**VILNIUS UNIVERSITY**
**FACULTY OF CHEMISTRY AND GEOSCIENCES**
**INSTITUTE OF GEOSCIENCES**
**DEPARTMENT OF HYDROGEOLOGY AND ENGINEERING GEOLOGY**

**Eveliina Kukka-Maaria Vanhala**
Geology
Master's Thesis

# THEORETICAL HYDRAULIC CONDUCTIVITY DETERMINATION OF LITHUANIAN SOIL SAMPLES

Scientific advisor
Assistant dr. Vytautas Samalavičius

Vilnius 2024

**VILNIAUS UNIVERSITETAS**
**CHEMIJOS IR GEOMOKSLŲ FAKULTETAS**
**GEOMOKSLŲ INSTITUTAS**
**HIDROGEOLOGIJOS IR INŽINERINĖS GEOLOGIJOS KATEDRA**

**Eveliina Vanhala**
Geologijos studijų programa
Magistro baigiamasis darbas

# TEORINIS FILTRACIJOS KOEFICIENTO NUSTATYMAS LIETUVOS GRUNTŲ MĖGINIUOSE

Mokslinis vadovas
Asistentas dr. Vytautas Samalavičius

Vilnius 2024

## TABLE OF CONTENTS

**ABBREVIATIONS**

| | |
|---|---|
| FSa | Fine sand |
| grMSa | Gravelly medium sand |
| grSa | Gravelly sand |
| grSaFM | Gravelly, medium graded, slightly silty-clayey sand |
| grSaFP | Gravelly, poorly graded, slightly silty-clayey sand |
| grSaFW | Gravelly, well-graded, slightly silty-clayey sand |
| grSaG | Gravelly, gap-graded sand |
| grSaM | Gravelly, medium graded sand |
| grSaP | Gravelly, poorly graded sand |
| MSa | Medium sand |
| O | Organic soil |
| saClL-SiL | Sandy, low plasticity clay-silt |
| SaFM | Medium graded, slightly silty-clayey sand |
| SaFP | Poorly graded, slightly silty-clayey sand |
| SaFU | Uniformly graded, slightly silty-clayey sand |
| saGr | Sandy gravel |
| saGrM | Sandy, medium graded gravel |
| saGrW | Sandy, well-graded gravel |
| SaM | Medium graded sand |
| SaP | Poorly graded sand |
| saSi | Sandy silt |
| saSiL | Sandy, low plasticity silt |
| SaU | Uniformly graded sand |
| siFSa | Silty fine sand |
| SiL | Low plasticity silt |
| siMSa | Silty medium sand |
| siSa | Silty sand |

## INTRODUCTION

Soils have the ability to let fluid flow through interconnected pores between particles. This feature is called permeability (Craig, 2004; Taylor, 1948). The hydraulic conductivity of soil refers to how easily fluid can move through the porous medium (Alyamani & Şen, 1993), and the hydraulic abilities of soil affect factors like water storing and soil stability, and have a significant role in many fields such as geotechnical design, contaminant migration, and waste disposal (Jang et al., 2011).

Hydraulic conductivity can be determined in the field or through laboratory tests. *In situ* tests are usually complex and more expensive than laboratory testing. The disadvantage of laboratory tests is that often samples are disturbed, resulting in the loss of their original internal structure. In addition to field and laboratory testing, empirical methods have also been developed to estimate and predict the hydraulic conductivity of soils. These methods utilize soil properties such as porosity, grain size, soil texture and bulk density (Chapuis, 2012). Besides empirical equations, machine learning offers an alternative method to predict the hydraulic conductivity of soils. It is especially useful when the relationship between soil properties is complex or not linear, or involves multiple parameters (Li et al., 2022); machine learning can use pattern recognition to find relationships between parameters and can be trained to predict new results (Nemes et al., 2006; Twarakavi et al., 2009).

In this thesis, a database of Lithuanian soil samples is created from the soil samples gathered by the investigations in the Department of Hydrogeology and Engineering Geology at Vilnius University. The database is then used to assess the theoretical hydraulic conductivity of Lithuanian soil samples by comparing the saturated hydraulic conductivity values obtained through laboratory testing to three empirical formulas and seven machine learning models.

**Study aim:** Modelling the theoretical hydraulic conductivity of soil samples using empirical formulas and machine learning methods.

**Study objectives:**

- Creating a database for Lithuanian soil samples to use in this study and in the future.
- Utilizing empirical formulas in calculating hydraulic conductivity and assessing their usability and limitations.
- Tuning machine learning models to find the best parameters to use in hydraulic conductivity determination.

**Novelty:** In this work, machine learning methods to study the permeability of Lithuanian soil samples are used for the first time and compared to classical empirical formulas.

**Thesis structure:** This study consists of 45 pages, 31 figures and 8 tables. The work is divided into six major parts: Introduction, theoretical framework and previous research history, geological-hydrogeological setting, materials and methods, results and discussion, and conclusions.

# 1. THEORETICAL FRAMEWORK AND PREVIOUS RESEARCH HISTORY

## 1.1 Saturated and unsaturated soil

Groundwater can be divided into two main zones: the unsaturated and the saturated zone. The unsaturated extends the zone from the ground surface down to the water table. Under the water table lies the saturated zone. In the saturated zone, all pore spaces between soil grains are saturated with water. In the unsaturated zone, however, the soil pores also contain air. Between the two zones, at the water table level, also lies the capillary fringe (Fig. 1). In the capillary fringe zone, the water is drawn upward by capillary forces. The thickness of the capillary fringe is soil-dependent; small pore sizes usually indicate to a thicker capillary fringe than larger pore sizes (Fitts, 2002; Yolcubal et al., 2004).

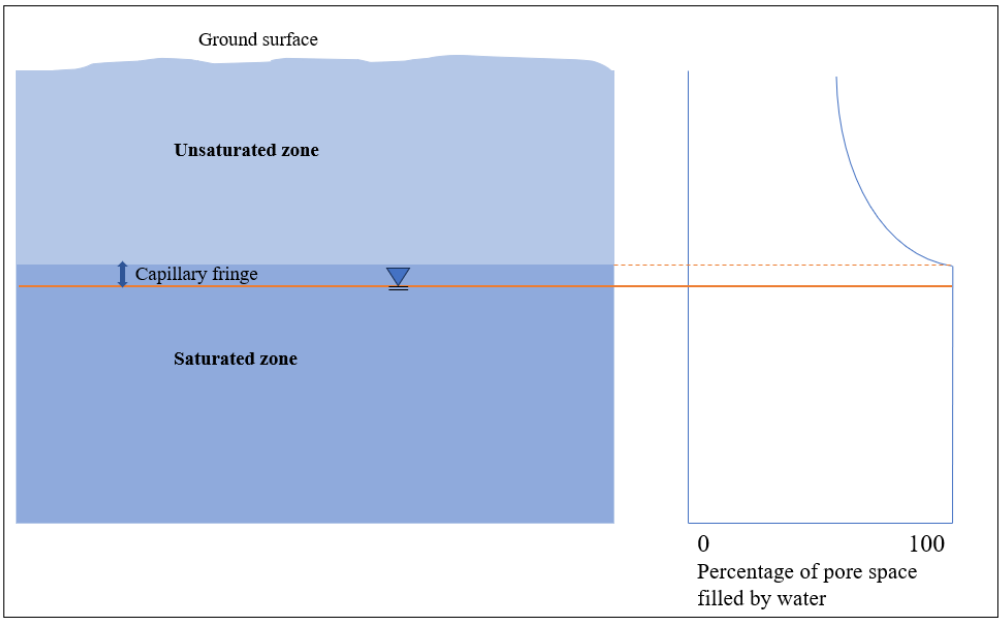



**Fig. 1.** A schematic figure of the unsaturated and saturated zones. Modified from (Fitts, 2002, p. 6) and (Freeze & Cherry, 1979, p. 40).

The soil samples of the database used in this study were collected from various depths, but their testing was conducted under saturated soil conditions. In theory, in the saturated zone the soil-water content is constant and doesn't vary with depth unless the soil is vastly heterogenous. In a uniform soil, the soil-water content is equal to the porosity of the soil (Yolcubal et al., 2004). In estimating and determining the permeability of the unsaturated zone, the presence of air needs to be considered, which makes it a more complex subject than permeability determination of saturated soil. In the unsaturated zone, hydraulic conductivity varies depending on the volumetric water content. Pore-water pressure is not constant and varies with the water content. When water content in the soil decreases, the passages for water to travel through are smaller and more indirect. In theory, hydraulic conductivity decreases when water content decreases (Fitts, 2002). However, matric suction, meaning the difference between pore-air and pore-water pressures, needs to be considered as well, because differences in matric suction affect hydraulic conductivity (Fredlund & Rahardjo, 1993).

Unsaturated permeability will not be further discussed in this thesis, and when "hydraulic conductivity" or "K" are mentioned, they refer to saturated hydraulic conductivity.

## 1.2 Darcy's law and hydraulic conductivity

How water flows underground is dependent on the physical and hydraulic properties of the soil in question (Yolcubal et al., 2004). French hydraulic engineer Henry Darcy's experiments on saturated sands in 1856 set the foundation for water flow characteristics and hydraulic conductivity determination. Darcy investigated the flow of water through homogenous sands and proved that in steady flow conditions the rate of flow is in line with the hydraulic gradient. This empirical principle is called Darcy's law. Darcy's law is valid in the circumstances of linear flow in sands. Errors happens in turbulent conditions of high velocity (Bear, 1972; Freeze & Cherry, 1979; Taylor, 1948).

Darcy studied the relation of flow rate (Q) and the hydraulic head loss of the column (called the hydraulic gradient) (Fig. 2). He concluded that the rate of flow Q through a porous medium is proportional to the cross-sectional area (A) of the column and the hydraulic gradient (i) and inversely proportional to the length (L). The hydraulic gradient is calculated from the change in hydraulic heads (Δh) (Bear, 1972).

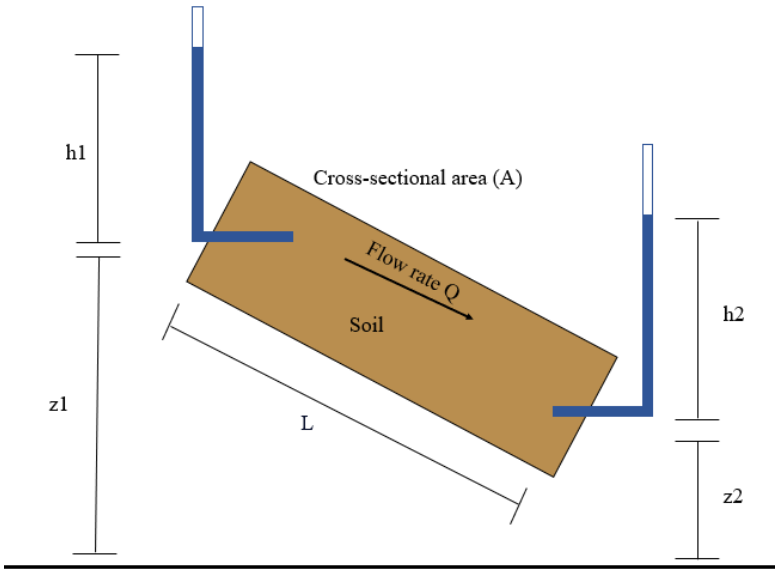



**Fig. 2.** A schematic illustration of Darcy's law. Modified from (Kaliakin, 2017, p. 251).

There are several ways to express Darcy's law mathematically. One such representation is presented in Equation 1 (Hiscock, 2005).

$$Q = -KA\,\frac{\Delta h}{\Delta l} \qquad (1)$$

In this equation, K is hydraulic conductivity, A is the cross-sectional area of the porous medium, and Δh/Δl is the hydraulic gradient (i). To solve hydraulic conductivity by Darcy's law, the equation can be rearranged. This formula is presented in Equation 2.

$$K = -Q\,\frac{\Delta l}{\Delta h\, A} \qquad (2)$$

In Equation 2, K is hydraulic conductivity, Q refers to flow rate, and Δh/Δl is the hydraulic gradient. The negative signs indicate that the flow is towards the decreasing hydraulic head. Hydraulic conductivity has the units of length per time, for example, meters per day (m/d) (Hiscock, 2005).

To summarize, Darcy's law is an empirical law that states that the flow of water in a porous medium is directly proportional to the hydraulic gradient, and the flow rate decreases to the direction of falling hydraulic head. However, Darcy's law is not valid in every case. It is only valid for laminar flow with a low Reynolds number[1] instead of turbulent flow. Very coarse soils, for instance, might have a turbulent flow (Leppäranta et al., 2017). In groundwater conditions, water flow through porous media is often considered to be laminar because water velocities are usually small. Darcy's law is a good measurement of hydraulic conductivity in sands, but in small hydraulic velocities, such as clays, it loses accuracy (Yolcubal et al., 2004). Darcy's law also assumes the soils to be homogenous (isotropic), which is often untrue. For instance, soils might have a layered profile due to e.g. deposition processes, and these layers, consisting of different soil matrices, have different hydraulic conductivity abilities (Leppäranta et al., 2017; Yolcubal et al., 2004).

In general, coarser and fractured soils have higher hydraulic conductivity values than fine-grained soils (Hiscock, 2005).

## 1.3 Soil properties

### 1.3.1 Permeability

As explained in the beginning of this thesis, all soils are permeable, which means fluid can flow around soil particles (Craig, 2004; Taylor, 1948), and hydraulic conductivity is the ability of soil to transmit water, measured by the rate at which water can move through the medium (Alyamani & Şen, 1993). The fluid can be either in liquid or gaseous form, but in hydrogeology, the liquid in most scenarios is water (Head, 1994).

In some older publications the terms *permeability* and *hydraulic conductivity* are used interchangeably. In theory, hydraulic conductivity (K) and permeability (k) are proportional to each other. Hydraulic conductivity (K) is dependent on both the fluid and the porous medium it flows through. The fluid properties are density and viscosity. Permeability (k) is sometimes also called *intrinsic permeability*, since it is dependent only on the characteristics of the medium, not of the fluid (Bear, 1972; Freeze & Cherry, 1979; Hiscock, 2005).

Permeability is not an absolute measure of soil, but instead depends on different factors. The affecting factors are grain size distribution, grain shape and texture, void ratio, degree of saturation, mineralogical composition, soil fabric, nature of fluid, type of flow and temperature. Grain size, shape, texture, and mineralogical composition vary depending on the soil, and void ratio and degree of saturation – while also related to soil characteristics – can vary depending on the chosen soil testing method. Nature of fluid, type of flow and temperature are connected to the permeating fluid, and soil fabric indicates to the soil *in situ*. The nature of fluid refers to the fluid that is flowing in the soil, and the variables are density and viscosity. For water, density and viscosity vary to some extent. Soil fabric is an important concept in the determination of permeability because soils are often anisotropic and consist of different layers. Intrusions, laminations, and other discontinuities affect permeability; hence, laboratory permeability analyses might differ from the actual permeability of the soil in nature, since in laboratory methods the original soil fabric is often disturbed (Head, 1994).

The hydraulic conductivity of different soil types can vary over 13 orders of magnitude (Freeze & Cherry, 1979). While soils are heterogenous in nature, some generalisations can be made. Typical

[1] Reynolds number $R_e$ is a dimensionless parameter that tells whether water flow is laminar or turbulent. Flow is considered laminar if $R_e$=1-10 (Fitts, 2002).

permeability value for gravel is $>10^{-2}$ m/s (864 m/d), and for clay $<10^{-8}$ m/s (0.000864 m/d) (Kaliakin, 2017). Materials with $<10^{-9}$ m/s are considered practically impermeable, and soils with $10^{-9} – 10^{-5}$ m/s have low or very low permeability, while soils with $>10^{-3}$ m/s have high permeability (Carter & Bentley, 1991).

### 1.3.2 Grain size distribution

The distribution of grain sizes is one of the most important characteristics of soil. Hydraulic conductivity is the measure of the ease of fluid flow through a porous medium. Thus, the composition of the porous medium is of interest (Alyamani & Şen, 1993). This composition of soil is typically represented by a grain size distribution curve that also demonstrates the gradation of soil (Craig, 2004). Grain size distribution has an impact on permeability especially when it comes to fine-grained particles, because smaller particles have smaller void spaces between them, thus allowing less room for water to flow, increasing resistance, and decreasing permeability. The shape and texture of grains have an effect as well, because irregularly shaped particles create complex flow paths and are rougher on the surface. These factors increase the resistance of flow as well as the gradation level of the soil (Fig. 3) (Fitts, 2002; Woessner & Poeter, 2020). In a recent study, the effects of grain size and shape on hydraulic conductivity of sands were studied and it was found that sands that had the same gradation characteristics but different shapes yielded different values of hydraulic conductivity (Cabalar & Akbulut, 2016).

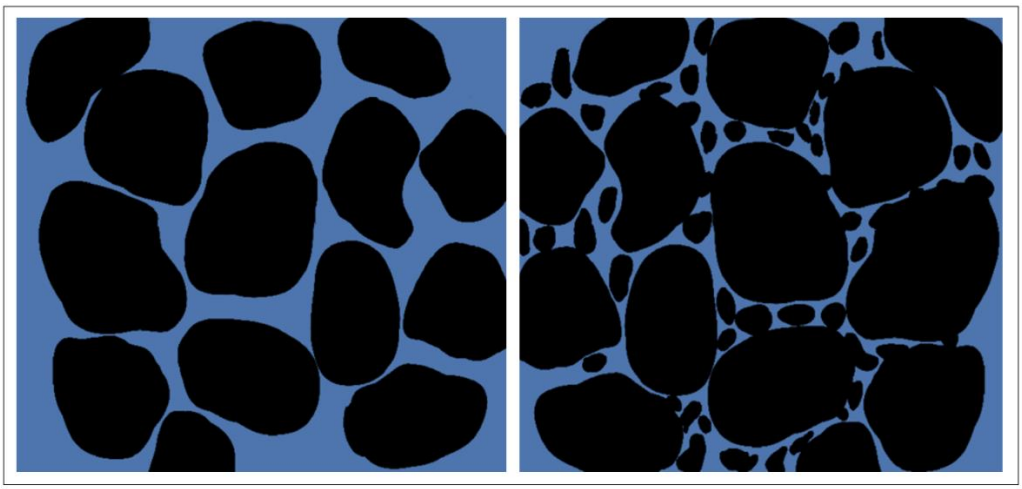


**Fig. 3.** A picture of a well-graded material (left) versus poorly sorted material (right). Modified from (Fitts, 2002, p. 26).

Soil parameters often used in the determination of hydraulic conductivity are the so-called $D_{xx}$ values, meaning the grain diameters. The $D_{xx}$ values correspond to “percentage finer”, meaning that, for instance, a $D_{60}$ particle size means that 60% of the soil is finer than that particular grain size (Bear, 1972; Craig, 2004; Freeze & Cherry, 1979). Many empirical formulas to determine hydraulic conductivity utilize different $D_{xx}$ values, most commonly the $D_{10}$ value (Chapuis, 2012).

Characteristics relating to grain size distribution and grain diameters are the uniformity coefficient ($C_U$) and the coefficient of curvature ($C_C$). These parameters can be calculated with $D_{xx}$ values (Equation 3 and 4) (International Organization for Standardization, 2004).

$$C_U = \frac{D_{60}}{D_{10}} \quad (3)$$

$$C_C = \frac{{D_{30}}^2}{D_{10} \times D_{60}} \quad (4)$$

The uniformity coefficient $C_U$ and the curvature coefficient $C_C$ track the gradation value of the soil and help quantifying the shape of the grading curve. The limit values for different gradation characterizations in the European standard EN ISO 14688-2 are presented in Table 1 below.

**Table 1**. $C_U$, $C_C$ and the shape of grading curve. Modified from (International Organization for Standardization, 2004).

| Shape of grading curve | $C_U$ | $C_C$ |
|---|---|---|
| **Multi-graded** | >15 | 1 < Cc < 3 |
| **Medium-graded** | 6-15 | < 1 |
| **Even-graded** | <6 | < 1 |
| **Gap-graded** | Usually high | Any (usually < 0.5) |

In theory, in a completely uniform soil where all soil particles are of the same size, the uniformity coefficient and curvature coefficient would be 1. The shape of the grading curve can help understand soil permeability, as uniform soils might have more straightforward routes for water to travel (Woessner & Poeter, 2020).

