 doi:10.20944/preprints202011.0451.v1

Explainable AI Framework for Multivariate Hydrochemical Time Series

Michael C. Thrun[1], Alfred Ultsch[1], Lutz Breuer[2]

[1]Databionics Research Group, University of Marburg, Germany
[2]Institute for Landscape Ecology and Resources Management (ILR), Justus Liebig University Giessen
Correspondence to: Michael C. Thrun: (mthrun@informatik.uni-marburg.de)

Abstract. The understanding of water quality and its underlying processes is important for the protection of aquatic environments enabling the rare opportunity of access to a domain expert. Hence, an explainable AI (XAI) framework is proposed that is applicable to multivariate time series resulting in explanations that are interpretable by a domain expert. The XAI combines in three steps a data-driven choice of a distance measure with explainable cluster analysis through supervised decision trees. The multivariate time series consists of water quality measurements, including nitrate, electrical conductivity, and twelve other environmental parameters. The relationships between water quality and the environmental parameters are investigated by identifying similar days within a cluster and dissimilar days between clusters. The XAI does not depend on prior knowledge about data structure, and its explanations are tendentially contrastive. The relationships in the data can be visualized by a topographic map representing high-dimensional structures. Two comparable decision-based XAIs were unable to provide meaningful and relevant explanations from the multivariate time series data. Open-source code in R for the three steps of the XAI framework is provided.

Keywords: Explainable AI; Cluster Analysis; Swarm Intelligence; Machine Learning System; High-Dimensional Data Visualization; Decision Trees

1 Introduction

Human activities modify the global nitrogen cycle, particularly through agriculture. These practices have unintended consequences; for example, terrestrial nitrate losses to streams and estuaries can impact aquatic life [1]. A greater understanding of the variability in water quality and its underlying processes can improve the evaluation of the state of water bodies and lead to better recommendations for appropriate and efficient management practices [2].

Accordingly, the objective here is to describe the water quality in terms of nitrate (NO_3) and electrical conductivity (EC) in the Schwingbach catchment (Germany) using environmental variables typically related to chemical water quality. Electrical conductivity is a measure that reflects water quality as a whole because it indicates the number of ions dissolved in the water. NO_3 in water bodies is partially responsible for the phenomenon of eutrophication [3]. Eutrophication occurs when an excess of nutrients (including NO_3) leads to the uncontrollable growth of aquatic plant life, followed by the depletion of the dissolved oxygen [3,4]. For decades, water quality has mainly been measured through manual grab sampling of water samples and subsequent chemical analysis in the laboratory. Due to limited resources, high-resolution measurements on the order of days, hours, or even minutes were not available for a long time. With the advancement of deployable, *in situ* measuring techniques, such as UV spectrometry, a new era of field monitoring has been established [5]. However, we still lack reproducible open-source code based methodological approaches interpretable by domain experts with which to analyze the resulting large datasets [6,7].

For example, Miller et al. argue that most AI researchers are building XAIs for themselves [8] rather than for the intended domain expert and, hence, are often not meaningful for the domain expert. Thus, the XAI framework is introduced using a dataset from the Schwingbach catchment for which a domain expert will interpret the explanations that the XAI provides. In consequence, the proposed XAI framework's goal is to provide meaningful and relevant explanations to a domain expert based on the given data. Miller

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

argues that an essential property of explanations should be that they are contrastive [9]. This means that a domain expert would not only ask why an event happened but also why that event happened rather than an alternative [9]. It follows that interesting features describing water bodies should be distinguishable for the explanations describing these water bodies to be relevant. This work proposed to use class-wise mirrored-density plots (MD plots) [10] to investigate if generated explanations are tendentially contrastive in interesting features (see. 2.4.3 for details).

The whole XAI framework is a mix of swarm intelligence, game theory, neural networks, and supervised decision trees. These are combined in a sophisticated manner to detect meaningful relationships in data, which makes it a comprehensive new tool for interpretable machine learning or so-called explainable AI systems [11]. The main contribution of this paper is the proposition of an XAI that reveals cluster structures in time series with a solely data-driven approach. In this context, data-driven means that the authors aim to refrain from making explicit or implicit assumptions about the existence and type of data structures. The found cluster structures are verified with independent approaches and then used in decision trees.

Overall, this work shows how to search for days with similar behavior by using a transparent and open-source framework. The results explain similar environmental, and in particular water quality, situations although NO_3 stream concentrations "integrate" many processes varying in space and in time [12]. Finally, relevant and meaningful explanations could be used to predict future NO_3 and EC values.

Related Works

There are two approaches for the explanation of machine learning systems: prediction, interpretation, and justification that is used for sub symbolic ML systems (defined in [13]) and interpretable approaches for symbolic ML systems (defined in [14]), which are explained through reasoning [15]. For the former, a well-known example is LIME [16], which approximates any classifier or regressor locally with an interpretable model. In the sense of the latter, explanation represents a distinct approach to extract information from the learned model [17]. Typical interpretable ML systems or so-called XAIs comprise combinations of neural networks with rule-based expert systems [18,19], Bayesian networks with rule mining [20], hybrids of clustering and fuzzy classification [21] or neuro-fuzzy classification [22], interpretable decision sets [23] or decision tables [24], decision tree clustering [25] or clustering combined with generative models [26]. Two of the most recent XAI approaches are the unsupervised decision tree clustering eUD3.5 [27] and a hybrid of k-means clustering and a top-down decision tree [28].

Loyola-González et al. present the eUD3.5 algorithm, which uses a split criterion based on the silhouette index[27]. The silhouette index compares every object of a cluster to its homogeneity within a cluster with the heterogeneity to other clusters [29]. In eUD3.5, a node is split only if it's possible descendants have a better split criterion than the best split criterion found so far. This leads to a decision tree which is based on the cluster structures. A cluster is associated with the class having the most members in the cluster. In eUD3.5, 100 different trees are generated, their performance is evaluated, and the best performing tree is kept. The user can specify the number of desired leaf nodes (stop criterion). If the algorithm produces more leaf nodes than specified by the user, then leaf nodes are combined using k-means. The authors claim to have similar performance to k-means and better performance than other conventional decision tree clustering algorithms [27].

Dasgupta et al use a hybrid of k-means with a decision tree called iterative mistake minimization algorithm (IMM) [28]. The k-means method provides the labels and cluster centers with which the decision tree is built top-down using binary splits in which each node of the tree is associated with a portion of the input data[28]. If this portion of input data contains more than one cluster center, then it is split so that the fewest points are separated from its cluster center: this type of optimal split is found by searching over all pairs using dynamic programming. The IMM algorithm terminates at a leaf node whenever all points have the same label resulting in a number of leaves equal to the number of clusters [28]. Dasgupta et al. did not provide any source code in their work [25]. Therefore, in the first part of the IMM algorithm [25], the k-means clustering is used through the toolbox published in [61]. Then, feature importance in a k-means

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

clustering is measured using the algorithm provided in the R package "FeatureImpCluster" available on CRAN [80]. The importance per feature is measured by the misclassification rate relative to the baseline cluster assignment due to a random permutation of feature values [80].

More general, time series clustering is either raw-data-based, feature-based, or model-based [30]. Often, the Euclidean metric on data points or Kullback-Leibler dissimilarity on distributions are used. Cluster analysis is performed via medoid or agglomerative methods of conventional clustering algorithms [30]. Clustering algorithms are often sensitive to outliers and noise [31,32]. Such approaches also imply that the relevant cluster structures are spherical or of other specific structures because a global optimization function has to be used [33]. Then, evaluation is commonly performed on the basis of within-cluster variance [30]. Other common approaches to the evaluation of time series are shape-based, meaning that the shapes of two time series are matched according to specific dissimilarity measures [34]. The approaches have in common that they optimize a global objective function that defines the data structures, and they do not investigate if cluster structures exist.

2 XAI Framework

Contrary to the approaches introduced above, this work applies an alternative approach to the optimization of a global objective function in the task of projection by Databionic swarm (DBS) [35] and subsequent projection-based clustering [36]. In contrast to most other clustering algorithms, the topographic map visualization [37] identifies if the data contains no structures based on clusters. Outliers can be interactively marked in the visualization [38] after the automated clustering process in the case that they are not recognized sufficiently in the automatic clustering process [38]. The clustering itself is one module of the XAI framework presented in Figure 1.

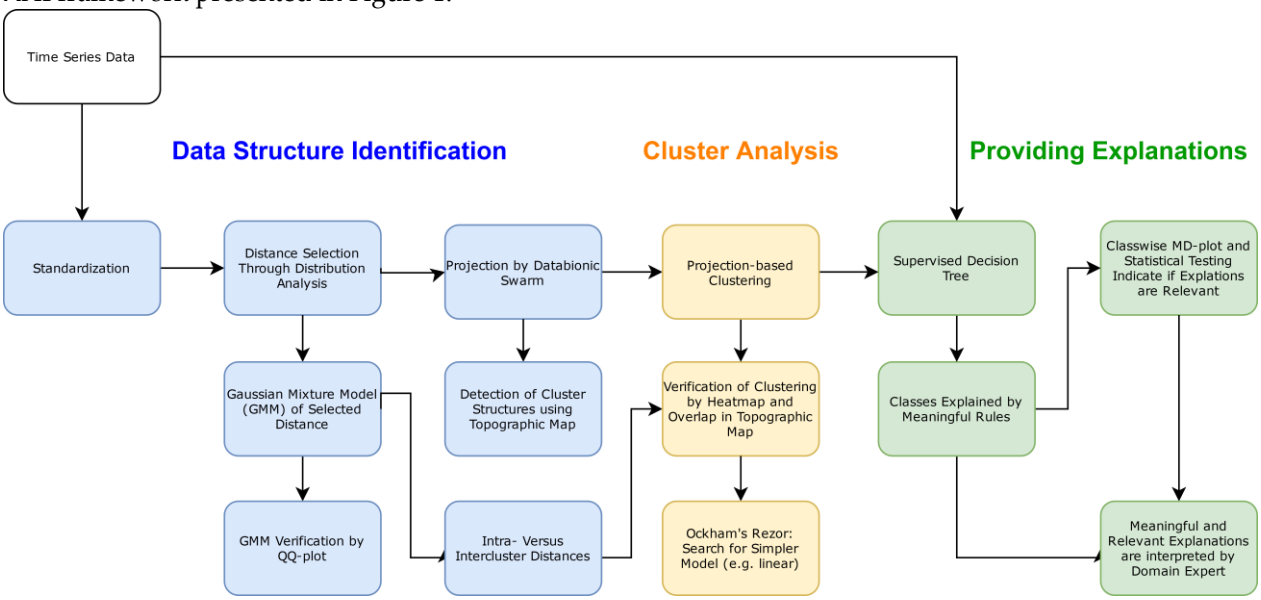

Figure 1: Framework of Explainable AI for multivariate time series without implicit assumptions about data structures (data-driven). The framework has the three main steps of the identification of data structures, Cluster Analysis and providing explanations. Each step has several modules explained in the methods section.

There are three main steps in Figure 1. First, the data structures have to be identified (blue). Next, the cluster analysis is performed (orange). In the third step, explanations (green) are provided using the labels of the clustering and the multivariate time series data. Each step has several modules that are connected to each other and colored similar to the step. Arrows outline the connections between the modules.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

In the first step, the time series data is aggregated appropriately (e.g., daily) and then standardized. In step II, various available distance metrics are applied to the data and investigated for multimodality. If a distance distribution is multimodal, it can be modeled by a Gaussian Mixture. This distance should be preferred for the cluster analysis in the second step. If not, distance is multimodal, the framework continuous with the Euclidean distance. Next, the chosen distances are used in the projection in the module followed by the structure visualization through a topographic map.

The cluster analysis in step II is composed of three modules. This visualization of the topographic map enables the user to choose the number of clusters and the setting of the Boolean parameter. The number of clusters can be set as the number of visible valleys in the topographic map. The clustering can be further evaluated by the model of the distance distribution (last arrow between steps I and II), because using the Gaussian mixture model of step I hypothesizes that the intracluster distances of the clustering should be smaller than the Bayesian boundary defined by the Gaussian Mixture. Projection-based clustering can be changed to another clustering algorithm accordingly to the preference of the user. The clustering can be validated by the topographic map [36]. Additionally, it is preferable to search for linear models and to validate the clustering externally (e.g., by a heatmap).

