Spatial prediction of soil qualitative variables using logistic regression and fuzzy techniques. Study region: Dobrovăţ basin (Central Moldavian Plateau)

Cristian-Valeriu PATRICHE[1*], Bogdan ROŞCA[1], Radu Gabriel PÂRNĂU[2] and Dan Laurenţiu STOICA[3]

[1] Romanian Academy, Department of Iaşi, Geography Group, 8 Carol I, 700505, Iaşi (Romania)
[2] Faculty of Biology, "Alexandru Ioan Cuza" University of Iaşi, 20A Carol I, 700505, Iaşi, (Romania)
[3] Faculty of Geography and Geology, "Alexandru Ioan Cuza" University of Iaşi, 20A Carol I, 700505, Iaşi, (Romania)

* Correspondence to: Cristian-Valeriu Patriche, Romanian Academy, Department of Iaşi, Geography Group, 8 Carol I, 700505, Iaşi, Romania, E-mail: pvcristi@yahoo.com

doi: 10.4316/GEOREVIEW.2012.21.1.54

GEOREVIEW

Article history
Received: August 2012
Received in revised form: October 2012
Accepted: November 2012
Available online: January 2013

ABSTRACT: The present study attempts to test the performance of two statistical approaches, namely the binary logistic regression and fuzzy techniques for spatial prediction of soil types. The study area is Dobrovăţ basin, located in NE Romania, within the Central Moldavian Plateau. The input parameters are the digital elevation model, slope, topographic wetness index, mean annual temperatures and precipitations. The logistic regression approach proved successful in estimating the spatial probabilities of Aluviosols, Chernozems, Preluvosols and Luvosols and generally failed in predicting the locations of Phaeozems. The fuzzy approach, implemented through SoLIM software, proved successful in predicting the occurrences of Aluviosols and Luvisols. On the whole, both methods managed to assign the same soil type as in the soil survey map for 55-56% of the basin. Though it is clear that the approaches need to be further improved, they do present, in the authors opinion, potential for the purpose of predicting soil qualitative variables.

KEY WORDS: statistical methods, SoLIM, spatial probability, soil types, Romania

1. Introduction

Digital soil mapping refers to the application of mathematical and statistical methods for studying the spatial distribution of soils and their properties. Compared to the classical approach, digital soil mapping has several advantages: the capability of rendering the spatial continuity of soils; the prediction of soil properties inside (*interpolation*) and outside (*extrapolation*) the sampling area; the explanation of the inferred spatial distributions.

There are many methods potentially useful for digital soil mapping purposes, which can be broadly grouped into mathematical methods (inverse distance weighting, spline functions, global

/ local polynomial functions, triangulated irregular netwoks etc.) and statistical methods (global / local regression, logistic regression, classification and regression trees, kriging, regression – kriging, fuzzy techniques, neural networks etc.). A thorough synthesis of digital soil mapping techniques is provided by McBratney A.B. et al. (2003).

The application of such methods implies the use of a certain statistical software and / or a GIS software. There are many statistical packages providing acces to a wide range of statistical and mathematical methods, such as Excel/XLSTAT, Statistica, SSPS, Matlab, Minitab etc. GIS software can be grouped into commercial software, such as ArcGIS, TNTmips and open source software, such as SAGA-GIS, R, GRASS etc or the more specialised SoLIM, Vesper etc.

Dobrovăţ basin is situated in North-Eastern Romania, in the Central Moldavian Plateau, covering a surface of about 186 km^2. The relief is characterized by the presence of cuesta flanks, structural plateaus at altitudes exceeding 300m and large floodplains. It has a temperate climate, with mean annual temperatures of 8.1-9.8^{o}C and mean annual precipitations of 550-612 mm (Patriche C.V., 2005). The northern part of the region is covered by oak and beech forests, with a large extent of Luvisols, while the southern half is dominated by agricultural lands and Chernozems (Pirnău R.G., 2011).

