INFORMATION VISUALIZATION APPROACH TO FORM BALANCED GROUPS

by

KENAN KOC

Submitted in fulfilment of the requirement for the degree of Doctor of Philosophy
to the School of Computing

October 2022

DEDICATION

To my family

ACKNOWLEDGMENTS

This may be the hardest part of the thesis to write up. Now, it is time to thank each person touching my life to complete the long journey. This research would not have been possible without the immense support I have received throughout the four years of conducting it.

My first thanks go to my two supervisors. I would like to firstly express my sincere gratitude to my supervisor **Dr Sara Johansson Fernstad**, for her various forms of support, guidance, patience, compassion, and wisdom. I would like to thank her for her continued mentorship and guidance from the earliest stages of my PhD, helping me understand what it takes to become a researcher. And I would like to thank **Dr A. Stephen McGough** for his help and advice, which he gives whenever I need it.

Also, I would like to thank **the Ministry of National Education of the Turkish Republic** for its financial support during my doctorate and master of science studies.

I'd like to acknowledge all my family, my close friends, Alexander Macquisten, Paul Omoregbee, other colleagues and lab mates for a cherished time spent together in social settings.

And finally, thanks to my wife, **Kafiye Koc**, who endured this long process with me, always offering support and love.

ABSTRACT

Abilities such as visualization and interaction play essential roles in data mining since they can help people grasp and explore information more easily through their ability to bring out complex and multivariate patterns in data. The research presented in this thesis exploits and demonstrates the powerful combination of visual representation, domain knowledge and machine learning techniques to support challenges related to forming balanced groups. Teamwork is of substantial interest in academia and industry since interpersonal skills count in modern society. A team can, for example, be defined as a group gathered around a common project. Education is one of the domains in which group studies are important, and research studies are done to increase the effectiveness of group studies. In this case, forming appropriate groups for tasks at hand may be overwhelming for educators, as several factors may affect the quality of the group output. The current research supports educators in the group formation process and explores how to form groups systematically with less bias. In this thesis, a holistic framework called GroupVis, is presented in which exploration, clustering and grouping are considered user workflow aspects of group formation. In the GroupVis, each of these aspects is designed as a module, and each module contains visualization and computational methods within itself. The framework is designed with a top-down flow, where the result of the higher modules acts as input to the lower modules. As part of the GroupVis research a novel glyph was designed and evaluated, as a method that supports the comparison of balanced patterns in multivariate data. The three main modules of the GroupVis support group formation with functions such as analysing the data attribute field, exploring the cluster field visually with different settings, creating desired type groups, and interactively examining and modifying the resulting groups. The effectiveness of the framework and its modules is evaluated through semi-structured interviews with target users as well as through a heuristic survey with domain experts in education. The approaches presented here were designed and developed to be practical and applicable for the formation of groups in a wide range of domains, with the educational setting being one of these domains where we recognized their usefulness.

TABLE OF CONTENTS

LIST OF TABLES

Table (*Chapter no- Figure no*) Page

LIST OF FIGURES

Figure (*Chapter no-Figure no*)

Page

LIST OF ALGORITHMS

This page intentionally left blank.

Chapter 1 : INTRODUCTION

1.1 OVERVIEW

Research in a wide variety of fields, including health, education, and finance, has been spurred on by the rapid growth of large data sets and our desire to extract the data's inherent information. It is a common goal for these fields to help us better understand data, what it contains, and how to find the particularly important parts of that data. The primary areas required to attain this goal include classical statistics, data visualization, and the merged fields of data mining, pattern recognition, machine learning, and artificial intelligence.

Visual analytics solutions provide technology that combines the strengths of human and electronic data processing [1]. Information transformation is the overarching goal of visual analytics, much like how information visualization alters how we see databases. Visual analytics makes our information and data processing process transparent for analytical purposes. Instead than focusing only on the outcomes, visualizing these processes offers ways to communicate about information and approaches.

Using visual representations, information visualization (InfoVis) takes advantage of humans' remarkable capacity to view, examine, and comprehend data at multiple levels of abstraction. More specially, people can use InfoVis approaches to swiftly spot patterns, trends, outliers, and clusters and obtain insights.

The visualization process has been viewed as a pipeline that converts raw data into representations interpretable by the human perceptual system and amenable to additional processing. The representations are in the form of an organized set of graphical markers such as colours, size and space, which carry information in datasets [2]. Graphical images can be interactive, which allows individuals to manipulate the views in real-time, such as zooming in on a relevant object. Visualization techniques are commonly used in data mining. Brodbeck et al. [3] posit "it is recognized that data mining algorithms alone are not enough. Without expressive visualizations and interfaces, it is hard to achieve the necessary flexibility to understand results, generate new hypotheses and test them on the fly in a natural interactive environment". When designed with intended audience in mind, data visualisation has the potential to make a significant contribution to teaching and communication [4].

Teamwork or group-working is a substantial interest in academia and industry since interpersonal skills count in modern society. It is worth noting that team or group are used interchangeably throughout the thesis. Likewise, teamwork and groupwork mean the same thing. Teamwork has a common purpose of developing successful, mutual interactions among team members to achieve team goals. Nonetheless, planning teamwork activities in an informed manner becomes critical; particularly, the establishment of groups is the most crucial factor influencing the overall quality of group activities [5].

This thesis focuses on the problem of creating balanced groups and addresses this issue in the context of information visualization. One of the most productive areas to study a group formation problem is education. Due to the shift from instructor-centred pedagogy to constructivist pedagogy, learning designs have started to incorporate cooperative learning such as project-based, case-based, or problem-based scenarios.

These learning designs promote knowledge development [6]. The allocation of pupils to groups has been the subject of substantial study in the educational literature, as group formation is a critical component of providing a good collaborative experience [7, 8]. The majority of allocation approaches are based on each student's ability or performance level in the class. These are followed by the characteristics such as personality, learning style, interest, and ethnicity [9].

As well as such attributes, considering the factors affecting the formation of the group, such as the characteristics of the groups, the size of the groups, it seems that forming a group is far away from easy. Further, students' data is required to create these desired groups.

The availability and accessibility of educational data repositories, such as educational statistics data sets, has given a foundation for educational data mining. Researchers that use data from such archives can skip time-consuming tasks like data collection. This type of data is usually collected from learning management systems, including Blackboard and Canvas, which are commonly used for managing educational material, monitoring learners, and customizing learning and teaching processes. Student attributes such as off-task behaviour, subject-learning outcomes, educational interests can easily be inferred in the data sets [9].

There exist computational techniques supporting group formation based on various characters in collaborative learning contexts. Most of them are largely based on probabilistic models and work with a fixed set of parameters or features. They are heavily dependent on the nature of the problem domain. Generally, these models look like “black-box”, so their decision-making process is not transparent. Moreover, some have difficulties putting all individuals into appropriate teams when working with limited data.

In this thesis, a group building framework is presented in order to address the aforementioned limitations of current approaches as well as add expert knowledge to the group building process. The framework combines information visualization and machine learning methods and domain knowledge for making the process more systematic. Further, this thesis touches on the topic of visual analytics. Visual clustering is a core visual analytics topic, and this analysis algorithm was integrated into the grouping framework presented.

The provided mechanism includes various visualization methods alongside computational techniques applied to the data at hand. Target users can explore and understand the attribute space of students’ data via visual means. Based on gained insights and their opinion of students, the users can intervene with the computational methods or make changes to the outcome of the methods through the utilities provided in the tool. The approaches presented here were designed and developed to be useful and applicable for the formation of groups in a wide range of domains, with the educational setting being one of these domains where we recognized their usefulness.

1.2 MOTIVATION AND RESEARCH QUESTIONS

Many institutions, including educational bodies, collect their data using various technologies. For example, universities collect their students' data through tools such as learning management systems.

The motivation of the research presented in this thesis is to help target users systematically form groups using attributes from the data they deem appropriate for the task at hand. In this process, using information visualization methods is the main focus of this thesis.

To our best knowledge, there has been no research examining the role of visualization in the context of building balanced collaboration teams.

In particular, the research project is guided by the following research questions:

RQ1: How to create balanced project groups?

RQ2: How to make the group formation process more systematic and less biased?

RQ3: How can domain knowledge be integrated into the group building process?

RQ4: How to use a glyph-based visualization approach as an integral part of investigating group features and evaluating the goodness of formed groups?

1.3 CONTRIBUTIONS

This research aims to investigate how a framework can be designed in a way that is acceptable and feasible with target users to support forming balanced groups. From this starting point, the GroupVis framework is proposed, which is an interactive framework utilizing the combined benefit of visualization and data mining.

Two modified clustering algorithms were implemented to improve the quality of forming groups regarding heterogeneity and homogeneity as the grouping algorithms utilize the clustering results as a basis.

The research ultimately serves as an interdisciplinary approach that combines multiple methods from different domains. The methods are specific to the participants, context and research investigation needs. To summarize, the main contributions of the thesis are:

- A novel glyph visualization for balancing/unbalancing structure in heterogeneous multivariate data
- A group formation module (including algorithms, visualization methods and measurements) built on top of equal size clustering methods to generate heterogeneous/homogenous groups
- Providing a new way of using student data to support learning and teaching in an educational context

The terms balance, heterogeneity, fairness of groups is used interchangeably throughout the thesis. In such a group formation, individuals are assigned to teams, so that differences across teams regarding each attribute are a reasonable minimum level. Thus, the average academic performance of the teams must be approximately equal. The challenge in constructing groups from a pool of individuals, with each individual assigned to no more than one group, is to create teams that are as comparable as possible to the entire pool of individuals. It ensures that the groups are similar to one another because they are made to mirror the overall pool. Each group will have a diversity comparable to the variety of the pool as a whole. Conversely, it is not expected that diversity will be similar to the variety of the whole pool for groups that will be formed homogeneously.

1.4 THESIS OUTLINE

This section provides an outline of the thesis, alongside a brief overview of each chapter. It is organized as follows:

Chapter 2 (Background) presents a contextual review of the literature gathered from the three domains of information visualization, machine learning and education. Here, firstly the overview of information visualization is provided, then multivariate visualization approaches are discussed; finally, the use of InfoVis methods in the education domain, alongside existing group formation techniques, are presented.

Chapter 3 provides a novel glyph-based design for investigating and comparing balance among data attributes in multivariate data. Two usability studies were carried out. An early user study was carried out with the purpose of analysing the performance of various current glyph designs, especially the Bar, Star, Whisker, and Ring, in the context of showing structures of balance or unbalance. Then, to assess the usability of the PeaGlyph design, a second user study was done in a manner identical to the first. The PeaGlyph was tested against the first study's top two performing glyphs.

Chapter 4 introduces the GroupVis framework that comprises the methods introduced in previous chapters. This framework is divided into three components. The first presents the' Exploration module' for analysing data space, the second one is the 'Cluster module' that generates balanced size clusters to be used in a further step, and the final is the 'Group Module' that includes group formation approaches, alongside visual methods for the exploration of formed Groups as well as evaluation of the goodness of the groups in terms of heterogeneity or homogeneous. The chapters 5, 6, and 7 detail these three main components of the GroupVis framework.

Chapter 5 presents the initial module of the GroupVis framework. The module supports the exploration of given data sets through multiple projections. The module consists of several low dimensional embedding methods: Multidimensional scaling (MDS), t-distributed stochastic neighbour embedding (t-SNE), Self-organizing maps (SOM), and includes some information visualisation methods. Additionally, the chapter provides a semi-automatic visual guide to help users select an optimal scatterplot from among the projections given for the following task (i.e., Clustering task).

Chapter 6 uses a visual analytics approach, in which two modified clustering methods, Fuzzy c means (FCM) and K-means, are explored for constructing balanced clusters. Throughout the chapter, balanced clustering means that the obtained clusters have roughly equal sizes. It illustrates an interactive visual analysis procedure for the exploration of clustering results using Multiple Coordinated Views approach [10]. The output clusters originating from this module will be inputs for the following module (i.e., Grouping module).

Chapter 7 introduces the grouping component of the GroupVis, which is the final component of the system. The module offers two major grouping algorithms, several visual approaches, and metrics showing group similarities as well as visual representations of 'balanced/unbalanced' between groups that have been generated.

Chapter 8 evaluates the GroupVis with the heuristic survey and the semi-structured interviews. In this evaluation session, target users tested the functionalities of the tool and its usability regarding a list of visualization heuristics. The users' feedback and heuristics findings were discussed in this chapter.

Chapter 9 (Conclusion) provides a summary of the key findings, including results of the quantitative and qualitative evaluations across the case studies and user interviews in this study. In addition, it summarized the thesis contributions and discussed future works and recommendations for how the research should proceed.

Publication(s)

Chapter 3 was published the under the title 'PeaGlyph: Glyph design for investigation of balanced data structures' [11] in *Information Visualization* journal.

Demonstration

The videos below show the components of the GroupVis tool.

Overview
https://youtu.be/4LkKzpwXGwI

Attribute Module
https://youtu.be/7H6pMFQa5SA

Cluster Module
https://youtu.be/A0BAYE-54w4

Group Module
https://youtu.be/T4_S1aXzSm0

Chapter 2 : BACKGROUND

The literature review includes four main sections, which together form the basis of the project. This chapter begins with a review of information visualization concepts and principles. It is accompanied by multivariate data visualization methods, including size reduction charts. This is followed by a summary of information visualization techniques in the context of education. Finally, current studies in the literature on establishing collaborative working groups in education are mentioned.

2.1 REVISITING INFORMATION VISUALIZATION PRINCIPLES

This part will present the principles of information visualization, including data attributes, marks and channels, and visual perception of encoding channels.

The term "data visualization" can be defined in several ways. Most definitions focus on the connection between data and computer technology in order to transform data into a visual form. Card, Mackinlay, and Schneiderman [12] define data visualization as “the use of computer-supported, interactive, visual representations of data to amplify cognition” (p. 6). Moreover, Friendly [13] defines it as “information which has been abstracted in some schematic form, including attributes or variables for the units of information”. The field of visual analytics and information/data visualisation are closely interlinked [14] and that the research presented in the thesis use a combination of information visualisation and visual analytic approaches.

2.1.1 *DATA ATTRIBUTES*

An attribute can be thought of as a data field containing information about a data instance's characteristics or properties. Data attributes are often classified into different types, and there exist various classification taxonomies in visualization and data mining literature. Figure 2-1 shows the different types of attributes. The primary distinction is between categorical and ordered classification, and the ordered classification is further subdivided into ordinal and quantitative categories [15].

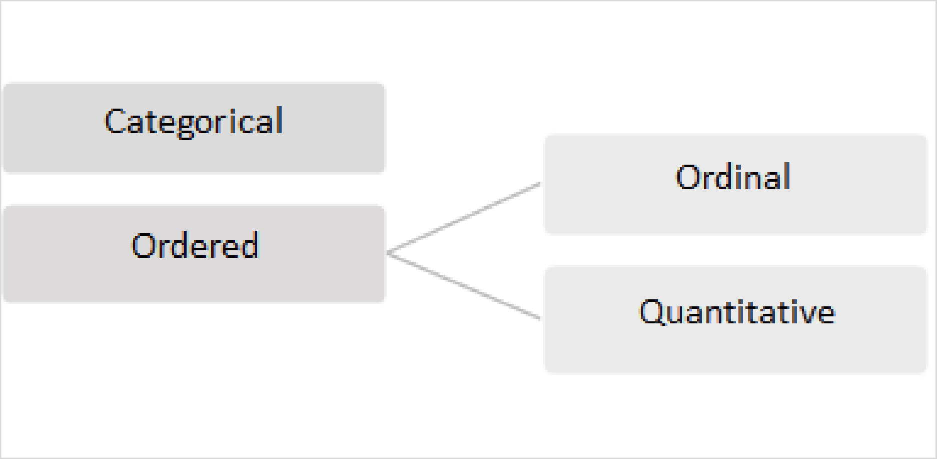


Figure 2-1 The attribute types by Munzner [15]

There is no implicit ordering in categorical data. The ordering is not implicit in the attribute itself, but any externally arbitrary ordering can be imposed upon categorical data. We do not need any arithmetic comparison to rank ordinal data (i.e. ratings). Quantitative data are represented by numbers and support arithmetic comparison.

The variety of data attributes limits the applicability of a single visual method to all types of data. The further subsection will address the visual variables that can be used to depict specific types of data attributes.

2.1.2 *MARKS AND CHANNELS*

A visual representation includes a number of visual building blocks that depict pieces of information. Each block, which is commonly called a mark or geometric primitive, is a fundamental unit for constructing any visualization. These are represented by a set of visual channels, which are retinal (visual) and planar (locational) variables proposed by Bertin [16]. His six retinal variables are size, shape, value, orientation, hue and texture. In addition, there are planar variables that are a pair of two coordinates (x, y), which locates any graphical artefact on 2D plane. In spite of the fact that Bertin lists each of these variables separately, successful representation can contain a combination of several visual variables. Put simply, the appearance of a mark is controlled by channels, such as position, size, colour separately or a combination of them. In this thesis, visual channels or variables refer to both retinal and planar variables.

Bertin's visual variables include six key channels, and since then, the researchers have continued to add new ones to the set. Recently, Chen and Floridi [17] categorized a large collection of over 30 visual channels.

Some of these channels are better perceived than others and lead to more accurate judgments than other channels with the same quantitative information. Maguire [18] provided a comprehensive overview of how these channels are processed by our visual system.

The comprehension of the visual channels and knowing how to organize them effectively will increase the chances of communicating information by visual means. All visual variables are not equally effective at representing information; in other words, some are more effective than others at mapping specific attribute types of data [15]. For instance, the length is a more appropriate channel for numerical data representation than the colour hue. The numerical values [1.0, 2.0] can be represented by using two parallel bars; one is twice the length of the other. Establishing such a relationship between colour hues is not possible.

Four levels of how visual variables are perceived are described as follows [16] :

> *Selective.* Channels that enable viewers to isolate encoded data and ignore others, select it from a group, e.g., planar, size, brightness, texture, colour, and orientation variables
>
> *Associative.* Channels that enable viewers to perceive them as a group, texture, colour, orientation, and shape.
>
> *Ordered.* Channels that facilitate visual ranking of data, e.g., planar, size, brightness, and texture.
>
> *Quantitative*. Channels that enable viewers to obtain extraction of ratios, e.g. planer and size.

The size variable, for example, is selective, associative, ordered and quantitative (with limited), whereas the shape is selective and associative, neither quantitative nor ordered. The attributes of the visual channels are examined in order to analyse the range of visual encoding alternatives.

2.1.3 *PERCEPTION OF VISUAL CHANNELS*

When constructing a visualization, it is critical to know which channel can most accurately represent which data type so that users, on average, are able to obtain the most accurate interpretation of visual objects. In this regard, Cleveland and McGill [19] studied the visual channels and ranked the various channels according to how effective they are at representing data. Since then, this ranking was extended to cover ordinal and categorical (nominal) data types by Mackinlay [20], as shown in Figure 2-2.

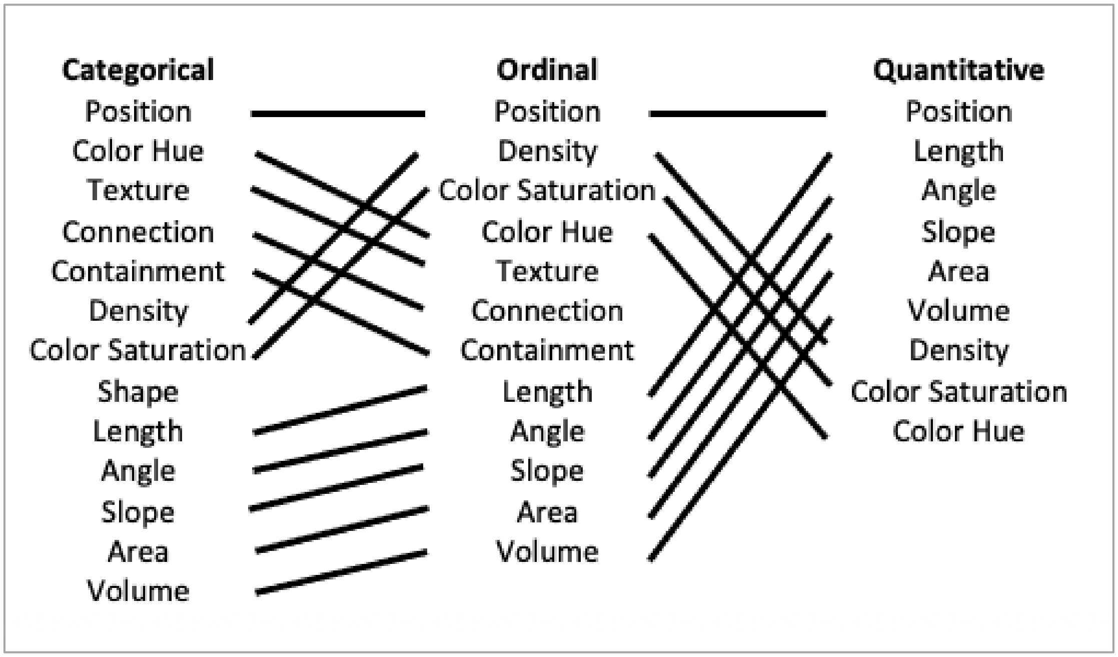


Figure 2-2 Reproduction of perception of visual channels ranking by Mackinlay [20]

A visual variable's ranking in Figure 2-2 can frequently be utilized to measure its relative effectiveness. For example, the position has a higher ranking than the area for encoding quantitative data. The accuracy ranking of position and length in quantitative perceptual tasks is higher than colour hue and density. In another study [21], the visual variables such as colour saturation, luminance, length, area, brightness have been investigated with regard to stimulus magnitude against perceived magnitude to determine their power. The findings showed that length was the most accurate channel while colour saturation was the weakest performer and produced a more perceived effect than the intended stimulus. Additionally, some channels are perceived as stronger than others due to their pre-attentive processing [22] since the pop-out speed of visual variables is not always the same. A good example of this is colour, as its pop-out effect is higher compared to others, and this may be an advantageous property for directing the viewer’s attention to critical features inside a visualization. Maguire [18] reviewed visual qualities in detail in his publication. Chung et al. [23] studied on the perceptual orderability of visual channels, and showed that certain visual channels are perceived as more ordered while others are perceived as less ordered.

It is critical to remember that such design criteria should ideally correspond to perceptual principles. Visual channels are often used together to present multiple data attributes at once. The organization of different channels can also create various visual effects for information that is being conveyed. In this case, one channel representing a data attribute may interfere with the other one portraying data attribute value. For example, encoding one data attribute into the chrome channel and the other into luminance may reduce the user's ability to correctly interpret each attribute independently. Hence, in order to inform effective visualizations, a better knowledge of the interplay between multiple visual channels is required [24].

2.2 MULTIVARIATE VISUALIZATION

Datasets with only one and two dimensions are referred to as univariate and bivariate, respectively; whereas multivariate represents datasets with multiple dimensions. A collection of variables describes each record in a multivariate set. A standard structure of multivariate data is defined as a MxN matrix where M denotes rows representing data items and N columns including variables. Multivariate data visualization is a class of visualization techniques that provide greater insight into such datasets.

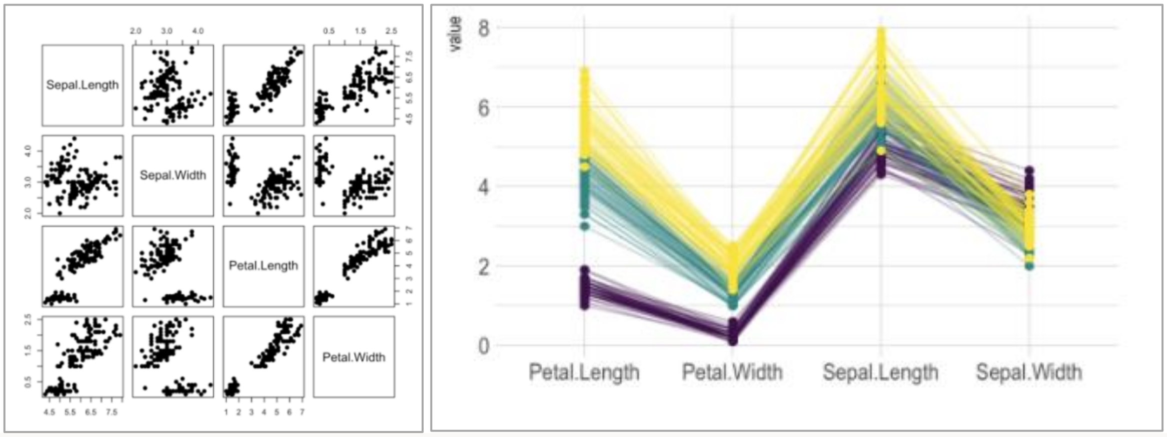


Figure 2-3 the scatterplot matrix at the left, the parallel coordinates at the right. Both methods showing *Iris dataset* [25]

A variety of visualization methods, including parallel coordinate plots [26], scatterplot matrices (SPLOMs) [27], as seen in Figure 2-3, treemap [28], and glyph-based visualization approaches [29], can be used to represent multivariate data and reveal insights into the data. There are advantages and disadvantages of each of these approaches. For example, like all high-dimensional plots, parallel coordinates (PCP) require fine-tuning processes such as scaling and sorting to expose information. They are particularly useful for comparing many variables at the same time and visualizing their relationships.

However, it is likely that geographical information in data will be lost through the use of these techniques. A data set is represented visually in glyphs by a collection of small visual primitives. The visual elements of a glyph such as size, colour and shape are mapped to numerous attributes of a data item. With this versatility, it is a highly ideal tool for communication as well as for enabling multi-dimensional analysis [30]. Aside from that, it is also possible to position glyphs either independently of or in conjunction with one another. Keim [31] provided an overview of multivariate approaches as well as a categorization scheme, which is depicted in in Table 2-1.

Geometric	Icon-based	Pixel-oriented	Hierarchical	Graph-based
Scatterplot matrix [27], parallel coordinates [26] etc.	Stick figures [32], colour icons [33], etc.	Circle segments [34], Spiral techniques [35] etc.	Treemap[28], dimension stack[36] etc.	Graph visualizations[37] etc.

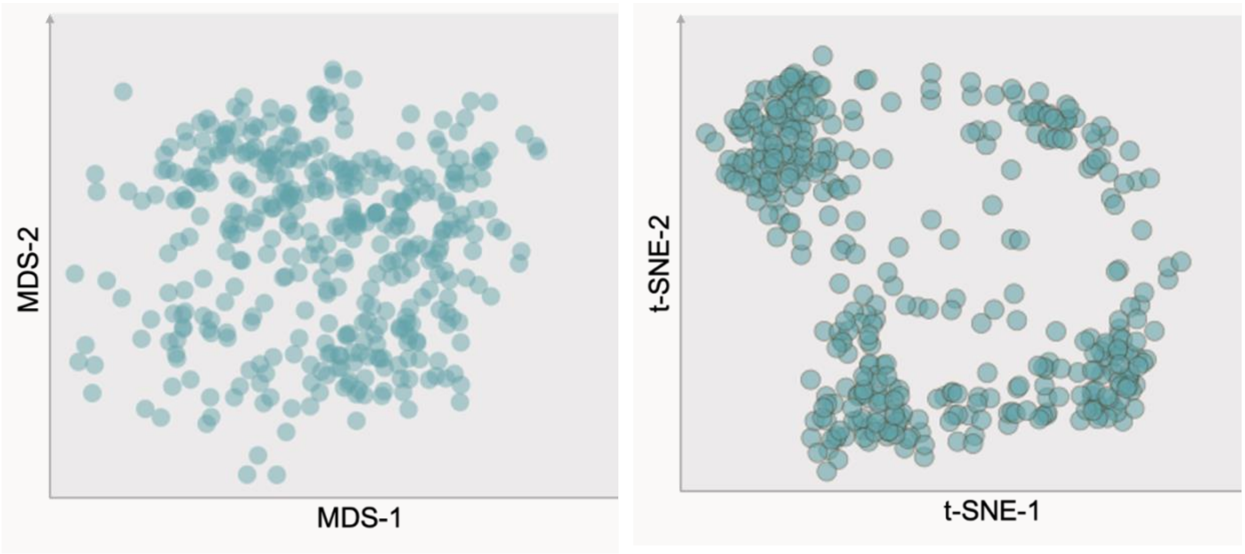
Table 2-1 Keim's classification scheme with some multivariate visualization methods

Besides the visualisation approaches to visualise multivariate data, interactivity is an effective means for analysing such data [38]. Shneiderman [39] offered the most widely known visual information seeking mantra, which outlines the essential components of interacting with visually presented data. The mantra includes three aspects: Overview, zoom and filter, then details-on-demand. An overview gives you a general idea or "picture" of how the data looks. In both zooming and filtering, unnecessary information is removed from the data representation to reduce the complexity of the view, allowing for additional data organisation. Detail-on-demand delivers this additional information without the need for a shift in perspective. A simple action can provide these details, such as an onmouse-over or selection [40].

2.2.1 *DIMENSION REDUCTION PLOTS*

Multiple dimensionality is a term that is most commonly used to refer to datasets with a high degree of dimensionality. It can be challenging for a human to grasp multi-dimensional concepts, and high-dimensional data may be challenging to visualize in its entirety since each dimension has so many possible values.

In proportion to the dimensionality of high-dimensional data, the difficulty of developing an effective visualisation, which promotes understanding of such data, grows in proportion to it [30]. Various approaches, collectively referred to as dimension reduction plots, have already been studied for displaying two-dimensional projections of multivariate data, such as the principal component analysis (PCA) [41] model, the objective of which is to discover a lower-dimensional subspace that captures the majority of the variability in the dataset, t-SNE, in Figure 2-4 (b), which has been shown to generate interesting low-dimensional clusters of data faithful to the distributions in the original data space [42], multi-dimensional scaling (MDS) [43], in Figure 2-4 (a), in which the goal is to approximate high-dimensional data in a low-dimensional (often two-dimensional) space by arranging points in such a way that their distance in the low-dimensional are near to their distances in the original dimension. The two-dimensional projections often provide informative views of data. In general, such methods attempt to reduce the difference between observed distances and lower dimension distances. t-SNE uses a probabilistic framework for transformation of observed and lower dimensional distances taking their local variance in data into account. MDS, however, does not include any of this type of local structure embedding at all.



a) MDS projection [43]

b) t-SNE projection [42]

Figure 2-4 The dimensional reduction techniques (*User Knowledge Modelling Dataset* [44] was used)

To this end, when it comes to data exploration, two-dimensional reduction methods have often been employed as an initial step [45-47].

2.2.2 *SELF-ORGANIZING MAPS*

Kohonen [48] proposed the self-organizing map, a type of artificial neural network. SOM visualises the relationships between instances in a high-dimensional dataset via dimensionality reduction based on attribute similarities. SOM produces both clusters as well as topology-preserved mapping of prototype vectors on a low- dimensional grid structure [49].

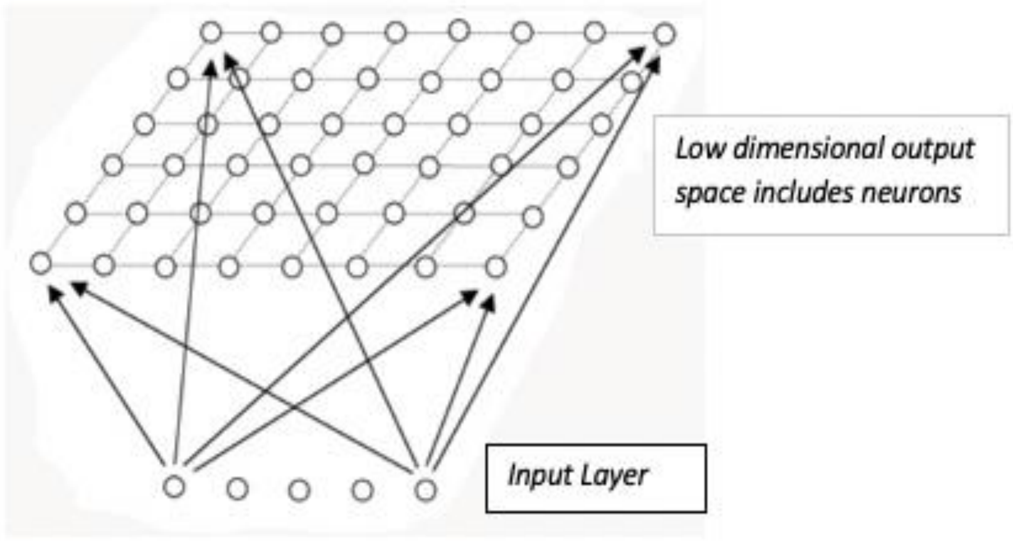


Figure 2-5 The architecture of Self organizing map [48]

Each neuron in the input layer is completely coupled to the nodes in output space, and the output of the SOM is a low-dimensional grid of nodes shown as circles in Figure 2-5. SOM is commonly used as a data visualization method in various applications. Qian et al. [50] used SOM to visualize and validate attribute relations between commercial materials. The study shows how SOM can be utilized for clustering with stock data [51]. This technique is rarely used in educational research. Silva et al. [52] aided in the formation of study groups by combining the SOM and k-means approach. In their study, k-means algorithm is performed on the output map to obtain the desired number of groups. Recently, Ahmad et al. [53] used SOM for clustering student browsing behaviours. In our research, SOM is used as a projection method to help investigate patterns and clusters in the multivariate data.

2.2.3 *QUALITY METRICS FOR SCATTERPLOTS*

A scatterplot is helpful for gaining a fast overview of the data and for highlighting any problems, unique qualities, or other noteworthy aspects of the data. One common task that supports the analysis in Scatter plot views is searching data groups and partitions [54]. Visual quality metrics narrow the projection selection space for users by filtering out views that provide little information and a low signal-to-noise ratio. Behrisch et al. [55] present a comprehensive overview of quality metrics for the field of information visualisation.

There exist some quality metrics identifying clusters, as seen in Figure 2-6, or group patterns in the scatterplots in the literature [56]. Such cluster-preserving projections quantify scatterplots for the visibility of dense groupings of data records. The ranking measures cover both labelled and unlabelled data that can be applied to scatterplots. In relation to the task, several factors have been taken into account to identify good views. Clumpiness [57], a term coined by Tukey and collaborators, indicates the degree to which data points in a 2D embedding are concentrated locally while simultaneously expanding globally. Tukey and Tukey presented Scagnostics (i.e. the term means scatterplot diagnostics) [58] for characterizing a large of scatterplots regarding 2D distributions in these plots, using measures of density, skewness, shape, outliers, and texture.

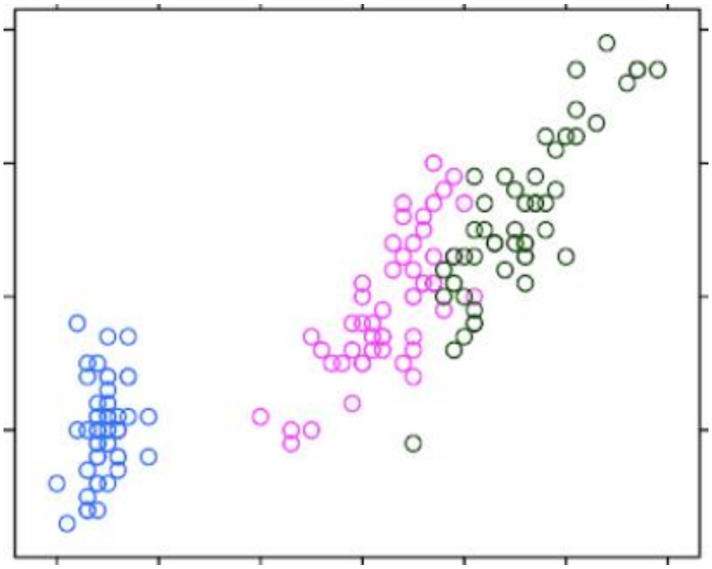


Figure 2-6 Scatter plot – Groups according to classified data [25] *(Fisher's Iris dataset was used)*

Several quality metrics for scatterplots of classified and unclassified data are presented by Sips et al. [59]. Unlike these studies, Sedlmair and Aupetit proposed an evaluation framework for evaluating such measures through a broad set of human judgments [60].

2.3 EDUCATIONAL DATA MINING AND VISUALIZATION IN EDUCATION

Many education institutions have used tools to collect massive amounts of data from educational settings, such as intelligent tutoring systems and different learning environments [61, 62].

Educators or administrators often use this type of data when developing a new program or making learning more effective. However, the huge volumes of data and complicated relationships between variables make data-driven decisions difficult. The field of study known as educational data mining (EDM) is primarily concerned with developing strategies for investigating data sets collected in learning environments. EDM often employs traditional data mining techniques such as classification, regression, and clustering to provide mechanisms for optimising the learning process [63]. EDM objectives are classified according to end users, namely learners, educators, administrators and researchers.

Information visualization methods are often coupled with EDM to simplify and convey meaningful information in these data sets, as it is hard to acquire the necessary flexibility in understanding findings and generating new hypotheses without the use of expressive visualizations and user interfaces [3]. Exploring the sequence of learning activities [64], tracing commonly taken paths by students when attempting to accomplish a certain goal [65], exploring and understanding user interaction in educational systems [66], visualizing the activities in study groups [67], visually inspect students' understanding and performance in their courses [64, 68, 69] are among EDM tasks. Lacefield et al. [70] explored how machine learning, predictive analytics, and data visualisation can be used to analyse student information system records in real time to detect at-risk students for advising and provide academic coaching. In another study by Nguyen et al. [71] educational assessment data of students was visualized by interactive word-stream and word-cloud method to show student context-related topic evolution over time. They applied a natural language processing technique to the mass amount of this text data. A motivating study related to group activities in education was conducted by Kharrufa et al. [72]. They presented an interactive visual evaluation tool, *Grouper Spinner*, which employed a radar chart for recognizing students' learning behaviours during group activities. A set of indicators forming the radar chart help teachers to keep track of a variety of group learning behaviours in their classes.

	Methods					
Literature	2D	3D	Geometrically-transformed	Iconic Display	Dense Pixel Display	Stacked Display
Mazza and Milani [68]					■	
Chiritoiu et al. [65]			■			
Johnson et al. [66]						■
Kay et al. [67]	■		■			
Kharrufa et al. [72]			■			
Vaitsis et al.[73]			■			
Raji et al. [74]			■			
Siirtola et al. [75]			■		■	
Saqr et al.[76]	■		■			
Santos and Brodlie [77]		■	■			
Chen et al. [64]	■		■			

Table 2-2 The use of multivariate visualization methods in education using Keim's taxonomy [31]

Visualization methods make analysis and interpretation easier for both instructors and students. For instructors, the methods are often used for training students and facilitating their learning. In addition to increasing the efficiency and quality of teachers' work, educational sets that use visualisation techniques also help students better understand and retain the materials that are presented to them [2]. Besides, the visual designs help decision-makers make more informed decisions based on that data. Sivanand and Frank [78] reviewed the current information visualization tools and practices in education. In their survey, they grouped the uses of visualization in education based on the sources from which the data is obtained. Firat and Laramee [79] categorized papers in which the purpose of interactive visualization is to teach students a subject while helping to facilitate learning. Unlike their works, we grouped the examples of visualizations used in education that aim to help lecturers or teachers make educational decisions based on student data sets collected through using Keim's taxonomy [31] in Table 2-2.

2.4 FORMING COLLABORATIVE TEAMS IN EDUCATION

Among the various instructional approaches to improve student performance mentioned in the education literature, collaborative learning is one of the well-documented. The benefits of collaborative learning have also been extensively researched [80, 81].

In the context of collaborative learning, educators emphasize the potential benefits of learning through interaction. Thus, the adequate formation of groups is crucial to promote the occurrence of meaningful interactions between group members, which leads to better learning situations [82]. As stated by Gibbs [83], group formation under appropriate conditions enhances peer learning and promotes providing students with an opportunity to clarify and deepen their grasp of subjects through discussion and repetition with peers. Many factors influence group formation, including the demographics of the group members, such as their age, gender, and race; group's size, group allocation procedures, and other differences between individuals. In addition, without careful formation and planning, group work can frustrate students and instructors and feel like a waste of time [84].

According to Wang et al. [85], in order for a group to work well in a given learning environment, teachers need to identify particular student characteristics such as gender, ability or psychological features such as learning styles [86, 87], self-efficiency [88] and the group types (homogeneous and heterogeneous) for learning activity. However, various teaming criteria and different goals make the team formation problem complicated and time-consuming [89]. Members are typically assigned to groups using one of four methods: random assignment, self-selection, particular criteria or task assignment [90]. The way student groups are formed has a marked impact on the quality of the end product [90].

Numerous lecturers organise groups using some type of random appointment mechanism. The procedure is frequently used since it is reasonably simple to give and requires little preparation. However, there are certain disadvantages to using the random appointment method. Students may feel as if they have no say in the selecting process. They may also be concerned about being allocated to a group that contains people that are incompatible with one another. In some cases, students are requested to establish groups on their own. Under these circumstances, pupils are more likely to know other students in their class and to choose to collaborate with them. Students who do not know anyone else in the class, on the other hand, may find it challenging, and it may be perceived as unfair to the rest of the class. Specific criteria refer to the grouping process leading to homogeneous and heterogeneous groups. Homogeneity within a group means that students are grouped according to the same personality traits, skills or preferences; however, in the case of heterogeneity, students should differ in terms of background, ideas, personality, ethnicity, and gender [80]. Furthermore, there are computer-based methods aimed at automating this process so that group formation can be done efficiently and effectively.

Among the algorithms used in the studies, there are different computational techniques, including Bayesian Network, swarm intelligence algorithms, machine learning techniques and ontologies are other computational techniques [8]. Group size has a marked impact on the quality of the end product that is assessed, as Kerr and Bruun [91] stated that individual motivation decreases if the group size increases. Four to six students seem to be ideal [83], with groups of eight or more creating significant problems, such as the difficulties of reaching decisions, allocating tasks, monitoring progress and the likelihood of ‘social loafing’ in groups where students can hide more easily.

2.5 DISCUSSION

After reviewing the many methods and tools that deal with the grouping problem, the following main conclusions are reached.

Existing methods and tools generally:

- heavily depend on the nature of the problem domain.
- are unable to indicate the importance or impact of each individual factor on the decision made by the algorithm.
- are black-box models (no transparency in the decision-making process) and too complicated for any human to comprehend. The explain-ability of the decision-making process is crucial as they might need to know the rationale behind the decision that the algorithm has made- European Union's General Data Protection Regulation (GDPR, EU 2016/679) highlights the importance of being able to explain decision-making systems [92].

Moreover, most systems have difficulties in putting all individuals into appropriate teams when working with limited and incomplete data.

When it comes to educational context, organizing different types of student groupings seems a time-consuming task and imposes an extra workload on educators [93] . While filling the above-mentioned gaps in our group creation tool, it is among our motivations to provide teachers with a practical and easy way. Information visualization methods are commonly used because of their ability to facilitate providing insights into complex student data sets and exploring students' learning activities. However, using information visualization methods to form balanced groups is an unmet need in the literature that needs to be addressed. The proposed framework in this research aims to deal with the weaknesses mentioned above by bringing visualization, machine learning and human knowledge together to generate collaboration groups with a “measurable” degree of balancing or unbalancing. Additionally, this approach provides a new way of using student data to support learning and teaching in the educational context.

Chapter 3 : VISUAL INVESTIGATION TO SUPPORT FORMATION OF BALANCED GROUPS

This chapter contains materials previously published by Koc et al.[11] , and it introduces a novel glyph-based visualization, PeaGlyph, which aims to support the understanding of balanced and unbalanced data structures, for instance, by using a frequency format through countable marks and salient shape characteristics. The glyph was designed mainly for relevance tasks for investigating properties of balanced and unbalanced groups, such as looking-up and comparing values.

3.1 INTRODUCTION

Recent research has argued that not only the relational attribute values across data items but also the dispersion of attribute values of each data item in the set are functions to be considered in several application fields. An example of this is related to behavioural decision-making in marketing. Chernev [94] investigated the role of the balanced dispersion of the properties (i.e., value, ease of use, quality) of the products on the consumer' decision when choosing among alternatives.

As in the example, for many data analysis tasks, the balancing between data attributes is at least as important as the actual values of items. At the same time, a comparison of values is implicitly desired for these tasks. Even with statistical methods available to measure the level of balance, human judgment and domain expertise plays an important role in judging the level of balance and whether the level of unbalance is acceptable or not. Accordingly, there is a need for techniques that improve decision-making in the context of multivariate balanced group formation that can be used as a visual complement to statistical analysis.

Glyph-based visualization methods are commonly used to depict multivariate data sets and can be utilized both independently and as part of a composition of a set of data records. These methods provide flexible and useful abstractions for exploring and analysing multivariate data sets. Munzner [15] compares views with glyphs, and states that glyphs are small, nested, schematic regions; in contrast to views that are large, stand-alone and highly detailed. Due to their usual small graphical appearance, glyphs can be used in various settings, such as within node-link diagrams, treemaps, tables, or geographic maps. Borgo et al. [95] stated that a glyph as a sign could potentially gain greater attention and stimulate more cognitive activity during visualization than other forms of visual design.

3.1.1 *CONTRIBUTION*

This chapter presents PeaGlyph, a novel glyph design for communication of ‘balanced’ or ‘unbalanced’ dispersion of attribute values and the comparison of values. The PeaGlyph design was guided by established glyph design guidelines and an evaluation of existing glyphs in related literature. The limitations derived from the quantitative and qualitative results of the experiments were addressed in the glyph design by using countable objects (i.e., filled, unfilled, or part filled circles) to map related attribute values, rather than a pure length channel, as humans are good at counting and judging the relative frequencies of discrete objects [96]. Filled and unfilled circles were used to increase the perceptual salience, to ensure that the glyph structure was kept simple and since symmetrical elements are useful for communicating information.

The performance of the novel PeaGlyph was then compared to the best ‘performers’ of an initial study through evaluation. The results from the study are encouraging, and the proposed design may be a good alternative to the traditional glyphs for depicting multivariate data and allowing viewers to form an intuitive impression as to how balanced or unbalanced a set of objects are in terms of their multivariate attributes.
The main contributions of this work are:

- a flexible novel glyph representation for multivariate data, designed for investigation of structures with balanced and unbalanced attribute dispersion in data;
- two experimental evaluations that compare the usability of existing glyphs and PeaGlyph in the context of revealing balanced and unbalanced objects;
- utilization of a graph visualization approach to measuring the effectiveness of visualization methods from a cognitive load perspective.

In addition to these inputs, a number of design concerns, including the scalability of the PeaGlyph, are discussed in detail.

3.2 RELATED WORK

The work presented in this paper is mainly related to the subjects of glyph design and usability experiments on data glyphs. Previous work in these areas is summarized in this section.

