https://doi.org/10.17221/81/2017-SWR

Digitization and Mapping of National Legacy Soil Data of Montenegro

Edin SALKOVIĆ[1], Igor DJUROVIĆ[1], Mirko KNEŽEVIĆ[2], Vesna POPOVIĆ-BUGARIN[1] and Ana TOPALOVIĆ[2]

[1]Faculty of Electrical Engineering and [2]Biotechnical Faculty, University of Montenegro, Podgorica, Montenegro

**Corresponding author: edins@ac.me*

Abstract

Salković E., Djurović I., Knežević M., Popović-Bugarin V., Topalović A. (2018): Digitization and mapping of national legacy soil data of Montenegro. Soil & Water Res., 13: 83–89.

This paper describes the process of digitizing Montenegro's legacy soil data, and an initial attempt to use it for digital soil mapping (DSM) purposes. The handwritten legacy numerical records of physical and chemical properties for more than 10 000 soil profiles and semi-profiles covering whole Montenegro have been digitized, and, out of those, more than 3000 have been georeferenced. Problems and challenges of digitization addressed in the paper are: processing of non-uniform handwritten numerical records, parsing a complex textual representation of those records, georeferencing the records using digitized (scanned) legacy soil maps, creating a single computer database containing all digitized records, transforming, cleaning and validating the data. For an initial assessment of the suitability of these data for mapping purposes, inverse distance weighting (IDW), ordinary kriging (OK), multiple linear regression (LR), and regression-kriging (RK) interpolation models were applied to create thematic maps of soil phosphorus. The area chosen for mapping is a 400 km^2 area near the city of Cetinje, containing 125 data points. LR and RK models were developed using publicly available digital elevation model (DEM) data and satellite global land survey (GLS) data as predictor variables. The digitized phosphorus quantities were normalized and scaled. The predictor variables were scaled, and principal component analysis was performed. For the best performing RK model an R^2 value of 0.23 was obtained.

Keywords: digital elevation model; georeferencing; kriging; multiple linear regression; parsing; soil phosphorus

From 1959 to 1984 extensive field and laboratory work was performed in Montenegro in order to create a soil map at a 1 : 50 000 scale covering the whole country. Numerical data on chemical and physical properties of soil were collected within a larger project that involved a similar collection of soil properties for former Yugoslavia, as back then Montenegro was one of its member states.

In recent times, due to technological development, there has been significant interest in digitizing these numerical records of soil properties in all countries that were members of Yugoslavia, in order to obtain modern digital soil databases which could, in turn, be used by governments and researchers (Vrščaj *et al.* 2005; Hengl & Husnjak 2006). Arrouays *et al.* (2017) presented a survey of broader international efforts in digitizing legacy soil data.

Although a substantial work about the soils of Montenegro based on these records was published (Fuštić & Đuretić 2000), only non-georeferenced numerical data on 1800 profiles were digitized and presented in it. The authors provided us with their digital version of the data and we managed to georeference the data for almost all profiles. However, we also digitized and georeferenced more than 1200 additional profiles with numerical data that were not presented in the book. In another research project, the legacy maps for Montenegro were scanned and merged into a single raster map, and the ordinal numbers of the profiles that were marked on the

https://doi.org/10.17221/81/2017-SWR

maps were extracted as a digital georeferenced point data set. The same project produced a polygon map of soil classes, based on the raster map, but did not digitize the numerical records.

We have digitized numerical soil data of more than 10 000 soil profiles and semi-profiles for Montenegro. Where possible, we have also georeferenced the data. Besides the numerical data, numerous metadata were inserted that enhance the database. The final produced database is filtered and validated. As such, it represents the basis for digital mapping research of Montenegro's soils.

Several digital soil mapping (DSM) algorithms (McBratney *et al.* 2003) were applied to a subset of digitized data. These algorithms were used to create thematic maps of soil phosphorus for a designated area in Montenegro. The main purpose of the created maps is to provide an initial assessment of the suitability of digitized data for creating interpolated thematic soil maps. Phosphorus is an important nutrient for plants (Schachtman *et al.* 1998). The factors which define the contents of particular phosphorus fractions, the mobility of phosphorus, and its availability to the biota are the main interest of agrochemical and eco-chemical research (Topalović *et al.* 2006). Digital mapping of soil phosphorus is an active field of research (Wang *et al.* 2009; Xiao *et al.* 2012; Liu *et al.* 2013; Rubæk *et al.* 2013; Yang *et al.* 2013; Roger *et al.* 2014; Sarmadian *et al.* 2014; Keshavarzi *et al.* 2015).

