Bulletin UASVM Agriculture, 68(1)/2011
Print ISSN 1843-5246; Electronic ISSN 1843-5386

Comparing Linear Regression and Regression Trees for Spatial Modelling of Soil Reaction in Dobrovăț Basin (Eastern Romania)

Cristian Valeriu PATRICHE[1], Radu Gabriel PÎRNĂU[2], Bogdan ROŞCA[1]

[1] Faculty of Geography and Geology, Geography Department, “Alexandru Ioan Cuza” University of Iaşi, 20A Carol I, 700505, Iaşi (Romania)
Romanian Academy, Department of Iaşi, Geography Group, 8 Carol I, 700505, Iaşi (Romania), email: pvcristi@yahoo.com , roscao@gmail.com
[2] Iaşi County Office for Soil Survey, 3 Dumbrava Roşie, Iaşi (Romania), 700471, email: radupirnau@yahoo.com

Abstract. Our study compares the performances of two statistical methods, namely multiple linear regression and classification and regression trees, for deriving spatial models of soil reaction in the surface horizon. The applications were carried out within a 186 km^2 hydrographic basin situated in eastern Romania. Statistical models were computed from a sample of 235 soil profiles, scattered in the eastern half of the basin. An independent sample of 237 expeditionary pH measurements was used to validate the results within the interpolation area, whereas an independent sample of 50 soil profiles was used to validate the results within the extrapolation area (the western half of the basin). The predictors included geomorphometrical parameters, derived from a 10x10 m digital elevation model, X and Y coordinates of soil profiles and the main soil types for the regression trees approach. The stepwise selection procedure indicated Y coordinate, digital elevation model, wetness index and surface ratio as the best predictors for soil reaction. The correlation between observed and predicted pH values for the training sample suggests a much higher quality of the regression trees spatial model. However, the validation using the two independent samples points out the instability of this model and recommends the regression model more reliable.

Keywords: linear regression, regression trees, pH, extrapolation, interpolation.

INTRODUCTION

During the last decades, statistical methods have developed exponentially, supported by the rapid evolution of precision instruments and computers, allowing the implementation of complex, manually unapproachable methods, as well as a fast and accurate processing of large amounts of data. Soil science has continuously and consistently benefited from application of statistical methods (McBratney *et al.*, 2003, Lagacherie *et al.*, 2006), leading to the individualization of a new branch within this field of science, called pedometrics, which is defined as “application of spatial statistics for the purpose of spatio-temporal modelling of soil data. It especially focuses on soil survey, precision agriculture applications, mapping of soil pollutants and other environmental applications” (Heuvelink, 2003).

The purpose of our study is to test the performances of two statistical methods, namely multiple linear regression (MLR) and classification and regression trees (CART), for spatial prediction of soil reaction both inside (interpolation) and outside (extrapolation) the main sampling zone.

MATERIALS AND METHODS

Dobrovăț basin is situated in eastern Romania, within the Central Moldavian Plateau, covering a surface of about 186 km^2. The monocline structure of the surface geological layers has conditioned the formation of a cuesta landscape (Ioniță, 2000), with steeper slopes (>20^o) oriented towards North and West. The highest altitudes, exceeding 350 m, correspond to structural plateaus located mainly in the northern part of the region, while the lowest altitudes, under 170 m, are encountered along the main floodplains from the southern part of the region. The climate is temperate continental, with mean annual temperatures of 8.1-9.8^{o}C and mean annual precipitations of 550-612mm (Patriche, 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, 2011).

The pedological database, consisting in georeferenced soil profiles and associated analytical data, was provided by Iasi County Office for Soil Survey. As one can easily see in figure 1, there is a great contrast in soil profiles coverage throughout the basin. This suggested us to test methods suitable for extrapolation, specifically to use the much consistent data from the eastern half of the basin (235 profiles) and test it against the data available for the western half (50 profiles). In addition, 237 expeditionary pH measurements were available for the eastern part of the basin, which we used as an independent validation sample for interpolation.

