УДК 631.4
doi: 10.31251/pos.v1i3.30

Check for updates

ОЦЕНКА ИНФОРМАТИВНОСТИ ПОЧВЕННЫХ ПРИЗНАКОВ

© 2018 В.А. Рожков

Адрес: ФГБНУ Почвенный институт им. В.В. Докучаева, Пыжевский перулок, 7, стр.2, г.Москва, 109017, Россия. E-mail: rva39@mail.ru

В задачах классификации информативность признаков означает относительный вклад каждого из них в разделение почв, горизонтов, образцов. Чем в большей мере данный признак почвы отличает ее от другой, тем он более информативен. Высокая вариация значений признака может скрывать неоднородность выборки и потенциальное деление ее на группы объектов. Но есть и более сложные многомерные оценки, обладающие свойствами критериев, использующие не только варьирование значений признаков, но и их взаимосвязь между собой. Это метод главных компонент, многомерные дисперсионный, кластерный и другие статистические анализы. Формализация концепций, определений и методов классификаций почв является наиболее актуальной проблемой почвоведения. От создания многочисленных списков почв пора приступать к научному подходу на основе числовых информационных технологий и математики. Количественная оценка информативности почвенных признаков является первоочередной задачей на этом пути.

***Ключевые слова**: информативность признаков; система информативности признаков; показатели и критерии информативности; численная классификация почв; формализация классификации.*

***Цитирование:** Рожков В.А. Оценка информативности почвенных признаков // Почвы и окружающая среда. 2018. № 1(3). С.143 – 150.*

ВВЕДЕНИЕ

Проблема выбора признаков (свойств, состава, отношений и др. показателей) объектов классификации и их классов – по составу, числу и шкалам - является принципиальной в задачах классификации и классифицирования (распознавания). Основная идея направлена на формализацию описаний и организацию единого пространства признаков, определяющих тот или иной таксономический уровень.

По существу, речь идет о формировании образа объекта исследования – горизонта, профиля, почвы, структуры почвенного покрова. Система информативных признаков (СИП) обеспечивает построение классификации, соответствующей поставленным целям.

Доступные коммерческие пакеты программ типа STATISTICA недостаточно полно и четко обращают внимание пользователей на учете шкал признаков, определяющих допустимые виды формальных операций и методов статистической обработки данных, что снижает достоверность результатов и выводов. Поэтому приводится таблица 1 описаний шкал, чтобы сократить время их поиска в других публикациях.

Э.Мах (1838-1916) – идеолог монизма – выдвинул идею сокращения пространства показателей: «Задача науки – искать константу в естественных явлениях, способ их связи и взаимозависимости. Ясное и полное научное описание делает бесполезным повторный опыт, экономит тем самым на мышлении. При выявленной взаимозависимости двух феноменов, наблюдение одного делает ненужным наблюдение другого, определенного первым. Также и в описании может быть сэкономлен труд благодаря методам, позволяющим описывать один раз и кратчайшим путем наибольшее количество фактов… Всякая наука имеет целью заменить, т.е. сэкономить опыт, мысленно репродуцируя и предвосхищая факты…» (Новейший философский словарь, 2001, с. 608).

Кроме экономии на анализах минимизация описаний несет важную методологическую функцию. У.Р. Эшби отмечает: «…если человек не справляется с огромными потоками информации, то выход надо искать не только в увеличении памяти электронно-вычислительных машин. Нужно идти по пути Ньютона: искать обобщения, находить способы компактного выражения информации об окружающем мире, опираясь на объективные законы природы» (Лук, 1965, с. 11).

Вопрос состоит в том, как оценить информативность каждого признака, чтобы построить их минимальную информативную комбинацию.

Таблица 1

Шкалы значений признаков (Рожков, 2011)

Названия шкалы	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, индексы 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) асимметрия, 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

* 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

ОСНОВНЫЕ АЛГОРИТМЫ

Описываются наиболее апробированные и доступные методы оценок для создания систем информативности признаков.

