International Journal of Data Science and Analytics (2021) 12:249–271
https://doi.org/10.1007/s41060-021-00264-2

REGULAR PAPER

Conventional displays of structures in data compared with interactive projection-based clustering (IPBC)

Michael C. Thrun[1,3,4] · **Felix Pape**[2] · **Alfred Ultsch**[1]

Received: 20 October 2020 / Accepted: 10 May 2021 / Published online: 8 June 2021

Abstract
Clustering is an important task in knowledge discovery with the goal to identify structures of similar data points in a dataset. Here, the focus lies on methods that use a human-in-the-loop, i.e., incorporate user decisions into the clustering process through 2D and 3D displays of the structures in the data. Some of these interactive approaches fall into the category of visual analytics and emphasize the power of such displays to identify the structures interactively in various types of datasets or to verify the results of clustering algorithms. This work presents a new method called interactive projection-based clustering (IPBC). IPBC is an open-source and parameter-free method using a human-in-the-loop for an interactive 2.5D display and identification of structures in data based on the user's choice of a dimensionality reduction method. The IPBC approach is systematically compared with accessible visual analytics methods for the display and identification of cluster structures using twelve clustering benchmark datasets and one additional natural dataset. Qualitative comparison of 2D, 2.5D and 3D displays of structures and empirical evaluation of the identified cluster structures show that IPBC outperforms comparable methods. Additionally, IPBC assists in identifying structures previously unknown to domain experts in an application.

Keywords Cluster analysis · Interactive machine learning · Visual analytics · Structures · Human-in-the-loop

1 Introduction

The term "visual analytics" was introduced as "the science of analytical reasoning facilitated by interactive visual interfaces" [1]. However, it was pointed out that, based on current practice, a more specific definition would be that visual analytics combines automated analysis techniques with interactive displays of structures in data for an effective understanding, reasoning, and decision making based on large and complex datasets [2]. In systems such as the visual cluster rendering system (VISTA) [3], the objective is to display the dataset in such a way that it would be easy for a human to manually cluster data and verify existing clustering results visually.

Projections from high-dimensional data spaces into two or three dimensions are typical methods used in visual analytics [4]. If the output space is two dimensions, the result is a scatter plot. Scatter plots generated by a projection method are *the* state-of-the-art methods in cluster analysis to visualize data structures [5–7]. The goal of such a scatter plot is to display the distance or to a certain extent density-defined structures in the data. However, the Johnson–Lindenstrauss lemma [8, 9] states that the two-dimensional similarities in a scatter plot cannot coercively represent high-dimensional distances. Projections of several datasets with distance and density-based structures yield a misleading interpretation of the underlying structures [10, 11]. One particular problem of these systems is the special case in which the dataset does not possess any cluster structures at all. Systems for visual analytics usually involve a large number of parameters that are usually left to be fine-tuned by the human-in-the-loop. At the core of these problems lies the assumption that the distances in the scatterplot are directly proportional to the distances of the data points in a high-dimensional space.

✉ Michael C. Thrun
mthrun@informatik.uni-marburg.de

[1] Databionics Research Group, Philipps-University of Marburg, 35032 Marburg, Germany

[2] Philipps-University of Marburg, 35032 Marburg, Germany

[3] Deptartment of Hematology, Oncology and Immunology, Philipps-University of Marburg, Marburg, Germany

[4] IAP-GmbH Intelligent Analytics Projects, In den Birken 10A, 29352 Adelheidsdorf, Germany

In this work, we propose an interactive, parameter-free, open-source display of structures in data based on the generalized U-matrix visualization method which is based on a simplified emergent self-organizing map [12]. This method displays the high-dimensional distances between the points. The density properties of the data space can also be incorporated. The method leads to a landscape like representation "called topographic map" on top of a scatterplot. Emerging structures such as walls, ridges and separate valleys can then be used for either clustering or the assessment that there are no cluster structures in the data.

The interactive projection based clustering (IPBC) method proposed here can be used with any projection method for the display and identification of cluster structures in the data. This work shows that

- Comparable interactive in 2D and 3D displays are misleading for a large variety of structures in data
- IPBC outperforms comparable methods in task of interactive identification of structures in data
- IPBC is the only visual analytics approach that accounts for the Johnson–Lindenstrauss lemma through the topographic map

Prior benchmarking showed that automatic PBC is always able to find the correct cluster structure, while the performance of the best of the 32 clustering algorithms varies depending on the dataset [11]. In this work, IPBC is compared to publicly available visual analytic methods for clustering artificial and high-dimensional datasets from real-world experiments. This work is an extension of the manuscript initially presented at DSAA 2020 to four examples [13].

2 Related works

Many visual analytics systems are either developed for a specific commercial solution or are no longer used.. To compare IPBC to other methods, we restricted our comparisons to publicly available systems that are still in use. In interactive clustering approaches scatter-plots are used in interactive principal component analysis (iPCA) [14], Clustrophile 2 [15], Morpheus [16], and Clustervision [17]. iPCA is an interactive system for PCA-based visual analytics. However, this system seems to no longer be publicly available. Clustervision enables the human-in-the-loop to choose from a variety of clustering techniques and parameters and then ranks the clustering results utilizing five quality metrics, additionally showing the scatter plots of the projected data [17]. Unfortunately, the web interface was not working for new data. Clustrophile 2 displays structures in data using heatmap visualizations of discrete clusters with scatterplots for dimensionality reduction [15]. It also enables what-if analyses through direct manipulation of the dimensionality reduction scatterplots [18]. However, the software was not accessible.

In systems such as the visual cluster rendering system (VISTA) [3], the objective is to display the dataset in such a way that it would be easy for a human to manually cluster the data and verify existing clustering results visually. While VISTA is often intended to import an existing clustering to validate or modify the clustering, it is also able to produce clusterings by itself through its interactive display of structures in data.

It should be noted that linear affine mapping, which is used by VISTA to render a 2D plot, has some disadvantages. While "gaps" in the visualized point clouds represent real gaps in the data, clusters can overlap, and outliers can even produce fake clusters. The mapping technique used is called α-mapping. To mitigate the previously mentioned effects, the display of structures in data is made dynamic by enabling the human-in-the-loop to modify the projection plane interactively. This allows the continuous observation of the dataset from different perspectives.. This α-mapping aims to preserve the N-dimensional information of the dataset in 2D space using a N-parameter-adjustable display of structures in data, which maps to 2D star coordinates as introduced in [19]. The two essential properties of this mapping are its linearity and its adjustability. Linearity means that gaps in the display of the structures in the data always correspond to gaps within the data space, and the adjustability enables changing the weight of each dimension, which indicates how significant the given dimension is in the display of the structures in the data. By changing a weight continuously, the effect of its corresponding dimension on the cluster distribution can be observed. Since VISTA smoothly animates the weight changes for one or even multiple dimensions, this method can also provide insights on the clustering process.

The first of its interactive options is the manual subset selection: the user uses a freehand drawing tool to create an enclosed region on-screen to select the point in this area of the 2D display. First, the whole dataset is defined as one subset. The clusters are defined as subsets from then on. The next operation is to enable the merging and splitting of subsets which enables the precise redefinition of cluster borders or the subdivision of subsets. Another option the user has, is to import already existing domain knowledge about the dataset.

The ability to define hierarchical cluster structures can be used to zoom in on a subset and define a sub-layer within it. This approach is used to model cluster details at different levels, for example, when clusters can be found that are distinct from one another but are very similar to one another compared to other clusters. Since the manual definition of cluster hierarchies is not found in the other visual analytics

approaches, it should be noted that this can potentially be quite an interesting feature. However, it cannot be compared to other solutions.

Graphical clustering toolkit (gCLUTO) by [20] is also an approach to 3D assisted clustering. In gCLUTO, the display of structures in data is used after applying a nonvisual clustering algorithm to verify the clustering results. One of the most interesting parts of this system from a visual analytics standpoint is the so-called mountain visualization. The focus lies in understanding high-dimensional datasets. The mountain visualization in a 3D display aims to provide a low-dimensional graphical representation of the clusters, including their relationship to another, their internal similarity, size, and standard deviation. This terrain consists of a horizontal plane that rises in peaks in several locations. Each peak in the plane represents a cluster from the clustering results, and its location, height, volume, and color represent various characteristics. The most visually recognizable feature is the distance between peaks which represents the relative similarity of the clusters using the similarity measure selected for the clustering algorithm. This means that similar clusters are represented by nearby peaks, and clusters that are dissimilar are displayed as distant peaks. This display of structures in data is achieved by using multidimensional scaling on the cluster midpoints to find a mapping that minimizes the data's distortion. The height of a peak is proportional to the internal similarity of the corresponding cluster. This internal similarity is calculated from a similarity function set during clustering. The internal similarity is therefore the average result of the similarity function for all pairs of points within the cluster. The volume of a peak represents the number of objects within the cluster, and the color visualizes the standard deviation of the cluster's objects; red corresponds to a large standard deviation while blue corresponds to a small standard deviation. In summary, this plot is used to visualize the relative similarity of classically calculated clusters as well as their size, internal similarity, and internal deviation. With this display of the structures in the data, the human-in-the-loop should quickly gain insights into the clustering results, which in turn should help to improve the parameters of the clustering algorithm. gCLUTO itself is not a solution for clustering, but rather a method to find a valid clustering schemes by trying many possibilities. Here a good display of structures in data for clustering must be found by the human-in-the-loop, which is especially true for high-dimensional data, and is not a trivial task.

Using the categorization from [21] for visual analytics systems, the IPBC, U-matrix, and clustering methods [22] belong to the dimension reduction and clustering groups. In both fields, there are also other current approaches such as the efforts of [23] who cluster trajectory data with interactive self-organizing maps. Hossain et al. [24], on the other hand, tried to improve clustering by allowing the system to learn from a domain expert by incorporating user feedback to steer the result of their scatter/gather clustering algorithm. This algorithm starts with a basic k-means clustering, which can be refined in multiple scatter and gather steps. However, its performance is limited in the high-dimensional data because it relies on two- or three-dimensional data plots created for the user to steer the algorithm. The last two methods demonstrate current approaches in visual analytics. The first is the specialization of a specific kind of data [23, 25], which also works on the interactive clustering of movement data. Other areas in domain-specific visual analytics are data from bioinformatics or climate change research. The IPBC approach differs from these methods since it is not fine-tuned to a specific use case, and it works on any numerical dataset.

Another approach in visual analytics is the restriction of a specific number of dimensions like the scatter/gather approach by [24]. Algorithms such as this are pure clustering approaches restricted on two or three dimensions because their display of structures in data is based on the dimensionality of the input data. This makes structures with more than three-dimension hard to detect and makes it nearly impossible to understand for a domain expert not explicitly trained in these displays of structures in data.

2.1 Displaying information in 2D versus 3D

In line with the definitions of Kraus et al. [26], in this work, 2D displays of the structure in the data are defined as the static presentation of two dimensions of information on the screen, e.g., a scatter plot of data in a two-dimensional plane defined by Cartesian coordinates x and y. In the case of a 3D display of the structures in the data, the observer is looking on a screen at a projection of a visualization defined by three dimensions that can be rotated in an arbitrary direction, e.g., a scatter plot of data in a three-dimensional space defined by Cartesian coordinates x, y and z.

In general, the displays of information in 3D versus 2D have been debated many times in many scenarios because both display approaches for information visualization have advantages and disadvantages [25–28]. For example, a 3D interface for browsing an image folder not only does not offer any real benefits to user interaction, but also creates an unnecessary cognitive load on the users, which can lead to a lower performance and frustration [29]. In contrast, 3D shaded displays can add a significant amount of information to the visualization of high-dimensional structures in data [30]. 3D displays tend to be more comprehensible for the task of identifying clusters [26]. 3D displays are most useful when the viewer is free to change the viewpoint relative to the scene, to peer around occlusions or to view the scene from a different perspective [31]. However, 3D displays "are often vulnerable to artifacts caused by the rendering of

depth-related information, such as line-of-sight ambiguities, occlusion, and perspective distortion"[26].

Hence, it is easy create a bad 3D implementation [27]. Thus, Munzer strongly argues that most tasks involving abstract data do not benefit from 3D displays, and that such displays of structures in data must be justified. Tory and Munzer stated that simpler features like colored points support the best performance; 2D colored landscapes performed second best and may therefore be suitable for some applications [32] but 3D landscapes that redundantly encoded the data using color and height were slower than 2D landscapes using color alone, with no difference in accuracy [32]. Subsequent research by Tory et al indicates that dot-based displays may lead to better user performance in memory retrieval tasks [33]. In both publications contours were shown through the use of color bands without a deeper meaning for the colors. Hence, their findings cannot be generalized and contradict those of vision research.

For example, Marx et al confirmed a benefit of color over grayscale images for object detection [34]. Moreover, global color manipulation has an effect on the perception of natural scenes and the recognition of objects [35]. For instance, the performance is mostly similar to that of grayscale for stimuli in which the color hue is inverted [34]. Consequently, it can be argued that the performance of 3D compared with 2D depends on the chosen color band for the 3D landscape. Furthermore, attention is dominated by objects in natural sciences rather than lower features [36]. Therefore, vision research motivated the generation of topographic maps with an accurate color mapping [37] that looks similar to a natural scene using the CIELab color space [38] which is only very slightly outperformed by the UCS but has the advantage of a simple color distance calculation equation [39].