### 1.3.3 Void ratio and porosity

Void ratio is one of the most important factors affecting permeability. Permeability increases significantly with the increase of void ratio (Dolzyk & Chmielewska, 2014). Void ratio refers to the number of void spaces between soil grains. Void ratio (e) is closely tied to porosity (n). Porosity is expressed by the ratio of volume of void space to the total volume. Porosity is a dimensionless unit often expressed as the range of $0 < n < 1$. Void ratio – also dimensionless – is the volume of voids divided by the volume of solids. The relationship between porosity *n* and void ratio *e* is presented in Equations 5 and 6 (Fitts, 2002; Woessner & Poeter, 2020).

$$n = \frac{e}{1+e} \quad (5)$$

$$e = \frac{n}{1-n} \quad (6)$$

If void ratio is now known, porosity can also be calculated using the coefficient of uniformity ($C_U$), as presented in Equation 7 (Vukovic & Soro, 1992, as cited in Odong, 2008).

$$n = 0.255(1 + 0.83^{C_U}) \quad (7)$$

In the saturated zone all spaces are filled with water, including voids. This means that in theory the water content of soil is equal to the porosity of the soil (Yolcubal et al., 2004). Void ratio and porosity are influenced by grain sizes and shapes. Unconsolidated sediments that consist of rounded and angular particles have higher porosity values than consolidated sediments (Hiscock, 2005). In general, it can be said that the bigger the void spaces, the easier flow. However, high porosity does not always equal to high permeability; the voids between soil particles need to be interconnected for fluid to be able to flow through (Head, 1994.)

Soil compaction and soil density affect the available flow path of water, too. If the soil is compacted (i.e. the density per unit area is higher), the solids in the soil are packed into a smaller volume area and the grains are closer together. This means less pore space and less room for water to move. Higher density of soil might indicate lower hydraulic conductivity (Moorberg & Crouse, 2021).

## 1.4 Methods for determining hydraulic conductivity

There are multiple ways to measure and determine the permeability of soils. The methods can be described as either direct or indirect. Direct measurements indicate to tests conducted in the field or laboratory, and indirect methods are comprised of calculations based on soil properties (Head, 1994).

### 1.4.1 Direct methods

Hydraulic conductivity can be directly determined *in situ* with e.g. slug tests or pumping tests (Freeze & Cherry, 1979). Two main laboratory tests for saturated hydraulic conductivity measurement are the constant-head method and the falling-head method. The constant-head method is meant for soils with higher permeability (e.g. sands), and the falling-head test is better for soils with lower permeability, like silt and clay (Craig, 2004; Freeze & Cherry, 1979).

The constant head method is one of the most popular laboratory methods. In the constant-head method, the soil sample is placed in a cylinder and let it get fully saturated with water (Fig. 4). The sample is then subjected to a steady vertical flow of water under a constant head difference. The hydraulic gradient and water flow volume per unit time are measured, and then Darcy's law can be applied to calculate hydraulic conductivity (Craig, 2004).

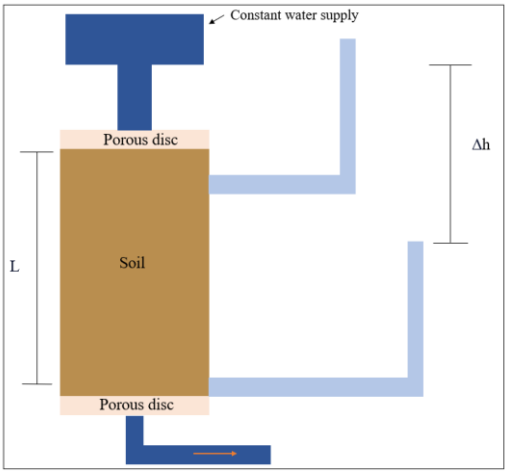



**Fig. 4.** A schematic picture of the constant-head method. Modified from (Craig, 2004, p. 33).

The falling-head method is usually applied for finer soils. The soil sample is placed inside a cylinder and a standpipe is attached to the top of the sample (Fig. 5). The soil is fully saturated with water. Then water from the standpipe is sent flowing through the sample soil into the reservoir below. The head loss between locations h1 and h2 and the time it takes for water to flow through are recorded (Craig, 2004).

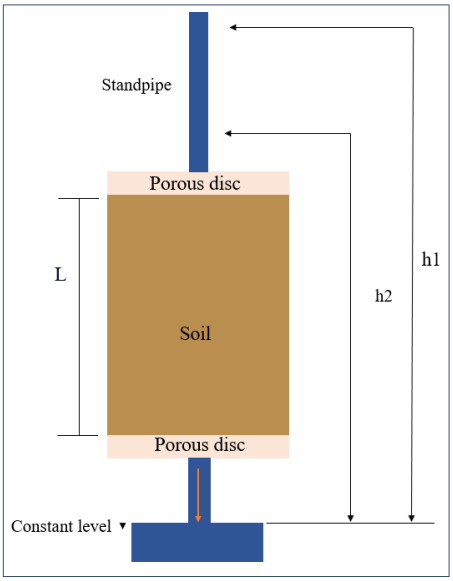



**Fig. 5.** A schematic picture of the falling head method. Modified from (Craig, 2004, p. 33).

Sometimes, for laboratory testing, soil samples representing the soil matrix are hard to obtain, and they might not properly represent the soil as a whole (Pap & Mahler, 2019). Laboratory tests are usually performed on small samples, and - depending on the sampling method and process - the samples are disturbed and remoulded, and they lose their original fabric and texture (Fitts, 2002). Other disadvantages of laboratory (and field) tests is that they can be costly and time-consuming (Boadu, 2000).

### 1.4.2 Indirect methods

Several indirect methods concerning soil permeability determination have been invented, most notably formulas utilizing grain size characteristics and other soil parameters. While these methods are theoretical or empirical in nature, some properties of the soil are still required. Without *in situ* investigations the soil texture cannot be tested, but certain geometric properties like porosity can be determined experimentally and used as parameters that might reflect the actual geometry of the soil matrix (Bear, 1972).

The correlation between grain size distribution and hydraulic conductivity has been studied for over a century, and many empirical determination formulas for hydraulic conductivity make use of grain size in the calculations - for example, Hazen, Slichter, USBR, Kozeny-Carman, and Terzaghi equations. A reliable permeability estimation method should consider porosity or void ratio and some characteristic grain size (Chapuis, 2012). Measuring pores and their diameters instead of diameters of grains would be more telling of the hydraulic properties of soil, but pore size distribution is more difficult to determine. That is why the estimation of soil hydraulic properties is more commonly based on grain size distribution, which is easier to measure (Pinder & Celia, 2006). In recent years, several studies have investigated the relationship between grain size and hydraulic conductivity and applied

empirical formulas to calculate hydraulic conductivity (see e.g. Chakraborty et al., 2006; Salarashayeri & Siosemarde, 2012; Pucko & Verbovšek, 2015; Onwe et al., 2016; Ann et al., 2022).

However, even these indirect methods, like empirical formulas, require some soil parameters for the determination, and some parameters can be complex to acquire. They also have some limitations and low accuracy (Singh et al., 2020). Limitations can include, for example, the range of applicability of formulas only to certain soil types (Chapuis, 2012).

Statistical modelling can help determine hydraulic conductivity and other soil properties. In a study from 2011, a statistical regression model was made by studying the relationship between saturated hydraulic conductivity and grain size parameters (Pliakas & Petalas, 2011), and in another study, a statistical hydraulic conductivity model was developed using the effective diameter $D_{10}$ and standard deviation (Chandel et al., 2022). While statistical models are often used to quantify the relationship between the input and output parameters, they often have limitations; the relationship between the parameters used for the prediction and the soil in question might not stay constant, but instead vary (Van Looy et al., 2017).

This is where machine learning can be more appropriately applied. Machine learning algorithms can adapt to the changes in the relationship between several parameters by using pattern recognition, and they can be trained to predict new results (Nemes et al., 2006; Twarakavi et al., 2009). In the Preface of *Introduction to Machine Learning* (Alpaydin, 2010, para. 2), one of the problems machine learning can help with was described as following: *"—when the problem to be solved changes in time, or depends on the particular environment. We would like to have general-purpose systems that can adopt to their circumstances, rather than explicitly writing a different program for each special circumstance*". This statement is true for the problems concerning hydrogeological modelling. Soil is often vastly heterogenous, which makes determination and prediction of different parameters, such as hydraulic conductivity, difficult especially in larger-scale investigations.

In chapter *3. Materials and Methods,* three empirical formulas and seven machine learning models used in this thesis to determine hydraulic conductivity will be introduced.

# 2. GEOLOGICAL-HYDROGEOLOGIAL SETTING

## 2.1 Soil conditions in Lithuania

Quaternary deposits in Lithuania formed during the last glaciation, during which Lithuania was covered by glacial ice, and they are the most common surface formations and extend across the majority of the territory (Karmaza & Baltrūnas, 2004). These glacial deposits vary depending on their formation processes; moraine loams are the most common surface formation, sand and gravel layers have been formed due to melting glacial water, and clays and silts have been layered in glacial lakes. Some aeolian sand dunes formed in this time period can be found in Southern Lithuania (Bičkauskas et al., 2011). The thickness of the Quaternary deposits can be up to 300 meters, but the average thickness is 130 meters. The deposits are the thinnest in the northern part of the country where the thickness varies from 10 to 30 meters (Guobyte & Satkunas, 2011).

Around 60 % of Lithuania's surface is covered by moraine loams, clays and other low-permeability soils. The filtration properties of moraine loams are of interest as many infrastructural works like roads and railways have been and are being built on top of these soils. Moreover, in Northern Lithuania there are karst regions in Biržai and Pasvalys districts, where there are moraine loams on top of Devonian sulphate carbonate rocks. Water infiltrating through the moraine loams in these areas causes karst processes and sinkholes to develop and progress. Overall, the filtration capabilities of soils affect their mechanical properties and foundation stability (Klizas et al., 2015). The filtration features in moraine loams and other lithological units in Lithuania have been studied extensively in the 21st century (Klizas & Šečkus, 2007; Klizas, 2014; Klizas et al., 2014; Klizas et al, 2015; Klimašauskas et al., 2020).

## 2.2 Lithuanian soil sample data

For this thesis, a database of Lithuanian soil samples was created. The samples in the database have been gathered from multiple locations in Lithuania between 2018 and 2022. The main source of information was the internal archive of geological investigations in the Department of Hydrogeology and Engineering Geology at Vilnius University. The database was created to integrate sampling data in a way that is accessible and functional for this study and further studies concerning soil permeability in Lithuania.

The locations of the samples used in this work can be seen in Fig. 6 and 7.

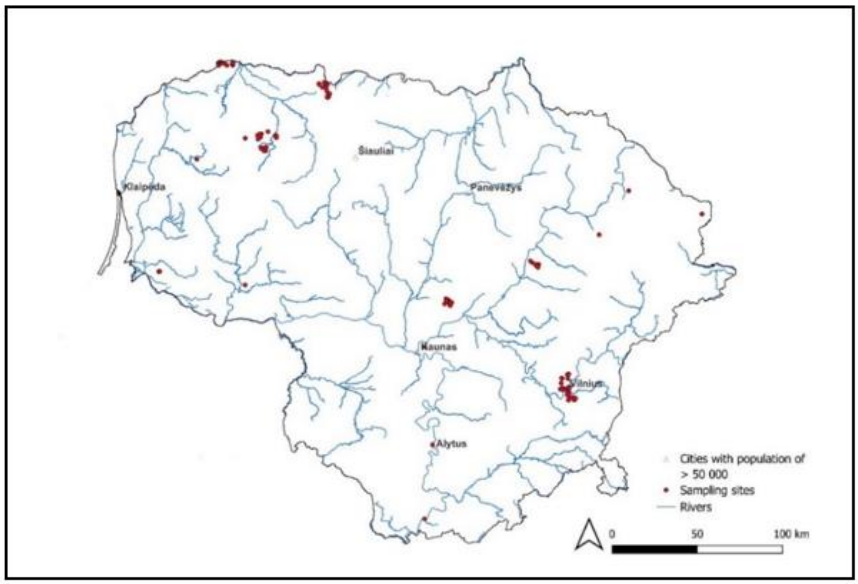



**Fig. 6.** Sampling site locations.

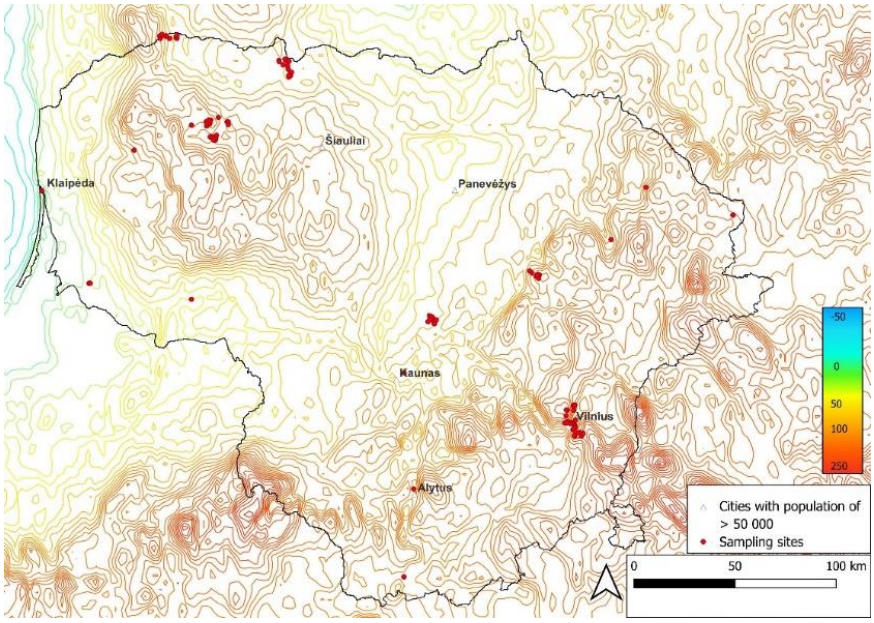



**Fig. 7.** Sampling site locations with a surface elevation model.

The soil parameters in the samples were acquired through laboratory test results and/or geological engineering reports that included hydrogeological testing, prepared for different clients and projects.

## 2.3 Soil sample database

The database was created using Microsoft Excel[2]. While Microsoft Excel is essentially a spreadsheet tool, for small to medium-sized datasets it is proved to be usable as a database. It has features such as filtering, sorting, and data validation, which are all useful functions in data manipulation. Excel's import and export functions to different data formats ensure easy

[2] Microsoft Excel. https://www.microsoft.com/en-us/microsoft-365/excel

transformation to other programs. As the data in this study was used in machine-learning via Python, Excel was a suitable choice for creating the database.

The database consists of 246 entries. The entries include, most importantly, hydraulic conductivity values and grain size distributions of each sample. The hydraulic conductivity values in 206 samples were acquired with the International Organization for Standardization (ISO) standard for the constant-head method (LST CEN ISO/TS 17892-11:2005, LST CEN ISO/TS 17892-11:2005/AC:2006 or LST EN ISO 17892-11:2019), and 40 samples were acquired with a KFZ-type constant-head filtrometer (Klizas, 2003). The database originally has over 500 samples, but for this thesis' purpose, the data was cleaned of missing or insufficient information entries.

Besides hydraulic conductivity, several other parameters from the laboratory reports were added to the database. These include density of the soil (ρ g/cm$^3$), water content (w), degree of saturation (Sr) and void ratio (e). These parameters have values taken before the constant-head testing has taken place, and after testing. Water temperature (°C), reference water temperature (°C), and site information (site name, coordinates, well number, sample number, depth of sampling) were also included. The hydraulic conductivity information in the laboratory reports was presented in meters per day (m/d). The tests have been conducted in different water temperatures, but the hydraulic conductivity values have been later adjusted to 10 °C.

### 2.3.1 Hydraulic conductivity values

In the sample entries used for this study, the lowest hydraulic conductivity value is 0.05 meters per day (m/d) and the highest is 27.90 m/d. If the conductivity values from constant head and filtrometer method are examined separately, the hydraulic conductivity values by the ISO standard method vary between 0.05-27.90 m/d, and KFZ filtrometer values range from 0.1 to 15.8 m/d (Table 2).

**Table 2.** Hydraulic conductivity data from the database used in this study.

| **Method** | **Number of entries** | **Min K*** | **Max K** | **Average K** | **Q1*** | **Q2* (median)** | **Q3*** | **Standard deviation** |
|---|---|---|---|---|---|---|---|---|
| All | 246 | 0.05 | 27.90 | 4.12 | 0.56 | 2.50 | 6.01 | 4.77 |
| ISO standard | 206 | 0.05 | 27.90 | 3.86 | 0.48 | 2.15 | 5.67 | 4.78 |
| KFZ | 40 | 0.10 | 15.80 | 5.44 | 1.73 | 4.00 | 8.30 | 4.54 |
| * K=hydraulic conductivity in meters per day, m/d<br>*Q1= 25 % quartile<br>*Q2= 50 % quartile<br>*Q3= 75 % quartile | | | | | | | | |

The distribution of hydraulic conductivity is illustrated in Fig. 8. Almost half of all the samples have < 2 m/d hydraulic conductivity. The distribution of hydraulic conductivity goes down with the number of entries; the least number of samples have the largest hydraulic conductivity of >20 m/d. Of the 110 entries that have a hydraulic conductivity of <2 m/d, a more specific distribution is presented in the pie chart on the left of the graph. Of the 110 values having a hydraulic conductivity of < 2 m/d, more than 50 % fall into the smallest hydraulic conductivity category of 0.05-0.5 m/d.

Soils with 0.05 m/d are medium-permeability soils, but 0.5 m/d is already considered a high-permeability soil (Carter & Bentley, 1991). Hence, the majority of the soils in this database are high-permeability soils.

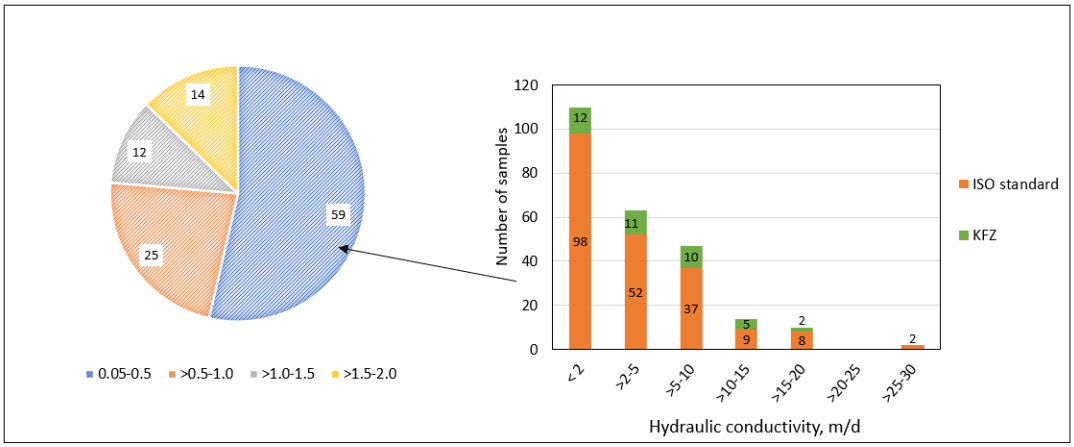



**Fig. 8.** Hydraulic conductivity distribution in the database.

In the database, the upper and lower depth of each sample is given (except for three samples that did not have depth information). The average thickness of the acquired soil samples is 0.35 meters, with the smallest sample thickness being 0.1 meters and largest 1.3 meters. The depth of the acquired soil samples vary from 0.1 meters to 24.5 meters from ground surface. The average depth of sampling is at 8.0 meters. Sampling depth plotted against the number of samples is depicted in Fig. 9 a. Most of the samples are acquired from 0.1-8.0 meter depth. Average hydraulic conductivity values plotted against depth of sampling is depicted in Fig 9 b.