Коэффициент вариации может оказаться хорошим индикатором неоднородности выборки и существования классов объектов. В почвоведении использовался корреляционный метод оценки информативности показателей: как указывал Max, из двух высоко коррелирующих признаков достаточно оставить один, другой не добавляет информации.

О тесноте сопряженности признаков можно судить по дендрограмме их сходства. Важно, что дендрограммы, в отличие от корреляции, можно рассчитывать и для качественных признаков, т.е. номинальных, бинарных, порядковых и пр.

Некоторое указание дает простая вариация значений признаков: слабая вариация означает слабую относительную информативность и наоборот.

$$V = S/M$$

где S – среднее квадратическое отклонение, а M – среднее арифметическое значение почвенного признака. Задается в % или в долях единицы.

Предполагается, что чем больше величина коэффициента, тем информативнее признак. Опыт показывает, что при $V \leq 30\%$ признак можно считать однородным, обычно с нормальным распределением. Вариация $V < 100\%$ свидетельствует об одновершинном распределении значений признака. И только при $V > 100\%$ кривая распределения многовершинная, а, значит, вероятно разделение классов по этому признаку. Показатель информативности, основанный на коэффициенте вариации, характеризует внутренние свойства признаков, а расчеты выполняют только в арифметических шкалах (Рожков, 2011).

Мерой информативности может служить коэффициент корреляции между признаками:

$$Rjl = \frac{1}{(n-1)Sj*Sl} \sum_{n} (xij - Mj) * (xil-Ml)$$

где Xij - i-тое значение j-го признака.

Из двух сильно коррелирующих признаков целесообразно оставить лишь один, не имеющий высоких корреляций с другими признаками почв. Этот метод также применим лишь в арифметической шкале.

Для демонстрации некоторых других расчетов в таблице 2 приведены реальные почвенные данные в арифметической шкале, хотя алгоритмы последующих расчетов применимы для любых шкал (Рожков, Симакова, 1974).

Таблица 2

Характеристика образцов из профиля дерново-подзолистой (Пд) почвы

№№ п.п.	Горизонты	Значения признаков						
		pH	Гумус %	ГК, мг-экв	ил	физ.глина	Вынос, %	
					%		ила	Ca-Mg
		1	*2*	*3*	*4*	*5*	*6*	*7*
1	$A_{пах}$	5,0	1,8	4,0	13	30	-56	-46
2		5,0	2,2	4,6	13	35	-54	-89
3		5,0	1,8	3,8	9	40	-62	-52
4	A_2	4,8	0,4	2,6	9	30	-69	-76
5		4,4	0,3	3,1	14	31	-52	-76
6		5,4	0,5	2,2	7	34	-69	-57
7	A_2B	4,2	0,5	4,2	26	44	-10	-42
8		4,1	0,4	5,1	39	54	36	-25
9		4,9	0,5	4,7	35	54	3	-33
10	B	4,7	0,3	2,6	33	52	17	-18
11		4,2	0,4	4,4	32	48	1	-9
12		4,3	0,4	4,3	31	49	16	-17

На рисунке 1 приведена дендрограмма сходства этих объектов.

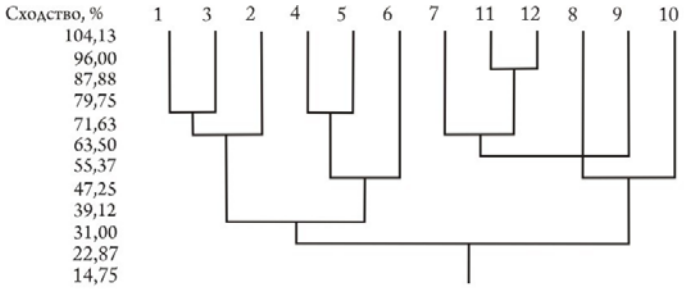
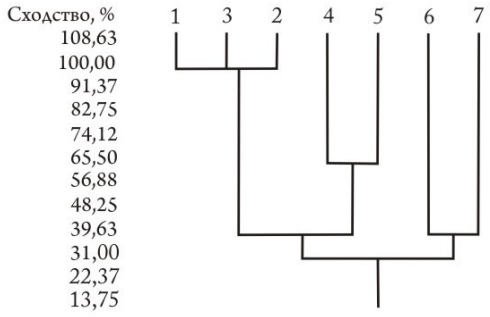


Рисунок 1. Дендрограмма сходства описаний образцов Пд почвы

Четко выделяются группы образцов из горизонтов А (1-3) и A_2 (4-6). Горизонты A_2B и В перемешаны, что вполне объяснимо трудностью отбора «чистых» образцов.

