ARCHIVES OF ENVIRONMENTAL PROTECTION
vol. 36 no. 2 pp. 25 - 34 2010
PL ISSN 0324-8461

THE BAYESIAN MODEL OF THE INTERDEPENDENCIES BETWEEN SOIL SORPTION FEATURES

STANISŁAW GRUSZCZYŃSKI

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

Keywords: Belief network, soil sorption feature, interdependencies.

Abstract: The paper presents a qualitative, Bayesian model used to determine some interdependencies between sorption features for mineral soils in southern Poland. Sorption properties are very important, crucial for measure of fertility, nutrient retention capacity, and the capacity to protect groundwater from contamination. Cation exchange capacity (CEC) is a commonly applied indicator of the soils conditions or vulnerability. Base saturation (BS) is an important element of hazard degree assessment in soils lying within reach of impact of acidifying agents. The considered soils represented different valuation classes and differed in their typology. The Bayesian model is used for interdependences assessment.

INTRODUCTION

The soil sorption properties (Cation Exchange Capacity – CEC and Base Saturation – BS) are regarded as important soil valuation criteria. It is assumed that soils characterized by higher cation exchange capacity (the CEC value) retain applied nutrients better than those with smaller values. Similarly, base saturation indicating percent saturation of cation exchange capacity by base cations is an important soil condition indicator. Complete vulnerabity estimation except for CEC and BS, needs to introduce the intensity parameter. It informs us about the sorption aspect resulting, for instance, from soil density, soil moisture and structure variation.

Traditionally, the characteristic of soil sorption properties (CEC and BS) is most often determined using the Kappen method. It involves determination of ions with base reaction and ions with acid reaction in a soil sample water extract. The sum of their capacity gives the computed cation exchange capacity, while the sum of ions with base reaction, given in percent, constitutes base saturation. The determination is relatively complex and time-consuming, therefore in some cases its substitute is being used (e.g. the extent of methylene blue sorption), or empirical equations [4, 7]. This is so because their parameters transform other soil properties, which are easier for laboratory testing. In soil science, these equations are called PedoTransfer Functions – PTF [6].

However, in some circumstances, it is enough to have only qualitative information, referring to the scale and state of soil sorption. This approach was successfully applied,

for example, in making the database of the soils of the European Union, where so-called PedoTransfer Rules (PTR) were adopted. The idea of constructing such a qualitative model results from the general knowledge of the relationship between sorption properties and the content of clay fraction and organic carbon, on one side, and the soil reaction on the other. Such a model can successfully be formalized. An efficient tool for the formalization can be a properly designed Bayesian belief network.

In Bayesian belief network, a prior probability distribution is assigned to each variable and then the strength of the dependence between each pair of variables is defined. A Bayesian network [1–3] is a probabilistic model that represents a set of random variables and their conditional independencies via a directed acyclic graph. Formally, Bayesian networks are directed acyclic graphs whose nodes represent variables, and whose edges encode conditional interdependencies between the variables. Generalizations of Bayesian networks that can represent and solve decision problems under uncertainty are called influence diagrams. A Bayesian network could represent the probabilistic relationships between reasons and effects.

MATERIAL AND METHODS

The soil material originated from studies carried out in the scope of implementation of a project concerning determination of soil quality in reclaimed areas [5]. In the scope of these studies, samples were taken from 67 soil pits in order to determine some physical and chemical properties. This was aimed to develop a database containing characteristics of soils in various (almost all) valuation classes. Soil pits were made in soils belonging to various valuation classes, from class II to VI of arable land. As regards typology, the soils represented Luvisols, Cambisols and Gleyic Phaeozems. Open pits were made in Southern Poland. Within the carried out investigations on 381 soil samples numerous physical and chemical properties were determined. The following analytical methods were employed for the purposes of determining the properties used in this work:

- grain-size distribution – Casagrande areometric method,
- MH, maximum hygroscopicity – in vacuum desiccator over 10% sulphuric acid,
- CEC and BS, cation exchange capacity and base saturation – the Kappen method,
- pH(1) soil reaction in water suspension, pH(2) soil reaction in KCL suspension – electrometrically,
- OC, organic carbon – in the CS-500 Eltra elementary analyzer, after having deducted inorganic carbon determined in the Scheibler apparatus,
- SBM, methylene blue sorption – the Peter-Markert method according to Myślińska [8].