3.2.1 *GLYPH ENCODINGS*

Different data attributes can be encoded by a set of retinal (visual) channels such as shape, colour, size, and orientation. Bertin [16] proposed the categorization of semantic relevance for determining the suitability of different channels in representing certain types of information. Cleveland and McGill [19] performed experimental studies to order visual channels based on how accurately they can be perceived. The most accurate method is the position along a scale, followed by interval length, slope angle, area, volume, and shading or colour saturation. In addition, adjacent properties of a glyph are generally easier to relate to and compare than nonadjacent ones. Ward [97] provided a data mapping taxonomy using these attributes. In one-to-one mapping, each data variable attribute is encoded into a distinct graphical primitive (i.e., visual channel). One-to-many mappings (redundant mappings) in which an attribute variable is mapped to more than one visual channel can be useful to improve accuracy. A many-to-one mapping represents multiple data attributes via the same kind of visual channel.

The appropriate design of glyphs is a crucial factor for their usability, and a well-designed glyph can enable efficient and effective visual communication like other encoding methods. For effective glyph visualization, appropriate visual channels should be carefully chosen and combined. As a glyph is likely to be composed of a set of visual channels, the channel composition may affect how individual channels are perceived. Maguire et al. [98] proposed a set of design principles of visual encoding based on the (findings) in perception and visible search areas. These are guidelines on semantic relevance, channel composition, pop-out effect and visual hierarchy.

Furthermore, Chung et al. [99] proposed criteria for glyph design in the context of sorting glyphs visually. Borgo et al. [95] provided an extensive overview of glyph visualization research. Several principles of perceptual organization, such as similarity, connectedness and closure, were outlined by Gestalt psychologists to assist the information design [97] and help readers save processing time. The glyph design described in this paper is based on these design guidelines and principles.

3.2.2 *GLYPH APPLICATIONS*

A large body of work exists for the application of different data glyph designs across multiple disciplines, from meteorological glyphs through geographically mapped medical data to sports visualization. This paper does not aim to provide an exhaustive summary of all but to provide examples of the diversity and flexibility of the concept of glyphs. Chernoff [29] is an early example of glyph-based visualization that uses human facial features to map multiple data dimensions into a single icon. Keogh et al. [100] utilized colour bitmaps for depicting time series data. Cao et al. designed a treemap-like icon [101] where each feature value was mapped to colour-coded cells, and then the cells were packed to produce individual icons. In PeopleGarden [102], the participants in a discussion group on web message boards were mapped as flowers in which the number of petals in the flower glyph represented the frequency of features. Glyph-based encoding has also been widely used in many analytic applications. Pearlman et al. [103] utilized data glyphs to understand large data sets in depth and diversity. SoundRiver by Jänicke et al. [104] is another example in which movie audio and video contents were depicted in glyphs. Ropinski and Preim [105] investigated the use of glyphs in medical visualization. Legg et al. [106] delivered MatchPad, where actions and events in sports were mapped to data glyphs. Similarly, the effectiveness of using glyphs in sports event analysis was shown by [99]. For visualizing temporal geoinformation, Drocourt et al. [107] examined several icon-based visual designs. Visualizations making use of environmental cues are quite common. An example of such data-driven glyph design is the botanical tree metaphor by [108]. However, the fruits and leaves are highly abstract representations (mainly coloured dots), and their shape does not change according to the data characteristics. The OECD's Better life index visualization was developed by [109], where environmental cues were used to visualize multidimensional data about country characteristics. Recently, Khawatmi et al. [110] developed a web-based application that enables users to create interactive glyph-oriented microscopy data representations.

3.2.3 *GLYPH USABILITY STUDIES*

A large number of usability studies have been conducted for different glyph types and design variations to evaluate their performance in order to select the best performers for various tasks, as well as to provide design guidelines for effective information encoding in the glyph design space. Examples of well-researched data glyphs are Chernoff faces, profiles (or Bar glyphs) [15], and Star glyphs [111], which have been used in various applications [104, 112]. These user studies can be divided into two categories: comparing data glyphs against each other and testing design variations of the same glyph category. Fuchs et al. [113] reviewed experimental studies on data glyphs and provided a systematic overview of the types of glyphs, the design characteristics, data and tasks. In their review, they found that synoptic tasks (i.e., similarity search, visual search, trend detection) were more commonly preferred for the studies rather than elementary tasks. Moreover, they reported that a high number of studies used three categories of visual variables, Position/Length, Colour and Orientation, to depict data in glyphs. Face glyphs received utmost research attention, so they were evaluated frequently. Blascheck et al. [112] presented a perception study to assess how quickly participants performed a simple data comparison task for small-scale visualisation on a smartwatch. Their research evaluated three common glyph types in smartwatches: Bar glyphs, Donut, and Radial bar glyphs. Their results showed that Bar and Donut encodings were preferred in small physical display spaces when quick data comparison is needed. Fuchs et al. [114] compared the performance of four glyphs for time-series data under a controlled experiment. They chose the Line, Stripe, Clock, and Star glyph for their study. Lee et al. [115] compared the ability of four different visualisation approaches for binary data sets. Two of these visualisations were Chernoff faces and star glyphs. The other encodings used a spatial arrangement of the objects based on a model of human mental representation (i.e. similarity) and distinctive features. The experiments confirmed that participants were faster, more confident and more accurate when an appropriate data visualisation was made available. Li et al. [116] used metaphor-based representations in their experiment and compared RoseShape glyphs against abstract polygons to visualise multidimensional data about the educational level in America. The results indicated that participants were more accurate when working with more realistic faces. Testing design variations of a glyph also drew attention. Fuchs et al. [117] conducted three experiments to compare the Star glyph with its variations in the detection of data similarity. They used Star glyphs with contours (outlines) and reference structures (i.e., tick marks and gridlines) or without them as plain Star glyphs. Based on their findings, they provide design considerations regarding the use of contours and reference structures on Star glyphs. Klippel et al. [118] experimented with examining the shape characteristics of Star glyphs on classification tasks. They varied assignments of attributes located along the horizontal and vertical axis to obtain different shapes based on the same data. Miller et al. [119] reproduced Klippel's study with the same settings, including coloured axes and the number of dimensions encoded. They used two different ordering strategies of star-glyph axes, similarity-based (homogeneous shape) and dissimilarity-based (spike shape), to gauge which one better supported users in visual clustering tasks.

3.3 DESIGN AND IMPLEMENTATION OF PEAGLYPH

This section will describe PeaGlyph, a novel glyph design that was built to facilitate the comparison of data values as well as represent overall structures in which the 'balance' between glyphs and encoded values is of interest. The design is based on established glyph design principles [95, 97-99] and on the results and feedback from the study presented in the '*Evaluation of glyph designs'* section. PeaGlyph has been designed to be used either as a stand-alone visualization or in combination with other visualization methods, such as scatterplots, tables and maps. Furthermore, it is able to encode both categorical and numerical values, as well as meta-data for attributes, as detailed in the following subsections.

3.3.1 *GLYPH DESIGN*

The visual features of a pea inspired the PeaGlyph design. The overall appearance of a pea consists of the combination of the shape with boundary details of a pea- pod, and the seeds, as exemplified in Figure **3-1** (left). These two aspects are the main features of the glyph available for mapping data. In its basic form, each PeaGlyph represents either a record in the data set or an aggregation of a group of objects (such as the centroid representation of a cluster). The data attributes of a multivariate data set are each represented by a pea-pod in the glyph, such that the number of pea-pods in a glyph corresponds to the number of attributes in the represented data set. Figure 3-1 (right) displays an example where the values of six attributes are displayed in a PeaGlyph with six pea-pods.

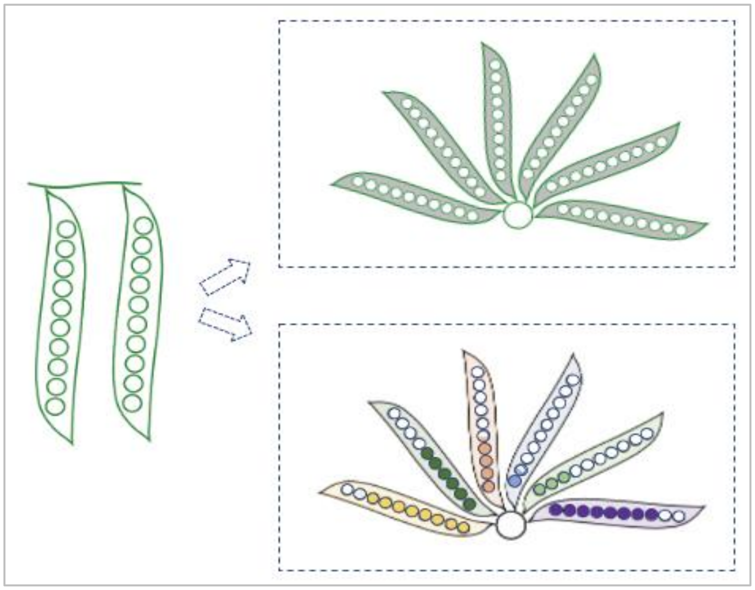


Figure 3-1 Examples of the basic the PeaGlyph design. (Left) Each pea-pod includes a maximum of ten seed. Continuous values of attributes are depicted into discrete form: filled, semi-filled or empty circles. (Top right) Abstraction of seed and pods for six data attributes with a circular layout. (Bottom right) Using colour to distinguish between attributes.

The basic glyph structure was intentionally kept simple with symmetrical elements that are useful for communicating information, as per the symmetry principle. The seeds and pods are highly abstract representations, as shown in Figure 3-1. The seeds are coloured circles, and their leaves are coloured with a light tone to show the shape of the pods. If there is a requirement to emphasize additional features according to the data, variation in colour or texture can be used for the pods, as described in Table 3-1, which summarizes the visual channels available for data representation in the PeaGlyph.

In the PeaGlyph, the data of a numerical attribute is represented through ten circles (seeds) in a pod, a number that can easily be related to percent values. Furthermore, while more seeds could be used to represent more detail, a smaller number of objects can often be easier to comprehend [120]. Based on the underlying data value, a number of circles will be filled, with all circles filled representing the maximum value of the attribute and no circles filled representing the minimum value. There is also an option to use semi-filled circles to represent values at a higher granularity. In Figure 3-1 (bottom right), it is visible from the three filled circles that the value for the upwards pointing pea-pod (attribute) is 45% of the attribute's maximum value, while the value of the attribute to the right is 15%. In a small setting, such as glyphs, filled and unfilled circles were preferred, as colour is a highly salient visual channel [98], making the value representation stand out from the background of pods. The result of the initial evaluation, as described in the Evaluation of glyph design section, indicated that the Bar glyph was the best performing representation for identification and comparison of values. Bar glyph utilizes length and position as the main visual channel for comparison of values, with the minimum value position as a point of reference for comparison. The PeaGlyph makes use of a similar abstraction in terms of employing the minimum and maximum values as points of reference for the comparison of values. Additionally, the combination of coloured and empty circles, when compared to the representation only by length or position, enables the user to also use the number of empty circles as an indication to judge and compare values. Research in human perception shows that humans are generally much better at perceiving, counting, and judging the relative frequencies of discrete objects as long as their total number is not too large [96]. To further facilitate the comparison of values across attributes, the glyph encodes an equal number of seeds for each pea-pod that represents a numerical attribute.

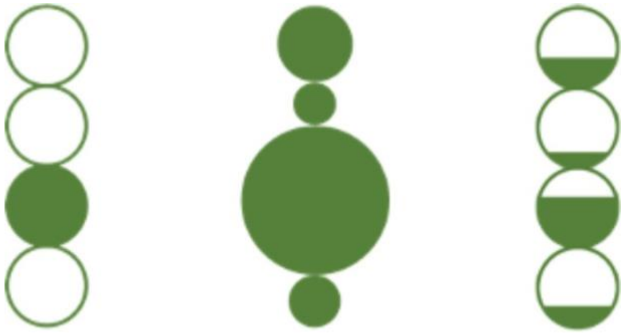


Figure 3-2 Alternative approaches to represent categorical attributes in PeaGlyph, for a data attribute with four categorical values. (Left) Representation of the categorical value of a single record. (Centre) Using pea-size to represent category frequencies for a group of records. (Right) Using semi-filled peas to represent category frequencies for a group of records.

Categorical attributes can also be depicted through the pea analogy by representing each unique category of an attribute by a seed, thus creating a pea-pod with as many seeds as there are categories for the corresponding attribute. Figure 3-2 displays examples of a categorical attribute with four possible values. If a single record is represented by the PeaGlyph, the seed representing the categorical value of that record would be filled (Figure 3-2, left). Different approaches can be taken for categorical attributes where the PeaGlyph represents a group of records. One alternative is to size each categorical seed according to the relative frequency of the corresponding category in the group (Figure 3-2, centre) or to use semi- filled circles where the fill level corresponds to the relative frequency of the category (Figure 3-2, right).

As described earlier, the circles representing the data value of an attribute are enclosed by a pea-pod shaped frame to further indicate that they are of the same attributes. As suggested in Table 3-1, the orientation of the pea-pod can be used to represent, for example, attribute meta-data with a small number of categories or relationships between adjacent attributes. Besides the shape of the pea-pods, each attribute can be encoded with a distinct colour to make differences between attribute levels clear, thus supporting the discriminability principle [97]. Attribute colouring can also be used to represent categories or groups of attributes, assigning a distinct colour to each group, which supports the similarity principle [95]. The combination of filled/unfilled circles, colour and shape as individual channels were selected as they normally do not affect each other in an integrated encoding, which is supported by the design principle of separability [99].

Further, there are other visual channels, which can be added to the list of PeaGlyph given in Table 3-1, such as varying curvature, texture, angle between leaves and length of leaves, to represent different types of data. However, these channels have not been implemented and tested.

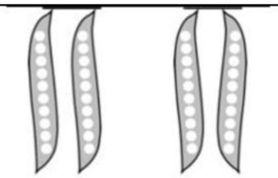
Key features	Typedness	Info
Glyphs overall size	Area	
The orientation of the peapod	Direction	 The natural structure of a pea contour can be used as a visual channel to encode orientation if needed
Colour hue/saturation	Categorical or numerical	
Seeds	Numerical or categorical (after normalisation)	Empty, fully filled or semi-filled circles, i.e. showing percentage or frequency falling into the interval identified

Table 3-1 Possible visual channels of the PeaGlyph

It is worth noting that the appropriateness of colouring as a method for distinguishing individual attributes will depend on the number of attributes of interest and should be decided on a case-by-case basis. There is a limitation in the number of distinguishable colours [98], and the hue channel should be used effectively [121]. If the number of attributes to be shown is high, a single colour may be preferred, and at this point, the pea-pod frame still helps in perceiving each attribute separately.

3.3.2 *LAYOUTS*

The examples provided in this paper are mainly focused on a radial layout, through this utilizing the strength of Whisker and Star glyphs in terms of evaluating structures of balance in the data. The glyph design can, however be used flexibly, as exemplified in Figure 3-3.

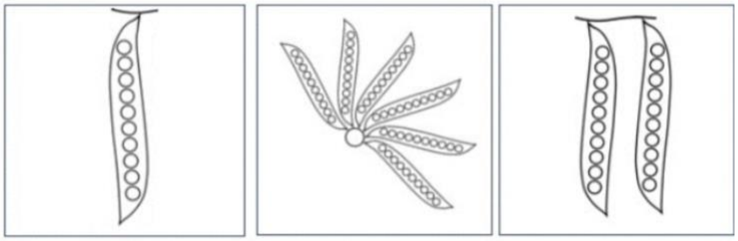


Figure 3-3 Possible design layouts of PeaGlyph

Pea-pods can be lined side by side or form a radial layout with a specific angle between each pea, as well as it can easily be adapted to a hierarchical layout. This flexibility can allow the user to choose the most suitable layout according to the task at hand. Building on the concept of the proximity principle [95], each data record, or each group of records, is represented by a set of pea-pods positioned close together in a glyph structure. To further emphasize belonging, an outer circle enclosing the pods can be used to visually separate the records or groups of records from each other, as seen in Figure 3-4.

3.3.3 *SCALABILITY*

PeaGlyph was designed for moderately sized multivariate data sets and thus has scalability limitations similar to the glyph designs compared in *Section 3-4*. For individual glyphs, the visual scalability is mainly related to the number of attributes. The 'interesting' attributes of data, or the representative attributes of the clusters in clustering analysis, are mapped to the pods of the PeaGlyph. Figure 3-4 shows examples of PeaGlyph with a different number of attributes displayed. It is likely that pea-pods will overlap after a certain number of attributes. This can, to some extent, be mitigated by not drawing the pod-shaped background, resulting in a level of overlap similar to that of the Star or Whisker glyph. To further prevent occlusions, the linear layout may be preferred instead of the radial. The visual scalability in relation to the number of glyphs that are displayed in a limited display space (i.e. the number of records or clusters) is highly dependent on the layout of the glyphs and directly comparable to the limitations of other glyph designs and approaches to overcome these limitations include the use of different layouts and interactive features. In an interactive system, the detailed attribute values may be shown as a tooltip to the users when hovering the mouse over the glyph of interest. Also, different interactive methods such as zoom in and pan to focus on glyphs can be used to facilitate the readability of the glyphs in small settings.

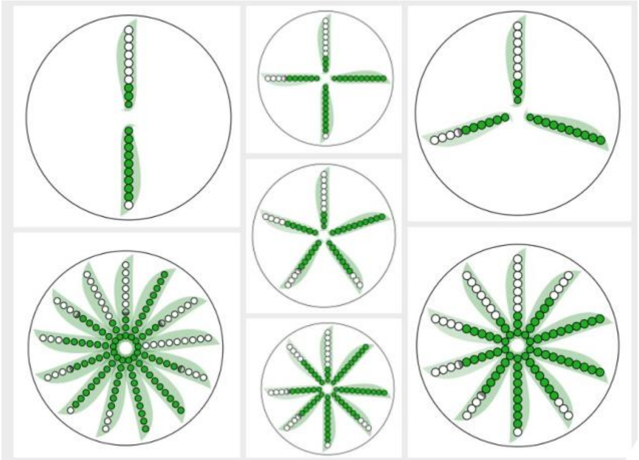


Figure 3-4 PeaGlyph with different number of attributes. The figure at the bottom left includes 13 attributes.

3.4 EVALUATION OF GLYPH DESIGNS

An initial user study was conducted with the goal of examining the performance of different existing glyph designs in the context of revealing structures of balance or unbalance in data. The findings from this study were to be used to guide the design of the PeaGlyph visualization for the exploration of balanced structures.

3.4.1 *VISUALIZATION METHODS*

Based on the literature, four potential glyphs were chosen to compare their performance for a set of tasks. These were Bar, Star, Whisker and Radial Bar glyphs, as displayed in Figure 3-5. In the study, each glyph presents a record in a data set, and each data value is mapped to an appropriate visual channel to encode the relevant information. For all glyphs used in the study, line lengths express the value of an attribute, and colour represents different attributes in the data. For example, in a student data set, each record is a student shown as a glyph in which attributes such as 'writing' and 'presentation' is depicted in colour, and their values are encoded in length.

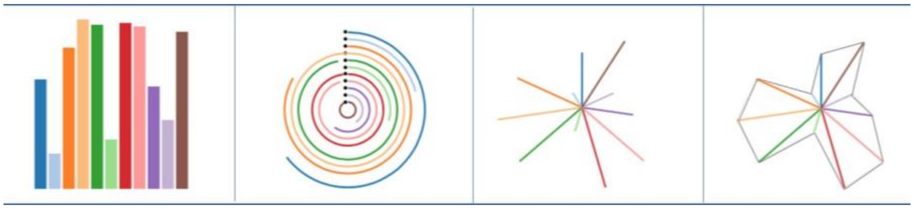
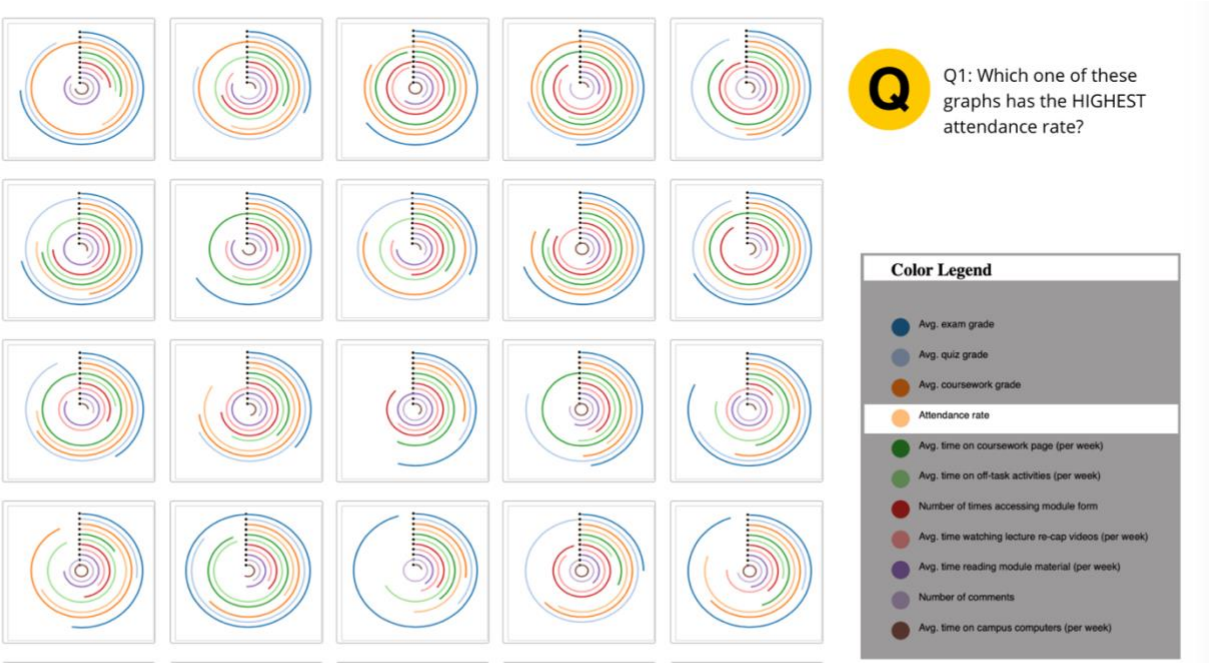


Figure 3-5 Chosen Glyphs for the user study from left to right: Bar, Ring, Whisker and Star Glyph-depicting same data values

Bar glyph was selected as it is one of the most utilized glyphs and visualization methods due to its linear layout. The data have multiple attributes, where each is represented by a unique colour, and the numerical values of the data items are represented by length/height. Whisker glyph also represents data values by length but uses line segments radiating out from a central point rather than parallel bars, as shown in Figure **3-5**. Besides, two variations of the aforementioned glyphs were used, Ring and Star, to assess whether additional graphic features such as contours and basement layout affect the task-dependent performance of the glyph designs. A ring glyph is a variation of the Bar glyph that utilizes a circular design, where each Bar has a different radius, so each Bar is judged by its angle. Star glyph and Whisker glyph have the same layout, but the Star encoding has contours between attribute lines.

3.4.2 *EXPERIMENT OVERVIEW*

The experiment was designed as a within-subject study. It included two main tasks, where each question was tested across the four glyph designs. In total, each participant performed ten different questions, which were repeated for each glyph design, and no data set was used more than once for each question. The questions were multiple-choice questions, and participants had to select the option that they thought was correct among a set of glyphs. The performance of the visualization methods, in terms of accuracy and response time when performing the tasks, was analysed for each main task. The study was conducted individually by using an online experiment builder, Gorilla [122], with an interactive interface as in Figure 3-6. Ethical approval was received prior to the study. Prior to the experimental phase, a short questionnaire was used to collect information about the participants and their previous experience with data analysis. Short background information was provided to ensure that all participants possessed the basic knowledge needed to interpret the visual representations and understand the tasks. This was followed by a training period, including a small number of test tasks using the different visualization methods. The training was used as a way for the participants to become familiar with the tasks, visualization methods and experimental environments. For the experimental phase, the tasks and visualization methods were counterbalanced using a Latin-square procedure, resulting in a unique ordering for each participant and, hence, reducing the potential learning impact on the results. The participants were able to take breaks between each task but were asked not to take a break while answering questions since the response time was measured. The system recorded the participants' answers and the time it took to answer the questions. The answers and response times were stored and later used to analyse the results. Upon completing each task, the participants were asked to rate their confidence in the chosen answer using a five-point Likert scale (1=low confidence, 5= high confidence). At the end of the study, participants were asked to select their preferred glyph for each of the two tasks. Besides that, some participants provided overall feedback on the study through email after completing the study.



Q

Q1: Which one of these graphs has the HIGHEST attendance rate?

Color Legend

- ● Avg. exam grade
- ● Avg. quiz grade
- ● Avg. coursework grade
- ● Attendance rate
- ● Avg. time on coursework page (per week)
- ● Avg. time on off-task activities (per week)
- ● Number of times accessing module form
- ● Avg. time watching lecture re-cap videos (per week)
- ● Avg. time reading module material (per week)
- ● Number of comments
- ● Avg. time on campus computers (per week)

Figure 3-6 Screenshots from the initial experiment. The image was taken from Task-1.

3.4.3 *TASKS AND DATA*

The tasks of the study were aimed to represent tasks of relevance for investigating aspects of balanced and unbalanced data structures. This included the identification of the attributes with the highest and lowest values and detection of balanced and unbalanced structures of the visual encoding, where the difference of the encoded values of attributes is minimum for a ‘balanced glyph’ and maximum for an ‘unbalanced glyph’. In other words, the dispersion between encoded attributes is lower than that of others for a ‘more balanced glyph’ and vice versa. The study tested the following hypotheses:

H1. Bar glyphs will perform better than the other methods for identifying and comparing values.

H2. Whisker glyphs will perform better than the other methods for tasks involving comparison of the overall structure of the glyphs or shape comparison, rather than identifying values.

Artificially generated data sets were created to fully control the patterns and attribute values in the study. For this study, each record was produced with 11 distinct attributes.

3.4.4 *FIRST TASK*

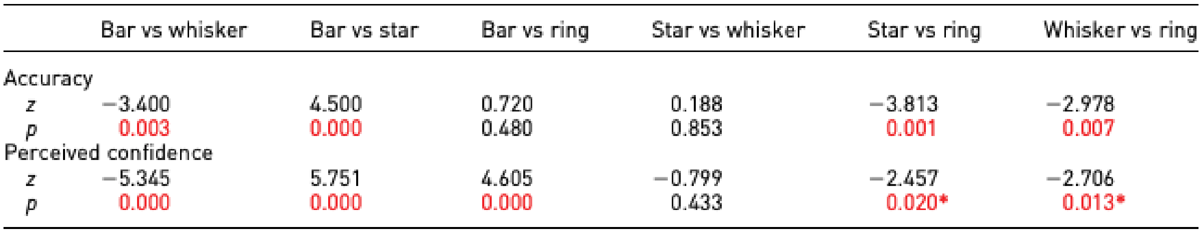
In the first task (Task-1), which relates to H1, participants were asked to find the glyphs displaying the max and min values of an attribute. The distance was kept identical for the different glyphs and, therefore, the same uniform small multiple layouts were used for all. As a consequence, it was essential to set a fixed aspect ratio for each glyph. A square aspect ratio was chosen for each glyph and a square framework to create a fairer comparison. For each trial, the same type of glyph showing different data was drawn at a resolution of 96 x 96 pixels, which is the same setting as in [114]. The glyphs were randomly laid out in an N-by-N grid, as displayed in Figure 3-6.

Results. 27 participants were recruited for the research study by leveraging the university network (i.e., sending e-mail to potential users in the university). Eighteen of them were male and nine females. 26 of them completed the study. Four of the participants were not included when evaluating the results, as their response times during tasks were too long or too short (being three standard deviations from the mean) compared to others. The participants were Master and PhD students from varying domains, recruited directly by the authors. A large portion of the participants were in fields other than computer and data science, and they reported that they had little or no experience in visualization and data analysis. The largest age group among participants was 29-38 years, followed by 20-28 years. In addition to the results presented in this section, the descriptive statistics of the results are provided as supplemental material. Friedman test was used to analyse the main effect, followed by a post- hoc test using Wilcoxon signed-rank test with Bonferroni correction for pairwise comparison.

Statistical testing confirmed significant differences for accuracy ($X^2(22) = 20.563, p<0.05$). shows post-hoc results with significant differences highlighted in red using $p<0.0125$ following Bonferroni correction. The difference in accuracy was significant when comparing Bar with Whisker (µBar=3.52, µWhisker=2.78). Meanwhile, the Star showed a performance close to the Whisker (µStar=2.66), with no significant difference. The Ring glyph (sometimes called a radial bar glyph) resulted in significantly better performance than the Star and Whisker.

Despite having different layouts, the Ring performed as well as the Bar glyph, and the difference was not significant (µRing= 3.38). According to the Friedman test, there were no significant differences in task completion time ($X^2(22) = 4.909, p=0.179$). Even so, the participants spent a long time answering the questions using the Star and Ring (µStar = 37498, µRing= 42965) compared to the Whisker and Bar (µWhisker=29067, µBar =3259), with time measured in milliseconds. The existence of speed and accuracy trade-off for the Ring glyph should be considered as to whether it is a good candidate for visualization. When looking at the perceived confidence ($X^2(22) = 28.262, p<0.01$) in Figure 3-7, the participants felt more confident answering the questions in Bar glyph, followed by Ring, Whisker and Star. The post- hoc analysis shows that all differences of Bar against other designs in perceived confidence were statistically significant. The results in the Star vs Ring (p = 0.02) and Whisker vs Ring (p=0.013) are near the significance threshold, which may be an example of a type-2 error caused by the Bonferroni measure being too strict

[123]. The mean values with 95% confidence intervals (CIs) are displayed in Table 3-3 and Figure 3-7. For mean accuracy, the confidence intervals for Whisker and Star almost overlapped completely. There was also an accuracy overlap between the Bar and Ring, while the CI of Bar glyph did not overlap with Whisker and Star. However, for the mean 'time', the intervals overlap for all four glyph designs.



	Bar vs whisker	Bar vs star	Bar vs ring	Star vs whisker	Star vs ring	Whisker vs ring
Accuracy						
z	−3.400	4.500	0.720	0.188	−3.813	−2.978
p	0.003	0.000	0.480	0.853	0.001	0.007
Perceived confidence						
z	−5.345	5.751	4.605	−0.799	−2.457	−2.706
p	0.000	0.000	0.000	0.433	0.020*	0.013*

Table 3-2 Post-hoc results (z- and p-values) from Wilcoxon signed rank test for different data glyphs. Significant differences are highlighted in red using p < 0.0125 in Task-1. (* denotes results near the significance threshold).

	Accuracy	Response time (ms)
Whisker	[2.34, 3.20]	[22594, 35549]
Bar	[3.25, 3.84]	[24435, 40703]
Star	[2.38, 3.07]	[24317, 50227]
Ring	[3.03, 3.79]	[33557, 53604]
The values in the bracket show [min, max] scores of the relevant methods.		

Table 3-3 Showing numerical values of CIs for mean of Accuracy and Response time in Task-1.

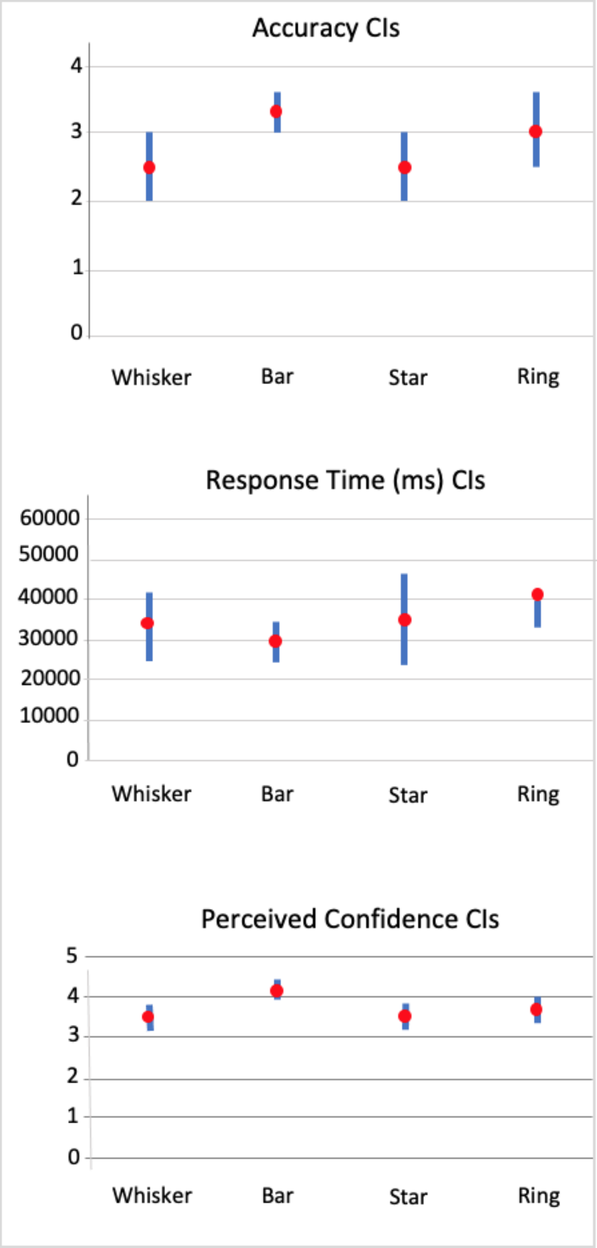


Figure 3-7 Confidence intervals of mean perceived confidence in Task-1. (95% confidence intervals adjusted for four data glyphs). This figure refers to Table 3-3

In summary, these results partially support the first hypothesis (see H1), with the Bar glyph displaying the best performance, while the Ring glyph was the second-best performer.

Visualization efficiency Different visualization methods prompt different amounts of cognitive load. To create better visualization as well as make more accurate assessments, it is important to understand this concern. Huang et al. [124] proposed a three-dimensional method of measuring visualization efficiency (E) (see Equation -1), where PRE signifies ‘user preference’, RT is ‘response time’, and RA denotes ‘response accuracy’, and these values are normalized using the z-score normalization to adjust them to a common scale. We used a likert scale to obtain user preference (PRE). In this way, after answering each question given in the tasks, the participants rated their perceived confidence for the answer they gave.

Equation -1.

$$E = \frac{Z_{RA} + Z_{PRE} - Z_{RT}}{\sqrt{3}}$$

The higher the score, the better the visual efficiency of the visualization. Using this approach, the efficiency scores of the compared glyphs - Whisker, Bar, Star and Ring - are 0.345, 0.171, -0.170 and 0.158, respectively. Thus, the Whisker glyph performed best in terms of efficiency, followed by Bar, Ring, and Star. Interestingly, although Star and Whisker glyphs are visually similar, the Star glyph’s efficiency score was considerably worse, while there was only a slight difference between the scores of the Bar and Ring glyphs. In terms of visualization efficiency, these results do not support H1.

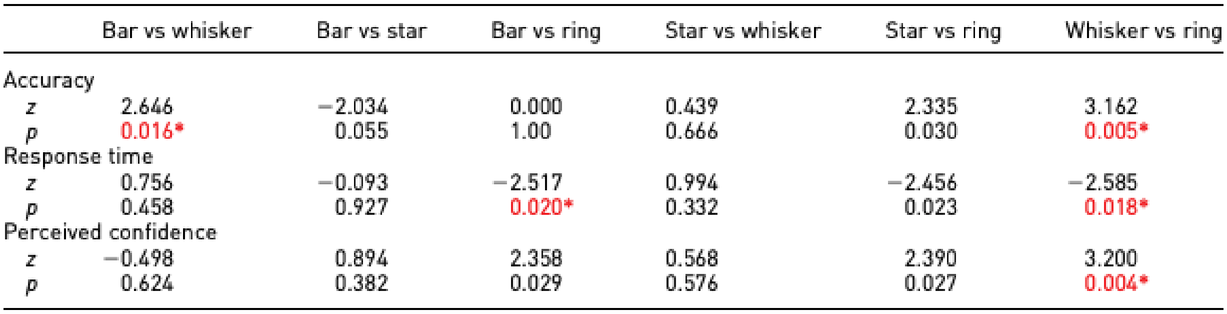
3.4.5 *SECOND TASK*

In the second task (Task-2), which relates to H2, participants were asked to select glyphs with a large variance, which indicates that numbers in the set are far from the mean and from each other, while a small variance indicates the opposite. In the experimental context, the glyphs with large variance were called unbalanced glyphs, and the glyphs with slight variance were called balanced glyphs. The participants answered the following questions in this task:

Q1. Which one of the glyphs represents the most balanced group?

Q2. Which one of the glyphs shows the least balanced group?

Results. 21 participants completed this task. In terms of accuracy score, the results were significant ($X^2(21) = 11.400$, p=0.001).



	Bar vs whisker	Bar vs star	Bar vs ring	Star vs whisker	Star vs ring	Whisker vs ring
Accuracy						
z	2.646	−2.034	0.000	0.439	2.335	3.162
p	0.016*	0.055	1.00	0.666	0.030	0.005*
Response time						
z	0.756	−0.093	−2.517	0.994	−2.456	−2.585
p	0.458	0.927	0.020*	0.332	0.023	0.018*
Perceived confidence						
z	−0.498	0.894	2.358	0.568	2.390	3.200
p	0.624	0.382	0.029	0.576	0.027	0.004*

Table 3-4 Post-hoc results (z-and p-values) from Wilcoxon signed rank test, with significant differences highlighted in red using p < 0.0125 in Task-2. (* denotes results near the significance threshold, which may affected by the strictness of the Bonferroni adjustment [32])

Table 3-4 shows post-hoc p-values for the task in which significant differences are highlighted in red using p < 0.0125 following Bonferroni correction. In particular, Whisker and Star glyphs (µWhisker= 1.66, µStar= 1.61) showed better performance than Bar and Ring (µRing= 1.33, µBar= 1.33). The number of accurate answers for Whisker was higher compared to Bar glyph, while there was no significant difference between the response times. Similarly, the Star glyph performed better compared to the Bar and Ring. When looking at response times ($X^2(21)$ =14.086, p=0.003), there was a significant difference. The participants spent significantly more time using the Ring glyph (µRing= 17126) compared to others (µBar= 6949, µWhisker = 7940, µStar= 7021). The Ring glyph showed worse performance than the Bar glyph in this task; however, the participants were perceived to have more confidence using the Bar glyph, followed by Whisker and Star. The statistical testing confirmed significant differences in perceived confidence ($X^2(21)$ =10.775, p=0.01).

	Accuracy	Response time (ms)
Whisker	[1.44, 1.89]	[6139, 9740]
Bar	[1.11, 1.55]	[5020, 8877]
Star	[1.39, 1.84]	[6001, 8042]
Ring	[1.11, 1.55]	[8524, 25728]
The values in the bracket show [min, max] scores of the relevant methods.		

Table 3-5 Showing numerical values of CIs for mean of Accuracy and Response time in Task-2.

While CIs, in Figure 3-8, of accuracy for Whisker and Bar overlap slightly, the mean accuracy for Whisker was higher than for Bar glyph. Star and Whisker showed quite similar performance in Table 3-5. The CI for response time shows a clearly longer response time for the Ring glyph compared to the other designs, which largely overlap with each other.

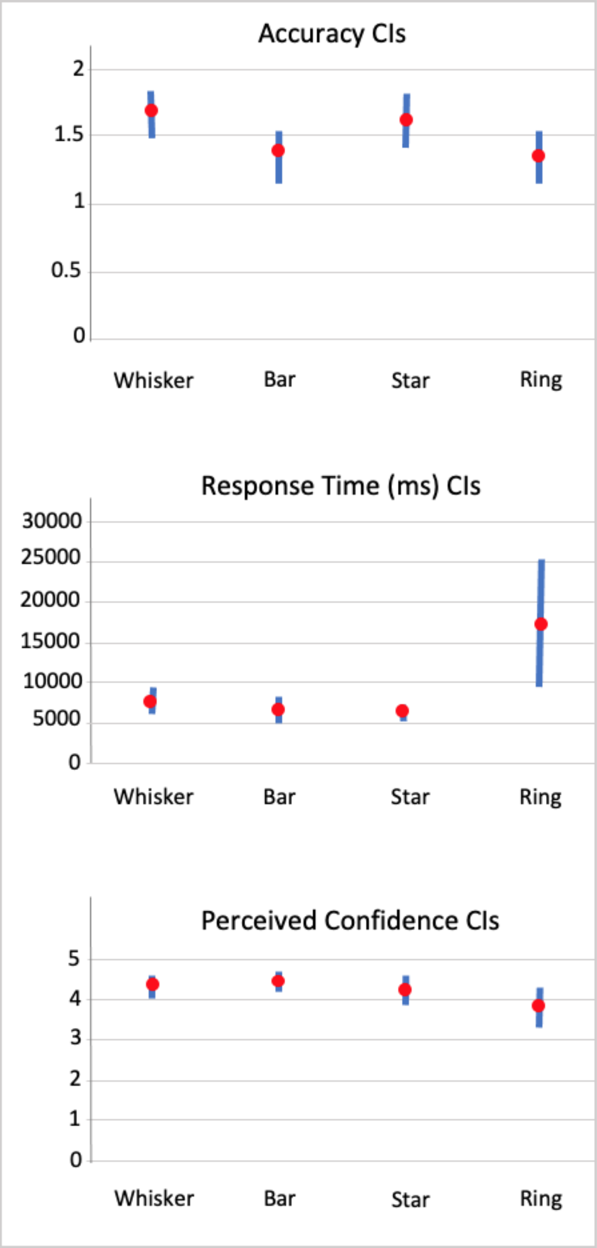


Figure 3-8 Confidence intervals for Task-2. (95% confidence intervals adjusted for four data glyphs). This figure refers to Table 3-5.

Visualization efficiency Using the approach suggested by Huang et al.[114], the efficiency of the Whisker (-0.183) and Star glyph (-0.191) is higher than for the Ring glyph (-0.229), and the Bar glyph (-0.278) had the lowest score for visual efficiency. Thus, the cognitive load was less using the Whisker and Star glyph compared to Bar and Ring. The results support H2 at large, as Whisker and Star have better accuracy than Bar and Ring; and showed better response times than Ring and more or less equal confidence.

3.4.6 *DISCUSSION*

In the evaluation, the performance of four glyph designs was compared: namely Bar and Whisker as baseline glyphs and their variations, Star and Ring. The focus was on two relevant tasks for the investigation of balanced data structures. The first task involved the comparison of feature values within and between glyphs, and the second task focused on comparing whether the combined records formed balanced objects in terms of attribute values. The goal is to identify possible usability problems and evaluate the performance of the glyphs for the tasks at hand. The results paralleled the findings in the literature, confirming that the comparison of attributes for the circular versions is not straightforward. Based on the study results presented here, Bar glyphs showed better performance against the other glyph variations for the identification and comparison of values (Task-1), thus supporting H1. In specific cases (e.g. when ordering of dimensions is not desirable), attribute values may be less easily perceivable and comparable between Bar glyphs. While considering the radial bar layout, it may be harder to perceive the length of the lines due to the angle and perspective. This problem is even more apparent for small size visualization methods (i.e. glyphs). Furthermore, equal distances in data space should be perceived equally when encoding a data variable to a glyph property [95]. This is clearly the case with the Bar glyph, which uses a combination of length and position in relation to a reference point, compared to the Radial glyph, which encodes equal values with different length bars. While Whisker and Star glyphs use perceptually uniform representations, the comparison of values may be complicated by the directional variation in relation to the reference point at the centre of the glyph.

Whisker and Star glyphs performed best for tasks requiring comparison of balanced/unbalanced structures (Task-2), which support H2 at large. For these glyphs, identification of balance/unbalance is related to evaluating the symmetry of the glyph shape, with a symmetric shape generally corresponding to a balanced structure. The facilitation of pattern perception using simple and symmetric shapes is supported by visualization guidelines [95]. The novel glyph design, PeaGlyph, described in the previous section, was designed taking the limitations of existing glyphs into account in the investigation of balanced and unbalanced data structures and comparison and identification of data values. In the following section, the performance of PeaGlyph is compared to the best performing glyphs in this initial study.

3.5 EVALUATION OF THE NEW GLYPH DESIGN

In order to evaluate the usability of the PeaGlyph design, a second user study was conducted following a similar approach to the initial study. The PeaGlyph was compared against two best performing glyphs from the first study, namely the Bar glyph and Whisker glyph.

3.5.1 *DATA*

For this study, the Sustainable Society Index data set (SSI) was used, in which 154 countries are included and it presents the level of sustainability of countries. SSI is a scoring system developed by Social Society Foundation to measure human wellbeing, environmental well-being and economic well-being every two years. ‘Human wellbeing’ indicators grouped into basic needs, personal development and well-balanced society are described by nine quantitative attributes. Each indicator has the interval [0.0 - 10.0] in the SSI scoring system. It is worth noting that a real dataset was preferred over a synthetic dataset for this study, as we want to see the performance of the glyphs in the study when comparing realistic data structures. Also, SSI dataset has enough number of features (i.e., 21 indicators) to be used for testing the performance of the glyph designs.

3.5.2 *EXPERIMENTAL DESIGN*

The experiment was designed the same way as the initial study, using a within-subject design and two main tasks where each question was tested across the glyph designs. The questions were multiple-choice questions, and participants selected the option that they thought was correct among a set of glyphs. The performance of the visualization methods was analysed for each main task in terms of accuracy and response time. The study was conducted individually by using the online experiment builder, Gorilla. Ethical approval was received prior to the study. In this second user study, each glyph represents a country in the data set, and each colour represents an attribute of the country. These attributes are Sufficient food, Sufficient drink, Safe sanitation, Education, Healthy life, Gender equality, Income distribution, Population growth, Good governance. The colour scheme used to encode each attribute was the *Tableau-10* from Tableau colour palettes [125].

The following hypotheses were tested:
H1. PeaGlyph performs equally well or better, in terms of accuracy and response time than Bar and Whisker glyphs, for questions related to finding the highest or lowest values.
H2. PeaGlyph performs equally well or better than Bar and Whisker glyphs for questions in which users are expected to compare balanced and unbalanced structures between glyphs.

After the study, each participant completed a qualitative survey regarding their glyph preferences:
Q1: Which of the three glyphs shown above is the best to compare values? (i.e., finding the highest or lowest indicators)
Q2: Which glyph shown above is better to show the unbalanced structure of glyph?

3.5.3 *FIRST TASK*

The first task (Task-1) aimed to compare the performance of the three glyph designs (Figure 3-9) under elementary lookup tasks where participants focus on a single attribute of a glyph and read individual values. Here the participants were asked to find the glyph in a set of glyphs that displayed the maximum and minimum values of an attribute.

The sample questions from this task were as follows:
Q1. Which one of these glyphs represents the highest education indicator?
Q2. Which one of these glyphs represents the lowest rate in safe sanitation indicator?

