DOI: 10.1515/jwld-2016-0029

JOURNAL OF WATER AND LAND DEVELOPMENT
2016, No. 30 (VII–IX): 127–139
PL ISSN 1429–7426
Available (PDF): http://www.itp.edu.pl/wydawnictwo/journal; http://www.degruyter.com/view/j/jwld

Received 09.06.2016
Reviewed 11.08.2016
Accepted 12.09.2016

A – study design
B – data collection
C – statistical analysis
D – data interpretation
E – manuscript preparation
F – literature search

Adaptive modelling of spatial diversification of soil classification units

Krzysztof URBAŃSKI [ABCDEF], Stanisław GRUSZCZYŃSKI [ABCDEF]

AGH University of Science and Technology, Faculty of Mining Surveying and Environment Engineering, Department of Environmental Management and Protection, al. Mickiewicza 30, 30-059 Kraków, Poland; e-mail: urbanski@agh.edu.pl, sgrusz@agh.edu.pl

For citation: Urbański K., Gruszczyński S. 2016. Adaptive modelling of spatial diversification of soil classification units. Journal of Water and Land Development. No. 30 p. 127–139. DOI: 10.1515/jwld-2016-0029.

Abstract

The article presents the results of attempts to use adaptive algorithms for classification tasks different soils units. The area of study was the Upper Silesian Industrial Region, which physiographic and soils parameters in the form of digitized was used in the calculation. The study used algorithms, self-organizing map (SOM) of Kohonen, and classifiers: deep neural network, and two types of decision trees: Distributed Random Forest and Gradient Boosting Machine. Especially distributed algorithm Random Forest (algorithm DRF) showed a very high degree of generalization capabilities in modeling complex diversity of soil. The obtained results indicate, that the digitization of topographic and thematic maps give you a fairly good basis for creating useful models of soil classification. However, the results also showed that it cannot be concluded that the best algorithm presented in this research can be regarded as a general principle of system design inference.

Key words: *adaptive algorithms, self-organizing map (SOM), soil classification, Upper Silesian Industrial Region*

INTRODUCTION

The development of information technologies is accompanied by a gradual change of attitude towards cartographical documentation of natural objects, including soils. Various digital modelling methods are replacing traditional, static and discrete structure of diversification image. Among properties of the new paradigm, particularly important is possibility to create a continuous or quasi-continuous soil characteristics. Regardless of the accepted solution, a manifest feature of this solution is relativization of soil classification. The basic feature of any soil classification is a multidimensionality of units: there is no single measurable property which would suffice to unambiguously assign a pedon to a specific soil class [JENNY 1941]. Such assignment involves a comparison of pedon similarity to the standards which make up a reference scale; a correct classification requires a vector of pedon features and evaluation of its similarity to the standards. This can be formulated as the following thesis: *a set of soil unit standards (soil types, soil valuation classes, land capability units, site types) creates reference points in a multidimensional space of features, and the pedon classification criterion is its similarity to a specific standard; as in the multidimensional space the pedon can occur in the neighbourhood of many standards, it is important to determine the degree of similarity to an appropriate class of units.*

The basic difficulty in implementation of this concept in digital systems is scarcity of data. Regardless of the observed soil feature, it is represented by values in specific points of space, surrounded by land of unknown, though by assumption similar, properties. Remote sensing, which under some circumstances provides a continuous image of surface diversification, does not allow to determine the soil profile properties. The basic alternative in these conditions is a form of modelling of spatial diversification on the

basis of point observations or features assigned to the soil units contours, linking them with land physiography. An example of such solution if the map of soils of European Union [JONES *et al*. (ed.) 2005] which is a combination of spatial information, soil profile standards, pedotransfer rules and pedotransfer functions. Many years ago the Digital Soil Map Working Group undertook the task of digital unification of cartographic and soil documentation of EU countries [DOBOS *et al.* 2006]. There are other concepts based on similar assumptions: SoLIM [ZHU *et al.* 2001] and SCORPAN [MCBRATNEY *et al.* 2003] derived from the Hans Jenny's concept [JENNY 1941].

PURPOS, INPUT DATA AND INVESTIGATED OBJECT

This paper aims at evaluating the usefulness of differentiation models of land capability unit on the basis of information from the digitization of cartographic materials: soil and agricultural maps and topographic maps [STRZEMSKI *et al*. 1964]. The algorithms used in these models belong to a broad category of adaptive (evolutionary) algorithms, methods of computational intelligence and machine learning.

Despite many years of (recently very keen) interest in adaptive modelling and machine learning methods, their choice for a specific application still remains a subjective issue. Based on some research [TADEUSIEWICZ 1993] a conclusion can be drawn that the model architecture is less important than its optimization (learning). However, it should be noted that some architectures are linked with strictly specific optimization methods. There is widespread belief that some algorithms perform well in some applications, and some perform poorly. A known example is handwriting recognition or voice analysis where deep learning techniques (stacked autoencoder model) are used with success, whereas other models fail. Thus, it can be presumed that creation of the model should be preceded by experiments with various algorithms in order to select the class of models most suitable to achieving the modelling goal. The problem of modelling the location of soil units on the basis of digitized cartographic materials is, at least in our conditions, a task with some practical value, at least in the future until the decisions to digitize the soil maps are made.

The area which is the source of data is the region known as GOP (Upper Silesian Industrial Region). We analysed only agricultural land: arable land (R) and permanent grassland (TUZ) – meadows and pastures (Ł and Ps). On the area of 2596 km^2 we distinguished 6.4 million units forming a grid made up of 20 × 20 m squares. A training set comprising 100,000 units and a validation set of 124,000 units were randomly drawn from the total set. In relation to each unit the following parameters were determined and included as spatial database: horizontal coordinates of the centre (x and y), elevation of the centre (z), land type (R or TUZ); environment: 25 elevations of the unit neighbourhood: centre elevations of all unit patches and the unit itself on 1 ha (100 × 100 m square), number of patches included in the neighbourhood which centre elevations exceed the unit elevation; grain size: 8 values (0–25, 25–75, 75–125, >125 cm below the surface), two for each layer of average content of silt (0.05–0.002 mm) and clay (below 0.002 mm) fraction for the mechanical group, and evaluation of the groundwater level according to the model based on levels of watercourses and reservoirs.

The key issue in models based on key data is components of input data vector which configuration could be characteristic for separated classes to the highest degree possible. In the task involving the use of materials from digitization of analogue maps, the possibility of building a very extensive data vector is limited. An important issue in this regard is whether this model includes the coordinates of the point in which the soil is classified. Of course, horizontal coordinates x and y unequivocally define a point on the surface, creating a risk of referencing of coordinates to the class labels which would mean that the model has acquired a memory. In such case the model would reflect correctly the location of soil units included in the set used for optimization, and the error of designating labels in other points would be random. On the other hand, introducing the coordinates into the input vector is justified due to the role of factors which can be called contextual and result from the geographical location. There are many factors, such as groundwater table, where absence of context makes a correct interpretation impossible. To solve this dilemma, the trials were conducted with two versions of data: A – data vector with horizontal coordinates, B – data vector without horizontal coordinates. The results given in the next part of the paper relate only to the data included in the validation set, so the evaluation principally does not include a risk of using models which overfit the data.

EVALUATION OF DIVERSIFICATION OF PHYSIOGRAPHIC AND SOIL DATA VECTORS ON THE BASIS OF SOM ALGORITHM

The self-organizing map (SOM) data clusterization algorithm developed by Kohonen [TADEUSIEWICZ 1993] is, to some degree, an alternative to the known procedure of multidimensional scaling (MDS). The result of both algorithms is a diagram (map) representing interrelations of objects characterized by many features. Contrary to the SOM, the MDS has two serious limitations: quantitative, due to hardware requirements the maximum number of clusterized vectors generally does not exceed 1500–2000; and qualitative – adding new vectors to the data requires a restart of the MDS procedure.