The results of determinations were used to develop the Bayesian, qualitative, empiri-

Table 1. Basic statistical characteristics of soil samples used in the research: Clay – colloidal clay content in %, OC – organic carbon content in %, MH – maximum hygroscopicity in % by weight, SBM – methylene blue sorption in cmol/kg, Sor – CEC determined using the Kappen method, in cmol/kg, pH(1) – pH measured in water suspension, pH(2) – pH measured in KCl suspension, BS – base saturation in %

Statistics	Clay	OC	MH	SBM	Sor	pH(1)	pH(2)	BS
Average	13.2	0.37	3.23	8.7	12.2	6.43	5.51	76.1
Standard dev.	10.6	0.46	2.42	6.5	6.4	0.98	1.03	18.1
Minimum	0	0.0	0.1	0.1	1.2	3.6	3.3	9.6
Maximum	57	3.06	20.9	37.6	41.4	8.6	8.0	97.9

cal model interrelations between properties of soils. Table 1 specifies elementary measures of central tendency and dispersion of data contained in the database.

BAYESIAN MODEL OF SORPTIVE RELATIONS

Two circumstances connected with the issue of CEC and BS estimation are worth consideration: correlations between these factors and thus variables forming them, and considerable fuzzy relations between valuation of both components and qualitative classification of soils. In other words, oftentimes linguistic, imprecise and fuzzy evaluations are sufficient, apart from the cases of absolute need to use numerical values of estimations of both quantities. In these conditions, the Bayesian model may constitute a good tool to perform an analysis of correlations determining sorption effects in soil.

The Bayesian model, which belongs to the group of graphical models, is known in reference literature as the Bayesian network (*Bn* – Bayesian network, or belief network), and by assumption it reflects statistical relations between the system elements. This concluding model assumes [1, 2] the existence of domain **X**, for which the following are specified: probability distribution Ω and attributes $a_1, a_2, ..., a_n$ for a_i:**X**$\rightarrow A_i$ for $i = 1, 2, ..., n$. Conventionally, it is assumed for simplicity that the attributes are solely nominal, and we are interested only in the relationship between them. In practical applications of this model, it is common to use different terminology: attributes are referred to as variables, and the domain is constituted by the set of all possible attributions of variable values [2].

In recent years, the *Bn* networks have become popular as the concluding systems, mainly due to the fact that computers considerably facilitate rather complex calculations required in order to estimate probability of a certain event. At the same time, packages make the *Bn* design easier and their utilization have been popularized as well [3].

The key problem involved in generating the *Bn* cause-effect networks is the algorithm applied to reconstruct the form of relationships between the system variables. In general, two approaches are possible here, depending on certain circumstances:

1. Construction of a *Bn* only with participation of experts, based on their belief, while an expert may design graph of the network and parameters for its processing; or else a database will be used to determine these parameters after having designed the graph.
2. Construction of the network as a whole – as a reflection of knowledge contained in the database only, in practice without direct participation of an expert. There are many algorithms allowing to perform this operation [1–3], from extremely simplified to complex ones.

In the *Bn* networks, individual variables are bound up with conditional distribution tables determined by an expert or as a result of applying the learning procedure. This allows to update individual distributions after having observed a specific value of any variable, that is to establish distributions a posteriori. Each *Bn* network node has an attributed table of variable state probability on condition of the state of its predecessors (so-called „parents") in the network. Concluding in the Bayesian networks is executed by propagation of successive conditional distributions at observed value of a specific variable, according to the formula [2]:

$$P(x_1,...,x_n) = \prod_{i-1}^{n} P(x_i \mid Predecessor(x_i)) \qquad (1)$$

The *GeNIe* computational system has been used in the paper, developed at Pittsburgh University for research purposes. The obtained model was evaluated using the *Netica* application from *NorSys*.

Discretization

Most of packages used to generate the *Bn* network require discrete variables. Certainly, this involves the need to convert continuous variables into nominal variables.