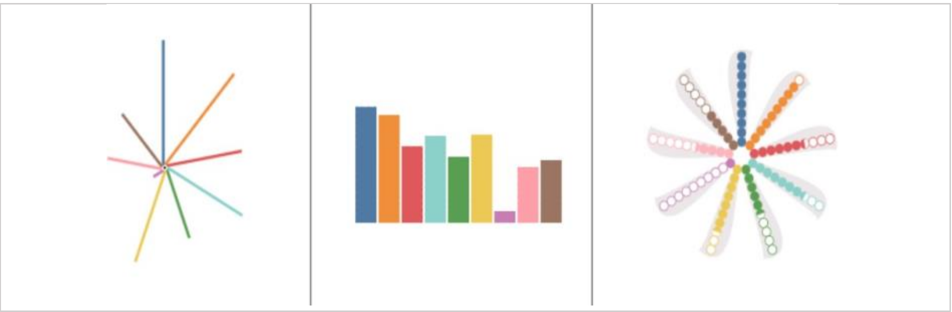


Figure 3-9 The glyphs used in follow-up study, from left to right: Whisker, Bar, PeaGlyph, depicting the same data attributes. (The dataset was used in the visual designs [126])

Results. 20 participants were recruited from various domains. 9 of them were females, and 11 were males. The analysis was performed in two stages. First, the performance of the three glyphs was considered, as this was the primary research question. To check for normality, a Shapiro-Wilk test was run on each distribution. Since the data was not always normally distributed, a non-parametric Friedman's test was used with a standard statistical level of $p < 0.05$ to determine the statistical significance between conditions.

	Accuracy	Response time (ms)
PeaGlyph	[3.44, 4.35]	[26160, 44920]
Whisker	[2.93, 3.86]	[19192, 49346]
Bar	[3.47, 4.32]	[23854, 36730]
The values in the bracket show [min, max] scores of the relevant methods.		

Table 3-6 Showing numerical values of CIs for mean of Accuracy and Response time in Task-1 of Experiment-2.

The Wilcoxon singed-rank test with a Bonferroni correction was used as post-hoc analysis to identify which particular differences between pairs of means were significant. Table 3-7 shows post- hoc p-values for the task in which significant differences for perceived confidence with performance are highlighted in red using $p<0.0167$ following Bonferroni correction.

The participants showed better performance with the PeaGlyph (μ=3.9) than the Whisker (μ=3.4), but the same as the Bar glyph (μ= 3.9) in terms of accuracy. Comparing the response time, Bar performed better concerning the average response time. While users spent less time obtaining the same number of correct answers with Bar, their intervals have a degree of overlap, as seen in Figure 3-10 and Table 3-6. Statistical testing did, however, not confirm significant differences for either accuracy ($X^2(20) = 3.825$, p=0.148) or response time ($X^2(20) = 0.900$, p=0.638). However, the perceived confidence ($X^2(20)$ =13.245, p<0.01) of the PeaGlyph was higher than others, which may indicate that the participants found it easy to use and understand.

	z	ρ
PeaGlyph vs Whisker	4.344	0.000
PeaGlyph vs Bar	3.240	0.004
Whisker vs Bar	-1.453	0.163

Table 3-7 Post-hoc results (z- and p-values) from Wilcoxon signed rank test. Significant differences for Perceived confidence with performance are highlighted in red using p< 0.0167 in Task-1.

Visualization efficiency When calculating efficiency scores for the visualization methods, the most efficient glyph is PeaGlyph with E=-0.125, followed by Bar (E=-0.449) and Whisker (E=- 0.527). This indicates that the cognitive load was lower using PeaGlyph. The results in part support H1, with PeaGlyph being the most efficient glyph design for this task and performing equally good as Bar and better than Whisker in terms of accuracy.

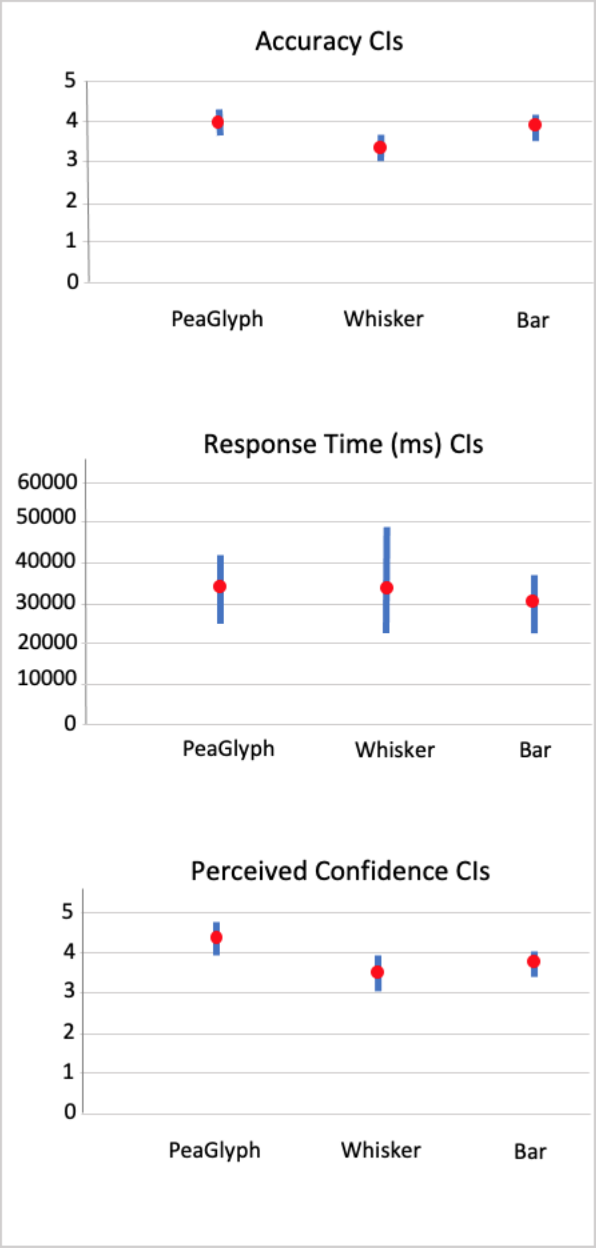


Figure 3-10 Confidence intervals for Task-1. This figure refers to Table 3.6.

3.5.4 *SECOND TASK*

This part was organized the same way as the second task (Task-2) in the first experiment. It aimed to compare the three glyph designs according to their performance under a synoptic task where the overall structure of glyphs is evaluated in terms of ‘balanced’ and ‘unbalanced’. The participants were asked to select the most balanced and unbalanced glyph in the glyph sets given.

Results. All participants in Task-1 finished Task-2 as well. Regarding the number of accurate answers, the participants showed better accuracy performance with PeaGlyph (μ=1.2), followed by Whisker (μ=0.8) and Bar (μ=0.7). Statistical testing did however not confirm significant differences for accuracy (X^2(20) = 0.840, p=0.689).

	z	ρ
PeaGlyph vs Whisker	3.954	0.001
PeaGlyph vs Bar	1.573	0.132
Whisker vs Bar	-4.200	0.000

Table 3-8 Post-hoc results (z- and p-values) from Wilcoxon signed rank test. Significant differences for Response time are highlighted in red using p< 0.0167 in Task-2

Regarding the response time (X^2(20) = 17500, p<0.01), the test confirms significant differences. Table 3-8 shows post-hoc p-values for the task with significant differences for *Response time* highlighted in red, with a Bonferroni correction resulting in a significance level set at p<0.0167. Whisker glyph (p<0.001) did well in terms of response time, and the difference was significant against the PeaGlyph and Bar, while its confidence interval has a small degree of overlap with Bar in Figure 3-11, with Table 3-9.

	Accuracy	Response time (ms)
PeaGlyph	[0.64, 1.76]	[13630, 33612]
Whisker	[0.47, 1.13]	[6818, 13869]
Bar	[0.36, 1.04]	[12867, 27960]
The values in the bracket show [min, max] scores of the relevant methods.		

Table 3-9 Showing numerical values of CIs for mean of Accuracy and Response time in Task-2 of Experiment-2.

In addition, the testing did not confirm significant differences for the perceived confidence (X^2(20) = 0.966, p=0.617). However, Bar had a slightly better confidence rate, followed by PeaGlyph and Whisker. The CIs of the perceived confidence is given in Figure 3-11, where the intervals almost entirely overlap.

Visualization efficiency. Finally, considering their efficiency for Task-2, PeaGlyph obtained a higher score (E= 0.324) and showed higher efficiency and less cognitive load compared to Bar (E= -0.603) and Whisker (E= -1.052). While not supported by any statistical significance, the results in part support H2, with PeaGlyph having a higher efficiency score and higher accuracy.

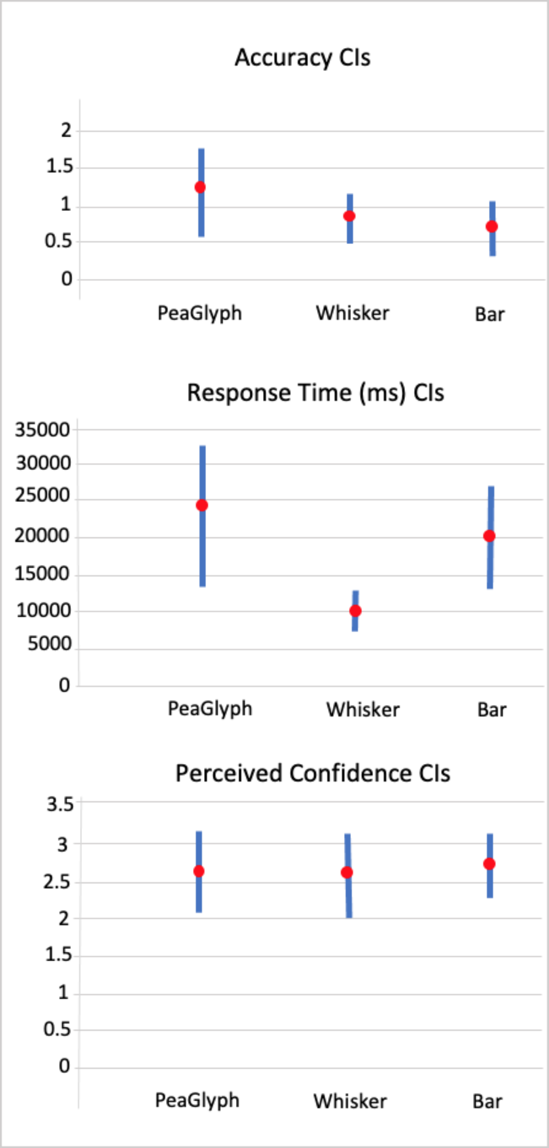


Figure 3-11. Confidence intervals for Task-2. This figure refers to Table 3.9.

3.5.5 *DISCUSSION*

Overall, the number of accurate answers increased with the new PeaGlyph design, although the participants were more familiar with Bar and Whisker. Meanwhile, the users generally felt more confident answering the questions using PeaGlyph. These are promising results and a good indicator of the performance and usability of the new glyph design. With regards to response time, the slightly worse performance may be due to the PeaGlyph being a new visualization method and, thus, possibly requiring more training than Bar and Whisker.

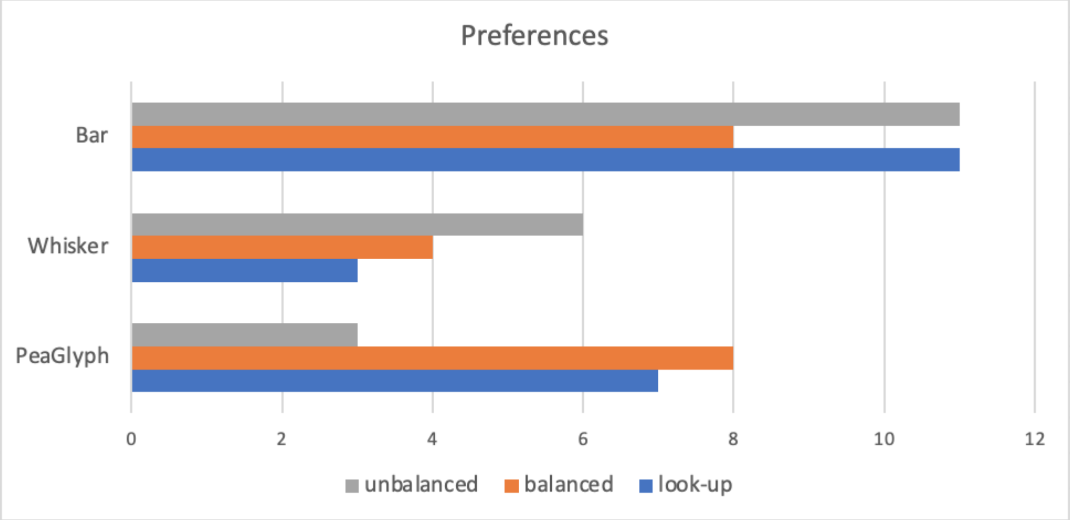


Figure 3-12 Participant preferences for the Task-1 and Task-2 in the Experiment-2

As seen in Figure 3-12, the participants widely preferred the Bar glyph followed by PeaGlyph for the look-up tasks (grey) and Whisker glyph was least preferred. Furthermore, PeaGlyph and Bar glyph were equally preferred for defining balanced structures, while PeaGlyph was least preferred for unbalanced structures. It is worth noting that the users' preferences generally support the quantitative results obtained from the usability tests in Task-2. In order to decide which design would be the most appropriate method for the tasks at hand, two user studies were conducted.

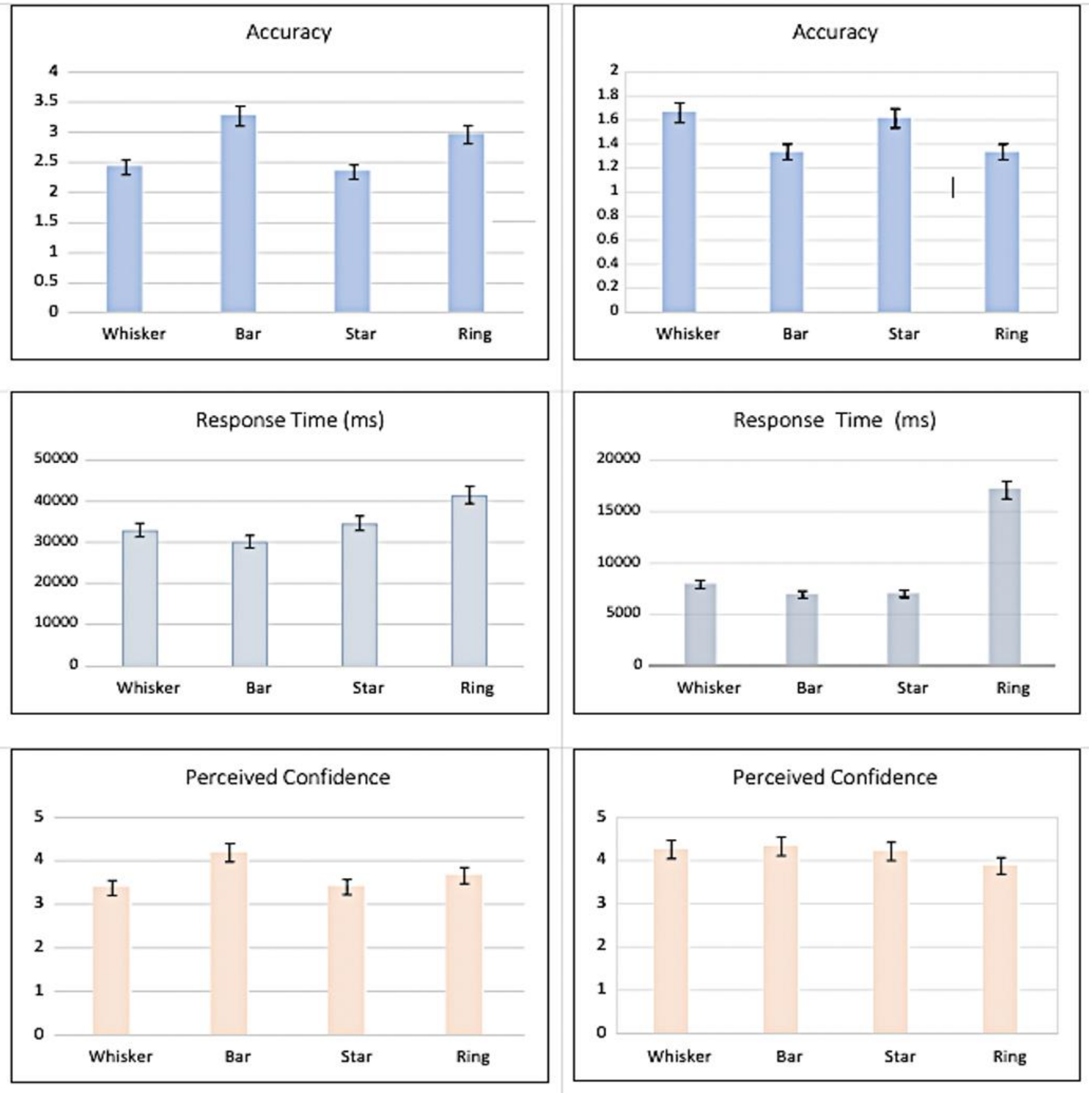


Figure 3-13 Showing Experiment-1 results. The charts illustrate the average accuracy and response time with the standard deviation. The left-side figures showing the results of the look-up task (Task-1), and the right-side showing the results of finding balance/unbalance objects (Task-2)

Figures 3-13 and 3-14 summarize the results from these studies in terms of accuracy and response time. The Bar glyph was the best performer for both experiments for lookup tasks in terms of accuracy and response time. The PeaGlyph showed similar performance as Bar for accuracy, although the response time increased. Compared to whisker glyph, the PeaGlyph had better accuracy, while the response time was more or less the same. For Task-2, the designs were evaluated in terms of how they help reveal balanced or unbalanced objects within formed multivariate groups. In the first evaluation, the Whisker was the best performer in terms of accuracy, although the Bar glyph was slightly better in terms of speed. For the second evaluation, PeaGlyph showed the best performance in terms of accuracy over Whisker and Bar. However, response time was better for Whisker glyph in this task compared to the PeaGlyph and Bar.

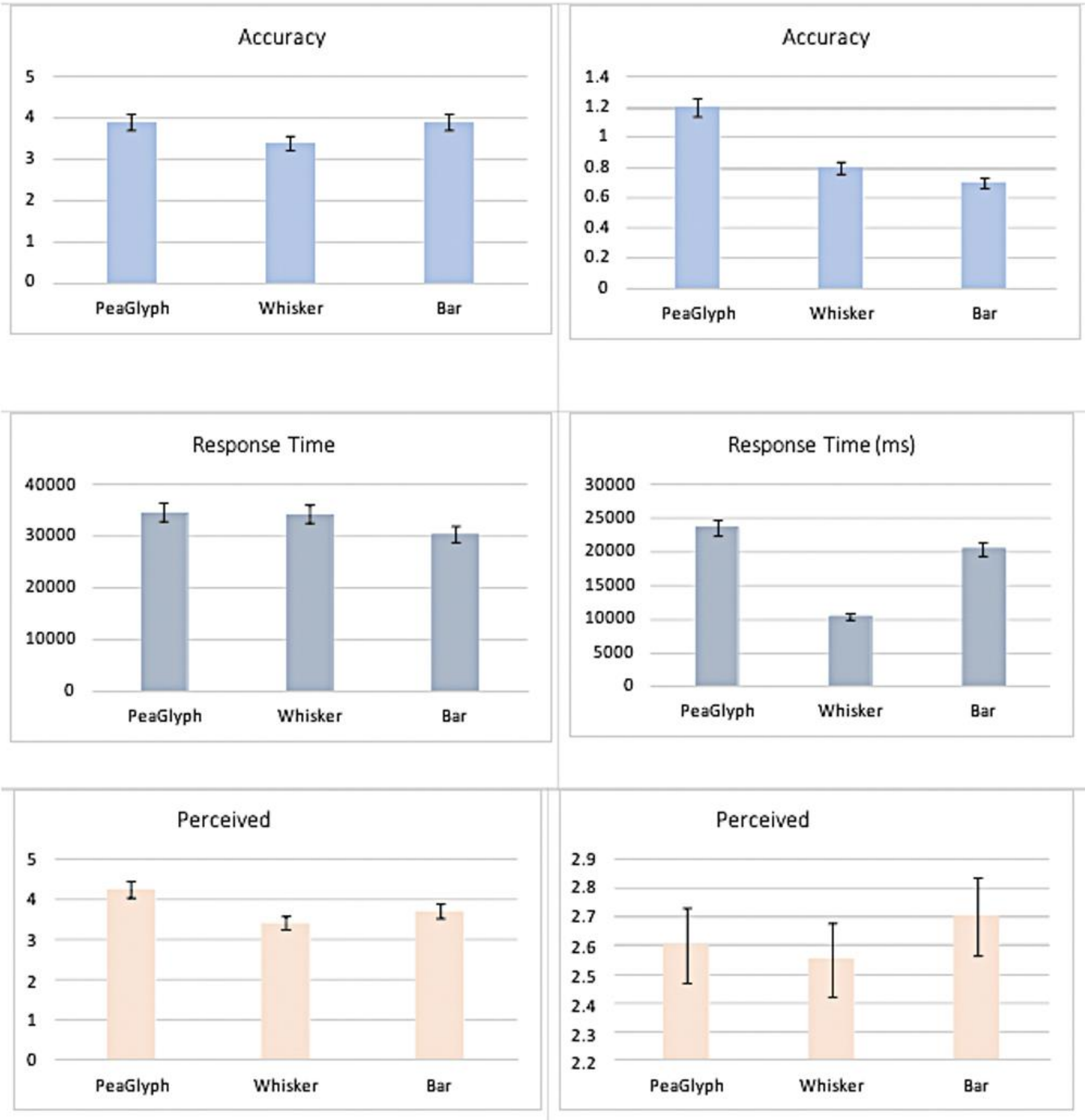


Figure 3-14. Showing Experiment-2 results. The charts illustrate the average accuracy and response time with the standard deviation. The left-side figures showing the results of the look-up task (Task-1), and the right-side showing the results of finding balance/unbalance objects (Task-2)

It is worth noting that in information visualization, the methods are typically evaluated by comparing their differences in accuracy and response time. This situation makes design evaluation difficult in choosing one visualization over another [114]. Thus, we also evaluated the visualization efficiency of the designs. The efficiency results showed that

PeaGlyph was the best option by obtaining the highest score in both tasks in the second experiment. It is worth noting that we want to keep the Experiment -1 and Experiment -2 consistent by not integrating the participant feedback about the glyph designs into study, although some of them emailed their thought after completing the Experiments. Instead, we asked them to answer which glyph design they would prefer in the given tasks, and Figure 3-12 shows their preferences according to given tasks.

In summary, the participants increased the accuracy of their responses using PeaGlyph; however, the task completion time increased marginally. In particular, PeaGlyph can be preferred for tasks where the overall accuracy level is more important rather than the response time. Furthermore, analysts may want to be able to interrogate data with a more complex nature. PeaGlyph comes with various visual channels that can be used to encode more complex data sets (see section ‘PeaGlyph Design’). We are aware that some channels are more efficient than others, so the visual designer can select among them to encode their data set.

3.6 USE CASE STUDY

This section demonstrates how the presented design(PeaGlyph) can be used to analyse the balanced and unbalanced within and between data attributes, as well as comparing data values. For this aim, A Better Life Index (*BLI1*) dataset was selected. As there will be groups of countries that are similar to each other and clearly different to other groups of countries, and as such it is a good representative for grouping and visualization of balanced/unbalanced structures. *BLI1* includes multiple dimensions of the well-being of the OECD’s member countries. The OECD selected 11 indications of life to measure, which are the availability of housing, income, and work as well as the overall quality of life (community, education, environment, governance, health, life satisfaction, security, and work-life balance) - the selection of dimensions is explained by [120] in detail. All of the indicators are normalised to the range of 0 to 1. These variables make it possible to compare its member counties from many different perspectives. For the 2014 OECD Better Life Index, Decancq [127] introduced the distributional Better Life Index (BLI2), and countries are ranked according to their BLI2 scores representing each country’s overall well-being, taking the distribution of well-being into account. In Figure 3-15, the countries are mapped onto two-dimensional space based on their BLI2 loss (i.e. A country with a larger BLI2 loss due to multidimensional inequality will have a smaller BLI2 score). The countries on the lower end of the figure have a smaller loss compared to the countries on the upper side. For example, the countries such as Austria and Netherlands have higher average scores for the various dimensions of life. To make it clear, the PeaGlyph depicted the various life attributes of three of the countries selected from different areas of the projection, as seen in Figure 3-16.

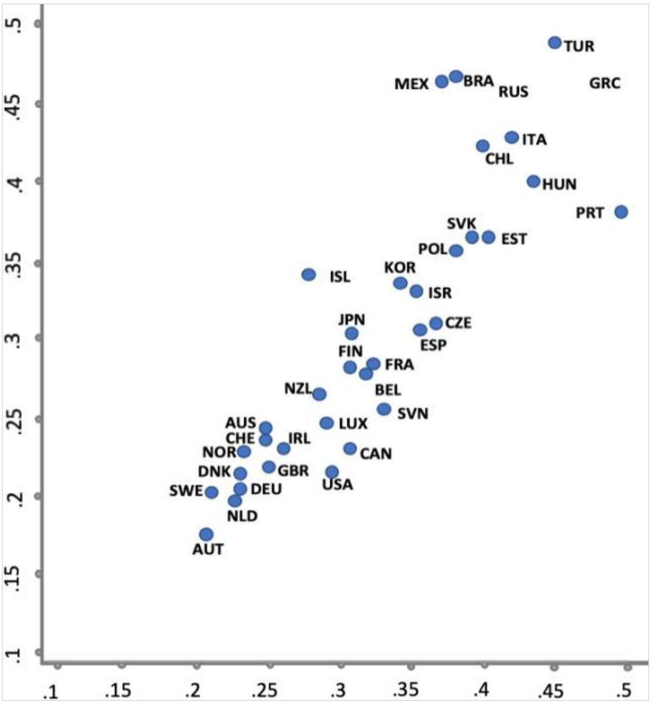


Figure 3-15 BLI2 loss due to inequity. Source: OECD(2014), adapted from Decancq [127]

The countries encoded into the PeaGlyph can be compared to see which ’life indicators’ (i.e. data attributes in our context) are more distinctive or similar across countries. PeaGlyph helps analysts identify the numerical value or ratio represented by each variable by counting the filled, semi-filled or empty circles, allowing for easy matching of multiple attributes across countries. For example, in Figure 3-16, the user can quickly detect that the ‘job (J)’ and ‘income (I)’ of the country glyph (TUR) are lower than these values for the middle (ISR) and bottom glyphs (SWE). However, the same country (TUR) has higher values than for both ’Housing (H)’ and ‘Work-life balance (W)’ of the ISR and SWE.

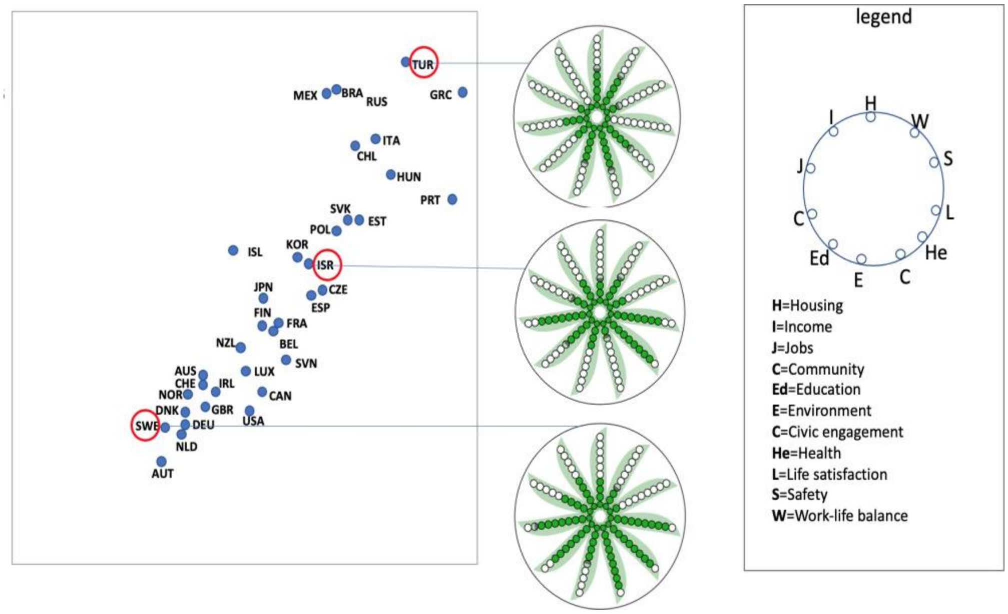


Figure 3-16 Comparing the multi- attributes across the countries in interest

Further, users can analyse the balance or imbalance between multiple attributes of each country, as shown in Figure 3-17. For this purpose, three of the projected countries were randomly selected, and their attributes were mapped to the PeaGlyph. Two versions of the glyph are displayed in the figure to demonstrate the flexibility of the design. The polygonal overlay structure on top of the PeaGlyph in the top part of Figure 3-17 serves to further emphasize the balance of attributes. The shape of the overlay is based on the underlying pea-based representation. The country glyph (PRT) at the top is perceived to be more balanced than the other country glyphs in the middle and bottom in terms of the distribution of its variables among themselves, as visible from both the top and bottom part of Figure 3-17. Also, the balance in (AUT) appears to be greater than that of the middle country (NZL).

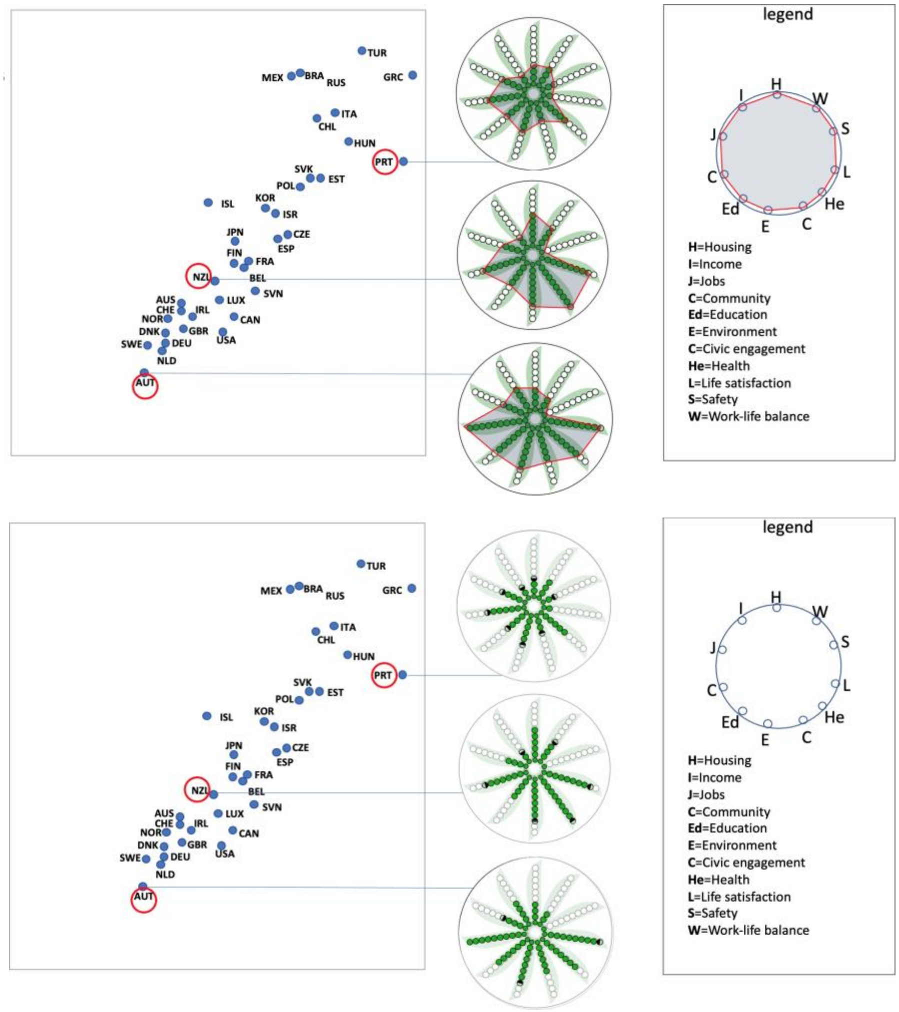


Figure 3-17 Representing balance/unbalance within the multi attributes of three countries of interest, using two design variations of the PeaGlyph. The top figure includes a polygonal overlay on the Glyphs, while the bottom figure fades out the empty peas.

Due to the strength of human visual perception, this type of comparison can be easily performed for moderately sized data using the PeaGlyph. Alternatively, it is necessary to consider the range and distribution together to decide on the balance or imbalance between the variables, which may be more cognitively demanding.

3.7 CONCLUSION

Glyph-based visualization is a common form of visual design that has attained great attention from researchers in the visualization domain. This chapter presents a new glyph design, PeaGlyph, which was designed based on established design principles and the results of a formal evaluation of four glyph designs. The PeaGlyph design aims to address the problems related to existing glyphs, as identified in the evaluation. The PeaGlyph is described in the paper, along with the introduction of design schemes for alternative usage and data types. Two usability studies are presented in this chapter. These compare glyph designs for tasks of relevance for investigating structures of balance and unbalance in data. For the first study, four glyph methods were compared: Bar and Whisker as the baseline glyphs, and their variations Star and Ring. The study compared these four glyphs in terms of look-up and comparison of values and revealing balanced objects. The performance of the novel PeaGlyph was then evaluated against the best performers of the first study, Bar and Whisker. The results were encouraging, and the PeaGlyph performed as well as Bar glyph in terms of accuracy, although the response time for the new design was higher than Bar. These results may be affected by the PeaGlyph being a new visualization method and, thus, possibly requiring more training to use efficiently compared to Bar.

Furthermore, the participants reported in their feedback a preference for the PeaGlyph and Bar over Whisker glyph for the look-up tasks. Whisker was the worst option for the studied balanced detection tasks. These results can be considered a good indicator of the potential of PeaGlyph as an intuitive data glyph. Finally, the utility of PeaGlyph is demonstrated through a use case example. Future work includes further evaluation of the glyph and additional use case testing in more realistic settings. Due to its visual simplicity and flexibility, the PeaGlyph has the potential to be used for a variety of scenarios. The future studies considered include cluster analysis and the formation of comparatively balanced teams [128], which will be presented in Chapter 6 and Chapter 7, respectively, in this thesis. The latter is particularly promising since the PeaGlyph has the ability to map summary attributes of groups as well as represent the balance across the attributes for each group. This allows facilitators to gain insight into the teams and change team members as needed to create the desired balance between teams. There is also a need to formally evaluate the number of data attributes that can be efficiently represented using the PeaGlyph design and the effect of the colour scale on the results.

Chapter 4 : GROUPVIS: AN OVERVIEW OF THE GROUP FORMATION FRAMEWORK

In this chapter, GroupVis, a group formation framework to make the grouping process in systematic way, is introduced in the abstract. Furthermore, a set of criteria was defined based on the findings in the literature to inform behind the framework design. The chapters that follow this chapter detail the main components of the GroupVis listed here, respectively.

4.1 INTRODUCTION

Teamwork is a matter of skills and cooperation in which the goal is mutual learning and knowledge sharing. Group formation is an essential process for the group development lifecycle. It is not uncommon for distinct terminology to be used in literature to refer to the ideas of group and team. A team can be defined as a group of people assembled around a common goal, the accomplishment of which requires a variety of forms of agreement and cooperation. In a variety of organisational settings, the ability to operate in a team has become a critical component of recruitment and hiring [8]. As Lockyer and Gordon stated the success or failure of the project organization and the quality and reliability of its product depends on the competency of its people [129].

Collaborative learning is an educational strategy that enables students how to work in teams and acquire subject-matter knowledge [93]. It becomes critical to identify characteristics of group construction that influence the effectiveness of individual and collaborative learning [130]. Additionally, the composition and formation of successful teams is critical for businesses to maintain their competitiveness [131].

Group formation is the process of allocating participants to groups and roles, which is not an easy undertaking. The instructor must select how to fill the groups. Assignments can be made by participants, moderators, or the system, and are based on parameters such as participant traits, educational level, and competence. A random selection of participants may result in an uneven group composition, therefore efficient groups may be improbable to form. There are features that have to be taken into account when creating effective study groups so that they can achieve the desired targets. Maqtary et al. [8] defined group formation as an atomic process affected by several factors. The authors categorized the factors into two categories: member and group attributes. The group attributes describe group characteristics such as homogeneous and heterogeneous or mixture as a whole, whereas member properties such as subject knowledge, learning skills describe the people in groups. When all of these aspects are considered, proposing an effective and pedagogically acceptable group formation becomes a challenging matter [132].

Computer-based approaches have been provided in the group formation context to make this process effectively. Among used techniques, evolutionary algorithms have been utilized most frequently, then followed by machine learning methods [8]. Besides, these studies leveraged a fixed number of attributes to form desired groups. The most common chosen attribute in these studies was knowledge [132]. Learning styles and personality traits are other commonly used attributes influencing group outcomes.

Unlike existing methods, this research presented here considers visualization solutions as an aid to support group formation tasks. To our best knowledge, this is the first attempt to form balanced groups using a visual-based approach. We believe that visually depicting the diversity within groups and group balance can be employed in conjunction with existing methodologies as a complementary and adaptable approach. The study presents a framework that includes three well-defined components to make the group formation process more systematic, supported by humans in the loop design. Each component is organized according to the determined workflow, and each includes a number of visual and computation methods to address the grouping problem. The framework is used in forming project groups in the educational domain and evaluated by some domain experts.

The contributions of this work are summarised as follows:

- Providing a novel and modular framework bridging human, machine learning and visualization methods to form balanced groups in a systematic way, supported by a semi-automated approach.
- Proposing a new way of using teaching and learning data sets through visual analytics
- Design guidelines that can guide the development of a teacher tool were obtained through semi-structured interviews and heuristics.

The following sections will discuss the visual methods and the computation methods in the framework presented to aid the grouping problem in education context.

4.2 RELATED WORK

This section is divided into two subsections to examine the feature space and the methods used in group formation in the literature. The first subsection highlights the attributes used to create the groups. In the next subsection, the computational methods commonly used for grouping in the literature are presented. These methods are grouped under four categories. Afterwards, the papers falling into these categories are listed in order of their publication date.

4.2.1 *ATTRIBUTE SPACE OF GROUP FORMATION*

Computer-assisted grouping is an automated process in which students are grouped based on data and constraints provided by the teacher (e.g., type and size of learning group) [8]. Belbin's role theory [130, 133] (or Belbin's role balance) is often used as an instrument in this context and states that balanced groups, where all roles are present, perform better and have a positive impact on the quality of teamwork compared to unbalanced groups. The Felder-Silverman model [134] is another approach that generates four dimensions of learning styles. Besides them, clustering algorithms for group formation [135] are perfectly suited to group formation problems since group formation attempts to split the student data into multiple groups. Balance is achieved on the basis of a set of specific criteria that apply to the learning objective in question. In other words, the desired balance in terms of diversity can be achieved by including students with different levels of achievement and characteristics in each group. Also, among all formed groups, the distribution of the attributes (performance level) of learners should be as similar as possible for forming competitively balanced teams [128]. Computer-supported collaborative learning (CSCL), which creates a collaborative learning environment [136], makes use of networking technology to facilitate social and instructional interaction between learners in small groups and learning communities. CSCL environment offers a variety of functions, including group creation, monitoring, and managing activities such as evaluation [137]. The studies on efficient group formation are surrounded by the following research purposes: determining the most suitable features to be used in creating groups, developing techniques that will optimize the group creation process, or transferring new methods from different domains to create groups. The current study falls under the category of transfer of new methods since it applies an information visualization approach to the group formation problem.

4.2.2 *COMPUTATIONAL APPROACHES TO GROUP FORMATION*

Group formation as an application problem has been handled in many ways, as shown in Table 4-1, which indicates that genetic algorithms and machine learning methods are more commonly used as part of group formation compared to other approaches.

A mathematical approach was proposed by Graf and Bekele [138] to maximize the heterogeneity of generated groups that were created based on the characters and performance of students, which are '*group work attitude, interest for the subject, achievement motivation, self-confidence, shyness, level of performance in the subject, and fluency in the language of instruction'*. The authors collected the student data records and tested their proposed algorithm for this specific dataset. According to Wi et al. [139], having a good leader in the role of team manager, as well as having competent workers collaborate as team members, is critical to the success of an institution's business activities. They provided a framework for analysing the knowledge of all applicants for team formation and utilised a genetic algorithm and social network measure to choose a group manager and team members from a pool of candidates in their work.

	Methods			
Literature	Genetic algorithms	Ontologies (i.e. semantic ontologies)	Machine learning algorithms (i.e. clustering)	Other methods (i.e. ant colony, generic systems etc.)
Graf and Bekele[138]	■			
Christodoulopoulos and Papanikolaou [140]			■	
Ounnas et al. [141]		■		
Isotani et al. [132]		■		
Wi et al.[139]	■			
Craig et al. [142]				■
Strnad and Guid [143]	■			
Abnar et al.[144]	■			
Moreno et al.[145]	■			
Arias-Báez et al.[146]				■
Srba and Bielikova [147]			■	
Zheng and Pinkwart [148]				■
Akbar et al.[149]			■	

Table 4-1 Several computational methods extracted from relevant studies in group formation. The computational methods commonly used in group formation are grouped under four categories. The papers falling into these categories were listed in order of publication date from oldest to newest.

Christodoulopoulos and Papanikolaou [140] have used an algorithmic approach to form homogeneous and heterogeneous groups, and their experiment focused on the usage of low complexity algorithms. Isotani et al. [132] proposed an ontology engineering solution to the problem of group creation in their paper. The technique makes use of ontology to express collaborative learning and the process that occurs throughout the learning. They used learning theories in the ontology framework to provide support for making pedagogical decisions, like creating project teams in a learning environment. Craig et al. [142] developed a mathematical model that included person attributes, group-formation criteria and fitness metrics that allowed them to generate reasonably optimal groups in accordance with the instructor’s requirements. Strnad and Guid [143] proposed a novel fuzzy-genetic analytic model for the problem of team building. Their suggested methodology obtains fuzzy values for specific features using an employee database that contains information about the employee's experiences and specialisations. These attributes(values) were used to form project teams. Arias-Báez et al.[146] presented a generic system that aims to create teams based on grouping criteria, including personal characteristics and knowledge about relevant collaboration capabilities and context. In another study by Akbar et al. [149], topic selection is combined with team building; students are assigned to a topic in which they are interested or to a team of students who share their interest. They conducted their study using a hierarchical k-means technique.

Among the methods described, probabilistic models like genetic algorithms are more commonly used to form groups in collaborative learning environments as such algorithms are capable of effectively generating groups from a huge number of variables [82]. The computational methods are quite complicated, and their decision-making process is not interpretable. In addition, it is widely acknowledged that assembling an effective team is not always doable. The presence of numerous criteria and the intricacy of their combination necessitates a lengthy process, and the formation of functional teams is sometimes not assured [89]. The study presented here aims to meet this gap by dividing the problem into well-defined components and utilizing information visualization methods in these components to make the process explainable for target users.

4.3 SYSTEM DETAILS

The present study offers a framework for group formation in order to facilitate the process of forming balanced teams as well as to address its relevant problems and also make this procedure more systematic and less biased, as illustrated in Figure 4-1.

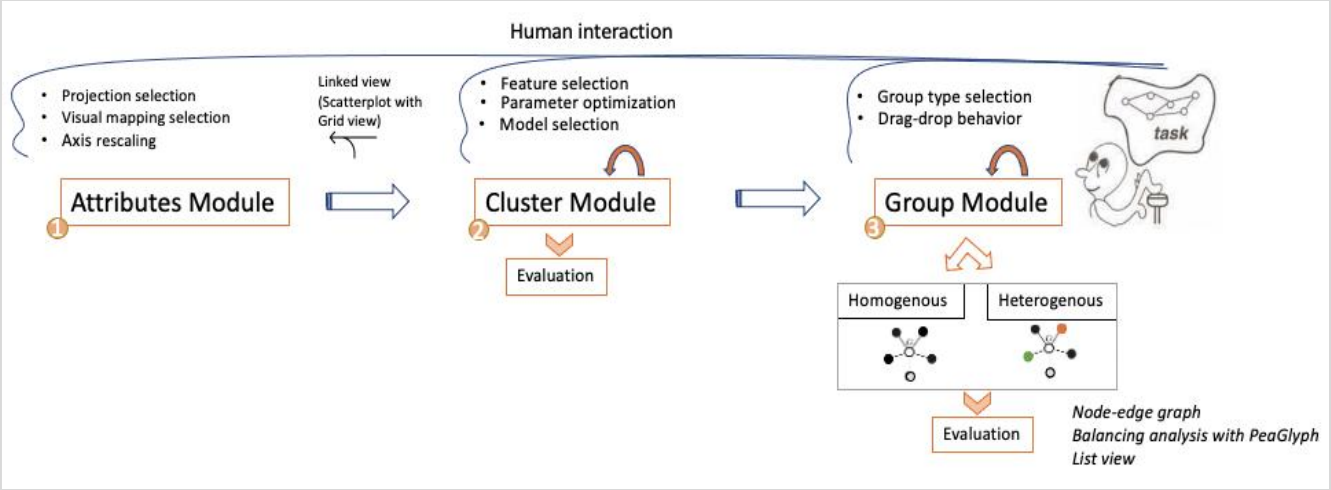


Figure 4-1 Proposed framework for group formation problem

The overall task of grouping is formulated as three well-defined sub-tasks, namely the modules: Attribute$^{(1)}$ (Chapter V), Cluster$^{(2)}$ (Chapter VI), Group$^{(3)}$ (Chapter VII), and each of them implements a set of interactive visual representations and computational methods, alongside some interactive methods, which are chosen based on underlying tasks. The detail of the framework is described in the following sections.

The system architecture is given in Figure 4-2. In the system, JavaScript is used for data preparation and implementation of the algorithms and the metrics mentioned above. The front end communicates with generated data.

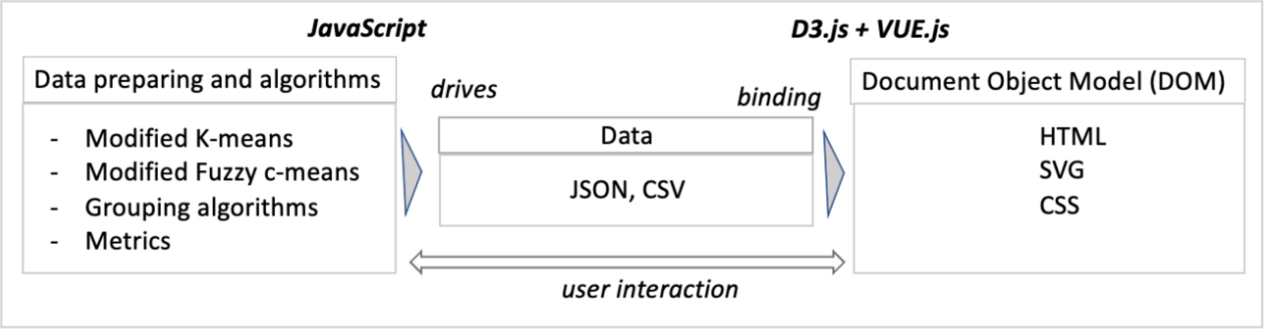


Figure 4-2 The proposed system architecture

D3.js binds the data to DOM elements and renders dynamic visualizations in web browsers. VUE.js is used for building user interfaces. In this system, users can interact with the model itself through changing feature sets or selecting a different clustering number (k), which leads to obtaining different clustering. Based on the changes made, the groups is regenerated to reflect the changes in clustering outcomes. Users can also compare the results obtained from given clustering algorithms in this framework.

