*Calibration and Reliability in Groundwater Modelling: Credibility of Modelling*
(Proceedings of ModelCARE 2007 Conference, held in Denmark, September 2007). IAHS Publ. 320, 2008.

# Grouping of soils according to their soil water retention characteristics

**B. TÓTH[1], A. MAKÓ[1], L. GUADAGNINI[2], A. AZZELLINO[2] & A. GUADAGNINI[2]**

1 *Department of Plant Production and Soil Science, Georgikon Faculty of Agriculture, University of Pannonia, Deák F. u. 16., H-8360 Keszthely, Hungary*
toth.brigitta@georgikon.hu

2 *Dipartimento di Ingegneria Idraulica Ambientale, Infrastrutture Viarie e Rilevamento, Politecnico di Milano, Piazza Leonardo da Vinci 32, I-20133 Milano, Italy*

**Abstract** Soil water retention capacity is usually estimated by pedotransfer functions which, in turn, are very problematic to apply for soil maps containing only categorical-type soil properties. Most of the soil information in Hungary is available in the format of soil maps reporting categorical-type information and often lacking data about measured hydrophysical properties. Here we are mainly concerned with the possibility of providing predictions of soil water retention categories starting from continuous- and/or categorical-type information. First we used Factor Analysis (FA) to synthesize the information contained in the soil variables. We then employed the factors extracted in the context of cluster analysis (CA) to differentiate soil groups that can be characterized by means of typical soil water retention values. Finally, we used multinomial logistic regression to address the relationship between the soil water content and the other soil properties used in FA.

**Key words** soil water retention; factor analysis; cluster analysis; multinomial logistic regression

## INTRODUCTION

Hydraulic properties that are most commonly used in the hydrophysical depiction of soils include hydraulic conductivity and soil water retention capacity. In Hungary, detailed soil maps (1:10 000 scale) covering about 60% of the agricultural areas are available (Tóth & Máté, 2006). These maps contain information about the soil type, soil subtype, physical and chemical properties (e.g. texture, organic matter and calcium carbonate content, and pH) of the soil horizons. However, they only include very limited information about the soil hydrophysical characteristics. A reliable and detailed knowledge of these characteristics is relevant, as these have significant impact on key soil properties, including soil fertility, aeration and thermal regime, drainage, irrigation and cultivability. As direct measurements of these properties are costly and labour and time intensive, an alternative which is often employed is to infer them from available soil hydrophysical characteristics using a proper PTF (pedotransfer function) (e.g. Bouma, 1989). Starting from a PTF derived on the basis of a given data set, one can further develop PTFs for homogeneous sub-sets of soil by preliminary grouping (Pachepsky & Rawls, 2005). In this context, it is very common to use the soil texture class as a preliminary grouping factor before creating the PTF (e.g. Quisenberry *et al*., 1993). A variety of site-specific PTFs is available for predicting soil water retention. A main drawback in their routine application is that they require soil information in continuous-type format (e.g. Gupta & Larson, 1979). Batjes (1996) proposed a methodology, denoted as pedotransfer rule (PTR), to be used to provide predictions of the available water content (AWC) of the soil using only categorical-type soil information, including the main soil units of the FAO-UNESCO world soil map, textural class and organic matter class. Using similar concepts, other types of PTRs have been proposed (e.g. Daroussin & King, 1996).

The possibility of predicting the soil water retention categories from soil map information would help to prepare soil drainage, irrigation or sewage-disposal maps (Tóth *et al*., 2006). Unfortunately, neither the available PTF nor the PTR can be applied in analysing the Hungarian soil map information, since for the former, not all the needed information content is available and the latter is not accurate enough for our dataset. With the distinctive aim to provide a methodology for estimating the soil water retention distribution on the basis of data included in soil maps of the type which is available in Hungary, we now focus on the relationship between the continuous-type soil information content of our database and soil water retention categories. This study is conducive to further develop a predictive model (PTR) to estimate soil water retention categories directly from categorical-type soil map information.

