Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Department of Soil and Environment

Benchmarking a random forest for predicting preferential flow in soils

Helena Jordà Guerra

Master's Thesis in Soil Science
Soil and Water Management – Master's Programme

Examensarbeten, Institutionen för mark och miljö, SLU
2013:15

Uppsala 2013

SLU, Swedish University of Agricultural Sciences
Faculty of Natural Resources and Agricultural Sciences
Department of Soil and Environment

Helena Jordà Guerra

Benchmarking a random forest for predicting preferential flow in soils

Supervisor: John Koestel, Department of Soil and Environment, SLU
Examiner: Nicholas Jarvis, Department of Soil and Environment, SLU

EX0430, Independent project/degree project in Soil Science – Master´s thesis, 30 credits, Advanced level, A2E
Soil and Water Management – Master's Programme, 120 credits

Series title: Examensarbeten, Institutionen för mark och miljö, SLU
2013:15

Uppsala 2013

Keywords: preferential flow, random forest, benchmarking, undisturbed soil, soil database

Online publication: http://stud.epsilon.slu.se

Abstract

Preferential flow processes are important to fully understand flow and solute transport in the vadose zone and implement adequate management practices. Physical models are difficult to use at large scales to predict soil susceptibility to preferential flow. Instead, pedotransfer functions might be applied. The strength of preferential flow can be measured by the relative 5% arrival time obtained from breakthrough curve (BTC) experiments. I used a database containing 560 BTC experiments to build random forests to predict the relative 5% arrival time and analyse the importance of soil properties and site factors on predicting this feature. The coefficient of determination for a 10-fold cross-validation was 70%, whereas the benchmarking process obtained a coefficient of 27%. Sand contents between 0.80 and 0.92 reached the highest importance and were strongly related to weak preferential flow. High importance was also observed in silt contents lower than 0.11, and clay contents between 0.04 and 0.08, which were strongly correlated to high preferential flow. In addition, experimental conditions such as flow rate, column diameter, the use of fixed drippers and column venting had 20% importance. This study revealed that texture can broadly predict soil susceptibility to preferential flow, while other site and soil factors can later refine this estimate. However, the dataset lacked land use information and a broader range of experimental conditions. I consider that enlarging the database is a key factor to obtain better predictions and to further understand how soil and site characteristics influence soil susceptibility to preferential flow.

Popular science summary

The water movement in the soil has already been studied since the 1950s due to its importance for the transport of pollutants through the soil. Water can move fast through preferential paths while avoiding or moving slower through other spaces of the soil profile. This is called preferential flow. This process is of great importance because it can make organic pollutants, pesticides and heavy metals move faster than expected and reach groundwater levels in concentrations exceeding the permitted limits. These phenomena need to be taken into account in models used to decide on agricultural and environmental policies in order to avoid the contamination of water bodies. This research has focused on investigating the accuracy with which we can predict preferential flow and on studying the effect of soil properties, such as soil texture (classification according to grain size), and site factors, for instance land use and irrigation type, on these predictions.

A regression technique called “random forest” was applied on a soil database including information about soil properties, site factors and an indicator of preferential flow in soils. The results showed that we can predict preferential flow with an accuracy of 66%.

Texture characteristics were found to be the most important soil properties to predict preferential flow. Sandy soils tend to exhibit weak preferential flow. The strength of preferential flow increases when silt and clay contents exceed 10%. This study also shows that characteristics related to rain intensity and other soil properties (bulk density and organic carbon content) are also important to predict this kind of process. Land use factors did not reach high importance in this study, probably because data regarding these characteristics was not often available.

The lack of information is a limiting factor for predicting the preferential movement of water through the soil. Enlarging the database with more soil examples and obtaining broader information concerning site factors is essential to obtain better predictions. This will help soil scientists to construct better models to foresee the movement of water and contaminants in the soil and adjust management practices to reduce the threats to the environment and human health.

Table of contents

Introduction

Preferential flow involves all circumstances when water and solutes move through specific regions of the soil pore space while avoiding other zones (Hendrickx and Flury, 2001). This is a process of great importance because it may cause metal and organic pollutants in water to move through the soil faster than expected. This may lead to unforeseen contamination of groundwater and, eventually, surface water bodies. Therefore, estimating soil's susceptibility to preferential flow in large-scale areas is essential to support good management practices.

Classic physical models have been used to describe water and solute transport in the unsaturated zone. However, using these models present different difficulties. On the one hand, the Navier-Stokes equations are models at a pore scale, which are difficult to use at large scales. On the other hand, models at the Darcian scale, such as Richard's and convection-dispersion equations, approximate the soil to be homogeneous, which usually is not the case. Therefore, these physical models are not good to predict preferential flow in soils.

Instead, preferential flow may be estimated by several indicators studied by Knudby and Carrera (2005). Nevertheless, the direct measurement of these parameters at large management scales is only possible with expensive monitoring studies. Current research has been focusing on elaborating pedotransfer functions (PTF) (Wösten et al., 2001) that estimates preferential flow from soil properties and site factors (Koestel et al., 2012; Jarvis et al., 2009; Shaw et al., 2000). These features are usually already available in soil databases and they are commonly cheaper to obtain from fieldwork. Hence, these PTF could be an alternative to the usage of classic physical models.

The aims of this study are (1) to investigate how accurate soil's susceptibility to preferential flow might be estimated and (2) to evaluate the importance of different soil properties, site factors and experimental conditions on predicting preferential flow. For those purposes, a machine learning technique called random forest was used on an existing soil database.

This thesis gives an overview on previous knowledge on preferential flow and on machine learning theory, which is included in the literature review. Then, the material and methods section contains information about the used database and the process of building the random forest. Finally, the results from the analyses are presented and discussed, and several conclusions are given.

Literature review

Preferential flow

Water flow in the unsaturated zone can be mathematically described by the Navier-Stokes equations, which are based on the concept of a fluid continuum filling the pore scale (Hendrickx and Flury, 2001). Using these equations to describe water flow at large scales would imply knowing the porous structure of the whole soil and calculating the movement of individual water molecules, which is too laborious and would require too much computation power.

To solve this problem it is necessary to move to a larger scale. Water flow in the unsaturated zone has also been described by Richard's equation, which combines Darcy's law with an equation of continuity for water mass. The convection-dispersion equation (CDE) has been used to calculate solute transport in the soil. This equation involves three physical processes occurring in the soil: convection, diffusion and dispersion. The mobile-immobile model (MIM) has also been used to describe solute transport in the vadose zone. This model partitions the soil water in two domains: a mobile and an immobile phase. It assumes that there are domains in the soil where water flows following the CDE, while the remaining pore space is filled with stagnant water. This model also considers solute exchange between the two domains by diffusion processes (Nielsen et al., 1986). Both Richard's and CDE are based on the conjecture that soil water pressure and solute concentration can be described for a representative elementary volume of soil. However, heterogeneities in the soil can produce variations in soil hydraulic properties that invalidate this assumption (Jarvis, 2007).

Due to soil heterogeneities, and also because preferential flow is very sensitive to initial and boundary conditions, describing preferential flow at pore and Darcian scales is very difficult (Jarvis et al., 2012).

What causes preferential flow?

Root channels, earthworm burrows, fissures and cracks are common examples of macropores (Hendrickx and Flury, 2001). A macropore is, by definition, a pore of large dimensions. Pores with an equivalent cylindrical diameter larger than about 0.3-0.5 mm have been termed macropores (Jarvis, 2007). Macropore flow is a type of preferential flow and consists of the preferential movement of water through macropores. Tomographic techniques such as computed tomography (Heijs et al., 1996) and other imaging methods like magnetic resonance imaging (Van As and van Dusschoten, 1997) have been found to be useful in order to visualize macropore structures that may cause preferential flow (Hendrickx and Flury, 2001).

Unstable flow is another sort of preferential flow. It can be observed in coarse-textured soils caused by textural layering, air entrapment, water repellency or unstable wetting (Hendrickx and Flury, 2001). Finally, funnel flow is the lateral redirection of water due to textural boundaries where water moves through the pathway that offers less resistance and avoids impeding layers (Hendrickx and Flury, 2001).

Preferential flow can be inferred in different ways. One way is by performing breakthrough curve (BTC) experiments (Hendrickx and Flury, 2001). BTC experiments consist of infiltrating a solution through a defined volume of soil and measuring the solute concentration at the outlet. The unexpected early arrival of tracer concentration and long tailing are indications of preferential pathways.

However, it is also important to be able to quantify preferential flow phenomena. Different indicators of flow and transport connectivity were analysed by Knudby and Carrera (2005). Their results, deduced from computer simulations of BTC experiments, suggest that the relative 5% arrival time calculated from BTC experiments is a good indicator of preferential flow. The relative 5% arrival time is the ratio between the time when 5% of the solute reaches the outlet and the mean arrival time of the solute. Smaller relative 5% arrival times indicate earlier arrivals of tracers and stronger preferential flow.

Furthermore, later research has focused on identifying soil and site attributes that may help to predict preferential flow. As discussed before, describing preferential flow at the pore scale would be too laborious. Indeed, there is no need to describe flow for each macropore, if it can be explained for larger areas in terms of measurable soil properties (Jarvis et al., 2012) and site factors (Jarvis et al., 2009).

Especially in cultivated topsoil, macropore flow is related to soil aggregation, because it affects the formation of macropores. Soils aggregation is hierarchical, so that bigger aggregates consist of groups of smaller aggregates separated by planes of weakness and these smaller aggregates by much smaller aggregates. The lower the order in the structure, the stronger the aggregates because they do not comprise the pore space between higher order aggregates. In addition, higher orders in the structure hierarchy should be related to stronger preferential flow because the bigger the aggregates, the larger and widely spaced they are (bigger pores) (Jarvis, 2007). Soil properties such as texture and organic matter have an effect on soil aggregation. Consequently, they have an indirect effect on macropore flow and transport (Jarvis, 2007). Minimum clay contents of 8 and 9% have been found to be associated with strong preferential flow when using the relative 5% arrival time as an indicator (Koestel et al., 2012). In structureless soils, high matrix permeability is also related to preferential flow (Jury and Flühler, 1992). Soil and crop management practices such as tillage may also have an impact on soil structure and pore connectivity, which are important factors for macropore flow. For instance, soil compaction due to field operations degrades the aggregate hierarchical structure. As a result, preferential flow is enhanced

because water flows are concentrated through the macropores that remain connected (Jarvis, 2007). Furthermore, land use history has been reported to affect water repellency degree in soils, which might lead to wetting front instability that causes unstable flow (Sonneveld et al., 2003). Finally, hydrologic initial and boundary conditions, such as rain duration and intensity or initial soil water content, have complex effects causing preferential flow (Jarvis, 2007).

The estimation of soil's susceptibility to preferential flow from soil properties and site factors constitutes a PTF. This type of PTF could be used to predict, for instance, nutrient losses from agricultural fields or metal mobilisation from mining areas. At present, the CDE is still implemented in software programs used to carry out environmental and agricultural policies as Jury and Flühler (1992) had already critised in 1992. Due to the inaccuracy of this model for describing preferential flow, it is necessary to elaborate more reliable models. Hence, the elaboration of these PTF, which could be implemented in new models as an alternative to the CDE, is of great importance.

Machine learning

Arthur Samuel (1959) defined machine learning as the "field of study that gives computers the ability to learn without being explicitly programmed". In order to illustrate this definition, one might look to the example given by Andrew Ng (Andrew Ng, 2009) about an automatized car which "learns" how to drive. This car has a system of artificial neural networks, a machine learning technique, which learns to steer by watching a person drive. The system is trained capturing images of the road ahead and recording the steering directions given by the driver. Using a learning algorithm, the system is then instructed to output the same steering direction as the human driver for each image. After the system has been trained, it starts to drive the car. It captures images of the road ahead and fits them to its neural networks. The steering direction from the most confident network is used to control the vehicle.

In the example given above, the algorithm uses as input the digitalised images of the road ahead captured by the system when the human drives. The output would be the steering directions given by the same driver. Machine learning tries to extract the algorithm from input and output data (Alpaydin, 2004). It may not be possible to obtain an algorithm that describes the processes completely, but we might still produce good and useful approximations. These approximations may be useful to detect patterns and regularities in data, which is thought to help understand processes and to make predictions (Alpaydin, 2004). Machine learning has many applications present in our daily lives such as webpage ranking, automatic translation, name entity recognition and speech recognition (Semola and Vishwanathan, 2008). Webpage ranking is an example of data mining, when machine learning is applied to large datasets in order to find patterns. Name entity and speech recognitions are cases of pattern recognition (Alpaydin, 2004).