One may assume – while analyzing regression examination results – that the network of connections determining sorption properties of soils includes the following variables: colloidal clay content, organic carbon content (*OC*), maximum hygroscopicity (MH), methylene blue sorption (*SBM*), reaction (pH), and of course cation exchange capacity (*CEC*) and base saturation (*BS*). In view of model construction requirements, these variables had to be discretized by rating their individual values in proper classes. Discretization resulted in the conversion of raw variables into modified variables: clay, carbon, higr, mblue, react, cec and bs. Table 2 shows the digitizing principles. Note that some variables are classified (named) with certain exaggeration (for example those related to soil reaction), but this results solely from the need to ensure the model communicativeness.

Table 2. The criteria of discretization the Bayesian model variables

Orig. var.	Units	Discr. var.	Values and limits of variables
Clay content	%	clay	ABSENT(0), LOW(1–5), MED(6–15), HIGH(> 15)
Carbon cont.	%	carbon	ABSENT(0), LOW(0.01–0.5), MED(0.51–1.0), HIGH(> 1.0)
MH	%	higr	VLOW(< 2), LOW(2.1–4), HIGH(4.1–8),VHIGH(> 8)
SBM	cmol/kg	mblue	VLOW(< 5), LOW(5.1–10), HIGH(10.1–20), VHIGH(> 20)
CEC	cmol/kg	cec	VLOW(< 5), LOW(5.1–10), HIGH(10.1–20), VHIGH(> 20)
Reaction	pH	react	VACID(< 5), ACID(5–6), NEUTR(6–7), ALK(7–8), VALK(> 8)
BS	%	bs	LOW(< 50), MED(50–75), HIGH(> 75)

The model graph

Construction of a Bayesian model requires many tests to be performed. Employment of an algorithm producing network graph in an automated procedure has the value of objectivism, but it also carries with it the risk of occurrence of incidental connections between variables. The graph of the model shown in Figure 1 is one of possible versions of the concluding system. From this point of view, the 'mblue' and 'react' variables function as diagnostic variables, connected with the variables we are interested in, which are difficult to observe.

Figure 2 illustrates the distribution of probabilities for variables at determined (by assumption observed) values of the 'react = NEUTR', 'mblue = VHIGH' variables. This configuration results in the $p(HIGH) = 0.99$ value being indicated for the 'bs' variable, and $p(HIGH) = 0.59$ and $p(VHIGH) = 0.4$, respectively, for the 'cec' variable. In Figure 3, the configuration has changed, since it has been observed that 'react = VACID'. In this situation, the following estimate has been obtained for the 'bs': $p(LOW) = 0.12$, $p(MED) = 0.7$, and $p(HIGH) = 0.18$. Also, the estimates of the 'cec' variable probability value have changed: $p(HIGH) = 0.27$, $p(VHIGH) = 0.7$, respectively, with further consequences.

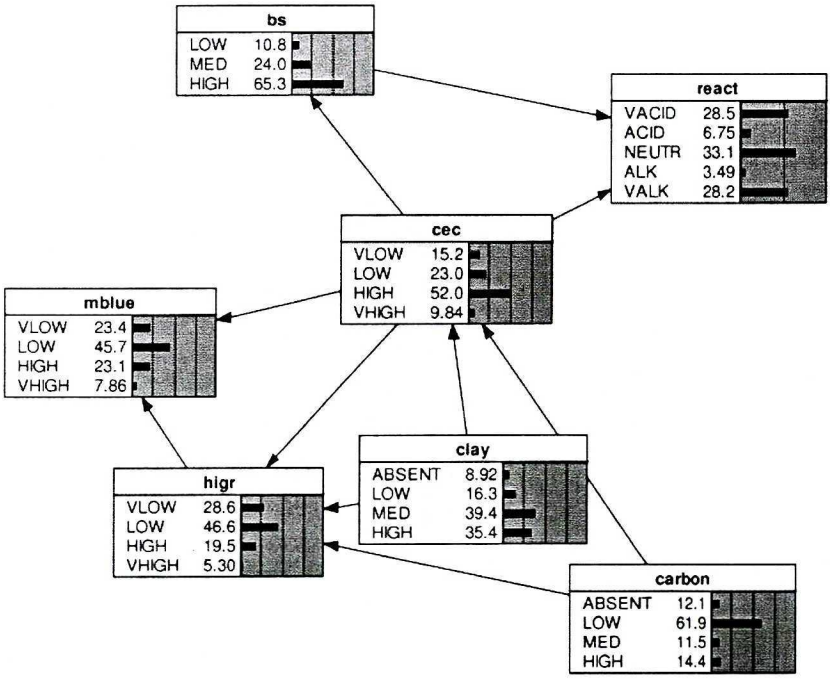

Fig. 1. The structure of a *Bn* network for analyzing sorption properties of soils

