Utrecht University

Master Thesis Results
Faculty of Geosciences
Supervisor: Dr. Alraune Zech

TNO innovation for life

Using Artificial Intelligence to predict Saturated Hydraulic Conductivity from Particle Size Distribution

A comparative analysis of 6 Machine Learning Algorithms

Valerie de Rijk, BSc

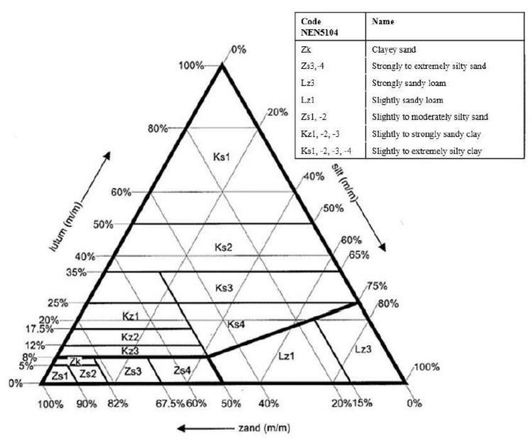
Abstract

- **Particle Size Distribution (PSD)** is a sorted list of minimal diameter of a material.
- Saturated hydraulic conductivity (**K_s**) is currently predicted through empirical formulas using the PSD or direct measurement.
- We evaluated the performance of 6 Machine Learning techniques (*Artificial Neural Network* (**ANN**), *Decision Tree* (**DT**), *Linear Regression* (**LR**), *Random Forest* (RF), *Support Vector Regression* (**SVR**) and *XGBoost* (**XG**)) to predict K_s from PSD on a dataset of 3400 samples from the Dutch shallow subsurface.

Table 1: Descriptive Statistics for the full dataset.

	Log K_s [m/d]	d10 [mm]	d50 [mm]	Lutum [%]	Silt [%]	Sand [%]
Mean	-0.62	0.10	0.20	6.31	13.80	77.42
Min	-6.70	4.21 e-4	2.74 e-3	0	0	0
Median	0.37	0.10	0.19	1.00	2.92	96.04
Max	2.08	0.50	0,98	83.27	77.37	100
Variance	4.39	0.01	0.02	131.79	427.67	1111.67

Figure 1: Visual representation of Soil Classes. [1]



- Soil lithologies are unequally distributed. sand (**Zs,** 2806 samples), clay (**Ks**,560 samples), silt(**Kz/Lz**, 181 samples)
- Models are trained for seperate soil classes as well as for the Full dataset (3547 samples) with and without outliers
- The Deep Drilling dataset consists of 164 samples, predominatly from the L_z/K_z category.

Table 2: R^2 and MSE for the Test Dataset of Different Algorithms

	RF Full PSD	ANN Full PSD	XG Full PSD	RF PSD-derived	ANN PSD-derived	XG PSD-derived
R^2	0.92	0.92	0.92	0.9	0.89	0.9
MSE	0.57	0.56	0.57	0.63	0.64	0.63

