Application of *k*-Nearest Code for the Improvement of Class Pedotransfer Functions and Countrywide Field Capacity and Wilting Point Maps

Markéta MIHÁLIKOVÁ, Svatopluk MATULA and František DOLEŽAL

Department of Water Resources, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, Prague, Czech Republic

Abstract

Miháliková M., Matula S., Doležal F. (2014): **Application of *k*-Nearest code to the improvement of class pedotransfer functions and countrywide Field Capacity and Wilting Point maps**. Soil & Water Res., **9**: 1–8.

The database of soil hydrophysical properties in the Czech Republic (HYPRESCZ) which contains the data needed for derivation of pedotransfer functions for soil water retention curves was used for the estimation of field capacity and wilting point of agricultural land resource on a countrywide scale. The results were combined with the existing Soil Texture Map of the Czech Republic to create four new maps, namely the Map of Field Capacity and the Map of Wilting Point for the topsoil and subsoil separately. From the total number of 1048 relevant database entries, only about a half included reliable wilting point data. The *k*-Nearest computer code employing the *k*-Nearest neighbour technique was used for estimation of the missing wilting points, which made it possible to use all entries. The estimation uncertainty was assessed in terms of standard deviations and the root mean square error. Finally, two sets of class pedotransfer functions were derived and found sufficiently comparable: (*i*) the functions estimating the soil water retention curve in the whole range, derived solely from the database entries containing the measured wilting points, and (*ii*) the functions estimating the field capacity and wilting point only, derived from all database entries, including the *k*-Nearest neighbour estimated data.

Keywords: field capacity; *k*-Nearest neighbour technique; pedotransfer functions; wilting point

We are currently witnessing a climate change which affects both ecological and economical aspects of our lives. The age of water scarcity is coming and an efficient improvement of landscape water management is necessary, including a better utilization of soil water.

The area of the Czech Republic (CR) due to its geographical position with drainage divided into three seas is fully dependent on precipitation: an inflow to the area is only about 3–6% of a total runoff (Ministry of Agriculture of the Czech Republic 2011). Agricultural land resource in the CR is currently about 54% of the whole area, with decreasing tendency. Arable land makes about 38% of the whole area; more than half is registered as Less Favoured Area. Water demand in agriculture is supposed to increase due to changes in precipitation distribution during a year on the account of a climate change (Cenia 2012). On the other hand and due to a climate change, disastrous flood events occurred during last two decades, both river-floods and flash-floods. Both technical and organizational prevention of flood events is of high importance. This study is supposed to contribute to the assessment of present situation and modelling of future scenarios.

Both agricultural and hydrological models need input information about soil water retention; at least for the basic hydrolimits like field capacity (FC) and wilting point (WP). These hydrolimits are described well in the literature, i.e. in Cassel and Nielsen (1986). When both FC and WP are known, the available soil water capacity can be calculated as their difference.

Direct measurement of the soil water retention data on a regional scale is still difficult, despite the recent progress in measurement methods. Pedotransfer functions (PTFs) have therefore become a useful alternative to expensive and time consuming measurements (e.g. McBratney *et al*. 2011).

PTFs based on grouping of the soils according to their particular characteristics were called "class PTFs" by Wösten *et al.* (1990). Class PTFs estimate average hydrophysical properties for certain defined soil groups, often for soil textural classes. However, more input data in addition to the texture may not imply in better results as it is reviewed by Wösten *et al.* (2001). They showed, too, that groupings according to genetic soil types or similar functional behaviour of different soil horizons can be more efficient. However the specification of local soil conditions in the case of class PTFs is limited compared to continuous PTFs. The characteristic of the defined soil group is the only input to the class PTFs, thus the whole set of PTFs is presented in one table. Class PTFs rather than continuous PTFs are being employed to assess the hydrophysical data in regional studies.

An alternative approach to the commonly used methods of PTFs derivation (such as regression methods or the artificial neural networks) is to use non-parametric techniques, such as the *k*-Nearest neighbour (*k*-NN), introduced in this context by Nemes *et al.* (2006a). Non-parametric techniques use pattern-recognition and data similarity instead of fitting the data by equations. The *k*-NN technique belongs to the "lazy learning" algorithms. It locates and retrieves the most similar objects (the nearest neighbours) in the reference data set to a target object. The development (reference) data set is stored as such until the time of performing calculations. It means, that new data can be simply added to the stored data.

