Norwegian University of Life Sciences

Master's Thesis 2024 60 ECTS

MINA (Faculty of Environmental Sciences and Natural Resource Management)

Can 3-D X-Ray tomography improve the estimation of saturated hydraulic conductivity of soils?

Einar Emil Låker

Environment and natural resources

Can 3-D X-Ray tomography improve the estimation of saturated hydraulic conductivity of soils?

Author: Einar Emil Låker

Table of Contents

1. Abstract

Saturated hydraulic conductivity (K_{sat}) is one of the most fundamental parameters in soil hydrology. It governs the rate of saturated flow through porous media and functions as a scaling factor for unsaturated flow. Knowledge of K_{sat} is key to understanding the movement of water in soils, transport, and recharge of groundwater, suspended and dissolved transport in soils, and soil-air water exchange. In hydrology and climate modelling K_{sat} is often estimated through pedotransfer functions. A large effort has been committed to the development of these models, using an array of differing algorithms and methods. However, estimating K_{sat} has been somewhat troublesome, since the commonly measured soil properties, such as soil texture, bulk density, and organic matter content, used as predictor variables in PTFs do not explain K_{sat} variation well. Instead, K_{sat} is largely controlled by pore-network characteristics especially in highly structured soils.

Using an extended, methodologically homogeneous dataset of commonly measured soil physical properties, 3-D X-ray computed tomography imaged pore-network parameters, and quasi-continuous particle-size measurements using the Integral Suspension Pressure method, we assess the benefits of using combined soil textural and structural information on the estimation of K_{sat}. Using this dataset, we have built models that estimate K_{sat} using a boosted random forest algorithm (XGboost) and used explanatory model analysis to tune and evaluate the models. Three input data scenarios included (i) basic soil inputs only (ii) imaged pore metrics only, and (iii) their combination. Using or adding imaged pore metrics as inputs greatly improved the K_{sat} estimations that were reflected, for example, by the respective coefficients of determination, evaluated using a cross-validation scheme (R2 = -0.07/0.30/0.48 for the three input scenarios respectively). 3-D imaging of soil and the subsequent characterization of its pore-space may serve multiple research purposes, but such data are still not routinely collected due to cost of measurement and general lack of access to equipment. Our study confirms, however, that when collecting such metrics will become economically feasible through e.g. better automation of image processing using tools like SoilJ, having those metrics will show great potential towards improving the estimation of the soil's water transport properties.

2. Acknowledgements

Thank you to the ever-brilliant Attila Nemes for his guidance and enthusiasm; for his willingness to let me work on ‘his baby’; and for taking a enthusiastic data-nerd under his wing.

Thank you to David Hirmas who was an absolute delight to work with and who’s perspective brought this project to height which would otherwise have been impossible.

Thank you to everyone related to the SOILSPACE program, without whom this project would never have been possible.

Thank you to my parents, for their love, support, and good humour.

And thank you to my classmates who endured rants about statistics and modelling, and breathed light into my daily tasks

3. Preface

This thesis does not follow the traditional format. Due to the nature of the findings and the intention to later rework them into a scientific article it was determined in a joint decision with the main advisor Attila Nemes that the text would be split into two parts: a background and a draft article. First the underlying ideas needed to understand the work and the article are presented in the “Background” part of the text, then the draft article is presented with a new main title and separately numbered headings.

.

Background

4. Soil and it's place in the world.

The role of soils in matters of great importance; plant growth, climate, nutrient recycling, water supplies/purification, and as a habitat; cannot be overstated (Weil & Brady, 2017, pp. 20–21). The qualities of soils are the principal regulator of plant ecosystems, which props up the rest of the terrestrial ecosystems. 99% of the calories the world consumes comes from the soil, the last percentage being seafood (Fig. 1) (FAOSTAT, 2023). In the modern world what is not mined is *grown*, thus understanding the dynamics of soil is a fundamental necessity for an evidence-based approach to interacting with the natural world.

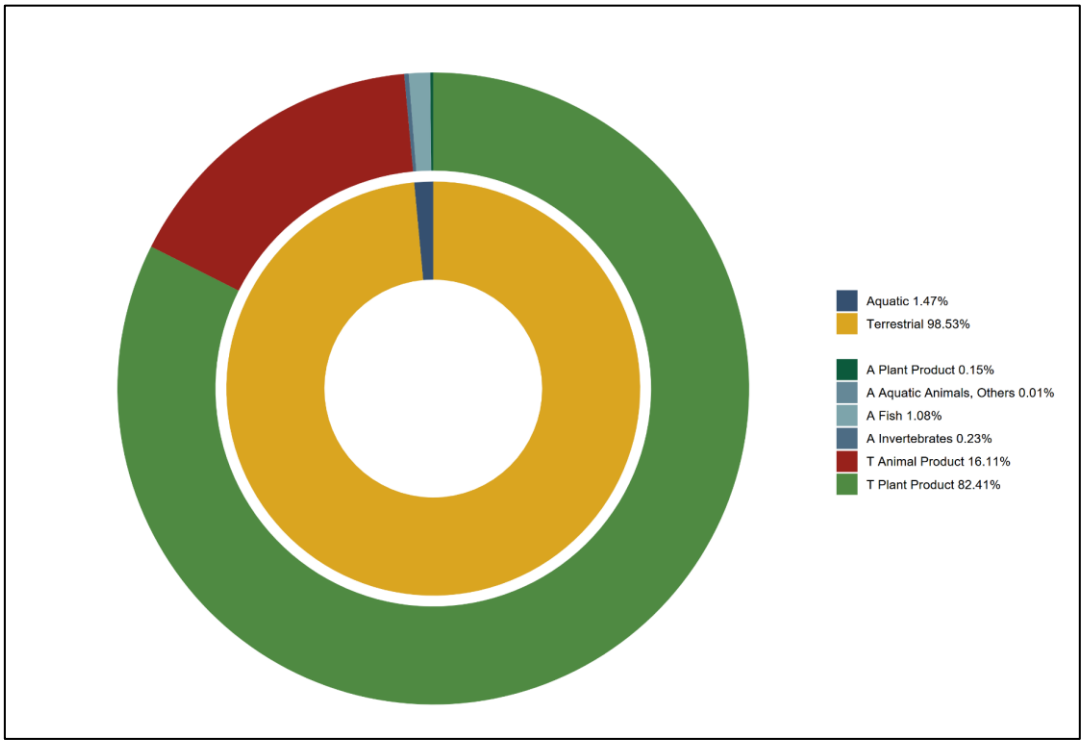


Fig. 1 Donut chart representing the ratios of calories consumed in the world for the year 2021. Inner donut represents aquatically or terrestrially produced, and the outer donut represents the breakdown of these categories. The data was retrieved from the FAOSTAT Food Balances database (FAOSTAT, 2023).

Soils are also in crisis, just like the rest of the earth system. Mismanagement, land cover change and climate change have caused damage to the earth's soils on a scale that is truly unfixable within our lifetimes. Each year we lose 3% of our capacity for food production to soil erosion, and by 2050 90% of the earths topsoil (FAO, 2022). The American Midwest has lost *56 700 000 000* metric tons of soils since the 1860s (Thaler et al., 2022), and some fields in China are losing 15 tons per hectare *each year* (Nearing et al., 2017). Not only are we

losing soils, the soils that are left are slowly degraded as their fertility wains (McLauchlan, 2006) and pollution festers (Khan et al., 2021). Finally climate change brings with it both desertification and increased rainfall, exacerbating these problems even further.

5. Soil physics and the issues at hand

Soil physics is a scientific effort trying to describe the physical properties of soils. The field of soil physics encompasses a large variety of topics and 122 years of history; starting in 1902 with Edgar Buckingham at the USDA Bureau of Soils (Nimmo & Landa, 2005). The topics range from:

- physical characteristics (texture, structure, porosity, density, aggregates size distribution & stability),
- behaviour of soils (compaction, consolidation, stress-strain relations),
- volumetric and its constituents (mineral particles, water, air, solutes, organic matter, biota),
- water retention characteristics, hydraulic conductivity,
- air permeability, and air volume.

Soil texture and the texture triangle (Fig. 2) is the part of soil physics more people have seen or heard about, although it's still quite niche. They describe the fractions of particle sizes grouped by sand, silt, and clay. Using this characterization we further group soils into soil texture classes (The names on the triangle in Fig. 2). Texture however only describes the most basic parameters of soil, the fractions of particle sizes.

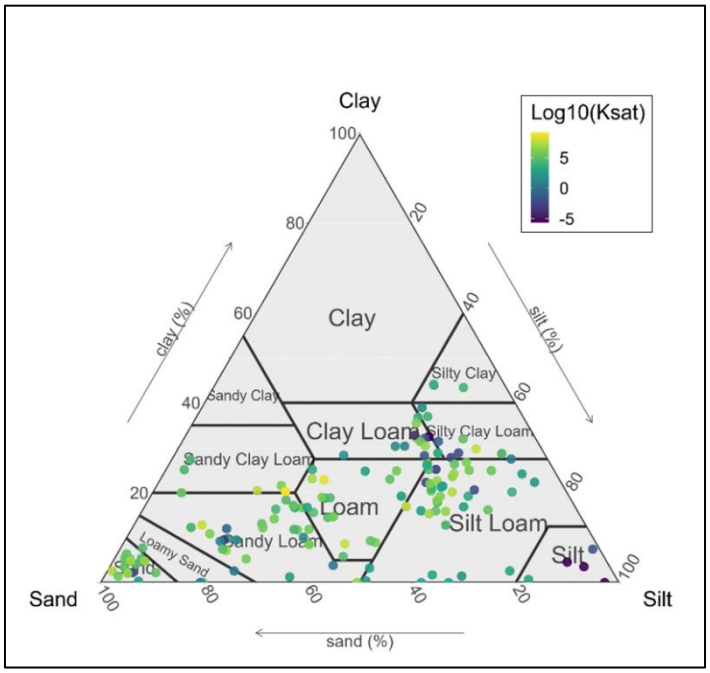


Fig. 2 Soil texture triangle with USDA classes and the SOILSPACE dataset plotted as points with the colour representing saturated conductivity

Soils are not simply the brown tinted matter under our shoes, but also the structures, pores, liquids, and biota within it. Soil often includes three phases of matter: solid, liquid and gas. The relationship between these is the often the controlling factor in physical characteristics of soils. The gas fraction found in a sample of soil often has a larger impact on the dry weight and bulk density of the sample than

the density of the solid material. Furthermore, we believe that the importance of soil pores and pore system structure has been underestimated as a integral part of their functions. Following this the effect of erosion, compaction, and the loss of biodiversity and organic matter on the pore system is also overlooked.

6. Hydrology, water flow in porous media

One of the most important relationships in the natural world is between soils and water the water within them. Soils are an integral part of the hydrological cycle, both as a transport medium, a hydrological pool, and a natural filter. Of the terrestrial water not locked away as ice or in groundwater 1/3rd is soil moisture (Wetzel, 2001). Soil water controls many biotic and abiotic processes; consequently, our ability to describe the mechanisms that control this relationship determines our ability to predict the outcomes of these processes. Using the language of physics and hydrology soil scientist try to determine the laws of soil hydrology. Possibly the most famous law of hydrology was published by Darcy in *Les fontaines publiques de la ville de Dijon* (1856, app. D). In what has later been termed Darcy's Law he describes the elements impacting the flow of a fluid through a porous medium.

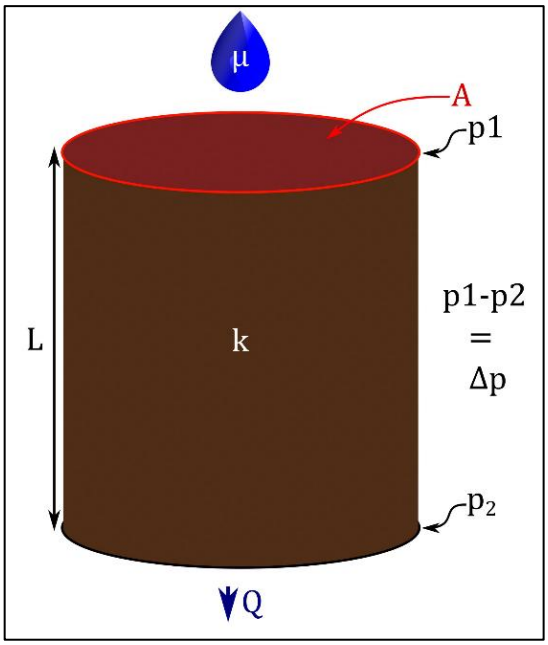


Fig. 3 Illustration of darcy's law

$$Q = -\frac{A \times k \times \Delta P}{\mu \times L}$$

Q, the rate of flow (volume per time) is given by the area (A), conductivity of the medium (k), the viscosity of the liquid moving through the medium (μ), length of the medium (L), and pressure drop (ΔP).

A simplified version of Darcy's law for slow viscous fluids (like water where $\mu \approx 1$) can be created like this:

$$Q = -\frac{A \times k \times \Delta p}{L} = Ak\frac{\Delta p}{L} = Aik$$

The hydraulic gradient or "i" replaces $\frac{\Delta p}{L}$. "A" and "i" are here relatively simple to calculate or control, however "k", conductivity, the scalar for the "ease" with which the liquid can move through the medium is a bit more complicated. To measure it we usually move "k" to the left, resulting in this formula:

$$k = -QiA = -\frac{\Delta V}{\Delta t} iA$$

In the lab we can then keep "i", and "A" constant and measure Q, or the amount of water that moves through the medium in a set period of time; this method is called the *Constant-head method* (Klute, 1986, Chapter 28.4). We mostly measure and talk about this k in saturated conditions; saturated hydraulic conductivity is shortened to K_{sat}. K_{sat} can also be used as the scaling vector for non-saturated conductivity. The determination of this through the Constant-head method is however relatively expensive and resource intensive, although it can be argued that it's no more resource intensive than measuring soil texture qualities, especially when it comes to man hours.

6.1. *Vadose zone hydrology*

The vadose zone is the space above the water table, this zone is also called the unsaturated zone. Water still clings to this zone through adhesion and capillary action, but any surplus water percolates down to the water table, or out of the area, with enough time. Not only does the drainage impact droughts and floods, but the dynamics of the vadose zone is often a defining factor for the water table. Understanding the hydrological elements of the vadose zone is also important when talking about the movement of contaminants, and the larger topic of the soil water budget. The depth and thickness of this zone is very variable, it's not uncommon, in dryer areas, to have a vadose zone layer is several hundreds of meters thick, for comparison wet areas can have one that's only a couple of centimetres deep, and in *wet*lands, as the name sort of entails, it can be non-existent (Tindall et al., 1999, Chapter 1).

However, with heavier or sustained rainfall the vadose zone can often become saturated with water, as long as the percolation of the water downwards is slower than the accumulation of water. In adverse scenarios this can cause flooding, land slides and soil erosion. It's really under these saturated conditions we see K_{sat}, *saturated conductivity,* in its truest form.

Not only is the vadose zone a controlling factor in water movements, but it plays a big part in transport of chemicals, surface runoff, groundwater recharge and evapotranspiration. The vadose zone also includes the root zone of the vegetation above it, acting both as a growing medium and water source. Roots also play a big role in chemical transport. Because of this important connection between the root- and vadose-zone, the vadose zone often becomes a defining factor in plant growth. Ideally in agricultural systems we would find a very thick and decently moist vadose zone allowing for deep roots and good water availability (Tindall et al., 1999, Chapter 14).

7. 'Pedotransfer Functions' – *a necessary imperfection.*

Accurate measurements and parameterization of soils are necessary to inform our soil related decisions. However, collecting the data for this is both expensive and labor-intensive, additionally we often also need these data in a range of spatial and temporal dimensions. The longest running and most spatially diverse datasets are usually national or regional soil surveys. They are however often limited to a smaller set of metrics. Pushing into the future of big-data models we're also experiencing an ever-increasing need for larger datasets. To reduce the amount of measurements we need to do we create *knowledge rules* called "Pedotransfer Functions" (PTFs); which can take a set of measured variables and estimate the variables we haven't measured (Van Looy et al., 2017). This not only saves time, but also allows us to make more "complete" datasets from less expansive ones. It follows that the effectiveness, and especially the confidence intervals, of these PTFs has a large impact on, not only the models efficacy, but also our ability to trust any large-scale soil data. One of the ways soil scientists try to generate these models is through empiricism. By collecting data from the real world and building statistical models based on the patterns in the data we're able to build relatively good estimators for our estimands and potentially arrive at new discoveries. The downside here is that we do not necessarily know the exact mechanisms behind pattern, despite this the models can often, very accurately, describe their estimands.

K_{sat} being one of the most important measurements in soils, and one of the more laborious and expensive parameters to measure, make it the focus of many PTFs. The first attempt was

done in 1984 by Cosby et al. The paper used simple linear regression models to estimate the *mean* values for a set of hydraulic properties, one of them being the average K_{sat} for different soil texture types, in the later literature this type of PTF is referred to as a "Class PTF". However, as hydraulic properties can vary immensely within the same texture class, even for soil samples that essentially have the exact same texture. Later models are trying to do something a bit more difficult; estimating the actual K_{sat} of a singular sample, this approach is referred to as a "continuous PTF".

8. Predictive modelling

Predictive models, such as PTFs, are models that use a set of one or more predictors (input) to generate a set of one or more estimands (output). These kinds of models have become a part of everyday life. Most peoples first experience with models like this were probably weather forecasts, and lately with the advent of *large language models* (LLMs) these predictive models have been launched into zeitgeisty super-stardom.

Predictive models usually come in two forms: classification, and regression. Classification models try to pick what category of the input set belongs to; "is this painting by Van Gogh or Mondrian?" Regression models try to estimate the "missing" value of a variable based on the set; "how big will this dog become?" In the environmental sciences both models are used frequently, making the basis of climate forecasts or land cover classification.

There are many ways predictive models try to achieve their estimands like Linear (Regression) Models (LM / LRM), and Machine Learning (ML). ML models can further be split into models like Neural Networks (NN), Support Vector Machine (SVMs), and Decision Tree (DT) based. The reason for the considerable number of differing models is both historical, *better models have been developed over time*, and because of application fitness, *certain models are better at specific things*. The choice of model *should* fundamentally come down to the needs of the task at hand.

