ГРУНТОЗНАВСТВО

UDC 631.4:551.4:004.942

COMPARATIVE ESTIMATION OF THE ACCURACY OF SIMULATION MODELING OF SOIL COVER AND FORECAST OF CARTOGRAMS OF AGRO-INDUSTRIAL GROUPS OF SOILS

V. R. CHERLINKA, Y. M. DMYTRUK, V. S. ZAHAROVSKYY

Yuriy Fedkovych Chernivtsi National University,
Institute of Biology, Chemistry and Bioresources, Department of Soil Science,
str. Lesia Ukrainka, 25, Chernivtsi, Ukraine, 58012,
e-mail: v.cherlinka@chnu.edu.ua

The main goal of the mathematical experiment was to compare the accuracy of the construction of predicative maps, depending on the type of input data, in particular the soil map and the complete or abbreviated (without definitions by composition of grain size) variants of the cartograms of agro-industrial soil groups. The tasks were solved: by building a digital relief model (DEM); digitization of cartographic materials; generation of a set of maps of morphometric and other derived characteristics; the analysis of the connections and the role of the mentioned parameters in the variability of the soil cover; creation of predicative map-versions of soils and cartograms of agro-industrial groups of soils. Object of research: a fragment of the territory of the Chernivtsi region with complex geomorphological conditions. Main methods used: correlation analysis; the principal component method; predicative algorithms Decision Trees, Random Forests and K-Nearest Neighbors. On the basis of the correlation analysis, the tightness of the connection and the role of predictors (independent variables) in the variability of the soil cover were assessed, and the analysis of the main components involved the selection of 9 basic ones: absolute altitude; topographic moisture index; the amount of solar radiation per unit area; steepness of slopes; longitudinal and maximum curvature of the topographic surface; accumulation, length and distance to water flow. The quality of predicted cartographic materials was estimated using the Cohen's kappa coefficient. Differences in the qualitative characteristics of the obtained simulated map-versions are established and it is shown that the morphometric parameters of the relief and its derivatives are a reliable basis for predicative modeling. An extended assessment of the quality of the map-models is made, depending on the type of input data and it is shown that the most accurate predictor cartogram of complete agro-industrial soil groups is used with the set of predictors used. Differences in the quality of predictive soil maps were established by using 3 types of predicative algorithms and it was shown that classification models, in particular, Decision Trees and Random Forests, which allowed obtaining up to 93% of the coincidence of real and model data, were the most suitable for such tasks. The possibilities of constructing forecast maps of soils using a standard set of materials that can be accessed by soil scientists in modern Ukrainian realities are shown: soil and topographic maps in conjunction with free full-featured software - GRASS and Quantum geoinformation systems, Easy Trace vectorizer and R-Statistic, language and environment for statistical computing.

Key words: soil map, cartogram of agro-industrial groups of soils, training data set, simulation, morphometric parameters, DEM, predicative algorithms.

Introduction. Consideration of the situation regarding the relevance of large-scale soil mapping materials in Ukraine (Polchina et al., 2004; Achasov et al. 2015; Cherlinka, 2017) shows that there will be no quick solution to existing problems in the near future. Nearly a quarter of the country's territory (in particular, the mountainous systems of the Carpathians and the Crimea, plain-covered areas of forest vegetation, areas of the number of settlements, etc.) have never been covered by continuous soil-based surveying. In modern economic conditions, it is not worthwhile to expect to allocate funds for actualization of existing materials and to study white spots. Similar problems exist not only in Ukraine or in a number of other developing countries, but also in countries such as Australia (Bui and Moran, 2003). Therefore, the logical step is to fill the gaps in cartographic information with predicted data. Indeed, over the past decades, the number of such studies devoted precisely to the simulation of the spatial location of taxonomic soil units has considerably increased (Bui and Moran, 2003; McBratney et al., 2003; Scull et al., 2003; Walter et al., 2006; MacMillan, 2008; Browning and

Duniway, 2011; Caten et al., 2013; Brungard et al., 2015; Malone et al., 2016; Heung et al., 2016, 2017). In this case, a wide range of mathematical methods is used: from multifactorial regression analysis, kriging, neural networks to different types of classification trees (Florinsky, 2016). The general idea underlying the application of such methods is to use the reference points of the landscapes and soil taxa associated with them (Lagacherie et al., 2001). The main source of predictors in this direction of simulation is the digital model of relief (DEM), the analysis of which allows to distinguish a number of geomorphological and related parameters. Since model variables (types of soils) do not refer to the numerical, but to the categorical type of data, and the indicators derived from the DEM are usually numeric, the use of advanced mathematical methods only allows us to establish the non-obvious, on the first view, dependence between all these parameters (Giasson et al., 2008; Kempen et al., 2009; Debella-Gilo and Etzelmüller, 2009; Hengl, 2009; Malone et al., 2016; Cherlinka, 2017).