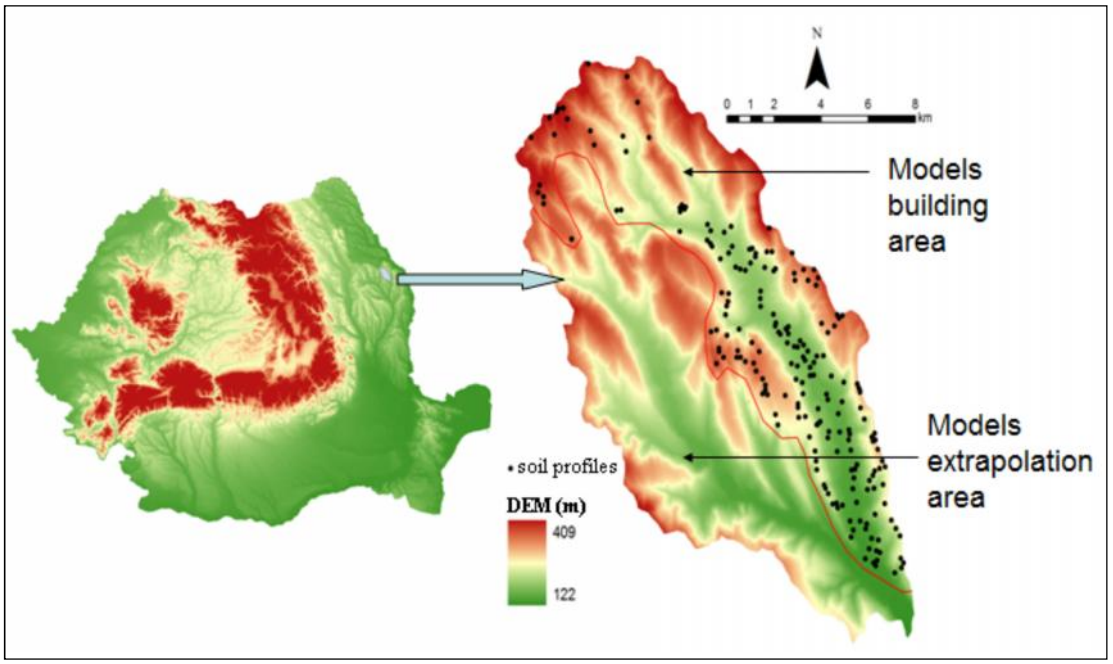


Figure 1. Geographical position of Dobrovăţ basin and the locations of soil profiles. This figure is available in colour online at www.georeview.ro

The objective of the present study is to predict the occurrence of the main soil types in the study area using several terrain variables as predictors and 2 statistical methods, namely the logistic regression and fuzzy techniques. The approach therefore attempts to reproduce an existing soil types map and to extrapolate the model in the neighbouring area. The intention is to see if such an approach could be useful for assisting the soil surveyor in his work, by providing a first approximation of soils distribution in the area of interest.

2. Materials and methods

The variables taken into account in the analysis are the 20x20m digital elevation model (DEM), slope, SAGA-GIS topographic wetness index, mean annual temperatures and precipitations, land cover and soil distribution map. The mean annual temperatures and precipitations were extracted from regression models computed for the larger area of the Moldavian Plateau (Patriche C.V., 2005). The initial rasters were resampled from 90x90m to 20x20m using the bilinear algorithm, in order to match the resolution of the other input variables. The land cover was derived from 1:5000 topographic maps and ortophoto images, while the soil map was the result of a large scale (1:5000) soil survey (Pîrnău R.G., 2011). In addition, there was available a consistent soil profiles database for the eastern half of Dobrovăţ basin, including a number of 246 soil profiles and associated analytical data, provided by Iaşi County Office for Soil Survey.

The logistic regression (Afifi A.A., Clark V., 1998) is a type of regression analysis which estimates the occurrence probability of a binary expressed (0, 1) event / qualitative variable (P) as a function of a linear combination of predictors (z). The binary values are transformed according to a logistic function, in order to derive a continuous probability range, from 0 to 1.

$$P = \frac{1}{1+e^{-z}}$$

$$z = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n$$

The regression coefficients are computed using the maximum likelihood estimation (Harrel F.E., 2010). In order to avoid the multicollinearity problem, the stepwise procedure was applied in order to select the relevant predictors. The quality of the logistic regression model can be tested in several ways: using the likelihood ratio, the pseudo coefficients of determination (McFadden, Cox and Snell, Nagelkerke), the area under the receiver operating characteristic (ROC) curve, the percentages of correctly classified observations.

The logistic regression analysis started from the soil map of the basin. A dense grid of points was overlaid and soil types were recorded for each point. Also, the predictors' values were extracted for each grid point and the location of points, in the eastern half or in the western half of the basin, was coded in order to split the sample into working sample and independent validation sample. This database was then exported into Excel/XLSTAT 2010 trial version for analysis. In a fist stage, the presence / absence of each considered soil type was coded (1/0). Because absences are much more numerous than presences, random samples were extracted from the absences, for each soil type, with a similar size in respect to the presences.