MATERIAL AND METHODS

Data entry. At the beginning of our research we received Excel files containing data for 1000 pages of the original notebooks. There were three notebooks containing physical properties and three containing chemical properties. The data were entered by about 100 persons.

For the purposes of easier handling and inspection, we scanned the pages of the notebooks into separate image files. We also created a computer program that allowed the researchers to quickly view both the scanned image and the corresponding Excel file, for a particular notebook and page.

Originally, the data had not been written in both types of notebooks at the same time, nor following the same procedure, which made it challenging to match the corresponding entries for identical profiles in both types of notebooks with a computer program.

Processing of entered data. For data processing and validation, and for populating the database, we used the Python programming language. About 4000 lines of code were written in over 50 scripts (available at https://bitbucket.org/edin1/montenegro_soil_data_parser).

Final database creation. Overall, the processing steps for creating the final database were as follows: parse all Excel files and create separate databases (SQLite files) for every notebook; add coordinates to those databases; combine corresponding databases of chemical and physical properties for every pair of books, where possible, into a single, combined database; and merge all combined databases into a single, final database.

Numerous data consistency checks were put at appropriate places in the above procedure, e.g. checks that the depths of the horizons for particular profiles were not overlapping, the horizons were in the appropriate order, the numbers of profiles were in increasing order, spelling checks, etc.

A detailed workflow of our digitization procedure can be found here: https://bitbucket.org/edin1/montenegro_soil_data_parser/src/master/parse_all.bat

Data georeferencing. The previously produced point data set contained the coordinates of individual soil profiles marked on the raster map. The legacy raster map was created by merging 38 scanned map sections that cover the whole territory of Montenegro. Figure 1 shows the whole merged map; an example section on the map, named "Cetinje 1" (42.25°–42.5°N, 18.83°–19.08°E); and a marked profile (25 m) located in that section.

As the original maps were created manually, and the scanning process was not perfect, the raster map and, consequently, the point data set had small but visible deviations from Google's satellite map. The coordinate reference system (CRS) encoded in the raster maps and hence the point data set were a slightly modified version of EPSG:31276 MGI/Balkans zone 6.

Mapped area. A 400 km^2 area near the city of Cetinje was chosen for mapping (42.29°–42.48°N, 18.83°–19.07°E). The chosen area is contained in the "Cetinje 1" section of the legacy map. Due to the visible deviations from Google's satellite map, we manually adjusted the CRS embedded in the raster map. This was done by a researcher modifying the proj4 parameters in QGIS in small increments. He would then visually inspect the overlap of various topographic objects (mainly bodies of water) between the raster map and Google's satellite map for "Cetinje 1" section. Final proj4 parameters were recorded for the overlap that was deemed best. These parameters were subsequently used for DSM algorithms.

Summers in Cetinje are dry and warm, with an average temperature of 20°C, while its winters are mild