The *k* denotes the number of entries (soils) which are selected as the nearest and then used for estimation of the target property. The value of *k* and the criterion for selecting the nearest neighbours have a large impact on the quality of the results.

The code *k*-Nearest was developed by Nemes *et al.* (2008). It employs the *k*-NN technique for estimation of the volumetric soil water content at matric potentials –33 kPa and –1500 kPa, associated with FC and WP, from contents of sand, silt, and clay and optionally from the bulk density and/or organic matter content.

The code optionally uses a "bootstrap" technique, a non-parametric method that generates alternative (replica) data subsets from a single data set. The replica data sets are of the same size as the original reference data set. Estimations of soil water contents for required soils with known predictors are then made using these replica data sets: each from the number of replica data set makes estimations for each required soil. The resulting estimation for a particular soil is then calculated as the average of the bootstrap estimations for that soil. The main benefit of it is that the standard deviation (SD) as a measure of estimating uncertainty can be calculated even when no validation data are available (Nemes *et al.* 2006b).

The objectives of this study are: (*i*) to utilize all relevant and available data from the CR for deriving the most appropriate class PTFs for estimation of FC and WP, (*ii*) to evaluate whether the *k*-Nearest code is a suitable tool for the estimation, (*iii*) to create new maps of FC and WP for the CR, based on these class PTFs and the existing soil maps.

MATERIAL AND METHODS

The data contained in the HYPRESCZ database (Matula *et al.* 2010; Miháliková *et al.* 2013) were used in this study. The HYPRESCZ is the database of soil hydrophysical properties in the CR and its structure is based on that of the HYPRES (HYdraulic PRoperties of European Soils) database (Wösten *et al.* 1999). The database gathers available data from many regions throughout the CR, needed for derivation of PTFs for soil hydrophysical characteristics estimation.

The data stored in HYPRESCZ are of different quality and completeness. For the purpose of this study, all data actually available were assessed, with the result that 1048 database entries (including replicates) were found suitable. Besides the measured soil water retention curves, the following predictors were required: particle size distribution according to FAO/USDA categories (content of clay below 0.002 mm, silt between 0.002 and 0.05 mm, and sand between 0.05 and 2 mm), dry bulk density, and organic matter content. Moreover, geo-referencing information was required and at least crude soil type and horizon classification. The soil water retention curves were mostly measured in a sand or kaoline box and a pressure plate apparatus. Figure 1 shows the distribution of the relevant HYPRESCZ database entries in a textural triangle.

Five groupings of USDA texture classes defined by Němeček *et al.* (2011) were used as classes of PTFs in this study as they were used for mapping units in Soil Texture Map of the Czech Republic (Kozák *et al.* 2010). See Figure 1 for distribution of available data in the textural triangle.

First, the procedure of derivation of the class PTFs was used, which is described in details by Wösten *et al.* (1999). In this study, FC was associated with the matric head h_{FC} = –50 cm to be comparable with HYPRES (Wösten *et al.* 1999). The WP was associated with h_{WP} = –15 000 cm. However, there were only 560 database entries from which both FC and WP could

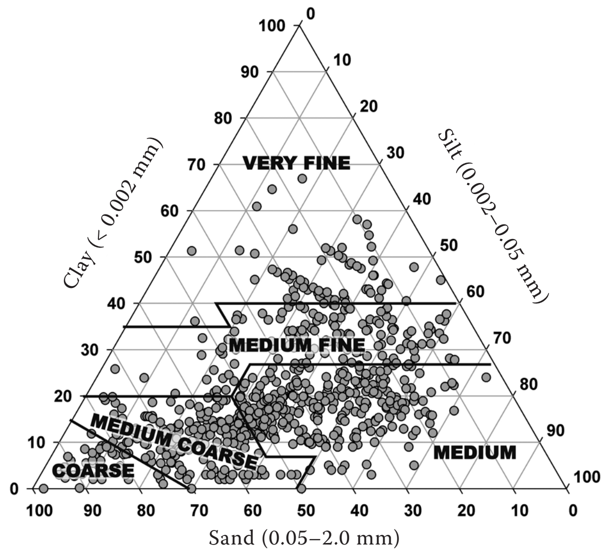

Figure 1. Distribution of available data entries within the texture triangle; groupings of USDA texture classes are according to Němeček *et al.* (2011)

be obtained in this way, because the remaining 488 retention curves in HYPRESCZ were incomplete. These "short" curves were only measured up to a matric head lower (in the absolute value) than $|h|$ = 15 000 cm, while they still possessed all other necessary predictors. They could be used for FC estimation but not directly for WP estimation, because the retention curve extrapolation with the van Genuchten Eq. (1) (van Genuchten 1980) outside the range of measurement is generally not recommended.