The fuzzy techniques, applied through SoLIM 5.0 software (Zhu A.X. et al., 2001), were used for the estimation of the spatial occurrence probabilities of the main soil types in Dobrovat basin, as well as for achieving an approximate map for soil distribution in the area.

In a first stage, *membership functions* were derived for each considered soil type. These functions were further used to achieve the *global spatial occurrence probabilities*, which can be extrapolated in the nearby area (Zhu A.X. et al., 2010). The membership functions are optimality functions, expressing the probability of having a certain soil type given the values of a certain factor. The SoLIM software allows the user to specify also local conditions favoring the occurrence of a certain soil type (*local probabilities*) and also local conditions excluding the

presence of a certain soil type, in which case the areas are designated probability values of zero or close to zero.

By means of fuzzy techniques, the software combines the global and local probabilities into continuous spatial distribution of probabilities, each pixel being described in terms of membership probabilities for one soil type or another (e.g. there are 70% chances that the respective pixel is represented by a Chernozem, but also 60% chances that it represents a Phaeozem). Finally, the probabilities associated with soil types can be further combined into a hardened map, in which each pixel receives the code of that soil type showing maximum probability of occurrence in that particular location.

3. Results and discussion

Analysing the overall accuracy (percentages of correctly classified points) and the area under ROC curve, as quality parameters for logistic regression models (table 1), it is found that Aluviosols and Chernozems are the best predicted soil types, followed by Preluvosols and Luvosols. A fair model was found for Erodosols and Regosols. The occurrence of Phaeozems is much less predictable, with respect to the given predictors, in which case the logistic regression model is weak.

Table 1. Quality parameters for logistic regression models

Soil type	Overall accuracy (%)		Area under ROC curve
	Working sample	Validation sample	
Aluviosols	86.5	90.1	0.902
Chernozems	86.8	71.4	0.935
Phaeozems	61.6	56.9	0.660
Preluvosols	73.3	50.3	0.857
Luvosols	70.3	89.1	0.763
Erodosols / Regosols	66.1	62.2	0.731

The spatial probability of Aluviosols was computed from altitude, topographic wetness index and mean annual temperature. In the case of Chernozems, the predictors which entered the logistic regression model are the altitude, mean annual precipitations, topographic wetness index and slope. In the case of Phaeozems, the only predictor is the slope, which points out again that the occurrence of this soil type is less predictable. The Preluvosols were modelled as a function of mean annual temperatures and precipitations, altitude and slope, while the occurrence of Luvisols took into account the mean annual temperatures, altitude, slope and topographic wetness index. Finally, the Erodosols and Regosols were modelled as a function of slope and mean annual temperatures. The spatial probability distributions of all considered soil types are given in figure 2.