There are two main procedures in machine learning: supervised and unsupervised learning. In supervised learning, a training set composed of inputs and corresponding outputs is fit to a learning algorithm. This algorithm has as output a function, which historically has been called the hypothesis. Given a new set of input data, this function will return the expected output (Figure 1) (Andrew Ng, 2009). The example about the automatized car is a case of supervised learning because inputs and outputs are provided to make future predictions, which in this case are the right steering directions to keep the car on the road. Classification and regression problems are examples of supervised learning. On the other hand, unsupervised learning only relies on data input and tries to identify the most frequent patterns, which in statistics is called density estimation (Alpaydin, 2004).

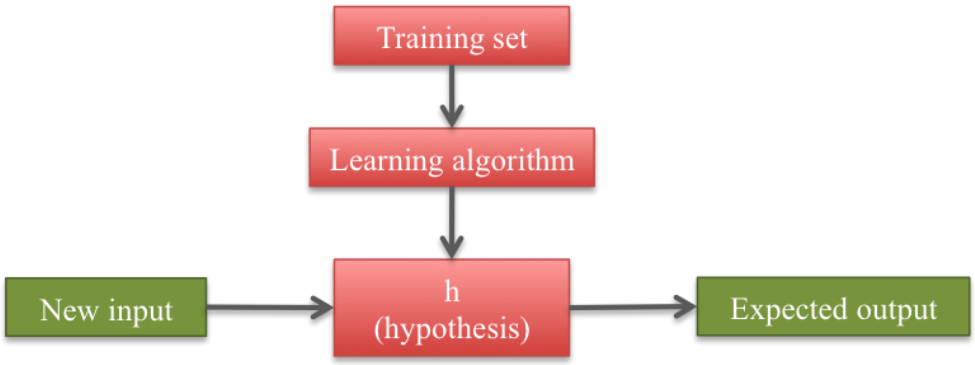


Figure 1. Supervised learning procedure (after Andrew Ng, 2009).

In supervised learning, the learning algorithm finds a particular hypothesis (e.g. a line with formula y=2x+3) belonging to a hypothesis class or function type (e.g. linear functions) to approximate as closely as possible the unknown function that describes the process (Alpaydin, 2004). The election of a hypothesis class, also called model selection, is one assumption to be made when looking for the hypothesis. The hypothesis is a function of x and parameterised by θ. After selecting the hypothesis class, the training set is fit to the hypothesis class (by giving values to its parameters θ, a process that in modelling is called calibration) and the inconsistent hypotheses are rejected. However, the training set may not be enough to reach a unique solution and many other consistent hypotheses are still left. This is called an ill-posed problem (Alpaydin, 2004). Therefore, more assumptions need to be made to reach a unique solution. For instance, in linear regression, an assumption is made when choosing the linear function that minimizes the squared error between the real and predicted values. The set of assumptions necessary to reach a unique solution is called inductive bias of the learning algorithm (Alpaydin, 2004).

The accuracy with which the model predicts the right output for a new data set is called generalisation (Alpaydin, 2004) and the process to evaluate this accuracy is called validation. The dataset used to evaluate the generalisation error is called the validation set (Alpaydin, 2004). For a good generalisation it is necessary that the

model we select is as complex as the function underlying the data. If the model is less complex than the function, there will be patterns in the data that the model fails to fit. This is called under-fitting (Alpaydin, 2004). On the other hand, if the model is more complex than the function, it will fit particularities and noise in the data instead of the true trend. This second situation is known as over-fitting (Alpaydin, 2004). The concepts of under-fitting and over-fitting are exemplified in figure 2.

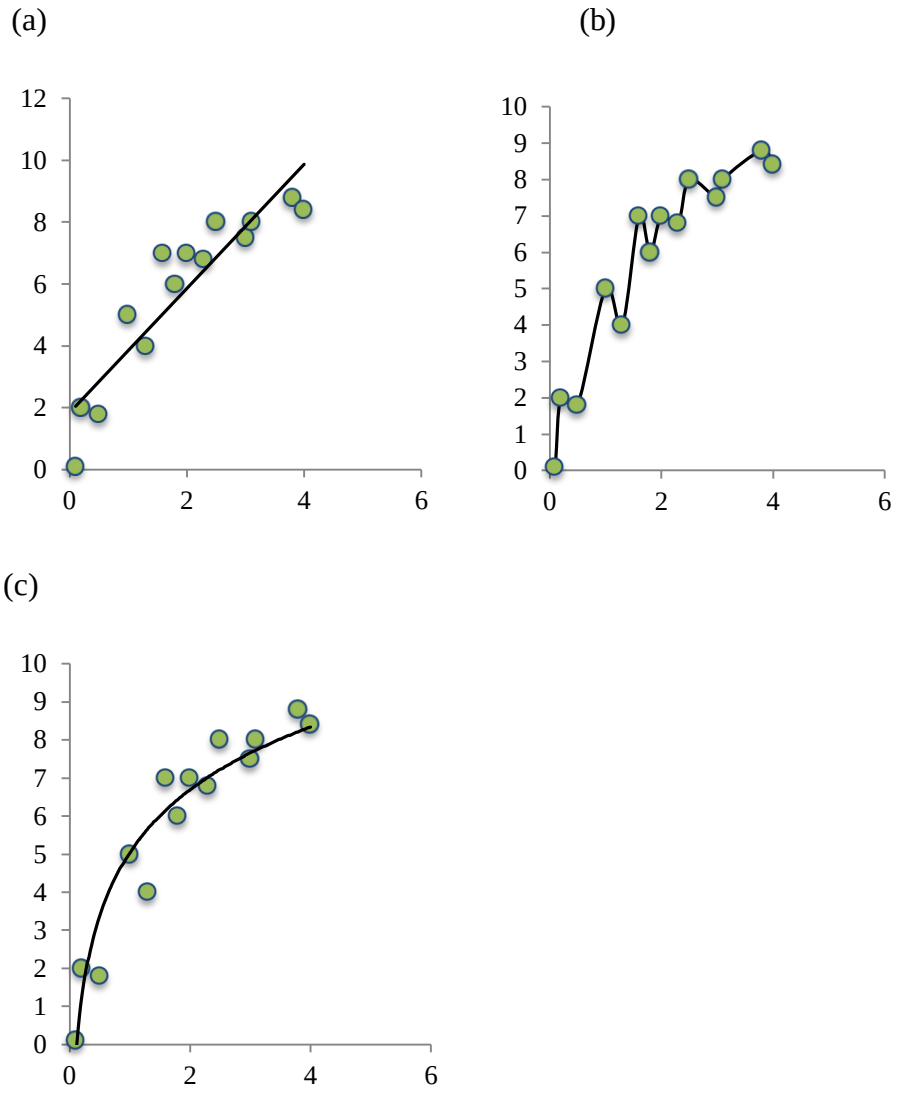


Figure 2. Examples of fitting a model to the same training set: a) case of under-fitting, b) case of over-fitting, c) real trend.

As we increase the number of samples in the training set, the generalisation error decreases. The generalisation error also decreases with increasing complexity of our model, until over-fitting occurs and the error increases again (Alpaydin, 2004). Therefore, the complexity of the model, the sample size and the

generalisation error are important factors when fitting a learning algorithm to the training data.

Random forests

"Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" (Breiman, 2001). A tree predictor, more popular called as decision tree**,** is a non-parametric method which Alpaydin (2004) defined as "hierarchical model for supervised learning whereby the local region is identified in a sequence of recursive splits into smaller number of steps". A decision tree is composed of internal decision nodes and terminal leaves. An internal decision node is a node where the tree splits into different branches according to a test function $f_m(x)$ with a finite number of outcomes. A terminal leaf is reached when the branch cannot be divided into more branches and it defines a specific region in the input space where instances falling in this region have the same output label (Figure 3a). The input space is the space that contains all input instances. The output label of a branch can be a class code for classification trees (e.g. "child", "adult" or "old person" if we try to estimate the life stage of people, shown in Figure 3b) or a numeric value for regression trees (e.g. values from 0 to infinity if we estimate the height of a tree) (Alpaydin, 2004).

a) b)

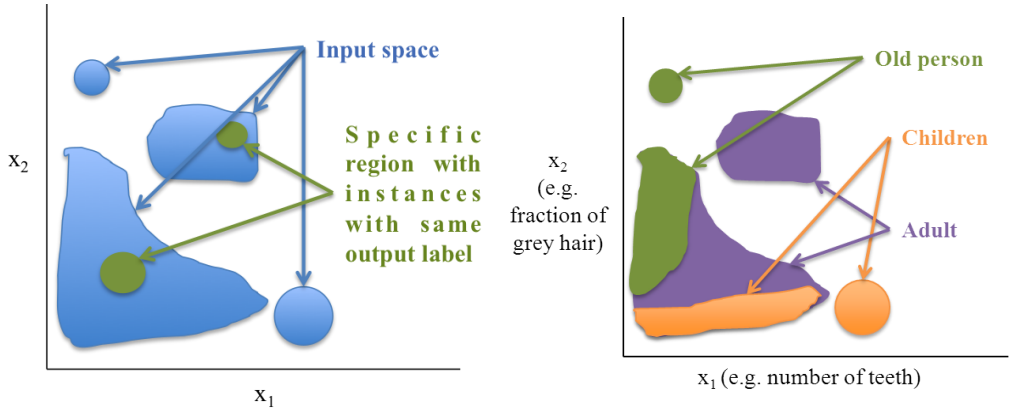


Figure 3. a) Example of an input space and a specific region with instances with same output label. In this case, two features (x_1 and x_2) compose the input instances; b) same example but with specific input variables (x_1 = number of teeth and x_2 = fraction of grey hair) and output labels (children, adult and old person).

A tree can be grown until there is only one sample coherent with the test-function at the end of the branch. However, generating such large trees may lead to over-fitting. On the other hand, small trees risk under-fitting the data (see Figure 2). The optimal tree size should be chosen according to the available data (Hastie et

al., 2009). There are different approaches to limit the growth of a forest, such as defining a minimum number of samples at a node to split it (Hastie et al., 2009). Defining criteria to stop the growth of a tree generates a pre-pruned tree. Pruning is a technique used to adjust the size of a tree. Pre-pruning stops the growth of a tree at the desired size. On the contrary, post-pruning removes unnecessary sub-trees after having grown the full tree (Alpaydin, 2004). Pruning is important because it avoids the process of splitting branches when few instances are left, which would cause an increase in generalisation error due to over-fitting (Alpaydin, 2004).

According to Breiman (2001), random forests do not over-fit data, so theoretically we could run as many trees as we desire. As we increase the number of trees in the random forest, the greater the chance we have that all predictors are equally picked for the bootstrap samples at the splitting steps and the higher the random forest predictive power will be. However, the forest size is limited by the processing capacity of the computer, and the higher the number of trees we build the more time is required to run the program. Therefore, it is necessary to limit the forest size.

There are different ways to validate a random forest. One of them is k-fold cross-validation (Nilsson, 1998). This validation consists of dividing the dataset into k equal-sized subsets. Each subset is used to validate a random forest built on the other k-1 subsets, and an error rate is calculated. This error rate is the number of misclassification errors made on the validation set divided by the number of features of the same set. The average of all error rates is an estimation of the error rate expected on new data of the forest built on the training set (Nilsson, 1998).

Material and methods

Dataset

I used a dataset composed of 560 BTC experiments from 59 peer-reviewed articles. This dataset was divided into two sets. The training set was composed of 454 BTCs from 52 articles from the dataset compiled by Koestel et al. (2012). These data was used to build the random forest and for the validation process. The benchmarking set was composed by 106 BTCs from 7 new articles added to the dataset by Koestel et al. (2012). These data were reserved for an additional benchmarking procedure, which is explained later in this section. The 7 articles added to the database are summarized in Table 1. Only BTC experiments under steady-state conditions, with texture data available and with fittings to BTC raw data with R^2 over 95% were used to build trees in this study. Data on soil properties, experimental conditions and site factors were assembled and organised in Excel tables and later transferred to a MySQL database. BTC raw data were obtained from the authors of the articles or by digitalising the curves included in the articles. For that purpose, the program Plot Digitizer 2.6.2 was used. The

relative 5% arrival times were obtained from the corresponding mobile-immobile model transfer-functions as explained in Koestel et al. (2012). This was done because for the majority of BTCs only MIM parameters were available and this way the relative 5% arrival times were estimated in a consistent way.