The general simulation procedure involves the allocation of a certain portion of the data from the population under study for machine learning and the subsequent simulation is already based on this data. Feng and Michie (1994) characterizes this process through such stages: generation of training data set; training algorithm; creation of classification rules; testing on a complete set of data. In our case, the main task of constructing a training sample for the subsequent construction of a forecast ground map (or any other map with categorical data) is the choice of such points, the spatial location of which would most fully cover the variation of taxonomic units of soil and their predictors. Modeling a model on this sample allows you to establish relationships and relationships between these all parameters and then transfer the resulting results to the entire study area. It also enables extrapolation of results beyond the existing ground maps, since a set of predictors is obtained on the basis of DEM, which covers the entire territory.

By constructing a set of training data clearly distinguish 2 approaches (Brungard et al., 2015; Heung et al., 2016; Heung et al., 2017): data on soil cuts laid out in field conditions and a sample of clearly defined contours of ground maps. The first approach has good prospects, but requires a large established database of verified soil cuts, with which there are currently problems. Ukraine is currently at the beginning of the path to establishing such a data bank throughout the country with complete and comprehensive information on the soil profiles (Postanova Prezydiji Nacionaljnoji akademiji .., 2017). Therefore, we use a different approach, as more relevant in the immediate time perspective and easier to implement in the current modeling environment.

Note that a number of predictive algorithms, especially when using large sets of training sample, give a high degree of coincidence of predictive and real classification units, which does not always correspond to such accuracy in the entire volume of data. When using similar inputs data, different results can be obtained. In this case, we mean, at the same time, soil maps and cartograms of agroindustrial groups of soils that are more often used in production conditions.

Accordingly, the task of this study was to cover the predicative modeling variants using as inputs a soil maps and cartograms of their agroindustrial groups and highlight those methods that give the best results as a result of forecasting. This is important given that predicative maps are interesting not only as an object of scientific study, but as an important tool for obtaining information on soil cover, in locations where studies have not yet been conducted. Therefore, the higher the degree of coincidence of the forecast data with the real map, the more grounded will be the conclusions about the information, localized in "white spots" of large scale maps. The latter is relevant and important in light of the optimization of normative monetary valuation of land and other scientific and practical tasks of the present.

Accordingly, the purpose of our work was to study the input data options and their impact on the qualitative characteristics of simulative soil maps by conducting a mathematical experiment using a typical set of materials that can be potentially available to ordinary soil scientist or scientist in contemporary Ukrainian realities. We refer to them large-scale soil and topographical maps, cartograms of agro-industrial groups of soils (M1:10000) and free software – geographic information systems GRASS (GRASS Development Team, 2017) and Quantum (QGIS Development Team, 2015), vectorizer Easy Trace (EasyTrace group, 2015); language and environment for statistical computing R-statistic (R Development Core Team, 2017).

Materials and methods. In accordance with the stated goal, we identified the following tasks: a) digitization and attribution vector information of cartographic materials; b) creation DEM with a resolution equal to 20 m; c) analysis of digital elevation models and extraction from them in the GIS GASS of set of maps of morphometric and other derivative characteristics; d) generation of training samples according to the described methodological approaches; e) creation in R-statistik of simulation models using 3 types of predicative

algorithms both for areas with available soil information and for those where it is not represented; g) analysis of the obtained results and conclusions about the optimal source material for predictive modeling.

As an object, a fragment of the territory of Ukraine (Fig. 1a) within the boundaries of the Chernivtsi region was selected (Fig. 1b), confined to the Prut-Dniester interfluve (Northern Bukovina) with contrasting geomorphological conditions and administratively owned by the Kitsman district (Fig. 1c). This area has different administrative subordination and economic use, and when it was selected, typical problems that often arise in the work of this nature were solved (Cherlinka and Dmytruk, 2014; Cherlinka, 2015; Cherlinka, 2017). The coordinate system of the project was selected SC 1963 (zone X2), 6 scanned sheets of topographic maps M 1:10000, in particular M-35-124-Vg-{1,2,3,4}, M-35-124-Vb-3 and M-35-124-V-v-2 (Fig. 1d) were georectified using by created vector mathematical basis, and the georectified of cartograms of agro-industrial groups of soils was carried out to the characteristic points of the locality and the administrative boundaries of existing rural councils: Nepolokivtsi (Nepolokivtsi) – "A", Beregomet (Beregomet and Revakhivtsi) – "B", and Dubivtsi (Dubivtsi) – "C" of Kitsman district of Chernivtsi region. Informative soil materials were based on a series of archival soil maps of the collective farm "Soviet Ukraine" (soil survey of 1957 year and correction in 1974). After the consolidation of the nomenclature list of soils into a single system and harmonization of contours and types of soils, the information was digitized and preliminary data on the percentage of coverage of the territory by soil surveys were obtained: from 4424,32 hectares of the total area for 1086,84 hectares, or 24,86%, data fully are absent (Fig. 2, Table 1). "White" spots are confined to the territory of settlements and forest areas that are within the boundaries of the mentioned village councils.

A step of 20 m was chosen as the base resolution of the DMP, which, with a relatively high accuracy of the reproduction of the topography, also provides a satisfactory coincidence of the areas of vectorized and rasterized soils. To process the data, the tools of the free software were used: georegistration of the cartographic material - GIS Quantum (QGIS Development Team, 2015), digitization - Easy Trace (EasyTrace group, 2015), preparation of morphometric parameters and generation of DEM – GRASS GIS (GRASS Development Team, 2017) and the simulation of soil maps - language and environment for statistical computing R-statistic (R Development Core Team, 2017). Based on the digital model of relief with a resolution of 20 m, a number of morphometric characteristics of the relief were provided as predictors: slope and aspect, curvature of the surface (longitudinal and maximum), solar radiation and relief forms. Additional maps of hydrological indicators were also generated: the topographic wetness index, accumulation, direction and length of water streams and the distance to them.

