Available online at www.sciencedirect.com
ScienceDirect
Geoderma 144 (2008) 212–224
GEODERMA
www.elsevier.com/locate/geoderma

# Optimisation of pedotransfer functions using an artificial neural network ensemble method

L. Baker [a,*], D. Ellison [b]

[a] Dundee, United Kingdom
[b] University of Abertay Dundee, Bell Street, Dundee, DD1 1HG, Scotland

Received 30 July 2006; received in revised form 7 October 2007; accepted 7 November 2007
Available online 26 December 2007

**Abstract**

Soil hydraulic properties, mainly saturated and unsaturated hydraulic conductivity and water retention, are crucial input parameters in any modelling study on water flow and solute transport in soils. However, direct measurement techniques remain relatively time consuming, labour intensive and expensive. Fortunately, they may be predicted by forming a mathematical relationship between relatively easily collected soil survey parameters, such as soil texture, bulk density and organic matter content, and less readily available soil properties, such as water retention or hydraulic conductivity. These mathematical relationships, pedotransfer functions (PTFs), allow the transfer of data we have into data we need. In recent years many PTFs have been created with the aid of artificial neural networks (ANNs).

We describe a PTF modelling method that combines a number of individual ANNs – the ensemble method – and compare directly the results obtained with those achieved by a competing single ANN method. The ensemble method is shown to produce significantly more accurate and robust PTFs when compared to single ANN methods, under the same conditions. These ANN–PTF ensembles have been optimised to produce maximum benefits from the ensemble method, whilst minimising data correlations between training and test data. Consideration has been given to how much data is required in the training and testing phases of modelling, and how many individual ANNs should be combined to produce the ensemble.

We also demonstrate that the current terminology used to describe various portions of the dataset in the single ANN method is insufficient when describing such portions in the ensemble method. As a consequence, new terminology is introduced. Furthermore, we establish that data may be recycled, i.e. used in both the training and testing phases of the ANN–PTF ensemble with virtually no loss of precision.

This report shows that, for the water retention data investigated here, the ensemble method requires significantly less data than does the single ANN method – more than 2 1/2 times less – to produce results of equivalent precision. This is a crucial result because, since ANN–PTFs formed from local data produce more accurate predictions than those built from data spread from a wider area, the concept of data conservation becomes a critical factor in ANN–PTF construction.


*Keywords:* Artificial neural network; Ensemble; Pedotransfer function; Water retention

## 1. Introduction

Soil hydraulic properties, mainly saturated and unsaturated hydraulic conductivity and water retention, are crucial input parameters in any soil modelling study of water flow and solute transport. Thus, it is crucial that the flux and storage of water at the land surface are accurately modelled in order to understand and deal with the transport of water, solutes and environmental contaminants. Despite the progress being made in direct measurement of soil hydraulic properties, the majority of these techniques

*Abbreviations:* %Cl, percentage of clay; %OM, percentage of organic matter; %Sa, percentage of sand; %Si, percentage of silt; ANN, artificial neural network; BD, dry bulk density; C, coarse soil texture class; dp, decimal place(s); FAO, Food and Agriculture Organisation; F, fine soil texture class; FC, field capacity; HYPRES, database of Hydraulic Properties of European Soils; M, medium soil texture class; ME, mean error; MF, medium-fine soil texture class; PTF, pedotransfer function; RI, relative improvement; RMSE, root mean squared error; VF, very-fine soil texture class.

* Corresponding author.
*E-mail address:* eelrekab@gmail.com (L. Baker).

0016-7061/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.geoderma.2007.11.016

remain relatively time consuming, labour intensive and expensive. Since good predictions may be accurate enough for many applications (Wösten et al., 2001), large databases – such as HYPRES, the database of Hydraulic Properties of European Soils (Wösten et al., 1999) – are being constructed for the development of predictive models, such as pedotransfer functions (PTFs), of agricultural and environmental soil functioning.

A PTF (Bouma and van Lanen, 1987) is a mathematical relationship between two or more relatively easily collected soil survey parameters, such as soil texture, bulk density, organic matter content, etc., and the less readily available soil hydraulic properties. This relationship is used in the estimation of non-measured soil parameters from one or more measured ones, i.e. in the words of Bouma (1989), to transfer the data we have (soil survey parameters) into the data we need (soil hydraulic properties). The different types of PTFs have become important tools in quantifying the most important physical and biological processes in soils, providing a measure of correspondence between measured and simulated functional soil behaviour.