$$\theta = \theta_r + \frac{(\theta_s - \theta_r)}{\left(1 + (\alpha|h|)^n\right)^{1-1/n}} \quad (1)$$

where:

$|h|$ – absolute value of the actual matric head (cm)

θ – actual soil water content (cm^3/cm^3)

θ_r – model parameter expressing the residual soil water content (cm^3/cm^3)

θ_s – model parameter expressing the saturated soil water content (cm^3/cm^3)

α – shape factor (1/cm)

n – shape factor (–)

On the other hand, the set of the "long" retention curves from the regular 560 entries was not representative enough; especially the Medium Fine and Very Fine textural groupings were poorly represented. The use of information contained in the remaining 488 entries could substantially improve the representativeness of the database (Figure 2).

Hence, the *k*-Nearest code (Nemes *et al.* 2008) was used as appropriate method for estimation of the missing 488 WPs. The code works with two data files: a reference data set and an application data set. The reference data set, like in the regression method of the PTFs estimation, contains both the predictors and the measured points of retention curves. An original reference data set, distributed with the software, contains data for North American soils. The FC included is a water content associated with a matric potential of –33 kPa. It was replaced by a new reference data set containing the chosen suitable data from HYPRESCZ database, where FC is associated with the matric head of –50 cm. At the start of the work, the application data set is filled in by the user with the predictors of the soils to be estimated. The *k*-Nearest software makes the addition of newly estimated data and experimentation with different reference data sets easy and very fast. The SD of the estimates is obtained by the bootstrap technique.

Coincidently with the missing WPs, the values of FC were estimated, too, and compared with known measured data, thus the reliability of the estimations was evaluated in terms of SD. Root Mean Squared Error (RMSE) for the FC estimating reliability was calculated as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\theta_{meas} - \theta_{est})^2} \quad (2)$$

where:

θ_{meas} – measured soil water content at FC (cm^3/cm^3)

θ_{est} – soil water content estimated by the *k*-NN method (cm^3/cm^3)

N – number of data pairs

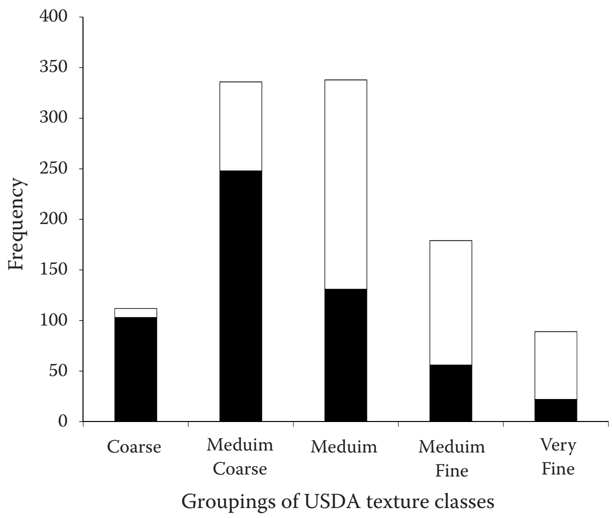

Figure 2. Frequency of data entries within particular groupings of USDA texture classes; the numbers of incomplete entries with missing (and later estimated) wilting points are hatched

For evaluation of estimating uncertainty, the terminology introduced by Wösten *et al.* (2001) is very useful. The estimating accuracy is evaluated if measured and estimated data come from the same data set, and the estimating reliability is evaluated if measured and estimated data come from independent data sets.

Finally, all available data, measured and estimated (altogether 1048 entries), were assorted into the five groupings of USDA texture classes (Figure 1). All data were further processed as data set for class PTFs derivation. Moreover, the common class PTFs were also derived from those 560 complete database suitable entries (using the Wösten's methodology mentioned above) and results were compared.

RESULTS AND DISCUSSION

Firstly, class PTFs were derived from the dataset of 560 entries only ("long" measured retention curves up to the WP). The class PTFs were derived for the whole range of the soil water retention curve (see fitted van Genuchten parameters in Table 1).