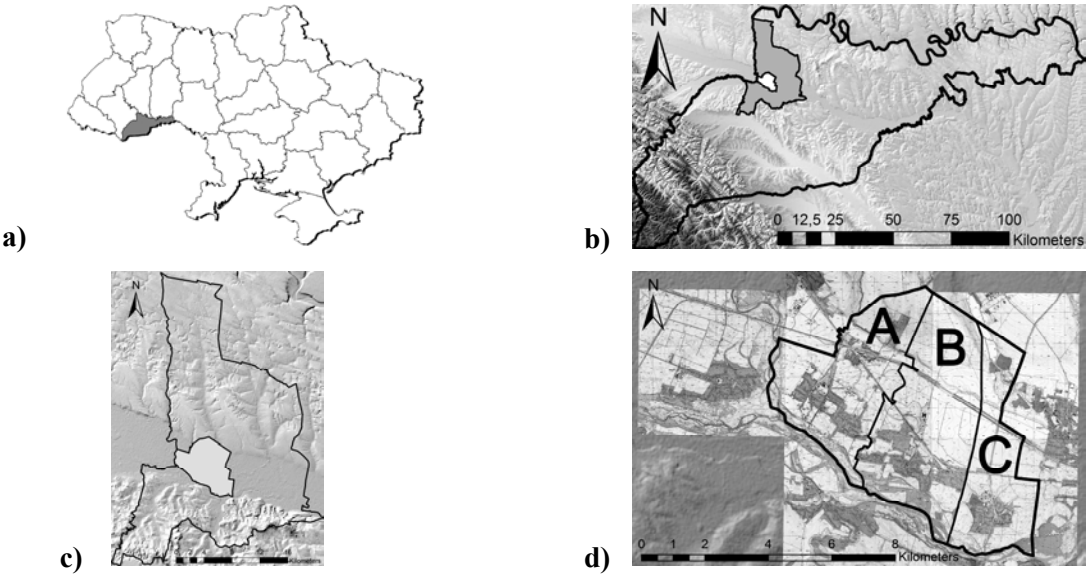

Fig. 1. Geographical location of the research area within Ukraine (a), Chernivtsi region (b), Kitsman district (c), and the scheme of the test ground (d) * for the background data used SRTM – NASA's Shuttle Radar Topography Mission

To create simulation models of soil cover, we wrote a script in the language R-statistic (R Development Core Team, 2017), which includes a number of adaptations for solving the tasks and implements 14 basic types of predicative algorithms, of which 3 were used in this study, in particular: 1. Decīšīon Trees - DT (Venables and Riley, 2002). 2. Random Forests - RF (Breiman, 2001; Cutler et al., 2012). 3. K-Nearest Neighbor - KNN Liu, 2011).

The quality of the obtained models was evaluated on the basis of the Cohens kappa index κ (Landis and Koch, 1977; Li and Zhang, 2007; Grinand et al., 2008; Kuhn, 2008; Hengl, 2009; Malone et al., 2016), which in this case shows the degree of correspondence between the original and the simulated data.

Results and discussion. First of all, it should be noted that the use of digitized soil materials that are available in the system of the State Geocadastre of Ukraine is impossible given the very large number of errors contained therein. The most important of these are faults in geo-rectification, which lead to the fact that the lowlands or thalweg are located on the peaks or slopes of the hills. The second type of error is attribution errors when the soil name on the map does not match its vector counterpart. Therefore, for the qualitative modeling and correct analysis of the results obtained, it is necessary to re-vectorize.

In this experiment, we did not use the methodology for creating a training sample (Dobos and Hengl, 2009; Hengl, 2009), but a randomized, weighted, which shows much better results in our studies. Unlike the median-weighted, by randomized-weighted approach, has no any problem with reducing the actual number of points in the training data set and allows you to get precisely those proportions between the training cells and the total sample size, which is conditioned by the conditions of the planned experiment, in particular coverage of 35% of the area of the surveyed soils.

Three main cartographic sources were also prepared: 1) an original map of soils; 2) cartogram of agro-industrial groups of soils; 3) shortened cartogram of agro-industrial groups of soils. The last map includes agro-industrial groups of soils without their division in granulometric composition. Thus, variants with a different number of prediction elements were received: 29, 22 and 16 for each of the maps respectively (Table 1). Such a set of experiment variants, in combination with 3 predicate algorithms, allowed to obtain 9 sets soil cover simulations, the analysis of which revealed quite interesting patterns.