Another way of addressing the 3D versus 2D display discussion "is to apply more than a century of research into the psychology of space perception meaning that 3D is not a simple unified concept in perceptual terms. We can have degrees of three dimensionality, using different spatial "depth cues" and we can choose to increase or decrease the amount and type of 3D information" [40]. In the literature, the term 2.5 D is usually used to describe a pseudo-3D visualization; that is, it is not a true 3D space, but perhaps a rendering making the scene look 3D [31]. In this case, research indicates that subjects' performance deteriorates in both physical and virtual systems as their freedom to locate items in the third dimension increases [41]. However, the performance in virtual 2.5 D displays was slightly better than that in 2D displays (Fig. 3b in [41].) A reported user evaluation outlined the benefit of interactive visual data mining of text documents in a 2.5 D display [42]. For precise tasks, combination 2.5D displays were better than strict 2D or 3D displays [43]. Consequently, perceived issues with 3D (c.f. [44]) are solvable and successful, and specialized information visualizations can be built [27]. In our work, a 2.5D display is defined as the static top view of a 3D display (e.g. top view of the topographic map described below).

In summary, three visual analytic methods are compared. VISTA displays information in 2D. gCLUTO displays information in 3D and interactive projection-based clustering (IPBC) displays information in 2.5D but has the separate option to provide 3D displays after the clustering process is finished.

3 Interactive projection-based clustering (IPBC)

The IPBC method is, at its core, a combination of a nonlinear dimensionality reduction method with clustering aimed at preserving high-dimensional neighborhoods within its structure, which leads to no limitations on the dimensionality of the input data for clustering while retaining most of the important information about the structures of the dataset within its 2.5D and 3D. (Automatic) projection-based clustering (PBC) itself works in three steps. First, a nonlinear projection method projects the data into a two-dimensional plane. Second, the Delaunay graph [45] between projected points is computed. Each vertex of the graph is weighted with the high-dimensional distance between the two corresponding high-dimensional points. Third, the shortest paths between every two projected points, which are computed with the Dijkstra algorithm [46], are used in an automatic clustering process. The automatic clustering process requires the number of clusters and a Boolean parameter defining the structure type as the input (details in [47]). Both parameters can be derived from the displays of the structures in the data of the topographic map which is described below.

3.1 The display of the topographic map

The task of clustering is performed interactively through the use of the topographic map generated by the generalized unified distance matrix (U-matrix) of an arbitrary projection method. A topographic map (landscape) is constructed on top of a scatterplot generated by any type of projection method as follows: the wall of the Voronoi cells (VCs) in the scatterplot between the projections xp and yp of two high-dimensional data points x and y represents the borderline where the affiliation to x, i.e., its closest neighbor, changes to y [48]. At these borderlines, the distance between x and y, d(x,y), in the high-dimensional space is displayed as the height of the wall of the VCs. This structure, called

the abstract U-matrix [49], faithfully represents the high-dimensional distance structure of the data on top of the scatterplot. However, these walls are infinitely thin, and adjacent walls have very different heights. To practically visualize the high-dimensional distance structure on top of a scatterplot a simplified method of self-organized maps (SOM) and the U-matrix is used [50]: the scatterplot is sampled using a finite grid on which each grid point is an artificial neuron representing in its weight vector a point in the data space. This is regarded as a two-dimensional SOMs where the representations (BMUs) of the data points are fixed [12]. Using Kohonen's learning algorithm for SOMs the interpolating weight vectors can be adapted [12].

In the literature, the information stored in the weights of the neurons of the SOMs is visualized by the U-matrix which represents the folding of the high-dimensional space performed by an SOM projection [51]. At the end of this learning phase, the generalized U-matrix is constructed as the sum of the high-dimensional distances at each neuron [12, 52]. The U-matrix [51] or one of its variants [52–55] represents the distances between neurons as U-heights by using proportional intensities of gray values, color hues, shapes or sizes. For example, every neuron can correspond to a pixel [53]. The U-height corresponding to a gray value of each pixel is determined by the maximum unit distance from the neuron to its four neighbors (up, down, left, and right). The larger the distance is, the lighter the gray value is. In summary, a visualization of the U-matrix presents the high-dimensional structures of a dataset [56, 57] and can be generalized to be computed for every projection method [12]. After an extensive literature survey, it was argued that the best visualization of high-dimensional structures is to encode the U heights as a topographic map with hypsometric tints [37].

The topographic map can be interpreted as follows: the projected points and their mapping to high-dimensional data points predefine the display of the topographic map. The color scale of the topographic map is chosen to display various valleys, ridges, and basins: blue indicates small high-dimensional distances and high densities (sea level), green and brown indicate middle high-dimensional distances and densities (small hills) and white indicates large distances and small densities (snow and ice of tall mountains) [58]. The valleys and basins indicate clusters, and the watersheds of hills and mountains indicate the borderlines of clusters [58]. The color scale is combined with contour lines. The topographic map with hypsometric tints can be 3D printed such that through its haptic form, it is even more understandable by domain experts [58]. The topographic map looks like a 3D landscape and can be presented either as a top view in 2.5D or interactively rotated in 3D. The additional information encoded on top of the 2D display has been proven to be informative and useful [47, 59, 60].

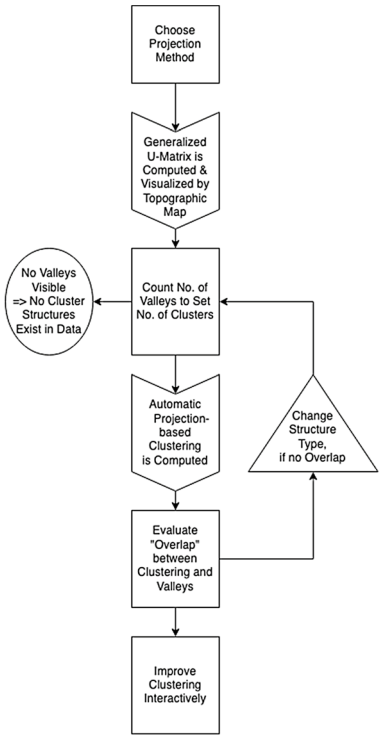

Fig. 1 High-level process flow of IPBC for which the interface is presented in Figs. 1 and 2. Clusters and valleys do not overlap if a cluster is either divided into separate valleys or several clusters lie within the same valley, c.f. [11]. The human-in-the-loop can modify clusters or outliers interactively after the (optional) automatic projection-based clustering (PBC). If no valleys are visible, the dataset does not possess any natural cluster structures

3.2 Interactive construction of display of structures in data

After the IPBC method is executed by the user, a Shiny interface [61], as shown in Fig. 1, opens with the first menu (1). The human-in-the-loop or user can perform the following actions: first, an arbitrary linear (e.g., PCA) or nonlinear projection method (e.g., NeRV [4]) can be selected to project high-dimensional data onto a two-dimensional plane (2). Some projection methods have adjustable parameters (e.g., NeRV); some have none (e.g., Pswarm, and MDS), which can be set in (3). In addition to the parameters of the projection method, no other parameters must be set by the user. In the next step, the generalized U-matrix [50] is computed (4) and can be visualized in 2.5D by the top view of the topographic map [58]: see (10) in Fig. 1. The structure of the topographic map is toroidal; i.e., the borders of the map are cyclically connected, which allows the problem of projected points on borders and, consequently, boundary effects to be avoided. Display options can be set with (5-8): (5) changes the size of the points, (6) allows the user to extend the topographic map by x neurons in each direction

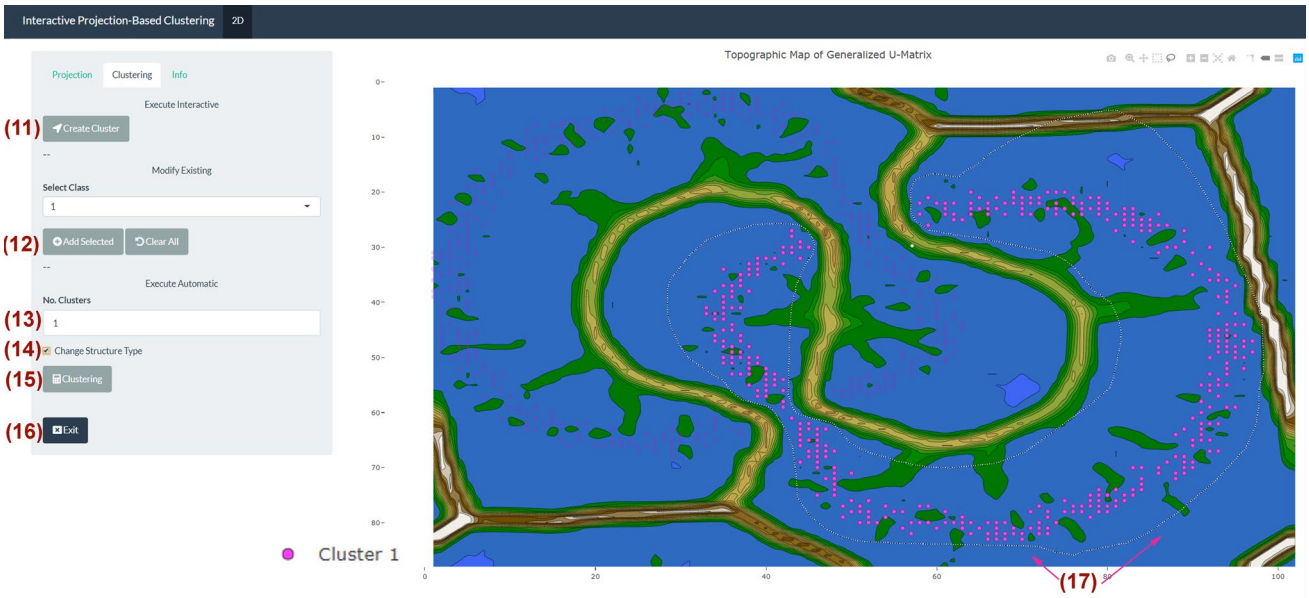

Fig. 2 Screenshot of the interface of the "Clustering" menu of the IPBC method after holding the left mouse button and framing a valley but before clicking on "Create Cluster" in (11). The human-in-the-loop can frame points with the mouse (8) and by clicking "Add Cluster" (9), a new cluster is separated from a given cluster

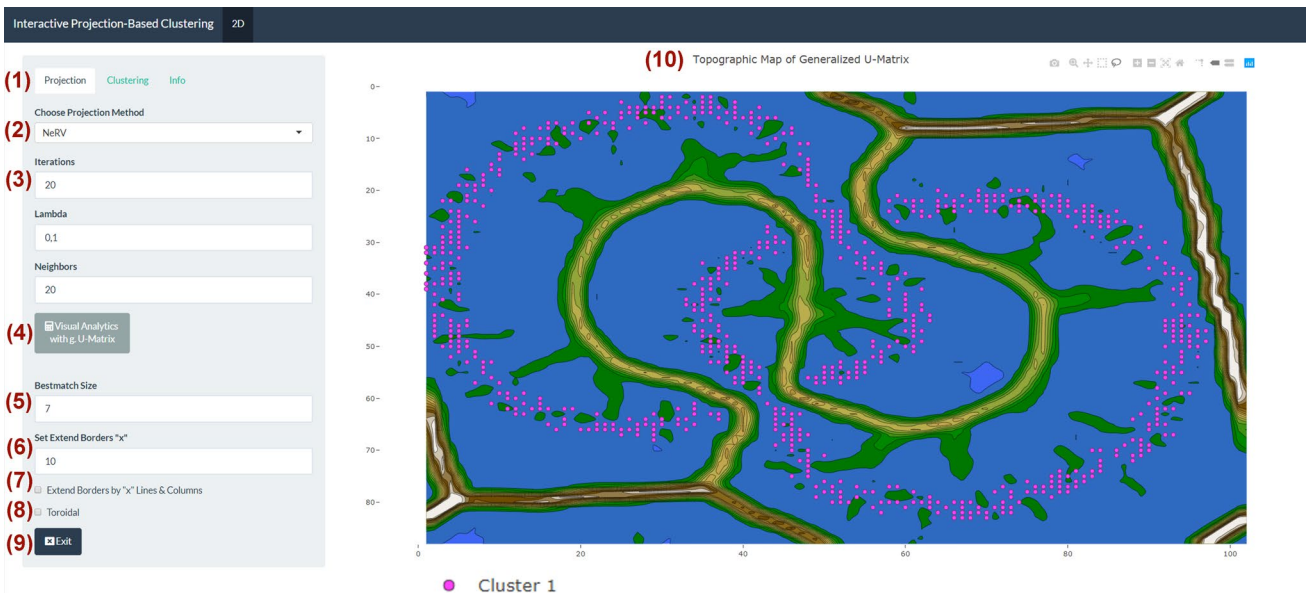

Fig. 3 Screenshot of the interface of the "Projection" menu of the IPBC method after loading the Chainlink dataset in Listing 1 and clicking on the button in (4) resulting in a topographic map shown in (1) of the NeRV projection selected previously in (2). The human-in-the-loop can select the projection method (2) and by clicking (4) a new topographic map is visualized

(for the def. please see [12]) for which then the extension of the topographic map is activated or deactivated with (7), (8) provides a four-tiled view of the toroidal topographic map which means that each point in the 3D landscape as well as the landscape itself is presented four times. Then, the interactive clustering procedure can be started. Selecting "Clustering" in the top menu (1) guides the user to the interactive clustering of Fig. 2. Here, interactive tools from plotly [62] can be used (17) to create clusters manually (11) and to modify them (12). For every projection method, automatic projection-based clustering is activated by clicking button (15) after setting the number of clusters as the number of visible valleys with (13), and the cluster structure type [11] can be changed optionally with (14) in Fig. 2.