Secondly, class PTFs were derived from both measured and *k*-NN-estimated data (1048 entries together). Only values of the FC and the WP were derived for five groupings of USDA texture classes, for topsoil and subsoil separately. These values were finally used for mapping. They were ascribed to particular mapping units of the existing maps. For the topsoil, the texture of the layer of 0–25 cm according to the Soil Texture Map (Kozák *et al.* 2010) was considered. For the subsoil, the texture of the layer at about 75 cm below the topsoil bottom was taken from the Map of Soil Parent Materials (*ibid.*).

The accuracy and reliability of estimation were evaluated. The estimation accuracy in terms of the average SD is 0.021 cm^3/cm^3 for FC and 0.023 cm^3/cm^3 for WP.

For a better insight into the distribution of the SD values obtained, we plotted their histograms (Figure 3). Most SD values for both FC and WP lie in the intervals from 0 to 0.02 and from 0.02 to 0.04 cm^3/cm^3.

The average estimation reliability (calculated by the bootstrap technique) in terms of the average SD was found to be 0.015 cm^3/cm^3 for FC and 0.018 cm^3/cm^3 for WP. As shown in Figure 4, most values of SD for both FC and WP still lie in three most left intervals from 0 to 0.03 cm^3/cm^3 but, in contrary to Figure 3, the frequency distributions are not monotonously decreasing and show modes between 0.02 and 0.03 cm^3/cm^3.

The average standard deviations of the estimated FC and WP values for individual groupings of the USDA soil texture classes based on all 1048 database entries were investigated, too (Figure 5). The SD for the FC varies independently, but the SD for WP is significantly higher for very fine grouping of USDA soil texture classes. It is caused by less data representing this grouping (Figure 2). Thus, the estimation of the WP in heavy soils is supposed to be less reliable than for medium and coarse soils. Within the coarse grouping of USDA texture classes, higher SD can be observed for both the FC and the WP. This is caused, on the contrary, by very few estimated data.

The estimating reliability was assessed in terms of RMSE. The measured and independently estimated FCs from the data set with missing WPs but measured FCs were evaluated. The total RMSE was 0.08 cm^3/cm^3 and when assessed within particular groupings of USDA soil texture classes, it was distributed evenly.

Table 1. Class PTFs for the retention curves (van Genuchten parameters) of the CR derived for groupings of USDA texture classes

Texture class		θ_s	θ_r	α	n
Topsoil	Coarse	0.400	0.111	0.0265	1.3911
	Medium Coarse	0.384	0.035	0.0375	1.1798
	Medium	0.396	0.010	0.0103	1.1526
	Medium Fine	0.444	0.010	0.0508	1.0899
	Very Fine	nd	nd	nd	nd
Subsoil	Coarse	0.375	0.071	0.0657	1.3383
	Medium Coarse	0.355	0.010	0.0276	1.1641
	Medium	0.370	0.010	0.0072	1.1939
	Medium Fine	0.395	0.010	0.0066	1.1321
	Very Fine	0.405	0.010	0.0023	1.1328

nd – not determined due to lack of data

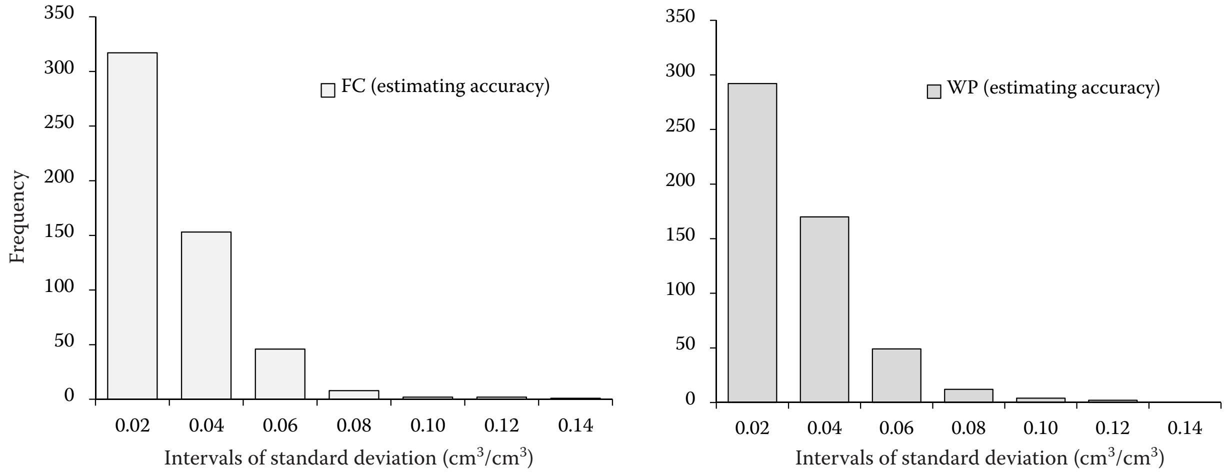

Figure 3. Distribution of SD characterizing the accuracy of field capacity (FC) and wilting point (WP) estimation

The RMSE of FC estimating reliability is significantly higher than SD. However, based on observation of SD can be declared that RMSE of the WP estimating reliability would be more or less very similar to the FC estimation. It was concluded that the estimations are reliable with satisfactory deviation and can be used for mapping in a regional study.

The performance of data set where data were obtained partly with using *k*-NN technique was compared with performance of data set analyzed with class PTFs only. The comparison was divided into five groupings of USDA texture classes and separately for topsoil and subsoil layers (Figure 6).

The differences between the two sets of PTFs are small. The average absolute difference is 0.013 cm^3/cm^3 for the FC and 0.021 cm^3/cm^3 for the WP. These differences can be caused e.g. by different sizes of the data sets. For example, there were not enough retention curves data measured up to –15 000 cm for the topsoil of the Very Fine textural grouping, and thus the class PTFs could not be derived in the usual way for this class. This problem was solved by employing the *k*-NN technique for the estimation of missing data. In general, we can conclude that the application of the *k*-NN algorithm was successful and very little additional error was introduced, while a much larger and more representative data set could be used.

Finally, we compared the newly created maps of FC and WP with two previous attempts to map the retention capacity of the soils of the CR. There are two such maps: the Map of the Water Retention Capacity of Soils in the Czech Republic (published in Hrnčiarová *et al.* 2009, the methodology described in Novák *et al.* 2008a) and the Map of the Available Water Capacity (Novák *et al.* 2008b). The former map shows "the amount of water in mm that can be held

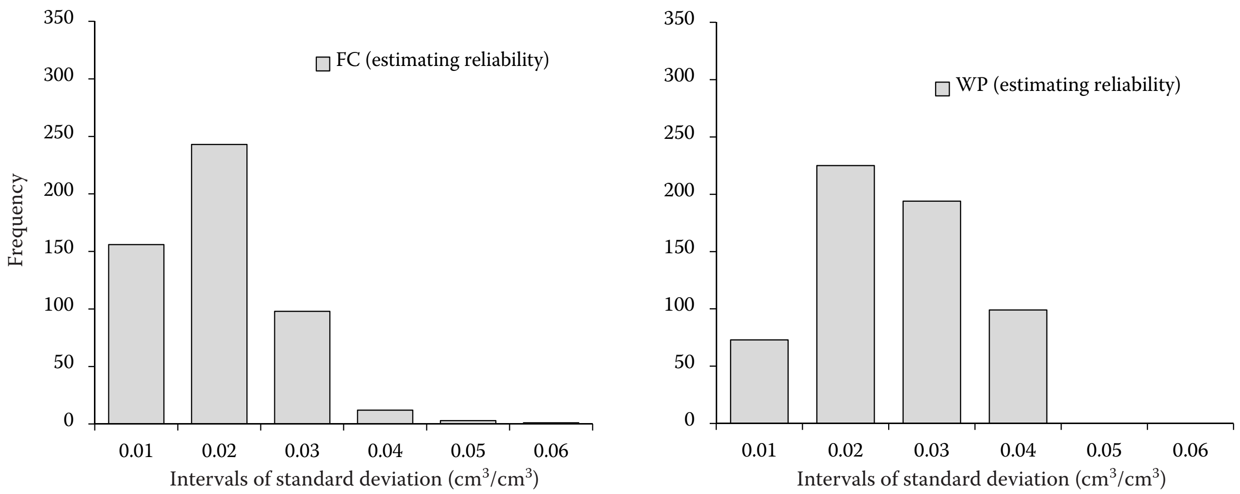

Figure 4. Distribution of SD characterizing the reliability of field capacity (FC) and wilting point (WP) estimation

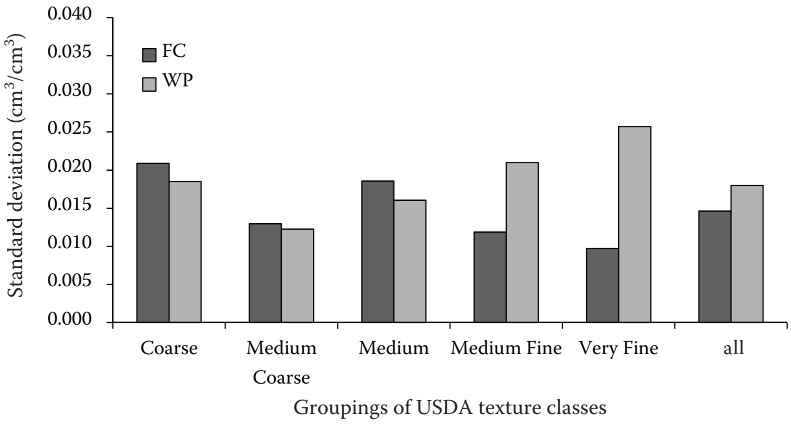

Figure 5. Average standard deviations (SDs) of estimating reliability within each particular grouping of USDA texture classes and for all data together

in the soil profile after ample rains". It corresponds to the field capacity. The Map of the Available Water Capacity estimates the storage of available water in mm for the entire soil profile. Both maps are based on the maps of the Ecological Site Valuation Units, which arose between 1974 and 1980 as a result of assessing all agricultural land resource in the country for the purpose of creating its catalogue prices. A code was assigned to each particular area containing the basic information about climate, soil type, parent material, slope steepness and exposition, skeleton content, and profile depth. The information about soil texture was not explicitly coded. The values of FC and WP were estimated for particular mapping units based on an expert guess, without explicit recourse to soil water retention curves. The authors of these maps estimate their uncertainty at about 25 and 30%, respectively. On the contrary, the maps created in this study estimate FC and WP as points on the soil water retention curve and are based on measured data. The approximate knowledge of the two hydrolimits for two different depths, depicted altogether by four

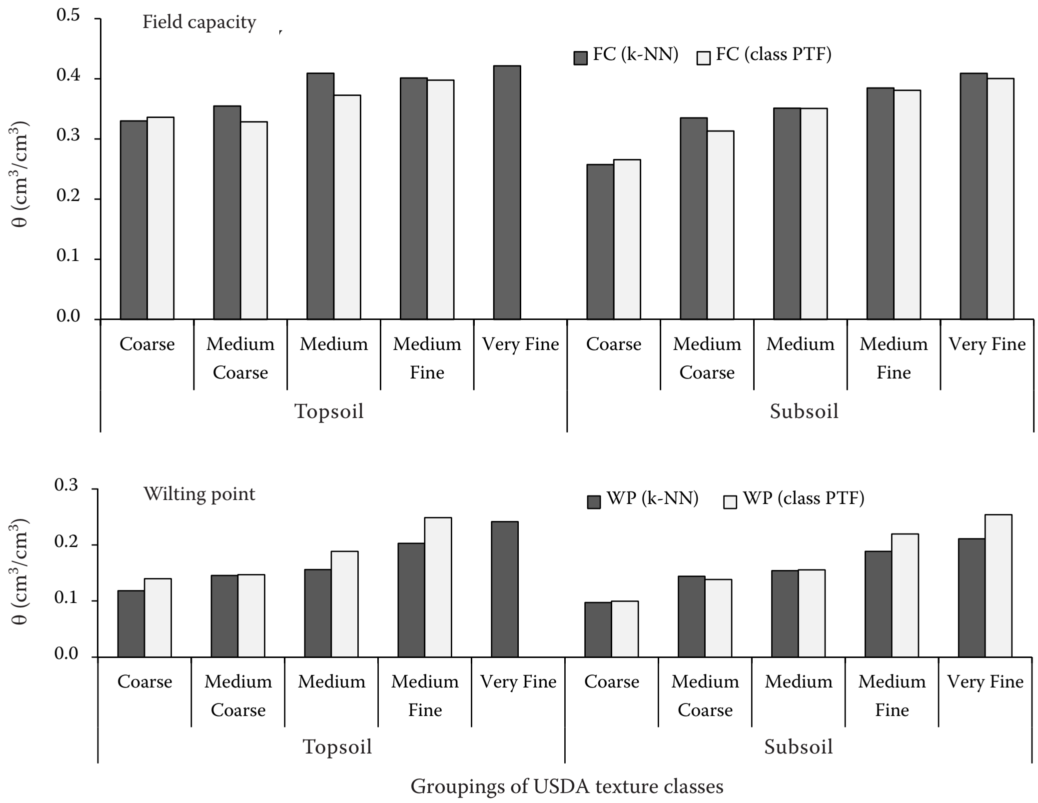

Figure 6. Comparison of two data sets performance: data of FC (a) and WP (b) obtained partly by *k*-NN technique (*k*-NN) and data obtained by class PTFs only (class PTF) (values are indicated for topsoil and subsoil separately; values for "Very Fine" grouping of USDA soil texture classes, topsoil, were not estimated by class PTFs due to lack of data)

separate new maps, allows an overall characterization of any command area and can provide useful inputs to models. When the previous maps were compared with the new ones, it was found out that the areas identified on the dry and the wet ends of the legend are similar. However, the previous maps depicted the soils in the CR as somewhat more apt to drought than the newly produced maps do. For example, the maps created by the present study do not assume FC lower than 30% vol. (in the topsoil) or 2% vol. (in the subsoil) at any place, while the Map of the Water Retention Capacity (Hrnčiarová *et al.* 2009) labelled the corresponding range (250–300 mm of water in a 1000 mm thick soil profile) as a higher medium water retention capacity. This difference can be attributed to different sources of data and different inference procedures used, as well as to different definitions of the field capacity.

The maps created in this study were certified by the Ministry of Agriculture of the Czech Republic for use in public administration (Matula *et al.* 2011) and are available online at http://katedry.czu.cz/en/kvz/.

CONCLUSIONS

The available soil water retention data from the HYPRESCZ database (Matula *et al.* 2010; Miháliková *et al.* 2013) were made more complete with the help of the *k*-NN technique as implemented in the *k*-Nearest code (Nemes *et al.* 2008). The code was employed for estimation of FC and WP in terms of volumetric soil water contents at pressure heads of –50 cm and –15 000 cm, respectively. The estimations were reliable with acceptable uncertainty, which was demonstrated in terms of SD and by comparison with the class PTFs derived as usual. In this way, more database entries than otherwise, including the incomplete retention curves, could be utilized for mapping.

The resulting class PTFs were combined with the existing soil texture maps (Kozák *et al.* 2010): the Soil Texture Map of the Czech Republic (for topsoil from 0 to 25 cm) and the Map of Soil Parent Materials (for subsoil from 25 to 100 cm). As a result, new maps of the CR were obtained, Map of the Field Capacity and Map of the Wilting Point in topsoil and subsoil layer. The maps cover the whole area of the CR but their validity is limited to agricultural land resource.

These new maps also compare well with the previous soil water retention capacity maps of the country (Novák *et al.* 2008a, b; Hrnčiarová *et al.* 2009) but suggest that the previous maps underestimate the retention capacity to some extent.

Acknowledgements. This study was supported by the Ministry of Agriculture of the CR, Project No. 1G58095, and by the Ministry of Education, Youth and Sports of the CR, Project No. 6046070901. The authors are grateful to their project partners from Department of Soil Science and Soil Protection (Czech University of Life Sciences Prague) for cooperation within this study, namely J. Janků, J. Kozák, late J. Němeček, and K. Němeček. The authors are also grateful to their project partners from the Research Institute for Soil and Water Conservation in Prague, namely M. Vlčková, Z. Kulhavý, and J. Vopravil. Furthermore, the authors thank other colleagues for sharing their data and valuable comments, namely K. Báťková, V. Kuráž, M. Tesař, A. Prax, and J. Haberle.

References

Cassel D.K., Nielsen D.R. (1986): Field capacity and available water capacity. In: Klute A. (ed.): Methods of Soil Analysis: Part 1. Physical and Mineralogical Methods. 9. 2nd Ed. American Society of Agronomy – Soil Science Society of America, Madison, 901–926.