8.1. Machine learning

There is probably no statistical topic more popular currently than Machine Learning (ML). Large language and image generating models are improving rapidly and their use have become prevalent. However, the general understanding of ML and these models is still quiet poor. Models like these are simply “black boxes”, their methods impenetrable to the user. Hidden behind their black exterior, which is easily misleading for the general public, are simply advanced statistical models trying to estimate the most likely answer. Today’s language models often beats the classic Turing test; even convincing one Google AI-engineer that it had consciousness (Tiku, 2022), although most experts agree that it’s not possible for a traditional computer to ever become conscious (Huckins, 2023).

The basic structure of these more famous models are neural networks, which are notorious for giving good predictions but opaque explanations for how that prediction came to be. However, there are also “logical decision tree”-based ML models, which, although not truly transparent, can be much easier to grasp. Depending on the model size a very basic “decision tree”-based model might be as simple as one decision tree with a few layers (Fig. 4).

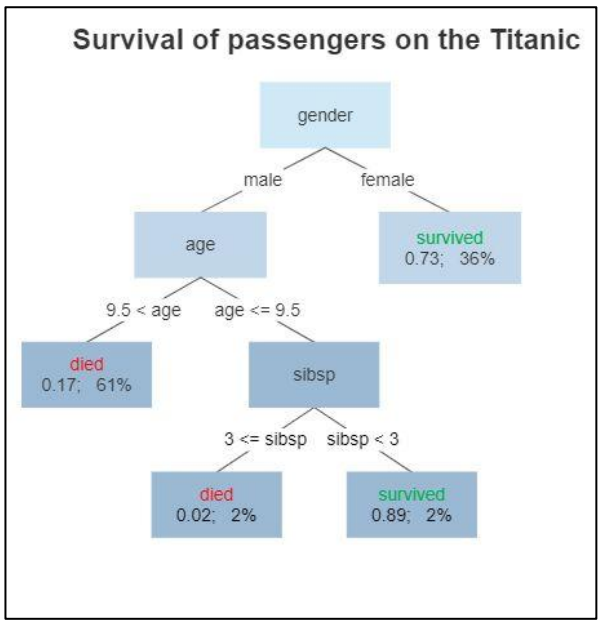


Fig. 4 Decision tree classifying the likelihood of a passenger surviving the sinking of the Titanic. (Gilgoldm, 2020)

The classic example is classifying whether a passenger on the titanic (probably) survived the shipwreck based on the characteristics of said passenger. Looking at characteristics like registered gender, age and number of siblings, we can using a decision tree estimate the likelihood of survival (Fig. 4 (Gilgoldm, 2020)). Starting at the uppermost node we follow the branches corresponding with the passengers personalia down until we end in one of the last leaf-nodes, which gives us the likelihood of survival.

8.2. Random Forest models

Although traditionally these logical decision trees were made by hand, we can now use computers to generate them based on a training dataset. Generating an array of these trees and assembling them into an ensemble model is called a *random forest*. A random forest allows the rather limited structure of a singular logical decision tree to become a part of a more flexible whole. Taking the most common or average result from the set of trees allows us to do both regression and classification, and often gives us a more accurate result than a singular decision tree, when doing predictions for more complex topics.

8.3. Exploratory modelling

Modelling of this kind is not only useful for generating estimates, but we can also use them for exploratory analysis of datasets, or put more simply: we can use them to find interesting patterns in our data. One of the main ways we do this is generating a model and then analysing how the predictors interact and produce the estimate. This works especially well on larger datasets with non-obvious and non-linear relationships between predictors and estimands. This can be used to inform our hypotheses, further analysis and finally as an indicator of how things work in real life.

8.4. *Common issues with ML models*

Building ML models comes with many common pitfalls, and a firm hand is necessary to create a model that's without any glaring issues. Overfitting, making a model that is *too well fitted* to the training dataset, and because of this, not fit for the job it's intended for, is a very common issue. Fig. 5 has examples for both regression and classification examples. A model like the ones on the left will have a really good performance on the training dataset but have worse performance than the "right fit" models when running predictions for new datasets.

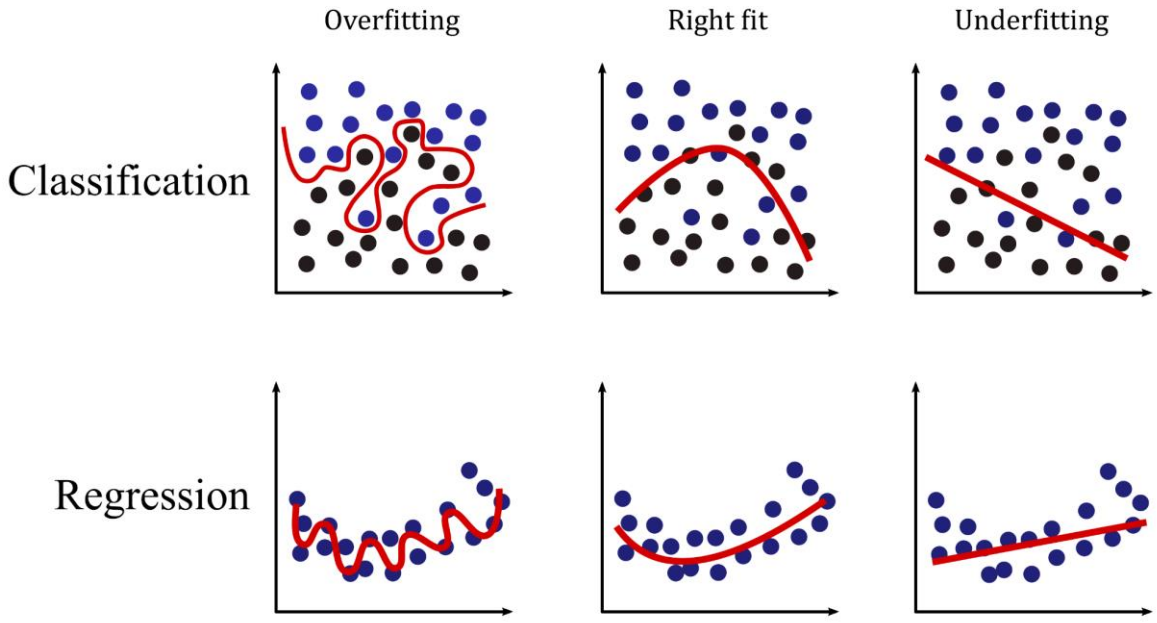


Fig. 5 Examples of over, under and correct fit of both a classification and regression model

We can correct for this by always working with both a training and testing set, that both have similar ranges of values. This last part, often called *stratifying*, is done by making sure both datasets have a representative amount of data in the differing quantiles for one or several variables. This is also a small part of having unbiased data, however most of the work towards having unbiased data is in how the data is collected. The old adage of "garbage in, garbage out" is especially true for ML models, as there is often very little intervention possible after the data is already collected, and any modeller or team of modellers without the necessary knowledge and experience with the type of data can easily miss poor data and build models around them. There have been a few famous examples of this, like a racist sentencing AI (Brannon, 2024), GPT-4 being more likely to suggest the death penalty for defendants speaking African American English (Hsu, 2024) and other similar cases.

9. The algorithms building our models

9.1. Boosting algorithms

Often generating these ML algorithms once will not get us satisfactory results; so, we take these models and try to improve them in some sort of formulaic way. For random forests this is usually this is either done through bagging or boosting a model. The clearest difference here is the difference between parallel and sequential improvement. Bagging, the parallel and more traditional method works by training several models at once and then putting the models together to create a better version of these models. Boosting, the sequential method, works by first creating one model, then assigning weights to different parts of the training dataset, where the estimated values that were the furthest away from the measured values are weighted with greater importance. Then this is fed into the next model which is now adjusting for the weighted parameters. This self-correcting type of model building gives better results but is also more likely to become overfitted to the training dataset.

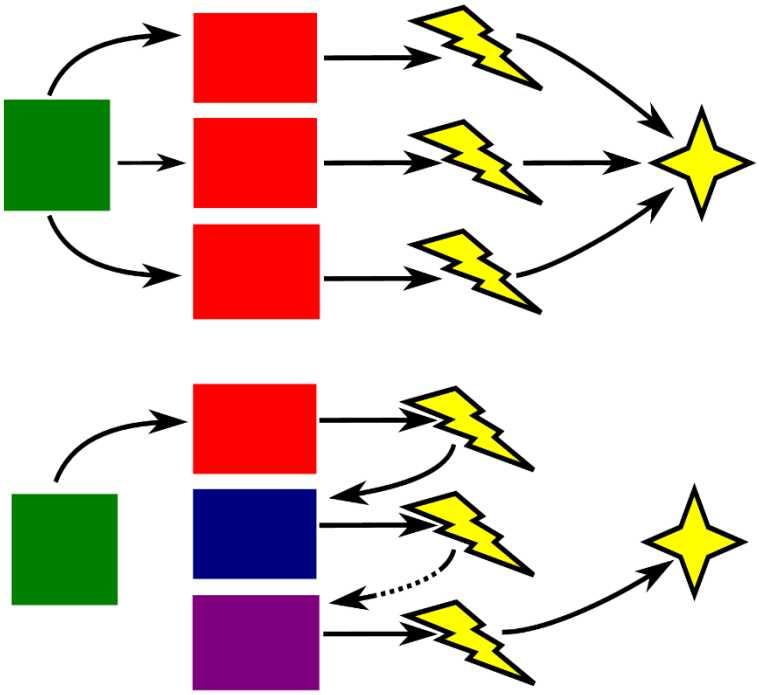


Fig. 6 Drawing of a parallel and sequential model building method.

We are using the XGboost (or eXtreme Gradient boost) model. It's a regression tree based gradient boosted model; a regression tree being a decision tree where each leaf (a ending node on the decision tree) has a continuous score. The score for these regression trees is then added together to give the estimate. The model starts out with a singular regression tree, then the difference between the estimated and measured values (residuals) based on this tree are calculated as a gradient, then a tree that corrects for this gradient is added to the ensemble. This process is repeated looking at the ensemble of trees, rather than the singular tree. This is done until the model is no longer being improved by adding more trees (Chen & Guestrin,

2016), depending on the task the final number of trees is usually in the hundreds or thousands for models around our size.

Additionally, to make sure that we're not overfitting the model, we've manually tuned the step size, or "learning rate" as it's called for XGB-models, to a place where we get similar residuals for both the training set and the testing set. The step size is how small of a change is allowed in each boosting repetition, if the value is too small the model will be overfitted, but if it's too large we would be leaving performance on the table.

9.2. Tuning and Anova races

When building XGboost models the hyperparameters like *number of trees*, *min_n* (minimum number of datapoints needed for a node be split further) and *mtry* (minimum number of predictors for each tree), can be automatically tuned for the model. For this task we can use something like the *Anova race algorithm*. "Racing" here means creating a set (often 20) of models in parallel and adjusting their hyperparameters for each step until some of the models are statistically different from the others, using an a *Anova* test on the RMSE of these models, giving the model its name. In a step where one or more of the models are different enough for it to be statistically significant, the worse models are eliminated and does not go on to the next step. This is done until one model is left or after a set number of steps, if it is the latter the best model is chosen although it's not statistically different from the competitors.

10. Model agnostic testing and dataset level explanations

One issue in the ML-sphere has always been testing. Making models has become easier and easier, but testing them has always been more tricky, there is a pretty good reason for this: we have to invent the model first, and then the tests. However with the advent of model *agnostic* tests, test that that can be used independently of the model type, we are now in a world where different models can be tested using the same metrics and plotting. This not only makes it easier to make and compare different models, but it also makes it easier for others not involved in the making of the model to test and critique the models. Which is important if models are to be used in the real world.

There are two levels of model agnostic analysis, instance level and model level. Although looking at how a particular estimate was reached can be useful (instance level), especially for diagnostics, it does not give us the larger understanding of how and why a model works. Dataset level analysis lets us look at things like performance metrics, what variables have the largest impacts, and how variables affect the model.

10.1. Model performance metrics

Previously we've talked about training datasets, the data that the model is trained on, in addition we usually have a testing dataset. They are usually parts of a larger dataset, that has been split into two. This allows us to evaluate the performance of the model after creating it. The most common metrics are error and performance measurements. In the former category we are looking at both MAE, MAD, and MSE and RMSE.

MAE is the mean absolute error and is calculated as:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

We take the sum of each absolute value of the difference between the prediction y_i and the measured value x_i and divide this by the number of values n.

MAD is the *median absolute deviation* and is calculated as:

$$MAD = median(|X_i - \tilde{X}|)$$

where $X = \{x_1, x_2 \dots x_n\}$ and $\tilde{X} = median(X)$. MAD is then the median value of the difference between a measurement and the median true value and is a measurement of statistical dispersion.

MSE is the *mean square error* and is calculated as:

$$MSE = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n}$$

It is quiet similar to MAE however we square the values instead of taking the absolute value, because of this MSE is more sensitive to large errors, which is often useful. RMSE is the square root of MSE, which shows the MSE value in the value range of the original dataset, which makes it easier to interpret.

Our main performance metric R^2 or *the coefficient of determination.* It shows how much of the variability in the dataset can be explained by the model as a fraction of 1, put more simply if we get an R^2 of 0.8 the model can explain 80% of the variability of the dataset. We're using a version where R^2 can go into the negative values. The formula is as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

SS_{res} being the sum of the squares of the residuals, or difference between the measured value and the estimated value, and SS_{tot} is the sum of squares for the observed data. Models that on average gives residuals bigger than the difference between the measured value and the mean measured value, also sometimes called a baseline model, will have a R^2 less than 0. ('Coefficient of Determination', 2024)

10.2. Residual diagnostics

10.2.1. Reverse cumulative absolute residual distribution

Residuals can be used for more than just giving us a simple numeric measure of performance, we can also plot them in a few ways to gain insight into our model's performance. One way to do this is to look at the *reverse cumulative distribution of the absolute residuals*. In Fig. 7 we can see such a graph. The residuals are sorted in order of lowest to largest and mapped out in a *reverse cumulative* step diagram. The X-axis shows the proportion of the residuals and the Y-Axis show absolute value of the residual. The resulting line tells us a lot about how the residuals are distributed between the samples. The head (blue) and tail (red) of the plot is of special interest. A sharp fall in the head, like in Fig. 7 tells us that a large part of the residuals is small (here below 0.5). The shape and extent of the tail tells us about the estimates with the largest residuals, how large these are, and the number estimates that have a large residual.

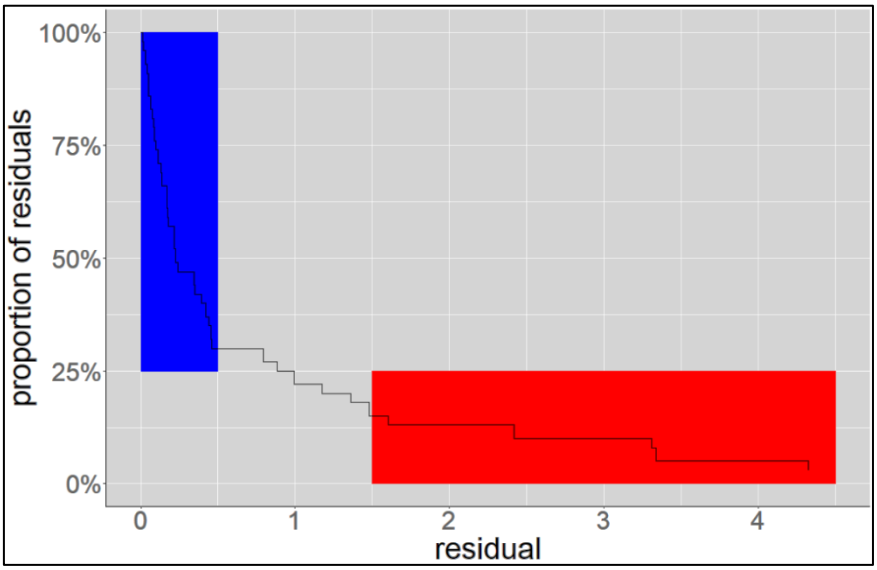


Fig. 7 Reverse cumulative distribution of absolute residuals.

When testing differing models against each other comparing these lines is a quick way to look deeper into their results. Models might have similar R^2 values but differing curves, where one model might have a very steep head and a long tail another model might have a much shorter tail, but a more gradual head. Depending on the task at hand either might be a better fit. Specifically, the former would be better, for most cases except for outliers and the latter would be a better allrounder. As the difficulty of prediction increases so does the likelihood of a tail, especially when trying to predict values with a range over orders of magnitude.

10.3. Variable importance measures

Variable importance measures are an important step in optimizing and understanding models. Variable importance measures can be used early in development to simplify the model by removing low impact variables, it can also be used to tune models during development, and it might be used at the end of modelling to figure out what's the most impactful on the model and potentially what impacts the estimand in real life (Burzykowski & Biecek, 2020).

10.3.1. Variable importance

There are many ways to estimate the impact of a variable in a machine learning model, but *RMSE loss*, is for our purposes the most interesting, as it's model agnostic, allowing us to run the same a wide range of different models for testing purposes, and because it's commonly used, reliable and easily understandable (Burzykowski & Biecek, 2020). RMSE loss is the increase in RMSE when a variable is removed, an important variable will have a higher RMSE loss. We calculate this for a single estimate at a time and take the average of these. We can also use graphing options like boxplots to show the range of RMSE loss from each estimate and for each variable.

