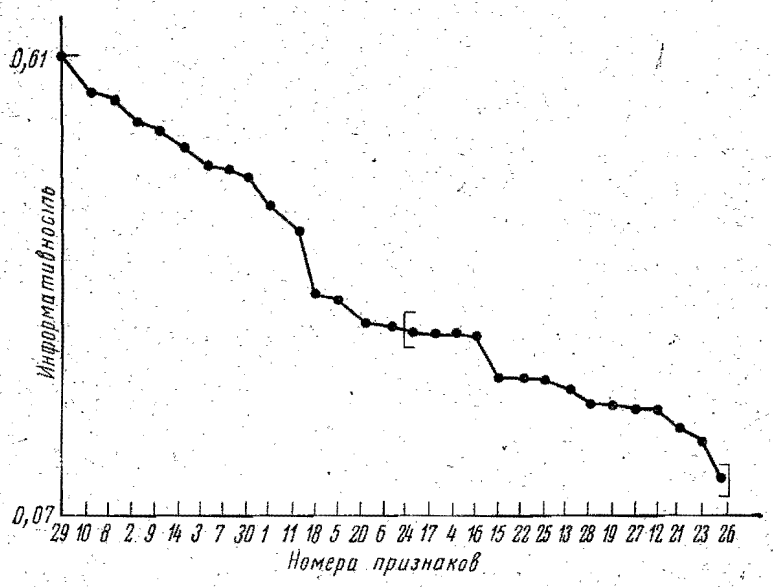

Рис. 2. Информативность признаков в ранжированном ряду (в скобки заключены признаки, которые можно исключить)

Как отмечалось, этот метод оценки информативности основан на учете вариабельности признаков: чем она больше, тем выше информативность (без учета их взаимосвязи).

Применяя выражение (13), мы вычислили корреляционную матрицу признаков (из-за большого размера она не приводится), исключая последовательно признаки, имеющие связи с другими, сначала на уровне $\geqslant 0{,}9$, потом $\geqslant 0{,}8$ и т. д., как это описано в работе [15]. Минимальную информативную комбинацию признаков, еще не изменяющую дендрограмму, составили всего 9 признаков: 8, 13, 14, 16, 21, 23, 25, 26, 29. Исключение 16-го признака, имеющего наибольшую из всех других связь, равную 0,655, уже ведет к смешению образцов почв III и IV, V и VI.

Как можно видеть, разными методами получены и разные информативные комбинации признаков. И, вероятно, они не исчерпывают все возможные другие комбинации. В частности, в последней комбинации признаки под номерами 26, 23, 21 имеют, согласно табл. 3, самые низкие ранги информативности: 30, 29, 28 соответственно. Естественно предположить их слабый вклад в отличие почв и исключить из анализа, тем самым объединив корреляционный метод и метод учета вариабельности

Таблица 4

«Экзаменационная» выборка

№ признака	Индекс признака	Почва					
		I	II	III	IV	V	VI
1	H	1,24	0,62	0,19	0,34	0,59	0,23
2	C	6,81	2,46	0,57	0,76	1,90	0,62
3	N	0,38	0,27	0,03	0,08	0,25	0,33
4	R_{880}	16,4	19,5	40,9	27,5	35,5	39,5
5	ΔR	4,1	6,0	12,4	9,5	10,5	11,4
6	ΣR	11,3	13,8	30,3	18,8	25,1	29,4
7	D_{384}	1,15	0,14	0,16	0,07	0,18	0,15
8	$Fe_{пр}$	35	460	370	430	110	230
9	R_{ρ}	183	82	36	63	45	65
10	$Mn_{к}$	1	31	1	19	1	1
11	Гигр.	3,64	1,50	0,64	1,14	1,24	0,81
12	$pH_{сол}$	4,4	6,7	4,4	7,0	4,3	4,8
13	B	13	21	17,3	23,6	22,6	17,3
14	Mn	635	763	390	878	1253	1070
15	Cr	10	20,3	17,6	25,3	17,3	20,3
16	Cu	6,2	7,1	4,0	6,3	3,0	4,2
17	V	9,0	23,9	16,0	27,3	16,6	17,3
18	Ti	698	1700	1820	3026	1940	1650
19	Ni	13,3	23,6	15,3	22,6	16,3	17,6
20	Y	7,2	22	14	21	17	15
21	Zr	303	497	463	433	480	450
22	Yb	1,0	1,8	1,6	2,4	1,9	1,8
23	Sc	4,6	4,9	3,8	5,4	4,4	4,7
24	Co	2,3	4,4	3,3	5,1	7,6	5,2
25	Sr	7,7	81	81	80	82	84
26	Ba	46	57	45	51	52	53
27	Ca	29	66	54	50	45	48
28	Fe	66	103	104	105	121	126
29	Инв.	39,1	30,4	4,7	6,3	3,9	1,8
30	Кат.	27,7	12,7	3,8	9,3	4,7	4,3