Table 1. Primary source publication and other information on the BTC experiments collected and added to the meta-database from Koestel et al. (2012).

Primary reference	# of BTCs	Tracer	Type of soil or porous medium	USDA texture class	Land use
(Bejat et al., 2000)	4	chloride	unknown	sandy clay loam silt loam clay	managed grassland arable
(Kluitenberg and Horton, 1990)	9	chloride	aquic hapludoll*	loam clay loam	arable
(Koestel et al., 2013)	69	tritium	alfic argiudoll*	sandy loam loam	arable
(Pot et al., 2011)	18	bromide	albeluvisol**	silt loam	arable
(Tabarzad et al., 2011)	5	chloride	calcixerillic xerochrepts typic calciorthids	clay loam sandy loam silt loam loam	arable
(Zhou et al., 2011)	10	chloride	unknown	sandy clay sandy loam	unknown
(Zurmühl et al., 1991)	2	chloride	typic haplaquept	loamy sand sandy loam	arable

*Classification according to the system of the United States Department of Agriculture (USDA).
**Classification according to the World Reference Base (WRB).

The investigated BTC experiments were performed in soils with a wide range of sand, silt and clay contents. The organic carbon content was frequently less than 0.05, with extreme values reaching 0.4, whereas the Hassink-Dexter index (ratio between clay and organic carbon contents) was commonly from 1 to 50. Finally, bulk density took values mostly from 1.20 to 1.70 $g \cdot cm^{-3}$. The experiments were mostly performed on short columns 15-50 cm long and 1-20 cm of diameter. The applied flow rate varied from 0.01 to 2232 $cm \cdot d^{-1}$, although the most common values ranged from 0.75 to 50 $cm \cdot d^{-1}$.

About 40% of the experiments used fixed drippers as irrigation device, although rotating drippers and ponding were also widely used to irrigate the columns, accounting for 50% of the experiments. In about 60% of the experiments, the soil was open to atmospheric pressure at the bottom, whereas suction was applied for

the rest. Only 30% of the experiments were carried in columns with sealed walls and about 48% of the experiments were vented, which means that they had been saturated from the bottom before starting the experiment to allow air to escape from the pore system. Anionic tracers were used in 80% of the experiments. Chloride and bromide were the most commonly used tracers.

Most experiments (58%) were conducted in arable soils, the second most common land uses were managed grasslands and forest soils, both accounting for 12% of the experiments. The rest of land uses were other kind of grassland or were unknown (6%). About 50% of the soils were ploughed and 30% had no management. Finally, around 70% of soils were trafficked and for most of them (85%) manure was not applied. Still, most articles did not state directly if the soils were trafficked or if manure was applied, so these site factors were inferred.

Random forest

Random forests are grown on bootstrap samples, which are generated by drawing instances from the original sample (Alpaydin, 2004). Our original sample was composed of 454 BTCs obtained from the articles. Soil properties, experimental conditions and land-use factors are used as input data, while shape-measures are used as output. In this study, the relative 5% arrival time has been investigated as indicator of preferential flow. In the tree-based model, the input data, also called predictors, are used to construct test functions in the internal decision nodes. The relative 5% arrival times are the values that go to the terminal nodes.

In this study, I called predictors to the collected data selected to grow the random forest and predict the relative 5% arrival times. Continuous data were discretized into many predictors. For example, the clay content was discretized to 99 predictors as [C>0.01], [C>0.02] … [C>0.98], [C>0.99]. The combination of discretized predictors and categorical predictors gave a total of 526 predictors (**p**). These predictors are Boolean predictors because they divide the dataset into two groups: TRUE (the ones which comply with the predictor) and FALSE (the ones which does not meet the predictor). Table 2 gives an overview on the predictors.

Table 2. Overview of predictors for the dataset.

Predictor	Abbreviation	Units	Range	Type of predictor
Clay content	C	$g \cdot g^{-1}$	[0 .. 1]	Soil property
Silt content	U	$g \cdot g^{-1}$	[0 .. 1]	
Sand content	S	$g \cdot g^{-1}$	[0 .. 1]	
Bulk density	ρ	$g \cdot cm^{-3}$	[0.4 .. 2]	
Organic carbon	OC	$g \cdot g^{-1}$	[0.001 .. 0.501]	
Hassink-Dexter index (clay content /	$n_H D$	-	[0.1 .. 1000]	

organic carbon)				
WRB soil horizon letter		-	A ; B ; C ; E; A; pedogenically unaltered parent material	
Column length	L	cm	[3.16 .. 100]	Experimental condition
Column diameter	d	cm	[3.16 .. 100]	
Flow rate	q	$cm \cdot d^{-1}$	[0.316 .. 3160]	
Experiment pre-treatment		-	[Column was saturated from below]	
Electrical charge of tracer		-	[anionic]	
Tracer application type		-	[applied manually as a pulse ; applied as pulse with irrigation device ; applied as a step]	
Irrigation type		-	[ponding ; rotating drippers ; fixed drippers ; tension disk]	
Lower boundary condition		-	[seepage face]	
Column wall sealing		-	[soil-wall gap was sealed]	
Land use		-	[arable ; forest ; grassland]	Site factor
Land use type		-	[perennial land use]	
Site management		-	[site was tilled ; site was not tilled]	
Soil compaction		-	[site was trafficked]	

Each tree *t* of the forest was built following the same procedure. For each branch *k* of *t*, at each splitting step, a bootstrap sample of the predictors has been randomly generated. The predictors contained in the sample have been used as test-functions. Considering that I used Boolean predictors (with answer 'TRUE' or 'FALSE') each branch has been split into 2 branches. The predictor, p_k, from the bootstrap sample selected to split the branch was the one minimizing the sum of

variances of the relative 5% arrival time of the two resulting branches (Eq. 1), which is similar to an ANOVA test:

$$f_m = \frac{\sum_{y_i \in R_{T,k}} w_i \left(y_i - \mu_{T,k} \right)^2}{(n_{T,k} - 1) \sum_{y_i \in R_{T,k}} w_i} + \frac{\sum_{y_i \in R_{F,k}} w_i \left(y_i - \mu_{F,k} \right)^2}{(n_{F,k} - 1) \sum_{y_i \in R_{F,k}} w_i}$$

Equation 1. Test-function for calculating the sum of variance in the relative 5%-arrival time corresponding to predictor, p_k.

where $R_{T,k}$ are all the BTCs *i* in branch *k* for which the predictor *p* was TRUE, and $R_{F,k}$ are the BTCs for which the predictor *p* was FALSE. Moreover, $\mu_{T,k}$ is the mean value of relative 5% arrival times in branch *k* from BTCs for which predictor *p* was TRUE, while $\mu_{F,k}$ is the mean value of relative 5% arrival times in branch *k* from BTCs for which the predictor *p* was FALSE. Furthermore, y_i is the relative 5% arrival times for BTCs *i* and w_i is a weighting factor. Finally, $n_{T,k}$ and $n_{F,k}$ are the number of BTCs for which predictor *p* was TRUE and FALSE, respectively.

The chosen predictor would be the one that has a minimum value for the sum of variances. This means that the two resulting branches have relative low variance on their relative 5% arrival time values respectively, which implies similar relative 5% arrival time values for each branch. At each splitting step, the square root of the number of predictors ($\sqrt{526} \approx 23$) where compared, as recommended by Hastie et al. (2009).

A weighting factor has been introduced to reduce bias because all 560 BTCs are not equally distributed among the 59 articles. Therefore, BTCs from the same publication, which were usually obtained for identical experimental conditions and from similar soil types with similar soil properties, were assigned a lower weight. This way, I did not favour any soil types or experimental conditions. The weighting factor also includes a constraint. This constraint is that eligible predictors must contain BTCs from at least 3 different publications in both 'TRUE' and 'FALSE' resulting branches, respectively. This also implies that at least 3 BTCs have to be consistent with both resulting branches. This constitutes a pre-pruning technique, because we stop the growth of the branch of a tree when less than 6 BTCs belong to this branch.

Validation and benchmarking of the random forest

A ten-fold cross-validation was applied to validate the random forest. For each subset, a random forest was grown. To each leaf in each tree, I assigned the mean value of relative 5% arrival time of the BTCs contained in that leaf.

I also performed a benchmarking process to the random forest. The benchmarking process relates to the case where there are two different sources of data to estimate

the same variable, and tries to correct inconsistencies between the different estimates (OECD, 2002). The benchmarking set was used to evaluate the estimated relative 5% arrival times resulting from the random forest built on the BTC training data.

Predictor importance and partial dependence

The predictor importance was evaluated using two measures. Firstly, the importance was evaluated by the reduction in data variability before splitting the branch and the data variability of the two resulting branches, $\Delta var_{p\in k,t}$ (Eq. 2):

$$\Delta var_{p\in k,t} = \sum_{y_i\in R_k} y_i - \mu_k \,^2 - \sum_{y_i\in R_{T,k}} y_i - \mu_{T,k} \,^2 + \sum_{y_i\in R_{F,k}} y_i - \mu_{F,k} \,^2$$

Equation 2. Calculation for the variability reduction of resulting branches from parent branch associated to predictor p for branch *k* in tree *t*, $\Delta var_{p\in k,t}$.

Data variability was calculated as the total sum square of the difference between the values of relative 5% arrival time, y_i, of each BTC corresponding to branch *k*, and the mean value of relative 5% arrival time, μ_k, of all BTCs belonging to branch *k*. A high reduction in variability in the resulting branches with respect to the parent branch means that the predictor is able to split the data into 2 groups whose relative 5% arrival times are more homogenous than in the parent group. Thus, the more reduction of variability when splitting the branch for a predictor, the more important this predictor was. The variability reduction per predictor p per tree *t* was calculated as the sum of the variability reduction from all branches *k* in tree *t* where this predictor was chosen, Δvar_{p_t} (Eq. 3):

$$\Delta var_{p_t} = \sum_{k\in t} \Delta var_{p\in k,t}$$

Equation 3. Calculation for the variability reduction associated to predictor p for tree *t*.

where $k \in t$ is all branches *k* belonging to tree *t*. The reduction in variability per each predictor *p* for the whole random forest (RF) was expressed as the average of variability reductions from all trees *t* in the forest, Δvar_p (Eq. 4):

$$\Delta var_p = \frac{\sum_{t\in RF} \Delta var_{p_t}}{\sum_{t\in RF} t}$$

Equation 4. Calculation for the variability reduction associated to predictor p for whole RF.

where $t \in RF$ are all the tree *t* in the random forest RF. The average variability reduction of each predictor, Δvar_p, was normalised to the average of the predictor

with largest variability reduction, so that the result is expressed as a percentage (Eq. 5):

$$\%\Delta var_p = \frac{\Delta var_p}{\Delta var_{p\,\max p}} * 100$$

Equation 5. Calculation of the variability reduction normalised to the average of the predictor with largest variability reduction.

where $\Delta var_{p\,\max p}$ corresponds to the predictor with the largest average variability reduction in the whole random forest. The same weighting factor applied to grow the decision trees was also implemented into the calculation of the reduction of variability.

Moreover, the predictor importance was also evaluated by the number of times that the predictor was chosen to split branches. The more times the predictor was chosen, the more important it was.

The partial dependence of all predictors was also calculated. The partial dependence points out the direction of the predictor performance. A positive partial dependence indicates that the predictor is positively correlated with the relative 5% arrival time. That means that the relative 5% arrival times on the TRUE branch were larger than in the FALSE branch. On the contrary, negative partial dependence points out negative correlation to relative 5% arrival times, where relative 5% arrival times on the FALSE branch are larger than in the TRUE branch.

In order to evaluate the optimal number of trees to grow the forest I performed a statistical analysis on a large random forest of 15000 trees. I calculated the average reduction in variability as the importance measure of all predictors for the forest when it is composed of 1 tree, 2 trees and up to 15000 trees. As we increase the number of trees in the forest, the average reduction in variability of the predictors will vary until a certain number of trees, when it becomes stable. Moreover, I calculated the variance of these means. The variance is an estimator of the data variability. As the sample size increases, in this case the forest size, variability of the data will fluctuate until it becomes constant. Finally, I identified the 20 most important predictors of the random forest each time a new tree was added to the forest. When all these parameter values do not change substantially, adding a tree to the random forest will not change the results meaningfully. The forest size was chosen as the one for which these parameters remained stable.