4.4 GROUPVIS FRAMEWORK

This section described the criteria that guided our design of the GroupVis. It also presents the technical details of each component that collectively builds the framework.

4.4.1 *DESIGN REQUIREMENTS OF GROUPVIS*

This study treats visualization solutions as an aid to support group building tasks and aims to make the grouping process more systematic and visually explorable to achieve the desired type of groups. This also helps the target users understand and evaluate algorithmically generated decisions (i.e. groups).

It is necessary for the GroupVis to meet a variety of requirements that were defined in connection with the research objects of this thesis, as follows:

C1: The system should present a chained sequence of tasks in well-defined components where the output of a component can become the input to the next component (as seen in Figure 4-1).

C2: The visualization system should facilitate rapid information seeking, comparing of possible outputs and decision making for educators/lecturers at different granularities of the group formation process

C3: The target users should be able to incorporate domain knowledge into the input and/or output of automated methods to derive desired groups for the tasks at hand.

C4: It is necessary for the system to highlight problematically formed groups to help the user identify them and make adjustments to them.

C5: The visualizations should be intuitive for the intended users and need little learning and memorization (will be discussed in Chapter 8: Heuristics Evaluation of GroupVis).

To summarise, this chapter provided an introduction to the GroupVis module as well as an outline of the five primary criteria upon which it is founded. The following sections will discuss the three analytical modules in the order they are used in the group building process. Also, how the above listed requirements are handled will be presented in the following chapters.

Chapter 5 : EXAMINATION OF DATA ATTRIBUTES THROUGH MULTIPLE PROJECTION

The role of this module in the framework is to provide initial understanding of the whole data space to our target users. In a teaching and learning context, it could be around assisting lecturers/educators in understanding the knowledge space and diversity of student cohorts. The module components have been explained in the following subsections.

5.1 INTRODUCTION

The Attribute module in Figure 5-1 aims to help users explore data space with several visualizations, including Scatterplot view and Grid view, alongside common projection approaches multidimensional scaling, t-SNE, as well as Self-organizing maps.

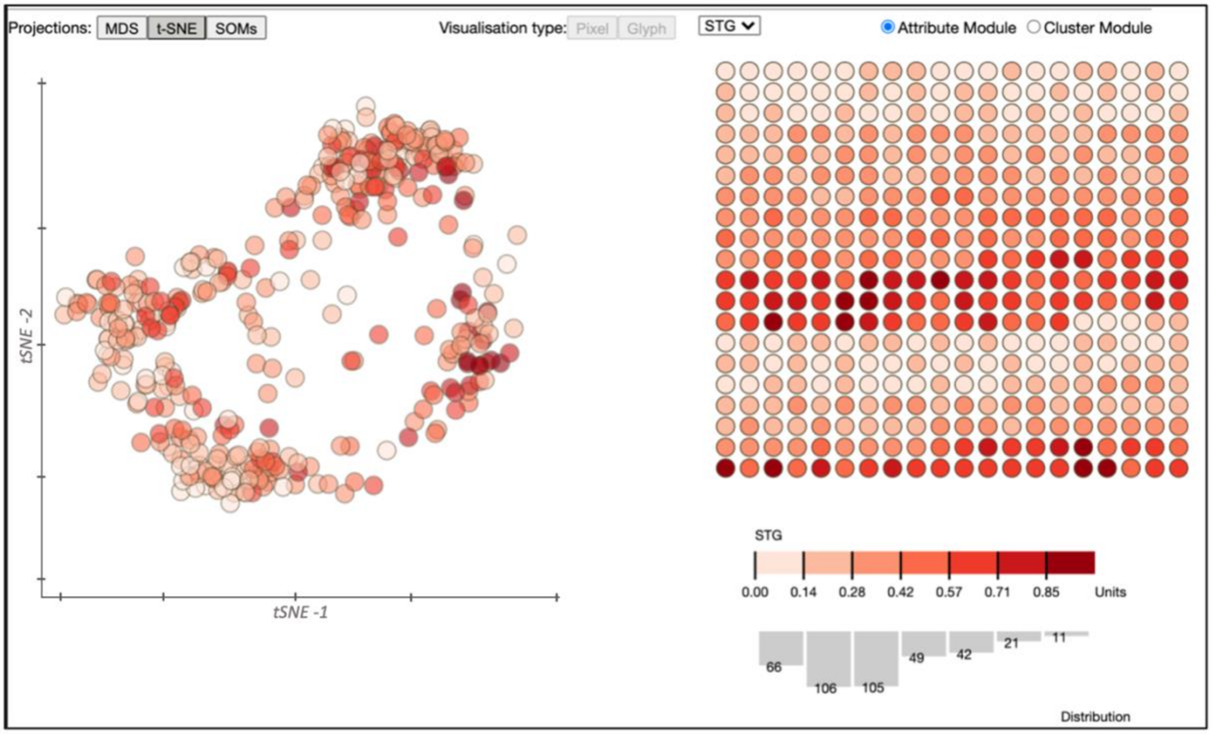

Projections: MDS t-SNE SOMs
Visualisation type: Pixel Glyph
STG
● Attribute Module ○ Cluster Module
tSNE -2
tSNE -1
STG
0.00
0.14
0.28
0.42
0.57
0.71
0.85
Units
66
106
105
49
42
21
11
Distribution

Figure 5-1 An overview of Attribute module that includes a Scatterplot and Grid view with the colour legend and the histogram representing the distribution of data instances

The following sections will describe the methods mentioned above in details.

5.2 SCATTERPLOTS

The Attribute view starts with a scatterplot, as seen in Figure 5-1. It is a versatile technique for displaying correlation on pairing axes and patterns of low dimensional data [150], as well as providing a high-level view of a huge amount of data [151]. Due to its adaptability, it has been used in various fields to present and explore relevant content [54]. Three frequently used scatterplot tasks are to select a subset of two dimensions, reduce the dimensionality to two dimensions using a dimensionality reduction approach, or display all dimensions pairwise. In the scatterplot, each data instance is encoded as a point, and their colours changes depending on the values of the selected attribute from the given drop-list widget. The Attribute module is designed to help explore data space and identify naturally formed clusters. For identifying data clusters, scatterplots have been found to have advantages over parallel coordinates [152]. Furthermore, projecting high-dimensional data into 2D scatterplots helps visually identify cluster patterns. It is worth noting that a scatterplot matrix may not efficiently visualize the complete attribute space, so it often provides useful information when pairwise dependencies are what is of greatest interest [153]. The projection approaches used in Attribute module of the GroupVis are described further in section 5.4 below. When increasing data size, cluttering is a common problem in the Scatter encodings, so to mitigate this case, the Grid-based visualization was placed adjacent to the scatterplot.

5.3 GRID-BASED VISUALIZATION

Unlike the Scatterplot, the Grid view is a grid full of circles (i.e. the grid layout uses circles for the grid cells) and represents the data instances of the Scatterplot in an organized manner. The order of grid circles is in the order of the data in its natural state (i.e. any point ordering is not specified for this tool prototype). The legend takes a scale based on the underlying data and derives a set of equally sized bins (or bands). The legend's colour scheme uses an array of light-to-dark reds, and darker colours represent higher attribute values, while lighter colours represent lower attribute values. Depending on a selected attribute, the colour of the grid circles is associated with the colour of the corresponding bins on the legend. In this module, the scatterplot and grid circles use the same colour scheme. Along with this colour-based legend, a histogram shows the frequency of instances falling in these bands. The legend is interactive, and legend-over behaviour emphasizes circle that fall inside a certain band while transparently de-emphasizing circles that do not fall within that band. As illustrated in Figure 5-1, the legend title, unit labels, suitable numerical formatting, and additional graphical elements utilised to highlight the breakpoints all contribute to the legend’s readability.

5.4 2D PROJECTIONS OF DIMENSIONALITY REDUCTION METHODS

The GroupVis supports common projection methods, namely Multidimensional scaling, t-SNE and Self-organising maps (SOM). In general, those methods are all aimed at lowering the number of attributes in a dataset while retaining as much variance as possible in the original dataset; however, they use distinct approaches to mapping [18, 19]. This prototype of GroupVis implemented the methods mentioned above; however, different projections methods could be added alongside them. In many cases, clusters exist in the lower-dimensional subspace of the original dimensions. The reduction of data dimension thanks to these methods reasonably preserves patterns and clusters in the data set.

In the current tool, MDS and t-SNE methods were implemented for enabling target users to analyse clustering results on the scatterplots. Also these methods preserve data structures within a large number of attributes when projecting onto 2D and 3D. MDS retains pair distance information in low-dimensional data space [43]. MDS can be categorized into classical MDS, metric MDS, and non-metric MDS. In this prototype, metric MDS has been implemented, and it tries to preserve the distance of points in the embedding space, so in this method, the difference in distances of points in the input and in the embedding space. The metric MDS uses the distance metric in its optimization, and it minimizes the cost function called ‘Stress’. t-SNE is a non-linear dimension reduction approach for effectively separating clusters, which is the problem at hand. Much of the local structure of the high-dimensional data may be captured effectively by t-SNE while also displaying global structure such as the presence of clusters on a variety of different scales.

As a result, t-SNE tends to produce a representation with discrete clusters for clusterable data, which are often in accord with the clusters formed by a devoted clustering technique. In the implementation of t-SNE for this current tool (system), the learning rate (i.e. epsilon) is set to 10, and roughly how many neighbours each point influences (i.e. perplexity) is set to 30, and iteration number is 1000, as seen Table 5-1.

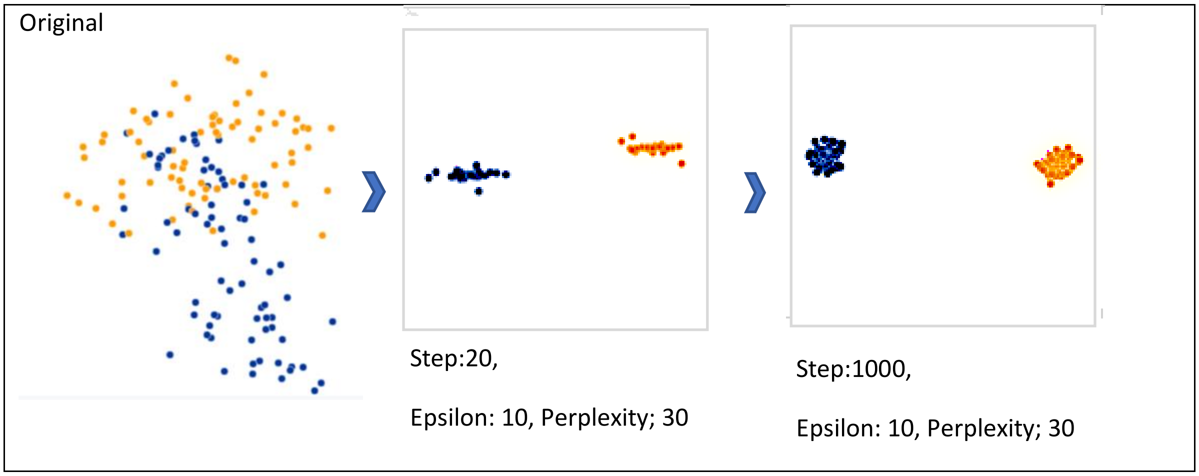


Table 5-1 The illustration of t-SNE with different settings. The dataset used was generated synthetically by drawing random samples from a multivariate normal distribution, and includes two classes.

In the implementation of t-SNE, the cost function was made by minimizing the Kullback-Leibler divergence (KL) (Equation -2) between the low dimensional (Q) and high-dimensional similarity (P) distributions, using a gradient-descent method. P and Q are two similar matrices. The measure of pairwise similarities in the high dimensional space is represented by the conditional probability $p_{j|I}$, *(i, j =1,2,..,N)* while the measure of pairwise similarities in the low dimensional space is the measure of pairwise similarities uses a student t-distribution (q_{ij}).

The cost function in t-SNE [42]:

Equation -2.

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \ log \frac{p_{ij}}{q_{ij}}$$

It is worth noting that the global geometry requires fine-tuning perplexity, so to deal with this problem and help users capture a good sense of clusters in the t-SNE scatterplot, the quantity metrics for t-SNE scatterplots were provided so users can choose the graph that shows the isolated clusters better among others, and continue their analysis through the selected one.

The tool enables users to pick a ‘good’ scatterplot view, in Figure 5-2, in which separated groups are more visible among generated scatterplots based on the data at hand through semi-automated guidance *(this meets **C4** on page 59).* The semi-automated guidance works on two metrics. Distance consistency metric (DSC) [59] (Equation -3) quantifies the proportion of data points with a violation of the centroid distance (CD). The centroid distance is the distance between a cluster member and its centroid, and it should be as small as possible in relation to the distance between all other centroids.

Equation -3.

$$\text{Distance consistency (DSC)} = \frac{\left|\mathrm{x} \in v(X) : CD\ (x, cent(c_x)) \neq true\right|}{\mathrm{k}}$$

x is a data point in the 2D projection and, $cent(c_x)$ is the 2D projection of the centroid of class c_x. The result is normalized to a score between 0 and 100.

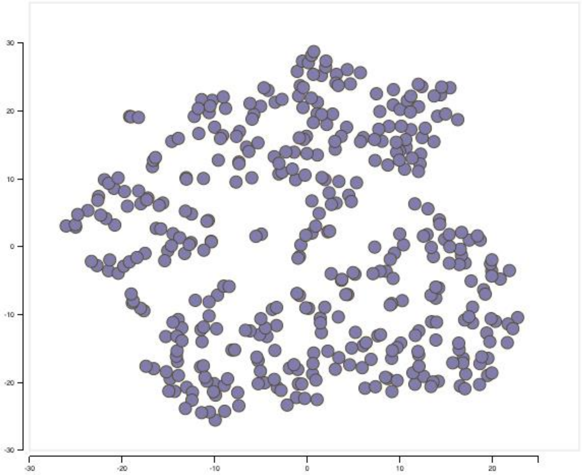
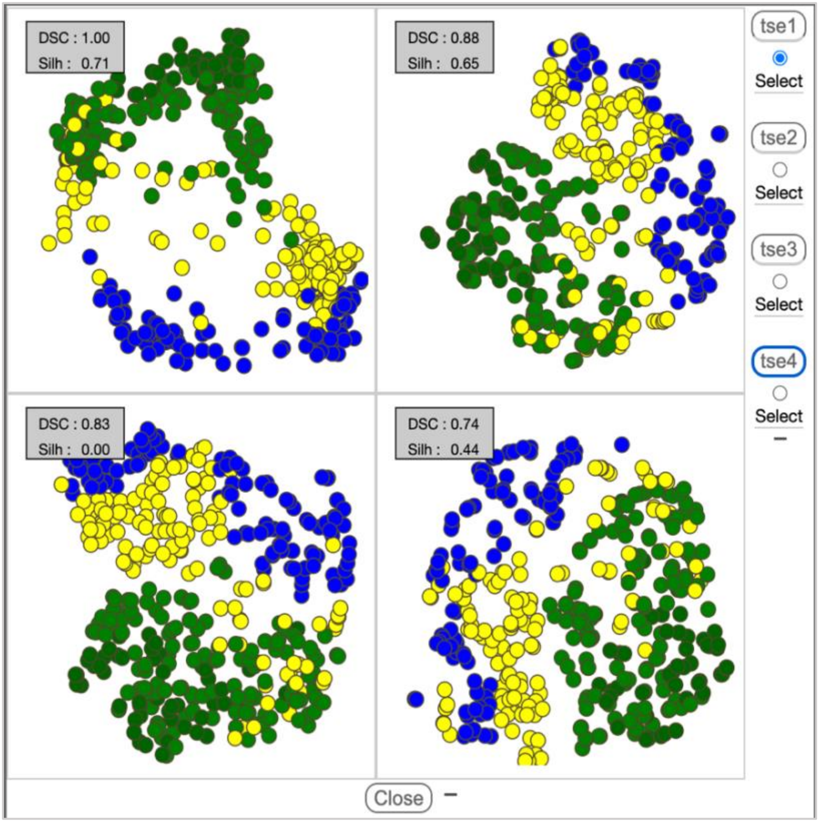


Figure 5-2 Multidimensional scaling scatterplot of original *unclassified* dataset. User Knowledge Modelling dataset [44] was used. (Source: Author's own)



DSC : 1.00
Silh : 0.71

DSC : 0.88
Silh : 0.65

DSC : 0.83
Silh : 0.00

DSC : 0.74
Silh : 0.44

Figure 5-3 Semi-automated visual guidance that displays the top four plots regarding cluster-preservation. The classified data (User Knowledge Modelling dataset [44]) was used and this dataset includes 4 classes (Very low, Low, Medium, High), which label students' overall performance over the given learning subjects. We used the colour scheme to depict the classes in the dataset- (Each points encodes a student). The clusters in the dataset are more clearly visible in the upper right figure when compared to the others so it obtains higher metrics scores (i.e. DSC and Silhouette coefficient). (Source: Author's own)

The second metric used is the silhouette coefficient, which was originally introduced for the purpose of evaluating clustering algorithms, and it quantifies both cohesion and separation between clustered instances [154]. The silhouette has a range of values between [-1, 1], and the larger value of the Silhouette means a better cohesion and separation. The technical details of silhouette coefficient will be discussed in Chapter 6. The guidance, as shown in Figure 5-3, displays the top four plots having visible useful partitions or clusters based on these metric scores, and users can select one from among them to use in their further analysis process. The number of presented plots was kept low (limited to 4 plots) so that it was aimed to make it easier for users to choose among them.

In the current implementation, Self-organizing map is mainly used as a projection method for multivariate data sets. The SOM is represented as a square grid, where each grid cell can represent several data items, with similar cells grouped together towards one area, and colouring is used to reveal global distance structures as well as local neighbourhood relations [155]. The authors in the study [156] employed Principal Component Analysis (PCA) to obtain three principal components of their multivariate data and they then mixed the components values into RGB colour channels with a linear function. Our approach follows their study by using the PCA approach and pairing the units with colour space. The code vectors (same dimension as the input vectors) of each node mi are used to derive the first principal component, whose values are continuous. These values are then mapped to an output range, which is determined by a custom interpolator function. The *viridis* colour scale, in Figure 5-4, was used as a primary option in the current tool, as the colour scale is perceptually uniform in hue and brightness [157]. Alternative scales with similar properties can be used to encode SOM nodes. In the scale, the higher values are encoded into darker colours, and the lower ones are in lighter colours. The advantage of our approach is that it is easy to be customized to the purpose of use.

sequentialScale =
d3.*scaleSequantial*(d3.interpolateViridis).domain([firstPrincipalComponents])

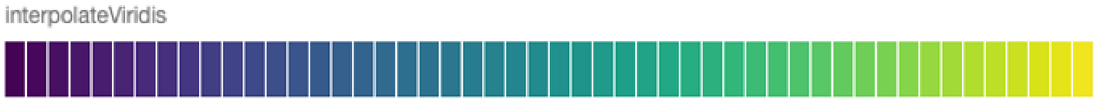


Figure 5-4 The scheme of *viridis* colour scale

Deciding the optimum grid size of SOM is also challenging as different grid sizes lead to different map presentations of the same data set. A thumb rule proposed by Vesanto and Alhoniemi [158] is that the optimal size of SOM is calculated (Equation -4) as follows:

Equation -4.

S = 5 $\sqrt{N}$, where N denotes the size of data samples

However, it is likely to obtain empty nodes as the size of SOM will grow as well when the number of data samples is huge. That causes the degradation of the accuracy with a decrease in interpretability. The issue was addressed by Shalaginov and Franke [159] by taking the statistical properties of the data set, such as variables and coefficients, into account. However, this leads to a smaller grid size of SOM for small or moderate data sets, which is likely to prevent the complete representation of structures in multivariate SOM. Growing SOM (GSOM) [160] is another approach in which, firstly, a small map is generated, and then new nodes are added to specific coordinates until specified conditions are satisfied. However, this approach needs domain knowledge such as the growth threshold and the spread factor that the end user may not have to be integrated into this process.

The current implementation considers the data size and generates a square grid of SOM in the following way:

Equation -5.

$s_{MxM} \cong \frac{1}{2}$ (5 $\sqrt{N}$), where *N* is sample size and *M x M* denotes a square grid.

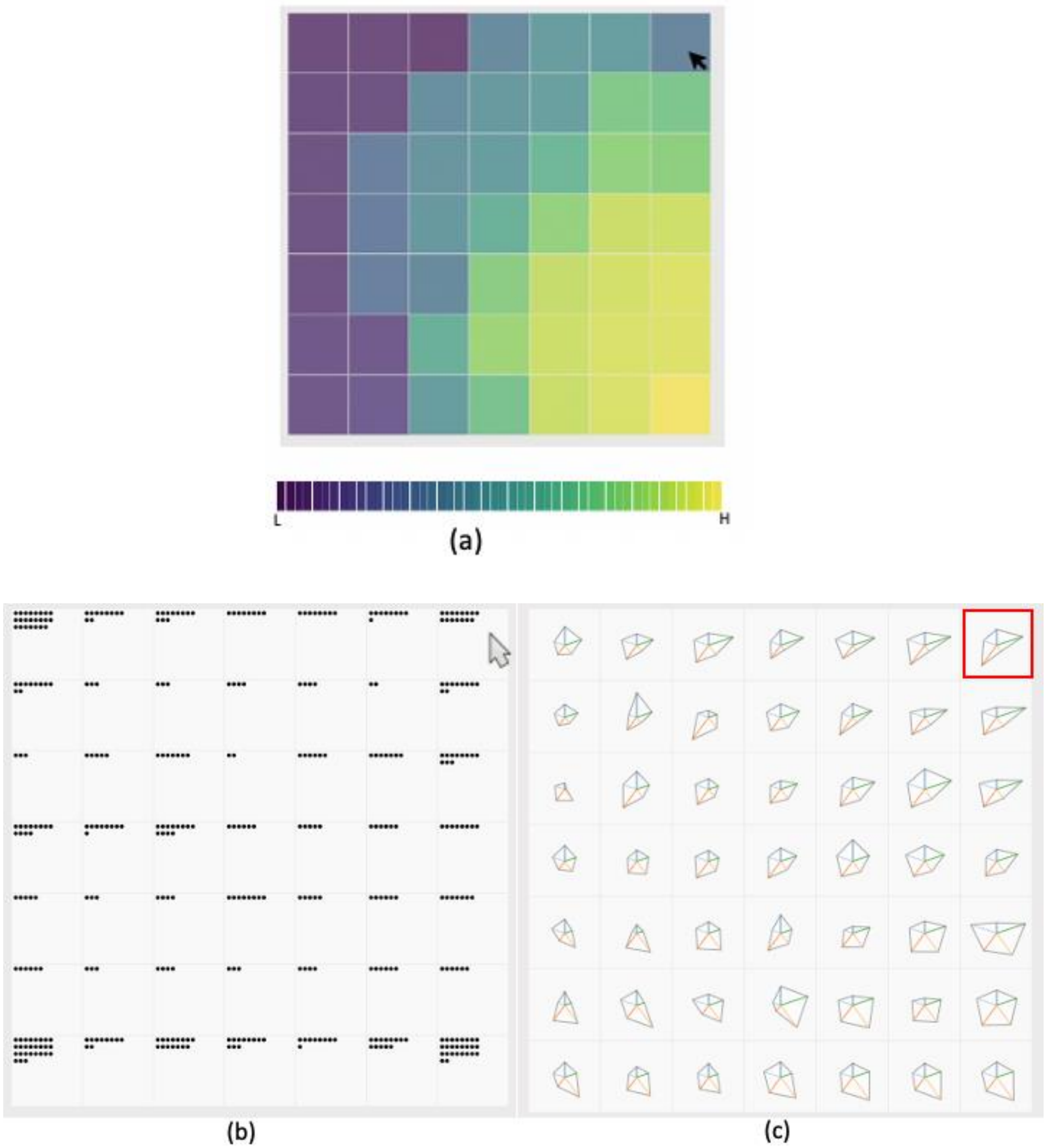


Figure 5-5 a) Self-organizing map depicting in the colour-map generated through Principal components of the code vectors of relevant cells b) along with ‘dots’ showing the distribution of objects among SOM cells c) along with ‘Star glyphs’ showing the mean of objects in relevant cells. (User Knowledge Modelling dataset [44] was used.)

The similar data samples fall into same node (cell) and the cells with similar attributes are adjacent to each other in the Self organizing map. For example, a user can detect the clusters formed, which are coloured with a colour scheme. The figures (b) and (c) in Figure 5-5 use the SOM layout (a) and giving extra information about the SOM. In other words, the pixel and Glyph-based methods are integrated into the top layer of SOM cells.

When looking at the node with the mouse pointer at the upper right side, figure (c) gives how many data sample takes place in that node and figure (b) takes average of multivariate data samples over their attributes in the node, and the glyphs encode these average values for each node. The size of the SOM nodes was taken into account in the glyph design selection. The SOM with the glyph visualization gives overall insight about the underlying values of the code vectors in the node. The figures (b) and (c) are used as complements to the traditional SOM map.

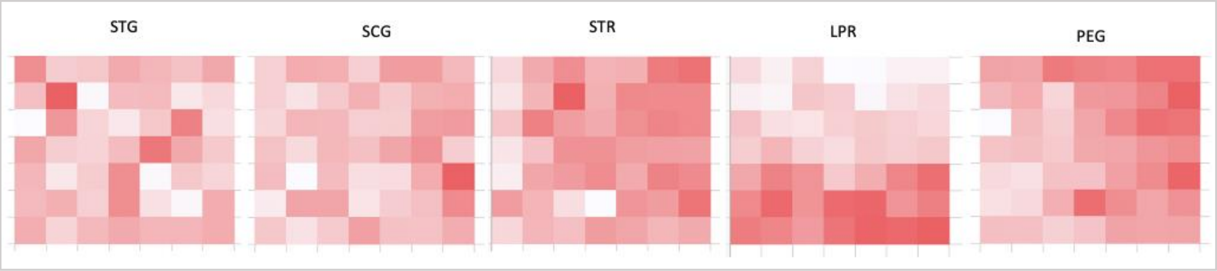


Figure 5-6 Each attribute in the dataset was shown with a separate SOM. The grid of each SOM is 7x7 as identical with Figure 5-5, whereas each cell depicting in a sequential colour scale was based on the relevant data attribute values normalized. In this colour scale, lower values depict on a lighter scale and higher values on a dark colour scale. (User Knowledge Modelling dataset was used. STG: The degree of study time for goal object materials, SCG: The degree of repetition number of user for goal object materials, STR: The degree of study time of user for related objects with goal object, LPR: The exam performance of user for related objects with goal object, PEG: The exam performance of user for goal objects)

The heatmaps above show the data features used. Heatmaps can be visualized side-by-side to detect patterns among attributes in the self-organizing map, as seen in Figure 5-6. For example, the heatmaps show a similar feature between STR (user's runtime rating for target objects-related objects) and PEG (user's exam performance for target objects-related objects), while LPR (user's exam performance for target-related objects) has an overall inverse relationship with these two features. Also, looking at the selected cell at the far right in the top row of each heatmap, STR and PEG have similar colour code (dark red). Again, both STG and SCG are coded in slightly lighter red for this cell. LPR, on the other hand, is encoded into almost white colour, which means it has a low value. These attributes together form a summary picture for this SOM cell, and the same inference can be easily obtained by looking at the Star glyph in this selected cell in Figure 5-5 (c).

5.5 CONCLUSION

To sum up, the Attribute module of the GroupVis was presented in this chapter, whose role in this framework is to provide initial understanding of the whole data space. The module consists of several projections and visual methods that aim to help the users investigate their data.

The dimensionality reduction methods, Multidimensional scaling, t-SNE and SOM methods, are implemented because they are able to preserve data structures within a large number of attributes when projecting onto 2D and 3D, while feature selection, which could be an alternative approach due to their explainability, would only preserve the structures of the two or three selected attributes when displayed in a 2D or 3D visualisation. With help of the dimensionality reduction projections, important insights can be gained by analysing these patterns (structures) in terms of clusters and much more. Moreover, the user can access all attributes of data items in dimensionality reduction views through the provided interactive methods (e.g., providing details of selected item with a tooltip).

Moreover, the quality metric views for t-SNE scatterplot was provided, as its various parameter settings often result in different projections. Thus, the software selects four plots from among all plots generated based on the metrics score to assist target users in making choice. Unlike the scatterplot of MDS and t-SNE, the SOM is represented as a square grid, where each grid cell can represent several data items, with similar cells grouped together towards one area. Two visual approaches are used with SOM, which are pixel-based and glyph based methods. Once the users understand their data by using given methods above, they can move on to the next stage (Clustering analysis) of the group formation process, which will be discussed in the next chapter.

Chapter 6 : EXPLORATORY VISUAL CLUSTER ANALYSIS USING PEAGLYPH

The Cluster module is intended to be used as an integral part of the GroupVis architecture. Users can explore cluster structures in a visual manner and analyse how these clusters are produced as well as their properties by utilising this module. This chapter introduces an interactive visual cluster analysis approach to make cluster analysis more intuitive. Through this approach, target users are provoked to ask new questions while exploring clustering outputs. The approach supports interactive exploratory analysis with multiple coordinated views to effectively help users analyse clustering results and facilitate practical cluster analysis for approximately equal-sized clusters coming from fuzzy and non-fuzzy clustering algorithms. The PeaGlyph is integrated as a complement to augment views by providing extra information and making the comparison of clusters and their relevant attribute values easier.

6.1 INTRODUCTION

When conducting data analysis, one of the most common tasks is the identification of groupings of dataset items that have common features with one another. Users will be able to obtain greater insight into their data, better comprehend it, and recognise trends as a result of this process. This, in turn, reduces the large dimensionality of the data. The term ‘clusters’ is used to refer to these conceptual groups.

In data mining, clustering is a technique that unsupervisedly discovers ‘natural’ structures that are hidden within data. The unlabelled data samples are automatically grouped into clusters, with samples from the same cluster being more similar than samples from other clusters. Clustering or clustering analysis is widely used in various fields, such as bioinformatics and image processing.

Numerous clustering techniques have been created, each with its own set of advantages and drawbacks. Due to the fact that traditional clustering methods do not put size constraints on clusters, they can result in severely unbalanced clustering [161]. As a result these techniques are incompatible with applications that need identical or balanced cluster sizes. A good example of this is in marketing campaigns. To ensure that each salesperson has the same workload for the purposes of fairness and efficiency, the provided consumers are partitioned into equal-sized clusters, with each cluster being assigned to a salesman [156]. Similarly, distributing students into equal size classes based on their abilities and tailoring teaching methods to the specific needs of each group. Most approaches treat the problem of obtaining equal sets as an optimisation problem.

Clustering is inherently a subjective method since ground truth labels in a data set (often) are not available [91], and a number of clusters are based on the task at hand or intended use of the results. Thus, automating the analysis of unsupervised learning tasks like clustering tasks is challenging, so practitioners may need to evaluate the validity of the analysis results using subjective factors. Moreover, it is an iterative process- there is no objectively ‘correct’ clustering algorithm for a particular problem- and often necessary to modify pre-processing and model parameters until the results achieve the desired priorities. Clustering outputs are used as inputs in the context of group formation, as discussed in Chapter 4.

Understanding clustering results and their reliability, exploring what features of the data set are responsible for clusters, and adjusting parameters to obtain better results are challenging tasks, especially for novice users with no background in related disciplines. Information visualisation methods play a prominent role in addressing these problems.

Visually analysing the balanced (equal or roughly equal) sized clusters that have emerged from the clustering algorithms taking the size constraint into account, can provide new insights from data, provoke users to ask further questions, and help understand the nature of the clustering structure as core or border points. The visualisations presented in this chapter allow users to interactively explore the border as well as core points to identify similarities (e.g. ‘similar’ should be understood as ‘which have close attributes‘) and distinct characteristics of clusters rather than necessarily to determine the optimal clustering algorithm for the data.

The PeaGlyph design that was presented in *Chapter 3* is used as a summary glyph of each cluster that encodes aggregated values of the data instances in generated clusters and as a complement to cluster views. Besides, the clustering results are visualised with the coordinated multiple views, which are the grid visualisation, scatterplot, and node-link diagram. Each visualisation technique offers its unique insight from the same clustering results, and interactive techniques allow target users to define the starting parameters of the provided clustering algorithms. In summary, the methods provided for cluster analysis are capable of providing valuable insights from the data and are worthy of further investigation.

The major contributions of this chapter are as follows:

- Methods for visual investigation of modified (size-constrained) fuzzy and hard clustering
- An interactive visual analysis procedure for exploration of clustering results, using the Multiple Coordinated Views approach
- PeaGlyph is integrated as a complement view that encodes averaged values of multivariate data samples in any cluster formed.

The proposed visualisation approaches and the clustering algorithms are discussed in detail in the rest of this chapter.

6.2 RELATED WORK

This part includes two subsections: the clustering methods with size constraints and visual tools or methods that are used for exploration clustering analysis.

6.2.1 *CLUSTERING WITH SIZE CONSTRAINTS*

Clustering algorithms developed to solve a particular problem, in a specialized field, usually make assumptions in favor of the application of interest. For example, the authors [162] observed automatic methods for text event detection, and stated that a widely used family of algorithms to detect events is based on clustering techniques. Xu and Wunsch [163] surveyed clustering algorithms and illustrated their applications in some benchmark data sets.

Clusters with predetermined sizes are required in a number of real-world scenarios. To illustrate, clustering with size constraint can be used to solve problems such as job scheduling when a set of jobs is assigned to different machines, considering each machine has a different capacity. This kind of clustering is also beneficial in establishing more relevant initial clusters and avoiding highly imbalanced clusters [164]. Finding clusters of roughly the same size using the k-means and Fuzzy-c-means algorithms is only possible when the data density is uniform. However, when the data distribution is not uniform, a single cluster covering can gain much more data instances than other clusters, and such cases may result in large differences in cluster sizes [165]. In practice, empty clusters are possible when using the typical K-means algorithm, especially in the case of multidimensional data sets with a larger number of clusters. Constraint k-means clustering was proposed by Bradley et al. [166] to assure that each cluster has a minimum number of objects in it by explicitly adding *k* constraints to clusters. Wagstaff et al.[167] presented an approach for constrained k-means clustering by the imposition of 'must-link' and 'cannot-link' constraints. They are focused on incorporating users' domain-specific knowledge into the clustering process. Malinen and Fränti [168] focused on obtaining clusters of balanced size and, at the same time, optimizing the mean square error (MSE). Ganganath et al. [169] demonstrated the modified k-means algorithm by setting size limitations on each cluster. They deliberately initialize the centroid of each cluster based on their prior knowledge, which decreases the likelihood of obtaining local minima and enables the extraction of clusters with preferred sizes. Additional data extracted from the data set can be used to create the limitations. In the study by Shalaginov and Franke [159], the 'soft clustering' problem was treated as a linear programming problem and solved using a heuristic technique. Instead of providing the precise size of each cluster, they discovered that utilizing a size range (e.g., the cluster size should not exceed 50) enhanced clustering performance. In the study by Höppner and Klawonn [165], the authors updated the model's objective function to take into account clusters of equal size. They proposed that such additional constraints into the objective function cause clusters to cover the same number of data objects. They suggested that the size of a cluster corresponds to the sum of the membership values for that cluster. However, their method does not always yield the desired cluster sizes. Recently, Li et al. [170] developed a cluster size-constrained fuzzy c-means algorithm intending to obtain the target size of clusters with varying data distributions. Their solution requires that the intended proportion of cluster sizes is known in advance for the clustering problem.

In our system, creating balanced size pools (clusters) in which individuals with similar characteristics take place is significant within the context of creating heterogeneous study (project) groups. The study groups will be formed by bringing together individuals from each pool to ensure that each student group has the diversity available in the data set.

6.2.2 *VISUAL CLUSTER ANALYSIS*

Data entities that are similar to each other in multidimensional space are grouped together. It can be difficult to understand whether their similarities are mainly in all dimensions or in sub-features. Also, it is hard to identify an ideal solution because of the lack of ground-truth labels, and human judgement is required to determine what may be regarded as a satisfiable clustering result in the first place [171]. Effective visual tools help successfully explore a clustering space as well as understand clustering results. The integration of cluster analysis with information visualisation techniques is denoted by the term ‘visual cluster analyses. In a comprehensive review of interactive clustering by Bae et al. [172], the reasons for favouring interactive clustering were described in four categories: improving the clustering quality, understanding final results, finding particularly interesting data in a particular context, and the subjective reasons of the clustering task.

Hierarchical Clustering Explorer [173] is an early form designed for exploring hierarchical clusters that uses a heatmap encoding to allow users to explore clustering results from pairs of clusters. XCluSim[174] supports interactive comparison of several clustering results in bioinformatics data, based on ’the visual information seeking mantra’ [39]- i.e. overview first, zoom and filter, then details on demand. It uses several graphical displays such as parallel sets, tabular set view and dendrogram to enable users to explore cluster distributions while comparing multiple clustering results. As a relevant example to the XCluSim, the Matchmaker [175] encoded heatmaps in dimensional axes of parallel coordinate plots and then reveales the relations between items in the heatmaps. In order to make the comparison more manageable, the item values in each dimension are rearranged based on their average values. Similarly, Younesy et al. [176] provided a design including an interactive heatmap with a query interface for analysing epigenomic data. For grouping data into subsets, the tool provides options for k-means clustering and querying. Demiralp [177] introduced Clustrophile, which supports iterative computing of clusters and user interaction.

It enables users to explore multiple choices of algorithm parameters through visualisations, including scatterplots and heatmaps. Clustervision [178] is another visual analytic tool that helps a user find the right clustering among the variety of clustering techniques and parameters available. It provides several coordinated visualisation methods for exploring clusters and comparing their results based on several quality metrics. The quality metric values were represented by a radar chart shown along with the corresponding cluster plot. Another efficient example by Cao et al. [101], who presented DICON, which adopted a treemap scheme for icon design to represent the multidimensional cluster, the also presented a layout algorithm to facilitate cluster comparison and interpretation. Unlike the studies mentioned, Fuchs et al. [179] presented a visualisation application for teaching clustering algorithms which educators and students can benefit from.

As mentioned in the previous sub-section, several different tools exist that integrate clustering methods in an interactive visualisation environment. In general, they perform cluster analysis on different types of datasets, including gene expression datasets, trajectory datasets and geoformation. A comparison of these different approaches are summarised in Table 6-1.

	2D	3D	Geometrically-transformed Display	Iconic Display	Dense Pixel Display	Stacked Display
Choo et al.[180]	■					■
Erra et al. [181]	■		■			
Lee et al.[182]	■		■		■	
Arin et al.[183]	■			■		■
Muller et al.[184]	■					
Tatu et al.[185]	■				■	
Xu et al.[186]		■	■			
Seo and Shneiderman [173]	■					
Schreck et al.[187]	■					
Cao et al.[101]	■		■	■		
Van et al.[188]	■		■			
L'yi et al.[174]	■		■			■
Kwon et al.[178]	■	■	■			
Demiralp[177]	■		■			

Table 6-1 The visualization techniques used in visual clustering analysis, given by Keim's taxonomy

As can be seen, some studies leveraged together more than two visual methods for clustering analysis. The absolute majority of visualizations used in clustering analysis rely on standard 2D and are followed by geometrically transformed displays. Iconic displays were the least preferred method compared to others. Apart from the visualizations, colour as an encoding (visual) channel is the most commonly used method for displaying clusters in analysis.

The clustering module (system) presented in this chapter utilizes the well-defined visual methods as well as the novel glyph design discussed above to enable users to explore the pool characteristics and compare them in terms of what makes the clusters (for example, identifying salient attributes in the clusters and use them to identify clusters). Also, the output of most clustering algorithms can be input into these visual methods. In the context of group formation, clustering results are provided as input to the grouping module; therefore, visualizing these results and allowing users to examine the results interactively will make this process transparent and understandable before the working groups are formed.

6.3 CLUSTERING AND EXPLORATORY ANALYSIS

This section will describe a new system for visual cluster analysis that has been developed as part of the research of this thesis. The system follows a visual analytic approach, and the outputs of it serve as input to Grouping module. The design is open-ended in terms of which clustering approaches to use. K-means (a hard clustering method) and Fuzzy c-means (FCM) (a soft clustering method) are being used to demonstrate the functionality of the system. Various visual methods including PeaGlyph are seamlessly used together in the workflow to analyse cluster results.

6.3.1 *CLUSTERING ALGORITHMS*

Finding hidden patterns in data can be accomplished by utilising a variety of clustering algorithms. Deciding between the cluster methods and parameters often depends on the data set and task. We chose two different clustering algorithms to show the effectiveness of our visualisation methods while analysing clustering results.

$$U_{mk} = \begin{bmatrix} 0.0 & 1.0 \\ 1.0 & 0.0 \\ . & . \\ 1.0 & 0.0 \end{bmatrix}$$

a) K-means metric where cluster number =2

$$U_{mf} = \begin{bmatrix} 0.1 & 0.9 \\ 0.7 & 0.3 \\ . & . \\ 0.4 & 0.6 \end{bmatrix}$$

b) Fuzzy c means membership metric where cluster number =2

Figure 6-1 The figure shows clusters columns (C=2) and N rows, where C represents the total number of clusters and N presents the total number of data points. In metric (a) ‘’1” means the record belongs to the corresponding column (cluster), but ‘0’ is not. The metric (b) presents the membership values of each record across clusters; and for each data point, the sum of its membership values should be 1.0.

Due to its simplicity, K-means clustering [189] is one of the widely used clustering methods, and this method ensures that each data point is associated with a single cluster. More precisely, the algorithm seeds random points as cluster centroids and then data records are grouped based on their distance to the centroids while minimizing the algorithm's objective function. After, the mean of the data items in each cluster becomes new centre point of the relevant cluster. Then, the data instances are reassigned accordingly. The process is repeated until there are no longer changes in the clusters and convergence has been reached.

When compared to K-means, Fuzzy c means (FCM) [190], a soft clustering method, is distinguished by the fact that each data point might belong to more than one cluster. The algorithm assigns a membership value to each member based on their distance to centre points (centroids). If a data record is closest to a particular centre, its membership value for that cluster is higher than others. The membership values are used to position the centre points of clusters. Figure 6-1 displays the difference in cluster membership assignment for hard and soft clustering. Both algorithms need prior knowledge of the number of clusters to create subsets accordingly. The appropriate number of clusters is related to the goals of the analyst and may be highly subjective. Besides, in Fuzzy c-means, fuzzifier value is critical since a high value results in information loss, whilst a low one results in the inclusion of false observations arising from random noise. In numerous investigations, the value of the fuzzifier was set to 2, which is a common setting so is it in this thesis.

6.3.1.1 *MODIFICATION OF K-MEANS*

We adapt the approach introduced by Ganganath et al. [169], where the conventional k-mean algorithm is initialized with selective initialization and size constraints are applied to clusters for obtaining preferred size clusters. They made the assumption that the users have prior knowledge of at least a few data points, which allowed them to selectively initialise the clusters.

In current settings, seen in Algorithm 6-1, we do not make any assumptions; instead the K-means++ [191] centroid initialization technique was used to ensure that the centroids are initialized in a more intelligent manner and that the quality of clustering improves. The update and termination steps from the k-means method were left unchanged. To obtain roughly equal size clusters, in Figure 6-2, the size limitation to be applied to the clusters is found as in *Equation -6*, where the cluster size constraint (ζ), the number of data instances (N) and the number of clusters (k) are shown.

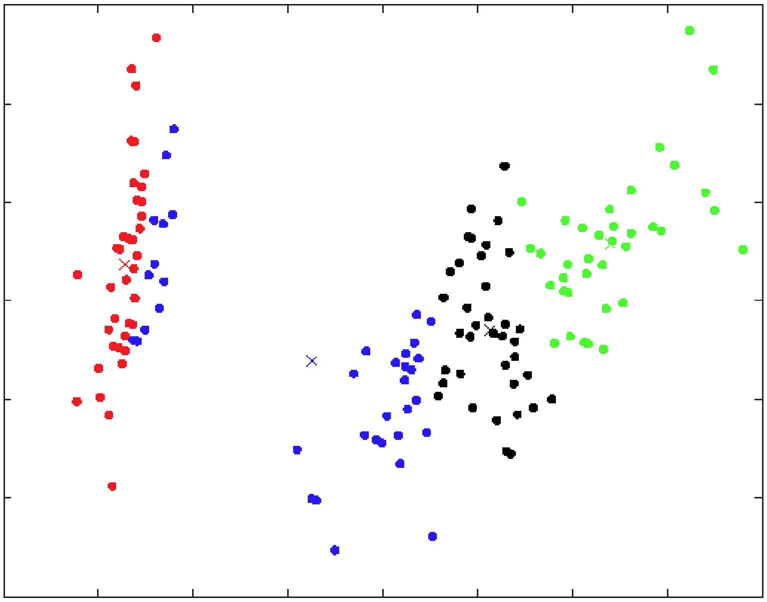


Figure 6-2 Representing balanced size clusters (38, 37, 38, 37), (k=4). Iris dataset is well-known dataset that contains 3 classes with 50 instances each. This data set was used to demonstrate how the clustering algorithms work with different k values.

Euclidean distance metric is used to calculate similarity which is basically the square root of the sum of squared differences between all data points to the cluster centroids. Besides, there are other metrics that are commonly used in data mining tasks. The city block (Manhattan) distance, for example, is the sum of absolute differences of 2-data points. The cosine similarity metric quantifies the similarity between two vectors in a multidimensional space by computing the cosine of the angle between them.

Equation -6.

$$\text{Cluster size constraint } (\boldsymbol{\zeta}) \cong \frac{N}{k}$$

Here the number of data instances (N) is divided by the cluster number (k). If not, the size of clusters will be $\boldsymbol{\zeta} \pm 1$. $|c_j|$ denotes the size of cluster c$_j$, and $1 \leq j \leq k$ **,** $|c_j| \leq \zeta$.

```
Algorithm : K means Algorithm
Input       $X = \{x_i\}_{i=1}^{N}$ the set of data points and $\boldsymbol{k} \ \boldsymbol{\varepsilon} \ \boldsymbol{N}$ the number of clusters
Output      The set of representations (R) and cluster sets (C)

function K-means(X, k):
    $R^o$ is initialized via K means ++ method
      repeat
          Sorting values of distance proximity
          Assigning data items to clusters, only if it satisfies $|c_j| \ \le \ \zeta_i$
              $C: c_j = \left\{x_p: |x_p - \mu_j|^2 < |x_p - \mu_i|^2, \forall_{i,} \ 1 \le j \le k \right\}$
          Updating the $R \ (1 \le j \le k)$
      until $| \mu_j{}^{(t+1)} - \mu_j{}^{(t)} |$ < ε
    return {R, cluster sets}
end function
```

Algorithm 6-1 Constraint sized K-means pseudo-code- ***(*:** the distances between data points and cluster centres are given by ascending order**,** and the algorithm was limited to Euclidean distance)*

In the assignment step, the distances from the centre of the clusters to each data point are added to the item in an ordered array. If the first set in an item list meets the size constraint, the item is assigned to that set. If the priority set is full, the process is iterated till it finds a cluster satisfying the size limitation among sorted clusters, and the item is assigned to this cluster. After obtaining the equal size clusters, if there are any unassigned members, they are similarly assigned to the appropriate clusters in the same way as earlier.

