Mathematical models application for mapping soils spatial distribution on the example of the farm from the North of Udmurt Republic of Russia

P M Dokuchaev1, J L Meshalkina1,2,3, A M Yaroslavtsev2,4,5

1Soil Science Faculty, Lomonosov Moscow State University, Leninskye Gory, GSP-1, bld.12, Moscow, 119991, Russia
2Department of Ecology, Russian Timiryazev State Agrarian University, Timiryazevskaya Str. 49, Moscow, 127550, Russia
3Dokuchaev Soil Science Institute, Pyzhyovskiy lane 7 building 2, Moscow, 119017, Russia
4RUDN University, Miklukho-Maklaya str.6, Moscow, 117198, Russia
5School of Natural Sciences, Far Eastern Federal University, Sukhanova St. 8, Vladivostok, 690090, Russia
E-mail: jlmesh@list.ru

Abstract. Comparative analysis of soils geospatial modeling using multinomial logistic regression, decision trees, random forest, regression trees and support vector machines algorithms was conducted. The visual interpretation of the digital maps obtained and their comparison with the existing map, as well as the quantitative assessment of the individual soil groups detection overall accuracy and of the models kappa showed that multiple logistic regression, support vector method, and random forest models application with spatial prediction of the conditional soil groups distribution can be reliably used for mapping of the study area. It has shown the most accurate detection for sod-podzolics soils (Phaeozems Albic) lightly eroded and moderately eroded soils. In second place, according to the mean overall accuracy of the prediction, there are sod-podzolics soils – non-eroded and warp one, as well as sod-gley soils (Umbrisols Gleyic) and alluvial soils (Fluvisols Dystric, Umbric). Heavy eroded sod-podzolics and gray forest soils (Phaeozems Albic) were detected by methods of automatic classification worst of all.

1. Introduction
Soil map is the main scientific document on the basis of which competent evaluation of land funds is possible, as well as the development of a system of practical measures for increasing soil fertility. Renewal of legacy soil maps with the actualization of soil cover state and properties is one of the most important requirements for conducting smart agriculture. At the same time, carrying out large-scale soil mapping by traditional means to the whole territory of Russia is not possible because of the high labor costs for field research and the limited number of specialists. Digital soil mapping (DSM) can be one of solutions to this problem.

At the heart of the DSM is the development of Dokuchaev’s ideas, which are expressed in the model SCORPAN [1]. This model considers the classification of soils and soil properties as a function of the soil forming factors and / or indicators, which in the DSM are called covariates. Information on covariates can be obtained from various sources: for example, remote sensing data [2], old soil maps
and survey materials [3, 4], digital elevation models [5, 6], data obtained by proximal sensing with measuring electrical, gravitational, magnetic, and electromagnetic fields [7, 8], etc.

Mapping of soils spatial distribution in DSM is realized utilizing algorithms specially developed for these purposes or used for the mathematical solution of many other problems. Two algorithms: multinominal logistic regression (MLR) [9, 10] and support vector machine (SVM) [11, 12], which create the probability vector finding a certain class of soil [9] for each point of the raster, are based on the predictor variables. They differ in functions used to construct those probability vectors. Three other algorithms: decision trees (C5) [10], random forests (RF) [10] and regression trees (RT) [13, 14], are based on so-called decision trees. In each node of such trees, classifying rules are formed, according to which the decision tree assigns a point to a class. The main difference between algorithms is the number of decision trees that are being built: for the decision tree algorithm, only one tree is constructed, and for other two algorithms - numerous.

The mapping process can be described as follows. First, a training set is formed, where for each known point on the Earth's surface soil classification is indicated, as well as the values of soil forming factors and covariates. Based on the training sample, a cartographic model is constructed for the study area. The cartographic model can be expressed in number of decision rules or in the form of coefficients in the equation estimates, depending on the chosen classification algorithm that specifies the spatial structure of a multidimensional space. Then, the diagnostics of soil classification is carried out throughout the study area using the obtained cartographic model, based on data of soil forming factors and indicators.

The main DSM feature is the ability to check sustainability of cartographic models obtained for the training set, and test the accuracy of predictions for an independent random sample. As a result, each map obtained by the DSM is characterized by the degree of uncertainty in the prediction.

The aim of the study was the construction of large-scale digital soil maps using five mathematical models, as well as comparison of digital soil maps with the existing large-scale soil map performed by UralGyprozem staff in 1984 (further called just “1984” map). The research tasks also included: the selection of the optimal pixel size of the image raster for modeling and the verification of stability of a model.