Appropriate clustering is found if mountains do not divide a cluster but separate clusters, and all the clusters lie in valleys [11]. Then, the IPBC method can be closed with (16). The high-level process-flow of the IPBC method is presented in Fig. 3. It includes the automatic PBC, which is optional because clusters can be marked exclusively in the interactive approach.

R libraries Shiny by [61] and plotly [62] are used in order to achieve interactive clustering in the 2.5D display. Shiny enables the building of interactive applications for data analysis. These features will be used to display information about the data or to add or remove clusters interactively. Plotly is a visualization tool that is used to produce the top view of the topographic map, lets the user interact with it, and send the user interactions to Shiny. This combination gives a high degree of interactivity with the the plot itself which is a property an method in visual analytics should have. Presently the identification of clusters is based on the top view of the 3D landscape in a 2.5D display. The 3D display can be visualized on the screen and interactively rotated if the function "plotTopographicMap" is called (see below) which is based on the R package "rgl" available on CRAN[63]. The following projection methods were used: t-SNE [64], NerV [4], Pswarm [65], and uniform manifold approximation projection [66]. The interactive method uses the 2.5D display of the topographic map, and can be accessed via the R package 'ProjectionBasedClustering' [67] on CRAN.

After closing the tool with the 2.5D display of the IPBC method, either 2.5D, or 3D high-resolution figures and STL files for 3D printing can be prepared in the last step by calling the plotTopograpahicMap for a 3D display or the TopviewTopographicMap function for a 2.5D display. Optionally, an island can be cut out interactively by encircling the most prominent mountain ranges. However this step is not necessary for the interactive clustering. The R code for these steps is presented in "Appendix A," Listing 1.

4 Evaluation

The evaluation section is divided into four sections. After the description of the datasets, the first two sections qualitatively evaluate the different displays of structures in data from artificial and real-world examples. The topographic maps are presented in the form of a 3D display in comparison with the 3D display from gCLUTO and the 2D display from Vista.

Table 1 Summary of the description and challenges of the 10 artificial and two natural datasets of the FCPS for cluster analysis because the FCPS offers a variety of real-world challenges [68].

Dataset Name	Number of Points	Number of Dimensions	Short Description of the Shapes	Challenge
Hepta	212	3	Six balls, each centered at each one of the six corners of a large octahedron with the 7th ball having a higher density at its center	Nonoverlapping convex hulls with varying intracluster distances
Chainlink	1000	3	Two intertwined chains	Linear nonseparable entanglements
Atom	800	3	Core enclosed by a hull	Completely overlapping convex hull
EngyTime	4096	2	Two Gaussian mixtures with different variances	Overlapping clusters separable only by density
GolfBall	4002	3	Empty sphere	No distance-based cluster structures
Lsun3D	404	3	One full sphere, two bricks perpendicular to each other, and outliers	Varying geometric shapes with noise defined by one group of outliers
Target	770	2	Circular disk enclosed by a circle with outliers in four corners	Overlapping convex hulls combined with noise defined by four groups of outliers
Tetra	400	3	Four close full spheres at the four corners of a tetrahedron	Narrow distances between the clusters
TwoDiamonds	800	2	Two rhombuses with one touching corner	Identification of the weak link in chain-like connected clusters
WingNut	1016	2	Two rectangles, each having a density that increases towards one corner in direction of the other rectangle	Short intercluster distances combined with vast intracluster distances
Tetragonula	236	13	Distance matrix easy associable with geographic origins of cases	Smooth transition between clusters and outliers, and the clusters have to be coherent with the geographic origins
Leukemia	554	12,692	Distance matrix with patient diagnosis for cases	Reproducing highly unbalanced class sizes

"FCPS is a collection of intentionally low-dimensional artificial datasets of user-defined sample sizes and an unique class labeling generated under the hypothesis that humans are most often able to group objects in two- or three-dimensional plots by eye. Additionally, two high-dimensional real-world datasets with a clear cluster structure are provided" [68]. These datasets are available in the R package "FCPS" on CRAN

The 2.5D displays of the same topographic maps of the benchmark datasets from Table 1 are provided in "Appendix C", Fig. 20. The third section presents the empirical evaluation of the identification of the displayed structures in the data by comparing the resulting interactively performed clusterings with previously given classifications and the fourth section presents an application.

4.1 Comparison approach

IPBC, VISTA and gCLUTO will be compared to each other using the SCADI dataset [69] and twelve benchmarking datasets called the "Fundamental Clustering Problems Suite" (FCPS) [68]. In this case, automatic projection-based clustering will not be used. The twelve datasets of the FCPS are extensively described in [68], and an overview is provided in Table 1. There are ten low-dimensional (3D) datasets called Hepta, Chainlink, Atom, EngyTime, GolfBall, Lsun3D, Target, Tetra, TwoDiamonds and WingNut. Additionally the FCPS provides two high-dimensional datasets called Tetragonula and Leukemia. The Tetragonula bee dataset from [70] consists of genetic data of 236 worker bees from 9 different species, each from an individual hive (see details in [68]). The Tetragonula dataset has the challenge of smooth transitions between clusters and outliers [68]. The Leukemia dataset is a microarray dataset that measures the gene expressions of 554 subjects who are either healthy or have one of the following illnesses: acute promyelocytic leukemia, chronic lymphocytic leukemia, or acute myeloid leukemia.

Additionally, the SCADI dataset was taken from [69]. It contains 206 attributes of 70 children with physical or motor disabilities. The classification refers to 7 different levels of self-care problems. As an application, the Boston Housing Dataset published in [71] was taken from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/machine-learning-databases/housing/). It consists of 506 data points of 14 features and is the only dataset investigated here without a prior classification scheme. Harrison et al showed that as the air pollution stored in the feature "NOX" increase, the home values of the owner-occupied homes storied in "MEDV" decrease [71].

Using the 2D, 2.5D and 3D displays, the identified structures in data will be compared with the prior classification. The Rand index R [72] corrected for chance by the method of Hubert and Arabie [73] with the following equation (Eq. 4, p.198 [73]) is used:

$$\mathrm{ARI} = \frac{R - \mathrm{ExpectedIndex}}{\mathrm{MaximumIndex} - \mathrm{ExpectedIndex}}$$

Given a contingency table, the mathematical details can be found in (Eq. 5, p.198 [73] and are implemented by the R package "phyclust" available on CRAN [74].

Additionally, the displays of structures in data leading to the clustering or verifying the clustering scheme will be shown. No default parameters other than the number of clusters for IPBC were changed. The number of clusters was estimated in the topographic map by counting the

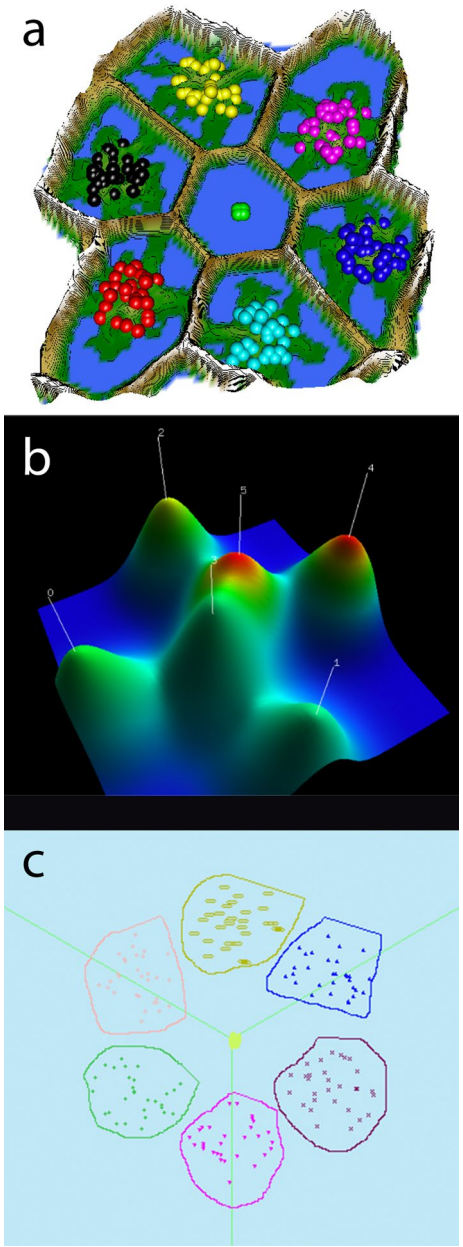

Fig. 4 Displays of the structures in the artificial Hepta dataset. **a** Topographic maps for which cluster analysis is performed interactively by IPBC based on the NeRV projection by using the top view. **b** Mountain visualization using gCLUTO graph-clustering with 6 clusters. **c** VISTA clustering

number of valleys. The parameters modified in gCLUTO are the similarity measure, the number of clusters, and the clustering algorithm. Because gCLUTO requires the number of clusters in advance, a different number of clusters will be tested, and the best number will be presented. For VISTA, the number of parameters increases with the number of dimensions, which leads to an impractically large number of parameters to print here.

4.2 Artificial datasets

For the first dataset called Hepta, the topographic map in Fig. 4a clearly shows seven distinct groups of data in different valleys, which are surrounded by mountains. In addition, behind each mountain, a new valley begins. This figure shows that there should be 7 clusters with low internal dissimilarities that are distinctively separate from one another. Using the matrix- and mountain-visualization and experimenting with different numbers of clusters and clustering approaches, six clusters seem to give the best display of structures in the data in Fig. 4b. Reducing the number of clusters shows that at least one cluster is massive, and the points within it have a high standard deviation. Increasing the number of clusters seems to only subdivide existing clusters into less clearly defined clusters or produces new exiguous clusters. The mountain visualizations show that there is, in every case, high internal deviation within the clusters, and they do not seem to be separated from one another. However, no such clustering could be found with gCLUTO. Graph clustering with six clusters gave the best combination of the matrix- and mountain-visualization, so it is used and presented in Fig. 4b. The seven cluster result is the default after importing the dataset into VISTA. The cluster assignment is visible in Fig. 4c.

The topographic map of the Chainlink dataset in Fig. 5a clearly shows two clusters in two separate valleys. One point which marked with an arrow is assigned to an incorrect cluster by the human-in-the-loop (user) as it lies in the incorrect valley. gCLUTO mountain- and matrix-visualization are both not suited to visualize linear nonseparable entanglements (Fig. 5b). Using mountain visualizations, the number of clusters is hard to guess. If roughly equally sized clusters were to be expected, which yields a low internal dissimilarity and produces visible groups in the matrix-visualization no less than 6 clusters should be used. With only minimal manipulation in VISTA (Fig. 5c), the user can see and understand that the Chainlink dataset consists of two interlocking rings, but the cluster assignment tools do not support selecting all the points of a cluster without selecting some points from the other. The workaround used here is to select 3 clusters: selecting the first half of a ring, rotating the projection so that the intersection lies in the already clustered part, and selecting the rest of the first ring and then the second ring as separate clusters. Afterward, these clusters could be merged manually into two clusters again; however, this step cannot be completed within VISTA. It should also be pointed out that some rare points can be found that are not labeled according to their surrounding group.

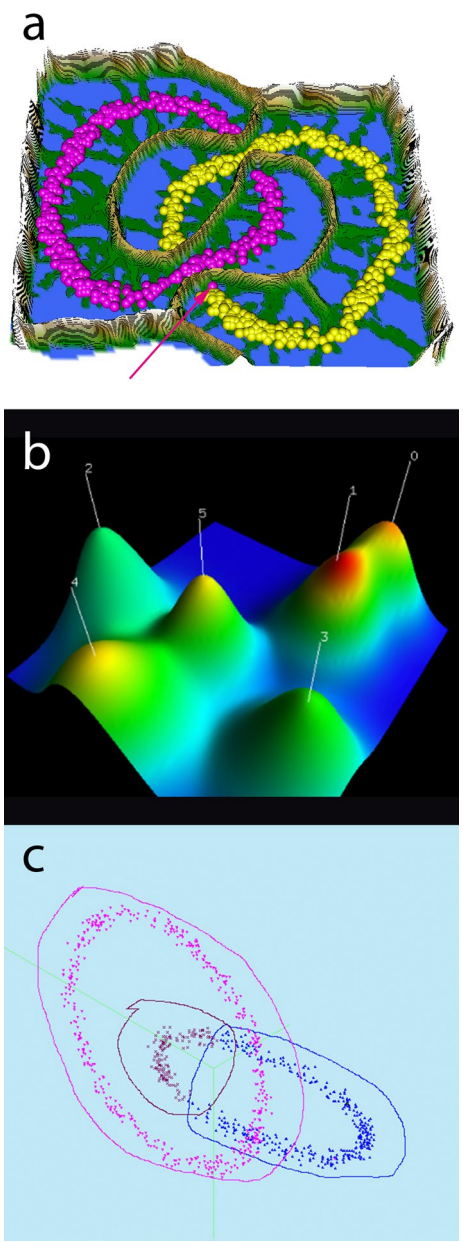

Fig. 5 Displays of the structures in the artificial Chainlink dataset. **a** Topographic maps for which cluster analysis is performed interactively by IPBC based on the NeRV projection by using the top view. The arrow points to the clearly misclassified point by the human-in-the-loop (user) because it lies in the incorrect valley. **b** Mountain visualization using gCLUTO shows 6 clusters for which then "repeated bisection" clustering is applied. **c** VISTA clustering

In the case of Lsun3D in Fig. 6 all three clusters and the outliers are distinctively separated in the topographic map

