

Cautionary notes on the use of the Rawls et al. (1982) soil hydraulic pedotransfer functions

Attila Nemes^{1,2}, Dennis Timlin², Bruno Quebedeaux¹

¹ University of Maryland, College Park, MD, USA

²USDA-ARS Crop Systems and Global Change Lab., Beltsville, MD, USA

Introduction

- Environmental simulation models use some form of soil hydraulic data
- Large scale / scenario based projects don't have measured data available
- Estimated values/parameters are needed (-> pedotransfer functions (PTFs))
- Estimations are error prone, and such errors will propagate in subsequent simulation processes

How it is currently solved in CEAP:

- measured data are not available/feasible
- estimation of -33 and -1500 kPa water retention (as approximates of FC and WP) by the Rawls et al. (1982) PTF equations are coded in the APEX model (*user's alternative in APEX: enter FC, WP values*)

??? Is this PTF suitable for this task ???

The Rawls et al. (1982) PTF:

Rawls et al. 1982. (Trans. ASAE 25(5): 1316-1320 & 1328)

- linear regression based
- point estimation on the water retention curve (12 points)
- uses inputs SSC, OM, BD(optional), θ_{33} (optional), θ_{1500} (optional)
- data collected from literature (26 listed sources), representing 32 states
- variables of interest / form of applied equations:

$$\theta_{33} \text{ (FC)} = 0.2576 - 0.002 * \text{SAND} + 0.0036 * \text{CLAY} + 0.0299 * \text{OM}$$

$$\theta_{1500} \text{ (WP)} = 0.026 + 0.005 * \text{CLAY} + 0.0158 * \text{OM}$$

- these two points can drive many soil/crop models by being trigger/cutoff points to processes and by indicating boundaries for optimal growing conditions.

How is a PTF usually tested?

- Commonly used testing of PTFs is insufficient

- most comparative studies report a maximum of 3 measures:

$$\text{RMSE}=0.055 \text{ cm}^3/\text{cm}^3 \quad R^2=0.72 \quad \text{ME}= -0.001 \text{ cm}^3/\text{cm}^3$$

- source and independence of test data (if any!)
 - applicability to an area/region is often taken for granted, never challenged
- Alternative, advanced possibilities
 - comparison of different techniques on the same data
 - other performance measures / data mining (e.g. distribution of errors)

Quick test of the Rawls et al PTF - methods:

- Apply Rawls' PTF to a subset of NRCS National Soil Survey Characterization (NSSC) data.
- Some basics of the NSSC database:
 - $N > 120\,000$, methodology mostly uniform
 - independent from the Rawls et al. (1982) data set
 - covers entire US (+ some foreign data)
- Selection criteria:
 - sand, silt, clay, OM, BD, θ_{33} , θ_{1500} available
 - limits in properties as described in Rawls et al 1982.
 - checked for apparent data errors, mismatches, inconsistencies
 - limited to the 48 contiguous States
 - $N=9395$ (split randomly to two subsets: "A": $N=4697$ and "B": $N=4698$)
 - test run against subset "A".

Quick test of the Rawls et al PTF – results (I):

FC: RMSE=0.073 cm³/cm³ ME= -0.006 cm³/cm³

WP: RMSE=0.043 cm³/cm³ ME= -0.005 cm³/cm³

AWHC (FC-WP): RMSE=0.055 cm³/cm³ ME= -0.001 cm³/cm³

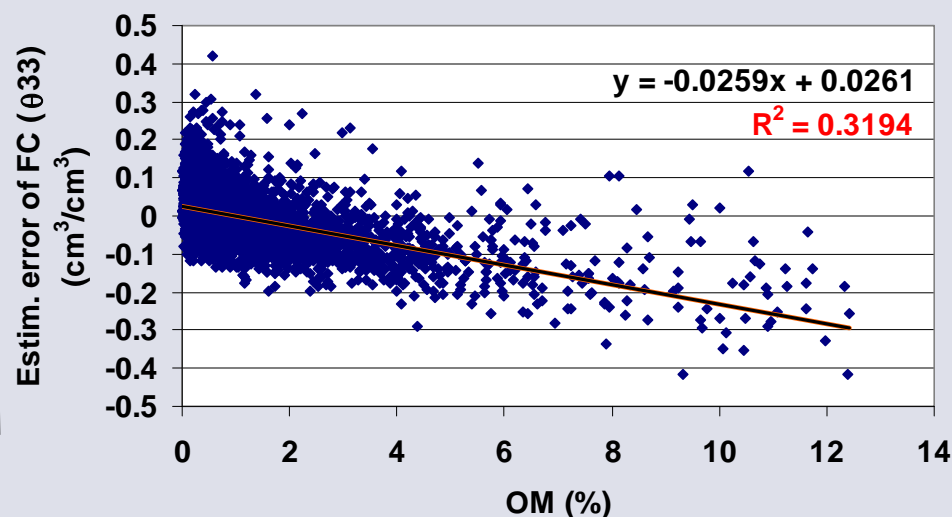
- acceptable accuracy (within range of many other PTFs)
- very good in terms of bias
- by default it should be applicable for the US as it originates from a large collection of US soils
- this is despite of known methodological differences between the development (Rawls) and test (NSSC) data sets.