## MATERIALS AND METHODS

### Data set

Our study makes full use of data currently available in the Hungarian Hydrophysical Database ver. 1. (HHD ver. 1). This is the largest database available in Hungary and contains information on basic and hydrophysical soil properties, including measured water retention values, physical and chemical characteristics for 3821 horizons. As such, it provides a unique opportunity to analyse in detail the relationship between basic soil properties and water retention capacity. In the context of our study we have only used the information about the water retention and those soil properties which are available on the soil maps at the 1:10 000 scale. The latter include, among other soil properties: soil texture, soil type and subtype, organic matter and calcium carbonate content, and pH in water. In order to study the relationship between these soil properties and soil water retention, we used the following information of the database: soil type and subtype; sand (0.2–2 mm), silt, (0.002–0.2 mm) clay (<0.002 mm), organic matter and calcium carbonate content; $pH_{(H2O)}$; and water retention values at pF 0, pF 2.5, pF 4.2 and pF 6.2. Soil types and subtypes are basic classification units of the Hungarian soil taxonomic classification system. The units are classified according to similar material and energy regimes. There are 40 soil types including 87 subtypes in the classification system (MÉM, 1989). The soil analyses were performed in compliance with standard methods (particle size distribution: Gee & Bauder, 1986; $CaCO_3$%: Nelson, 1982; $pH_{(H2O)}$: McLean, 1982; organic carbon content: Tyurin, 1931). The pF values were measured by the "sand box method" (sand/kaoline box and membrane apparatus) according to the Hungarian standard (Várallyay, 1973). The data set contains the particle size distribution, calcium carbonate and organic matter content, pH and the retention values in continuous-type format.

### Statistical analysis

We started by performing multivariate statistical analysis of the continuous soil variables (i.e. sand, silt, clay, organic matter and calcium carbonate content and pH), with the objective of providing further insights on the relationship between easily available soil properties and the soil water retention capacity. We performed a preliminary PCA (Principal Component Analysis), which resulted in coming up with five components accounting for 100% of the total variance. The components extracted through PCA were then rotated (i.e. performing FA – Factor Analysis) by using a VARIMAX rotation with Kaiser normalization criterion (Norusis, 1994). On the basis of the Factor loadings it was then possible to identify the most significant variables within each factor and to quantify the degree of correlations between the raw variables. The factors extracted from the soil properties were used as input variables for a subsequent CA (Cluster Analysis). We employed a K-means CA technique to differentiate five groups that could be characterized by similar soil properties. We then analysed the water retention properties of each identified group. Since our aim is to provide a workable method to predict categorical-type water content information, available water content data at different pF values (i.e. pF 0, pF 2.5, pF 4.2 and pF 6.2) were transformed according to a 4-category measurement scale by using quartiles as segmentation criterion. The key characteristics in water content of the five clusters were then studied upon overlaying the categorical water content variable to the identified clusters. Finally, MLR (Multinomial Logistic Regression) was employed to analyse the relationship between the soil water retention categories and the other soil properties. Since only three out of the five clusters provided by K-means CA displayed well recognizable and distinct water content characteristics in terms of the four available pF values, a 3-clusters classification was used as response variable of the MLR analysis. MLR analysis allows predicting the soil cluster upon using basic soil information. The advantage of this technique is that it is applicable for any combinations of discrete and continuous variables (Afifi & Clark, 1998). As basic soil information both continuous (i.e. silt, clay, sand, organic content and pH) and categorical (i.e. soil subtype codes) predictors were used.

## RESULTS

### Factor analysis

Five factors were extracted from the six variables, accounting for almost 100% of the variance (Table 1). Factor I (FI) is informative of about 43.1% of the total variance of the system; Factor II (FII) and Factor III (FIII) explain 28.1% and 12.8% of the total variance, respectively; Factor IV (FIV) and Factor V (FV) provide minor contributions (they account for 10.3% and 5.7% of the total variance, respectively). Table 1 shows the FA factor loadings.

In the light of FA results, to avoid multicollinearity, silt content is, in fact, correlated with clay and sand content and has the lowest correlation with the water content.

**Table 1** Key results of the Factor Analysis for the continuous soil variables. Factor loadings have been omitted when less than 0.4 (in absolute value).

| Variable | FI | FII | FIII | FIV | FV | Communalities |
|---|---|---|---|---|---|---|
| Sand content | –0.78 | –0.60 | – | – | – | 1.00 |
| Silt content | 0.97 | – | – | – | – | 1.00 |
| Clay content | – | 0.95 | – | – | – | 1.00 |
| Organic matter | – | – | 0.98 | – | – | 1.00 |
| Calcium carbonate | – | – | – | – | 0.94 | 1.00 |
| pH in water | – | – | – | 0.96 | – | 1.00 |
| % of variance | 43.1 | 28.1 | 12.8 | 10.3 | 5.7 | |
| Cumulative % of variance | 43.1 | 71.2 | 84.0 | 94.3 | 100.0 | |