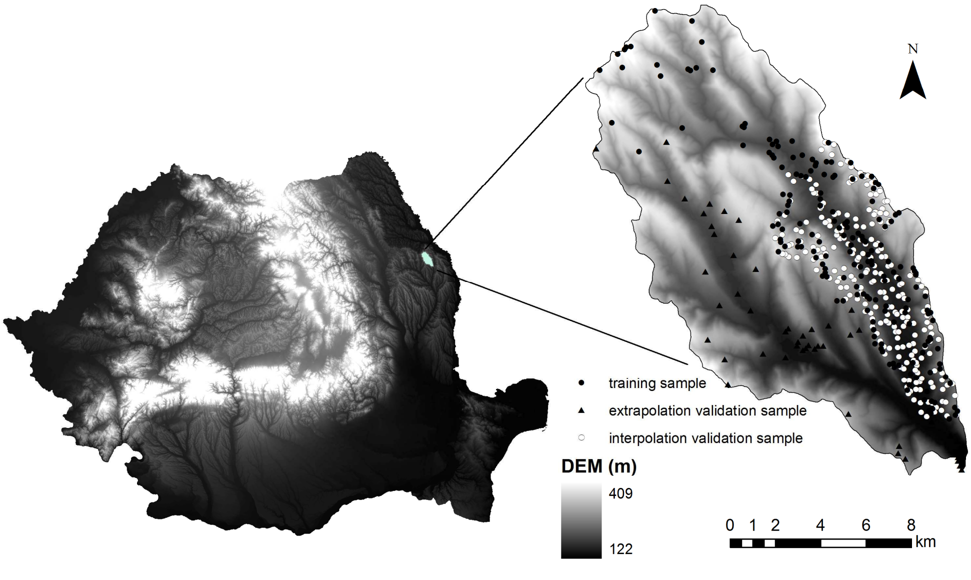

N
● training sample
▲ extrapolation validation sample
○ interpolation validation sample
DEM (m)
409
122
0 1 2 4 6 8
km

Fig. 1. Location of the study region within Romania and spatial distribution of soil profiles

Among the statistical methods suitable for extrapolation, we focused on the classical, wide-spread linear regression and on the less common classification and regression trees. Local interpolators, such as kriging or geographically weighted regression, are not fit for this purpose, as they depend on the existence of neighbouring data.

Statistical analysis was carried out using XLSTAT 2010 trial version software. The results were subsequently applied in GIS using ArcGIS 9.3 and TNTmips 6.9 software packages. Also, SAGA-GIS 2.0.3 was used to derive some of the geomorphometrical predictors.

Multiple linear regression is a method commonly used for computing spatial models of soil variables, especially in combination with ordinary kriging of residuals (regression-kriging). Details regarding the theory and application of linear regression can be found in numerous scientific papers approaching the use of statistical methods in geosciences (Johnston, 1978; Burrough and McDonnell, 1998; Hastie et al., 2001; Hengl et al., 2004; Kutner et al., 2004; Freund et al., 2006; Hengl, 2007).

Essentially, the linear regression method aims to explain the spatial distribution of a quantitative soil variable (dependent variable) by means of a linear combination of predictors, through the form of a regression equation:

$$\hat{y} = a + \sum_{i=1}^{n} b_i \cdot x_i \pm \varepsilon$$

where ŷ is the dependent variable (soil parameter), x_i are the predictors, n is the number of predictors, a is the intercept, b_i are the partial regression coefficients and ε is the standard error of estimate.

The regression equation form is determined by minimizing the sum of squares of differences between observed and predicted values (minimizing residual variance). The global approach, suitable for extrapolation, computes a single equation for the area of interest, based on all available data. For this reason, the resulting statistical model reflects the general spatial trend of soil parameters, as the method is unable to render spatial anomalies. These anomalies, induced by local factors and generally important for soil cover, interfere with model's quality to the extent that the explained variance rarely exceeds 60%. Typically, the solution to this problem is the coupling of regression model with a kriging model of residuals (Hengl, 2004, 2007).