https://doi.org/10.17221/81/2017-SWR

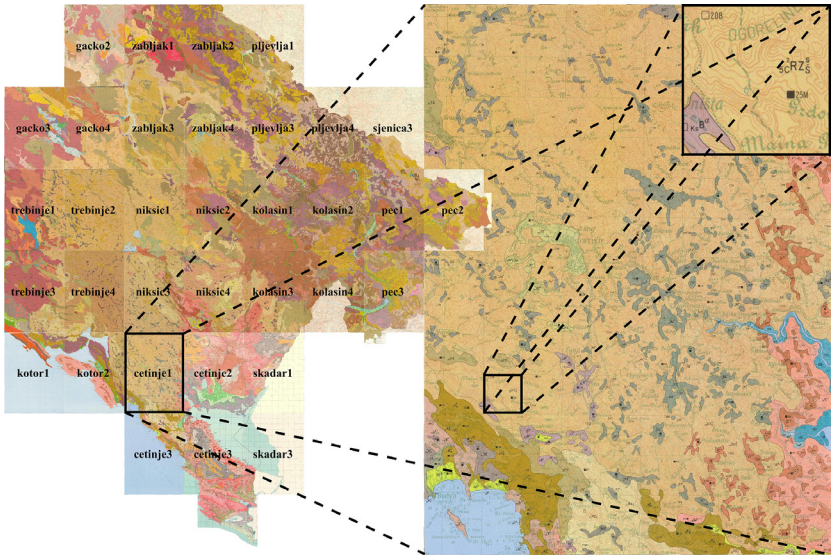

Figure 1. Legacy soil map of Montenegro, "Cetinje 1" map section, and profile 25 m in that section

and wet, with an average temperature of 2.1°C. With around 4000 mm of average yearly rainfall, Cetinje is one of the rainiest cities in Europe. However, due to the karst landscape, there are not many bodies of water in the city and its surroundings. The mapped area represents a part of the limestone plateau. There are several levels there: the first level is a perimeter of the Zeta plain at 150–200 m, the second is the area above Rijeka Crnojevića at 350–500 m, and the third are the surroundings of Cetinje and Katun Karst at 650–800 m. In terms of land cover, this is one of the poorest areas of Montenegro, due to the very slow formation of soil in the limestone, which is poor in clay, and the ever-present erosion. Rough relief forms with a very shallow soil profile are dominant in this karst area. In terms of vegetation, these soils are mostly covered with forest (rare assembly) and grass (pasture). Calcomelanosol is a dominant soil type, not only for the investigated area, but also for Montenegro. It occupies about 47% of the territory of Montenegro (Fuštić & Đuretić 2000). Prior to and at the time of the survey, the area around Cetinje had low agricultural development. Such an area with one dominant soil type, a small number of bodies of water, and low human activity was chosen as these factors are beneficial to most mapping algorithms.

There were 125 data points in this area. The mean nearest neighbour distance between them was 1127 m. The complete spatial randomness (CSR) test showed that the original sampling plan was mostly geographically representative of the chosen area.

Data samples. Only the topsoil horizons were considered. For all data points, only the phosphorus concentration values were extracted from the database and used in the map creation process. The concentration of soil available phosphorus (SAP) was expressed as the number of mg of P_2O_5 in 100 g of soil. SAP was determined by the ammonium lactate-acetate method (Egnér *et al.* 1960), used in several European countries. The soil samples were air-dried and passed through a 2-mm sieve before analysis. SAP was then extracted with a mixture of 0.1 mol ammonium lactate and 0.4 mol acetic acid, buffered at acidic pH (3.75). The phosphorus concentration in the extracts was determined spectrophotometrically.

Figure 2 shows a bubble plot of the recorded values of phosphorus concentration. Because the concentrations were skewed, they were log-transformed and scaled prior to applying the algorithms (Hengl 2009).

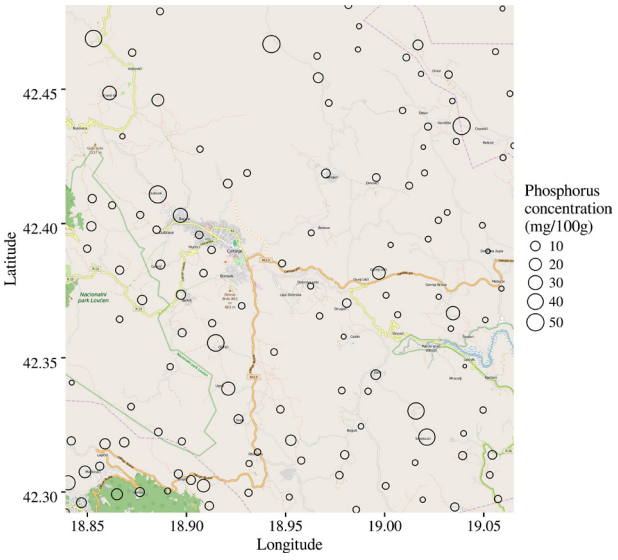

Figure 2. Concentrations of phosphorus in the designated area for the original sample points

