Methods of verification of soils prediction maps: a case study from Chernivtsi region, Ukraine

Vasyl CHERLINKA, Yuriy DMYTRUK, Dušan BARABAS

Abstract: *Knowing the spatial distribution of individual soil taxonomic units is a key factor in managing efficient land use not only for agriculture but also for forestry. The use of a comprehensive soil surveys held in past decades and based on fieldwork created the basis for the initial spatial representation of the soil fund structure. However, the spatial distribution of the soil cover was the result of fieldwork and the experience of the person who drew this map. Often this led to some errors in determining the types of soils and their boundaries. To date, there is a growing need for precise methods of land taxation, based on correct information on soil cover. In countries with a large area, such as Ukraine, field surveys still do not cover the whole territory, often the density of the allocation of soil pits was too low, which in some cases led to an incorrect demarcation of soil boundaries. Since such a problem is very urgent for Ukraine, the search and identification of probable problem soil maps by constructing their predicted versions, their comprehensive analysis and cross-validation is an important task. The conducted investigations revealed that morphometric parameters of the relief and their derivatives obtaining from the analyses of DEM are a reliable basis for the predictive modelling of the spatial distribution of soil cover with sufficiently high accuracy, and the methodology based on 11 types of prognostic algorithms would have a significant prospect in solving scientific and production problems. Very important in this process is the selection of predictors derived from the DEM, as well as the structure and distribution of the training dataset, based on which the model will be built later. Afterwards the results need to be validated, in our case, on the basis of the cross-validation of the models and by comparing the results with field survey. The article presents the results of 11 simulations, evaluates the quality of predictive algorithms and the models obtained. Therefore, several possible ways to check the cartographic and simulation results of the spatial distribution of soil taxonomic units were described, as well as their comparison with those actually existing in nature. The most reliable method of the 11 presented is a direct study of the soil in the field and comparing them with the soil map. It is recommended to use it in case of suspicion of poorly executed maps, although financially it is very expensive. More preferred is a set of modelling methods that is based on the data already collected. With reliable sources, they provide an opportunity to predict the soil in places where the survey was not conducted at all. Verification of the quality of the tested models was carried out on a fragment of the Ukrainian region within the boundaries of the Chernivtsi region, confined to the Prut-Dniester and Prut-Siret interfluves.*

Keywords: *soil map, simulation, morphometric parameters, DEM, prediction algorithms*

Introduction

It is important to have comprehensive information on the soil cover of a specific territory for high-productivity agricultural production, monitoring and environmental quality management. However, the analysis of published research shows (Jones et al. 2005) that large-scale soil surveys (for example, scale 1:10 000) have not been conducted for all states.

DOI: https://doi.org/10.33542/GC2019-2-04

So, in Ukraine, about 15 of the 60 million hectares remained unchecked. At the moment, only agricultural land was investigated. The areas of settlements, forests, mountainous regions in most cases remained off the attention of researchers. The results obtained in 1957-1961 are outdated, there are many inappropriate results, and the correction of these data was not carried out over 25 years (Cherlinka 2017a). The situation in other European countries is different, but similar problems often occur. In Slovakia, as in one of the countries, the "Comprehensive Survey of Agricultural Areas" took place in 1961-1970. The survey was focused on agricultural areas only in the form of field research. The basic unit was the agricultural production entity. The basis of the survey was three categories of probes, with a density of 1 probe per 7-18 ha, a selection of 70-180 ha and a special one probe for 3-4 thousand hectares. This survey has become the basis for processing maps of Bonitated Soil Units (BPEJ) at a scale of 1:5 000 and 1:10 000 including other parameters relevant to the management of agricultural land. The BPEJ maps included other parameters important for the management of agricultural land (Džatko 1974). The same nature of the soil mapping was performed in the Czech Republic as the two countries were in a common state. The first soil map in Poland at a scale of 1:500 000 originated in 1907 (Miklaszewski 1907). Map of soil-agricultural units was created at a scale of 1:5 000 in 1956 based on a field survey. This map work was also categorized by forest taxons. A map was adopted in 1999 that accepted the international classification of soils. It was processed at a scale of 1:1 000 000 with a further refinement (Białousz et al. 2005). Despite detailed land surveys, detailed maps of the forest soils cover are lacking. Note that such problems are inherent not only for Ukraine or fora number of other developing countries, but also, for example, in Australia (Bui and Moran 2003).

Manual allocation of slope steepness involved errors which propagated into delineation of soils of varying degrees of erosivity. This approach was used in Ukraine resulting in significant inaccuracies of spatial pattern of soil types. Filling the gaps and adjusting the boundaries of the soil map data in such conditions can be improved by simulation. The number of studies devoted to the modelling of the spiatial distribution of taxonomic soil units is increasing (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 multivariate regression analysis, kriging, neural networks to different types of classification trees (Florinsky 2012).

At the same time, in recent years attention has been increasingly paid to machine learning methods, such as the Classification and Regression Tree (CART) and Random Forests, while the proportion of classical methods such as regression kriging is decreasing (Keskin and Grunwald 2018). The general idea underlying the application of such methods is using the reference points of the landscapes with linked taxa associated with them (Lagacherie et al. 2001). Digital elevation model (DEM) is the main source of predictors in such a simulation as many geomorphometric parameters can be derived of the DEM and used as proxies of soil parameters (Kempen et al. 2009, Hengl et al. 2017, Marques al. 2018). The challenge in using DEMs is in defining relationship between quantitative measures of terrain topography and categorical variables defining as the soil types. Therefore, advanced mathematical methods are needed to establish the relationship between all these parameters which existence can be non-obvious at the first glance (Giasson et al. 2008, Kempen et al. 2009, Debella-Gilo and Etzelmüller 2009, Hengl 2009, Cherlinka 2017a, Malone et al. 2016).

The simulation results usually have a deviation from the real state of things, both for their own simulation, and for the probable errors on the soil maps, which are the inputs for calculating the input parameters. Therefore, the analysis of possible methods for verifying the results obtained is a necessary and important step for making substantiated conclusions about the quality of the performed soil prediction.

Research background

The general simulation procedure involves allocation of a certain portion of data from the population under study for machine learning and the subsequent simulation will be already based on these data. Feng and Michie (1994) characterize this process through certain stages: the generation of a training sample, learning the algorithm; creation of classification rules; testing on a complete set of data.

Thus, the main task of constructing a training sample for the subsequent creation of a forecast soil map (or any other map with categorical data) is the choice of such points, the spatial arrangement of which would most fully cover the variation of taxonomic units of soils and their corresponding predictors. Model training on this sample allows you to establish relationships between all these parameters and then transfer obtained results to the entire study area. Potentially, this also allows extrapolating or interpolating results beyond the existing soil maps, since a set of predictors is derived from a DEM that covers the entire territory.

In constructing of training dataset Brungard et al. (2015), Heung et al. (2016), Heung et al. (2017) clearly distinguish between 2 approaches: (i) based on a field research of soil pits and (ii) a sample of clearly defined polygons from soil maps. The first approach has good prospects, but requires a large established database of verified soil cuts. Large-scale mapping of soils requires a significant amount of such samplings, which in practice is costly.

When choosing a training dataset different authors use diverse arsenal of techniques, from simple mechanistic (Steers and Hajek 1979, Wright and Wilson 1979, McKay et al. 2000, Campling et al. 2002, McBratney et al. 2003, Walter et al. 2006, White 2006, Giasson et al. 2008, Hengl 2009, Caten et al. 2013, Malone et al. 2016), to theoretically substantiated (McBratney et al. 2003, Hengl et al. 2003). There are also different opinions about choosing the method for locating the points of the training sample (Walter et al. 2006).

For example, Bui and Moran (2003) describe how to select a certain percentage of data to cover the area of individual habitats of each soil class by randomized learning sample – area-weighted approach. Thus, for different map scales different percentages of area-weighted points were used, particularly: 15% for 1:500 000, 25% or 20% for 1:250 000; and 35% for 1:100 000. Walter et al. (2006) defined several possible strategies for creating a sample: a simple randomized selection (Wright and Wilson 1979), random transects with a fixed distance between the points (Steers and Hajek 1979), stratified sampling to investigate the differences between polygons at short distances (Walter 1990), cited by (Walter et al. 2006).

White (2006) offers one of the possible options to select educational samples based on an expert approach. In contrast, McKay et al. (2000) distinguishes three main ways of allocating a training set of data: randomized, stratified and based on the Latin hypercube. Hengl et al. (2003) tried to summarize all the methods and bring them the theoretical basis. In further studies, Hengl (2009) concluded that random selection of points has drawbacks due to variations in the quality of existing maps in their various parts, and since soil taxa have plane (polygonal) character, they suggested placing training points along the medial (median) axes of these landfills. Accordingly, it was substantiated that such a method of their location more fully described soil conditions, and to minimize classification errors within the boundaries of soil differentiation.