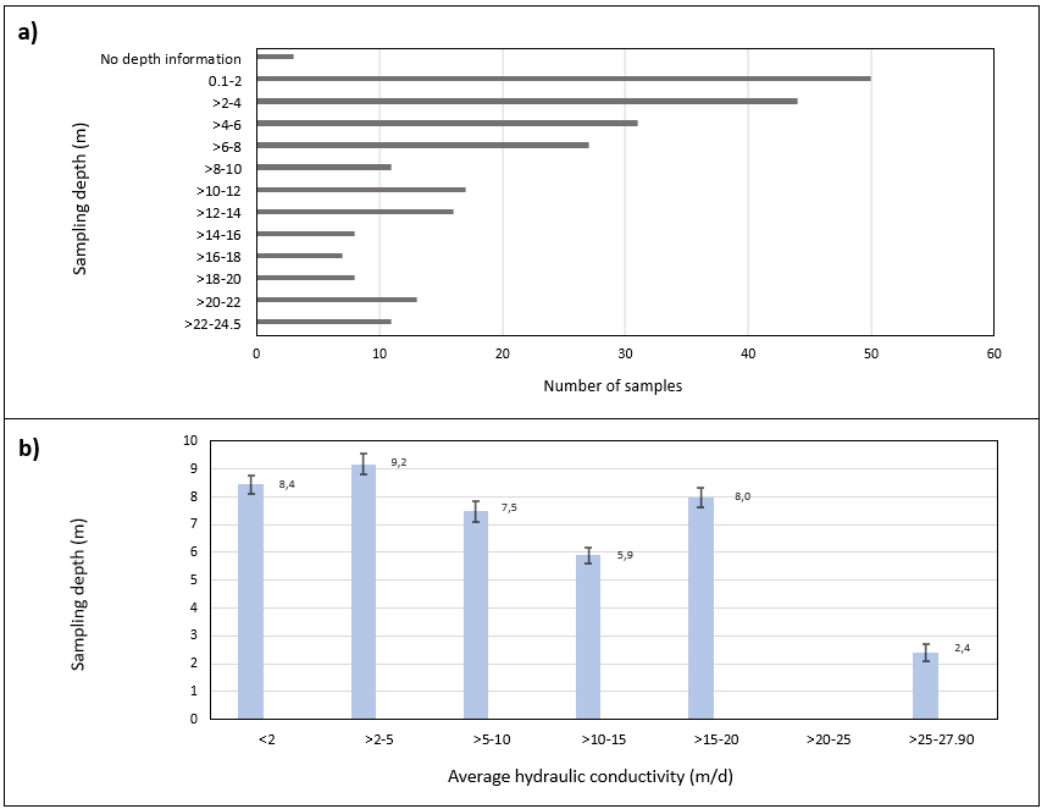



**Fig. 9.** Sampling depth versus number of samples (a) and average hydraulic conductivities versus depth of sampling (b).

Comparing the average values of hydraulic conductivity to depth (Fig. 9 b) shows that samples with smaller hydraulic conductivity have been acquired, generally, from deeper than the ones with higher hydraulic conductivity. For example, samples with a hydraulic conductivity of >25-27.90 m/d, on average have been acquired from 2.3 meters while samples with a hydraulic conductivity of <2 m/d are from 8.3 meters.

### 2.3.2 Grain size distribution

The grain size distribution of the soil samples has been acquired via sieving or sedimentation, as described in ISO 17892-4:2004/2016/2017. In the sieve method, the soil particles are separated by different sized sieves. The amount of soil particles retained on each sieve size is weighted, and the mass of different-sized soils can be transferred to percentages. The sieve method is applicable in soils with less than 10 % of fine soil (silt and clay). For soils containing more than 10% of fines, a sedimentation method via a hydrometer was used. A sedimentation method is a process where the differences in settling rate of soil is calculated, and this separates the particles sizes (International Organization for Standardization, 2016).

The soils have been classified according to the ISO 14688-2 standard, where their grain size distribution and gradation, plasticity and organic content are considered (International Organization for Standardization, 2004). In Fig. 10, the number of samples in each soil type have been presented.

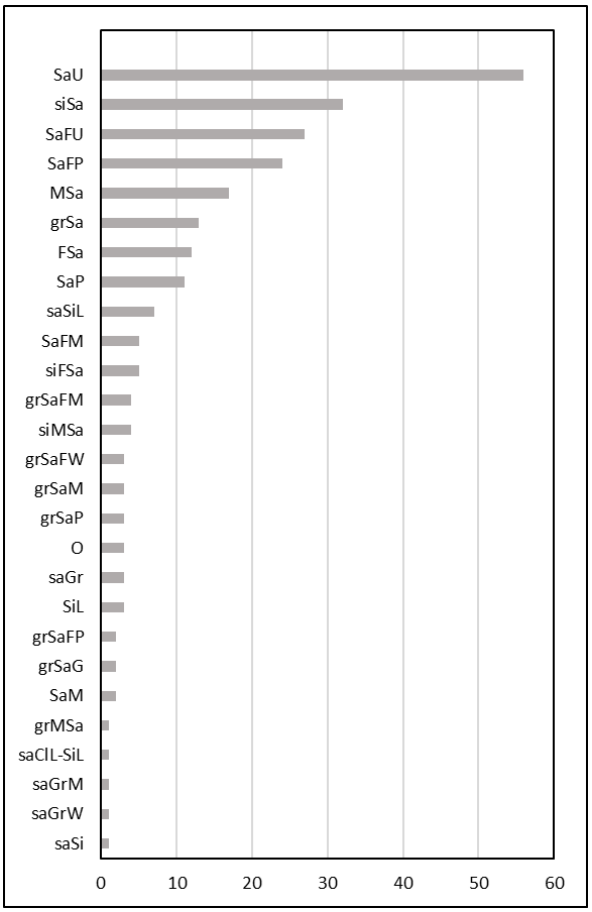



**Fig. 10.** Number of samples in each soil classification name.

Almost 60 samples are classified as uniform sands (SaU). The second largest category is silty sands (siSa). The third largest category is uniform, slightly silty-clayey sands (SaFU). There are three

samples classified as organic soils (O). The organic content is high enough for it to become the primary type of the soil, but in the soil type information for these samples, there is additional descriptive information: *'slightly silty-clayey, uniformly graded sand with a little organic matter', 'uniformly graded sand with a mixture of organic matter'*, and *'slightly silty-clayey, uniformly graded sand with a little organic matter'.*

All abbreviation meanings can be found in the abbreviation table in the beginning of this document.

### 2.3.3 Other parameters

Other parameters besides hydraulic conductivity and grain size distribution were also given in the laboratory reports. These are bulk density of the soil (ρ g/cm$^3$), water content (w), degree of saturation (Sr), and void ratio (e). The parameters include values from both before and after hydraulic conductivity testing. The minimum, maximum, average, and quartiles Q1, Q2, Q3 as well as standard deviation of the parameters in the database are presented in Table 3.

**Table 3.** Soil sample parameters.

| | **Number of samples** | **Min value** | **Max value** | **Average value** | **Q1*** | **Q2* (median)** | **Q3*** | **Standard deviation** |
|---|---|---|---|---|---|---|---|---|
| **e (before test)** | 156 | 0.40 | 1.82 | 0.74 | 0.62 | 0.74 | 0.83 | 0.18 |
| **e (after test)** | 156 | 0.42 | 1.83 | 0.77 | 0.65 | 0.75 | 0.86 | 0.20 |
| **ρ g/cm$^3$ (before test)** | 192 | 1.23 | 2.14 | 1.73 | 1.61 | 1.75 | 1.86 | 0.17 |
| **ρ g/cm$^3$ (after test)** | 175 | 1.42 | 2.20 | 1.90 | 1.84 | 1.91 | 1.98 | 0.12 |
| **w (before test)** | 206 | 0.002 | 0.50 | 0.12 | 0.05 | 0.13 | 0.18 | 0.08 |
| **w (after test)** | 205 | 0.1 | 0.64 | 0.26 | 0.21 | 0.25 | 0.30 | 0.07 |
| **Sr (before test)** | 156 | 0.006 | 1.07 | 0.43 | 0.19 | 0.46 | 0.64 | 0.26 |
| **Sr (after test)** | 156 | 0.52 | 1.30 | 0.90 | 0.83 | 0.91 | 0.98 | 0.11 |
| *Q1= 25 % quartile<br>*Q2= 50 % quartile<br>*Q3= 75 % quartile | | | | | | | | |

# 3. MATERIALS AND METHODS

## 3.1 Data preparation

Before conducting hydraulic conductivity calculations, some preparatory needed to be done for the soil samples. These preparations are introduced in the subchapters below.

### 3.1.1 Grain size diameters

The soil samples in the database included grain size distribution charts with absolute grain size information. Grain size diameters ($D_{xx}$ values) are often needed to calculate hydraulic conductivity with empirical equations. For this work, $D_{xx}$ values ranging from $D_{10}$ to $D_{90}$ were acquired mathematically. From absolute percentages given in the grain size distribution charts, cumulative percentages were calculated, as can be seen in an example soil sample in Table 4 and Fig. 11. By using a Python code by the Department of Hydrogeology and Engineering Geology at Vilnius University, grain sizes were assigned a grain size class and $D_{xx}$ values were acquired by interpolating grain size classes to their corresponding cumulative percentages. In the example soil sample distribution in Table 4, the $D_{10}$ value, where 10% of the soil is finer, falls between grain size classes 0.06 mm and 0.2 mm. Interpolating 0.06 and 0.2, and their corresponding cumulative percentages 5.81 and 20.54 together gives a mathematical estimate of the $D_{10}$ value which, in this case, is 0.0998 mm.

**Table 4.** An example soil sample distribution**.**

| Grain size (mm) | Grain size class | Absolute % | Cumulative % |
|---|---|---|---|
| <0.06 | 0.01 | 5.81 | 0 |
| 0.06-0.2 | 0.06 | 14.73 | 5.81 |
| 0.2-0.63 | 0.2 | 46.23 | 20.54 |
| 0.63-2.0 | 0.6 | 20.18 | 66.77 |
| 2.0-4.0 | 2 | 6.62 | 86.95 |
| >4.0 | 4.75 | 6.43 | 93.57 |
| All | 10 | 100 | 100 |

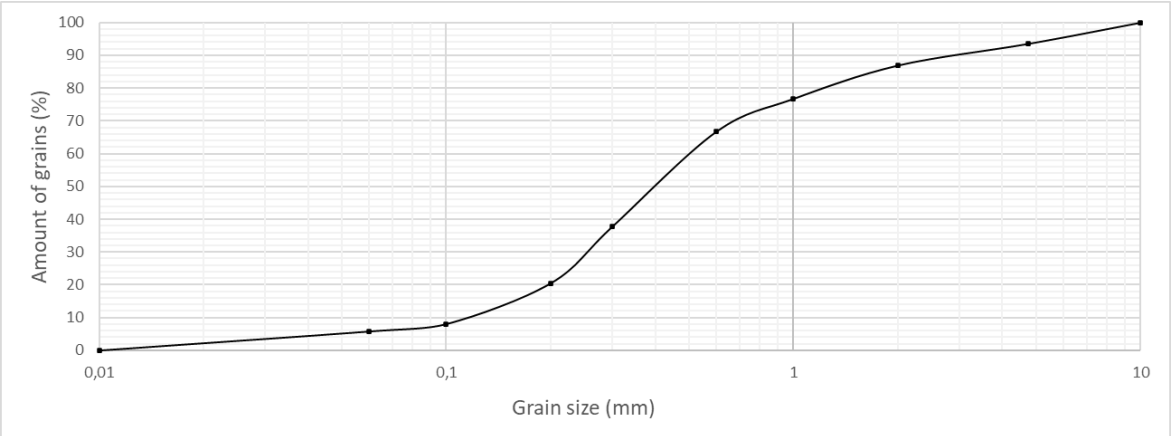



**Fig. 11.** Grain size distribution of an example soil sample.

$D_{xx}$ values can also be manually acquired from the grain size distribution curves, as can be seen from Fig. 11. It displays the same soil sample as presented in Table 4 above. Following the grain size percentages on the right side of the plot, the $D_{10}$ value of the soil can be found at the 10 % mark. This method, however, does not give precise $D_{xx}$ values and is a slower, manual method and difficult to utilize in large datasets.

### 3.1.2 Machine learning data

The machine learning method testing was carried out in Jupyter Notebook[3]. It is an open-source, interactive platform where documents containing code, data and visualizations can be shared. Jupyter supports various programming languages. The machine learning code for this study was written in Python by assistant dr. Vytautas Samalavičius. The code utilizes different libraries and databases, such as Pandas, Matplotlib and NumPy. Pandas[4] is a general Python library supporting data analysis and manipulation and has tools for e.g. reading, writing and merging data, Matplotlib[5] is for creating Python visualizations and graphs, and NumPy[6] supports numerical computing. On top of these, specific modules from Scikit-Learn[7] were imported. Scikit-Learn is an open-source library for Python containing algorithms for different machine-learning analyses.

In the soil sample information in the database, there are both *before* and *after testing* values for different soil parameters. The *before* values were aqcuired before the constant-head hydraulic conductivity testing, and *after* values were measured after hydraulic conductivity testing where the soil has been subjected to a steady water flow. For example, bulk density ($\rho$ g/cm$^3$) in one sample was 1.68 g/cm$^3$ before the test, and 1.97 g/cm$^3$ after the test. This means that during the constant-head hydraulic conductivity testing, the sample underwent some compaction. In the machine learning part of the study, the *after* values were used in the prediction, as they might reflect the soil‘s natural behaviour after it has been saturated with water slightly better than the values taken before testing.

The parameters that were transferred to Jupyter Notebook from the database as .CSV data were: water content, soil density, saturation degree, void ratio, $D_{xx}$ values from $D_{10}$ to $D_{90}$, and six grain size classes. In the database, there are originally more than ten grain size categories, depicting grain size information in a more precise way – for instance, in some samples, the grain size interval 0.06-0.2 mm is further divided into two classes: 0.06-0.106 mm and 0.106-0.212 mm, and samples containing a large amount of fine grains also have the distribution of fines depicted more closely (<0.002 mm, 0.002-0.0063 mm, 0.0063-0.02 mm, 0.02-0.063 mm). For the purpose of this work, the grain size distribution information was unified into six classes: <0.06 mm, 0.06-0.2 mm, 0.2-0.6 mm, 0.6-2 mm, 2-4.75 mm and >4.75 mm.

The preprocessing of the imported database included data standardization. Standardization means that the data is scaled so that all of the parameters and features of the data have the same scale, for example from zero to one. The data was standardized to resemble a normally distributed data (Gaussian with zero mean and unit variance) (Scikit-Learn, 2024a).

A specific data column (actual hydraulic conductivity values) in the database was set out as the target (output) variable in the machine learning code that the model aims to predict. The database was then split into test and train sets, where 25 % were test set values and 75 % train set values. This

[3] Jupyter Notebook. https://jupyter.org/
[4] Pandas. https://pandas.pydata.org/
[5] Matplotlib. https://matplotlib.org/
[6] NumPy. https://numpy.org/
[7] Scikit-Learn. https://scikit-learn.org/

means that, for example, in the case of a dataset consisting of 100 values, 25 samples would be reserved for the test set and 75 samples would be reserved for the train set. It is necessary to split data into test and train sets to accurately assess the performance of the machine learning model. The train set data is used to train the model, while the separate test set data is used to evaluate how well the model works on data on data it hasn‘t seen (Alpaydin, 2010).

## 3.2 Empirical formulas

Empirical equations for hydraulic conductivity determination have been invented for well over a century. Most equations utilize grain sizes in the determination process, and, depending on the equation, porosity function of the soil, sphericity, and other parameters (Říha et al., 2018). Three empirical equations used in this thesis by Hazen, Slichter and USBR will be introduced below.

### 3.2.1 Hazen

One of the first people to find relation between soil porosity and hydraulic conductivity was Allen Hazen (Říha et al., 2018). In 1892 and later in 1911, Hazen developed an empirical formula for predicting permeability of saturated sands based on the $D_{10}$ particle size (where 10% of the soil is finer). Hazen's equation is based on the observations of loose sands with a uniform texture. While the equation was originally developed for the design of sand filters for water purification, it is still used to estimate hydraulic conductivity. Hazen's formula is presented in Equation 8 (Carrier, 2003).

$$K = C_H \times {D_{10}}^2 \quad (8)$$

In this equation, K is hydraulic conductivity in cm/s, $C_H$ is Hazen's empirical coefficient, and $D_{10}$ is the effective grain size $D_{10}$ in centimetres. Hazen found that $D_{10}$ is the effective size that has the most effect on the soil's hydraulic properties (Wenzel & Fishel, 1942). Hazen's empirical coefficient $C_H$ is often assumed to be 100, but based on a study, the reported range of coefficient values in geotechnical textbooks is 1-1000 (Carrier, 2003). The coefficient depends on the uniformity coefficient $C_U$, the shape and chemical composition and compactness of the soil and the purity of it. Hazen said that the value of $C_H$ decreases as the uniformity coefficient $C_U$ increases (Hazen, 1905; Wenzel & Fishel, 1942). With units in cm/s, the coefficient can be seen to vary from 40 to 150 in most sands, where 40 is for finer and 150 for coarser soils (Fitts, 2002).

Hazen's empirical coefficient is set for water temperature of 10 C°. The coefficient due to temperature can be calculated with $C_H$= (0.7-0.3 T), where T is temperature C° (Hazen, 1892).

The limitations of Hazen's equation are that it only considers grain size as a hydraulic conductivity parameter. The equation is only applicable to grain size $D_{10}$ of 0.1-3.0 mm and a uniformity coefficient $C_U \leq 5$. Moreover, for the equation to be considered accurate, the sample needs to be loose and not compacted (meaning that the void ratio *e* meets its maximum conditions) (Chapuis, 2012).

### 3.2.2 Slichter

In 1899, Charles Slichter determined the hydraulic conductivity of sands with the assumption that all grains are spherical and distributed regularly within the soil matrix. His equation of hydraulic

conductivity determination includes a coefficient that is based on the correlation of laboratory measured porosity and hydraulic conductivity values. The coefficient is a computed value for a given porosity and the laboratory-confirmed hydraulic conductivity. As an example, for 26 % porosity, the coefficient is 84.3 (Graton & Fraser, 1935; Slichter, 1899). Slichter's studies showed that same size spherical grains can have porosity from 25.95% to 47.64% depending on the compactness of the soil (Wenzel & Fishel, 1942).

Different empirical equations to calculate hydraulic conductivity can be turned into a general formula (Vukovic and Soro, 1992, as stated in Odong, 2008). This is presented in Equation 9.

$$K = \frac{g}{v} \times C \times f(n) \times {d_e}^2 \quad (9)$$

In this general formula, K is hydraulic conductivity in m/s, g is gravitational acceleration (9.81 $m^2/s$, v is kinematic viscosity ($m^2/s$), C is a dimensionless sorting coefficient, f(n) is a porosity function and $d_e$ is the effective grain diameter, for example $D_{10}$.

Slichter's formula, using this general formula, becomes:

$$K = \frac{g}{v} \times 1 \times 10^{-2} \times n^{3.287} \times {D_{10}}^2 \quad (10)$$

In this equation, K is hydraulic conductivity in m/s. Slichter's original coefficient based on porosity has been transformed into $1 \times 10^{-2} \times n^{3.287}$. Slichter's formula is applicable for $D_{10}$ value between 0.01 mm and 5 mm (Odong, 2008).

### 3.2.3 USBR

The United States Bureau of Reclamation's (USBR) equation for hydraulic conductivity utilizes the effective grain size $D_{20}$ in the determination. Originally, the USBR equation was presented in a table containing approximate permeability coefficients of various soils based on the $D_{20}$ grain size, acquired from field tests. The table presents most examples for sizes from coarse silt to coarse sand; gravel, clay and fine silt all have just one example permeability coefficient per each (Justin et al., 1945; Urumović et al., 2020).

Utilizing the general equation presented earlier (Equation 9), the USBR formula can be written as presented in Equation 11 (Odong, 2008).

$$K = \frac{g}{v} \times 4.8 \times 10^{-4} \times {D_{20}}^{0.3} \times {D_{20}}^2 \quad (11)$$

Another, commonly used way to express the USBR equation is presented in Equation 12 (Urumović et al., 2020).

$$K = 0.36 \times {D_{20}}^{2.3} \quad (12)$$

In Equation 12, hydraulic conductivity is K in cm/s and $D_{20}$ is the effective grain size diameter expressed in millimetres. The coefficient value 0.36 in Eq.12 is the same value as the general equation's part $\frac{g}{v} \times 4.8 \times 10^{-4}$ when viscosity 0.0131 $m^2/s$ is used for 10 C° temperature.