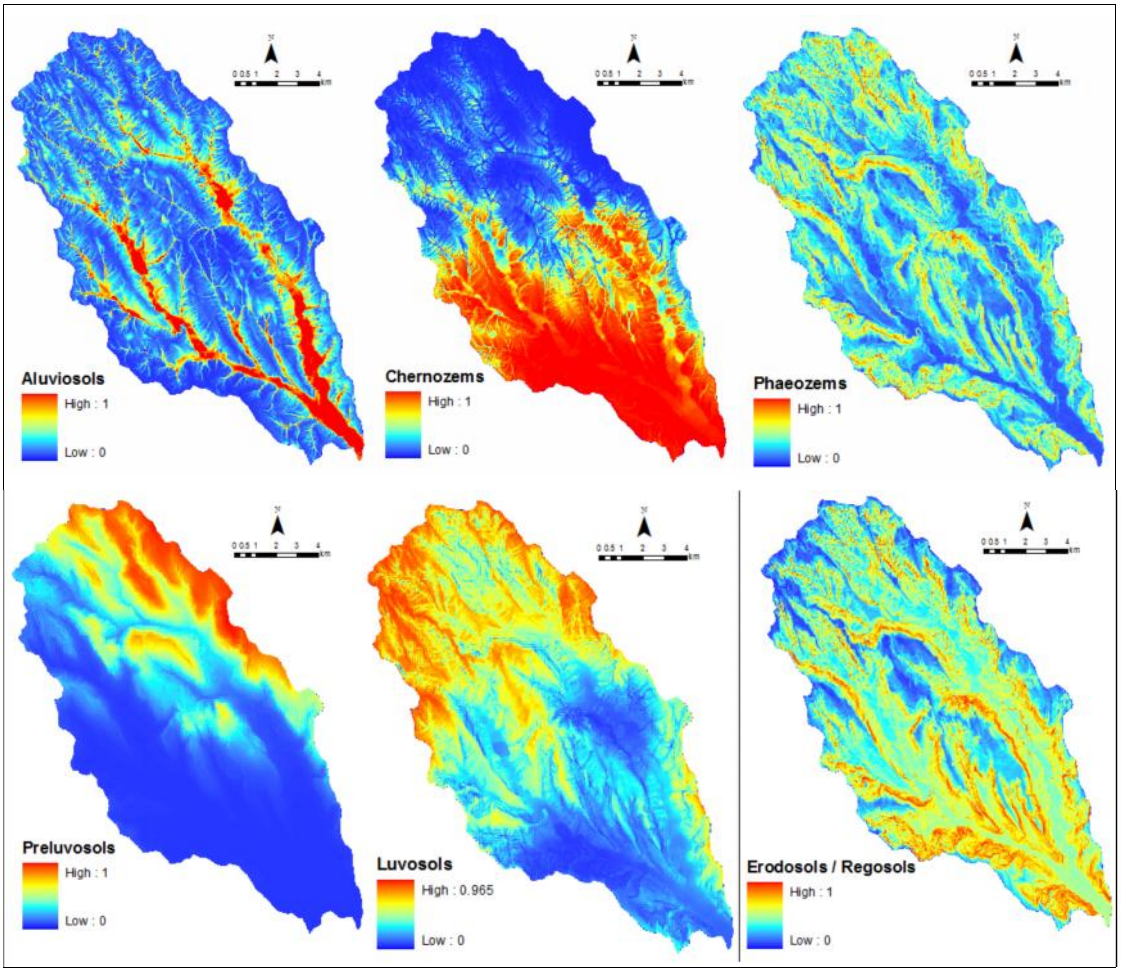


Figure 2. Spatial occurrence probabilities for soil types derived from logistic regression models. This figure is available in colour online at www.georeview.ro

The individual probabilities for soil types' occurrence were combined within SoLIM software in order to derive the predicted spatial distribution of soils according to logistic regression analysis. Comparing this distribution with the real distribution (figure 3) certain common features may be noticed. Both estimated and real soil maps were converted into raster layers using the soil type codes as pixel values. By subtracting the two representations, the resulting zero values are associated with pixels bearing the same soil type code. In this manner, it was found that for only 56.5% of the basin the logistic regression approach assigned the same soil type as in the real soil map, meaning that the analysis must be further improved.