After validation in step II, the labels of step II's resulting clustering are used in supervised decision trees in step III for the training of the un-preprocessed but aggregated data. Then, meaningful explanations are defined by paths in the decision tree. Class-wise distribution analysis and statistical testing of interesting features are performed to assure that explanations are relevant to the domain expert. In the last module, a domain expert interpret relevant and meaningful explanations (c.f. [8]). The analytic procedure details can be found in the methods sections, which are organized according to these three steps, as illustrated in the titles in Fig. 1. The result sections are organized the same way. The R packages used in this work are summarized in SI H, table 6.

2.1 Step I: Identification of Data Structures

2.1.1 Collection of Multivariate Time Series Data

Data used in this work have been collected in 2013/2014 in the Schwingbach Environmental Observatory (SEO), in central Germany. The mixed developed landscape is mainly characterized by agricultural land use (44%) and forests (48%). The Schwingbach is a small headwater stream draining the landscape. An in-situ hyperspectral UV-spectrometer (ProPS, Trios, Rastede, Germany, wavelength range 200–360 nm, path length 5 mm, solar panel supplied) was used to measure absorption spectra every 15 min at the gauging station of the catchment's outlet. Prior to measurements, air blasts (5 s) were blown on the lens to prevent the optics from biofouling. Wavelengths spectra at 200–220 nm were utilized for the calculation of nitrate concentration, following calibration with local stream water matrix (see [6] for further details). All other variables used in this machine learning approach (Table 1) were also monitored at the same high-frequency or aggregated to 15 min intervals.

Discharge and stream water temperature was recorded by pressure transducers (Diver DCX, Schlumberger Water Services, ON, Canada) at two gauging stations at the outlet (q13) and upstream (q18) of the first-order stream of the Vollnkirchener Bach. Groundwater depths in three piezometers were measured with similar pressure transducers. The piezometers were located in different positions along the Schwingbach. GWl3 was measured in the lowland close to the outlet of the stream, GWl25 was recorded at a hillslope, and GWl32 upstream in the riparian zone). Electromagnetic induction sensor (5TE) attached to EM50 data loggers (Decagon, Labcell LTD, Alton, UK) installed at 0.1 m depth in the riparian zone were used to gauge soil moisture and soil temperature. All meteorological data were collected at a climate station 4 km from the outlet (Campbell Scientific Inc., CR1000 data logger, Loughborough, UK).

In total, the dataset contains 32,196 data points for 14 different variables. Data gaps due to technical problems and data quality control reduced the available data coverage to two growing seasons (05 March 2013 to 24 September 2013, n=15,475 measurements; 27 April 2014 to 23 October 2014, n=16,721

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

measurements). In Table 1, abbreviations and SI units of all variables are provided. The data was published earlier by Aubert et al. [6]. However, Aubert et al. used a high-frequency temporal analysis. In comparison, this work focuses on the average daily measures for each variable, resulting in a low-frequency analysis. Further technical information on the SEO, the analytical procedures applied, the coding of abbreviations, and the experimental design, in general, are outlined in detail by [6],[7],[39].

Table 1: Measured environmental variables with abbreviations and units. The probability density distributions of the transformed dataset are visualized in the supplementary section.

Variable	Abbreviation	SI Unit
Soil temperature	St24	°C
Groundwater level 3=lowland, 25=hill slope, 32= upstream in riparian zone	GWl3 GWl25 GWl32	m
Soil moisture	Smoist24	m^3/m^3
Rainfall	rain	mm/d
Discharge	q13 q18	L/s
Electric conductivity (EC)	Con47	mS/m
Solar radiation	Sol71	W/m^2
Air temperature	At47	°C
Streamwater temperature	Wt18 Wt13	°C
Nitrate (NO_3)	nnit13	mg/L

Four percent of the data are missing. For each day, the measurements were aggregated by the mean of all available measurements for that day. Then, missing values (i.e., days) were interpolated using the seven-nearest-neighbors approach. Distance measures are sensitive to the variance in the distribution of features. For example, the Euclidean metric weights feature more if they have values above 1. Therefore, the variance of features is usually standardized before a cluster analysis is performed.

The variables q13 and q18 were log-transformed. All variables, with the exception of rainfall, were normalized to values between zero and one through a robust normalization procedure [40] improved by [33]. Correlating variables have to be detected before further data evaluation; otherwise, these variables will be over-weighted in assessing the following distance matrices. The discharges correlated linearly with each other ($r=0.95$, $p(S=347{,}270, N=351) <0.001$), and q13 were therefore excluded from the analysis. The air temperatures Wt13 and Wt18 also correlated linearly ($r=0.99$, $p(S=18{,}386, N=351) <0.001$); hence, Wt13 was removed as well.

The outliers in the rainfall variable were detected via ABC analysis [41]. ABC analysis is a method used to compute precise limits to acquire subsets in skewed distributions by exploiting the mathematical properties pertaining to the distribution. The data containing positive values are divided into three disjoint subsets, A, B, and C, with subset A comprising very profitable values, i.e., the largest data values ("the important few"). Subset B contains values for which the yield equals the effort required to obtain it, and subset C contains the non-profitable values, i.e., the smallest datasets ("the trivial many"). The R package is called 'ABCanalysis'. Then, the rain was normalized with respect to the minimum value in group A. All other points in group A were capped by defining the upper bound for rainfall.

2.1.2 Distance Selection

Usually, partitioning and hierarchical clustering algorithms require a distance metric because they seek to find groups of similar objects [42] (i.e., objects with small distances between them). If no specific distance

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

measure is used, most algorithms use the Euclidean distance metric, and the user is not always able to manually change the distance metric (c.f. 54 common algorithms in the R package 'FCPS'). Projection-based clustering with the Databionic swarm (DBS) has the advantage that a specific distance metric can be selected by the user, which is then used in the dimensionality reduction and clustering part of the algorithm. However, the choice of distance remains undiscussed in prior work. We propose that a user selects a distance metric based on the specific properties of the specific data set's distance distribution.

The Hellinger point distance measure is selected because clear multimodality is visible in the probability density distribution. Several metrics were investigated using the R package 'parallelDist' and the MD-plot function [10] in the R package 'DataVisualizations'. The detailed mathematical definitions can be found in SI F. The probability density distribution is modeled with a Gaussian mixture model and verified visually with QQplot as described in [43] with the R package 'AdaptGauss'. The Bayes boundary of two modes with the highest weights separates the intra-cluster distances from the inter-cluster distances. It provides a data-driven hypothesis about the similarity of data points (i.e., days in the example of this work). In this sense, days that a cluster analysis partitions to the same cluster are similar if their intra-cluster distances are lower than the Bayes boundary.

2.1.3 Projection

The swarm-based projection method of the Databionic swarm algorithm is used to project the distance matrix of data into a two-dimensional plane [33,35]. Similar to the nonlinear and focusing projection methods of ESOM [44], CCA[45], t-SNE[46], or NerV[47], the dimensionality reduction by the swarm first adapts to global structures. As time progresses, structure preservation shifts from global optimization to the preservation of local neighborhoods. This learning phase requires an annealing scheme and usually require parameters to be set. However, by exploiting concepts of self-organization and emergence, swarm intelligence, and game theory, this projection method is parameter-free [33,35]. The intelligent agents of the swarm, called DataBots [48], operate on a toroid grid, where positions are coded into polar coordinates to allow for the precise definition of their movement, neighborhood function, and annealing scheme.

In contrast to other focusing projection methods (e.g., [45,46,49]) the size of the grid and the annealing scheme are data-driven. During the learning phase, each agent moves across the grid or stays in its current position in the search for the most potent scent. The equation (c.f. Eq. 18 in [35]) that mathematically defines the scent uses information stored in the distance matrix (c.f. step II and SI F). Hence, agents search for other agents carrying data with the most similar features to themselves with a data-driven decreasing search radius [35]. Every agent's movement is modeled using a game-theory approach, and the radius decreases only if a Nash equilibrium is found [50,51]. Contrary to ant-based clustering algorithms, DataBots do not move data. Instead, each DataBot possesses a scent, defined by one high-dimensional data point.

2.1.4 Structure Visualization by Topographic Map

Projection points near to each other are not necessarily near in the high-dimensional space (vice versa for faraway points), but in planar projections of data, these errors are unavoidable (c.f. Johnson-Lindenstrauss Lemma [52]). Hence, the topographic map identifies data structures based on a projection. First, the generalized U-matrix [33,53] is calculated on this projection using emergence through an unsupervised artificial neural network called a simplified (because parameter-free) emergent self-organizing map [37]. The generalized U-matrix generates the visualization of a topographic map with hypsometric tints, which can be vividly described as a virtual 3D landscape with a specific color scale chosen with an algorithm defining the contour lines [54]. The topographic map addresses the central problem in clustering, i.e., the correct estimation of the number of clusters. It allows the assessment of the number of clusters by inspecting the 3D landscape.

The topographic maps correspond to high-dimensional distance and density structures. Hypsometric tints are surface colors that represent ranges of elevation. The contour lines are combined with a specific

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

color scale. The specific color scale is chosen to display various valleys, ridges, and basins: blue colors indicate small distances (sea level), green and brown colors indicate middle distances (low hills), and shades of white colors indicate vast distances (high mountains covered with snow and ice). Valleys and basins represent clusters, and the watersheds of hills and mountains represent the borders between clusters. In this 3D landscape, the borders of the visualization are cyclically connected with a periodicity (L,C).

2.3 Step II: Cluster Analysis

In step II, projection-based clustering is applied by calculating the shortest paths [55]) of the Delaunay graph between all projected points weighted with high-dimensional Hellinger point distances. This is possible because it was shown that the U-matrix is an approximation of the abstract U-matrix [56], which is based on Voronoi cells. Voronoi cells define a Delaunay graph where the edges between every projected point are weighted by the corresponding data points' high-dimensional distances.

The clustering approach itself involves two choices. For this dataset, the compact approach is used, where the two clusters with the minimal variance (S) are merged together until the number of clusters defined by the topographic map is reached. The other approach for connected structures and a general discussion of cluster structures can be found in [36].

Let $c_r \subset I$ and $c_q \subset I$ be two clusters such that $r, q \in \{1, \dots, k\}$ and $c_r \cap c_q = \{\}$ for $r \neq q$ and

$$\Delta Q(j,l) = \frac{k * p}{k + p} * D^*(j,l) \qquad (1)$$

where

(l, j)	the data points in the clusters be denoted by $j_i \in c_q$ and $l_i \in c_r$;
k	the cardinality $\lvert c_q \rvert$ of the first set;
p	the cardinality $\lvert c_r \rvert$ of the second set;
D^*	the high-dimensional distance based on weighted shortest paths in the Delaunay graph;
ΔQ	the merging cost between two the clusters $c_r, c_q \subset I$

Then, the variance (S) between two clusters (c_r and c_k) is defined as

$$S(c_r, c_k) = \sum_{i=1, j=1, j \neq i}^{k,p} \Delta Q(j,l) \qquad (2)$$

In praxis, the choice of the Boolean parameter of compact versus connected can be evaluated in step I using the topographic map as specified in Figure 1: If a cluster is either divided into separate valleys, or several clusters lie in the same valley of the topographic map, the compact (or connected) clustering approach is not appropriate for the data. An extensive discussion of this behavior can be found in [36]. Additionally, the clustering can be improved further using an interactive interface (c.f. provided in the R package 'ProjectionBasedClustering' [38]. The ultrametric portion of the distance [57] can be visualized by a dendrogram allowing the alternative selection of the number of clusters: Large changes in the fusion levels of the ultrametric portion of the distance indicate the best cut.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

Projection, topographic map of structures, and cluster analysis are available in R language on CRAN in the package 'DatabionicSwarm'[33]. It should be noted that the cluster analysis can also be performed independently to the projection method, with, for example, k-means but most often results in a focus on specific cluster structures [36]. This can be preferable if prior knowledge about the data exists that leads a user to a specific choice of a global cluster criterion.

2.3.1 Validation of Clustering and Ockham's Razor

Engle et al. state: "Cluster heatmaps are commonly used in biology and related fields to reveal hierarchical clusters in data matrices. Heatmaps visualize a data matrix by drawing a rectangular grid corresponding to rows and columns in the matrix and coloring the cells by their values in the data matrix. In their most basic form, heatmaps have been used for over a century [58]. In addition to coloring cells, cluster heatmaps reorder the rows and/or columns of the matrix based on the results of hierarchical clustering. (…). Cluster heatmaps have high data density, allowing them to compact large amounts of information into a small space [59]." [60]. For distance matrices, the procedure can be performed as described above, meaning that the clustering reorders the distances, each pixel represents a distance value, and the clusters are divided by black lines. Further, the clustering is valid if mountains do not partition clusters indicated by colored points of the same color and colored regions of points [58].