Структура этих двенадцати образцов, определенная своеобразием семи их описаний, представлена на рисунке 2.



Сходство объектов			
i	1	2	%
1	1	3	100
2	1	2	100
3	4	5	69
4	6	7	47
5	1	4	46
6	1	6	31

Рисунок 2. Сходство 7 описаний 12 образцов из генетических горизонтов Пд почвы

Свойства под номерами 1-3 (pH, гумус и ГК - гидролитическая кислотность) характеризуют горизонты одинаково, потому что их сходство рано 100%. Предположительно из них можно сохранить только одно, а какое – решение за исследователем или за применением дополнительных методов.

Достоинство этого метода в том, что он применим для значений признаков в любой шкале – от номинальной до абсолютной.

Метод главных компонент (МГК) позволяет оценить обобщенную информационную нагрузку признаков. Можно исключить признаки, не проявившие существенного вклада в разнообразие объектов. Однако это возможно, когда имеет место дифференциация выборки на классы. В противном случае веса признаков будут примерно одного порядка.

Метод выдает статистическую характеристику признаков, корреляционную матрицу описаний, собственные числа и вектора, а также распределение образцов в пространстве ГК.

i	**Среднее**	**Среднее квадратическое отклонение**
1	4,7	0,4
2	0,8	0,7
3	3,8	0,9
4	21,7	11,9
5	41,8	9,5
6	-2 ,9	38,9
7	-43,3	23,4

Корреляционная матрица признаков:
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***

Собственные числа – это дисперсии свойств на ГК1 и ГК2. Их сумма более 70% от общей, что обещает четкое разделение групп образцов.

i	Собственные числа	%
1	4,69	67
2	1,26	85

Собственные вектора корреляционной матрицы:

ГК1	-0,34	-0,24	0,26	***0,45***	***0,43***	***0,45***	***0,40***
ГК2	0,09	*-0,73*	-0,67	0,001	-0,04	0,02	-0,04

Примечание: **жирным шрифтов** выделены информативные ГК1, *подчеркнутый курсив* – информативные ГК2

Параметры новых координат ГК1 и ГК2 являются весами признаков и указывают (выделены жирным курсивом) наиболее информативные из них. На ГК1 это содержание ила, физической глины и вынос ила и Ca^{+2}+Mg^{+2} (%). На ГК2 наиболее информативным оказался гумус (выделен подчеркнутым курсивом).

Следовательно, из трех признаков (1-3), на 100% сходных на дендрограмме рис. 2, следует оставить второй (гумус), а исключить признаки 1 и 3. Именно выделенные признаки с наибольшими весами определили распределение образцов в плоскости первых двух главных компонент (рис. 3):

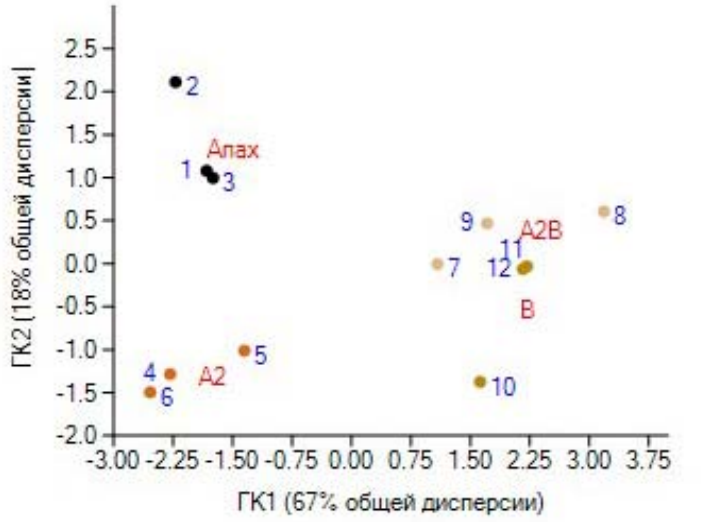


Рисунок 3. Образцы из таблицы 2 в плоскости первых двух главных компонент

Картина полностью согласуется с дендрограммой свойств (рис. 2): компактные группы 1-3, 4-6 и более размытые 7-9 и 10-12.