In the SOM algorithm, as a result of the iterative procedure, the vectors representing objects are arranged on a two-dimensional table (map consisting of

cells storing information on the configuration of vectors' value) according to the similarity principle: similar object vectors activate the same or similar map cells in space; dissimilar vector activate cells distant from each other. With some simplification it can be supposed that the SOM algorithm allows to evaluate the set of factors which differentiate the objects, grouping such factors in clusters suitable in the process of further classification. The data used for clusterization do not have, or are intentionally deprived of, labels (classes) in order to avoid grouping based on a feature determined according to the remaining data vector components. The image labels can be then used in the map calibration. Thus, the obtained image of object groups location on the two-dimensional surface allows to evaluate the similarity of objects characterized by many features.

RESULTS OF CLUSTERIZATION OF SOIL DATA VECTORS FOR INPUT DATA IN VERSION A

The training set comprising 100,000 vectors, proportional to number of complexes in the entire GOP set was used to generate the SOM, and the presented result of calibration relates to the validation set. We used a prototype of randomly initiated map consisting of 625 cells (25 rows and 25 columns) organized in a "honeycomb" structure (a single cell has six adjacent cells). Figure 1 includes a diagram symbolically presenting distances between vectors which activate the cells in the manner used in the MATLAB package. The darker zone between cells indicate data vectors with similar parameters, the brighter zone between cells indicate significant distances between the vectors, thus separating data clusters. A dozen or so clusters can be distinguished on the SOM surface. In order to identify them, we specified the location of

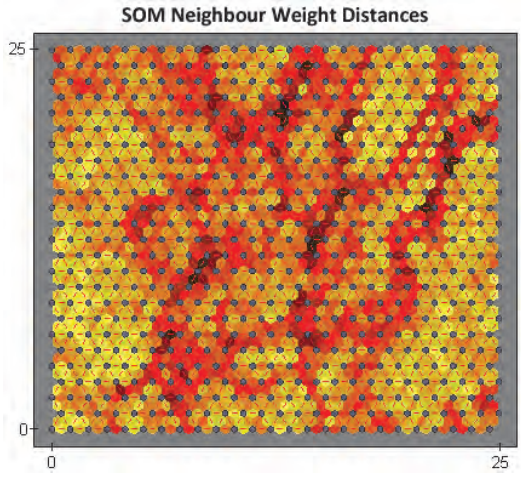

Fig. 1. Version A SOM 25 × 25; darker cells indicate longer distances between date vectors, thus separating the clusters of similar objects (brighter cells); the distance map for data containing coordinates; source: own study

data vectors of individual land capability units of soils.

Figure 2 presents a set of histograms illustrating the number of observations of individual complex labels (after the map calibration) on the SOM. Distribution of these observations shows similarity between data vectors from different complexes. By interpreting a domination of specific labels in individual cells as an indicator that a certain cell belongs to a certain class, it is possible to show the location of complexes on the SOM (Fig. 3). Typical is the mixing of labels of individual complexes, which is probably a result of horizontal coordinates being included in the data vector. Indirectly, it can be inferred that the classification task is relatively complicated and requires the use of rather large classification architectures.

The calibrated SOM image shows in the central map band the presence of durable grassland vectors, separating two groups of arable land vectors. Noticeable is the presence of two clusters of the K2 complex, and a relatively small set of Z1 vectors, indicating that this complex is significantly different. It is difficult to draw conclusions about the vector interrelations on the basis of this image, although in general the location of clusters belonging to individual complexes corresponds to expectations. It is known that the division into complexes generally does not form an ordinal scale, but the SOM situates individual clusters in terms of similarity. It can be noticed that the K2 complex is distinctly separated from the K3, though close to K4 and K5.

RESULTS OF CLUSTERIZATION OF SOIL DATA VECTORS FOR INPUT DATA IN VERSION B

A more compact map structure was expected in version B of the SOM because the coordinates were excluded. In addition to the configuration, dominating in the vector is lithological and hydrological information. Also in this case, a dozen or so clusters can be distinguished in the SOM (Fig. 4).

Figure 5 with histograms of complex vectors location on the SOM indicates a more compact diversification. With coordinates absent, the image is affected only by the components connected with grain size, land elevation, neighbourhood configuration, location relative to the supposed groundwater table.

Figure 5 presents histograms of location of individual complexes' vectors on the SOM which is also synthetically presented in Figure 6. It indicates heterogeneity of complexes in terms of physical properties or land relief. The image includes three complex K2 clusters, three K6 clusters, two permanent grassland clusters. The K8 clusters are "stuck" within K2, similarly to K4.

The SOM image has a significant informative importance. It shows that regardless of the input vector composition, the modelling of diversification of

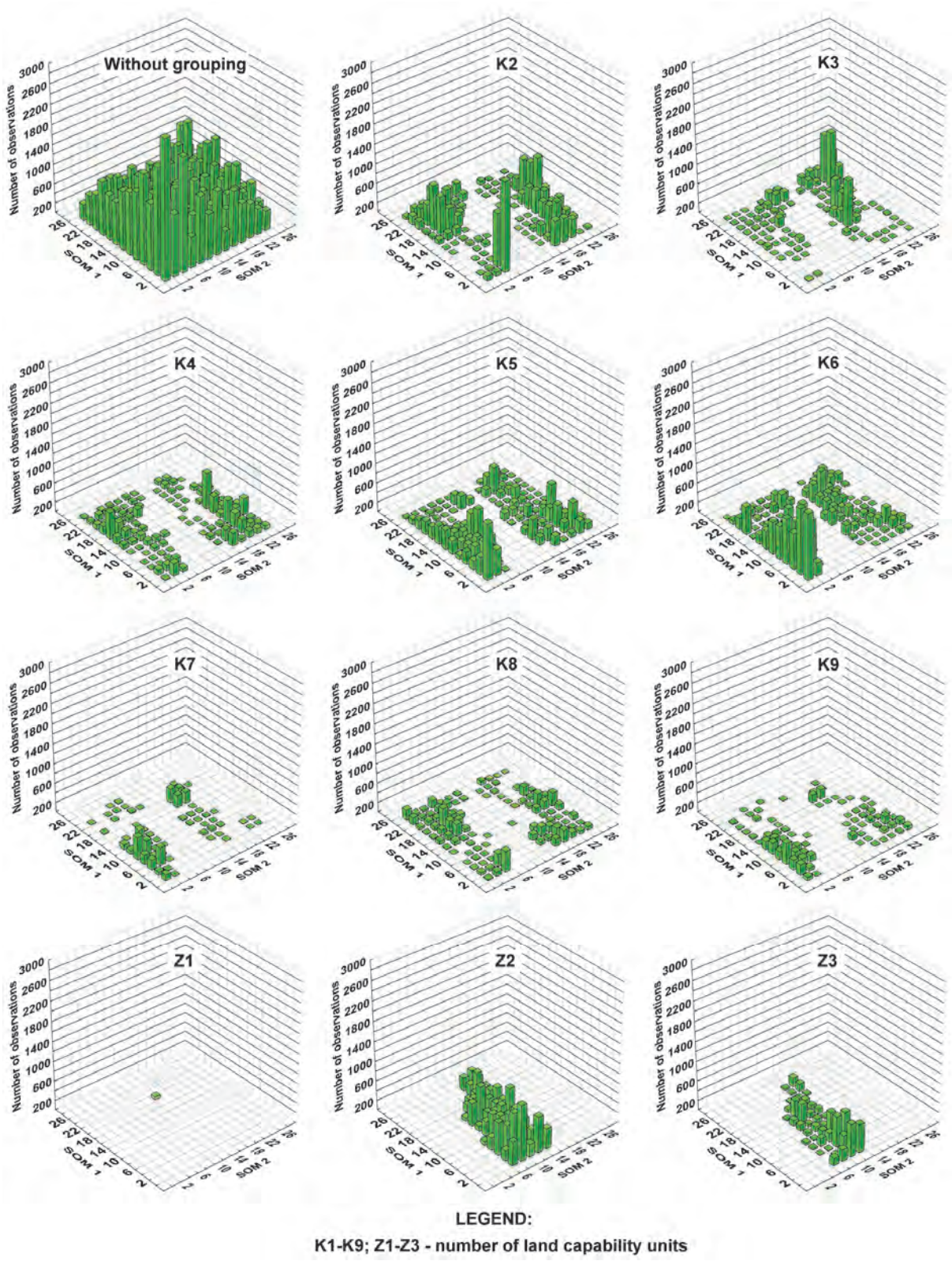

Fig. 2. Histograms of vectors' location on the self-organizing map (SOM), version A; agreed vertical scale; the histograms show the location of complexes' vectors on the SOM; the top chart on the left-hand side includes location of all validation set vectors; source: own study