A heatmap visualizes the homogeneity of clusters and the heterogeneity of intercluster distances if the clustering is appropriate. The R package 'DataVisualizations' is used.

Ockham's razor states that if two models are applicable, the less complex one should be used [61]. Therefore, the authors suggest investigating if simpler models can represent the data structures and provide meaningful and relevant explanations. A simpler projection approach assuming linear cluster structures and a simpler clustering [62] approach assuming spherical cluster structures is applied to the data. Moreover, spherical cluster structures are tested with the Silhouette plot using the R package' DataVisualizations '.

2.4 Step III: Providing Meaningful and Relevant Explanations

Explainability should follow the Gricean maxims of quality, relevance, manner, and quantity [63], for the usage of decision trees in this work. They summarized to meaningful and relevant explanations which should be then interpreted by a domain expert [8].

2.4.1 Decision Trees for Identified Data Structures

The conventional usage of decision trees is usually supervised, requiring a prior classification (e.g. [64]) but, alternatively, can also be unsupervised using split evaluation criteria that do not require a prior classification (e.g. [25]). Supervised decision trees are, for example, the classification and regression tree (CART) [65] or globally optimal classification and regression trees [66]. Here, we propose a third approach by using the clustering labels provided by cluster analysis in step III instead of a prior classification. Contrary to common usage, the decision tree is exploited here to explain the identified data structures. The supervised decision trees are computed using the labels and unprocessed or back-transformed data.

2.4.2 Extracting Meaningful Explanations

The maxim of quality states that only well-supported facts and no false descriptions should be reported. Quality will be measured by the accuracy of supervised decision trees representing the clustering (see section 2.5). The maxim of manner suggests to be brief and orderly and to avoid obscurity and ambiguity [63]. For the explanation to be in an appropriate manner, the standardization has to be either back-transformed to provide SI units of measurement or unprocessed data has to be used. The maxim of quantity states that neither too much nor too few explanations should be presented [63]. This work specifies the statement in the sense that the number of explanations should follow the Miller optimum of 4-7 [67,68]. Then, explanations are meaningful to a domain expert (c.f. discussion [8]).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

However, Decision tree algorithms do not aim at meaningful explanations [63,69]. Therefore a transformation of the decision tree into rules is necessary [63,69]. Here, rules are extracted from the decision tree by following each path from the root to the leaf. Exemplary, the R package "rpart" and the package "evtree" are applied. The number of rules measures the property of meaningfulness.

2.4.3 Evaluating the Relevance of Explanations

The maxim of relevance requires that only rules relevant to the expert are listed. Typically, explanations are especially relevant if they are tendentially contrastive (c.f. [9])). Suppose an explanation based on a clustering of the data is relevant (i.e., reveals to the domain expert relevant high-dimensional structures for similar days). In that case, the classes defined by such explanations should contain samples of different environmental states and be based on different processes. The property of relevance is qualitatively evaluated by class mirrored-density plots class (MD plots) [10]. Additionally, statistical testing of class-wise distributions of features can be performed to ensure that the classes defined by rules are tendentially contrastive and, in consequence, relevant.

The Mirrored-Density plot (MD-plot) introduced in [10] visualizes a density estimation in a similar way to the violin plot [70]. The MD-plot uses for density estimation, the Pareto density estimation (PDE) approach [71]. It can be shown that comparable methods have difficulties in visualizing the probability density function in case of uniform, multimodal, skewed, and clipped data if density estimation parameters remain in a default setting [10]. In contrast, the MD plot is particularly designed to discover interesting structures in continuous features and can outperform conventional methods [10]. The MD plot does not require any adjustments in density estimation parameters, which makes the usage compelling for non-experts. The class MD-plot is available in the R package 'DataVisualizations'. The class MD plot visualizes the density of each class of an interesting feature separately and is used to show the relationship of the clusters of the XAI systems to NO_3 and EC concentrations.

3 Results

An overview of the analysis is provided in Fig. 1. For clarity, the rest of this chapter is subdivided into three sections: the first section consists of the selection of an appropriate distance metric, extracting the first hypothesis from the distribution of distances (3.1) and providing the topographic map. However, the projected points in the topographic map are already colored by the clustering of the second step. The second section presents the projection-based cluster analysis method and validates the clustering (3.3). The third section presents and evaluates explanations.

3.1 Step I: Structure Identification

In supplementary Information A (SI A), the probability density distributions of the 12 finally selected and preprocessed variables are visualized with mirrored-density plots [10] (see also 2.4.3 Step III). SI A shows the result of an appropriate standardization of features resulting in similar variances.

Exemplary, Fig. 2 presents the probability density estimation (PDE) [71] of the distance feature *df* of the Hellinger point distance matrix of the preprocessed data in black and its Gaussian mixture model (GMM) in red. Specific definitions can be found in SI F. The Hellinger point distance [72] in the R package 'parallelDist' was chosen for cluster analysis because the distribution of distances is statistically not unimodal according to Hartigan's dip test [73] (with $p(D = 0.006385, N= 61425)<0.001$). The distance distribution can be modeled through the Gaussian mixture model (GMM) using the expectation-maximization (EM) algorithm [74]. The distance distribution and GMM are visualized in Fig. 2. The QQ-plot verifies the GMM in Fig. 3. This serves as an indication of the existence of high intercluster distances (distances between different clusters) and outlier distances as well as small intracluster distances (distances within each cluster), meaning that a distance-based cluster structure can be found. The Bayes boundary of the GMM in Fig. 2 is 0.39.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

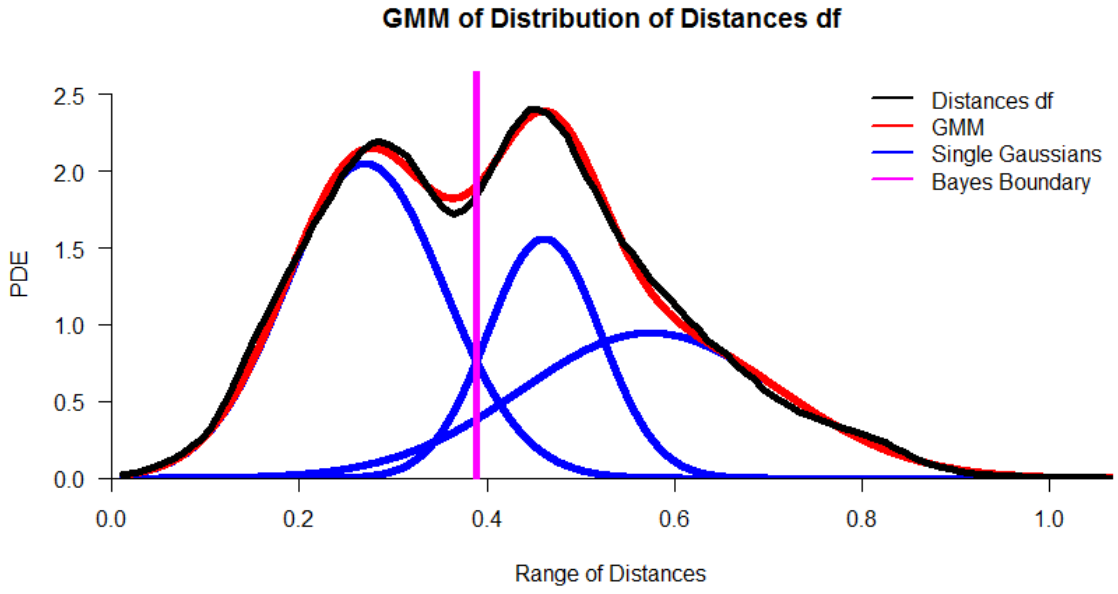

Figure 2: Distribution analysis of the distances using a Gaussian mixture model (GMM) using the R package 'AdaptGauss '. The black line indicates the estimated distribution of the distance feature df (defined in SI F). The distribution is estimated using PDE. The blue line depicts the single Gaussian distributions ("modes") of the model and the red line the overall model, i.e., the superposition of single Gaussians to a mixture. Bayes Boundary in magenta separates the first mode from the second mode and the third mode leading to the hypothesis that the first mode should consist of intra-cluster distances if clustering is performed. PDE=Pareto Density Estimation (Ultsch, 2005).

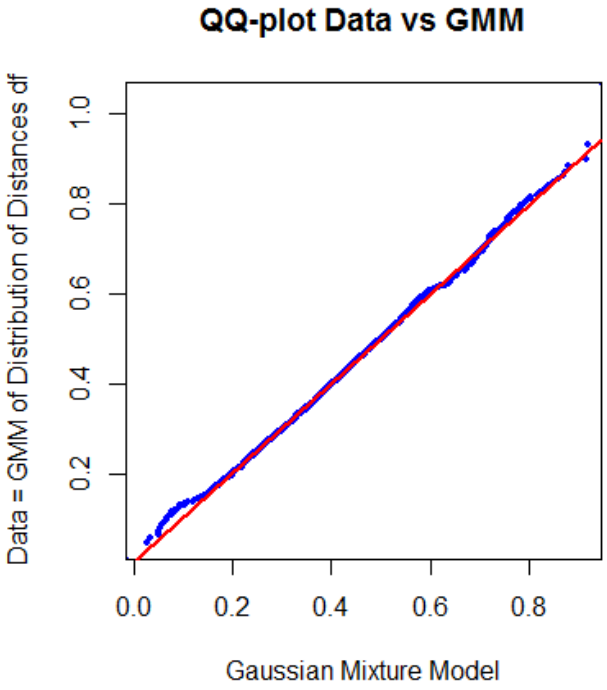

Figure 3: Quantile-Quantile plot (QQ plot) visualizes a good match between the distance and the GMM through a straight line. The plot is generated using the R package 'AdaptGauss'.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

Next, the topographic map of high-dimensional structures evaluates the clustering by indicating which points are in the high-dimensional space far away (brown/white hills) or near (blue seas, green grassland). The topographic map with hypsometric tints generated with the R package 'GeneralizedUmatrix' is toroidal, meaning that the grid's borders are cyclically connected with a periodicity defined by the size of the grid of the projection of the Databionic swarm. In Fig. 4, a cutout island of the topographic map is shown. Every point symbolizes a day. The high-dimensional distances of the low-dimensional projected points are visualized. The topographic map shows three valleys and basins indicating clusters and watersheds of hills and mountains shown by borderlines between clusters. Thus, the number of clusters is equal to the number of valleys. Without the cluster analysis providing the labels, all projected points would have the same color.

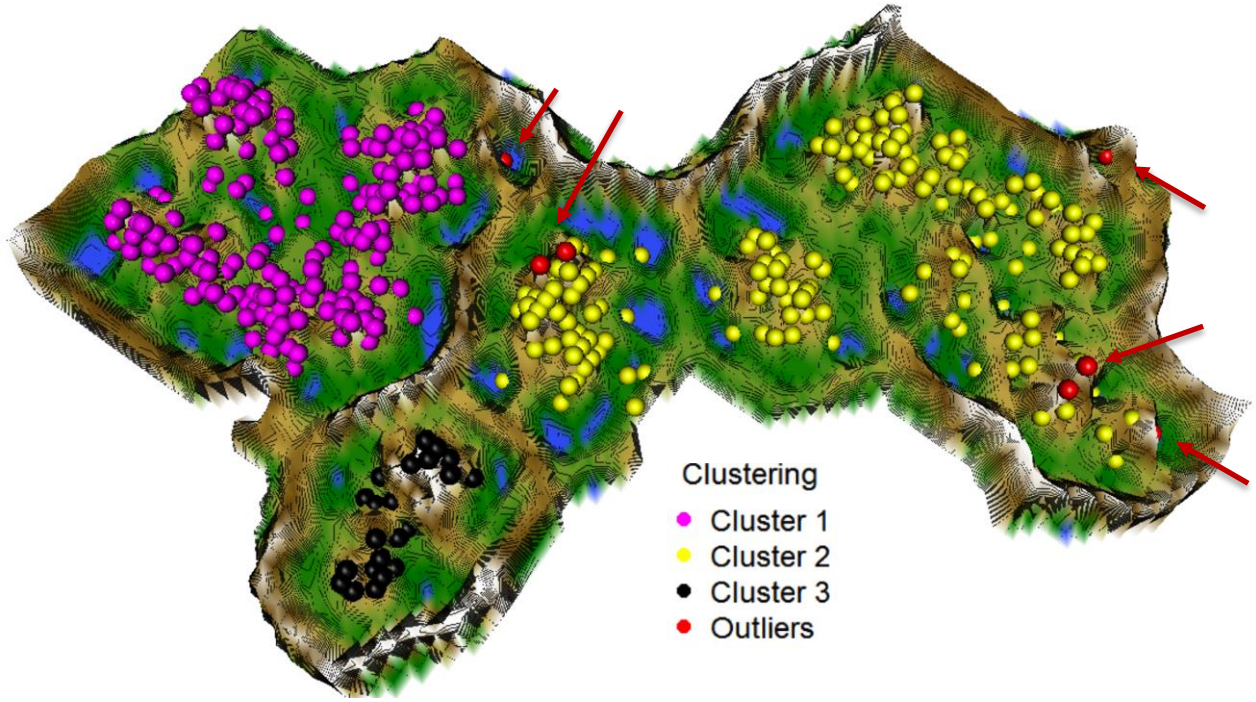

Figure 4: In the topographic map of high-dimensional structures, every point symbolizes a day and is colored by the independently performed clustering. The color of the points is defined by the labels of projection-based clustering. Clusters lie in valleys. The topographic map shows two main clusters (magenta and yellow points), a smaller cluster (black points). In addition, seven single outliers (marked in red and by red arrows) in the hydrology dataset are disregarded before comparable XAIs are applied. Visualization of high-dimensional data structures is generated using the R package "GeneralizedUmatrix".