The range of applicability for the USBR method is medium sands with $C_U < 5$ (Cheng & Chen, 2007). In the ISO 14688-1 classification, the range for medium-grained sand is 0.2-0.63 mm and for sands in general, 0.063-2 mm (International Organization for Standardization, 2004).

## 3.3 Machine learning methods

Machine learning methods are useful at learning complex and non-linear relationships between parameters, which makes them useful in soil sciences (Li et al., 2022). The methods can be either supervised or unsupervised. Supervised learning aims to learn a relationship between input parameters and output value that is the correct or desired value given by the user (Alpaydin, 2010). In the case of hydraulic conductivity determination, the output value is the laboratory-acquired hydraulic conductivity. In unsupervised learning, there is only input data.

The objective of machine learning is the prediction of new cases, not to replicate the existing data. The more the algorithm sees test samples – based on which the prediction will be produced – the more the underlying function between parameters is known. *Generalization* is the action of how well a model can make accurate predictions on new, unseen data. *Underfitting* happens when the chosen model is too simple and does not represent the real relationship between input and output values, and *overfitting* happens when the model is too complex for the data and the model may not learn the existing function between parameters (Alpaydin, 2010). In Fig. 12 a, an illustration shows a line that separates two datasets in a precise way, and in Fig. 12 b, the same datasets are separated by a simpler, curved line. Despite the few misclassified data points, this type of decision surface might generalize data better while the decision surface on Fig. 12 b is prone to overfitting and would not generalize new, unseen data as well (Mohri et al. 2018).

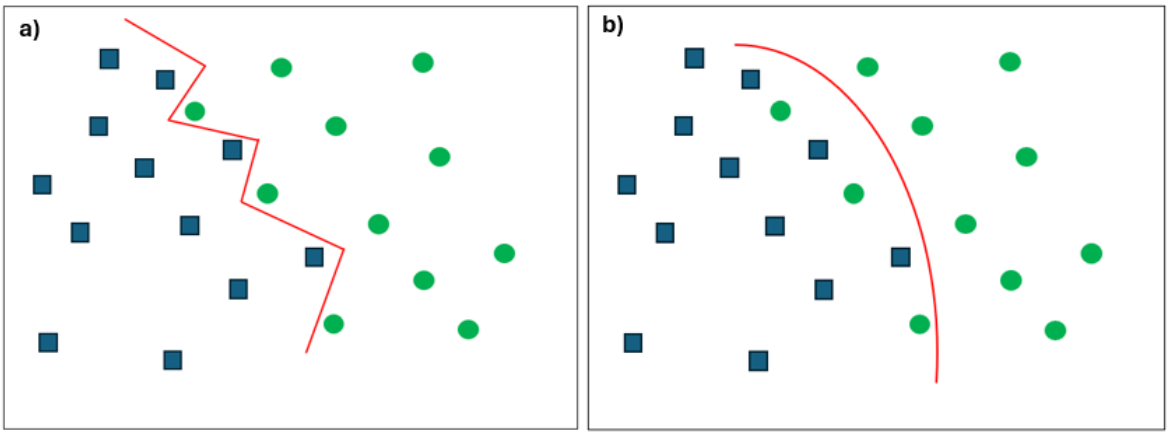



**Fig. 12**. Two types of decision surfaces. Modified from (Mohri et al. 2018, p. 8).

In this study, seven different machine-learning regression models were used to assess the performance of machine-learning methods in predicting hydraulic conductivity. These models will be presented in the subsections below.

### 3.3.1 Linear regression

Linear regression is the simplest form of regression model. It predicts the desired output value as a weighted sum of the input values. The advantage of linear regression models is that the prediction procedure is simple and assumes a linear relationship (Molnar, 2022). The straight line depicted in

linear regression plots shows the best fit that minimizes the residual sum of squares between the data points and predicted values (Scikit-learn, 2024b). The best-fitting line is called a regression line. A schematic graph of a linear regressor is depicted in Fig. 13 below. In the figure, the diagonal line is the regression line, The red, vertical lines from the datapoints represent the distance from the data points to the prediction line – hence, they present the errors in prediction. The best-fitting line aims to minimize the sum of squared errors in prediction (Lane, n.d.).

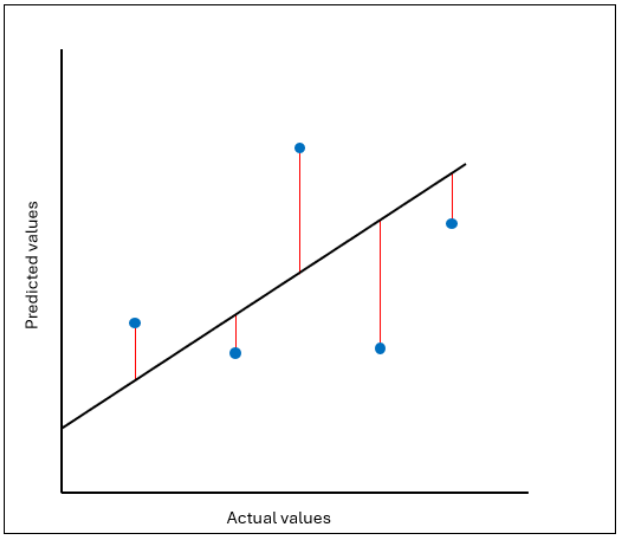



**Fig. 13.** Linear regression model. Modified from (Lane, n.d., p. 464).

Linear regression assumes that the parameters are independent of each other. If the parameters are related to each other or portray similar information, the best-fitting line can become sensitive to errors as it cannot determine the importance of different parameters. This problem is called multicollinearity (Scikit-learn, 2024b).

### 3.3.2 Ridge regression

Ridge regression is an extension of the linear regression model, and it aims to answer some of the problems of the simple linear regression (Hoerl & Kennard, 1970). Linear regression aims to find the best-fitting line between independent parameters and the one dependent variable (output) by minimizing the sum of squared errors in prediction. If the independent parameters are correlated, the model might not be able to predict the importance of parameters correctly. Ridge regression aims to control the importance or weight of some parameters over others; it is designed to keep the weights of different parameters small (Scikit-learn, 2024b). This is done by adding a penalty to the regression coefficients. The penalty makes the regression coefficients smaller and helps with the multicollinearity problem of linear regression (Wu, 2021).

### 3.3.3 Support Vector Regression (SVR)

Support vector regression is an extension of the support vector machine (SVM). The support vector machine is a learning system that operates in a high-dimensional space and can produce predictions based on a subset of support vectors. The support vector regression model depends on a selected portion of the training data, not taking into consideration data points that are too close to the prediction values (Basak et al., 2007). SVR works in a n-dimensional space, where *n* indicates to the

number of independent variables used in the prediction (Vapnik, 1995 as cited in Veloso et al., 2022). The support vector machine can generalize unseen data well and can work with non-linear data (Smola & Schölkopf, 2004). The advantage of SVRs is that they can handle complex data, but if the number of features exceeds the number of samples, overfitting of data can happen (Scikit-learn, 2024d).

### 3.3.4 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is among the simplest machine-learning algorithms. It is a nonparametric method, where a new predicted datapoint is placed closest to a class of existing points. The determination of similarity between the samples is based on distance (Hechenbichler & Schliep, 2004). In Fig. 14, an illustration of the KNN model is depicted.

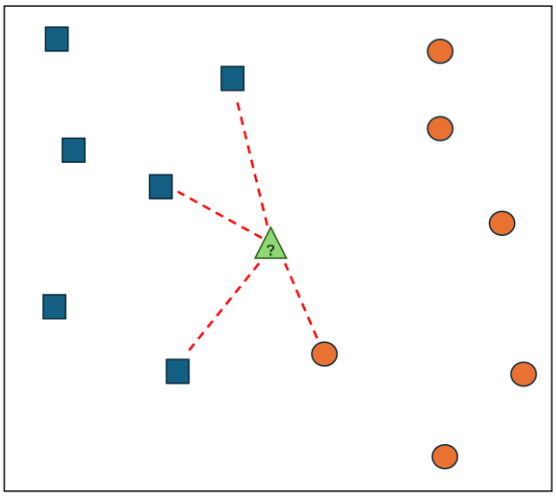



**Fig. 14.** A schematic model of the KNN algorithm with K=4 and two data classes. Modified from Imandoust & Bolandraftar (2013, p. 606).

The number of nearest neighbours is user-defined and the only parameter that is tuned or specified in the training of KNN models (Araya & Ghezzehei, 2019). In Fig. 14, a KNN model with four neighbors is shown with two different classes of data.

### 3.3.5 Decision tree

Decision trees are a hierarchical method used for both classification and regression tasks. They are non-parametric and sequential and are constructed by dividing data into subsets (Alpaydin, 2010; Kotsiantis, 2013). The aim of decision trees is to predict a value of the target variable by learning decision rules from the data parameters (Scikit-learn, 2024d). The subsets are further classified into decisions based on the values of the input data. The decision tree continues to split and grow until all values of the input data are addressed (Fig. 15). The error of the decision tree is calculated by the total number of misclassified data points divided by the number of data points (Kotsiantis, 2013).

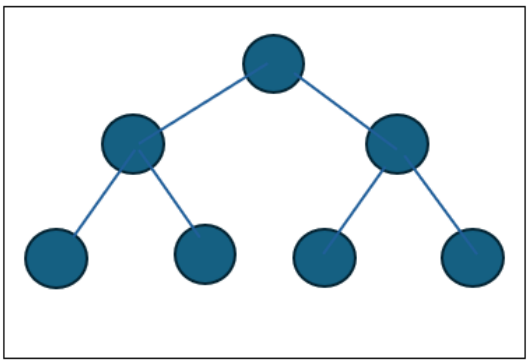


**Fig 15.** A schematic picture of a hypothetical decision tree. Modified from (Alpaydin, 2010, p. 197).

The decision tree algorithm aims to generalize and use pattern recognition in the process; it determines which questions are the best at separating data into different classes. One of the disadvantages is that working with a large dataset - that leads into a large decision tree - may lead to bad generalization performance and overfitting (Kotsiantis, 2013).

### 3.3.6 Random forest

Random forest regression is a method that is derived from decision trees. Random forests are ensemble decision trees where, instead of having one decision tree, there are several decision trees (Breiman, 2001). The model is called random because the subsets of data are chosen randomly to build the tree, and the features that split the data into further branches are also random. Because all decisions are made in a random manner, the algorithm can better generalize and avoid overfitting (Hastie et al., 2009, Wang et al., 2019).

A schematic graph of the Random forest prediction is depicted in Fig. 16.

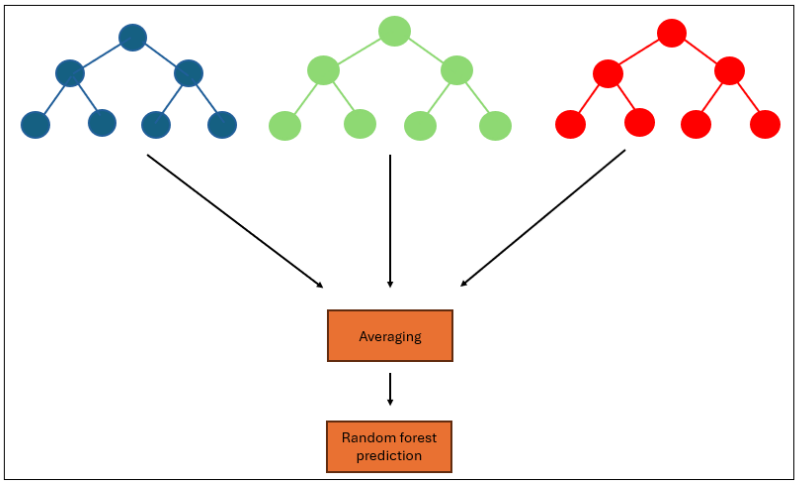



**Fig. 16**. A schematic picture of the Random forest prediction. Modified from (Sahour et al., 2021, p.747).

### 3.3.7 Gradient boosting

Boosting algorithms are ensemble models, just like random forests. But gradient boosting has a different way to approach the model building; it adds new models along the prediction process based on the errors of previous models. The aim is to reduce errors (Natekin & Knoll, 2013). Boosting

regression combines the predictions of many weak decision trees to produce one stronger model (Hastie et al., 2009). The disadvantage of gradient boosting machines is that they require a lot of memory to be able to store all the information, and this can become a problem with large datasets (Natekin & Knoll, 2013).

# 4. RESULTS AND DISCUSSION

## 4.1 Empirical equations

### 4.1.1 Hazen

The hydraulic conductivity in the soil samples were calculated using Hazen's empirical formula introduced in *3. Materials and Methods.* Hazen's values for hydraulic conductivity were calculated using Equation 8. As the actual, laboratory-acquired values were indicated in meters per day (m/d), Hazen's values were converted into the same units.

As presented in *Materials and Methods*, the value for Hazen's empirical coefficient is generally thought to be 100, but can be changed according to soil type. In this study, hydraulic conductivity values were calculated in two ways to investigate how different coefficients change the results. For additional coefficient division, values 40, 100 and 140 were chosen – 40 to indicate to fine soils and 140 to gravelly soils. The soils were divided according to their soil name abbreviations. If *si* or *cl* (silt, clay) were as a prefix in the name (e.g. siSa – silty sand), the coefficient was set to 40, and if *gr* (gravelly) was a prefix in the name, the coefficient was set to 140. The rest of the samples in-between had the value of 100.

In Fig. 17 is illustrated how actual hydraulic conductivity values acquired via the constant-head method correlate with empirical Hazen values. In the graph, the black dotted line is the calibration line. In the ideal case where actual hydraulic conductivity values are the same as calculated values, the linear regression lines would follow the calibration line. It is shown that Hazen's equation overestimates hydraulic conductivity (K). Also depicted in the figure is the difference in values of Hazen's equation with different coefficients; using a coefficient of 100 to each sample versus using a varying coefficient 40, 100 or 140. Hazen's values with $C_H100$ seem to be slightly closer to the actual *K* values than the ones with a varying coefficient.

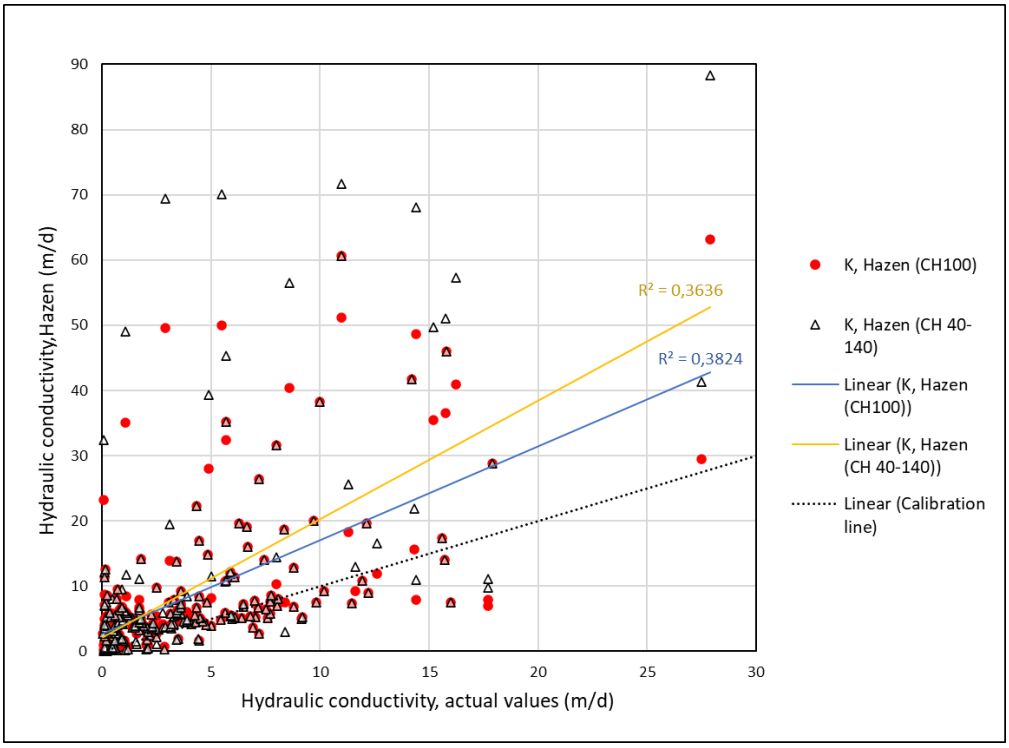



**Fig. 17.** Actual hydraulic conductivity values vs. Hazen values.

Average hydraulic conductivity corresponding to sampling depth is depicted in Fig. 18 a. In laboratory-measured values, hydraulic conductivity decreases with depth consistently. Hazen's values show a similar trend; however, Hazen's equation overestimates hydraulic conductivity, and values between different depths change drastically. In this graph, too, is shown how Hazen's formula with varying coefficients calculates higher values than $C_H100$.

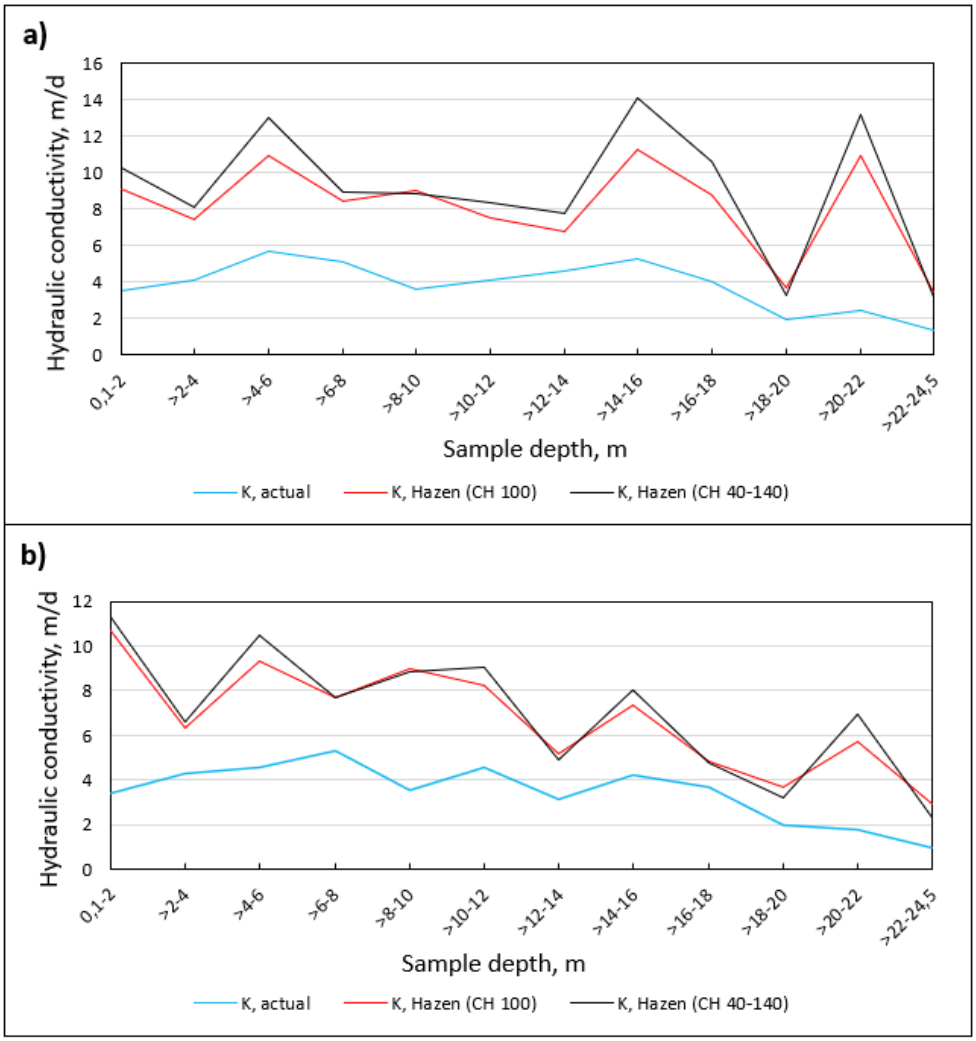



**Fig. 18**. Average hydraulic conductivity values vs. sampling depth (a) and average hydraulic conductivity values of samples with $C_U < 5$ vs. sampling depth (b).