For 12 classes of soils, Malone et al. (2016) used sample dataset based on 1000 points, of which 70% use to internal validation of model, and 30% reserved for external validation on the same sample. Then the results of "learning" are transferred to a complete set of data, that is, in this case, the number of points of the study sample is strictly deterministic and is an average of only 700/12 = 58 points per class. Similar strategies were used by Campling et al. (2002) or Giasson et al. (2008). McBratney et al. (2003) gave clearer guidance on the number of points in the training dataset (relative to the previously rasterized map): $0.0001M < x < 0.001M$ points located randomly. Caten et al. (2013) state that a set of training points less than 5% of their total number is not representative, and more than 20% is excessive and requires unnecessary computing time without improving the quality of the final models.

Hengl et al. (2018) generated a global predicted gridded soil map SoilGrids250m based on 150,000 soil pits used for training and a stack of 158 remote sensing-based soil covariates derived from MODIS land products, SRTM DEM derivatives, climatic images and global landform and lithology maps. All of this were used to fit an ensemble of machine learning methods – random forest and gradient boosting and/or multinomial logistic regression (Hengl et al. 2017).

A new method for the digital mapping of taxonomic soil units via fuzzy taxonomy and fuzzy clustering was proposed by Horáček et al. (2018). Fuzzified taxonomic soil information from 106 soil pits with 75 geomorphometric parameters (potential environmental covariates of soil units) derived from a 10 m LIDAR DEM was used for the input (training) data for territory of 104.5 km^2. It was shown that in the zonal dataset the absolute match of the control soil pits (60) and the results of the modelling is only 26%, but the absolute mismatch is only 10%. The remaining 64% showed a partial match.

Thus, a number of predictive algorithms give a high (up to 100%) coincidence of forecasted and real classification units when using large volumes of the study sample. However, this does not always correspond to such accuracy in the entire volume of data. Therefore, it is useful to evaluate the options for constructing a training sample with the use of several perspective results of a weighted approach in predicting the best results not only on training data but also on real data sets. This is important given that prediction maps are interesting as the object of scientific study and as an important tool for obtaining information on soil cover in locations where no studies have been conducted yet. Therefore, the higher the degree of coincidence of predictive data with real maps, the more substantiated will be the conclusions on the information localized in unsampled (i.e. white spots) of large-scale soil maps.

Another study (Teng et al. 2018) had used dataset, containing 38 756 observations and their covariates for whole territory of Australia, but training and a validation was set by random sampling. Two-thirds were assigned to the training set and the remaining profiles were used for the validation. The accuracy of classification of forecasting soils obtained at these models does not exceed 69%, and often was much smaller.

Note also that almost all of these authors consider the verification of the results obtained from the purely mathematical side. We propose a so-called cross-validation, the various variants of which will be discussed in the section «Results and discussion», which allows to compare model data with real soil surveys and existing maps. This potentially allows to track errors not only in the model experiment, but also yet inherited in existing soil maps. For example, Minár (2003) showed large disparities by means direct comparing the information from existing detailed soil maps and hundreds of new soil pits in several regions of the Western Carpathians. On the level of soil types, his results showed a full match for 16-18% of pits, partial match for 30-49%, and absolute mismatch for 34-54%. Naturally, there is a strong demand for high-precision soil information for a variety of purposes, for example, precision farming.

The purpose of our work was to compare the methods of verifying the simulated data and give recommendations regarding the optimal combination of verification methods, the methodology of creating a training set of data and the algorithm itself for constructing the model. At the same time, we tried to make the study process as transparent and reproducible as possible, therefore, we propose an approach involving purely open-source software.

Methods and Data

The approach of presented research comprised the following tasks: (a) vectorising raster maps and assigning attributes the geographic objects; (b) construction of DEM with a resolution of 5 m based on the vectorised contours; (c) geomorphometric analysis of the DEM generating relevant raster layers for soil prediction; (d) generating a training dataset according to the described methodological approaches; e) the establishment of a network of soil pits in field research; (f) simulation and generating soil models (map-versions or map-models) using 11 basic types of predictive algorithms; (g) analysis of methods for verification of quality of predictive modelling and map of the soil cover.

The object of interest involved a part of the territory of Ukraine (Fig. 1a) within the boundaries of the Chernivtsi region (Fig. 1b) in the west of the state. The first polygon belongs to the administrative area of the city of Chernivtsi (Fig. 2a), confined to the Prut-Dniester interfluve. Specific problems of spatial modelling of soils from Ukrainian perspective were subject to research in this area in various studies (Cherlinka and Dmytruk 2014, Cherlinka 2015, Cherlinka 2017a). A map sheet of a topographic map M 1:2 000 is the most detailed source of topographic information on this area (Fig. 3a). The map was georeferenced in GIS Quantum (QGIS Development Team 2015) using the coordinate system Pulkovo 1942 CS63 Zone X2 (Evenden and Warmerdam 1990), EPSG code 7826. The second test area is located in the Glybotsky district (Fig. 2b) of the Chernivtsi region (Prut-Siret interfluve), for which a topographic map M-35-136-G-g-3 was selected. The map was georeferenced similarly as the first map.

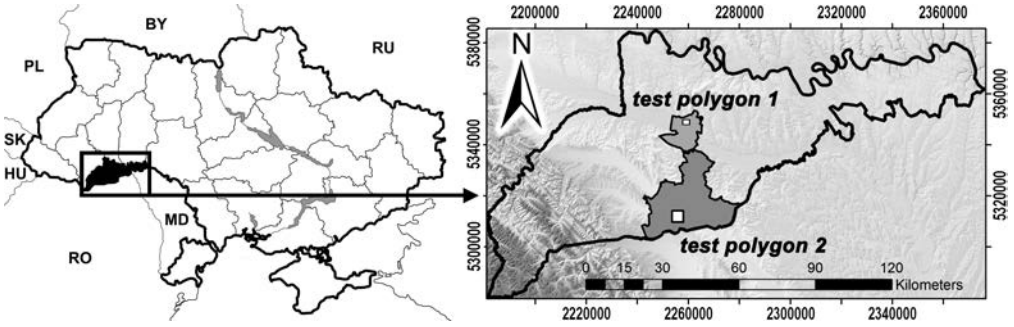

Fig. 1. *Geographical location of the research areas within Ukraine (a) and Chernivtsi region (b) (*for background was used SRTM data – NASA's Shuttle Radar Topography Mission)*

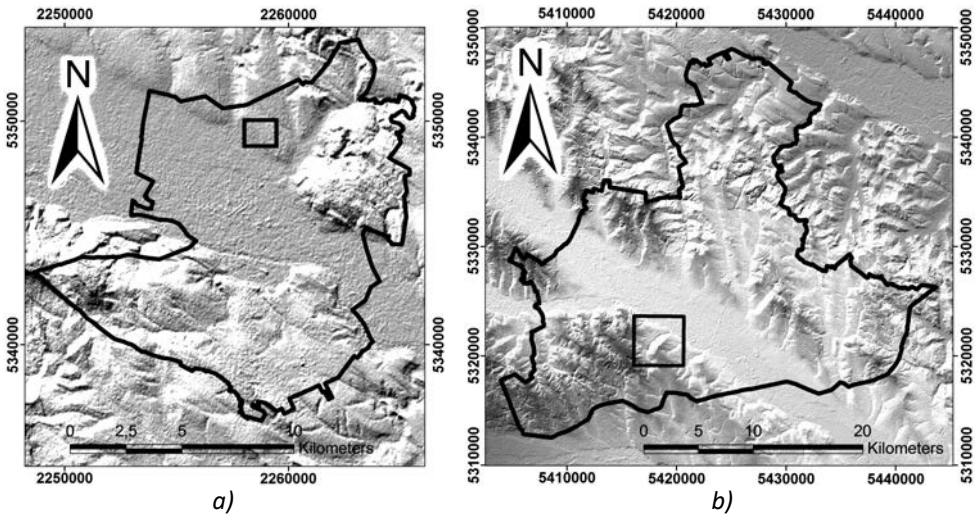

Fig. 2. *Disposition of the research polygons in Chernivtsi city (a) and Glybotsky district (b) (*for background was used SRTM data – NASA's Shuttle Radar Topography Mission)*

Informative soil materials for the first area was based on the archival agro-industrial soil map of the research area created in 1993 (Fig. 3b). The test area was selected for comprising the full soil catena and entire coverage (Fig. 3c). The soil map for second area (Fig. 3d) was created in 1974 and it comprises a big gap (no soil data).

We used open-source software for data processing. Digitization and vectorising was performed in Easy Trace (EasyTrace group 2015), preparation of maps of geomorphometric parameters was done in GRASS GIS (GRASS Development Team 2017) and the construction of a prediction model of soil cover was conducted in R which is a language and environment for statistical computing (R Development Core Team 2017). The DEM for both areas was interpolated at the spatial resolution of 5 m from contour lines using the v.surf.rst module based on regularized splines with tension and smoothing (Mitášová and Mitáš 1993) in GRASS GIS. Tuning the interpolation parameters based on Hofierka et al. (2007). The DEM cell size of 5 m was chosen because with a relatively high accuracy reproduction of the topography, it also provides a practical coincidence of the areas of vectorized and rasterized soils (Cherlinka 2017d). This resolution also enables to express minimum area of soil being mapped on the map of 1:10 000 scale which is 0.3 ha corresponding to 120 pixels of 5 x 5 m which allows for creating a complete training dataset. This is difficult to achieve with a lower detail of the map, i.e. coarser resolution. Also, we used the agronomic classification of soils in Ukraine for which such spatial scale is required for agronomic suitability.