https://doi.org/10.17221/81/2017-SWR

Predictor variables. The variables used for prediction were terrain attributes generated from satellite-based digital elevation model (DEM) data, with a spatial resolution of 3'' (90 m) (Ferranti 2014). The SAGA GIS software (Conrad *et al.* 2015) was used to generate the slope, aspect, and plan curvature terrain attributes from elevation, which were then, together with elevation, used as predictor variables. Sarmadian *et al.* (2014) and Keshavarzi *et al.* (2015) reported that these terrain attributes can be successfully used as predictors for phosphorus. Satellite data from the global land survey (GLS) data sets were used as additional predictor variables, namely multispectral scanner (MSS) bands 1, 2, 3, and 4 for the year 1975 (USGS 2008a), and thematic mapper (TM) bands 1, 2, …, 7 for the year 1990 (USGS 2008b). These particular GLS data sets were chosen because they are among the oldest publicly available satellite data sets for the territory of Montenegro, i.e. they are the closest to the time period of the original soil data collection.

All predictor variables were scaled. Principal component analysis was performed using the R package GSIF (Hengl *et al.* 2014), and only the principal components that explained 80% of variance were used for prediction (King & Jackson 1999).

Mapping algorithms and procedure. Map creation algorithms were applied using the R programming language (R Core Team 2015) and its libraries. The algorithms that were used are: inverse distance weighting (IDW) (Shepard 1968), ordinary kriging (OK) (Hengl 2009), multiple linear regression (LR) (McBratney *et al.* 2003; Hengl 2009), and regression-kriging (RK) (Cressie 2015). The variograms used for OK and RK were, in general, obtained automatically from the data using the R library automap (Hiemstra *et al.* 2008).

Training and validation. For the purposes of model training and validation, several setups of filtering and data set splitting were tested. The training data set was used only for calculating the variogram model. Because of the small number of points, it was difficult to find a good setup, as even the choice of the random "seed" value used for splitting the set into training and validation subsets strongly influenced the final results. Only the maps for the setup which gave the best results (across all models) are shown. This setup involved removal of outliers (values out of the 2 standard deviations range around the median value) for a total of 115 points remaining; using 100% of the data points for both training and validation; using a "hand-tuned" variogram for OK, as the automap generated one gave very poor results; and using leave-one-out (LOO) validation. In other setups that were tested, outliers were included, and the automap was used even for OK. The prediction reliability indicators that were recorded are the root mean square error (RMSE) and the coefficient of multiple correlation of determination R^2.

Table 1. Prediction reliability indicators with leave-one-out (LOO) validation, removed outliers, and 100% of data used for both training and validation

	IDW	OK	LR	RK
RMSE	0.9519	0.9807	0.8870	0.8744
R^2	0.0894	0.0297	0.2064	0.2287

IDW – inverse distance weighting; OK – ordinary kriging; LR – multiple linear regression; RK – regression-kriging; RMSE – root mean square error; R^2 – coefficient of multiple correlation of determination

RESULTS

With respect to creating a thematic map of phosphorus, for the best setup, Figure 3 shows the generated maps for all algorithms. Table 1 shows the prediction reliability indicators for all algorithms.