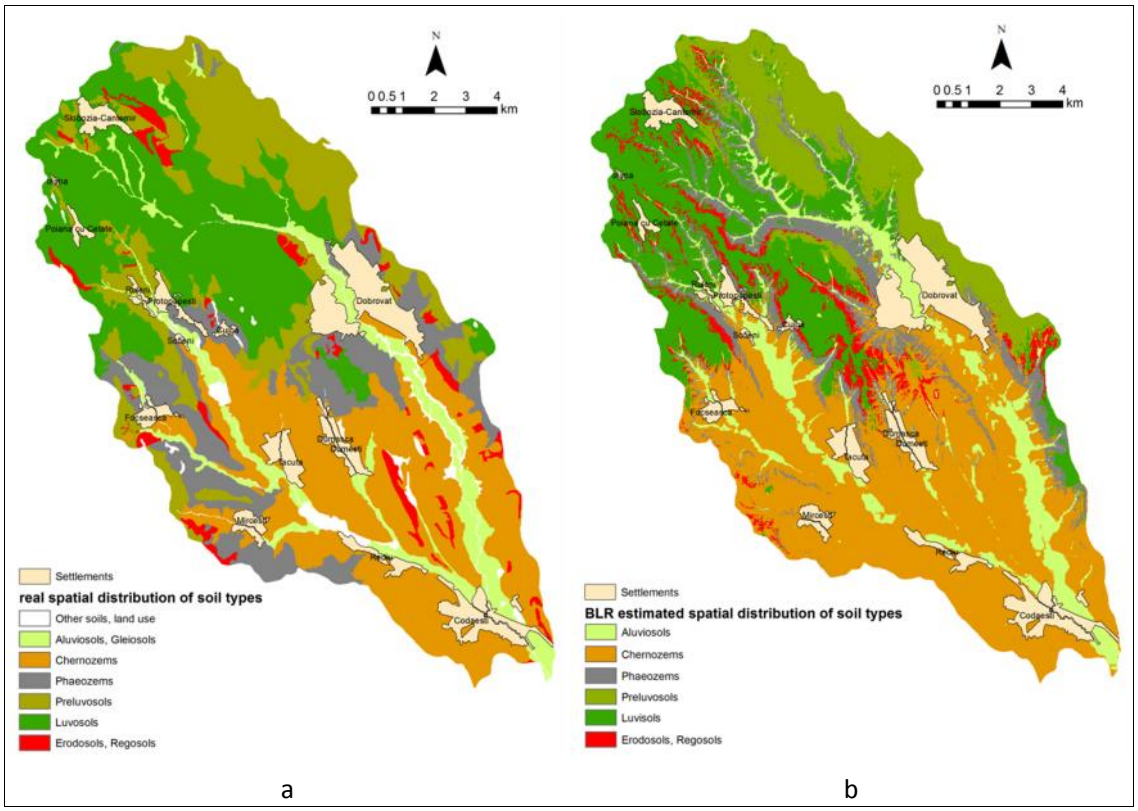


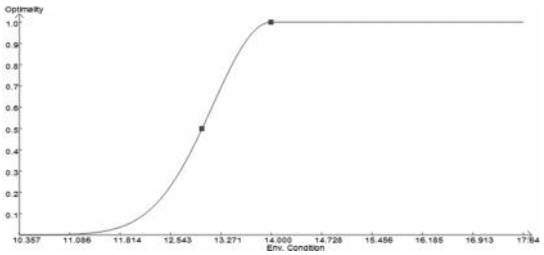
Figure 3. Real (a) and logistic regression estimated (b) spatial distribution of soil types in Dobrovăţ basin. This figure is available in colour online at www.georeview.ro.

The fuzzy approach began with the establishment of membership functions. These were defined starting from the histograms of predictors computed for each soil type. The histograms which succeeded the most to differenciate the soil types were further selected (e.g. topographic wetness index for Aluviosols, slope for Erodosols etc.). The histograms were used to identify threshold values for soil distribution (e.g. Luvisols occur predominantly at altitudes higher than 300m and they are not present at altitudes lower than 250m), which were further employed to build membership functions. In addition, the spatial probability models for A molic horizon and Preluvosols, achieved through logistic regression, were used in the case of Chernozems, Phaeozems and Preluvosols.

The variables used for building the membership functions (figure 4) are the following:

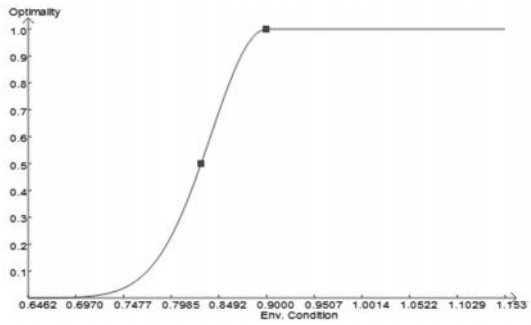
- Topographic wetness index for Aluviosols and Gleysols;
- Occurrence probability for A molic horizon and mean annual temperatures for Chernozems;
- Altitude and A molic horizon probability for Phaeozems;
- Occurrence probability for Preluvosols (logistic regression);
- Altitude and land cover for Luvosols (maximum probability under forest vegetation);
- Slope for Erodosols and Regosols.