6.3.1.2 *MODIFICATION OF FUZZY C MEANS*

The conventional Fuzzy c-means approach does not control the size of clusters by its own inherent mechanic, so it is necessary to explicitly add size constraints explicitly to clusters. Recently, Chakraborty and Das [192] proposed a variation of the original FCM method to obtain clusters with specified sizes. The original paper contains the specifics on how the implementation was carried out. Unlike their work, in the current implementation presented, seen in Algorithm 6-2, we initialized the Fuzzy c mean algorithm via the Fuzzy c means ++ by [193] for improving the quality of the clustering method. Secondly, it makes no such assumption and even if the number of data samples cannot be divided by the number of clusters.

```
Algorithm : Fuzzy c means Algorithm
Input   $X = \{x_i\}_{i=1}^{N}$ the set of data points and $\boldsymbol{k \, \varepsilon \, N}$ the number of clusters
Output  U: the matrix of membership degrees, R: the set of representatives

function FCM (X, k):
    $U^{O}$ is initialized via Fuzzy c -means++ method
    repeat
        Update the membership matrix U
        Update the set of centroid p
        Calculate the value of the objective function J using
            $J(U,P) = \sum_{\mathrm{i=0}}^{\mathrm{n}} \sum_{j=1}^{k} u_{ij}^{m} \left|x_i - p_j\right|^2$
    until | $J^{t+1} - J^{t}$ | < e
  return {U, R, cluster sets}
end function
```

Algorithm 6-2 Constraint sized Fuzzy c-pseudo-code

After executing the standard FCM algorithm, the potential belongingness matrix *U* = {u_{ij}} is obtained. Then optimization problem with size constraint is solved by using the following way:

Equation -7.

$$maximize \;\; f(x) = \sum_{i=1}^{N} \sum_{j=1}^{k} m_{ij} y_{ij},$$

for *i= 1,2, ..,N and m_{ij} is the membership value of i -point for the cluster j.* A data instance is included into only one cluster with

Equation -8.

$$\textstyle\sum_{j=1}^{k} y_{ij} = 1, \; y_{ij} \text{ is a solution matrix.}$$

For example, given membership matrix *(U)* in Table 6-2, where columns represent clusters *(0,1,.., k)* and rows denote the data points (*0,1,.., N).* Looking at the membership matrix below, *item-0* goes to the *cluster-1* (*c1)*, as its membership value is the highest among other clusters in this row.

	c0	*c1*	*c2*	*c3*
item-0	0.12	0.65	0.15	0.08
item-1	0.03	0.70	0.07	0.2
item-2	…	…	…	…
item-3	0.55	0.24	0.15	0.06
item-4	0.42	0.02	0.01	0.55

Table 6-2 A sample of a membership matrix

Here the solution matrix (y) (Table 6-3) is given for the membership matrix (U) above. *1s* in the matrix show the assigned clusters, whereas *0s* show the unassigned clusters. Each row should have only an '*1'*; in this way, a data instance is included in an only cluster.

	c0	*c1*	*c2*	*c3*
item-0	0	1	0	0
item-1	0	1	0	0
item-2	…	…	…	…
item-3	1	0	0	0
item-4	0	0	0	1

Table 6-3 The solution matrix for the membership matrix given in Table 6-2.

The sum of each row must be *1* as a data point is included in only one cluster. The sum of each column gives the number of data items within the corresponding cluster.

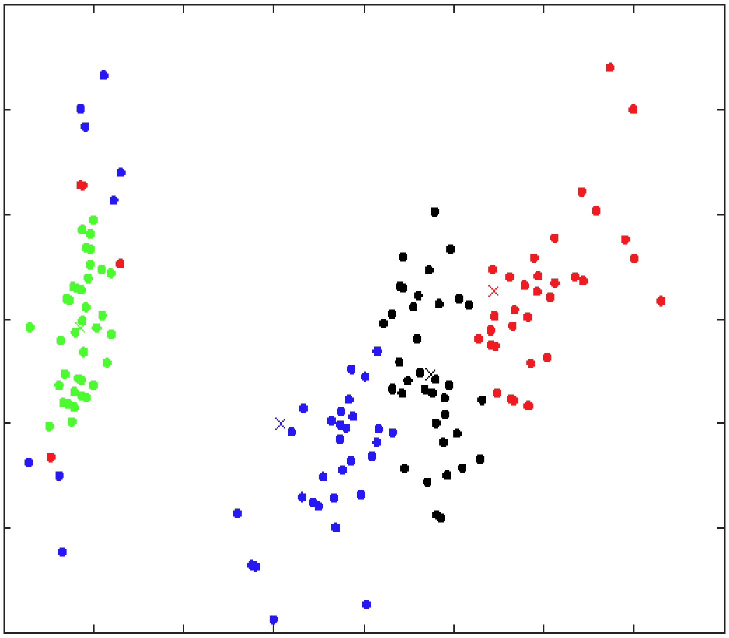
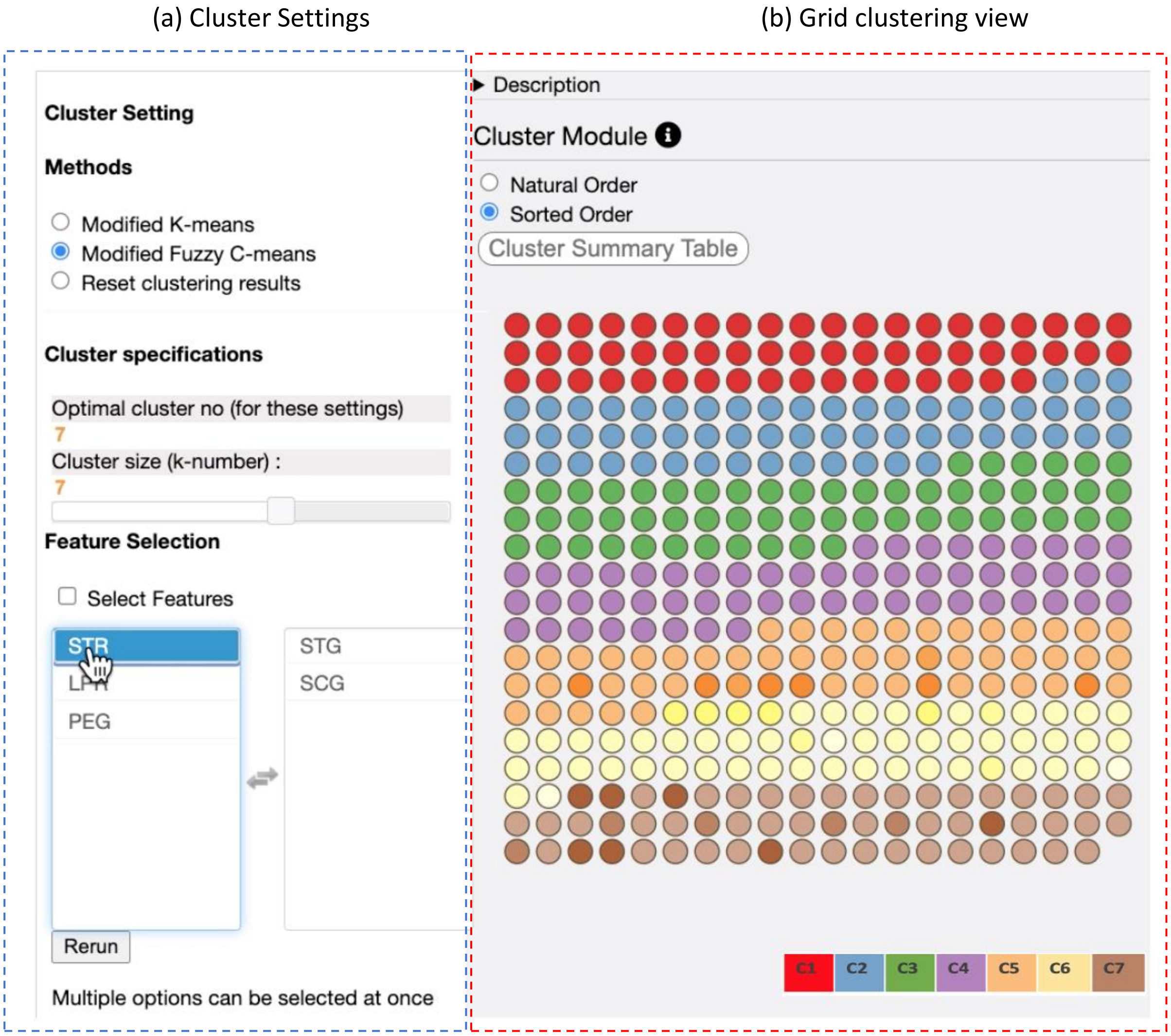


Figure 6-3 Representing balanced size clusters (38, 37, 38, 37) with the variation of Fuzzy c-means (Iris data, k=4). Looking at the cluster on the left in the figure, the algorithm redistributed points with a lower belonging score in the cluster for creating balanced size clusters.

In the study by [192], their assumption is that the number of data items *(N)* is exactly divisible by the number of clusters (k). In real-world cases, this is not always possible; hence our implementation did not assume it. Furthermore, in our implementation, the points far from cluster centres (i.e. the points located at the boundary of the cluster) are redistributed rather than points closer to cluster centres, as can be seen in Figure 6-3.

6.3.2 *VISUALIZATION SYSTEM OVERVIEW*

As previously noted in ‘Background’, visualisation of data is critical for cluster analysis. This section presents several well-structured visualization methods to help effective and interactive inspection of clustering results.



(a) Cluster Settings

Cluster Setting

Methods

○ Modified K-means
● Modified Fuzzy C-means
○ Reset clustering results

Cluster specifications

Optimal cluster no (for these settings)
7

Cluster size (k-number) :
7

Feature Selection

□ Select Features

STR
LPA
PEG

STG
SCG

Rerun

Multiple options can be selected at once

(b) Grid clustering view

▸ Description

Cluster Module

○ Natural Order
● Sorted Order

Cluster Summary Table

Figure 6-4 The overview of Cluster module . The legend at the bottom of the Grid view represents the colour scheme the clusters were encoded into.

The system shown in Figure 6-4 delivers a single clustering result but utilises a variety of visualisation techniques to simultaneously convey many aspects of the result.

Our approach is similar to previous research by Long and Linsen [188] and achieves a seamless balance of overview and details [194]. Moreover, the system enables users to select one of the given clustering algorithms and flexibly define parameters of the relevant method like cluster numbers to account for the diversity of analysis questions. The proposed approach attempts to make it easier to explore variations between data points inside clusters (i.e. presenting variability within clusters) and to reason about clustering instances (i.e. What data attributes are informative in defining a given cluster).

Additionally, the system enables the application of clustering and projection techniques to filtered data subsets.

We focused on several requirements to make the system applicable to clustering tasks. The details of the system components are described in the following sections. The system is similar to existing studies in that it coordinates visuals in order to investigate clustering results. Additionally, the system supports iterative and interactive data exploration by allowing users to examine the algorithmic parameters, as well as given visuals. Further, extra functionalities, e.g. additional clustering algorithms, reordering approaches, and more visualizations, can be added to the system in order to support the different clustering tasks.

6.3.2.1 *SETTING PANEL*

The tool presents optimal cluster number that best capture the segmentation of the data set in terms of the attributes, using the Silhouette coefficient metric [195]. Users can directly use this optimal number provided for the clustering algorithms. Besides, the users themselves have the ability to set the cluster number by using the slider given. The silhouette coefficient (*Equation -9*) is a commonly used approach for determining the quality of clusters that incorporates both cohesion and separation. For item *i's,* the silhouette coefficient is written as:

Equation -9.

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

Where α_i is the average distance between *item -i* and *all other items* in its cluster (*c_A*), and b_i is the minimum of the average distance of the *item -i* to *all items* of any cluster (not containing *c_A*). *Equation -10* calculates the average silhouette coefficient of *all items*, which provides an overall estimate of the goodness of the clustering.

Equation -10.

$$s_k = \frac{1}{N}\sum_{i=1}^{N} s_i$$

where N represents the total number of data instances contained within data set, and s_k is the average silhouette coefficient, which is calculated by taking the average of the silhouette coefficients of *all items*. When it comes to the silhouette coefficient, its score ranges between -1 and 1. A greater score indicates a higher degree of clustering quality, which means there are smaller dissimilarities within clusters, but higher dissimilarities between clusters. If desired, the user can set a different value to the clustering number instead of the proposed number using the slider web widget. The color-coded clustered data samples are represented by the scatterplot, in Figure 6-6.

Another property of the Clustering module allows users to create a new feature set by choosing from the attributes in the dataset to be used in clustering analysis by using their domain knowledge. As the data may have a large number of attributes and the irrelevant ones can ruin the clustering [196]. Under such circumstances, selecting the most discriminative or representative attributes of a sample inevitably becomes an important issue.

6.3.2.2 *ENCODING CLUSTER ITEMS*

Colour and positions, which are highly efficient visual cues [93], are commonly used for representing and identifying clusters in scatterplots. Data encodings can be correctly deciphered with the help of a well-designed colour scheme [197].

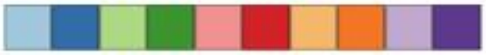


Figure 6-5 Ten-class paired from *ColorBrewe2*, these do not imply magnitude differences between classes and is used to create the primary visual differences between classes.

The system provided in this chapter uses a qualitative hue scheme [198], as shown in Figure 6-5, to mark the different cluster points on the Grid view in Figure 6-4 and the scatterplot of MDS or t-SNE in Figure 6-6.

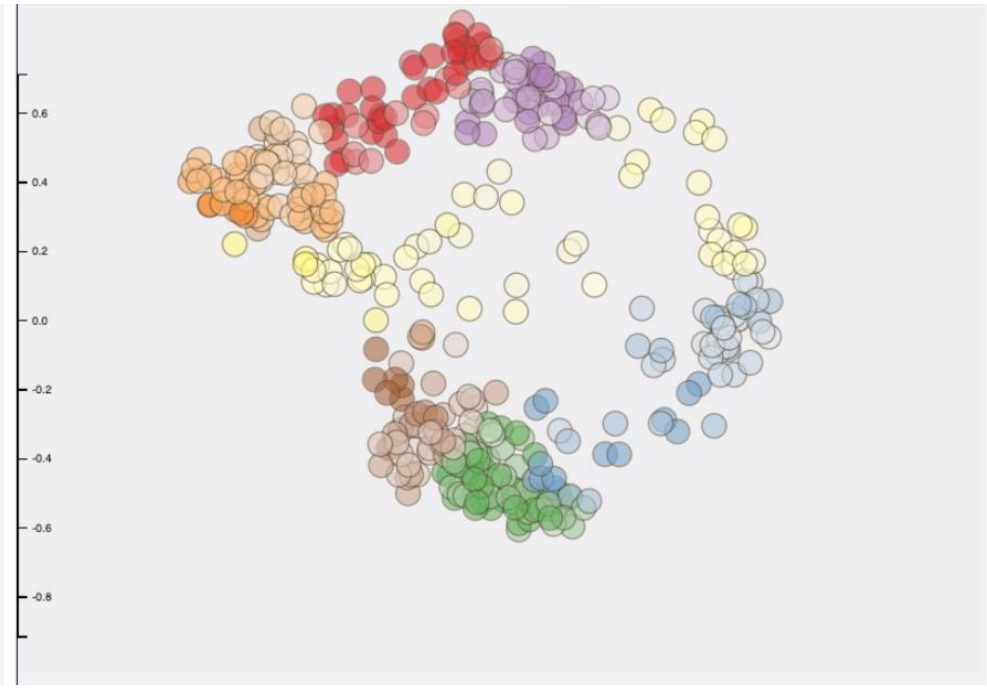


Figure **6-6** The MDS scatterplot showing the clusters formed based on the settings in Figure 6-4. (User Knowledge Modelling dataset [44] was used.)

Alongside the scatterplot projection, the grid visualization is presented, in which the data instances in the scatter view are given in an organized manner to mitigate the overplotting problem of the scatter plot. Each circle in the Grid view is colour-coded with the corresponding cluster.

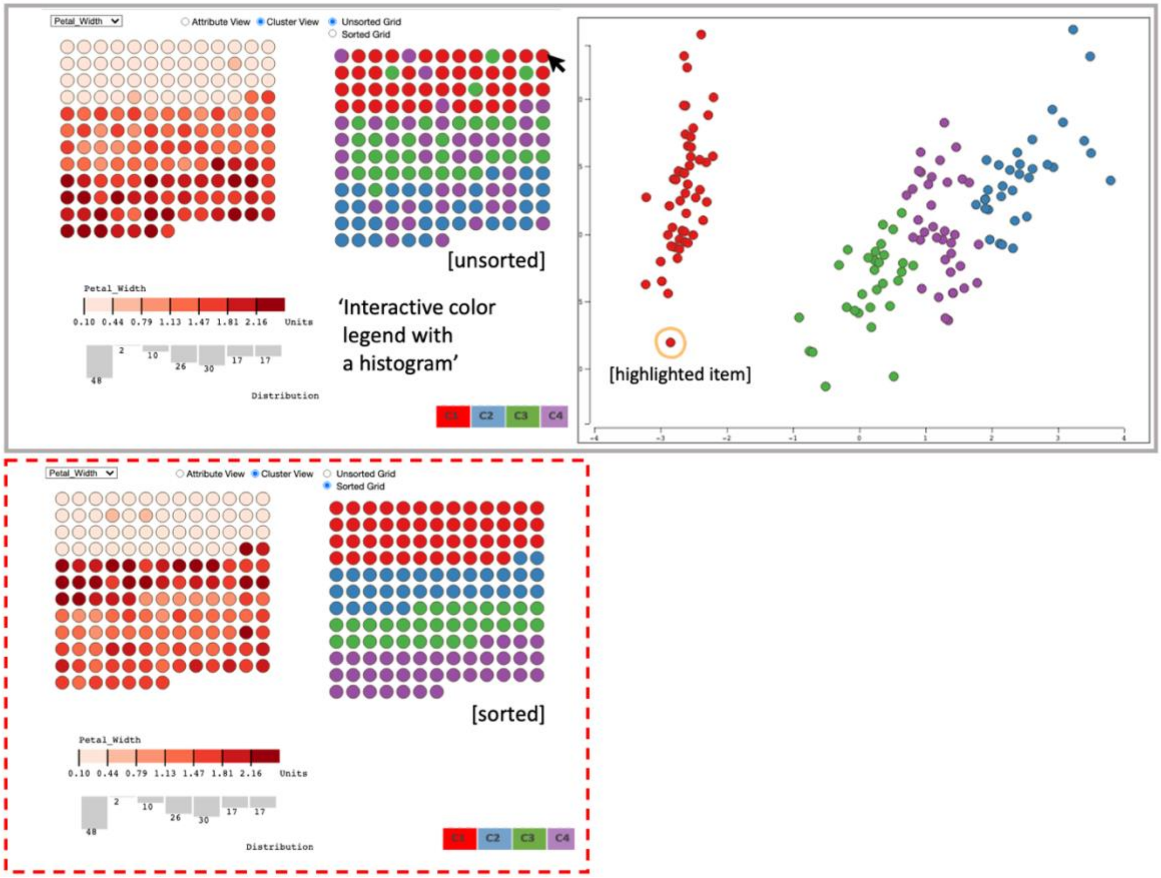
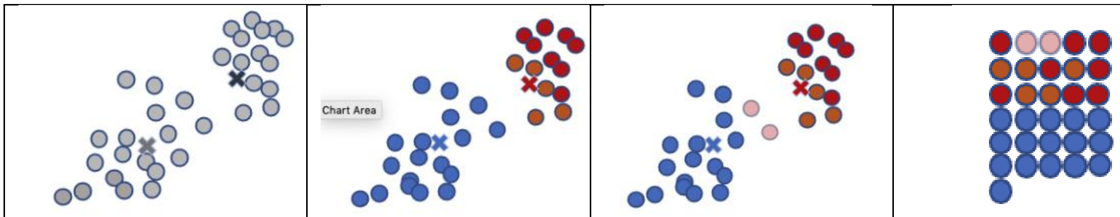

Petal_Width
○ Attribute View ● Cluster View ● Unsorted Grid
○ Sorted Grid
[unsorted]
Petal_Width
0.10 0.44 0.79 1.13 1.47 1.81 2.16 Units
48 2 10 26 30 17 17
Distribution
'Interactive color legend with a histogram'
[highlighted item]
C1 C2 C3 C4
Petal_Width
○ Attribute View ● Cluster View ○ Unsorted Grid
● Sorted Grid
[sorted]
Petal_Width
0.10 0.44 0.79 1.13 1.47 1.81 2.16 Units
48 2 10 26 30 17 17
Distribution
C1 C2 C3 C4

Figure 6-7 The Grid view (on the left side) has two options, 'unsorted'-on the top left and 'sorted'-on the bottom left. When selecting sorted, the items are organized in terms of color-coded clusters, at the same time the Grid view on the leftmost is sorted out according to the order of the relevant items

The 'Sorted order' option helps to see the clusters more clearly compared to the 'unsorted' on the grid view, as on the scatterplot shown in Figure 6-6. The hue scheme for a given scatterplot is used in the Grid view to encode clusters, and the data items are encoded in the same colour of the cluster to which they are assigned.

Two different views, the Scatterplot and Grid view of the same data were presented side by side. Correspondingly, if the user wants to see the details of the data samples, they can click on the circles in the grid view or on the scatterplot, and the detail of the relevant sample is given via the tooltip. The Grid view is quite scalable as the circles in pixel size can be scaled down in case too many data points need to be compared. This approach may provide an overview even for large amounts of comparison values. The data instances that are usually located in areas close to the distal border are not assigned to the nearest cluster due to the size constraints; instead, they are distributed to other neighbouring clusters (nearby clusters). The re-assigned items are encoded in a lighter tone of the hues of the clusters to distinguish them from other data instances in Figure 6-8.



1)Initial cluster results with different sized clusters

2)The cluster results are encoded in different colours

3)The points near border are re-assigned (shown with a light tone of colour red as the points belong to the blue cluster with lower degree of belonging so the points are re-assigned to red to provide balanced size of clusters.

4)All points projected on the scatterplot are given in the Grid view and the re-assigned points are shown in the a light tone of assigned cluster. The order of the data items within clusters is not significant.

Figure 6-8 Sematic representation of fuzzy grid view

In this way, users can recognize that although the items are in the same set, the degree of belonging to the cluster may differ. They can also distinguish data instances with lower degrees (i.e. usually located in the border regions) from the others within the cluster by following the lighter colour-coded data instance on the Grid view.

6.3.2.3 *VISUALIZATION OF CLUSTER SUMMARY*

A summary cluster glyph encodes the average of all attributes of the members in a cluster. The PeaGlyph in the cluster summary table depicts the summary values of each cluster. The summary glyphs are created to make a quick comparison between clusters and help understand what instances are dominant for the grouping of data points. Each instance value is mapped to a pod with filled/unfilled peas, then combined to form the PeaGlyph to facilitate quantitative comparison over clusters in case the peas encode the normalized values. The glyph gives the ability to explore salient features of clusters and assess the differences and similarities of clusters regarding their attributes.

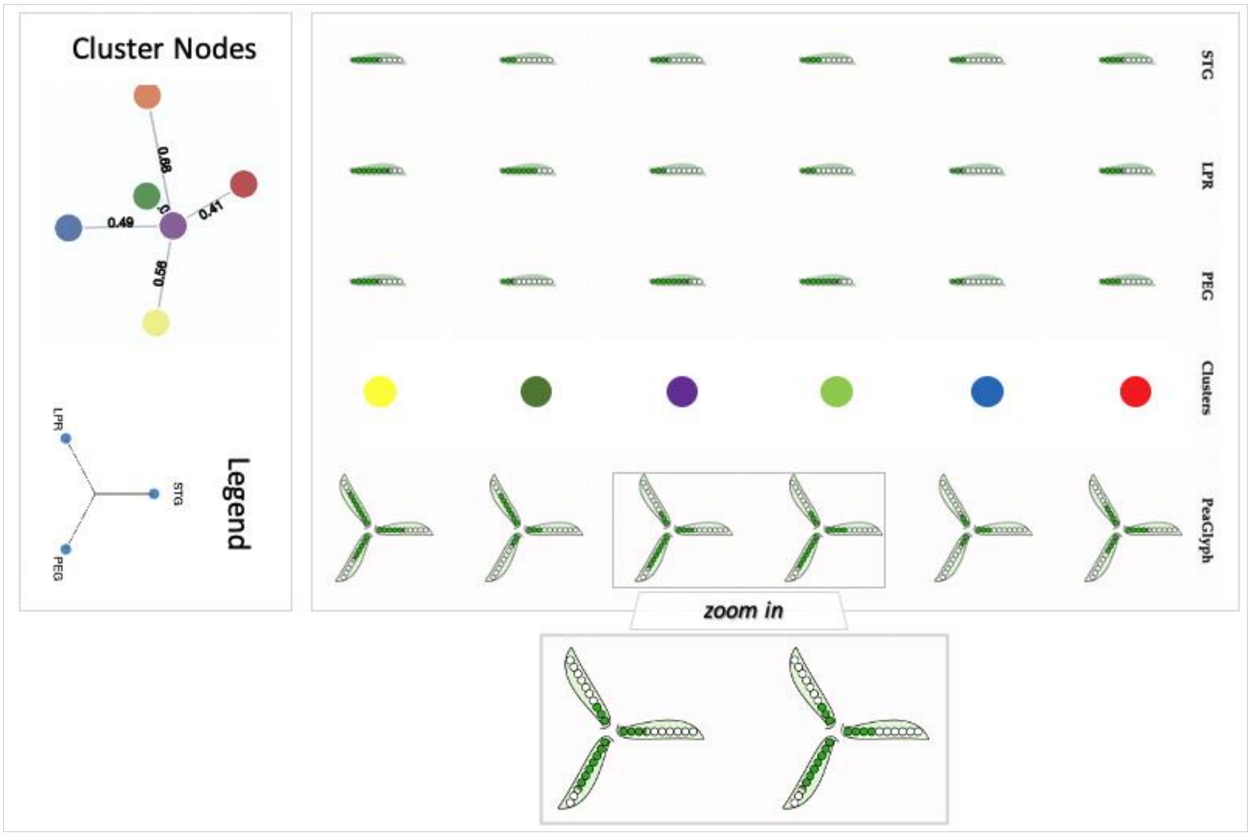


Figure 6-9 The overview of ‘Cluster Summary Table’. The figure was generated based on the dataset (Chapter 7) used to create student groups, (Settings: K means, cluster number: 6, and the data attributes used for clustering: STG, PEG, and LPR)

The overall shapes of peas capture the similarity and difference clusters visually, which means the shapes of similar clusters appear similar or vice versa. In Figure 6-9, *pea pods* in the columns encode attribute values, the coloured circles in the Cluster column indicate *hues* for which the clusters are encoded, and each *PeaGlyph* combines the attributes that summarize a cluster to form a summary glyph.

6.3.2.4 *CLUSTER SIMILARITY VIEW*

It is possible to depict gauged similarity values between clusters in diverse clustering findings using a variety of ways. For example, a colour-coded similarity matrix can display how frequently each pair of items is clustered together or the number of items are shared by each pair of clusters [199]. Typically, this sort of visualisation is limited to comparing a pair of clustering results. A graph layout was employed in the study [174] to display cluster results in more scalable manner, in which nearly identical findings are clustered together with thicker links. For displaying similarity in the graph layout view they used physical distance, which provides a perceptual benefit. When additional results are presented to the graph view, the size of the nodes has been reduced in order to maintain the scalability of the view. A heatmap (matrix diagram) visualization can be used to show the results of a clustering result. In general, the number of clusters is represented by the heatmap's columns, and the features are represented by its rows. In each cell, the colour code represents the normalised average feature value for clusters, as shown in the Clustrophile tool by Demiralp [177], in Figure 6-10.

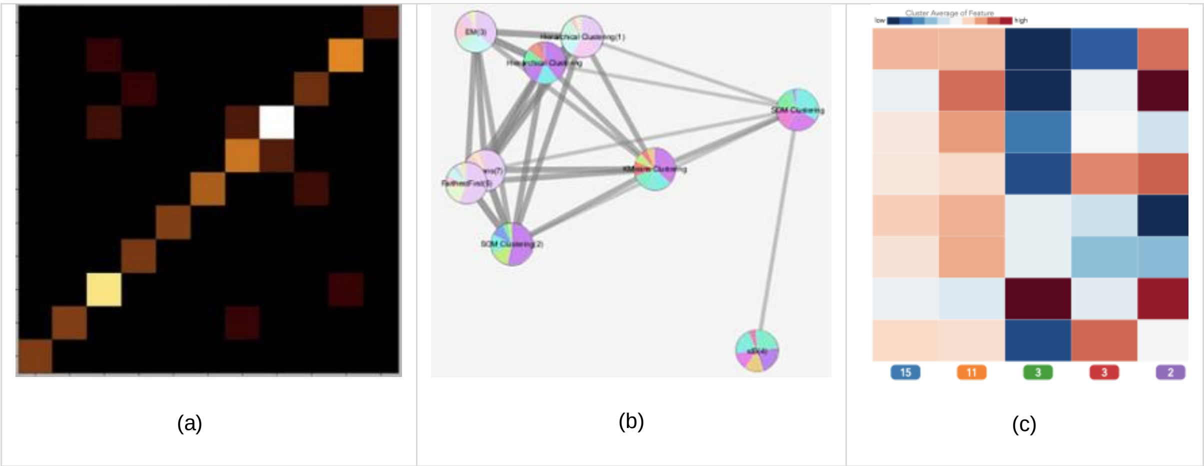


Figure 6-10 a) A pairwise stability matrix for density based clustering and K-means [199] b) A graph layout (a force-directed layout) showing similarity overviews of the different clustering algorithms [174] c) A matrix diagram of a discrete clustering in Clustrophile by Demiralp [177].

In the node-link graph, general visual variables are encoded into nodes (data instance) and links (their relationship). In the graph, each node represents a cluster, and its colour is matched with the related to the colour of a cluster, making it easy for the users to differentiate the nodes of the graph.

The length of the connection between the nodes indicates their distance from each other (i.e. the quantitative nature of the similarity of the clusters). This visualization component visualizing clustering results places similar clusters closer together, making it easier to detect similarities between them, which is supported by the Gestalt principle of proximity.

The fact that the connection between two nodes is shorter than the others indicates that these two clusters-coded nodes are more similar to each other. According to the different cases at hand, different measurements may work better than others. The current implementation uses Euclidean distance as the similarity metric. In the graph, when the user selects the relevant cluster node, that node becomes the central node, and other cluster nodes are connected to the central node based on their similarity to that central node. The node graph may be thought of as a starting point for exploring cluster results.

6.3.2.5 *INTERACTIVITY*

The visualization methods are supported by several interaction techniques such as selecting, zooming & panning, and highlighting. For example, when selecting a point of interest in the '*Grid view',* the point is highlighted in the scatterplot, alongside a tooltip that presents key attributes associated with the data item. When hovered over a point in given scatterplot, *the zoom and pan behaviour* appear by default, which facilitates seeing different areas of the zoomed region of the scatterplot. Besides, when a point is selected, the remaining points are shown with a low opacity without losing the context. The clustering stage itself is not interactive; however, after getting the clustering results users can request another run of the clustering phase with different parameters and feature subsets to achieve desired cluster results.

6.3.2.6 *EVALUATING CLUSTERS*

Typically, what constitutes a *'good'* clustering is determined by criteria relevant to a specific domain or application. The Cluster module is designed to assist users in comprehending clustering findings by allowing them to examine multiple synchronised views at the same time. Besides, it has two evaluation metrics to evaluate the quality of clusters in a quantitative way.

The one of measurements implemented in this module is the *Within-cluster sum of squares (WSS)*, which sums the average squared distance of all points in a cluster to the cluster centroid. The higher the WSS of a cluster, the greater the variation in the data inside the cluster. The other measurement in the module is the *Between-cluster sums of squares (BSS),* which sums the average squared distance between all cluster centroids. Having a large number indicates that clusters are spread apart, whilst having a small value indicates that clusters are near together.

In Figure 6-11, it can be seen that when *k* is set to 3 for Iris dataset, the clustering output contains 3-cluster divisions, and a single group (the iris-setosa) is clearly distinguished from the others (the iris-versicolor and iris-virginica groups), whereas the others are not. For this case, the scores of WSS and BSS for traditional K-means and Fuzzy-c means seems close to each other. However, if clusters of roughly equal size are to be created, we can see that the scores of these metrics may be worsened as seen in Figure 6-12. The dataset is not properly clustered by using these two clustering methods as in the previous case. This is not surprising to see, since the dataset is inherently composed of 3 classes of equal size, but the two classes are not well separable. At this point, it is useful to reiterate that our priority is not to obtain more optimal clusters, but to provide clusters of balanced size as input to the following 'group formation'' stage.

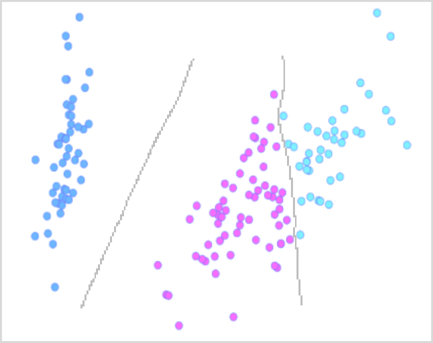
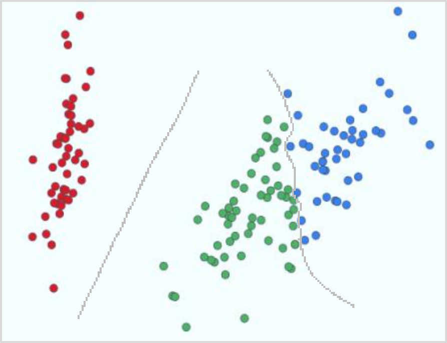
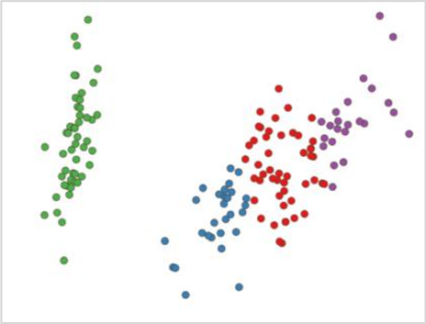
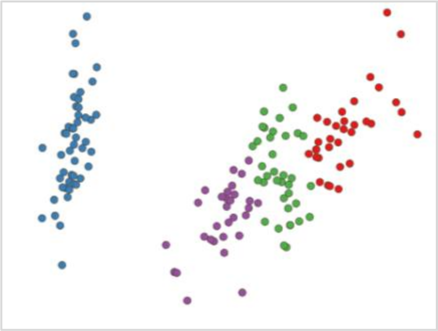
	K-means		Fuzzy c-means	
k=3				
	Within cluster variance	11.57 %	Within cluster variance	11.68 %
	Between cluster variance	88.43 %	Between cluster variance	88.32 %
k=4				
	Within cluster variance	8.42 %	Within cluster variance	8.50 %
	Between cluster variance	91.58 %	Between cluster variance	91.50 %

Figure 6-11 The comparison of K-means and Fuzzy c-means in terms of WCSS and BCSS. The algorithms were applied on the *Iris dataset* with different cluster numbers (k) and the clusters encoded into hue.

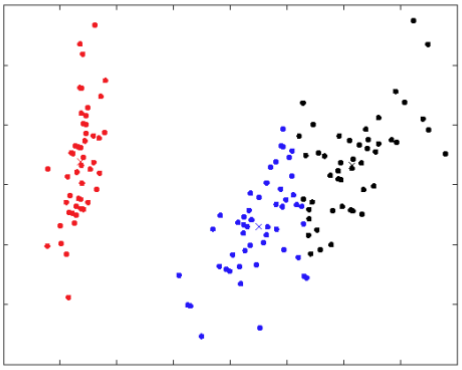
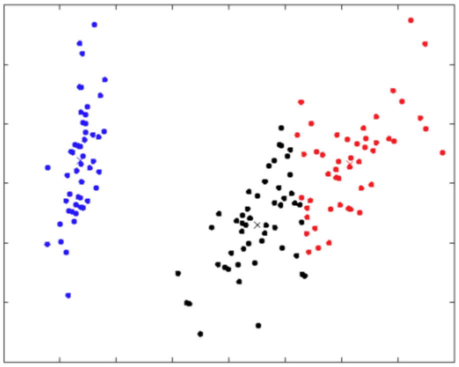
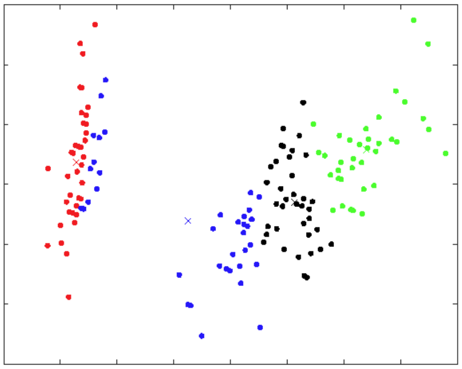
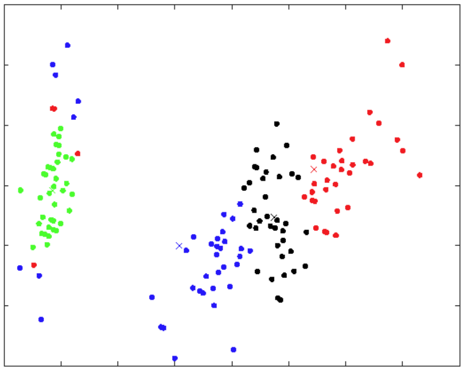
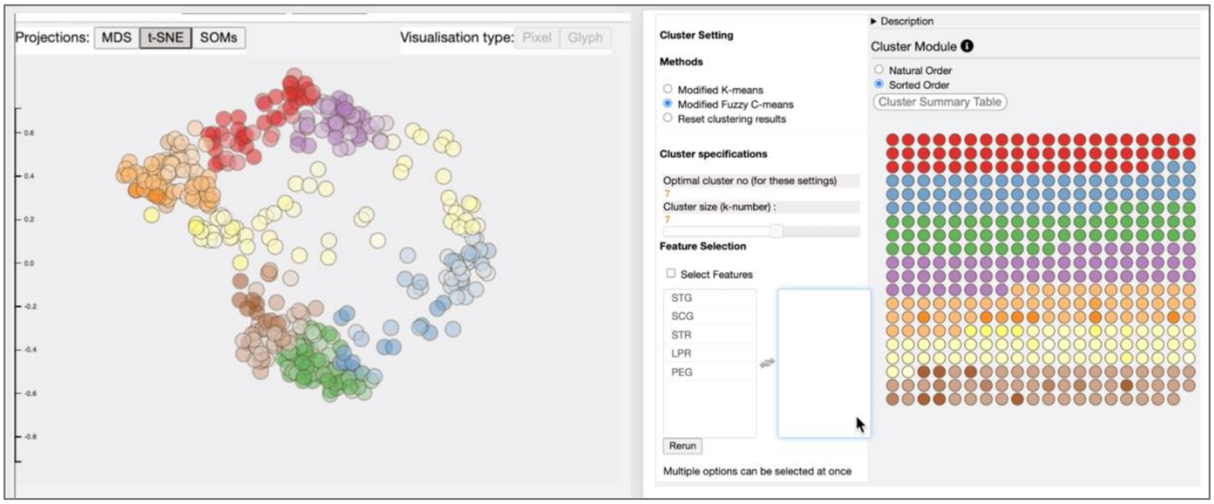
	Variation of K-means (roughly eq. size clusters)		Variation of Fuzzy c-means(roughly eq. size clusters)	
k=3				
	Within cluster variance	29.38 %	Within cluster variance	29.39 %
	Between cluster variance	70.61 %	Between cluster variance	70.60 %
	Iteration number for optimal solution	25	Iteration number for optimal solution	11
k=4				
	Within cluster variance	34.27 %	Within cluster variance	38.35 %
	Between cluster variance	65.72 %	Between cluster variance	61.64 %
	Iteration number for optimal solution	127	Iteration number for optimal solution	168

Figure 6-12 The comparison of the variations of K-means and Fuzzy c-means in terms of WCSS and BCSS. The algorithms were applied on the *Iris dataset* with different cluster numbers (k) and the clusters encoded into hue.

6.4 USE CASE STUDY

This section demonstrates how the Clustering module can be used to detect student clusters and the characteristics of the clusters. For this aim, the User Knowledge Modelling dataset was selected. The data set includes five learning attributes (*see Appendix -A*) that presents the students' knowledge status about the subject of Electrical DC machines.

After analysing the data set in Attribute module, she wants to analyse student clusters in dataset using one of the given clustering algorithms, and decided to use Fuzzy-c means algorithm. The algorithm requires the number of clusters to be specified, and so she continued with the optimal cluster number (k = 7) the tool provided for the dataset (for the five quantitative attributes), and also she used all attribute in data set, and applied the algorithm on them, in Figure 6-13.



Projections: MDS [✓ t-SNE] SOMs

Visualisation type: Pixel Glyph

Cluster Setting

Methods

○ Modified K-means
● Modified Fuzzy C-means
○ Reset clustering results

Cluster specifications

Optimal cluster no (for these settings)
7

Cluster size (k-number) :
7

Feature Selection

□ Select Features

STG
SCG
STR
LPR
PEG

Multiple options can be selected at once

▸ Description

Cluster Module

○ Natural Order
● Sorted Order

Figure 6-13 Cluster module. Left: t-SNE scatterplot showing clusters formed, encoded in colour scheme. Right: Cluster settings including algorithms, feature set area, cluster selection slider, and the Grid view.

The module showed the clusters formed in the Scatterplot and Grid view (Figure 6-13). In the grid view, she observed the organized clusters and their members, and the same colouring scheme is used at both Scatterplot and Grid view. When looking at the t-SNE scatterplot she observed that the clusters which are similar to each other in terms of their properties are located next to each other. To understand the similarity ratio between these clusters, he examined the Similarity view (Figure 6-14). With the help of the Similarity view, she compared the clusters formed visually and quantitatively, and found that the clusters located next to each other on the Scatterplot are also coded as two nodes, close to each other on this node-link graph, with a short distance between them.

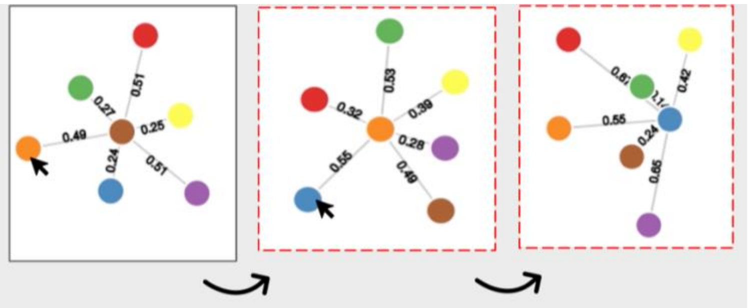


Figure 6-14 The length of links between the nodes depicts the similarity ratio. Each node colour represents the relevant cluster. Based on the selected cluster (placed in centre of the graph), the similarity score is calculated for each cluster. When clicking any cluster node in the graph, the selected one will be the centre node.

To understand the properties of these clusters and their differences in terms of the attributes, she opened the Cluster Summary view (Figure 6-15).

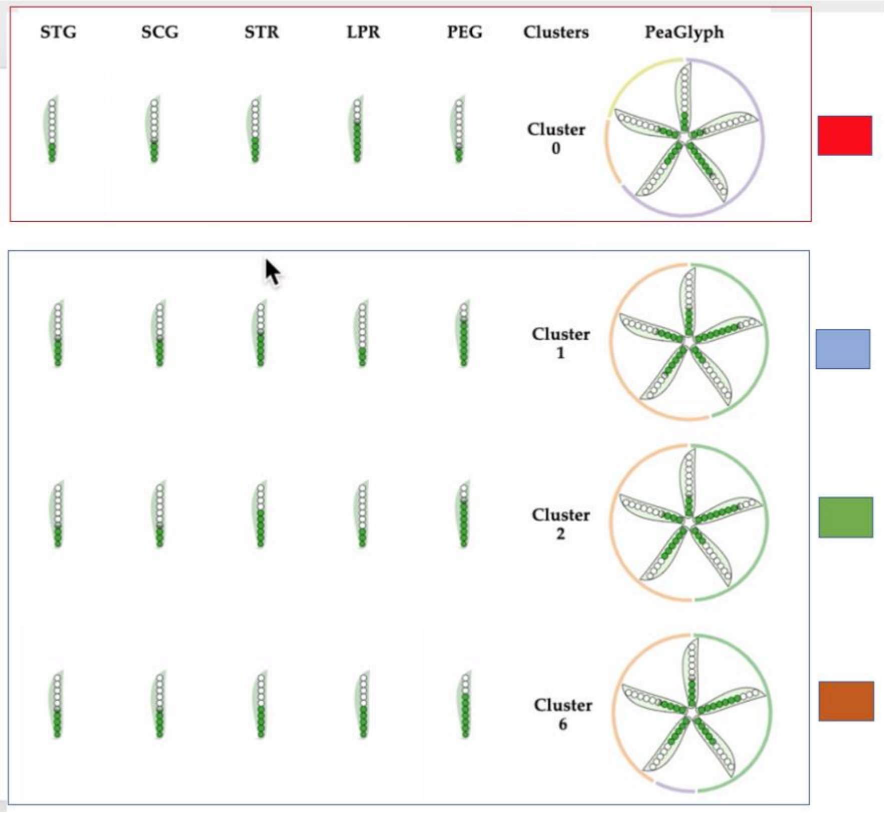


Figure 6-15 Comparing the clusters of interest through PeaGlyph. The clusters in Blue, Green, Brown having similar properties when compared to the Red-coded cluster.

By looking at the summary glyph, she got an overview of the members (students) of each cluster, as each summary glyph reflects the average values of its members' attributes. When compared the clusters' attributes coded in the PeaGlyph, she observed that LPE and PEG attributes in the clusters in red are quite different in the clusters in blue, green, and brown. There are also slightly small differences among other. Then, she found the students in the red coloured- cluster are overall good in terms of LPR (the exam performance of user for related objects with goal object) but not good in PEG (the exam performance of user for goal objects) when compared with the other clusters. Based on the observations, she understood she would customize her teaching materials based on the clusters. As an example, she would prepare additional materials to help the students in relevant clusters improve their lower-valued qualifications. She thought that the balanced (heterogeneous groups) would be formed by mixing the students in the red coloured cluster with the students in the blue, green and brown clusters, so that they can complement each other while working together. Also, she would divide the clusters into smaller homogeneous groups or combine similar groups to form a larger homogeneous group. In this way, she comprehend how Cluster module serves as an input to the Grouping module and the rationale behind the grouping algorithms (homogeneous and heterogenous) presented in the thesis.