Linear regression and regression-kriging have been extensively used for mapping soil quantitative parameters, such as clay content, soil moisture, soil depth, soil reaction, organic carbon content (Odeh and McBratney, 2000; Florinsky et al., 2002; Xie et al., 2004; Cheng et al., 2004; Hengl et al., 2007; Ziadat, 2010).

Working with many predictor variables arouses the multicollinearity problem, which is caused by the presence of significant correlations among the predictors. This redundancy may negatively affect the interpretation of partial regression coefficients. The minimization of multicollinearity may be achieved by using principal components analysis (PCA) to extract orthogonal predictors from the original data (Hengl, 2004) or by filtering the predictors using the stepwise approach, the latter being applied in our analysis.

Classification and regression trees (CART, Breiman et al., 1984) identifies optimum break points within predictor variables, separating them in groups inside which the values of the dependent variable are as homogeneous as possible. At first step, the method selects the predictor on the basis of which the dependent variable may be best separated into two groups and identifies the optimum break point. Each of the two resulting groups are further separated into two sub-groups on the basis of another (or the same) predictor. Following this logic, the method generates a tree-like structure by means of which the dependent variable is optimum divided into a number of groups, characterized by maximum internal homogeneity and maximum external differentiation.

CART may be used to explain and predict both qualitative variables (classification trees), such as soil classes (Lagacherie and Holmes, 1997; Mendonça-Santos et al., 2008), soil drainage classes (Ciatella et al., 1997) and quantitative variables (regression trees), such as soil profile depth, total organic carbon (McKenzie and Ryan, 1999; Ryan et al., 2000), clay content, silt content, cation exchange capacity (McBratney et al., 2000; Bishop et al., 2001; Park and Vlek, 2002).

Though less used for modelling soil parameters, CART method presents a series of important advantages compared to the classical linear regression: predictor – predictand relationship is non-linear; it is a non-parametric method; it may easily integrate qualitative variables both as predictors and dependent variable. Among the disadvantages, we may mention the subjectivity involved in choosing the optimum tree size, the discrete output (the method does not generate a continuous series of values, but a finite number equal to the number of the terminal nodes).

RESULTS AND DISCUSSION

The predictors used for explaining the spatial distribution of soil reaction in the surface horizon are mostly geomorphometrical predictors, derived from a 10x10 m digital elevation model (DEM): slope angle and aspect, surface ratio, convergence index, SAGA wetness index, mean, profile and plan curvature, flow accumulation. In addition, we used the X and Y coordinates of soil profiles and, for CART approach, the main soil types. The surface ratio represents the ratio between the real surface area and the plan projected area of a pixel. The SAGA wetness index is similar to the topographic wetness index (TWI), but it is based on a modified catchment area calculation, using a formula proposed by Böhner et al. (2006). The convergence index evaluates the convergent or divergent nature of the flow passing through a cell, while flow accumulation represents the number of upstream pixels converging into a pixel.

Using stepwise MLR, the optimum spatial model for soil reaction in the surface horizon is described by the following equation:

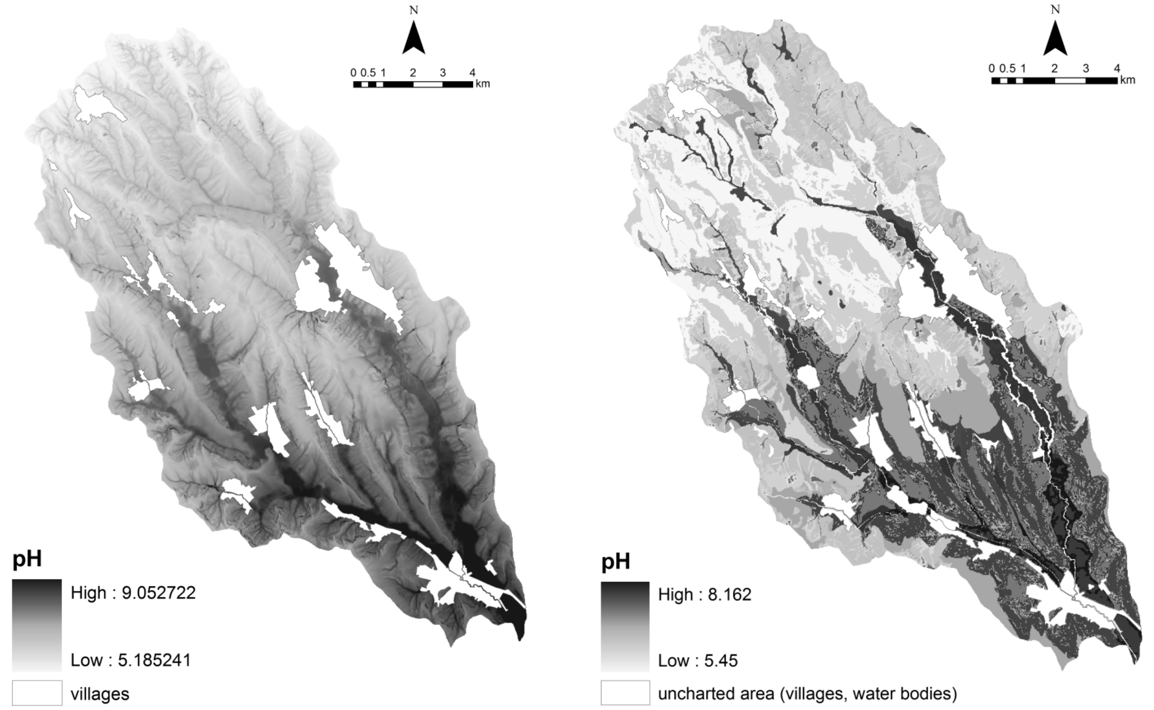
$$pH = 43.659 - 0.000075 \cdot Y - 0.00428 \cdot DEM + 0.0871 \cdot WI + 9.0076 \cdot SR$$

where WI is the wetness index and SR is the surface ratio. The model explains 56% of pH spatial variance, with a root mean square error (RMSE) of 0.585 (fig. 4).

According to the standardized regression coefficients, the predicted pH spatial distribution (fig. 2a) depends mainly on altitude and Y coordinate and, secondary, on wetness index and surface ratio. The regression equation indicates the increase of soil acidity from south to north and from lower to higher altitudes, explained by the increase of precipitations and the dominance of forests in the northern part of the basin. Also, the equation point out the increase of soils basic character in areas with high potential humidity (floodplains) and on steeper slopes (surface ratio being highly correlated to terrain slope), due to the more intense erosion processes which bring to surface the more basic soil material from underlying horizons.

For the application of CART, we tried to use the same predictors in order to facilitate the comparison of the two methods. However, the Y coordinate was eliminated in this case because it led to unrealistic north-south discontinuities in pH spatial distribution. Instead, because the method allows the ready integration of qualitative data, we tested the use of main soil types' spatial units as predictor.

The computed regression tree includes 20 terminal nodes (fig. 3), the predicted pH spatial distribution (fig. 2b) depending mainly on soil types spatial units and DEM and secondary, on wetness index and surface ratio. The first predictor partition separates basic soils (Fluvisols, Gleysols, Chernozems, Regosols) from acid soils (Phaeozems, Luvisols), then another differentiation is made between the more basic Fluvisols and Gleysols and the less basic Chernozems and Regosols, on one hand, and between the more acid Luvisols and the less acid Phaeozems, on the other hand.



a b

Fig. 2. Spatial model of soil reaction in A horizon derived by multiple linear regression (a) and classification and regression trees method (b)