SOM 2 \ SOM 1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
25	K4	K5	K5	K6	K6	K8	K4	K4	K4	K2	K8	K8	K2	K2	K2	K2	K4	K2	K2	K6	K6	K3	K3	K3	K3
24	K5	K4	K5	K5	K6	K9	K4	K4	K5	K5	K2	K8	K2	K2	K2	K2	K2	K2	K2	K5	K3	K3	K3	K3	K3
23	K5	K5	K5	K5	K5	K6	K4	K5	K5	K5	K5	K2	K2	K2	K2	K2	K2	K2	K3	K5	K3	K6	K6	K3	K6
22	K8	K5	K5	K5	K5	K5	K4	K4	K2	K5	K4	K2	K5	K2	K5	K2	X	K5	K5	K6	X	K6	K5	K3	K3
21	K2	K2	K2	K4	K2	K5	K5	K5	K5	K2	K2	K2	K4	K3	K3	X	K3	K3	X	X	K6	K6	K3	K5	K3
20	K2	K2	K2	K2	K2	K2	K5	K6	K6	K6	K7	X	K6	K3	K5	K3	K7	X	K6	K5	K5	K5	K4	K3	K3
19	K2	K8	K2	K2	K2	K2	K6	K7	K5	K3	K6	K5	K6	K3	K3	X	K6	K7	K6	K6	K6	K5	K5	K3	K3
18	K2	K2	K4	K2	K2	K8	K3	K6	K5	K3	K3	K5	K5	K3	K6	X	K6	K7	K6	K6	X	K5	X	K3	X
17	K5	K4/ K5	K5	K5	K8	K2	K8	K5	K3	K3	K3	K3	K6	K6	K6	K6	X	K7	X	X	Z2	X	Z2	Z2	Z2
16	X	X	X	K8	K5	K5	K6	K6	K2	X	X	K6	K6	K5	Z2/ Z3	Z3	Z2	Z2	Z2	Z2	Z1	Z2	Z2	Z2	Z2
15	Z2	Z2	X	K5	K9	X	Z3	X	Z3	X	Z2	K6	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	X	Z2	X
14	Z2	Z2	Z2/ Z3	X	Z3	Z2	Z2	Z3	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z3	Z3	Z2	Z2	Z2	Z2	X	K5	K6	K2
13	Z2	Z3	Z3	Z2	Z3	Z2	Z2	Z3	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z3	Z3	Z2	K5	K6	K3	K2
12	Z3	Z3	Z2	Z3	Z2	Z3	Z2	Z3	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z3	Z3	Z2	Z2	X	K6	K6	K3	K2
11	Z3	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z3	Z2	Z2	Z2	Z3	Z3	Z2	Z2	Z2	K5	K5	K3	K5	K2
10	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z3	Z2	Z2	Z2	Z3	Z2	X	K3	K6	K6	K2	K2	K2
9	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	K4	Z3	K6	X	X	X	X	X	K4	K4	K2/ K3	K6	K3	K8	K6
8	X	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	Z2	K5	K5	K5	K5	K5	K3	K2	K3	K8	K2	K2	K2	K3	K6	K2
7	X	X	X	X	X	Z2	K5	X	K5	K2	K5	K5	K5	K6	K6	K4	K8	K2	K4	K2	K2	K8	K6	K2	K2
6	K2	K4	K2	K2	K5	K5	K6	K6	K6	K5	K8	K5	K4	K5	K5	K8	K2	K2	K8	K4	K6	K9	K2	K6	K2
5	K2	K2	K4	K5	K6	K6	K6	K5	K6	K6	K6	K5	K5	K5	K5	K2	K2	K2	K8	K2	K2	K2	K2	K6	K2
4	K5	K4	K4	K6	K6	K6	K6	K6	K5	K6	K6	K6	K8	K6	K6	K8	K8	K2	K2	K2	K2	K2	K4	K2	K2
3	K5	K2	K5	K6	K6	K6	K6	K6	K6	K6	K6	K7	K5	K9	K5	K2	K2	K2	K2	K2	K2	K2	K2	K2	K2
2	K6	K5	K4	K6	K5	K6	K6	K6	K6	K7	K9	K6	K6	K5	K4	K4	K4	K4	K5	K4	K2	K2	K6	K2	K8
1	K6	K5	K6	K6	K6	K6	K6	K6	K6	K6	K6	K6	K6	K5	K5	K5	K4	K4	K6	K6	K6	K5	K5	K8	K8

Fig. 3. Location of complexes on the self-organizing map (SOM) for input data in version A; source: own study

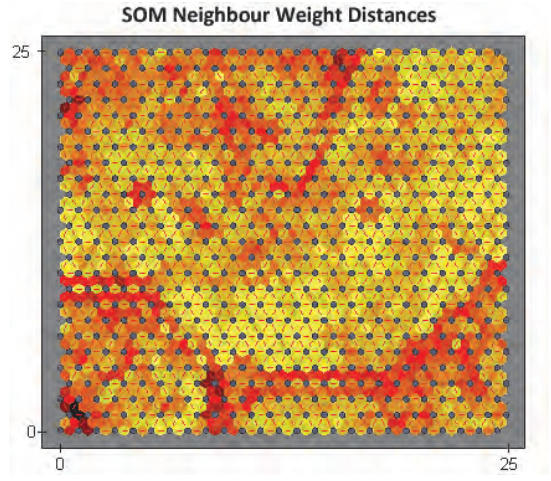

Fig. 4. Version B SOM 25 x 25; darker cells indicate longer distances between date vectors, thus separating the clusters of similar objects (brighter cells); the distance map for data containing the without coordinates; source: own study

complexes requires algorithms suitable for classification of non-linearly separable data. Natural candidates are classifiers and algorithms of computational intelligence and machine learning.

CLASSIFICATION MODELS USED IN RESEARCH

Classifiers are tools which implement the algorithms of vector data labelling. Vectors can represent an ordered set of quantitative or qualitative data, and labels are tags of classes of these objects. The classification algorithms are subject of many published papers, they also have an extended systematics of their own [ABE 2010; BALDI *et al*. 2010; BREIMAN *et al*. 1984; CHMIELNICKI *et al*. 2010; DEHZANGI *et al*. 2010; LANDGREBE *et al*. 2007).

In most general terms, the classifiers obtain the desired properties during the process of optimization, called learning. The process involves an intentional modification of randomly initiated processing algorithm on the set on isolated correctly classified cases. The main progress in the machine learning field results from appearance of new classifier architectures, but mostly from new learning methods. From the point of view of technical application of classifiers, the most important property is probably their suitability under conditions of absence of linear separability of data vectors, which manifests itself for instance as occurrence of many clusters of vectors which belong to the same class.