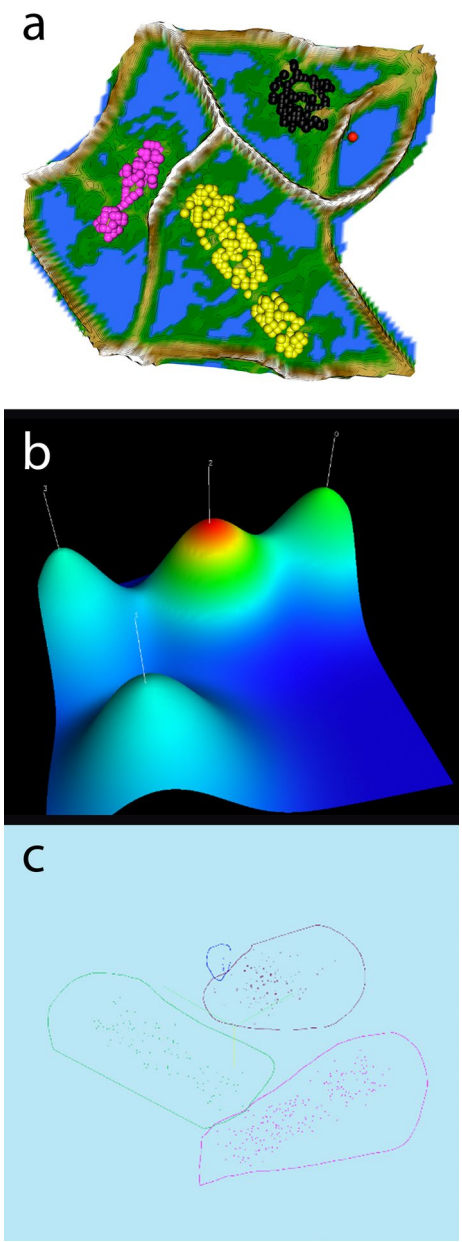

Fig. 6 Displays of structures in the artificial Lsun3D dataset. **a** Topographic maps for which cluster analysis is performed interactively by IPBC based on the t-SNE projection by using the top view. **b** Mountain visualization using gCLUTO with agglomerative clustering. **c** VISTA clustering

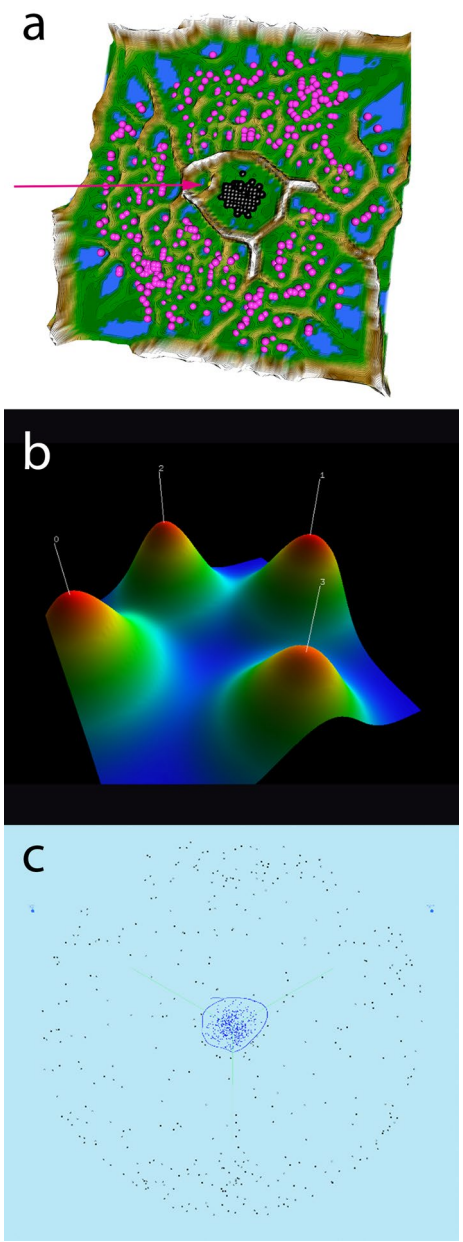

Fig. 7 Displays of structures in the artificial Atom dataset. **a** Topographic maps for which cluster analysis is performed interactively by IPBC based on the NeRV projection by using the top view. One outlier lies in a volcano which is marked by a red arrow. **b** Mountain visualization using gCLUTO for which then "repeated bisection" clustering is applied. **c** VISTA clustering

(a), gCLUTO with bisection clustering (b) and VISTA (c). It seems that varying geometric shapes with noise defined by outliers does not pose a problem for these methods. This finding is in contrast to conventional clustering algorithms such as spectral clustering of subspace clustering which fail to reproduce such cluster structures [47].

The displays of the structures in the data of the artificial datasets Atom (Fig. 7a), GolfBall (Fig. 8a), and Target (Fig. 9a) all have their points distributed within a sphere around the origin. All were clustered with repeated bisection in gCLUTO and seemed to produce similar mountain plots (Figs. 7b, 8b, 9b, respectively), which seemingly correctly indicates the similarities between these datasets, yet the red color indicates that the identified clusters seem to have a high standard deviation. Additional, for all these datasets, the mountain plots for 4 or 5 clusters look better. However, GolfBall has no existing cluster structure and Atom has

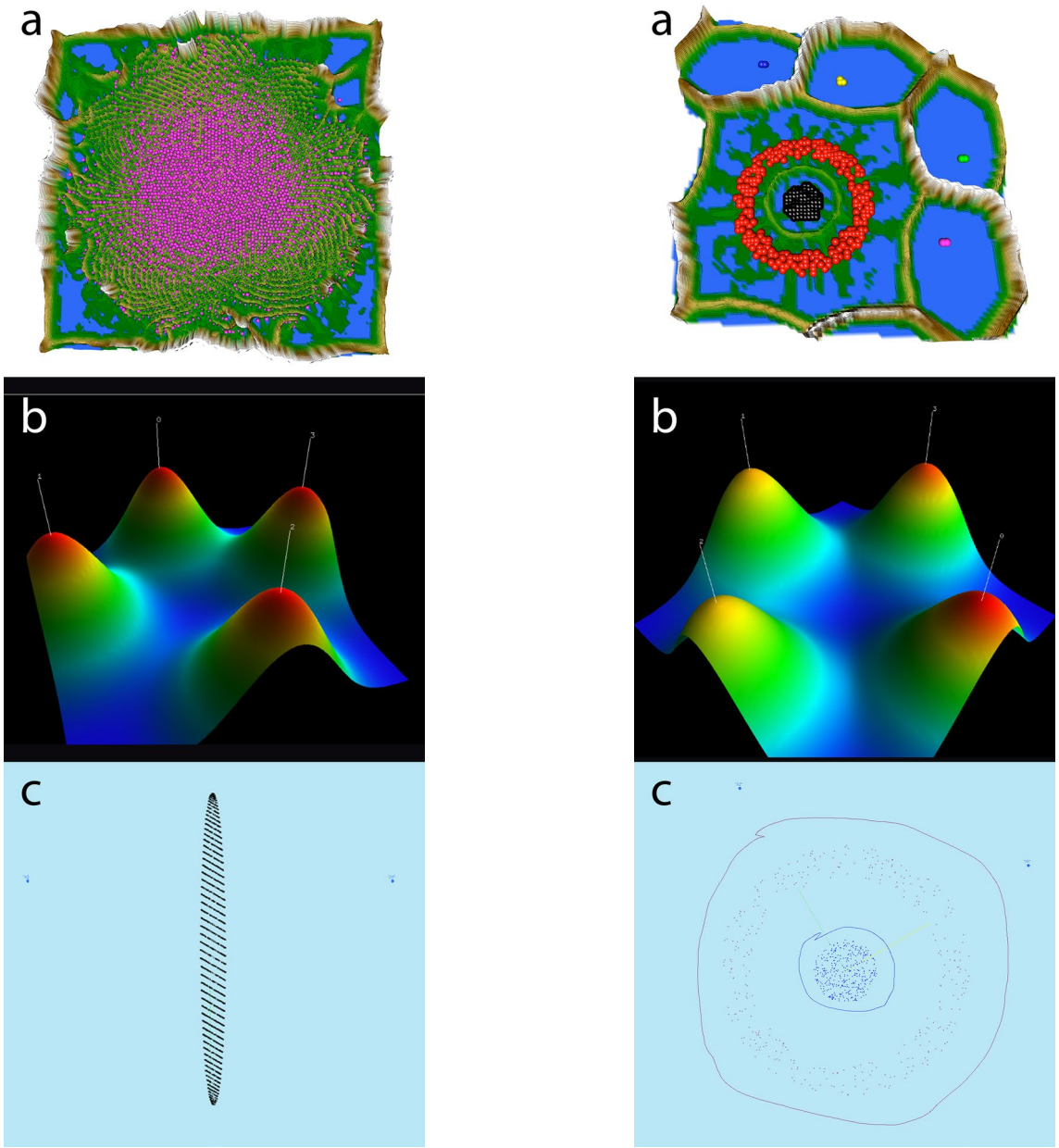

Fig. 8 Displays of structures in the artificial GolfBall dataset. **a** Topographic maps for which cluster analysis is performed interactively by IPBC based on the NeRV projection by using the top view. **b** Mountain visualization using gCLUTO for which then "repeated bisection" clustering is applied. **c** VISTA clustering

Fig. 9 Displays of structures in the artificial Target dataset. **a** Topographic maps for which cluster analysis is performed interactively by IPBC based on the NeRV projection by using the top view. **b** Mountain visualization using gCLUTO for which then "repeated bisection" clustering is applied. **c** VISTA clustering

only two clusters. In contrast, the same datasets in VISTA (Figs. 7c, 8c, 9c, respectively), show three different results. While these three datasets seem to be easily understood visually in VISTA, Atom becomes hard to cluster in Fig. 7c. This phenomenon is due to the clusters having overlapping convex hulls. The workaround used for the linear nonseparable entanglements of Chainlink does not work in Vista here, as points from outside the central cluster always overlap with the central cluster. In the case of GolfBall in Fig. 8c, one can find discrete rings within the dataset on which the points reside, which would suggest a large number of small clusters. In Fig. 9c, Target, due to being a two-dimensional dataset, looks similar to a scatterplot, and the clusters and can simply be visually separated. The topographic map of Atom in Fig. 7a shows one outlier incorrectly, in addition to this error, it is able to separate the two clusters. The outlier, which is marked with a red arrow, lies in a volcano and is slightly concealed by a hill and serves as an example that

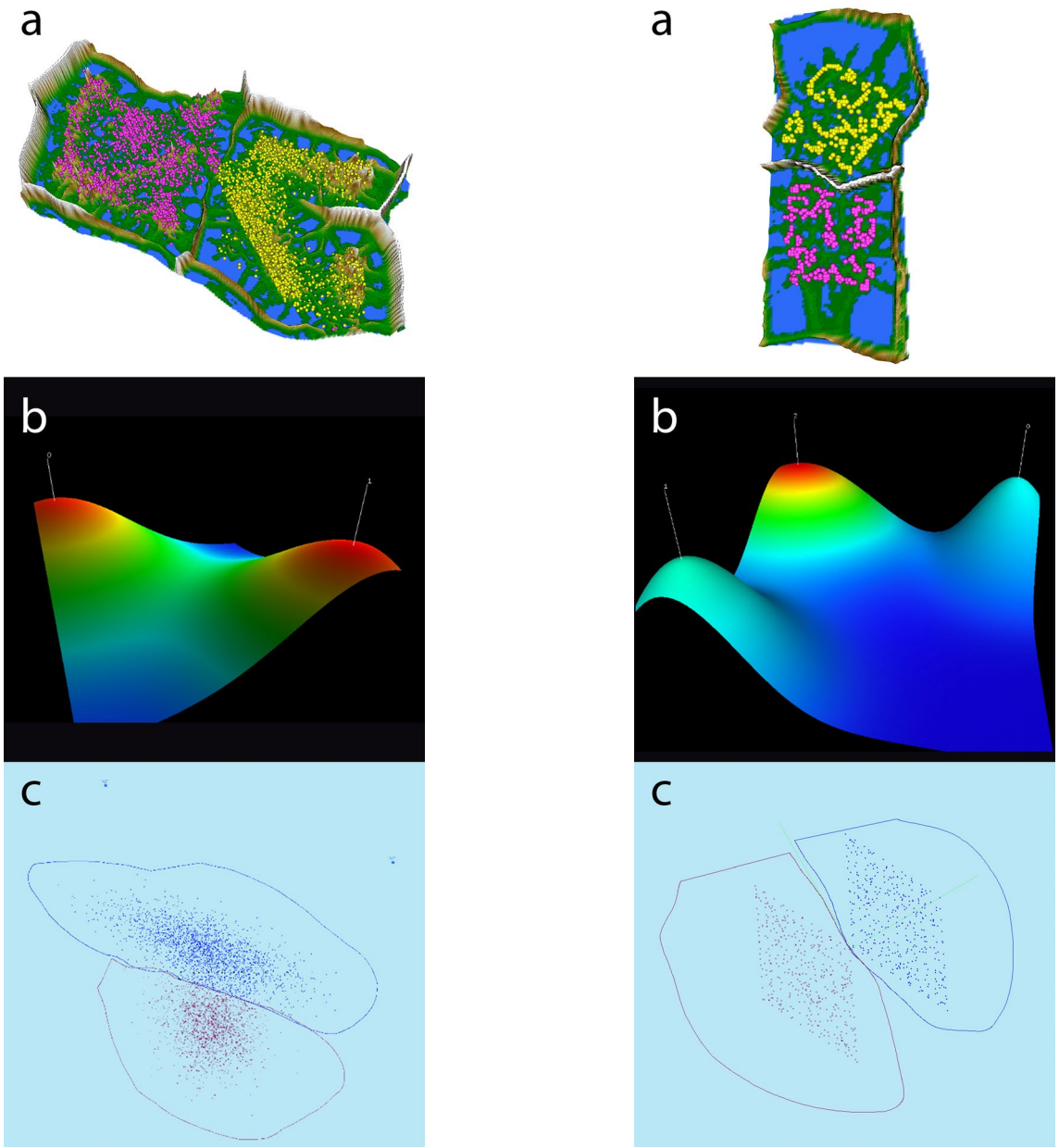

Fig. 10 Displays of structures in the artificial EngyTime dataset. **a** Topographic maps for which luster analysis is performed interactively by IPBC based on the Pswarm projection by using the top view. **b** Mountain visualization using gCLUTO for which then "repeated bisection" clustering is applied. **c** VISTA clustering

Fig. 11 Displays of structures in the artificial TwoDiamonds dataset. **a** Topographic maps for which cluster analysis is performed interactively by IPBC based on the uniform manifold approximation projection by using the top view. **b** Mountain visualization using gCLUTO for which then "repeated bisection" clustering is applied.. **c** VISTA clustering

occlusion can hide information [44]. The 3D view of the topographic map must be interactively moved to obtain a full overview of the structures in the data. Contrary to Fig. 7a of the topographic map of Atom, Fig. 8a of the topographic map of GolfBall shows no cluster structures at all because every point lies in its own valley. In Fig. 9a the topographic map of Target clearly distinguishes the two clusters from the outliers but the separation of the two clusters is not by a high mountain range.