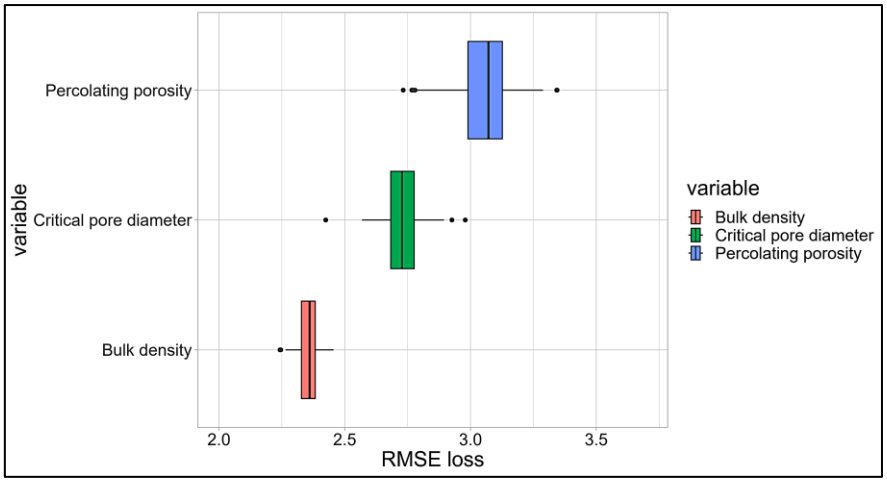


Fig. 8 Example of variable importance plot for a K_{sat} PTF using percolating porosity, critical pore diameter and bulk density

10.3.2. Shapley- & SHAP-values

Shapley-values is an idea borrowed from cooperative game theory, where L.S. Shapley described a way to determine the effects of each player in a game with a coalition of players (Shapley Lloyd Stowell, 1953), translated into the language of machine learning: what is the effect of each singular predictor on final estimate of the model. This idea has been implemented many times with slight differences (Sundararajan & Najmi, 2020), since the first implementation for machine learning by Lindeman et al. in 1980, which introduced the concept as a way to break down the effects of differing variables on R^2. In 2010 Strumbelj & Kononenko reintroduced the idea, but now to explain the effects of the variables for individual estimates, as a solution to the "attribution problem" or the issue of how to describe *local variable importance*. Strumbelj & Kononenko coined *conditional expectations Shapley* as the name for their idea. However, it was only with the 2017 paper by Lundberg & Lee and the 2018 release of Lundbergs accompanying python package (*Shap/Shap*, 2016/2024) that it was more widely adopted in machine learning spaces. The Lundberg & Lee paper introduces the term SHAP, or *Shapley Additive Explanations,* for conditional expectations Shapley, and with the proliferation of the Shap package and its derivatives "SHAP" has become the term most commonly used.

We can use a simple classification case as an example. We can make a model that tries to classify[1] whether a painting is by Mondrian or Van Gogh based, and as parameters we use:

- the country the artist was born in,
- what century the painting was made,
- what the main color of the painting is,
- the genre of painting.

Both Mondrian and van Gogh are dutch artists, *mostly* working in different centuries (1800s and 1900s), with very different styles and choices when it came to what and how they chose to paint. Here (Fig. 9) we can use one of each artists most famous paintings as an example. Mondrian's *Composition with Red, Blue, and Yellow* (1930 (A)) is famous for it's striking color composition, especially the large red square in the top right corner. The parameters for this painting would be "The Netherlands, 1900s, Red, Abstract". Van Gogh's self-portrait (1887 (B)) would be parameterized as "The Netherlands, 1800s, Green/Blue, Self-Portrait".

To look at the *total* effect of each parameter we can look at whether the likelihood of each answer *changes* if the parameter was changed or stayed the same, for each possible combination

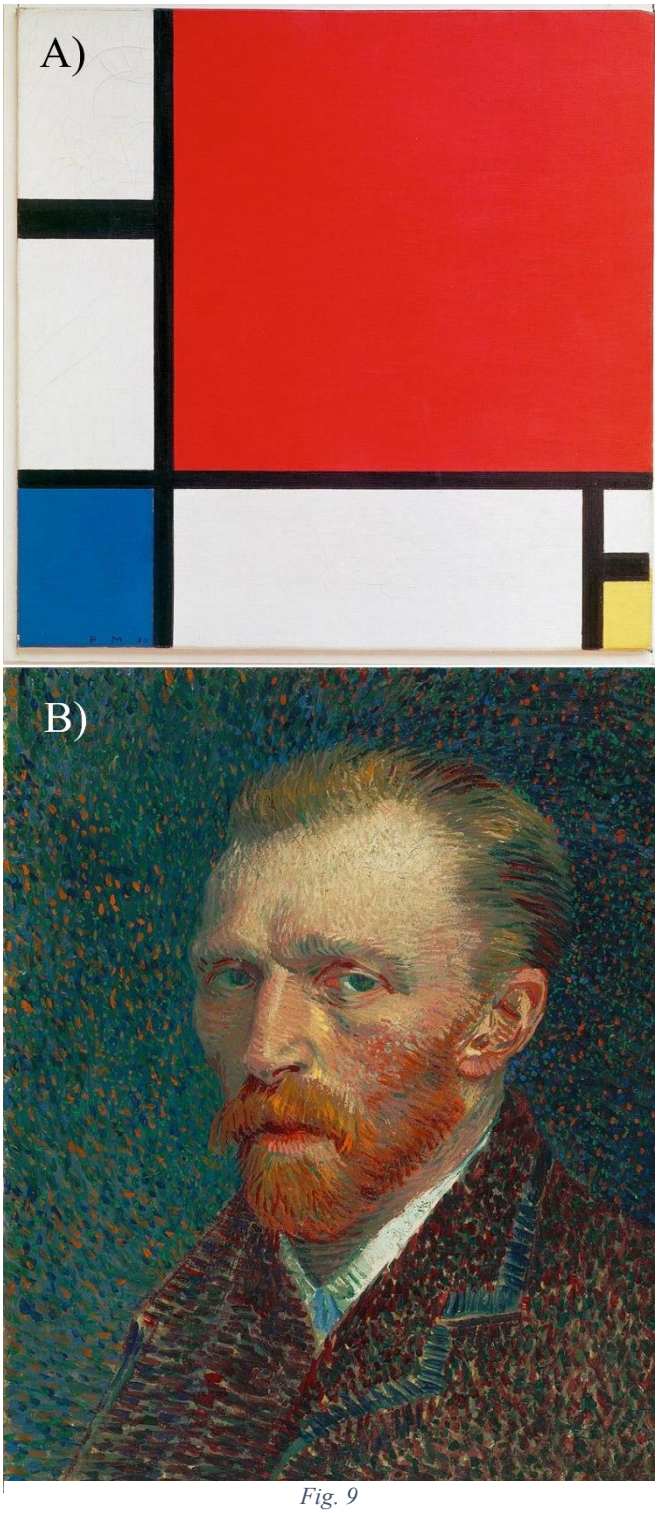


Fig. 9
A) Piet Mondrians "Composition II in Red, Blue, & Yellow", 1930
B) A self-portrait by Vincent van Gogh, 1887

[1] Note that this model would need to be a probability-based classification model, as SHAP values can only be calculated for numerical estimates.

of the other parameters in the model. If we do this for the artist country of origin, the impact should be 0, as both artists are Dutch. The impact of what century the painting comes from should however be much larger as Mondrian has only active for a few years in the 1890s, in what is otherwise van Gogh's century. The main color will have a lesser effect as both are known to have used strong colors in their paintings, however (with the exception of the few snowy landscapes van Gogh's painted) Mondrian's use pure white would probably make an impact. Lastly the genre would also probably have a large effect on the model, Mondrian is mostly famous for his work with abstract paintings, a genre that didn't exists during van Gogh's time, which in turn would be a dead giveaway to who had painted the painting.

However, as the complexity of a model increases calculating the effect of each parameter becomes more and more resource intensive, especially when we take continuous variables into account. The SHAP method therefore uses the *Monte Carlo method,* also known as repeated random sampling, estimating the values by using a large set of random values instead of every possible value.

Let us estimate the SHAP value "$\hat{\phi}$" for variable j out of p variables in model $f(x)$ for a number of Monte Carlo iterations $(m, \dots, M)$ as:

$$\hat{\phi}_{x,j} = \frac{1}{M}\sum_{m=1}^{M}(\hat{f}(x_{+j}^{m}) - \hat{f}(x_{-j}^{m}))$$

where, working from the right-side in, we add some random values, designated by "z", replacing the original input value for a variable "x", to a random number of variables in two slightly different sets of the variables:

$$x_{+j}^{m} = (x_{(1)}, \dots, x_{(j-1)}, x_{(j)}, z_{(j+1)}, \dots, z_{(p)})$$

and

$$x_{-j}^{m} = (x_{(1)}, \dots, x_{(j-1)}, z_{(j)}, z_{(j+1)}, \dots, z_{(p)})$$

The only difference between the two is that variable the value for j is also *random* in x_{-j}^{m}. The subtraction gives us $\Delta\hat{f}x_j^m$, sometimes written as ϕ_j^m in the literature (Molnar, 2022). It's important to note here that the order of the variables, the number of randomized variables and the position of variable J is random for each repetition (m), this is done to reduce the likelihood of another variable having an impact on the estimate. $\Delta\hat{f}x^m$ is then the difference between the estimations by f for the variable sets where variable j is the original value or a

random value. The summation of the M (Monte Carlo) iterations and it's divis 'ion on M (again number of iterations) gives us the estimated $\hat{\phi}_{x,j}$ for variable j and model x. (Burzykowski & Biecek, 2020, Chapter 8; Molnar, 2022, Chapter 9; Štrumbelj & Kononenko, 2014)

Doing this for each variable gives us a numerical break-down of the effect of each variable in the same numerical scale as the estimand.

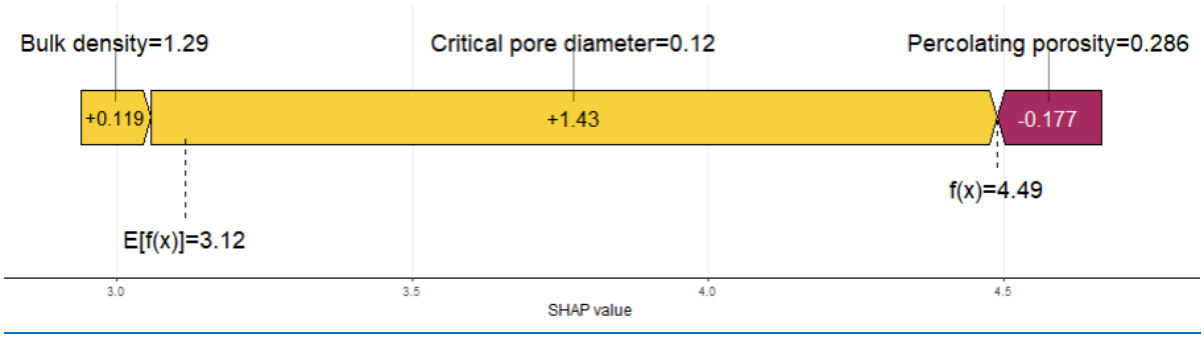


Fig. 10 Force plot of shap values for a estimate of K_{sat} using 3 variables, percolating porosity, critical pore diameter and bulk density. f(x) is the estimate, E[f(x)] is the baseline value. Values next to the name is the measured value for each variable

10.3.3. Graphing SHAP values

Graphing SHAP values are a reliable way to look at the effects of variables and their interactions at both the instance and dataset level.

On the instance level we often use the force plot (Fig. 10), or the waterfall equivalent (Fig. 11) for models with a larger number of variables. These graphs take the estimated SHAP value ($\hat{\phi}$) for each variable used in the estimate and plot them together showing how they in total combine to be the estimate, or at least very close to the estimate, due to the slight inaccuracy of the Monte-Carlo method.

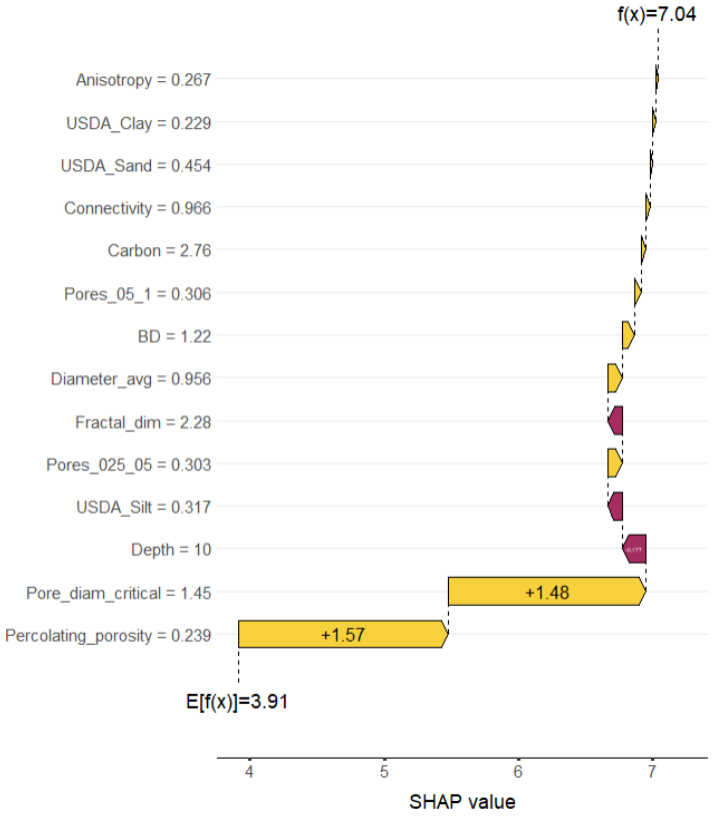


Fig. 11 Waterfall break down of SHAP scores for a XGB model estimating K_{sat} using a large number of variables. f(x) is the estimate, E[f(x)] is the baseline value for the SHAP values. Values next to the name is the measured value for each variable.

These instance level measures of importance can also be generalized to the whole dataset. One way to do this is taking the average of the

absolute SHAP-value. We need to take the absolute, as this shows what variables have the largest effect on the result, the same variable can have both a large positive and negative SHAP value dependent its value. This average absolute SHAP-value is another way to estimate the dataset level variable importance.

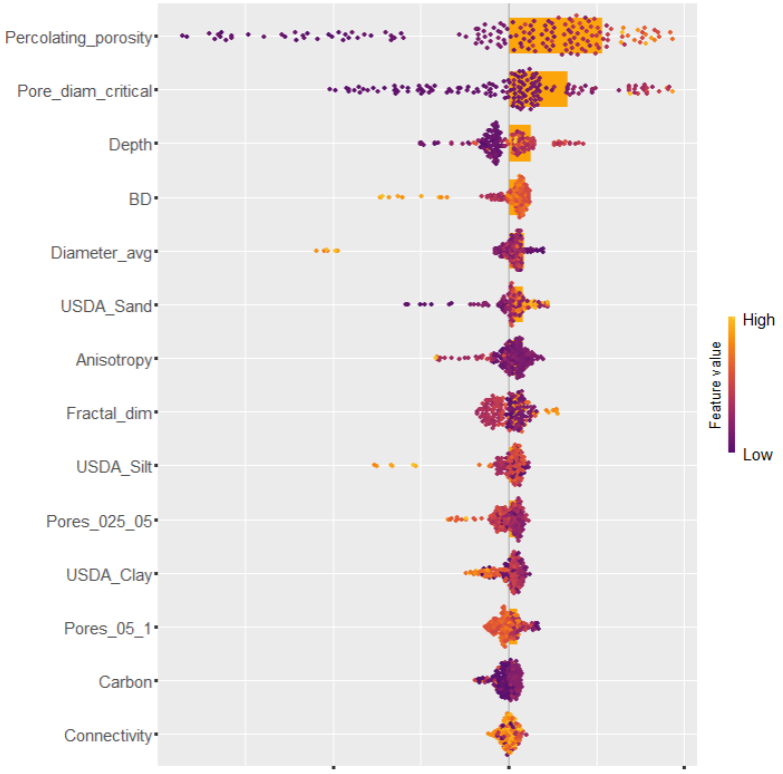


Fig. 12 A SHAP value "Beeswarm" plot of the same XGB model as Fig. 11. Showing the dataset level distribution of SHAP scores and feature/variable value. The orange bars show the average absolute SHAP-value for each variable.

A *beeswarm* plot (Fig. 12) is often shown together with the absolute average SHAP value (orange bars). What has been a until recently a somewhat esoteric plot-type, has found popularity in plotting SHAP-values, as it can show both the original values, and SHAP-values succinctly. It also allows the reader to see groupings, the range of SHAP-values for each variable and conceptualize the effects of differing variables on the model.

The force plot (Fig. 10) can also be extended to a whole dataset, into a dataset force plot. Following the principles of the original instance level force plot (Fig. 10), each instance is mapped out along the X-axis with SHAP-effects mapped around the Y-axis. These can be grouped by statistical similarity of the SHAP values like in Fig. 13, which can be an indicator of overfitting or similarity between samples.