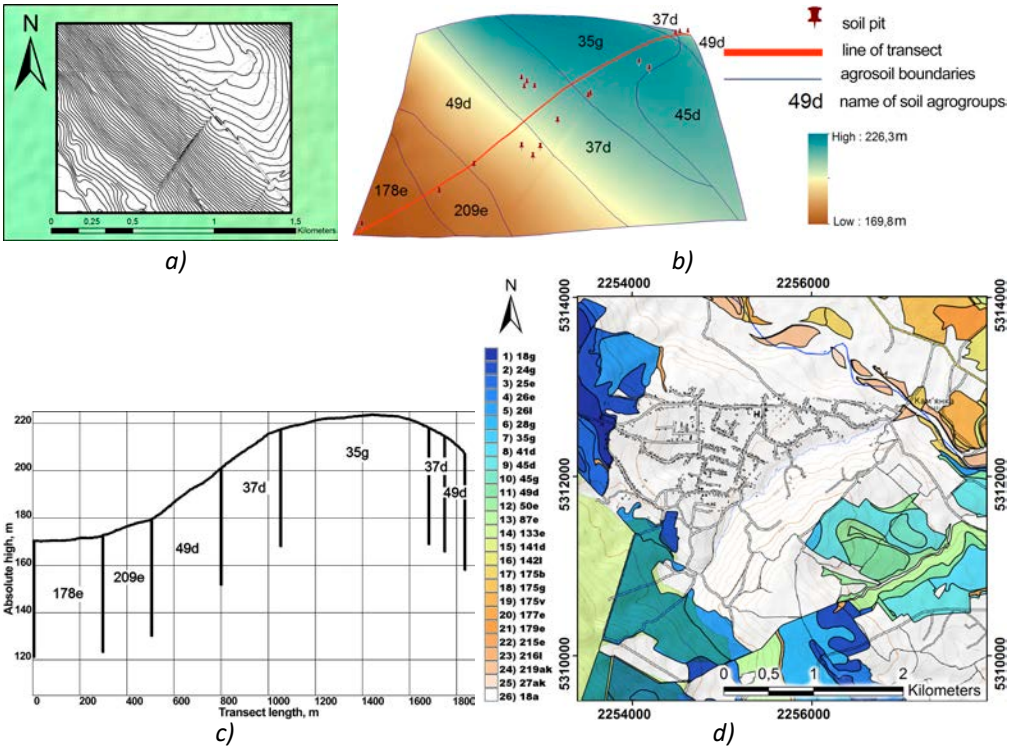

Fig. 3. *Topographic data of test polygon 1 (a) and location of agro-industrial soil groups/point of soil pits (b) along the transect on the investigated catena (c) and soils map of polygon 2 with gaps in soil survey (d)*

In the process of choosing predictors for the model, we used the results of McBratney et al. (2003) regarding the SCORPAN model. However, given the inaccessibility of many specific predictor data, we have focused on the most accessible, in particular, hydrological, climatic and morphological properties of a terrain. Hence, DEM was the basis for the allocation of a certain number of its characteristics, particular: altitude (*dem*), *slope* angle and slope *aspect* (module r.slope.aspect as defined by Hofierka et al. (2009)), DEM curvatures (*profc*, *planc*, *longc*, *minic*

and *maxic* respectively by the module r.param.scale based on Wood (1996)), data on global solar radiation for day 180 at 2 p. m. (*rad*) (module r.sun by Hofierka and Šúri (2002)), landform type (*gmf*) by r.geomorphon (Jasiewicz and Stepinski 2013). Additional maps of hydrological indicators were also generated: topographic wetness index (*twi*) (Moore et al. 1993) in r.topidx, accumulation (*flowaccum*) and the direction of water flows (*flowdirect*) in r.terraflow (Arge et al. 2003), length of flowlines *flowlength* (Mitášová and Hofierka 1993) in r.flow and the distance to them (*diststream* by r.stream.distance by Jasiewicz and Metz (2011)). Cherlinka (2017b) analysed these predictors in detail and showed that for conditions used in the presented work, the minimum acceptable set of the predictors consists of the following parameters, which were used as basic: *soil* is a soil mapping unit; *dem, twi, rad, slope, longc, maxic, flowaccum, flowlength, diststream* are the above described parameters. To date, there are fuzzy c-means clustering analyses that show that there are more optimal ways to select predictors, which will be used in our future research.

Simulation models of soil cover was performed by a script described in Cherlinka (2017a, b, c). The code includes a number of adaptations for solving a set tasks and implements 14 basic types of predictive algorithms, of which 11 were used in this study, in particular: 1) Multinomial Logistic Regression – MLR (Giasson et al. 2008, Kempen et al. 2009, Debella-Gilo and Etzelmüller 2009, Hengl 2009, Cherlinka 2017a, Malone et al. 2016); 2) Neural Networks – NN (Venables and Ripley 2002, Ripley and Venables 2016); 3) K-Nearest Neighbors – KNN (Kuhn 2008, Liu 2011); 4) Random Forests – RF (Breiman 2001, Cutler et al. 2012); 5) Nonlinear Discriminant Analysis – NDA (Huberty and Olejnik 2006, Kuhn 2008); 6) Support Vector Machines – SVM (Venables and Ripley 2002, Kuhn 2008, Hastie et al. 2009); 7) Linear Discriminant Analysis – LDA (Huberty and Olejnik 2006, Kuhn 2008); 8) Partial Least Squares Discriminant Analysis – PLS (Kuhn 2008, Hair et al. 2010); 9) Penalized Logistic Regression – PLR (Kuhn 2008, Hilbe 2009); 10) Nearest Shrunken Centroids – NSC (Venables and Ripley 2002, Kuhn 2008, Hastie et al. 2009); 11) Bagged Trees – BGT (Hastie et al. 2009, Peters et al. 2009, Kuhn and Johnson 2013).

The example of a lines of code in R for the different models shows the main principle (Tab. 1). The training dataset in all variants with map data was created by randomly-weighted average method on the basis of a detailed analysis (Cherlinka 2017c). In general, the modelling process involves the creation of a training dataset, the use of which for the "training" of predictive algorithms allows for ascertaining prognostic maps in the future. The proposed technique (Dobos and Hengl 2009), named “median” in this paper, showed good results. Therefore, it was used as an etalon for comparison with other methods, in particular with the median and randomized weighted averages (Cherlinka 2017c). The calculations revealed that in the etalon median training dataset the ratio of the training pixels and their percentage distribution does not correspond to the proportions of the areas of the soil on the original map. Therefore, based on this etalon median sampling, we created a median-weighted average sample, a set of pixels that fully corresponds to the structure of the soil areas of the original map. In this case, considering the limited amount of data from the etalon sample and the need to adhere to the exact proportions between the individual pixels of the soil, the size of the training dataset decreases with respect to the standard. Therefore, the randomized-weighted averages approach proposed in this article is devoid of the problem with decrease in the actual number of pixels of the training dataset relative to the median-weighted approach. Then it allows getting precisely those proportions between the training pixels and the total sample size, which is conditioned by the conditions of the planned experiment, for example, 5%, 10%, 15% or other needed proportions. The implementation of such calculations is done using the written script on Python in the GRASS GIS environment (Cherlinka 2017c).

Tab. 1. *R code example*

Type of model	R code
MLR	mlr.soilpredict <- multinom(soil ~ dem + twi + rad + slope + longc + maxic + flowaccum + flowlength + diststream, traindata, maxit = 5000)
NN	NN.soilpredict.train <- nnet(soil ~ dem + twi + rad + slope + longc + maxic + flowaccum + flowlength + diststream, data = traindata, size = 9, decay = .1, maxit = 5000)
KNN	KNN.soilpredict.train <- train(soil ~ dem + twi + rad + slope + longc + maxic + flowaccum + flowlength + diststream, data = traindatastatd, method = "knn", metric = "Kappa", preProc = c("center", "scale"), na.action=na.exclude)
RF	RF.soilpredict.train <- randomForest(soil ~ dem + twi + rad + slope + longc + maxic + flowaccum + flowlength + diststream, data = traindatastatd, ntree = 5000, mtry = 5, na.action=na.exclude)
NDA	NDA.soilpredict.train <- mda(soil ~ dem + twi + rad + slope + longc + maxic + flowaccum + flowlength + diststream, data = traindatastatd, subclasses = 7, na.action=na.exclude)
SVM	SVM.soilpredict.train <- train(soil ~ dem + twi + rad + slope + longc + maxic + flowaccum + flowlength + diststream, data = traindatastatd, method = "svmRadial", metric = "Kappa", preProc = c("center", "scale"), na.action=na.exclude)
LDA	LDA.soilpredict.train <- train(soil ~ dem + twi + rad + slope + longc + maxic + flowaccum + flowlength + diststream, data = traindatastatd, method = "lda", metric = "Kappa", preProc = c("center", "scale"), na.action=na.exclude)
PLS	PLS.soilpredict.train <- train(soil ~ dem + twi + rad + slope + longc + maxic + flowaccum + flowlength + diststream, data = traindatastatd, method = "pls", metric = "Kappa", preProc = c("center", "scale"), na.action=na.exclude)
PLR	PLR.soilpredict.train <- train(soil ~ dem + twi + rad + slope + longc + maxic + flowaccum + flowlength + diststream, data = traindatastatd, method = "plr", metric = "Kappa", preProc = c("center", "scale"), na.action=na.exclude)
NSC	NSC.soilpredict.train <- train(soil ~ dem + twi + rad + slope + longc + maxic + flowaccum + flowlength + diststream, data = traindatastatd, method = "pam", metric = "Kappa", preProc = c("center", "scale"), na.action=na.exclude)
BGT	BGT.soilpredict.train <- bagging(soil ~ dem + twi + rad + slope + longc + maxic + flowaccum + flowlength + diststream, data = traindatastatd, nbagg = 100)