In previous research, it is mentioned that the applicability of Hazen's equation is $D_{10}$ 0.1-3.0 mm and $C_U \leq 5$. All the samples in the database have a $D_{10}$ value that falls in the range, but only some of the samples have a sorting coefficient under 5. If we exclude the samples with $C_U > 5$ as well as the samples that do not have a depth information, we are left with 191 samples. The average hydraulic conductivity of those samples based on depth is depicted in Fig. 18 b.

If we compare Fig. 18 a and 18 b, we can see that the largest overestimations of K values are missing from Fig. 18 b where all samples fall within the range of applicability.

The six most common soil types in the database are uniform sand (SaU), silty sand (siSa), uniformly graded, slightly silty-clayey sand (SaFU), poorly graded, slightly silty-clayey sand (SaFP), medium sand (MSa) and gravelly sand (grSa). Calculated the average hydraulic conductivity values of each of these soil types is depicted in Fig. 19a. The trend is that gravelly sand samples have the highest K values, which is plausible since large-grained sands have more pore spaces between grains

and thus allow more water to pass through. Silty sand has lowest K values. In silty sand's case, the actual average K value is 0.53 m/d. Hazen ($C_H$ 100) yields a value of 0.93 m/d, and with Hazen ($C_H$ 40-140), it is 0.37 m/d. With a varying coefficient, silty sand has the coefficient 40. Calculated with this coefficient, we get a value that underestimates the actual K value.

In Fig. 19 b, only samples that fall within the range of applicability are taken into consideration. Besides gravelly sand (grSa), the graph gives similar values to Fig. 19 a. Samples labelled SaU, SaFU and MSa all had a $C_U \leq 5$, but siSa, SaFP and grSa had samples with Cu > 5, and they were excluded from this graph. The average hydraulic conductivity value in gravelly sand is much higher than in Fig. 19a, but only five samples named gravelly sand (grSa) fulfil the applicability limits and thus does the sampling size in this soil type is not very large.

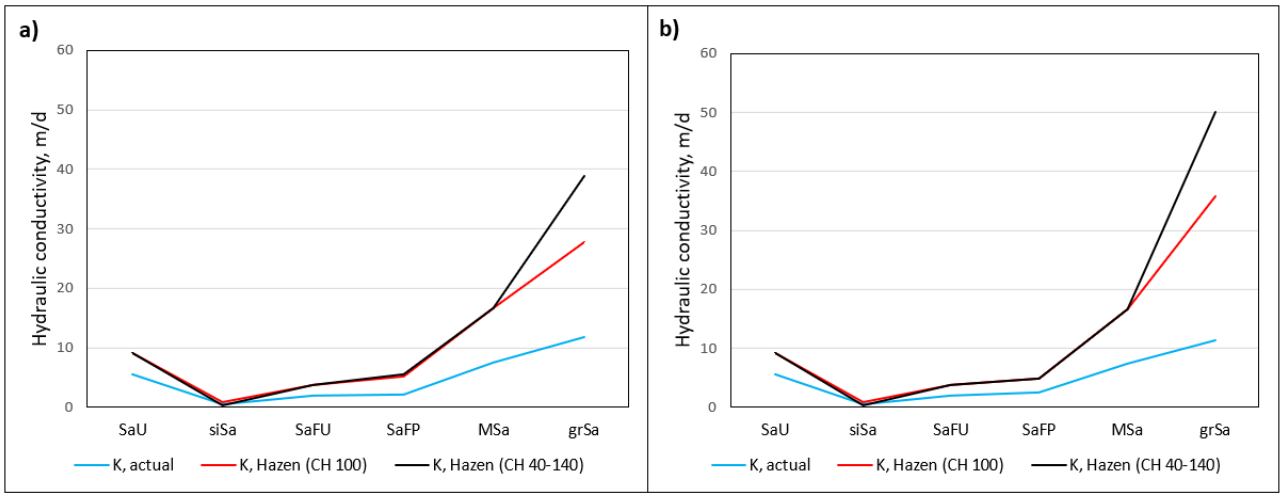



**Fig. 19.** Average hydraulic conductivity vs. soil classification (a) and hydraulic conductivity vs. soil classification with samples $C_U < 5$ (b).

In Fig. 20, soils are divided into a rough division of three soils – fine, medium and coarse – based on their soil type. In the grain size distribution tests performed to the sample soils, some samples have a gradation characteristic in the soil type abbreviation and some do not, as well as a description of the coarseness of the soil – see, eg. siSa (silty sand) vs, siFSa (silty, fine sand) and siMSa (silty medium sand); grSa (gravelly sand), grSaP (poorly graded, gravelly sand) and grSaM (medium graded, gravelly sand). In Fig. 19 a and 19 b above, the soil classifications presented are the six most common, absolute soil name abbreviations in the database. In Fig. 20, soil classifications are bundled together to form three groups. Soils like silt and clay were included in fine soils, as well as soils with an prefix *silty* (e.g. silty fine sand, siFSa). In coarse soils, gravels and soils including the prefix *gravelly* were included (e.g. gravelly sand, grSa).

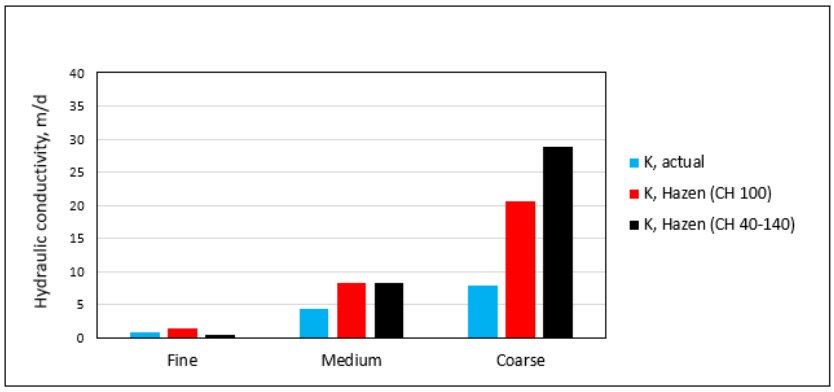



**Fig. 20.** Average hydraulic conductivity vs. soil types.

Using a constant coefficient $C_H100$ seems to yield slightly better results than varying coefficients (Fig. 20). $C_H(100)$ still overestimates all actual K values, but with $C_H(40\text{-}140)$, a coefficient of 40 underestimates hydraulic conductivity values while 140 overestimates them.

### 4.1.2 Slichter

C.S. Slichter's equation for hydraulic conductivity was presented in Chapter *3. Materials and Methods.* Slichter's values for hydraulic conductivity were calculated using Equation 10, and as the actual hydraulic conductivity values were indicated in m/d, Slichter's values were converted into the same units.

The applicability range of Slichter's formula is $0.01\text{mm} \leq D_{10} \leq 5$ mm. All the samples in the used in this study fall into this range. Slichter's values for hydraulic conductivity compared to those acquired via laboratory methods can be seen in Fig. 21. The red dotted line in the graph is the calibration line. The trend is that Slichter's equation yields lower values than laboratory-aqcuired values.

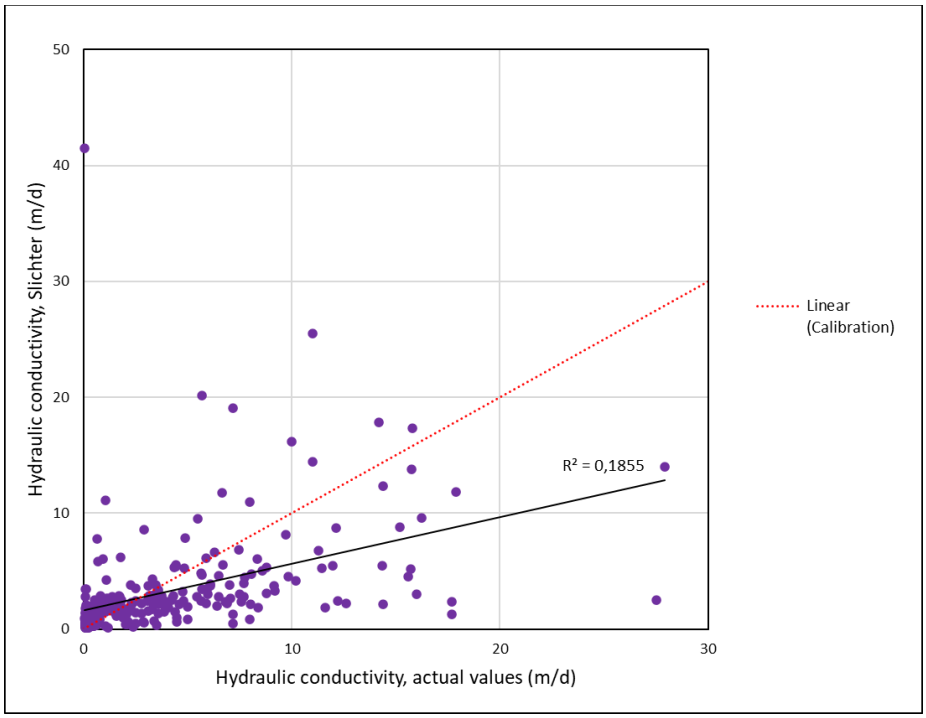



**Fig. 21.** Actual hydraulic conductivity values vs. Slichter's values.

Average hydraulic conductivity values depending on depth are presented in Fig. 22. It can be seen that Slichter's equation yields lower K values than actual, laboratory tested K, except for between depths >8-10 m and >20-22 m.

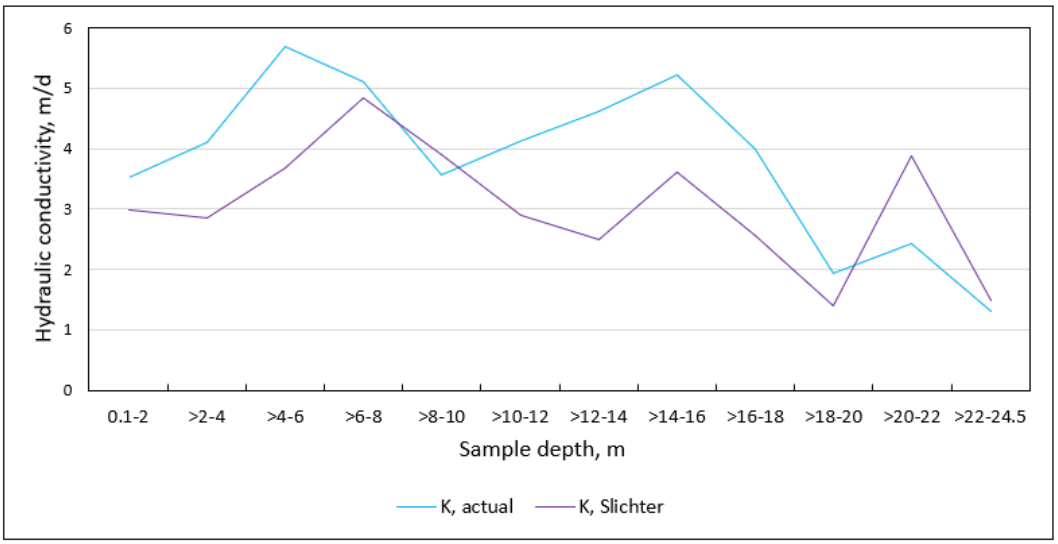



**Fig. 22.** Average hydraulic conductivity vs. sampling depth.

While comparing the average hydraulic conductivity values in the six most common soil names, the shape of the graph is fairly similar in both Slichter and actual K values (Fig. 23). The biggest differences are in coarser soil types like gravelly sand (grSa) and medium sand (MSa). Even with finer soil types, Slichter's equation gives smaller values – e.g. average siSa (silty sand) value is 0.52 m/d according to the constant-head laboratory test, and 0.46 m/d via Slichter's method, and SaFU (slightly silty-clayey, uniform sand) values are 1.90 m/d and 1.88 m/d, respectively.

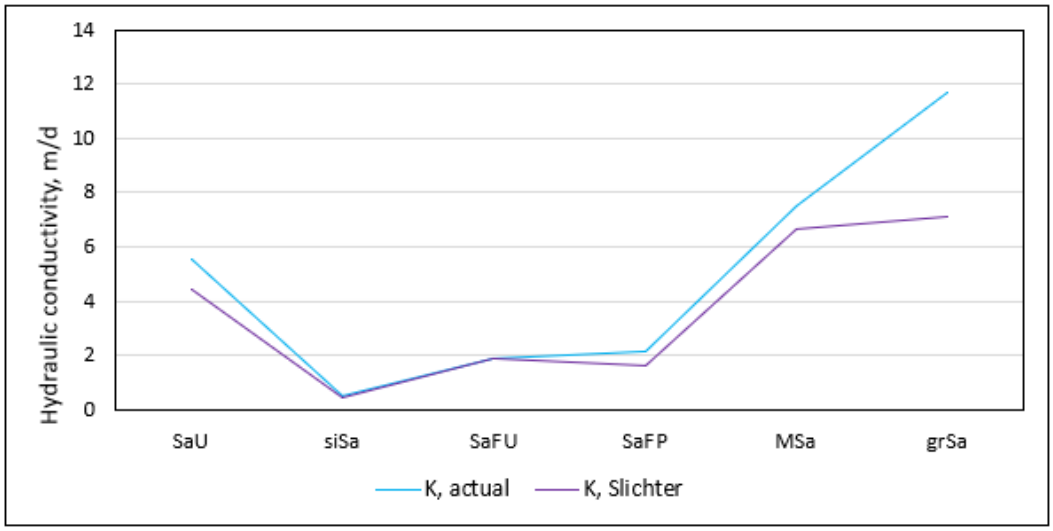



**Fig. 23.** Average hydraulic conductivity vs. soil classifications.

Similar trend can be seen in Fig. 24 (below). All samples have been divided into three types – fine, medium and coarse soils. Their average values are depicted in the graph. The actual K results are higher than with Slichter's method.

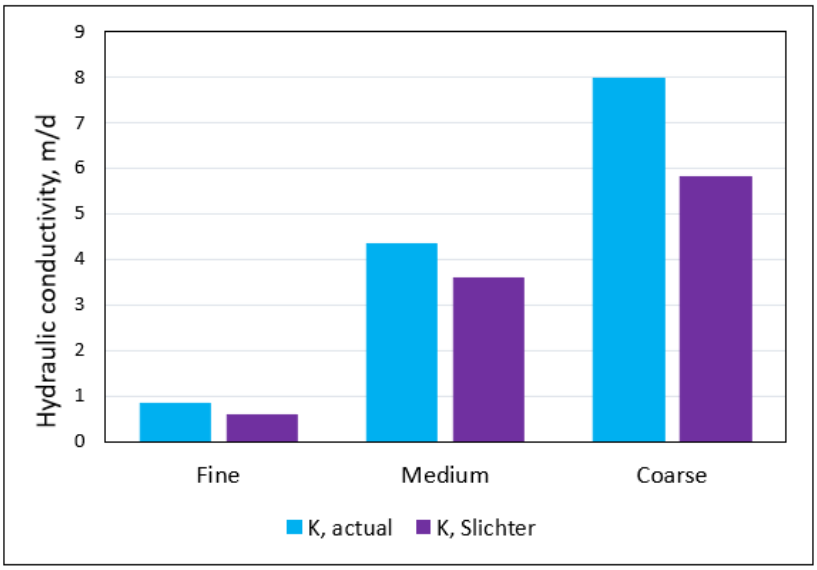



**Fig. 24.** Hydraulic conductivity vs. soil types.

### 4.1.3 USBR

The United States Bureau of Reclamation (USBR) equation was presented in Chapter *3. Materials and Methods*. USBR‘s values for hydraulic conductivity were calculated using Equation 12, and as actual hydraulic conductivity values were presented in m/d, USBR‘s values were converted into the same units. The range of applicability for the USBR equation is medium sands with $C_U < 5$.

USBR’s values plotted against actual hydraulic conductivity values are presented in Fig. 25. The red dotted line in the graph is the calibration line. It can be seen that USBR‘s equation slightly overestimates hydraulic conductivity.

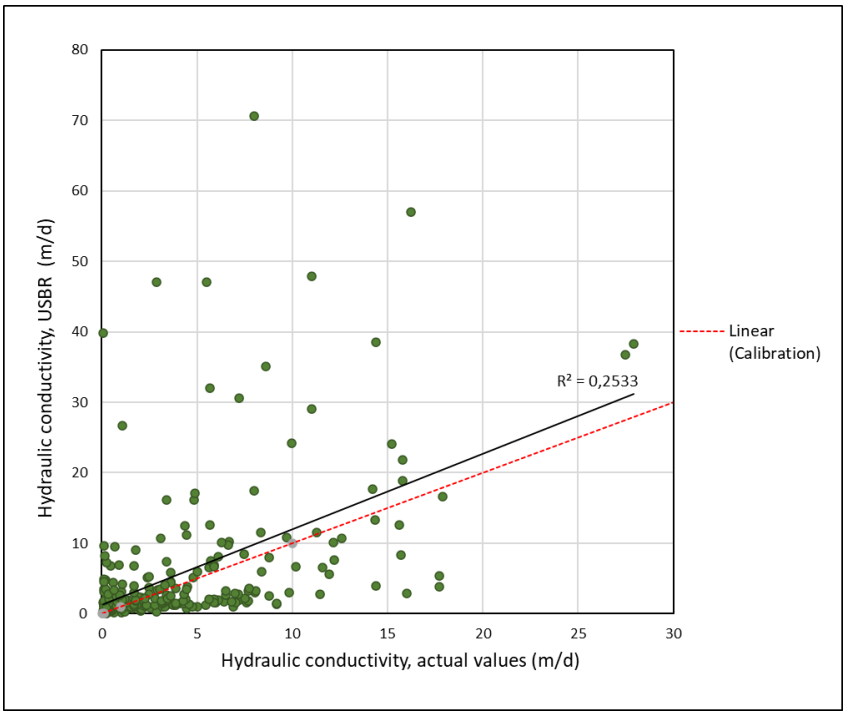



**Fig. 25.** Actual hydraulic conductivity values vs. USBR values.

Average hydraulic conductivity values plotted against depth is illustrated in Fig. 26 a and 26 b. In Fig. 26 a, all samples were used and in Fig. 26 b, the range of applicability was taken into consideration, meaning that only medium sands (0.63-0.2 mm) with $C_U$<5 were included. Only 78 samples could be taken into consideration with these restrictions. In both figures, USBR's equations give the largest overestimations of hydraulic conductivity in the samples acquired from deeper depths. However, one of the reasons for such differences might be the small number of samples taken from deeper depths as opposed to those taken from 0-12 meters.

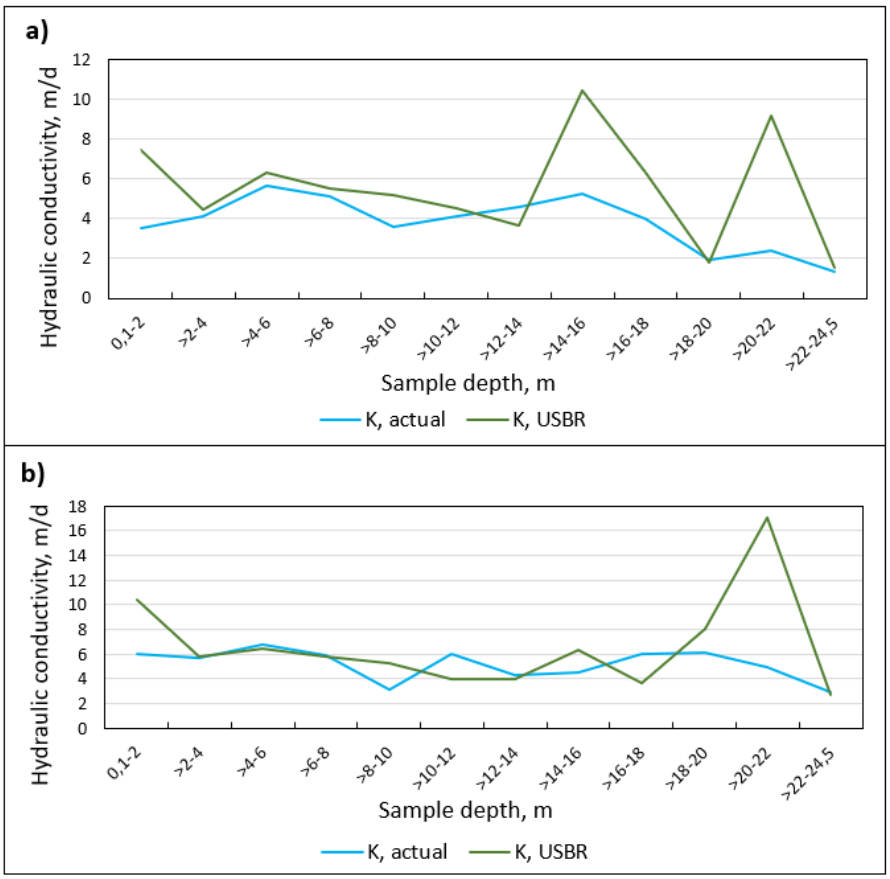



**Fig. 26.** Hydraulic conductivity vs depth (a) and hydraulic conductivity vs. depth with samples $C_U$ < 5 (b).

Average hydraulic conductivity plotted against six most common soil classifications is presented in Fig. 27. In this graph, all samples were considered and no graph with the range of applicability was made since it would only include medium sands. USBR's formula underestimates hydraulic conductivity in silty sands (siSa) and slightly silty-clayey, uniform sands (SaFU), and uniform sands (SaU), but overestimates K in slightly silty-clayey, poorly graded sands (SaFP), medium sands (MSa) and gravelly sands (grSa).