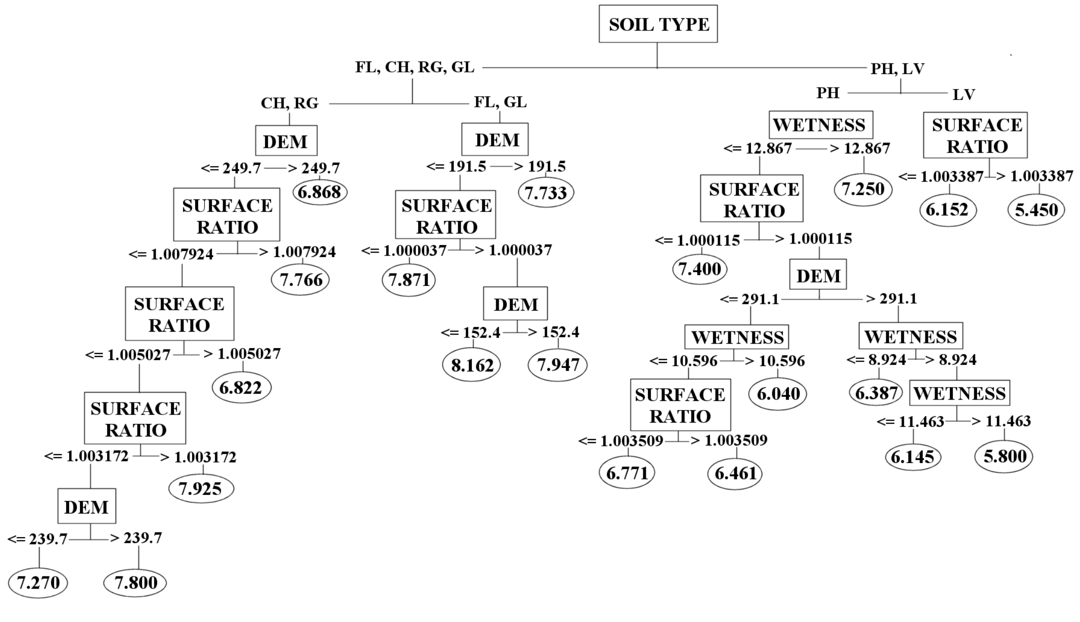
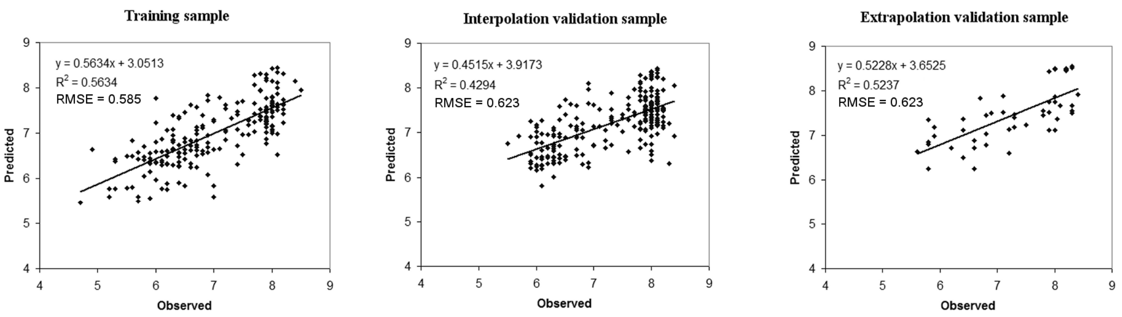
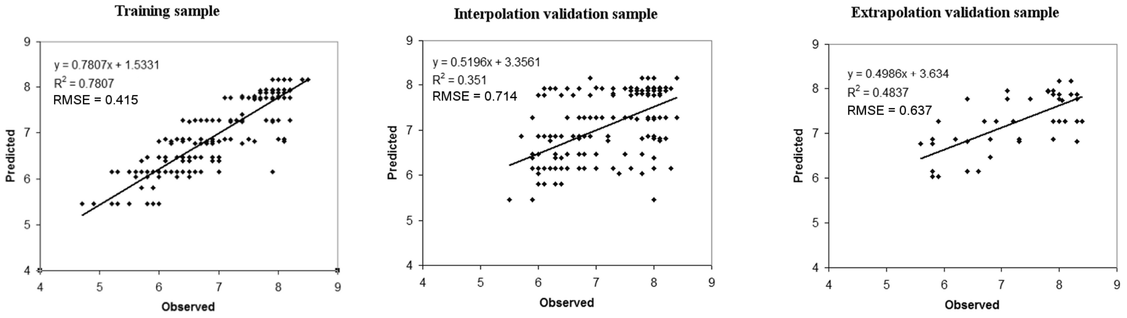