To evaluate the quality of the model, the Cohen's kappa index was used (Landis and Koch 1977, Foody 2004, Li and Zhang 2007, Grinand et al. 2008, Congalton and Green 2008, Kuhn 2008, Hengl 2009, Malone et al. 2016, Marques al. 2018), because kappa statistics is a common measure of classification accuracy. The evaluation of the statistical significance of the difference in accuracy between two soil maps (source and predicted) has often been based on the comparison of the kappa coefficient calculated for each map. In general, the kappa coefficient of agreement for a thematic map is based on the comparison of the predicted and actual soil class labels for each case in the model dataset and can be calculated from:

$$\kappa = \frac{p_0 - p_c}{1 - p_c},$$

where p_0 is the proportion of cases correctly allocated and p_c is the proportion of agreement that is expected by chance. So, the derived coefficient kappa provides an estimate of the accuracy of the map which together with that derived from another map is the basis of most map comparisons. In order to maintain consistent nomenclature when describing the relative strength of agreement associated with kappa statistics, the following labels were proposed for the corresponding ranges of kappa (Landis and Koch 1977) and used in our research (Tab. 2).

Tab. 2. *Ranges of kappa*

Kappa Statistic	Strength of Agreement	Kappa Statistic	Strength of Agreement
<0.00	Poor	0.41-0.60	Moderate
0.00-0.20	Slight	0.61-0.80	Substantial
0.21-0.40	Fair	0.81-1.00	Almost Perfect

In order to systematically approach the analysis of possible ways of verification of real, cartographic and modelled data, we have developed an appropriate scheme (Fig. 4). In particular, they can be divided into three groups, depending on the type of input information: the existing soil map (polygon data), reliable field survey of soils (point data) and their combination. These data groups are the basis for constructing three corresponding prediction models.

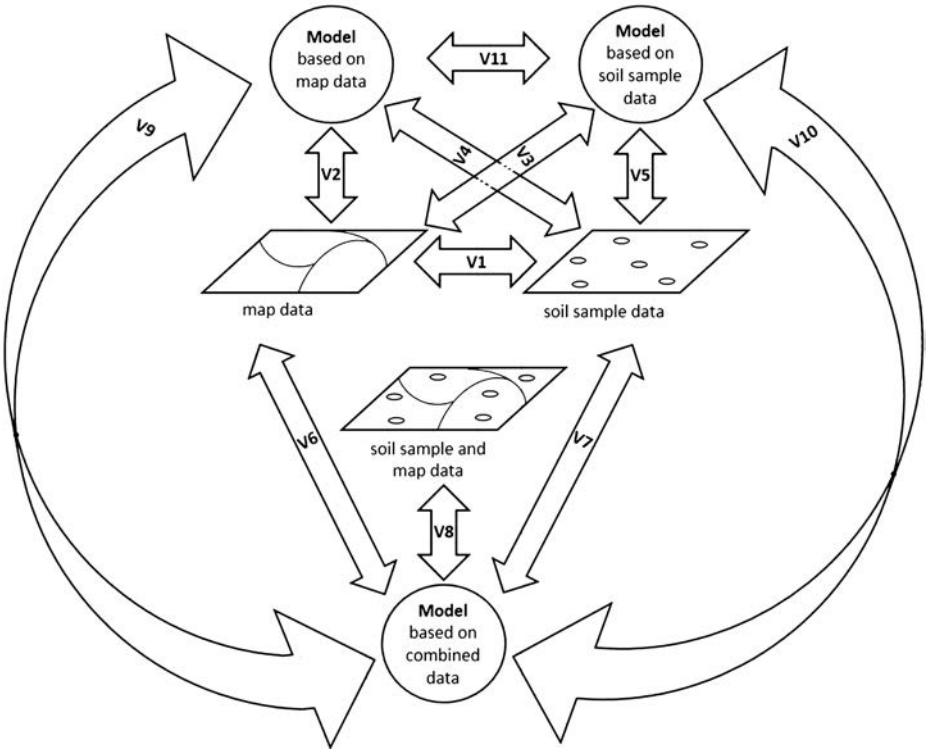

Fig. 4. *Types of verification real, cartographic and modelled data*

Results and discussion

Verification of real, cartographic and modelled data

The analysis of paths of verification of real data, mapped data, and modelled soil data shows that there are several possible options for this process (Fig. 4). Note that the only type of verification that does not require simulation matches with the first version (V1) and is dedicated to compare the field surveys with cartographic material. The models of the first type ($Model_{map}$), the second type ($Model_{pit}$) and the third type ($Model_{combined}$) are created with the algorithms described in the section on Methods and data and the types can be cross-verified in several ways, as among themselves (V9-V11), and with data sources (V2-V8). In this case, $Model_{combined}$ can be validated with 5 validation options (V6-V10), and for two other types of model are only 4: V2, V4, V9, V11 and V3, V5, V10, V11 – for $Model_{map}$ and $Model_{pit}$ respectively. The need for one or another validation option is determined when planning an experiment depending on its purpose and in extreme cases all 11 methods can be involved.

Specifically, in our case, we planned to investigate the accuracy of the existing map of agrogroups of soils, since there were certain doubts about its reliability, in particular the accuracy of soil boundaries and soil types. Thus, capabilities and options for direct and inverse verification can be traced and used for this purpose. For Ukraine, a major problem is the availability even at least of outdated soil maps, which, moreover, often contain significant errors. Therefore,

in this case, the main thing was to check the quality of the soil map according to the given soil pits: inverse verification by type V1. That is why the following scheme of laying the soil pits on the catena was planned (Fig. 3b), which provided control of the soil boundaries and the accuracy of the definition of the soil on map. Since the simplest and most obvious comparison is the comparison of the soil diagnosed in the field conditions with soil on map (V1) we have summarized the data in Tab. 3. The table shows qualitatively correlation of the actual investigated soils with the soils depicted in the maps. Let us recall that the correspondence between the codes of agro-industrial groups of soils and their names is given in Tab. 4. The expected inaccuracies of the map quality was confirmed, because the soil map coincided with the field soil only in few cases (highlighted bold: soil pits 12, 13 and 16). In all other cases, we observe differences (sometimes quite strong) between the genetic types of real soils and the data from the map indicating most likely lacking expertise of map authors. Soils identified us in the field did not match the map data in 15 of the 18 researched cases. Given the relatively flat landscape and simple geomorphology, such mapping errors can be considered as blunders. In addition, field surveys and mapped data have only one aggregate soil group, in the area of which there are just three coincident soil pits. All other agrogroups do not coincide in quality and number: according to our research, 8 agrogroups are in real against 6 on the map.

The presence of such gross differences indicates that further modelling based on map data ($Model_{map}$) in our case makes sense only to assess the quality of predictive algorithms by covering the range of research albeit with false but complete data. Accordingly, data validation under $Model_{combined}$ and its corresponding variants (V6-V10) is impossible here, since such a model can only be constructed in the case of a complete match of field studies with mapped data. Another model ($Model_{pit}$) and a set of corresponding variants verification (V3, V5, V10 and V11) can be in principle constructed, as a training dataset will use field diagnostics that is absolutely reliable and based on our personal observations. However, the simulation of soil cover for test polygon 1 cannot be accurate a priori, since the purpose of laying soil pits in our case is not subject to the objectives of large-scale soil mapping (which involves a completely different layout of soil pits) and the goal is the control of accuracy of soil boundaries and the allocation of soil types themselves.

Modelling based on cartographic data

Since the results of field studies test area 1 showed that the continuation of the work is possible only with $Model_{map}$ and $Model_{pit}$, we tried to get the most useful information from this situation. The accuracy of the models in cases of estimation of cartographic and real data, as noted above, is determined by means of the Cohen's kappa index. The index determines how precisely the model describes the conditions for the placement of soils, and, accordingly, to reproduce the soil map. Consider closer the $Model_{map}$. For the low semantic and spatial accuracy of the existing agrosoil map, this simulation has a purely academic interest in connection with the validation and ranking of a set of prediction algorithms. Obtained 11 soil cover simulations (Fig. 5) show quite interesting regularities about the quality of simulation even when analysing only the visual characteristics of the obtained maps. The evaluation of their numerical characteristics reveals that two classifications models (Random Forest and Bagged trees) (Fig. 6) are the most accurate.