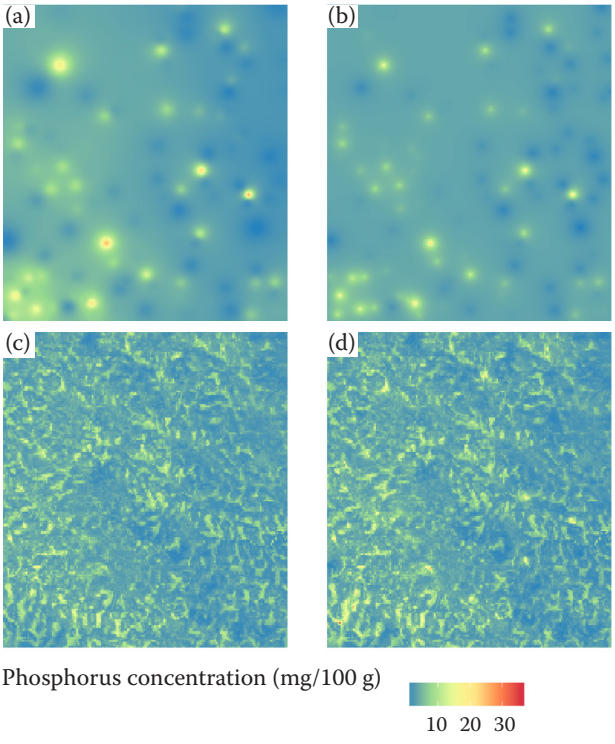

Figure 3. Generated maps for all models: inverse distance weighting (IDW) (a), ordinary kriging (OK) (b), multiple linear regression (LR) (c), regression-kriging (RK) (d)

https://doi.org/10.17221/81/2017-SWR

Table 2. Prediction reliability indicators for other validation setups

	IDW	OK	LR	RK
10-segment validation, 100% of data used for both training and validation				
RMSE	0.9720	0.9886	0.9549	0.9403
R^2	0.0491	0.0149	0.0807	0.1086
LOO validation, 100% of data used for both training and validation				
RMSE	0.9720	0.9850	0.9434	0.9338
R^2	0.0490	0.0219	0.1028	0.1210
10-segment validation, 45% of data used for training and 55% for validation				
RMSE	0.9675	0.9277	0.8882	0.8128
R^2	–0.1765	–0.0853	0.0051	0.1680
LOO validation, 45% of data used for training and 55% for validation				
RMSE	0.8796	0.8963	0.8214	0.7731
R^2	0.0547	0.0185	0.1757	0.2699

IDW – inverse distance weighting; OK – ordinary kriging; LR – multiple linear regression; RK – regression-kriging; RMSE – root mean square error; R^2 – coefficient of multiple correlation of determination; LOO validation – leave one out validation

For some of the other setups that were tested, only the prediction reliability indicators are shown in Table 2.

DISCUSSION

Concerning the quality and reliability of the digitization process, as well as the resulting database of soil properties, it can be considered as mostly finished. Although enhancements are certainly possible, we believe that the database in its current state can be used as a reliable replacement for the original handwritten data. Furthermore, the number of georeferenced profiles is approximately half the number of digitized profiles reported by Arrouays *et al.* (2017) (6500) for four times larger Croatia.

All of the profiles could not be georeferenced because a significant number of them had not been marked on the legacy map. On the other hand, a significant number of profiles that were marked on the map were not present, or not appropriately marked in the notebooks.

Although the area chosen for mapping is small, and hence not representative of the whole dataset, the results are still useful as they can serve as a gauge of the highest attainable result for the whole map, provided a similar setup. This is because, as was already discussed, the mapping area was purposely chosen to be suitable for DSM algorithms that use terrain attributes as ancillary predictors. Also, for such a small area, the results can be easily checked and validated by a future ground-truth campaign.

Chemistry of soil phosphorus is very complex and the variability of phosphorus concentration is high in agricultural soils (Topalović *et al*. 2006). As such, phosphorus is not as easily mappable as some other basic soil properties, such as clay, pH, organic matter, etc. However, we already mentioned that some researchers reported obtaining good results when mapping phosphorus using terrain attributes only. Also, there is growing interest in agricultural development in the mapped area, and a reliable map of phosphorus could be helpful in making a fertilization program for the area.

In terms of the generated thematic maps of phosphorus, it is obvious that their quality is not high. The R^2 value for the best case is 0.23, while a desirable value, in general, should be close to 0.80 (Hengl 2009). Sarmadian *et al.* (2014) reported an R^2 of 0.48 for mapping phosphorus by regression-kriging (RK). Keshavarzi *et al.* (2015) reported an R^2 of 0.68 for mapping phosphorus by a neural network model. However, Roger *et al.* (2014) reported an R^2 value between 0.20 and 0.25 for various phosphorus forms.