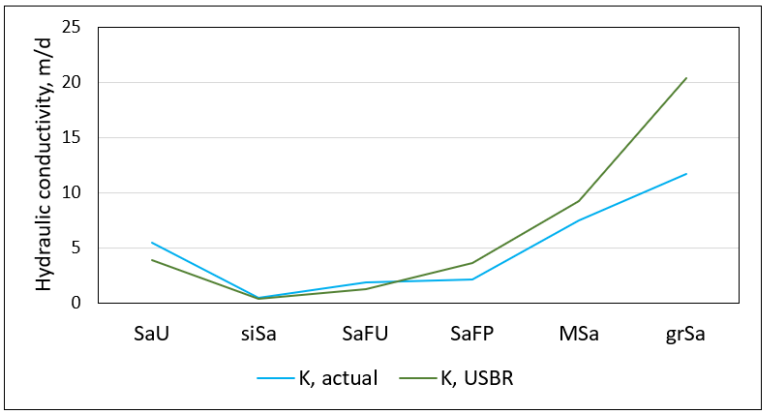



**Fig. 27.** Average hydraulic conductivity vs. soil classification.

In Fig. 28, average hydraulic conductivity values vs. soil types (fine, medium, and coarse soils) are illustrated. USBR's equation yields much higher hydraulic conductivity values in coarse soils than in medium and fine soils. In fact, in medium and fine soils, USBR's equation slightly underestimates hydraulic conductivity as opposed to the actual values, but are clearly closer to actual values than with coarse soils.

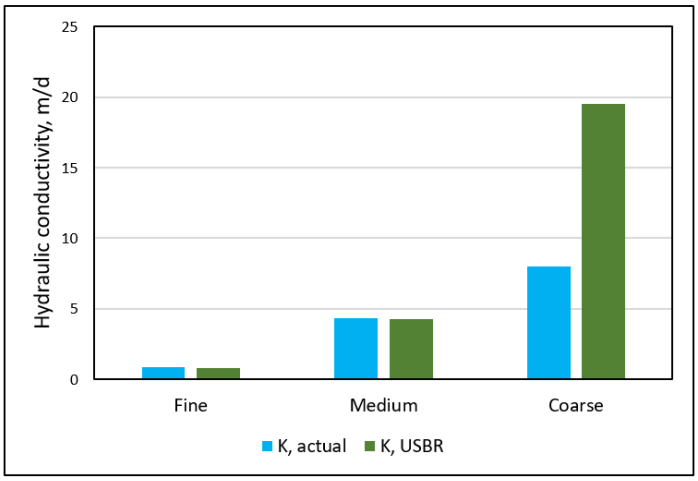



**Fig. 28.** Average hydraulic conductivity vs. soil types.

### 4.1.4 Comparison of formulas

Hazen's hydraulic conductivity values were inspected using a fixed coefficient of 100, and a varying coefficient 40, 100 or 140 depending on the soil type. Using $C_H(100)$, the hydraulic conductivity values were slightly closer to the actual K values than with a varying coefficient. Slichter's equation yielded almost consistently lower hydraulic conductivity values than the laboratory-tested values. USBR's formula both underestimated and overestimated hydraulic conductivity depending on soil type.

Hazen's, Slichter's and USBR's hydraulic conductivity values plotted against actual values are depicted in a calibration plot below (Fig. 28). The black dotted line in the graph is the calibration line. Also R-squared ($R^2$) values are shown. R-squared value shows the variance in the data and indicates how well the data fits in the regression model. It ranges from 0 to 1, with 1 being the perfect fit (Fahrmeir et al., 2013). Looking at the R-squared values of the empirical formulas, we can see that Hazen's equation gives the highest value (0.3824) and Slichter's equation the lowest value (0.1855),

meaning that Hazen's equation is – from the three investigated formulas – the best fit for the diagonal trendline where actual and calculated hydraulic conductivity values are the same (Table 5).

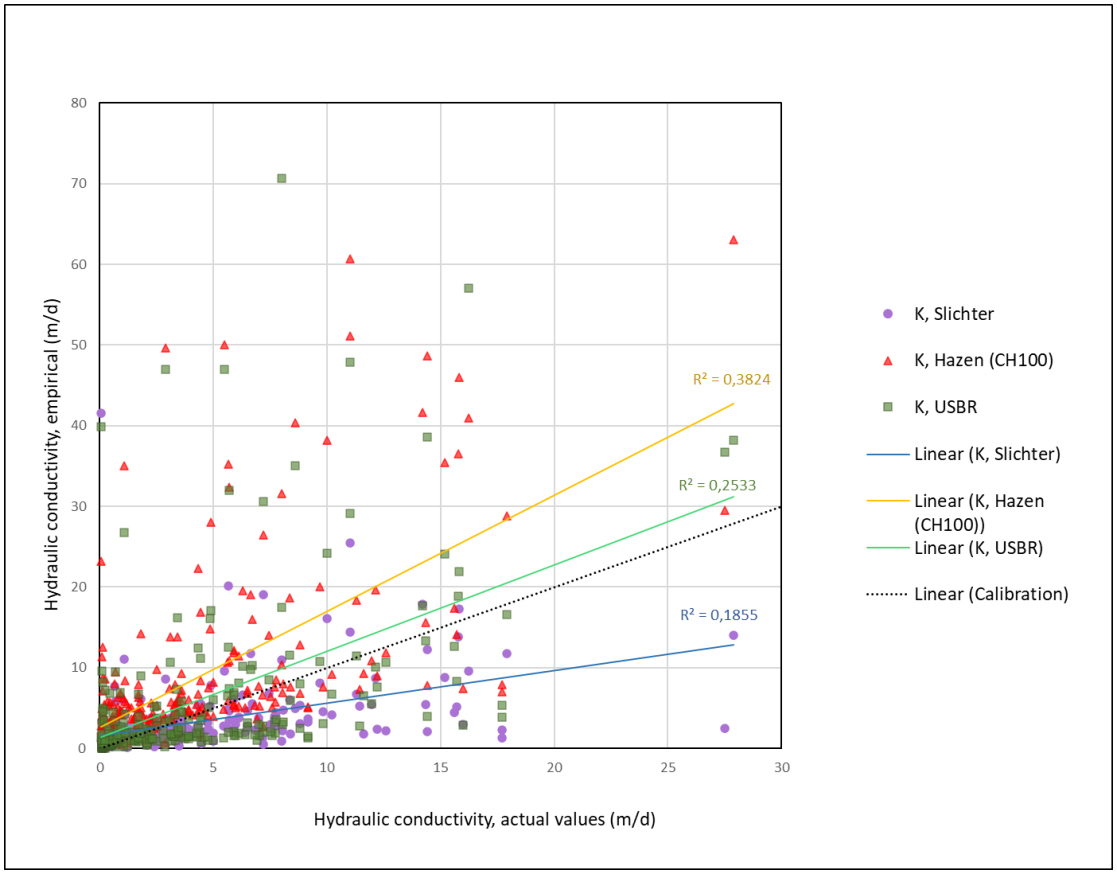



**Fig. 28** Actual hydraulic conductivity values plotted against empirical ones.

**Table 5**. R-squared values of calibration plots.

| **Method** | **$R^2$, linear** |
|---|---|
| Hazen ($C_H$ 100) | 0.3824 |
| Slichter | 0.1855 |
| USBR | 0.2533 |

In hydraulic conductivity tests made for Quaternary gravels in Northeast Slovenia, it was found that USBR's equation overestimates K values the most. The researchers said that USBR's values are not in line with other empirical methods and, thus, do not compare well. Hazen's method yielded the highest values, and Slichter's method yielded the lowest values of hydraulic conductivity. However, in their study, the samples were Quaternary gravels, and this means that only Slichter's equation offers an applicability range where the gravels fall into, which might have affected the results. It was also found that Slicther's empirical equation gave two magnitudes lower values than those acquired with field pumping tests (Pucko & Verbovšek, 2015). The results in this thesis match the results by Pucko and Verbovšek; Hazen had the highest estimated hydraulic conductivity values and Slichter had the lowest.

Different empirical hydraulic conductivity methods were tested on sand samples in another study (Odong, 2008). In this study, the overall results showed that USBR and Slichter gave, in all cases, lower values than other methods, and these two methods were seen as inaccurate. Hazen's equation was the second-best most accurate of seven empirical methods Odong (2008). Other studies

have also reported that USBR‘s and Slichter‘s formulas underestimate actual hydraulic conductivity (Ann et al., 2022; Cheng & Chen, 2007; Onwe et al., 2016). The observations of these studies of Slichter‘s equation underestimating hydraulic conductivity is in line with this thesis‘ results, except for USBR that overestimated K in coarser soils.

## 4.2 Machine learning

In this part of the study, seven different regression models were used to determine hydraulic conductivity. The performance of these models was tracked by investigating their mean absolute error (MAE), standard deviation (SD) and R-squared ($R^2$) values. Mean absolute error (MAE) shows the average error in the model in units that relate to the variable that is being studied. Mean absolute error is calculated by summing the absolute values of errors before dividing the total error by the number of data points included (Willmott & Matsuura, 2005). The R-squared value is also called the coefficient of determination. Value of 1 makes a perfect fit between the predicted and actual values. $R^2$ is calculated by subtracting the sum of squared errors (Alpaydin, 2010).

To acquire the best results, four important points needed to be inspected:

- $R^2$ value of the test set should be high.
- $R^2$ value of the trained set should be higher than the test set.
- $R^2$ standard deviation value should be as small as possible.
- MAE (mean absolute error) should be as small as possible.

Multiple tests were performed by changing the parameters included in the testing and observing their results. The available parameters were water content, soil density, saturation degree, void ratio, $D_{xx}$ values from D10 to D90, and six grain size classes: <.0.06 mm, 0.06-0.2 mm, 0.2-0.62 mm, 0.6-2 mm, 2-4.75 mm, >4.75 mm. The entries that had hydraulic conductivity testing done via the ISO standard were included in the machine learning testing (208 samples).

After numerous tests, it was found that the best results are obtained by focusing on the grain size distribution and $D_{xx}$ values. In general, using the distribution of smaller grain sizes (grain size classes <0.06 mm, 0.06-0.2 mm and 0.2-0.62 mm) yielded better results than coarser soils. Using <0.06 mm and 0.06-0.2 mm categories together gave better prediction results than using either category separately. However, including the next grain size class (0.2-0.62 mm) in the prediction gave worse results – hence, using <0.06 mm and 0.06-0.2 mm together was regarded the best grain size category to take into consideration. These findings back up the results of previous research where the importance of smaller grains in the hydraulic abilities of a soil matrix has been studied.

By using these two best-working soil classes and pairing them with $D_{xx}$ values, in almost every $D_{xx}$ category from $D_{10}$ to $D_{90}$, there was at least one model that had high $R^2$ value for the test set, higher $R^2$ value for the train set, and a low $R^2$ standard deviation. Using several $D_{xx}$ values together gave better results than using only one.

Highest correlation and prediction in the testing was found by including the following parameters: <0.06 mm, 0.06-0.2 mm, water content, $D_{60}$ and $D_{70}$. By using these parameters, the dataset was divided into a test set of 52 samples and a train set of 156 samples. The numerical values of this test can be found in Table 6 below. The highest prediction was obtained with Random forest, followed by Gradient boosting; both of these models gave a $R^2$ value of over 0.40 for the test set along with higher train set values and a low standard deviation between the two. Looking at Table 6,

we can see that while ridge and linear regression models gave relatively high correlation for the test set, linear models couldn‘t predict new datapoints correctly.

**Table 6.** Prediction results using parameters: 0.06 mm, 0.06-0.2mm, water content, $D_{60}$, $D_{70}$.

| Model | MAE (test set) | MAE (train set) | MAE (SD) | $R^2$ (test set) | $R^2$ (train set) | $R^2$ (SD) |
|---|---|---|---|---|---|---|
| Random forest | 2.28 | 2.43 | 0.25 | 0.47 | 0.51 | 0.10 |
| Gradient boosting | 2.32 | 2.46 | 0.23 | 0.44 | 0.48 | 0.08 |
| K-neighbors | 3.04 | 2.70 | 0.28 | 0.11 | 0.35 | 0.10 |
| Ridge | 2.42 | 3.10 | 0.13 | 0.39 | 0.15 | 0.29 |
| Linear | 2.40 | 3.13 | 0.11 | 0.38 | 0.14 | 0.29 |
| SVR | 2.45 | 2.70 | 0.48 | 0.18 | 0.19 | 0.10 |
| Decision tree | 2.67 | 2.97 | 0.46 | 0.15 | 0.09 | 0.19 |

Visual graphs for Random forest and Gradient boosting regression models for the same test can be found in Fig. 29 a and 29 b. In both models, the train set follows the calibration line relatively well. Visual graphs for all of the models performed in this test can be found in Appendix 1.

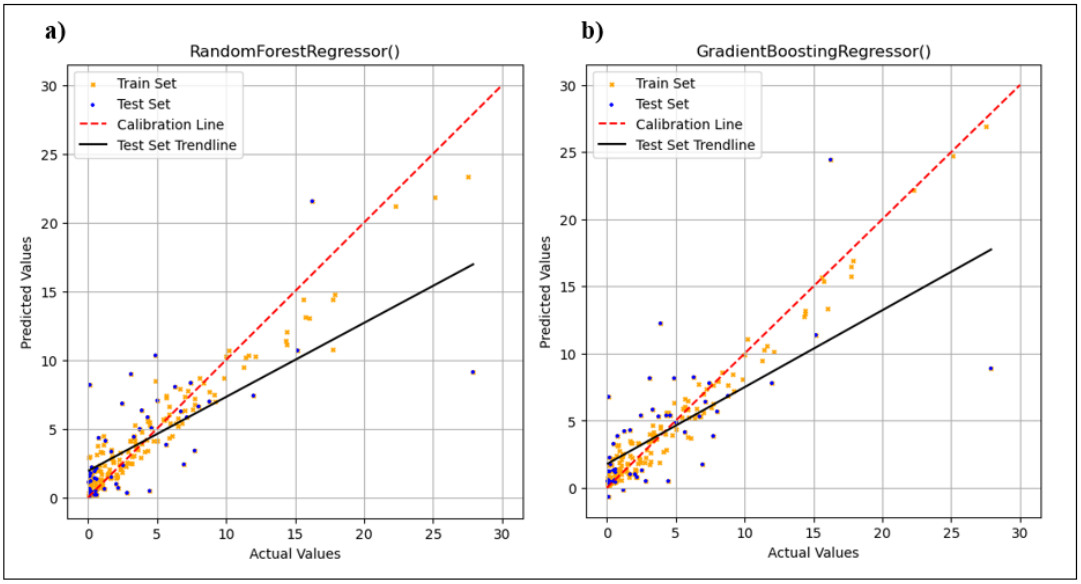



**Fig 29.** Random forest (a) and Gradient boosting (b) regression models of the best-fitting test.

A recent research studied six machine-learning models to predict staurated hydraulic conductivity, and found that Random forest and Gradient boosting models worked the best (Adjuik et al., 2023). A research studying supervised machine learning methods to model hydraulic conductivity and 3D hydrostratigraphy also found that Random forest algorithm gave the best results (Tilahun & Korus, 2023), and another research compared different machine learning methods to an empirical model to find the best-performing models for hydraulic conductivity, and found that Random forest performed the best (Singh et al., 2021). These studies are in line with this thesis' results: Random forest yielded the highest correlation results along with Gradient boosting algorithm in most tests. As explained in Chapter *3.3 Machine learning methods*, random forest and gradient boosting algorithms apply the mechanisms of a regular decision tree, but are developed further; random forests gather several decision trees and average their results (Wang et al. 2019), and gradient boosting creates new models based on the errors of previous models (Natekin & Knoll, 2013). The

abilities of these models to reduce errors in the prediction process is likely the reason why they work better than the other models in this study, as well. Decision tree was the worst at predicting new data in the test introduced in Table 6 (test set $R^2$ 0.15, train set 0.09). Looking at Fig. 30, it can be seen that the trained datapoints follow the calibration line perfectly. This means that while the model can find the relationship between predicted and actual values, it cannot predict the real, underlining function between the two and is subjected to overfitting – hence the low $R^2$ value.

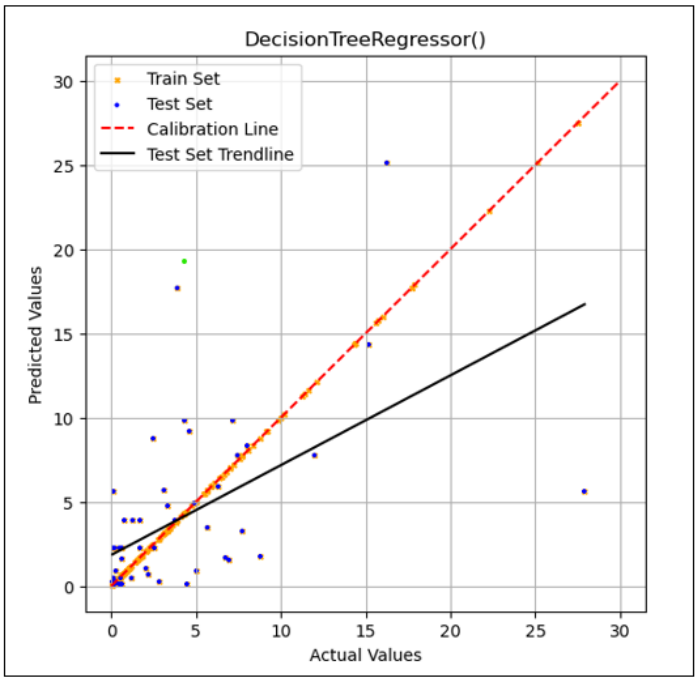



**Fig. 30.** Decision tree regression model of the best-fitting test.

The second-best performance was achieved with the same parameters as in the abovementioned test, only water content parameter was excluded. The results of the test can be seen below in Table 7. It is similar to the test with the highest values: Random forest and Gradient boosting algorithms were the best match. K-Neighbors could also predict hydraulic conductivity, but did not reach as high $R^2$ values as the other two best models.