3.2 Step II: Cluster Analysis

In Figure 4, the labels of the clustering are hereafter visualized as the colors of the projected points. In addition to the two main clusters (magenta and yellow points) and one outlier cluster (black points), seven outliers can be identified as volcanoes or within the valleys indicated by red arrows in Fig. 4. Next, the authors created a heatmap in order to verify the clustering. The heatmap shows intra- versus intercluster distances ordered by each cluster (Fig. 5). Blue colors symbolize small distances, and yellow and red colors represent large distances. The heatmap depicts clusters' homogeneity because the pattern of blue and teal colors is present for intra-cluster distances and yellow to the red color pattern for inter-cluster distances. The median intra-cluster distances of clusters 1, 2, and 3 are 0.24, 0.36, and 0.31, respectively and are below the Bayes boundary of the GMM in Fig. 2 of 0.39. The average intercluster distance is 0.48 and above the Bayes boundary of 0.39. These results indicate that the intracluster distances are smaller than the intercluster distances. This means that days within each cluster are evidently more similar to one another than days between clusters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

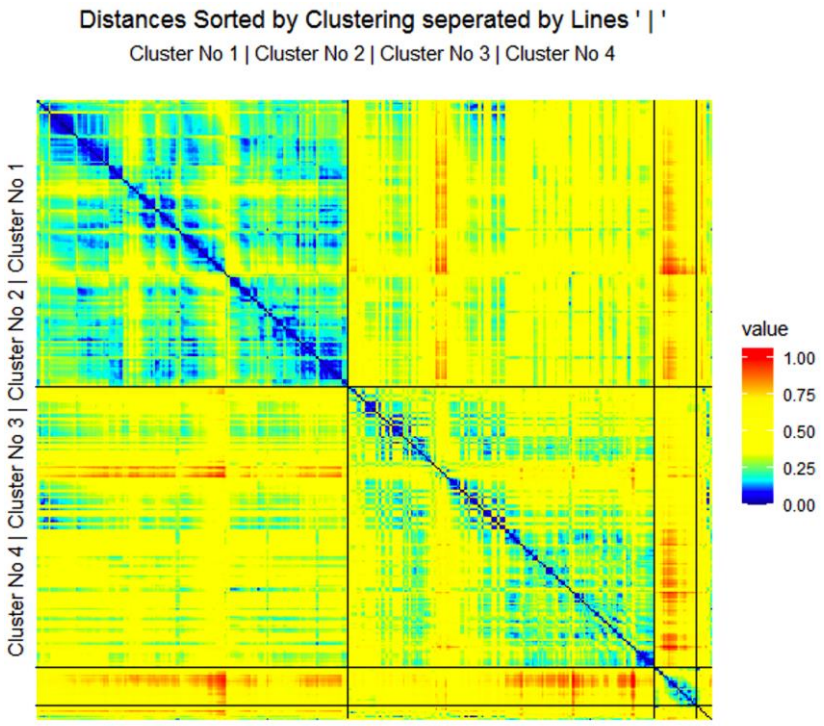

Figure 5: The four clusters have distinctive distances, as shown by the heatmap. The black lines divide the distances between the data points belonging to a cluster. The outliers are summarized in cluster 4. There are small distances within each cluster and large distances between the clusters. The heatmap was generated with the R package 'DataVisualizations'.

To check for a possibility of a simpler model, a linear projection by the method projection pursuit [75] using a clusterability index of variance and ratio (c.f. [76]) is applied on the dataset. The linear projection does not reveal clear structures, even if the generalized U-matrix is applied to visualize high-dimensional distance structures in the two-dimensional space (Figure SI G, Fig. 11). Therefore, it can be assumed that the structures cannot be separated linearly, motivating the usage of more complex and elaborate methods. The clustering can be reproduced with an accuracy of 86% using hierarchical clustering as described by Ward [77] if the seven outliers are disregarded because the method is sensitive to outliers [78]. Silhouette plots (SI D, Fig. 9) indicate inappropriate values for this clustering procedure if a spherical cluster structure is assumed.

3.3 Step III: Providing Explanations

For the explanation of the clustering as described in step II, non-standardized features have to be used because the system should explain the clustering to the domain expert and not the data scientist [8]. The clusters are explained by applying the evtree [66] and CART algorithm [65,79]. The evtree decision tree is shown in Fig 6a, and the CART decision tree for the data is visualized in Fig. 6b. Both decision trees agree on the same features sets and relations for each cluster except for cluster three, for which Rain <0.2 is not required to differentiate from cluster one and two in evtree although that makes cluster 3 less meaningful. The boundaries vary slightly between CART and evtree. None of the outliers could be explained by either evtree or CART. CART has a lower error and improves the meaningfulness of cluster three. Therefore, the rules are extracted from the CART tree instead of the evtree by following a path from root the leaf. For example, rule R1 explains cluster one of figure 4 (magenta). The combination of rules and clusters describes the classes which could predict future NO_3 and EC values (Table 2). The description of class 2 gains more detail if maximum likelihood plots of rain and water temperature (Wt18) are used (SI E, Fig. 10).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

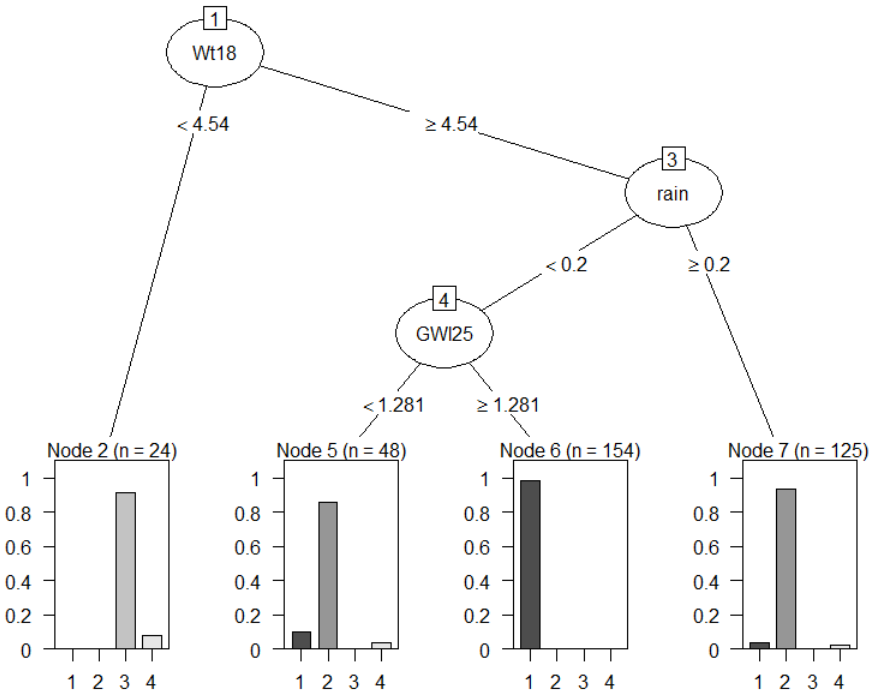

Figure 6a: Globally optimal classification and regression trees (evtree) analysis visualizes a decision tree for the dataset using the labels of the three clusters identified by projection-based clustering. The error of class 1 is 15%, class2 is 6.4%, and class 3 is 8.3%. Outliers are summarized in class 4. The rules are quite similar to Fig 6B but have a higher error. The tree was generated using the R package' evtree.

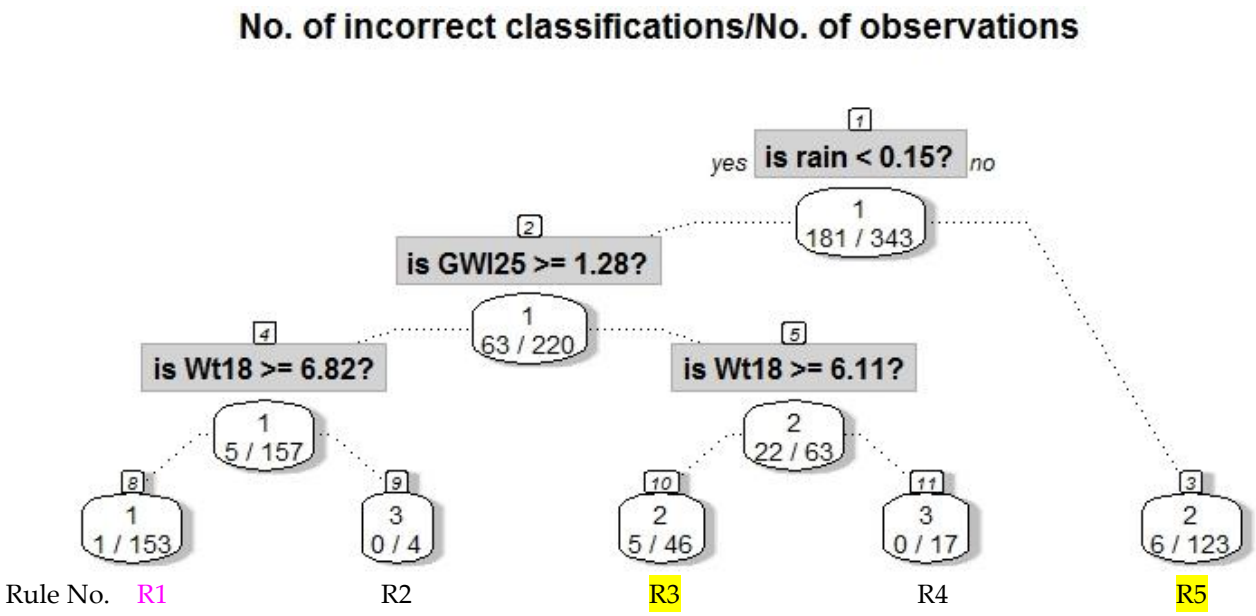

Figure 6b: Classification and regression tree (CART) analysis visualizes a decision tree for the dataset using the labels of the three clusters identified by projection-based clustering. Applying the algorithm to the labels of the clustering in combination with the dataset results in 12 misclassified points (3.5% of daily observations). 8 outlier points are in class 4 for which nodes can be derived. The leaves are identified with rule numbers used in table 2 and colors of Fig. 4. This error is lower than in Fig. 6a. For units of measurements and abbreviations, please see table 1. The tree was generated using the R package 'rpart'.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

Table 2: Explanations based on rules derived from the decision tree of Fig. 6b. Abbreviations: rainfall intensity (rain), soil temperature (St24), soil moisture (Smoist24), and water level at point 3 (GWl3). All values are expressed as percentages. For units of measurement, please see table 1. Class 2 R5 is extended by SI E, Fig. 10. The color names of the projected points of Fig. 4 are mapped to the rules of this table.