Таблица 5

Показатели сходства 6 образцов «экзамена» с исходными объектами и их диагностика, %

№ образца	$Q = 30$						$Q = 5$					
	1	2	3	4	5	6	1	2	3	4	5	6
1	87	63	57	53	60	57	63	60	44	47	44	40
2	85	60	58	52	60	56	60	58	53	46	52	48
3	87	59	54	50	56	53	72	43	29	31	25	24
4	65	93	73	83	78	74	59	96	68	82	58	66
5	64	93	72	84	77	76	60	98	71	83	59	67
6	64	94	70	83	76	72	61	97	72	83	59	66
7	61	71	92	78	84	87	41	67	95	79	73	76
8	59	72	94	77	83	88	40	69	97	80	73	75
9	60	74	93	81	83	89	44	74	93	86	72	74
10	56	84	77	94	78	79	46	79	84	95	75	79
11	54	81	74	93	74	77	46	82	80	92	68	72
12	58	83	78	93	78	81	50	83	82	95	71	76
13	62	75	83	77	90	85	48	55	65	65	85	82
14	63	74	82	78	93	83	49	56	69	72	89	71
15	60	75	85	80	93	87	39	59	74	72	93	82
16	60	76	88	82	87	94	40	67	75	75	76	93
17	60	75	88	81	84	93	39	70	78	78	72	91
18	58	73	89	81	85	93	46	68	78	80	82	90
	87	94	94	94	93	94	72	98	97	95	93	93
Образец	3	6	8	10	15	16	3	5	8	10	15	16
Почва	I	II	III	IV	V	VI	I	II	III	IV	V	VI

для оценки информативности. Действительно, их исключение практически не повлияло на дендрограмму, а минимальная информативная комбинация стала включать 5 признаков: 8, 14, 16, 25, 29. Теоретически число признаков можно сократить до одного. В данном случае для сохранения общего вида дендрограммы оказалось возможным исключить и 25-ый признак (ранг 22 по табл. 3). Однако следует иметь в виду, что дендрограмма не является единственной целью анализа. Последнее исключение существенно исказило порядок величин в матрице показателей сходства, и тройки образцов из одних почвенных разностей уже имели между собой сходство порой даже меньше, чем с образцами других почв. Это могло привести к ошибкам диагностики образцов «экзамена», а потому за окончательную минимальную информативную комбинацию (систему информативных признаков) были приняты указанные 5 признаков.

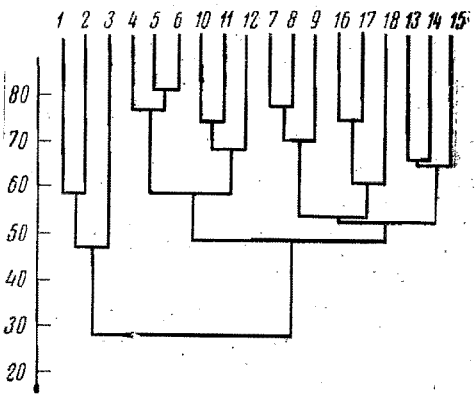

Рис. 3. Дендрограмма по минимальной комбинации признаков ($q=5$)

Нижний треугольник в табл. 2 составлен из показателей сходства 18 исходных объектов по 5 признакам, на рис. 3 приведена соответствующая дендрограмма. Как следует из табл. 2 и рис. 3, в преобладающем большинстве сходство всех объектов уменьшилось. Так, если при $q=30$ признакам объекты 1 и 3 имели 65% сходства, то при $q=5$ оно составило лишь 44% (это максимальная разница, отмеченная для образцов торфянистых почв). Наоборот, для пар объектов 4—5, 4—6 и 5—6 сходство увеличилось на 5%. Однако здесь важнее, чтобы сходство образцов одной почвы было существеннее, чем их сходство с образцами других почв. Это необходимо, как указывалось, для более точной диагностики новых объектов.