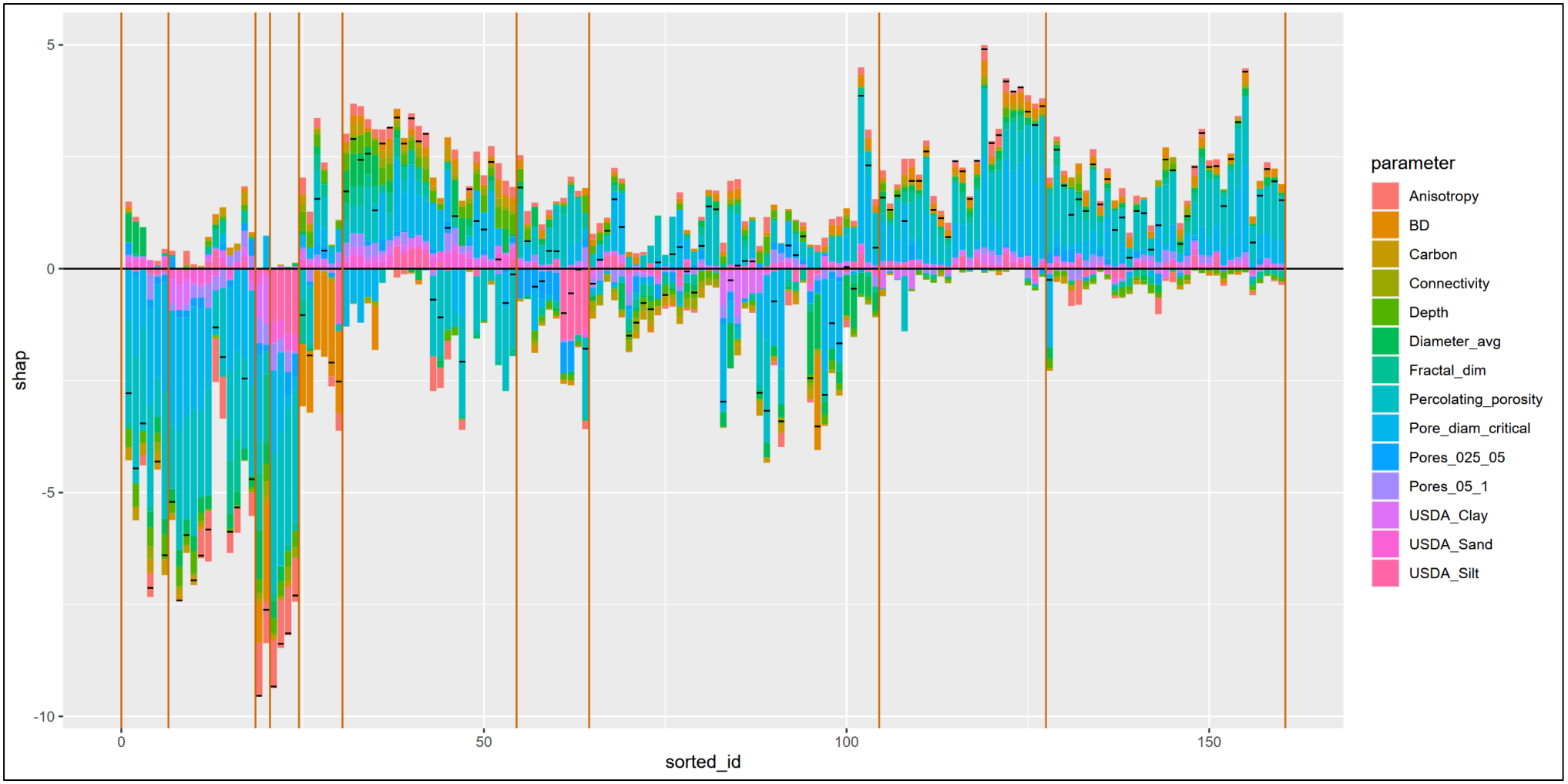


Fig. 13 Dataset level force SHAP plot of the same XGB model as Fig. 11. Orange lines deliniate between different groupings. Groupings are sorted by similarity. The black lines represent the final estimate for each prediction.

Lastly differing 'SHAP-interaction' plots are often used when looking at the effects of differing variables. There are many ways to do this however we'll just look at 2 different examples. The simplest is just plotting the SHAP value for a variable as a function of the variable value (Fig. 14). This can be used to look at how a parameter impacts the estimated value, which can lead to insights into how a variable might impact the estimand in real life.

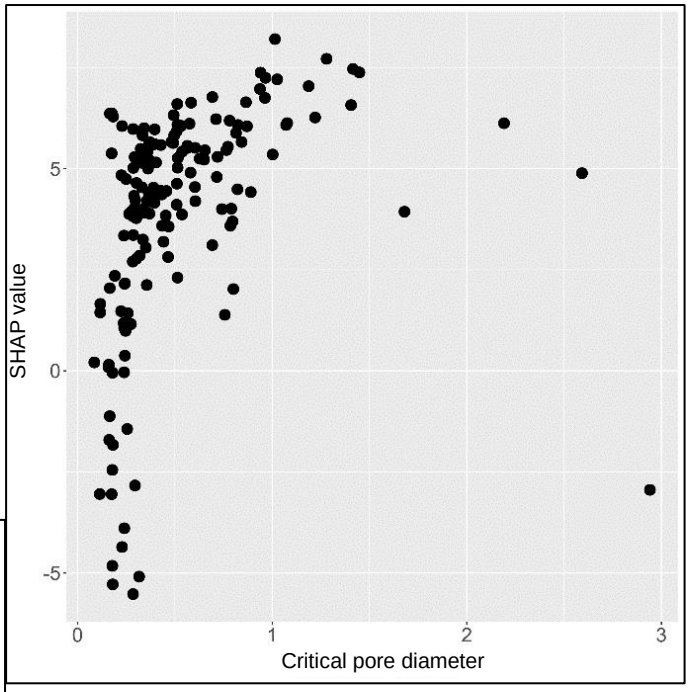


Fig. 14 SHAP-interaction plot for critical pore diameter. The SHAP values from critical pore diameter is plotted agains the measured critical pore diameter value. This is from a model only using critical pore diameter and percolating porosity.

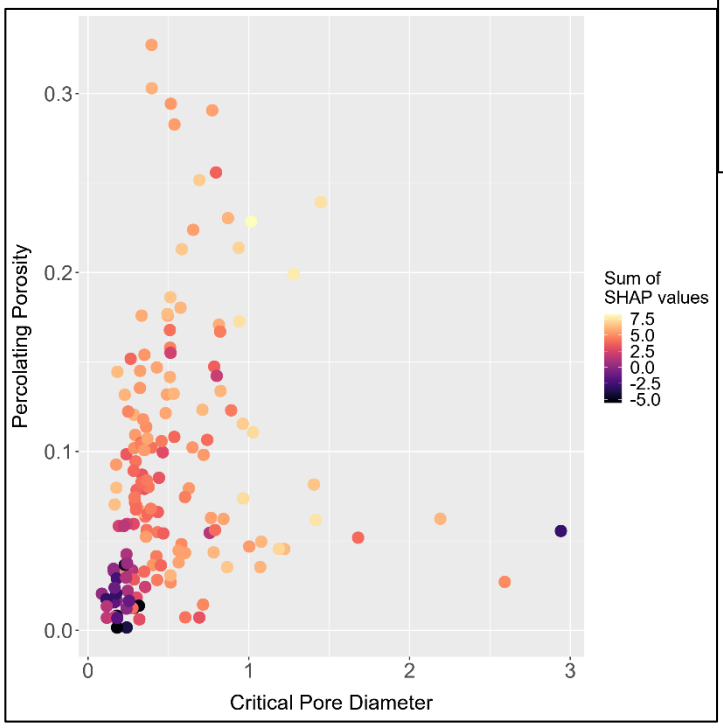


Fig. 15 SHAP-interaction plot from the same model as Fig. 14. Plotting critical pore diameter and percolating porosity against each other and using the sum of the SHAP values (corresponding to each instance) as point colour.

Further we can use the same principles to look at how the interaction of two differing variables interact with the result. A simple way to do this is by plotting a simple scatterplot with the colour as the sum of the corresponding SHAP-values (Fig. 15). This allows us to look at the interactions between the two different variables. As an example we can see that the estimated K_{sat} values in the simple model in Fig. 15, which only uses the variables present in the figure, is ver y low for values with a low percolating porosity & critical pore diameter.

10.4. Variable dependence

In a similar vein as SHAP-values 'partial dependence'-profiles show the effect a variable has on a model as a function of its value. It's often used for comparisons between models, similar

'partial dependence'-profiles between several models are a reassuring sign that the models are stable and well fitted to the task at hand. Obvious differences call into question the legitimacy of one of the models, as an example a seeing a non-linear relationship in certain models could make it obvious that simplifying the model to a linear one would *not* be a good idea. Furthermore, with empirical models it's often a good idea to look at the 'partial dependence'-profiles on the fringes of the datasets, where a flexible model is more likely to over fit to the few available values (Burzykowski & Biecek, 2020, Chapters 17–18).

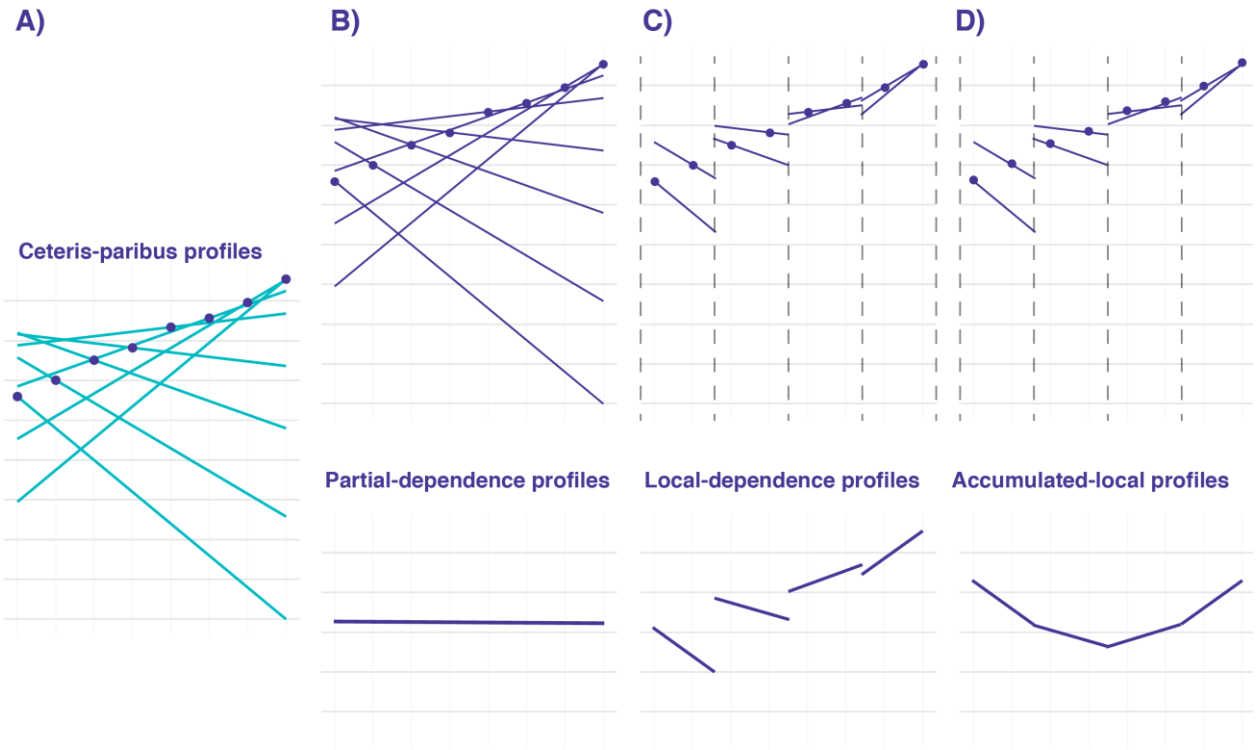


Fig. 16 Comparison of Ceteris-paribus profiles, Partial-dependence profiles, Local-dependence profiles, and Accumulated local. Borrowed from Burzykowski & Bieceks, 2020 (Figure 18.4). The purple points in are the sampled points that profiles are based on.

There are however a few different ways to calculate and show 'partial dependence'-profiles. Looking at Fig. 16 we see Ceteris-paribus profiles and the 3 main ways to show 'partial dependence'-profiles, traditional partial-dependence, local-dependence, and accumulated-local-dependence. The basis for all of these 'partial dependence'-profiles is calculating the Ceteris-paribus profiles, which is done by selecting a parameter and seeing how the estimation of the model changes as this parameter changes, which leaves us with a series of curves or lines (aka profiles) like in Fig. 16 A). 'Partial dependence'-profiles then becomes 3 separate ways to simplify these profiles into one singular profile. Traditional partial dependence, the most common and oldest of these, has been used since at least the late 1970s (Harrison & Rubinfeld, 1978). However, it has it's issues, especially when working with

models that have correlated explanatory variables (Burzykowski & Biecek, 2020), which is usually the case when working with data from the real world. A simple example could be taken from limnology. Usually, when measuring fish, we take both length and weight measurements, as long and thin fish can indicate something like a change in the food availability, however it's impossible to get a fish that weighs 2 kg but is only 5 cm long, although both are values that aren't unlikely to be found within the same dataset. The ceteris-paribus-profile that would be made by this discrepancy is misleading, and it's impact on the partial dependence profile would again make this metric misleading. This is where local-dependence and accumulated-local profiles come in, they only take the parts of the ceteris-paribus-profiles that are around the measured values, as shown in part C) & D) in Fig. 16. The difference between these two is that the accumulated plot considers the previous point into account during calculations. We can see all 3 'Partial dependence'-profiles for the same model as in Fig. 17. Where the density of values is high can we see that the shape of the profiles are quite similar.

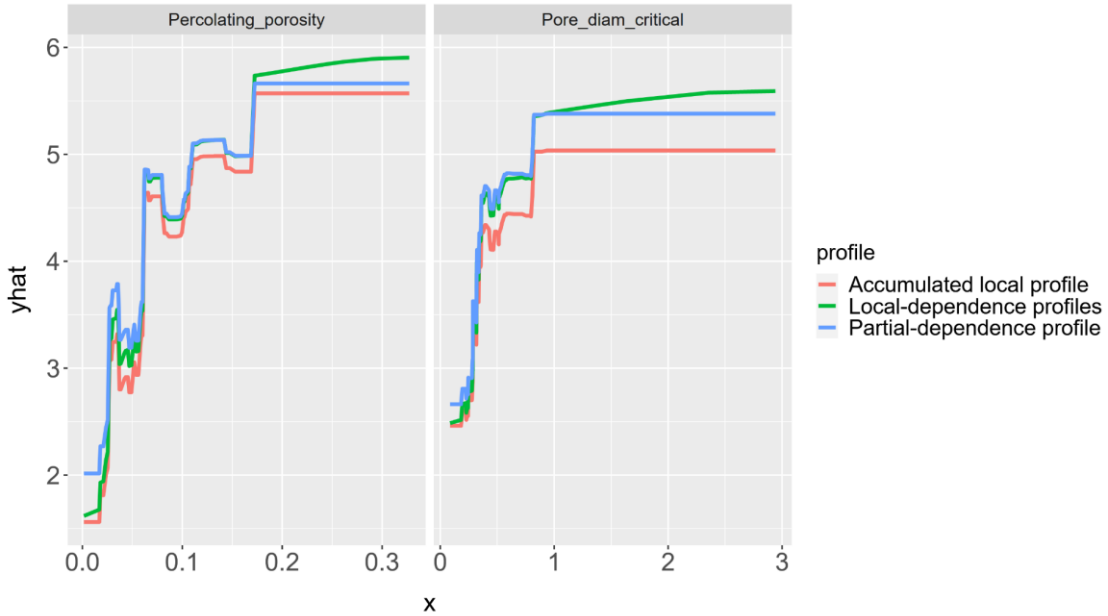


Fig. 17 Example of all 3 Partial-dependence profiles, here for the same model as Fig. 14. (N points = 160)

10.5. The impacts of testing

Integrating these different analyses allows us to gain a lot of insight into “black box” models and extends their explanatory properties. These analytics are also important during model development, as in Burzykowskis & Bieceks (2020) circular lifecycle of model development (Fig. 18). It’s based on these analytics that we build our models, validate them, and break them down to gain insights into what improvements we might gain from the new datatypes presented in the this text.

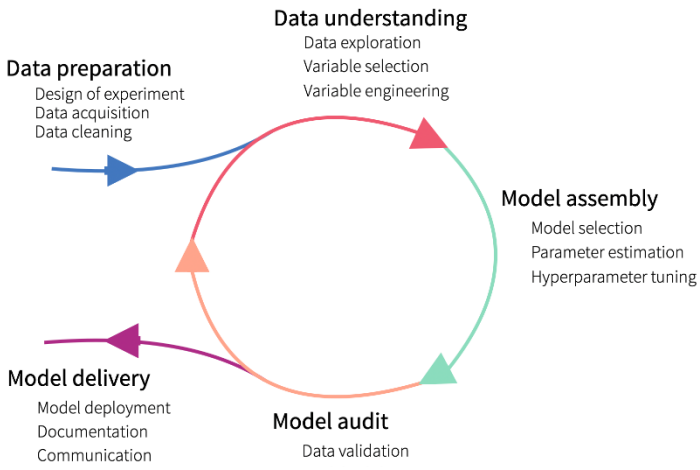


Fig. 18 (Tomasz Burzykowski & Przemyslaw Biecek, 2020)

11. Our model selection criteria

Formalizing our model selection criteria is an good way to ensure the strength and efficacy of the model. We've put together a ranking of both softer and harder criterxia and applied them to different models as a way to choose the right model for the job.

Model Criteria	Linear Models	Support Vector Machines	Random Forest	Xtreme Gradient Boosting	Neural Networks
Shareability*	xxx	x	xx	x	x
Interactiveness	xxx	xx	x	x	x
Low model input requirements	xx	x	xx	x	x
Ease of running the model	xxx	x	x	x	x
Transparency*	xxx	xx	xx	xx	
Human understandable structure	xxx	xx	xx	x	
Interpretability of variable effects	xxx	x	x	xx	
Ease of production*	xxx	xx	xx	xx	xx
Computer power needed for production	xxx	x	x		
Simplicity in building the model	xxx	xx	xx	xx	xx
DALEX compatibility	xx	xx	xx	xx	xx
Model strength & flexibility*	x	xxx	xx	xxx	xx
Prediction strength	x	xx	xx	xxx	xxx
Handling of mixed data	x	xx	xx	xx	
Nonlinearity		xxx	xxx	xxx	xxx
Variable selection		xx	xx	xxx	xx

Rows with * are the main categories and are informed by the subcategories underneath them. These metric are based in (Burzykowski & Biecek, 2020; Saloranta et al., 2003; Van Looy et al., 2017). The criteria are intended to inform the choice of model, however certain criteria have more weight than others. Additionally, we've highlighted in pink some of the criteria that are below the cutoff point for being useful in our case. Note that there a few different models that were considered for the task but were cut as they would not work with the dataset or had implementations that were strenuous to work with.

As seen in the table above we've narrowed our criteria down into categories to make it more easily understandable.

11.1. Shareability

For any model to have an impact, and even just to be scientifically valid, it must be able shareable, otherwise it cannot be cross validated, nor will it ever be used by others. Because of this, shareability is an important metric for any model. We've divided this category into: "interactiveness", low model input requirements, and ease of running the model.