The following array of simulations of soil cover is interesting in terms of its correspondence to the original map and, accordingly, predictive power for areas with no information (for research, where relevant). Since the algorithms analyse the entire spectrum of prediction parameters, producing classification rules, then at a high level of coincidence of model and real data, one can speak of a certain level of statistical reliability of the results in the areas of "white spots". For this reason, our conclusions in this regard are quite encouraging in the view of the range of values of Cohen's κ (Fig. 6).

Tab. 3. *Correlation between nature and map data for test polygon 1*

Sequence number of pit	Agrosoil in nature	Agrosoil map
17	*29d	37d
1, 2	*33d	35g
7	*33d	49d
6	37d	49d
12, 13, 16	**37d**	**37d**
11	*37e	37d
5	*38e	37d
3	*40d	45d
4	*40d	35g
15	*121e	178e
8, 9, 10, 18	*209d	49d
14	*209d	209e

**not present on agrosoil map*

Tab. 4. *The correspondence of codes of agro-industrial groups of soils with their names*

Code*	Name of agro-industrial groups of soils	Available on area
18a	Humus podzolic and podzolic humus gleyed sandy soils	2
18g	Humus podzolic and podzolic humus surface-gleyed loamy soils	2
24g	Humus podzolic surface-gleyed weakly eroded soils	2
25e	Humus podzolic surface-gleyed medium eroded heavy loamy soils	2
26e	Humus podzolic surface-gleyed strongly eroded heavy loamy soils	2
26l	Humus podzolic surface-gleyed strongly eroded light loamy soils	2
27ak	Humus podzolic gleyed sandy loam drained rocky soils	2
28g	Humus podzolic surface-gleyed drained light loam soil	2
29d	Light gray and gray podzolic soils	1
33d	Light gray and gray podzolic gleyed soils	1
35g	Light gray and gray podzolic surface-gleyed soils	1, 2
37d	Light gray and gray podzolic weakly eroded soils	1
37e	Light gray and gray podzolic weakly eroded soils	1
38e	Light gray and gray podzolic medium eroded soils	1
40d	Dark gray podzolic and weakly regressed soils	1
41d	Podzolic chernozem and weakly regressed and dark gray strongly regressed medium loam soils	2
45d	Dark gray podzolic and podzolic chernozem gleyed	1, 2
45g	Dark gray podzolic and podzolic chernozem gleyed light loam soils	2
49d	Dark gray podzolic, podzolic chernozem and regressed weakly eroded soils	1, 2
50e	Dark gray podzolic and regraded soils and podzolic chernozems, and regraded medium eroded heavy loamy soils	2
87e	Chernozem unsalonized and slightly salonized on dense clay strongly eroded heavy loamy soils	2
121e	Meadow-chernozem soils and their slightly saline and slightly solodized varieties	1
133e	Meadow, chernozem-meadow soils and their slightly saline and slightly solodized varieties heavy loamy soils	2
141d	Meadow marsh, swamp and muddy marsh non-dried medium-sandy soils	2
142l	Meadow marsh, swamp and muddy marsh dried light clay soils	2
175b	Humus non-deep gleyed clay-(bind)-sandy soils	2
175g	Humus non-deep gleyed light loam soil	2
175v	Humus non-deep gleyed sandy soils	2
177e	Humus non-deep gleyed heavy loamy soils	2
178e	Sward deep gley soils and their podzolic variants	1
209d	Alluvion chernozems and meadow chernozem soils	1
209e	Alluvion chernozems and meadow chernozem soils	1
215e	Eroded soil and the outputs loose (sand and loess) heavy loam soil species	2
216l	Eroded soil and the outputs quaternary clay light clay soil	2
219ak	Modern riverbed sediments and sandy rocky soil	2

**a – sandy; b – clay-sandy; d – medium loam; e – heavy loam; g – light loam; k – stony; l – heavy clay*

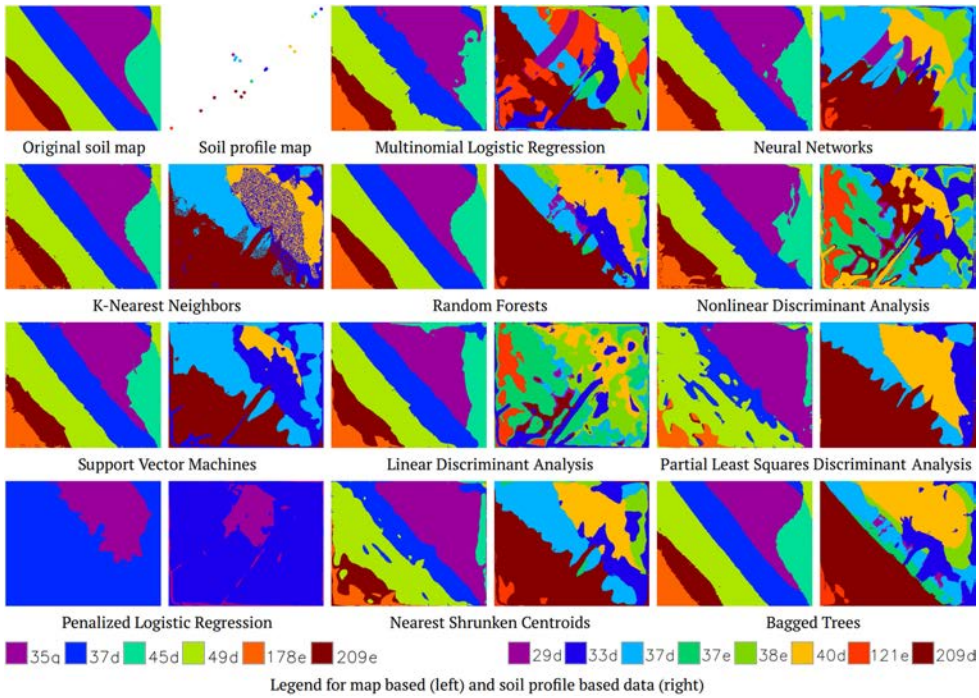

Fig. 5. *Results of mathematical modelling of soil cover of first test area. Signatures are grouped together in pairs of patterns on the basis Model$_{map}$ (left) and Model$_{pit}$ (right). The soil codes are explained in Table 2*

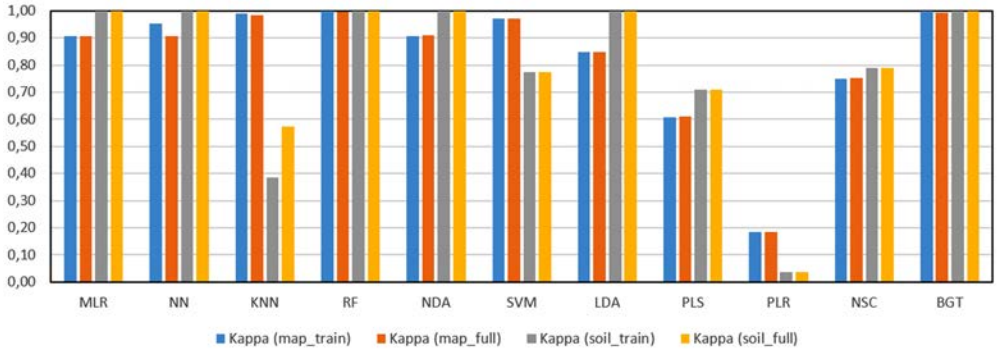

Fig. 6. *The distribution of the value of the index κ depends on the type of simulation model*

Based on the rank of the models in order to increase the quality of the prediction by κ of the main data set, the PLR algorithm was the worst result among others. Following in ascending order placed PLS, NSS, LDA, NDA, MLR, NN, SVM, KNN, RF and BGT. The last two algorithms (RF, BGT) belong to the classifications type, and their high results indicate the greatest suitability of this kind of approaches in mapping the soil cover on the basis of the cartographic sampled dataset. It should be noted that a number of algorithms have crossed the prediction quality limit of 90% (MLR, NN, SVM, KNN, RF and BGT), of which the last two (RF and BGT) practically reproduce the original soil map (Fig. 5). This confirms the previous research in Cherlinka (2017c) who demonstrated that a randomized-weighted sample of learning dataset is optimal for problems of reproduction (re-creation) and simulation of soil cover maps. This is true if we use only cartographic data as input. In other cases, as Horáček M. et al. (2018) show, other approaches may be successful.

The obtained results showed the kappa index of most of our map data based models exceeds the averaged values reported in other similar studies. For example, Hengl (2009) considered 51-67% a good level of modelling quality. Grinand et al. (2008) obtained κ = 67-87% for the training dataset and about 30% – for the main dataset. For small-scale soil maps Giasson et al. (2008) received the following values κ 37-54%. Malone et al. (2016) reported 35-40% of the kappa index. According to the ranges given Landis and Koch (1977), the PLR model reached the worst kappa – a slight strength of agreement (κ = 0.01-0.20), PLS, NSC – substantial strength (κ = 0.61-0.80), and at the best ones (LDA, NDA, MLR, NN, SVM, KNN, RF and BGT) – almost perfect strength of agreement (κ = 0.81-0.99). Accordingly, the evaluation of the quality of maps based on simulation can follow being above the levels reported in the above mentioned literature. In addition, we believe that there is still some potential for increasing the total value of κ, in particular through a more thorough selection of predictors of the model. For example, fuzzy c-means clustering analyses for selecting predictors as shown Horáček et al. (2018) and the expansion of their number by incorporating remote sensing data, anthropogenic deposits map and more.

A significant beneficial effect of this kind of modelling is the ability to fill gaps on existing cartographic materials with data from prediction map-versions and, thus, obtaining maps of continuous soil cover. It does not detract from the value of the conclusions on the ranking of predictive algorithms and the choice of how to obtain a training dataset. In the case of analysing correct cartographic materials with gaps in the soil data, proposed approach allows to obtain statistically reliability data. In the absence of field surveys in certain areas, these data can be used to solve applied problems of soil science, agronomy, land management, cadastre etc.

To check the previous conclusion, we conducted simulations for a test site 2 that is characterized by large gaps in cartographic soil information and conducted additional field survey of soils. The obtained forecast soil map using the algorithm of Random Forests has a kappa index equal 0.867, which means 86.7% of the coincidence of predicted and mapped soils (Fig. 7).

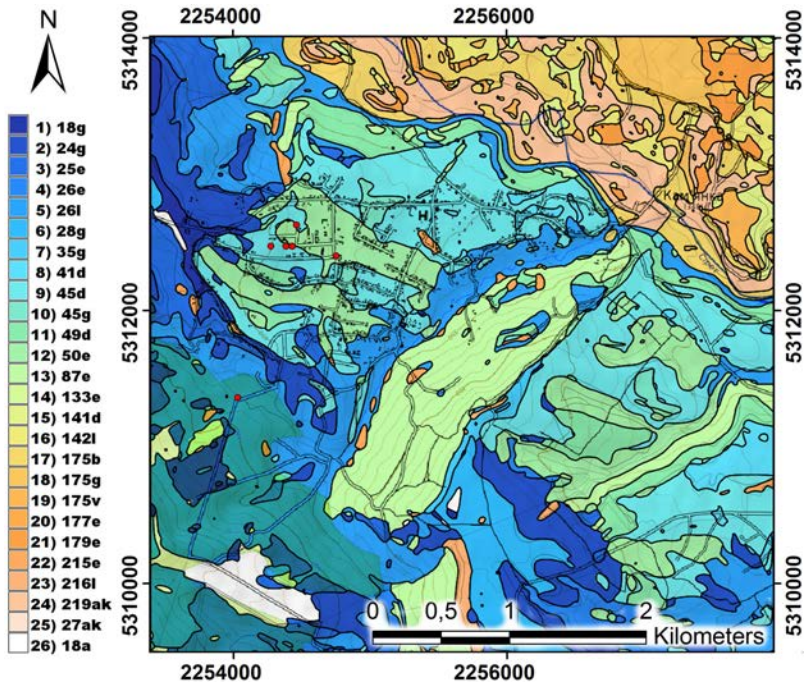

Fig. 7. *Predicted map of soils of the test site 2 using the Random Forest algorithm. Red dots are the locations of the soil pits*

To reaffirm assume that in the gaps on the map, this forecast has a similar reliability, we laid a series of soils pit that allowed a direct verification according to option V4. Results are encouraging, since 5 out of 6 soil pits coincided with the simulated soil meaning that the reliability is high. The high coincidence of the real and simulated soils indirectly testifies to the high quality of the soil map of the second test area, although it was created 44 years ago.

Note that the average simulation time using script R for test area 1 was about 52 minutes and approximately 3 hours for research area 2 on a test platform with an operating system Debian GNU/Linux 9 (Stretch with the kernel 4.9.0-3-amd64 x86_64) and a processor Intel Core I7-5700HQ CPU@3.50GHz (16 Gb RAM). Such computing time expenditures are small and allow you to analyse and simulate, if necessary, much larger areas than in this study.

Modelling based on soil pits data

Equally interesting is the detailed analysis $Model_{pit}$. As it was already mentioned, the structure of the source data, in this case, was not designed for the purpose of large-scale soil mapping, but for the verification of soil boundaries and the accuracy of the definition of soil types on a map. This fact imposed certain limitations on the quality of the expected results, since the points of laying soil pits are not typical points of the area (the "keys" in the terminology of soil mapping).

The use of 11 prediction algorithms allowed obtaining a series of forecast maps (Fig. 5), which generally allow for a conclusion that only 18 soil pits with extremely non-optimal allocation are insufficient to simulate the soil cover for such a territory. Hence, on average, one soil pit spatially supports 9.28 ha, which is insignificant for this category of complexity of the landscape. The high values of Cohen's κ (Fig. 6) should not be misleading: 6 algorithms out of 11 showed that there are 100 % matches of training and complete data. In fact, this only happened because the sampling itself consists of only 18 pits, and in such a small amount of data this coincidence is not indicative. Of course, even with such data, we can make certain conclusions about the quality of the algorithms. Steadily high as in the previous model, 100% of kappa was reached by RF and BGT. A 100% match for the MLR, NN, NDA, and LDA algorithms was achieved. Relatively high kappa parameters are in PLS and NSC, declined slightly in SVM, and in PLR and KNN, they dropped sharply.

Visual inspection of the resulting simulations provides a more interesting aspect. Given certain experience from previous soil studies, only RF and BGT are the most similar to real from the maps obtained by modelling based on field data. Obviously, cross-validation by means the kappa (in cases where it is 100%) will be formally valid, but we would be careful to recommend such predictive maps for use.

It has been established that in order to cover all possible cross-validation options for a field survey-based, map-based and purely model data, certain conditions need to be met. In particular, these are the following conditions: 1) the number of planned soil pit should meet the objectives of large-scale soil survey; 2) the scheme of their location should correspond to the key points of the studied area, which are allocated on the basis of geomorphological analysis. This enables to get the most reliable and correct results. The use of soil maps as a source of information to fill the gaps in research with predictive data is possible only with the high quality of these maps, otherwise the maps errors will be present in the prediction models.

Conclusions

The presented results demonstrated that there are 11 possible ways of verifying real, mapped and modelled soils, which can be divided into three groups depending on the type of input information: polygon data (soil maps), point data (reliable field survey of soils) and their combination. This data grouping provides possibilities for constructing the corresponding three sets of prediction models that include 11 types of basic predictive algorithms. Despite the relative simplicity of our scheme of cross-validation variants, it allows us to fully appreciate the quality of the work performed. This concerns soil surveys in field conditions, existing map data and modelling itself. Depending on the purpose, you can choose a specific set of cross-validation combinations.

An analysis of the soil cover simulations from the point of view of their correspondence to the original soil map, permit us suggests some algorithms for the creation of prediction maps that will be contained forecast soil in areas with missing information with the results of a certain level of statistical reliability. The performance of the output soil models based on the Cohen's κ can be ordered from the lowest to the highest kappa as follows: PLR, PLS, NSS, LDA, NDA, MLR, NN, SVM, KNN, RF and BGT. The high results of the last two algorithms indicate that they are most suitable for soil prediction based on the cartographic training dataset. Randomized-weighted sample of training data was found to be optimal for improving reproduction and modelling of soil cover maps.

It is shown that in modelling the soil cover the location points of field observations should be optimal and correspond to the method of such surveys in order to obtain cartographic results that are acceptable in soil science. Such a conclusion can be made, despite the formally high values of Cohen's κ. This was expected, and due, in the first place, to the orientation of the scheme of pits for verifying the existing agro-industrial soil map.

These results clearly outline the scope and technical characteristics of future research: 1) the number of soil pits should correspond to the plan of large-scale soil survey; 2) the scheme of soil pits allocations should correspond to the key points of the area set on the basis DEM and detailed geomorphological analysis; 3) the use of soil maps as a source of information to fill the gaps in research with predictive data is possible only with the high quality of these maps, otherwise the maps errors will be present in the models. Only in such a case, it is advisable to apply the full range of possible cross-validations of field, cartographic and model data and obtaining the most reliable and correct results from the mathematical and soil science side.

References

ARGE, L., CHASE, J. S., HALPIN, P. et al. 2003: Efficient flow computation on massive grid terrain datasets. *GeoInformatica,* 7(4), 283-313. DOI: https://doi.org/10.1023/A:1025526421410.

BIAŁOUSZ, S., MARCINEK, J., STUCZYŃSKI, T., TURSKI, R. 2005: Soil survey, soil monitoring and soil database in Poland. In Jones, R. J. A., Houšková, B., Bullock, P., Montanarella, L. eds. *Soil Resources of Europe (2nd edition).* Luxembourg (Office for Official Publications of the European Communities), pp. 263-273.

BREIMAN, L. 2001: Random forests. *Machine learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324.

BROWNING, D. M., DUNIWAY, M. C. 2011: 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. *Applied and Environmental Soil Science*, 2011, 1-12. DOI: https://doi.org/10.1155/2011/421904.

BRUNGARD, C. W., BOETTINGER, J. L., DUNIWAY, M. C. et al. 2015: Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*, 239-240, 68-83. DOI: https://doi.org/10.1016/j.geoderma.2014.09.019.

BUI, E. N., MORAN, C. J. 2003: A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray-Darling basin of Australia. *Geoderma*, 111(1), 21-44. DOI: https://doi.org/10.1016/s0016-7061(02)00238-0.