Метод главных компонент включает в расчет и обобщает сочетание двух выше приведенных подходов при выборе наиболее информативных свойств почв при создании СИП.

Самый простой метод оценки информативности признаков многомерных объектов основан на очевидности. На рисунке 4 представлены дендрограммы объектов, построенные по убывающим количествам признаков. На нем видно, что сокращение числа показателей с 38 до 13 не изменяет вида дендрограммы, следовательно, исключенные показатели не несут информации о структуре взаимоотношений множества объектов. Дальнейшее исключение признаков ведет к резкой дезорганизации этой структуры, т.е. исключение признаков необходимо прекратить.

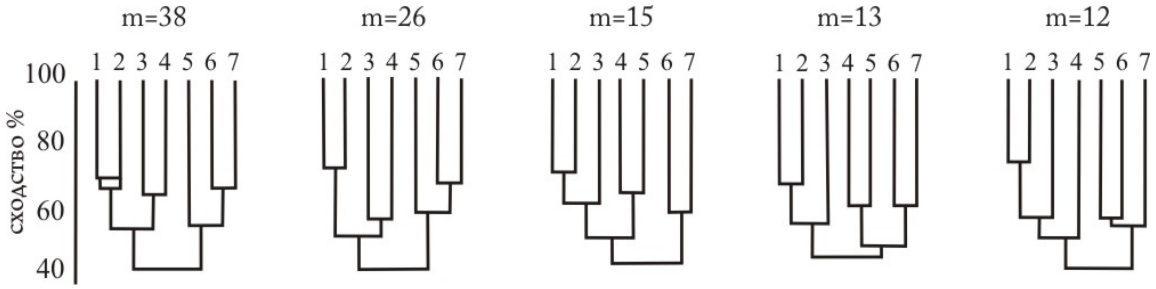


Рисунок 4. Оценка информативности признаков методом исключения

Если позволяет объем исходных данных (число объектов существенно больше числа признаков), то применимы средства многомерной статистики.

Для проверки отношений классов объектов используют многомерные статистические критерии. Один из них основан на сопоставлении расстояний Махаланобиса между классами по полному и сокращенному набору признаков (Рао, 1968; Рожков, Симакова, 1973):

$$F = \frac{(n1 * n2 - 1) * n1 * n2 * (D2p - D2q)}{q - p * (n1 + n2)(n1+n2-2)+n1 * n2 * D2p}$$

где F-критерий **имеет f1 = q - p** и **f2=n1+n2-q-1** степеней свободы; n1 и n2 - число объектов в сопоставляемых классах; q и p -исходное и сокращенное число признаков; (q > p);

где D2q и D2p -расстояния Махаланобиса: D2q = (Mq - Mp)' S-1 * (Mq - Mp).

Если F ≤ Fα, f1, f2, то это означает, что исключение признака не привело к потере информации. Иначе исключение признаков прекращается.

В основу следующего метода положен критерий потери информации, рассчитываемым с использованием многомерного дисперсионного анализа (Рожков, 2011). Решением является оценка потери информации сопоставлением варьирования (сходства) признаков по полному и сокращенному набору показателей:

$$\chi^2 f = - n (p+k)/2 * \ln (\lambda q / \lambda p)$$

где f = p(k-1), k – число классов; λq = |W|/|T| - отношение определителей матриц внутри- и межклассового варьирования (сходства) q признаков.

Эти подходы требуют значительного объема данных – по крайней мере, превосходства числа объектов над числом выбранных признаков.