Для проведения диагностической классификации в качестве «экзамена» взято 6 новых образцов (по одному из каждой почвы), для которых определены те же 30 признаков (табл. 4). «Экзамен» необходим не только для проверки диагностирующей способности найденной иерархической структуры, но и для подтверждения обоснованности сделанной с ее помощью группировки. В связи с этим в табл. 5 полностью приведены результаты «экзамена». Из них следует, что по 30 признакам диагностика безошибочна, а группировка объектов обоснована: показатели сходства к соответствующим повторностям по почвам (в табл. 5 подчеркнуто) существенно выше, чем к другим образцам. В целом это справедливо и для 5 признаков, но здесь сходство образцов торфянистой почвы не вполне очевидно, а максимальная близость для пойменной луговой почвы сместилась с 6-го на 5-ый образец. Однако и здесь результаты в целом безошибочны, рассматривать ли их в среднем по почве или по наибольшему сходству с одним из образцов повторностей, хотя эту комбинацию в общем случае и следовало бы несколько расширить за счет включения дополнительных признаков.

Литература

1. *Бейли Н.* Математические методы в биологии и медицине. «Мир», 1970.
2. *Васильев В. И.* Распознающие системы. «Наукова думка», 1969.
3. *Верховская Л. А., Голубева В. А., Коган Р. И.* Выбор информативной комбинации признаков для различения двух геологических объектов. М., 1972.

4. *Дорофеюк А. А., Лумельский В. Я.* Реализация алгоритмов обучения распознаванию образов «без учителя» на ЭВМ. В сб.: Алгоритмы обучения распознаванию образов. «Сов. Радио», М., 1973.
5. *Мучник И. Б., Петренко Е. С.* Программы для решения задач распознавания образов методом потенциальных функций. В сб.: Алгоритмы обучения распознаванию образов. «Сов. Радно», М., 1973.
6. *Рожков В. А., Симакова М. С.* Статистическое исследование профилей дерново-подзолистых почв на покровных суглинках. Почвоведение, 1973, № 12.
7. *Рожков В. А.* Метод главных компонент и его применение в почвоведении. Почвоведение, 1975, № 10.
8. *Сонечкин Д. М.* Об объективной классификации метеорологических явлений и ситуаций с помощью ЭВМ. Метеорол. и гидрол. 1968, № 5.
9. *Фридланд В. М., Рожков В. А.* Применение математических методов и ЭВМ для классификации почв. Географический сб., № 5. М., ВИНИТИ, 1975.
10. *Bidwell O. W., Hole F. D.* Numerical taxonomy and soil classification. Soil Sci., 1964, v. 97, № 1.
11. *Bidwell O. W., Marcus L. F., Carkar P. K.* Numerical classification of soils by electronic computer. Intern. longr. Soil Sci. Trans. 8th (Bucharest), 1964, v. V, p. 251—252.
12. *Campbell N. A., Mulcahy N. G., McArthu W. M.* Numerical classification of soil profiles on the basis of field morphological properties. Austral. J. Soil Res., 1970, v. 8, № 1.
13. *Cuanalo De La C. H. E., Webster R.* A comparative study of numerical classification and ordination of soil profiles in a locality near Oxford. J. Soil Sci., 1970, v. 21, № 2.
14. *Grigal D. F., Arneman H. F.* Numerical classification of some forest Minnesota soils. Soil Sci. Soc. Amer. Proc., 1969, v. 33, № 3.
15. *Sarkar P. K., Bidwell O. W., Marcus L. F.* Selection of caracteristics for Numerical classification of soils. Soil Sci. Soc. Amer. Proc., 1966, v. 30, № 2.
16. *Sokal R. R., Sneath P. H. A.* Principle of numerical taxonomy. San Francisco, 1963.
17. *Rayner J. H.* Numerical methods. J. Soil Sci., 1966, v. 17, № 1.
18. *Rayner J. H.* Numerical approach to soil systematics. Soil Ecosyst. London, 1969, p. 31—38.
19. *Russell J. S., Moore A. W.* Use of a numerical method in determining affinities between some deep sandy soils. Geoderma, 1967, № 1, p. 47—68.

Почвенный ин-т
им. В. В. Докучаева

Дата поступления
22.XII.1975 г.

V. A. ROZHKOV, N. V. PROSHINA

A TENTATIVE NUMERICAL SOIL TAXONOMY

The well-known work (P. K. Sarkar a. oth., 1966) on the use of dendrograms and the elimination of noninformative characteristics is developed in this paper. The broadening of the considered approach is accomplished by including various methods of dendrogram construction, by evaluating the informativity of the characteristics, and by completing the study of new objects by means of a diagnistic classification.