I grew the random forest, performed the validation and benchmarking of the random forest, and basic visualisation of the variability reduction and partial dependence on the results using a Matlab library written by John Koestel.

Results and discussion

Analysis on optimal random forest size

Figure 4 shows the 20 most important predictors regarding the reduction in variability as the number of trees in a random forest increases up to 15000 trees. For forests with 150 trees (≈ 5 ln (#trees)) or more, the 20 most important predictors remain roughly the same. If we look at the predictor mean importance in figure 5, it is approximately at 3000 trees (≈ 8 ln (#trees)) when the mean reaches about the same value as for larger forest sizes. Finally, figure 6 displays the variance in importance as the forest size enlarges, and it approximately becomes constant around 3000 trees as well.

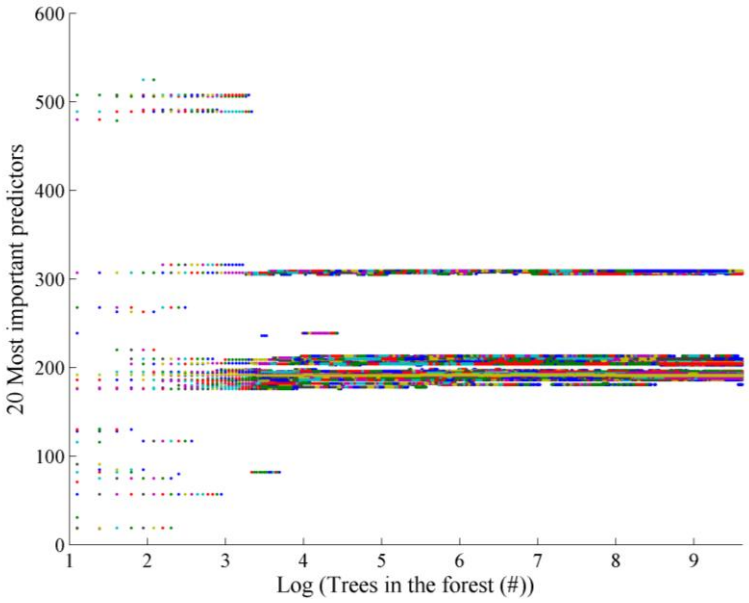


Figure 4. Plot of the 20 most important predictors (evaluated as variability reduction) as we increase the forest size up to 15,000 trees.

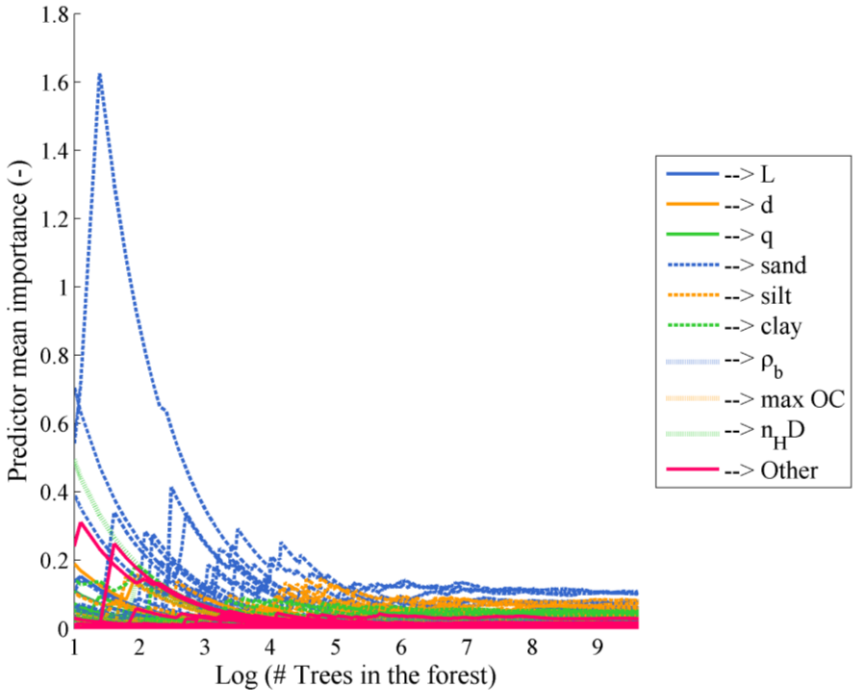


Figure 5. Plot of the predictor mean importance (evaluated as variability reduction) as we increase the forest size up to 15,000 trees.

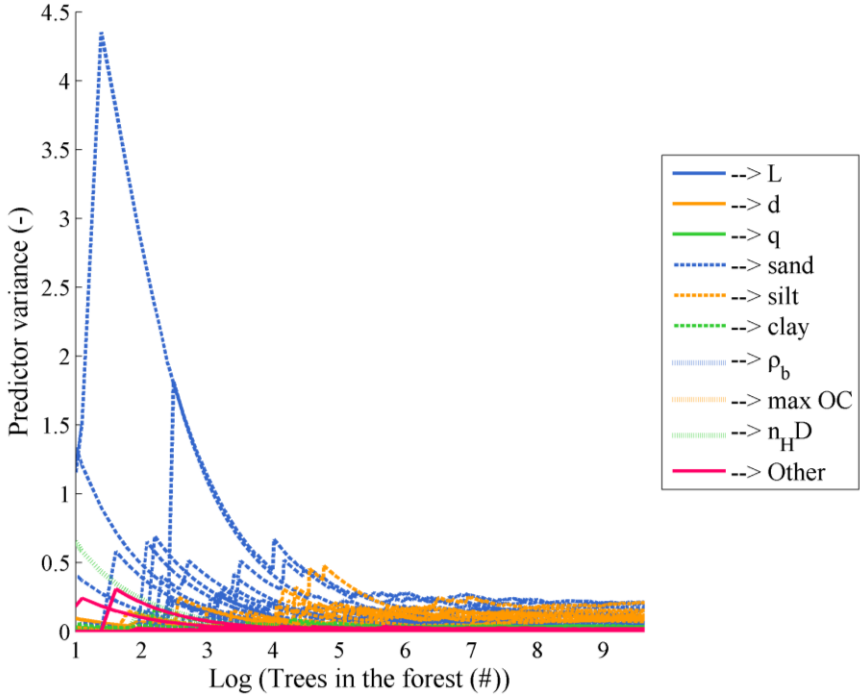


Figure 6. Plot of the predictor variance in importance (evaluated as variability reduction) as we increase the forest size up to 15,000 trees.

If we only take into account the identity of the 20 most important predictors, 150 trees would be enough to grow the random forest. However, the mean importance and its variance show that it is necessary to grow between 2000-3000 trees to get similar results regarding the values of predictor importance. Considering all 3 plots and to be on the safe side, I decided to grow a random forest of 5000 trees.

Validation and benchmarking

The cross-validation results are shown in Figure 7. The dotted line displays the line where estimated values equal real values. The blue line shows the best regression linear model that fits the data. According to this regression model, the random forest explains approximately 66% of the variation in the training data (Figure 7a). The plot shows that for small relative 5% arrival times, the random forest is overestimating its values. On the other hand, the random forest underestimates large relative 5% arrival times. This bias might be attributed to both the method and the data. If this bias is removed, the random forest can explain 88% of the variation in the training data (Figure 7b). Still,

a) b)

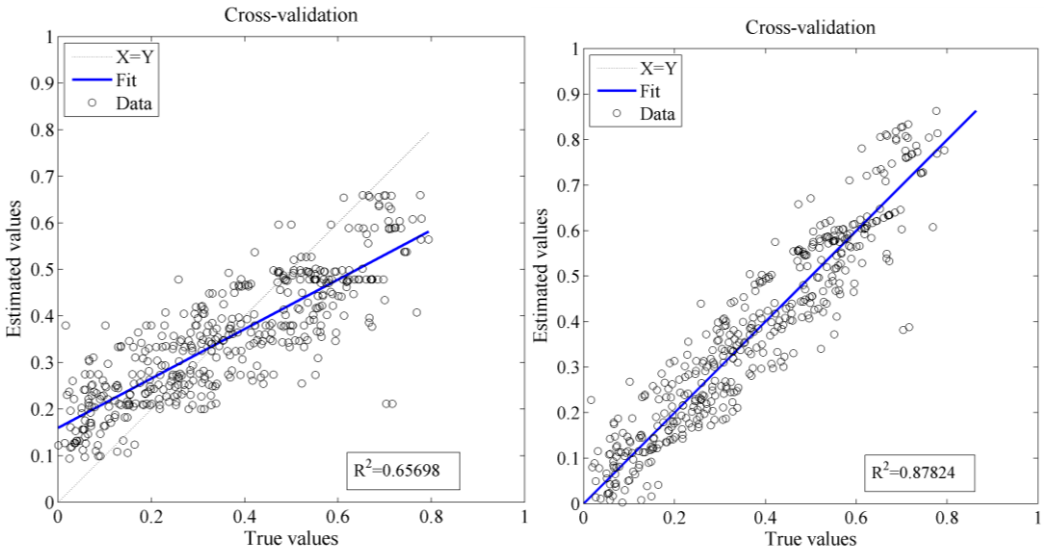


Figure 7. a) Plot of relative 5% arrival times estimated values for cross-validation against relative 5% arrival time true values, b) same plot but with bias removed from the estimated values of relative 5% arrival time.

The results from the benchmarking process reveal that the random forest is able to predict 22% of the variation of the data that has not been used in the training set (Figure 8a).

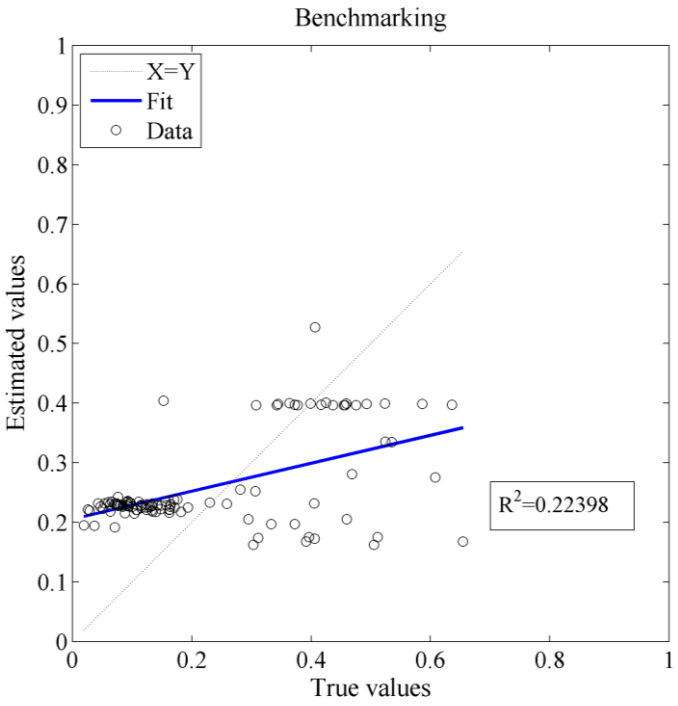


Figure 8. Plot of relative 5% arrival times estimated values for benchmarking against relative 5% arrival time true values.

These results show that the model describes the training data well, but it is not so accurate when describing data from new sources as it only explains 27% of the variation in the benchmarking set. In order to validate any model, it is essential to use a data set that has not been used to calibrate it. The cross-validation technique does that, but when randomly choosing data to generate the 10 different subsets, it takes data without considering which article they belong to. Therefore, it might happen that 2 or more different subsets contain data from the same article. It may also occur that the subset used to validate the random forest contains data from sources used in the other 9 subsets. On the contrary, the benchmarking set is made up of data from 7 articles that have not been used to build the random forest. Experiments from the same source are usually performed in the same way, and in soils with similar properties and land use. This fact can explain the better predictions of relative 5% arrival time from the cross-validation process, as the validation set contains data similar to the ones used to build the random forest.

In other words, the generalisation error is larger for the benchmarking procedure because the benchmarking set may have characteristics that were not present in the training set, so the model fails to fit BTCs with these characteristics. In order to reduce this error, it is necessary to increase the sample size, which in this case would be the number of experiments in the database. It could also be interesting to build random forests according to specific characteristics, such as soil types, initial and boundary conditions or land uses.