Similar to the Target dataset, EngyTime, TwoDiamonds and WingNut, are 2-dimensional datasets and therefore can be seen as scatterplots in VISTA, which leads to clear results (Figs. 10c, 11c, 12c, respectively). gCLUTO's mountain visualization uses repeated bisection and suggests the same number of clusters in these

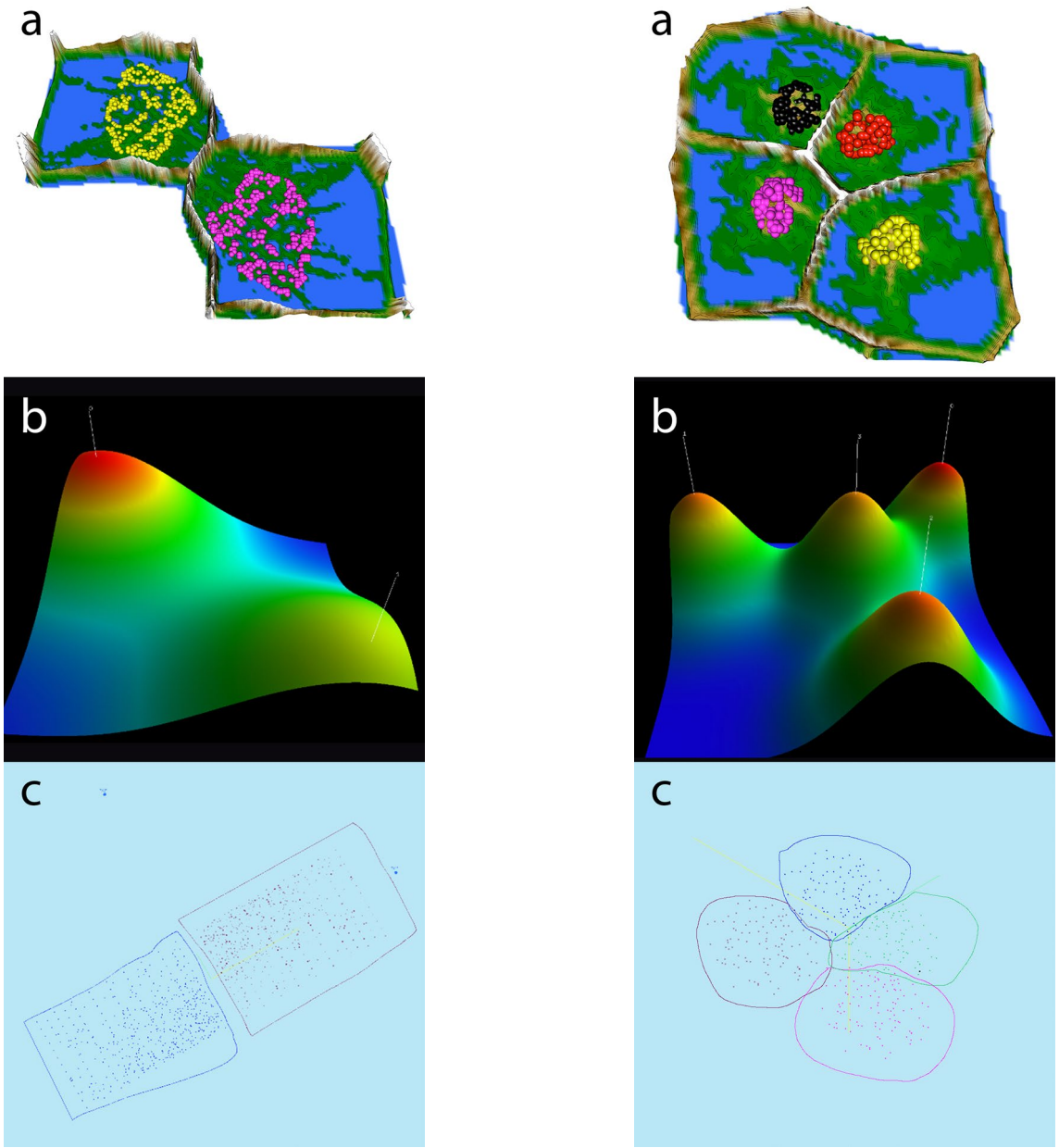

Fig. 12 Displays of structures in the artificial WingNut dataset. **a** Topographic maps for which cluster analysis is performed interactively by IPBC based on the uniform manifold approximation projection by using the top view. **b** Mountain visualization using gCLUTO for which then "repeated bisection" clustering is applied. c) VISTA clustering

Fig. 13 Displays of structures in the artificial Tetra dataset. **a** Topographic maps for which cluster analysis is performed interactively by IPBC based on the uniform manifold approximation projection by using the top view. **b** Mountain visualization using gCLUTO with agglomerative clustering. **c** VISTA clustering

cases (Figs. 10B, 11b, 12b, respectively), except for Two-Diamonds in Fig. 12b, for which a three cluster solution with one cluster with high deviation seems to be the most likely. In Figs. 11a and 12a the topographic maps and Two Diamonds and WingNut, respectively, as well as those in Fig. 10a (right) for EngyTime outline the given two clusters.

For the final artificial dataset, Tetra in Fig. 13c , VISTA provides a great interactive display of structures in data that can only be partially utilized to create proper clustering. Similar to Chainlink and Atom (Fig. 5c and 7c, respectively), here, we can see the clusters during animation, but they become not clearly separable in a

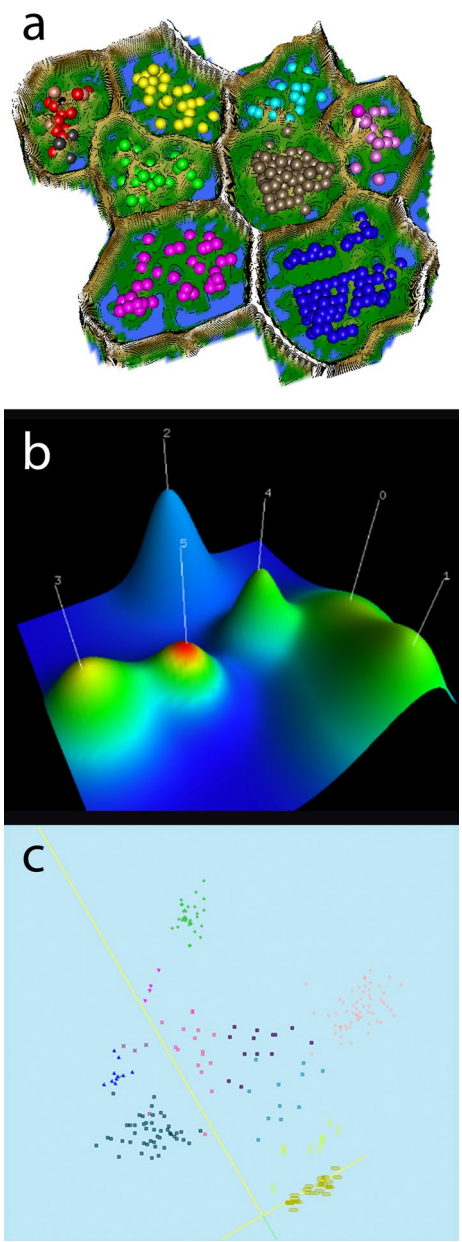

Fig. 14 Displays of structures in the natural Tetragonula dataset. **a** Topographic maps for which cluster analysis is performed interactively by IPBC based on the Pswarm projection by using the top view. **b** Mountain visualization using the agglomerative clustering in gCLUTO with 6 clusters. **c** VISTA clustering

still image. gCLUTO on the other hand, quickly provides a clear display of the structures in the data with 4 clusters of similar sizes and densities which would be expected in this case in Fig. 13b. The topographic map of Tetra in Fig. 13a shows that the four clusters lie in four distinct valleys separated by mountain ranges.

4.3 Real-world examples

The topographic map of the Tetragonula datasets shown in Fig. 14a has eight valleys and some outliers. The marking of the outliers is not easy. For eight or more clusters, the mountain visualization in Fig. 14b shows clusters close to one another and sometimes also exiguous clusters of just a few points. This persists in all available cluster methods. Therefore, using gCLUTO, one would assume six or seven clusters depending on the clustering method. Agglomerative gCLUTO clustering would suggest six clusters, while gCLUTO's "repeated bisection" clustering function produces a better display of the structures in the data with seven clusters. Both produce useful and quite similar mountain visualizations (not shown), but due to Occam's razor, the six-cluster solution will be used (Fig. 14b). Since the Tetragonula dataset has some outliers, it is a question whether this or the number of clusters was known previously. Then, solutions using eight or nine clusters would still produce good or at an least expected display of the structures in the data. In Fig. 14c, for VISTA many dimensions only seem to rotate points around one another. The most substantial effects of alpha changes can be seen in the first dimension and decrease with the number of dimensions. This finding is not surprising since MDS created the points from a distance matrix. It does not appear that meaningful clusters could be found in this display of structures in data. Another approach for the Tetragonula dataset is to reduce the number of dimensions. As shown in Fig. 14c, only the first four dimensions are loaded into VISTA. Here changing the α-mapping shows groups of data quickly. The clusters are estimated by selecting a group of points and then further modifying the α-mapping to see if it splits into other clusters if the alpha values are changed. A group that stays together even when the projection is changed is defined as a cluster. It should be noted that this can only be done since it is known that this dataset was produced with MDS from a distance matrix.

With the Leukemia dataset VISTA becomes unusable (see Fig. 15c). The high number of dimensions leads to more UI elements of the tool than what can be distinguished. The process collapses to pure guessing of the alphas for every dimension without containing any properly displayed structures in the data at all. With gCLUTO, the mountain plot using repeated bisection suggests 6 clusters with clearly separated centroids and a mostly low deviation in Fig. 15b. Only the topographic map of the leukaemia dataset in Fig. 15a shows 6 valleys more clearly, some of which clearly show outliers with just individual points within them.