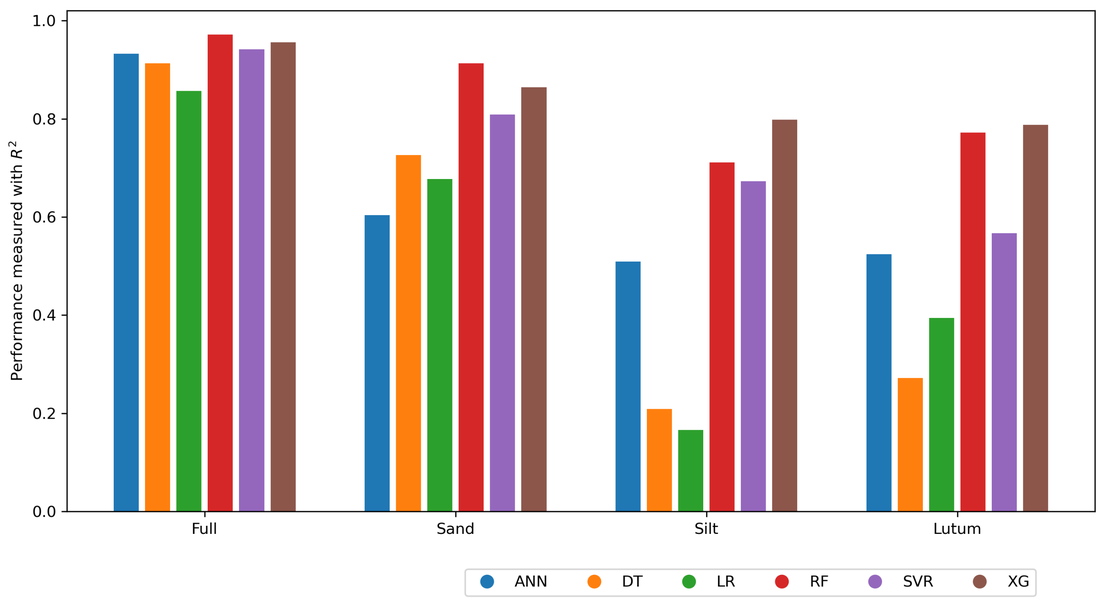


Figure 3: Model Performance, evaluated by R^2, for all algorithms and subsets of data.