6.5 CONCLUSION

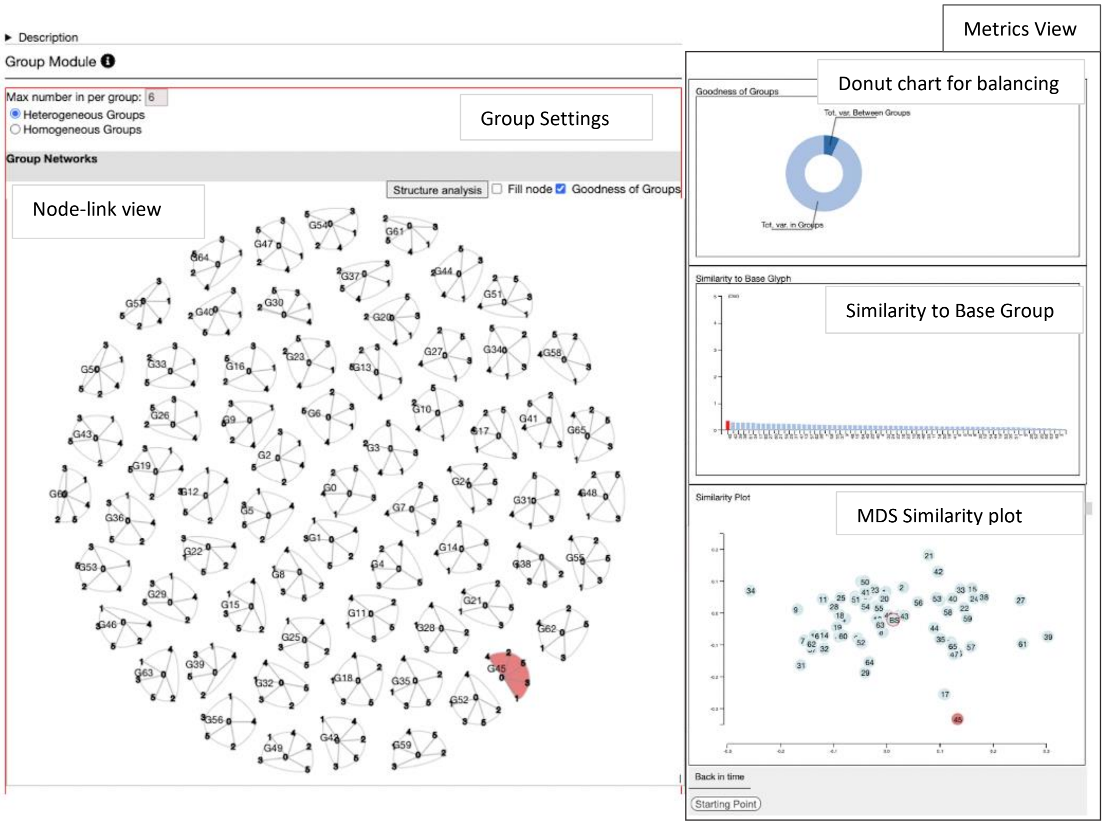
Clustering analysis is often used in a variety of areas because of its usefulness for exploring a dataset without necessarily requiring ground truths from the data. As part of the GroupVis, the Cluster module was introduced in this chapter so that target users could grasp the clustering process and explore the balanced size clusters before proceeding to the group creation process, which is based on the clusters gathered from this module. The approach provided supports interactive cluster exploratory analysis with multiple coordinated views that make the analysis more intuitive. The scatterplot provides a good overview of clustering results, and the Grid view presents the scatterplot instances in an organized way to deal with the overlapping dots. The PeaGlyph was also leveraged as a complement to augment views by providing extra information as well as making the comparison of clusters and their relevant attribute values easier. Besides the visualization techniques, two partitioning clustering methods, K-means and Fuzzy-c, which take into account size constraints, were applied to generate clusters with balanced dimensions. However, despite the fact that what defines a good clustering is typically determined by domain-specific and application-specific factors, two objective metrics were added into the system in order to assess the quality of clustering results. The system also includes another metric that guides users to select an appropriate clustering number based on the underlying feature set. The approaches shown here would be evaluated with datasets of varying sizes and types, as well as with multivariate datasets. If, for example, the offered methods are incompatible with the hierarchical dataset, it may be necessary to incorporate new clustering methods into the system. This is on our to-do list as a future work. Moreover, future research includes expanding the visualizations to cope with different multivariate data sets, including temporal and hierarchical, and performing user studies to evaluate the visual methods in this study for the clustering analysis task.

Chapter 7 : GROUP FORMATION THROUGH COMBINED INTERACTIVE AND SEMI-AUTOMATED APPROACHES

This chapter presents the creation of different types of groups by combining interactive and semi-automatic approaches. The module consists of ‘*Grouping settings*’, ‘*Node-link view’*, *'Structure group analysis view’*, and ‘*Metrics view*’. Each will be discussed in detail in the respective subsections.

7.1 INTRODUCTION

The third module of the framework is Group Module, as seen in Figure **7-1**, which splits individuals from the equally sized cluster into structurally defined groups to facilitate collaboration.



► Description

Group Module

Max number in per group: 6

● Heterogeneous Groups

○ Homogeneous Groups

Group Networks

Structure analysis ☐ Fill node ☑ Goodness of Groups

Figure 7-1 Overview of group view. The main view of Group module includes 3 parts, Group setting, Node-link view, and Metric view. Group setting allows to change Group type and appearance of given views. Node-link view represents groups formed. Metric view includes 3 different metrics to visually show the goodness of formed groups in terms of homogeneity and heterogeneity. (User Knowledge Modelling data was used.)

The groups formed are represented as node-link graphs where nodes depict group members, and edges show the relationship among group members. As there is no data attribute describing the relationships of members, equal size undirected links are utilized for encoding. All formed groups are represented with a force-layout diagram in which each group is an individual node-link graph in Figure 7-2. There is a central node and the other group members are aligned circular around the centre node, as shown in Figure 7-3. The centre node (member) for each group is randomly selected among the members of the groups. Although the tool does not aim to assign any role to the members in this version, the centre node may be used to map the team leader of a group.

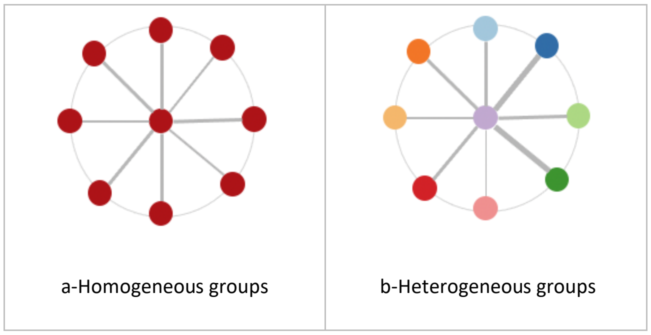


Figure 7-2 Representing two types of groups supported by the software. The groups are encoded in node-link graph where nodes are group members and the width of links (edges) between the centre node and other members representing relation (similarity) – The thicker edge, the more similar member pairs in terms of their attributes.

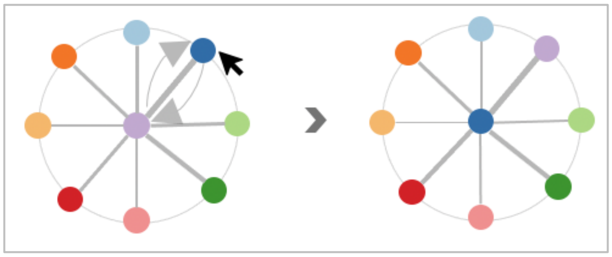
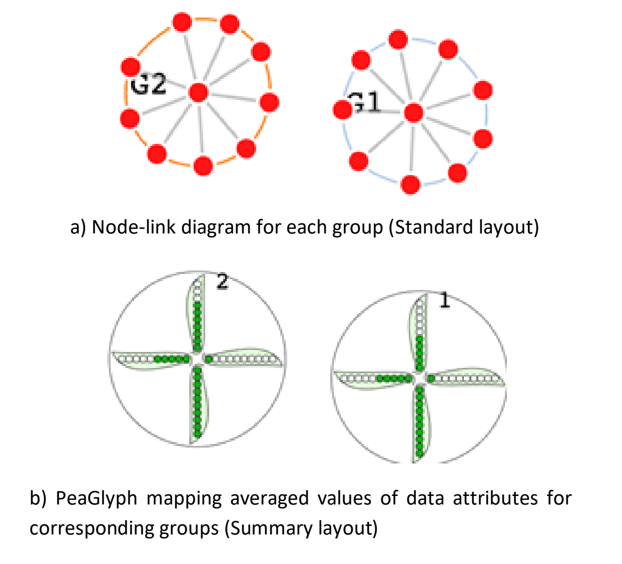


Figure 7-3 Making the node (member) of interest a central node. in this way, users can easily compare members in a group with the selected member (centre)

The mean of attribute values of each formed group is mapped to PeaGlyph, as shown in Figure 7-4 (b), due to its ability to the comparison of attribute values as well as in the judgement of balancing/unbalancing among attribute groups (*this meets **C2** on page 59*).



a) Node-link diagram for each group (Standard layout)

b) PeaGlyph mapping averaged values of data attributes for corresponding groups (Summary layout)

Figure 7-4 Node-link graphs with relevant PeaGlyph. Two groups are formed with 10 members. Each member has 4 attributes. At the bottom figures, the attribute numbers are mapped to pea pods, and the averaged values of attributes within relevant groups are mapped to peas, which are abstracted as full-filled circles, semi-filled circles or empty ones. The averaged values are scaled and approximated with 10-dots.

Node-link diagrams can express a wide of information types. They are widely used to map social networks[200] where human relationship matters, and the diagram abstracts particular entities and relationships among them. Similarly, a node-link diagram maps the overall group members and the interrelationships (i.e. similarity) of group members in our case. As an alternative, matrix-based representation could be preferred; however, the node-link diagrams are more readable and familiar than matrix representations for small graphs [201].

For each group, a PeaGlyph is generated to summarize the group attributes, and in this way, the representations complement each other. The position of a node-link diagram depicting a group in the *Standard graph layout* in Figure **7-4** (a), also indicates the position of the PeaGlyph related to that group in the *Summary layout* Figure **7-4** (b). In this way, users easily 'lookup' [15] the groups of interest quickly as they already know both what they are looking for and where is it. Also, both layout representations allow users to compare the groups from different aspects (i.e. the relations of members, the attributes of the groups, balancing among the groups).

7.2 GROUPING ALGORITHMS

Two different algorithms were built on the output of clustering algorithms, and based on the user preference heterogeneous and homogeneous groups can be generated.

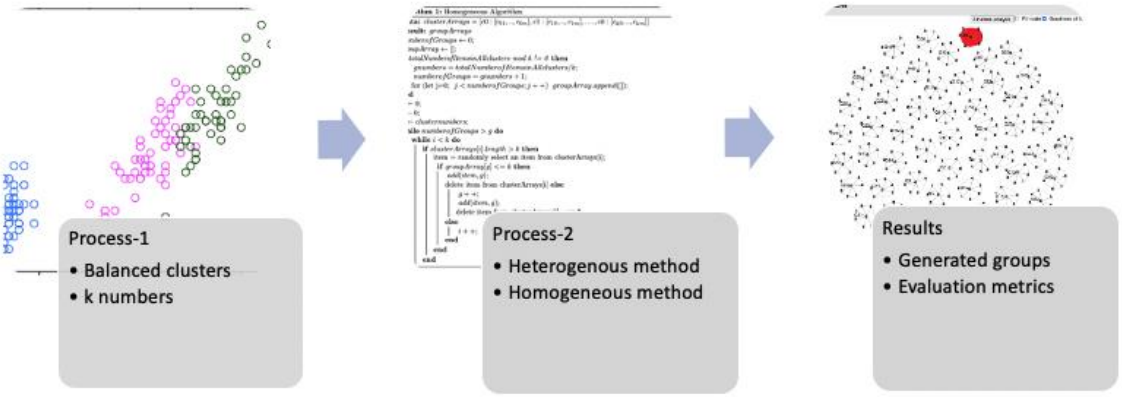


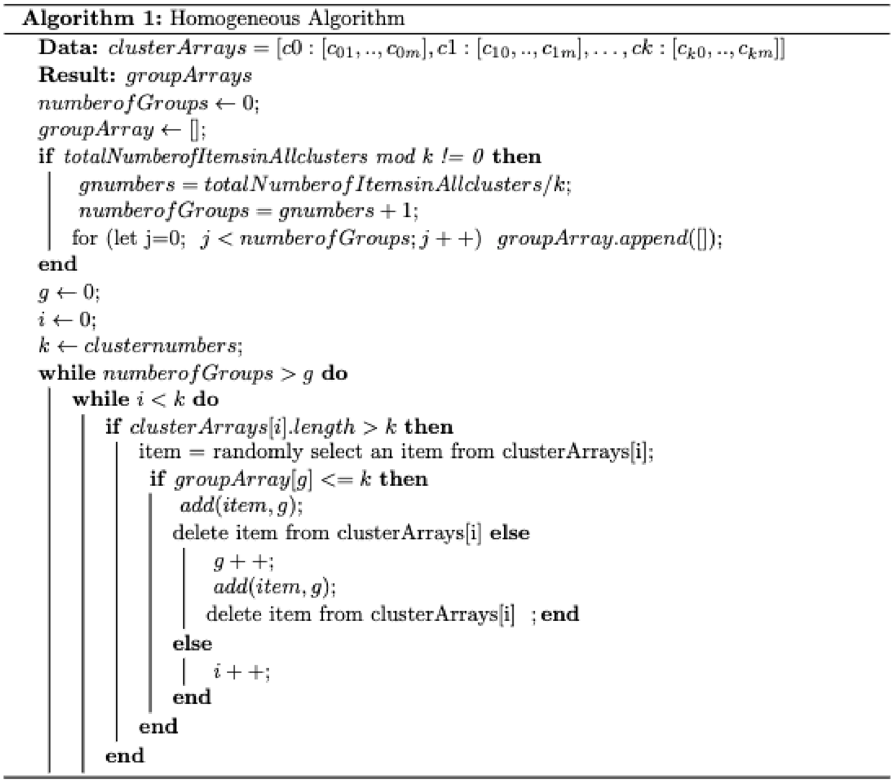
Figure 7-5 The workflow: the clustering outputs are transmitted to the group module

This workflow in Figure 7-5 summarizes the processes between the Cluster module and Group module *(this meets **C1** on page 58).* The generated clusters, including a number of clusters, are provided as an input to the grouping algorithms (i.e. heterogeneous and homogeneous) for obtaining the desired type of groups and each group will include roughly (k number of members), as shown in Algorithm 7-1 and Algorithm 7-2. Finally, the goodness of formed groups is evaluated with a set of metrics, including highlighting 'problematically formed groups' to get the attention of target users in further exploration. The degree to which groups are homogeneous or heterogeneous may be characterized a group feature. According to [202], the definition of heterogeneity,

> 'Heterogeneous group is a group where all the possible values of the learner space are present.'

In the educational context, a homogeneous group is a group in which the members are similar to one another in several attributes according to American Psychological Association [203]. When more than one criterion is used, it becomes even harder to define group heterogeneity. In order to reach some level of heterogeneity, a random procedure may be useful, especially if the level of group heterogeneity does not need to be the highest feasible.

In the present implementation, the rationale for forming heterogeneous groups is to pick a member from the clusters to provide desired heterogeneity of the groups, considering the quality of each formed group is relatively similar. However, homogenous groups are formed by dividing each cluster into several more homogeneous clusters. The output of group formation algorithms includes groups, their members, and relationships between them. For visually depicting these outcomes, a traditional node-link diagram was used where the members are connected to each other in a group, but there is no connection between groups by default.



```
Algorithm 1: Homogeneous Algorithm
  Data: $clusterArrays = [c0 : [c_{01}, .., c_{0m}], c1 : [c_{10}, .., c_{1m}], \ldots, ck : [c_{k0}, .., c_{km}]]$
  Result: $groupArrays$
  $numberofGroups \leftarrow 0$;
  $groupArray \leftarrow []$;
  if $totalNumberofItemsinAllclusters \bmod k$ != 0 then
    $gnumbers = totalNumberofItemsinAllclusters / k$;
    $numberofGroups = gnumbers + 1$;
    for (let j=0;  $j < numberofGroups; j++$)  $groupArray.append([])$;
  end
  $g \leftarrow 0$;
  $i \leftarrow 0$;
  $k \leftarrow clusternumbers$;
  while $numberofGroups > g$ do
    while $i < k$ do
      if $clusterArrays[i].length > k$ then
        item = randomly select an item from clusterArrays[i];
        if $groupArray[g] <= k$ then
          $add(item, g)$;
          delete item from clusterArrays[i] else
            $g++$;
            $add(item, g)$;
            delete item from clusterArrays[i]  ; end
          else
            $i++$;
          end
        end
    end
```

Algorithm 7-1 Specified for generating homogeneous groups

In this homogeneous algorithm, the formed clusters are presented as input data to the algorithm. After the total number of groups to be formed is determined, homogeneous groups are formed, as seen in Figure **7-6**, by randomly selecting elements as much as the number of clusters from each cluster. This process continues until there is no unassigned data instance.

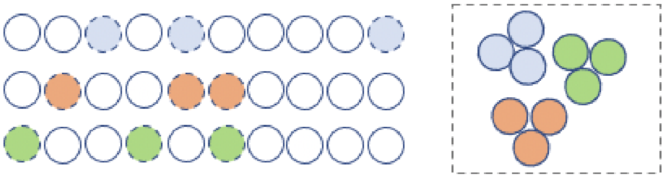
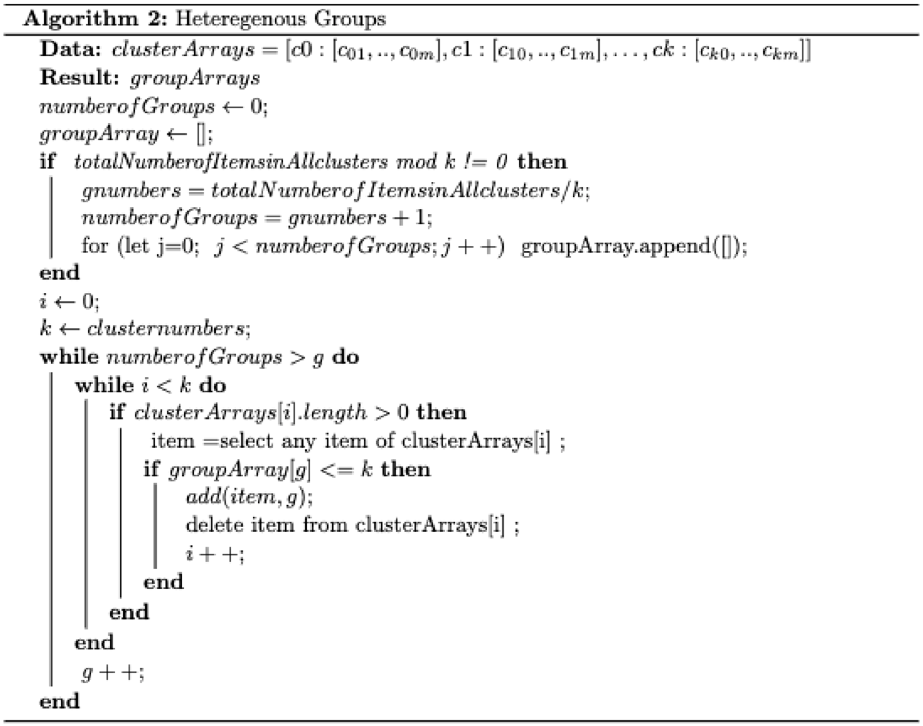


Figure 7-6 The homogeneous group algorithm divides the clusters horizontally (randomly selected data items) to form groups with similar members within themselves.



```
Algorithm 2: Heteregenous Groups
Data: $clusterArrays = [c0 : [c_{01}, .., c_{0m}], c1 : [c_{10}, .., c_{1m}], \ldots, ck : [c_{k0}, .., c_{km}]]$
Result: $groupArrays$
$numberofGroups \leftarrow 0$;
$groupArray \leftarrow []$;
if $totalNumberofItemsinAllclusters$ mod $k$ != 0 then
    $gnumbers = totalNumberofItemsinAllclusters/k$;
    $numberofGroups = gnumbers + 1$;
    for (let j=0;  $j < numberofGroups; j++$)  groupArray.append([]);
end
$i \leftarrow 0$;
$k \leftarrow clusternumbers$;
while $numberofGroups > g$ do
    while $i < k$ do
        if $clusterArrays[i].length > 0$ then
            item =select any item of clusterArrays[i] ;
            if $groupArray[g] <= k$ then
                $add(item, g)$;
                delete item from clusterArrays[i] ;
                $i++$;
            end
        end
    end
    $g++$;
end
```

Algorithm 7-2 Specified for generating heterogeneous groups

Heterogeneous groups are formed by randomly selecting one element from each cluster. Since each group will contain as many elements as the number of clusters, the groups to be formed, as seen in Figure **7-7**, by taking elements from each cluster will provide heterogeneity within groups and a balance between the groups to be created.

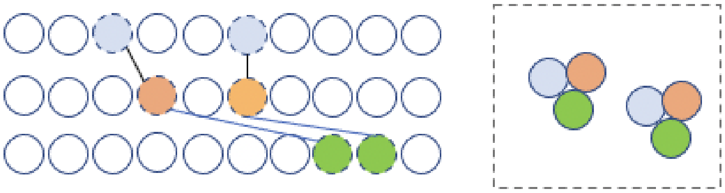


Figure 7-7 The heterogeneous group algorithm divides the clusters vertically (randomly selected data items) to form groups with different members within themselves, ensuring balance between the groups.

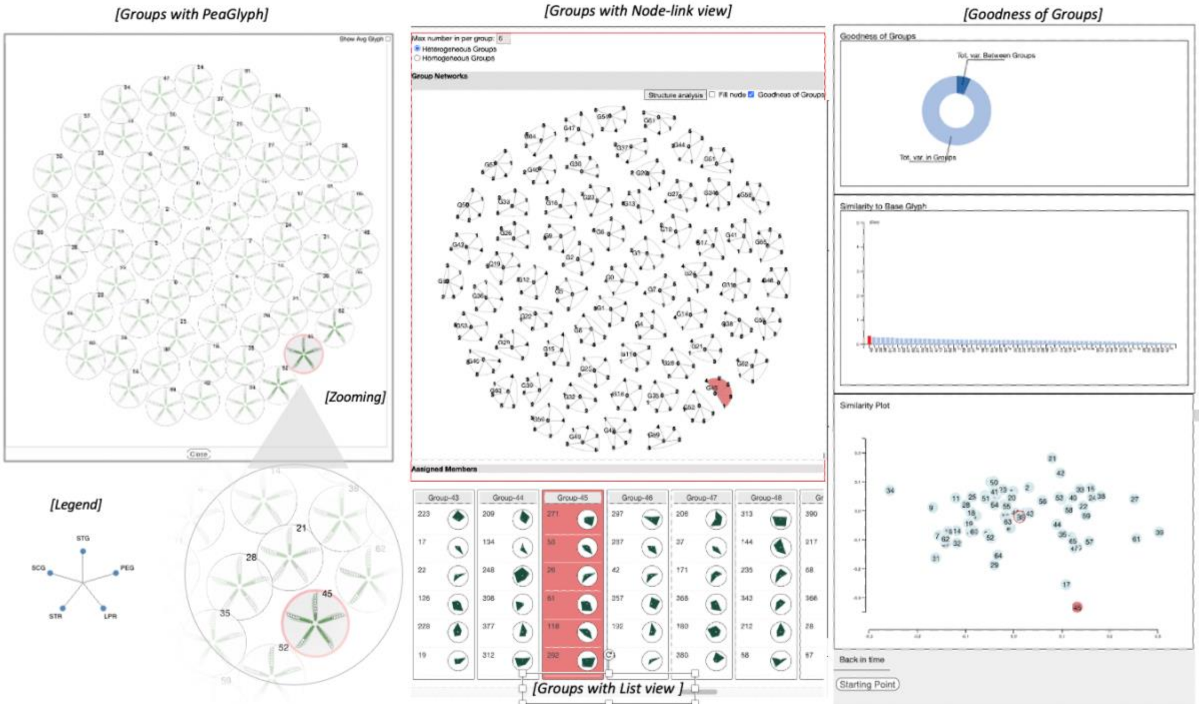


Figure 7-8 Structure view (left): The summary of formed groups is encoded into PeaGlyph Group layout (middle): The formed groups are depicted as Node-link graphs. List view (bottom): The formed groups are given in the List format. Metric view (right): It visually depicts the goodness of formed groups in terms of homogeneity and heterogeneity. The groups highlighted in red are detected as outliers by the metrics.

Besides the visualization of the network of the formed groups, the listed view represents groups with the group members, as shown in Figure 7-8. The view was considered as a locomotive for all formed visualizations. This view enables users to interactively exchange individuals between groups *(this meets **C3** on page 59).* The changes within the List view are reflected in the charts that encode the metric results, as well as the relevant visualizations in the module.

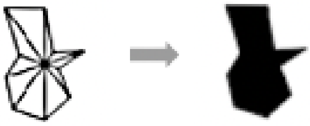


Figure 7-9 Representing star glyph with contour at the left, and the shape filled on the right by Fuchs et al.[117]

Each group member is coupled with a polygon shape (filled star-glyph shape), in Figure **7-9**, in the List view, and the shape reflects the underlying data values hence the data instances with similar data attributes have similar shapes as star glyphs are commonly used for similarity search and grouping analysis task [117, 204]. The filled star glyph is employed for a rapid overview and detecting similarities, not for providing precise values of data attributes.

7.3 GOODNESS OF GROUPS

Two different metrics are presented to measure the quality of heterogeneous and homogeneous of formed groups.

7.3.1 *ANALYSIS OF GOODNESS OF GROUPS*

The statistics here aim to measure the quality of the formed groups in terms of 'heterogeneity and homogeneity'. More specifically, given heterogeneous groups, the dispersion within-group is expected to be high; however, it is low for homogeneous groups because such groups consist of similar members in attribute space.

Suppose we have *k* groups; we use the index *j* for these, and each group consists of a sample of size *nj*. For the sample elements the index *i* will be used. Then, the total sample consists of all the elements :

$$\left\{ x_{ij} \colon 1 \ \leq i \ \leq \ n_{j,} \, 1 \ \leq j \leq k \right\}$$

The abbreviation $\bar{x}_j$ will be used for the mean of the *j*-th group sample (called the group mean) and $\bar{x}$ for the mean of the total sample (called the total or grand mean).

Let the sum of squares for the *j*-th group be;

Equation -11.

$$SS_j = \sum_i \left(x_{ij} - \ \bar{x}_j\right)^2$$

Equation -12.

SS$_T$ is the sum of the squared deviations from the grand mean for the total sample;

$$SS_T = \sum_j \sum_i \left(x_{ij} - \ \bar{x}\right)^2$$

Equation -13.

SS$_B$ is the weighted sum of the squared deviations of the group means from the grand mean;

$$SS_B = \sum_j n_j \left(\bar{x}_j - \bar{x}\right)^2$$

Equation -14.

SS_W is the sum of the squared means across all groups

$$SS_W = \sum_j SS_j = \sum_j \sum_i \left(x_{ij} - \bar{x}_j\right)^2$$

The relationship between the three types of sum of squares can be summarized by the following equation;

Equation -15.

$$SS_T = SS_W + SS_B$$

Finally, we define the following degrees of freedom (df) and the mean squares (MS) as:

Equation -16.

$$df_T = n - 1 \text{ and } MS_T = SS_T \frac{1}{df_T}$$

$$df_B = k - 1 \text{ and } MS_B = SS_B \frac{1}{df_B}$$

$$df_W = n - k \text{ and } MS_W = SS_w \frac{1}{df_w}$$

The donut view, in Figure **7-8**, shows the degree of heterogeneity or homogeneity of the formed groups by providing the proportion of the total dispersion within the groups against the total dispersion between the groups.

Proximity of groups to overall (grand)

Imbalance between formed groups is measured depends on the distance of their attributes to the grand average, in Figure **7-10**, The data distance is found using the Euclidian metrics. For quantitative attributes, the average values of quantitative attributes across all team members would be identical to the population-wide average values.

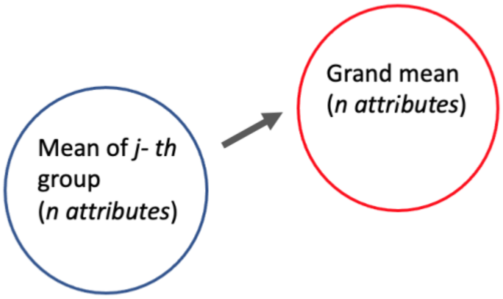


Figure 7-10 The mean of ***JTH*** group sample are compared with the grand mean, thereby determining the similarity of the group averaged values to the population-wide averaged values. In this way, the groups are positioned relative to the reference point (grand mean) on the Similarity plot in Figure 7-8.

An imbalance in quantitative attributes in a team can be measured numerically by determining how far the team average differs from the population average [205].

We have k groups, j index is used: $\{ x_{ij}: 1 \leq i \leq n_{j,} 1 \leq j \leq k\}$

$\bar{x}_j$ will be used for the mean of the j -th group sample, and $\bar{x}$ for the mean of the total sample (grand mean). Then, the similarity between $(\bar{x}_j)$ and grand mean $(\bar{x})$ is calculated with following equation;

Equation -17.

$$d\ (\bar{x}_j\ , \bar{x}) = \sqrt{\sum_j (\bar{x}_j - \bar{x})^2}$$

Based on the similarity measurements of the team to *grand mean* $(\bar{x})$, the teams differing from the majority of a set of the team are found, and the system identifies them as outliers *(this meets **C4** on page 59),* and through using colour highlighting, the system makes the information stand out and get the attention of users. In particular, the ‘outliers’ are more or less than 2.5 standard deviation distance from the grand mean. The distance values are sorted in descending order, and the Bar chart view encodes these sorted distances in length channel. The scatterplot of MDS for multivariate distance data takes place alongside the Bar view. In the graph, the point representing *grand mean* is positioned in the middle zone of the plot. Other items representing the groups formed are positioned relative to the grand mean. Some interactive elements, such as highlighting and zooming with panning, are supported by the tool, allowing users to communicate with information that has been encoded in a visual format. Besides, the web utilities (i.e., sliders, checkbox, etc.) and mouse-over and single-clicking behaviour, commonly used in web-based tools, are available as interactive options to complete the tasks. Grouping module logs each operation performed by users while swapping groups members, enabling them to undo/redo a single operation. Further, the linked views paradigm is used. In this way, several views are linked together so that when to be interacting with one view, the other views will update and show the results of such an interaction. For example, user can exchange members among the formed groups by using the *List view*. The changes made on groups are reflected on the *Group view* and the *Structure view*, as well as the metrics scores are updated according to the changes, so they are observed on the *Metrics view*.

7.4 CONCLUSION

In this chapter, the overview of the GroupVis framework was presented. The design criteria on which the GroupVis tool is based have been created to make the group formation process systematic and explorable for target users. This framework includes three main task-centric components: the Attribute module, the Cluster module, and Grouping module, whose computational and visualization methods were explained in detail in this chapter. The next chapter will present the heuristic evaluation of the three main components of the GroupVis framework, which have been explained in detail in the Chapters 5, 6 and 7.

Chapter 8 : HEURISTIC EVALUATION OF GROUPVIS FRAMEWORK

8.1 INTRODUCTION

A thorough examination of tools, including their visualisation techniques, is critical for providing effective support to tool users, and ensure its usefulness and usability. Performance and subjective characteristics are frequently cited as relevant indicators of data visualisation usability [206]. The usability of the software presented in this thesis was evaluated by considering heuristic criteria. Heuristic evaluation methods [207] are widely used which allows finding potential problems in user interface, and prioritizes the problems to tackle [208, 209]. Recently, Wall et al. [210] proposed a heuristic technique for quantifying the potential benefit of a visualisation in terms of data comprehension. Their methodology ICE-T includes four value components: *Insight, Confidence, Essence* and *Time.* The participants are expected to evaluate the visualizations using the individual heuristics within each component. Each heuristic for a visualization has a 7-level ranking ranging from 1 -strongly disagree to 7 -strongly agree, and N/A -not applicable. A heuristic approach based on the methodology suggested by Wall et al, was used to evaluate the GroupVis, and that the evaluation was conducted through online interview and survey.

8.2 EVALUATION PROCESS

Data. We used the User Knowledge Modelling Dataset [44] to evaluate the functions of the software (GroupVis). The data set is described by 6 different attributes, one of which is categorical (the knowledge level of user), and the remaining ones are quantitative, namely "STG (the degree of study time for goal object), SCG (the degree of repetition number of user for the materials), STR (the degree of study time of user for related objects with goal object), LP (the exam performance of user for related objects with goal object), PEF (the exam performance of user for goal object)". The data set contains *403 instances*.

Recruiting Participants. *GroupVis* has been designed to be used mainly in colleges and universities, but that it could be usable in different settings as well. The first phase of the study was a demonstration of the tool functionality, and the second phase included a semi-structures interviews (*see Appendix -A*) and discussions where the participants had a chance to test out the tool and give feedback.

Demonstrating Process. The interviews were conducted with two college teachers of science and mathematics, four university lecturers from different domains and one education specialist, details shown in Table **8-1**. In the sessions, the participants took control of the software using the feature of the Microsoft Team tool, and used the software (GroupVis) to create study groups.

During the session, the participants were encouraged to explore the dataset and to generate clusters and visually explore them, as well as to create balanced groups and to identify groups that may be unbalanced compared to others. The dataset in CSV file format was uploaded into the software (web-based) using file upload library before starting the session. The evaluation of the tool was done online via Microsoft Teams and each interview session was recorded as this allowed us to fully focus on interview rather than note taking. The individual interviews took on average 40 minutes. The participants provided their feedback on the views and functions of the GroupVis. After the interview session, each participant was asked to complete the visualization heuristics survey and return it to us. The survey (see Appendix-B) includes 19-heuristic items to be rated via Likert scale and also provided the option of typing comments about each heuristic.

Participants	Subject
P1	Mathematics (Secondary education)
P2	Science (Secondary education)
P3	Statistics (University)
P4	Computing (University)
P5	Assessment and evaluation (University)
P6	Educational Technology (University)
P7	Specialist in education (Department of Education)

Table 8-1 The participants' codes with their domains

Interpreting and Reporting. One of the aims was to explore the potential benefits of GroupVis to educators in the context of forming balanced groups, as well as quantify its usability. The questions of the interview were guided by these goals. Participants were interviewed about their own experiences with cooperative learning, how to use GroupVis in a classroom setting, its perceived usefulness, their grasp of visual approaches, and their feedback on the tool's ease of use and areas for development. The analysis was shaped by codes as given in Table **8-2**, based on research and the interview questions.

Themes:	Collaboration	Usability	Visualizations	Computational methods
Codes:	Education, grouping, team	Interface, colour(s), ease to use, intuitiveness, guidance, wording	Scatterplot, Grid views, PeaGlyph, Node-link diagrams, metric views, SOM map	MDS, t-SNE, SOM, k-means, fuzzy c-means, Homogenous groups heterogenous groups

Table 8-2 The Coding sets for analysing the interview transcripts

The themes and codes were specified related to the research goal prior to the interview, and these created a framework of the interview questions. Then, the excerpts that fit the codes were found in the interview transcripts.

Collaboration learning: At the beginning of the interview the participants were encouraged to share their own experiences and views on educational approaches, including group-based learning. All of them agree with the statement that students become better problem solvers in better formed groups. They used collaboration method if the team project is as part of their courses. They asked their students to form groups of 8-10 people. One lecturer (P5) allows larger study groups (i.e., more than 10). They did not use any specific method or tool for creating groups even though they admit that their approach is problematic from a pedagogical point of view.

The participants provided feedback and comments that demonstrate the usefulness of the system, as well as indicating potential areas of improvement. The following sections summarize the main comments made by participants. The heuristic evaluation form used in the thesis can be seen in *Appendix -B*.

Overview: P3 described using the tool : “the application can be especially useful for classroom use. The versatile display of the data make it more useful’. Similarly P1: “I enjoyed using the tool and it is presented with an easy interface.” P6 said that “I have found it easy to use. I believe that it can be easily used when creating groups, especially experimental studies”. P4 ‘if you have a data set and want to create groups; and make sure your groups are competitive, it will quickly create the groups.’

P2 and P7 said that they struggled with the technical names or abbreviations at first. After the demonstration, P7 said ‘it became clear’. Also she stated that a short tutorial could be prepared for its use in classroom. Likewise, P6 said : ‘some abbreviations (i.e. MDS and t-SNE) are not understandable’.

P2 suggested an alternative usage of the tool: “When I consider the use of this tool in the classroom, maybe it can give us insights about the readiness of students. I think it could definitely help if you have got a new class and you do not know the students in detail there. This would actually give you a visual way of seeing students' strengths and weaknesses". P1 described "As a data analysis program it is superb'.

Attribute Module

(see Appendix -B)

Self-organizing map. P2, P5, and P7 found Self-organizing maps confusing compared to other views. When we explained what the method does and its structure, they said it helped them for seeing of data groups. P7 said: 'the colours made it easier to see them (the groups formed)'. In addition, all of them found the use of star shape and SOM together more informative than SOM with an empty cell.

Grid view. All participants found the Grid view useful for comparison of multiple data entities regarding attributes.

Quality metrics. The participants found the semi-automated guidance quite useful, except P1 who said that 'I do not think the semi-automated guidance would be entirely necessary for what the tool is being used for unless it was for a management perspective and managing lots of groups.'

Cluster Module

(see Appendix -B)

Colour scheme. All participants liked colour scheme used for showing clustering results. For example, P1 said 'when I look at the colours, I can deduce that they have different entities. However, P4 stated that 'I thing providing various colour options to be selected by the users who have colour vision deficiency can make it accessible for'.

PeaGlyph. All participants agreed that clustering summary table alongside Node-link diagram is quite easy to understand. When we showed them the table with the numbers instead of the pea glyph, they preferred the pea glyph version. *A quote from* P6 , ' the cluster summary table allows you to compare clusters' and added 'the pea-like shapes made values easily comparable when compared to the numbers in another table.' P4 was the only one who commented on the node-link diagram. He mentioned node-link shows the relationship between clusters in "a simple way", and it can serve as a jumping off point for further investigation.

Cluster setting panel. While testing the tool out, they were asked to select some features in the features of the test data set and then to run one of the clustering methods by choosing a certain number of clusters. As P2 said "I initially found the setting panel complex as there were too many options to be used". After trying it out, they *got familiar* *with the options*. P7 suggested that "giving a short description right next to the options can help choose one of the options".

Grouping Module

(see Appendix -B)

Settings. All participants referred to the value of the grouping options (homogeneous and heterogeneous). According to P7, "predictably, I think many teachers would find it to be really useful."

PeaGlyph. P1 stated that 'I can see how balanced attributes of across groups. Exactly, that's probably the best way of looking at it.' P4 said 'Apart from the tool, I think the pea-like shapes can be used to analyse whether the groups are balanced or not'.

Metrics (metric views). P7 stated that ''they convey the data in concrete way for the users like me who are not good with numbers". The participants were intrigued by highlighting problematically formed groups. P2 said: 'S*ee it and go for that grou*p makes the process quicker'. No one gave any negative feedback on them.

List view. The participants liked the list views, especially the ability of swapping the students between the group lists. As P7 said: I liked it because the final decision about students still belongs to instructors'. Similarly, P4 : "this is what I want to see in this kind of tool".

8.3 HEURISTIC EVALUATION

The ICE-T methodology was adapted for heuristic evaluation of the GroupVis tool. Some of the components were renamed to cover the heuristics in our evaluation set. Hence, the term 'Understanding' (U) for the Insight, Time (T), Intuitiveness (I) for the Confidence, and Essence & Guidance (E&G) instead of 'Essence'. The participants rated the visualizations using the individual heuristics within each component. The participants' ratings are broken down by the four value components. The summary ratings of the participants on the heuristics with respect to each of the components are shown in Table 8-3.

GroupVis received an overall cumulative score of 6.27, in which the maximum score is 7.0. In the original paper [210] of the method, visualizations with a score of 4 or less are identified candidates for redesign but a score of 5 or higher represents valuable and good visualizations. Figure **8-1** represents the participants' scores for each component. The component *E&G* scores relatively higher and followed by component *U* (Understanding). Overall ratings of the participants are consistent among the components. The participant P7 did not respond to the heuristic form.

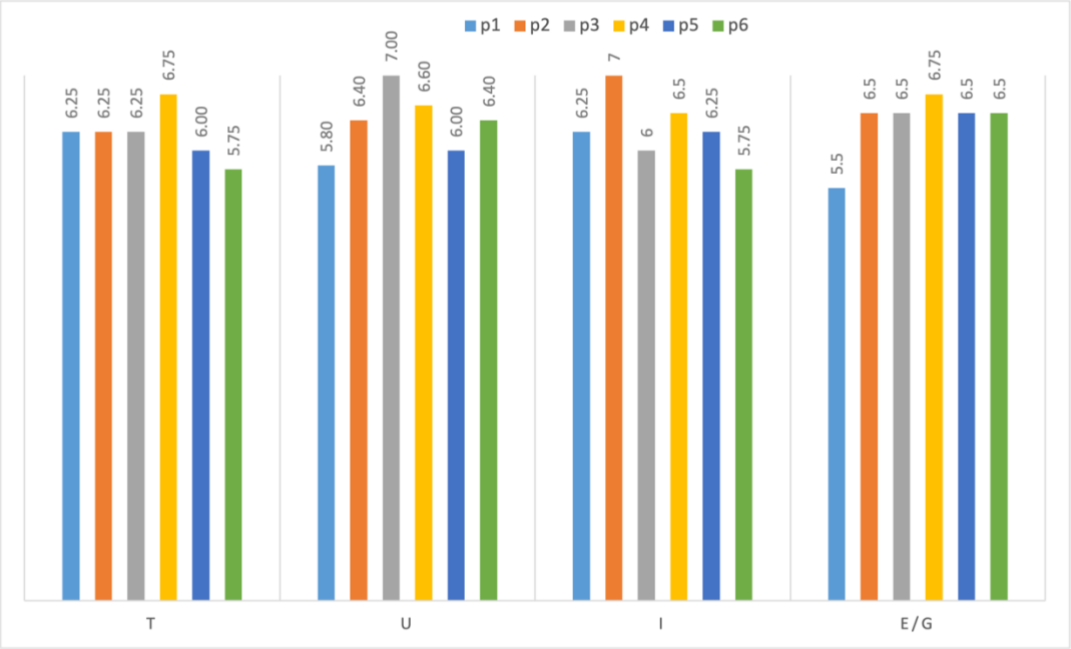


Figure 8-1 Showing average scores of participants' (p) ratings for each component. Also, it represents the consistency among participants' ratings. (T: Time, U: Understanding, I: Intuitiveness, E/G: Essence and Guidance)

The study results verified the efficiency of our design principles. The techniques given have facilitated data discovery and balanced group formation according to the quantitative results as shown in Table 8-3.

In the evaluation session, the most striking feature of the tool was that the groups which were relatively unbalanced were automatically highlighted by the tool so that the participants identified them easier. This was followed by the PeaGlyph that was reported as being useful in analysing the cluster features and formed groups, as the participants said it was easier with PeaGlyph to see the properties of the clusters and to compare the clusters on these attributes. They also used very similar expressions for this glyph design to analyse the created groups.

Also, the participants highly rated the interaction techniques supported by the tool and found them practical. However, there are some points indicated by the result to be improved. At first, the users found the SOM structure a bit confusing, which was revealed in the interview session we made with the participants. Another point that the participant evaluated as relatively weak is the transitions between views. In addition, as one of the participants did not fully understand the two heuristic items (*H7* and *H11*) in the form, he noted that he gave 4 points to these two questions.

Comp.	Heuristic items	µ	σ	max	min
Time	H1 -The tool interface provides a meaningful spatial organization of the data.	6.33	0.52	7	6
	H2- The visualization avoids complex commands and textual queries by providing direct interaction with the data representations	**6.67**	0.84	7	6
	H3- The visualization supports smooth transitions between the views	**5.66**	0.82	7	5
	H4- The polygon shapes for each member are intuitive and help users to notice similar/different members across groups at a glance.	6.33	0.52	7	6
Understanding	H5- Visualizations expose individual data cases and their attributes (e.g., you can easily see whatever the items in the dataset represent as well as easily compare their attribute values)	**6.67**	0.52	7	6
	H6- The visualization provides useful interactive capabilities to help investigate the data in multiple ways	6.50	0.55	7	6
	H7- The visualization helps generate data-driven questions	6.00	1.10	7	4
	H8- The visualization shows multiple perspectives about the data	6.17	0.41	7	6
	H9- The visualization provides useful interactive capabilities to help investigate the data in multiple ways.	6.50	0.55	7	6
Intuitiveness	H10- The representations in the cluster summary table are intuitive of what constitutes a cluster and which attributes differ among clusters	6.33	0.52	7	6
	H11- The relationship between Grid view and Scatterplot is intuitive	6.17	1.33	7	4
	H12- The representations of the metrics are intuitive.	6.33	0.52	7	6
	H13- The representations in the detail view are intuitive, and provide clear information about differences between groups	6.33	0.52	7	6
	H14- The Self organizing map has an intuitive structure	**5.50**	1.22	6	3
Essence & Guidance	H15- The dimension reduction view provides an objective indication of the quality of the plots that helps users choose from among the plots provided	5.83	0.75	7	5
	H16- The colouring of the SOM cells helps reveal cluster structures	6.17	0.75	7	5
	H17- PeaGlyph helps reveal balanced /unbalanced groups features as well as compare them.	6.50	0.84	7	5
	H18- Highlighting the groups in this view that were marked as outliers by the metric was a useful guidance for the starting point of the analysis.	**7.00**	0.00	7	7
	H19- The visualization provides a comprehensive and accessible overview of the data	6.33	0.51	7	6

Table 8-3 Including the four components, and the constituent heuristics for each component. The figure also shows summary (average) ratings for the three visualizations on each of the heuristics, as well as the standard deviation of each rating.

8.4 DISCUSSION

The participants offered their opinions on GroupVis' existing capabilities as well as suggestions for how the tool could be improved in the future. At the beginning, all participants stated that collaborative learning is beneficial, and they used this learning approach in their modules. They also agree that the way they form groups has a huge impact on the output of the groups. They found the GroupVis useful, as it helps to create different type groups of students.