Quick test of the Rawls et al PTF – results (II):

Correlation coefficient (R^2) between estimation errors and input variables

- optimally $R^2 = 0$

	FC	WP	AWHC
Sand	0.067	0.004	0.085
Silt	0.009	0.003	0.029
Clay	0.079	0.028	0.056
OM	0.319	0.319	0.089



Why would two databases that were thought to be representative to the same area show such great bias compared to each other?

Recovery of the Rawls et al. data (I):

Re-generate Rawls data set and derived PTF

- Access to overall data source from W. Rawls (N=4515)
- Selection criteria applied
- Recovered most of the data set (N=2528 out of original N=2541)

- Statistical properties:

	Rawls et al. (1982) (N=2541)					Rawls et al. (recovered, N=2528)				
	min	max	mean	std	med	min	max	mean	std	med
Sand [%]	0.10	99.00	56.00	N/D	N/D	0.20	99.00	55.21	31.19	58.65
Silt [%]	0.10	93.00	26.00	N/D	N/D	0.00	91.00	26.76	22.92	21.20
Clay [%]	0.10	94.00	18.00	N/D	N/D	0.00	93.40	18.03	15.36	14.10
BD [g/cm ³]	0.10	2.09	1.42	N/D	N/D	0.01	2.02	1.43	0.23	1.47
OM [%]	0.10	12.50	0.66	N/D	N/D	0.00	14.99	1.10	1.91	0.34
θ ₃₃ [v/v]	N/D	N/D	N/D	N/D	N/D	0.01	1.38	0.24	0.14	0.24
θ ₁₅₀₀ [v/v]	N/D	N/D	N/D	N/D	N/D	0.00	0.56	0.12	0.09	0.10

1.104 / 1.724 = 0.64 !!

PTF equations redeveloped: **coefficients close only if OC is used for OM!!!**

Recovery of the Rawls et al. data (II):

original data sources: reports from 60's, 70's (23 of 26 sources found)

Applied sampling methodology, and data format used in the Rawls data set:

State/ region	Org. amendments as... org.matter/org.carbon	Water content: gravim./volum.	Sample used: disturbed/non-dist.
CA	oc	volumetric	disturbed
FL	oc	volumetric	non-disturbed
GA	om	volumetric	disturbed
LA-II	om	volumetric	disturbed
LA-II	no data (reported =0)	volumetric	disturbed
MA-I	oc	gravimetric	WP dist. FC not.
MA-II	oc	gravimetric	?
MA-III	oc	gravimetric	?
MO	om	volumetric	disturbed if gravelly
MN	no data (reported =0)	gravimetric	?
ND	no data (reported =0)	volumetric	disturbed
N-Central Reg.	no data (reported =0)	volumetric	?
OH	oc	volumetric	non-disturbed
SC	no data (reported =0)	volumetric	disturbed
SD	no data (reported =0)	volumetric	non-disturbed
GA	?	?	?
NJ	?	?	?

