ISSN 1064-2293, Eurasian Soil Science, 2011, Vol. 44, No. 12, pp. 1289–1303. © Pleiades Publishing, Ltd., 2011.
Original Russian Text © V.A. Rozhkov, 2011, published in Pochvovedenie, 2011, No. 12, pp. 1411–1424.

GENESIS AND GEOGRAPHY OF SOILS

Formal Apparatus of Soil Classification

V. A. Rozhkov

Dokuchaev Soil Science Institute, Russian Academy of Agricultural Sciences, per. Pyzhevskii 7, Moscow, 119017 Russia
E-mail: rva39@mail.ru

Received December 20, 2010

Abstract—Mathematical tools that may be applied for soil classification purposes are discussed. They include the evaluation of information contained in particular soil attributes, the grouping of soil objects into a given (automatically determined) number of classes, the optimization of the classification decisions, and the development of the models and rules (algorithms) used to classify soil objects. The algorithms of multivariate statistical methods and cluster analysis used for solving these problems are described. The major attention is paid to the development of the systems of informative attributes of soil objects and their classes and to the assessment of the quality of the classification decisions. Particular examples of the solution of the problems of soil classification with the use of formal mathematical methods are given. It is argued that the theoretical and practical problems of classification in science cannot find objective solutions without the application of the modern methods of information analysis. The major problems of the numerical taxonomy of the soil objects described in this paper and the appropriate software tools for their solution should serve as the basis for the creation of not only formal soil classification systems but also the theory of soil classification.

DOI: 10.1134/S1064229311120106

INTRODUCTION

The introduction of mathematical methods in biology is credited to M. Adanson's work *Families of Plants* (1763) [7]. In this classical study, Adanson described the principles of plant classification on the basis of the degree of similarity (frequency of coincident characteristics) between related plant taxa. Statistical methods are widely used for soil classification purposes. However, the methods of multivariate statistics have limited use, though they can suggest efficient algorithms for image identification and cluster analysis. The idea of numerical taxonomy was developed in the 1960s. The theory and methods of numerical taxonomy are described in a special monograph [35]. Later, generalizing monographs [21, 32, 36] and numerous papers on the problems of numerical taxonomy in soil science were published. It should be noted that the advances of numerical taxonomy in other sciences, including biology and geology, often found their application in soil science.

The problems of soil classification were actively discussed in the 1960s–1970s. At present, a large number of publications are devoted to the problem of the correlation between different soil classification systems. In such publications, the lists of soil names in a given classification are compared with analogous lists in other classification systems. This is an important problem of a "common language" in soil science. However, the principles and goals of soil classification are more important. The particular classification schemes and soil names may change, whereas the principles of the classification will remain intact.

The necessity to introduce mathematical methods for solving the problems of soil classification is widely recognized. To popularize this approach and the use of numerical taxonomy, pedologists refer to the words of A. Whitehead (1925), who argued that a transition from classification to mathematics is a necessary step in the cognition of various phenomena [8]. As noted by Baily (cited from [3]), the scientific value of classifications that have no quantitative basis is arguable. The theory of such classifications is entirely intuitive; it lies in the domain of art rather than science. The notion of numerical taxonomy implies the assessment of the relations or the degree of similarity between taxonomic units by numerical methods and the use of these relations for arranging the units in a taxonomic order [35]. Taxonomic units are understood as the particular objects and their classes (taxa). In essence, numerical taxonomy is "a subdivision of a given set of multidimensional objects into classes, so that each class is represented by an isolated group of points in the space of selected parameters" [2]. Thus, numerical taxonomy can be perceived as a formally (with the use of mathematics) constructed taxonomy allowing the identification of objects on the basis of their characters.

The logic of classification decisions and the requirements for classification systems in dependence on their functions are developed fairly well. However, there are no definite rules to construct soil classifica-

tion systems. At the same time, the experience in this field makes it possible to suggest a sequence of necessary steps in the development of numerical soil taxonomy with the use of the already tested methods. Such a sequence should serve as a framework for the development of the theory of soil classification.

The active development of remote sensing methods and digital soil mapping with the use of geographic information systems has served as a new stimulus for the application of mathematical methods in soil science [12, 25]. The major attention is focused on the elaboration of adequate technical equipment and on the unification of procedures for the automated treatment of the corresponding data. Less attention is paid to the diverse methods of data analysis used in soil science. It should be noted that the methods to evaluate the quality and information capacity of digital cartographic materials differ from the methods of data valuation that are traditionally used in soil science. It is a challenge for pedologists not just to use the advantages of geoinformation technologies and procedures for automated image recognition but also to introduce their own efficient methods to discriminate between different soil objects. The purpose of this paper is to characterize the main algorithms that can be used to classify different soil objects, including aggregates, thin sections, soil horizons, and soil profiles.

FORMALIZATION OF SOIL CLASSIFICATION

The term classification has three meanings: (a) the process of the creation of a classification, (b) the classification system (the result of this process), and (c) the procedure of the use of this system for the identification of particular soil objects (or soil correlation).

Formally, the logic of classification can be described in terms of the set theory. The set of objects A is subdivided into classes $A = A_j$, where $j = 1, 2, ..., k$ (k is the number of classes) in such a way that (1) $A_j \neq \emptyset$, which means that all the classes are not empty (they contain at least one object); (2) $A_j \cap A_l = \varnothing$ (where $i, j = 1, 2, ..., k$, and $j \neq i$), which means that the classes do not intersect (they have no common objects); and (3) $A_j = A$, which means that the union of classes gives us the initial set of objects.

Separate classes represent equivalence classes. Such classes have the properties of reflectivity (xRx), symmetry ($xRy \longrightarrow yRx$), and transitivity (xRy & $yRz \longrightarrow xRz$, where R denotes some relations (similarity, difference, likeness, etc.). Thus, a classification system is a system of equivalence classes.

Typological regionalization is also characterized by equivalency relations, though individual regionalization does not possess the property of transitivity; i.e., (xRy & yRz) do not necessarily mean (xRz) (the neighbor of my neighbor is not necessarily my neighbor).