According to the findings, the tool can be used for data analysis, alongside creating balanced groups. Besides, the visuals of the tool are intuitive (*this meets **C5** on page 59*) and the process-flow steps to form balanced groups is logical. Although some users expressed some concerns such as abbreviations, technical names, about the usability of the tool's interface, they understood the relevant points after a short presentation. Thus, we think to provide adaptive user interfaces [211] based on the users' profile or experiences in the future version of the software, as well as to prepare a tutorial for classroom use of the tool functionalities. Although the tool's main aim is to support educators for obtaining competitively balanced groups, the some participants expressed that the tool can be used for different tasks such as exploring student readiness. The PeaGlyph was the most prominent visualization in the interviews with the participants. They found it very useful both in comparing data values and in giving information about the balance. This is supported by the score of participants' score for it in Table 8-3. Similarly, highlighting unbalanced groups to draw users' attention to unbalanced groups was the tool's top rated feature, while SOM's intuitiveness scored the least average relatively. Since there is no score of 4 or less in the table, it can be said that the visuals are quite effective, which supports the feedbacks of participants in the interview sessions. Surprisingly, six of the participants did not want to use any tool or method in this context. Although they are aware of the existence of some methods, they are not willing to learn to use them. Thinking about the ease of use and the time it would save, they said that they could use this tool in their own modules to form student teams. As a limitation, as a general student dataset is used in this section to illustrate the functions of the tool, it will not be possible to interact with the students in the groups created by GroupVis to see their progress and experience. In our next agenda we are considering using our own dataset and analyse the experiences and results of the students in GroupVis-created groups compared to the students in non-GroupVis-created groups.

Chapter 9 : CONCLUSION

The research presented in this thesis has introduced a group formation framework that includes three main components: data exploration, clustering and grouping. These components, or modules, are ordered following the workflow of the users in the context of group formation. Each component is provided with relevant computational methods and visualization methods; which enables users to perform module-related tasks that as a whole leads to the creation of balanced groups in a more systematic and interpretable way. More specifically, the main focus of the work has been on the use of visualization methods, along with algorithms, to help establish 'balance' between groups as well as to explore the distribution within groups. Further, the combination of computational and visualization methods designed for group formation in such a way that domain knowledge of the users is made of use. This section will serve as an overview of the contributions, followed by discussions of the findings reached from the work and recommendations for future study directions.

9.1 OVERVIEW OF CONTRIBUTIONS

The contributions of this thesis haven been highlighted in the Chapters above. The following ones are the most significant research contributions made by this research study:

- *A novel glyph visualization for balancing/unbalancing structure in multivariate data*

  Chapter-3 presented PeaGlyph, which was designed based on established design principles and the results of a formal evaluation of four glyph designs. The chapter was published in the Information Visualization journal, under the title: PeaGlyph: Glyph design for investigation of balanced data structures. The performance of novel PeaGlyph shows that it is a feasible option for representing multivariate data and allows viewers to acquire an intuitive sense of how balanced or imbalanced a set of objects is. The glyph based visualization can represent features of aggregated data and can be used both as an enhancement to existing visualization methods, and as a stand-alone visualization method. Further, we also evaluated the effectiveness of the PeaGlyph in showing the balance and imbalance between groups formed in Chapter -8.

- *A group formation framework (including algorithms, visualization methods and measurements) built on top of the equal size clustering methods to generate heterogenous/homogenous groups*

The final software prototype is itself a research output, as the software was implemented as a web-based tool. Based on the findings in the balanced group formation literature, the issues and requirements were defined and in order to help surmount these issues as well as meet the requirements a group formation framework was proposed. The frameworks is made of three main components which are ordered as a user workflow for creating balanced groups in a more systematic manner. Balanced grouping is about the task of grouping individuals into proper teams to make teams similar across multiple attributes. The groups created as problematic are automatically detected and the user may be interested in examining further or changing members between the groups to obtain the desired distribution within the groups. Even while putting people in groups and making sure those groups are balanced and fair is not a new challenge, we believe this is the first time it has been approached from a visual way assisted by machine learning approaches, making it a significant contribution to the field.

- *By providing a new way of using student data sets (i.e. the learning data of students in software engineering module) to support learning and teaching in educational context*

  The application scope of digital technologies in various formats is expanding. The collection of data through the technologies offer an opportunity to solve real world problems. For example, data collected from sports games can enable us to monitor players' abilities, deficiencies, and make predictions about their performance for future games. Likewise, student data sets collected while using digital learning systems are frequently used to develop strategies or make decisions to optimize students' performance. In this research, the usefulness and ease of use of the tool by educators in process of creating balanced groups in cooperative learning was demonstrated. With help of the GroupVis, educators can create desired type groups, semi-automatically using the student attributes related to the task-based knowledge, or they can modify the automatically formed groups based on their knowledge about them. The formed groups are given in different visualization methods. The interactivity enables the users to make desired changes on the groups. Consequently, the tool as a whole provides a new way of using student data to support learning and teaching in education context.

9.2 DISCUSSION

This thesis has primarily concerned the application of information visualization and visual analytics approaches in the context of balanced group formation. As previously discussed in the thesis, it offers a framework in which balanced groups can be generated in more systematic way and the output groups are more explainable compared to the purely automated methods in this context, as the modules in the framework is supported by efficient visual methods. The GroupVis helps decision-makers both to understand why the computational algorithm end up with a decision and to analyse the individuals before engaging them in project groups. The findings in the evaluations in *Chapter VIII* shows that the tool can be used efficiently within the group formation context in educational settings. The users found it easy to use and the visualization methods along with computational methods are intuitive. The strengths and weaknesses of the methods of each component of the GroupVis were discussed in the relevant sections.

Apart from the matters mentioned in the chapters, it is an important issue to provide the data in an appropriate format. The log datasets from technology products are large, multivariate, and complex, so pre-processing is often necessary in order to use them in an implementation. In addition, the extent to which the data obtained from the learning systems are reliable and how some usage measures such as time spent, error rates should be modelled to make the data more accurate is a separate issue and is beyond the scope of this thesis. Additionally, some data attributes in the data sets may be missing and need to be handled carefully.

In the software presented in this thesis, users can make arrangements such as not separating pairs of friends, and splitting people who cannot work together into separate groups, using the interactive features of the tool. If such constraints are presented as external inputs to the algorithms during group creation, it may be more practical for users, especially in cases where the number of individuals is large, saving users time.

The tool has a modular structure and is open-ended that means new computational methods for clustering and grouping can be added. Also, the provided visual methods, i.e. the PeaGlyph and Grid view, are generic and may be useful for a range of analytical tasks. Specific guidance was provided in the *attribute module* and *grouping module,* similarly user specific guidance can be provided in cluster analysis or user can guide cluster algorithm to improve cluster quality. For the latter, the knowledge of the domain in unsupervised learning is necessary and may not be suitable for those out of the field. It is also worth saying that the purpose of this research was to obtain the desired clusters to solve the task at hand, rather than to increase the qualities of the clusters. More specifically, the expectation is to determine the parameters related to the task by the user, and then to obtain clusters of almost equal size based on these parameters (i.e. cluster number, data attributes). Two types of glyph based methods, the Star glyph and PeaGlyph, were used in the implementation, due their intuitiveness. Their visual complexity will remain constant as the number of data elements increases, however, an increasing number of attributes likely impact the usability of them. The selection of attributes due to the task content may reduce the visual complexity. As mentioned before, exact number of variables depends not only the glyph size but also layout being used. Also, we are aware of these glyphs are designed for moderately sized multivariate data set.

The findings from the visual heuristics surveys and the interviews with the experts showed that the visual methods of the tool are intuitive and tool interface are easy to use. However, having a demonstration of how the tool can be used in the relevant environments can be beneficial for the effective use of the available functionalities. In addition, simplifying the terms used or re-naming them differently will make the tool more accessible for non-domain users, this concern was expressed by the participants in the interviews. Similarly, if the interface can be customized according to the domain of the user, expert users in the domain can continue with the existing method and explanations, while non-domain users can continue the task with a simpler interface. Balanced grouping is a real-world problem and we can encounter it in different fields as discussed above. We chose the field of education as the application area and demonstrated the ability of the tool to assist the teacher in creating balanced groups. However, evaluation of the tool by participants from different fields to find problems with visual heuristics and ease of use can provide clues for making the methods and tool abilities more useful. In addition, the mechanism that supports the group formation can be useful in an environment where students do not know each other.

9.3 FUTURE WORK

The evolution of information facilities and increase of data being logged present an opportunity to solve real-world problems. Data analysis and information visualization methods have reached a certain maturity, there is an increasing demand to integrate them into applications. Creating balanced groups is a real world issue and apart from the use case here, it has been explained in the chapters above that there is a problem of creating balanced groups in different domains. For example, the golf scramble problem [212] is among these problems, in which players are separated into four equal-sized groups based on their integer-valued handicaps defined by skill level, and then one player from each group is selected to create a team in a ‘equitable (balanced)' manner. Further, testing the applicability of this tool for different users from different domains to solve similar problems (i.e. the golf scramble problem) [212] is among our plans.

As a future work, how to use the GroupVis in the Computer-supported collaborative learning environment, and its effect on students engagement and performance in group work will be investigated. In the Group module, the random selection method built on balanced size clusters provides a reasonable level of fairness (balancing); however, the design of optimally balanced teams can be achieved through a variety of methods like heuristics approach. Making such improvements is in our future agenda. Moreover, it has been stated in the literature that mixed groups can be more successful in solving some problems than the groups that are formed in homogeneous or heterogenous way. Consequently, a *blended approach* may be added in the next version so that a group can be produced as a mixed group that is homogeneous in some traits and heterogeneous in some traits.

Due to time and access constraints, we were only able to conduct online video calls with participants to hold interviews throughout the pandemic. It is therefore part of our future agenda to conduct a contextual research [15] to observe the current functions of the tool in the real environment of the target users. Further, more candidates from a variety of backgrounds should be recruited for usability testing as the research might be biased because of the low diversity of user groups. Including questions from the *technology acceptance model* [213] in the evaluation process can help us see the attitudes of target users towards the prototype of the software.

Creating competitively balanced teams with players assigned according to specific positions (roles) is a desirable factor for gamers in online games, and creating such groups with aid of the visual methods could be an interesting research direction.

BIBLIOGRAPHY

1. Keim, D., Andrienko, G., Fekete, J.D., Görg, C., Kohlhammer, J. and Melançon, G. "Visual analytics: Definition, process, and challenges." In Information visualization, pp. 154-175. Springer, Berlin, Heidelberg, 2008

2. Liu Z-J, Levina V and Frolova Y. Information visualization in the educational process: Current trends. International Journal of Emerging Technologies in Learning (iJET) 2020; 15: 49-62.

3. Brodbeck, Dominique, Riccardo Mazza, and Denis Lalanne. "Interactive visualization-A survey." In Human machine interaction, pp. 27-46. Springer, Berlin, Heidelberg, 2009.

4. Franconeri, Steven L., Lace M. Padilla, Priti Shah, Jeffrey M. Zacks, and Jessica Hullman. "The science of visual data communication: What works." Psychological Science in the public interest 22, no. 3 (2021): 110-16

5. Wessner M and Pfister H-R. Group formation in computer-supported collaborative learning. In: Proceedings of the 2001 international ACM SIGGROUP conference on supporting group work 2001, pp.24-31.

6. Vygotsky LS and Cole M. Mind in society: Development of higher psychological processes. Harvard university press, 1978.

7. Paredes Barragán P, Ortigosa A and Rodríguez Marín P. A method for supporting heterogeneous-group formation through heuristics and visualization. Journal of Universal Computer Science 2010.

8. Maqtary N, Mohsen A and Bechkoum K. Group formation techniques in computer-supported collaborative learning: A systematic literature review. Technology, Knowledge and Learning 2019; 24: 169-190.

9. Mingers J and O'Brien FA. Creating student groups with similar characteristics: a heuristic approach. Omega 1995; 23: 313-321.

10. Scherr M. Multiple and coordinated views in information visualization. Trends in information visualization 2008; 38: 1-33.

11. Koc K, McGough AS and Johansson Fernstad S. PeaGlyph: Glyph design for investigation of balanced data structures. Information Visualization 2022; 21: 74-92.

12. Card, S., Mackinlay, J., and Shneiderman, B. (Editors), Readings in Information Visualization: Using Vision to Think, Morgan Kaufmann Publishers, San Francisco, CA (1999), 295-305.

13. Friendly M. A brief history of data visualization. Handbook of data visualization. Springer, 2008, pp.15-56.

14. Cook KA and Thomas JJ. Illuminating the path: The research and development agenda for visual analytics. 2005. Pacific Northwest National Lab.(PNNL), Richland, WA (United States).

15. Munzner T. Visualization analysis and design. CRC press, 2014.

16. Bertin J. Semiology of Graphics. Madison, Wis. Univ. of Wisconsin Press, 1983.

17. Chen M and Floridi L. An analysis of information visualisation. Synthese 2013; 190: 3421-3438.

18. Maguire E. Systematising glyph design for visualization. DPhil Thesis 2015.

19. Cleveland WS and McGill R. Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American statistical association 1984; 79: 531-554.

20. Mackinlay J. Automating the design of graphical presentations of relational information. Acm Transactions On Graphics (Tog) 1986; 5: 110-141.

21. Stevens S. Psychophysics: Introduction to its perceptual, neural, and social prospects. 2000 ed. New York: John Willey and Sons 1975.

22. Ware C. Information visualization: perception for design. Morgan Kaufmann, 2019.

23. Chung, David HS, Daniel Archambault, Rita Borgo, Darren J. Edwards, Robert S. Laramee, and Min Chen. "How ordered is it? on the perceptual orderability of visual channels." In Computer Graphics Forum, vol. 35, no. 3, pp. 131-140. 2016.

24. Smart S and Szafir DA. Measuring the separability of shape, size, and color in scatterplots. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems 2019, pp.1-14.

25. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of eugenics 1936; 7: 179-188.

26. Inselberg A and Dimsdale B. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In: Proceedings of the First IEEE Conference on Visualization: Visualization 90 1990, pp.361-378. IEEE.

27. Carr, Daniel B., Richard J. Littlefield, W. L. Nicholson, and J. S. Littlefield. "Scatterplot matrix techniques for large N." Journal of the American Statistical Association 82, no. 398 (1987): 424-436.

28. Shneiderman B. Tree visualization with tree-maps: 2-d space-filling approach. ACM Transactions on graphics (TOG) 1992; 11: 92-99.

29. Chernoff H. The use of faces to represent points in k-dimensional space graphically. Journal of the American statistical Association 1973; 68: 361-368.

30. Chung DH. High-dimensional glyph-based visualization and interactive techniques. Swansea University (United Kingdom), 2014.

31. Keim DA. Information visualization and visual data mining. IEEE transactions on Visualization and Computer Graphics 2002; 8: 1-8.

32. Grinstein G, Pickett R and Williams MG. Exvis: An exploratory visualization environment. In: Graphics Interface 1989, pp.254-261.

33. Levkowitz H. Color icons-merging color and texture perception for integrated visualization of multiple parameters. In: Proceeding Visualization'91 1991, pp.164-170. IEEE.

34. Ankerst M, Keim DA and Kriegel H-P. Circle segments: A technique for visually exploring large multidimensional data sets. In: Visualization 1996.

35. Keim DA. Databases and visualization. ACM SIGMOD Record 1996; 25: 543.

36. LeBlanc J, Ward MO and Wittels N. Exploring n-dimensional databases. In: Proceedings of the First IEEE Conference on Visualization: Visualization90 1990, pp.230-237. IEEE.

37. Becker RA, Eick SG and Wilks AR. Visualizing network data. IEEE Transactions on visualization and computer graphics 1995; 1: 16-28.

38. Cook D. Incorporating exploratory methods using dynamic graphics into multivariate statistics classes: curriculum development. Quality Research in Literacy and Science Education. Springer, 2009, pp.337-355.

39. Shneiderman B. The eyes have it: A task by data type taxonomy for information visualizations. The craft of information visualization. Elsevier, 2003, pp.364-371.

40. Craft B and Cairns P. Beyond guidelines: what can we learn from the visual information seeking mantra? In: Ninth International Conference on Information Visualisation (IV'05) 2005, pp.110-118. IEEE.

41. Jolliffe I. Principal Component Analysis. books. google. com. 2002.

42. Van der Maaten L and Hinton G. Visualizing data using t-SNE. Journal of machine learning research 2008; 9.

43. Cox MA and Cox TF. Multidimensional scaling. Handbook of data visualization. Springer, 2008, pp.315-347.

44. Kahraman HT, Sagiroglu S and Colak I. The development of intuitive knowledge classifier and the modeling of domain dependent data. Knowledge-Based Systems 2013; 37: 283-295.

45. Rhodes JS, Cutler A, Wolf G, et al. Supervised visualization for data exploration. arXiv preprint arXiv:200608701 2020.

46. Groenen PJ and Franses PH. Visualizing time-varying correlations across stock markets. Journal of Empirical Finance 2000; 7: 155-172.

47. Wattenberg M, Viégas F and Johnson I. How to use t-SNE effectively. Distill 2016; 1: e2.

48. Kohonen T. The self-organizing map. Proceedings of the IEEE 1990; 78: 1464-1480.

49. Maltarollo VG, Honório KM and da Silva ABF. Applications of artificial neural networks in chemical problems. Artificial neural networks-architectures and applications 2013: 203-223.

50. Qian, Jimin, Nam Phuong Nguyen, Yutaka Oya, Gota Kikugawa, Tomonaga Okabe, Yue Huang, and Fumio S. Ohuchi. "Introducing self-organized maps (SOM) as a visualization tool for materials research and education." Results in Materials 4 (2019): 100020.

51. Wanner, Franz, Wolfgang Jentner, Tobias Schreck, Andreas Stoffel, Lyubka Sharalieva, and Daniel A. Keim. "Integrated visual analysis of patterns in time series and text data-workflow and application to financial data analysis." Information Visualization 15, no. 1 (2016): 75-90.

52. Silva HM, Silva CA and Gorgônio FL. A self–organizing map based strategy for heterogeneous teaming. Applications of Self-Organizing Maps 2012.

53. Ahmad, Nor Bahiah, Umi Farhana Alias, Nadirah Mohamad, and Norazah Yusof. "Principal component analysis and self-organizing map clustering for student browsing behaviour analysis." Procedia Computer Science 163 (2019): 550-559.

54. Sarikaya A and Gleicher M. Scatterplots: Tasks, data, and designs. IEEE transactions on visualization and computer graphics 2017; 24: 402-412.

55. Behrisch, Michael, Michael Blumenschein, Nam Wook Kim, Lin Shao, Mennatallah El ‐ Assady, Johannes Fuchs, Daniel Seebacher et al. "Quality metrics for information visualization." In Computer Graphics Forum, vol. 37, no. 3, pp. 625-662. 2018 ..

56. Tatu, Andrada, Peter Bak, Enrico Bertini, Daniel Keim, and Joern Schneidewind. "Visual quality metrics and human perception: an initial study on 2D projections of large multidimensional data." In Proceedings of the international conference on advanced visual interfaces, pp. 49-56. 2010

57. Friedman JH and Tukey JW. A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on computers 1974; 100: 881-890.

58. Tukey JW and Tukey PA. Computer graphics and exploratory data analysis: An introduction. In: Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics 1985, pp.773-785.

59. Sips, Mike, Boris Neubert, John P. Lewis, and Pat Hanrahan. "Selecting good views of high ‐ dimensional data using class consistency." In Computer Graphics Forum, vol. 28, no. 3, pp. 831-838. Oxford, UK: Blackwell Publishing Ltd, 2009.

60. Sedlmair M and Aupetit M. Data ‐ driven evaluation of visual quality measures. In: Computer Graphics Forum 2015, pp.201-210. Wiley Online Library.

61. West DM. Big data for education: Data mining, data analytics, and web dashboards. Governance studies at Brookings 2012; 4: 1-10.

62. Sin K and Muthu L. APPLICATION OF BIG DATA IN EDUCATION DATA MINING AND LEARNING ANALYTICS--A LITERATURE REVIEW. ICTACT journal on soft computing 2015; 5.

63. Romero C and Ventura S. Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 2010; 40: 601-618.

64. Chen, Qing, Xuanwu Yue, Xavier Plantaz, Yuanzhe Chen, Conglei Shi, Ting-Chuen Pong, and Huamin Qu. "Viseq: Visual analytics of learning sequence in massive open online courses." IEEE transactions on visualization and computer graphics 26, no. 3 (2018): 1622-1636.

65. Chiritoiu MS, Mihaescu C and Burdescu DD. Students activity visualization tool. In: Educational Data Mining 2013 2013, Citeseer.

66. Johnson M, Eagle M and Barnes T. Invis: An interactive visualization tool for exploring interaction networks. In: Educational Data Mining 2013 2013, Citeseer.

67. Kay, Judy, Nicolas Maisonneuve, Kalina Yacef, and Peter Reimann. "The big five and visualisations of team work activity." In International Conference on Intelligent Tutoring Systems, pp. 197-206. Springer, Berlin, Heidelberg, 2006 .

68. Mazza R and Milani C. Exploring usage analysis in learning systems: Gaining insights from visualisations. In: AIED'05 workshop on Usage analysis in learning systems 2005, pp.65-72. Citeseer.

69. Rueda U, Larrañaga M, Elorriaga JA, and Arruarte A. Validating DynMap as a mechanism to visualize the student's evolution through the learning process. In: International Conference on Intelligent Tutoring Systems 2004, pp.864-866. Springer..

70. Lacefield WE and Applegate EB. Data Visualization in Public Education: Longitudinal Student-, Intervention-, School-, and District-Level Performance Modeling. Online Submission 2018.

71. Nguyen, Huyen N., Caleb M. Trujillo, Kevin Wee, and Kathleen A. Bowe. "Interactive Qualitative Data Visualization for Educational Assessment." In The 12th International Conference on Advances in Information Technology, pp. 1-9. 2021.

72. Kharrufa, Ahmed, Sally Rix, Timur Osadchiy, Anne Preston, and Patrick Olivier. "Group Spinner: recognizing and visualizing learning in the classroom for reflection, communication, and planning." In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 5556-5567. 2017.

73. Vaitsis C, Nilsson G and Zary N. Visual analytics in healthcare education: exploring novel ways to analyze and represent big data in undergraduate medical education. PeerJ 2014; 2: e683.

74. Raji, Mohammad, John Duggan, Blaise DeCotes, Jian Huang, and Bradley Vander Zanden. "Visual progression analysis of student records data." In 2017 IEEE Visualization in Data Science (VDS), pp. 31-38. IEEE, 2017.

75. Siirtola H, Räihä K-J and Surakka V. Interactive curriculum visualization. In: 2013 17th International Conference on Information Visualisation 2013, pp.108-117. IEEE.

76. Saqr M, Fors U and Tedre M. How the study of online collaborative learning can guide teachers and predict students' performance in a medical course. BMC medical education 2018; 18: 1-14.

77. Dos Santos S and Brodlie K. Gaining understanding of multivariate and multidimensional data through visualization. Computers & Graphics 2004; 28: 311-325.

78. Sivanand A and Frank B. Information Visualisation in Education: A Review of Current Tools and Practices. Proceedings of the Canadian Engineering Education Association (CEEA) 2016.

79. Fırat EE and Laramee RS. Towards a survey of interactive visualization for education. Proc Computer Graphics and Visual Computing 2018: 91-101.

80. Slavin RE. Cooperative learning and achievement: Theory and research. 2013.

81. Felder RM and Brent R. Cooperative learning. Active learning: Models from the analytical sciences 2007; 970: 34-53.

82. Cruz WM and Isotani S. Group formation algorithms in collaborative learning contexts: A systematic mapping of the literature. In: CYTED-RITOS International Workshop on Groupware 2014, pp.199-214. Springer.

83. Gibbs G. The assessment of group work: lessons from the literature. Assessment Standards Knowledge Exchange 2009: 1-17.

84. Centre for Teaching Excellence UoW. Implementing Group Work in the Classroom, (accessed 2022 2022).

85. Wang D-Y, Lin SS and Sun C-T. DIANA: A computer-supported heterogeneous grouping system for teachers to conduct successful small learning groups. Computers in Human Behavior 2007; 23: 1997-2010.

86. Cohen EG and Lotan RA. Working for equity in heterogeneous classrooms: Sociological theory in practice. Teachers College Press, 1997.

87. Sternberg RJ. Styles of thinking and learning. Language Testing 1995; 12: 265-291.

88. Bandura A, Freeman WH and Lightsey R. Self-efficacy: The exercise of control. Springer, 1999.

89. Stavrou G, Adamidis P and Papathanasiou J. Computer Supported Team Formation. In: International Conference on Decision Support System Technology 2018, pp.119-131. Springer.

90. Bekele R. Computer-assisted learner group formation based on personality traits. Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky, 2005.

91. Kerr NL and Bruun SE. Dispensability of member effort and group motivation losses: Free-rider effects. Journal of Personality and social Psychology 1983; 44: 78.

92. Goodman B and Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation". AI magazine 2017; 38: 50-57.

93. Le H, Janssen J and Wubbels T. Collaborative learning practices: teacher and student perceived obstacles to effective student collaboration. Cambridge Journal of Education 2018; 48: 103-122.

94. Chernev A. Extremeness aversion and attribute-balance effects in choice. Journal of consumer research 2004; 31: 249-263.

95. Borgo R, Kehrer J, Chung DH, et al. Glyph-based visualization: Foundations, design guidelines, techniques and applications. In: Eurographics (State of the Art Reports) 2013, pp.39-63.

96. Wilke CO. Fundamentals of data visualization: a primer on making informative and compelling figures. O'Reilly Media, 2019.

97. Ward MO. Multivariate data glyphs: Principles and practice. Handbook of data visualization. Springer, 2008, pp.179-198.

98. Maguire, Eamonn, Philippe Rocca-Serra, Susanna-Assunta Sansone, Jim Davies, and Min Chen. "Taxonomy-based glyph design—with a case study on visualizing workflows of biological experiments." IEEE Transactions on Visualization and Computer Graphics 18, no. 12 (2012): 2603-2612 .

99. Chung, David HS, Philip A. Legg, Matthew L. Parry, Rhodri Bown, Iwan W. Griffiths, Robert S. Laramee, and Min Chen. "Glyph sorting: Interactive visualization for multi-dimensional data." Information Visualization 14, no. 1 (2015): 76-90.

100. Keogh, Eamonn, Li Wei, Xiaopeng Xi, Stefano Lonardi, Jin Shieh, and Scott Sirowy. "Intelligent icons: Integrating lite-weight data mining and visualization into GUI operating systems." In Sixth International Conference on Data Mining (ICDM'06), pp. 912-916. IEEE, 2006.

101. Cao, Nan, David Gotz, Jimeng Sun, and Huamin Qu. "Dicon: Interactive visual analysis of multidimensional clusters." IEEE transactions on visualization and computer graphics 17, no. 12 (2011): 2581-2590.

102. Xiong R and Donath J. PeopleGarden: creating data portraits for users. In: Proceedings of the 12th annual ACM symposium on User interface software and technology 1999, pp.37-44.

103. Pearlman J, Rheingans P and des Jardins M. Visualizing diversity and depth over a set of objects. IEEE Computer Graphics and Applications 2007; 27: 35-45.

104. Jänicke, Heike, Rita Borgo, John SD Mason, and Min Chen. "SoundRiver: semantically ‐ rich sound illustration." In Computer Graphics Forum, vol. 29, no. 2, pp. 357-366. Oxford, UK: Blackwell Publishing Ltd, 2010 .

105. Ropinski T and Preim B. Taxonomy and usage guidelines for glyph-based medical visualization. In: SimVis 2008, pp.121-138.

106. Legg, Philip A., David HS Chung, Matthew L. Parry, Mark W. Jones, Rhys Long, Iwan W. Griffiths, and Min Chen. "MatchPad: interactive glyph-based visualization for real ‐ time sports performance analysis." In Computer graphics forum, vol. 31, no. 3pt4, pp. 1255-1264. Oxford, UK: Blackwell Publishing Ltd, 2012.

107. Drocourt, Yoann, Rita Borgo, Kilian Scharrer, Tavi Murray, S. I. Bevan, and Min Chen. "Temporal visualization of boundary ‐ based geo ‐ information using radial projection." In Computer Graphics Forum, vol. 30, no. 3, pp. 981-990. Oxford, UK: Blackwell Publishing Ltd, 2011.

108. Kleiberg E, Van de Wetering H and Van Wijk JJ. Botanical visualization of huge hierarchies. In: IEEE Symposium on Information Visualization, 2001 INFOVIS 2001 2001, pp.87-94. IEEE.

109. Stefaner M., Rasuch F., Leist J., Paeschke M., Baur D., Kekeritz T. OECD Better Life Index. Paris, France: The Organisation for Economic Co-operation and Development; 2017.

110. Khawatmi, Muhammed, Yoann Steux, Sadam Zourob, and Heba Sailem. "ShapoGraphy: a glyph-oriented visualisation approach for creating pictorial representations of bioimaging data." bioRxiv (2021).

111. Siegel JH, Farrell EJ, Goldwyn RM, and Friedman HP. The surgical implications of physiologic patterns in myocardial infarction shock. Surgery 1972; 72: 126-141.

112. Blascheck, Tanja, Lonni Besançon, Anastasia Bezerianos, Bongshin Lee, and Petra Isenberg. "Glanceable visualization: Studies of data comparison performance on smartwatches." IEEE transactions on visualization and computer graphics 25, no. 1 (2018): 630-640.

113. Fuchs, Johannes, Petra Isenberg, Anastasia Bezerianos, and Daniel Keim. "A systematic review of experimental studies on data glyphs." IEEE transactions on visualization and computer graphics 23, no. 7 (2016): 1863-1879.

114. Fuchs, Johannes, Fabian Fischer, Florian Mansmann, Enrico Bertini, and Petra Isenberg. "Evaluation of alternative glyph designs for time series data in a small multiple setting." In Proceedings of the SIGCHI conference on human factors in computing systems, pp. 3237-3246. 2013.

115. Lee MD, Butavicius MA and Reilly RE. Visualizations of binary data: A comparative evaluation. International Journal of Human-Computer Studies 2003; 59: 569-602..

116. Li Y-N, Li D-J and Zhang K. Metaphoric transfer effect in information visualization using glyphs. In: Proceedings of the 8th International Symposium on Visual Information Communication and Interaction 2015, pp.121-130.

117. Fuchs, Johannes, Petra Isenberg, Anastasia Bezerianos, Fabian Fischer, and Enrico Bertini. "The influence of contour on similarity perception of star glyphs." IEEE transactions on visualization and computer graphics 20, no. 12 (2014): 2251-2260.

118. Klippel A, Hardisty F and Weaver C. Star plots: How shape characteristics influence classification tasks. Cartography and Geographic Information Science 2009; 36: 149-163.

119. Miller, Matthias, Xuan Zhang, Johannes Fuchs, and Michael Blumenschein. "Evaluating ordering strategies of star glyph axes." In 2019 IEEE Visualization Conference (VIS), pp. 91-95. IEEE, 2019.

120. Boarini R and d'Ercole MM. Going beyond GDP: An OECD perspective. Fiscal Studies 2013; 34: 289-314.

121. Borland D and Ii RMT. Rainbow color map (still) considered harmful. IEEE computer graphics and applications 2007; 27: 14-17.

122. Anwyl-Irvine, Alexander L., Jessica Massonnié, Adam Flitton, Natasha Kirkham, and Jo K. Evershed. "Gorilla in our midst: An online behavioral experiment builder." Behavior research methods 52, no. 1 (2020): 388-407.

123. Armstrong RA. When to use the B onferroni correction. Ophthalmic and Physiological Optics 2014; 34: 502-508.

124. Huang W, Eades P and Hong S-H. Measuring effectiveness of graph visualizations: A cognitive load perspective. Information Visualization 2009; 8: 139-152.

125. Heer J and Stone M. Color naming models for color selection, image editing and palette design. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 2012, pp.1007-1016.

126. Witulski N and Dias JG. The sustainable society index: Its reliability and validity. Ecological Indicators 2020; 114: 106190.

127. Decancq K. Measuring multidimensional inequality in the OECD member countries with a distribution-sensitive Better Life Index. Social Indicators Research 2017; 131: 1057-1086.

128. Rubin PA and Bai L. Forming competitively balanced teams. IIE Transactions 2015; 47: 620-633.

129. Lockyer KG and Gordon J. Project management and project network techniques. Pearson Education, 2005.

130. Meslec N and Curşeu PL. Are balanced groups better? Belbin roles in collaborative learning groups. Learning and Individual Differences 2015; 39: 81-88.

131. Andrejczuk, Ewa, Rita Berger, Juan A. Rodriguez-Aguilar, Carles Sierra, and Víctor Marín-Puchades. "The composition and formation of effective teams: computer science meets organizational psychology." The Knowledge Engineering Review 33 (2018).

132. Isotani, Seiji, Akiko Inaba, Mitsuru Ikeda, and Riichiro Mizoguchi. "An ontology engineering approach to the realization of theory-driven group formation." International Journal of Computer-Supported Collaborative Learning 4, no. 4 (2009): 445-478.

133. Belbin RM. Team roles at work. Routledge, 2012.

134. Felder RM and Silverman LK. Learning and teaching styles in engineering education. Engineering education 1988; 78: 674-681.

135. Maina EM, Oboko RO and Waiganjo PW. Using machine learning techniques to support group formation in an online collaborative learning environment. International Journal of Intelligent Systems & Applications 2017; 9: 26-33.

136. Matazi I, Messoussi R and Bennane A. The design of an intelligent multi-agent system for supporting collaborative learning. In: 2014 9th International conference on intelligent systems: Theories and applications (SITA-14) 2014, pp.1-8. IEEE.

137. Sun G and Shen J. Teamwork as a service: a cloud-based system for enhancing teamwork performance in mobile learning. In: 2013 IEEE 13th International Conference on Advanced Learning Technologies 2013, pp.376-378. IEEE.

138. Graf S and Bekele R. Forming heterogeneous groups for intelligent collaborative learning systems with ant colony optimization. In: International conference on intelligent tutoring systems 2006, pp.217-226. Springer.

139. Wi, Hyeongon, Seungjin Oh, Jungtae Mun, and Mooyoung Jung. "A team formation model based on knowledge and collaboration." Expert Systems with Applications 36, no. 5 (2009): 9121-9134.

140. Christodoulopoulos CE and Papanikolaou KA. A group formation tool in an e-learning context. In: 19th IEEE international conference on tools with artificial intelligence (ICTAI 2007) 2007, pp.117-123. IEEE.

141. Ounnas A, Davis H and Millard D. A framework for semantic group formation. In: 2008 Eighth IEEE international conference on advanced learning technologies 2008, pp.34-38. IEEE.

142. Craig M, Horton D and Pitt F. Forming reasonably optimal groups: (FROG). In: Proceedings of the 16th ACM international conference on Supporting group work 2010, pp.141-150.

143. Strnad D and Guid N. A fuzzy-genetic decision support system for project team formation. Applied soft computing 2010; 10: 1178-1187.

144. Abnar S, Orooji F and Taghiyareh F. An evolutionary algorithm for forming mixed groups of learners in web based collaborative learning environments. In: 2012 IEEE international conference on technology enhanced education (ICTEE) 2012, pp.1-6. IEEE.

145. Moreno J, Ovalle DA and Vicari RM. A genetic algorithm approach for group formation in collaborative learning considering multiple student characteristics. Computers & Education 2012; 58: 560-569.

146. Arias-Báez MP, Pavlich-Mariscal JA and Carrillo-Ramos A. Forming adapted teams oriented to collaboration: Detailed design and case study. Dyna 2013; 80: 87-95.

147. Srba I and Bielikova M. Dynamic group formation as an approach to collaborative learning support. IEEE transactions on learning technologies 2014; 8: 173-186.

148. Zheng Z and Pinkwart N. A discrete particle swarm optimization approach to compose heterogeneous learning groups. In: 2014 IEEE 14th international conference on advanced learning technologies 2014, pp.49-51. IEEE.

149. Akbar S, Gehringer E and Hu Z. Poster: Improving formation of student teams: A clustering approach. In: 2018 IEEE/ACM 40th International Conference on Software Engineering: Companion (ICSE-Companion) 2018, pp.147-148. IEEE.

150. Friendly M and Denis D. The early origins and development of the scatterplot. Journal of the History of the Behavioral Sciences 2005; 41: 103-130.

151. Sedlmair M, Munzner T and Tory M. Empirical guidance on scatterplot and dimension reduction technique choices. IEEE transactions on visualization and computer graphics 2013; 19: 2634-2643.

152. Holten D and Van Wijk JJ. Evaluation of cluster identification performance for different PCP variants. In: Computer Graphics Forum 2010, pp.793-802. Wiley Online Library.

153. Kononenko I and Kukar M. Machine learning and data mining. Horwood Publishing, 2007.

154. Joia, Paulo, Danilo Coimbra, Jose A. Cuminato, Fernando V. Paulovich, and Luis G. Nonato. "Local affine multidimensional projection." IEEE Transactions on Visualization and Computer Graphics 17, no. 12 (2011): 2563-2571 .

155. Kaski S, Venna J and Kohonen T. Coloring that reveals cluster structures in multivariate data. Australian Journal of Intelligent Information Processing Systems 2000; 6: 82-88.

156. Yang Y and Padmanabhan B. Segmenting customer transactions using a pattern-based clustering approach. In: Third IEEE International Conference on Data Mining 2003, pp.411-418. IEEE.

157. Nuñez JR, Anderton CR and Renslow RS. Optimizing colormaps with consideration for color vision deficiency to enable accurate interpretation of scientific data. PloS one 2018; 13: e0199239.

158. Vesanto J and Alhoniemi E. Clustering of the self-organizing map. IEEE Transactions on neural networks 2000; 11: 586-600.

159. Shalaginov A and Franke K. A new method for an optimal som size determination in neuro-fuzzy for the digital forensics applications. In: International Work-Conference on Artificial Neural Networks 2015, pp.549-563. Springer.

160. Alahakoon D, Halgamuge SK and Srinivasan B. Dynamic self-organizing maps with controlled growth for knowledge discovery. IEEE Transactions on neural networks 2000; 11: 601-614.

161. Tang, Wei, Yang Yang, Lanling Zeng, and Yongzhao Zhan. "Optimizing MSE for clustering with balanced size constraints." Symmetry 11, no. 3 (2019): 338.

162. Wanner, Franz, Andreas Stoffel, Dominik Jäckle, Bum Chul Kwon, Andreas Weiler, Daniel A. Keim, Katherine E. Isaacs, Alfredo Giménez, Ilir Jusufi, and Todd Gamblin. "State-of-the-Art Report of Visual Analysis for Event Detection in Text Data Streams." In EuroVis (STARs). 2014.

163. Xu W. Xu R., Wunsch D. Survey of clustering algorithms, IEEE Transactions on Neural Networks 2005; 16: 645-678.

164. Banerjee A and Ghosh J. Scalable clustering algorithms with balancing constraints. Data Mining and Knowledge Discovery 2006; 13: 365-395.

165. Höppner F and Klawonn F. Clustering with size constraints. Computational Intelligence Paradigms. Springer, 2008, pp.167-180.

166. Bradley PS, Bennett KP and Demiriz A. Constrained k-means clustering. Microsoft Research, Redmond 2000; 20: 0.
167. Wagstaff, Kiri, Claire Cardie, Seth Rogers, and Stefan Schrödl. "Constrained k-means clustering with background knowledge." In Icml, vol. 1, pp. 577-584. 2001.

168. Malinen MI and Fränti P. Balanced k-means for clustering. In: Joint iapr international workshops on statistical techniques in pattern recognition (spr) and structural and syntactic pattern recognition (sspr) 2014, pp.32-41. Springer.

169. Ganganath N, Cheng C-T and Chi KT. Data clustering with cluster size constraints using a modified k-means algorithm. In: 2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery 2014, pp.158-161. IEEE.

170. Li J, Horiguchi Y and Sawaragi T. Cluster size-constrained fuzzy C-means with density center searching. International Journal of Fuzzy Logic and Intelligent Systems 2020; 20: 346-357.

171. Cavallo M and Demiralp Ç. Clustrophile 2: Guided visual clustering analysis. IEEE transactions on visualization and computer graphics 2018; 25: 267-276.

172. Bae J, Helldin T, Riveiro M, Nowaczyk S., Bouguelia M., and Falkman G. Interactive clustering: A comprehensive review. ACM Computing Surveys (CSUR) 2020; 53: 1-39.

173. Seo J and Shneiderman B. Interactively exploring hierarchical clustering results [gene identification]. Computer 2002; 35: 80-86.

174. L'Yi, Sehi, Bongkyung Ko, DongHwa Shin, Young-Joon Cho, Jaeyong Lee, Bohyoung Kim, and Jinwook Seo. "XCluSim: a visual analytics tool for interactively comparing multiple clustering results of bioinformatics data." BMC bioinformatics 16, no. 11 (2015): 1-15.

175. Lex, Alexander, Marc Streit, Christian Partl, Karl Kashofer, and Dieter Schmalstieg. "Comparative analysis of multidimensional, quantitative data." IEEE Transactions on Visualization and Computer Graphics 16, no. 6 (2010): 1027-1035.

176. Younesy, Hamid, Cydney B. Nielsen, Torsten Möller, Olivia Alder, Rebecca Cullum, Matthew C. Lorincz, Mohammad M. Karimi, and Steven JM Jones. "An interactive analysis and exploration tool for epigenomic data." In Computer Graphics Forum, vol. 32, no. 3pt1, pp. 91-100. Oxford, UK: Blackwell Publishing Ltd, 2013..

177. Demiralp Ç. Clustrophile: A tool for visual clustering analysis. arXiv preprint arXiv:171002173 2017.

178. Kwon, Bum Chul, Ben Eysenbach, Janu Verma, Kenney Ng, Christopher De Filippi, Walter F. Stewart, and Adam Perer. "Clustervision: Visual supervision of unsupervised clustering." IEEE transactions on visualization and computer graphics 24, no. 1 (2017): 142-151.

179. Fuchs, Johannes, Petra Isenberg, Anastasia Bezerianos, Matthias Miller, and Daniel Keim. "Educlust-a visualization application for teaching clustering algorithms." In Eurographics 2019-40th Annual Conference of the European Association for Computer Graphics, pp. 1-8. 2019.

180. Choo, Jaegul, Hanseung Lee, Zhicheng Liu, John Stasko, and Haesun Park. "An interactive visual testbed system for dimension reduction and clustering of large-scale high-dimensional data." In Visualization and Data Analysis 2013, vol. 8654, p. 865402. SPIE, 2013.

181. Erra U, Frola B and Scarano V. An interactive bio-inspired approach to clustering and visualizing datasets. In: 2011 15th International Conference on Information Visualisation 2011, pp.440-447. IEEE.

182. Lee, Hanseung, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. "iVisClustering: An interactive visual document clustering via topic modeling." In Computer graphics forum, vol. 31, no. 3pt3, pp. 1155-1164. Oxford, UK: Blackwell Publishing Ltd, 2012.

183. Arın İ, Erpam MK and Saygın Y. I-TWEC: Interactive clustering tool for Twitter. Expert Systems with Applications 2018; 96: 1-13.

184. Müller, Emmanuel, Ira Assent, Ralph Krieger, Timm Jansen, and Thomas Seidl. "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.

185. Tatu, Andrada, Fabian Maaß, Ines Färber, Enrico Bertini, Tobias Schreck, Thomas Seidl, and Daniel Keim. "Subspace search and visualization to make sense of alternative clusterings in high-dimensional data." In 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 63-72. IEEE, 2012.

186. Xu, Panpan, Nan Cao, Huamin Qu, and John Stasko. "Interactive visual co-cluster analysis of bipartite graphs." In 2016 IEEE Pacific Visualization Symposium (PacificVis), pp. 32-39. IEEE, 2016.

187. Schreck, Tobias, Jürgen Bernard, Tatiana Von Landesberger, and Jörn Kohlhammer. "Visual cluster analysis of trajectory data with interactive kohonen maps." Information Visualization 8, no. 1 (2009): 14-29 .

188. Van Long T and Linsen L. Visualizing high density clusters in multidimensional data using optimized star coordinates. Computational Statistics 2011; 26: 655-678.

189. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability 1967, pp.281-297. Oakland, CA, USA.

190. Bezdek JC, Ehrlich R and Full W. FCM: The fuzzy c-means clustering algorithm. Computers & geosciences 1984; 10: 191-203.

191. Arthur D and Vassilvitskii S. k-means++: The advantages of careful seeding. 2006. Stanford.

192. Chakraborty D and Das S. Modified fuzzy c-mean for custom-sized clusters. Sādhanā 2019; 44: 1-7.

193. Stetco A, Zeng X-J and Keane J. Fuzzy C-means++: Fuzzy C-means with effective seeding initialization. Expert Systems with Applications 2015; 42: 7541-7548.

194. Qu Z and Hullman J. Keeping multiple views consistent: Constraints, validations, and exceptions in visualization authoring. IEEE transactions on visualization and computer graphics 2017; 24: 468-477.

195. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 1987; 20: 53-65.

196. Zeng H and Cheung Y-m. Feature selection for clustering on high dimensional data. In: Pacific Rim International Conference on Artificial Intelligence 2008, pp.913-922. Springer.

197. Metsalu T and Vilo J. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic acids research 2015; 43: W566-W570.

198. Brewer CA, Harrower M, Sheesley B. ColorBrewer 2.0: color advice for cartography. The Pennsylvania State University,[En línea] Available: https://colorbrewer2 org/# type= sequential&scheme= BuGn&n 2009; 3.

199. Sharko, John, Georges G. Grinstein, Kenneth A. Marx, Jianping Zhou, Chia-Ho Cheng, Shannon Odelberg, and Hans-Georg Simon. "Heat map visualizations allow comparison of multiple clustering results and evaluation of dataset quality: Application to microarray data." In 2007 11th International Conference Information Visualization (IV'07), pp. 521-526. IEEE, 2007.

200. Ham Fv, Schulz H-J and Dimicco JM. Honeycomb: Visual analysis of large scale social networks. In: IFIP Conference on Human-Computer Interaction 2009, pp.429-442. Springer.

201. Ghoniem M, Fekete J-D and Castagliola P. A comparison of the readability of graphs using node-link and matrix-based representations. In: IEEE symposium on information visualization 2004, pp.17-24. Ieee.

202. Christodoulopoulos CE and Papanikolaou K. Investigation of group formation using low complexity algorithms. In: Proc of PING Workshop 2007, pp.57-60. Citeseer.

203. Ennett ST and Bauman KE. The contribution of influence and selection to adolescent peer group homogeneity: the case of adolescent cigarette smoking. Journal of personality and social psychology 1994; 67: 653.

204. Hu, Ruizhen, Bin Chen, Juzhan Xu, Oliver Van Kaick, Oliver Deussen, and Hui Huang. "Shape-Driven Coordinate Ordering for Star Glyph Sets via Reinforcement Learning." IEEE Transactions on Visualization and Computer Graphics 27, no. 6 (2021): 3034-3047.