Rule No. *Color*	**Class No.**	**No. of Days**	**Explanations**	**Short Description of Class for Subsequent Plots**
R1 *magenta*	1	162	rain < 0.15 and GWl25≥1.28 and Wt18 ≥6.86 => *Dry days, increased stream water temperature and groundwater levels*	*DryDaysWarmWater*
R3 & R5, Fig. 10 *yellow*	2	159	rain < 0.15 and GWl25 < 1.28 and Wt18 ≥6.11 or rain >= 0.15 and Wt18 ≥6.11 => Intermediate stream water temperature with either dry days and low groundwater levels or rainy days with a high level of water	Duality
R2 & R4 *black*	3	22	rain < 0.15 and GWl25 ≥ 1.28 and Wt18 < 6.86 or rain < 0.15 and GWl25 < 1.28 and Wt18 < 6.11 or => *Dry days with colder stream water and variable groundwater levels*	*DryDaysColdWater*
-	Uncla ssified	7	Excluded, because cannot be explained with decision trees	Outliers

3.3.1 Evaluating Explanations for different XAIs

Next, we investigate the NO_3 and EC probability density distributions per class. In the last section, the clusters were explained by rules to define classes. The class-dependent MD-plots of Fig.7a and Fig. 7b show that the classes depend on normal or high NO_3 levels (Fig. 7a) as well as on low, intermediate or high conductivity levels (Fig. 7b) because the distributions of classes differ significantly from one another, with the exception of NO_3 classes 2 and 3. It is confirmed by Kolmogorov–Smirnov tests (SI C) that the classes differ significantly from each other in the NO_3 and EC distributions, except for class 2 versus class 3 in NO3. However, class 2 and class 3 also differ significantly from each other in the variables of rain and Wt18 (water temperature) in SI E, Fig. 10. Therefore, the explanations are relevant to the domain expert.

Applying the eUD3.5 algorithm [27] to the unprocessed data identified three clusters and resulted in 541 rules that explain various overlaps in the data points of the three clusters. The seven outliers identified in our analysis were disregarded from the data before using eUD3.5. In comparison, the XAI framework proposed here provides five rules. Furthermore, the class MDplot for nitrate does not show different states of water bodies for eUD3.5 (SI H, Fig. 12, right), but one high state of electric conductivity can be identified (SI H, Fig. 12, left).

Dasgupta et al. did not provide any source code in their work [28]. Therefore, the first part of the IMM algorithm [28], the k-means clustering [62] was performed with the unprocessed data. The seven Outliers identified in our analysis were disregarded from the data. Measuring the feature importance for this clustering [80] indicates that it is based mainly on sol71 leading to the assumption that the second part of the IMM algorithm, the decision tree would favor this feature strongly to explain the clusters. Compared to the DBS clustering, the contingency table is presented in SI B, table 3, and does not show an overlap of clusters between projection-based clustering and the first part of the IMM algorithm, k-means. Additionally, the class MDplots are presented in SI H, Fig 13, which do not show different states of water bodies, meaning that IMM's explanation would not be relevant to the domain expert.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

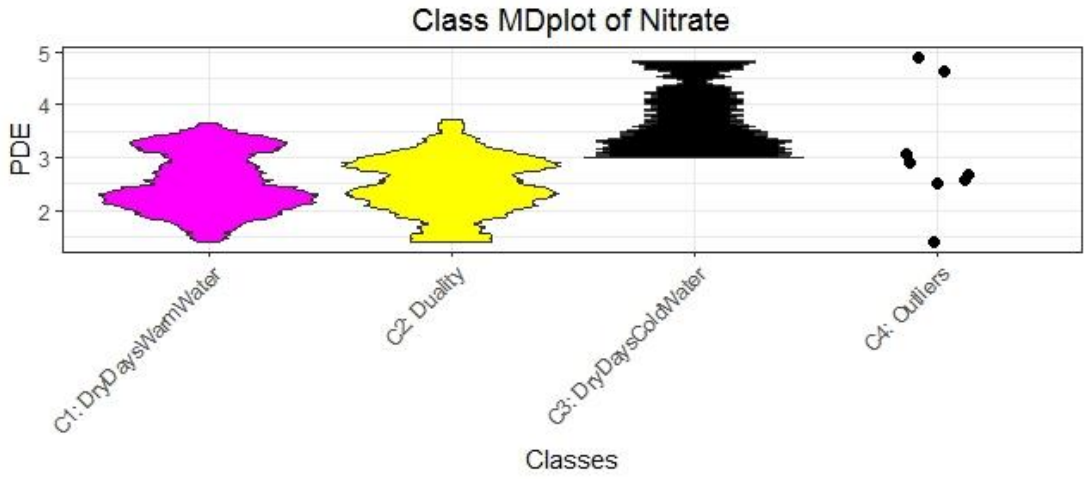

Figure 7a: Class-wise mirrored-density plot (MD-plot) of the three explained classes with regard to NO_3 and the outliers. There are two low to intermediate classes of N concentrations and one class of high N concentrations. Classes are colored similar to the clusters in Fig. 4. The MD-plot was generated using the R package 'DataVisualizations'.

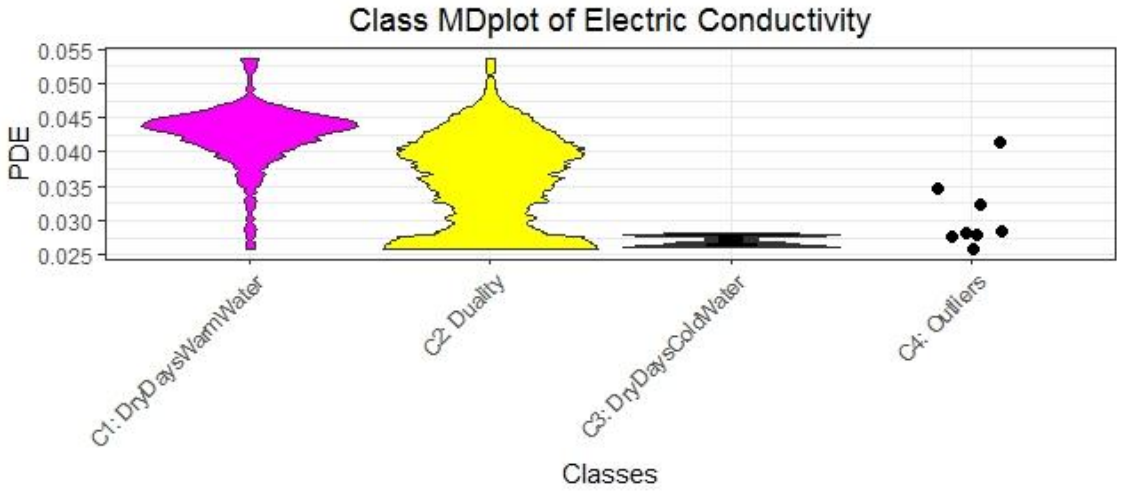

Figure 7b: Class-wise mirrored-density plots (MD-plot) of the three explained classes with regard to electrical conductivity C. There is a class of high concentration, a class of low to intermediate concentration, and a class of low C concentrations. Classes are colored similarly to the clusters in Fig. 4. The MD plot was generated using the R package 'DataVisualizations',

3.3.2 Interpreting Explanations

The acquired relevant and meaningful explanations by the XAI framework proposed in this work (Tab. 2) can be explained as follows: While water temperature governs the biological turnover of nitrogen compounds in the stream water, hydrological variables such a groundwater level determine how and whether terrestrial NO3 pools are connected to the stream system by activating flow pathways. Furthermore, the rainfall-runoff generation processes either concentrate or dilute the stream NO3 concentration, according to the difference in NO3 concentration in the stream and in the "new water" added to the stream system.

In the search for days with similar behavior, days with normal and high NO3 were identified. In 321 out of 343 days, the NO3 concentrations were normal (in the average range of [1, 3.5] mg/L). On such days, the concentrations of electric conductivity (EC) were either high (in the average range of [0.034, 0.055] mS/m)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

or intermediate to low (in the average range of [0.25, 0.045] mS/m). Normal NO3 and higher EC occurred on dry days with increased stream water temperature and higher groundwater levels. From a data-driven perspective, these days were highly similar to one another (c.f. cluster 1 in Fig. 4 and Fig. 5). The explanation for normal NO3 with low to normal EC concentrations is more complex and described by "duality": they likely had an intermediate stream water temperature (6.1°C<WT18<12.5°C) with either dry days (average rain < 0.15 mm) and low groundwater levels (<1.28 m) or rainy days with high groundwater levels (see SI E).

Simultaneously, high NO3 concentrations (in the average range of [3, 5.5] mg/L) and very low EC concentrations (in the average range of [0.025, 0.028] mS/m) occurred only if the stream water temperature was low on dry days. In particular, stream water temperature influences the activities of living organisms. The groundwater level (or head, in m) is the primary factor driving discharge in the Schwingbach catchment, while rainfall intensity triggers discharge and affects the leaching of nutrients [39].

4 Discussion

Selecting a suitable distance measure enabled to apply the process of clustering of the above-described dataset. It is assumed that cluster analysis is valid if intracluster distances are smaller (more similar to each other) than the intercluster distances. As a parameter-free clustering algorithm, the DBS was chosen. It enables the evaluation of the clustering with a topographic map in addition to the conventional heatmap. Projection-based clustering[36] with the projection of the Databionic swarm[35] is a flexible and robust clustering framework that has the ability to separate complex distance-based structures. The existence of data structures defining clusters and the number of clusters can be estimated prior to the clustering by the visualization of the topographic map. Such structures were identified by low intracluster distances and high intercluster distances of a Gaussian mixture model of the distance distribution and verified by the heatmap (Fig. 5) and topographic map (Fig. 4). However, a simpler linear model (SI G, Fig. 11) or spherical cluster structure were inappropriate (SI D, Fig.10, SI B, Table 3). It follows that most conventional approaches for clustering listed in [34] or recent XAIs [27,28] would be not appropriate to detect meaningful and relevant data structures. Statistical testing indicates that the distributions of interesting variables differ between classes (SI C and E). Further results imply that the explanations for the clustering were meaningful because brief rules were extracted by applying decision trees and using maximum likelihood plots. Overall, it can be deduced that this dataset contains linearly non-separable distance-based on non-spherical cluster structures that are meaningful and relevant to the domain expert.

The major difference of other XAIs to the approach followed here is that comparable approaches use unsupervised decision trees, whereas this work uses decision trees based on a clustering that is performed independently. This has two main advantages.

First, in Thrun and Ultsch, it was shown that k-means only can grasp spherical cluster structures, which is a severe restriction [36]. Moreover, the Silhouette plot is only useful in the case of spherical or ellipsoidal structures [29,81], leading to the assumption that eUD3.5 will prefer hyper ellipsoidal cluster structures and IMM will only work on specific cluster structures because it is based on k-means. In contrast, the XAI presented here outperforms k-means because it can find a large variety of cluster structures through the exploitation of emergence and self-organization [35]. These authors state that "Additionally, [clustering algorithms] may return meaningless results in the absence of natural clusters [82-84] in contrast to DBS, which will clearly indicate that no cluster structures exist" [35]. Moreover, DBS is able to discover small classes [35] whereas UD3.5 is not [27] (p. 52381).

Second, the compared approaches of eUD3.5 and IMM are limited to either use transformed data simultaneously providing no meaningful explanations to the domain expert, or to use non-transformed data. If unprocessed data with SI units are used, the results showed that comparable XAI would not provide meaningful explanations because for eUD3.5, many rules are presented for each cluster which do not cover the clusters well enough, and because IMM would focus on one feature. Moreover, neither IMM nor eUD3.5 provide a clustering that is relevant to the domain expert because the class-wise distributions of NO_3 and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

EC do not differ besides one class for EC (see. SI. H, Fig 12, 13). Hence, these classes do not define different states of water bodies.
In sum, the explanations are relevant and meaningful in our work, whereas for eUD3.5 and IMM the explanations are not relevant and meaningful.

5 Conclusion

No prior knowledge usable for a cluster analysis was available. Therefore, a machine-learned AI system relying on the Databionic swarm (DBS) method was used for the projection and clustering of environmental and water quality data. Explanations were provided in a three-step approach of "data structure identification" followed by cluster analysis for which explanations were provided using unprocessed data.

Explanations were provided by rules through a combination of supervised decision trees that trained on the labels of the clustering of preprocessed data. Contrary to a comparable XAI approach, the explanations of the XAI framework proposed here were meaningful and explained relevant content in a human-understandable way. The explanations suggest that the stream water quality data regarding NO_3 and EC can be described by a combination of one variable related to biological processes (water temperature) and two variables related to hydrological processes (rain and groundwater level). Our XAI provided explicit ranges of values and could enable future prediction of stream water quality. One comparable XAI (eUD3.5) failed to extract relevant and meaningful rules. Another XAI (IMM) failed because it focuses on specific cluster structures and features, hence relying on prior knowledge about data structure.