The resulting array of simulations of soil cover is interesting in terms of its correspondence to original maps, and, accordingly, predictive "force" in areas with no information. Since the algorithms analyze the entire spectrum of predicate parameters, producing classification rules, with a high degree of coincidence of model and real data, one can talk about a certain level of statistical accuracy of data in the areas of "white spots". The results we get in this regard can be called quite encouraging given the range of values obtained κ (Table 2). If we rank the models by criteria of increasing the quality of the prediction by the κ main data set, then the KNN algorithm showed the worst results among others. They follow in the order of increasing RF and DT. The last two algorithms belong to the classification methods, and their high results testify to the greatest suitability of such approaches in simulations of soil cover.

A step of 20 m was chosen as the base resolution of the DMP, which, with a relatively high accuracy of the reproduction of the topography, also provides a satisfactory coincidence of the areas of vectorized and rasterized soils. To process the data, the tools of the free software were used: georegistration of the cartographic material - GIS Quantum (QGIS Development Team, 2015), digitization - Easy Trace (EasyTrace group, 2015), preparation of morphometric parameters and generation of DEM – GRASS GIS (GRASS Development Team, 2017) and the simulation of soil maps - language and environment for statistical computing R-statistic (R Development Core Team, 2017). Based on the digital model of relief with a resolution of 20 m, a number of morphometric characteristics of the relief were provided as predictors: slope and aspect, curvature of the surface (longitudinal and maximum), solar radiation and relief forms. Additional maps of hydrological indicators were also generated: the topographic wetness index, accumulation, direction and length of water streams and the distance to them.

In general, the prediction for full cartograms of agro-industrial groups of soils work better than for abbreviated cartograms, and especially for soil maps. Since the soil map is the most diverse on the soil taxons, the dropping of the kappa value becomes clear. The KNN algorithm stands out slightly, which showed somewhat different results: it has the best prediction of 78.51% for the soil map, and the worst 68.82% for the full agrogroups of soils. In any case, Decision Trees and Random Forest are the most powerful algorithms that can simulate the distribution of soil deviations and agro-industrial groups of soils with κ equal to 82.66-93.09% in case with 35% saturation of the training data set with the source data (Fig. 3). In general, for these algorithms, it is characteristic of almost 100% coincidence of the calculated data from the training sample with their real values.

Table 1.

Variants of the encoding of soil taxons of the territory of simulation

Id soil	Id agrosoil	Id agrosoil simple	Standart agrosoil code	Code of soil	Name of soil
0	0	0	nodata	nodata	
1	1	1	40 d	1 l	Temno-siri lisovi
2	2	2	51 d	10 l	Chornozemy opidzoleni serednozmyti z pliamamy 30-50% sylnozmytykh
3	2	2	51 d	11 l	Chornozemy opidzoleni sylnozmyti
4	3	3	133 e	12 ad	Chornozemno-luchni mocharysti
5	4	3	133 l	13 a	Luchni hlyboki vyluhovani
6	5	3	133 d	14 a	Luchni pyluvato-serednosuhlynkovi
7	6	3	133 e	15 a	Luchni pyluvato-vazhkosuhlynkovi
8	6	3	133 e	16 a	Luchni hleiovi
9	7	4	142 e	17 a	Luchno-bolotni osusheni
10	7	5	141 e	18 a	Bolotni pyluvato-vazhkosuhlynkovi na davnomu aliuviiu
11	7	5	141 e	19 al	Bolotni pyluvato-vazhkosuhlynkovi na suchasnomu aliuviiu
12	8	6	49 d	2 l	Temno-siri lisovi slabozmyti
13	7	5	141 e	20 d	Bolotni pyluvato-vazhkosuhlynkovi na suchasnomu deliuviiu
14	9	6	139 l	21 d	Bolotni mocharysti
15	10	7	176 d	22 a	Dernovi hlyboki karbonatni
16	11	8	175 v	23 al	Dernovi karbonatni supishchani
17	12	7	176 g	24 al	Dernovi karbonatni pishchano-lehkosuhlynkovi
18	13	9	181 d	25 al	Dernovi karbonatni hleiovi namyti
19	14	10	215 e	26 l	Slabozadernovani skhyly yariv ta krutykh ustupiv pyluvato-vazhkosuhlynkovi na lesovydnykh suhlynkakh
20	14	10	215 e	27 a	Slabozadernovani skhyly yariv ta krutykh ustupiv pyluvato-vazhkosuhlynkovi na davnomu aliuviiu
21	14	10	215 e	28 l	Vykhody porid
22	15	11	219 ak	29 al km	Ruslovi vidklady
23	16	12	41 g	3 l	Chornozemy opidzoleni pyluvato-lehkosuhlynkovi
24	17	12	41 d	4 l	Chornozemy opidzoleni pyluvato-serednosuhlynkovi
25	18	13	209 d	5 dl	Chornozemy opidzoleni hleiuvati namyti
26	19	6	49 g	6 l	Chornozemy opidzoleni slabozmyti pyluvato-lehkosuhlynkovi
27	20	6	49 d	7 l	Chornozemy opidzoleni slabozmyti pyluvato-serednosuhlynkovi
28	21	6	49 e	8 l	Chornozemy opidzoleni slabozmyti z pliamamy 10-30% serednozmytykh
29	22	16	50 d	9 l	Chornozemy opidzoleni serednozmyti

Fig. 2. A digital model of the relief of the research area draped with the original soil map (the soil numbers correspond to their serial numbers in the nomenclature list in Table 1)

Table 2.

The distribution of the index κ depending on the input data and the type of simulation model

Modeling options	Type of simulation model					
	DT		RF		KNN	
	κ_t	κ	κ_t	κ	κ_t	κ
Soil map	86,91	83,74	86,90	82,66	82,31	78,51
Cartograms of full agrogroups	99,96	93,09	100,00	92,90	83,46	68,82
Cartograms of reduced agrogroups	99,94	91,54	100,00	90,71	80,43	70,77

*** κ_t - kappa of training data set, κ - kappa of main data set**

Concerning the lower quality of the prediction of soils in comparison with agrogroups of soils, it can be assumed that the used set of predictors of the model does not fully describe the definitions of the distribution of soils on the elements of the relief. Therefore, the study of this issue will be the subject of our next research.

An analysis of the level of comparability of our results on the quality of simulation with similar studies shows that the kappa of our models exceeds the averaged values from literary sources. So, in the work (Hengl, 2009) 51-67% is considered a good indicator. In work Grinand et al. (2008) κ=67-87% for the study sample and is about 30% for the main data set. For small-scale soil maps Giason et al. (2008) obtained κ 37-54%, and Malone et al. (2016) its value ranges from 35-40%. According to the ranges given by Landis and Koch (1977), the results we have obtained and the results described above have, in the worst cases, a significant convergence (κ=0.61-0.80), and in the best cases, almost complete convergence (κ=0,81-0,99). Accordingly, this allows us to assess the quality of simulation card versions as good and not below the level of similar literary data. In addition, we believe that there is still some potential for increasing the overall κ, in particular by more accurately selecting the model's predictors and extending their number by incorporating Earth remote sensing data, anthropogenic deposits maps, and more. A significant beneficial effect of this kind of modeling is the ability to fill gaps on existing cartographic materials with data from predicative map-versions and, thus, obtaining composite soil maps. This certainly does not exclude the need for a large-scale soil survey of such areas, but in the absence of the possibility of its carrying out, it allows to obtain at least some scientific data with a certain level of statistical reliability.

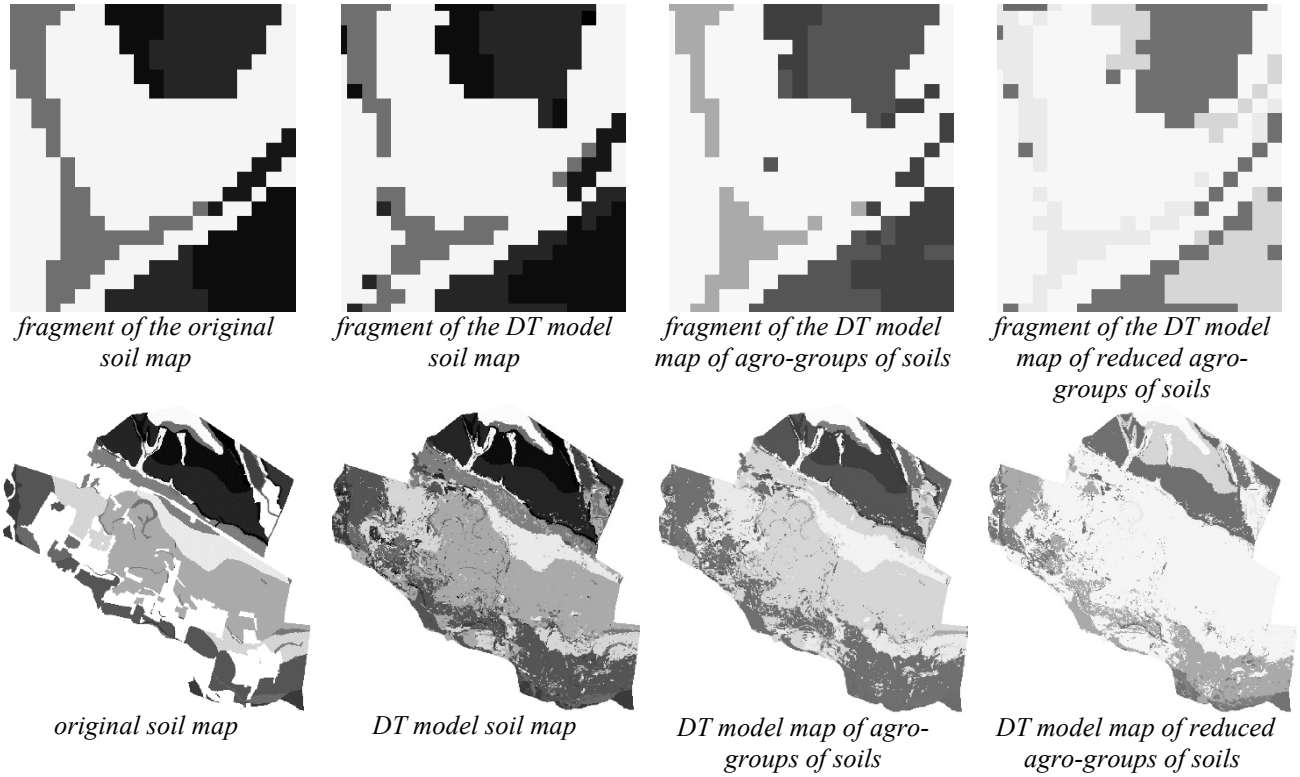

Fig. 3. Results of simulation of soil maps and agricultural production groups by the Decision Trees algorithm

It also allows it to be used in applied problems of soil science, agronomy, land management and land management, that is, areas where the need for such data is most acute.

Conclusions. The conducted mathematical experiment has found that there is a significant influence of initial cartographic materials on the qualitative characteristics of simulative ground maps, which are obtained through simulation using a typical set of materials that can be potentially available to ordinary soil scientist or specialist in modern Ukrainian realities. It is shown that the morphometric parameters of the relief and its derivatives are a reliable basis for the predictive modeling of the spatial distribution of soil taxons with sufficiently high accuracy, and the presented method has a significant perspective in solving main scientific and production problems. An expanded estimation of the quality of simulative soil maps has been made at different variants of the initial data and it has been shown that the most promising use is the use of data of complete agro-industrial groups of soils. The differences in the quality of predictive soil maps were established using three types of predictive algorithms and it has been proved that classification models are the most suitable for such problems, in particular Decisions Trees and Random Forests.

The author acknowledge the financial support the National Scholarship Programme for the support of mobility of students, PhD students, university teachers, researchers and artists was established by the approval of the Government of the Slovak Republic. NSP is funded by the Ministry of Education, Science, Research and Sport of the Slovak Republic. The programme is managed by SAIA, n. o. (Slovak Academic Information Agency) [2016/2017:id17680].

References

1. A. B. Achasov Dani dystancijnogho zonduvannja jak osnova kartoghrafuvannja gruntiv: ekonomichnyj aspekt. / A. B. Achasov, Gh. V. Titenko, V. I. Kurilov // Visnyk Kharkivsjkogho nacionaljnogho universytetu imeni V. N. Karazina. — Serija: Ekologhija. – 2015. Vyp. 10. – S. 60–66. URL http://journals.uran.ua/visnukkhnu_ecology/article/download/25458/33191
2. Breiman L. Random forests. / L. Breiman // Machine learning. - 2001. – Vol. 45, № 1. – P. 5-32. URL https://doi.org/10.1023/A:1010933404324
3. Browning, D. M. Digital soil mapping in the absence of field training data: A case study using terrain attributes and semiautomated soil signature derivation to distinguish ecological potential / D. M. Browning, M. C. Duniway // Applied and Environmental Soil Science. – 2011. URL https://doi.org/10.1155/2011/421904
4. Machine learning for predicting soil classes in three semi-arid landscapes / C. W. Brungard, J. L. Boettinger, M. C. Duniway [et al.] // Geoderma. – 2015. – Vol. 239. – P. 68–83. URL https://doi.org/10.1016/j.geoderma.2014.09.019
5. Bui E. N. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray–Darling basin of Australia / E. N. Bui, C. J. Moran // Geoderma. – 2003. – Vol. 111 (1). – P. 21–44. URL https://doi.org/10.1016/s0016-7061(02)00238-0
6. An appropriate data set size for digital soil mapping in Erechim, Rio Grande do Sul, Brazil / A. T. Caten, R. S. D. Dalmolin, F. d. A. Pedron [et al.] // Revista Brasileira de Ciência do Solo. – 2013. – Vol. 37 (2). – P. 359–366. URL https://doi.org/10.1590/s0100-06832013000200007
7. Cherlinka V. R. Problemy stvorennja, gheorektyfikaciji ta vykorystannja krupnomasshtabnykh cyfrovykh modelej reljjefu / V. R. Cherlinka, Ju. M. Dmytruk // Gheopolytyka y ɔkogheodynamyka reghyonov. – 2014. – Vol. 10 (1). – P. 239-244. URL http://geopolitika.crimea.edu/arhiv/2014/tom10-v-1/040cherlin.pdf
8. CherlinkaV. R. Adaptacija velykomasshtabnykh kart gruntiv do jikh praktychnogho vykorystannja u GhIS. In: Aghrokhimija i gruntoznavstvo. Mizhvidomchyj tematychnyj naukovyj zbirnyk. – 2015. – Vyp. 84. TOV «Smughasta typoghrafija», Kharkiv, pp. 20–28. URL http://agrosoil.yolasite.com/resources/2015-AiG-84-pp20-28.pdf
9. Cherlinka V. Using Geostatistics, DEM and Remote Sensing to Clarify Soil Cover Maps of Ukraine. In: Dent, D., Dmytruk, Y. (Eds.), Soil Science Working for a Living: Applications of soil science to present-day problems. Springer-Verlag GmbH, Cham, Switzerland, 2017. – Ch. 7, pp. 89–100. URL https://link.springer.com/chapter/10.1007/978-3-319-45417-7_7
10. Cutler A. Random Forests / A. Cutler, D. R. Cutler, J. R. Stevens. – Springer US, Boston, MA, 2012. – pp. 157–175. URL https://doi.org/10.1007/978-1-4419-9326-7_5
11. Debella-Gilo M. Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS / M. Debella-Gilo, B. Etzelmüller // Examples from Vestfold County, Norway. Catena. – 2009. – Vol. 77 (1). – P. 8–18. URL https://doi.org/10.1016/j.catena.2008.12.001
12. Dobos, E., Hengl, T., 2009. Soil mapping applications. In: Hengl, T., Reuter, H. I. (Eds.), Geomorphometry: Concepts, Software, Applications. Vol. 33 of Developments in Soil Science. Elsevier, Amsterdam, Ch. 20, pp. 461–479. URL https://doi.org/10.1016/s0166-2481(08)00020-2
13. EasyTrace group, 2015. Easy Trace 7.99. Digitizing software. URL http://www.easytrace.com
14. Feng, C., Michie, D., 1994. Machine learning of rules and trees. Machine learning, neural and statistical classification, 50–83. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.27.355&rep=rep1&type=pdf

15. Florinsky, I. V., 2016. Digital Terrain Analysis in Soil Science and Geology, 2nd Edition. ACADEMIC PRESS / Elsevier, Amsterdam. URL https://doi.org/10.1016/c2015-0-02363-2
16. Giasson, E., Figueiredo, S. R., Tornquist, C. G., Clarke, R. T., 2008. Digital soil mapping using logistic regression on terrain parameters for several ecological regions in Southern Brazil. In: Hartemink, A. E., McBratney, A. B., de Lourdes Mendonça-Santos, M. (Eds.), Digital Soil Mapping with Limited Data. Springer Netherlands, Amsterdam, Ch. 19, pp. 225–232. URL https://doi.org/10.1007/978-1-4020-8592-5_19
17. GRASS Development Team, 2017. Geographic Resources Analysis Support System (GRASS GIS) Software. Version 7.2. URL http://grass.osgeo.org
18. Grinand C. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context / C. Grinand, D. Arrouays, B. Laroche, M. P. Martin // Geoderma. – 2008. – Vol. 143 (1). – P. 180–190. URL https://doi.org/10.1016/j.geoderma.2007.11.004
19. Hengl, T., 2009. A practical guide to geostatistical mapping, 2nd Edition. Office for Official Publications of the European Communities, Luxembourg. URL http://www.academia.edu/download/40396676/A_Practical_Guide_to_Geostatistical_Mapping.pdf
20. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping / B. Heung, H. C. Ho, J. Zhang, A. Knudby, C. E. Bulmer, M. G. Schmidt // Geoderma. – 2016. – 265. – P. 62–77. URL https://doi.org/10.1016/j.geoderma.2015.11.014
21. Heung, B. Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes / B. Heung, M. Hodúl, M. G. Schmidt // Geoderma. – 2017. – Vol. 290. – P. 51–68. URL https://doi.org/10.1016/j.geoderma.2016.12.001
22. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach / B. Kempen, D. J. Brus, G. B. M. Heuvelink [et al.] // Geoderma. – 2009. – Vol. 151 (3). – P. 311–326. URL https://doi.org/10.1016/j.geoderma.2009.04.023
23. Kuhn M. Building Predictive Models in R Using the caret Package / M. Kuhn // Journal of Statistical Software. – 2008. – Vol. 28 (5). – P. 1-26. URL https://doi.org/10.18637/jss.v028.i05
24. Lagacherie P. Mapping of reference area representativity using a mathematical soilscape distance / P. Lagacherie, J. M. Robbez-Masson, N. Nguyen-The, J. P. Barthès // Geoderma. – 2001. – Vol. 101 (3-4). – Vol. 105–118. URL https://doi.org/10.1016/s0016-7061(00)00101-4
25. Landis J. R. The measurement of observer agreement for categorical data / J. R. Landis, G. G. Koch // Biometrics. – 1977. – Vol. 33 (1). – P. 159–174. URL https://doi.org/10.2307/2529310
26. Li W. A Random-Path Markov Chain Algorithm for Simulating Categorical Soil Variables from Random Point Samples / C. Zhang, W. Li // Soil Science Society of America Journal. – 2007. – Vol. 71 (3). – P. 656–668. URL https://doi.org/10.2136/sssaj2006.0173
27. Liu, B., 2011. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, 2nd Edition. Springer-Verlag GmbH, London New York Dordrecht. URL https://doi.org/10.1007/978-3-642-19460-3
28. MacMillan, R. A., 2008. Experiences with applied DSM: protocol, availability, quality and capacity building. In: Hartemink, A. E., McBratney, A. B., de Lourdes Mendonça-Santos, M. (Eds.), Digital Soil Mapping with Limited Data. Springer Netherlands, Amsterdam, pp. 113–135. URL https://doi.org/10.1007/978-1-4020-8592-5_10
29. Malone, B. P., Minasny, B., McBratney, A. B., 2016. Using R for Digital Soil Mapping. Progress in Soil Science. Springer International Publishing. URL https://doi.org/10.1007/978-3-319-44327-0
30. McBratney A. B. On digital soil mapping / A. B. McBratney, M. L. M. Santos, B. Minasny // Geoderma. – 2003. – Vol. 117 (1-2). – P. 3-52. URL https://doi.org/10.1016/s0016-7061(03)00223-4
31. Poljchyna S. M. Zastosuvannja suchasnoji systemy klasyfikaciji gruntiv FAO/WRB do karty gruntovogho pokryvu Chernivecjkoji oblasti / S. M. Poljchyna, V. A. Nikorych, O. A. Danchu // Gruntoznavstvo. – 2004. – Vol. 5 (1–2),. – P. 27–33. URL http://arr.chnu.edu.ua/jspui/bitstream/123456789/471/1/Nikorich.pdf
32. Postanova Prezydiji Nacionaljnoji akademiji ..., 2017. Orghanizacijna struktura, porjadok formuvannja ta funkcionuvannja Gruntovo-informacijnogho centru Ukrajiny. Postanova Prezydiji Nacionaljnoji akademiji aghrarnykh nauk Ukrajiny. 20.09.2017 r. Protokol #13. URL http://issar.com.ua/downloads/postanova_vid_20_veresnya_2017_protokol_no13_organizaciyna_struktura_poryadok_formuvannya_gic.pdf
33. QGIS Development Team, 2015. QGIS Geographic Information System. URL http://qgis.osgeo.org
34. R Development Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing. URL http://www.r-project.org
35. Scull P. Predictive soil mapping: a review / P. Scull, J. Franklin, O. A. Chadwick, D. McArthur // Progress in Physical Geography. - 2003. – Vol. 27 (2). – P. 171–197. URL https://doi.org/10.1191/0309133303pp366ra
36. Venables, W. N., Ripley, B. D. 2002. Modern Applied Statistics with S, 4th Edition. Vol. 53 (1) of Statistics and Computing. Springer-Verlag, New York. URL http://dx.doi.org/10.1007/978-0-387-21706-2
37. Walter, C., Lagacherie, P., Follain, S., 2006. Integrating pedological knowledge into digital soil mapping. In: Lagacherie, P., McBratney, A. B., Voltz, M. (Eds.), Digital Soil Mapping: An Introductory Perspective. Vol. 31 of Developments in Soil Science. Elsevier, Amsterdam, Ch. 22, pp. 281–301. URL https://doi.org/10.1016/s0166-2481(06)31022-7

ПОРІВНЯЛЬНА ОЦІНКА ТОЧНОСТІ СИМУЛЯТИВНОГО МОДЕЛЮВАННЯ ҐРУНТОВОГО ПОКРИВУ ТА ПРОГНОЗНИХ КАРТОГРАМ АГРОВИРОБНИЧИХ ГРУП ҐРУНТІВ

В. Р. Черлінка, Ю. М. Дмитрук, В. С. Захаровський

Основною метою математичного експерименту було порівняння точності побудови предикативних карт залежно від різновиду вхідних даних, зокрема ґрунтової карти та повного і скороченого (без дефініцій по гранулометричному складу) варіантів картограми агровиробничих груп ґрунтів. Поставлені завдання вирішувалися: шляхом побудови цифрової моделі рельєфу (ЦМР); оцифруванням картографічних матеріалів; генерацією набору карт морфометричних та інших похідних характеристик; аналізом тісноти зв'язків та ролі згаданих параметрів у мінливості ґрунтового покриву; створенням предикативних карт-версій ґрунтів та картограм агровиробничих груп ґрунтів. Об'єкт досліджень: фрагмент території Чернівецької області зі складними геоморфологічними умовами. Основні використані методи: кореляційний аналіз; метод головних компонент; предикативны алгоритми Decision Trees, Random Forests та K-Nearest Neighbors. На основі кореляційного аналізу було оцінено тісноту зв'язку та роль предикторів (незалежних змінних) у мінливості ґрунтового покриву, що з залученням аналізу головних компонент дозволило обрати з них 9 базових: абсолютна висота; топографічний індекс вологості; кількість сонячної радіації на одиницю площі; крутість схилів; поздовжня та максимальна кривизна топографічної поверхні; акумуляція, довжина та відстань до водних потоків. Якість прогнозних картографічних матеріалів оцінено за допомогою індексу kappa Когена (Cohen's kappa coefficient). Встановлено відмінності у якісних характеристиках отриманих симулятивних карт-версій і показано, що морфометричні параметри рельєфу та його деривати є надійним базисом предикативного моделювання. Зроблено розширену оцінку якості карт-моделей залежно від типу вхідних даних і показано, що найбільш точною при використаному наборі предикторів є прогнозна картограма повних агровиробничих груп ґрунтів. Встановлено відмінності у якості прогнозних ґрунтових карт при використанні 3 типів предикативних алгоритмів та показано, що найбільш придатними для такого роду задач є класифікаційні моделі, зокрема Decision Trees та Random Forests, застосування яких дозволило отримати до 93% співпадіння реальних та модельних даних. Показано можливості щодо побудови прогнозних карт ґрунтів з використанням типового набору матеріалів, які можуть бути доступними ґрунтознавцю в сучасних українських реаліях: карти ґрунтова та топографічна і безкоштовне повнофункціональне програмне забезпечення – геоінформаційні системи GRASS та Quantum, векторизатор Easy Trace і мова статистичних розрахунків R-Statistic.

Ключові слова: ґрунтова карта, картограма агровиробничих груп ґрунтів, навчальна вибірка, симуляція, морфометричні параметри, ЦМР, предикативні алгоритми.

Отримано редколегією 18.11.2017