The topographic map the SCADI dataset in Fig. 16a shows data inhomogeneity; several points lie alone in

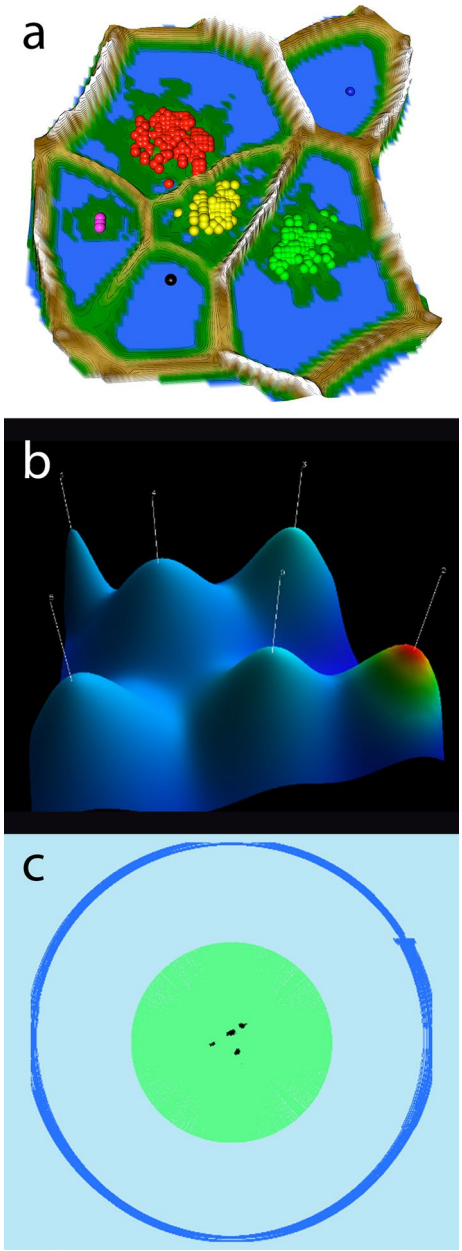

Fig. 15 Displays of structures in the high-dimensional Leukaemia dataset. **a** Topographic maps for which cluster analysis is performed interactively by IPBC based on the NeRV projection by using the top view. **b** Mountain visualization using gCLUTO with agglomerative clustering. **c** VISTA clustering

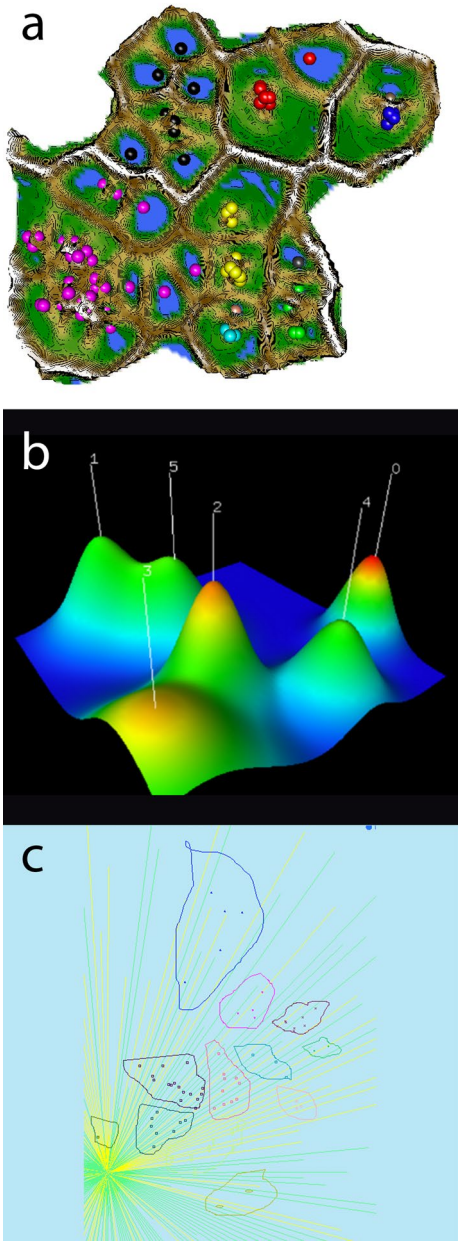

Fig. 16 Displays of structures in the natural SCADI dataset. **a** Topographic maps for which cluster analysis is performed interactively by IPBC based on the NeRV projection by using the top view. **b** Mountain visualization using the agglomerative clustering in gCLUTO. **c** VISTA clustering

a valley. In Fig. 16b, using agglomerative clustering in gCLUTO gives clusters with the lowest standard deviation, which are also distinct from each other according to the mountain plots. The likely number of clusters seems to be within the range of three to six. The cluster with ID 0 has the highest standard deviation but is never split while for less than seven clusters, but with seven or more clusters, the clusters begin to be less clearly defined in the mountain plot. Therefore, the six clusters with repeated bisection clustering from CLUTO are used. In Fig. 16c, VISTA requires many user inputs that do not impact the display of the structures in the data. Groups of data points are visible when animating changes in a specific dimension, but when viewing other dimensions, other groups may appear. It is not possible for the user to distinguish which groups are more relevant or occur more often due to

Table 2 Comparison of the clusterings using the adjusted Rand index on the twelve benchmark datasets for various clustering challenges

Name of Dataset/Method	gCLUTO	VISTA	IPBC
Chainlink: nonoverlapping convex hulls with varying intracluster distances	0.2083	0.992	0.996
Hepta: linear nonseparable entanglements	0.763	0.990	1
Atom: completely overlapping convex hull	0.003	0.946	0.998
Engytime: overlapping clusters separable only by density	0.0349	0.815	0. 856
Golfball: no distance-based cluster structures	0	0	1
lsun3d: varying geometric shapes with noise defined by one group of outliers	0.472	1	1
Target: overlapping convex hulls combined with noise defined by four groups of outliers	0.001	0.999	1
Tetra: narrow distances between the clusters	1	0.777	1
Twodiamonds: identification of the weak link in chain-like connected clusters	0.210	1	0.995
Wingnut: short intercluster distances combined with vast intracluster distances	0.885	1	1
Tetragonula: smooth transition between clusters and outliers, and the clusters have to be coherent with the geographic origins	0.979	0.8365	0.979
Leukemia: reproducing highly unbalanced class sizes	0.643	No result	0.998

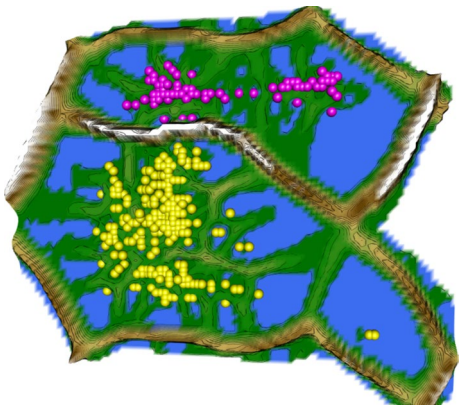

Fig. 17 Topographic map of the NeRV projection of the natural Boston Housing Datasets shows two clusters within IPBC. Cluster one consists of owner-occupied homes in yellow and cluster two in magenta

the number of parameters the user can adjust is too high. For VISTA, this is a direct consequence out of the high number of dimensions. The display of the structures in the SCADI dataset does not have clearly defined clusters, but due to the animations, while changing the α-mapping, the user can still try to find groups, that behave similarly to one another and are close to one another. Additionally, it is difficult to repeat the clustering process to achieve the same results.

4.4 Cluster evaluation

Table 2 shows the adjusted Rand index [72] between the given correct clustering of a dataset and a result for each approach. The results are the ones for which the displays of the structures in the data have been shown. The best values are marked in bold. For the Leukemia dataset, clustering with VISTA was impossible because the number of UI elements for the multiple dimensions overlaid the data points. Automatically switching through all the dimensions would take a long time without generating any usable display of the structures in the data due to UI overlap. Additionally, the adjusted Rand indices for the SCADI natural dataset are 0.7206 for gCLUTO, 0.1392 for VISTA, and 0.787 for IPBC.

4.5 Application: Boston Housing

The Boston Housing Dataset was clustered with IPBC and the results are shown in Fig. 17. Welch's two sample *t*-test for the MEDV variable indicates that cluster one has a greater MEDV (median value of homer owner-occupied houses) values than cluster two ($t = 10.181$, df $= 240.41$, p-value $< 2.2e{-}16$)). Additionally, The Welch two sample t-test indicates that cluster one has greater NOX (air pollution) values than cluster two (($t = -21.342$, df $= 401$, p-value $< 2.2e{-}16$)). Adapting the XAI procedure described in [75] to this dataset, the clustering can be explained by one rule stating that TAX values below 568 define cluster one and cluster TAX values above 558 cluster two with an accuracy of 100%.

4.6 Computation time

Evaluation of the computation time of the interactive projection-based clustering is based on the computations of four components: projection method, generalized U-matrix, topographic map and clustering. Obviously, the computation time will vary depending on the projection method, hence, we will use the method of multidimensional scaling (MDS) (for the computation time of other projection methods, the reader is referred to the corresponding publications). The four components are evaluated in combination disregarding the time

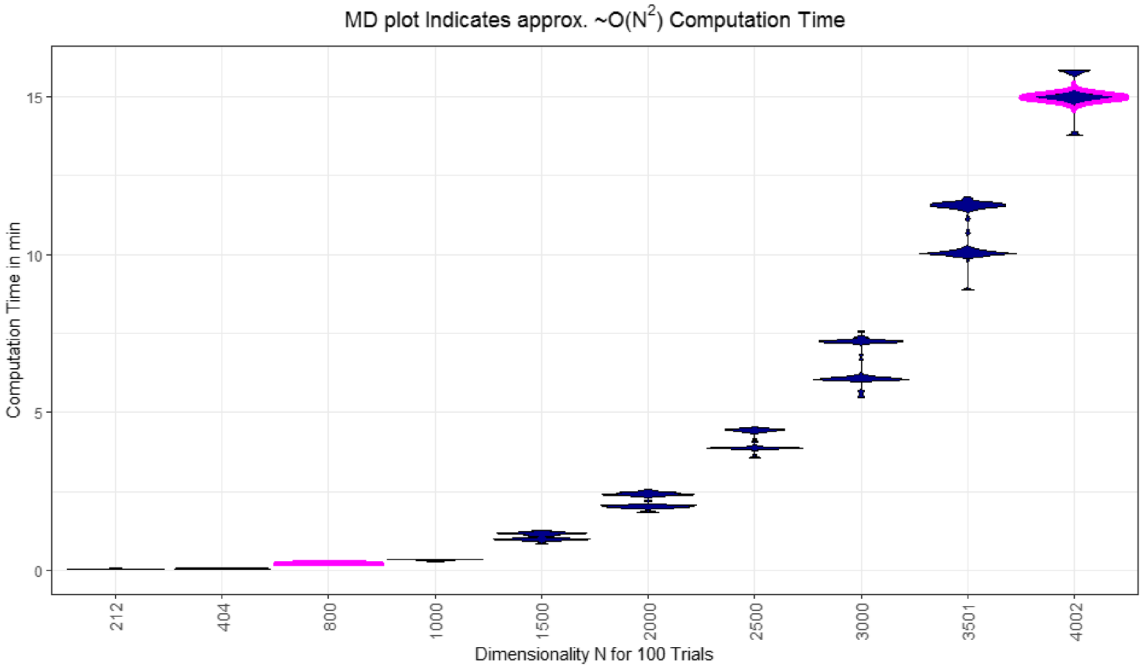

Fig. 18 MD plot of the computation time for IPBC with an increasing number of cases N indicating computation time of $O(N^2)$

required for user-interactivity. In total, 100 trials for a given number of cases N or dimensionality d of the data points are evaluated in two mirrored density plots (MD plots), which are presented in "Appendix B", Figs. 18 and 19, respectively. The MD plots indicate that the computation time increases linearly with dimensionality and quadratically with the number of cases. The main portion of the computation time is required to compute the internal sESOM algorithm [12] prior to U-matrix calculation, the display of the structures in the data in either 2.5D or 3D and clustering. Details about the evaluation procedure are described in "Appendix B".

5 Discussion

Comparison studies between different interactive clustering approaches are seldom done. VISTA [3] and gCLUTO [20] are compared with selected conventional clustering algorithms; iPCA [14], and Clustrophile 2 [15] provide user studies instead of a benchmarking study; Clustervision [17] provides only results on specific datasets without any comparison at all. Contrary to prior works, this work compared its results to accessible software. Automatic benchmarking with 32 conventional clustering algorithms was performed previously in [11].

While the more visualization- and animation-intense VISTA delivers only slightly worse results for low dimensional datasets, compared with the IPBC method, it is evident by the first clustering attempt for the Tetragonula and SCADI datasets that it fails to find any meaningful structure as soon as the number of dimensions becomes too high. For 40 and over 200 dimensions, it is also no longer easy to use since there are too many variables the user needs to adjust to obtain a projection of the data. The exceptions are clusters with low-intercluster distances for which VISTA seems to be inappropriate. Furthermore, VISTA shows clusters even if they do not exist in the data. For higher dimensionality, the advantage of the topographic map becomes apparent. Although the single display of structures in the data is more computationally intensive to create, it does not require additional adjustments of parameters from the user the higher the number of dimensions in the dataset is.

For the high-dimensional datasets, the IPBC method gives the best results. For the gCluto and IPBC methods give results of similar quality on the SCADI and Tetragonula datasets. Both methods also provide displays of the structures in the data that showed that the SCADI dataset does not seem to have an as clearly defined clusters as the Tetragonula dataset. While gCLUTO relies more on computational results and letting the user find the correct parameters by giving displays of the structures in the data primarily for error checking, the IPBC method lets the user perform the clustering herself or himself. In contrast, the gCLUTO method fails in case of very high-dimensionality in the Leukemia dataset for which IPBC provides accurate results. gCLUTO has some restrictions regarding the use of the Euclidean distance

that probably led to the failures at clustering the Hepta dataset, which is partly based on varying density. This insight and also the substandard performance for linear nonseparable clusters in the Chainlink and Atom datasets show that gCLUTO seems to be built with implicit assumptions about the input. It is surprising that for overlapping clusters separable only by density, gCLUTO shows the correct number of clusters but is unable to cluster the dataset correctly. Here the authors tried every option available in gCLUTO but were not able to improve the result in Table 2. It is possible that the density-based EngyTime dataset cannot be clustered with the distance metrics provided in gCLUTO. Moreover, gCLUTO shows a distinct cluster structure in Fig. 8b for a dataset without any natural clusters similar to datasets with natural cluster structures, for example in Fig. 13b.

Overall, the IPBC method seems to be the most versatile approach, without compromising on the quality of the results and without much need for adjustment on a specific dataset. IPBC also works very well on high-dimensional datasets. This finding makes the IPBC a valuable method for any interactive cluster analysis since it produces relatively fast results without first having to test several different parameters of projection methods or clustering methods. However, the user must select the appropriate projection method. If an appropriate projection method is selected, neither the curse of dimensionality up to $d = 7.700$ (Leukemia data set) nor the number of cases up to $n = 4096$ (EngyTime dataset) effects the outcome of IPBC. It should be noted that some projection methods will take a longer time to compute for a larger number of cases.

In addition, IPBC provides even more insights about the data than the other approaches, where the different border heights between clusters provide insight into which clusters are closer to one another, which for a gene dataset could be used to argue earlier for the closer relatedness of these groups. Additionally, IPBC yield clusters that have some points with higher dissimilarity to the rest of the cluster than others but are still clearly part of the cluster. Such information cannot be as easily obtained by the other solutions.