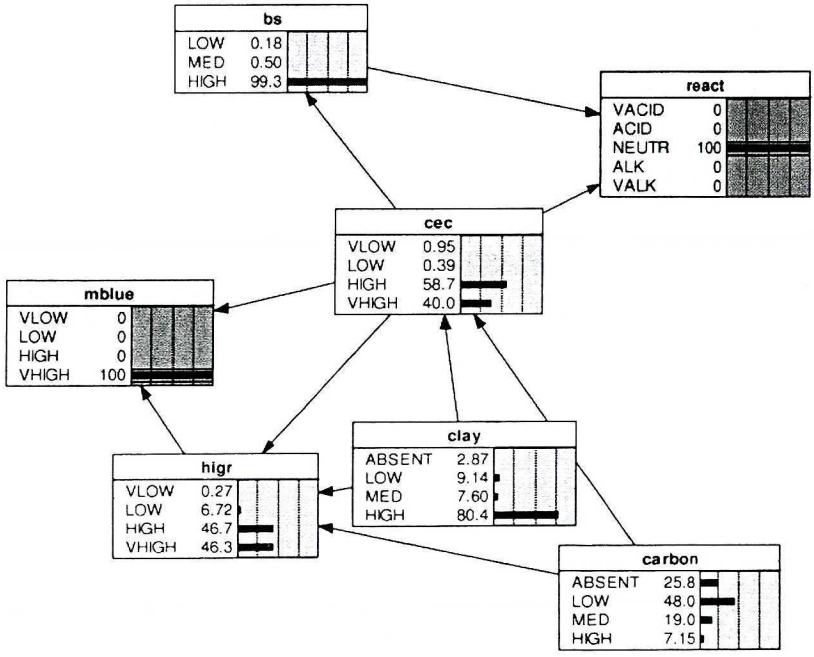

Fig. 2. Probability distributions for variables of the *Bn* model variables at determined 'react = NEUTR', 'mblue = VHIGH' variables

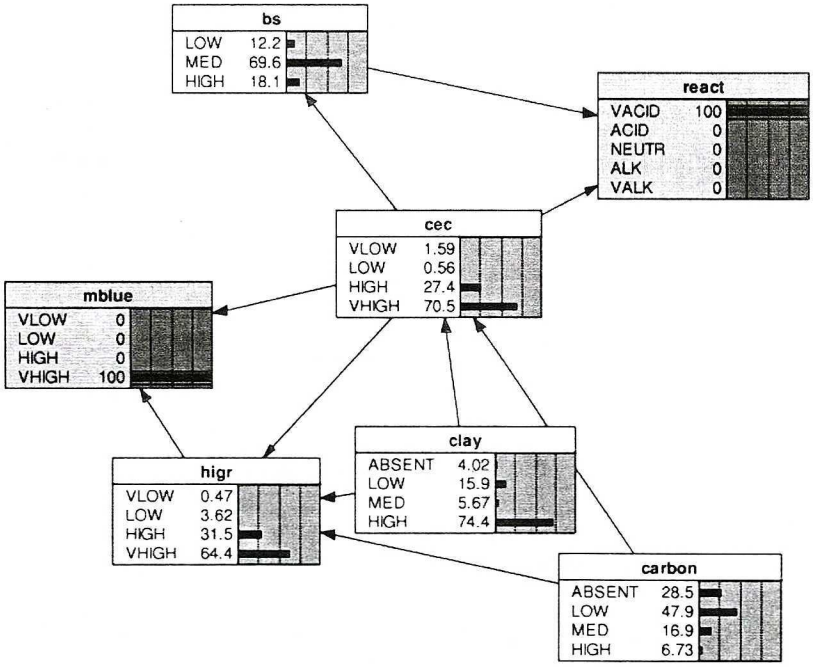

Fig. 3. Probability distributions for variables of the *Bn* model variables at determined 'react = VACID', 'mblue = VHIGH' variables