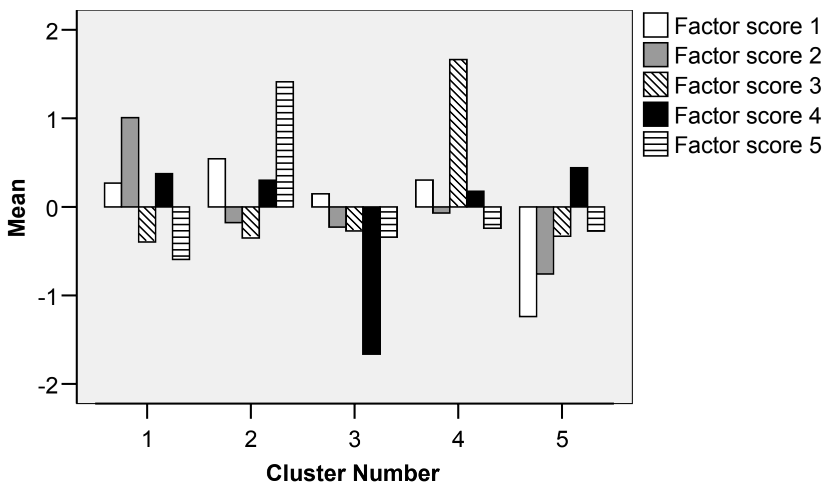


**Fig. 1** The soil properties of the clusters characterized by the calculated factor scores.

### K-means cluster analysis

Five homogeneous soil groups were identified by means of K-means CA, on the basis of the factor scores of the continuous-type soil properties which are available from the soil maps. Figure 1 depicts the soil characteristics of each cluster, comparing the cluster mean values with the global mean (i.e. the global mean of the data set). Since the factors are standardized, "0" is the global mean and deviations from zero indicate values which are larger and/or lower than the global mean. Cluster number 1 is characterized by high clay and low sand and calcium carbonate content. Cluster number 2 contains soils with high silt but low sand and high calcium carbonate content and pH. Soil samples characterized by low pH and calcium carbonate content belong to cluster 3. Cluster number 4 mainly groups soils with high organic matter content. Cluster number 5 contains soils with high sand, low silt and clay content and relatively low organic matter content but with pH higher than the global mean.

Figure 2 depicts the distribution of the retained water content for different pF values within each identified cluster. It appears quite clear that some clusters display a relatively homogeneous distribution of water content values (e.g. cluster 1 and 2 with respect to water content at pF 0 and clusters 2 and 3 for the water content at both pF 2.5, 4.2 and 6.2). On the basis of these results obtained by the analysis for five clusters we re-run the clustering and it was concluded that a 3-clusters classification was more reasonable. This classification was then used as input to the MLR. Figure 3 shows the water content characteristics of the identified 3 clusters.

## Multinomial logistic regression

Table 2 reports the results of the MLR analysis for the prediction of the three soil clusters. Albeit variable, the model estimation efficiency (defined as the proportion of correct predictions over the