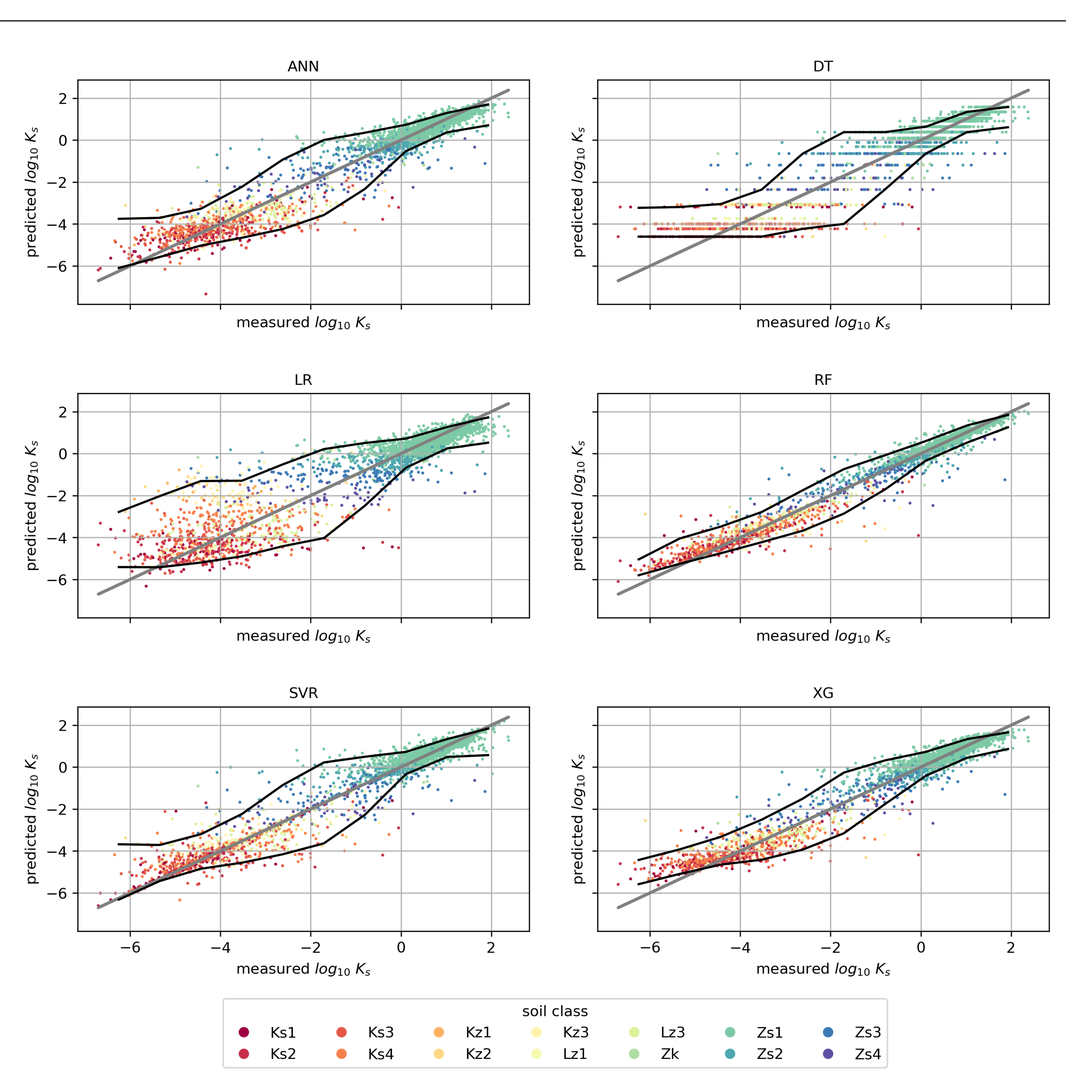


Figure 2: Predicted K_s values by Algorithms (y-axis) against the measured K_s data (x-axis). The dotted line represents a 1:1 line. The black lines represent the 5th and 95th quantiles

Results

- RF and XG are the best performing algorithms. All algorithms outperform the five best performing empirical formulas.
- Predictive ability is similar to other studies that used different soil parameters. [3,4]
- All algorithms perform best when trained on the Full Dataset, rather than on subsets of soil classes
- Using PSD-derived variables (d_{10}, d_{50}, d_{60}) yields good results (Table 2)
- Porosity prediction from the PSD is not accurate. .

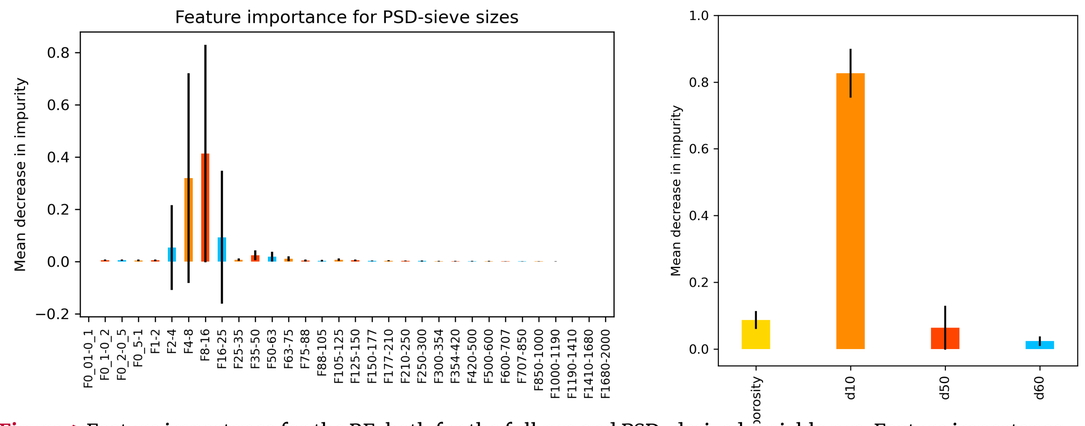


Figure 4:Feature importance for the RF, both for the full run and PSD-derived variable run. Feature importance indicates which feature variable has most influence in determining the target variable.

Conclusion

- Promising and cheap technique for geo-technical, geo-ecological as well as geothermal exploration, since performance of RF and XG on the Deep Drilling dataset is in line with results from petrophysical analysis.
- A large and diverse training dataset leads to good probability of succesfull prediction in other areas.
- Importance of d_{10} calls for further exploration in this parameter.

References

1) Stichting Koninklijk Nederlands Normalisatie Instituut, 2019, Geotechnical investigation, testing and classification of soil - part 2: Principles for classification.

Using Artificial Intelligence for Saturated Hydraulic Conductivity prediction