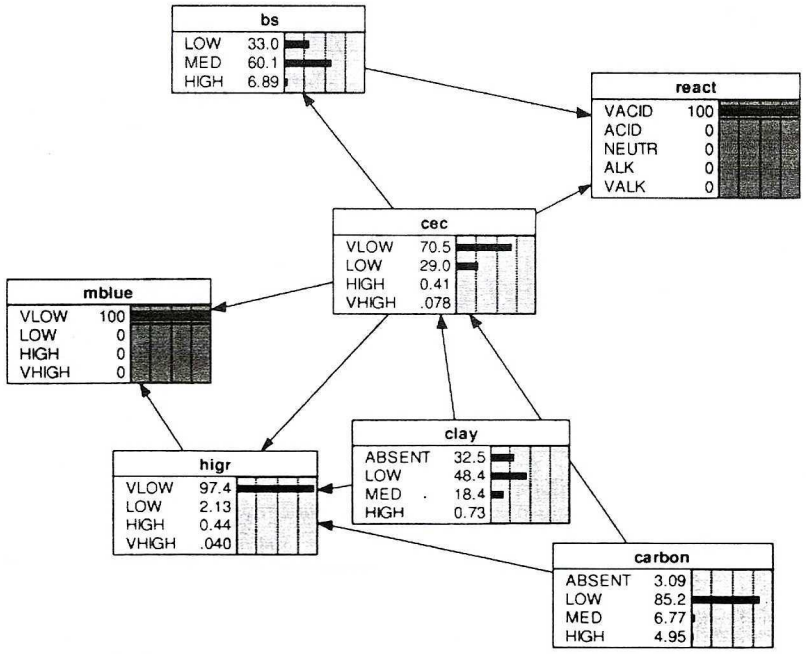

Fig. 4. Probability distributions for variables of the *Bn* model variables at determined 'react = VACID', 'mblue = VLOW' variables

Figure 4 illustrates the situation in which 'react = NEUTR', and the observed value 'mblue = VLOW'. As a consequence of these changes, we obtain the following configuration for the 'bs': $p(LOW) = 0.13$, $p(MED) = 0.36$, and $p(HIGH) = 0.51$. Significant changes have occurred in the distribution of probabilities for the 'cec' variable, which assume the following values: $p(VLOW) = 0.67$ and $p(LOW) = 0.32$.

There is a problem of qualitative valuation of the obtained models. The following indicators are used conventionally in models valuation: error frequency coefficients, error matrixes, and indicators taking into consideration probability distribution for the states of the variable we are interested in. The Netica Package offers the following three indicators which analyze the distribution of probability indications: L_I (logarithmic loss), computed using formula:

$$L_I = \frac{1}{n} \cdot \sum_{i=1}^{n} [-\log(P_c)] \tag{2}$$

L_B (Brier, quadratic loss), computed using formula:

$$L_B = \frac{1}{n} \sum_{i=1}^{n} [1 - 2 \cdot P_c + \sum_{j=1}^{m} (P_j)^2] \tag{3}$$

and C_S (spherical payoff), computed using formula:

$$C_S = \frac{1}{n} \sum_{i=1}^{n} \frac{P_c}{\sqrt{\sum_{j=1}^{m} (P_j)^2}} \tag{4}$$

Lower limit of indicator $L_I = 0$; this value shows the best efficiency. The L_B indicator value ranges within (0; 2), where 0 indicates the best efficiency. The C_S indicator value ranges within (0; 1), where $C_S = 1$ shows best efficiency.

Tables 3, 4 and 5 contain indicators for correct estimates made using the *Bn*. They show serious scattering of results, and difficulties with building of a reliable model, since the indicators have been obtained for the training set, assuming complete knowledge of the explanatory variables distribution (apart from predicted variables).