The presented results were obtained using three algorithms made available on Oxdata (currently www.h2o.ai) in order to facilitate processing of huge amounts of data (BigData) in a computation cloud. In many elements, the package interface refers to water (h2o, Sparkling Water, Flow).

The DL algorithm (deep learning as an alternative to shallow networks) has an architecture similar to classical multilayer perceptron (MLP), an iteratively optimized classifier, usually with sigmoidal transfer functions. Differences from classical model comprise: more than one hidden layer with many processing units each (also possible in MLP, but usually much less developed and seldom in two layers which require a different optimization method), RLU-transfer functions (Rectifier Linear Unit – value range <0, ∞>, contrary to MLP where the output values are <0, 1> or <–1, 1>), and regularization of processing parameters [CANDEL *et al*. 2015; LECUN *et al*. 2015].

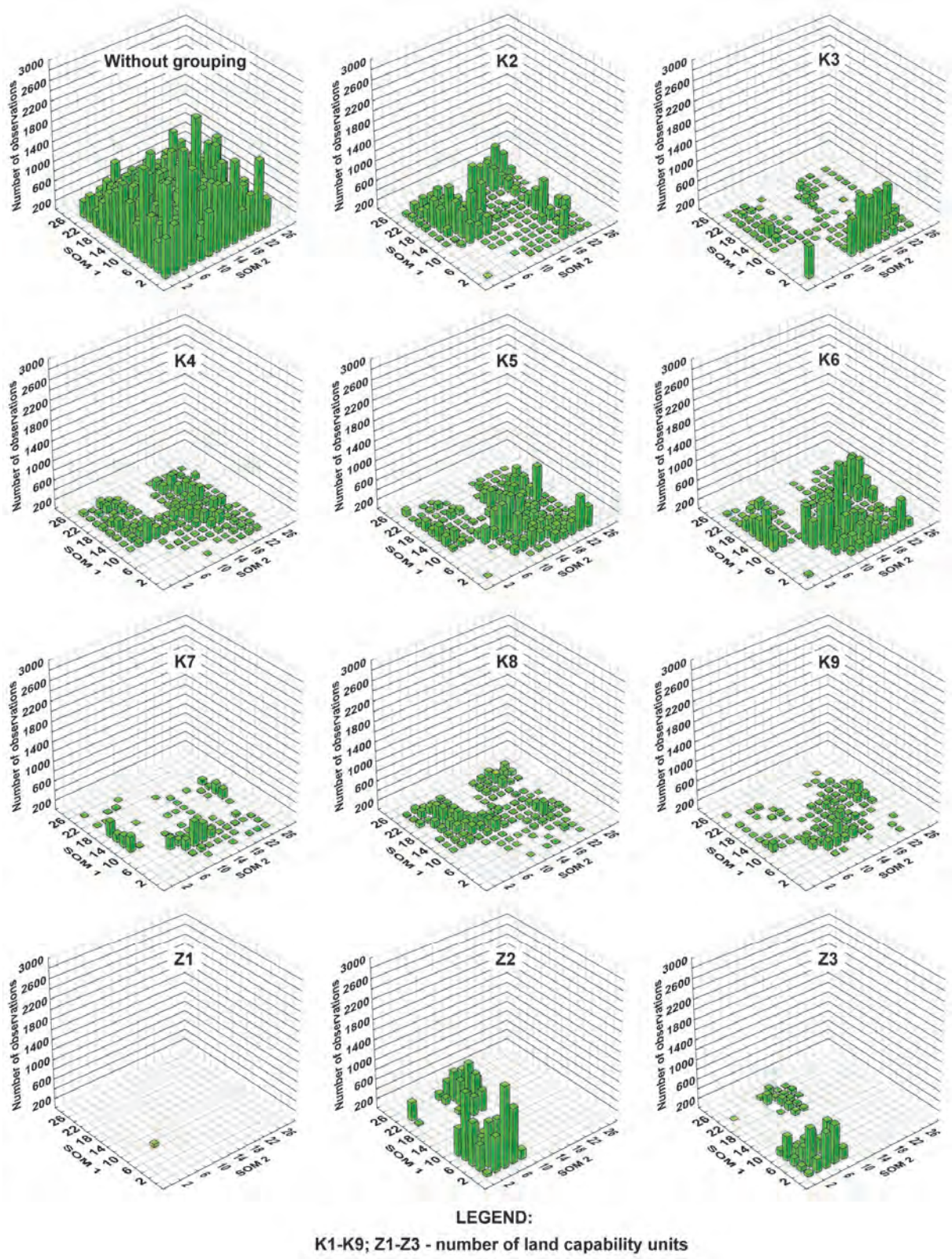

Fig. 5. Histograms of vectors' location on the self-organizing map (SOM), version B; agreed vertical scale; the histograms show the location of complexes' vectors on the SOM; the top chart on the left-hand side includes location of all validation set vectors; source: own study

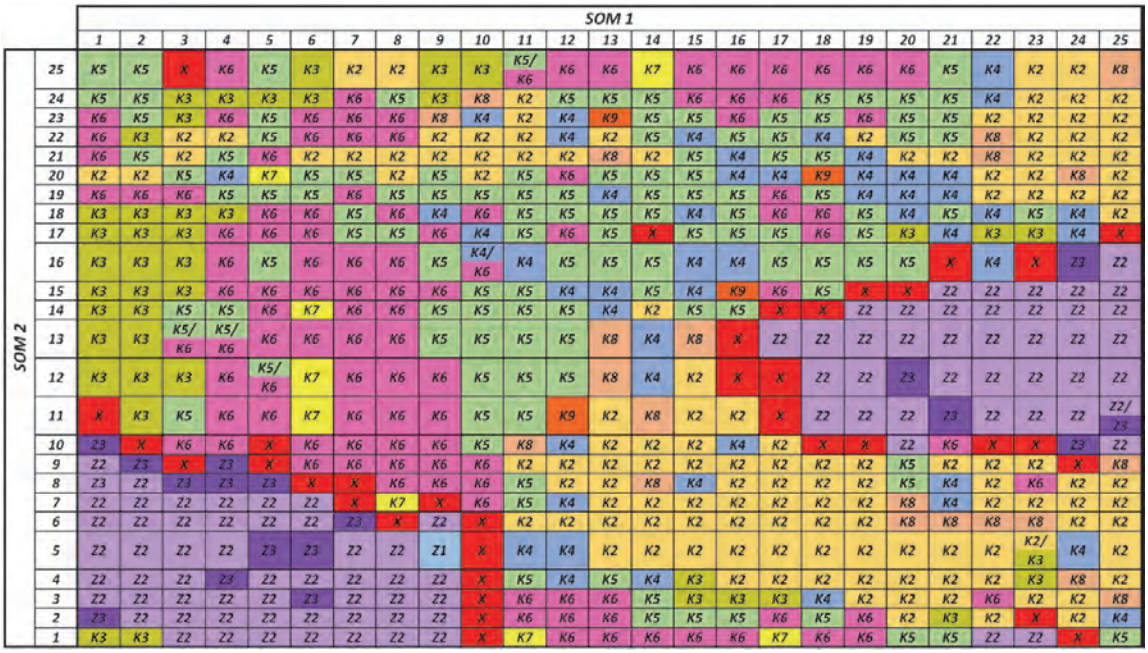

Fig. 6. Location of complexes on the self-organizing map (SOM) for input data in version B; source: own study

The DRF algorithm (Distributed Random Forest – as a reference to random tree algorithms) is an adaptive decision algorithm (classification and regression) in which a large number of random trees (built on the basis of data analysis) with a relatively low identification capability, treated as a set of classifiers, generate a relatively small identification error. The successive model (tree) components are constructed on the basis of data groups randomly drawn from the training set. In fact, this model represents a set of weak classifiers with individually weak classification capabilities which as an ensemble however give very correct indications [GISLASON *et al*. 2006].