ОБСУЖДЕНИЕ И ДОПОЛНЕНИЕ РУЗУЛЬТАТОВ

Качество классификаций также может служить мерой информативности признаков. Наиболее распространенным из них является отношение среднего внутриклассового сходства объектов к межклассовому: чем оно больше, тем более четкое разделение объектов на классы (Рожков, 2011).

Сравнение двух ординатных классификаций проводится с помощью полихорического показателя связи Чупрова по количеству совпадающих объектов по классам обоих разбиений (Рожков, 1989). Сравнение иерархических классификаций (дендрограмм) проводится методом, предложенным другими исследователями ранее (Sokal, Rohlf, 1962).

Сложность алгоритмов и требований, отсутствие доступных компьютерных программ, видимо, объясняют недостаточно широкое приложение этих методов, хотя они позволяют порой в 2-5 раз уменьшить затраты на лабораторные анализы почв.

В реальных задачах удавалось сократить пространство признаков в зависимости от условий на 40-80%. Кроме экономии на анализах минимизация описаний несет важную методологическую функцию экономии мышления и генерации новых идей.

Сочетание коэффициента вариации, корреляции и ГК позволило сократить число признаков в распознавании почв поймы Москва-реки на 75% (с 30 до 5) (Рожков, Прошина, 1977, с. 106-116). Также была показана возможность сокращения числа признаков при автоматической классификации почв поймы р. Оби на 46 %. (Шеремет, Рожков, Афанасьева, 1981).

Примеры применения разных приемов создания СИП излагались также в ранее опубликованных работах (Рожков, Симакова, 1973; Рожков, Симакова, Юшкевич, 1988; и др.).

ЗАКЛЮЧЕНИЕ

Информативность признаков является относительной в том смысле, что конкретные классы объектов различаются с точностью до данного набора признаков и не обладают таким свойством при другом их наборе.

Исключение малоинформативных показателей имеет сегодня материальную основу: анализы стали довольно дорогими, а опыт показывает, что объем анализов можно сократить более чем вдвое. Избыточны привычные всем таблицы горизонт-свойства, однако еще только в будущем предстоит нашей науке освоить цифровую культуру в исследованиях.

Применение математических методов требует четких формулировок и конкретной постановки задачи, что само по себе оказывает влияние на почвоведа, заставляя его более глубоко осмыслить решаемую проблему. Применение математики - это не просто использование количественных методов, а, главное, строгий язык, стиль мышления. Сбор данных должен быть подчинен требованиям последующей обработки математическими методами.

Почвенный институт им. В.В. Докучаева поможет в освоении методов формализации классификации, как развития традиционных подходов составления экспертных списков почв. Институт заинтересован в распространении своих программных средств с целью их внедрения и дальнейшей апробации для создания полностью формализованной классификации почв России.

ЛИТЕРАТУРА

1. *Лук А.* Формула гениальности // *Знание – сила. 1965. № 3. с. 11.*
2. Новейший философский словарь. Минск: Интерпрессервис; Книжный дом, 2001. 1280 с.
3. *Рао С.Р.* Линейные статистические методы и их применение. М.: Наука. 1968. 548 с.
4. *Рожков В.А.* Почвенная информатика. М.: Агропромиздат, 1989. 222 с.
5. *Рожков В.А.* Формальный аппарат классификации почв // *Почвоведение. 2011. № 12 с. 1411-1424.*
6. *Рожков В.А., Прошина Н.В.* Опыт численной таксономии почв // *Почвоведение. 1977. № 8. С. 106-116.*
7. *Рожков В.А., Симакова М.С.* Статистическое исследование профилей дерново-подзолистых почв на покровных суглинках // *Почвоведение. 1973. № 12. С. 110-120.*
8. *Рожков В.А., Симакова М.С., Юшкевич С.Х.* Таксономический анализ почв методами численной классификации // *Почвоведение. 1988. № 5. С.7-14.*
9. *Шеремет Б.В., Рожков В.А., Афанасьева Т.В.* Применение математических методов для классификации и диагностики почв поймы Средней Оби // *Вестник МГУ. Сер. 17. Почвоведение. 1981. № 1. С. 11-20.*
10. *Sokal R.R., Rohlf F.I.* The comparison of dendrogramms by the objective methods // *TAXON. 1962. № 11. P. 33-40.* doi: 10.2307/1217208.