Table 3. The error matrix for Bayesian model of sorption properties: distribution of the CEC classes indications; symbols: 'Pr' – predicted class, 'Obs' – observed class

	Pr(VLOW)	Pr(LOW)	Pr(HIGH)	Pr(VHIGH)
Obs(VLOW)	62	4	0	0
Obs(LOW)	4	65	13	0
Obs(HIGH)	0	28	169	5
Obs(VHIGH)	0	1	9	21

Table 4. The error matrix for Bayesian model of sorption properties: distribution of the BS classes indications; symbols: 'Pr' – predicted class, 'Obs' – observed class

	Pr(LOW)	Pr(MED)	Pr(HIGH)
Obs(LOW)	15	28	0
Obs(MED)	6	81	6
Obs(HIGH)	0	27	218

Table 5. Indicators of the Bayesian network indications quality for the CEC and BS valuation

Indicator	CEC	BS
Error rate [%]	16.8	17.6
Logarithmic loss	0.32	0.36
Quadratic loss	0.21	0.23
Spherical loss	0.88	0.86

The Bayesian network indicates probability distribution for states, related to certain configuration of variables. Besides certain cases of close function-type relations it does not allow to determine what exactly will be the value of the variable. However, a diagnostic model of this type may be a useful tool helping to make correct technical decisions, for example during reclaiming works, in particular considering suitably extensive database used to build it. Making of decisions concerning neutralizer doses, aimed to reach certain saturation state may be based on a similar model.

Quite a serious problem in making adaptation models is defining their generalization properties. Adaptation models often contain many free parameters, being justed in the training procedure. With a relatively small amount of data, there is a risk of adjusting a model to a concrete set, which then manifests in a disproportional growth of the error in the classification of the data outside the training set. A commonly applied way of the

Table 6. Results of the Bayesian network cross-validation (10-Folds method) for the CEC and BS valuation (ER – error rate in %)

			CEC			BS		
Sample set	ER	L_I	L_R	C_S	ER	L_I	L_R	C_S
Training 1	15.7	0.31	0.20	0.88	16.9	0.33	0.21	0.87
Valid 1	15.8	∞	0.26	0.86	23.7	0.74	0.45	0.74
Training 2	15.7	0.32	0.21	0.88	15.5	0.32	0.21	0.88
Valid 2	23.7	∞	0.31	0.83	44.7	∞	0.58	0.66
Training 3	14.3	0.29	0.18	0.89	18.1	0.38	0.24	0.86
Valid 3	39.5	∞	0.48	0.72	13.6	0.23	0.15	0.90
Training 4	15.5	0.31	0.19	0.89	16.9	0.35	0.23	0.87
Valid 4	28.9	∞	0.40	0.78	23.7	0.48	0.31	0.81
Training 5	16.9	0.31	0.21	0.88	16.9	0.34	0.21	0.87
Valid 5	15.8	∞	0.30	0.83	26.3	0.69	0.43	0.74
Training 6	18.1	0.33	0.21	0.87	14.3	0.31	0.20	0.88
Valid 6	26.3	∞	0.39	0.79	52.6	∞	0.66	0.60
Training 7	17.2	0.32	0.21	0.88	19.5	0.40	0.26	0.85
Valid 7	15.8	∞	0.23	0.87	0.0	0.02	0.02	0.99
Training 8	16.0	0.33	0.21	0.88	19.5	0.41	0.26	0.84
Valid 8	21.1	0.34	0.22	0.88	0.0	0.00	0.00	1.00
Training 9	16.6	0.33	0.21	0.88	18.9	0.39	0.25	0.85
Valid 9	18.4	0.34	0.22	0.87	5.3	0.11	0.07	0.96
Training 10	15.5	0.31	0.20	0.88	17.8	0.35	0.23	0.87
Valid 10	23.7	∞	0.45	0.77	23.7	∞	0.43	0.76
Tr. av.	16.2	0.32	0.30	0.88	17.4	0.36	0.23	0.86
Val. av.	23.7	–	0.33	0.74	21.4	0.32	0.31	0.81

assessment of this risk is cross validation [1]. The results of this procedure for the discussed model are presented in Table 6. It is natural that the validation usually gives worse results of the correctness of classification than it takes place with the application of data used in the training. This is also true in this case, although the results of the validation are relatively little significant, they are, generally, different than the assessment of the model, based on the training data. Looking through the indicators of the classification correctness, the model can be regarded as reflecting real relationships between soil features.