The GBM algorithm (Gradient Boosting Machines), just like DRF leading to a set of "weak" classification trees, using random data sampling and so-called greedy optimization algorithms as well as boosting in less identifiable areas [CLICK *et al*. 2015; FRIEDMAN 2000; NATEKIN, KNOLL 2013].

In both cases the maximum number of declared trees was 100 and their depth 25. The output in each model was a softmax function which is considered a reflection of probability of a given label. The datasets used were analogous as in the SOM, with (version A) and without coordinates (version B).

In the light of two configurations of input data and three types of models, the principle of choosing the most favourable model must be specified. The basic model evaluation is percentage of correct indications of the validation set. This measure however fails in case of quantitatively unbalanced sets (domination of one class). Similarly, the evaluation of multi-class models is also a sort of problem. Both circumstances occur in this particular case: quantitative unbalance of classes (class size varies from a few dozen to more than 20 thousand data records) and multi-class structure (11 complex-classes).

The analysis used a method suggested in publications on statistics, modifying the evaluation indicators of two classes. Four types of classified indications are distinguished in the binary evaluation: *TP* (true positive – correct indication of discriminated class), *TN* (true negative – correct indication of alternative class), *FP* (false positive – false indication of discriminated class, indications of examples which belong to alternative class as belonging to discriminated class), *FN* (false negative – false indication of alternative class). The transformation of multi-class classification takes place according to this division: all classes, one after another, play the part of the discriminated class; the alternative class comprises all classes together which are not the discriminated class. The weakness of this approach is intensification of the problem quantitative class unbalance.

The numbers of above-mentioned indications can be used to estimate the quantitative identification indicators of individual classes. Thus, each model is evaluated many times, depending on their number.

Indicator to estimate the classification precision within the analysed class:

1) precision, positive predictive value, *PPV*: $PPV = TP/(TP + FP)$;

Indicator of estimating the ability to identify the discriminated class

2) recall, true positive rate, sensitivity, *TPR*: $TPR = TP/(TP + FN)$;

General indicator $F1$ of identification correctness:

3) $F1 = 2PPV \cdot TPR/(PPV + TPR)$.

All indicators have similar informative properties: value close to 0 (zero) means a poor quality of the model in terms of identification of discriminated class; value close to 1 (one) means good identification properties.

RESULTS AND DISCUSSION

The basic information on percentages of correct indications of used classification models was shown in Table 1. Column one indicates the complex symbol, column two includes number of examples of a given complex in the validation set. In version A (input vector with coordinates) the best results are obtained with the DRF algorithm (total of 30 incorrect indications) and the GBM. The DL algorithm gives the worst results in both versions. The table also indicates that the biggest potential problems are related to the K8 complex (despite a relatively large share in the training and validation set). Complex Z1 is not identified by the DL algorithm at all, but is well discriminated by the remaining two algorithms.

Table 1. Percentages of correct indications of the validation set elements by the classification algorithms by complexes: K

K	*N*(Kx)	Version A			Version B		
		DL	DRF	GBM	DL	DRF	GBM
K2	21 269	84.80	99.96	99.66	92.74	99.49	99.04
K3	9 678	86.56	100.00	99.48	78.65	99.26	98.64
K4	9 571	78.46	99.99	99.29	67.13	95.73	98.03
K5	18 626	85.95	99.97	98.87	77.38	98.60	97.72
K6	23 067	84.92	99.99	99.23	84.97	98.03	97.97
K7	3 810	86.30	99.97	99.97	81.23	99.08	99.71
K8	7 348	59.93	99.96	98.68	22.81	88.49	94.07
K9	3 994	44.17	99.85	99.10	23.11	83.10	92.84
Z1	62	0.00	100.00	100.00	0.00	100.00	100.00
Z2	19 593	90.04	99.99	99.93	91.68	99.92	99.71
Z3	7 308	71.63	100.00	99.75	55.93	96.46	96.92
% val	124 326	81.93	99.98	99.40	76.83	97.48	98.04

Explanations: K2–K9 and Z1–Z3 = land capability units of soils; DL = Deep Learning neural network with 2 hidden layers, 200 units each; DRF = distributed random forest (100 trees, depth 20), GBM = Gradient Boosting Machine (100 trees, depth 20), *N* = the number of cases representing of land capability unit.
Source: own study.

In addition to the general indicator of standards recognition, the evaluation of suitability of multi-class classifiers also requires analysing the share of false indications and instances of membership in individual classes. To some extent this is possible using the *PPV*, *FPR* and *F*score indicators. Table 2 includes the indicators for both versions of input vectors of the DL model. Taking into account the distribution of individual indicator values, the model should be evaluated as poor. The best *F*score values do not exceed 0.7, not to mention the total insensitivity of the model to the Z1 vectors. In comparison with poor usefulness and accuracy of the DL model, the Distributed Random Tree (DRF) algorithm performed much better. In the version with coordinates (A) the total number of erroneous indications in more than 120 thousand data records was about 30. Even very poorly represented Z1 complex (62 indications) was fully recognized, which at low level of alternative indication error gives a relatively high *F*score = 0.80. *F*score is much lower in version B because of a large number of erroneous indications, despite a relatively high share (97.4%) of correct indications of the entire validation set. The DRF algorithm is in fact represented by a set of classifiers (random trees) which according to numerous observations has better properties than a single classifier.

The GBM algorithm has an identification boosting element which in successive model optimization cycles increases the weights of incorrectly identified data vectors. It performs slightly worse than DRF, but definitely better than DL. Please note that the effect of each algorithm is to some extend random, so it can be supposed that modification of the model structure (size) and initialization could lead to slightly different results, although the relations between the correctness of their indications will remain similar.

A graphical comparison of indications by individual models in presented in Figure 7. It is a set of three-dimensional histograms on which both horizontal axes are the scale of complexes. In the best classifier all observations should be located on the main axis of the charts. Presence of results outside the main axis denotes erroneous indications. The chart of distribution of complexes indications by individual models (Fig. 8) presents indirectly the differentiation of the models credibility.

The quality of models can be also shown on the diagrams of location of the *PPV* and *TPR* points in the Cartesian coordinate system (Fig. 9). In the optimum classifier the points corresponding to individual classes will lie as close to the (1, 1) coordinate as possible. The diagrams indicate the advantage of classification tree algorithms, particularly DRF which classifies the vectors which contain point coordinates.

The indications results of the DRF model in version A allow an attempt to generalize the interrelations of objects subject to classification. As the *softmax* function was used as the models output, the models can be treated as a representation of evaluation of probability distribution of models membership in individual classes. Here, the classifier is an algo-

Table 2. Positive predictive value (PPV), true positive rate (TPR) and *F*score values of correct identification of land capability units by the three models