CAMPLING, P., GOBIN, A., FEYEN, J. 2002: Logistic modelling to spatially predict the probability of soil drainage classes. *Soil Science Society of America Journal*, 66(4), 1390-1401. DOI: https://doi.org/10.2136/sssaj2002.1390.

CATEN, A. T., DALMOLIN, R. S. D., PEDRON, F. D. A., RUIZ, L. F. C., SILVA, C. A. D. 2013: An appropriate data set size for digital soil mapping in Erechim, Rio Grande do Sul, Brazil. *Revista Brasileira de Ciência do Solo*, 37(2), 359-366. DOI: https://doi.org/10.1590/s0100-06832013000200007.

CHERLINKA, V. R. 2015: Adaptation large-scale maps of soils to their practical use in GIS. *Agrochemistry and Soil Science*, 84, 20-28. [ЧЕРЛІНКА В. Р. 2015: Адаптація великомасштабних карт ґрунтів до їх практичного використання у ГІС. *Агрохімія і ґрунтознавство*, 84, 20-28].

CHERLINKA, V. 2017a: 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*. Switzerland (Springer-Verlag GmbH, Cham), pp. 89-100. DOI: https://doi.org/10.1007/978-3-319-45417-7_7.

CHERLINKA, V. R. 2017b: Morphometric parameters of relief as basis for predictive modeling of spatial distribution of soil cover. *Agrochemistry and Soil Science*, 86, 5-16. [ЧЕРЛІНКА, В. Р. 2017b: Морфометричні параметри рельєфу як базис для предикативного моделювання просторового поширення ґрунтових відмін. *Агрохімія та ґрунтознавство*, 86, 5-16].

CHERLINKA, V. R. 2017c: Variations in the predictive efficiency of soil maps depending on the methods of constructing training samples of predicative algorithms. *Ecology and Noospherology*, 28(3-4), 55-71. DOI: https://doi.org/10.15421/031716. [ЧЕРЛІНКА, В. Р. 2017c: Варіації прогнозної ефективності ґрунтових карт залежно від способів побудови навчальних вибірок предикативних алгоритмів. *Екологія та ноосферологія*, 28(3-4), 55-71].

CHERLINKA, V. R. 2017d: Influence of resolution of digital elevation models on the quality of predicative simulation of soil cover. *Gruntoznavstvo,* 18(1-2), 79-95. [ЧЕРЛІНКА, В. Р. 2017d: Вплив роздільної здатності цифрових моделей рельєфу на якість предикативної симуляції ґрунтового покриву. *Ґрунтознавство*, 18(1-2), 79-95].

CHERLINKA, V. R., DMYTRUK, Y. M. 2014: Problem in creating, georectifications and using of large-scale digital elevation models. *Geopolitics and ecogeodynamics of regions* 10(1), 239-244. [ЧЕРЛІНКА, В. Р., ДМІТРУК, Ю. М. 2014: Проблеми створення, георектифікації та використання крупномасштабних цифрових моделей рельєфу. *Геополитика и экогеодинамика регионов*, 10(1), 239-244].

CONGALTON, R. G., GREEN, K. 2008: *Assessing the accuracy of remotely sensed data: principles and practices (2nd edition)*. Boca Raton, FL (CRC press). DOI: https://doi.org/10.1201/9781420055139.

CUTLER A., CUTLER D. R., STEVENS J. R. 2012: Random Forests. In Zhang, C., Ma, Y. eds. *Ensemble Machine Learning*. Boston, MA (Springer), pp. 157-175. DOI: https://doi.org/10.1007/978-1-4419-9326-7_5.

DEBELLA-GILO, M., ETZELMÜLLER, B. 2009: Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modelling integrated in GIS: Examples from Vestfold County, Norway. *Catena,* 77(1), 8-18. DOI: https://doi.org/10.1016/j.catena.2008.12.001.

DOBOS, E., HENGL, T. 2009. Soil mapping applications. In Hengl, T., Reuter, H. I. eds. *Developments in Soil Science*. Amsterdam (Elsevier), pp. 461-479. DOI: https://doi.org/10.1016/s0166-2481(08)00020-2.

DŽATKO, M. 1974: Metodika a prax vyčleňovania pôdno-ekologických jednotiek. *Studijni informace,* 74(5), 1-30.

EASYTRACE GROUP 2015: *Easy Trace 7.99. Digitizing software.* Ryazan (Easytrace) Retrieved from: http://www.easytrace.com.

EVENDEN, G., WARMERDAM, F. 1990: *Proj. 4 – Cartographic Projections Library. Source code and documentation*. Chicago (Open Source Geospatial Foundation). Retrieved from: http://trac.osgeo.org/proj.

FENG, C., MICHIE, D. 1994: Machine learning of rules and trees. In Michie, D., Spiegelhalter, D. J., Taylor, C. C. eds. *Machine learning, neural and statistical classification*. Chichester (Ellis Horwood Limited), pp. 50-83.

FLORINSKY, I. V. 2012: *Digital Terrain Analysis in Soil Science and Geology*. Amsterdam (Academic Press/Elsevier). DOI: https://doi.org/10.1016/c2010-0-65718-x.

FOODY, G. M. 2004: Thematic map comparison: Evaluating the Statistical Significance of Differences in Classification Accuracy. *Photogrammetric Engineering & Remote Sensing*, 70(5), 627-633. DOI: https://doi.org/10.14358/PERS.70.5.627.

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*. Netherlands, Amsterdam (Springer), pp. 225-232. DOI: https://doi.org/10.1007/978-1-4020-8592-5_19.

GRASS DEVELOPMENT TEAM 2017: *Geographic Resources Analysis Support System (GRASS GIS) Software. Version 7.2*. Retrieved from: http://grass.osgeo.org.

GRINAND, C., ARROUAYS, D., LAROCHE, B., MARTIN, M. P. 2008: Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. *Geoderma,* 143(1), 180-190. DOI: https://doi.org/10.1016/j.geoderma.2007.11.004.

HAIR, J. F., BLACK, W. C., BABIN, B. J. ANDERSON, R. E. 2010: *Multivariate data analysis (7th edition)*. Prentice Hall, Upper Saddle River (Pearson Education).

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. 2009: *Springer Series in Statistics: The elements of statistical learning: Data mining, inference, and prediction (2nd edition).* New York (Springer). DOI: https://doi.org/10.1007/978-0-387-84858-7.

HENGL, T. 2009: *A practical guide to geostatistical mapping (2nd edition)*. Luxembourg (Office for Official Publications of the European Communities).

HENGL, T., DE JESUS, J. M., HEUVELINK, G. B. et al. 2017: SoilGrids250m: Global gridded soil information based on machine learning. *PLoS one*, 12(2), DOI: https://doi.org/10.1371/journal.pone.0169748.

HENGL, T., ROSSITER, D. G., STEIN, A. 2003: Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Australian Journal of Soil Research,* 41(8), 1403-1422. DOI: https://doi.org/10.1071/SR03005.

HEUNG, B., HO, H. C., ZHANG, J., KNUDBY, A., BULMER, C. E., SCHMIDT, M. G. 2016: An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma,* 265, 62-77. DOI: https://doi.org/10.1016/j.geoderma.2015.11.014.

HEUNG, B., HODÚL, M., SCHMIDT, M. G. 2017: Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes. *Geoderma*, 290, 51-68. DOI: https://doi.org/10.1016/j.geoderma.2016.12.001.

HILBE, J. 2009: *Logistic Regression Models – Texts in Statistical Science.* Boca Raton (Chapman & Hall/CRC, Taylor & Francis Group).

HOFIERKA, J., CEBECAUER, T., ŠÚRI, M. 2007: Optimisation of interpolation parameters using cross-validation. In Peckham, R. J., Jordan, G. eds. *Digital Terrain Modelling – Lecture Notes in Geoinformation and Cartography.* Berlin Heidelberg (Springer), pp. 67-82. DOI: http://dx.doi.org/10.1007/978-3-540-36731-4_3.

HOFIERKA, J., MITÁŠOVÁ, H., NETELER, M. 2009: Geomorphometry in GRASS GIS. In Hengl, T., Reuter, H. I. eds. *Developments in Soil Science.* Amsterdam (Elsevier), pp. 387-410. DOI: https://doi.org/10.1016/S0166-2481(08)00017-2.

HOFIERKA, J., ŠÚRI, M. 2002: The solar radiation model for Open source GIS: implementation and applications. In Ciolli, M., Zatelli, P. eds. *Proceedings of the Open source GIS-GRASS users conference.* Trento, Italy (Dipartimento di Ingegneria Civile e Ambientale, Universita' di Trento), pp. 1-19.

HORÁČEK, M., SAMEC, P., MINÁR, J. 2018: The mapping of soil taxonomic units via fuzzy clustering – A case study from the Outer Carpathians, Czechia. *Geoderma*, 326, 111-122. DOI: https://doi.org/10.1016/j.geoderma.2018.04.012.

HUBERTY, C. J., OLEJNIK, S. 2006: *Applied MANOVA and Discriminant Analysis (2nd edition)*. Hoboken, New Jersey (John Wiley & Sons). DOI: https://onlinelibrary.wiley.com/doi/book/10.1002/047178947X.