"Interactiveness", or how easy it is for a user to use the model, inspect the model and interact with it as they'd like, is a very soft metric. However, it's one of the fastest ways to get someone to believe the model. Black-box models, as most of these are, have a big problem in their interactions with the end user. They ask for a lot of trust, without really assuring the user of why they should, trying to mitigate this will be an overarching concern in many of these criteria. A model that we can take apart and play with is more trustworthy, as it allows the user to understand it's internal mechanics. An example of a model with very little interactiveness, could be a neural network trained on the MNIST database ('MNIST Database', 2024). It's an common exercise to create ML model that can recognize letters based on this database, but the end product only ever really allows the user to input these very constrained images, and a neural network model based on this would have neither a structure that can be picked apart, nor would it give a understandable reasoning for it's result, creating very little "interactiveness". A simple linear model or a simple logistical tree model are examples of the opposite. Not only can we often throw a lot of varying fake data at it, but we can also piece together an understanding of how the model works by looking at it's structures, or at least a graphical version of them.

When it comes to ease of use, how easy it is for the user to *input* the data is an important metric. If the model needs specialized tools for inputting the data, user-performed transformations, unusual file types or data structures, it becomes a lot more difficult to work with the model. Most of the options we've shown only really let the user input the data as simple data frames, which can be imported as CSVs, which is a very simple and common format. However a linear model

can be even easier than this, as a person, with some time on their hands and a ruler, could print out a graphical representation of the model and give a pretty good estimate with some simple arithmetic. Our model works with quite advanced data, parameterized Xray CT images from SoilJ (Koestel, 2018), so we can expect the end users to be a bit more data-savvy, however the users ease of use should not leave our mind when deciding the type of model and how we choose to set it up.

In this vain, the ease of running a model is also important. In our case with R-object models, we're only really looking at two things, (1) how easy is it to get the model onto a computer, and (2) do we need any special packages to do so? As we've chosen to make all of our more complex models with tidymodels (Kuhn & Wickham, 2020) (and with analysis trough the DALEX package (Biecek, 2018) in mind), any of the models produced here can be run using the standard "predict" function from the base R package "stats". Getting the model is also as easy as retrieving the model-object and loading it into an environment.

11.2. Transparency

Transparency of the model goes hand in hand with shareability, and they do share a bit of an overlap, however both are important metrics with differing driving forces behind them. It's important to note here that, as we've chosen to focus on a model agnostic approach for our model analysis, we have some leeway to use less structurally transparent models and still retain a lot of explainability from the models. With this in mind have we broken this category into human understandable structure and interpretability of results.

A common issue with ML models is that the more advanced models have a less and less transparency as their sizes increase and as their structures become more difficult to understand. The current prime example is the deep-learning neural networks which is the basis of the large language models like the GPTs. Some of these LLMs can have more than 100 billion parameters, although their size might not improve performance (Leffer, 2023). Luckily we are working with much smaller models, which allows us to use more human understandable structures, with

slightly lower performance. SVMs and ‘Random Forest’-based models, only sacrifice some of their understandability for performance, as long as we’re looking at the lower ranges of parameters and sizes.

Interpretability of the results, or how we ended up with a specific result, is also important to look at. For most ML models, apart from neural networks, this can relatively easily be understood for smaller models, but it’s becomes more and more difficult as the size of the models increase. We’ve again chosen a model agnostic method to simplify the task of “explaining” each estimate (10.3.2 Shapley- & SHAP-values), again allowing us to be more flexible, as we can “simplify” the larger and more complex structures into more interpretable metrics, without losing too much information, or at least making it obvious when we’re losing detail.

We cast aside neural networks because of their general lack of transparency. Our model agnostic approach to analysis allows us to compare the results of such a model with our other models. Ultimately the structure of neural networks does not lend themselves to explanatory modelling the way SVMs and ‘Random Forest’-based models do.

11.3. Ease of production

Not only did we keep in mind the end user when it came to model choice, but also us, the model makers, and the process of creating the model. As model choice often comes down to whatever has previously been used by the user or close colleagues (Melsen, 2024), the ease of production or adaptation is often a very biased metric. However, we’ve tried to keep this to a minimum here, and use 3 modeller centric metrics: computer power needed for production, simplicity of building the model, and DALEX compatibility.

Although non of the models named here are extremely taxing to create, on our higher end modern computer, it’s still an important metric. The same model is probably created with slight variations hundreds of times. And consequently, a model that takes 1-2 minutes to make, can,

accumulatively, take up to several hours to fine tune. This needs to be seen in context with model performance, as this trade off might still be worth it in the end (like here).

Simplicity of building the models is also an important metric when thinking about how much work it takes to build the final model. Some models like linear models, are very simple to put together, often fitting into a singular line of code, while a more complex model with more hyperparameters, like the Xtreme Gradient Boosting model, take comparatively a lot more time to code and tune. However, the Tidymodels package (Kuhn & Wickham, 2020) does streamline a lot of this work.

Knowing that we wanted to use the DALEX package to compare different models in the experimental phase of the project and use the same metrics to investigate the final model, made the compatibility with the package important. A few models were considered early on, but disregarded as they did not work with the framework, which would make it harder to argue for them if we picked them over other models, as their metrics would not be easily comparable.

11.4. Model strength & flexibility

Finally model strength and flexibility quickly became important metrics for our models, as early linear models showed big weaknesses in predictive power. We've broken these down into, prediction strength (10.1 Model performance metrics), ability to handle mixed data, ability to deal with nonlinear relationships, and it's strength in variable selection.

Prediction strength quickly became one of the main issues when comparing the models, as an early linear model had an R^2 for the testing data of -0.5, meaning the estimates the model created was *a lot* worse than just guessing the average value of the dataset. This also informed the next thing, the need for nonlinearity in the models, which any of the other models discussed allows for.

We also wanted to focus on models that were well-suited for variable selection, as what variable was going to be the most impactful was a previously undetermined factor. Thus, Models that more easily will let more impactful predictors shine, like LLMs, Boosted Random Forests and SVMs, were preferred.

12. 3D-Xray computed tomography in soil science

12.1. CT-scanning in soil science

The presence of paramagnetic elements in soil has made X-rays the preferred method over MRI and NMR (Macedo et al., 1998). These tomographies, coming from the Greek *tómos* = 'slice' or 'section' and the French *graphie* = 'to write' ('Tomography', 2023), are ways to image an object by sections, and then piece them together into a singular object, allowing us to see not only the exterior of an object but crucially the interior. These methods are more commonly known for their use in the medical sciences, but X-ray imaging has been used in soil science since the 1982 when Petrovic et al. used it to quantify the change in bulk density using a medical CT system. It's an non-invasive & non-destructive way to image the geometry and topology of the pore networks in samples, which is entirely necessary if we want to investigate them (Helliwell et al., 2013; Koestel et al., 2018). It's also been used to estimate a large array of differing soil variables ex: bulk density, layer detection, porosity, pore network structures, volumetric water content (Helliwell et al., 2013). However, their accuracy is fundamentally halted by their penetrating power, which makes the fidelity of the images a function of the sample size. But, with the ever-present onset of new technology, we've seen large improvements in X-ray technology (Mooney et al., 2012).

12.2. Soil-J: Quantitative descriptions of soil images

John Koestels SoilJ (2018), a 3-D Xray image processing tool, is our way to parametrize these 3D-images so we can use them in our ML-models. It allows the user to automate the processing of cylindrical soil volumes, by including modules for different image segmentation & correction, morphological metrics, and most importantly for us, analytics for percolation. This automation allows for expedited analysis of these X-ray images, which had previously had to be done by trained professionals.

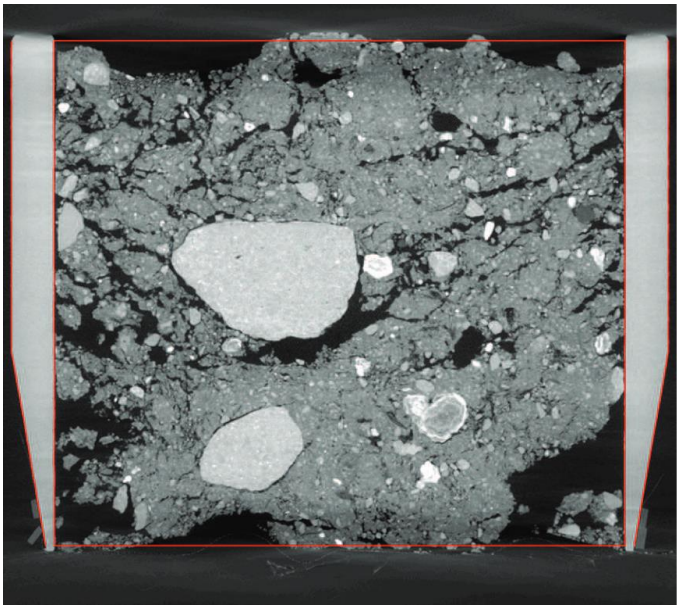


Fig. 19 X-ray cross section of a soil sample showing pores (black) soil and some gravel. This image is from the SOILSPACE project and has been borrowed from (Koestel, 2018)

This increases the accessibility of X-ray tomography for the soil sciences. Fig. 19 is from the paper publishing SoilJ (Koestel, 2018), and shows off the automatic sample detection. Fig. 20 shows off some of the pore parameterization that can be done with the package (Koestel et al., 2018).

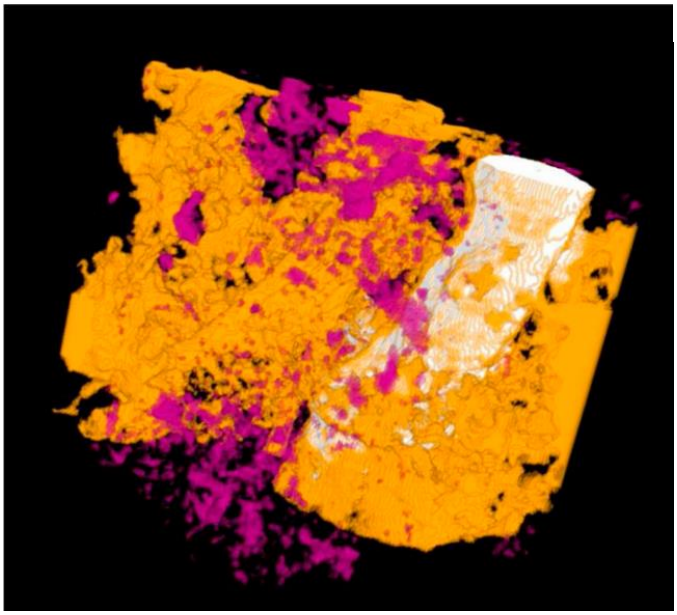


Fig. 20 An image showing percolating porosity (yellow + white) and the largest pore (white) which is used to calculate critical pore diameter. Purple shows non-percolating pores. The image is borrowed from (Koestel et al., 2018)

13. Hypothesis

Looking at previous K_{sat} PTFs we might've hit a wall basing our parameters on commonly measured soil parameters like soil texture, bulk density, and organic matter content; following this pattern no large advances have been achieved over the last 20 years. A common critique is that these techniques do not consider the pores of the soil. There is a particularly good reason for this, as trying to measure the pore-structure within a soil sample has been basically impossible; however, 3D-xray tomography provides an attractive, previously not fully available technique. The soil space project lead by NIBIO[2] had as an aim to "*Quantify the complex 3D soil pore system*", this thesis is a continued exploration of the data built by this project. Using their very robust dataset we attempt to build a new type of PTFs which are based on imaged parameters. We hypothesize that:

1. A model based on imaging parameters & traditional parameters will outcompete a traditional model.
2. The imaged parameters will have a significant impact on the resulting models based on variable importance measures.
3. A model based purely on imaged parameters will show merit in K_{sat} estimation.

14. Our aim – the why

Our aim is to improve our understanding of soil properties and extend the toolkit available to the people working with soils. We will attempt this by building a set of 3d-xray image based pedotransfer-functions for K_{sat} estimation and test them against a traditional model. In this we hope to also shed light on the importance of soil pores and their structure for K_{sat} in soils.

[2] The partners in the project was: NIBIO (The Norwegian Institute of Bioeconomy Research, Ås), Rutgers University (The State University of New Jersey, New Brunswick), SLU (Swedish University of Agricultural Sciences, Uppsala) and NMBU (Norwegian University of Life Sciences, Ås)

Improving the estimation of saturated conductivity through 3D-xray tomography

1. Abstract

Pedotransfer functions (PTFs) for continuous K_{sat} estimation have only seen small improvements over the last years despite of improvements in the machine learning space. This suggests that we need novel approaches to improve K_{sat} PTFs. One such approach is using parameterized 3D-xray tomography of soil samples. Here we use an extended methodologically homogeneous set of with both imaged parameters and traditional soil parameters such as texture, bulk density, organic carbon content and sampling depth to create a set of 3 models, (i) basic soil inputs only (ii) imaged pore metrics only, and (iii) their combination. These 3 XGboost-models are then compared to investigate the efficacy of these models. The traditional model had an R^2 of 0.07, the pure imaging model had one of 0.30 and the combined model got 0.483. This shows that imaged parameters heavily improve K_{sat} estimation, and can even be used alone, suggesting that they are a good basis for new PTFs.

2. Introduction

Saturated hydraulic conductivity, K_{sat}, is one of the most important soil metrics, being a critical metrics for hydrological models. However, estimating K_{sat} is still a difficult task after 40 years of research into the issue (Cosby et al., 1984; Van Looy et al., 2017; Zhang & Schaap, 2019). Despite the ever increasing strength of machine learning models we've not seen any big strides forwards in the last 20 or so years (Zhang & Schaap, 2019). It's suggested, by some of the leading voices in the 'pedo-transfer function'-sphere, that better parameterization of the pore space can be a critical part in improving the performance of pedo-transfer functions (PTFs) (Koestel et al., 2018; Van Looy et al., 2017; Zhang & Schaap, 2019), as it might be indirect nature of the parameters currently in use that is holding back progress. 3D-Xray tomography has been used on soil samples since 1982 (Petrovic et al.), and Koestel et al. showed in 2018 that it parameters based on images of ring samples have a good correlation with $K_{sat.}$ Additionally

taking a 3D-Xray image can be done in less than half an hour, and with very little human intervention, comparatively it often takes days of work to measure the common soil metrics used in current PTFs. With this in mind we attempt to find out if these imaged parameters benefit PTFs and if a X-ray based PTF could be made with the data available to us. We do this by creating a set of 3 types of PTFs, a “traditional” PTF, using texture fractions, bulk density and carbon content, a model only using a imaged parameters, and a model using imaged parameters, the traditional parameters and depth data. We then investigate the effects of adding the imaged parameters, what imaged parameters have the biggest impact, and how well a PTF based on images performs.

3. Method

The data used is from the SOILSPACE project. Over a few years (2015-2019) 178 samples were collected in aluminium rings, with the internal dimensions of Ø6.5 cm*6 cm. The samples were collected from places ranging from close to the southern most tip of Norway, all the way up to north of Narvik, additionally a larger set of samples were collected from the Skuterud site. These samples where then sent to SLU to be scanned with their GE Phoenix v|tome|x 240 X-ray scanner, before being ran through a series of traditional measurements. This gives us a internally consistent set of data where each variable has been measured on each sample. The X-ray scans was sent through SoilJ, which parameterized the data giving us our imaged parameters. The dataset was further filtered to remove any samples with issues making them non-compatible with our ML-models, leaving us with 160 samples. After filtering out any imaged variables that had a p-coefficient ∓ 0.75 against any of the traditional or other imaged variables, after this filtration we were left with:

Fig. 21 Map of Norway showing approximate position of sampeling sites

- Anisotropy: How *directional* the pore volumes are in the scale of 0-1. 0 is isotropic (without directionality) and 1 is anisotropic (with directionality) (Doube, 2020)
- Average pore diameter: The average diameter of the pores in the sample in mm. (Koestel, 2018)
- Critical pore diameter: The bottleneck in the largest pore in mm. This can be conceptualized as the smallest sphere that could pass through the sample. (Koestel, 2018)
- Fractal dimensionality: A measure of geometrical complexity. The higher the fractal dimensionality the 'rougher' the object is. Fractal dimensionality stems from Mandelbrot's seminal work "How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension" (1967), where he discusses the issue of the Coastline paradox ('Coastline Paradox', 2024). The paradox states that the exact length of a coastline can

only be approximated as it is functionally a fractal, and its length will thereby be determined by the length of the measuring stick. Taking coastlines as an example South Africa's coastline (very smooth) gets a fractal dimensionality like 1.02, Great Britan which has a very rough coast gets 1.25 (Mandelbrot, 1967), maybe the most extreme would be Norway with 1.52 (Feder, 2013, p. 8). Here it is a measure of the roughness of the pore space.

- Percolating porosity. The total volume of pores that is connected with both the top and the bottom, in mm^3. (Koestel, 2018)
- Pore metrics: We also have fractional data of pore size, with fraction of pores between 0.25-0.5 and 0.5-1.0.
- Γ-connectivity: The probability of a random pore voxel being connected to the rest of the pores given as a value between 0 and 1, where 1 would indicate all pore voxels being connected (Jarvis et al., 2017). This is also sometimes referred to as "connection probability" in the literature.

The models were created in R using the Tidymodels (Kuhn & Wickham, 2020) and DALEX (Biecek, 2018) frameworks. A series of XGboost models were created using the '*finetune*' package, from the Tidymodels package, to automatically adjust all parameters, except for learning rate, using the 'Anova race' method. A set of learning rates was then tested, and the best performing learning rate, based on R^2, was used for each of the three models. When creating the pure image model a series of models with differing constellations of the imaged parameters were made, with the best performing model being one which used only 'critical pore diameter' and 'percolating porosity'.

The final 3 models included these metrics:

- Traditional model includes: texture (USDA sand, silt and clay fractions), bulk density, and soil carbon content (%).
- Image enhanced model includes: texture, bulk density, soil carbon content (%), depth, average pore diameter, percolating porosity, critical pore diameter, anisotropy, fractal dimensionality, pore metrics, and Γ-connectivity.
- Pure imaging model: percolating porosity, and critical pore diameter.