When colouring the data in the benchmarking according to the publication (Figure 9), we can observe that the data is clustered according to this feature.

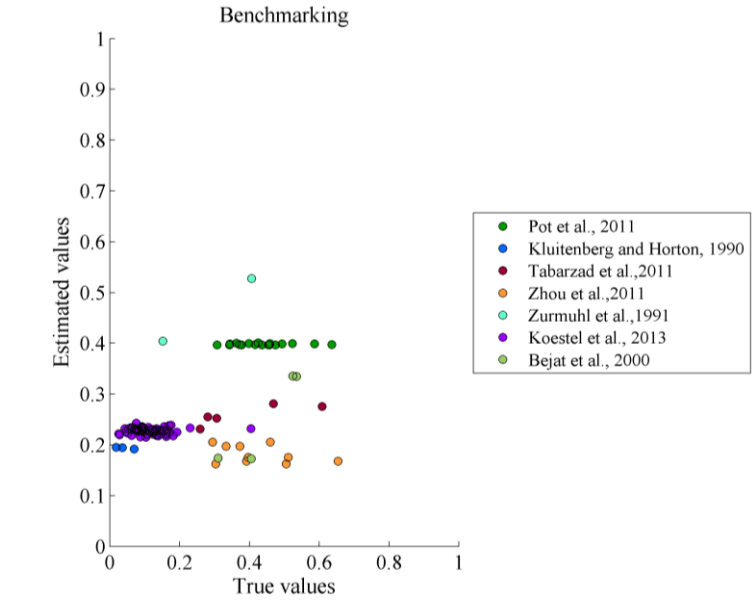


Figure 9. Plot of relative 5% arrival times estimated values for benchmarking against relative 5% arrival time true values according to publication.

The benchmarking experiments in the same publication have differences mainly regarding soil properties and flow rate, and therefore their relative 5% arrival times vary as well, but the estimated values are very similar. For instance, Kluitenberg and Horton (1990), Tabarzad et al. (2011), Zhou et al. (2011) had large differences in flow rate for their respective experiments. All publications show differences in organic carbon content and bulk density, and only Pot et al. 2011 does not contain experiments performed in soils with different textures. It seems like the model is not fitting some variation in the data regarding soil properties and flow rate. Concerning soil texture, the training set seems not to cover some texture combinations present in the benchmarking set. For instance, the combination of approximately 0.5 sand, 0.3 silt and 0.2 clay contents is very common in the benchmarking set (Figure 10b), corresponding to Koestel et al. (2013), but it is not so frequent in the training set (Figure 10a). Therefore, small differences in texture around these values are not well described by the model and consequently it predicts similar relative 5% arrival times (see Koestel el al. (2013) in Figure 9). In addition, water saturation had been found to be positively correlated to the strength of preferential flow in the study by Koestel et al. (2013). However, soil saturation was often not stated in the articles so it was not included as a predictor. The inclusion of this variable as a predictor might improve the predictions of the random forest.

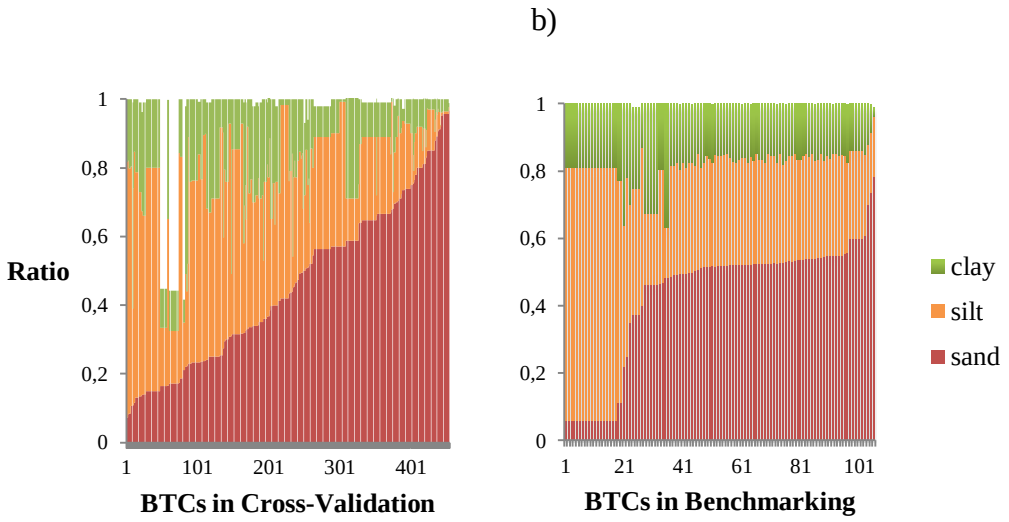


Figure 10. Sand, silt and clay ratios for the investigated BTC experiments. a) 454 BTCs included in the cross-validation; b) 106 BTCs included in the benchmarking set.

Furthermore, the training set contains values of flow rate mainly below 50 $cm \cdot d^{-1}$, which might lead to less accurate predictions for extreme values. Flow rates in Kluitenberg and Horton (1990) experiments are 1392, 672 and 288 $cm \cdot d^{-1}$ (encircled in red in Figure 11), much larger than 50 $cm \cdot d^{-1}$. This may explain the inaccurate predictions for their relative 5% arrival times.

Finally, land use information for Zhou et al. (2011) is unknown, which might account for their relative 5% arrival time's erroneous estimations.

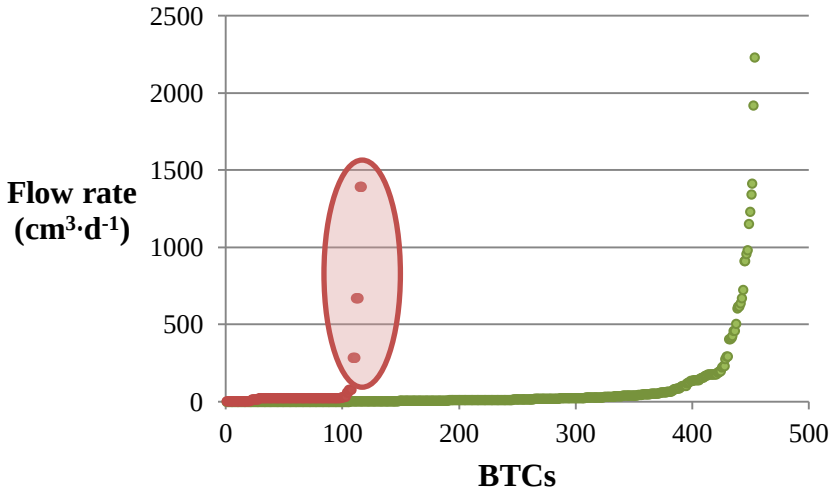


Figure 11. Flow rate corresponding to BTCs in training set (green) and BTCs in the benchmarking set (red).

The variation in true values of relative 5% arrival times between publications may also stem from the difficulties of performing BTC experiments. BTC experiments performed ostensibly in the same way in the same soil column may lead to different results because of unexpected incidents such as air bubbles occasionally blocking the outflow tubes. Different results may also come from problems with devices or instruments used to measure the tracer concentration at the outlet.

Predictor importance analyses

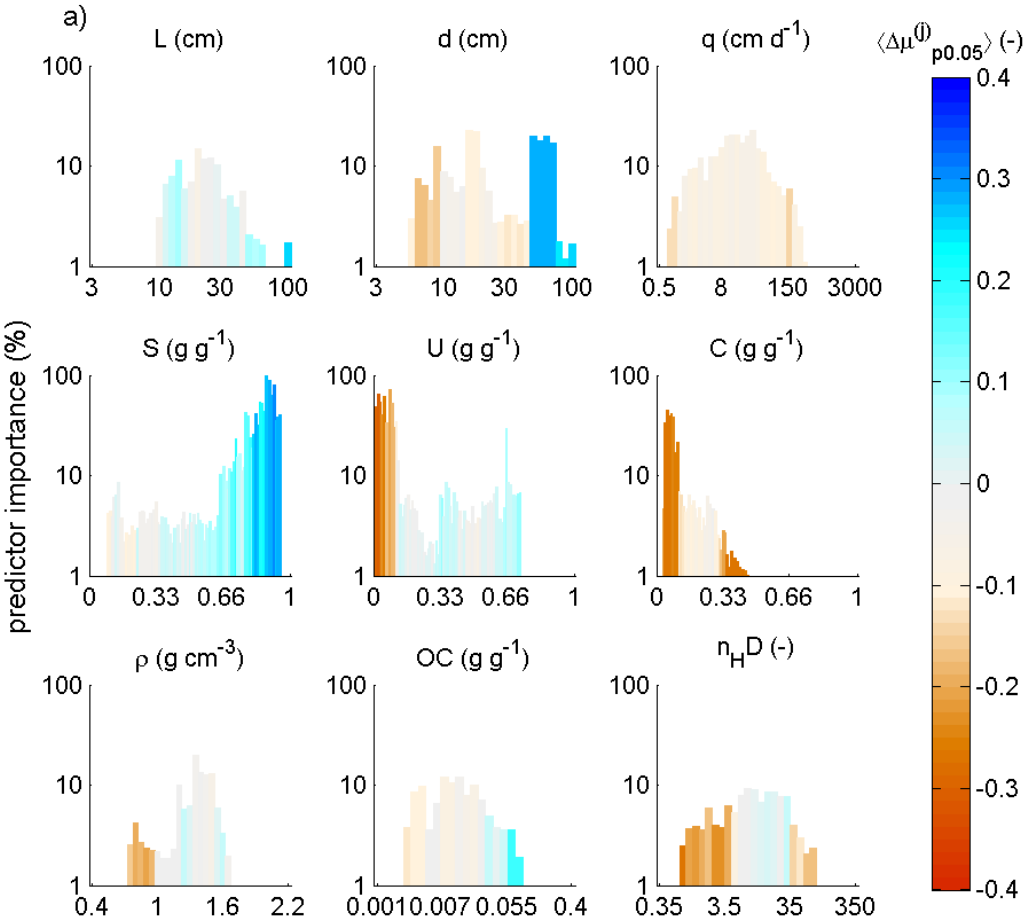
The variability reduction as an importance measure and the partial dependence of the different analysed predictors are shown in Figure 12. Regarding soil properties, texture is the most important predictor. Texture had already been considered as an important feature for predicting preferential flow (Jarvis, 2007; Koestel et al., 2012; Koestel et al., 2013). In my study, sand contents between 0.80 and 0.92 have an importance above 50% to predict preferential flow. Specifically, sand contents from 0.88 to 0.90 reach the highest importance among all predictors, with values above 90%. These large sand contents are positively correlated with the relative 5% arrival time, which implies weak preferential flow. Sandy soils tend to have a weak aggregate structure. Conversely, low silt and clay contents, from 0.01 up to 0.11 and from 0.04 up to 0.08, respectively, are strongly negatively related to the relative 5% arrival time with importance above 40%. This means that a minimum content of 10% silt and clay is necessary for preferential flow to occur. For contents over these values, preferential flow is strong and these parameters are no longer important for predicting preferential flow. This agrees with other studies stating that minimum clay contents are necessary for preferential flow to occur (Koestel et al., 2012).

The remaining investigated soil properties were found to be moderately important. Bulk density and organic carbon have low importance on predicting preferential flow, and their partial dependence is neutral. This can imply positive partial dependence for some trees and negative partial dependence in others. However, low values of bulk density were negatively correlated to the relative 5% arrival times, meaning stronger preferential flow. Bulk density is correlated to soil porosity and has also been included by Stolf et al. (2011) in a PTF to estimate macroporosity in soils. Their study shows that bulk density is negatively correlated with macroporosity. Hence, we could expect this parameter to be important in order to explain macropore flow. Small values of bulk density imply higher porosities, which may mean more macropores and might imply greater macropore flow. However, bulk density does not say anything about connectivity between macropores, which might explain its low importance to predict soil's susceptibility to preferential flow in this study. Concerning the Hassink-Dexter index, its importance is also moderate, despite a negative dependence for values between 1 and 3.5. Since clay content has a high importance and organic carbon a low importance, the Hassink-Dexter index reaches a moderate importance. The

negative partial dependence is related to the negative partial dependence of low clay content predictors.