Aluviosols

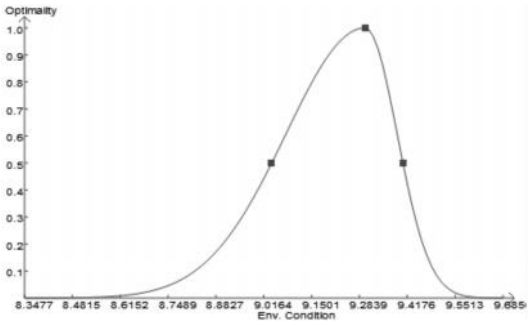


Topographic wetness index

Chernozems

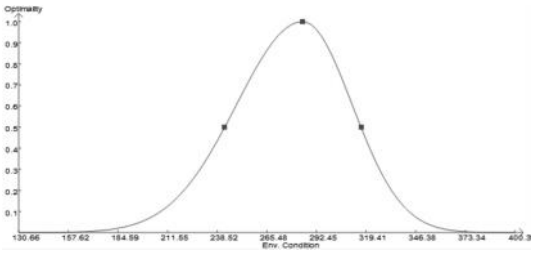


Occurrence probability for A molic horizon

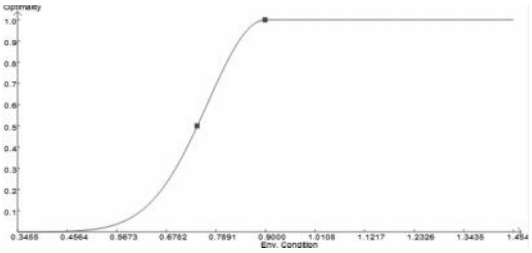


Mean annual temperature

Phaeozems

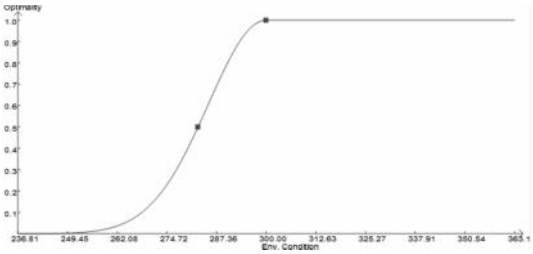


Altitude

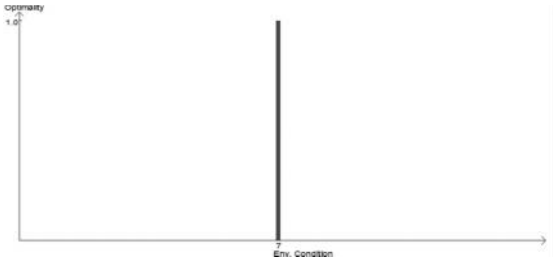


Occurrence probability for A molic horizon

Luvosols



Altitude



Land cover (presence of forests)

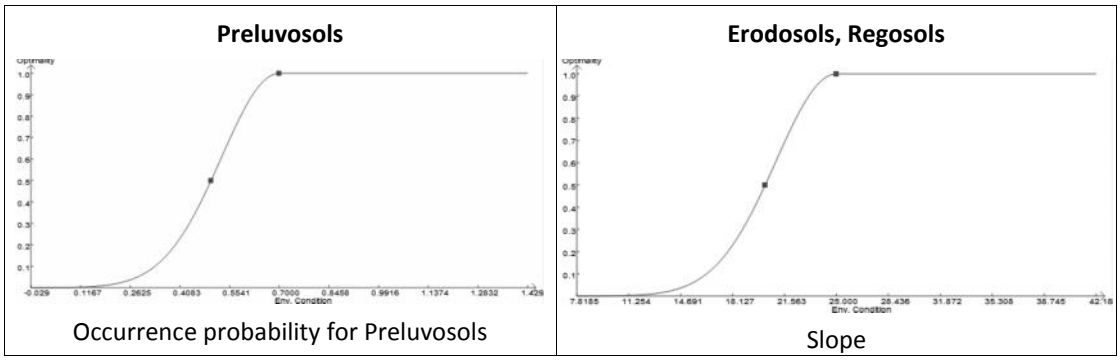


Figure 4. The membership functions used for deriving global occurrence probabilities for soil types.