We used the 'DALEX' model-agnostic framework when analysing our models. 'Explainer'-objects were made for each model and these were used to get model performance metrics, SHAP-values and other variable importance metrics. This is also the basis for our visuals.

4. Result

4.1. Performance metrics:

The image enhanced model is the highest performing model achieving an average R^2-value of 0.483, this is followed by the Pure Imaging model achieving an average R^2-value of 0.3, and the traditional model performed poorly, only achieving an average R^2-value of 0.07. For comparison the R2-values for the models on the dataset they were trained on is also showed. The value of 0.996 for the Image Enhanced model suggest overfitting to the training set, despite the model being tuned to reduce the difference between the different R^2-values. The differences (Traditional model = +0.34, Image Enhanced = +0.513, Pure Imaging = +0.364) suggest, together with the relatively good performance, that the Pure Imaging model is the least overfitted.

Model:	**Traditional model**	**Image enhanced**	**Pure imaging**
Hyperparameters			
Learning rate**	0.01	0.005	0.01
Number of predictive features	6 (4***)	15 (13***)	2
Model performance metrics*			
Testing set			
RMSE	2.736	2.064	2.407
MAE	2.118	1.545	1.761
MAD	1.756	1.227	1.469
R^2	0.070	0.483	0.300
Training set			
RMSE	1.068	0.204	1.828
MAE	0.792	0.140	1.337
MAD	1.336	0.102	1.035
R^2	0.410	0.996	0.664

* all performance metrics are the mean for 10 models with the same hyper parameters but models and test/training sets are seeded differently
** tuned to reduce difference in performance between training and testing sets
*** number of predictive features when the 3 texture variables are counted as one

Fig. 22 Model performance metrics for the traditional, image enhanced and pure imaging models.

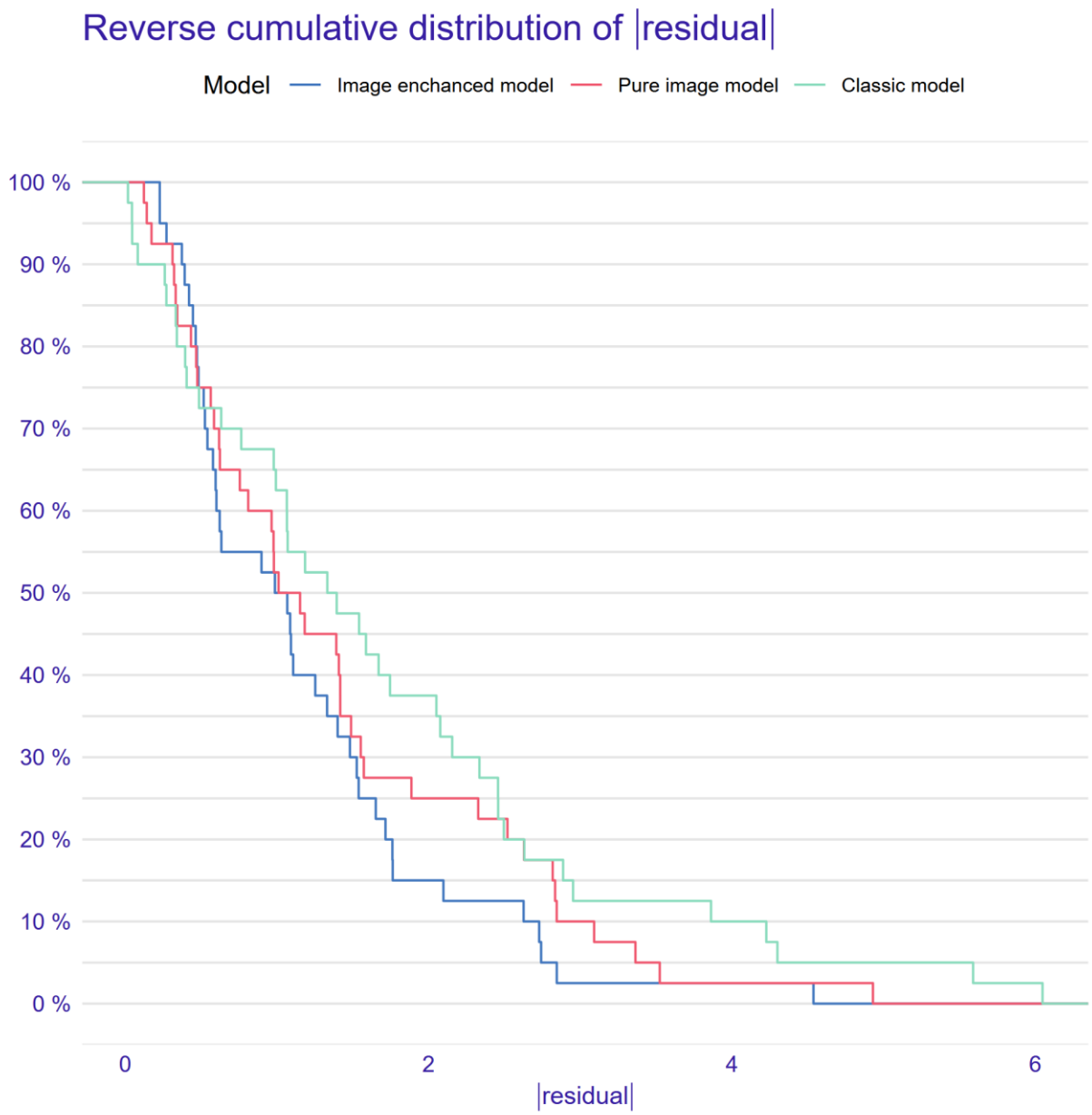


Fig. 23 Reverse cumulative distribution of residuals from the from the best performer from each of the 3 models.

In Fig. 23 we show the different distributions of residuals for the 3 models. The tail of the traditional model is especially large and long. The heads of both image models are pretty good with minor discrepancies, but they diverge especially in the last tertile, where the pure imaging model is less performant. The long tail at the end of both is probably caused by some kind of outlier data. It should be noted that these are the best performing models, and might be so because the testing data for each of them lacks diversity, this is especially true for the traditional model as it's performance is a lot higher than the average performance for this model type.

4.2. *Average SHAP-values for the image enhanced model*

The average SHAP-value for all the parameters in the image enhanced model was calculated for a set of 20 identical models trained on differently seeded training and test sets and models. Percolating porosity and critical pore diameter are the only parameters significantly different from the rest of the variables (Fig. 24, right upset plot), with the adjusted P-value between percolating porosity and critical pore diameter being ~0.002. Percolating porosity and critical pore diameter achieve average SHAP-values of 0.905 and 0.807 each, which is 4 times higher for all other variables except for bulk density.

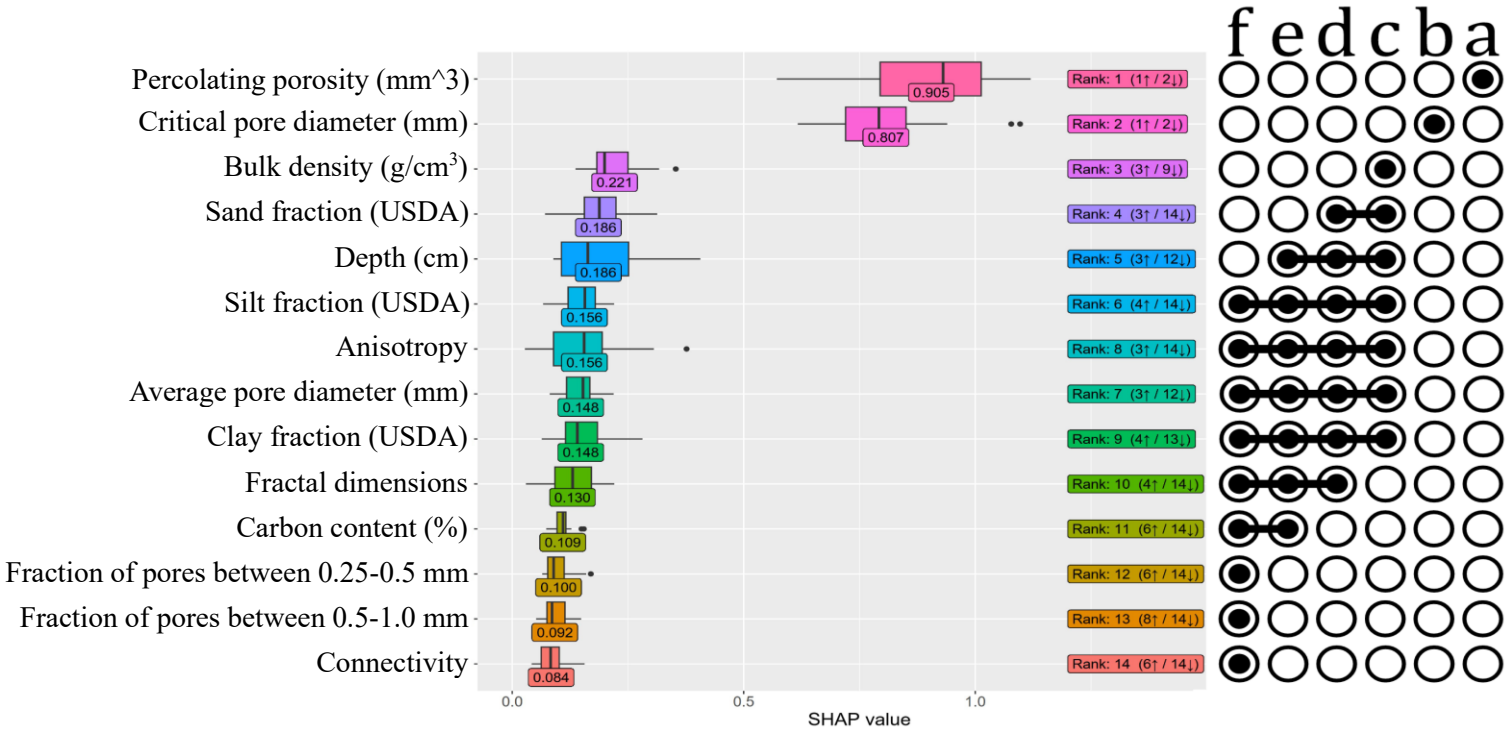


Fig. 24 Plot of SHAP score distribution – showing variable importance, model stability and average, max, and min rank for each variable. Furthest to the right is an upset table showing significance through Tukey letters. N = 20. All SHAP values can be found in Appx. 4

4.3. *The Pure Imaging Model*

The pure imaging model achieves an average R^2 0.30 and a max R^2 of 0.47 against the testing set. The model only has 2 predictive parameters allowing us to plot the results (and in that the limits model) as a raster. Fig. 25 is such a raster for the best performing pure imaging model.

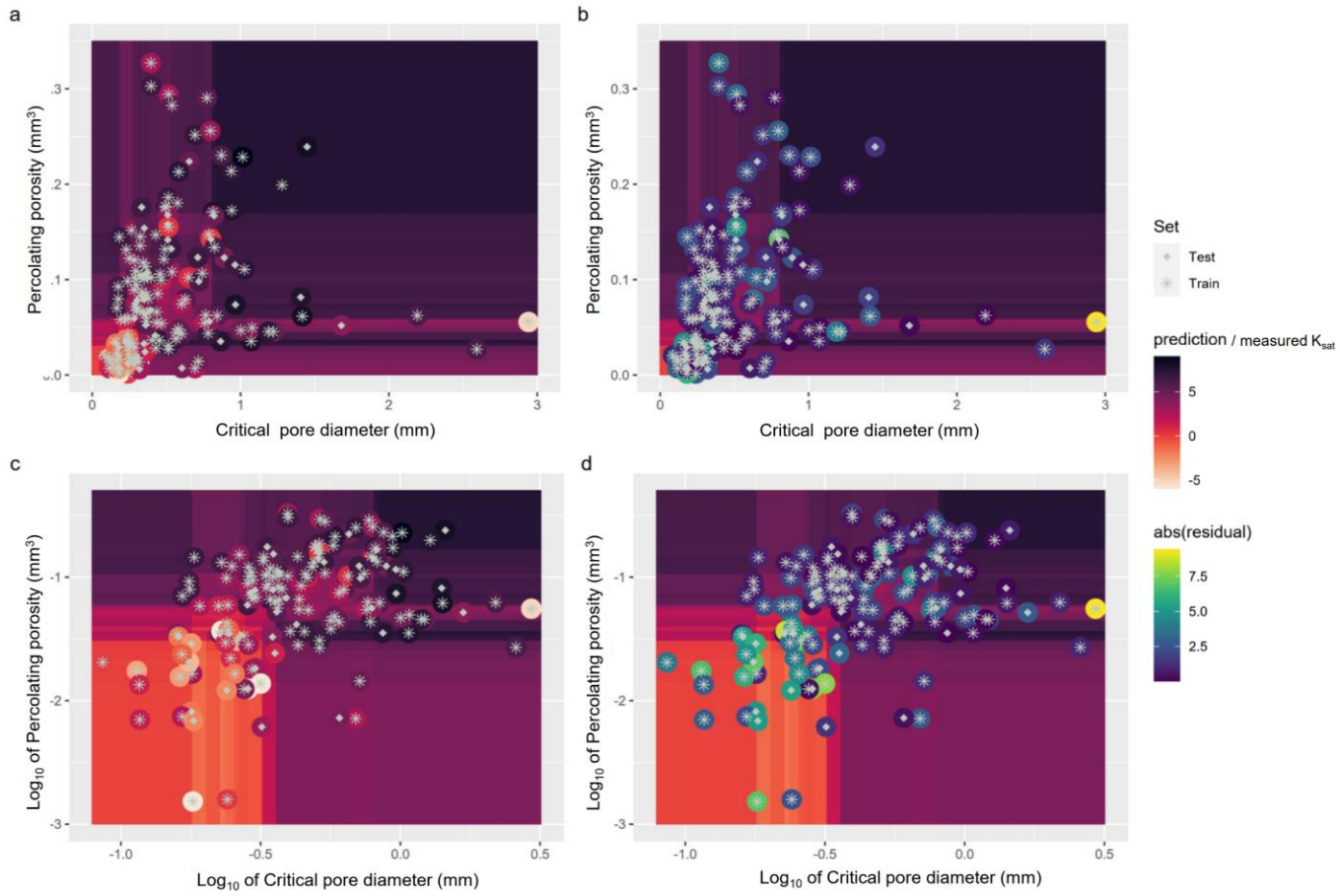


*Fig. 25 (A larger version of this image can be found in Appx. 6.) A 2D representation of the 'Pure imaging model' shown with the measured K_{sat} (a & c) and the residual for each prediction (b & c) plotted as points. Additionally, each point has a diamond or a 8 pointed star showing what set in the dataset (testing or training) each point is part of. The raster plot in the background of each image is the model predictions for an imputed dataset with evenly distributed values (503*503 values) around the limits of the original dataset, in turn this shows close to every possible prediction for the model in that space. The lower two graphs (c & d) are log transformed to better represent the large number of values with low values for the two predictive parameters. R^2 for the model is 0.47 against the testing set and 0.72 against the training set.*

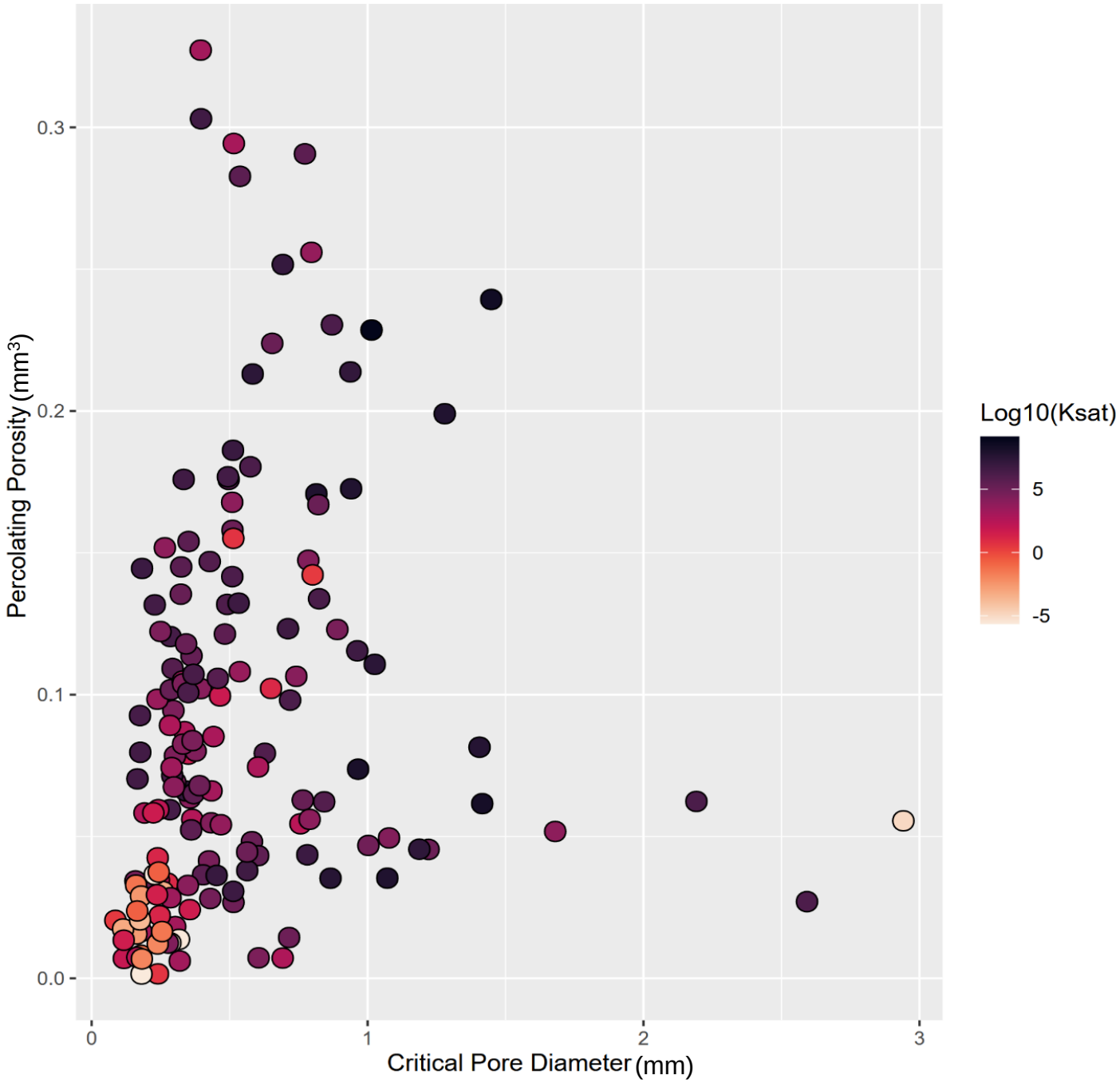


Fig. 26 $\log_{10}(K_{sat})$ mapped as a colour onto a scatter plot with percolating porosity and critical pore diameter as the axes.