K	Version A			Version B		
	PPV	TPR	*F*score	PPV	TPR	*F*score
Deep Learning model						
K2	0.8327	0.4888	0.6160	0.7139	0.4852	0.5778
K3	0.7936	0.2942	0.4292	0.7578	0.2256	0.3477
K4	0.8273	0.2642	0.4004	0.7974	0.1911	0.3083
K5	0.8080	0.4615	0.5875	0.7717	0.3698	0.5000
K6	0.8648	0.5021	0.6354	0.7955	0.4517	0.5762
K7	0.8282	0.1310	0.2263	0.7912	0.0995	0.1768
K8	0.6219	0.1819	0.2815	0.5958	0.0571	0.1042
K9	0.6937	0.0752	0.1356	0.5978	0.0317	0.0602
Z1	X	0.0000	X	X	0.0000	X
Z2	0.8920	0.4643	0.6108	0.8443	0.4131	0.5548
Z3	0.7284	0.2032	0.3177	0.7123	0.1307	0.2209
Distributed Random Forest model						
K2	0.9999	0.9987	0.9993	0.9423	0.9200	0.9310
K3	0.9999	0.9970	0.9985	0.9828	0.7659	0.8609
K4	0.9997	0.9972	0.9984	0.9923	0.7494	0.8539
K5	0.9999	0.9984	0.9992	0.9708	0.8767	0.9214
K6	0.9997	0.9990	0.9994	0.9724	0.9007	0.9352
K7	0.9995	0.9927	0.9961	0.9808	0.5523	0.7067
K8	0.9986	0.9973	0.9980	0.9921	0.6784	0.8058
K9	0.9997	0.9928	0.9963	0.9991	0.5146	0.6793
Z1	1.0000	0.6739	0.8052	1.0000	0.0194	0.0381
Z2	1.0000	0.9985	0.9992	0.9869	0.9126	0.9483
Z3	0.9997	0.9962	0.9980	0.9979	0.6933	0.8181
Gradient Boosting Machines model						
K2	0.9913	0.9743	0.9828	0.9693	0.9197	0.9438
K3	0.9939	0.9342	0.9631	0.9869	0.7647	0.8617
K4	0.9950	0.9318	0.9623	0.9901	0.7539	0.8560
K5	0.9911	0.9696	0.9802	0.9734	0.8758	0.9220
K6	0.9923	0.9759	0.9840	0.9847	0.9007	0.9408
K7	0.9956	0.8397	0.9110	0.9857	0.5539	0.7092
K8	0.9932	0.9126	0.9512	0.9836	0.6916	0.8122
K9	0.9980	0.8432	0.9141	0.9904	0.5422	0.7007
Z1	1.0000	0.0769	0.1429	1.0000	0.0194	0.0381
Z2	0.9991	0.9642	0.9814	0.9886	0.9124	0.9490
Z3	0.9982	0.9089	0.9514	0.9920	0.6943	0.8169

Explanation: X = not determinable value (value "0" in the denominator).
Source: own study.

rithm transforming the data vector into the vector of distribution of probability of membership in individual complexes. One should expect that the distributions will indicate indirectly the mutual proximity of complexes.

Figure 10 is a representation of the SOM in which the input data vectors were probabilities of complexes. The clusters of calibrated SOM are more compact than in the image obtained from the raw data vectors. The number of clusters corresponds to the number of discriminated data classes (complexes) which indicates the transformation of distributed representation of raw vectors into the set of linearly separable data.

Analogously to some statistical models, the adaptive algorithms allow to make a ranking of relative importance of variables which make up the input vector (Fig. 11). The chart indicates that in addition to the land use which is the main differentiating criterion also important are: grain size (mainly the factor depending on the content of floating parts), groundwater table and land relief factors (grade, difference in height between the neighbourhood and the observation point). The neighbourhood elevation is less important.

The advantage of data-based models lies in their informative value: depending on the range of input data, they can be a useful tool to predict the soil and land response to anthropogenic factors (fertilization, drainage, surface deformation) or to natural factors (temperature increase). Their advantage and strength is the base of construction: actual data; their disadvantage is a "black box" character, although current studies to a large extent allow also to detect quantitative and qualitative relationships stored in their processing structures [GISLASON *et al*. 2006].

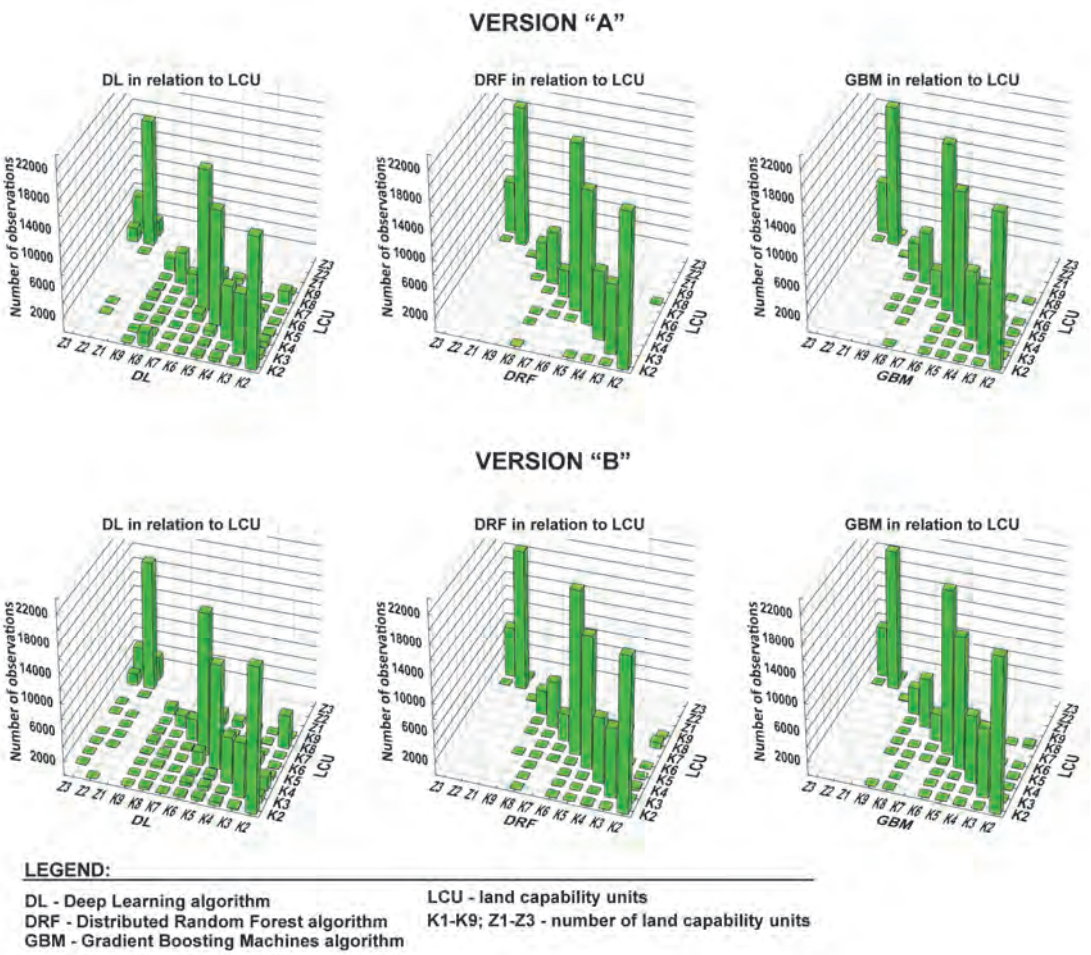

Fig. 7. Histogram of conformity of indications to the standards: version A with XYZ coordinates, and version B without coordinates; classification algorithms: DL, DRF and GBM; source: own study