Fig. 3. The regression tree computed for soil reaction in A horizon

The correlations between observed and predicted pH values are displayed in figure 4. Apparently, CART method performs better than MLR, the explained variance being much higher ($R^2 = 0.78$) and RMSE lower (0.415), for the training sample, than the corresponding values for linear regression ($R^2 = 0.56$, RMSE = 0.585). However, the quality of the CART model drops significantly outside the training sample. For the interpolation validation sample, the explained pH variance is only 35%, with a RMSE of 0.714, while for the extrapolation validation sample, these parameters have values of 48% and 0.637 respectively. On the other hand, the regression model is much more stable, therefore more reliable, the differences among the quality parameters for the three samples being much more reduced.



Linear regression



Classification and regression trees

Fig. 4. Validation of pH spatial models

CONCLUSIONS

Our analysis shows that the classical linear regression approach is better suited than the classification and regression trees method for explaining and mapping soil reaction in the surface horizon within Dobrovăț basin. Though the correlation between observed and estimated pH values is higher in the case of CART method, for the training sample, the model is much more unstable than the regression one, its quality dropping significantly outside the training area. However, because its important advantages (non-linearity, non-parametric nature, easy integration of qualitative information), the CART approach is worth being further investigated.

Acknowledgements.This study was carried out with support from project POSDRU/89/1.5/S/49944, coordinated by "Alexandru Ioan Cuza" University of Iaşi and from project TE 172/2010, financed by the CNCSIS.

REFERENCES

1. Addinsoft. XLSTAT Tutorial. http://www.xlstat.com/en/support/tutorials/

2. Bishop, T. F. A., McBratney, A. B. and B. M. Whelan. (2001). Measuring the quality of digital soil maps using information criteria. Geoderma. 105: 93– 111.

3. Böhner, J. and T. Selige. (2006). Spatial prediction of soil attributes using terrain analysis and climate regionalization, SAGA - Analysis and Modelling Applications. Göttinger Geographische Abhandlungen, Vol.115, 130pp.

4. Breiman, L., Friedman, J. H., Olshen, R. and C. J. Stone. (1984). Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California.

5. Burrough, P. A. and R. A. McDonnell. (1998). Principles of Geographical Information Systems. Oxford University Press.

6. Cheng, X. F., Shi, X. Z., Yu, D. S., Pan, X. Z., Wang, H. J. and W. X. Sun .(2004). Using GIS spatial distribution to predict soil organic carbon in subtropical China. Pedosphere. 14(4): 425—431.

7. Cialella, A. T., Dubayah, R., Lawrence, W. and E. Levine. (1997). Predicting soil drainage class using remotely sensed and digital elevation data. Photogramm. Eng. Rem. S. 63: 171–178.

8. Cimmery, V. (2010). User Guide for SAGA (version 2.0.5). Volume 2. http://sourceforge.net/projects/saga-gis/files/SAGA - Documentation/SAGA 2 User Guide/SAGA_User_Guide_Vol2_Cimmery_version_2.0.5_20101209.pdf

9. ESRI. ArcGIS Desktop 9.3 Help. http://webhelp.esri.com/arcgisdesktop/9.3/

10. Florinsky, I. V., Eilers, R. G., Manning, G. R. and L. G. Fuller. (2002). Prediction of soil properties by digital terrain modelling. Environ. Modell. Softw. 17: 295– 311.

11. Freund, R. J., Wilson, W. J. and P. Sa. (2006). Regression Analysis. Statistical Modeling of a Response Variable. Elsevier.