One of the reasons for the low R^2 value obtained might be that the mean nearest neighbour distance between sample points is large, 1127 m, compared to e.g. 300 m in Keshavarzi *et al.* (2015). This is strengthened by Sarmadian *et al.* (2014), where a variogram was presented that shows a decreasing

https://doi.org/10.17221/81/2017-SWR

geostatistical correlation of phosphorus values after 1000 m, and independence above 1500 m. Moreover, for our data, the sampling was not done on a regular rectangular grid (Hengl *et al.* 2002). Another reason could be the imprecision of the coordinates and the low resolution of the utilized DEM data (Keshavarzi *et al.* 2015). On the other side, Roger *et al.* (2014) suggested that the low R^2 value they obtained might have been due to land use, or even due to geological factors and parent material. For our mapped area, the parent material is uniform (limestone). Land use should not be a factor of low R^2, because there was low agricultural activity in the area and because soil sampling was done from soil profiles opened at places which best represent the properties of individual mapped units and away as much as possible from the direct effect of human activities (Fuštić & Đuretić 2000). However, the spatial variability of soil properties that influence phosphorus sorption and desorption, such as particle size distribution, pH, Fe and Al oxides, could be the main factor of low R^2.

Acknowledgements. This research is supported in part by the Ministry of Science of Montenegro through the BIO-ICT Centre of Excellence in Bioinformatics.

References

Arrouays D., Leenaars J.G., Richer-de-Forges A.C., Adhikari K., Ballabio C., Greve M., Grundy M., Guerrero E., Hempel J., Hengl T., Heuvelink G., Batjes N., Carvalho E., Hartemink A., Hewitt A., Hong S.-Y., Krasilnikov P., Lagacherie P., Lelyk G., Libohova Z., Lilly A., McBratney A., McKenzie N., Vasquez G.M., Mulder V.L., Minasny B., Montanarella L., Odeh I., Padarian J., Poggio L., Roudier P., Saby N., Savin I., Searle R., Solbovoy V., Thompson J., Smith S., Sulaeman Y., Vintila R., Rossel R. V., Wilson P., Zhang G.-L., Swerts M., Oorts K., Karklins A., Feng L., Ibelles Navarro A.R., Levin A., Laktionova T., Dell'Acqua M., Suvannang N., Ruam W., Prasad J., Patil N., Husnjak S., Pásztor L., Okx J., Hallett S., Keay C., Farewell T., Lilja H., Juilleret J., Marx S., Takata Y., Kazuyuki Y., Mansuy N., Panagos P., Van Liedekerke M., Skalsky R., Sobocka J., Kobza J., Eftekhari K., Alavipanah S. K., Moussadek R., Badraoui M., Da Silva M., Paterson G., Conceição Gonçalves M. da, Theocharopoulos S., Yemefack M., Tedou S., Vrscaj B., Grob U., Kozák J., Boruvka L., Dobos E., Taboada M., Moretti L., Rodriguez D. (2017): Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ, 14: 1–19.

Conrad O., Bechtel B., Bock M., Dietrich H., Fischer E., Gerlitz L., Wehberg J., Wichmann V., Böhner J. (2015): System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geoscientific Model Development Discussions, 8: 2271–2312.

Cressie N. (2015): Statistics for Spatial Data. New York, John Wiley & Sons.

Egnér H., Riehm H., Domingo W.R. (1960): Untersuchungen über die chemische Bodenanalyse als Grundlage für die Beurteilung des Nährstoffzustandes der Böden. II. Chemische Extraktionsmethoden zur Phosphor-und Kaliumbestimmung. Kungliga Lantbrukshögskolans Annaler, 26: 204–209.

Ferranti J. de (2014): Worldwide 3" DEM. Available at http://www.viewfinderpanoramas.org/dem3.html (accessed May 2015)

Fuštić B., Đuretić G. (2000): The Soils of Montenegro. Podgorica, University of Montenegro. (in Montenegrin)

Hengl T. (2009): A Practical Guide to Geostatistical Mapping. Amsterdam, University of Amsterdam.

Hengl T., Husnjak S. (2006): Evaluating adequacy and usability of soil maps in Croatia. Soil Science Society of America Journal, 70: 920–929.

Hengl T., Rossiter D. G., Husnjak S. (2002): Mapping soil properties from an existing national soil data set using freely available ancillary data. In: Proc. 17th World Congress of Soil Science. Bangkok, IUSS: 1140-1–1140-10.