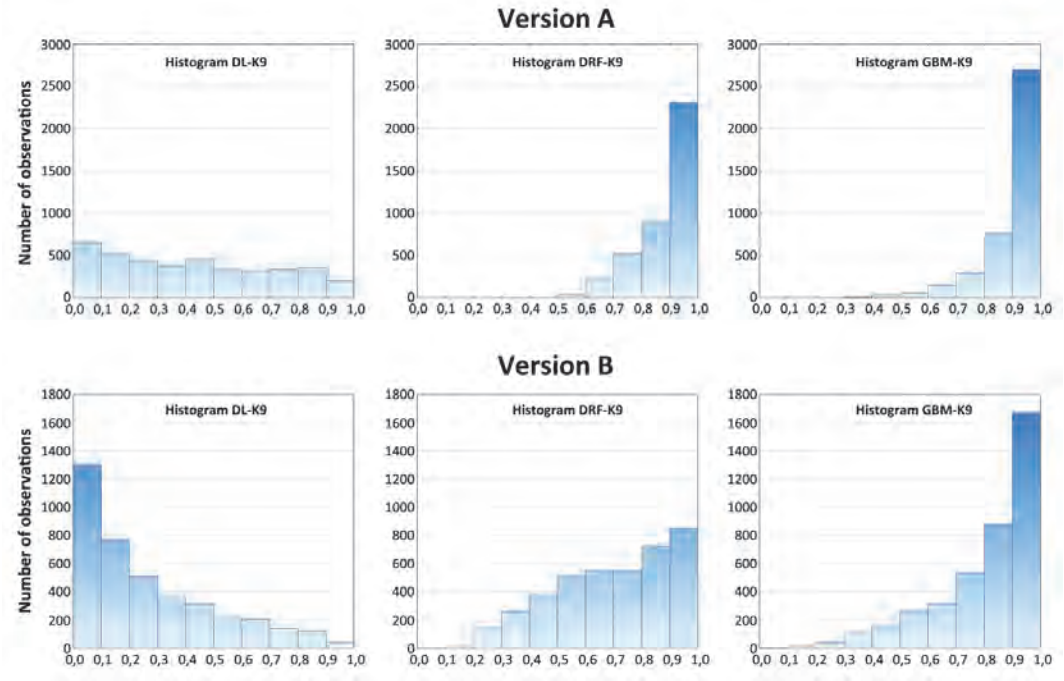

Fig. 8. Example of histogram of distribution of indications for land capability units number 9: version A with XYZ coordinates, and version B without coordinates; classification algorithms: DL, DRF and GBM; source: own study

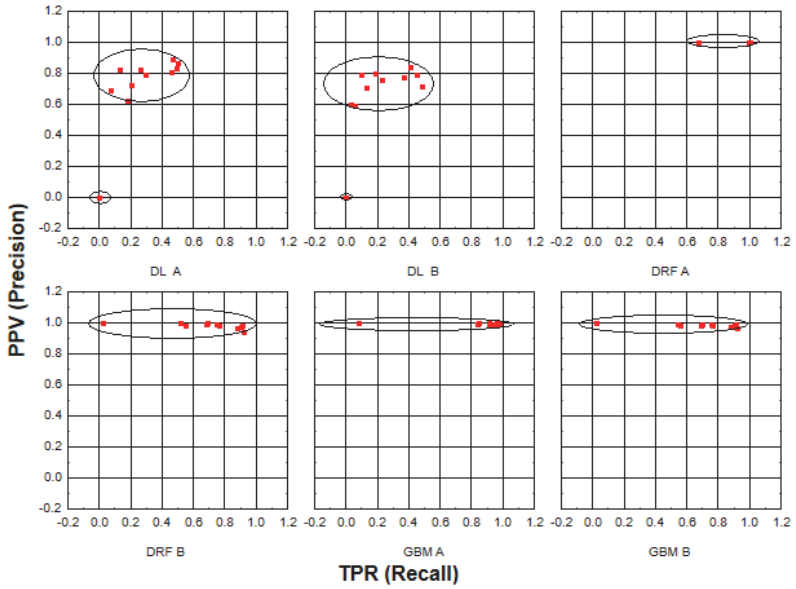

Fig. 9. Diagrams of location of the positive predictive value (PPV) and true positive rate (TPR) points in the Cartesian coordinate system; source: own study

SOM 2 \ SOM 1	1	2	3	4	5	6	7	8	9	10
10	K4	K4	K4	K4	K4	K7	K7	K7	K5	K5
9	K4	K4	K4	K4	K6	K7	K5	K5	K5	K5
8	K6	K6	K6	K6	K6	K6	K5	K5	K5	K5
7	K6	K6	K6	K6	K6	K6	K5	K5	K5	K5
6	K6	K6	K6	K6	K6	K9	K9	K5	K5	K5
5	Z3	Z3	K8	K8	K9	K9	K9	K9	K5	K5
4	Z3	Z3	K8	K8	K9	K9	K9	K9	K3	K3
3	Z3	Z3	K8	K8	K8	K9	K9	Z1	K3	K3
2	Z2	Z2	K8	K2	K2	K2	K2	K2	K3	K3
1	Z2	Z2	Z2	K2	K2	K2	K2	K2	K3	K3

Fig. 10. Representation of the self-organizing map (SOM) in which the input data vectors were probabilities of complexes; source: own study

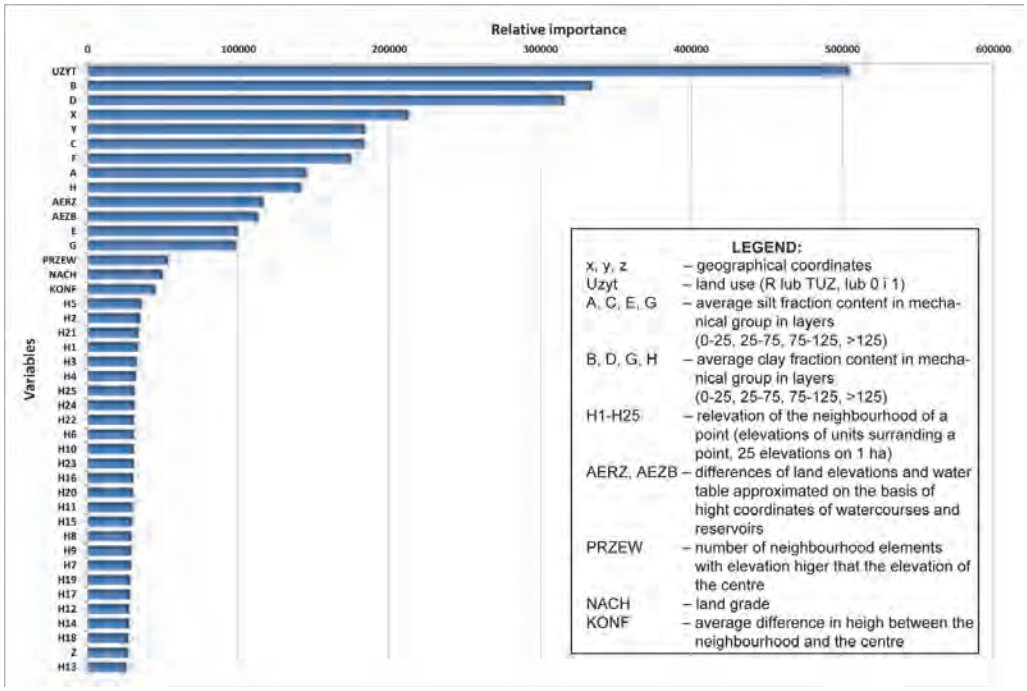

Fig. 11. Evaluation of relative importance of input variables; source: own study

CONCLUSION

Digital Soil Mapping is considered in the EU as basic information resource linked with the soil management policy on the European scale [DOBOS *et al.* 2006]. In addition to the possibility of digitizing the existing soil and cartographic documentation by means of copying the thematic maps, worth considering is a gradual transition to the technique without spatial data on the soils. Abandonment of existing analogue materials is hard to imagine; to the contrary, they should be a basis of the system as an *a priori* information in relation to further, detailed field studies, with full morphological and physiographical information, and with credible spatial reference. Relativization of soil information is under these conditions inevitable. Modelling is a solution of this problem used worldwide; it often uses data fathoming algorithms: computational intelligence, fuzzy inference, machine learning. The presented results indicate that digitization of thematic and topographic maps provides a rather good basis to develop useful soil classification (qualification) models.

In this case the Distributed Random Forest algorithm was the best solution, particularly DRF which classifies the vectors which contain point coordinates. Among of variables which make up the input vector, the most important was land use. Slightly less importance gained: grain size (mainly the factor depending on the content of floating parts), groundwater table and land relief factors. The neighbourhood elevation was the least important.