2. Material and methods
The key site is located 8 km to South-West from town Glazov, which is located on the north of Udmurt Republic of Russia. Soil-climatic conditions there are typical for the eastern part of the Eastern European Plain. The climate is moderately continental; the precipitation-evaporation ratio is 1.00-1.33. Vegetation is represented mainly by fir-spruce and pine forests with a significant admixture of birch.

The length of the site from north to south is 4.5 km, and from west to east - 5.2 km (figure 1). The total area is about 19 square km. Maximal elevation difference in the study area is 62 m. The territory is rugged with ravines. Forests occupy 24% of the investigated area and are located along the slopes of deep ravines. There are sod-gley soils (Umbrisols Gleyic) on the ravine slopes. The river Kypka flows from the North-East to the South-West of the study area. Accordingly, floodplain meadows with alluvial soils (Fluvisol Dystric, Umbric) occupy about 5% of the area. Anthropogenic load is significant here: 65% of the territory is occupied by arable land, 5% are hayfields, 6% are household plots of the village of Kypka. Most of the sod-podzolic soils (Albeluvisol Umbric) on arable land are represented by eroded categories. Small flat areas in the lower part of the slopes can be occupied by gray forest soils (Phaeozems Albic). All soils of the area under consideration are with medium- and heavy-loamy soil texture.

On the basis of a topographic map with a scale of 1: 50000, a DEM with a 5 meter resolution was built. Based on the DEM, 35 maps of the morphometric relief variables were constructed, such as slope exposition, steepness of slopes, catchment area, flow power index, topographic index, horizontal curvature, vertical curvature, etc. [6].
During the field survey of 2014-2015, there were 166 survey points laid (figure 1), with the full description of the soil profile. The survey points covered the entire territory and were laid considering mapping of all elementary landscapes positions of the investigated territory.

![Study area](Image)

**Figure 1.** Study area (scale 1:50 000). Legend 1. Ravines. 2. Settlement. 3. Forests. 4. Arable land. 5. Flood plain and water. 6. Sampling points.

The database consisted of 166 profiles, and during simulation procedures it was necessary that for each classification group there should be at least 10 observations, therefore, 8 large soil groups were selected for digital soil mapping: sod-podzolics non-eroded soils, sod-podzolics lightly eroded soils, sod-podzolics moderately eroded soils, sod-podzolics heavy eroded soils, sod-podzolics wrap soils, gray forest soils, sod-gley soils and alluvial soils. Such a set of conditional soil groups allowed to reflect all presented types of soils and distribution of eroded soils in the study area. The latter is important for the proper planning of agricultural activities.

One-way analysis of variances was used for covariates selection: soil groups acted as a factor, and responses were covariates. The input data for modeling were the following covariates: 25 out of 35 morphometric variables, land use, NDVI index.

Modeling of soils spatial distribution was based on five algorithms: multiple logistic regression, decision trees C5, random forest, regression trees and support vector machines. Models stability was estimated using cross-validation. In this method, a subset of observations is divided randomly into two subsets: the "training" sample and the "test" sample. The model is built using data from training sample, whereas the data from the second sample are used to assess the model quality and stability. For each algorithm, 100 iterations were carried out. As a result, for each algorithm 100 digital maps were constructed. To produce each map, 146 points (88%) of the points were selected randomly from the training sample, and the remaining 20 points (12%) were used as testing one.

For each test sample, a confusion matrix was made between the actual soil groups identified during the field studies and the soil groups predicted for these points as a result of modeling. The sum of the diagonal elements values ($X_{ii}$) reflects the total number of identically classified points, and the ratio of this quantity to the total number of points (N) is called overall accuracy [15, 16]:

$$A_o = \frac{\sum_{i=1}^N X_{ii}}{N} = \sum_{i=1}^N P_{ii};$$

$$P_{ii} = \frac{X_{ii}}{N}$$

(1)

The accuracy of individual classes prediction is calculated similarly. Traditionally, they are divided by the total number of class predictions. This measure of prediction accuracy indicates the likelihood
that the observation was classified correctly and called "producer's accuracy", because the model creator is interested in how well a particular class can be predicted.

The kappa coefficient or simply kappa is another statistical measure of the coincidence between the observed and predicted classes. The calculation is based on the difference between how many coincidences are present ("observed" coincidence is the percentage of the overall prediction accuracy), it is possible to expect a coincidence that has turned out just randomly.