The creation of any classification system begins from the formulation of its goals. These goals should be clearly specified and should not be self-contradictory; philosophical "universal" goals are not appropriate in this case. The goals of classification are taken into account upon the choice of soil characteristics. From the entire diversity of these characteristics, those that fit the preset goal and reflect the existing notions in the best way should be chosen. The formulation of the goal of classification is a subjective process. To limit this subjectivity, we need to specify this goal very precisely. The system of soil characteristics (attributes) used in the classification should be adequate for the conceptual model of the object of the classification. To achieve success, it is important to exclude the diversity of different aspects that may be potentially reflected in the classification. In this aspect, a soil classification differs from a soil database. A soil classification is based on information highly relevant to a particular goal, whereas a soil database may be designed to satisfy a broad circle of different aspects that might interest particular users. In other words, a classification can be considered a compact information system containing the maximum information about the particular classes of soil objects (soils, soil horizons, soil samples, thin sections, aggregates, etc.) in the space of their specially chosen attributes. These attributes may include information on the morphology and composition of soil objects, and information about the factors of soil formation contained in the description of these objects may also be included.

The particular values of soil attributes can be characterized with the help of different scales. The theory of scales for the description of various objects and its particular applications are considered in the general theory of measurements [14] and in particular fields of science [8]. One of the basic requirements for the design of such scales is the possibility to transform the values of the described attributes into some characters that allow arithmetic operations with them and may be processed by mathematical methods. The possibility of such a transformation is tightly related to the possibility to predict some attributes from the values of other attributes. The transformation is possible if it does not disturb the prediction. For a particular attribute, this means that the transformation does not change its major characteristics, so that the relations between the particular values of the attribute are preserved after some arithmetic operations with them.

Table 1 summarizes data on the types of scales that can be used for soil attributes and on the operations that can be performed with them. The types of soil attributes have been widely discussed in the literature [9, 17, 21], and there is no need for additional comments on this table. It should be stressed, however, that the types of scales should be taken into account for the

Table 1. Scales of attribute values

Type of scale	Allowable procedures							Examples
	transformation*	operations**					statistical treatment	
		1	2	3	4	5		
Nominal scale (classification scale)	One-to-one	+	–	–	–	–	(1) Frequency distribution, (2) determination of modal class	Color, structure, and names of soils and horizons; shape of boundaries; etc.
Ordinal	Monotonous continuous	+	+	–	–	–	(1), (2), (3) estimation of the median, (4) estimation of centiles, (5) rank correlation	Degree of podzolization, degree of soil cultivation, soil water content, soil bulk density, etc.
Interval	$y(x) = ax + b$ $a > 0$	+	+	+	–	–	(1–5), (6) estimation of the expectation, (7) dispersion, (8) asymmetry, (9) moments	Temperature, absolute age, etc.
Differential	$y(x) = ax + b$ $a = 1$	+	+	+	+	–	1–9	Soil characteristics determined from the difference in the sums of some indices
Relational	$y(x) = ax$ $a > 0$	+	+	+	–	+	All the methods	Depth, thickness
Absolute	$y(x) = x$ $a = 1$	+	+	+	+	+	All the methods	Number of samples, number of horizons, etc.

Notes: * Allowable changes in the values of the attributes and their use as arguments in equations.

** (1) Equal (=), unequal (≠); (2) more (>), less (<); (3) $(x_1 - x_3)/(x_2 - x_3)$; (4) $(x_1 - x_2)$; and (5) x_1/x_2, where x_1 is the value of the attribute.

appropriate choice of the method of mathematical treatment of the data.

The information capacity (informativeness) of the attributes used in a classification means their capacity to separate a given object (or class) from other objects (classes). The assessment of this capacity may be performed for the cases when the particular classes of objects are unknown, as well as for the cases when they are preliminarily specified. In the first case, for a multidimensional sample $X = x_{ij}$, where $i = 1, 2,..., n$ (n is the number of objects), and $j = 1, 2, ..., m$ (m is the number of attributes), the method of principal components can be applied [17, 24]. The eigenvalues (λ_{lj}) and eigenvectors (υ_{lj}, where $l = 1, 2,..., k$ is the number of the determined values and vectors) of the correlation matrix of the sample are calculated. This method can be illustrated by the following example.

The description of a profile of a podzolic soil contains seven attributes (the pH, the carbon content, the acidity, the contents of clay and physical clay, and the characteristics of the removal of clay and the removal of Ca + Mg) determined for the four major horizons (Ap, A2, A2B, and B). A fragment of the results of the calculations is shown in Fig. 1.

The first principal component describes 67% of the variance of the attributes in the sample. The highest values in the first eigenvector correspond to the four latter attributes of the soil (particle-size distribution data on the soil horizons). Hence, these attributes specify the differentiation of the horizons along the ordinate axis (the projection of the attributes of the horizons onto the first principal component). In the second principal component, the highest role is played by the carbon content; it mainly (0.73) specifies the differentiation of the horizons along the abscissa axis.

In the further analysis, only the most informative attributes can be used, though some information may be lost in this case. This approach has already been tested in various problems of economics [1]. The visible differentiation of the objects in the plot can be used as a criterion of the quality of the analysis. It can be seen that the Ap (samples 1–3) and A2 (samples 4–6) horizons are clearly separated, whereas the A2B (samples 7–9) and B (samples 10–12) horizons partly overlap. The dendrogram developed for these objects (Fig. 2) shows the same regularities in a more distinct manner.

The grouping of the horizons with respect to their similarity on the dendrogram is analogous to that in the coordinates of the principal components. Figures 1 and 2 illustrate the ordinate and hierarchical classifications, respectively. Dendrograms represent one of the forms of visualization of the results of classifications. In the development of a dendrogram, it is important to take into account the scales used to describe the attributes, because these scales specify the choice of the measure of similarity [21].

Dendrograms can also be used for the visual assessment of the informativeness of the attributes (Fig. 3).

Correlation matrix of the properties of soddy-podzolic soils:

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	Eigenvalues of the matrix	Principal component loads, %
1	4.69	67.0
2	1.26	85.0

Eigenvectors of the matrix:

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

Position of the soil samples in the coordinates of the principal components:

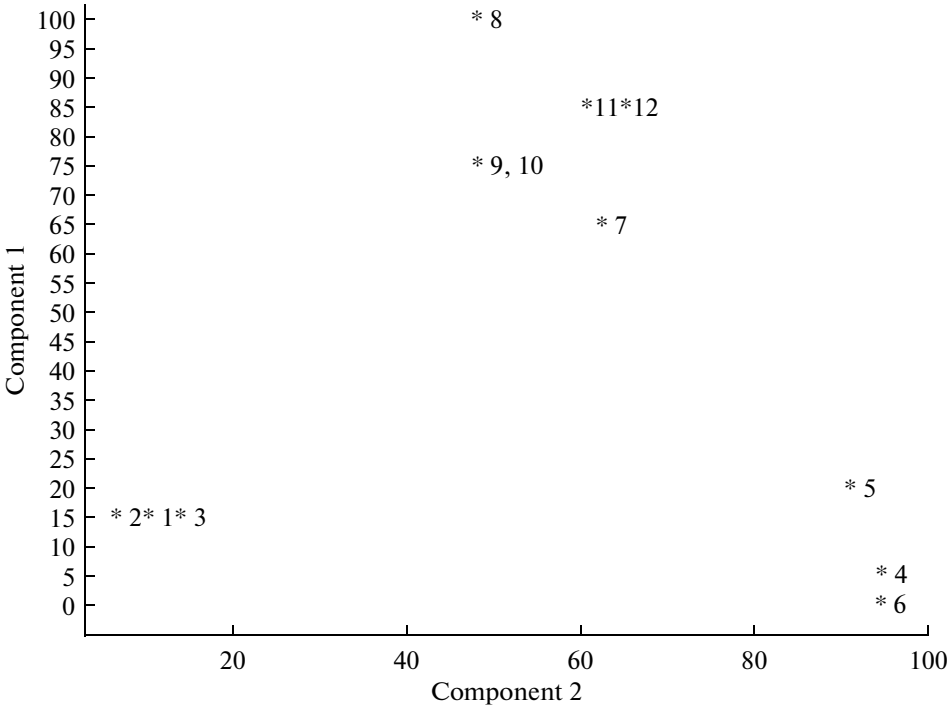

Fig. 1. The principal component analysis of data on 12 samples of soddy-podzolic soils characterized by 7 attributes; samples (1–3) are from the A1 horizon; (4–6), from the A2 horizon; (7–9), from the A2B horizon; and (10–12), from the B horizon.

The dendrogram of similarity developed for several objects on the basis of data on 38 attributes does not change upon the exclusion of the low-informative attributes until only 12 of the attributes are left. However, for large samples, this approach can be replaced by a more stringent comparison of the hierarchical structures.

When we have no information on the classes of objects, the determination of the low-informative attributes for their exclusion from the analysis is performed by relatively nonrigorous methods. Thus, we may take into account the degree of variation of the values of a given attribute. The low variation corresponds to the low informativeness of the attribute and vice versa. Another approach is based on the analysis of the correlation between the analyzed attributes. If we have two tightly correlated attributes, we may leave only one of them for the classification procedure, because the addition of the second attribute does not increase the amount of information. Dendrograms of similarity (Fig. 3) can also be used to control the possibility of the exclusion of some attributes. If two

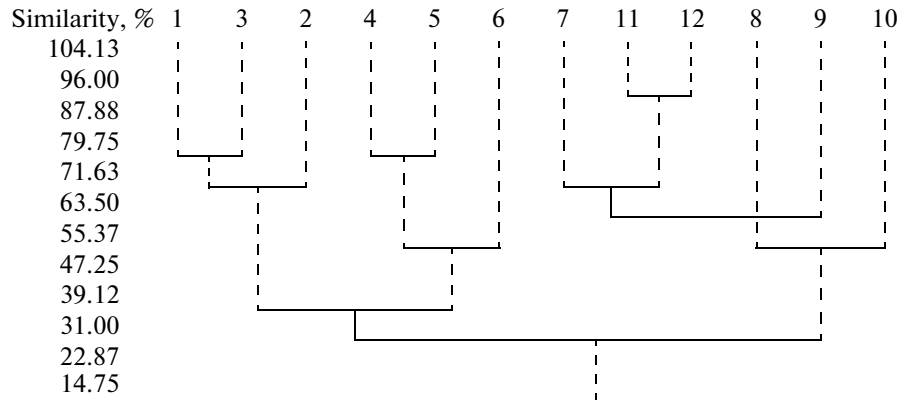

Fig. 2. Dendrogram of similarity for the 12 samples (see Fig. 1) characterized by 7 attributes.

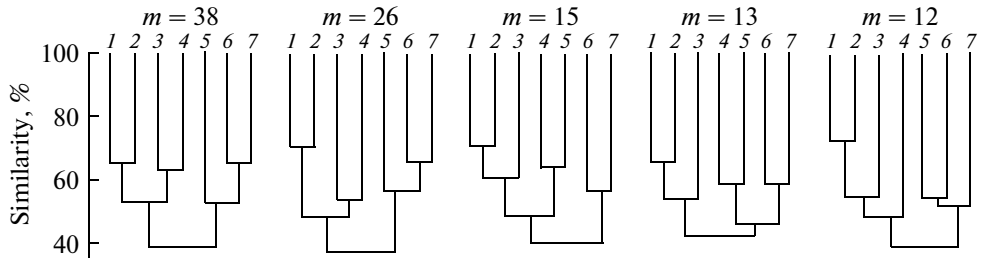

Fig. 3. Assessment of the informativeness of the attributes by the method of dendrograms (for seven arbitrarily selected objects).

attributes are characterized by their high similarity, one of them can be excluded. An advantage of the method of dendrograms is that they may be developed for any scale of the attributes (not only arithmetic scales can be used).

More stringent criteria of informativeness can be applied if the classes of objects are already specified (by expert judgment or by some formal methods). If a sample is large, the informativeness of the attributes can be assessed by the methods of multivariate statistics (http://lem.edu.mhost.ru/doc/presntations/Rozhkov.pdf; [21]). For this purpose, the number of objects in the classes should exceed the number of attributes. The loss of information upon the exclusion of some attributes is estimated from the comparison of the similarity of the classes for the full (p) and reduced (q) sets of the attributes:

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

where $f = p(k-1)$, $n = n_1 + n_2$ (the number of objects in the classes), k is the number of classes, and $\lambda q = |W|/|T|$ is the ratio of the matrix determinants for the intra- and interclass variances with p and q attributes.

Similar results can be obtained if we compare the Mahalanobis distances for the compared classes characterized by the full and reduced sets of the attributes:

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

where F is the distribution with $f_1 = q - p$ and $f_2 = n_1 + n_2 - q - 1$ degrees of freedom; n_1 and n_2 are the numbers of objects in the compared classes, and q and p are the initial and reduced numbers of attributes, respectively ($q > p$); and D_q^2 and D_p^2 are the Mahalanobis distances. If $F \le F_{\alpha\, f1,f2}$, then the exclusion of a given attribute has not led to the loss of information.

The corresponding calculations are performed by the cyclic examination of all the attributes. The least informative attribute is excluded from the sample, and the procedure is performed again until the loss of information (according to the χ^2 criterion) becomes significant. The exclusion of the low-informative attributes is cost-efficient, as it allows us to reduce the number of necessary soil analyses by more than two times.

The space of the attributes can also be reduced via the compression of the data. For this purpose, the dis-

Table 2. Approximation of data on the bulk elemental composition (%) of the humus-illuvial podzol

Genetic horizon	Depth (x), cm	Bulk forms, %	Analysis	Calculated	Difference	Difference, % of the analytical data
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

Note: Coefficients of the polynomial $P_{2,2}(x, y) = 1.20 - 86.0x + 16.0x^2 + 11.0 - 9.5y + 1.9y^2$, where x is the depth and y is the ordinal number of the oxide.

tribution of the attribute values in the soil profiles is approximated by the following polynomials[1] [20, 21]:

$$P_m(x) = \sum_{k=0}^{m} a_k x^k,$$

where m is the degree of the polynomial.

For example, a polynomial

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

represents a good approximation of the field water capacity (w, %) in the profile of a soddy-podzolic soil [20].

Tables containing data on the composition and properties of soils $f(x, y)$ can be approximated by the following polynomials [15]:

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

where n is the degree of the approximating polynomial according to the variable x, $\underline{s}$ is the number of approximation nodes according to the variable x (the number of columns in the table), m is the degree of the approximating polynomial according to the variable y, r is the number of approximation nodes according to the variable y, and A denotes the coefficients of the polynomial.

Table 2 illustrates the possibility of the two-dimensional approximation of soil attributes. More accurate approximations of the data may also be found if we use some other algorithms. However, the mere possibility of the compression of the initial data is important. This approach is convenient if we need to present the data in a uniform way (as distributions of attributes in the soil profile or along the catena) and to perform data interpolation. The parameters of the polynomials can serve as the initial data for the further analysis, as has been tested in many works in soil science [5, 13, 15, 17, 29, 30].

Numerical taxonomy has its own means to evaluate the quality of classification decisions and to compare different classification systems. Figure 4 contains an example of the separation of soil objects into classes (classification) and the formulas to calculate several criteria of the quality of this separation. The initial data were obtained from different sources and include information on the contents of humus and exchangeable calcium in the A1 horizons of seven different soils. They are grouped in clusters in the upper part of the figure. A special program was used to separate 27 soil samples into 2, 3,...12 classes (groups) successively (the lower part of the figure) with the calculation of six indices of the quality of this separation (QC1–QC6). The first index (the quality coefficient (QC)) represents the ratio of the number of correctly determined classes to the total number of classes (in this case, we know the initial classes represented by the genetic soil types). If we do not have the initial division into classes, other criteria should be applied: QC2 (the difference in the mean values of the similarity of the samples within the groups and between the groups), QC3 (the mean similarity within the groups), QC4 (the ratio of the determinants of the matrices of the intragroup variation and the total variation), QC5 (the spur of the matrix of the mean intragroup similarity), and QC6 (the determinant of the matrix of the mean intragroup similarity).

[1] Power polynomials serve as basic functions in different formal procedures of data approximation, regression, presentation of linear discriminant functions, linear transformation of the vectors of attributes into principal components, and canonic correlation; they are used as models in the design of complex technical systems and in their self-organization [4, 12].

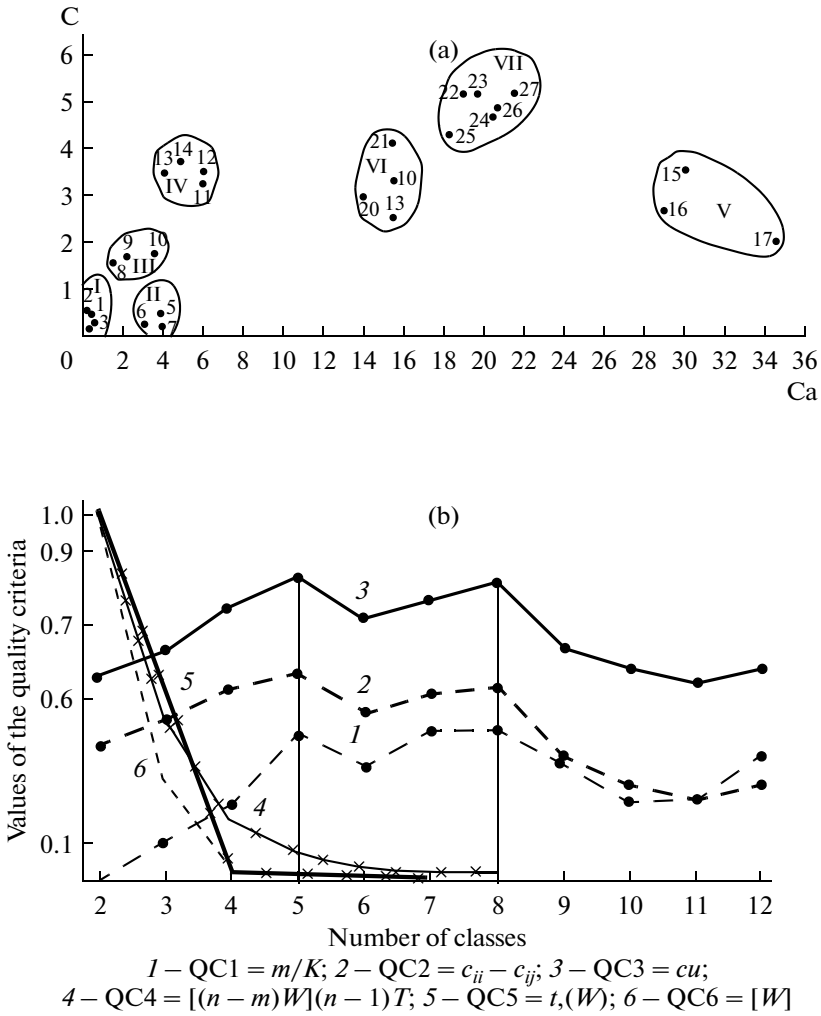

Fig. 4. Quality criteria for classification systems (explanations in the text): (a) ordinate representation of soil groups (I—brown desert, II—gray-brown (semidesert), III—soddy medium-podzolic, IV—soddy strongly podzolic, V—chestnut, VI—light gray forest, VII—gray forest); 1–17 are the numbers of the soil profiles; (b) changes in the quality of the classification with an increase in the number of classes (groups).

The use of the first three indices (QC1–QC3) gives us similar results. According to them, the separation of the samples into five and eight groups (instead of the preset seven classes) is characterized by good quality coefficients. However, if we use other criteria, we can say that the best quality of the classification is achieved, when the samples are separated into seven groups. However, in practice, the second criterion (QC2) is usually applied, as it is calculated in the course of the separation (classification) procedure.

Thus, different classification decisions can be compared on the basis of specially calculated quality coefficients. However, other special approaches can also be used for this purpose. For different ordinate classifications of the same soil objects (soils), the coefficient of association and the polychoric correlation index can be used [21, 31].

The algorithm of the calculation of the coefficient of the association of the ordinate classifications. The initial data is as follows: k is the number of compared classifications, $q(k)$ is the number of classes in them, $n(k, q_{max})$ is the number of objects in the classes, and $m(k, n)$ are the numbers of objects by the classes.

(1) A table of the paired associations of classifications (n_{ij}) is constructed.

(2) The coefficient of the association (CA) is calculated:

$$CA = (\omega_1 + \omega_2 - t/(2n - t), \quad (0.1)$$

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

Table 3. Numbers of objects coinciding in classifications I and II (by separate classes*)

Classification II	Classification 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	$\omega_1 = 75$	

* The explanations of the particular indices are given in the text.

$j = 1, \ldots k_1$ is the number of classes in the first classification;

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

$i = 1, \ldots k_2$ is the number of classes in the second classification; and

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

The algorithm of the calculation of the polychoric correlation of the ordinate classifications.

The initial data are the same as in the previous algorithm.

The Chuprov polychoric correlation (PC) is calculated:

$$\mathrm{PC} = \varphi k^{-0.25}, \text{ where } k = (k_1 - 1)(k_2 - 1);$$

and k_1 and k_2 are the numbers of classes in the two classifications, respectively.

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

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

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

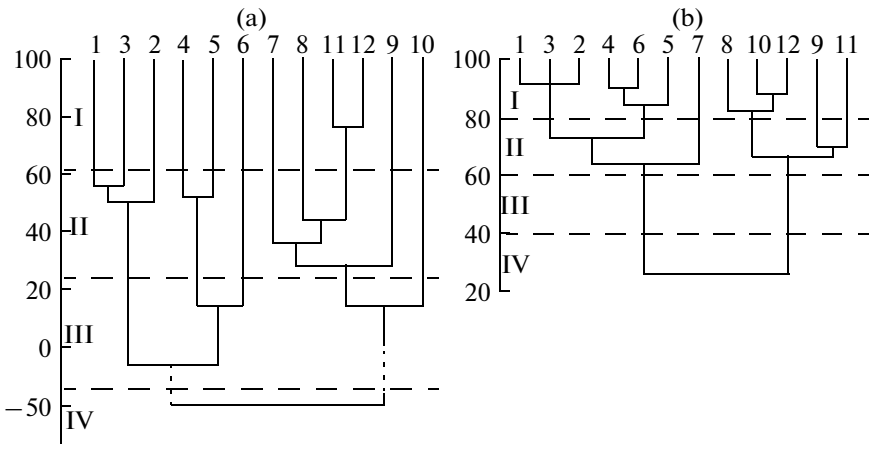

Fig. 5. Comparison of hierarchical classifications according to (a) different sets of attributes and (b) different similarity indices.

Table 4. Matrices cV for dendrograms A (lower triangle) and B (upper triangle) (see Fig. 5)

A/B	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	–

Note: $R = 0.5$ at $R_{0.99} = 0.35$.

For example, we have obtained two classifications for 107 objects ($n = 107$). In the first (I) classification, these objects are separated into four classes, and, in the second (II) classification, they are separated into ten classes. Table 3 contains data on the number of coinciding objects in the classes of both classifications and on the results of the required summing operations.

On the basis of these data, we obtain the CA value: $CA = 75 + 84 - (31 + 42)/2.107 - (31 + 42) = 0.61$. This means that classifications I and II agree at the level of 61%.

For the polychoric correlation, we obtain $k = 27$, $\varphi^2 = 2.59$, $PC = 0.71$, and $\chi^2 = 277$, which points to the high conjugation of these classifications ($\chi^2_{table} = 55.5$ with the probability of 0.999).

A comparison of hierarchical classifications can be performed by the method of Sokal and Rohlf [33] with the use of special software to compare dendrograms. Figure 5 and Tables 4 and 5 illustrate the algorithm to compare two dendrograms for the above-described twelve soil horizons constructed with the use of two different sets of attributes (or with the use of different indices of similarity).

The range of the similarity indices of the dendrograms is subdivided into equal numbers of intervals. Four intervals are recommended when the number of objects $n \le 10$; when the number of objects $n \ge 100$, more than 10 intervals should be specified.

Then, the matrices of similarity of cV_a and cV_b sizes $[n, n]$ are constructed (Table 4). These matrices show the numbers of the intervals that characterize similarity between the particular objects. For example, for the first object of dendrogram A (the lower triangle in the table), its similarity with objects 2 and 3 and with objects 4–6 lies within intervals 2 and 3, respectively; its similarity with objects 7–12 lies within interval 4. These intervals are indicated in the first column of the matrix. For dendrogram B (the upper triangle), the similarity of the first object with objects 2 and 3 lies within interval 1 (the upper row of the matrix), etc.

To measure the degree of similarity between the two dendrograms, the coefficient of correlation R is used. To calculate it, the matrix is transformed into the correlation table. When the number of intervals is less than four, nonparametric indices are used instead of R. In our case, $R = 0.5$ with the probability of 0.99.

Until recently, this was the only approach to compare dendrograms [24]. However, it can be modified with the use of the PC and CA indices calculated for hierarchical classifications (Table 5).

The agreement between the matrices cV_a and cV_b is judged from the number of coinciding elements of these matrices. In this case, the PC equals 0.36, and

Table 5. Indices of agreement between the dendrograms (the number of classes in classifications C_1 and C_2 is equal to four)

C_1/C_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) $PC = 0.36$, $\varphi^2 = 0.38$.

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

2) $CA = 0.36$.

Table 6. Coding of the morphological descriptions of soil profiles

Code	Indications of hydromorphism	Texture	Groundwater depth, cm
0	Hydromorphic features are absent	Sand	>200
1	Brown and ocherous mottles, fine iron nodules, ocherous veins	Alteration of sandy (loamy sandy) and loamy layers	150–200
2	Bluish tint with ocherous mottles against the brown (or gray) background color	Loamy sand	100–150
3	Bluish mottles against the brown (or gray) background color and, in some cases, ocherous mottles	Light loam	50–100
4	Ocherous (rusty) horizon with iron–manganic concentrations	Medium loam	0–50
5	Iron–manganic nodules and laminae against the mottled ocherous–bluish background color pattern	Heavy loam	
6	Bluish (dirty bluish) horizon with ocherous mottles	Clay	
7	Thin peat horizon	Silty sand, peat	
8	Peat with moderately and highly decomposed plant residues	Peat	

this correlation is significant with the probability of 0.99. The CA also equals 0.36, but the authors have not suggested methods to estimate its probability; thus, this index has just an illustrative meaning.

This is not a stringent method to compare dendrograms, because the number of similarity intervals is not strictly specified. It remains an arbitrary characteristic. A more objective approach consists of the separation of objects in the dendrograms with respect to the quality indices and the comparison of the dendrograms according to the PC and CA indices.

This approach can be illustrated by the comparison of three classification systems of floodplain soils in the middle course of the Ob River [26].

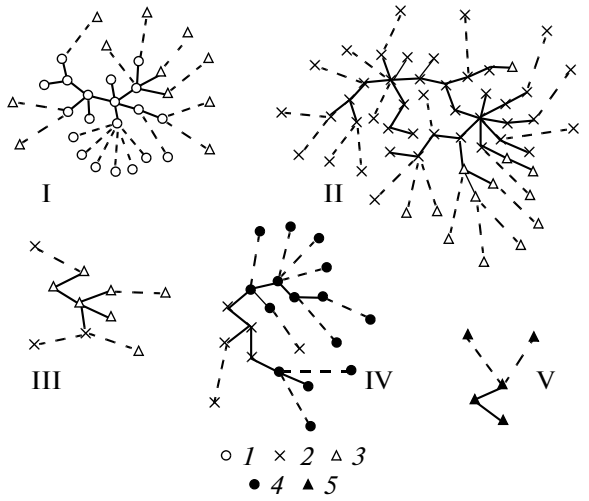

Fig. 6. Results of an automated computer-generated grouping of soils. Soils: (*1*) soddy, (*2*) meadow, (*3*) soddy-meadow, (*4*) meadow-bog, and (*5*) bog (swampy).

In the field, the soils were classified according to the system suggested by Dobrovol'skii [10]. Then, they were reclassified according to the system suggested by Shrag [28]. Each of the pits was described with the use of nine attributes characterizing the degree of the soil hydromorphism and data on the texture of the Asod, A1, B, and D(C) horizons and on the groundwater level. (Table 6).

For an automated classification procedure, the values of these attributes were encoded. These codes do not represent some ranged scales; they give us designations of the particular soil attributes in the numerical form; i.e., they represent nominal scales for the considered attributes (the degree of hydromorphism, the soil texture, and the groundwater depth). The results of the automated grouping (classification) are shown in Fig. 6. The conventional signs shown on the diagrams denote the names of the soils according to the classification system of Dobrovol'skii.

All the soils were separated into five compact groups that mainly consist of the same soil subtypes (according to the classification of Dobrovol'skii). However, as the changes in the soil properties upon the transition from a given soil type to another type have a gradual character, it is quite natural that the group of soddy soils (group 1) also includes soddy meadow soils that differ from the soddy soil proper in the appearance of some hydromorphic features (attributes). A large group of meadow soils (2) includes 35 pits; some of them were initially classified as soddy meadow soils with a significant degree of hydromorphism. The group of meadow-bog soils (4) includes some meadow soils with increased hydromorphism. The distribution of all the studied 126 soil pits by the taxa of the three classification systems (including the automated numerical classification) is shown in Table 7.

Table 7. Classification of soils on the Ob River floodplain

Class	Alluvial soil	Numbers of classified objects (soil profiles)
		Classification by G.V. Dobrovol'skii
1	Soddy	1–3, 11–13, 31, 32, 34, 37,40, 59, 66, 88, 89, 99
II	Soddy-meadow	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	Meadow	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	Meadow-bog	14, 16, 24, 51, 64, 69, 90, 92, 112, 114, 117, 119, 121–124
V	Bog (swampy)	8, 91, 111, 115, 120
		Classification by V.I. Shrag
I	Layered	2, 59, 99, 103
II	Granular-layered	1, 11–13, 29, 31, 40, 42, 61, 66, 67, 89, 96
III	Granular	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	Swampy layered and granular-layered	15, 33, 38, 43–45, 72, 74, 75, 97, 101, 106, 110
V	Swampy granular	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	Silty gleyed	14, 120
VII	Silty swamp	8, 91, 111, 112, 115, 119
		Automated computer-generated classification
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

Table 8 contains the results of the comparison between the three systems according to the coefficients of association and polychoric correlation. It can be seen that a reliable agreement is only observed between the automated classification and the classification according to G.V. Dobrovol'skii. Other compared pairs of classifications are weakly associated (according to the CA and PC indices).

This can be explained by the fact that the soil attributes selected for the computer-based processing better fit the classification criteria suggested by Dobrovol'skii than the criteria used by Shrag. However, it is more important that we have compared different classifications with the use of quantitative methods. It can be supposed that such a comparison of different classifications tested in different regions of the country will help us to solve many disputable problems.

If the soil classes (taxa) are already specified, we can examine the relationships between them (the degrees of separation, inclusion, or intersection) with the help of discriminant analysis (Fig. 7).

The methods of multivariate statistical analysis can be applied to formalize the results of the classification and find the position of new objects (i.e., to classify them) in the system of existing classes. These methods can be applied at any level of the organization of soil systems [23].

In the given example, we have three classes of objects (soils) that consist of 8, 14, and 12 individual soils, respectively. They are described by four attributes. The results of the calculations with the use of the method of multivariate statistical analysis are shown in Fig. 8. It contains data on the Mahalanobis distances between the classes and the corresponding F criteria with freedom degrees $f_1 = m$ (number of attributes), and $f_2 = n_i + n_l - m - 1$.

The linear discriminant function (LDF) describing the maximum differences between classes 1 (I) and 2

Table 8. Criteria of agreement between the different classifications: the coefficients of association (CA) and polychoric correlation (PC, in parentheses)

Classification	V.I. Shrag	Computer-generated
G.V. Dobrovol'skii	0.27 (0.30)	0.53 (0.56)
V.I. Shrag	–	0.25 (0.12)

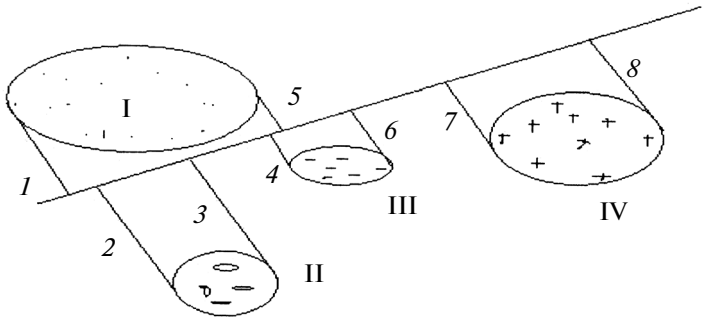

Fig. 7. The linear discriminant function $MS^1 = L_0 + L_1 X_1 + L_2 X_2 + ... + L_M X_M$.

Numbers of classes		Mahalanobis distance	F	f2
i	l	D2		
1	2	223.43	241.7	17
1	3	371.00	371.0	15
2	3	18.35	25.9	21

Coefficients of linear discriminant functions:

.....Between classes 1 and 2:
B[0] = 492.59 B[1] = 1.46 B[2] = 2.41 B[3] = 1.04 B[4] = 11.17
.....Between classes 1 and 3:
B[0] = 508.53 B[1] = 0.73 B[2] = 2.58 B[3] = −0.35 B[4] = 19.17

.....Between classes 2 and 3:
B[0] = 108.46 B[1] = 0.54 B[2] = 0.78 B[3] = −0.25 B[4] = 3.91

Projections of the objects onto the dividing plane:

Between classes 1 and 2
Class 1
41 41 42 36 40 36 34 41
Class 2
4 3 1 2 0 7 0 2 2 5 5 5 2 6

Between classes 1 and 3
Class 1
57 56 58 52 57 50 49 57
Class 3
3 1 3 5 3 5 4 1 3 7 0 3

Between classes 2 and 3
Class 2
23 22 17 18 17 26 17 20 19 24 24 24 19 25
Class 3
6 4 7 9 7 9 8 1 4 13 0 8

Fig. 8. Multivariate statistical analysis (linear discriminant functions); $f1 = 4$.

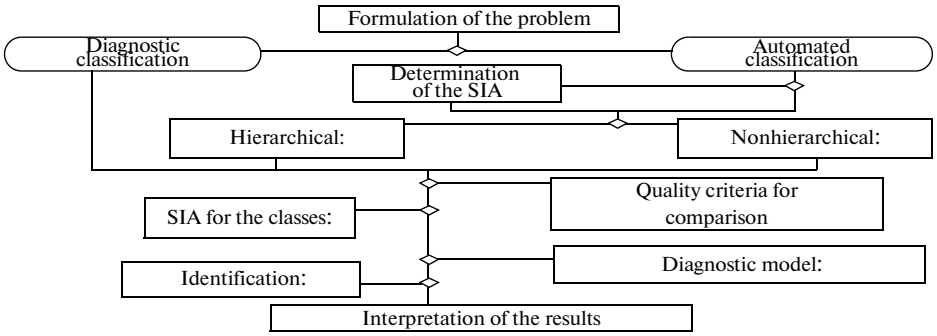

Fig. 9. Problems of numerical taxonomy.

(II) according to attributes $x_1 - x_4$ has the following form:

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

The differences between these classes are reliable according to Fisher's test F_{f1f2}, which is also confirmed by the fact that the projections of the objects onto linear axes (the L_{ilj} values) do not intersect.

If we have learning samples analogous to those given in this example, the classification of new objects is performed via the calculation of the L_{ijl} values, which should be compared with the projections of the objects in the specified classes. For classes 1 and 2, they correspond to the intervals of 34–42 and 0–7, respectively. If the projection of the object to be classified does not fit the existing intervals, it is classified into the closest interval. In some cases, however, its classification will be uncertain. In particular, this concerns the cases when the projections of the existing classes do intersect.

The classification of new objects according to the LDF system is performed via calculating projections for each of these functions; the object is attributed to the class in which it falls most often (the voting method).

There are variants of discriminant analysis that take into account the inequality of covariance matrices, the LDFs for three and more classes, and other formal rules of classification, including binary scales of the attributes. The classification of new objects can be based on different indices of similarity with the objects in the learning classes: the nearest neighbor, the mean similarity with all the objects in the classes, the intervals of similarity, etc.

The particular algorithms and methods are realized within the framework of general problems of numerical taxonomy (Fig. 9). Two different groups of the problems can be specified: the diagnostic classification implies the identification (classification) of new soils relative to the existing (learning) classes, whereas the automated classification implies the grouping of soils by automatically determined classes of hierarchical or other (nonhierarchical) order. The criteria of the quality of the classification can be determined both for preset and newly constructed classifications. For the selected classifications, more compact systems of informative attributes (SIA) can be determined.

The differences between the classes can be determined with the use of linear discriminant functions that represent diagnostic models for classifying new objects. There are also other methods to classify (identify) new objects, i.e., to determine the classes to which they should be attributed.

The methods described in this paper are realized in the MERON software package especially designed to solve the problems of the numerical taxonomy of soils.

CONCLUSIONS

The formalization of notions and procedures is the most efficient tool to achieve objectivity in the field of soil classification and to compare and evaluate the existing classification systems. This is a necessary stage for the further development of the theory of pedology and its practical applications. The problem of soil classification is one of the acute problems of pedology. A system of axioms of soil classification should be developed. The author believes that the examples of application of formal numerical methods to solve the particular problems of soil classification that are discussed in this paper should initiate new ideas and hypotheses in this field, though the results obtained with the use of these methods are not always easily visualized and interpreted.

A general scheme to develop the numerical taxonomy of soils is important. At the first stage, the goals of the classification should be specified. Then, we should determine the optimum system of informative characteristics (attributes) for the preset or automatically determined taxa (classes) of soil objects arranged along the ordinate in a hierarchical way. The criteria of the quality of the obtained classification should be examined, and the formal rules (diagnostic models) for the identification of new objects should be devel-

oped. Thus, we may create a quantitatively substantiated and visualized classification system that is devoid of the hidden subjectivity factor.

A separate problem concerns the use of the methods of numerical soil taxonomy in digital soil mapping. These are tightly interrelated problems; it is probable that the methods of numerical taxonomy will be widely applied in digital soil mapping. The informativeness of direct and indirect indicators should be examined, and the quality of the soil interpretation should be evaluated. It is important to formalize soil cover patterns [23], to solve the problem of inclusions, and to adapt the methods of fuzzy logic for the purposes of digital soil mapping.

REFERENCES

1. S. A. Aivazyan, Z. I. Bezhaeva, and O. V. Staroverov, *Classification of Multidimensional Observations* (Statistika, Moscow, 1974) [in Russian].
2. A. G. Arkad'ev and E. M. Bravermann, *Supervised Computer-Based Classification of Objects* (Nauka, Moscow, 1971) [in Russian].
3. N. T. J. Bailey, *The Mathematical Approach to Biology and Medicine* (J. Wiley and Sons, London–New York–Sydney, 1967). Translated under the title *Matematika v biologii i meditsine* (Mir, Moscow, 1970).
4. V. M. Balyk, N. S. Kalutskii, and R. D. Kulakova, "Structural-Parametric Self-Organization of Complex Technical Systems," Delovaya Rossiya, 58–63 (2007).
5. V. A. Bobrov, "Approximation of Some Soil Functions and an Integral Way to Calculate the Amount of Humus in Soil," in *Methods of Biology and Soil Science* (Nauka, Alma-Ata, 1976), pp. 103–111 [in Russian].
6. *The Great Russian Encyclopedia* (Izd. BSE, Moscow, 2005), Vol. 1 [in Russian].
7. S. W. Boul, F. D. Hole, and R. J. McCracken, *Soil Genesis and Classification* (Ames, 1973). Translated under the title *Genezis i klassifikatsiya pochv,* (Progress, Moscow, 1977).
8. Yu. A. Voronin, *Fundamentals of Similarity Theory* (Nauka, Novosibirsk, 1991) [in Russian].
9. G. N. Vysokos and V. A. Rozhkov, "Scales of Soil Features and the Choice of Similarity Measures for Soil Objects," in *Soil and Agrochemical Studies with the Use of Computers* (Tr. Pochv. Inst. im. V.V. Dokuchaeva), pp. 30–39 (1981).
10. G. V. Dobrovol'skii, *Soils of Floodplains in the Center of the Russian Plain* (Izd. Mosk. Gos. Univ., Moscow, 1968) [in Russian].
11. E. N. Knyazeva and S. P. Kurdyumov, *Basics of Synergetics. Human Constructing Himself and His Future* (KomKniga, Moscow, 2007) [in Russian].
12. I. K. Lur'e, *Geoinformation Mapping* (KDU, Moscow, 2010) [in Russian].
13. T. B. Makhlin, "Approximation of Johnson's Curves of Element Distribution in Soils," Pochvovedenie, No. 6, 123–130 (1973).
14. J.. Pfanzagl, *Theory of Measurement* (Physica Verlag, Wurzburg–Wien, 1971). Translated under the title *Teoriya izmerenii* (Mir, Moscow, 1976) [in Russian].
15. L. Yu. Reintam, "Correlation and Regression between the Properties of Brown Forest, Pseudopodzolic, and Soddy-Podzolic Soil Types," Pochvovedenie, No. 1, 116–132 (1971).
16. I. V. Reshetukha, "Approximation of Tabulated Function $f(x, y)$ by the Method of Least Squares with a Polynomial $P_{n,m}(x, y)$," in *Program Support of Computers MIR-1 and MIR-2*, Vol. 2 (Naukova dumka, Kiev, 1976), pp. 126–129 [in Russian].
17. V. A. Rozhkov, "Algebra of the WRB (Formalized Concept)," in *Experimental Information in Soil Science: Theory and Methods of Standardization* (Proc. All-Russia Conf.) (Izd. Mosk. Gos. Univ., Moscow, 2005), pp. 73–82 [in Russian].
18. V. A. Rozhkov, *Algorithms and Programs for a Computer-Based (MIR-2) Objective Soil Classification* (Izd. VASKhNiL, Moscow, 1976) [in Russian].
19. V. A. Rozhkov, "Method of Principal Components and Its Application in Soil Science," Pochvovedenie, No. 10, 141–151 (1975).
20. V. A. Rozhkov, "On the Mathematical Formalization of Distribution of Substances in the Soil Profile," Byul. Pochv. Inst. im. V.V. Dokuchaeva, Iss. IV, 66–73 (1972).
21. V. A. Rozhkov, *Soil Informatics* (Agropromizdat, Moscow, 1989) [in Russian].
22. V. A. Rozhkov, "Efficient Coding of Soil Data," Byul. Pochv. Inst. im. V.V. Dokuchaeva, No. 53, 32–35 (1972).
23. V. A. Rozhkov and E. B. Skvortsova, "Tectology of Soil Megasystems: Universal Principles of Organization and Analysis of Data," Pochvovedenie, No. 10, 1155–1164 (2009) [Eur. Soil Sci. **42** (10), 1073–1082 (2009)].
24. R. R. Sokal, "Clustering and Classification: Background and Current Directions," in *Classification and Clustering*, J. van Ryzin (Ed.), (Academic Press, New York, 1977), pp. 7–19.
25. A. M. Chandra and S. K. Ghosh, *Remote Sensing and Geographic Information Systems* (Alpha Science Int. Ltd., 2006).
26. B. V. Sheremet and V. A. Rozhkov, "An Experience in Numerical Taxonomy of Floodplain Soils of the Ob River according to Their Morphological Features," Vest. Mosk. Univ., Ser. 17: Pochvoved., No. 3, 41–50 (1977).
27. L. L. Shishov, V. A. Rozhkov, and V. S. Stolbovoi, "Information Base of Soil Classification," Pochvovedenie, No. 9, 9–20 (1985).
28. V. I. Shrag, *Floodplain Soils, Their Reclamation and Agricultural Use* (Rossel'khozizdat, Moscow, 1969) [in Russian].
29. J. D. Colwell, "A Statistical-Chemical Characterization of Four Great Soil Groups in Southern New South Wales Based on Orthogonal Polynomials," Austral. J. Soil Sci. Res. **8** (3), 221–238 (1970).
30. K. T. Erle, "Application of the Spline Function on Soil Science," Soil Sci. **114** (5), 333–338 (1972).

31. L. A. Goodman and W. K. Kruskal, "Measure of Association for Cross Classification," J. Amer. Statist. Assoc, **49**, 732–764 (1954).
32. J. J. de Gruijter, "Numerical Classification of Soils and Its Application in Survey," Soil Surv. Papers, No. 12, 117 pp. (1977).
33. R. Protz, R. W. Arnold, and E. W. Presant, "The Approximation of the True Modal Profile with Use of the High Speed Computer and Landscape Control," Trans. 9th Int. Congr. Soil Sci., Adelaide, 1968, Vol. 4, 193–204.
34. R. R. Sokal and F. I. Rohlf, "The Comparison of Dendrograms by the Objective Methods," Taxon., No. 11, 33–40 (1962).
35. R. R. Sokal and P. H. A. Sneath, *Principles of Numerical Taxonomy* (W.H. Freeman and Co., San-Francisco, 1963).
36. R. Webster, *Quantitative and Numerical Methods in Soil Classification and Survey* (Clarendon Press, Oxford, 1979).

SPELL: 1. Reshetukha, 2. Naukova, 3. Conf, 4. Byul