However it may not be inferred that the best algorithm in the described studies can be treated as a general rule of constructing the inference systems, but it can be imagined that this issue is to some extent regional, despite numerous advantages of random trees (scalability, low hardware requirements, resistance to overfitting).

Acknowledgements

Dofinansowano ze środków Wojewódzkiego Funduszu Ochrony Środowiska i Gospodarki Wodnej w Lublinie
Cofinanced by Voivodeship Fund for Environmental Protection and Water Management in Lublin

REFERENCES

ABE S. 2010. Support vector machines for pattern classification. New York. Springer Verlag. ISBN 978-1-84996-098-4 pp. 473.

BALDI P., BRUNAK S., CHAUVIN Y., ANDERSEN C., NIELSEN H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. Vol. 16 p. 412–424.

BREIMAN L., FRIEDMAN J.H., OLSHEN R.A., STONE C.J. 1984. Classification and regression trees. Wadsworth, Belmont. ISBN 0-412-04841-8 pp. 368.

CANDEL A., PARMAR V., LEDELL E., ARORA A. 2015. Deep learning with H_2O [online]. [Access 05.02.2016]. Available at: https://h2o-release.s3.amazonaws.com/h2o/rel-slater/9/docs-website/h2o-docs/booklets/DeepLearning_Vignette.pdf

CHMIELNICKI W., STĄPOR K., ROTERMAN-KONIECZNA I. 2010. An efficient multi-class support vector machine classifier for protein fold recognition. Advances in Intelligent and Soft Computing. Vol. 74 p. 77–84.

CLICK C., MALOHLAVA M., PARMAR V., ROARK H., CANDEL A. 2015. Gradient boosted models with H_2O [online]. [Access 05.02.2016]. Available at: https://h2o-release.s3.amazonaws.com/h2o/rel-slater/9/docs-website/h2o-docs/booklets/GBM_Vignette.pdf

DEHZANGI A., PHON-AMNUAISUK S., DEHZANGI O. 2010. Using random forest for protein fold predition problem: An empirical study. Journal of Information Science and Engineering. Vol. 26. No. 6 p. 1941–1956.

DOBOS E., CARRÉ F., HENGL T., REUTER H.I., TÓTH G. 2006. Digital Soil Mapping as a support to production of functional maps. EUR 22123 EN. Office for Official Publications of the European Communities. Luxembourg pp. 68.

FRIEDMAN J. 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics. Vol. 29. Iss. 5 p. 1189–1232.

GISLASON P.O., BENEDIKTSSON J.A., SVEINSSON J.R. 2006. Random forests for land cover classification. Pattern Recognition Letters. Vol. 27. Iss. 3 p. 294–300.

JENNY H. 1941. Factors of soil formation – A system of quantitative pedology. New York, USA. McGraw-Hill pp. 281.

JONES A., MONTANARELLA L., JONES R. (ed.) 2005. Soil atlas of Europe. Luxembourg. Office for Official Publications of the European Communities pp. 128.

LANDGREBE T.C.W., DUIN R.P.W. 2007. Approximating the multiclass ROC by pairwise analysis. Pattern Recognition Letters. Vol. 28. Iss. 13 p. 1747–1758.

LECUN Y., BENGIO Y., HINTON G.E. 2015. Deep learning. Nature. Vol. 521 p. 436–444.

MCBRATNEY A.B., MENDONÇA SANTOS M.L., MINASNY B. 2003. On digital soil mapping. Geoderma. Vol. 117. Iss. 1–2 p. 3–52.

NATEKIN A., KNOLL A. 2013. Gradient boosting machines, a tutorial. Frontiers in Neurorobotics. Vol. 7. Article 21. DOI 10.3389/fnbot.2013.00021.

STRZEMSKI M., BARTOSZEWSKI Z., CZARNOWSKI F., DOMBEK E., SIUTA J., TRUSZKOWSKA R., WITEK T. 1964. Instrukcja w sprawie wykonywania map glebowo-rolniczych w skali 1: 5000 i 1: 25000 oraz map glebowo-przyrodniczych w skali 1: 25000. Załącznik do Zarządzenia nr 115 Ministra Rolnictwa z dnia 28 lipca 1964 r. w sprawie organizacji prac gleboznawczo- i rolniczo-kartograficznych. Dz. Urz. Min. Rol. Nr 19 poz. 121.

TADEUSIEWICZ R. 1993. Sieci neuronowe [Neural networks]. Warszawa. Akademicka Oficyna Wydawnicza. ISBN 83-85769-03-X pp. 195.

ZHU A.X., HUDSON B., BURT J., LUBICH K., SIMONSEN D. 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Science Society of America Journal. Vol. 65. Iss. 5 p. 1463–1472.

Krzysztof URBAŃSKI, Stanisław GRUSZCZYŃSKI

Adaptacyjne modelowanie przestrzennego zróżnicowania jednostek klasyfikacyjnych gleb

STRESZCZENIE

Wraz z rozwojem technologii informatycznych następuje stopniowa zmiana podejścia do dokumentacji kartograficznej obiektów przyrodniczych, w tym gleb. Podstawowymi cechami dowolnej klasyfikacji, których przedmiotem są gleby, jest wielowymiarowość jednostek (nie ma pojedynczej właściwości, możliwej do wyznaczenia w drodze pomiaru, która wystarczałaby do jednoznacznego przypisania pedonu do określonej klasy w stosowanych skalach klasyfikacyjnych gleb), w związku z czym właściwe wydaje się wykorzystanie do tego celu dostępnych komputerowych metod przetwarzania danych. Modelowanie przestrzennego zróżnicowania gleb na podstawie przesłanek natury fizjograficznej, odtworzonych na podstawie digitalizacji istniejących materiałów kartograficznych, jest podstawą do tworzenia przestrzennych baz danych przechowywanych w wersji cyfrowej. Inaczej niż w typowej kartografii tematycznej zawierającej treści glebowo-siedliskowe, modele te wskazują na prawdopodobieństwo *a priori* występowania określonej jednostki glebowej w określonym położeniu, nie zaś bezwzględną przynależność terenu do niej. Taka interpretacja wymaga zbudowania stosownego algorytmu wiążącego czynniki natury geologicznej i fizjograficznej z jednostkami glebowymi. Do tego celu często wykorzystuje się tak zwane algorytmy adaptacyjne, umożliwiające elastyczne tworzenie modeli zależności bazujących na danych.

W pracy przedstawiono dwa warianty doboru parametrów przetwarzania danych fizjograficzno-glebowych potencjalnie przydatnych do tych celów. Wykorzystano dane pochodzące z bazy danych fizjograficzno-glebowych z rejonu GOP (Górnośląski Okręg Przemysłowy) uzyskanych w wyniku digitalizacji materiałów kartograficznych. Analizie poddano wyłącznie tereny użytków rolnych: gruntów ornych (R) i trwałych użytków zielonych (Ł i Ps). Na obszarze o powierzchni 1 km^2 wyodrębniono 6,4 mln jednostek tworzących siatkę kwadratów o rozmiarach 20 × 20 m. Wykorzystane zostały algorytmy samoorganizującej mapy (SOM) Kohonena oraz klasyfikatory – głęboka sieć neuronowa, oraz dwa rodzaje drzew decyzyjnych – rozproszony las losowy (ang. Distributed Random Forest) i wzmacniane drzewa losowe (ang. Gradient Boosting Machine). Szczególnie algorytm rozproszonego lasu losowego (algorytm DRF) wykazał bardzo wysoki stopień zdolności generalizacyjnej w modelowaniu zróżnicowania kompleksów glebowych.

Słowa kluczowe: *algorytmy adaptacyjne, Górnośląski Okręg Przemysłowy (GOP), klasyfikacja gleb, samoorganizująca mapa (SOM)*