In summary, it is visible in Table 2 that a user is only able to cluster datasets in gCLUTO with very specific cluster structures of low-intercluster distances for which VISTA seems less appropriate. With VISTA it is also impossible to cluster datasets of very high-dimensionality. The results showed that IPBC outperformed VISTA and gCLUTO in terms of the adjusted Rand index on the clustering benchmark set of twelve datasets and the SCADI dataset. Applying IPBC to the Boston Housing Dataset reproduced the results in [71] as follows. Statistical testing showed that the cluster one consisted of houses with a lower market value and higher air pollution than cluster two. A clustering structure was not reported in [71]. Additionally, an explainable AI procedure [75] explained the clustering by one rule stating that the taxes were lower in cluster one than in cluster two.

6 Conclusion

This work investigated various displays of the structures in data and introduced interactive projection-based clustering (IPBC) for visual analytics. Contrary to iPCA [14], Clustrophile 2 [76], Clustervision [77], Morpheus [16], and VISTA [3], it interactively shows the high-dimensional density and distance-based structures in the third dimension on top of the scatter plot of projected points in either a 2.5D display or a 3D display of a topographic map. With this approach for information display, IPBC is the only visual analytics approach that accounts for the challenge stated in the Johnson–Lindenstrauss lemma [8, 9] that points in a scatter plot cannot accurately visualize high-dimensional distances. Unconventional for interactive clustering approaches (e.g. [76, 77]), the visual displays of IPBC are a parameter-free and open-source. The interactive projection-based clustering was compared with two available methods: gCLUTO which displays information about the structures in data in 3D and VISTA which displays information about the structures in data in 2D. IPBC outperformed these two accessible methods in twelve artificial and natural benchmark datasets as well as two additional natural datasets. The clustering performed by IPBC reproduced domain knowledge of one application and extended it further. IPBC can be accessed as a module in the R package "ProjectionBasedClustering" on CRAN. Further research on IPBC is required with the goal of performing a user study to investigate if 2.D displays or 3D displays serve better for the cluster identification task..

Appendix A: Implementation details

By default the IPBC algorithm starts with the NeRV projection of the parameter settings defined by Venna et al. [79] which are shown in Fig. 3: iterations=20, lambda=0.1, neighbors=20. Linear projection methods are included (PCA, ICA, projection pursuit) but are not recommended (see [47] for details). Accessible nonlinear projection methods are MDS, NeRV, Sammons mapping, uniform manifold approximation projection, Pswarm and t-SNE. The existence and number of parameters depend on the projection method. The algorithms yielding the topographic map does not require any parameters. The recommendation of the automatic projection-based clustering procedure requires the setting of the Boolean parameter structure type (checkbox,

Fig. 2) and the number of clusters that can be derived from the number of valleys in the topographic map. Listing 1 presents an example of IPBC.

```
> install.packages(ProjectionBasedClustering,dependencies = T)
> require(FCPS)
> data("Chainlink")#from FCPS
> library(ProjectionBasedClustering)
> V = interactiveProjectionBasedClustering(Hepta$Data)
> imx = interactiveGeneralizedUmatrixIsland(V$Umatrix,V$Bestmatches,V$Cls)
> require(GeneralizedUmatrix)
> GeneralizedUmatrix::plotTopographicMap(V$Umatrix,V$Bestmatches,V$Cls, Imx = imx)
> GeneralizedUmatrix:: TopviewTopographicMap(V$Umatrix,V$Bestmatches,V$Cls, Imx = imx)
```

Listing 1: Exemplary source code for applying IPBC to the Chainlink dataset. Visualization of the results can be either performed as a top view of the topographic map in 2D using plotly or with rgl with interactivity in 3D.

Appendix B: Computation time

Computations were performed in R 4.0.1 on an iMac Pro 2017 with the specification of 18-core Intel Xeon W, 256 GB RAM with the R package 'parallel' in R-core for parallel computation.

The computation time in minutes is shown via MD plots [80] in Fig. 19 for an increasing number of cases N and in Fig. 19 for an increasing number of dimensions d in 100 trials.

Mirrored-density (MD) plot visualize density estimation in a similar way to violin plots [80]. It can be shown that comparable methods have difficulties in visualizing the probability density function in the case of uniform, multimodal, skewed, and clipped data if the density estimation parameters remain in a default setting [80]. In contrast, the MD plot is particularly designed to discover interesting structures in continuous features and can outperform conventional methods[80]. The MD-plot is available in the R package 'DataVisualizations' on CRAN.

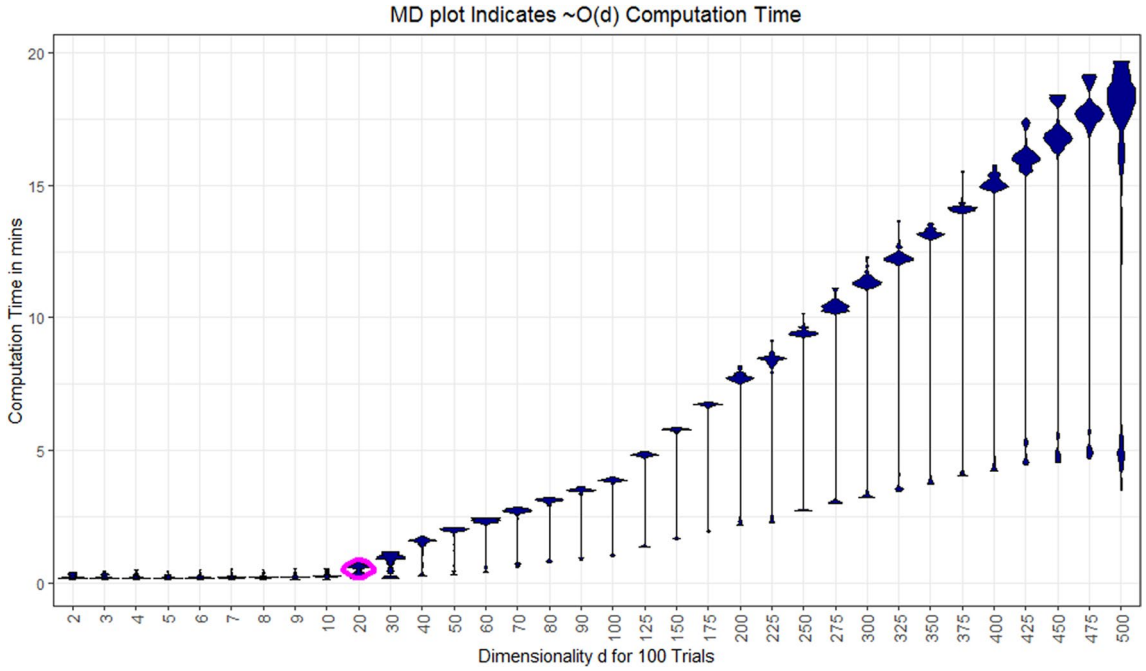

Fig. 19 MD plot for IPBC with an increasing dimensionality indicating a computation time of O(d)

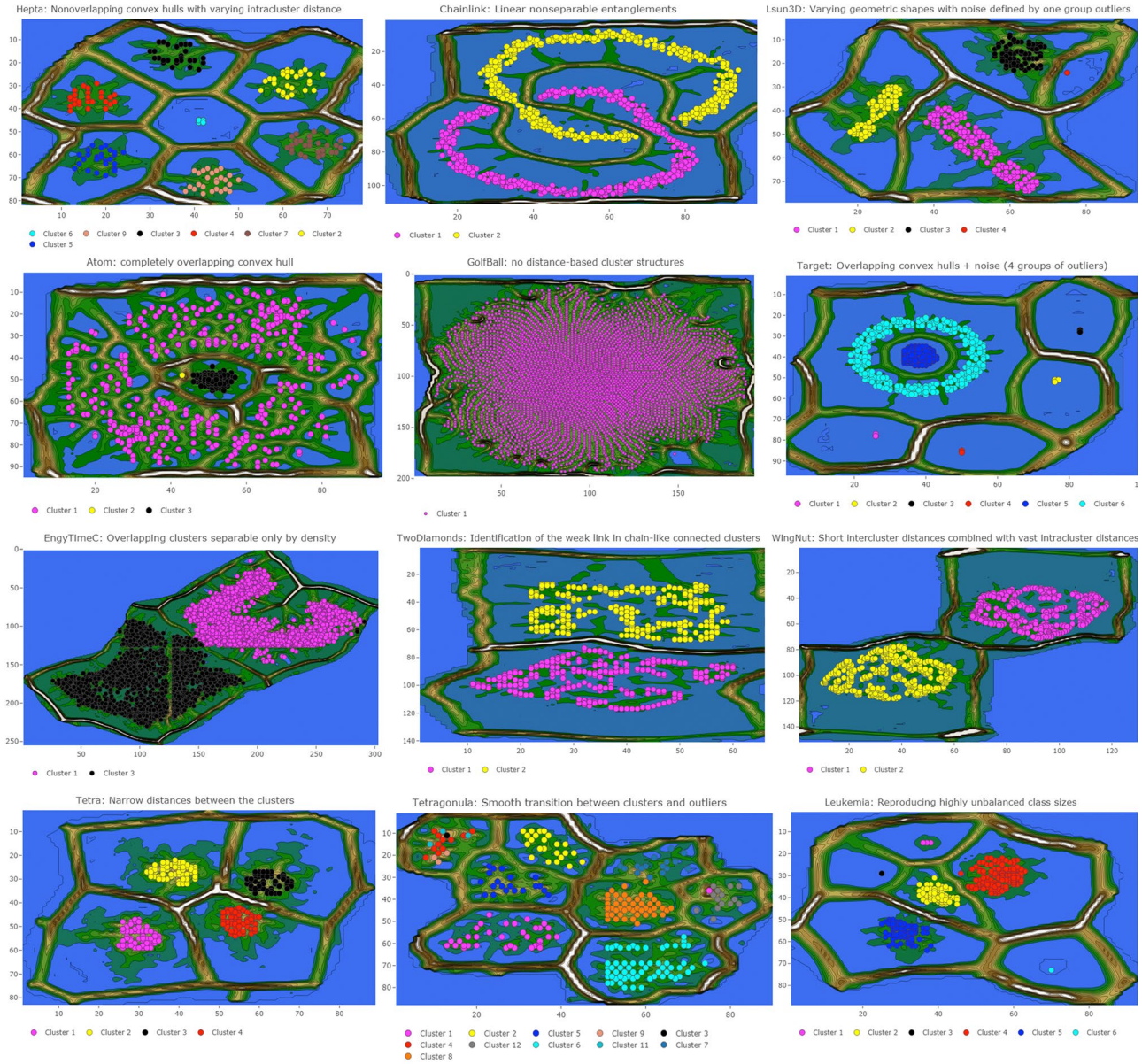

Fig. 20 Topographic maps in 2.5D display using plotly for the 10 artificial and two natural datasets of the FCPS for cluster analysis because the FCPS offers a variety of real-world challenges [68]

In Fig. 16, the specific datasets of FCPS datasets are Hepta($N = 212$), Lsun3D($N = 404$), Atom($N = 800$), Chainlink($N = 1000$), Golfball($N = 4004$). Additionally several samples of GolfBall are taken between $N = 1500$ and $N = 4000$ (see [78] for details).

In Fig, 16 $N = 554$ data points of the Leukemia dataset with varying dimensionality between $d = 2$ and $d = 500$ are used. The MDS transformation is applied to transform the published distance matrix into a dataset with a priorly specified dimensionality priorly.

Appendix C: Topographic maps in the 2.5 Displays for FCPS

In Fig. 20, the ten topographic maps for the FCPS datasets are presented in 2.5D display and used in the interactive clustering process of the IPBC method. The TopviewTopographicMap function is used here.

Acknowledgements The authors wish to thank Tim Schreier and Luis Winckelmann for programming the first version of the plotting functionality in plotly and Hamza Tayyab for the suggestion to apply IPBC to the Boston Housing Dataset.

Authors' contributions MCT was responsible for conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing—original draft, and writing—review and editing. FP was responsible for the evaluation of the comparable methods, methodology, software, writing—original draft, writing—review and editing. AU was responsible for the project administration, supervision, writing—review and editing

Funding Open Access funding enabled and organized by Projekt DEAL. No funding was received for this study.

Code availability IPBC is available through CRAN: https://CRAN.R-project.org/package=ProjectionBasedClustering.

Declarations

Conflicts of interest The authors have no conflicts of interest to declare that are relevant to the content of this article. ***Availability of data and material*** Data is available through [78] and for Boston housing through UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/machine-learning-databases/housing/).

Ethics approval and consent to participate Dr. Cornelia Brendel, in accordance with the Declaration of Helsinki, obtained patient consent for the Leukemia dataset and the Marburg local ethics board approved the study (No. 138/16).

Consent for publication All authors give their consent for publication of this work.