JASIEWICZ, J., METZ, M. 2011: A new GRASS GIS toolkit for hortonian analysis of drainage networks. *Computers & Geosciences,* 37(8), 1162-1173. DOI: https://doi.org/10.1016/j.cageo.2011.03.003.

JASIEWICZ, J., STEPINSKI, T. F. 2013: Geomorphons – a pattern recognition approach to classification and mapping of landforms. *Geomorphology,* 182, 147-156. DOI: https://doi.org/10.1016/j.geomorph.2012.11.005.

JONES, R. J. A., HOUŠKOVÁ, B., BULLOCK, P., MONTANARELLA, L. 2005: *Soil resources of Europe (2nd edition)*. Luxembourg (Office for Official Publications of the European Communities).

KEMPEN, B., BRUS, D. J., HEUVELINK, G. B. M., STOORVOGEL, J. J. 2009: Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. *Geoderma,* 151(3), 311-326. DOI: https://doi.org/10.1016/j.geoderma.2009.04.023.

KESKIN, H., GRUNWALD, S. 2018: Regression kriging as a workhorse in the digital soil mapper's toolbox. *Geoderma,* 326, 22-41. DOI: https://doi.org/10.1016/j.geoderma.2018.-04.004.

KUHN, M. 2008: Building Predictive Models in R Using the caret Package. *Journal of Statistical Software,* 28(5), 1-26. DOI: https://doi.org/10.18637/jss.v028.i05.

KUHN, M., JOHNSON, K. 2013: *Applied Predictive Modeling*. New York (Springer). DOI: https://doi.org/10.1007/978-1-4614-6849-3.

LAGACHERIE, P., ROBBEZ-MASSON, J. M., NGUYEN-THE, N., BARTHÈS, J. P. 2001: Mapping of reference area representativity using a mathematical soilscape distance. *Geoderma*, 101(3-4), 105-118. DOI: https://doi.org/10.1016/s0016-7061(00)00101-4.

LANDIS, J. R., KOCH, G. G. 1977: The measurement of observer agreement for categorical data. *Biometrics* 33(1), 159-174. DOI: https://doi.org/10.2307/2529310.

LI, W., ZHANG, C. 2007: A Random-Path Markov Chain Algorithm for Simulating Categorical Soil Variables from Random Point Samples. *Soil Science Society of America Journal*, 71(3), 656-668. DOI: https://doi.org/10.2136/sssaj2006.0173.

LIU, B. 2011: *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data (2nd edition)*. Heidelberg, Dordrecht, London, New York (Springer-Verlag GmbH). DOI: https://doi.org/10.1007/978-3-642-19460-3.

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*. Amsterdam (Springer Netherlands), pp. 113-135. DOI: https://doi.org/10.1007/978-1-4020-8592-5_10.

MALONE, B. P., MINASNY, B., MCBRATNEY, A. B. 2016: *Using R for Digital Soil Mapping*. Progress in Soil Science. Switzerland (Springer International Publishing). DOI: https://doi.org/10.1007/978-3-319-44327-0.

MARQUES, K. P., DEMATTÊ, J. A., MILLER, B. A., LEPSCH, I. F. 2018: Geomorphometric segmentation of complex slope elements for detailed digital soil mapping in southeast Brazil. *Geoderma Regional*, 14, 1-9. DOI: https://doi.org/10.1016/j.geodrs.2018.e00175.

MCBRATNEY, A. B., SANTOS, M. L. M., MINASNY, B. 2003: On digital soil mapping. *Geoderma,* 117(1-2), 3-52. DOI: https://doi.org/10.1016/s0016-7061(03)00223-4.

MCKAY, M. D., BECKMAN, R. J., CONOVER, W. J. 2000: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics,* 42(1), 55-61. DOI: http://dx.doi.org/10.1080/00401706.2000.10485979.

MIKLASZEWSKI, S. 1907: *Gleby ziem polskich + mapa 1:1 500 000*. Warszawa (Wyd. Gebethnter i Wolf).

MINÁR, J. 2003: Detailed physical-geographical (geoecological) research and mapping in the landscape ecology. *Ekologia (Bratislava)/Ecology (Bratislava)*, 22(2), 141-149.

MITÁŠOVÁ, H., HOFIERKA, J. 1993: Interpolation by regularized spline with tension: II. Application to terrain modeling and surface geometry analysis. *Mathematical Geology*, 25(6), 657-669. DOI: https://doi.org/10.1007/BF00893172.

MITÁŠOVÁ, H., MITÁŠ, L. 1993: Interpolation by regularized spline with tension: I. Theory and implementation. *Mathematical Geology,* 25(6), 641-655. DOI: https://doi.org/10.1007/BF00893171.

MOORE, I. D., GESSLER, P. E., NIELSEN, G. A., PETERSON, G. A. 1993: Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal*, 57(2), 443-452. DOI: https://doi.org/10.2136/sssaj1993.572npb.

PETERS, A., HOTHORN, T., HOTHORN, M. T. 2009: Package 'ipred'. 0.8-7. *The R Foundation for Statistical Computing.* Retrieved from: https://CRAN.R-project.org/pack-age=ipred.

QGIS DEVELOPMENT TEAM 2015: *QGIS Geographic Information System.* Retrieved from: https://qgis.osgeo.org.

R DEVELOPMENT CORE TEAM 2017: R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Retrieved from: https://www.r-project.org.

RIPLEY, B., VENABLES, W. 2016: *R-package nnet: Feed-forward neural networks and multinomial log-linear models – version 7.3-12*. Retrieved from: https://cran.r-project.org/package=nnet.

SCULL, P., FRANKLIN, J., CHADWICK, O. A., MCARTHUR, D. 2003: Predictive soil mapping: a review. *Progress in Physical Geography,* 27(2), 171-197. DOI: https://doi.org/10.1191/0309133303pp366ra.

STEERS, C. A., HAJEK, B. F. 1979: Determination of map unit composition by a random selection of transects. *Soil Science Society of America Journal,* 43(1), 156-160. DOI: https://doi.org/10.2136/sssaj1979.03615995004300010030x.

TENG, H., ROSSEL, R. A. V., SHI, Z., BEHRENS, T. 2018: Updating a national soil classification with spectroscopic predictions and digital soil mapping. *Catena*, 164, 125-134. DOI: https://doi.org/10.1016/j.catena.2018.01.015.

VENABLES, W. N., RIPLEY, B. D. 2002: *Statistics and Computing: Modern Applied Statistics with S (4th edition)*. New York (Springer-Verlag). DOI: https://dx.doi.org/10.1007/978-0-387-21706-2.

WALTER, C. 1990: *Estimation de propriétés du sol et quantification de leur variabilité à moyenne échelle: cartographie pédologique et géostatistique sur un secteur du sud de l'ille-et-vilaine – PhD thesis*. Rennes (Institut National de la Recherche agronomique). DOI: https://doi.org/10.13140/RG.2.1.2794.640.

WALTER, C., LAGACHERIE, P., FOLLAIN, S. 2006: Integrating pedological knowledge into digital soil mapping. In Lagacherie, P., McBratney, A. B., Voltz, M. eds. *Developments in Soil Science.* Amsterdam (Elsevier), pp. 281-301. DOI: https://doi.org/10.1016/s0166-2481-(06)31022-7.

WHITE, R. E. 2006: *Principles and practice of Soil science: The Soil as a natural resource (4th edition)*. Oxford, UK (Blackwell Publishing). DOI: https://doi.org/10.1017/S0014479-706303791.

WOOD, J. D. 1996: *The geomorphological characterisation of digital elevation models – PhD thesis*. UK, Leicester (University of Leicester).

WRIGHT, R. L., WILSON, S. R. 1979: On the analysis of soil variability, with an example from Spain. *Geoderma,* 22(4), 297-313. DOI: https://doi.org/10.1016/00167061(79)900-26-0.

Acknowledgement: This work (programming and mathematical modelling) was partially funded by 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 [2016/2017:id17680]. 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). We also acknowledge the project VEGA 1/0963/17: "Landscape dynamics in high resolution" for financial support funded by the Ministry of Education, Science, Research and Sport of the Slovak Republic. The authors thank the two anonymous reviewers for their valuable comments. Also, authors gratefully thank Michal Gallay for the constructive recommendations and English corrections.

Authors´ affiliations

Doc. Vasyl Cherlinka, D.Sc.
Yuriy Fedkovych Chernivtsi National University
Institute of Biology, Chemistry and Bioresources
Department of Agrotechnologies and Soil Science
Lesya Ukrainka Str. 25
58012 Chernivtsi
Ukraine
v.cherlinka@chnu.edu.ua

Prof. Yuriy Dmytruk, D.Sc.
Yuriy Fedkovych Chernivtsi National University
Institute of Biology, Chemistry and Bioresources
Department of Agrotechnologies and Soil Science
Lesya Ukrainka Str. 25
58012 Chernivtsi
Ukraine
y.dmytruk@chnu.edu.ua

RNDr. Dušan Barabas, CSc.
Pavol Jozef Šafárik University in Košice
Faculty of Science
Institute of Geography
Jesenná 5
040 01 Košice
Slovakia
dusan.barabas@upjs.sk