12. Hastie, T., Tibshirani, R. and J. Friedman. (2001). The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer Series in Statistics. Springer-Verlag, New York.

13. Hengl, T. (2007). A Practical Guide to Geostatistical Mapping of Environmental Variables. JRC Scientific and Technical Research series. EUR 22904 EN, Luxembourg.

14. Hengl, T., Reuter, H. I. and L. Rodriguez-Lado. (2007). Digital Soil Mapping at work: interpolation of soil parameters for the Danube river basin, In Hengl, T. et al. (eds.) Status and prospect of soil information in southeastern Europe: soil databases, projects and applications. JRC Scientific and technical reports, EUR 22646 EN. pp. 129-137.

15. Hengl, T., Heuvelink, G. B. M. and A. Stein. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma. 120: 75–93.

16. Heuvelink, G. B. M. (2003). The Definition of Pedometrics, In Grunwald, S. (ed.) Pedometron, IUSS, 15: 11-12.

17. Ioniță, I. (2000). The cuesta relief from the Moldavian Plateau (in Romanian). Corson Press, Iaşi.

18. Johnston, R. J. (1978). Multivariate Statistical Analysis in Geography. Longman, New York.

19. Kutner, M. H., Nachtsheim, C. J., Neter, J. and W. Li (eds.). (2004). Applied Linear Statistical Models. 5th Edition. McGraw-Hill.

20. Lagacherie, P., McBratney, A. and M. Voltz (eds). (2006). Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science, 31, Elsevier, Amsterdam.

21. Lagacherie, P. and S. Holmes. (1997). Addressing geographical data errors in a classification tree soil unit prediction. Int. J. Geogr. Inf. Sci. 11: 183– 198.

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

23. McBratney, A. B., Odeh, I. O. A., Bishop, T. F. A., Dunbar, M. S. and T. M. Shatar. (2000). An overview of pedometric techniques for use in soil survey. Geoderma. 97: 293–327.

24. McKenzie, N. J. and P. J. Ryan (1999). Spatial prediction of soil properties using environmental correlation. Geoderma. 89: 67– 94.

25. Mendonça-Santos, M. L., Santos, H. G., Dart, R. O. and J. G. Pares. (2008). Chapter 34. Digital Mapping of Soil Classes in Rio de Janeiro State, Brazil: Data, Modelling and Prediction. In Hartemink A. E. et al. (eds.) Digital Soil Mapping with Limited Data. Springer, pp. 381-396.

26. Microimages Inc. (2000). Reference Manual for the TNT products V6.4.

27. Odeh, I. O. A. and A. B. McBratney. (2000). Using AVHRR images for spatial prediction of clay content in the lower Namoi Valley of eastern Australia. Geoderma. 97: 237-254.

28. Park, S. J. and L. G. Vlek. (2002). Prediction of three-dimensional soil spatial variability: a comparison of three environmental correlation techniques. Geoderma. 109: 117-140.

29. Patriche, C. V. (2005). The Central Moldavian Plateau between Stavnic and Vaslui Rivers. A Physical Geography Study (in Romanian). Terra Nostra Press, Iaşi.

30. Pîrnău, R. G. (2011). Land use and agricultural soils quality in Dobrovăț hydrographic basin (in Romanian). Ph.D. dissertation, “Alexandru Ioan Cuza” University of Iaşi.

31. Ryan, P. J., McKenzie, N. J., O’Connell, D., Loughhead, A. N., Leppert, P. M., Jacquier, D. and L. Ashton. (2000). Integrating forest soils information across scales: spatial prediction of soil properties under Australian forests. Forest Ecol. Manag. 138: 139-157.

32. Xie, X. L., Sun, B., Zhou, H. Z. and A. B. Li. (2004). Soil organic carbon storage in China. Pedosphere. 14(4): 491-500.

33. Ziadat, F. M. (2010). Prediction of soil depth from digital terrain data by integrating statistical and visual approaches. Pedosphere. 20(3): 361-367.