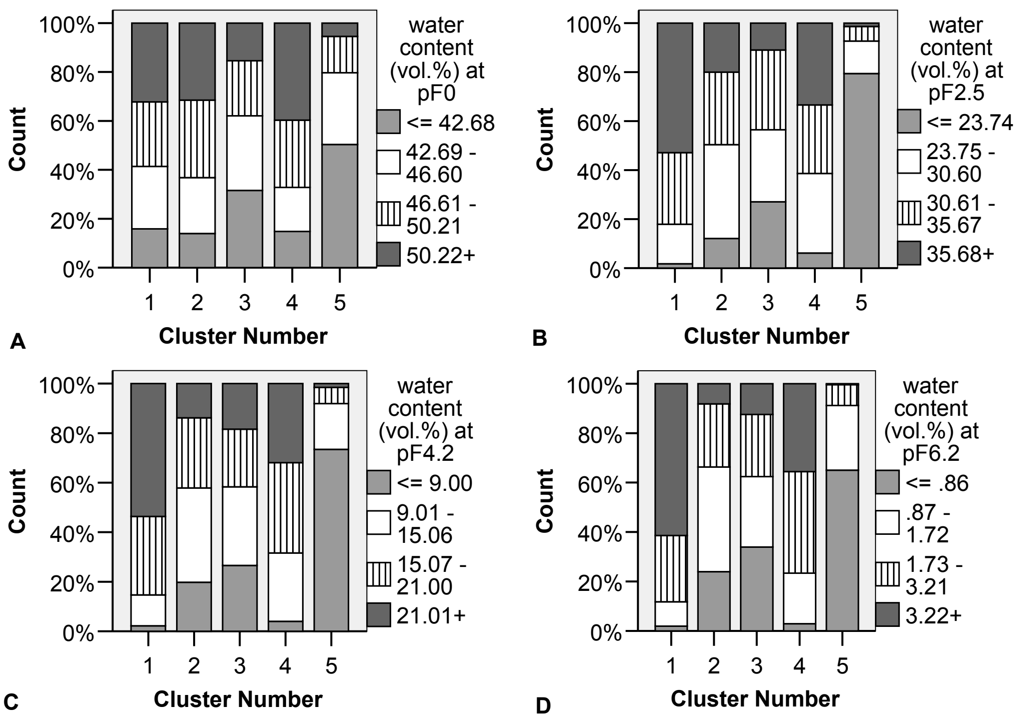


**Fig. 2** Water retention distribution within the clusters (A: pF 0; B: pF 2.5; C: pF 4.2; D: pF 6.2).

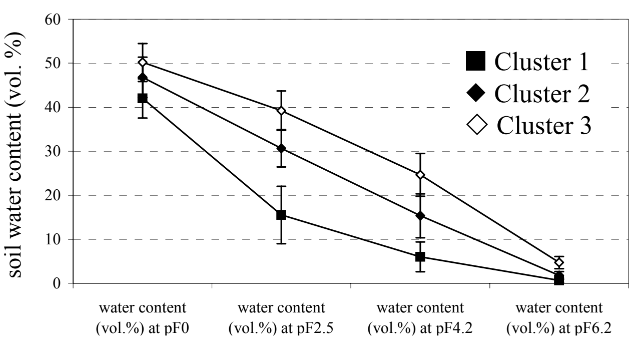


**Fig. 3** Water retention characteristics (at pF 0, pF 2.5; pF 4.2, pF 6) for the adopted 3-clusters solution. Mean curve + / – one standard deviation.

total number of cases analysed) is quite large for all the clusters. We note that the model reacted by associating continuous predictors with some soil subtype codes. Cluster 2, which is characterized by the lowest water content at all pF values, is directly correlated with the sand content and with the presence of some subtype codes (i.e. 31, 132 and 381 (MÉM, 1989)) and inversely correlated with clay and organic matter. Cluster 3, which has the highest water content, is directly correlated with clay, organic matter, pH and subtype code 395, and inversely correlated with sand, calcium carbonate and the presence of subtype code 402.

**Table 2** Summary of the results of the Multinomial Logistic Regression. Only significant predictors ($P<0.05$) are shown; (+) indicates a direct correlation, (–) indicates inverse correlation.

| | | Soil properties entered into the MLR (soil codes are defined with the Hungarian classification system (*MÉM*, 1989) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Cluster number | Clay | Sand | Organic matter | $CaCO_3$ | pH | Subcode 31 | Subcode 132 | Subcode 381 | Correct Predictions (%) |
| 1 *vs* 2 | 1 | (+) | (–) | (+) | | | (–) | (–) | (–) | 87.8 |
| | 2 | (–) | (+) | (–) | | | (+) | (+) | (+) | 82.2 |
| | | Clay | Sand | Organic matter | $CaCO_3$ | pH | Subcode 395 | Subcode 402 | | Correct Predictions (%) |
| 3 *vs* 2 | 3 | (+) | (–) | (+) | (–) | (+) | (+) | (–) | | 78.2 |
| | 2 | (–) | (+) | (–) | (+) | (–) | (–) | (+) | | 82.2 |

This model suggests that the clay, sand, organic matter content, calcium carbonate and pH influence the soil water retention. The effect of the calcium carbonate on the soil water retention is indirect, as it influences the soil aggregation and the stability of the soil structure. These properties, in turn, strongly influence the pore size distribution, which ultimately governs the soil water retention. Similarly, while soil pH does not have a direct effect on soil water retention, it is strongly connected to the soil subtype. Similarly to what was noted for pH, it is reasonable to argue that the information of other continuous predictors might be contained in other soil subtypes.

A MLR analysis was then performed upon adopting only the soil subtype as predictor. The results of the model (not shown) clarify that only taking into account the information about the soil subtype allows the correct discriminate slightly less than 90% of soils belonging to cluster 1 (i.e. intermediate water content) from cluster 2 (i.e. lower water content). The model is characterized by low performances for the highest water content (i.e. cluster 3). However, these results appear quite promising as they indicate that acceptable predictions may be obtained upon using the categorical information available on the soil subtype. Further analysis is needed to clarify the water holding mechanism of soil subtypes.