We denote this by $\theta_2$:

$$\theta_2 = \sum_{i=1}^{r} \frac{x_{ij} - x_{ij}}{N^2},$$  \hspace{1cm} (2)

The index kappa is defined as:

$$\hat{\kappa} = \frac{\theta_1 - \theta_2}{1 - \theta_2}$$  \hspace{1cm} (3)

The kappa is equal to 1 when the two maps coincide or when the results of the field survey and the soil map are completely identical. The kappa is zero if $\theta_1$ coincides with $\theta_2$ and a purely coincidental match of the two maps is observed or if the soil map coincides with the results of the field survey. In most cases, the kappa lies from 0 to 1 and shows how similar maps between each other or how close the soil map and survey results.

All indicators produced as a result of modeling iterations were collected in a table in which for cross-validation: kappa coefficient and overall accuracy [15, 16]. For each model, the mean values, medians, standard deviations and the boundaries of the above characteristics were calculated. For each model, two specific implementations of models were selected for further investigation: with average and with maximum value of the kappa. To select the optimal pixel size, curves were plotted showing the ranking of the kappa model depending from pixel size.

![Figure 2. Soil map of 1984 (scale 1:50 000). Conditional soil groups: 1. Alluvial soils (Fluvisols Dystric, Umbric), 2. Sod-gley soils (Umbrisols Gleyic). Sod-podzolics soils (Phaeozems Albic) 3. non-eroded, 4. lightly eroded, 5. moderately eroded, 6. heavy eroded, 7 gray forest soils (Phaeozems Albic).](image)

A comparison of the obtained soil maps with the UralGiprozem map of 1984 was carried out for 200 random points (figure 2). The following maps were compared: the soil map of 1984, made by the specialists of UralGiprozem, and five digital soil maps obtained on the basis of models with an average kappa value. For each pair of maps being compared, the overall accuracy and kappa were calculated. The following software was used in purpose of the investigation: Quantum GIS 2.8.2, SAGA GIS 2.1.2, Multispec and R software environment 3.3.1 [17].
3. Results and discussion

For the selection of the optimal image raster pixel size, a complete simulation was performed for 5 m, 20 m and 50 m pixels. The simulation results for multiple logistic regression are shown in figure 3. For all five algorithms, decreasing pixel size lead to increasing in model’s kappa values. Further results are presented for a pixel size of 5 meter.

![Figure 3. Series of decrease for multiple logistic regression model kappas in cross-validation for different pixel sizes.](image)

Modeling the spatial distribution of soil groups by five algorithms showed (table 1) that the highest average model kappa 0.44 was observed using multiple logistic regression, while the lowest average kappa, equal to 0.26, was detected for model produced with regression trees algorithm. In general, the best prediction was obtained with models produced by multiple logistic regression, support vector machines and random forest methods. Models produced with regression tree algorithm, showed the worst results.

|       | MNLR | C5  | RF  | RT  | SVM |
|-------|------|-----|-----|-----|-----|
| Mean  | 61   | 56  | 59  | 49  | 60  |
| Minimum | 30 | 35  | 35  | 30  | 30  |
| Maximum | 85 | 75  | 85  | 70  | 85  |
| Mean  | 0.44 | 0.32| 0.40| 0.26| 0.41|
| Minimum | 0.06| 0.00| 0.00| 0.00| 0.00|
| Maximum | 0.77| 0.64| 0.75| 0.52| 0.79|

Table 1. Characteristics of the distributions of overall accuracy and kappa for 100 iterations of cross-validation in modeling based on 5 algorithms.

Maps for different kappa values and overall accuracy were analyzed. It turned out that in most cases the best results from the expert's point of view were found for maps with the average kappa and overall accuracy values.

From an expert point of view, digital soil maps based on decision tree models are closest to the map of 1984, compared to other digital maps. High level of resemblance was observed with maps produced by multiple logistic regression and the support vector machines models.

Based on visual comparison, the least degree of similarity of the soil map was observed with a map created with regression trees. However, a quantitative comparison with the map of 1984 showed
satisfactory coincidence with digital maps: the kappa was 0.27-0.28, and the overall accuracy was 53-54%.

The most accurate detection by models was shown for sod-podzolics lightly eroded and moderately eroded soils (table 2). In second place, according to the mean overall accuracy of the prediction, there were sod-podzolics soils non-eroded and wrap one, as well as sod-gley soils and alluvial soils. Heavy eroded sod-podzolics soils and gray forest soils were detected worst of all. The latter were not determined on the average by any algorithm.