In recent years PTFs constructed by artificial neural networks (ANNs) have proven popular with many researchers, and have yielded results that are at least as good as other techniques and overcome some of the statistical assumptions hard-wired into PTFs. ANN–PTFs have been developed by researchers such as Pachepsky et al. (1996), Tamari et al. (1996), Schaap and Bouten (1996), Koekkoek and Booltink (1999), Minasny et al. (1999) and Minasny and McBratney (2002). The overall conclusion made by these (and other) investigators was that when the number of input parameters is greater than three, ANNs usually perform better than regression techniques, particularly when uncertainties in the quality of the data were small. All of these investigators used single ANNs to model PTFs, however, a few researchers are beginning to use ANN ensembles – combinations of ANNs – to develop PTF models. Publications include those by Schaap and Leij (1998a, b), Schaap et al. (1998), Schaap et al. (2001), Nemes et al. (2003), Dimopoulos et al. (2004), Jeong and Kim (2004), Minasny et al. (2004), Quanet al. (2004), Baker (2005), Parasuraman et al. (2006) and Spencer et al. (2006).

The recent popularity of ANN–PTFs is, however, a little deceiving. ANNs are very data hungry, and such methods have only become possible since the development of databases of soil hydraulic properties. However, for many modellers, it has become relatively easy to simply pass all or most of the available data to the ANN(s) with little regard to the precision, resolution and quality of the process.

The aims of this paper are to demonstrate that under identical conditions the ensemble method produces significantly more accurate PTF models than single ANN methods. Additionally, it will demonstrate mathematically, statistically and empirically how to optimise the model in terms of the number of ANN ensemble members, the number of soil horizons per ANN and the number of tests to perform for stringent testing. Furthermore, this paper will show that current terminology used when discussing ANN–PTFs is insufficient, and will introduce new terms with which to refer to data and results based on tests performed on ensembles with these data.

## 2. Materials and methods

### 2.1. The HYPRES database

The HYPRES database was constructed to overcome a lack of data regarding soil hydraulic properties, and brought together into one central database the existing hydraulic data that resided at 20 institutions within 12 European countries. HYPRES Version 2.0 comprises around 25 Mb of data held in six data tables and represents 95 different soil types according to the modified Food and Agriculture Organisation (FAO) soil legend (CEC, 1985) used in the 1:1,000,000 Soil Geographical Database of Europe. There are 1791 soil profiles with a total of 5560 horizons. The soil hydraulic data were derived by various methods; however, the measurement of pressure head values ($h$) was standardised to produce soil water retention ($\theta$) values at 14 pressure heads for each soil horizon: 0, −10, −20, −50, −100, −200, −250, −500, −1000, −2000, −5000, −10,000, −15,000 and −16,000 cm $H_2O$. Wösten et al. (1998) divided the soil horizons into the six FAO texture classes; five mineral – coarse (C), medium (M), medium-fine (MF), fine (F) and very-fine (VF) – and one organic.

### 2.2. Data selection criteria

Soil water retention data were extracted from HYPRES for each of the 5 FAO mineral texture classes, and exported into Microsoft Excel 97 SR-2 spreadsheets (Microsoft Corporation).

Percentages of sand (%Sa), silt (%Si) and clay (%Cl), dry bulk density (BD) and percentage of organic matter (%OM) were used as the input parameters to PTF models of water retention. The HYPRES standardised value of −250 cm $H_2O$ was used as the output parameter. This is equivalent to −24.5 kPa and, since the many possible definitions of field capacity (FC) report that the FC resides at a point between −5 kPa and −33 kPa on the water retention curve (McKeague et al., 1984), this is therefore considered as the water retention FC, and denoted $\theta_{FC}$.

Also, duplicate horizons were discarded to reduce target noise in the model, since two physically identical soil horizons (same texture, BD and %OM), may differ in hydraulic characteristics (different $\theta_{FC}$). This difference could be due to many factors, from pre-processing fitting errors or differing measurement techniques, to errors in measurement, calculation or insertion into the database. The discrepancy in hydraulic characteristic represents a source of irreducible error, hence when duplicates exist, one (chosen randomly) is retained, whilst the remaining are discarded.

In addition, there were a number of soil horizons for which the proportions of sand, silt and clay did not sum to 100. These were regarded as ‘erroneous’ and were not included in the dataset for modelling.

When these criteria have been applied to, and the data extracted from, HYPRES, the data detailed in Table 1 resulted. In all, 2764 soil horizons were selected for modelling. Approximately one-third and one-quarter of the horizons belong to the M and C texture classes, respectively, whilst