**Table 7.** Prediction results using parameters: <0.06 mm, 0.06-0.2mm, $D_{60}$, $D_{70}$.

| Model | MAE (test set) | MAE (train set) | MAE (SD) | $R^2$ (test set) | $R^2$ (train set) | $R^2$ (SD) |
|---|---|---|---|---|---|---|
| Random forest | 2.34 | 2.38 | 0.32 | 0.45 | 0.48 | 0.10 |
| Gradient boosting | 2.34 | 2.51 | 0.26 | 0.36 | 0.44 | 0.11 |
| K-Neighbors | 2.72 | 2.82 | 0.33 | 0.24 | 0.31 | 0.10 |
| Linear | 2.44 | 3.14 | 0.12 | 0.38 | 0.19 | 0.22 |
| Ridge | 2.45 | 3.11 | 0.13 | 0.38 | 0.20 | 0.22 |
| SVR | 2.57 | 2.77 | 0.42 | 0.14 | 0.16 | 0.08 |
| Decision tree | 2.72 | 3.02 | 0.66 | 0.29 | 0.05 | 0.19 |

In empirical equations, like the ones presented in this thesis, often effective diameter $D_{10}$ (or in USBR's case, $D_{20}$) is used in the prediction because it has been proven in empirical tests that the diameter with 10% grains passing tells about the hydraulic abilities of soil the best. This is because it is thought that the finer grains in the soil matrix control the available void spaces and thus, the

available progression route for water (Rehman et al., 2022). In this study, using $D_{10}$ as a parameter along with grain size information did give relatively good $R^2$ values for the test sets, but not for the train sets. For example, when examining only the grain size category <0.06 mm, using $D_{10}$ as a parameter gave $R^2$ 0.57 for the test set, and 0.31 for the train set. Similar trend could be detected in all grain size categories, except with grain size class >4.75 mm ($R^2$ test set 0.27, train set 0.51) and <0.06 mm + 0.06-0.2 mm ($R^2$ test set 0.31, train set 0.43), where the train set values were higher.

One recent study proposed a new mathematical prediction model for hydraulic conductivity, where a new gradation coefficient utilizing $D_{10}$, $D_{30}$, $D_{50}$ and $D_{60}$ as well as void ratio is taken into account. The study found the new equation to work better than existing empirical equations (Arshad et al., 2020). In this thesis' case, using larger $D_{xx}$ values gave better prediction in every grain size category, and using several $D_{xx}$ values at once were better at predicting hydraulic conductivity than using only one. Using all $D_{xx}$ values (from $D_{10}$ to $D_{90}$) along with grain size categories did give relatively high prediction values, too, however they did not exceed the test with the highest values presented earlier in Table 6. These findings highlight the fact that single grain diameter values might necessarily not depict the whole soil and several parameters are needed to produce more accurate results of the whole soil matrix.

Using other soil parameters than $D_{xx}$ and grain size classes (void ratio, saturation degree, bulk density, $C_U$ and $C_C$) did not predict well, despite the fact that, for instance, $C_U$ and $C_C$ tell about the gradation of the soil, and void ratio has been found to be an important soil parameter when assessing hydraulic conductivity. The parameters yielded relatively good results with the test set values, but did not perform well with the train set and could not predict new samples. One possible reason for this could be the size of the data and more notably, the number of samples available to use. 208 samples were included in the machine-learning part of the study. In the best-performing machine-learning model introduced earlier, 52 samples were used as a test set, and 156 new samples were trained. Adding several parameters might make the test sets even smaller, because not all samples in the database included every soil parameter. For example, using grain size information, $D_{10}$ and void ratio, the test set size drops to 40 samples and train set to 118 samples. If the database was bigger and, thus, there were more test set values to use, the algorithms could perhaps better learn the underlying function. Grain size distribution and $D_{xx}$ values were the only parameters that were available in every sample. The relationship between hydraulic conductivity and other parameters than grain size distribution and $D_{xx}$ might be more complex and need a larger dataset to find the correlation.

### 4.3 Comparison of results

The hydraulic conductivity determination results from both empirical equations and machine-learning are depicted below in Table 8. Their best achieved results are presented by their respective $R^2$ values. From the empirical equations, Hazen's formula performed the best. Machine learning methods exceeded the $R^2$ values of empirical equations in various different tests. The best achieved correlation was found when using grain size categories 0.06 mm and 0.06-0.2 mm, water content, $D_{60}$ and $D_{70}$. The best-performing model was Random forest.

**Table 8**. R-squared values of calibration plots.

| **Method** | **Model** | **$R^2$** |
|---|---|---|
| Machine learning | Random forest | 0.47 (test set)<br>0.51 (train set) |
| Hazen ($C_H$ 100) | Linear | 0.38 |
| USBR | Linear | 0.25 |
| Slichter | Linear | 0.19 |

The regression models of the methods presented in Table 8 are presented in Fig 31.

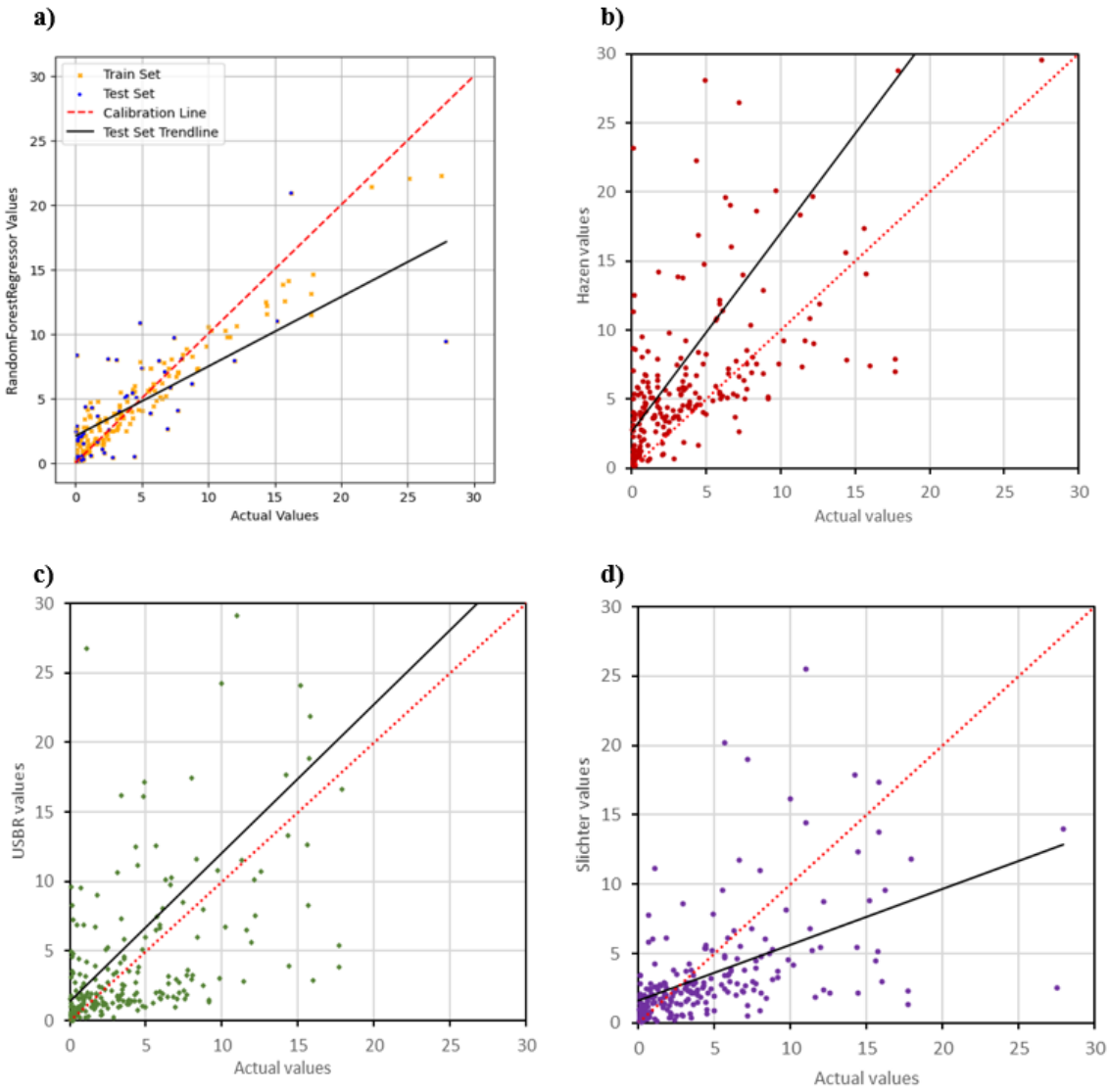



**Fig. 31.** Best-performing machine-learning model (a), Hazen‘s method (b), USBR‘s method (c), Slicther‘s method (d).

From Fig. 31 a, it can be seen that in the Random forest regression model, new trained values follow the calibration line fairly well, which means the algorithm can predict new samples with success. What is noteworthy is that the regression line in the machine learning model is for the test set, not for the train set. Hazen‘s equation (Fig. 31 b) overestimates actual K values the most, but still yields the best correlation from the three empirical equations. USBR‘s (Fig. 31 c) equation also slightly overestimates values, and Slichter‘s formula (Fig. 31 d) underestimates them.

Overall, the results acquired from empirical formulas and machine-learning testing narrate the fact that regression models performed by machine learning can determine hydraulic conductivity better than Hazen's, USBR's or Slichter's empirical formulas. This indicates that machine-learning methods can better adapt to the heterogeneity of soils and a varying soil matrix. The disadvantage of empirical formulas is that they should be used within their ranges of applicability and not all formulas fit all soils, and the formula should fit the soil in question (Chapuis, 2012).

The values of coefficient of determination ($R^2$) in the studies done in this thesis were not very high; even with the best-fitting prediction model, only a value of 0.51 was reached. There are multiple possible reasons for this, most notably the size of the database. The number of samples used in the machine learning part of this thesis is 208 samples. The fewer there are data points that can be used for prediction, the bigger the probability for error.

A recent research studied different predictive methods for saturated hydraulic conductivity of soils, assessed the quality of laboratory tests, and explained the most important mistakes that can happen during testing that affects the results of parameters and thus the prediction of K (Chapuis, 2012). In all fields where numerical information is handled and used in studies, the origin of said information should be critically examined. Especially in large databases where data has been gathered from numerous different tests, the validity of those tests should be considered. In the case of this thesis, for example the $D_{xx}$ values that were used are purely mathematical estimations aqcuired from interpolating cumulative grain size percentages instead of laboratory testing and thus might be a little different from the actual grain diameters.

Sometimes using soft computing tools (including machine learning) are not as transparent as using empirical formulas; sometimes machine learning algorithms work in a way that is difficult to transform into understandable, representative rules. The transparency of the function behind machine learning models is something that needs to be considered in the future of machine learning prediction (Naeej et al., 2017). After all, using machine learning modelling in e.g. hydraulic conductivity determination is to help get accurate information based on attainable soil parameters in a way that is more flexible than traditional methods, but studying the relationships between different parameters shouldn't be overshadowed by the more complex deduction processes of machine learning.

In the future, the database of Lithuanian soil samples can be further developed by adding new soil samples to create a wider database that can be used not only for studying hydraulic conductivity but for multiple hydrogeological and engineering practises and widen the range of possibilities of study based on e.g. regionality, depth and deposition processes.

## CONCLUSIONS

**1.** From the three empirical equations used in this study (Hazen, Slichter, USBR), Hazen's equation performed the best. However, the overall values of $R^2$ still remained relatively low ($R^2$ <0.4). The disadvantage of most empirical equations is their limited applicability range to certain soil types and grain sizes.

**2.** Machine learning methods performed better than empirical equations in both the test set and train set. Random forest and Gradient boosting algorithms performed the best from the seven models used. The highest achieved $R^2$ values for the test set and train set were 0.47 and 0.51, respectively.

**3.** Grain size information of fine soils and grain size diameters ($D_{xx}$ values) proved out to be the best parameters in the prediction of K in the case of this database. Other available parameters could, in many cases, find relatively good $R^2$ results for the test set values, but could not successfully produce new predictions.

**4.** Overall, the low coefficients of determination ($R^2$) might be due to the small number of samples available (246 samples inspected in the empirical equations, 208 inspected in the machine learning part). It is recommended to further expand the database of Lithuanian soil samples to establish a wider resource for future investigations of soil permeability in Lithuania.

## REFERENCES


Adjuik, T. A., Nokes, S. E., Montross, M. D., Sama, M. P., & Wendroth, O. (2023). Predictor selection and machine learning regression methods to predict saturated hydraulic conductivity from a large public soil database. *Journal of the ASABE*, *66*(2), 285–296. https://doi.org/10.13031/ja.15068

Alpaydin, E. (2010). *Introduction to Machine Learning* (2nd ed.). The MIT Press.

Alyamani, M. S., & Şen, Z. (1993). Determination of Hydraulic Conductivity from Complete Grain-Size Distribution Curves. *Groundwater*, *31*(4), 551–555. https://doi.org/10.1111/j.1745-6584.1993.tb00587.x

Ann, V., Romaní, A. M., & Butturini, A. (2022). Estimating the hydraulic conductivity in river unconsolidated sediments. A critical analysis of several grain-size empirical approaches. *Serie Correlacion Geologica*, *38*(1), 15–25. https://doi.org/10.5281/zenodo.7194569

Araya, S. N., & Ghezzehei, T. A. (2019). Using Machine Learning for Prediction of Saturated Hydraulic Conductivity and Its Sensitivity to Soil Structural Perturbations. *Water Resources Research*, *55*(7), 5715–5737. https://doi.org/10.1029/2018WR024357

Arshad, M., Nazir, M. S., & O'Kelly, B. C. (2020). Evolution of hydraulic conductivity models for sandy soils. *Proceedings of the Institution of Civil Engineers: Geotechnical Engineering*, *173*(2), 97–114. https://doi.org/10.1680/jgeen.18.00062

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support Vector Regression. *Neural Information Processing-Letters and Reviews*, *11*(10), 203–224.

Bear, J. (1972). *Dynamics of Fluids in Porous Media*. American Elsevier Publishing Company, Inc.

Bičkauskas, G., Brazauskas, A., Kleišmantas, A., & Motuza, G. (2011). *Bendrosios Geologijos pratybos [General Geology Exercise Book]*. Vilniaus universiteto leidykla.

Boadu, F. K. (2000). Hydraulic Conductivity of Soils from Grain-Size Distribution: New Models. *Journal of Geotechnical and Geoenvironmental Engineering*, *126*(8), 739–746. https://doi.org/10.1061/(asce)1090-0241(2000)126:8(739)

Breiman, L. (2001). Random Forests. *Machine Learning 45*, 5-32. http://dx.doi.org/10.1023/A:1010933404324

Cabalar, A.F. & Akbulut, N. (2016). Effects of the particle shape and size of sands on the hydraulic conductivity. *Acta Geotechnica Slovenica*, 13(2), 83-93.

Carrier, W. D. I. (2003). Goodbye, Hazen; Hello, Kozeny-Carman. *Journal of Geotechnical and Geoenvironmental Engineering*, *129*(11), 1054–1056. https://doi.org/10.1061/ASCE1090-02412003129:111054

Carter, M., & Bentley, S. P. (1991). *Soil Properties And Their Correlations*. John Wiley & Sons, Ltd.

Chakraborty, D., Chakraborty, A., Santra, P., Tomar, R. K., Garg, R. N., Sahoo, R. N., Choudhury, S. G., Bhavanarayana, M., & Kalra, N. (2006). Prediction of hydraulic conductivity of soils from particle-size distribution. *Current Science*, *90*(11), 1526–1531.

Chandel, A., Sharma, S., & Shankar, V. (2022). Prediction of hydraulic conductivity of porous media using a statistical grain-size model. *Water Supply*, *22*(4), 4176–4192. https://doi.org/10.2166/ws.2022.043

Chapuis, R. P. (2012). Predicting the saturated hydraulic conductivity of soils: a review. *Bulletin of Engineering Geology and the Environment*, *71*(3), 401–434. https://doi.org/10.1007/s10064-012-0418-7

Cheng, C., & Chen, X. (2007). Evaluation of methods for determination of hydraulic properties in an aquifer-aquitard system hydrologically connected to a river. *Hydrogeology Journal*, *15*(4), 669–678. https://doi.org/10.1007/s10040-006-0135-z

Craig, R. F. (2004). *Craig's Soil Mechanics* (7th Edition). Spon Press.

Dolzyk, K., & Chmielewska, I. (2014). Predicting the Coefficient of Permeability of Non-Plastic Soils. *Soil Mechanics and Foundation Engineering*, *51*(5), 213–218. https://doi.org/10.1007/s11204-014-9279-3

Fahrmeir, L., Kneib, T., Lang, S. & Marx, B. (2013*). Regression. Models, Methods and Applications*. Springer.

Fitts, C. R. (2002). *Groundwater Science*. Academic Press.

Fredlund, D. G., & Rahardjo, H. (1993). *Soil Mechanics for Unsaturated Soils*. John Wiley & Sons, Inc.

Freeze, R. A., & Cherry, J. A. (1979). *Groundwater*. Prentice-Hall, Inc.

Graton, L. C., & Fraser, H. J. (1935). Systematic Packing of Spheres - With Particular Relation to Porosity and Permeability. *Journal of Geology*, *43*(8, Part I), 785–909.

Guobyte, R., & Satkunas, J. (2011). Pleistocene Glaciations in Lithuania. In J. Ehlers, P. L. Gibbard, & P. D. Hughes (Eds.), *Quaternary Glaciations - Extent and Chronology. A Closer Look.* (Vol. 15, pp. 231–246). Elsevier B.V. https://doi.org/10.1016/B978-0-444-53447-7.00019-2

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* (2nd ed.).

Hazen, A. (1892). *Some Physical Properties of Sands and Gravels, with Special Reference to Their Use in Filtration. 24th Annual Report, Massachusetts State Board of Health*, 539-556.

Hazen, A. (1905). *The Filtration of Public Water-supplies* (3rd ed.). John Wiley & Sons.

Head, K. H. (1994). *Manual of Soil Laboratory Testing. Volume 2. Permeability, Shear Strength and Compressibility Tests* (2nd ed). John Wiley & Sons, Inc.

Hechenbichler, K., & Schliep, K. (2004). Weighted k-Nearest-Neighbor Techniques and Ordinal Classification. *Collaborative Research Center 386, Discussion Paper 399.*

Hiscock, K. M. (2005). *Hydrogeology. Principles and Practice*. Blackwell Science Ltd.

Hoerl, A. & Kennard, R. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics, 12(1),* 55-67.

Imandoust, S. B., & Bolandraftar, M. (2013). Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background. *Journal of Engineering Research and Applications*, *3*(5), 605–610.

International Organization for Standardization. (2004). *Geotechnical investigation and testing - Identification and classification of soil. Part 2: Principles for a classification (ISO 14688-2:2004).*

International Organization for Standardization. (2016). *Geotechnical investigation and testing – Laboratory testing of soil – Part 4. Determination of particle size distribution (ISO 17892-4:2016).*

Jang, J., Narsilio, G. A., & Santamarina, J. C. (2011). Hydraulic conductivity in spatially varying media — a pore-scale investigation. *Geophysical Journal International*, *184*(3), 1167–1179. https://doi.org/10.1111/j.1365-246X.2010.04893.x

Justin, J., Hinds, J., & Craeger, W. (1945). Earth, Rock-fill, Steel and Timber. In *Engineering for Dams*, 619–650. John Wiley & Sons.

Kaliakin, V. N. (2017). *Soil Mechanics. Calculations, Principles, and Methods*. Butterworth-Heinemann.

Karmaza, B., & Baltrūnas, V. (2004). Geological, geomorphological and hydrogeological heritage in Lithuania. *Polish Geological Institue Special Papers 13, Proceedings of the Conference: "Geological Heritage Concept, Conservation and Protection Policy in Central Europe,"* 183–190.

Klimašauskas, M., Šaulys, V., Baublys, R., & Survilė, O. (2020). Hydraulic conductivity of drainage ditch backfill with a lime additive in clay soils. *Environmental Engineering and Management Journal , 18*(3), 497–504.

Klizas, P. (2003). *Hidrogeologijos laboratoriniai darbai: mokomoji knyga [Laboratory work in hydrogeology: a textbook]*. VU I-kla.

Klizas, P. (2014). Geofiltration studies of clay at the future radioactive waste repository for Ignalina nuclear power plant. *Journal of Environmental Engineering and Landscape Management, 22*(03), 2019–2225. https://doi.org/10.3846/16486897.2014.903186

Klizas, P., Gadeikis, S., & Žilionienė, D. (2015). Evaluation of Moraine Loams' Filtration Properties. *The Baltic Journal of Road and Bridge Engineering, 10*(4), 293–298. https://doi.org/10.3846/bjrbe.2015.37

Klizas, P. & Šečkus, R. (2007). Filtration and geoelectrical investigations in the karst region of North Lithuania. *Geologija 59*, 77-81.