205. Solow, Daniel, Jie Ning, Jieying Zhu, and Yishen Cai. "Improved heuristics for finding balanced teams." IISE Transactions 52, no. 12 (2020): 1312-1323.

206. Zhu Y. Measuring effective data visualization. In: International symposium on visual computing 2007, pp.652-661. Springer .

207. Carpendale S. Evaluating information visualizations. Information visualization. Springer, 2008, pp.19-45.

208. Santos BS, Ferreira BQ and Dias P. Heuristic evaluation in information visualization using three sets of heuristics: an exploratory study. In: International Conference on Human-Computer Interaction 2015, pp.259-270. Springer.

209. Forsell C and Johansson J. An heuristic set for evaluation in information visualization. In: Proceedings of the International Conference on Advanced Visual Interfaces 2010, pp.199-206.

210. Wall, Emily, Meeshu Agnihotri, Laura Matzen, Kristin Divis, Michael Haass, Alex Endert, and John Stasko. "A heuristic approach to value-driven evaluation of visualizations." IEEE transactions on visualization and computer graphics 25, no. 1 (2018): 491-500.

211. Gullà, Francesca, Silvia Ceccacci, Michele Germani, and Lorenzo Cavalieri. "Design adaptable and adaptive user interfaces: a method to manage the information." In Ambient Assisted Living, pp. 47-58. Springer, Cham, 2015.

212. Dear RG and Drezner Z. Applying combinatorial optimization metaheuristics to the golf scramble problem. International Transactions in Operational Research 2000; 7: 331-347.

213. Lee Y, Kozar KA and Larsen KR. The technology acceptance model: Past, present, and future. Communications of the Association for information systems 2003; 12: 50.

APPENDICES

APPENDIX -A

Topic	Guiding Questions	Some Follow-up questions
Overview usage of tool (in education settings)	Do you find the tool applicable to the classroom settings?	How to use it in your classroom?
Collaboration learning (methods, significant attributes)	What are your thoughts about collaboration learning?	Are you preferring project-based learning in your classroom? Which grouping method is better over others? Which group characteristics (homogeneity/heterogeneity) may help most for students? While forming groups, which attributes may be significant?
Visual methods and interactions supported by the tool	How do you find the visuals of a particular component?	Which visual method or methods did you find useful in its own context or vice versa?
Aspects/points of tools that need to be improved (including technical names)	We want to know how to improve the tool's functionalities and make it usable and accessible to instructors /teachers.	What made it hard for you to form the desired teams via the tool?

APPENDIX -B

Group Formation Framework - Heuristic Survey

The goal of this study is to evaluate a tool called *GroupVis*, a JavaScript based web application designed to support the formation of balanced homogenous and heterogeneous collaboration teams. The tool has three modular structure- Attribute view, Clustering View and Grouping view. The first phase of the study will consist of a demonstration of the tool functionality. While the second phase is a semi-structured interview and discussion where you will have a chance to test out the tool and give feedback.

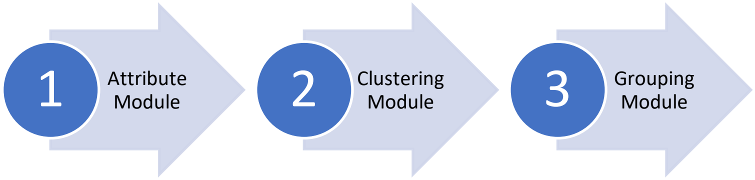
Presented here are a list of visualisation heuristics. You will be using to evaluate the tool, each heuristic is measured through a Likert scale. A description of each component is listed below alongside its heuristics, each component will be explained during the demonstration. While the discussion is recorded, there is a space below each heuristic if you wish to make any written notes.

The Dataset

This a real dataset about the students' knowledge status-

The dataset has 5 attributes, namely :

STG: the degree of study time for goal object materials,
SCG: the degree of repetition number of used for goal object materials
STR: The degree of study time of user for related objects with goal object
LPR: The exam performance of user for related objects with goal object
PEG: The exam performance of user for goal objects
UNS: The knowledge level of user (Very low , low, middle, high)



Time

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
						x	
Note:							

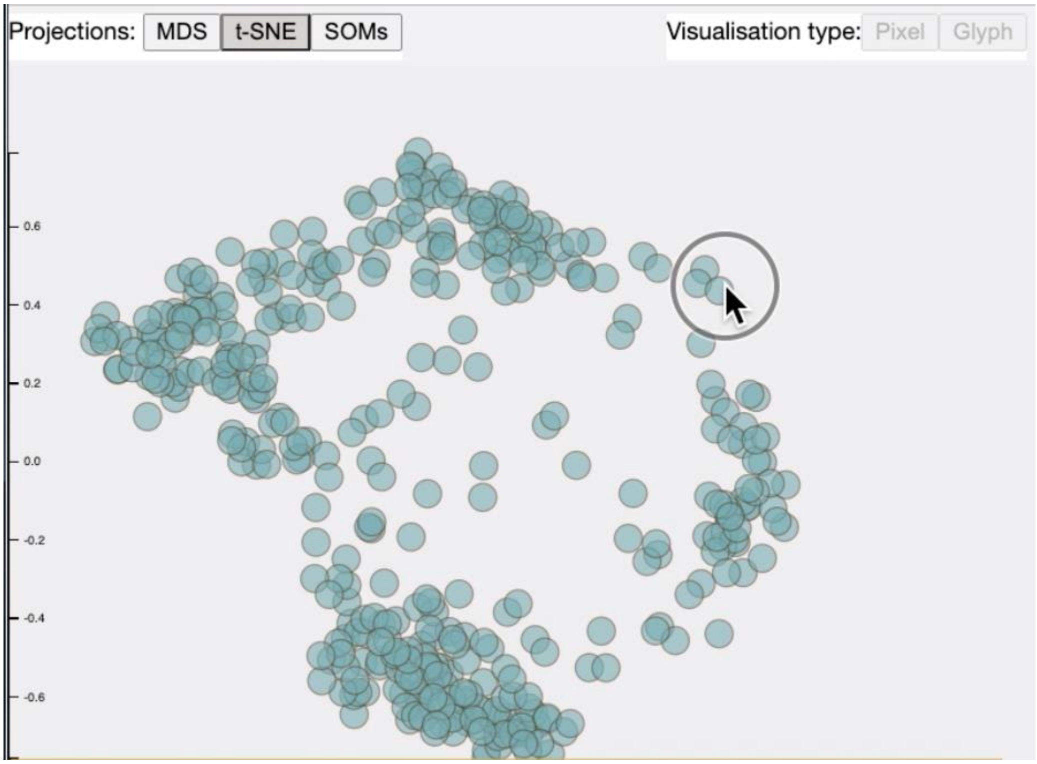


Figure 0-1 the scatterplot in Attribute module

H2. The visualization avoids complex commands and textual queries by providing direct interaction with the data representations							
How would you rate your agreement with the above statement ?							
Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
						x	

Note:

H3. The visualization supports smooth transitions between the views

How would you rate your agreement with the above statement ?

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
				x			

Note:

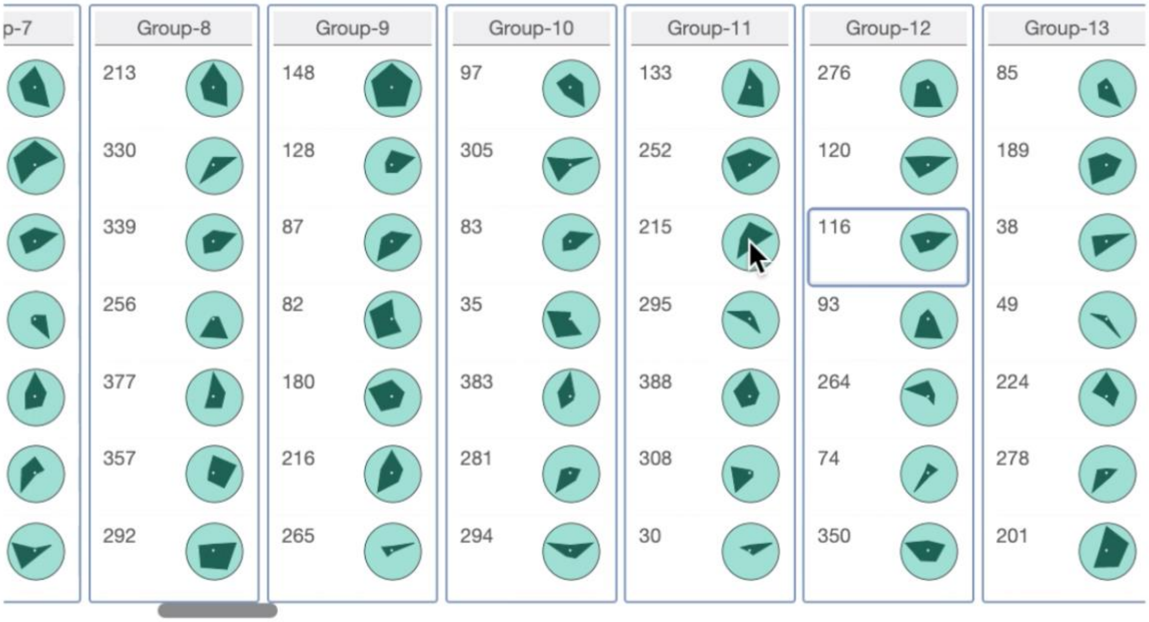


Figure 0-2 List view

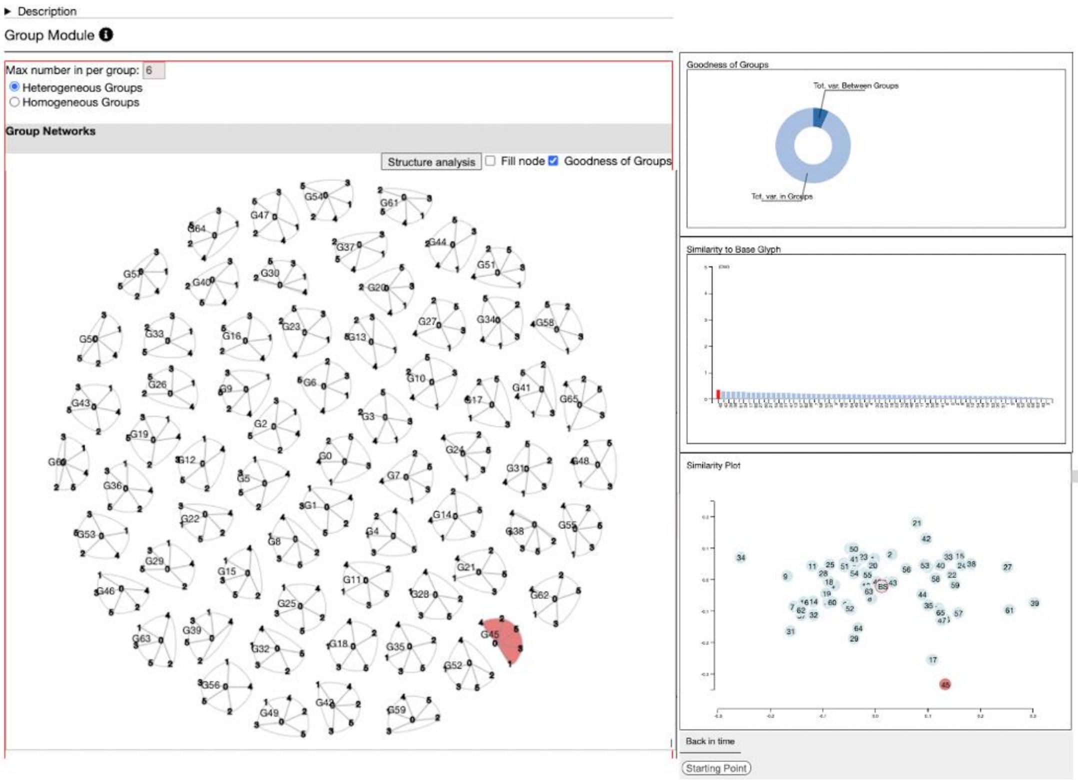
H4. The polygon shapes for each member are intuitive and help users to notice similar/different members across groups at a glance.

How would you rate your agreement with the above statement ?

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
					x		

Note:

Understanding



▸ Description

Group Module

Max number in per group: 6

● Heterogeneous Groups

○ Homogeneous Groups

Group Networks

Structure analysis □ Fill node ☑ Goodness of Groups

Figure 0-3 the components of Grouping module

H5. Visualizations expose individual data cases and their attributes (e.g., you can easily see whatever the items in the dataset represent as well as easily compare their attribute values) How would you rate your agreement with the above statement ?							
Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
						x	
Note:							

H6. The visualization provides useful interactive capabilities to help investigate the data in multiple ways How would you rate your agreement with the above statement ?							
Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor	Somewhat Agree	Agree	Strongly Agree	N/A

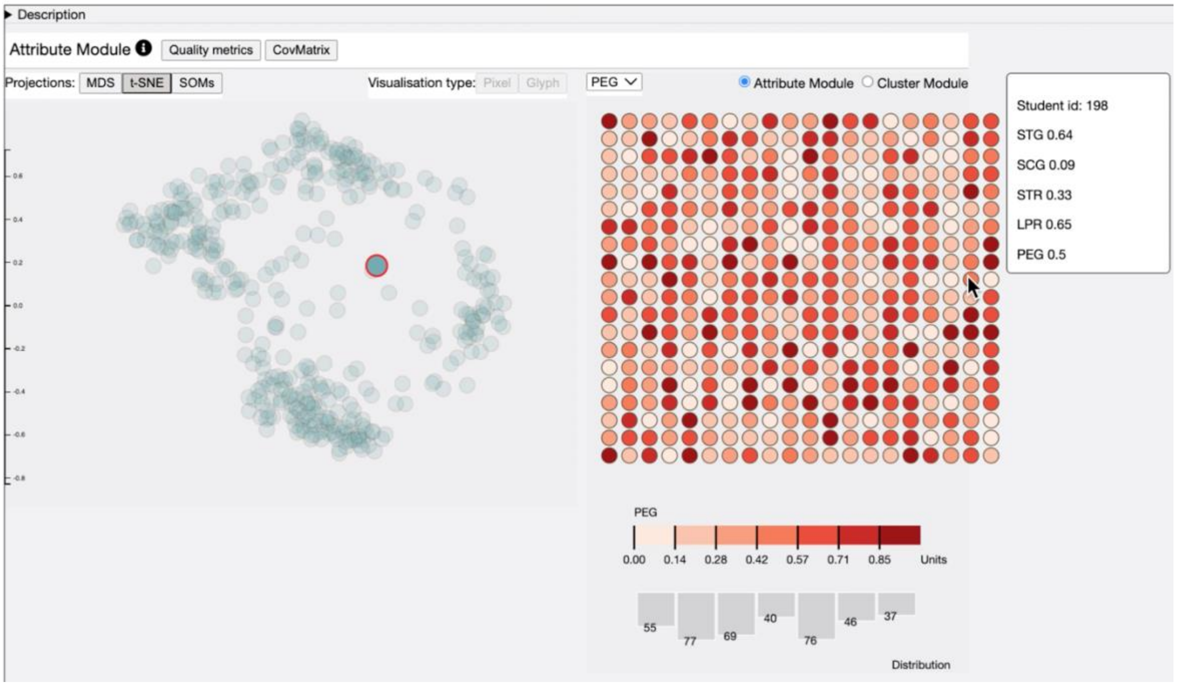
			Disagree				
						x	
Note:							

H7. The visualization helps generate data-driven questions

How would you rate your agreement with the above statement ?

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
						x	

Note:



Student id: 198
STG 0.64
SCG 0.09
STR 0.33
LPR 0.65
PEG 0.5

Figure 0-4 Attribute module

H8- The visualization shows multiple perspectives about the data

How would you rate your agreement with the above statement ?

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
						x	

Note:

H9- The visualization provides useful interactive capabilities to help investigate the data in multiple ways.

How would you rate your agreement with the above statement ?

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
						x	

Intuitiveness

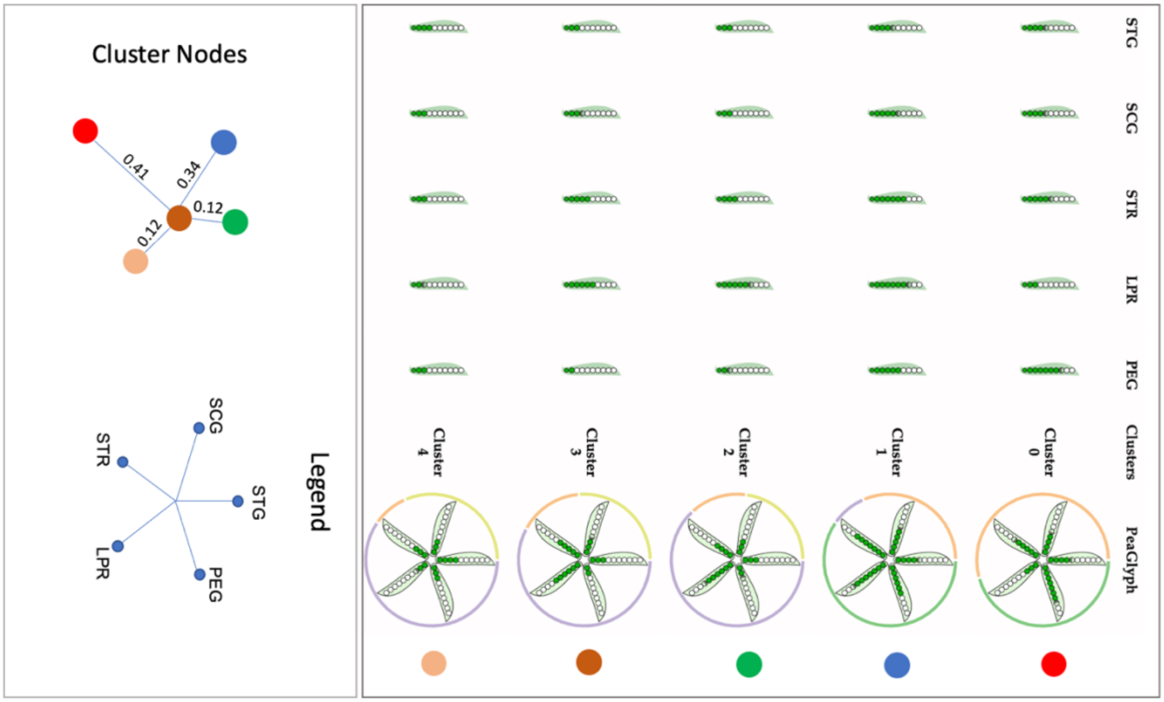


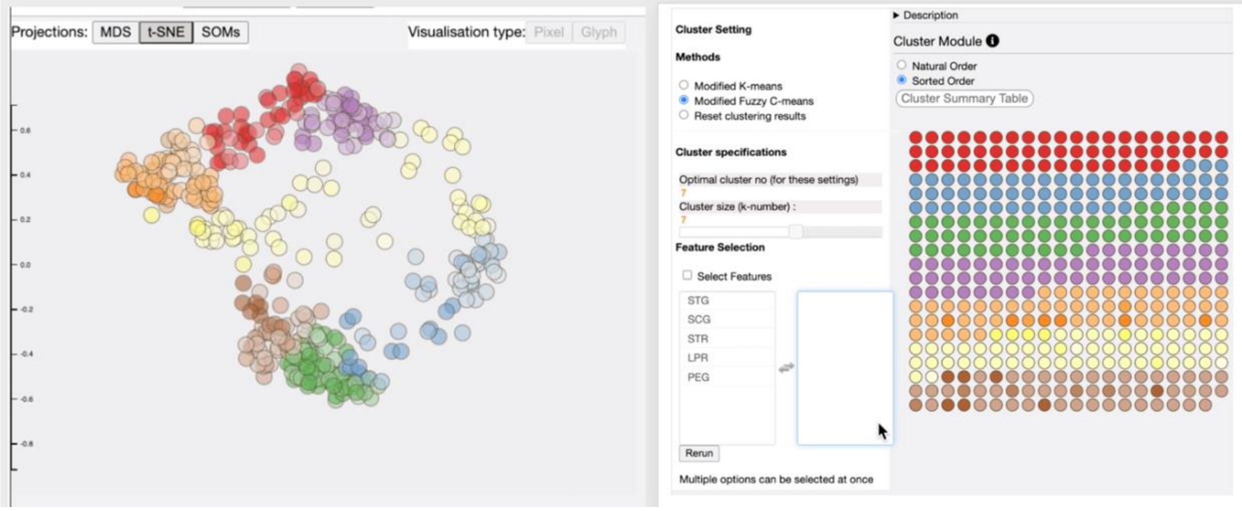
Figure 0-5 Cluster Summary Table integrating PeaGlyph alongside the Cluster Node-Link graph

H10- The representations in the cluster summary table are intuitive of what constitutes a cluster and which attributes differ among clusters

How would you rate your agreement with the above statement ?

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
						x	

Note:



Cluster Setting

Methods

○ Modified K-means
● Modified Fuzzy C-means
○ Reset clustering results

Cluster specifications

Optimal cluster no (for these settings)
7

Cluster size (k-number) :
7

Feature Selection

□ Select Features

STG
SCG
STR
LPR
PEG

Rerun

Multiple options can be selected at once

▸ Description

Cluster Module

○ Natural Order
● Sorted Order

Cluster Summary Table

Figure 0-6 Scatterplot alongside Cluster Module components

H11- The relationship between Grid view and Scatterplot is intuitive

How would you rate your agreement with the above statement ?

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
			x				

Note:

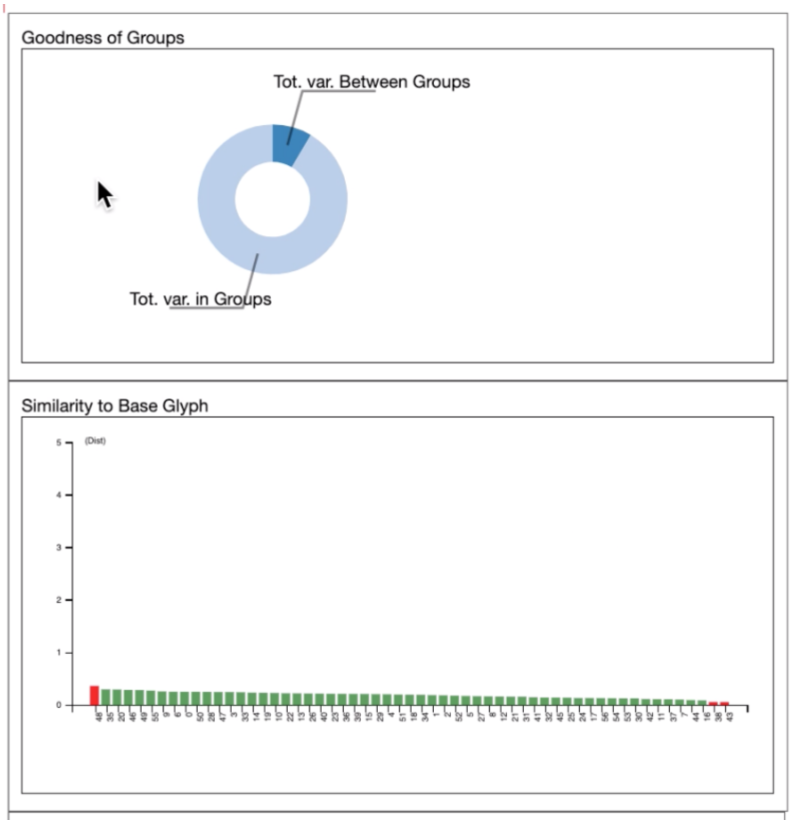


Figure 0-7 Metric View showing the goodness of formed clusters

H12- The representations of the metrics are intuitive.

How would you rate your agreement with the above statement ?

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
						x	

Note:

H13- The representations in the detail view are intuitive, and provide clear information about differences between groups

How would you rate your agreement with the above statement ?

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
						x	

Note:

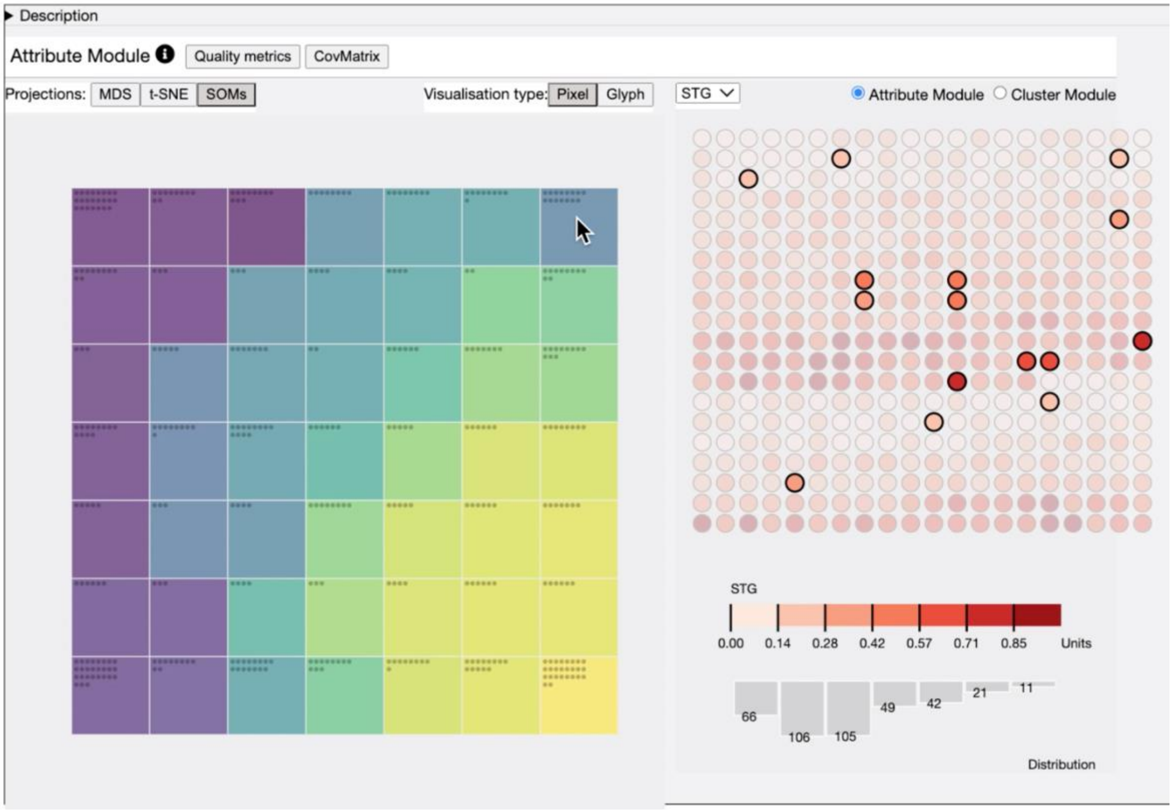


Figure 0-8 Self organizing map alongside Grid view with Interactive color legend

H14- The Self organizing map has an intuitive structure

How would you rate your agreement with the above statement ?

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
					x		

Note:

Essence & Guidance

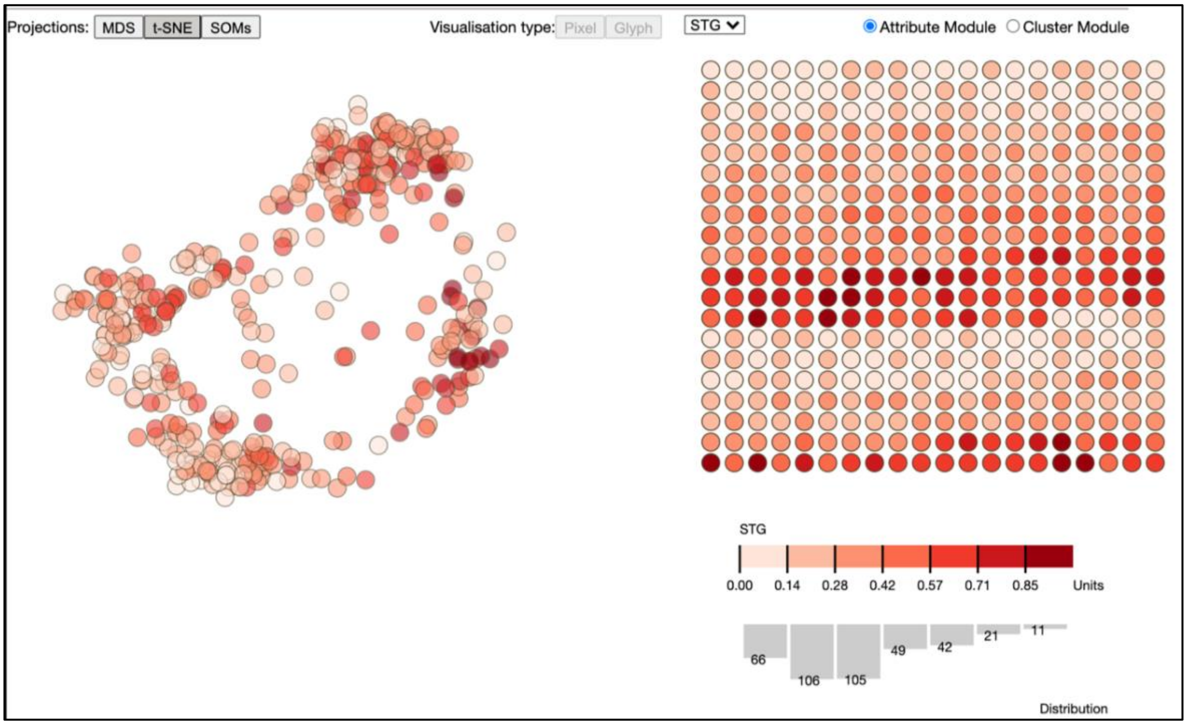

Projections: MDS t-SNE SOMs
Visualisation type: Pixel Glyph
STG
● Attribute Module ○ Cluster Module
STG
0.00
0.14
0.28
0.42
0.57
0.71
0.85
Units
66
106
105
49
42
21
11
Distribution

Figure 0-9 Overview of Attribute Module

H15- The dimension reduction view provides an objective indication of the quality of the plots that helps users choose from among the plots provided

How would you rate your agreement with the above statement ?

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
						x	

Note:

H16- The colouring of the SOM cells helps reveal cluster structures

How would you rate your agreement with the above statement ?

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
						x	

Note:

H17- PeaGlyph helps reveal balanced /unbalanced groups features as well as compare them.

How would you rate your agreement with the above statement ?

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
						x	

Note:

H18- Highlighting the groups in this view that were marked as outliers by the metric was a useful guidance for the starting point of the analysis.

How would you rate your agreement with the above statement ?

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
						x	

Note:

H19- The visualization provides a comprehensive and accessible overview of the data

How would you rate your agreement with the above statement ?

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	N/A
						x	

1) Have you found the tool easy to use in general ?

APPENDIX-C

EXPERIMENT -1

TASK -1

(on page 32)

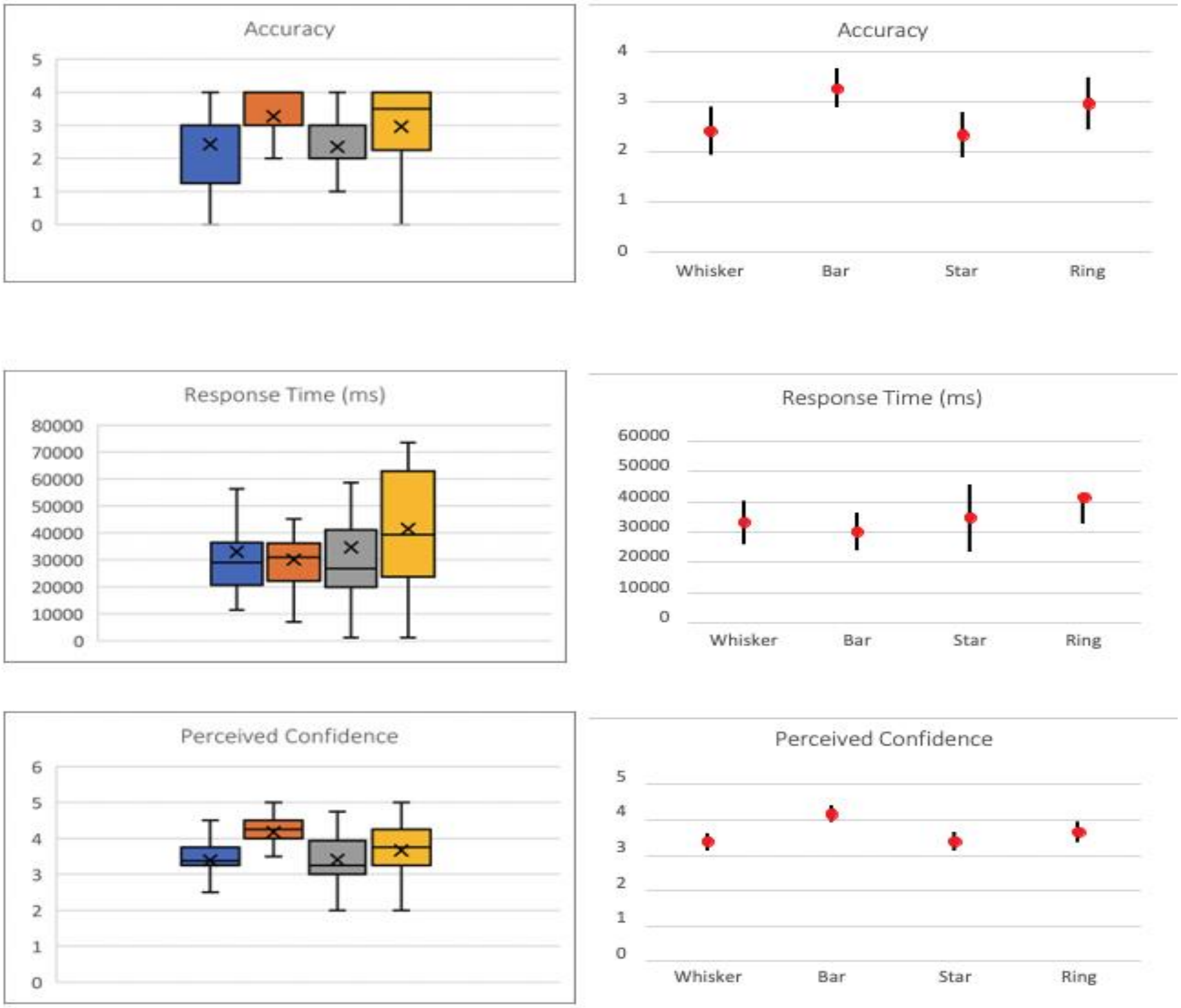


Figure 0-10 The result of Task -1 of the Experiment -1 in the Chapter -3. The Box plots on the left show the performance of the Whisker, Bar, Star, and Ring, respectively. The charts on the right show the confidence intervals of the visual designs in the Task -1.

TASK -2

(on page 35)

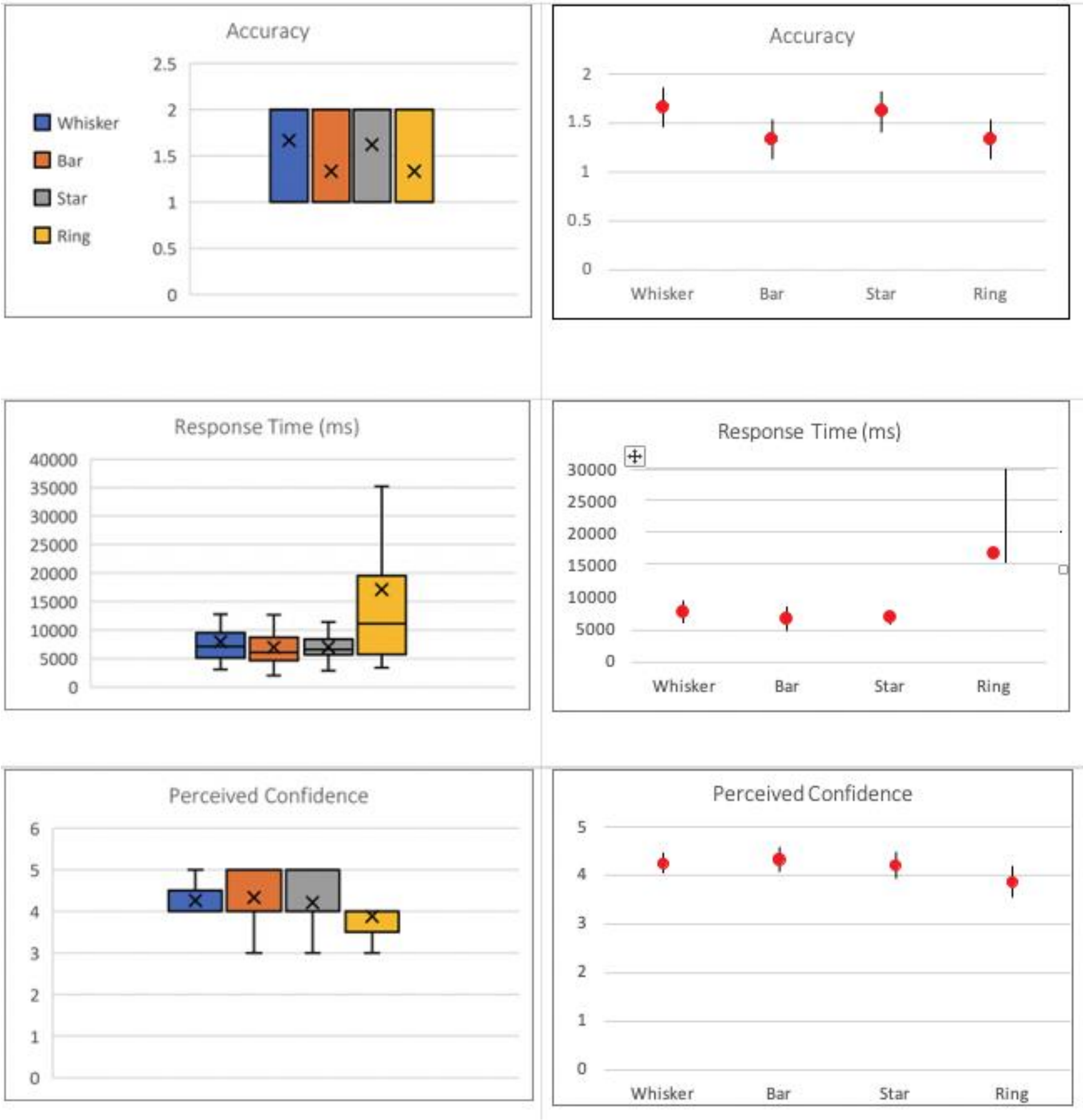


Figure 0-11 The result of Task -2 of the Experiment -1 in the Chapter -3. The Box plots on the left show the performance of the Whisker, Bar, Star, and Ring, respectively. The charts on the right show the confidence intervals of the visual designs in the Task -2.

EXPERIMENT -2

TASK -1

(on page 40)

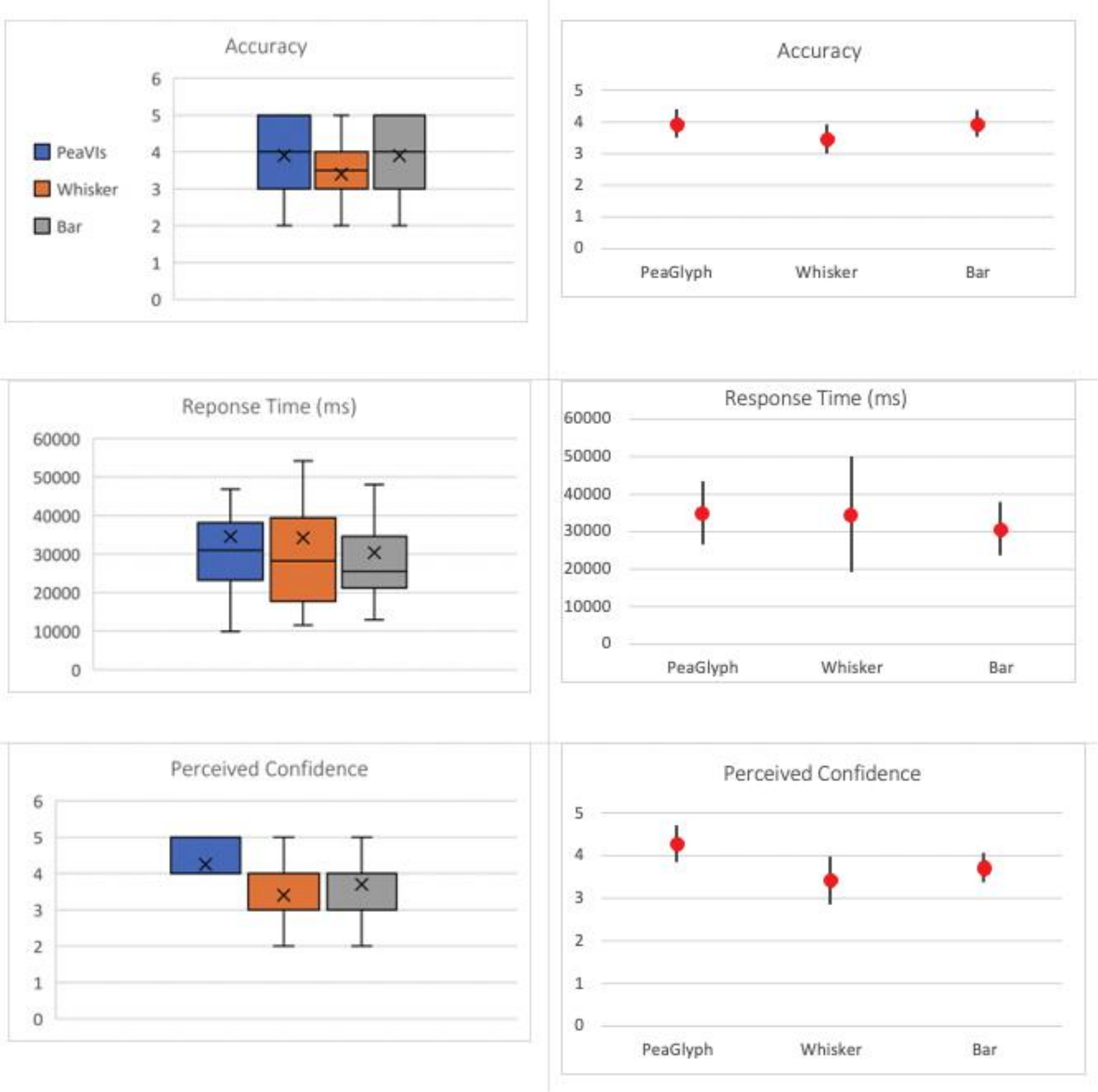


Figure 0-12 The result of Task -1 of the Experiment -2 in the Chapter -3. The Box plots on the left show the performance of the PeaGlyph, Whisker, and Bar, respectively. The charts on the right show the confidence intervals of the visual designs in the Task -1.

TASK -2

(on page 43)

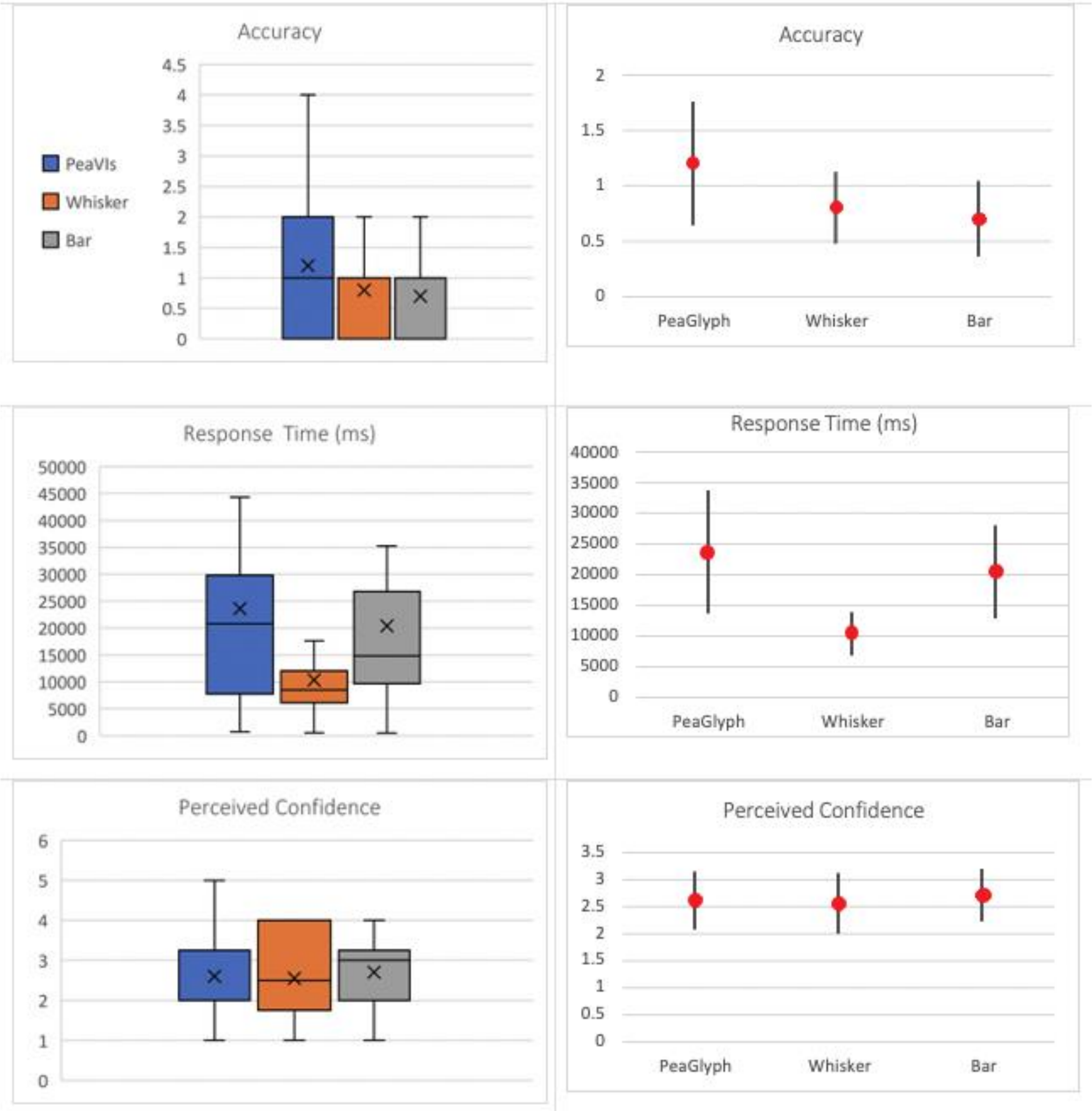


Figure 0-13 The result of Task -2 of the Experiment -2 in the Chapter -3. The Box plots on the left show the performance of the PeaGlyph, Whisker, and Bar, respectively. The charts on the right show the confidence intervals of the visual designs in the Task -2.