SUMMARY AND CONCLUSIONS

Problems on land use, soil conservation and environmental management require increasingly accurate information on soil properties and their geographical location. Qualitative models, particulary in GIS/LIS application as a soils inference system, are developed for estimation of attributes related to management, planning practices, environmental impact assessment, environmental risk assessment and other areas.

All the soil properties are linked with one another in numerous relationships. The activities, undertaken, e.g. to improve a certain soil property, can lead to the modification of many features. A comprehensive model of the functioning of the soil system would probably be useful, but nowadays we have to do with extremely simplified models, neglecting some conditions. The Bayesian belief network is helpful in the construction of such models. They were constructed with the objective to assess the impact on the environment of different soil management scenarios. The attempt presented in the paper shows the advantages and disadvantages of such modeling: despite significant simplifications and obvious removals of some aspects, the obtained result can be regarded satisfactory. A model of this type enables to bind comprehensively the observed parameters of the ecosystem and analyzing its behavior in case of the intervention in its structure. Indirectly, the model of this type shows the need of analyzing, if possible, many soil properties, because rudimentary data do not allow the estimation of the directions of potential threats connected with the intended activities, e.g. reclamation measures. An obvious disadvantage of this is the need of basing on a rather large set of empirical data, because, otherwise, there is a high risk to obtain a model insufficiently reflecting general relationships between soil properties.

REFERENCES

[1] Bishop C.M.: *Pattern Recognition and Machine Learning*, Springer, 2006.

[2] Cichosz P.: *Learning systems*, Wydawnictwa Naukowo-Techniczne, Warszawa 2000.

[3] Druzdzel M.J.: *A development environment for graphical decision-analytic models*, [in:] Proceedings of the 1999 Annual Symposium of the American Medical Informatics Association (AMIA-1999), Washington D.C., November 1999.

[4] Fooladmand H.R.: *Estimating cation exchange capacity using soil textural data and soil organic matter content: A case study for the south of Iran*, Archives of Agronomy and Soil Science, **54**(4), 381–386 (2008).

[5] Gruszczyński S., T. Eckes, T. Gołda, M. Trafas, P. Wojtanowicz, K. Urbański: *The principles for determination of industrial soils quality in the reclaimed areas*, Factual report for the final report on the implementation of the research project no. 4 T 12 E 041 29, Technical report, AGH-University of Science and Technology, Krakow 2008.

[6] Horn A.L., R.-A. Düring, S. Gäth: *Comparison of the prediction efficiency of two pedotransfer functions for soil cation-exchange capacity*, Journal of Plant Nutrition and Soil Science, 168, 372–374 (2005).

[7] McBratney A.B., B. Minasny, S.R. Cattle, R.W. Vervoort: *From pedotransfer functions to soil inference systems*, Geoderma, 109(1-2), 41–73 (2002).
[8] Myślińska E.: *Laboratory testing of soils (In Polish)*, Wydawnictwo Naukowe PWN, Warszawa 1998.

Received: January 19, 2009; accepted: November 24, 2009.

MODEL BAYESOWSKI WSPÓŁZALEŻNOŚCI MIĘDZY CECHAMI SORPCYJNYMI GLEB

Praca przedstawia jakościowy, bayesowski model niektórych współzależności między cechami sorpcyjnymi mineralnych gleb opróbowanych w południowej Polsce. Właściwości sorpcyjne są ważnymi cechami, współdecydującymi o poziomie nawożenia, zasobności w składniki pokarmowe oraz zdolności do ochrony wód gruntowych przed zanieczyszczeniem. Pojemność wymienna kationów jest powszechnie używanym wskaźnikiem stanu gleb i ich odporności na różnorodne wpływy. Stopień wysycenia kationami zasadowymi jest ważnym czynnikiem kształtowania ryzyka środowiskowego związanego z procesami zakwaszania gleb. Badane gleby reprezentują różne klasy bonitacyjne i typologię. Do oceny współzależności zastosowano model bayesowski.