The XAI framework presented here allows for unbiased detection of meaningful data structures in high dimensionality datasets. Such datasets become more and more available, not only in hydrochemistry but also in other environmental disciplines due to the technical innovation in monitoring equipment. Our Explainable AI provides a unique possibility to search for unknown structures and can provide meaningful and relevant explanations because it does not rely on prior knowledge about data structure.

Supplementary Information A: Features after Preprocessing

Variables were preprocessed such that metric distances can be used because the range of every feature is approximately between zero and one. The distribution of the features is shown by the MD-plot [10] (see also 2.4.3 Step III for definition) in Fig. 8 of the 12 variables used for DBS projection, visualization, and clustering [33]. The complete aggregated dataset consisted of 343 days**.** The mirrored-density plots (MD-plots) show that the range of a variable is approximately between zero and one because the normalization approach uses 1 and 99% quantiles instead of maxima and minima, thus allowing outliers to lie below zero or above one.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

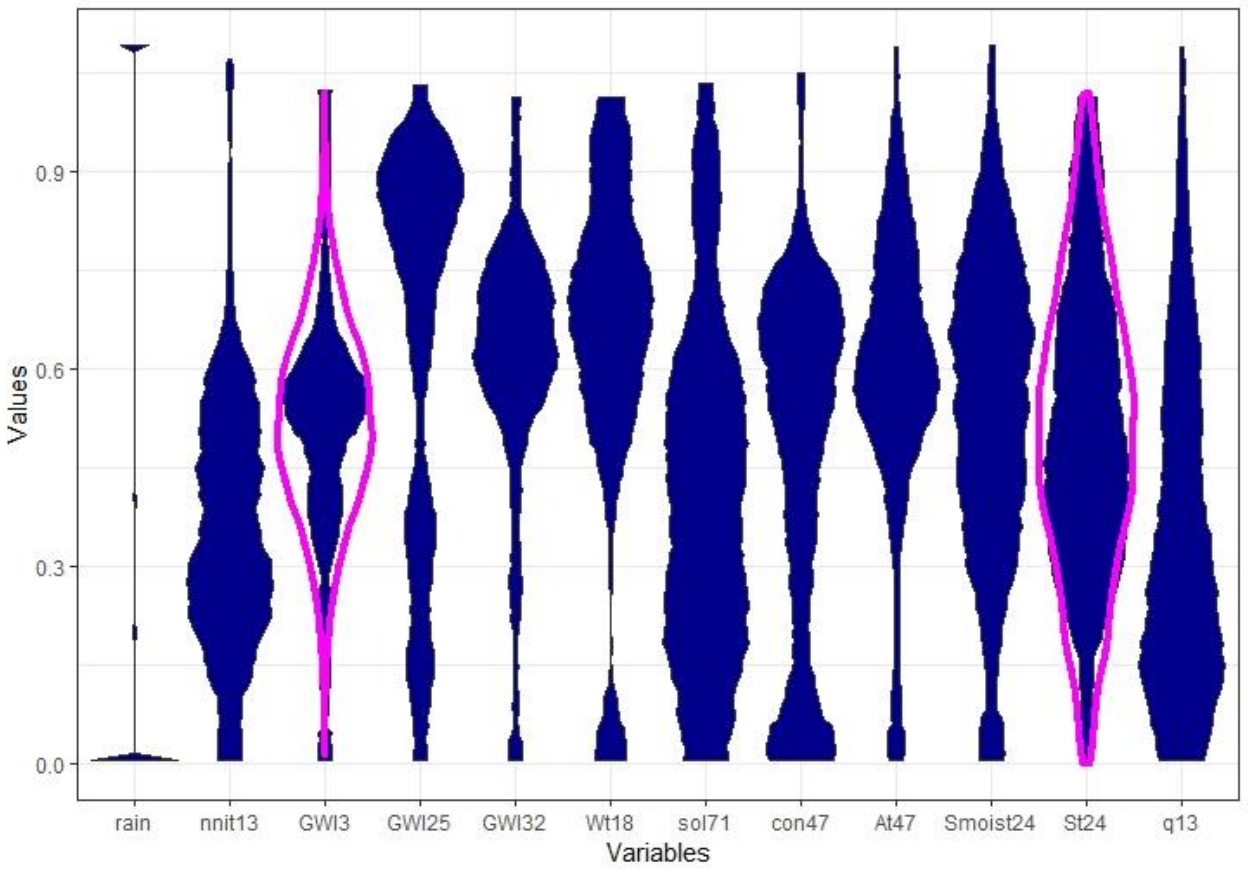

Figure 8: The distribution of variables after preprocessing is visualized using mirrored-density plots of the hydrology dataset. The magenta overlay marks features that are statistically not skewed or multimodal. The mirrored-density plot (MD-plot) was generated using the R package 'DataVisualizations'..

Supplementary Information B: Comparison to the K-means clustering approach

The clustering can be reproduced with an accuracy of 42% using the k-means algorithm [85] if the outliers are disregarded. The contingency table is presented below (Table 3).

Table 3: Contingency table of projection-based clustering versus k-means published in [62] and integrated in FCPS [86].

DBS/k-means	1	2	3	RowSum	RowPercentage
1	24	77	61	162	47.23
2	28	65	66	159	46.36
3	0	8	14	22	6.41
ColumnSum	52	150	141	343	0
ColPercentage	15.16	43.73	41.11	0	100

Supplementary Information C: Kolmogorov-Smirnov tests of clusters

Tables 4 and 5 compare the clustering achieved for conductivity and NO_3. The clusters should contain samples of different natures and are based on different processes. Given this assumption, it is valid to statistically test whether the NO_3 and EC distributions significantly differ between clusters. The Kolmogorov-Smirnov test (KS test) is a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

nonparametric two-sample test of the null hypothesis that two variables are drawn from the same continuous distribution [87]. For the first three clusters, the NO_3 and EC distributions significantly differ among clusters.

Table 4: KS test with test statistic (*D*) and p-value (*p*) for conductivity for the first three clusters.

Cluster No. (Sample Size)	C2 (159)	C3 (22)
C1 (162)	*D=0.13429, p=0.11*	*D=0.74074, p<0.001*
C2 (159)		*D=0.84906, p<0.001*

Table 5: KS test with test statistic (*D*) and p-value (*p*) for NO_3 for the first three clusters.

Cluster No. (Sample Size)	C2 (159)	C3 (22)
C1 (162)	*D=0.50769, p<0.001*	*D=0.98765, p<0.001*
C2 (159)		*D=0.83019, p<0.001*

Supplementary Information D: Silhouette Plot

The Silhouette plot of DBC clustering is presented in Fig.9 and demonstrates an inappropriate clustering w.r.t. spherical cluster structures.

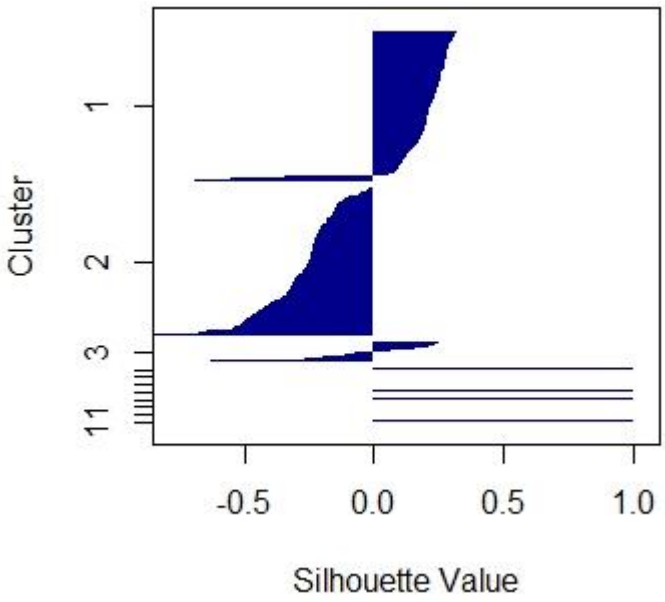

Figure 9: Silhouette plot of DBC clustering shows low values for the three main clusters, indicating inappropriate clustering with regard to expected spherical structures. The silhouette plot was generated using the R package 'DataVisualizations'.

Supplementary Information E: Distinction of Classes 1 and 2 in Regard to Rain and Water Temperature

Using the Kolmogorov-Smirnov test (KS test), which is a nonparametric two-sample test of the null hypothesis that two variables are drawn from the same continuous distribution [87], Class 1 significantly differs from Class 2 in the variable Wt18 (water temperature) with p<(162,159, D= 0.31982)<0.001 and in the variable rain with p<(162,159, D=

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

0.70498)<0.001. This is visualized in the class-wise maximum-likelihood plots of Fig. 10. Moreover, Fig. 10(right) shows that the water temperature in Class 2 is more likely to be lower than that in Class 1 and less likely to be lower than that in Class 3.

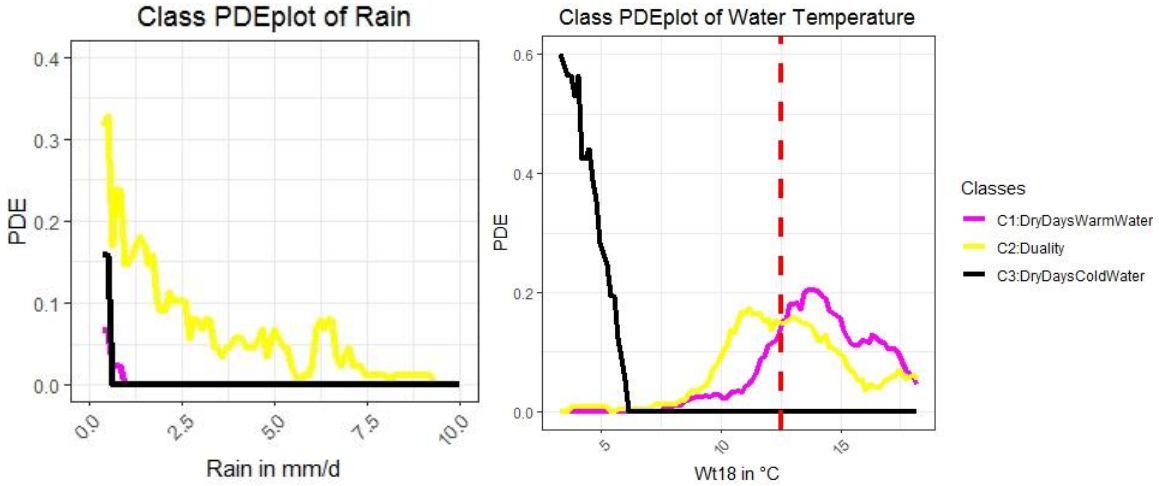

Figure 10: Class wise estimation of the probability density function using PDE allows for a more precise definition of Class 2, "Duality", because the plot shows that in Class 2, there are also rainy days with colder water than in Class 3. The red and dashed line in the right plot marks a temperature of 12.5°C. Classes are colored similarly to the clusters in Fig. 4.

Supplementary Information F: Distance Distributions

Let I be a finite subset of N high-dimensional points in a metric space with a distance function d(l,j), then the matrix $D = (D_{l,j})_{l,j \in I}$ is called a distance matrix of I (c.f. [88]) with each entry as $D_{l,j} = d(l,j)$ being the distance between two high-dimensional points of data. The distance matrix D satisfies four conditions, meaning that the diagonal entries are all zero ($d(l,l) = 0 \; \forall \; 1 \leq l \leq N$), positive ($d(l,j) > 0 \; \forall \; l \neq j$), symmetric $d(l,j) =$ d(j, l) and for any l,j $d(l,j) \leq d(l,k) + d(k,j), \forall k$ (triangle inequality). Using the definition above, we define the distance feature df as the upper (or lower) triangle of the symmetric distance matrix ($df = D_{lj}, \; \forall \; l > j \; , 1 \leq l \leq N, 1 \leq j \leq N$).

Given a finite dataset I of N cases, each described by d features, the Euclidean distance is defined as

d(l, j) $= \sqrt{\sum_i (l_i - j_i)^2}$ which can be modified to the Hellinger point distance with d(l, j) $= \sqrt{\sum_i \left(\sqrt{\frac{l_i}{\sum l_i}} - \sqrt{\frac{j_i}{\sum j_i}} \right)^2}$

c.f. ([72,89,90]). For details regarding the Hellinger point distance and the application to the data, we refer to parallelDist::parallelDist() [91] and the provided source code.

