Available online at www.sciencedirect.com

ScienceDirect

Environmental Modelling & Software 23 (2008) 254–255

Environmental Modelling & Software

www.elsevier.com/locate/envsoft

Software data news

Software to estimate −33 and −1500 kPa soil water retention using the non-parametric *k*-Nearest Neighbor technique

A. Nemes [a,b,*], R.T. Roberts [c], W.J. Rawls [c], Ya.A. Pachepsky [d], M.Th. van Genuchten [e]

[a] *University of Maryland, Department of Plant Science and Landscape Architecture, College Park, MD 20742, USA*
[b] *USDA — ARS Crop Systems and Global Change Laboratory, 10300 Baltimore Avenue, Bldg. 001, BARC-West, Beltsville, MD 20705, USA*
[c] *USDA — ARS Hydrology and Remote Sensing Laboratory, Beltsville, MD 20705, USA*
[d] *USDA — ARS Environmental Microbial Safety Laboratory, Beltsville, MD 20705, USA*
[e] *USDA — ARS George E. Brown, Jr. Salinity Laboratory, Riverside, CA 92507, USA*

Received 22 May 2007; accepted 31 May 2007
Available online 3 August 2007

Abstract

A computer tool has been developed that uses a *k*-Nearest Neighbor (*k*-NN) lazy learning algorithm to estimate soil water retention at −33 and −1500 kPa matric potentials and its uncertainty. The user can customize the provided source data collection to accommodate specific local needs. Ad hoc calculations make this technique a competitive alternative to publish pedotransfer equations, as re-development of such equations is not needed when new data become available.

Keywords: *k*-Nearest Neighbor; Soil water retention; Field capacity; Wilting point; Pedotransfer function; Uncertainty; Bootstrap

Software availability

Name of software: *k*-Nearest
Developers: USDA − ARS HRSL
Contact address: USDA − ARS Hydrology and Remote Sensing Laboratory, 10300 Baltimore Avenue, Bldg. 007, BARC-West, Beltsville, MD 20705, USA. Email: ralph.roberts@ars.usda.gov
Availability and Online Documentation: Free download with manual and supporting material at: http://www.ars.usda.gov/ba/anri/hrsl/computer models
Year first available: 2006
Hardware required: IBM compatible PC
Software required: MS Windows (tested on Windows 2000 and XP)
Programming language: MS Visual Basic
Program size: 6.3 MB

1. Overview and validation

Various ecological, hydrological, meteorological, engineering and other models use soil hydraulic parameters as input to describe the ability of soils to retain and transmit water and solutes. Soil hydraulic properties are costly to measure, and an array of estimation techniques, mostly termed pedotransfer functions (PTFs), exists to provide an affordable alternative to the measurements. One common feature of today's PTFs is that they are all based on some regression-type approach, which bars the users from including any additional data to improve estimates for their site-specific range of soil properties without redeveloping the equations.

We developed and validated an alternative approach that is based on the non-parametric '*k*-Nearest Neighbor' (*k*-NN) technique (e.g. Dasarathy, 1991). The technique is based on pattern-recognition rather than on fitting equations to data. Application of the *k*-NN means identifying and retrieving the most similar instances, based on their input attributes, to the target object from a known set of stored instances. This method was applied by Nemes et al. (2006a) to estimate two

[*] Corresponding author. USDA − ARS Crop Systems and Global Change Laboratory, 10300 Baltimore Avenue, Bldg. 001, BARC-West, Beltsville, MD 20705, USA. Tel.: +1 301 504 5177; fax: +1 301 504 5823.
E-mail address: attila.nemes@ars.usda.gov (A. Nemes).

1364-8152/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2007.05.018

key soil water contents – water content at −33 kPa that is often associated with field capacity, and water content at −1500 kPa that is often referred to as the wilting point for most plants. Nemes et al. (2006a) found the *k*-NN technique to be comparable with neural network models that are cited as probably the most advanced and accurate PTF development technique of the day. Subsequently, Nemes et al. (2006b) developed ensembles of predictions by performing multiple randomized subset-selections. They tested the sensitivity of this version of *k*-NN to seven different algorithm and data scenarios, using hierarchical sets of input attributes and data set sizes. It was concluded that the *k*-NN technique shows a large degree of stability and insensitivity to different settings and options. The presented software was built using the findings of Nemes et al. (2006a,b).

2. Features and capabilities

Similarly to classic PTFs, the software uses two data sets, i.e. the *reference data set* that is analogous to the development data set of classic PTFs, and an *application data set*, which can both be entered in different file formats.

Data are first normalized to avoid bias towards any of the inputs attributes and Euclidean distance – in terms of the input attributes – is used to identify the nearest known instances to the queried instance. Two 'design parameters' have been optimized by Nemes et al. (2006a): '*k*' is the number of nearest neighbors (hence the name *k*-Nearest Neighbor) that are actually utilized while formulating the output; and '*p*' a term used in calculating weights in a distance-dependent weighting of neighbors. Once the nearest neighbors are identified and their weights are assigned, the response attribute is formulated and outputted as the weighted average of the response attributes of the selected nearest neighbors.

The software uses sand, silt and clay contents according to the FAO/USDA particle-size classification system (USDA, 1951) and, as options to the user, soil bulk density (BD) and/or soil organic matter (OM) content as input.

We provide a quality assured data subset of near 40,000 samples of the February 2003 edition of the all-US NRCS-SCS Soil Characterization Database (Soil Survey Staff, 1997) as default reference data set. The user can opt to use the provided database as reference data or define an alternative data set, which can be partly or entirely independent of the above database. Such alternative data sets are subject to data format and consistency requirements, detailed in the Users Guide.

Running the algorithm can optionally be combined with the use of the bootstrap data subset selection technique (e.g. Efron and Tibshirani, 1993) that allows generating multiple replica data sets and can yield an uncertainty estimate for each queried sample, formulated as the standard deviation of the mean of bootstrap estimates.

Assuming the same relevance of all input attributes for various outputs, the technique can estimate several output attributes simultaneously. Estimated key soil water content values for both matric potentials (−33 kPa and −1500 kPa) – and optionally their uncertainty – are calculated and outputted simultaneously. An optional user/model/data header is provided along with the data in the alphanumeric output.

Having only two windows and a small number of options to set, the program has an easy to use interface with general help, as well as window specific help available.

3. Summary

A simple, yet robust PTF technique has been packaged in this software that has the capability to provide the wide range of users with reliable estimates of missing soil hydraulic information. Because the *k*-NN technique performs all estimation calculations ad hoc, the user is able to customize the reference data set without the need to redevelop any equations.

Acknowledgements

During completion of this study, Attila Nemes was affiliated with the University of California, Riverside, CA and the USDA-ARS Hydrology and Remote Sensing Laboratory, in Beltsville, MD.

References

Dasarathy, B.V. (Ed.), 1991. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press, Los Alamitos, CA, pp. 1–30.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability. Chapman and Hall, New York, USA.

Nemes, A., Rawls, W.J., Pachepsky, Ya.A., 2006a. Use of the non-parametric nearest neighbor approach to estimate soil hydraulic properties. Soil Science Society of America Journal 70, 327–336.

Nemes, A., Rawls, W.J., Pachepsky, Ya.A., van Genuchten, M.Th., 2006b. Sensitivity of the nearest neighbor approach to estimate soil hydraulic properties. Vadose Zone Journal 5, 1222–1235.

Soil Survey Staff, 1997. National Characterization Data. Soil Survey Laboratory, National Soil Survey Center, and Natural Resources Conservation Service, Lincoln, NE.

United States Department of Agriculture (USDA), 1951. Soil survey manual. In: Agricultural Handbook No. 18. U.S. Department of Agriculture, Washington DC, USA.