Distribution of samples:

OM/OC content represented as:

1109 OC
541 OM
633 no data (was =0)
 237 *unknown*

Water retention represented as:

1836 volumetric
447 gravimetric
 237 *unknown*

Sampling:

578 disturbed
 341 mixed
 968 non-disturbed
 633 *unknown*

Adjustments to the and redevelopment of PTFs:

- | | |
|----------------------------------|---------------|
| - disturbed/non-disturbed: | no adjustment |
| - OM/OC content: | adjusted |
| - gravimetric/volumetric: | adjusted |
| - no OM/OC data: | omitted |
| - no information on methodology: | omitted |

Step 1: ...original **linear regression** equations of Rawls et al. (1982)

Step 2: ...**linear regression**, recovered Rawls et al. data, OM/OC data corrected

Step 3: ...**linear regression**, all data corrected, missing data cases eliminated

Step 4: ...as in Step 3, but using a **k-Nearest Neighbor** pattern recognition technique (Nemes et al. 2006, 2008)

*Step 5: ... using the **k-Nearest Neighbor** pattern recognition technique and data of the NSSC 'B' subset. Step 5 reflects the typical level of noise in this type of data. (statistically identical to the test data but it is independent)*

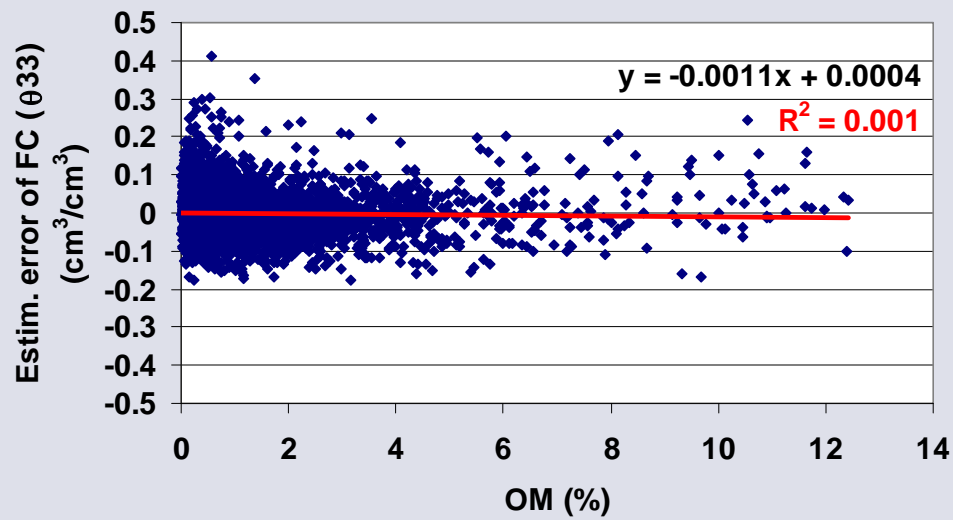
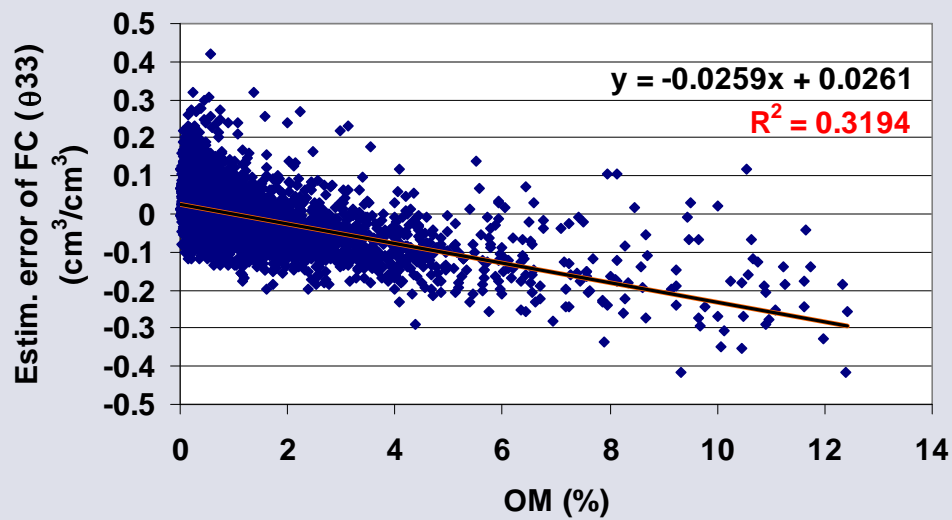
Data adjustments and redevelopment of PTFs:

Correlation coefficient (R^2) between estimation errors and input variables:

	Rawls et al. (1982) (as published) linear regression Step 1			Rawls N=2528 subset (OM/OC corrected) linear regression Step 2			Rawls N=1615 subset (all data corrected) linear regression Step 3			Rawls N=1615 subset (all data corrected) k-Nearest Neighbor Step 4			NSSC subset 'B' (N=4698) k-Nearest Neighbor Step 5		
	FC	WP	AWHC	FC	WP	AWHC	FC	WP	AWHC	FC	WP	AWHC	FC	WP	AWHC
Sand	0.067	0.004	0.085	0.068	0.002	0.079	0.116	0.020	0.097	0.027	0.021	0.006	<0.001	<0.001	<0.001
Silt	0.009	0.003	0.029	0.003	0.012	0.022	0.005	0.022	0.037	0.001	0.005	0.008	<0.001	<0.001	0.002
Clay	0.079	0.028	0.056	0.108	0.038	0.062	0.188	0.155	0.058	0.086	0.018	0.051	<0.001	0.002	0.002
OM	0.319	0.319	0.089	0.111	0.132	0.018	0.083	0.038	0.043	0.001	0.003	0.005	<0.001	<0.001	<0.001

RMSE and ME of estimations:

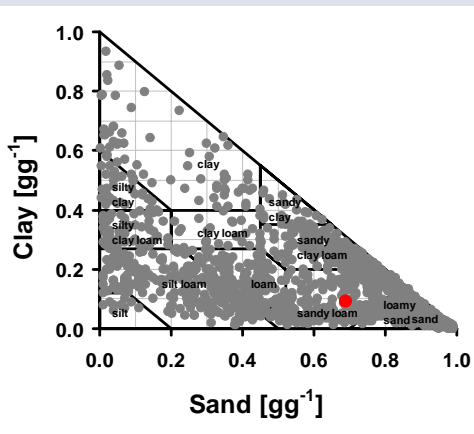
	Step 1			Step 2			Step 3			Step 4			Step 5		
	FC	WP	AWHC	FC	WP	AWHC	FC	WP	AWHC	FC	WP	AWHC	FC	WP	AWHC
RMSE	0.072	0.043	0.055	0.065	0.039	0.057	0.073	0.039	0.058	0.058	0.036	0.053	0.053	0.032	0.048
ME	-0.006	-0.005	-0.001	-0.008	0.010	-0.017	-0.011	<0.001	-0.011	-0.001	<0.001	-0.001	0.002	0.001	0.001



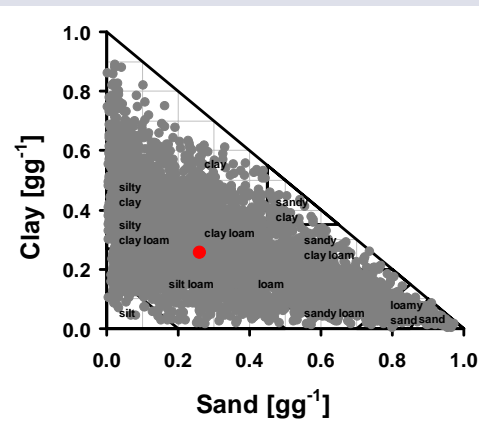
Why k-NN is better

- linear regression: [global solution](#) for entire data space
- k-Nearest Neighbor: [local solution](#) based only on the neighborhood of each sample

Rawls



NSSC



Degree of bias:

	Rawls		NSSC	
[%]	mean	med.	mean	med
sand	61.39	69.0	31.05	26.0
silt	24.03	12.3	41.21	40.7
clay	14.58	9.1	27.74	25.7

Additional attempts to improve performance:

- splitting data further by methodology did not yield improvement
- omitting extremes/outliers did not yield improvement
- “error in field capacity” vs. “clay content”: still somewhat correlated

suspected reason: different clay mineralogy, climatic differences

Rawls data - US E-SE Coast over-represented (~65%)

NSSC data - ~ area size / agron. importance represented

Summary of findings:

1. The PTFs of Rawls et al. (1982) deliver input-specific bias in the estimates, primarily because of misrepresentation of data, but also because of bias in the data collection towards coarse textured soils.
2. Correcting the data helped to overcome a significant degree of such bias.
3. Using a PTF solution that provides a local solution in the data space for each single query (e.g. the presented k-NN solution) helps to reduce bias due to the over/under-representation of soils in the base data.
4. Input variables selected for a PTF may not describe all correlations in the data, additional input variable(s) may be needed for that (e.g. clay vs. CEC)

Conclusions and outlook...

- APEX/EPIC users should be critical about the built-in PTF when used
- Find or develop a feasible alternative!
- In progress:
 - influence of clay mineralogy
 - functional evaluation using the APEX model

Questions?

- attila.nemes@ars.usda.gov
- dennis.timlin@ars.usda.gov