Supplement Information G: Linear Model

In Fig. 11, the generalized U-matrix is applied to a linear projection to visualize high-dimensional distance structures in the two-dimensional space. No structures are visible on the topographic map.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

Figure 11: Toroidal topographic map of a projection pursuit approach by [75] of the hydrology dataset. The linear projection does not reveal a linear structure, even if the generalized U-matrix is used to visualize high-dimensional distances of the two-dimensional projection [53]. This visualization was generated using the R package 'GeneralizedUmatrix" and the projection and clustering using "FCPS".

Supplement Information H: Class MDplots of eUD3.5 and k-means

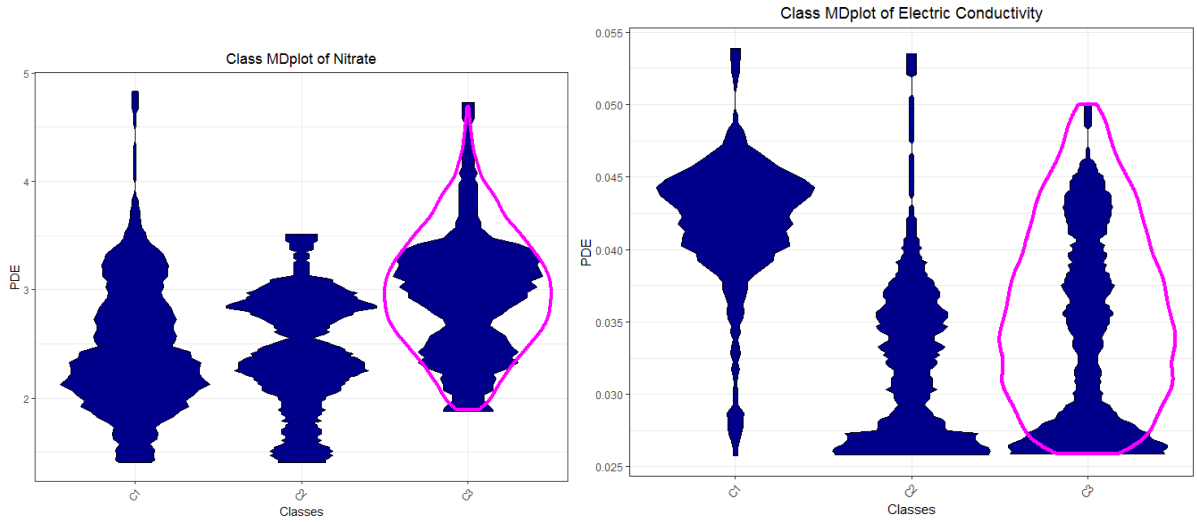

Figure 12: Class wise mirrored-density plot (MD-plot) of the three classes defined 541 rules of by eUD3.5 with regard to nitrate NO_3 (left) and electrical conductivity C (right). Seven Outliers identified in DBS were priorly disregarded. In the case of nitrate, no clear differences between the distributions of the classes are visible. There is one high to intermediate classes of C concentrations and classes of low to intermediate C concentrations. The MD-plot was generated using the R package 'DataVisualizations'.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2020 doi:10.20944/preprints202011.0451.v1

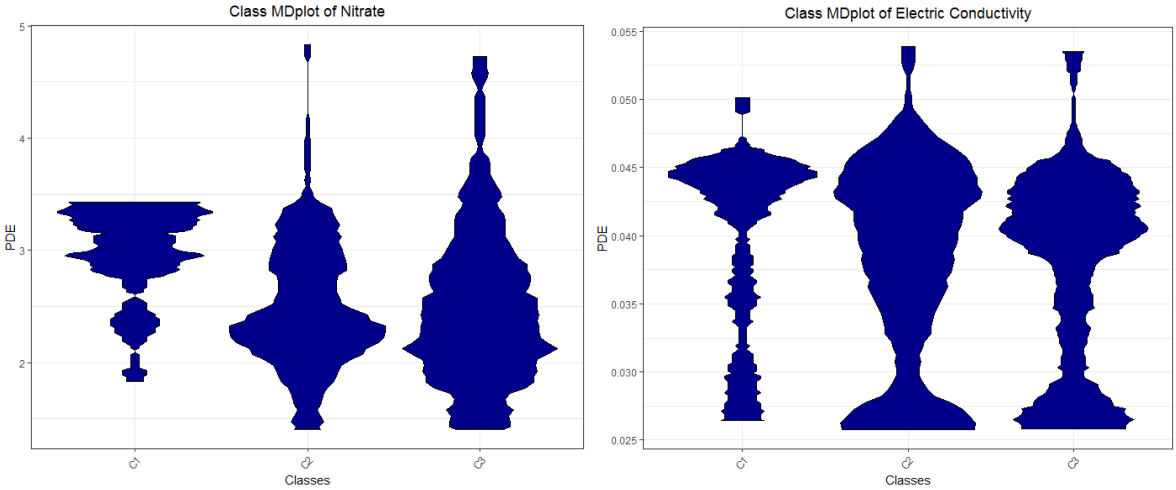

Figure 13: Class wise mirrored-density plot (MD-plot) of the three classes defined by k-means with regard to nitrate NO_3 (left) and electrical conductivity C (right). . Seven Outliers identified in DBS were priorly disregarded.. No clear differences between the distributions of the classes are visible. The MD-plot was generated using the R package 'DataVisualizations'.

Supplement Information I: Table of Used R Packages

Table 6: The following R packages available on CRAN are used in this work.

Name of Packag	Usage	Reference	Accessibility
ABCanalysis	Computed ABCanalysis for outlier detection	[41]	https://CRAN.R-project.org/package=ABCanalysis
DataVisualizations	Mirrored density plot (MD plot), density estimation, heatmap	[92]	https://CRAN.R-project.org/package=DataVisualizations
FCPS	54 alternative clustering algorithms for specific cluster structures	[86]	https://CRAN.R-project.org/package=FCPS
DatabionicSwarm	Projection algorithm that finds a large variety of cluster structures and can cluster data as a special of Projection-based clustering.	[35]	https://CRAN.R-project.org/package=DatabionicSwarm
parallelDist	Distance computation for many distance metrics	[91]	https://CRAN.R-project.org/package=parallelDist
AdaptGauss	Gaussian Mixture Modelling (GMM), QQ plot for GMM	[43]	https://CRAN.R-project.org/package=AdaptGauss
rpart	Supervised Decision Tree	[79]	https://CRAN.R-project.org/package=rpart

evtree	Supervised Decision Tree	[66]	https://CRAN.R-project.org/package=evtree
GeneralizedUmatrix	Provides the topographic map, enables to visualize any projection method with it	[53]	https://CRAN.R-project.org/package=GeneralizedUmatrix
ProjectionBased Clustering	Provides projection-based clustering, interactive interfaces for cutting tiled topographic map into islands and for interactive clustering	[38]	https://CRAN.R-project.org/package=ProjectionBasedClustering
FeatureImpCluster	"Implements a novel approach for measuring feature importance in k-means clustering"	[80]	https://CRAN.R-project.org/package=FeatureImpCluster

Code availability

Every function used in this manuscript is available in R packages on CRAN and is referenced throughout the text and summarized in table 6, SI I. The specific application of these functions to the analyzed data is available in https://github.com/Mthrun/ExplainableAI4TimeSeries2020/08AnalyseProgramme/. If not stated otherwise, no setting of parameters or changing of default parameters is necessary to reproduce the results above with the limitation that stochastic algorithms like most clustering and projection methods have a variance of results depending on the trial (c.f. discussion in [35]). The exact version of the model used to produce the results in this paper and the results of the eUD3.5 XAI are archived on Zenodo: DOI: 10.5281/zenodo.4274700 under GPL license.

Data availability

The raw data is available on GitHub: https://github.com/Mthrun/ExplainableAI4TimeSeries2020/90RawData/. Aggregated data is available at https://github.com/Mthrun/ExplainableAI4TimeSeries2020/09Originale/. The exact version of the model used to produce the results used in this paper is archived on Zenodo: DOI: 10.5281/zenodo.4274700 under GPL license.

Author contribution

Lutz Breuer devised the project and collected the data. Michael Thrun wrote the manuscript with major support from Lutz Breuer regarding all non-data science aspects. Michael Thrun designed the model and the computational framework and analyzed the data. Alfred Ultsch supervised the project, checked calculations and programs and contributed and revised the manuscript. All authors discussed the results and contributed to the final manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgments

We thank Alice Aubert for fruitful discussions regarding the interpretation of the results and Hamza Tayyab for compiling the C# Source code of eUD3.5 and applying it to the multivariate time series data.

References

1. Durand, P.; Breuer, L.; Johnes, P.J. Chapter 7: Nitrogen processes in aquatic ecosystems. In *European Nitrogen Assessment (ENA)*, al., S.e., Ed. Cambridge University Press: 2011; pp. 126-146.
2. Cirmo, C.P.; McDonnell, J.J. Linking the hydrologic and biogeochemical controls of nitrogen transport in near-stream zones of temperate-forested catchments: a review. *Journal of Hydrology* **1997**, *199*, 88-120.
3. Diaz, R.J. Overview of hypoxia around the world. *Journal of environmental quality* **2001**, *30*, 275-281.
4. Howarth, R.W.; Billen, G.; Swaney, D.; Townsend, A.; Jaworski, N.; Lajtha, K.; Downing, J.A.; Elmgren, R.; Caraco, N.; Jordan, T. Regional nitrogen budgets and riverine N & P fluxes for the drainages to the North Atlantic Ocean: Natural and human influences. In *Nitrogen cycling in the North Atlantic Ocean and its watersheds*, Springer: 1996; pp. 75-139.
5. Rode, M.; Wade, A.J.; Cohen, M.J.; Hensley, R.T.; Bowes, M.J.; Kirchner, J.W.; Arhonditsis, G.B.; Jordan, P.; Kronvang, B.; Halliday, S.J., et al. Sensors in the stream: the high-frequency wave of the present. *Environmental Science & Technology* **2016**, *50*, 19, doi:10.1021/acs.est.6b02155.
6. Aubert, A.H.; Thrun, M.C.; Breuer, L.; Ultsch, A. Knowledge discovery from high-frequency stream nitrate concentrations: hydrology and biology contributions. *Scientific reports* **2016**, *6*, doi:10.1038/srep31536.
7. Aubert, A.H.; Breuer, L. New seasonal shift in in-stream diurnal nitrate cycles identified by mining high-frequency data. *PloS one* **2016**, *11*, e0153138, doi:10.1371/journal.pone.0153138.
8. Miller, T.; Howe, P.; Sonenberg, L.; AI, E. Explainable AI: Beware of inmates running the asylum. In Proceedings of International Joint Conference on Artificial Intelligence, Workshop on Explainable AI (XAI); pp. 36-42.
9. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* **2019**, *267*, 1-38.
10. Thrun, M.C.; Gehlert, T.; Ultsch, A. Analyzing the Fine Structure of Distributions. *PLoS ONE* **2020**, *15*, e0238835, doi:10.1371/journal.pone.0238835
11. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138-52160.
12. Pellerin, B.A.; Downing, B.D.; Kendall, C.; Dahlgren, R.A.; Kraus, T.E.; Saraceno, J.; Spencer, R.G.; Bergamaschi, B.A. Assessing the sources and magnitude of diurnal nitrate variability in the San Joaquin River (California) with an in situ optical nitrate sensor and dual nitrate isotopes. *Freshwater Biology* **2009**, *54*, 376-387.
13. Ultsch, A. The integration of connectionist models with knowledge-based systems: hybrid systems. In Proceedings of SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218); pp. 1530-1535.
14. Ultsch, A.; Korus, D. Integration of neural networks and knowledge-based systems. In Proceedings of IEEE Int. Conf. Neural Networks, Perth, Australia.
15. Biran, O.; Cotton, C. Explanation and justification in machine learning: A survey. In Proceedings of IJCAI-17 workshop on explainable AI (XAI); pp. 8-13.