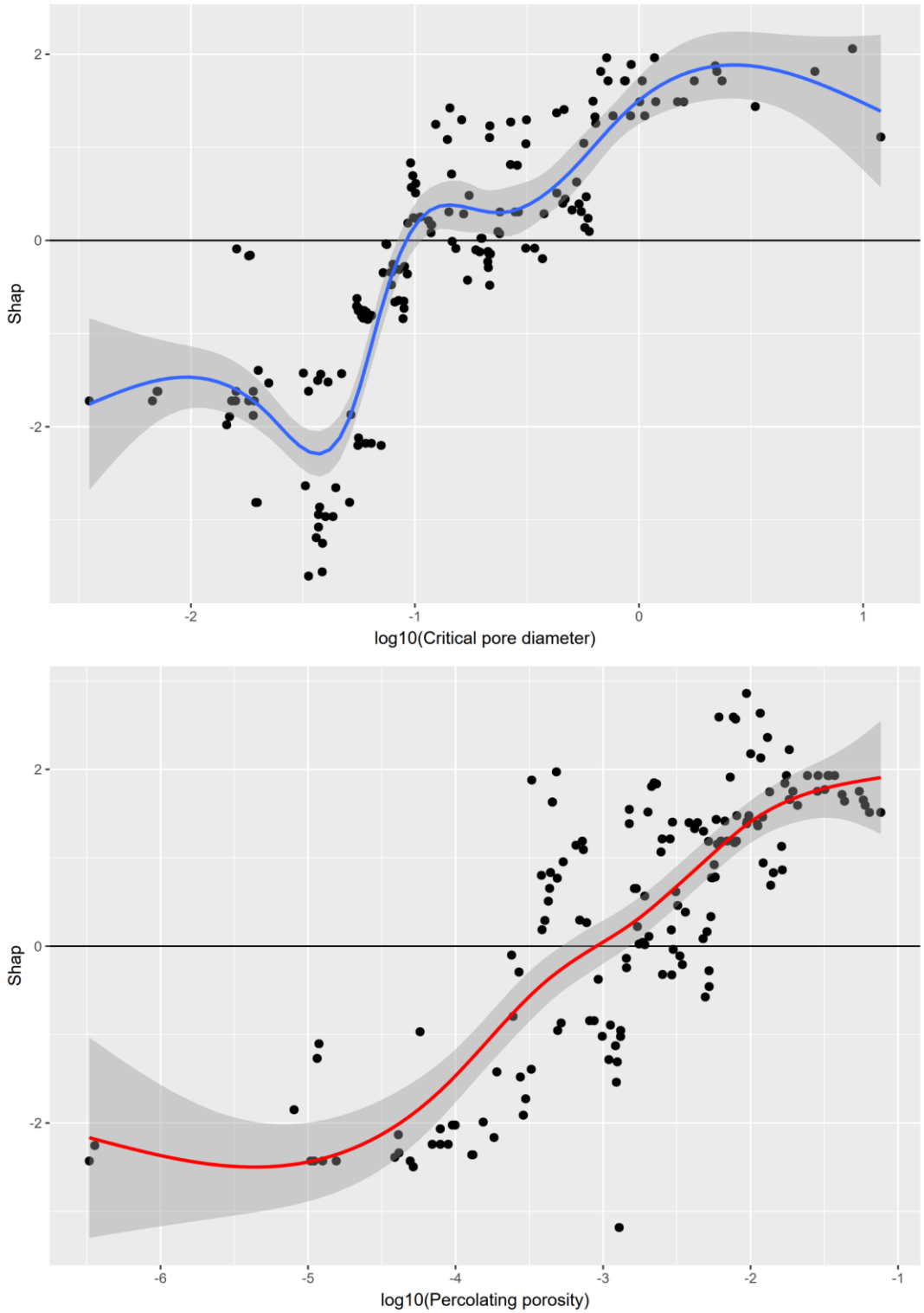


Fig. 27 Showing the SHAP values across the whole dataset (N = 160) for both parameters in the pure imaging model, log10 transformed. A gam model has been fitted to the points to show the trends in the data. A line at 0 has also been added to delineate between values that reduce or increase the final estimate.

5. Discussion

The differences in performance show that imaged parameters can be a great addition to, or the basis for new PTFs. Especially when we take into account that the SOILSPACE dataset (here n = 160) is much smaller than what is often used to train pedo-transfer models like ROSETTA (n = 1306) or the EU-HYDI (up to n = 3206) (Zhang & Schaap, 2019). Hopefully, if X-ray tomography of soil samples becomes more common, we'll see an increase in the efficacy of X-ray based PTFs. Based on this an argument for using metrics more closely related to the soil pores when building PTFs can easily be made. The average SHAP-values for the Image enhanced model (Fig. 12, Fig. 13, and Fig. 24) also support this, with the really high values for critical pore diameter and percolating porosity. Fig. 13 also shows how these parameters are almost always the most impactful.

There is also a notable pattern in the data here. The divergence of critical pore diameter and percolating porosity, which is likely to be part of the reason that the pure imaging model, which only uses these two parameters is as performant as it is. When these are both low we see very a very low K_{sat} and as either or both increase so does the K_{sat}, this is true both for the model and when plotting K_{sat} against the parameters (Fig. 26). It follows some intuitive sense that both metrics have an impact on soils conductivity. A sample with a large hole straight through the middle will let water fall through, as the size increases the capillary action also decreases allowing the water too more easily flow through the hole, increasing K_{sat}. The percolating volume also has this intuitive sense, the larger the volume becomes the more space is there for the water to percolate through. This relationship shows merits as a predictor for K_{sat}, both as a the main predictors of a larger set of predictors (Fig. 24), when they're the only predictors (Fig. 25). This suggests an exciting and real relationship between K_{sat}, and percolating porosity and critical pore diameter. Additionally looking at Fig. 27 we can see that there are trends in the data that when that at times seem almost linear.

It's important to note that our images have a resolution of 0.04 mm, any pores smaller than this will go unrepresented in the data. This might not be too much of a problem as with decreasing pore size, the small volume and increasing capillary action decreases the amount of water that can travel through them.

As with any large project that has a decent number of samples we can expect some errors to creep in. The samples all had to travel longer distances, like the ones between NMBU and SLU, as well as the transportation of the samples from the sampling locations to the labs. Depending on their treatment during said transportation there might be some unnatural outliers in the dataset, especially since the X-ray tomography was done in Uppsala, Sweden, and the other measurements were done in Ås, Norway.

We also concede that there is currently no uncorrelated data that can be used as a testing set for our models, however it's the hope of the authors that this text inspires others to create similar datasets so that this superior level of verification can be achieved.

6. Conclusion

We believe that these data show that parameterized CT-Xray tomography can be a powerful tool in estimating K_{sat} as soon as now. The models show the efficacy of the percolating porosity and critical pore diameter parameters as predictors for K_{sat}. Further research and especially larger datasets will hopefully improve these methods. X-ray tomography seems to have a lot of promising value in K_{sat} estimation, both in terms of efficacy and speed, as doing X-ray tomography on a sample is something that can be done in the background during a 30-minute lunch break, rather than being a arduous process taking several days like the measurements of common soil parameters or K_{sat} itself. The improvement, in both speed, accuracy, and non-invasiveness, of this kind of K_{sat} estimation of samples would hopefully also make the larger tasks of generalizing these metrics for larger spatial scales easier; in turn improving earth system models.

References:

Biecek, P. (2018). DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research*, *19*(84), 1–5.

Brannon, K. (2024, January 23). *AI sentencing cut jail time for low-risk offenders, but study finds racial bias persisted*. Tulane University News. https://news.tulane.edu/pr/ai-sentencing-cut-jail-time-low-risk-offenders-study-finds-racial-bias-persisted

Burzykowski, T., & Biecek, P. (2020). *Explanatory Model Analysis*.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Coastline paradox. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Coastline_paradox&oldid=1221118762#cite_note-1

Coefficient of determination. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Coefficient_of_determination&oldid=1210579823

Cosby, B. J., Hornberger, G. M., Clapp, R. B., & Ginn, T. R. (1984). A Statistical Exploration of the Relationships of Soil Moisture Characteristics to the Physical Properties of Soils. *Water Resources Research*, *20*(6), 682–690. Scopus. https://doi.org/10.1029/WR020i006p00682

Doube, M. (2020, May 29). *Anisotropy*. BoneJ; doube.net ltd. https://bonej.org/anisotropy

FAO. (2022, July 27). *FAO warns 90 per cent of Earth's topsoil at risk by 2050*. UN News. https://news.un.org/en/story/2022/07/1123462

FAOSTAT. (2023). *Food Balances* [CSV]. https://www.fao.org/faostat/en/#data/FBS

Feder, J. (2013). *Fractals*. Springer Science & Business Media.

Gilgoldm. (2020). *Survival of passengers on the Titanic* [Digital image]. Created for the wikimedia. https://commons.wikimedia.org/wiki/File:Decision_Tree.jpg

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81–102. https://doi.org/10.1016/0095-0696(78)90006-2

Helliwell, J. R., Sturrock, C. J., Grayling, K. M., Tracy, S. R., Flavel, R. J., Young, I. M., Whalley, W. R., & Mooney, S. J. (2013). Applications of X-ray computed tomography for examining biophysical interactions and structural development in soil systems: A review. *European Journal of Soil Science*, *64*(3), 279–297. https://doi.org/10.1111/ejss.12028

Henry Philibert Gaspard, D. (1856). *Les fontaines publiques de la ville de Dijon: Exposition et application des principes à suivre et des formules à employer dans les questions de distribution d'eau : Ouvrage terminé par un appendice relatif aux fournitures d'eau de plusieurs villes, au filtrage des eaux et à la fabrication des tuyaux de fonte, de plomb, de tôle et de bitume*. Victor Dalmont, éditeur.

Hsu, J. (2024, March 7). *AI chatbots use racist stereotypes even after anti-racism training*. New Scientist. https://www.newscientist.com/article/2421067-ai-chatbots-use-racist-stereotypes-even-after-anti-racism-training/

Huckins, G. (2023, October 16). *Minds of machines: The great AI consciousness conundrum*. MIT Technology Review. https://www.technologyreview.com/2023/10/16/1081149/ai-consciousness-conundrum/

Jarvis, N., Larsbo, M., & Koestel, J. (2017). Connectivity and percolation of structural pore networks in a cultivated silt loam soil quantified by X-ray tomography. *Geoderma*, *287*, 71–79. https://doi.org/10.1016/j.geoderma.2016.06.026

Khan, S., Naushad, Mu., Lima, E. C., Zhang, S., Shaheen, S. M., & Rinklebe, J. (2021). Global soil pollution by toxic elements: Current status and future perspectives on the risk assessment and remediation strategies – A review. *Journal of Hazardous Materials*, *417*, 126039. https://doi.org/10.1016/j.jhazmat.2021.126039

Klute, A. (Ed.). (1986). *Methods of soil analysis. 1: Physical and mineralogical methods / Arnold Klute, ed* (2. edition). American Society of Agronomy.

Koestel, J. (2018). SoilJ: An ImageJ Plugin for the Semiautomatic Processing of Three-Dimensional X-ray Images of Soils. *Vadose Zone Journal*, *17*(1), 170062. https://doi.org/10.2136/vzj2017.03.0062

Koestel, J., Dathe, A., Skaggs, T. H., Klakegg, O., Ahmad, M. A., Babko, M., Giménez, D., Farkas, C., Nemes, A., & Jarvis, N. (2018). Estimating the Permeability of Naturally Structured Soil From Percolation Theory and Pore Space Characteristics Imaged by X-Ray. *Water Resources Research*, *54*(11), 9255–9263. https://doi.org/10.1029/2018WR023609

Kuhn, M., & Wickham, H. (2020). *Tidymodels* (1.2.0) [Computer software]. https://cloud.r-project.org/web/packages/tidymodels/index.html

Leffer, L. (2023, November 21). *When It Comes to AI Models, Bigger Isn't Always Better*. Scientific American. https://www.scientificamerican.com/article/when-it-comes-to-ai-models-bigger-isnt-always-better/

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to Bivariate and Multivariate Analysis*. Scott, Foresman.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Macedo, A., Crestana, S., & Vaz, C. M. P. (1998). X-ray microtomography to investigate thin layers of soil clod. *Soil and Tillage Research*, *49*(3), 249–253. https://doi.org/10.1016/S0167-1987(98)00180-9

Mandelbrot, B. (1967). How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension. *Science*, *156*(3775), 636–638. JSTOR.

McLauchlan, K. (2006). The Nature and Longevity of Agricultural Impacts on Soil Carbon and Nutrients: A Review. *Ecosystems*, *9*(8), 1364–1382. https://doi.org/10.1007/s10021-005-0135-1

Melsen, L. (2024, March 11). *The sociology of modelling: How we shape a perception together*. https://doi.org/10.5194/egusphere-egu24-21063

MNIST database. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=MNIST_database&oldid=1220699331

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (Second edition). Christoph Molnar.

Mooney, S. J., Pridmore, T. P., Helliwell, J., & Bennett, M. J. (2012). Developing X-ray Computed Tomography to non-invasively image 3-D root systems architecture in soil. *Plant and Soil*, *352*(1–2), 1–22. https://doi.org/10.1007/s11104-011-1039-9

Nearing, M. A., Xie, Y., Liu, B., & Ye, Y. (2017). Natural and anthropogenic rates of soil erosion. *International Soil and Water Conservation Research*, *5*(2), 77–84. https://doi.org/10.1016/j.iswcr.2017.04.001

Nimmo, J. R., & Landa, E. R. (2005). The Soil Physics Contributions of Edgar Buckingham. *Soil Science Society of America Journal*, *69*(2), 328–342. https://doi.org/10.2136/sssaj2005.0328

Petrovic, A. M., Siebert, J. E., & Rieke, P. E. (1982). Soil Bulk Density Analysis in Three Dimensions by Computed Tomographic Scanning. *Soil Science Society of America Journal*, *46*(3), 445–450. https://doi.org/10.2136/sssaj1982.03615995004600030001x

Piet Mondrian. (1930). *Composition with Red, Blue, and Yellow* [Oil on canvas]. Kunsthaus Zürich. https://commons.wikimedia.org/wiki/File:Piet_Mondriaan,_1930_-_Mondrian_Composition_II_in_Red,_Blue,_and_Yellow.jpg

Saloranta, T. M., Kämäri, J., Rekolainen, S., & Malve, O. (2003). Benchmark Criteria: A Tool for Selecting Appropriate Models in the Field of Water Management. *Environmental Management*, *32*(3), 322–333. https://doi.org/10.1007/s00267-003-0069-3

Shapley Lloyd Stowell, L. S. (1953). Paper 17. A VALUE FOR n-PERSON GAMES. In *Contributions to the Theory of Games: Vol. Volume II* (pp. 307–317). Princeton University Press; JSTOR. http://www.jstor.org/stable/j.ctt1b9x1zv

Shap/shap. (2024). [Jupyter Notebook]. shap. https://github.com/shap/shap (Original work published 2016)

Štrumbelj, E., & Kononenko, I. (2010). An Efficient Explanation of Individual Classifications using Game Theory. *The Journal of Machine Learning Research*, *11*, 1–18.

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x

Sundararajan, M., & Najmi, A. (2020). The Many Shapley Values for Model Explanation. *Proceedings of the 37th International Conference on Machine Learning*, 9269–9278. https://proceedings.mlr.press/v119/sundararajan20b.html

Thaler, E. A., Kwang, J. S., Quirk, B. J., Quarrier, C. L., & Larsen, I. J. (2022). Rates of Historical Anthropogenic Soil Erosion in the Midwestern United States. *Earth's Future*, *10*(3), e2021EF002396. https://doi.org/10.1029/2021EF002396

Tiku, N. (2022, June 11). *The Google engineer who thinks the company's AI has come to life*. Washington Post. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

Tindall, J. A., Kunkel, J. R., & Anderson, D. E. (1999). *Unsaturated Zone Hydrology for Scientists and Engineers*. Prentice Hall.

Tomography. (2023). In *Wiktionary, the free dictionary*. https://en.wiktionary.org/w/index.php?title=tomography&oldid=76437143

van Gogh, V. (1887). *Self-Portrait* [Oil on artist's board, mounted on cradled panel]. Art Institute of Chicago. https://commons.wikimedia.org/wiki/File:Vincent_van_Gogh_-_Self-Portrait_-_Google_Art_Project_(454045).jpg

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y. A., Padarian, J., Schaap, M. G., Tóth, B., Verhoef, A., Vanderborght, J., Ploeg, M. J., Weihermüller, L., Zacharias, S., Zhang, Y., & Vereecken, H. (2017). Pedotransfer Functions in Earth System Science: Challenges and Perspectives. *Reviews of Geophysics*, *55*(4), 1199–1256. https://doi.org/10.1002/2017RG000581

Weil, R. R., & Brady, N. C. (2017). *The nature and properties of soils* (Fifteenth edition, global edition). Pearson.