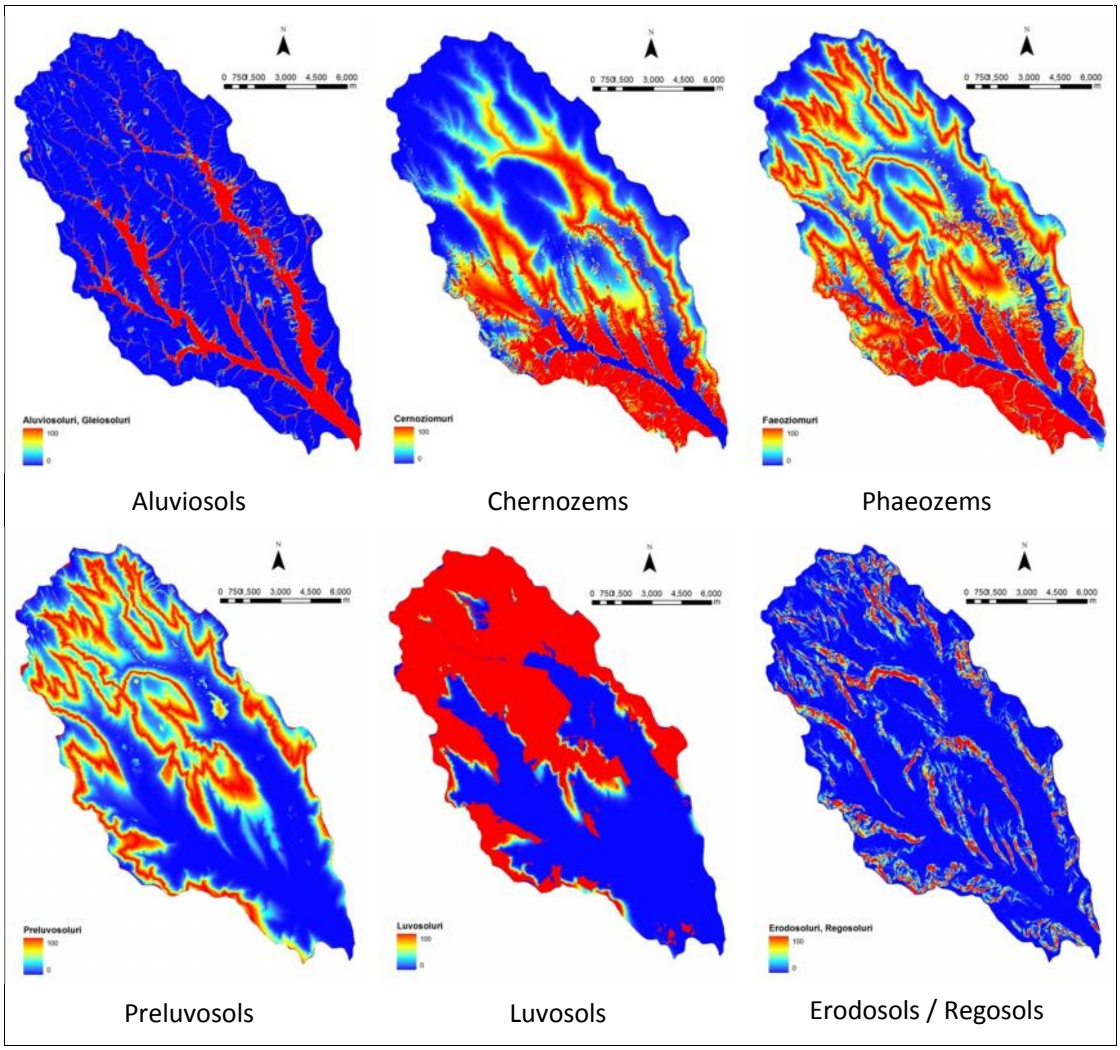


Figure 5. Spatial occurrence probabilities (%) for soil types computed from the membership functions. This figure is available in colour online at www.georeview.ro.

The spatial distribution of occurrence probabilities for soil types, according to specified criteria, are rendered in figure 5. The integration of these probabilities led to the hardened map displayed in figure 6b. The validation in respect to the 246 soil profiles indicate that the fuzzy approach managed to differentiate very well the Aluviosols (87% of Aluviosols profiles were correctly classified) and the Luvisols (79% of Luvisols profiles were correctly classified). In the case of the other soil types, the percentages of success are under 50%: Preluvosols – 41%, Chernozems – 33%, Phaeozems – 29%, Erodosols / Regosols – 13%. Comparing the predicted distribution with the real one (fig. 6), it was found that 54.8% of the basin was assigned the same soil type as in the real soil map.