## CONCLUSIONS

Our preliminary work leads to the following key conclusions:

– Soil grouping can be performed on the basis of soil water retention categories.
– Multinomial Logistic Regression analysis allows predicting soil water retention categories upon using basic soil information, in the format of both continuous (i.e. silt, clay, sand, organic content and pH) and categorical (i.e. soil subtype codes) variables.
– Categorical-type information can provide valuable and quantifiable information in predicting soil water retention categories.

**Acknowledgements** The present work was supported by the National Scientific Research Fund (OTKA) under grant no. T048302 and the scholarship of the Italian Ministry of Foreign Affairs.

## REFERENCES

Afifi, A. A. & Clark, V. (1996) (reprinted 1998) *Computer-Aided Multivariate Analysis,* third edn. Chapman & Hall, UK.

Batjes, N. H. (1996) Development of a world data set of soil water retention properties using pedotransfer rules. *Geoderma* **71**, 31–52.

Bouma, J. (1989) Using soil survey data for quantitative land evaluation. *Adv. Soil Sci.* **9**, 177–213.

Daroussin, J. & King, D. (1996) Pedotransfer rules database to interpret the Soil Geographical Database of Europe for environmental purposes. In: *The Use of Pedotransfer in Soil Hydrology Research in Europe. Workshop proceedings* (ed. by A. Bruand, O. Duval, J. H. M. Wösten & A. Lilly) Orleans, France, 10–12 Octobre 1996, 25–40.

Gee, G. W. & Bauder, J. W. (1986) Particle-size analysis. In: *Methods of Soil Analysis*. Part 1. (ed. by A. Klute), 383–412. American Society of Agronomy, Inc. Soil Science Society of America, Inc. Madison, Wisconsin. USA.

Gupta, S. C. & Larson, W. E. (1979) Estimating soil water retention characteristics from particle size distribution, organic matter percent, and bulk density. *Water Resour. Res.* **15**(6), 1633–1635.

McLean, E. O. (1982) Soil pH and lime requirement. In: *Methods of soil analysis*. Part 2. (ed. by A. L. Page), 199–224. American Society of Agronomy, Inc. Soil Science Society of America, Inc. Madison, Wisconsin, USA.

MÉM (Mezőgazdasági és Élelmiszerügyi Minisztérium) (1989) *Útmutató a nagyméretarányú országos talajtérképezés végrehajtásához. '88 melléklet.* (Guide to the implementation of the national large scale soil survey, '88 annex). Agroinform, Budapest (in Hungarian).

Nelson, R. E. (1982) Carbonate and gypsum. In: *Methods of Soil Analysis.* Part 2. (ed. by A. L. Page), 181–197. American Society of Agronomy, Inc. Soil Science Society of America, Inc. Madison, Wisconsin, USA.

Norusis, M. J. (1994) *SPSS Professional Statistics – User's Manual.* SPSS Inc., Chicago, USA.

Pachepsky, Y. & Rawls, W. J. (2005) *Development of Pedotransfer Functions in Soil Hydrology*. Developments in Soil Science, 30, Elsevier, Amsterdam, 512, The Netherlands. ISBN 0-444-5175-7.

Quisenberry, V. L., Smith, B. R., Philips, R. E., Scott, H. D. & Nortcliff, S. (1993) A soil classification system for describing water and chemical transport. *Soil Sci.* **156**, 306–315.

Tóth, G. & F. Máté (2006) Comments for developing a national soil information system. (Megjegyzések egy országos, átnézetes, térbeli talajinformációs rendszer kiépítéséhez.) *Agrokémia és Talajtan* **55**(2), 473–478 (in Hungarian).

Tóth, B., Makó, A., Rajkai, K., Szabóné, K. G., Hermann, T. & Marth, P. (2006) Use of soil water retention capacity and hydraulic conductivity estimation in the preparation of soil water management maps. *Agrokém. és Talajtan* **55**, 49–58.

Tyurin, I. V. (1931) A new modification of the volumetric method of determining soil organic matter by means of chromic acid. *Pochvovedenie* **26**, 36–47 (in Russian).

Várallyay, Gy. (1973) Soil moisture potential and new apparatus for the determination of moisture retention curves in the low suction range (0–1 atmospheres). *Agrokémia és Talajtan* **22**(1–2), 1–22 (in Hungarian with English summary).