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review, 39*(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Lane, D. M. (n.d.). *Online Statistics Education: A Multimedia Course of Study*. Retrieved March 2, 2024, from http://onlinestatbook.com/

Leppäranta, M., Virta, J., & Huttula, T. (2017). *Hydrologian perusteet [Basics of Hydrology]*. Helsingin yliopisto, Fysiikan laitos. http://doi.org/10.31885/2018.00021

Li, Y., Rahardjo, H., Satyanaga, A., Rangarajan, S., & Lee, D. T. T. (2022). Soil database development with the application of machine learning methods in soil properties prediction. *Engineering Geology*, *306*, 1–25. https://doi.org/10.1016/j.enggeo.2022.106769

Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2018). *Foundations of Machine Learning* (2nd ed.). The MIT Press.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*. Retrieved 2 March, 2024, from https://christophm.github.io/interpretable-ml-book/

Moorberg, C. J., & Crouse, D. A. (2021). *Soils Laboratory Manual: K-State Edition, Version 2.0*. New Prairie Press.

Naeej, M., Naeej, M. R., Salehi, J., & Rahimi, R. (2017). Hydraulic conductivity prediction based on grain-size distribution using M5 model tree. *Geomechanics and Geoengineering*, *12*(2), 107–114. https://doi.org/10.1080/17486025.2016.1181792

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics, 7, Article 21*. https://doi.org/10.3389/fnbot.2013.00021

Nemes, A., Rawls, W. J., & Pachepsky, Y. A. (2006). Use of the Nonparametric Nearest Neighbor Approach to Estimate Soil Hydraulic Properties. *Soil Science Society of America Journal*, *70*(2), 327–336. https://doi.org/10.2136/sssaj2005.0128

Odong, J. (2008). Evaluation of Empirical Formulae for Determination of Hydraulic Conductivity based on Grain-Size Analysis. *The Journal of American Science*, *4*(1), 1-7.

Onwe, IM., Akudinobi BEB., Aghamelu OP. (2016). Estimating the Hydraulic Conductivity of the Ajali Sandstone in the Udi Area of South Eastern Nigeria from Pumping Test and Grain Size Based Empirical Analysis. *Journal of Hydrogeology & Hydrologic Engineering*, *5*(3), 1-7. https://doi.org/10.4172/2325-9647.1000139.

Pap, M., & Mahler, A. (2019). Comparison of different empirical correlations to estimate permeability coefficient of quaternary danube soils. *Periodica Polytechnica Civil Engineering*, *63*(1), 25–29. https://doi.org/10.3311/PPci.13108

Pinder, G. F., & Celia, M. A. (2006). *Subsurface Hydrology*. John Wiley & Sons, Inc.

Pliakas, F., & Petalas, C. (2011). Determination of Hydraulic Conductivity of Unconsolidated River Alluvium from Permeameter Tests, Empirical Formulas and Statistical Parameters Effect Analysis. *Water Resources Management*, *25*(11), 2877–2899. https://doi.org/10.1007/s11269-011-9844-8

Pucko, T., & Verbovšek, T. (2015). Comparison of hydraulic conductivities by grain-size analysis, pumping, and slug tests in Quaternary gravels, NE Slovenia. *Open Geosciences*, *7*(1), 308–317. https://doi.org/10.1515/geo-2015-0032

Rehman, Z. ur, Khalid, U., Ijaz, N., Mujtaba, H., Haider, A., Farooq, K., & Ijaz, Z. (2022, December 20). *Machine learning-based intelligent modeling of hydraulic conductivity of sandy soils considering a wide range of grain sizes*. Engineering Geology; Elsevier B.V. https://doi.org/10.1016/j.enggeo.2022.106899

Říha, J., Petrula, L., Hala, M., & Alhasan, Z. (2018). Assessment of empirical formulae for determining the hydraulic conductivity of glass beads. *Journal of Hydrology and Hydromechanics*, *66*(3), 337–347. https://doi.org/10.2478/johh-2018-0021

Sahour, H., Gholami, V., Torkaman, J., Vazifedan, M., & Saeedi, S. (2021). Random forest and extreme gradient boosting algorithms for streamflow modeling using vessel features and tree-rings. *Environmental Earth Sciences*, *80*, 747. https://doi.org/10.1007/s12665-021-10054-5

Salarashayeri, A. F., & Siosemarde, M. (2012). Prediction of Soil Hydraulic Conductivity from Particle-Size Distribution. *International Scholarly and Scientific Research & Innovation*, *6*(1), 16–20.

Scikit-learn. (2024a). *Preprocessing data*. Retrieved February 28, 2024, from https://scikit-learn.org/stable/modules/preprocessing.html

Scikit-learn. (2024b). *Linear Models*. Retrieved March 1, 2024, from https://scikit-learn.org/stable/modules/linear_model.html

Scikit-learn. (2024c). *Support Vector Machines*. Retrieved March 5, 2024, from https://scikit-learn.org/stable/modules/svm.html

Scikit-learn. (2024d). *Decision Trees*. Retrieved March 16, 2024, from https://scikit-learn.org/stable/modules/tree.html

Singh, B., Sihag, P., Parsaie, A. & Angelaki, A. (2021). Comparative Analysis of Artificial Intelligence Techniques for the Prediction of Infiltration Process. *Geology, Ecology, and Landscapes*, 5(2), 109–18. doi:10.1080/24749508.2020.1833641.

Singh, V. K., Kumar, D., Kashyap, P. S., Singh, P. K., Kumar, A., & Singh, S. K. (2020). Modelling of soil permeability using different data driven algorithms based on physical properties of soil. *Journal of Hydrology*, *580*. https://doi.org/10.1016/j.jhydrol.2019.124223

Slichter, C. S. (1899). Theoretical Investigations of the Motion of Groundwater. In *United States Geological Survey 19th Annual report, 1897-98, Part 2*.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*, 199–222.

Taylor, D. W. (1948). *Fundamentals Of Soil Mechanics*. John Wiley & Sons, Inc.

Tilahun, T., & Korus, J. (2023). 3D hydrostratigraphic and hydraulic conductivity modelling using supervised machine learning. *Applied Computing and Geosciences*, *19*. https://doi.org/10.1016/j.acags.2023.100122

Twarakavi, N. K. C., Šimůnek, J., & Schaap, M. G. (2009). Development of Pedotransfer Functions for Estimation of Soil Hydraulic Parameters using Support Vector Machines. *Soil Science Society of America Journal*, *73*(5), 1443–1452. https://doi.org/10.2136/sssaj2008.0021

Urumović, K., Borović, S., Urumović, K., & Navratil, D. (2020). Validity range and reliability of the United States Bureau of Reclamation (USBR) method in hydrogeological investigations. *Hydrogeology Journal*, *28*(11), 1–35. https://doi.org/10.1007/s10040-019-02080-2

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., van der 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

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer.

Veloso, M. F., Rodrigues, L. N., & Filho, E. I. F. (2022). Evaluation of machine learning algorithms in the prediction of hydraulic conductivity and soil moisture at the Brazilian Savannah. *Geoderma Regional*, *30*, 1–12. https://doi.org/10.1016/j.geodrs.2022.e00569

Vukovic, M., & Soro, A. (1992). *Determination of Hydraulic Conductivity of Porous Media from Grain-Size Composition*. Water Resources Publications.

Wang, Y., Wu, X., Chen, Z., Ren, F., Feng, L., & Du, Q. (2019). Optimizing the predictive ability of machine learning methods for landslide susceptibility mapping using smote for Lishui city in Zhejiang province, China. *International Journal of Environmental Research and Public Health*, *16*(3), 1–27. https://doi.org/10.3390/ijerph16030368

Wenzel, L. K., & Fishel, V. C. (1942). Methods for determining permeability of water-bearing materials, with special reference to discharging-well methods, with a section on direct laboratory methods and bibliography on permeability and laminar flow, by V.C. Fishel. *United States Department of the Interior, Water Supply Paper 887*.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Climate Research*, *30*, 79–82.

Woessner, W. W., & Poeter, E. P. (2020). *Hydrogeologic Properties of Earth Materials and Principles of Groundwater Flow*. The Groundwater Project.

Wu, Y. (2021). Can't ridge regression perform variable selection? *Technometrics*, *63*, 263–271.

Yolcubal, I., Brusseau, M. L., Artiola, J. F., Wierenga, P., & Wilson, L. G. (2004). Environmental Physical Properties and Processes. In J. F. Artiola, I. L. Pepper, & M. L. Brusseau (Eds.), *Environmental Monitoring and Characterization* (pp. 207–239). Elsevier Science & Technology Books.

## SUMMARY

## VILNIUS UNIVERSITY
## FACULTY OF CHEMISTRY AND GEOSCIENCES

EVELIINA KUKKA-MAARIA VANHALA

**Theoretical hydraulic conductivity determination of Lithuanian soil samples**

Hydraulic conductivity is the ability of soil to transmit water and is measued by the rate which water can move through the porous medium. The hydraulic conductivity of soil is affected by numerous factors like soil physical properties and grain size, and has a significant role in fields like geotechnical design, contaminant migration and waste disposal. It can be determined directly *in situ* or through laboratory tests. Indirect methods include empirical formulas and machine learning modelling. Both utilize physical soil parameters in the determination of hydraulic conductivity. Machine learning is useful in the sense of computational capacity to process data that it can more easily find relationships between multiple parameters as well as produce new predictions by using complex algorithms, while empirical formulas require manual data processing.

This thesis investigates theoretical hydraulic conductivity determination methods of Lithuanian soil samples. The primary objectives of this study are creating a database for Lithuanian soil samples and assessing theoretical hydraulic conductivity methods by comparing them to laboratory-acquired values. The study is conducted by using three empirical formulas from Hazen, Slichter and USBR and tuning seven machine learning regression models to find the best parameters to use in the determination of hydraulic conductivity. The regression models used in the study are linear and ridge regression, support vector regression (SVR), K-Nearest Neighbors (KNN), Decision tree, Random forest and Gradient boosting.

The results reveal that from the three empirical formulas, Hazen‘s formula performs the best while Slichter‘s formula has the lowest correlation to actual hydraulic conductivity values. However, the overall accuracy of these empirical formulas remains low. From the six machine learning models, Random forest performed the best in multiple different tests and by using different parameters. The highest correlation is achieved by using grain size information of fine soils, grain size diameters $D_{60}$ and $D_{70}$, and water content. Overall, the machine learning models performed better than the empirical formulas. The results reveal that the machine learning models can adjust to the heterogenous nature of soils and find patterns between multiple soil parameters.

The overall correlation of both empirical formulas and machine learning models remain relatively low, and the main reason for this might be due to the small size of the database entries. It is encouraged to keep updating the soil sample database to gain a wider resource for future investigations of soil permeability in Lithuania.



**Keywords:** hydraulic conductivity, groundwater, empirical formulas, machine learning.

## SANTRAUKA

**VILNIAUS UNIVERSITETAS**
**CHEMIJOS IR GEOMOKSLŲ FAKULTETAS**

EVELIINA KUKKA-MAARIA VANHALA

**Teorinis filtracijos koeficiento nustatymas Lietuvos gruntų mėginiuose**

Filtracijos koeficientas nusako grunto gebą praleisti vandenį per porėtą terpę. Gruntų filtracijos koeficientas turi įtakos daugeliui veiksnių, tokių kaip grunto fizikinės savbės, granuliometrinė sudėtis. Šis parametras yra plačiai naudojamas tokiose srityse kaip geotechninis projektavimas, teršalų migracija, bei įvairių atliekų poveikio aplinkainustatymui. Filtracijos koeficientas gali būti nustatytatomas tiesiogiai *in situ* arba atliekant laboratorinius tyrimus.. Netiesioginiai metodai apima empirines formules ir mašininio mokymosi modeliavimą. Filtracijos koeficientui apskaičiuoti naudojami fizinikiniai grunto parametrai. Skaičiavimai empirinėmis lygtimis reikalauja daug rankinio darbo su duomenimis ir galiausiai turi būti supaprastinti į lygtis. Mašininis mokymasi privalumas yra galimybė kompiuterizuoti skaičiavimus ir rasti ryšius tarp kelių parametrų ir sukurti naujas prognozes, naudojant sudėtingus algoritmus.

Baigiamajame darbe nagrinėjami teoriniai filtracijos koeficiento skaičiavimo metodai naudojant Lietuvos grunto mėginius. Pagrindiniai šio tyrimo tikslai – sukurti Lietuvos gruntų mėginių duomenų bazę ir įvertinti teorinius filtracijos koeficiento skaičiavimo rezultatus, lyginant juos su laboratorijoje gautomis reikšmėmis. Tyrimas atliktas naudojant Hazen, Slichter ir USBR empirines formules, bei derinant šešis mašininio mokymosi regresijos modelius, kad būtų rasti tinkamiausi parametrai, kuriuos galima naudoti modeliuojant. Tyrime naudojami regresijos modeliai: tiesinė ir keterinė regresija, atraminių vektorių regresija (SVR), artimiausių kaimynai (KNN), sprendimų medžio, atsitiktinio miško ir gradiento didinimo regresijos.

Rezultatai rodo, kad iš trijų empirinių formulių Hazen formulė veikia geriausiai, o Slicther formulė turi mažiausią koreliaciją su faktinėmis hidraulinio laidumo vertėmis. Tačiau bendras šių empirinių formulių tikslumas išlieka mažas. Iš šešių mašininio mokymosi modelių atsitiktinio miško algoritmas tiksliausiai nustatė teorines vertes atliekant kelis skirtingus testus ir naudojant skirtingus parametrus. Geriausi statistiniai rodikliai pasiekti naudojant smulkaus grunto grūdelių dydžio informaciją, grūdelių dydžio skersmenis $D_{60}$ ir $D_{70}$ bei drėgnį. Mašininio mokymosi modeliai tiksliau nustatė filtracijos koeficientą nei empirinės formulės. Rezultatai rodo, kad mašininio mokymosi modeliai gali prisitaikyti prie nevienalyčio grunto pobūdžio ir rasti modelius tarp kelių grunto parametrų.

Bendra empirinių formulių ir mašininio mokymosi modelių koreliacija išlieka santykinai žema, o pagrindinė to priežastis gali būti mažo duomenų imti, todėl tokio pobūdžio tyrimai turėtų būti tęsiami plečiant duomenų bazę.

**Raktiniai žodžiai:** filtracijos koeficientas, požeminis vanduo, empirinės formulės, mašininis mokymasis.

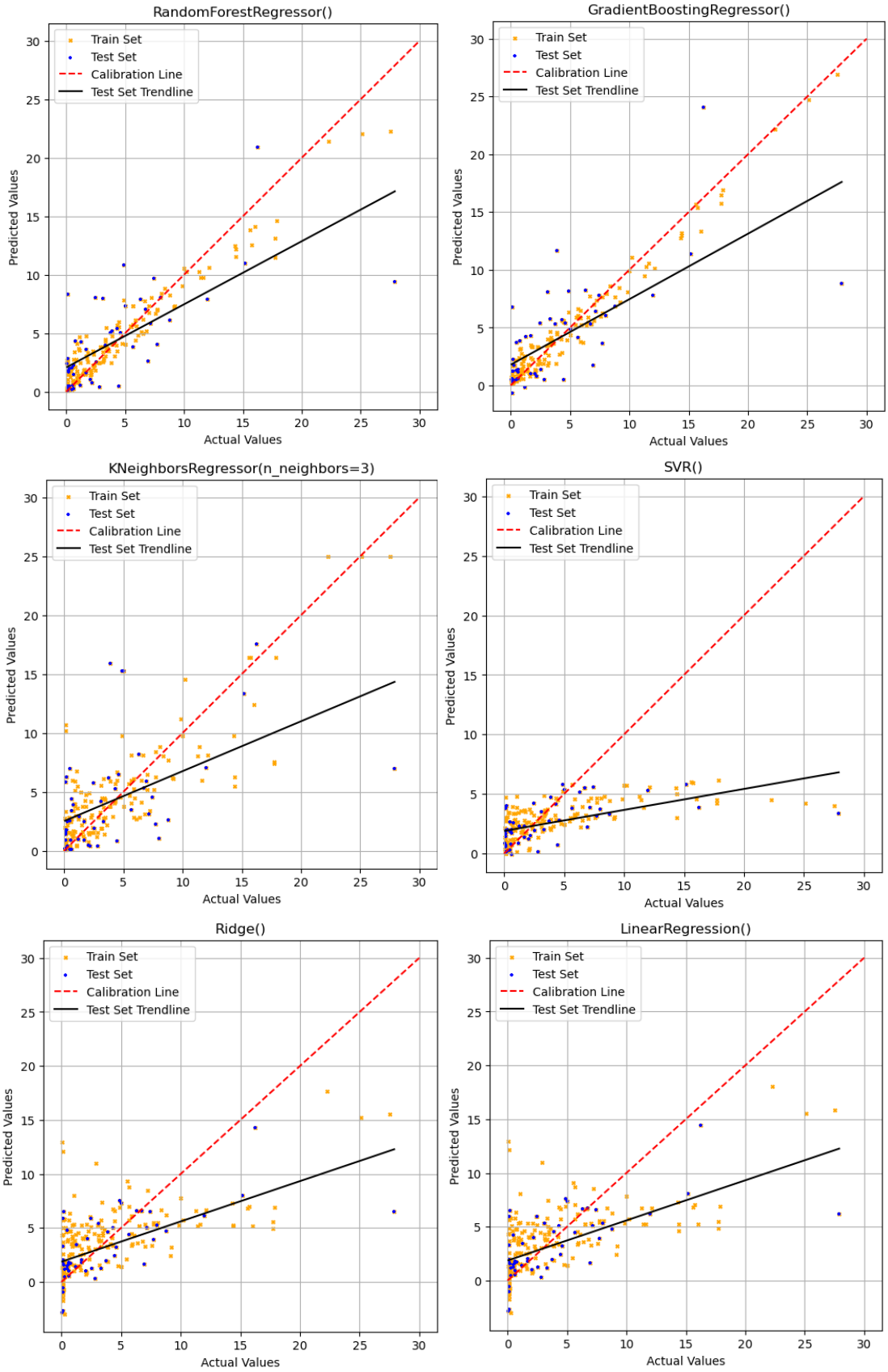
**APPENDICES**

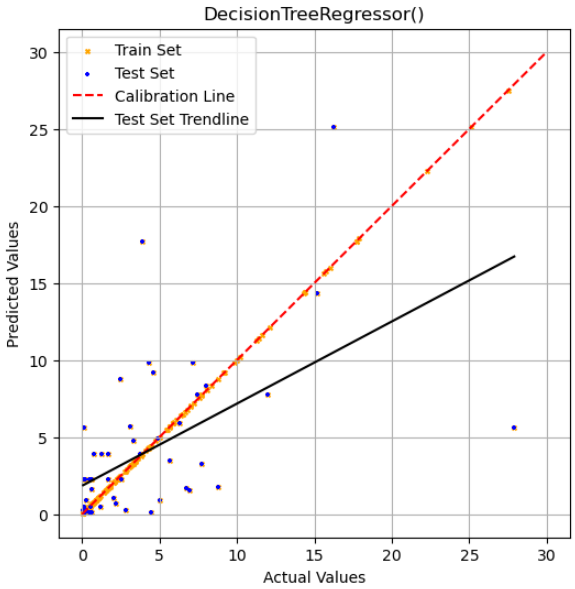
1. The best-fitting machine learning test

APPENDIX 1

**THE BEST-FITTING MACHINE LEARNING TEST**

Used parameters in the test: <0.06 mm, 0.06-0.2mm, water content, $D_{60}$, $D_{70}$.






| **Model** | **MAE (test set)** | **MAE (train set)** | **MAE (SD)** | **$R^2$ (test set)** | **$R^2$ (train set)** | **$R^2$ (SD)** |
|---|---|---|---|---|---|---|
| Random forest | 2.28 | 2.43 | 0.25 | 0.47 | 0.51 | 0.10 |
| Gradient boosting | 2.32 | 2.46 | 0.23 | 0.44 | 0.48 | 0.08 |
| K-neighbors | 3.04 | 2.70 | 0.28 | 0.11 | 0.35 | 0.10 |
| Ridge | 2.42 | 3.10 | 0.13 | 0.39 | 0.15 | 0.29 |
| Linear | 2.40 | 3.13 | 0.11 | 0.38 | 0.14 | 0.29 |
| SVR | 2.45 | 2.70 | 0.48 | 0.18 | 0.19 | 0.10 |
| Decision tree | 2.67 | 2.97 | 0.46 | 0.15 | 0.09 | 0.19 |