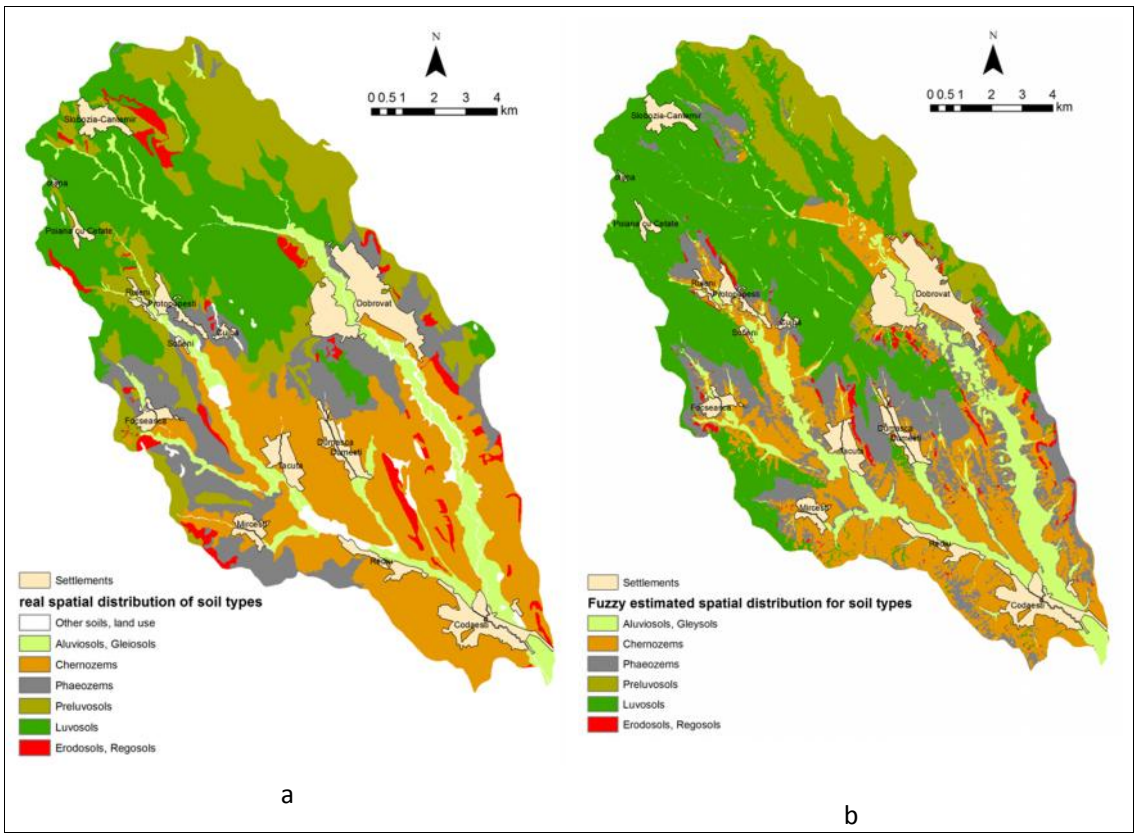


Figure 6. Real (a) and fuzzy estimated (b) spatial distribution of soil types in Dobrovăţ basin. This figure is available in colour online at www.georeview.ro.

4. Conclusions

The present study attempts to predict the main soil types in Dobrovăţ basin by means of logistic regression analysis and fuzzy techniques, using quantitative predictors such as altitude, slope, topographic wetness index, mean annual temperatures and precipitations. Both methods reveal that certain soil types, such as the Aluviosols, Chernozems, Luvisols, are more predictable than other, such as Phaeozems. The failure to make good predictions on the occurrences of Phaeozems may be due to the choice of predictors, the scale of the approach etc., but also to the Phaeozems themselves. It is generally recognized that this soil type is difficult to classify in the

field. The methods perform quite similar, as they assign the same soil type as the ones in the real soil map for 55-56% of the basin. The study certainly needs further improvements, by adding some new, relevant predictors, by expanding the region in order to take into account a broader geographical context etc. However, the results achieved so far prove that the methods presented show potential for use in predicting soil qualitative variables. They cannot substitute the soil survey, but they can provide first approximations of soils distribution in the area of interest.

References

Afifi A.A., Clark V. 1998. *Computer aided multivariate analysis*, Chapman Hall: London.

Harrel F.E. 2010. *Regression Modelling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis*, Springer Series in Statistics, Springer-Verlag, New York.

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

Patriche C.V. 2005. Podişul Central Moldovenesc dintre râurile Stavnic şi Vaslui – studiu de geografie fizică, Edit. Terra Nostra, Iaşi.

Patriche C.V., Pîrnău R., Roşca B., Vasiliniuc I. 2011. *Preliminary results regarding the application of statistical methods for spatial prediction of soil parameters in Dobrovăţ Basin (Central Moldavian Plateau),* Factori şi Procese Pedogenetice din Zona Temperată, S. Nouă, **10:** 95-102.

Pîrnău R.G. 2011. *Utilizarea terenului şi calitatea solurilor agricole din bazinul hidrografic Dobrovăţ*, teza de doctorat, Univ. "Alexandru Ioan Cuza" din Iaşi.

Zhu A.X., Hudson B., Burt J., Lubich K., Simonson D. 2001. *Soil mapping using GIS, expert knowledge, and fuzzy logic*, Soil Science Society of America Journal, Vol. **65**, pp. 1463-1472.

Zhu A.X., Yang L., Li B., Qin C., Pei T., Liu B. 2010. *Construction of membership functions for predictive soil mapping under fuzzy logic*, Geoderma **155**: 164–174.

* * * *ArcGIS Desktop 9.3 Help*, Environmental Systems Research Institute (ESRI), http://webhelp.esri.com/arcgisdesktop/9.3/

* * * *XLSTAT Tutorial*, http://www.xlstat.com/en/support/tutorials/.