References

1. Cook, K.A., Thomas, J.J.: Illuminating the Path: The Research and Development Agenda for Visual Analytics. PNNL, Richland (2005)
2. Keim, D.A., Mansmann, F., Thomas, J.: Visual analytics: how much visualization and how much analytics? ACM SIGKDD Explorations Newslett. **11**, 5–8 (2010)
3. Chen, K., Liu, L.: VISTA: validating and refining clusters via visualization. Inf. Vis. **3**, 257–270 (2004)
4. Venna, J., Peltonen, J., Nybo, K., Aidos, H., Kaski, S.: Information retrieval perspective to nonlinear dimensionality reduction for data visualization. J. Mach. Learn. Res. **11**, 451–490 (2010)
5. Mirkin, B.G.: Clustering: A Data Recovery Approach. CRC Press, Boca Raton, FL (2005)
6. Ritter, G.: Robust Cluster Analysis and Variable Selection. CRC Press, New York, NY (2014)
7. Hennig, C., Meila, M., Murtagh, F., Rocci, R.: Handbook of Cluster Analysis. CRC Press, New York, NY (2015)
8. Johnson, W.B., Lindenstrauss, J.: Extensions of Lipschitz mappings into a Hilbert space. Contemp. Math. **26**, 189–206 (1984)
9. Dasgupta, S., Gupta, A.: An elementary proof of a theorem of Johnson and Lindenstrauss. Random Struct. Algorithms **22**, 60–65 (2003)
10. Thrun, M.C.: Projection Based Clustering through Self-Organization and Swarm Intelligence. Springer, Heidelberg (2018)
11. Thrun, M.C., Ultsch, A.: Using projection-based clustering to find distance- and density-based clusters in high-dimensional data. J. Classif. (2020). https://doi.org/10.1007/s00357-020-09373-2
12. Thrun, M.C., Ultsch, A.: Uncovering high-dimensional structures of projections from dimensionality reduction methods. MethodsX **7**, 101093 (2020)
13. Thrun, M.C., Pape, F., Ultsch, A.: Interactive machine learning tool for clustering in visual analytics. In: 7th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2020). IEEE, Sydney, Australia, pp. 672–680 (2020)
14. Jeong, D.H., Ziemkiewicz, C., Fisher, B., Ribarsky, W., Chang, R.: iPCA: an interactive system for PCA-based visual analytics. Comput. Graph. Forum **28**, 767–774 (2009)
15. Cavallo, M., Demiralp, C.: Clustrophile 2: guided visual clustering analysis. IEEE Trans. Vis. Comput. Graph. **25**, 267–276 (2018)
16. Müller, E., Assent, I., Krieger, R., Jansen, T., Seidl, T.: Morpheus: interactive exploration of subspace clustering. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1089–1092 (2008)
17. Kwon, B.C., Eysenbach, B., Verma, J., Ng, K., De Filippi, C., Stewart, W.F., Perer, A.: Clustervision: visual supervision of unsupervised clustering. IEEE Trans. Vis. Comput. Graph. **24**, 142–151 (2017)
18. Demiralp, Ç.: Clustrophile: a tool for visual clustering analysis (2017). arXiv:1710.02173
19. Kandogan, E.: Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In: Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 107–116 (2001)
20. Rasmussen, M., Karypis, G.: gcluto: An interactive clustering, visualization, and analysis system. Technical Report: UMN-CS TR-04. University of Minnesota, Minneapolis, MN (2004)
21. Endert, A., Ribarsky, W., Turkay, C., Wong, B.W., Nabney, I., Blanco, I.D., Rossi, F.: The state of the art in integrating machine learning into visual analytics. Comput. Graph. Forum **36**, 458–486 (2017)
22. Lötsch, J., Lerch, F., Djaldetti, R., Tegder, I., Ultsch, A.: Identification of disease-distinct complex biomarker patterns by means of unsupervised machine-learning using an interactive R toolbox (Umatrix). Big Data Anal. **3**, 5 (2018)
23. Schreck, T., Bernard, J., Von Landesberger, T., Kohlhammer, J.: Visual cluster analysis of trajectory data with interactive Kohonen maps. Inf. Vis. **8**, 14–29 (2009)
24. Hossain, M.S., Ojili, P.K., Grimm, C., Muller, R., Watson, L.T., Ramakrishnan, N.: Scatter/gather clustering: flexibly incorporating user feedback to steer clustering results. IEEE Trans. Vis. Comput. Graph. **18**, 2829–2838 (2012)
25. Andrienko, G., Andrienko, N., Rinzivillo, S., Nanni, M., Pedreschi, D., Giannotti, F.: Interactive visual clustering of large collections of trajectories. In: 2009 IEEE Symposium on Visual Analytics Science and Technology, pp. 3–10. IEEE, (2009)

26. Kraus, M., Weiler, N., Oelke, D., Kehrer, J., Keim, D.A., Fuchs, J.: The impact of immersion on cluster identification tasks. IEEE Trans. Vis. Comput. Graph. **26**, 525–535 (2019)
27. Brath, R.: 3D InfoVis is here to stay: Deal with it. In: 2014 IEEE VIS International Workshop on 3DVis (3DVis). IEEE, pp. 25–31 (2014)
28. Schumann, H.: 3D in der Informationsvisualisierung. In: Proceedings Go-3D 2015 (2015)
29. Kyritsis, M., Gulliver, S.R., Morar, S., Stevens, R.: Issues and benefits of using 3D interfaces: visual and verbal tasks. In: Proceedings of the Fifth International Conference on Management of Emergent Digital EcoSystems, pp. 241–245 (2013)
30. Wang, B., Mueller, K.: Does 3D really make sense for visual cluster analysis? Yes! In: 2014 IEEE VIS International Workshop on 3DVis (3DVis). IEEE, pp. 37–44(2014)
31. Dwyer, T.: Two-and-a-half-dimensional Visualisation of Relational Networks. Citeseer, (2004)
32. Tory, M., Sprague, D., Wu, F., So, W.Y., Munzner, T.: Spatialization design: Comparing points and landscapes. IEEE Trans. Vis. Comput. Graph. **13**, 1262–1269 (2007)
33. Tory, M., Swindells, C., Dreezer, R.: Comparing dot and landscape spatializations for visual memory differences. IEEE Trans. Vis. Comput. Graph. **15**, 1033–1040 (2009)
34. Marx, S., Hansen-Goos, O., Thrun, M.C., Einhäuser, W.: Rapid serial processing of natural scenes: color modulates detection but neither recognition nor the attentional blink. J. Vis. **14**, 4–4 (2014)
35. Thrun, M.C.: Interaktion von Aufmerksamkeit und Erkennung bei globaler Farbmanipulation von natürlichen Szenen. Philipps University, Marburg (2014)
36. Stoll, J., Thrun, M.C., Nuthmann, A., Einhäuser, W.: Overt attention in natural scenes: objects dominate features. Vis. Res. **107**, 36–48 (2015)
37. Thrun, M.C., Lerch, F., Lötsch, J., Ultsch, A. Visualization and 3D Printing of Multivariate Data of Biomarkers. In: International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG). Plzen. Czech Republic, pp. 7–16 (2016).
38. Colorimetry. C.I.E. Vienna: Central Bureau of the CIE, 2004 20.06.2004. Report No.: 3 901 906 xx y.
39. Liu, Y., Heer, J.: Somewhere over the rainbow: An empirical assessment of quantitative colormaps. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2018)
40. Ware, C.: Designing with a 2½D attitude. Inf. Des. J. **10**, 258–265 (2000)
41. Cockburn, A., McKenzie, B.: Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 203–210 (2002)
42. Jacquemin, C., Folch, H., Nugier, S.: Ocean: 2 1/2d interactive visual data mining of text documents. In: Tenth International Conference on Information Visualisation (IV'06). IEEE, pp. 383–388 (2006)
43. Tory, M., Kirkpatrick, A.E., Atkins, M.S., Moller, T.: Visualization task performance with 2D, 3D, and combination displays. IEEE Trans. Vis. Comput. Graph. **12**, 2–13 (2005)
44. Munzner, T.: Visualization Analysis and Design. CRC Press, Boca Raton (2014)
45. Delaunay, B.: Sur la sphere vide. Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk 7, 1–2 (1934)
46. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische Math. **1**, 269–271 (1959)
47. Thrun, M.C., Ultsch, A.: Using projection based clustering to find distance and density based clusters in high-dimensional data. J. Classif. (2020)
48. González, D.L., Einstein, T.: Voronoi cell patterns: theoretical model and applications. Phys. Rev. E **84**, 051135 (2011)
49. Lötsch, J., Ultsch, A.: Exploiting the structures of the U-matrix. In: Villmann, T.H., Schleif, F.M., Kaden, M., Lange, M. (eds.) Advances in Self-Organizing Maps and Learning Vector Quantization. Advances in Intelligent Systems and Computing. Springer, Cham, pp. 249–257 (2014)
50. Ultsch, A., Thrun, M.C.: Credible visualizations for planar projections. In: 2017 12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM). IEEE, Nany, France, pp. 1–5 (2017)
51. Ultsch, A., Siemon, H.P. Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. In: International Neural Network Conference; Paris, France. Dordrecht , Netherlands: Kluwer Academic Press; 1990. p. 305–308.
52. Ultsch, A., Siemon, H.P.: Kohonen's self organizing feature maps for exploratory data analysis. In: Proceedings of the International Neural Network Conference (INNC-90). Kluwer Academic Press, Paris, France, pp. 305–308 (1990)
53. Kraaijveld, M., Mao, J., Jain, A.K.: A nonlinear projection method based on Kohonen's topology preserving maps. IEEE Trans. Neural Netw. **6**, 548–559 (1995)
54. Häkkinen, E., Koikkalainen, P.: SOM based visualization in data analysis. Artificial Neural Networks—ICANN'97. Springer, pp. 601-606 (1997)
55. Hamel, L., Brown, C.W.: Improved interpretability of the unified distance matrix with connected components. In: 7th International Conference on Data Mining (DMIN'11), pp. 338–343. (2011)
56. Ultsch, A.: Maps for the visualization of high-dimensional data spaces. In: Workshop on Self organizing Maps (WSOM), pp. 225-230. Kyushu, Japan (2003)
57. Ultsch, A.: U*-matrix: a tool to visualize clusters in high dimensional data. Fachbereich Mathematik und Informatik (2003)
58. Thrun, M.C., Lerch, F., Lötsch, J., Ultsch, A.: Visualization and 3D printing of multivariate data of biomarkers. In: International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision. Plzen, pp. 7–16 (2016)
59. Thrun, M.C., Ultsch, A.: Swarm intelligence for self-organized clustering. Artif. Intell. **290**, 103237 (2021)
60. López-García, P., Argote, D.L., Thrun, M.C.: Projection-based classification of chemical groups and provenance analysis of archaeological materials. IEEE Access **8**, 152439–152451 (2020)
61. RStudio Inc. : Shiny: Easy Web Applications in R. JSM, Boston (2014)
62. Sievert, C., Parmer, C., Hocking, T., Scott, C., Ram, K., Corvellec, M., Despouy, P.: plotly: create interactive web graphics via 'plotly. js'. R Package Version 4, 110 (2017)
63. Adler, D., Murdoch, D., Others, a. rgl: 3D Visualization Using OpenGL. 0.100.30 ed2019. p. R package.
64. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
65. Thrun, M.C., Ultsch, A.: Swarm intelligence for self-organized clustering. Artif. Intell. (2020). https://doi.org/10.1016/j.artint.2020.103237
66. McInnes, L., Healy, J., Melville, J.: Umap: uniform manifold approximation and projection for dimension reduction (2018). arXiv:1802.03426
67. Thrun, M.C., Ultsch, A.: Projection based clustering. In: International Federation of Classification Societies, pp. 250–251. Tokai University, Japanese Classification Society (JCS), Tokyo, Japan (2017)
68. Thrun, M.C., Ultsch, A.: Clustering benchmark datasets exploiting the fundamental clustering problems. Data Br. **30**, 105501 (2020)
69. Zarchi, M., Bushehri, S.F., Dehghanizadeh, M.: SCADI: a standard dataset for self-care problems classification of children with physical and motor disability. Int. J. Med. Inform. **114**, 81–87 (2018)

70. Franck, P., Cameron, E., Good, G., Rasplus, J.Y., Oldroyd, B.P.: Nest architecture and genetic differentiation in a species complex of Australian stingless bees. Mol. Ecol. **13**, 2317–2331 (2004)
71. Harrison, D., Jr., Rubinfeld, D.L.: Hedonic housing prices and the demand for clean air. J. Environ. Econ. Manag. **5**, 81–102 (1978)
72. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**, 846–850 (1971)
73. Hubert, L., Arabie, P.: Comparing partitions. J. Classifi. **2**, 193–218 (1985)
74. Chen, W.-C.: Overlapping codon model, phylogenetic clustering, and alternative partial expectation conditional maximization algorithm. Iowa State University, Ames (2011)
75. Thrun, M.C., Ultsch, A., Breuer, L.: Explainable AI framework for multivariate hydrochemical time series. Mach. Learn. Knowl. Extr. **3**, 170–205 (2021)
76. Cavallo, M., Demiralp, Ç.: Clustrophile 2: guided visual clustering analysis. IEEE Trans. Vis. Comput. Graph. **25**, 267–276 (2018)
77. Kwon, B.C., Eysenbach, B., Verma, J., Ng, K., De Filippi, C., Stewart, W.F., Perer, A.: Clustervision: Visual supervision of unsupervised clustering. IEEE Trans. Vis. Comput. Graph. **24**, 142–151 (2017)
78. Thrun, M.C., Ultsch, A.: Clustering benchmark datasets exploiting the fundamental clustering problems. Data Brief **30**, 105501 (2020)
79. Venna, J., Peltonen, J., Nybo, K., Aidos, H., Kaski, S.: Information retrieval perspective to nonlinear dimensionality reduction for data visualization. J. Mach. Learn. Res. **11**, 451–490 (2010)
80. Thrun, M.C., Gehlert, T., Ultsch, A.: Analyzing the fine structure of distributions. PLoS ONE **15**, e0238835 (2020)