Wetzel, R. G. (2001). *Limnology: Lake and river ecosystems* (3rd ed). Academic Press.

Zhang, Y., & Schaap, M. G. (2019). Estimation of saturated hydraulic conductivity with pedotransfer functions: A review. *Journal of Hydrology*, *575*, 1011–1030. https://doi.org/10.1016/j.jhydrol.2019.05.058

Appendix

model num	RMSE test	MAE test	MAD test	R^2 test	RMSE train	MAE train	MAD train	R^2 train
1	2.666	2.152	2.141	0.030	0.348	0.259	2.141	0.030
2	3.293	2.481	2.125	-0.342	0.336	0.247	2.125	-0.342
3	2.550	1.912	1.231	0.290	0.371	0.259	1.231	0.290
4	2.254	1.707	1.364	0.353	0.298	0.219	1.364	0.353
5	2.774	2.297	1.964	0.063	1.429	0.999	1.964	0.063
6	3.029	2.167	1.631	0.262	1.321	0.993	0.779	0.800
7	2.851	2.139	1.618	-0.340	1.461	1.125	0.861	0.803
8	2.751	2.245	1.993	0.152	2.144	1.591	1.244	0.536
9	2.479	1.920	1.782	-0.061	1.506	1.154	0.884	0.792
10	2.714	2.160	1.707	0.290	1.461	1.073	0.769	0.774
Mean	2.736	2.118	1.756	0.070	1.068	0.792	1.336	0.410

Appx. 1 Model performance metrics for the "traditional" model which only used the traditional parameters: texture measures, bulk density, & soil carbon content (%). "Test" indicates that the metric is against the testing against and vice-versa for testing.

model num	RMSE test	MAE test	MAD test	R^2 test	RMSE train	MAE train	MAD train	R^2 train
1	1.737	1.423	1.260	0.588	0.249	0.157	0.106	0.994
2	1.989	1.650	1.472	0.510	0.159	0.115	0.075	0.998
3	2.509	1.694	1.013	0.312	0.238	0.160	0.111	0.994
4	2.113	1.542	1.300	0.432	0.238	0.160	0.115	0.994
5	2.306	1.689	1.201	0.352	0.182	0.123	0.090	0.997
6	2.544	1.813	1.362	0.479	0.143	0.107	0.083	0.998
7	1.952	1.514	1.252	0.372	0.158	0.119	0.092	0.998
8	1.955	1.315	0.888	0.572	0.248	0.161	0.120	0.994
9	1.499	1.192	1.029	0.612	0.282	0.193	0.150	0.993
10	2.035	1.623	1.492	0.601	0.142	0.108	0.081	0.998
Mean	2.064	1.545	1.227	0.483	0.204	0.140	0.102	0.996

Appx. 2 Model performance metrics for the "Image enhanced" model which used traditional parameters, and all other parameters with a Pearson coefficient less than 0.7, culminating in: texture, bulk density, depth, average pore diameter, percolating porosity, critical pore diameter, anisotropy, fractal dimensionality, pore metrics, soil carbon content (%), & Γ-connectivity. "Test" indicates that the metric is against the testing against and vice-versa for testing.

model num	RMSE test	MAE test	MAD test	R^2 test	RMSE train	MAE train	MAD train	R^2 train
1	2.370	1.875	1.781	0.234	1.979	1.393	1.125	0.623
2	2.333	1.906	1.754	0.326	2.070	1.506	1.176	0.578
3	2.692	1.841	1.601	0.208	1.754	1.305	0.974	0.688
4	2.430	1.680	1.511	0.248	1.904	1.427	1.187	0.647
5	2.912	2.028	1.672	-0.033	1.484	1.078	0.805	0.783
6	2.946	1.954	1.362	0.301	1.425	1.055	0.773	0.767
7	1.799	1.410	1.083	0.467	1.742	1.227	0.794	0.720
8	2.342	1.567	1.294	0.386	1.996	1.583	1.333	0.598
9	1.815	1.488	1.099	0.431	1.932	1.355	1.059	0.657
10	2.432	1.861	1.538	0.430	1.999	1.440	1.122	0.577
Mean	2.407	1.761	1.469	0.300	1.828	1.337	1.035	0.664

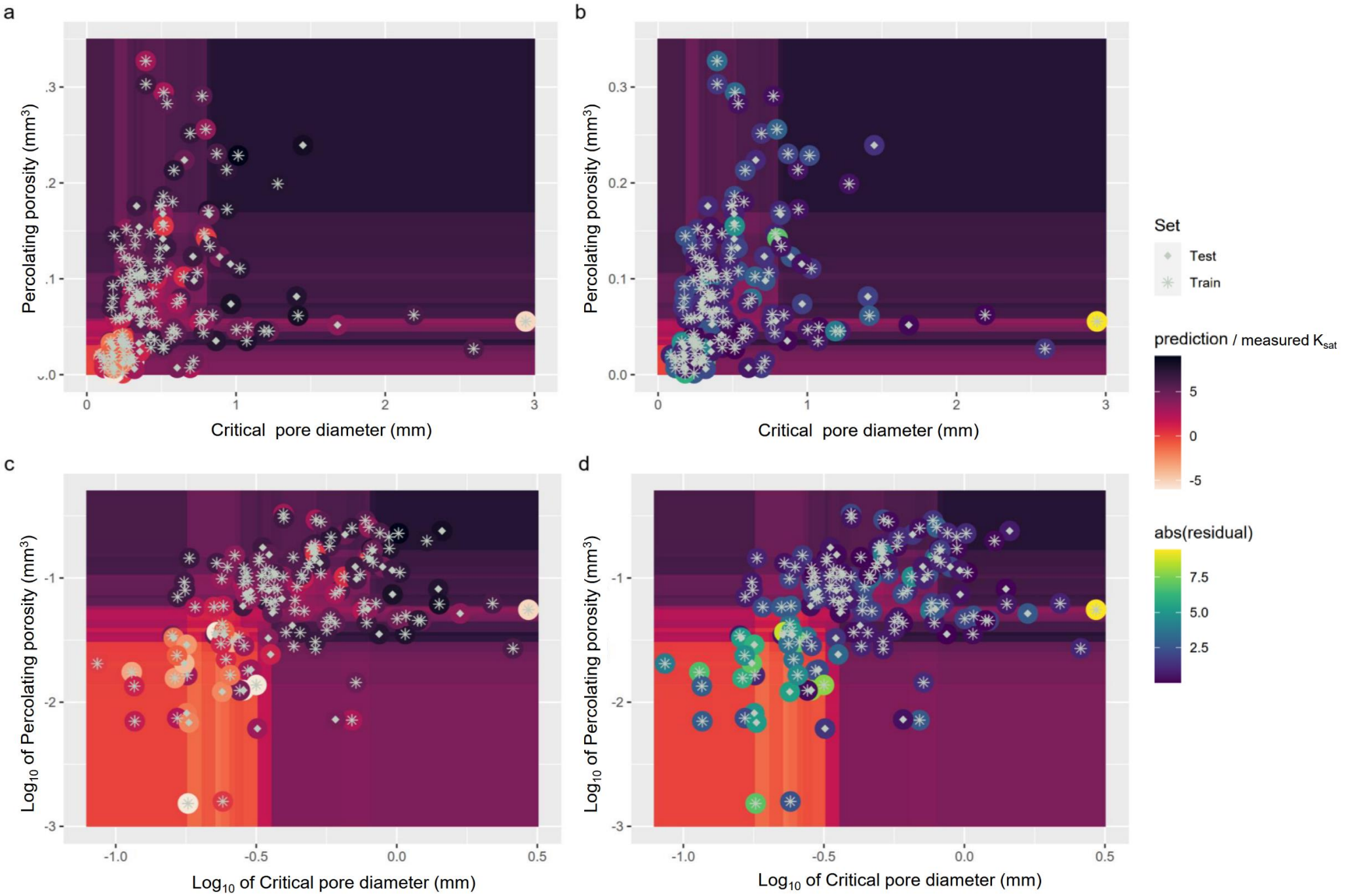
Appx. 3 Model performance metrics for the "Pure Image" model which only used the imaged parameters: Critical Pore Diameter and Percolating Porosity. "Test" indicates that the metric is against the testing against and vice-versa for testing

feature \ model number	1	2	3	4	5	6	7	8	9	10
Percolating Porosity	0.962 Rank:1	0.931 Rank:1	0.781 Rank:1	1.007 Rank:1	0.713 Rank:2	1.031 Rank:1	0.835 Rank:1	1.066 Rank:1	0.888 Rank:1	0.952 Rank:1
Critical Pore Diameter	0.817 Rank:2	0.786 Rank:2	0.676 Rank:2	0.818 Rank:2	1.078 Rank:1	0.761 Rank:2	0.722 Rank:2	0.679 Rank:2	0.808 Rank:2	0.852 Rank:2
Anisotropy	0.216 Rank:3	0.143 Rank:7	0.377 Rank:3	0.210 Rank:6	0.189 Rank:5	0.027 Rank:14	0.306 Rank:3	0.161 Rank:7	0.268 Rank:3	0.176 Rank:5
Bulk Density	0.201 Rank:4	0.240 Rank:3	0.179 Rank:6	0.251 Rank:4	0.314 Rank:3	0.137 Rank:5	0.187 Rank:6	0.193 Rank:4	0.175 Rank:6	0.353 Rank:3
Average Pore Diameter	0.185 Rank:5	0.081 Rank:12	0.149 Rank:9	0.151 Rank:9	0.113 Rank:10	0.122 Rank:7	0.105 Rank:10	0.176 Rank:5	0.166 Rank:8	0.164 Rank:7
Depth (top)	0.177 Rank:6	0.118 Rank:10	0.192 Rank:5	0.386 Rank:3	0.100 Rank:12	0.135 Rank:6	0.091 Rank:12	0.250 Rank:3	0.176 Rank:5	0.260 Rank:4
Fractal dimensionality	0.170 Rank:7	0.119 Rank:9	0.125 Rank:11	0.030 Rank:14	0.163 Rank:6	0.078 Rank:12	0.173 Rank:7	0.160 Rank:8	0.195 Rank:4	0.109 Rank:11
Silt fraction (USDA)	0.156 Rank:8	0.196 Rank:4	0.156 Rank:8	0.220 Rank:5	0.208 Rank:4	0.149 Rank:4	0.117 Rank:8	0.140 Rank:9	0.173 Rank:7	0.172 Rank:6
Clay fraction (USDA)	0.121 Rank:9	0.104 Rank:11	0.203 Rank:4	0.192 Rank:7	0.146 Rank:8	0.097 Rank:9	0.212 Rank:5	0.127 Rank:11	0.147 Rank:9	0.152 Rank:8
Fraction of pores between 0.5-1 mm	0.120 Rank:10	0.119 Rank:8	0.134 Rank:10	0.149 Rank:10	0.094 Rank:13	0.080 Rank:11	0.052 Rank:13	0.104 Rank:12	0.074 Rank:14	0.112 Rank:10
Fraction of pores between 0.25-0.5 mm	0.105 Rank:11	0.065 Rank:14	0.069 Rank:13	0.081 Rank:13	0.105 Rank:11	0.086 Rank:10	0.093 Rank:11	0.133 Rank:10	0.135 Rank:10	0.097 Rank:13
Sand Fraction (USDA)	0.100 Rank:12	0.192 Rank:5	0.176 Rank:7	0.186 Rank:8	0.151 Rank:7	0.198 Rank:3	0.220 Rank:4	0.169 Rank:6	0.112 Rank:11	0.131 Rank:9
Γ-Connectivity	0.080 Rank:13	0.067 Rank:13	0.059 Rank:14	0.097 Rank:12	0.042 Rank:14	0.120 Rank:8	0.047 Rank:14	0.064 Rank:14	0.103 Rank:12	0.094 Rank:14
Soil Carbon Content (%)	0.078 Rank:14	0.155 Rank:6	0.106 Rank:12	0.108 Rank:11	0.128 Rank:9	0.073 Rank:13	0.112 Rank:9	0.090 Rank:13	0.077 Rank:13	0.099 Rank:12
feature \ model number	11	12	13	14	15	16	17	18	19	20
Percolating Porosity	0.736 Rank:2	0.930 Rank:1	1.120 Rank:1	0.742 Rank:2	1.071 Rank:1	0.799 Rank:1	0.572 Rank:2	1.087 Rank:1	0.961 Rank:1	0.912 Rank:2
Critical Pore Diameter	0.850 Rank:1	0.739 Rank:2	0.714 Rank:2	0.910 Rank:1	0.787 Rank:2	0.684 Rank:2	1.097 Rank:1	0.617 Rank:2	0.797 Rank:2	0.940 Rank:1
Anisotropy	0.116 Rank:10	0.156 Rank:6	0.057 Rank:13	0.154 Rank:8	0.066 Rank:12	0.057 Rank:14	0.160 Rank:6	0.065 Rank:14	0.121 Rank:7	0.097 Rank:12
Bulk Density	0.215 Rank:5	0.249 Rank:4	0.189 Rank:4	0.307 Rank:3	0.184 Rank:9	0.198 Rank:4	0.227 Rank:4	0.317 Rank:4	0.150 Rank:6	0.144 Rank:4
Average Pore Diameter	0.117 Rank:9	0.117 Rank:10	0.154 Rank:5	0.202 Rank:5	0.219 Rank:4	0.117 Rank:10	0.159 Rank:7	0.136 Rank:9	0.173 Rank:4	0.162 Rank:3
Depth (top)	0.257 Rank:3	0.107 Rank:12	0.123 Rank:7	0.150 Rank:9	0.245 Rank:3	0.264 Rank:3	0.092 Rank:11	0.406 Rank:3	0.089 Rank:10	0.104 Rank:9
Fractal dimensionality	0.113 Rank:11	0.221 Rank:5	0.097 Rank:11	0.076 Rank:13	0.195 Rank:8	0.136 Rank:8	0.047 Rank:14	0.173 Rank:8	0.075 Rank:11	0.138 Rank:5
Silt fraction (USDA)	0.174 Rank:7	0.121 Rank:9	0.113 Rank:9	0.067 Rank:14	0.209 Rank:5	0.097 Rank:11	0.165 Rank:5	0.200 Rank:6	0.157 Rank:5	0.121 Rank:6
Clay fraction (USDA)	0.180 Rank:6	0.135 Rank:8	0.117 Rank:8	0.281 Rank:4	0.197 Rank:7	0.128 Rank:9	0.068 Rank:13	0.182 Rank:7	0.064 Rank:13	0.108 Rank:8
Fraction of pores between 0.5-1 mm	0.120 Rank:8	0.075 Rank:14	0.052 Rank:14	0.080 Rank:12	0.051 Rank:13	0.080 Rank:12	0.100 Rank:10	0.076 Rank:12	0.092 Rank:9	0.075 Rank:13
Fraction of pores between 0.25-0.5 mm	0.081 Rank:13	0.139 Rank:7	0.076 Rank:12	0.170 Rank:6	0.160 Rank:10	0.072 Rank:13	0.077 Rank:12	0.085 Rank:11	0.073 Rank:12	0.103 Rank:10
Sand Fraction (USDA)	0.238 Rank:4	0.252 Rank:3	0.190 Rank:3	0.157 Rank:7	0.202 Rank:6	0.176 Rank:5	0.252 Rank:3	0.313 Rank:5	0.242 Rank:3	0.071 Rank:14
Γ-Connectivity	0.068 Rank:14	0.090 Rank:13	0.148 Rank:6	0.086 Rank:11	0.047 Rank:14	0.156 Rank:6	0.101 Rank:9	0.065 Rank:13	0.050 Rank:14	0.102 Rank:11
Soil Carbon Content (%)	0.089 Rank:12	0.115 Rank:11	0.105 Rank:10	0.116 Rank:10	0.126 Rank:11	0.150 Rank:7	0.116 Rank:8	0.115 Rank:10	0.106 Rank:8	0.116 Rank:7

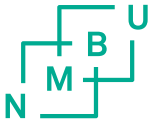
Appx. 4 SHAP-values and rank within model.

Variables	Depth	USDA Sand	USDA Silt	USDA Clay	BD	Soil Carbon Content (%)	Ksat	Volume	Average Pore Diameter	Fractal Dimensionality	Anisotropy	T-Connectivity	Critical Pore Diameter	Percolating porosity	Fraction of pores between 0.25-0.5 mm
Fraction of pores between 0.5-1 mm	-0.189	-0.344	0.234	0.383	-0.160	0.258	0.050	0.073	0.266	-0.414	0.052	0.346	0.149	0.237	-0.283
Fraction of pores between 0.25-0.5 mm	0.074	0.003	0.114	-0.235	-0.077	0.069	-0.305	0.071	-0.668	0.203	0.197	-0.235	-0.367	-0.172	
Percolating porosity	-0.414	0.149	-0.138	-0.093	-0.446	0.315	0.445	0.075	0.205	0.449	-0.309	0.731	0.169		
Critical Pore Diameter	-0.029	-0.023	-0.037	0.130	-0.026	0.024	0.281	-0.028	0.443	-0.155	-0.113	0.227			
Connectivity	-0.543	0.097	-0.140	0.041	-0.400	0.449	0.440	0.123	0.317	0.231	-0.192				
Anisotropy	0.347	-0.345	0.415	0.023	0.216	-0.284	-0.409	0.092	-0.013	-0.356					
Fractal dimensionality	-0.373	0.679	-0.541	-0.599	-0.295	0.133	0.192	-0.157	-0.388						
Average Pore Diameter	-0.016	-0.278	0.180	0.328	0.093	-0.024	0.107	0.071							
Volume	0.123	-0.238	0.249	0.092	0.006	0.077	0.025								
Ksat	-0.238	0.256	-0.295	-0.044	-0.259	0.171									
Carbon	-0.425	0.007	-0.124	0.232	-0.624										
BD	0.400	0.070	-0.080	-0.015											
USDA Clay fraction	0.096	-0.631	0.282												
USDA Silt fraction	0.178	-0.922													
USDA Sand fraction	-0.183														

Appx. 5 Pearson coefficients for the SOILSPACE database variables used.



Appx. 6 Larger format version of Fig. 25



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway