Characterising flow patterns in soils by feature extraction and multiple consensus clustering

Christina Bogner[a,*], Baltasar Trancón y Widemann[a,b], Holger Lange[c]

[a] *Ecological Modelling, BayCEER, University of Bayreuth, Dr.-Hans-Frisch-Straße 1–3, 95448 Bayreuth, Germany*

[b] *Programming Languages and Compiler Technology, University of Ilmenau, Ehrenbergstr. 29, 98693 Ilmenau, Germany*

[c] *Norsk Institutt for Skog og Landskap, P.O. Box 115, 1431 Ås, Norway*

Abstract

The quality of surface water and groundwater is closely related to flow paths in the vadose zone. Therefore, dye tracer studies are often carried out to visualise flow patterns in soils. These experiments provide images of stained soil profiles and their evaluation demands knowledge in hydrology as well as in image analysis and statistics. The classical analysis consists of image classification in stained and non-stained parts and calculation of the dye coverage (i.e. the proportion of staining). The variation of this quantity with depth is interpreted to identify dominant flow types. While some feature extraction from images of dye-stained profiles is necessary, restricting the analysis to the dye coverage alone might miss important information. In our study we propose to use several index functions to extract different (ideally complementary) features. We associate each image row with a feature vector (i.e. a certain number of image function values) and use these features to cluster the image rows to identify similar image areas. Because images of stained profiles might have different reasonable clusterings, we calculate multiple consensus clusterings. Experts can explore these different solutions and base their interpretation of predominant flow type on quantitative (objective) criteria.

Keywords: Feature extraction, multiple consensus clustering, dye patterns, dye tracer study, vadose zone hydrology

1. Introduction

Groundwater pollution by agrochemicals, degradation of soil quality and pollution of aquatic ecosystems by agricultural drainage waters have become major issues in the last decades (Jarvis, 2007). Flow paths in soils are closely related to these problems because water that infiltrates and eventually reaches groundwater has to pass the unsaturated (vadose) zone first. Two rather different flow types can be distinguished: water can percolate slowly through the soil matrix (uniform flow) or move rapidly through preferred pathways and

*Corresponding author: Tel.: +49 921 555655; Fax: +49 921 555799

Email address: christina.bogner@uni-bayreuth.de (Christina Bogner)

This is an author's post-print version of the article published in Ecological Informatics, **15**, 44–52 (2013) doi:10.1016/j.ecoinf.2013.03.001.

bypass a large portion of the soil (preferential flow). When preferential flow occurs along soil macropores like root channels or soil fissures, a larger amount of agrochemicals might be leached towards groundwater without degradation (Jarvis, 2007, and references therein).

In general, it is difficult to predict by modelling the type of water flow that will occur in a particular soil. Indeed, often parameter estimation in hydrological models does not yield a unique solution and predictions are uncertain. Therefore, dye tracer studies are carried out to visualise flow paths in soils directly (e.g. Flury et al., 1994; Forrer et al., 2000). First, a dye tracer solution is evenly applied onto the soil surface. After the infiltration of the tracer and appropriate waiting time (usually 24 hours), several vertical soil profiles are excavated and photographed. The images are rectified to correct for any geometrical or colour distortions. Thereafter, the classical image analysis consists of classification into stained and non-stained pixels yielding binary images of dye patterns. Usually, these dye patterns are used for a qualitative or a semi-quantitative description of flow: the number of stained pixels per depth, the so-called dye coverage, is determined and its shape is interpreted to identify the flow type.

Several recent studies went beyond the simple analysis of the dye coverage. Weiler and Flühler (2004), for instance, used the width of stained paths to divide the dye patterns in different types of preferential and uniform flow. However, the authors based their analysis on stereological methods that assume either isotropy of dye patterns or random sampling. The first assumption might be invalid because, in general, water moves vertically through the soil and properties of patterns might differ between horizontal and vertical directions. The assumption of random sampling is seldom valid because soil profiles are prepared in regular intervals and are parallel to each other. Furthermore, the class limits applied by Weiler and Flühler (2004) are probably site-specific and cannot be generalised. Kulli et al. (2003) identified zones of homogeneous flow regions via hierarchical clustering. They found the dye coverage and the mean width of stained paths especially useful for clustering. However, this might not be the case for other study sites.

Another approach proposed by several authors consisted in characterising images of dye patterns as a whole. Ogawa et al. (1999), for instance, analysed surface fractal characteristics of preferential flow paths. These authors and a related work by Baveye et al. (1998) showed that the image thresholding algorithm and image resolution strongly influenced the results. Wang et al. (2009) used different information measures to describe heterogeneities in water flow and solute transport. They transformed binary images of dye patterns in binary sequences based on median infiltration depths. Their results suggested that information and complexity grow with flow heterogeneity. However, analysing images of dye patterns as a whole can be problematic if the flow type varies with depth. This is often the case when physical properties (like soil texture) or density of macropores change between soil horizons.

In our study, we investigate how dye patterns vary with depth because, in the vadose zone, the vertical transport of water and solutes dominates. Therefore, we characterise binary images of dye patterns row by row. We assume that these images are representative for the transport system of a study area. To discern hydrologically relevant information from binary images some kind of feature extraction is necessary. Dye coverage, for example, is one such feature. However, restricting the analysis to one feature alone might miss important information. Thus, we propose to use several image index functions to extract different (ideally complementary) features like pattern fragmentation or width of stained objects. We

associate each image row with a feature vector (i.e. a certain number of image function values) and use these vectors to classify the image rows by clustering. Because images of stained profiles might have different reasonable clusterings, we calculate multiple consensus clusterings.

The evaluation of dye tracer studies does not only require hydrological expertise but also knowledge in image processing and statistics. To our knowledge, no agreed and effective method to analyse large collections of images exists in the soil hydrology community. The goals of our work are to (i) show that feature extraction using simple index functions can improve the analysis and (ii) adapt a clustering framework to assist the experts in the evaluation of these images.

2. Material and methods

2.1. Feature extraction via image index functions

2.1.1. Preliminaries and notation

Binary images show dye patterns classified in stained and non-stained areas. Here, we identify the stained pixels with the integer 1 and non-stained with 0. We introduce functions of binary vectors that we call *index functions* or *indices* to characterise binary images in a quantitative manner. To compare indices of different images, we require the following properties (for a mathematical specification of the properties see Trancón y Widemann and Bogner, 2012):

(1) Magnification Invariance: The range of an index should not depend essentially on the width of the image or the width of the photographed soil profile.

(2) Translation Invariance: The value of an index should not depend essentially on the horizontal position and the horizontal orientation of stained objects.

(3) Range Efficiency: The extreme cases of completely stained or completely non-stained image rows should approximate extreme values of the range of an index.