Hengl T., de Jesus J.M., MacMillan R.A., Batjes N.H., Heuvelink G.B.M., Ribeiro E., Samuel-Rosa A., Kempen B., Leenaars J.G.B., Walsh M.G., Gonzalez M.R. (2014): SoilGrids1km – global soil information based on automated mapping. PLoS ONE, 9: 1–17.

Hiemstra P.H., Pebesma E.J., Twenhöfel C.J.W., Heuvelink G.B.M. (2008): Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity Monitoring Network. Computers & Geosciences, 35: 1711–1721.

Keshavarzi A., Sarmadian F., Omran E.-S.E., Iqbal M. (2015): A neural network model for estimating soil phosphorus using terrain analysis. The Egyptian Journal of Remote Sensing and Space Science, 18: 127–135.

King J.R., Jackson D.A. (1999): Variable selection in large environmental data sets using principal components analysis. Environmetrics, 10: 67–77.

Liu Z.-P., Shao M.-A., Wang Y.-Q. (2013): Spatial patterns of soil total nitrogen and soil total phosphorus across the entire Loess Plateau region of China. Geoderma, 197: 67–78.

McBratney A.B., Mendonça Santos M.L., Minasny B. (2003): On digital soil mapping. Geoderma, 117: 3–52.

R Core Team (2015): R: A Language and Environment for Statistical Computing. Vienna, R Foundation for Statistical Computing.

https://doi.org/10.17221/81/2017-SWR

Roger A., Libohova Z., Rossier N., Joost S., Maltas A., Frossard E., Sinaj S. (2014): Spatial variability of soil phosphorus in the Fribourg canton, Switzerland. Geoderma, 217: 26–36.

Rubæk G.H., Kristensen K., Olesen S.E., Østergaard H.S., Heckrath G. (2013): Phosphorus accumulation and spatial distribution in agricultural soils in Denmark. Geoderma, 209: 241–250.

Sarmadian F., Keshavarzi A., Rooien A., Iqbal M., Zahedi G., Javadikia H. (2014): Digital mapping of soil phosphorus using multivariate geostatistics and topographic information. Australian Journal of Crop Science, 8: 1216–1223.

Schachtman D.P., Reid R.J., Ayling S.M. (1998): Phosphorus uptake by plants: from soil to cell. Plant Physiology, 116: 447–453.

Shepard D. (1968): A two-dimensional interpolation function for irregularly-spaced data. In: Proc. 23rd ACM National Conf. New York, ACM: 517–524.

Topalović A., Pfendt L.B., Perović N., Đorđević D., Trifunović S., Pfendt P.A. (2006): The chemical characteristics of soil which determine phosphorus partitioning in highly calcareous soils. Journal of the Serbian Chemical Society, 71: 1219–1236.

USGS (2008a): Collection Name: Global Land Survey, Epoch: 1975, Sensor name: Landsat MSS, Image Name: 60 meter scene p201r030_3dm19780706. Sioux Falls, United States Geological Survey.

USGS (2008b): Collection Name: Global Land Survey, Epoch: 1990, Sensor name: Landsat TM, Image Name: 60 meter scene p187r031_5dt19870724. Sioux Falls, United States Geological Survey.

Vrščaj B., Prus T., Lobnik F. (2005): Soil Information and soil data use in Slovenia. In: Jones R. J., Houšková B., Bullock P., Montanarella L. (eds): Soil Resources of Europe. 2nd Ed. Luxembourg, Office for Official Publications of the European Communities: 331–344.

Wang Y., Zhang X., Huang C. (2009): Spatial variability of soil total nitrogen and soil total phosphorus under different land uses in a small watershed on the Loess Plateau, China. Geoderma, 150: 141–149.

Xiao R., Bai J., Gao H., Huang L., Deng W. (2012): Spatial distribution of phosphorus in marsh soils of a typical land/inland water ecotone along a hydrological gradient. Catena, 98: 96–103.

Yang X., Post W.M., Thornton P.E., Jain A. (2013): The distribution of soil phosphorus for global biogeochemical modeling. Biogeosciences, 10: 2525–2537.

Received for publication March 29, 2017
Accepted after corrections September 18, 2017
Published online November 24, 2017