Поступила в редакцию 27.09.2018
Принята 08.12.2018
Опубликована 09.12.2018

Сведения об авторе:
Рожков Вячеслав Александрович – член-корреспондент РАН, главный научный сотрудник Почвенного интитута им.В.В. Докучаева РАН (Москва, Россия), *rva39@mail.ru*

Автор прочитал и одобрил окончательный вариант рукописи

ESTIMATION OF INFORMATION CONTAINED IN SOIL ATTRIBUTES

© 2018 V.A. Rozhkov

Address: V.V. Dokuchaev Soil Institute, Pyzhevskyi pereulok 7, s.2, Moscow, 109017, Russia.
E-mail: rva39@mail.ru

In classification tasks the information contained in particular soil attributes denotes the relative contribution of each of the attributes into soils, horizons and samples discrimination. The bigger such contribution of a given attribute is, the more information the attribute contains. High variation of its values can mask the heterogeneity of the data set, which may potentially consist of several groups of objects. At the same time there are more complex multivariate estimates with the properties required for criteria, which use not only attribute variation, but also their interrelationship. These methods are principle components analysis, multivariate analysis of variance and cluster analysis. Formalization of concepts, definitions and classification techniques is the most pressing issuein soil science, as the mere listing of soils should be substituted by the approach based on informatics technologies and mathematics. Quantitative estimation of the information contained in soil attributes is the primary task in this direction.

Keywords: *information contained in soil attributes; the system of information in attributes; indicators and criteria of information; numerical soil classification; formalization of soil classification*

How to cite: *Rozhkov V.A. Estimation of information contained in soil attributes // The Journal of Soils and Environment. 2018. 1(3): 143 – 150. (in Russian with English abstract).*

REFERENCES

1. *Louk A.* The formula of geniality, *Knowledge is power, 1965, Iss. 3, p. 11*. (in Russian)
2. Newest philosophical dictionary. Minsk: Interpressservice; Book House, 2001. 1280 p. (in Russian)
3. *Rao S.R.* Linear statistical methods and their application. Moscow: Nauka Pubs., 1968. 548 p. (in Russian)
4. *Rozhkov V.A.* Soil informatics. Moscow: Agroppromizdat Pubs., 1989. 222 p. (in Russian)
5. *Rozhkov V.A.* Formal apparatus of soil classification, *Eurasian Soil Science, 2011, Iss.12, p.1289 -1303*. doi: 10.1134/S1064229311120106.

6. *Rozhkov V.A., Proshina N.V.* Test on numerical soil classification, *Pochvovedenie, 1977, Iss. 8, p. 106-116.* (in Russian)
7. *Rozhkov V.A., Simakova M.S.* Statistical investigation of sod-podzolic soil profiles developed on clay loams, *Pochvovedenie, 1973, Iss. 12, p. 110-120*. (in Russian)
8. *Rozhkov V.A., Simakova M.S.,Yushkevich S.H.* Taxonomical analysis of soils by numerical classification methods, *Pochvovedenie, 1988, Iss.5, p.7-14.* (in Russian)
9. *Sheremet B.V., Rozhkov V.A., Afanasieva T.V.* The application of mathematical methods for classification and diagnostics of soils in the floodplain of the mid-Ob River, *Moscow University Soil Science Bulletin. Biological series, 1981, Iss. 1, p. 11-20.* (in Russian)
10. *Sokal R.R., Rohlf F.I.* The comparison of dendrograms by the objective methods, *TAXON, 1962, Iss.11, p. 33-40.* doi: 10.2307/1217208.

Received 27 September 2018
Accepted 08 December 2018
Published 09 December 2018

About the author:
Rozhkov Vyacheslav A. – Corresponding Member of the Russian Academy of Sciences, Principal Researcher in the Dokuchaev Soil Institute of the Russian Academy of Sciences (Moscow, Russia); *rva39@mail.ru*

The author read and approved the final manuscript