Regarding experimental conditions (Figure 12a), column length does not present a high importance in predicting preferential flow. However, column diameters around 16 cm reach importance values of over 20%. This may be explained by the fact that most experiments were conducted in columns with diameter values between 5.5 and 25 cm. Moreover, large column diameters are strongly positively correlated to the relative 5% arrival time. Flow rate was found to be rather important for predicting preferential flow. As Jarvis (2007) had already mentioned, macropore flow is strongly dependent on the surface boundary conditions. Intermediate values of flow rate reached moderate importance, above 20% when the flow rate was about 10-30 $cm^3 \cdot d^{-1}$, but with neutral partial dependence, which means that for some cases it was positively correlated to preferential flow, and for other it had a negative association. This agrees with what was stated in the literature review, that boundary conditions have a complex effect on preferential flow, and so has the flow rate. The remaining experimental conditions are represented in Figure 10c. Among those, the most important predictor is the irrigation using fixed drippers. This irrigation technique has a 28% importance and is moderately positively correlated to preferential flow. This might be explained by the fact that fixed drippers was the most common irrigation technique. Column venting reached 17% of importance. Irrigation techniques of tension disk and ponding, manual and pulse application of tracer, and seepage face have similar importance percentages (around 10%) and partial dependences around neutrality. The rest of experimental conditions have less importance to predict preferential flow.

With respect to site factors, illustrated in Figure 12b, all predictors show low importance percentages, even no importance for forest as land use, and almost neutral partial dependences. Only the non-tillage practice achieved an importance over 10%. This implies that, in this study, they are not important factors for predicting preferential flow. However, most articles did not give extensive information on management factors. Furthermore, most studies had been performed in soils from agricultural sites, so not much data was included on preferential flow in forest and other land uses. Therefore, lack of data for site factors might account for their low significance in this study.


a)
L (cm)
d (cm)
q (cm d-1)
S (g g-1)
U (g g-1)
C (g g-1)
ρ (g cm-3)
OC (g g-1)
nHD (-)
predictor importance (%)
100
10
1
3
10
30
100
0.5
8
150
3000
0
0.33
0.66
1
0.4
1
1.6
2.2
0.001
0.007
0.055
0.4
0.35
3.5
35
350
⟨Δμ(j)p0.05⟩ (-)
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4

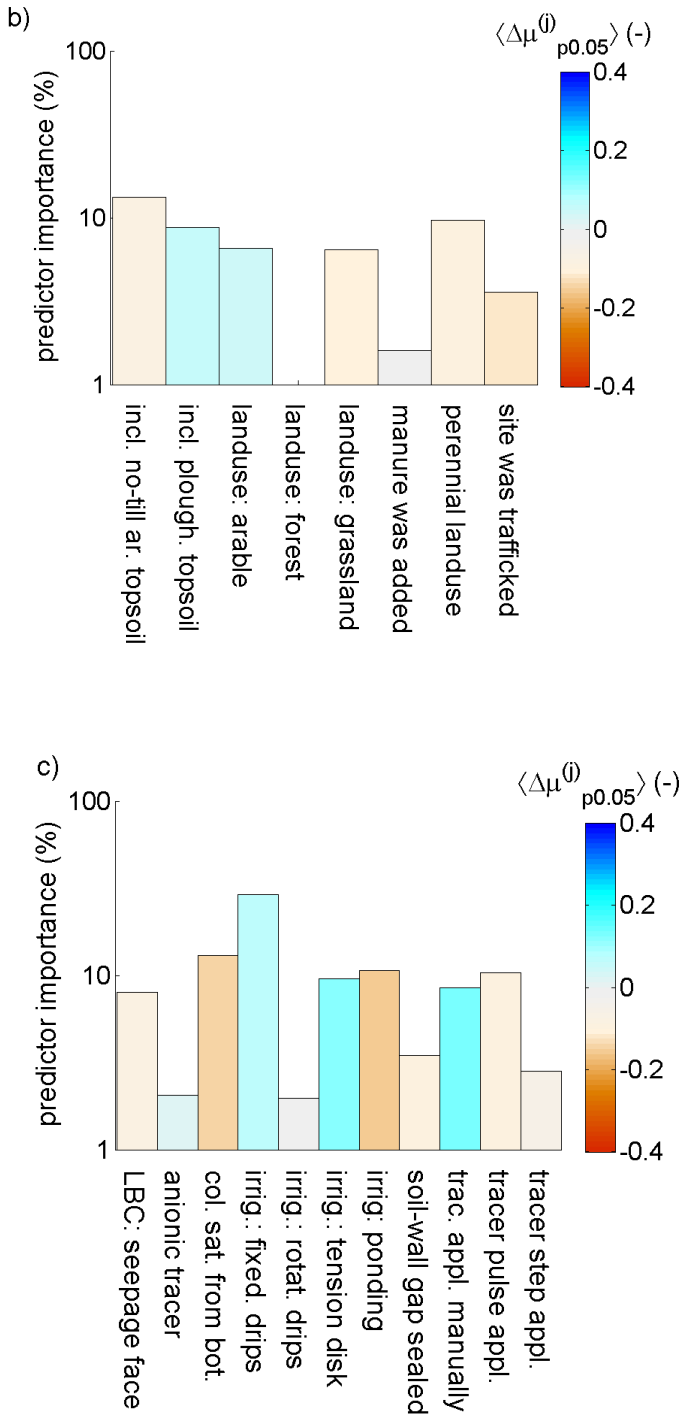


Figure 12. Predictor importance (evaluated as the reduction in variability) and partial dependence: a) soil property predictors (S, U, C, ρ, OC, n_HD) and some experimental conditions predictors (L, d, q,) b) site factor predictors, c) other experimental conditions predictors. See table 2 for abbreviations.

Further analysis of predictor importance

Finally, I created several plots in order to further analyse the predictor importance adding the number of trees where predictors were chosen to split branches as a second measure of the predictor importance. In the first place, I created histograms of the reduction of variability of the sibling branches respect to the parent branch of each predictor with respect the number of trees where they were chosen to split branches. The abscissa (x axis) represented the variability reduction from 0 to 7 (maximum value = 6.1571), which was divided into 350 equidistant classes, and the number of trees was placed into the ordinate (y axis). The histograms show the frequency distribution of the predictor importance. If we have a look at some of the histograms on predictor importance we can observe different patterns.

Sand and silt predictors with high variability reductions show similar histograms. Their histograms (Figures 13 and 14) are skewed to the left and with highest values around 1-3 variability reduction, meaning that they reach a very high relative variability reduction in most of the trees were they have been chosen to split a branch. They also show extreme variability reduction values of more than 5.

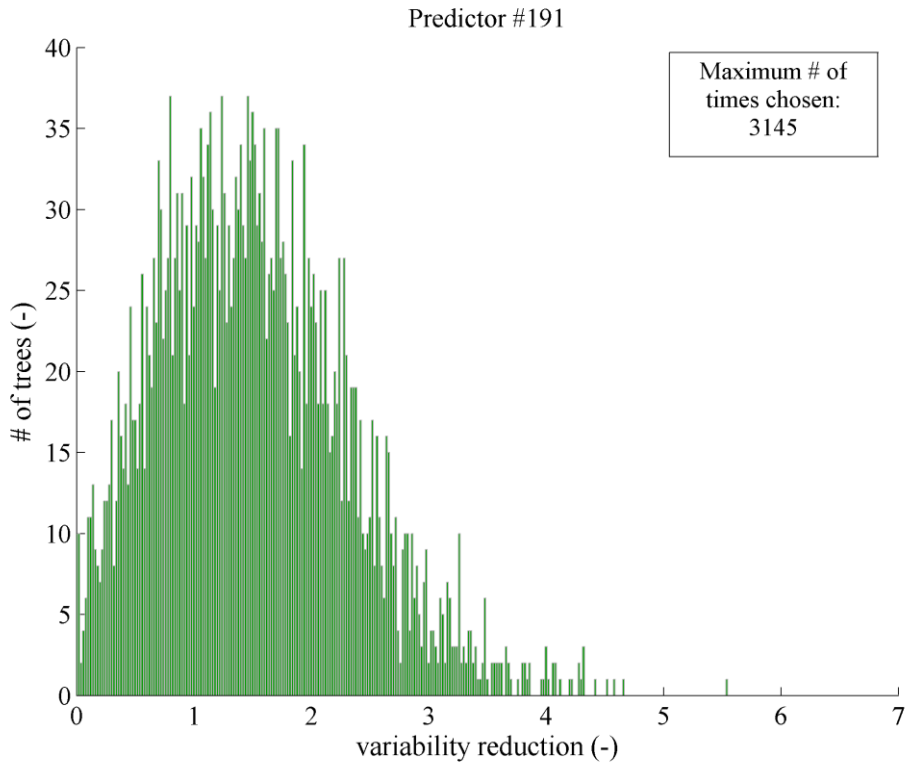


Figure 13. Histogram of the variability reduction per tree of predictor #191 (sand content>0.88 $g \cdot g^{-1}$).

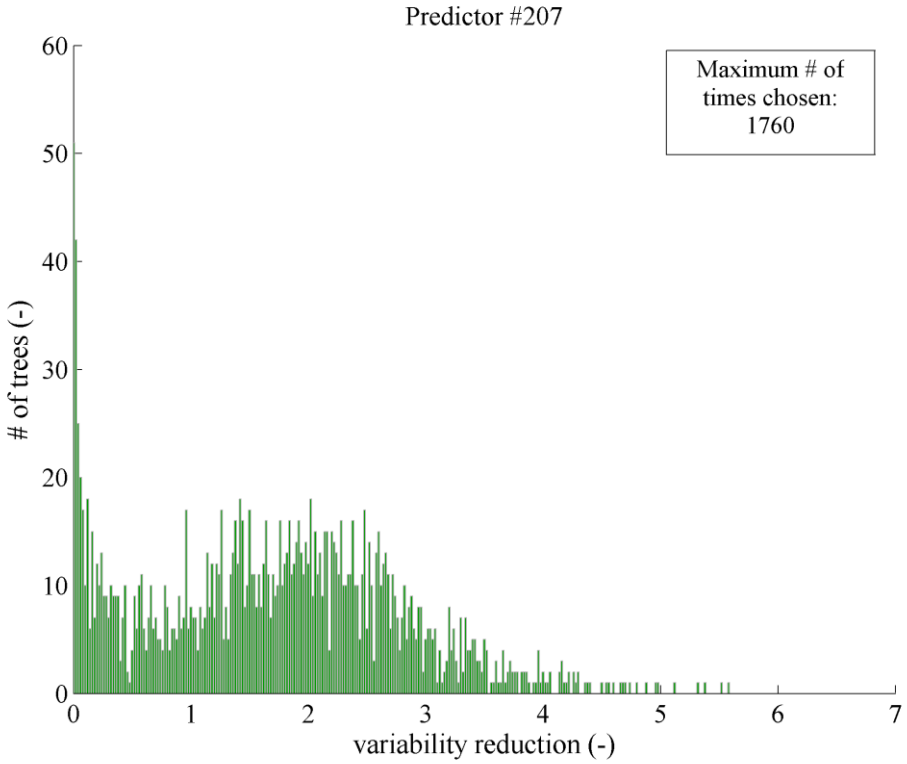


Figure 14. Histogram of the variability reduction per tree of predictor #207 (silt content>0.05 $g \cdot g^{-1}$).

Clay predictors with high variability reduction present different histograms. Figure 15 is an example of clay predictor importance histogram. Although their frequency for large relative variability reduction is also high (notice the difference in scale in y axis), they show high frequency at low relative values as well. This suggests that sand and silt predictors with high reduction in variability are always very important to predict preferential flow in all trees of the random forest where they are chosen. However, averagely important clay predictors are highly important in some trees but only moderately to little important in other trees of the forest. Moreover, clay predictors are chosen more often than silt and sand predictors.

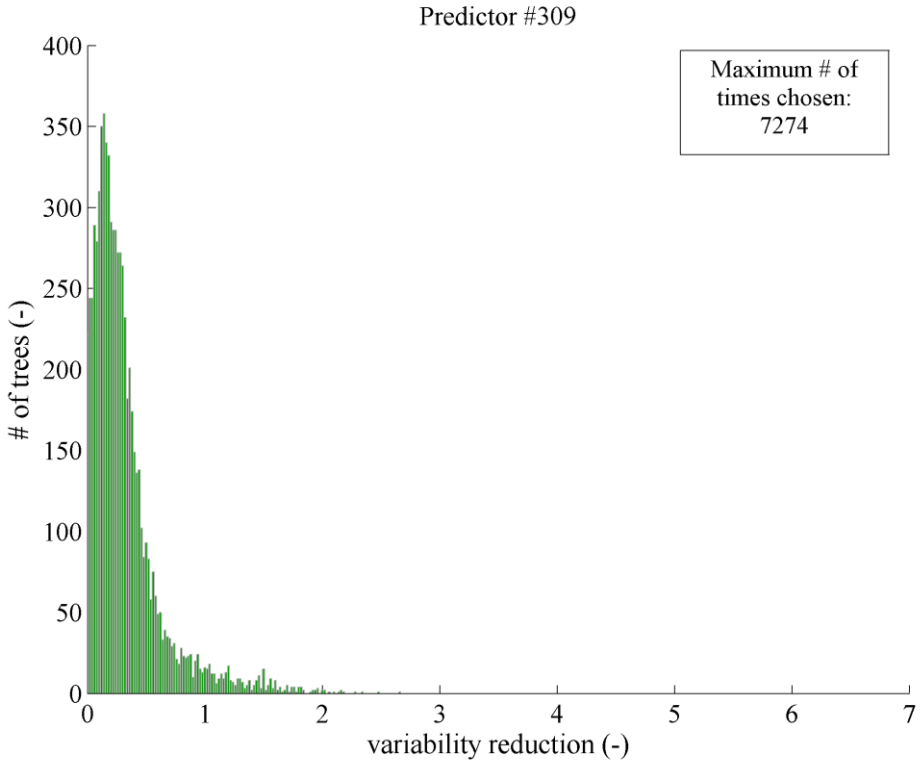


Figure 15. Histogram of the variability reduction per tree of predictor #309 (clay content>0.08 $g \cdot g^{-1}$).

Moderately important predictors such as flow rate around 20 $cm^3 \cdot d^{-1}$ and bulk densities of 1.30-1.40 $g \cdot cm^{-3}$ show histograms similar to clay predictors. Figure 16 present a plot strongly skewed to the left but it has been chosen to split branches in many trees (notice the difference in scale in y axis again). This fact increases the mean importance of predictors with the same kind of histogram.

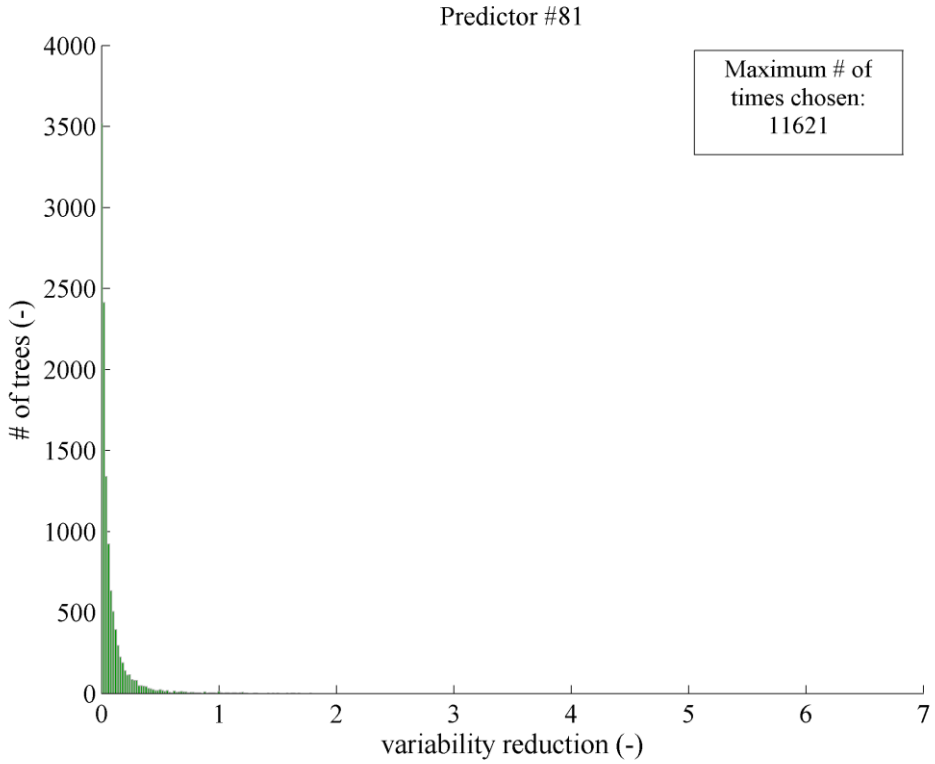


Figure 16. Histogram of the variability reduction per tree of predictor #81 (flow rate>20 $cm^3 \cdot d^{-1}$).

Predictors with low importance have, essentially, two types of histogram curve. Most unimportant predictors are strongly skewed to the left and with low number of trees on the y axis (Figure 17; notice the change in scale in y axis). Those are predictors that are clearly not important for estimating preferential flow. On the other hand, there is a group of predictors that follow the pattern showed in Figure 18, a little-skewed histogram with also low number of trees. This indicates that while their mean importance is moderate to low, it can reach high percentages in some trees. This is the case for less than 0.005 $g \cdot g^{-1}$ organic carbon contents and bulk densities around 0.75 and 0.95 $g \cdot cm^{-3}$.

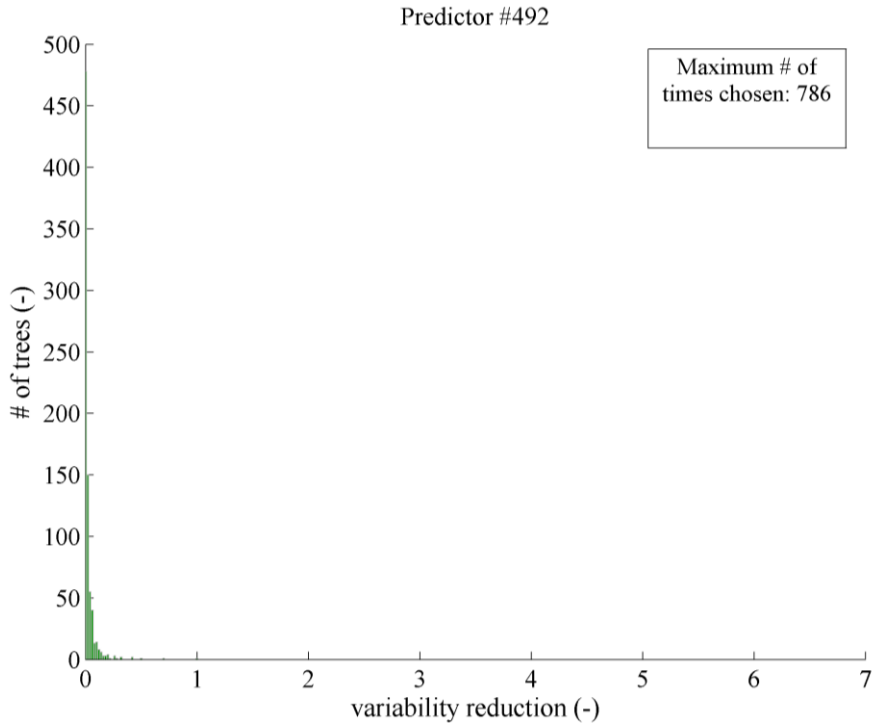


Figure 17. Histogram of the variability reduction per tree of predictor #492 (Hassink-Dexter index>100).

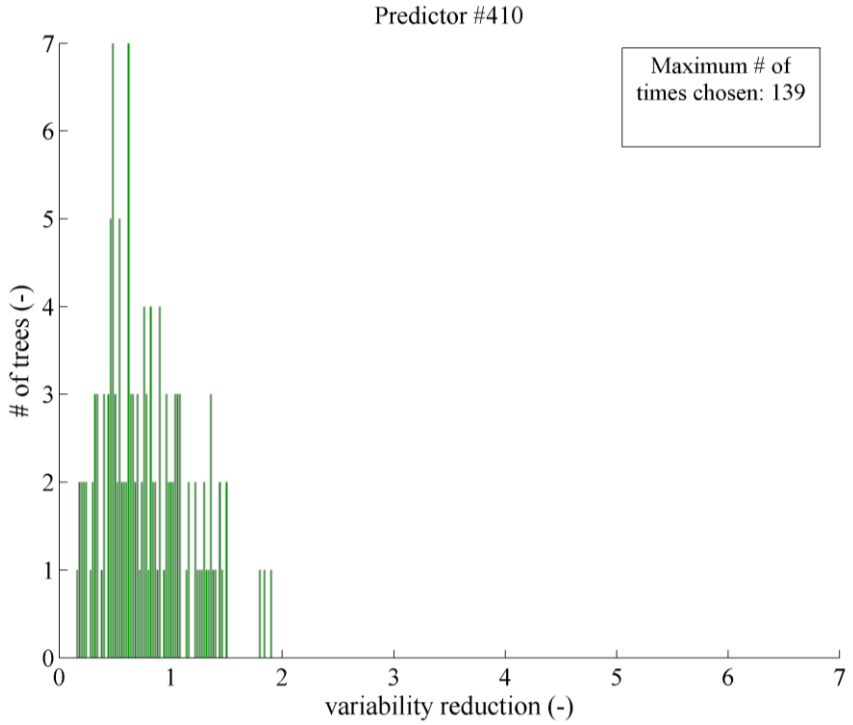


Figure 18. Histogram of the variability reduction per tree of predictor #410 (bulk density>0.85 $g \cdot cm^{-1}$).

Moreover, I plotted the variability reduction of the predictor against the number of times the same predictor was chosen to further investigate the predictor importance.

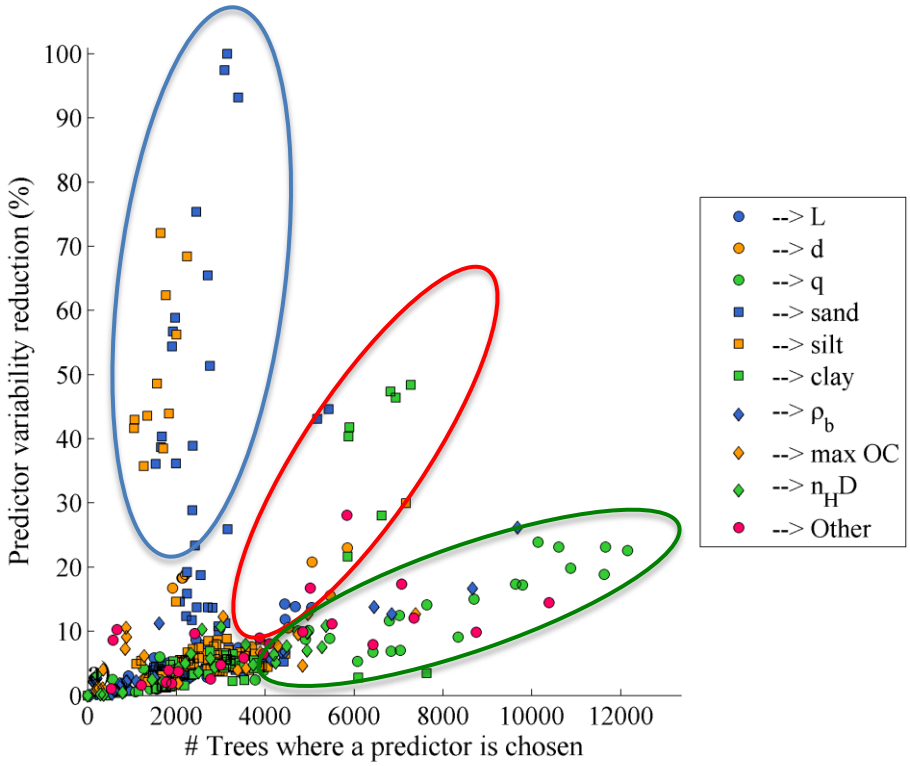


Figure 19. Plot of the predictor mean variability reduction (%) versus the number of times the same predictor is chosen.

The data highlighted in blue corresponds mainly to sand and silt predictors. Figure 19 shows that even if these predictors are chosen few times compared to other predictors, they still have a high importance in predicting the relative 5% arrival time. In contrast, flow rate predictors have generally a low mean variability reduction even when they are chosen large number of times. Finally, there is an intermediate tendency where the variability reduction increases with the number of times predictors are chosen and which corresponds, primarily, to some clay and sand contents.

These trends are also seen in Figure 20. To finalise the analysis on the predictor importance, I also plotted the number of times a predictor is chosen and the variability reduction against each group of predictors. Figure 20 displays the number of times a predictor is chosen and its variability reduction on the ordinates so we can compare these parameters between groups of predictors. For most predictors, both parameters follow the same pattern, so that when the variability reduction values increase the number of times also rises. The predictors describing

the diameter (d) are a good example of this behaviour and one can see that both curves have a similar shape. In the case of flow rate predictors, although the shape of both curves resembles one another, the increase in variability reduction is not as large as the one in number of times these predictors are chosen. The same happens with bulk density predictors. This is consistent with figure 19, which already showed the same behaviour. Furthermore, in the case of sand predictors, when increasing the sand content to the highest values (predictors #160-190), the reduction in variability grows drastically although the times that these predictors are chosen remain low. The same happens for the first silt predictors (predictors #203-215). This is also congruous with the results from figure 19. From the rest of experimental conditions, column venting, ponding as irrigation and immobile drippers were the most often chosen. Finally, the land use factors most chosen were the "arable land use", "grassland land use", "arable but not tilled" and "perennial land use".

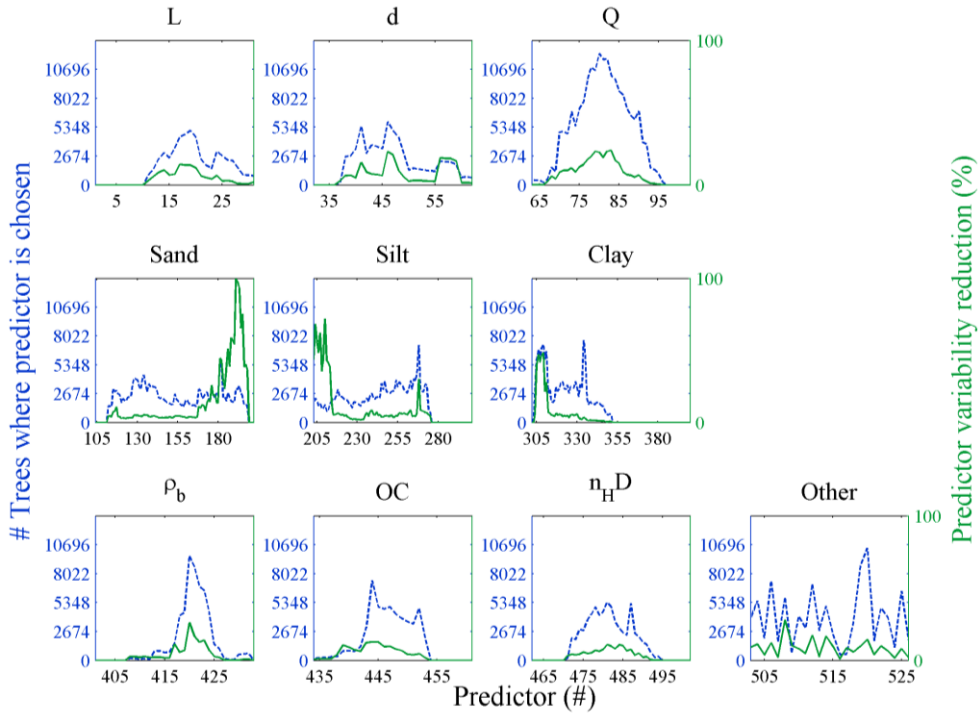


Figure 20. Plot of the number of times a predictor is chosen (on the left y axis) and the predictor mean importance (%) (on the right y axis) against the number of the predictor.

It would be logical to think that the more important a predictor is (explained by the reduction in variability), the more times it is selected to split a branch. However, figures 19 and 20 show two particular patterns that are contrary to this reasoning. There are some predictors, such as intermediate values of flow rate, which are chosen many times to split branches but reach moderate variability

reduction values. On the contrary, some sand and silt predictors are chosen a fewer number of times but attain higher variability reduction.

If we look into the way the model selects a predictor to split a branch, we see that the selected predictor is the one that minimizes the sum of variance of the resulting branches. So the splitting process does not consider the variance in the parent branch. On the other hand, the variability reduction is between the variability from the parent branch and the variability of the resulting two branches, being more important the larger the reduction in variability is.

Variance increases with sample size, which is larger at the root of the tree. Therefore, I assume that the variability in data is greater at the root of the tree and, as the tree is growing, variability in data in the resulting branches decreases. Given this assumption, I interpret these results so that the predictors with highest variability reductions are usually used to split branches of the tree in the beginning of the splitting process. Consequently, they are chosen fewer times because there are fewer branches at the root of the tree. On the other hand, predictors such as flow rate are mostly employed to split branches closer to the leaves, so that they are chosen more times but the reduction in variability is not that large.

Therefore, generally, large sand and small silt and clay contents (texture properties) indicate a low susceptibility of a soil to preferential flow in general terms. Then, given the texture, other characteristics such as flow rate, bulk density, organic carbon content and irrigation techniques, would describe the strength of preferential flow for the specified soil in more detail. This explains why these characteristics have more neutral partial dependences; as for some textures they are related to stronger preferential flow and to weaker preferential flow for others.

Conclusions

Given the complex interaction among factors influencing soil's susceptibility to preferential flow, we need statistical techniques capable of capturing this complexity. It seems like a random forest is a powerful tool that may help us to present a coherent explanation for the importance of these factors and to build PTF to predict the susceptibility of soils to preferential flow.

The cross-validation of the random forest reported more successful predictions of relative 5% arrival times than the benchmarking process. This was probably caused by the inclusion of similar soil properties and experimental conditions in the validation set with respect to the training set, which were not included in the benchmarking set.

Soil texture was identified as a dominant characteristic to predict soil susceptibility to preferential flow for the analysed BTC experiments. Furthermore,

the most important experimental conditions in this study are related to hydrologic boundary conditions, which are exemplified by flow rate and irrigation techniques. This study suggests that texture could be used to coarsely classify soils into general classes of susceptibility, which later would be refined using other site factors and soil characteristics, such as flow rate, bulk density, organic carbon and irrigation techniques.

One of the limitations of this study is the difficulty to find articles that give comprehensive information on soil properties, experimental conditions and land use factors. This has different implications. On the one hand, the correctness with which the random forest predicts the relative 5% arrival times is limited by the sample size of BTC experiments. Enlarging the database is necessary in order to achieve smaller generalisation errors of the random forest, which becomes a harder task if articles provide insufficient information. On the other hand, the lack of specific data, such as land use factors and water saturation, may lead to an underestimation of the importance of these predictors and overestimate the importance of those that are more abundant in the literature. In conclusion, and as stated by Jarvis et al. (2012), presenting a coherent picture of flow and solute transport in the dominant soils of the world is restricted by the lack of extensive information.

Acknowledgement

I very much thank my supervisor John Koestel for all the time dedicated to support me on the different subjects of this thesis.

References

Alpaydin, E., 2004. Introduction to machine learning. MIT Press, Cambridge, Mass.

Andrew Ng, S.U., 2009. Video Lecture 2 - An application of supervised learning - Autonomous deriving, Stanford Engineering Everywhere CS229 - Machine Learning. University of Stanford.

Bejat, L., Perfect, E., Quisenberry, V.L., Coyne, M.S., Haszler, G.R., 2000. Solute transport as related to soil structure in unsaturated intact soil blocks. Soil Sci Soc Am J 64, 818.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Hastie, T., Tibshirani, R., Friedman, J.J.H., 2009. The elements of statistical learning, Second edition. ed. Springer New York.

Heijs, A.W.J., Ritsema, C.J., Dekker, L.W., 1996. Three-dimensional visualization of preferential flow patterns in two soils. Geoderma 70, 101–116.

Hendrickx, J.M.H., Flury, M., 2001. Uniform and preferential flow mechanisms in the vadose zone, in: Conceptual Models of Flow and Transport in the Fractured Vadose Zone. The National Academies Press, pp. 149–187.

Jarvis, N.J., 2007. A review of non-equilibrium water flow and solute transport in soil macropores: principles, controlling factors and consequences for water quality. Eur. J. Soil Sci. 58, 523–546.

Jarvis, N.J., Moeys, J., Hollis, J.M., Reichenberger, S., Lindahl, A.M.L., Dubus, I.G., 2009. A conceptual model of soil susceptibility to macropore flow. Vadose Zone J. 8, 902.

Jarvis, N.J., Moeys, J., Koestel, J., Hollis, J.M., 2012. Preferential flow in a pedological perspective, in: Hydropedology. Elsevier, pp. 75–120.

Jury, W.A., Flühler, H., 1992. Transport of chemicals through soil: mechanisms, models, and field applications, in: Donald L. Sparks (Ed.), Advances in Agronomy. Academic Press, pp. 141–201.

Kluitenberg, G.J., Horton, R., 1990. Effect of solute application method on preferential transport of solutes in soil. Geoderma 46, 283–297.

Knudby, C., Carrera, J., 2005. On the relationship between indicators of geostatistical, flow and transport connectivity. Adv. Water Resour. 28, 405–421.

Koestel, J.K., Moeys, J., Jarvis, N.J., 2012. Meta-analysis of the effects of soil properties, site factors and experimental conditions on solute transport. Hydrol. Earth Syst. Sci. 16, 1647–1665.

Koestel, J.K., Norgaard, T., Luong, N.M., Vendelboe, A.L., Moldrup, P., Jarvis, N.J., Lamandé, M., Iversen, B.V., Wollesen de Jonge, L., 2013. Links between soil properties and steady-state solute transport through cultivated topsoil at the field scale. Water Resour. Res. 49, 790–807.

Nielsen, D.R., van Genuchten, M.T., Biggar, J.W., 1986. Water flow and solute transport processes in the unsaturated zone. Water Resour. Res. 22, 89–108.

Nilsson, N.J., 1998. Introduction to machine learning. An early draft of a proposed textbook. Robotics Laboratory, Department of Computer Science, Stanford University.

OECD, (Organisation for Economic Co-operation and Development), 2002. Glossary of statistical terms.

Pot, V., Benoit, P., Etievant, V., Bernet, N., Labat, C., Coquet, Y., Houot, S., 2011. Effects of tillage practice and repeated urban compost application on bromide and isoproturon transport in a loamy Albeluvisol. Eur. J. Soil Sci. 62, 797–810.

Samuel, A.L., 1959. Some studies in machine learning using the game of checkers [draughts]. IBM J. Res. Dev. 3, 210–229.

Semola, A., Vishwanathan, S.V., 2008. Introduction to machine learning. Cambridge University Press.

Shaw, J.N., West, L.T., Radcliffe, D.E., Bosch, D.D., 2000. Preferential flow and pedotransfer functions for transport properties in sandy Kandiudults. Soil Sci. Soc. Am. J. 64, 670–678.

Sonneveld, M.P.W., Backx, M.A.H.M., Bouma, J., 2003. Simulation of soil water regimes including pedotransfer functions and land-use related preferential flow. Geoderma 112, 97–110.

Stolf, R., Thurler, A. de M., Oliveira, O., Bacchi, S., Reichardt, K., 2011. Method to estimate soil macroporosity and microporosity based on sand content and bulk density. Rev. Bras. Cienc. Solo 35, 447–459.

Tabarzad, A., Sepaskhah, A.R., Farnoud, T., 2011. Determination of chemical transport properties for different textures of undisturbed soils. Arch. Agron. Soil Sci. 57, 915–930.

Van As, H., van Dusschoten, D., 1997. NMR methods for imaging of transport processes in micro-porous systems. Nmr Soil Sci. 80, 389–403.

Wösten, J.H.M., Pachepsky, Y.A., Rawls, W.J., 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. J. Hydrol. 251, 123–150.

Zhou, B., Jiang, Y., Wang, Q., Shao, M., 2011. Chloride transport in undisturbed soil columns of the loess Plateau. Afr. J. Agric. Res. 6, 4807–4815.

Zurmühl, T., Durner, W., Herrmann, R., 1991. Transport of phthalate-esters in undisturbed and unsaturated soil columns. J. Contam. Hydrol. 8, 111–133.