16. Ribeiro, M.T.; Singh, S.; Guestrin, C. " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; pp. 1135-1144.
17. Lipton, Z.C. The mythos of model interpretability. *Queue* **2018**, *16*, 31-57.
18. Ultsch, A.; Halmans, G.; Mantyk, R. CONKAT: a connectionist knowledge acquisition tool. In Proceedings of Proceedings of the Twenty-Fourth Annual Hawaii International Conference on System Sciences; pp. 507-513.
19. Ultsch, A.; Korus, D.; Kleine, T. Integration of neural networks and knowledge-based systems in medicine. In Proceedings of Conference on Artificial Intelligence in Medicine in Europe; pp. 425-426.
20. Letham, B.; Rudin, C.; McCormick, T.H.; Madigan, D. An interpretable stroke prediction model using rules and Bayesian analysis. In Proceedings of Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence.
21. Riid, A.; Sarv, M. Determination of regional variants in the versification of estonian folksongs using an interpretable fuzzy rule-based classifier. In Proceedings of 8th conference of the European Society for Fuzzy Logic and Technology (EUSFLAT-13).
22. Nauck, D.; Kruse, R. Obtaining interpretable fuzzy classification rules from medical data. *Artificial intelligence in medicine* **1999**, *16*, 149-169.
23. Lakkaraju, H.; Bach, S.H.; Leskovec, J. Interpretable decision sets: A joint framework for description and prediction. In Proceedings of Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; pp. 1675-1684.
24. Hewett, R.; Leuchner, J. The power of second-order decision tables. In Proceedings of Proceedings of the 2002 SIAM International Conference on Data Mining; pp. 384-399.
25. Basak, J.; Krishnapuram, R. Interpretable hierarchical clustering by constructing an unsupervised decision tree. *IEEE transactions on knowledge and data engineering* **2005**, *17*, 121-132.
26. Kim, B.; Shah, J.A.; Doshi-Velez, F. Mind the gap: A generative approach to interpretable feature selection and extraction. In Proceedings of Advances in Neural Information Processing Systems; pp. 2260-2268.
27. Loyola-González, O.; Gutierrez-Rodríguez, A.E.; Medina-Pérez, M.A.; Monroy, R.; Martínez-Trinidad, J.F.; Carrasco-Ochoa, J.A.; García-Borroto, M. An Explainable Artificial Intelligence Model for Clustering Numerical Databases. *IEEE Access* **2020**, *8*, 52370-52384.
28. Dasgupta, S.; Frost, N.; Moshkovitz, M.; Rashtchian, C. Explainable $ k $-Means and $ k $-Medians Clustering. In Proceedings of 37th International Conference on Machine Learning,, Vienna, Austria.
29. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **1987**, *20*, 53-65, doi:https://doi.org/10.1016/0377-0427(87)90125-7.
30. Liao, T.W. Clustering of time series data—a survey. *Pattern recognition* **2005**, *38*, 1857-1874.
31. Ma, Q.; Zheng, J.; Li, S.; Cottrell, G.W. Learning Representations for Time Series Clustering. In Proceedings of Advances in Neural Information Processing Systems; pp. 3776-3786.
32. Ferreira, L.N.; Zhao, L. Time series clustering via community detection in networks. *Information Sciences* **2016**, *326*, 227-242.

33. Thrun, M.C. *Projection Based Clustering through Self-Organization and Swarm Intelligence*; Ultsch, A., Hüllermeier, E., Eds.; Springer: Heidelberg, 2018; 10.1007/978-3-658-20540-9.
34. Aghabozorgi, S.; Shirkhorshidi, A.S.; Wah, T.Y. Time-series clustering–A decade review. *Information Systems* **2015**, *53*, 16-38.
35. Thrun, M.C.; Ultsch, A. Swarm Intelligence for Self-Organized Clustering. *Artificial Intelligence* **2020**, *290*, 103237, doi:10.1016/j.artint.2020.103237.
36. Thrun, M.C.; Ultsch, A. Using Projection based Clustering to Find Distance and Density based Clusters in High-Dimensional Data. *Journal of Classification* **2020**, 10.1007/s00357-020-09373-2, doi:10.1007/s00357-020-09373-2.
37. Thrun, M.C.; Ultsch, A. Uncovering High-Dimensional Structures of Projections from Dimensionality Reduction Methods. *MethodsX* **2020**, *7*, 101093, doi:10.1016/j.mex.2020.101093.
38. Thrun, M.C.; Pape, F.; Ultsch, A. Interactive Machine Learning Tool for Clustering in Visual Analytics. In Proceedings of 7th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2020), Sydney, Australia, 06-09 Oct.; pp. 672-680.
39. Orlowski, N.; Lauer, F.; Kraft, P.; Frede, H.-G.; Breuer, L. Linking spatial patterns of groundwater table dynamics and streamflow generation processes in a small developed catchment. *Water* **2014**, *6*, 3085-3117.
40. Milligan, G.W.; Cooper, M.C. A study of standardization of variables in cluster analysis. *Journal of classification* **1988**, *5*, 181-204.
41. Ultsch, A.; Lötsch, J. Computed ABC Analysis for Rational Selection of Most Informative Variables in Multivariate Data. *PloS one* **2015**, *10*, e0129767, doi:10.1371/journal.pone.0129767.
42. Bouveyron, C.; Hammer, B.; Villmann, T. Recent developments in clustering algorithms. In Proceedings of ESANN.
43. Ultsch, A.; Thrun, M.C.; Hansen-Goos, O.; Lötsch, J. Identification of Molecular Fingerprints in Human Heat Pain Thresholds by Use of an Interactive Mixture Model R Toolbox (AdaptGauss). *International journal of molecular sciences* **2015**, *16*, 25897-25911.
44. Ultsch, A. Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series. In *Kohonen maps*, 1 ed.; Oja, E., Kaski, S., Eds. Elsevier: 1999; pp. 33-46.
45. Demartines, P.; Hérault, J. CCA:" Curvilinear component analysis". In Proceedings of 15° Colloque sur le traitement du signal et des images, France, 18-21 September.
46. Van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **2008**, *9*, 2579-2605.
47. Venna, J.; Peltonen, J.; Nybo, K.; Aidos, H.; Kaski, S. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *The Journal of Machine Learning Research* **2010**, *11*, 451-490.
48. Ultsch, A. Clustering with DataBots. In Proceedings of Int. Conf. Advances in Intelligent Systems Theory and Applications (AISTA), Canberra, Australia; pp. p. 99-104.
49. Ultsch, A.; Behnisch, M.; Lötsch, J. ESOM Visualizations for Quality Assessment in Clustering. In *Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of the 11th International Workshop WSOM 2016, Houston, Texas, USA, January 6-8, 2016*, Merényi, E., Mendenhall, J.M., O'Driscoll, P., Eds. Springer International Publishing: Cham, 2016; 10.1007/978-3-319-28518-4_3pp. 39-48.

50. Nash, J.F. Equilibrium points in n-person games. *Proc. Nat. Acad. Sci. USA* **1950**, *36*, 48-49.
51. Nash, J.F. Non-cooperative games. *Annals of mathematics* **1951**, 286-295.
52. Johnson, W.B.; Lindenstrauss, J. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics* **1984**, *26*, 189-206.
53. Ultsch, A.; Thrun, M.C. Credible Visualizations for Planar Projections. In Proceedings of 12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM), Nany, France; pp. 1-5.
54. Thrun, M.C.; Lerch, F.; Lötsch, J.; Ultsch, A. Visualization and 3D Printing of Multivariate Data of Biomarkers. In Proceedings of International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG), Plzen; pp. 7-16.
55. Dijkstra, E.W. A note on two problems in connexion with graphs. *Numerische mathematik* **1959**, *1*, 269-271.
56. Lötsch, J.; Ultsch, A. Exploiting the Structures of the U-Matrix. In Proceedings of Advances in Self-Organizing Maps and Learning Vector Quantization, Mittweida, Germany, July 2–4; pp. 249-257.
57. Murtagh, F. On ultrametricity, data coding, and computation. *Journal of classification* **2004**, *21*, 167-184.
58. Wilkinson, L.; Friendly, M. The history of the cluster heat map. *The American Statistician* **2012**.
59. Weinstein, J.N. A postgenomic visual icon. *Science* **2008**, *319*, 1772-1773.
60. Engle, S.; Whalen, S.; Joshi, A.; Pollard, K.S. Unboxing cluster heatmaps. *BMC bioinformatics* **2017**, *18*, 63.
61. Sober, E. Let's Razor Ockham's Razor. In *Explanation and its Limits*, Knowles, D., Ed. Cambridge University Press: Cambridge, 1991; 10.1017/CBO9780511599705.006pp. pp. 73-94.
62. Leisch, F. A toolbox for k-centroids cluster analysis. *Comput Stat Data An* **2006**, *51*, 526-544.
63. Mörchen, F.; Ultsch, A. Efficient mining of understandable patterns from multivariate interval time series. *Data Min Knowl Disc* **2007**, *15*, 181-215.
64. Breiman, L. Random forests. *Mach Learn* **2001**, *45*, 5-32.
65. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and regression trees*; CRC press: 1984.
66. Grubinger, T.; Zeileis, A.; Pfeiffer, K.-P. evtree: Evolutionary learning of globally optimal classification and regression trees in R. *Journal of Statistical Software* **2014**, *61*, 1-29, doi:10.18637/jss.v061.i01.
67. Miller, G.A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* **1956**, *63*, 81.
68. Cowan, N. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences* **2001**, *24*, 87-114.
69. Mörchen, F.; Ultsch, A.; Hoos, O. Extracting interpretable muscle activation patterns with time series knowledge mining. *International Journal of Knowledge-based and Intelligent Engineering Systems* **2005**, *9*, 197-208.
70. Hintze, J.L.; Nelson, R.D. Violin plots: a box plot-density trace synergism. *The American Statistician* **1998**, *52*, 181-184.

71. Ultsch, A. Pareto density estimation: A density estimation for knowledge discovery. In *Innovations in classification, data science, and information systems*, Baier, D., Werrnecke, K.D., Eds. Springer: Berlin, Germany, 2005; Vol. 27, pp. 91-100.
72. Rao, C. Use of Hellinger distance in graphical displays. Multivariate statistics and matrices in statistics. In Proceedings of Proceedings of the 5th Tartu Conference; pp. 143-161.
73. Hartigan, J.A.; Hartigan, P.M. The dip test of unimodality. *The annals of Statistics* **1985**, *13*, 70-84.
74. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **1977**, *39*, 1-22.
75. Hofmeyr, D.; Pavlidis, N. Maximum clusterability divisive clustering. In Proceedings of 2015 IEEE Symposium Series on Computational Intelligence; pp. 780-786.
76. Steinley, D.; Brusco, M.J.; Henson, R. Principal cluster axes: A projection pursuit index for the preservation of cluster structures in the presence of data reduction. *Multivariate behavioral research* **2012**, *47*, 463-492.
77. Ward Jr, J.H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **1963**, *58*, 236-244.
78. Everitt, B.S.; Landau, S.; Leese, M.; Stahl, D. Hierarchical clustering. *Cluster Analysis, 5th Edition* **2011**, 71-110.
79. Therneau, T.; Atkinson, B.; Ripley, B.; Ripley, M.B. Package 'rpart'. *Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016)* **2018**.
80. Pfaffel, O. *FeatureImpCluster: Feature Importance for Partitional Clustering*, CRAN.R-project.org, 2020.
81. Herrmann, L. Swarm-Organized Topographic Mapping. Doctoral dissertation, Philipps-Universität Marburg, Marburg, 2011.
82. Handl, J.; Knowles, J.; Kell, D.B. Computational cluster validation in post-genomic data analysis. *Bioinformatics* **2005**, *21*, 3201-3212.
83. Jain, A.K.; Dubes, R.C. *Algorithms for Clustering Data*; Prentice Hall College Div: Englewood Cliffs, New Jersey, USA, 1988.
84. Cormack, R.M. A review of classification. *Journal of the Royal Statistical Society. Series A (General)* **1971**, 321-367.
85. Steinley, D.; Brusco, M.J. Initializing k-means batch clustering: A critical evaluation of several techniques. *Journal of Classification* **2007**, *24*, 99-121.
86. Thrun, M.C.; Stier, Q. FCPS: Fundamental Clustering Problems Suite in R. *SoftwareX* **2020**, *under review*.
87. Conover, W.J. *Practical nonparametric statistics*; John Wiley & Sons: New York, USA, 1971.
88. Neumaier, A. Combinatorial configurations in terms of distances. *Dept. of Mathematics Memorandum* **1981**, 81-09.
89. Legendre, P.; Gallagher, E.D. Ecologically meaningful transformations for ordination of species data. *Oecologia* **2001**, *129*, 271-280.
90. Conde, A.; Domínguez, J. Scaling the chord and Hellinger distances in the range [0, 1]: An option to consider. *Journal of Asia-Pacific Biodiversity* **2018**, *11*, 161-166.
91. Eckert, A. *parallelDist: Parallel Distance Matrix Computation using Multiple Threads*, 0.2.4; CRAN, 2018.

 doi:10.20944/preprints202011.0451.v1

92. Thrun, M.C.; Ultsch, A. Effects of the payout system of income taxes to municipalities in Germany. In Proceedings of 12th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena, Cracow, Poland; pp. 533-542.