Cenia (2012): Report on the Environment of the Czech Republic in 2011. Available at http://www1.cenia.cz/www/sites/default/files/Zprava_2011.pdf (accessed March 2013). (in Czech)

Hrnčiarová T., Mackovčin P., Zvara I. *et al.* (2009): Landscape Atlas of the Czech Republic. Ministry of the Environment of the Czech Republic, Prague and The Silva Tarouca Research Institute for Landscape and Ornamental Gardening, Pruhonice.

Kozák J., Němeček J., Borůvka L., Kodešová R., Janků J., Jacko K., Hladík J. (2010): Soil Atlas of the Czech Republic. 1st Ed. Czech University of Life Sciences, Prague.

Matula S., Miháliková M., Špongrová K., Vlčková M. (2010): Predictive soil-agrohydrological models of soil water retention in the Czech Republic and their integration to EU countries databases. [Final Report on the Project of Ministry of Agriculture of the CR, Project No. 1G58095.] Czech University of Life Sciences Prague, Prague. (in Czech)

Matula S., Miháliková M., Špongrová K., Janků J., Kozák J., Němeček J., Němeček K. (2011): National Maps of Hydrolimits. Certified by Ministry of Agriculture of the Czech Republic on May 25, 2011, Czech University of Life Sciences Prague, Prague.

McBratney A.B., Minasny B., Tranter G. (2011): Necessary meta-data for pedotransfer functions. Geoderma, **160**: 627–629.

Miháliková M., Matula S., Doležal F. (2013): HYPRESCZ – database of soil hydrophysical properties in the Czech Republic. Soil and Water Research, **8**: 34–41.

Ministry of Agriculture of the Czech Republic (2011): Report 2011 on Water Management in the Czech Republic. Available at http://eagri.cz/public/web/en/mze/water/publications/blue-report/report-2011-on-water-management-in-the.html (accessed March 2013).

Němeček J., Mühlhanselová M., Macků J., Vokoun J., Vavříček D., Novák P. (2011): Taxonomic Classification System of Soils of the Czech Republic. 2nd revised Ed. Czech University of Life Sciences Prague, Prague. (in Czech)

Nemes A., Rawls W.J., Pachepsky Y.A. (2006a): Use of a nonparametric nearest-neighbour technique to estimate soil water retention. Soil Science Society of America Journal, **70**: 327–336.

Nemes A., Roberts R.T., Rawls W.J., van Genuchten M.T., Pachepsky Y.A. (2006b): Software to estimate –33 and –1500 kPa soil water retention using the non-parametric *k*Nearest Neighbour technique. Version 1.00.00. The User's Guide. Available as a part of the software.

Nemes A., Roberts R.T., Rawls W.J., Pachepsky Y.A., van Genuchten M.T. (2008): Software to estimate –33 and –1500 kPa soil water retention using the non-parametric *k*-Nearest Neighbour technique. Environmental Modelling & Software, **23**: 254–255.

Novák P., Vopravil J., Chramostová B. (2008a): Mapping of soil water holding capacity of soils in the Czech Republic. In: Pálka B. (ed.): Proc. Conf. 5th Pedological Days on Topic The Soil as a National Wealth. October 15–16, 2008, Sielnica, 217–224. (in Czech with English abstract)

Novák P., Vopravil J., Khel T. (2008b): Mapping of available holding water capacity of soils in the Czech Republic. In: Pálka B. (ed.): Proc. Conf. 5th Pedological Days on Topic The Soil as a National Wealth. October 15–16, 2008, Sielnica, 225–229. (in Czech with English abstract)

van Genuchten M.T. (1980): A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Science Society of America Journal, **44**: 892–898.

Wösten J.H.M., Schuren C.H.J.E., Bouma J., Stein A. (1990): Functional sensitivity analysis of four methods to generate soil hydraulic functions. Soil Science Society of America Journal, **54**: 832–836.

Wösten J.H.M., Lilly A., Nemes A., Le Bas C. (1999): Development and use of a database of hydraulic properties of European soils. Geoderma, **90**: 169–185.

Wösten J.H.M., Pachepsky Y.A., Rawls W.J. (2001): Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. Journal of Hydrology, **251**: 123–150.

Received for publication July 12, 2013
Accepted after corrections August 19, 2013

Corresponding author:

Ing. Markéta Miháliková, Ph.D., Česká zemědělská univerzita v Praze, Fakulta agrobiologie, potravinových a přírodních zdrojů, katedra vodních zdrojů, Kamýcká 129, 165 21 Praha 6-Suchdol, Česká republika; e-mail: mihalikova@af.czu.cz