Table 2. Average overall accuracy of soil group predictions in modeling based on 5 algorithms.

| Soils                      | MNLR  | C5   | RF   | RT   | SVM  |
|----------------------------|-------|------|------|------|------|
| Sod-podzolics non-eroded   | 71.7  | 26.0 | 43.7 | 68.3 | 46.2 |
| Sod-podzolics lightly eroded | 78.4  | 86.2 | 77.6 | 68.8 | 75.9 |
| Sod-podzolics moderately eroded | 60.9  | 55.3 | 67.1 | 54.7 | 57.6 |
| Sod-podzolics heavy eroded  | 19.1  | 0    | 0    | 0    | 9.1  |
| Sod-podzolics wrap         | 45.4  | 5.2  | 23.5 | 18.6 | 42.1 |
| Alluvial soils             | 95.4  | 97.3 | 91.2 | 11.8 | 100  |
| Gray forest soils          | 0     | 0    | 0    | 0    | 0    |
| Sod-gley soils             | 43.7  | 8.2  | 53.4 | 0    | 63.3 |

In this study, a set of cartographic models for one territory was generated as a result of the DSM algorithms application. The use of mathematical models to describe the spatial distribution of soils is a promising approach in the development of soil cover mapping, which makes it possible to quickly obtain soil maps that are comparable in accuracy to maps constructed by traditional soil survey.

The resulting cartographic materials can form the basis for soil risks assessment and environmental development modeling scenarios.

Acknowledgement
This study has been conducted with support from grant №14-120-14-4266-SeSh and by grant according to the Agreement № 02.A03.21.0008 of the Russian Federation Ministry of Education and Science, grant №ScSh-10347.2016.11 of the Leading Scientific Schools, EU’s FP7 grant agreement №603542 (the project LUC4C) and grant № 15-16-30007 of the Russian Science Foundation.

References
[1] McBratney A B, Mendonca Santos M L and Minasny B 2003 Geoderma 117 pp 3–52
[2] Mulder V L, de Bruin S, Schaeppman M E and Mayr T R 2011 The use of remote sensing in soil and terrain mapping—A review Geoderma 162 pp 1–19
[3] Mayr T, Rivas-Casado M, Bellamy P, Palmer R, Zawadzka J and Corstanje R 2010 Digital Soil Mapping (New York: Springer Science + Business Media BV) pp 191–202
[4] Dobos E, Bialko T, Micheli E and Kobza J 2010 Digital Soil Mapping (New York: Springer Science + Business Media BV) pp 309–323
[5] Shary P A, Sharay L S and Mitusov A V 2002 Geoderma 107 pp 1–43
[6] Florinsky I V 2011 Digital Terrain Modeling in Soil Science and Geology (Amsterdam: Elsevier Academic Press) 432
[7] Huang J, Lark R M, Robinson D A, Lebrón I, Keith A M, Rawlins B, Tye A, Kuras O, Raines M and Triantafilis J 2014 Geoderma 232–234 pp 69-80
[8] Buchanan S and Triantafilis J 2009 Ground Water 47 pp 80–96
[9] Kempen B, Brus D J, Heuvelink G B M and Stoorvogel J J 2009 Geoderma 151 (3–4) 311–326
[10] Brungard CW, Boettinger J, Duniway MC, Wills S A and Edwards J 2015 Geoderma 239–240 pp 68–83
[11] Simin C, Rongqun Z, Liming L and Zhou De 2010 *Math. and Computer Modelling* **51** pp 1319–25
[12] Lamorski K, Pachepsky Y, Slawiński C and Walczak RT 2008 *Soil Sci Soc Am J* **72** pp 1243–47
[13] Lemercier B, Lacoste M, Loun M and Walter C 2012 *Geoderma* **171** pp 75–84
[14] Schmidt K, Behrens T and Scholten T 2008 *Geoderma* **146** (1–2) pp 138–146
[15] Samsonova V P and Meshalkina YuL 2011 *Moscow University Soil Science Bull* **66** (3) pp 83–85
[16] Hengl T, Toormanian N, Reuter H I and Malakouti M J 2007 *Geoderma* **140** pp 417–27
[17] R Core Team 2017. R: A language and environment for statistical computing. (Vienna: R Foundation for Statistical Computing) [http://www.R-project.org/](http://www.R-project.org/)