For convenience, we choose real numbers in the interval $[0, 1]$ for the range of an index. Scaling the images by a constant factor should not affect the indices (cf. property 1). Because the exact location of a dye tracer experiment is often chosen at random, the horizontal position of a flow path in the image is not important. Rather, we are interested in statistical properties of images and consider the stained patterns as horizontally translation invariant. Additionally, we want the index functions to be essentially independent of the horizontal orientation of the patterns. In other words, the value of an index function should be the same when flipping the stained patterns horizontally (cf. property 2).

We denote index functions by capital letters with subscripts, as in I_{D}, binary vectors by italic lower-case letters with an arrow, as in $\vec{r}$, and the length of a vector as $|\vec{r}| = m$, with $m \in \mathbb{N}$. To select a particular element of a binary vector, we write r_i, with $i = 0, \ldots, m-1$ being the index of its elements. The angle brackets $\langle\rangle$ are used for the inner product of two vectors.

In the following we list the proposed indices that we have used as feature extractors for binary images. However, any function satisfying properties 1 to 3 is a valid image index.

2.1.2. Dye coverage

The *dye coverage* is the classical index used to summarize the information of a binary image of a stained soil profile. It is calculated as

$$I_{\mathrm{D}}(\vec{r}) = \frac{1}{m}\sum\nolimits_i r_i \tag{1}$$

and shows the proportion of the soil profile stained by the tracer. As for all averaged values, I_{D} does not discriminate between different pattern configurations (e.g. whether a pattern is contiguous or fractionated).

2.1.3. Statistics of runs

We call (maximal) contiguous subvectors of equal values "runs". The runs of 1s show the widths of stained objects. Their number in the one-dimensional case equals the Euler number. Normalizing by the maximum possible number of runs in an image row of length m gives

$$I_{\mathrm{E}}(\vec{r}) = \frac{\left|\mathcal{R}_1(\vec{r})\right|}{\lceil m/2 \rceil}\,. \tag{2}$$

The function $\mathcal{R}_1(\vec{r})$ calculates the sequence of run lengths of 1s. The brackets $\lceil\,\rceil$ indicate the ceiling function that rounds up to the nearest integer. To summarize the distribution of runs in a robust manner we calculate their 5%, 50% (i.e. median) and 95% quantiles

$$I_{\mathrm{Q}\,0.05} = \frac{1}{m} Q_{0.05}\big(\mathcal{R}_1(\vec{r})\big)\,, \tag{3}$$

$$I_{\mathrm{Q}\,0.5} = \frac{1}{m} Q_{0.5}\big(\mathcal{R}_1(\vec{r})\big)\,, \tag{4}$$

$$I_{\mathrm{Q}\,0.95} = \frac{1}{m} Q_{0.95}\big(\mathcal{R}_1(\vec{r})\big)\,, \tag{5}$$

where the function Q_p calculates the p^{th} quantile.

2.1.4. Contiguity and fragmentation

To measure the contiguity of the stained pixels we define:

$$I_{\mathrm{C}}(\vec{r}) = \big\langle \mathcal{R}_1(\vec{r}), \mathcal{R}_1(\vec{r}) \big\rangle \;/\; \big(\textstyle\sum_i r_i\big)^2\,, \tag{6}$$

where the indeterminate case 0/0 is completed as 1. Among all vectors with k runs, I_{C} attains its minimum of $1/k$ for uniform run length. I_{C} can be interpreted as the reciprocal of a non-integer measure of the number of stained objects weighted by their size. Compared with the other indices, it behaves differently; it attains its maximum of 1 for completely stained and completely non-stained rows and decreases for more fragmented (i.e. more interesting) patterns. Therefore, for an easier interpretation we complement I_{C} and define *fragmentation*

$$I_{\mathrm{F}}(\vec{r}) = 1 - I_{\mathrm{C}}(\vec{r})\,, \tag{7}$$

I_{F} equals 0 for completely stained and non-stained image rows. Additionally, given two rows with the same amount of staining (i.e. equal I_{D}) I_{F} will be larger for patterns with smaller runs.

2.1.5. Metric entropy

The information content of an outcome x of a discrete random variable is defined as $h(x) = -\log_2 p(x)$ and is measured in bits. For a set of events X with probability of occurrence $p(x_1), p(x_2), \ldots, p(x_n)$, Shannon (1948) defined the average information content, also called Shannon's entropy:

$$H(X) = -\sum_{j=1}^{n} p(x_j) \cdot \log_2 p(x_j)\,. \tag{8}$$

Among all distributions with n possible events, H attains its maximum of $\log_2 n$ for the uniform distribution.

To apply equation (8) to binary vectors, we consider individual bits as independent realisations of a binary random variable. In this case the set of events X is the binary alphabet $A = \{0, 1\}$. We replace the theoretical probabilities by empirical frequencies, $p(0)$ and $p(1)$, and calculate Shannon's entropy as

$$H(\vec{r}) = -\big(p(0) \cdot \log_2 p(0) + p(1) \cdot \log_2 p(1)\big)\,. \tag{9}$$

More detailed structures can be captured by considering words of length L of a binary vector. Then, Shannon's entropy is calculated based on the frequencies of these words. A normalization by L yields the metric entropy

$$I_{\mathrm{ME}\,L}(\vec{r}) = \frac{1}{L} \cdot H\big(\mathcal{W}_L(\vec{w})\big)\,, \tag{10}$$

where H is the generalisation of Shannon's entropy in equation (9) for the alphabet A of size 2^L containing now all possible words of length L

$$H(\vec{w}) = -\sum_{x \in A} p(x) \cdot \log_2 p(x) \quad \text{where } p(x) = \frac{\mathcal{H}(\vec{w})(x)}{m - L + 1}\,. \tag{11}$$

$\mathcal{W}_L(\vec{r})$ is the sliding window transform that moves a window of length L through the vector $\vec{r}$ to get the different words. $\mathcal{H}(\vec{w})$ is the histogram function that counts the different words of length L. Metric entropy has a minimum vector length $m_0 = L$, but yields useful values only for $m \gg L$. For larger L we need more data to avoid finite size effects. The length of the image rows we use for our study equals 599 (cf. Section 3). Therefore, according to Wolf (1999), the maximum length for which the expected relative error does not exceed 5% is $L = 8$.

2.2. Multiple consensus clustering

We used the indices to partition rows of binary images into similar clusters with the k-means algorithm (MacQueen, 1967), one of the most popular clustering algorithms. Clustering is often referred to as an "ill-posed" problem (e.g. Jain, 2010) for the true underlying structure of the objects to be clustered is unknown. In other words, the data are not labelled like in genuine classification problems. For example, k-means always finds a partition (or clustering) of the data into k clusters, where k is specified by the user and does not necessarily reflect the

true underlying structure of the data. Additionally, this algorithm is sensitive to the initial assignment of cluster centroids and results for different initialisations can differ considerably (e.g. Arthur and Vassilvitskii, 2007).

Instead of searching for the best partition, we can calculate different clusterings. They can be grouped into an ensemble (i.e. a collection of individual clustering solutions) and then "aggregated" to give a better partitioning of the data than a single clustering (Hornik, 2005b). "Better" means here better separated clusters or a smaller value of the objective function. Additionally, aggregation (also called consensus) helps to overcome problems like the initial assignment of cluster centroids. The whole procedure is called consensus clustering, aggregation of clusterings or ensemble clustering (Strehl and Ghosh, 2003). However, Zhang and Li (2011) argued that building a consensus on possibly substantially different clustering solutions does not necessarily improve the result. Therefore, they extended this method by first clustering the single solutions (called input or base partitions) hierarchically to obtain "similar" ensembles and then calculating the consensus on each ensemble separately. This multiple consensus clustering allows for a multi-faceted view of a data set that might have different reasonable clusterings (Zhang and Li, 2011; Cui et al., 2007).

2.2.1. Data transformation

The distribution of the indices can influence the success of clustering. Pyle (1999), for instance, emphasized that some techniques had equal sensitivity to all values across the whole range. Indeed, for k-means all values are equally important because it uses Euclidean distances to infer their proximity. However, some values of the indices are more frequent and tend to dominate the distribution (e.g. zeros for completely non-stained image rows) thus possibly masking interesting structures. Therefore, we transformed the values to approximate a uniform distribution in the range $[0, 1]$. The transformation of the indices preserves their properties (cf. 2.1.1). Actually, every order-preserving transformation whose range lies in the interval $[0, 1]$ is a valid index transformation.

Another argument in favour of a data transformation is provided by information theory. The indices that we calculate might be (monotonic) "non-linear" functions of the "true" properties of patterns. Since we ignore the distribution of these properties, any monotonic transformation of the calculated indices represents the same properties. We chose a monotonic transformation that yields a distribution with maximum entropy (i.e. the uniform distribution).

2.2.2. Workflow

1. *Generation of input partitions:* We calculated input partitions of the transformed indices by k-means clustering. The number of clusters varied between 2 and 10, the initial cluster centroids were randomly assigned for every input partition.
2. *Comparison of partitions:* On this ensemble, we determined the similarity matrix where each entry equals the pairwise similarity of partitions (cf. Section 2.2.3).
3. *Hierarchical clustering:* Using the similarity matrix, we clustered the input partitions by the "bottom-up" (agglomerative) approach as suggested by Zhang and Li (2011) and cut the hierarchical tree into three subtrees. Obviously, this might be an over-simplified approach and Zhang and Li (2011) showed how to find the number of subtrees based on clustering quality. However, we want to perform Monte Carlo simulations (see below) and need to fix the number of subtrees for better comparison.

4. *Building consensus:* We calculated the optimal number of clusters and the final partitioning in each subtree by performing binary clustering (Li et al., 2010).
5. *Monte Carlo Simulations:* To infer how many input clusterings are needed for a "stable" multiple consensus clustering, we carried out Monte Carlo simulations with random initial cluster centroids. We generated 5, 10 and 20 partitions à $k = \{2, \ldots, 10\}$ resulting in 45, 90 and 180 input partitions all together. This experiment was repeated 100 times.

All calculations were performed in R (R Core Team, 2012) using the packages `clue` (Hornik, K., Ver. 0.3-44) (Hornik, 2005a) and `svd` (Korobeynikov, A., Ver. 0.2). In the following we describe steps 2 and 4 in detail.

2.2.3. Comparison of partitions with variation of information

To cluster the input partitions hierarchically, we need a measure of their similarity. Meilă (2007) proposed an information-based criterion called the *variation of information* (*VI*) to compare partitions of the same data set regardless of the number of clusters. Consider n points to be partitioned into K clusters. Then the probability that a randomly picked point will fall into the cluster c_k equals $p(c_k) = n_k/n$, and depends on the proportion of points in this cluster. Using equation (8) we can define what Meilă (2007) calls the entropy of clustering C

$$H(C) = -\sum_{k=1}^{K} p(c_k) \log p(c_k)\,. \tag{12}$$

Additionally, we need another notion from information theory to define *VI*, namely the mutual information

$$I(C, C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} p(c_k, c'_{k'}) \log \frac{p(c_k, c'_{k'})}{p(c_k) \cdot p(c'_{k'})}\,, \tag{13}$$

where $p(c_k, c'_{k'})$ is the joint probability that a point falls into cluster k in the partition C and into k' in the partition C'. This is the information one partition provides about the other. In other words, if we do not know to which cluster a point belongs in the partition C and we are given its cluster membership in the partition C', mutual information tells us how this knowledge reduces our uncertainty about the cluster membership in C on the average. Finally, *VI* is defined as

$$VI(C, C') = H(C) + H(C') - 2 \cdot I(C, C') \tag{14}$$

VI is a metric on the space of all clusterings. Furthermore, it is bounded by $VI(C, C') \leq 2 \cdot \log K^*$, where K^* is the maximum number of clusters.

2.2.4. Consensus via binary clustering

We combine partitions in each subtree of the hierarchical tree obtained in step 3 by performing a binary clustering (Li et al., 2010). A similar approach was proposed by Fallah et al. (2008). For simplicity, we explain the procedure for one single tree. Let $P = \{p_1, p_2, \ldots, p_n\}$ be a set of n data points and T the number of clusterings $\mathcal{C} = \{C^1, C^2, \ldots, C^T\}$ in the subtree. Let each partition have K_j clusters $C^j = \{C^j_1, C^j_2, \ldots, C^j_{K_j}\}$ and $q = \sum_{j=1}^{T} K_j$

the total number of clusters in all partitions. Then we can write the n points as q-dimensional vectors of the form

$$\begin{aligned} d_i &= (d_{i11}, \ldots, d_{i1K_1}, \ldots, d_{ijK_j}, \ldots, d_{iT1}, \ldots, d_{iTK_T}) \\ d_{ijl} &= \begin{cases} 1 & p_i \in C_l^j, \\ 0 & \text{otherwise} \end{cases} \qquad 1 \le j \le T, 1 \le l \le K_j, 1 \le i \le n\,. \end{aligned} \tag{15}$$

Thus, the data set P can now be represented by a $n \times q$ matrix D where every row consists of a binary vector containing the information about the clustering membership in different partitions.

To find the final number of clusters we first normalize the matrix D such that the vectors d_i have unit length. Then we calculate $\tilde{D}\tilde{D}^T$, where $\tilde{D}$ is the normalized matrix D and $\tilde{D}^T$ is its transpose. Note that $\tilde{D}\tilde{D}^T$ can be thought of as a similarity matrix, where the similarity between point i and point j is expressed by their inner product. Since we have normalized D, we get cosine similarities in $\tilde{D}\tilde{D}^T$. Additionally, $\tilde{D}\tilde{D}^T$ is also the mean association matrix S (sometimes called consensus matrix) that shows the proportion of times two points fall into the same cluster

$$S_{p_i,p_{i'}} = \frac{1}{T}\sum_{j=1}^{T} M_j(p_i, p_{i'})\,, \tag{16}$$

where $M_j(p_i, p_{i'})$ is the $n \times n$ association matrix for the clustering C^j that indicates whether two points fall into the same cluster

$$M_j(p_i, p_{i'}) = \begin{cases} 1 & p_i, p_{i'} \text{ belong to the same cluster} \\ 0 & \text{otherwise}\,. \end{cases} \tag{17}$$

Summing over all association matrices and normalizing by the number of clusterings T is indeed the same as first normalizing D that contains the information about the cluster membership of a point and then building the product $\tilde{D}\tilde{D}^T$.

We can now reorder the matrix $\tilde{D}\tilde{D}^T$ to have points belonging to the same cluster appear in blocks. Then $\tilde{D}\tilde{D}^T$ can be rewritten as

$$\tilde{D}\tilde{D}^T = L + E\,, \tag{18}$$

where L has a block structure

$$L = \begin{pmatrix} X_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & X_k \end{pmatrix}. \tag{19}$$

X_i are blocks of 1s of size $n_i \times n_i$ and E is a symmetric (error or perturbation) matrix with a small value in each entry. Following the results by Juhàsz (1989), $\tilde{D}\tilde{D}^T$ has k large (in absolute value) eigenvalues of order $n_1, n_2, \ldots, n_k$ and $n-k$ eigenvalues of order $\sqrt{n}$. Thus, we can deduce the final number of clusters k and their approximate sizes n_i, $i = 1, \ldots, k$, from the spectral properties of $\tilde{D}\tilde{D}^T$. In our case, because $\tilde{D}\tilde{D}^T$ is an inner-product matrix, it is positive semi-definite (i.e. its eigenvalues are non-negative). To distinguish between large

and small eigenvalues we use a criterion comparable to the explained variance in principal component analysis:

$$\sum_i \lambda_i = 0.9 \cdot n \,, \tag{20}$$

where λ_i denote the eigenvalues of $\tilde{D}\tilde{D}^T$. In other words, we infer the number of clusters by summing up the largest eigenvalues up to a chosen threshold, in this case 90%.

The final step consists of partitioning the binary matrix D by k-means using the determined number of clusters.

2.3. Post-processing of the clusterings

The clustering algorithm operates on the (unordered) set of index vectors for one or more images. Hence clusters are assigned to rows irrespective of their vertical proximity. Nevertheless, there is a strong tendency of cluster numbers to form vertical runs. In other words, rows close to each other are likely to be assigned to the same cluster, with a small amount of jitter on small spatial scales. Thus the clustering can almost immediately be visualised, for instance by colours as in Fig. 4.

Two post-processing issues need to be addressed: Firstly, the numbering of clusters is arbitrary due to their random initialization, resulting in uncontrolled assignments of colours and permutations between Monte Carlo repetitions. Secondly, visualisation is improved by smoothing out the jitter with a suitable vertical filter.

Cluster numberings between Monte Carlo repetitions were made compatible by choosing an arbitrary but fixed ordering criterion and sorting clusters accordingly. We chose to sort clusters in decreasing order of dye coverage of the cluster centroids. This choice has additional benefits: Because the dye coverage index varies smoothly with depth, spatially adjacent clusters are likely to be assigned adjacent numbers. This simplifies the allocation of visually informative colours. Furthermore, effective vertical smoothing can be achieved by a median filter with a symmetric window of suitable size z.

3. Results and discussion

3.1. Examples of indices

We calculated the indices for six images of dye patterns from a forest site in south-east Germany. The images (599 × 579 pixels) originated from a tracer study and represent each a vertical stained soil profile of approximately 1 m^2. A description of the study site and the image classification procedure are given by Bogner et al. (2008). Additionally, the Supplemental Material contains the original colour photographs (Supplemental Fig. S1).

Fig. 1 shows an example of a dye tracer image from this data set and three indices: the dye coverage I_{D}, the fragmentation I_{F} and the median run length $I_{\mathrm{Q}\,0.5}$. The latter has been transformed to enhance the details. As indicated by the circles and arrows, the minima and maxima of the indices highlight complementary properties of the dye patterns. Indeed, the dye coverage is large when the image row is mostly stained. On the other hand, the fragmentation increases for rows where the pattern consists of many isolated objects and the median run length is small when the single objects are mainly small.

The image indices can be used to compare images quantitatively without any further analysis. Ruidisch (2012), for example, used them to evaluate the influence of ridge tillage

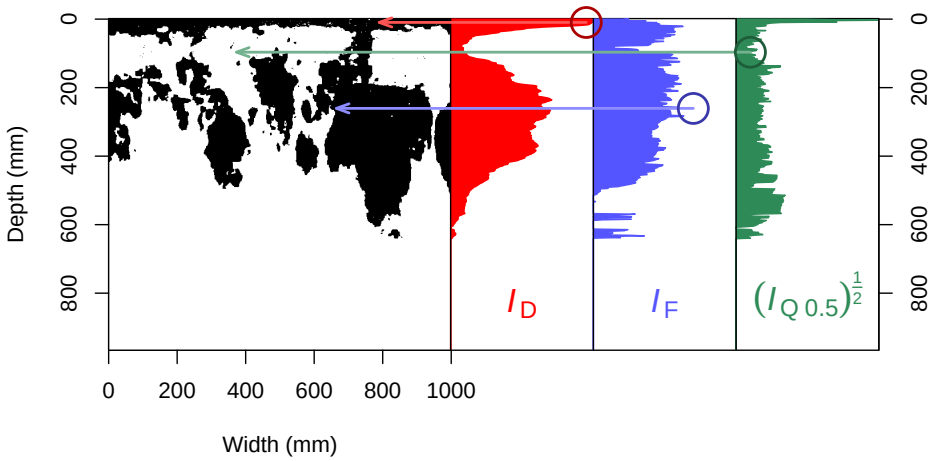

Figure 1: Example of a binary image and three indices (from left to right): dye coverage I_D, fragmentation I_F and the square root of the median run length $I_{Q\,0.5}$. The x axis of the indices contains the full range $[0, 1]$. For explanation of circles and arrows see Section 3.1.

with and without plastic mulching on flow patterns in an agricultural soil. They found different indices valuable to highlight the effects of topography, plastic mulching and the crop root system.

3.2. Monte Carlo simulations

The relationship between stability and accuracy of clustering ensembles depends on the data (Kuncheva and Vetrov, 2006). Because we are not given any external labels for our data set to assess the accuracy, we only evaluate the stability as a function of the number of input partitions (45, 90 and 180). Every Monte Carlo simulation yielded three subtrees that were sorted by the number of clusters in the consensus in increasing order (S_1, S_2 and S_3). The median number of clusters was 2, 5 and 10 for the three subtrees, respectively, regardless of the number of input partitions (Fig. 2). While for 5 input partitions the number of clusters varied (especially in the third subtree), it was rather stable for 90 input partitions (i.e. 10 input partitions à $k = \{2, \ldots, 10\}$). The variation of information decreased with increasing number of input partitions indicating more similar consensus partitions. For further analysis we consider one example (referred to as MC1) of consensus clustering for 90 input partitions.

3.3. Multiple clustering views

With the multiple consensus clustering we obtained three different views of the data set. Fig. 3 shows the splitting of consensus clusterings with increasing number of clusters for MC1. Additionally, Figs. S3, S4 and S5 summarize the variation of splitting between different Monte Carlo simulations. The partitioning in two clusters separated the stained (green cluster) from (mainly) non-stained rows in the soil profiles. It is an obvious solution that can be seen as a rough filter depicting the region of interest (i.e. stained rows) to be considered in more detail. With increasing number of clusters the non-stained rows remained mostly in the same cluster. While the splitting from $k = 2$ into $k = 5$ clusters was quite robust (Fig. S4), further branching into $k = 10$ clusters varied between Monte Carlo simulations (Fig. S5).

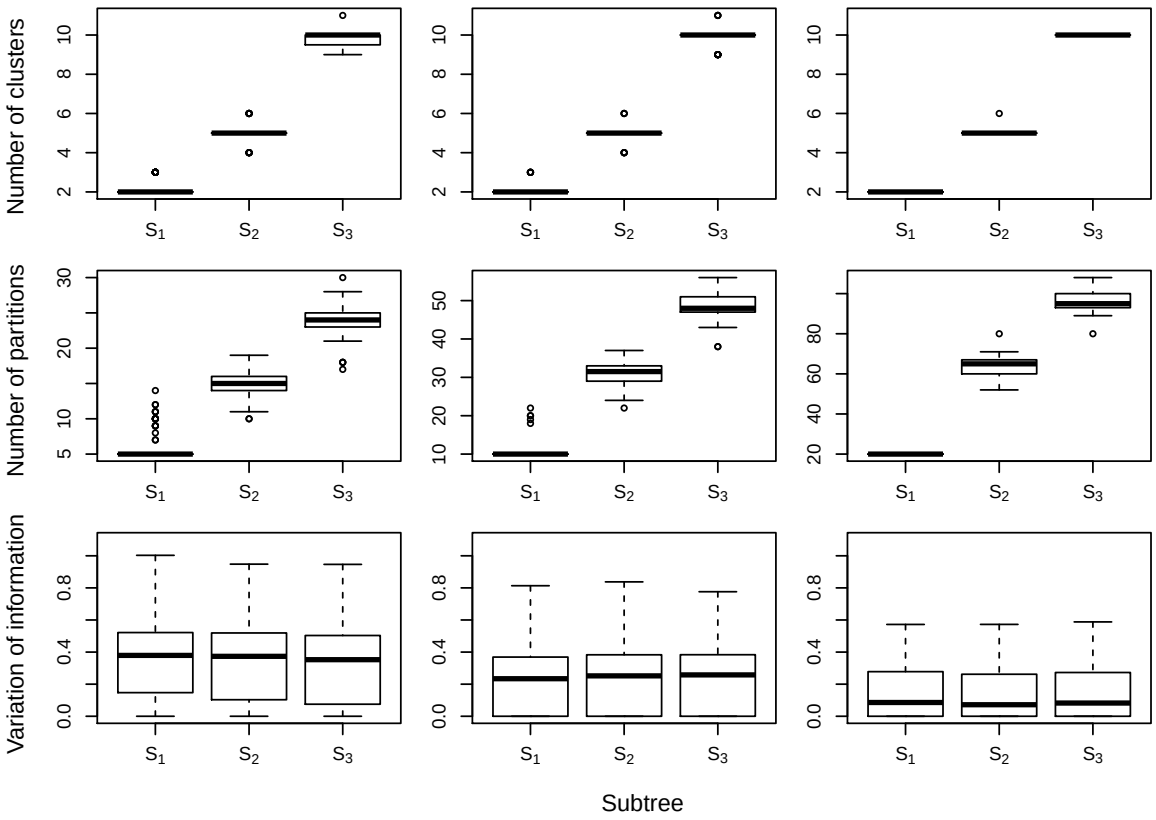

Figure 2: Results of the Monte Carlo simulations. From top to bottom: the number of clusters, the number of partitions in the consensus and the variation of information between consensus clusterings. The left column refers to the experiment with 45 input partitions, the middle and the right ones summarize the results of simulations with 90 and 180 input partitions, respectively. Every experiment was repeated 100 times.

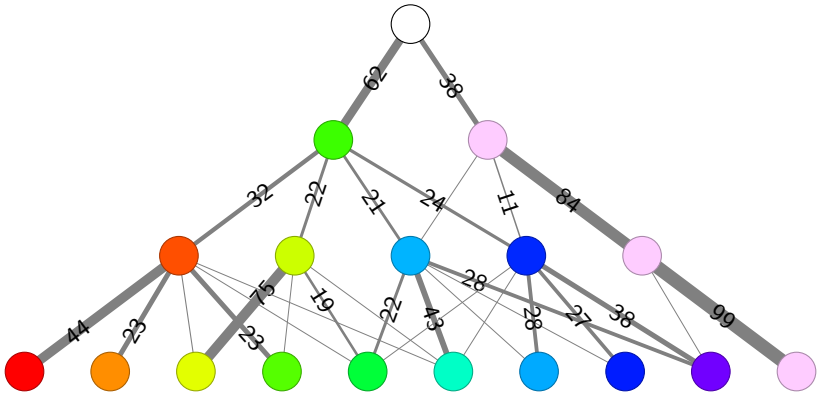

Figure 3: Splitting of consensus clusterings with increasing number of clusters. Circles show different clusters (for explanation of colours see Fig. 4), numbers indicate the proportions of splitting exceeding 10%, widths of connections between clusters were scaled accordingly. The clusterings were smoothed with a median filter of size $z = 17$.

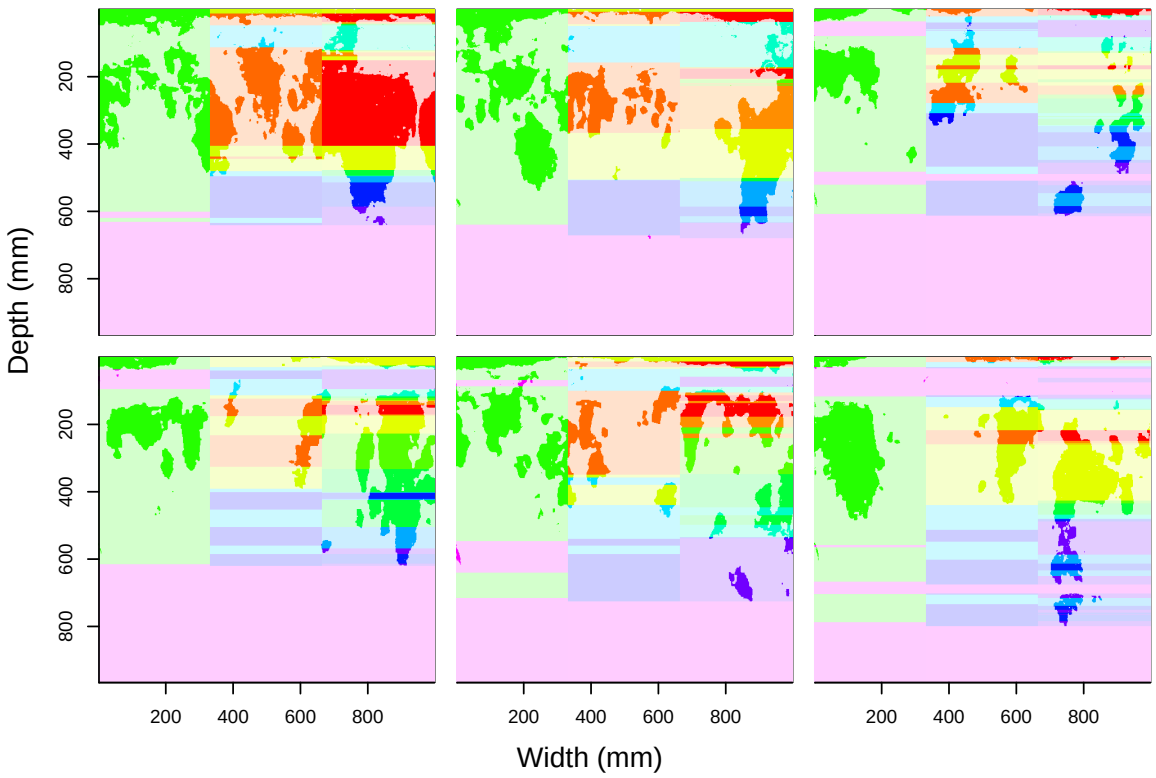

Figure 4: Merging of partitions with decreasing number of clusters in MC1. The horizontal coloured stripes indicate different clusters. From right to left in every image: 10, 5 and 2 clusters. The colours indicate which clusters were merged together when reducing the number of clusters: the colours of the clustering with $k = 5$ are mixtures of the clustering with $k = 10$ weighted by the cluster sizes, and the colours for $k = 2$ are mixtures of $k = 5$. The clusterings were smoothed with a median filter of size $z = 17$.

The increase of the number of clusters from five to 10 did not affect all clusters in the same way. At the top of the images, for example, the clusters remained essentially unchanged. On the other hand, the part between approximately 200 and 600 mm was split. This is particularly well illustrated in the third and fifth profiles when we compare the cluster borders (Fig. 4).

We can turn the perspective around and ask which clusters are merged together when we reduce the number of clusters. The colours in Fig. 4 show that the large orange cluster ($k = 5$) originates predominantly from the large red and orange clusters ($k = 10$). On the other hand, the blue clusters are merged together and the violet cluster of non-stained image rows remains quite stable. In other words, individual clusters do not merge/split with decreasing/increasing number of clusters, respectively, in a random way. Therefore, clustering solutions with larger number of clusters can be regarded as a more detailed view on the data set.

We chose the solution with $k = 5$ clusters to show the differences between clusters. Fig. 5 depicts the four most interesting indices as a matrix plot in MC1, coloured according to their cluster membership (see Fig. S2 for a matrix plot of all indices). The orange and yellow clusters have a larger dye coverage and longer stained runs than the blue ones. Image rows in the orange cluster have larger metric entropy compared to the yellow cluster. The fragmentation seems to be less discriminative. However, a matrix plot can only show

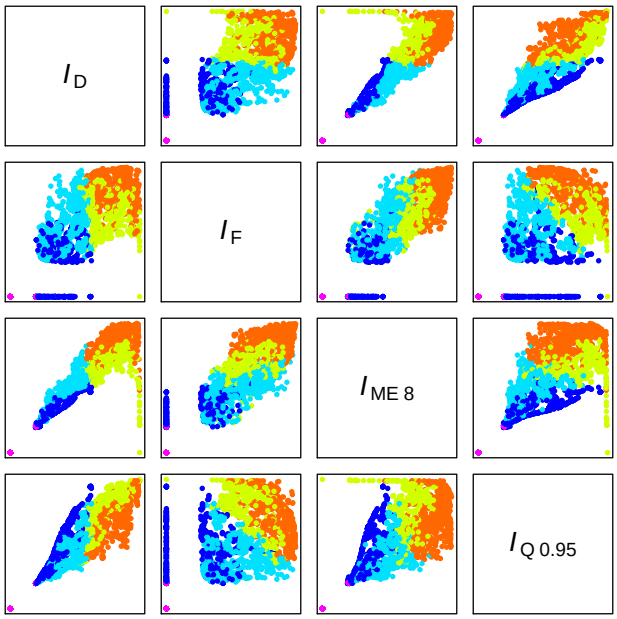

Figure 5: Matrix plot of dye coverage I_D, fragmentation I_F, metric entropy $I_{ME\,8}$ and median run length $I_{Q\,0.5}$ in MC1. The colour coding corresponds to the middle part of Fig. 4.

relationships between pairs of indices and we cannot directly judge the importance of an index in the seven-dimensional index space. It is not self-evident to find the same colours assembled together. This shows that the indices are related, however, in general not linearly.

3.4. Relationship between clusters and soil properties

To relate clusters to soil properties additional field and laboratory measurements are necessary. Bogner et al. (2008) and Bogner et al. (2010) identified the main flow types at the study site. They reported that root macropores constituted the main preferential flow paths, especially in the densely rooted upper soil. The soil below approximately 200 mm was dominated by heterogeneous matrix flow because there, the root macropores ended and the flow was forced to disperse into the sandy soil matrix. Therefore, at this study site, the distribution of soil fine fraction is less important than the distribution of roots. Indeed, the cluster boundaries do not coincide with boundaries of soil horizons (Fig. 7).

At the top of the soil profiles, our clustering solution separates a homogeneously stained area (yellow) from the area of converging flow (orange) where the flow concentrates into few preferential pathways (Fig. 6). In the orange cluster the metric entropy and the fragmentation increase, indicating more diversified patterns. Below, we discover a region of preferential flow (light blue) that is dominated by few stained objects and a decrease in dye coverage. Then, we find again an alternation of orange and yellow clusters. In other words, our clustering solution appears to group two pairs of different flow types, namely converging flow with heterogeneous matrix flow (orange cluster) and homogeneous matrix flow with parts of heterogeneous flow (yellow cluster), respectively, into common clusters. However, this is not a failure of the method: The indices reflect phenomenological properties of the images and so does the clustering based on them, but the mechanisms leading to similar patterns might be different.

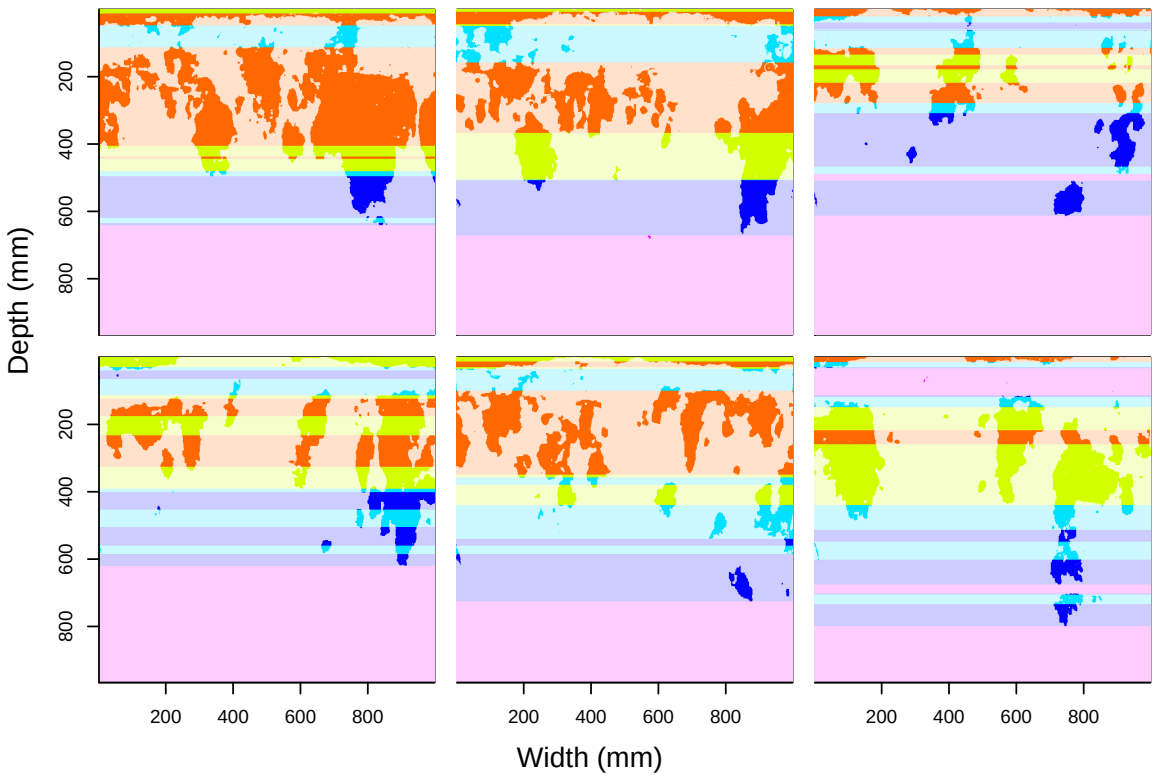

Figure 6: Clustering solution with five clusters (MC1). The horizontal coloured stripes indicate different clusters. The clustering was smoothed with a median filter of size $z = 17$.

In order to resolve this ambiguity, expert knowledge and additional field measurements are necessary. For example, the order and alternation of clusters is important because typical vertical sequences of patterns might exist. Actually, Flühler et al. (1996) described schematically a characteristic order of phenomenological flow regimes: *distribution flow* with strong flow convergence toward preferred flow paths, then *preferential flow* along these flow paths and finally *dispersive flow* where the flow is forced into the surrounding soil matrix. In our case, distribution flow is captured by the orange cluster in the topsoil, preferential flow by the light blue cluster in the topsoil and dispersive flow by the yellow and orange clusters in the subsoil. The blue clusters in the subsoil mark the end of dispersive flow.

4. Conclusions

We have proposed a mathematical framework for indices that characterise dye patterns in a quantitative manner. Because this framework is general, other functions satisfying properties 1, 2 and 3 are valid indices and can be added to the ones described here. The image indices can be used to explore a large number of dye tracer images. For example, image pre-classification based on a threshold of an index or the position of the maximum/minimum can be done in an automated way. In a more general way, the introduced indices are sufficiently universal to be useful to analyse other binary images, for example in remote sensing.

Furthermore, we have adapted a state-of-the art clustering framework to explore multiple views on image data sets where a definitive clustering solution is difficult to find. Experts

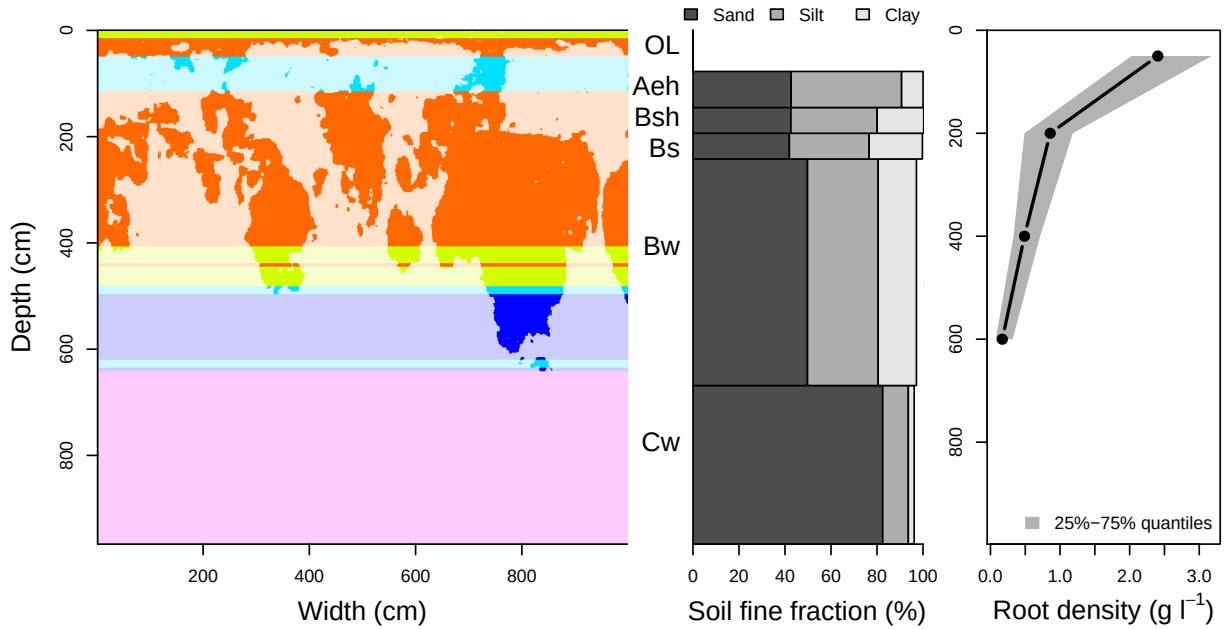

Figure 7: Relationship between clusters (MC1, $k = 5$) and soil properties. Left: example profile; center: distribution of soil fine fraction (sand, silt and clay) with depth (OL, Aeh, Bsh, Bs, Bw and Cw are names of soil horizons); right: distribution of fine roots with depth. The horizontal coloured stripes in the left figure indicate different clusters. The clustering was smoothed with a median filter of size $z = 17$.

can explore these multiple consensuses and base their interpretation of predominant flow types on quantitative (objective) criteria.

Acknowledgements

We thank Prof. Dr. Martin Schlather for statistical advice. Benjamin Wolf carried out a part of the field work.

AppendixA. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecoinf.2013.03.001.

References

Arthur, D., Vassilvitskii, S., 2007. k-means++: The advantages of careful seeding, in: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics. pp. 1027–1035. http://dl.acm.org/citation.cfm?id=1283494.

Baveye, P., Boast, C.W., Ogawa, S., Parlange, J.Y., Steenhuis, T., 1998. Influence of image resolution and thresholding on the apparent mass fractal characteristics of preferential flow patterns in field soils. Water Resour. Res. 34, 2783–2796.

Bogner, C., Gaul, D., Kolb, A., Schmiedinger, I., Huwe, B., 2010. Investigating flow mechanisms in a forest soil by mixed-effects modelling. Eur. J. Soil Sci. 61, 1079–1090.

Bogner, C., Wolf, B., Schlather, M., Huwe, B., 2008. Analysing flow patterns from dye tracer experiments in a forest soil using extreme value statistics. Eur. J. Soil Sci. 59, 103–113.

Cui, Y., Fern, X., Dy, J., 2007. Non-redundant multi-view clustering via orthogonalization, in: Seventh IEEE International Conference on Data Mining, pp. 133–142. http://dx.doi.org/10.1109/ICDM.2007.94.

Fallah, S., Tritchler, D., Beyene, J., 2008. Estimating number of clusters based on a general similarity matrix with application to microarray data. Stat. Appl. Genet. Mol. Biol. 7, Article 24.

Flury, M., Flühler, H., Jury, W.A., Leuenberger, J., 1994. Susceptibility of soils to preferential flow of water: a field study. Water Resour. Res. 30, 1945–1954.

Flühler, H., Durner, W., Flury, M., 1996. Lateral solute mixing processes – A key for understanding field-scale transport of water and solutes. Geoderma 70, 165–183.

Forrer, I.E., Papritz, A., Kasteel, R., Flühler, H., Luca, D., 2000. Quantifying dye tracers in soil profiles by image processing. Eur. J. Soil Sci. 51, 313–322.

Hornik, K., 2005a. A CLUE for CLUster Ensembles. J. Stat, Softw. 14, 1–25.

Hornik, K., 2005b. Cluster ensembles, in: Weihs, C., Gaul, W. (Eds.), Classification – The Ubiquitous Challenge. Proceedings of the 28th Annual Conference of the Gesellschaft für Klassifikation e.V. University of Dortmund, March 9–11, 2004, Springer-Verlag. pp. 65–72.

Jain, A.K., 2010. Data clustering: 50 years beyond k-means. Pattern Recogn. Lett. 31, 651–666.

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.

Juhàsz, F., 1989. On the theoretical backgrounds of cluster analysis based on the eigenvalue problem of the association matrix. Statistics: A Journal of Theoretical and Applied Statistics 20, 573–581.

Kulli, B., Stamm, C., Papritz, A., Flühler, H., 2003. Discrimination of flow regions on the basis of stained infiltration patterns in soil profiles. Vadose Zone J. 2, 338–348.

Kuncheva, L., Vetrov, D., 2006. Evaluation of stability of k-means cluster ensembles with respect to random initialization. IEEE T. Pattern. Anal. 28, 1798–1808. http://dx.doi.org/10.1109/TPAMI.2006.226.

Li, T., Ogihara, M., Ma, S., 2010. On combining multiple clusterings: an overview and a new perspective. Appl. Intell. 33, 207–219.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations, in: Le Cam, L.M., Neyman, J. (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley. pp. 281–297. http://projecteuclid.org/euclid.bsmsp/1200512992.

Meilă, M., 2007. Comparing clusterings – an information based distance. J. Multivariate Anal. 98, 873 – 895.

Ogawa, S., Baveye, P., Boast, C.W., Parlange, J.Y., Steenhuis, T., 1999. Surface fractal characteristics of preferential flow patterns in field soils: evaluation and effect of image processing. Geoderma 88, 109–136.

Pyle, D., 1999. Data preparation for data mining. Morgan Kaufmann Publishers, Inc., San Francisco, California (USA). ISBN: 1-55860-529-0.

R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0.

Ruidisch, M., 2012. Flow and Transport Processes as affected by tillage management under monsoonal conditions in South Korea. Ph.D. thesis. University of Bayreuth.

Shannon, C.E., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379–423.

Strehl, A., Ghosh, J., 2003. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. JMLR 3, 583–617.

Wang, K., Zhang, R., Hiroshi, Y., 2009. Characterizing heterogeneous soil water flow and solute transport using information measures. J. Hydrol. 370, 109–121.

Weiler, M., Flühler, H., 2004. Inferring flow types from dye patterns in macroporous soils. Geoderma 120, 137–153.

Trancón y Widemann, B., Bogner, C., 2012. Image analysis for soil dye tracer infiltration studies, in: Proceedings of the 3rd International Conference on Image Processing Theory, Tools and Applications, pp. 409–414. http://dx.doi.org/10.1109/IPTA.2012.6469517.

Wolf, F., 1999. Berechnung von Information und Komplexität in Zeitreihen – Analyse des Wasserhaushaltes von bewaldeten Einzugsgebieten. Ph.D. thesis. University of Bayreuth. In German.

Zhang, Y., Li, T., 2011. Extending consensus clustering to explore multiple clustering views, in: Proceedings of the Eleventh SIAM International Conference on Data Mining, pp. 920–931. http://siam.omnibooksonline.com/2011datamining/index.html.