Informatics

Visual Comparison of Multivariate Data Ensembles

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Visual Computing

eingereicht von

Anja Heim, BSc
Matrikelnummer 01226809

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof.Dipl.-Ing.Dr. Eduard Gröller
Mitwirkung: DI(FH) Dr. Christoph Heinzl

Wien, 24. November 2020

Anja Heim

Eduard Gröller

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Informatics

Visual Comparison of Multivariate Data Ensembles

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Visual Computing

by

Anja Heim, BSc
Registration Number 01226809

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof.Dipl.-Ing.Dr. Eduard Gröller
Assistance: DI(FH) Dr. Christoph Heinzl

Vienna, 24th November, 2020

Anja Heim

Eduard Gröller

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Anja Heim, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 24. November 2020

Anja Heim

Acknowledgements

The research leading to these results has received funding from the Austrian Research Promotion Agency (FFG) within the program line "TAKE OFF", FFG grant no. 874540 "BeyondInspection", and it has received funding from the Research Foundation Flanders (FWO) and the Austrian Science Fund (FWF) under the grant numbers G0F9117N and I3261-N36 "Quantitative X-ray tomography of advanced polymer composites" respectively.

Above all, I would like to thank my supervisor DI(FH) Dr. Christoph Heinzl for his constructive feedback, support and for the great amount of time and effort that he devoted to this thesis. Further I want to thank Ao.Univ.Prof. Dipl.-Ing. Dr. Eduard Gröller for his expert advises and inspiring viewpoints. My sincere thanks also go to my colleagues of the Research Group Computed Tomography of the FH OOE, who provided me with valuable insights into their work. A special thanks goes to DI Julia Maurer, who provided me with various material datasets and her profound expertise.

Kurzfassung

In sicherheitskritischen Bereichen wie der Aeronautik, aber auch in anderen Sparten wie der Freizeitindustrie, werden die Weiterentwicklungen entsprechender Produkte massgeblich durch die Verbesserung der eingesetzten Werkstoffe und Materialien geleitet. Um die Zieleigenschaften dieser neuen Materialien zu analysieren, werden durch bildgebende Verfahren wie Röntgencomputertomographie (XCT) Abbildungen der inneren Strukturen angefertigt, die anschließend durch Segmentations- und Quantifizierungsalgorithmen detailiert aufgeschlüsselt werden. Der genaue Aufbau der inneren Strukturen ist für MaterialwissenschaftlerInnen hierbei ausschlaggebend für die Charakterisierung von Werkstoffen und ein Vergleich mehrerer Materialkandidaten anhand ihrer Charakteristika somit unumgänglich für die Untersuchung verschiedener Herstellungs-, Optimierungsverfahren oder Eigenschaftsverhalten.

Derzeit sind MaterialwissenschaftlerInnen bei dem Vergleich von mehreren Materialien auf sequentielle Vergleiche angewiesen. Die Verteilungen der einzelnen Attribute verschiedener Materialsysteme müssen verglichen werden, weshalb diese Aufgabe typischerweise geistig aufwendig, zeitintensiv und dadurch fehlerbehaftet ist. Diese Arbeit zielt darauf ab, Fachexperten bei ihren täglichen Aufgaben in der Analyse großer Materialensembledatensätze zu unterstützen. Wir haben für diesen Zweck ein komparatives Visualisierungsrahmenwerk entwickelt, das durch eine Übersichtsvisualisierung und drei Detailvisualisierungstechniken ein ganzheitliches Bild über Ähnlichkeiten bzw. Unähnlichkeiten in den Daten liefert. Anhand der Verwendung der Dimensionsreduktionsmethode Multidimensionale Skalierung werden die individuellen Strukturen zusammengefasst und in einer tabellen-basierten Visualisierungtechnik, namens Histogramm-Tabelle, dargestellt. Die Informationen, in welchen Attributen sich die Strukturen am ähnlichsten sind und ihre genauen Ausprägungen, werden mittels statistischer Berechnungen evaluiert, deren Ergebnisse in einem Säulendiagramm und Kastengrafik visualisiert werden. Schlussendlich können auch die linearen Korrelationen zwischen den individuellen Charakteristiken genauer in einer Korrelationskarte exploriert werden. Wir präsentieren die Nutzbarkeit dieses Visualisierungssystems anhand von drei konkreten Anwendungsszenarien und überprüfen ihre Anwendbarkeit mittels einer qualitativen Studie mit 12 Materialexperten. Die aus unserer Arbeit gewonnenen Erkenntnisse repräsentieren einen signifikanten Schritt im Bereich der komparativen Materialanalyse von hochdimensionalen Daten und unterstützen Fachexperten dabei, ihre Arbeit einfacher und effizienter zu gestalten.

Abstract

In safety-critical areas such as aeronautics, but also in other sectors such as the leisure industry, the advancement of respective products is largely driven by the improvement of the materials used. In order to analyze the targeted properties of these new materials, data of the internal structures is generated, using imaging techniques such as X-ray computed tomography (XCT), which is then analyzed in detail using segmentation and quantification algorithms. For materials scientists, the exact design of the internal structures is crucial for the characterization of materials and a comparison of several material candidates based on their characteristics is therefore indispensable for the investigation of different manufacturing and optimization processes or property behavior.

Currently, material scientists are dependent on sequential comparisons when analyzing several material candidates. Distributions of the individual attributes across the material systems need to be compared, which is why this task is typically cognitively demanding, time consuming, and thus error-prone. This work aims to support domain experts in their daily tasks of analysing large ensembles of material data. For this purpose we developed a comparative visualization framework that provides a holistic picture of similarities and dissimilarities in the data by means of an overview visualization and three detailed visualization techniques. Using the dimension reduction method Multidimensional Scaling, the individual structures are summarized and rendered in a table-based visualization technique called Histogram-Table. Information, describing in which attributes the structures are most similar as well as their exact characteristics, is evaluated by statistical calculations, the results of which are visualized in a bar chart and box plot. Finally, the linear correlations between the individual characteristics can be explored in a correlation map. We present the usability of this visualization system by means of three concrete usage scenarios and verify its applicability by means of a qualitative study with 12 material experts. The knowledge gained from our work represents a significant step in the field of comparative material analysis of high-dimensional data and supports experts in making their work easier and more efficient.

Contents

CHAPTER 1

Introduction

1.1 Motivation

In industry, the search for improvement in order to produce customized materials and material systems is ever-increasing. There are numerous reasons for this continuous ambition: not only cost-efficiency is here the main driving force towards advances within the field, but also the need for highly function-oriented, light-weight and safe components and materials, which are required in application areas such as aerospace, automotive, defense, and others. To ensure continuous optimization as well as to discover novel materials, knowledge about the internal structure is crucial [HS17, NKUC20].

A candidate for advanced future materials, which is promising superior properties over conventional materials and which is applicable in numerous novel applications, is found in composite materials. Composites are heterogeneous material systems, that typically consist of a reinforcement material, which is carrying the loads, and a matrix material, which acts as a glue keeping the reinforcements in place. Due to the heterogeneous nature of this material system a detailed analysis is essential in order to qualify the material system for its targeted use. In order to prevent the failure of composites, especially their inherent structures, such as fibers or pores, need to be monitored and evaluated, e.g., in first part load tests or during the life cycle of the material. The properties of these interior structure such as orientation, length, volume etc., provide valuable insights and allow the expert to draw conclusions about the composites' prospective behaviour. Currently state of the art visualization tools support domain experts with standard visualization techniques targeted on datasets of single materials, employing scatterplot matrices, parallel coordinate plots and other techniques [WAL+14]. With these graphical mappings, interesting features and their attributes are explored individually and per material system, but the users are not provided with any guidance to see the bigger picture of one dataset including all interesting features and how they compare against similar datasets of other materials. Furthermore, the datasets describing specimens often consist of a large quantity of data, usually composed of thousands of features of interest, each represented by 10s to hundreds of different attributes. Comparing and visualizing more than one dataset, with this composition, at a time is not suitable with conventional visualization systems due to overlapping and visual clutter effects. We address this issue in the following chapters of this thesis.

1.2 Aim of Thesis

To support material scientists in understanding, discovering, designing, and optimizing materials, the overall goal of this work is the development of tailored visualization techniques, which enable not only the comparison of individual features such as fibers or pores according to their characteristics within an individual dataset, but more importantly a comparison across several specimens and all interesting feature characteristics at once. We therefore focused on the implementation of a tool that allows material scientists to quickly identify similar (or respectively dissimilar) datasets according to their feature characteristics. The resulting ensemble visualization framework should enable domain scientists to compare information, previously not possible to compose, by simplifying the identification of patterns in numerical multivariate data ensembles through a table-based visual analytic approach. In essence and as generalization, we are striving for a visual comparison of n-dimensional data in ensembles of complex data, in our case as generated in the quantitative analysis and visualization of "rich" XCT data (spatial data + derived quantitative data).

The design of our visualization framework was driven by the following task-specific questions, which a material engineer might formulate according to the provided data basis:

- Which specimens consists of the same type of objects? - For example, do the different materials consist of the same type of fibers, such as very short or very long fibers?
- Are there any similar groups of features within or across the datasets? - Are there individual areas in which the structures are similar, in the case that the materials to be compared are in general very inhomogeneous?
- In which attributes are the objects similar or dissimilar? - For example, are the similar pores all of the same thickness, or are they all of the same length?
- If the elements are similar in one attribute, can the similarity in other attributes be inferred? - If the fibers are very long, do they all have an increased volume?

We developed a visual analysis framework that helps material scientists to get answers to the questions above. In order to avoid losing focus on our goals regarding the development of respective visualization techniques, we defined three general main research questions that our tool should be able to visualize:

R1. Which datasets are similar or dissimilar?

R2. In which characteristics are the datasets similar or different?

R3. Is there a correlation between certain characteristics?

1.3 Methodological Approach

Based on the defined research questions, a visual analysis tool was developed that combines a compact overview with in-depth representations for the complete ensemble of data. To provide the observer with a concise summary of the complete ensemble a table-based visualization tool is offered. Each individual object, such as a fiber or a pore, is represented as a point with a specific position on a horizontal numerical line. The dimension reduction method Multidimensional Scaling is used to reduce each element to a specific position based on its n-dimensional attributes. Different data sets are plotted one beneath the other. To avoid overlapping of the individual objects, they are grouped into bins and their quantity within a bin is encoded using a sequential color scheme. More comprehensive information on the properties of the structures can be inspected more closely using, statistically calculated, detailed visualizations. A bar chart shows the similarity of each attribute by calculating the coefficient of variation. A box plot is provided to explore the exact characteristics of the individual attributes. Finally, a correlation map is presented to determine potential linear correlations between the individual attributes. In addition, the detailed visualizations react to the selections made by the users in the overview visualization and update automatically.

The main contribution of this thesis is the conception and development of a visual analysis tool for material scientists working on the design of safe and customized materials. We support their work using a tool that allows for a comparison of many different materials based on their segmented structures such as fibers, pores and others, as well as their attributes. Our highly abstracted design renders it possible to compare not only different areas of materials, but also samples that have been changed over time.

1.4 Outline of Thesis

The following chapters are structured as follows: In Chapter 2 we describe the backgrounds regarding the field of material science. We briefly explain which topics and goals this area is concerned with, how the required datasets are generated and how they are structured. In Chapter 3 we give an overview of research activities in the field of ensemble visualization. An ensemble is a collection of related datasets. Material science data can consist of collections of related datasets, and thus represents one or several ensembles. Knowledge about general problems and definitions in ensemble visualization is therefore necessary to incorporate the latest findings into the development process. In Chapter 4 various state of the art works are described, which influenced the development of our visualization framework. Especially systems for the visualization of material datasets and latest works from the field of ensemble visualization are described in this section. Chapter 5 describes our developed visual analysis framework in detail. Here, the used aggregation method Multidimensional Scaling and further computations are explained in detail. Furthermore, the developed visualizations and algorithms are introduced. In addition, it is described with which tools the analysis framework was implemented. Chapter 6 deals with the discussion of the developed analysis framework and examines possible use cases. For this purpose, three different usage scenarios are presented. During our work we also conducted a qualitative evaluation in cooperation with 12 material scientists, which is also described in this chapter. Chapter 7 contains a short summary about the developed visualization framework. Furthermore, current limitations are discussed on the basis of the evaluation and possible future extensions and improvements are suggested.

CHAPTER 2

Material Science Background

The structure of datasets being used in material science are rather complex. Each one describes several thousands of features of interest, typically fibers or pores, for which tens to hundreds of attributes are computed. Characteristics such as their length, orientation, diameter, and more describe each feature. This type of data can be defined as multivariate numerical data, since this data contains at each point multiple scalar values that represent simulated or measured quantities. To understand the data, it is often necessary to consider a large amount of them. In the literature such a data compilation is called ensemble data. By definition, ensembles are large data collections containing a number of individual, but often related datasets with slightly varying characteristics. In this context individual ensemble datasets are referred to as members or ensemble members [Sch16].

In recent years, a rapid growth in the generation of ensemble data in various disciplines can be observed [OBJ16, HHB16, Cro18, WHLS19]. This increase may be explained by the necessity to precisely model and understand real-world phenomena and structures. Especially in domains, where the generation of simulation data is crucial for understanding different connected phenomena as for example in climate research, large quantities of ensemble datasets are currently generated. The reason for this rapid increase in respective ensemble data is motivated by the availability of novel powerful computational techniques. These give users access to faster computations of simulation runs of respective simulation models, which are strongly affected by the chosen predefined configurations, such as input parameters, boundary and initial conditions, as well as phenomenological models. With the increased accessibility to advanced parallel computing and momentous improvements in computing power, it is possible to run the models with different configurations to generate multiple realizations of the same experiment within reasonable time. The different characteristics of the models can then be evaluated and enable a new way to gain insight into complex phenomenons.

Although most research in the field of ensemble visualization is done in the area of climate research, there are other domains that would benefit from the analysis of ensemble data. Recent advances in imaging and analysis techniques now also enable the generation of ensemble data. These novel data generation processes, originating from innovations in ultrasound or computed tomography, not only reduce scan times, but also generate images with higher resolutions. The resulting images allow the viewers a more precise recognition and thus interpretability and understanding of the data. Especially, also disciplines in natural science as for example the domain of material science benefits from investigating their ensemble data [HS17].

2.1 Material Science

Material science involves the subareas of understanding, discovery, design and use of (new) materials as well as material systems [HS17]. The scientific field of material science is thus responsible for understanding material's inner structures, and related thereto the material's properties and performances. The knowledge about material's characteristics and abilities allows domain experts the creation of tailored components, fulfilling unique requirements, which have to be preserved to be applicable in the targeted domains. Industries such as agriculture, construction, health care, automotive or aeronautics, depend on adopting these function-oriented, highly integrated components, that specifically meet their needs, since otherwise the improvement of their products and their competitive advantage would be severely limited.

2.2 Non-destructive Testing

The properties of materials are largely determined by its inner structure. For this reason a detailed examination is necessary to understand and ensure the qualitative criteria, which have to be fulfilled to enable a safe usage in the target application. Currently, there is a growing trend to reduce the weight of components, since less weight facilitates less costs and less energy consumption in manufacturing and beyond. As a result, the impact of defects in the material's structure becomes more serious, since even small abnormalities may lead to 100% material failures. Therefore, the recognition of defects and potential points of failure is becoming increasingly important [WDJ15].

Non-destructive testing is thus highly important to ensure the application-specific requirements with regards to the material's inner structures. Besides testing methods based on magnetic, optical or acoustic principles, radiographic examination has gained in importance in recent years: X-ray computed tomography (XCT) is not restricted to medical applications anymore, but becomes common practice in material testing [Sch18]. The goal of this non-destructive imaging technique is to produce a spatial representation (i.e., 3D volumetric datasets), which is generated from series of 2D X-ray penetration images, each of which taken at a different perspective typically around a circular scanning trajectory. To generate images in this examination method, the test object is irradiated by a sharply focused X-ray beam. The X-ray beam, which is attenuated by the scanned specimen, is then recorded by one or more detectors. While in medical CT-applications the source and the detectors rotate synchronously around the object, in industrial applications typically the test specimen is rotated, keeping X-ray source and detector at fixed positions. Each detector therefore determines the total absorption along the X-ray beam. This procedure requires large amounts of data to be stored and processed by the computer, especially when high resolution images are involved. Afterwards the individual sectional images are reconstructed to a single final image, i.e., the 3D reconstructed data volume. This way it is possible to non-destructively inspect the specimen to detect cavities, cracks or irregularities in its inner structure [WDJ15, Sch18].

When analyzing volumetric data, data exploration and visualization can become quickly cumbersome, especially with increasing volume sizes, which is due to the high information content inside the reconstructed volume data. Therefore, typically additional data processing pipelines including segmentation and quantification algorithms, e.g., as proposed by prior papers [SKK$^+$11, TMG$^+$10], are applied to the intensity-based volume, such that further secondary and derived information can be computed and made available in the form of attributes of specific features of interest. The resulting multivariate dataset for every segmented object describes interesting characteristics as length, orientation, Cartesian coordinates of start and endpoint, diameter, etc.

2.3 Composite Materials

Advanced composite materials such as fiber reinforced polymers (FRP) are one type of material, which gained popularity in recent years, because of its potential abilities for optimization in terms of stiffness, strength, density and lower cost with improved sustainability [RPML19]. These properties allow composite materials to be suitable in various application areas, thus making it a viable alternative over many conventional materials. Composite materials consist of a base matrix material and various filler materials. The base matrix material (also referred to as matrix) binds the filler materials, which contain the reinforcement materials and other application-specific filler materials. The reinforcements can be composed of different items, as for example sheets of fiber fabrics, particles or individual fibers. Figure 2.1 shows the classification of composite materials according to their internal structure.

According to Rajak et al. [RPML19], composites can be classified according to their fiber length, with long fiber reinforcements referred to as continuous and short fiber reinforcements as discontinuous fiber reinforcement composites. Furthermore, fibers can be located unidirectional or bidirectional in the matrix structure of continuous fiber composites. Because of this placement and the respective alignment of fibers, this class of composites takes the loads from the matrix to the fibers in an easy and effective way. The fibers placed in discontinuous fiber reinforced components may feature a preferred direction or they are randomly oriented. For effective load transfer and to restrict the growth of cracks, fibers used in this class must have sufficient length. Mechanical

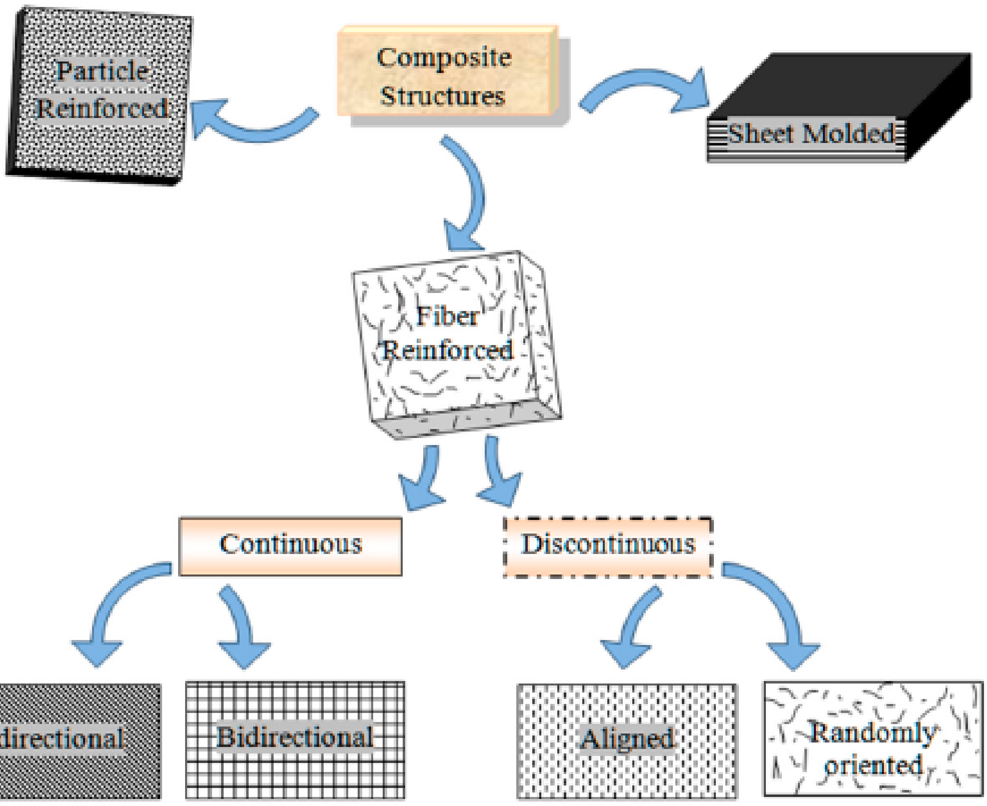


Figure 2.1: Rough classification of fiber reinforced polymers according to their fiber structure, taken from [RPML19].

properties and other structural behavior of composite materials can also be influenced by the arrangement and orientation of the fibers. For this reason, domain experts are typically interested in how the fibers are distributed with respect to length and orientation or which fiber properties are most prominent in spatial subregions when analyzing such fiber reinforced composite samples. This qualitative and quantitative analysis helps to predict final characteristics of the specimens in terms of stiffness, strength, ductility, and durability [WAL$^+$14].

Structures emerging in composites during the manufacturing process are pores. Porosity is a measure for quantifying the void volume in relation to an observation volume in a material. Porosity is calculated from the sum of the individual pores, which are described in literature as small cavities. Pores can be divided into different classes according to their size and appearance, in micro-, meso- or macropores. Isolated pores do not influence the mechanical properties of the material. However, when they aggregate in specific regions, they can significantly reduce the mechanical characteristics of a component. A detailed examination of their distribution and size is inevitable to assure the component's specified properties [Kie07, CSS14].

2.4 Analysis of Composite Materials

The workflow that is generally performed for the analysis of composite materials and the visualization techniques used by the experts, is described on the basis of three differing state-of-the-art papers.

Chung et al. [CEK+19] recently carried out an investigation to describe the material properties and characteristics of lightweight aggregate concrete and foamed concrete with the same density levels. In their analysis six different specimens, three of each type with different density levels, were compared based on an image analysis. The comparison was performed by positioning the 8-bit three-dimensional μ-CT images beside each other and the similarity was determined on the basis of their voxel intensities between 0 (black) and 255 (white). To get more detailed insights about the concrete's inner structure, the distribution of the pore sizes was visualized in a histogram. Thereby, one histogram was drawn for each type of concrete. For each type, the pore size distributions of the specimens with different density levels were superimposed. For studying the pore size distribution of air voids in cement-based materials, Chung et al. [CSR+20] used 2D and 3D imaging approaches. Focusing on the examination of the correlation between pore characteristics and the mechanical properties of the specimens, micro-computed tomography images were used to describe the inner structure of five test materials. The pore distributions of the test specimens were compared using a side by side volume rendering approach, after a segmentation of the voids has been computed using a watershed algorithm. The pore-size distributions of the different materials were compared using individual histograms for each specimen. Since the pore-size distributions were measured with different approaches, the varying results were superimposed in histograms. The correlation between parameters of the pore size distribution and further characteristics were visualized using scatterplots. Figure 2.2 and Figure 2.3 show visualizations of different materials used for exploration in both the works described above.

In the work presented by Maurer et al. [MHPK19] the damage propagation of fibre reinforced polymers was investigated. To get detailed insights about the deformations coming from different forces applied to the materials, four specimens with two types of fiber orientations were tested. Five load steps were performed on each material, resulting in five three-dimensional XCT datasets for each specimen. Since a detailed defect analysis was the aim of this work, their workflow was outlined as follows: First, a pore segmentation was performed on the XCT images in order to extract individual pores in the datasets. Then a defect classification was applied with respect to interesting characteristics of the pores. The four datasets were compared by using 3D rendering of the determined defect classes. Each defect class was assigned a unique color. Then a manual comparison was carried out by juxtaposition of the renderings of the same time steps. The number of defects per material and loading force was visualized using a stacked bar chart. Figure 2.4 and Figure 2.5 show some examples of the visualizations

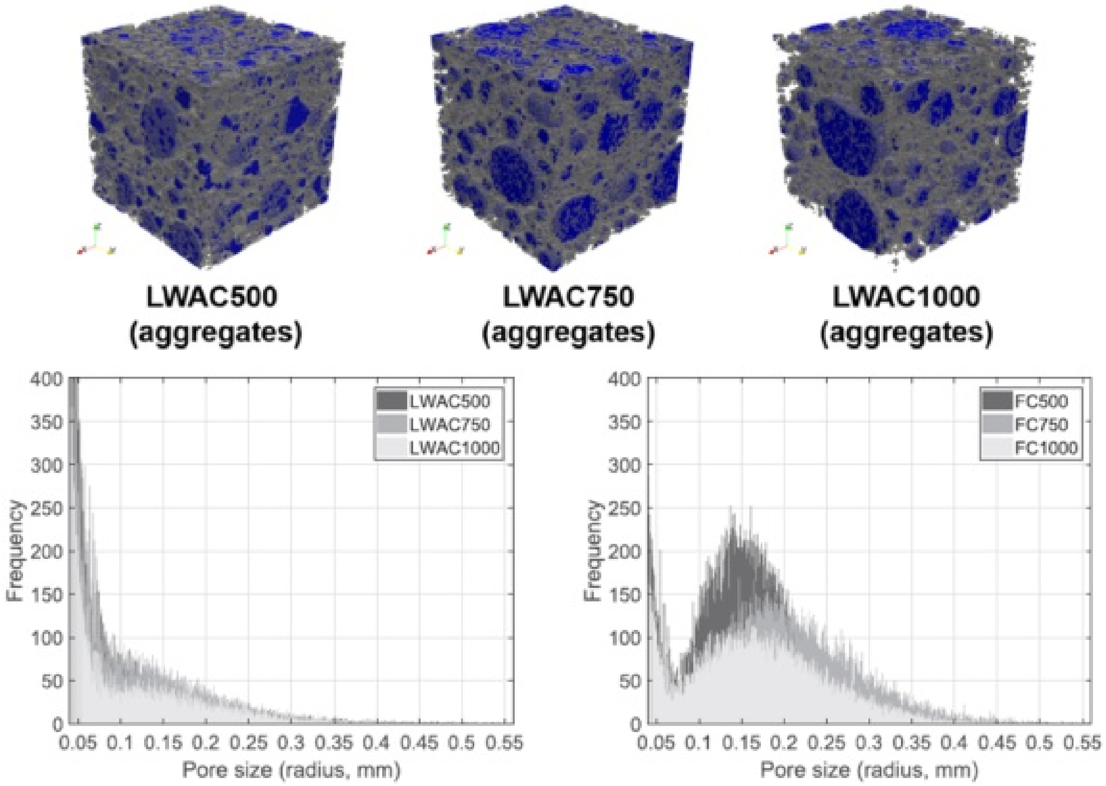


Figure 2.2: Visualizations used for the exploration of different composites, taken from [CEK+19].

used in this study.

In summary, the main goal in the analysis of composites is to investigate different material properties. This is done by examining varying numbers of specimens and sequentially comparing their inner structures. The main target is therefore comparison. This is mostly done using juxtaposed basic visualization techniques like simple 3D renderings or basic charts like scatterplots or line charts. The used visualization techniques, which are based on basic charts, have to manage several challenges common in comparison tasks, as stated by Gleicher et al. [GAW+11]:

1. **Data Complexity:** In this domain the comparison of very intricate but similar objects is necessary. The complexity of the observed structures can be described through:

   a) the number of objects the inner structure consists of, i.e., how many fibers/pores have to be visualized

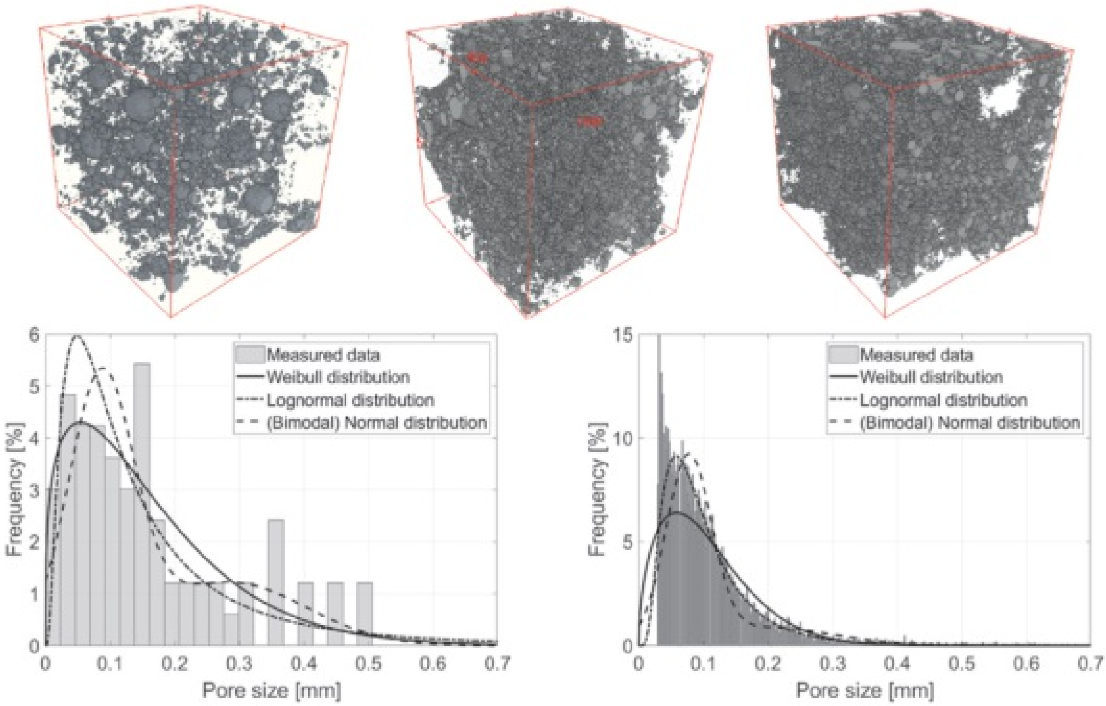


Figure 2.3: Visualizations used for the exploration of different composites, taken from [CSR+20].

b) the abstractness of the information of the data, which is given by the intensity values in the three-dimensional XCT data or in the form of a multitude of numerical attributes

c) the subtleness of the similarity or dissimilarity within the structures, in this case the difference in local intensities or numerical values.

2. **Scalability:** Depending on the complexity of the compared datasets and objects, the visualization technique should be able to scale appropriately. So independently from whether the three-dimensional intensities of the objects or their multivariate attributes are in focus, and from how many elements inside one or several specimens have to be compared, the visualization should adapt accordingly.

Regarding an adaptation to data complexity and scalability, the visualisation techniques currently used in the material-science domain are not too efficient. While visualizing voxel-based datasets in a three-dimensional rendering is quite effective, a side-by-side comparison of various renderings of different (but only slightly changing) datasets is rather limited. In this case, simple juxtaposition of renderings is only effective for a very small number of datasets. As can be seen from the visualization techniques used in the material-science approaches described in this section, at the moment only very few specimens can be simultaneously investigated. In the studies described above, no more than six specimens were examined at a time.

In case of investigating the multivariate characteristics of specimens, the visualization becomes even more challenging. While conventional XCT data is fixed to four different attributes (x, y, z, intensity), multivariate datasets consist of a multitude of characteristics, which is increasing the complexity. The use of conventional visualizations for one such dataset may be sufficient depending on the number of objects being visualized. Frequently used visualizations like scatterplots, where too many objects result in strong visual clutter, often make an effective understanding impossible. Comparing different specimens, each described through its own multivariate dataset, is even more difficult. Simple sequential examination of separate, juxtaposed charts, as currently common practice in the material-science domain, is time consuming, error-prone and therefore costly, since the users have to rely on their memory to make the comparison.

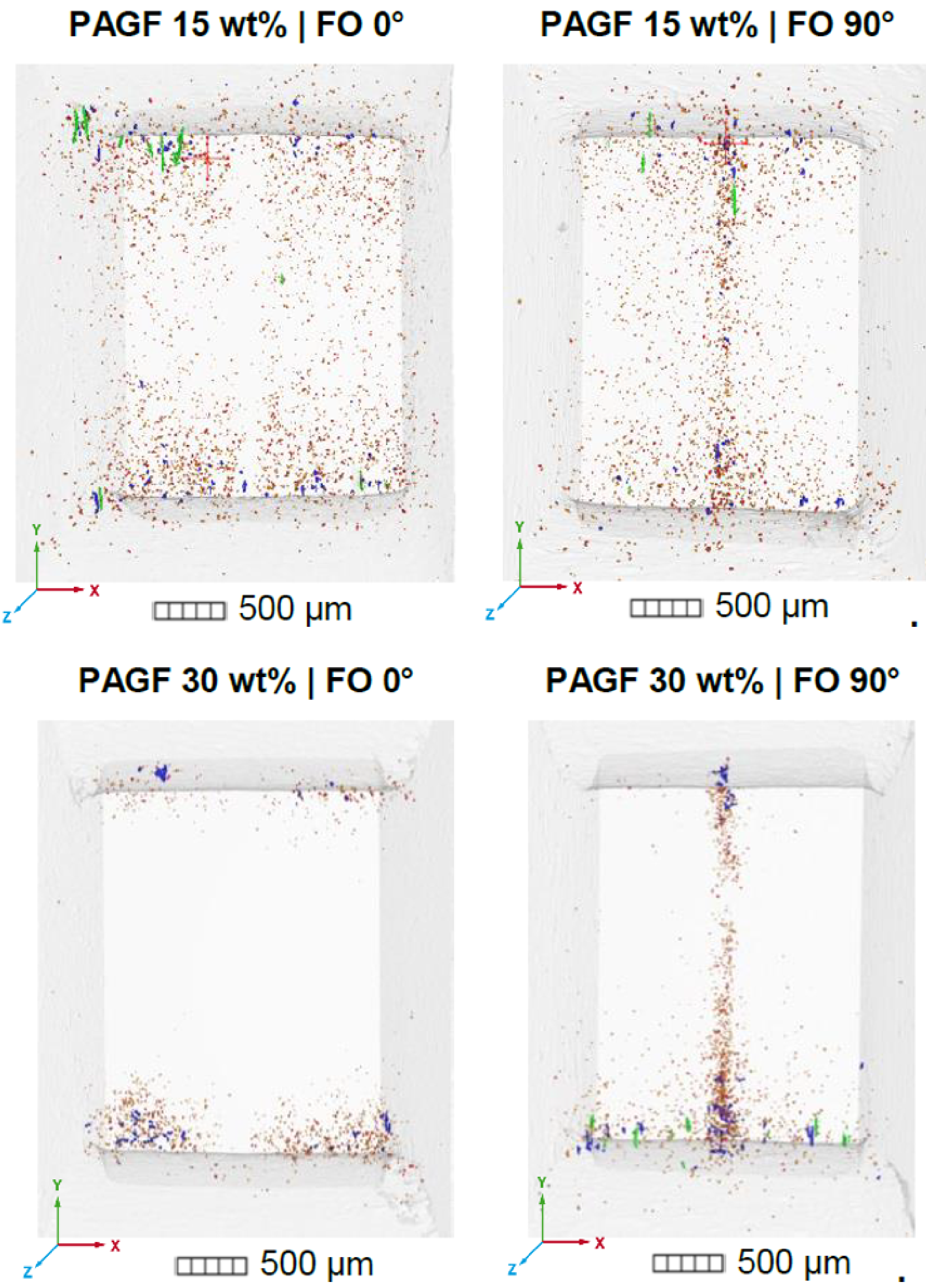
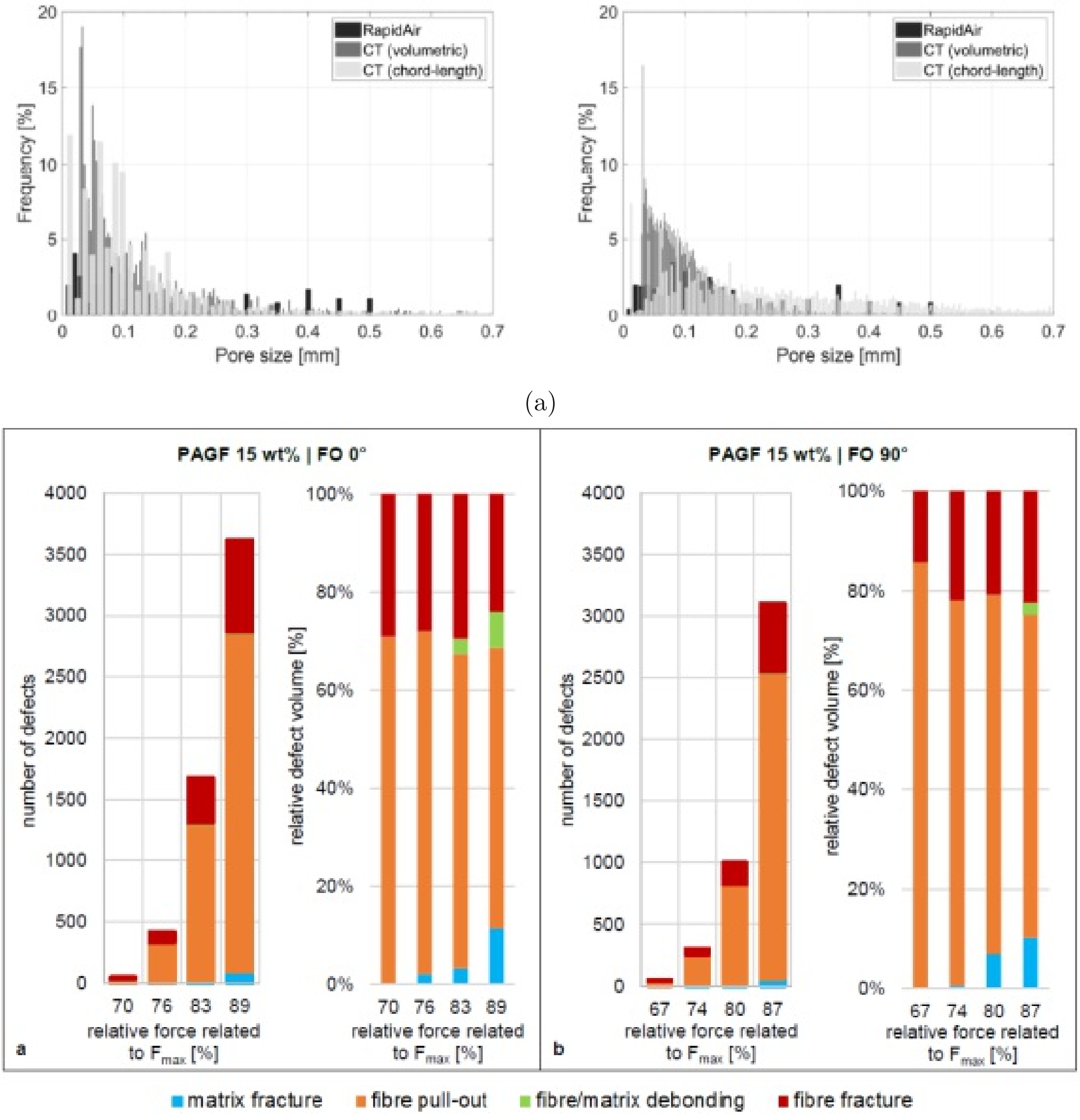


Figure 2.4: 3D rendering of various defects used in [MHPK19].



(a)

(b)

Figure 2.5: Extracts of the visualizations used in [MHPK19]: (a) histograms visualizing the distribution of pore sizes, (b) number of defects displayed via stacked bar charts.

CHAPTER 3

Ensemble Visualization Background

Visualization is necessary to lead scientists towards a more intuitive and more detailed understanding especially of complex data as well as sets of data generated for the analysis of complex phenomena. Visualization of such complex dataset series with spatial, temporal, and abstract data is also referred to as ensemble visualization. Due to the complexity of these data the analysis and comparison of many members of an ensemble is difficult, tedious, and error-prone, which is aggravated by often just subtle differences. Ensemble visualization addresses the three main challenges regarding visual analysis: (1) identifying one or several analysis tasks that reveal high-level relationships between members and between their variables, (2) handling the complex structures and connections of data ensembles, and (3) finding the appropriate visual representation at various levels of abstraction [Cro18].

3.1 Ensemble Analysis Tasks

Although exploratory visual analysis can be performed to gain understanding of the processes or phenomena hidden in the data, usually ensembles are generated to perform a specific analysis task. As stated by Wang et al. [WHLS19], analysis approaches mostly used in the existing literature are targeted to fulfill concrete predefined problems. Ensemble visualization is mainly used for visualizing complex simulation models and their respective outputs, but much less in traditional deterministic scientific domains, as medicine or material science, where datasets are gained by measuring or imaging techniques. Many ensemble visualization techniques are designed in order to fulfill analytic tasks, which are only applicable for this special type of data and limited to a specific application area. For completeness and to show the significant impact on ensemble visualization designs, these tasks are briefly listed below.

- **Overview first, zoom and filter, then details-on-demand**, stated by Shneiderman [Shn96], describes the sequence of individual analysis tasks from whose execution the viewer receives a clear overall picture of the data. Overview is one of the most important visual tasks, especially at the beginning of a visual exploration, when no prior information is available. It is designed to give a concise visual summary to jointly present the overall structure. Usually collective trends, like spatial or temporal patterns, of all ensemble members can be identified at one glance, when using the appropriate visual representations. Further outliers of the entire ensembles can be detected. Because of the complex structure of ensemble data, typically there is a need to perform certain levels of abstraction on each ensemble member to get a concise summary. Since overview is a very basic task, it is used in almost all works in one form or another. To get a compact summary, abstractions must be performed on the complex structure of the ensemble data. Thereby essential details are removed. In order not to lose important information, it is necessary to ensure that it remains accessible. Using zooming and filters, the viewer can select specific structures and display more detailed information for them. This ensures that all information contained in the data can be effectively accessed and examined.

- **Uncertainty Quantification** is an ongoing trend in ensemble visualization works. Here, the uncertainty of the ensemble data is modeled by abstracting the individual ensemble members as a probability density function.

- **(Visual) Parameter Space Analysis** is one of the most popular tasks and targets to build connections between ensemble data and model parameters. So, not only uncertainty in simulation inputs can be analyzed, but domain scientists are able to identify the relation between the input parameters and the resulting output.

- **Comparison** is a difficult and resource-intensive task. It enables viewers to identify similarities and differences, but also common patterns or trends. Regardless of the data domain or type, this task is often challenging, since it requires understanding of relationships among several objects. Comparative visualization techniques are categorized in: superimposition, juxtaposition, and explicit encoding. This classification can also be applied for ensemble visualizations with the difference that the objects to be compared are not limited to merely two data entities, but to collections of entities. Instead of comparing two members or time steps, two or more ensembles of values have to be taken into consideration.

- **Trend Analysis** tries to highlight the evolution of the data over time. The extraction of the temporal trend in ensemble data can be done, as in the traditional temporal analysis, by examining one individual member after another and performing a view composition of the gained information. Because of the additional information of ensemble data, also the whole or a collection of the members can be used to retrieve temporal similarities or differences. Many ensemble visualizations rely on summarizing the temporal evolution by using statistical measurements and visualize these high-level abstractions with two-dimensional time-series plots.

Aside these tasks, ensemble visualization also strongly builds on related dataprocessing, e.g., on Feature Extraction, which aims to extract sophisticated geometric or topological features, like flow transport behaviour or topological critical points, such as a source, sink, or saddle, from a simulation uncertainty field [WHLS19]. Furthermore, clustering is of importance for ensemble visualization, which describes the classification of individual ensemble members into separate groups, where similar members are in the same group. Thereby prominent patterns can be revealed. In the field of ensemble visualization, clustering is often applied to cluster members or similar objects inside the ensemble members. According to Wang et al. [WHLS19], the distance metrics used in various works were selected on the basis of emerging data characteristics, and are therefore very application-specific. Using ensemble visualizations for a variety of different data is mostly not feasible.

3.2 Ensemble Data Structure

In general, ensemble datasets contain similar attributes as conventional scientific data. These can be multivariate, multidimensional, spatio-temporal, etc. The main characteristic that distinguishes ensemble data from traditional scientific data, is their extra dimension, called member dimension [WHLS19]. This dimension is imposed by the individual members of the ensemble. In summary, ensemble data consists of many slightly altered copies, called members, of scientific data. For a detailed description of the structure of ensemble data, we formalize their data representation. Our formalization is based on that of Wang et al. [WHLS19], which was originally introduced for computer simulations. However, it can also be applied to general ensemble data as used in our work: For the generation of XCT images, domain scientists need to specify a set of N parameters, determining image properties of interest, like spatial resolution or geometric blur [Kie07]. These parameters are summarized in the following parameter space as:

$$P = \{p_1, p_2, ..., p_N\},$$

where each parameter $p_i \in P$ has a specific range of interest, defined by physical characteristics, determining image quality. Now we define a parameter vector x_i, which is a sample from this N-dimensional parameter space:

$$x_i = (x_1, x_2, ..., x_N)$$

A output result (i.e., an XCT scan) is generated, based on the specifications of this vector x_i. When so-called in situ tests are performed, a material is scanned several times, after a force, like pulling or shearing, is applied to it. This results in a series of images of a sample. The modification in the force can also be interpreted as different time steps. Therefore an in situ test generates a time series of XCT images, summarized as an ensemble member, here denoted as $S\left(x_i\right)$:

$$S\left(x_i\right) = m_i = \left\{i_1, i_2, ..., i_T\right\},$$

where i_i is the XCT scan for the corresponding time step. Each image i_i is composed of a three-dimensional volume storing intensity values. On each of the images different segmentation and quantification algorithms can be applied, resulting in a descriptive dataset, storing each detected feature, like fibers or pores, denoted as f_i:

$$i_i = \left\{f_1, f_2, ..., f_L\right\}$$

For each feature f_i various characteristics can be computed, like length, diameter, etc. The set of characteristics C for each feature f_i can be denoted as following:

$$C\left(f_i\right) = \left\{c_1, c_2, ..., c_V\right\}$$

For the sake of completeness, it should be noted that various ensembles can be generated, by simply changing the input vector x_i. In our work we do not focus on the effects of the input parameters on the different ensemble datasets. Our aim is the exploration and comparison of several XCT images, which are not necessarily connected. Therefore we do not take the temporal relationship into consideration, if there exists one at all.

As introduced by Wang et al. [WHLS19], we can identify five dimensions of ensemble data: **variable**, **location**, **time**, **member** and **ensemble**. While the dimension variable describes the multivariate characteristics of the features, the dimension location represents the individual features, like fibers or pores, in our scenario. The dimension time is described by the series of images. The number of scans of different specimens, also representing multi-valuedness, results in the member dimension. The variation of the input vector x_i results in multi-resolution ensemble datasets and defines the ensemble dimension. Since we neglect the dimension time and ensemble in our work, we focus on three of the five dimensions. Figure 3.1 shows the structure of ensemble datasets according to the five dimensions.

The comprehensive explanation above shows that the structure and relationship of ensemble data is quite complex. From the given structure of the ensemble datasets two main problems arise, which make a direct application of traditional visualization techniques difficult or even impossible. On the one hand, common representation methods cannot correctly represent the **member dimension**. On the other hand, the **multi-facetedness** of ensemble data is difficult to handle for conventional visualization techniques.

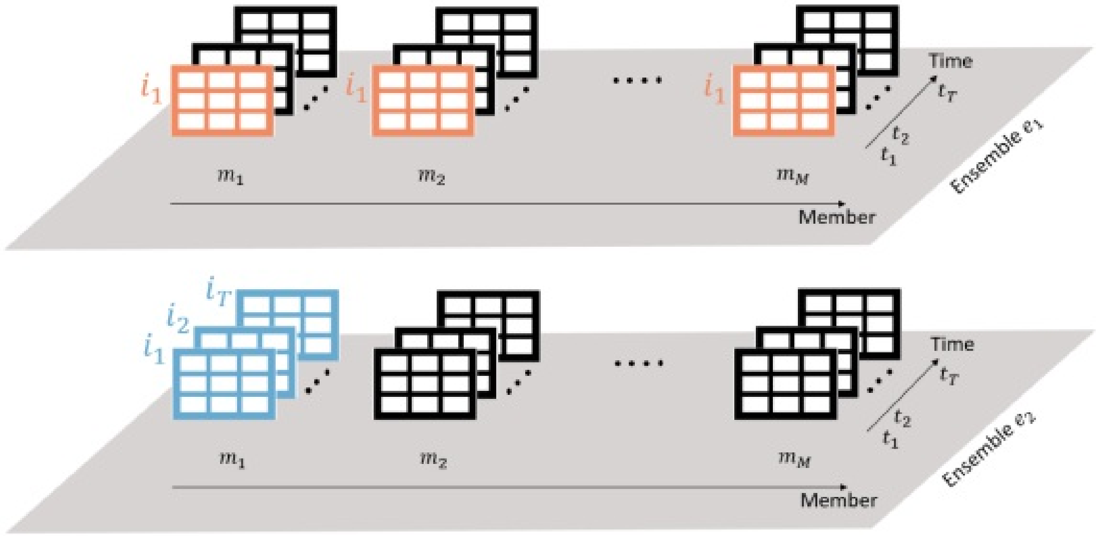


Figure 3.1: This figure shows the structure of ensemble datasets according to the five specified dimensions. Each ensemble dataset e_1 and e_2 has M members, each consisting of one material, recorded for T time steps. On the one hand our analysis tool is able to compare recordings of different materials (red): $m_1 = \{i_1\}, m_2 = \{i_1\}, ..., m_M = \{i_1\}$. On the other hand it is able to compare different time steps of one material (blue): $m_1 = \{i_1, i_2, ..., i_T\}$.

The Member Dimension is the crucial difference between traditional scientific data and ensemble data. This extra dimension can usually not be modeled with traditional visualization techniques. For example direct volume rendering techniques are able to render three-dimensional spatial information with one additional attribute. But rendering more than one descriptive characteristic for one position, is currently not easily possible. One straightforward solution to adapt the algorithm for ensemble data, could be to aggregate the different values at each position to a single value. The calculated mean values can then be visualized, resulting in rendering a mean volume. Of course, such a simple abstraction is not always plausible, or at worst can even be misleading.

Multi-facetedness of ensemble data is determined by the five dimensions defined above. Since the generation of ensemble data is still time consuming, specialists want to explore and understand it in its entirety, when it is finally available. For this reason, the simultaneous presentation of all these facets and their relations is a fundamental goal. Using traditional visualizations a maximum of two different of the five dimensions can be displayed. For example a parallel coordinates plot is able to visualize the variable and location dimension, since the parallel axes are used to display the individual characteristics, while the polylines represent the individual objects. However, extending this visualization to cover a further dimension like the member dimension, for example by adding objects from different members for comparison, is not possible without aggravating the problem of clutter.

3.3 Ensemble Visualization Design

As stated by Crossno et al. [Cro18], information gain of ensemble data is highly influenced by the domain knowledge, for which it is generated. The mental model of the user community has a huge impact on the visual representation. The cognitive image can either be supported or abstracted by the chosen representation and is therefore crucial to consider in the designing phase. Furthermore, the mental model also determines the necessary levels of abstraction of the data, ranging from highly abstract to very precise. We specified three different levels of abstraction: **entire ensemble**, **groups of members**, and **individual members**.

- **Entire Ensemble** – The representation at the level of the entire ensemble shows overall trends and behaviours. Furthermore high level relationships can be investigated like correlations between variables or temporal shifts.
- **Groups of Members** – The abstraction at group level can reveal group relationships such as similarities or differences between clusters.
- **Individual Members** – The representation displaying separate members and characteristics, can uncover commonalities and dissimilarities relative to groups, identifying outliers or anomalies.

Last but not least, the different levels of abstraction must be linked to each other in order to ensure a smooth transition between them and show their relationship to each other. Interactions can be used as an instrument within the interface, performing operations across the entire ensemble or on specific members, e.g. like filters to quickly reduce visual clutter or connect individual elements to specific groups.

To enable the representation of different levels of abstraction, ensemble visualization techniques usually rely on various aggregation techniques applied before or after mapping the data to visual representations. As stated by Wang et al. [WHLS19], the general visualization pipeline can be summarized as follows: Either an initial statistical aggregation step is performed before the visualization, or a visual composition is computed after the visualization. The combination of the two approaches is also common. Figure 3.2 shows a general ensemble visualization pipeline.

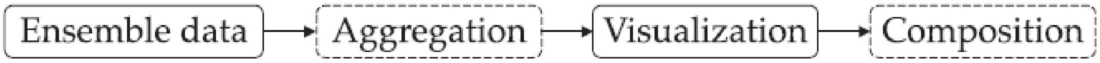


Figure 3.2: General ensemble visualization pipeline taken from [WHLS19].

.

3.3.1 Aggregation

The aggregation is usually generated using statistical methods, which provide summarized descriptions of the individual ensemble members as a result. The summaries are then mapped to appropriate visual encodings, creating a unified representation with the goal of exposing patterns or behaviors of the ensemble data. Commonly used statistical techniques include calculating mean value, variance, standard deviation or modeling probability distributions.

Further, clustering is a very frequently used aggregation method. Cluster-based techniques provide the possibility to group objects into subpopulations. This solves the question to what extent the individual subpopulations differ. However, if there is no great heterogeneity between the objects, i.e. they are very homogeneous, a clear division into discrete groups is difficult to achieve [SSW17]. Since it is not always possible to classify objects, like fibers and pores, unambiguously and correctly into discrete classes, another type of aggregation methods, called dimension reduction techniques, is used in this thesis.

Dimensionality reduction is required, since the dimensionality of the used datasets exceeds the available number of visual channels. By using a dimension reduction technique the data dimension is decreased, making a visualization in 3D, 2D, or even 1D possible. An example is the work of Weissenböck et al. [WAL+14], who use a Hilbert curve as an aggregation method, which transforms the three-dimensional XCT scan into a one-dimensional line. In general, there are many different methods of dimension reduction. A distinction can be made between linear and nonlinear methods [AHT20]. In linear techniques, as the name suggests, linear functions are used to project high-dimensional data into lower dimensions. Commonly used linear methods are for instance Principle Component Analysis (PCA) and Factor Analysis. Multidimensional Scaling (MDS) belongs to the type of non-linear dimension reductions. These techniques were designed to map complex non-linear features into a lower dimensional space. To preserve their structure, methods like MDS or t-distributed stochastic neighbor embedding (t-SNE) are necessary, since they are, in contrast to the linear methods, not bound to a linear approximation [SSW17, SNHS18]. As no assumption can be made about the linearity of the data investigated in this work, the use of linear projections may be too restrictive to obtain an accurate representation. Therefore, only non-linear dimension reduction methods were considered for aggregation in this thesis. One of the currently most popular nonlinear dimensional reduction methods is t-SNE [AHT20, CMK20]. t-SNE is based on matching distances between distributions. This method is very well suited to represent non-linear structures properly, but it also has some disadvantages. The visual clusters found by t-SNE are strongly dependent on the chosen parameterization of t-SNE. Therefore, an exact understanding of the parameters of t-SNE is important. Otherwise, clusters can be formed that do not occur in the original data, which can cause wrong conclusions to be drawn. Furthermore, t-SNE is also known to find clusters in data that are in fact randomly distributed. The main reason why t-SNE is not used in this thesis is that the distances between clusters do not encode any information. Since the similarity of the different objects in our overview visualization is visually encoded by their distances, the results of t-SNE cannot be used for our visualization. MDS was introduced by the authors Kruskal and Wish [KW78] as an unsupervised dimension reduction method, aiming to preserve similarity, dissimilarity, or distances between pairs of data points. Since the MDS globally preserves the pairwise spacing in low-dimensional space, the intrinsic information of the data remains available [AHT20]. MDS is very popular among dimension reduction techniques, because of its simplicity and many fields of application. It is used in various areas of pattern recognition, for seismic data, temperature studies and even for the analysis of viral diseases [SNHS18]. For these reasons we have chosen to use the MDS technique.

To facilitate an effective comparison of features and attributes, a reduction of the complexity of the respective datasets is necessary. Different abstraction techniques can be used to ease complexity, by turning the big data objects into simpler ones that are easier to compare. This strategy can be either applied on the complex objects themselves, or on the relationship between them. We should note here, that abstraction can be seen as a pre-processing step, since the simplified data still has to be compared.

3.3.2 Composition

The composition visually combines the separate visual encodings of the individual ensemble members. This approach also aims to uncover patterns or collective behavior, but instead of abstracting the data directly, their individual representations are merged. Composition is crucial to ensure the effectiveness of the comparison for the user. Three main techniques can be used to describe how different visual encodings can be combined. These techniques originate from comparative visualization and are described by Gleicher et al. [GAW+11] as follows:

- **Juxtaposition** – The elements are placed next to each other to support a direct comparison. This technique is known to provide a good overview and prevents the occurrence of occlusions. An effective comparison is only possible if a limited number of elements have to be compared side by side. The larger the number of objects to be compared becomes, the more difficult it gets to display them all side by side or to compare them for subtle differences.

- **Superimposition** – The elements are placed on top of each other. With this technique a direct comparison is achieved most easily, since the objects to be compared are positioned directly above each other. However, this positioning also causes visibility problems due to the mutual overlapping, which becomes increasingly pronounced the more elements are overlaid.

- **Visual Encoding** – Instead of showing the elements themselves, the differences between them are calculated and displayed. Depending on the realization of this technique, it is often possible to compare many elements at the same time with no significant impact on the readability of the visualization. By using this technique, however, the original information is lost for the viewer, since in the pre-processing step the information is abstracted or its difference is calculated. Furthermore, when using this technique, the validity of the used abstraction or difference determination must be checked if it is meaningful.

CHAPTER 4

Related Work

Since the purpose of this work is to optimize the working process of material engineers and enable a deeper understanding of their data, in the following subsections the visualization techniques developed for domain experts are described. Furthermore, an overview of various works in the field of ensemble visualization is given.

4.1 Visual Analysis in Material Science

In the field of visualization, a body of work already exists, focusing on the improvement of visualizations for material science data [HS17]. In the following works focusing on comparing the inner structure of one material will be discussed.

To simplify the exploration of composites, Fritz et al. [FHG+09] developed an improved volume visualization to better examine the individual objects. They focused on the improved visualization of the orientation of fibers and the roundness of the particles. For these characterizations, special transfer functions were calculated, which allowed the authors an easier exploration and improved volume rendering. Grau et al. [GVTA10] used an illustrative rendering method to achieve an improved volume representation of pores. Different visualization styles, such as the calculation of a topological graph and other illustrative effects, and interactive pore selection, allowed the viewers a better understanding of the connectivity of individual pores and their sizes. Both works give a good overview of the 3D structure, but do not provide any information about other characteristics. By relying on volume rendering, only one attribute at a time can be displayed for each individual object and not several at once, which is the aim in our thesis.

Weissenböck et al. [WAL+14] introduced a visual analysis framework to explore fiber reinforced polymers, focusing on the visualization of fiber properties as generated from segmented and quantified XCT computed dataset. Different visualization techniques are composed in the framework to give a broad overview and enable a detailed investigation of the multivariate data. By using parallel coordinates and a scatterplot matrix, the quantitative characteristics of the features of interest are shown and interaction possibilities are offered to focus the analysis on interesting features. Since fiber length and fiber orientation distributions are of special interest for domain experts, these distributions are visualized in separate 3D representations. Because of the large number of fibers, visual clutter is a recurring problem. To reduce clutter in 3D renderings, the visual abstractions of a blob and a meta data visualization support in revealing regions of fibers with similar characteristics. Summarizing, this advanced visualization toolkit supports crucial domain specific tasks such as identifying and visualizing classes of individual fibers, and provides supplementary investigation opportunities for the most important characteristics fiber length and orientation. Furthermore it enables the identification of regions with similar fiber characteristics. Figure 4.1 shows the FiberScout system.

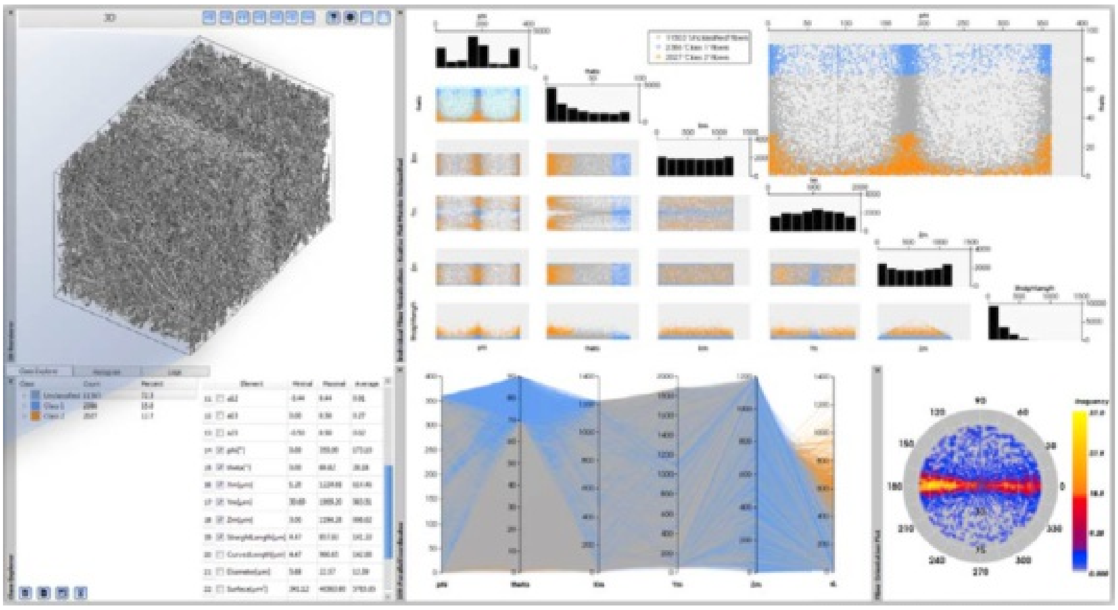


Figure 4.1: System Overview of FiberScout taken from [WAL+14].

.

Bhattacharya et al. [BWW+17] presented a tool to extract and visualize individual fiber bundles and weaving patterns from XCT scans: First, a coarse version of integral curves was used to trace sections of the individual fiber bundles. Then these fiber bundles were extracted and clustered using an hierarchical approach: first the data was clustered by orientation, then by proximity. After the extraction of the fiber bundles, either a volumetric representation or surface models were generated for further analysis. Aside a spatial representation of the fiber bundles, their properties could be visualized using the methods introduced in FiberScout [WAL+14]. Through the scatterplot matrix and the parallel coordinates plot, the domain experts were able to identify specific classes and to compute statistical information like the minimum, mean, or maximum of the characteristics of the class. This work expands the investigation possibilities of fiber reinforced polymers by allowing the users to examine even more complex fiber structures.

The study of Chiverton et al. [CIBP17] focuses on the understanding of the arrangement of steel fibers in concrete. For this purpose the orientation of the fibers and the distance between the fibers is used to develop models in order to understand the randomness of an orientation distribution both locally and across the entire volume. Multiscale entropy is used here to summarize the orientation and the spatial distribution into a single stochastic model. The model is then visualized using volume rendering, resulting in a detailed overview visualization, which helps to accurately identify differences in the properties of the material. Figure 4.2 shows a volume rendering of the entropy estimation giving an overview of the randomness of the fiber orientations.

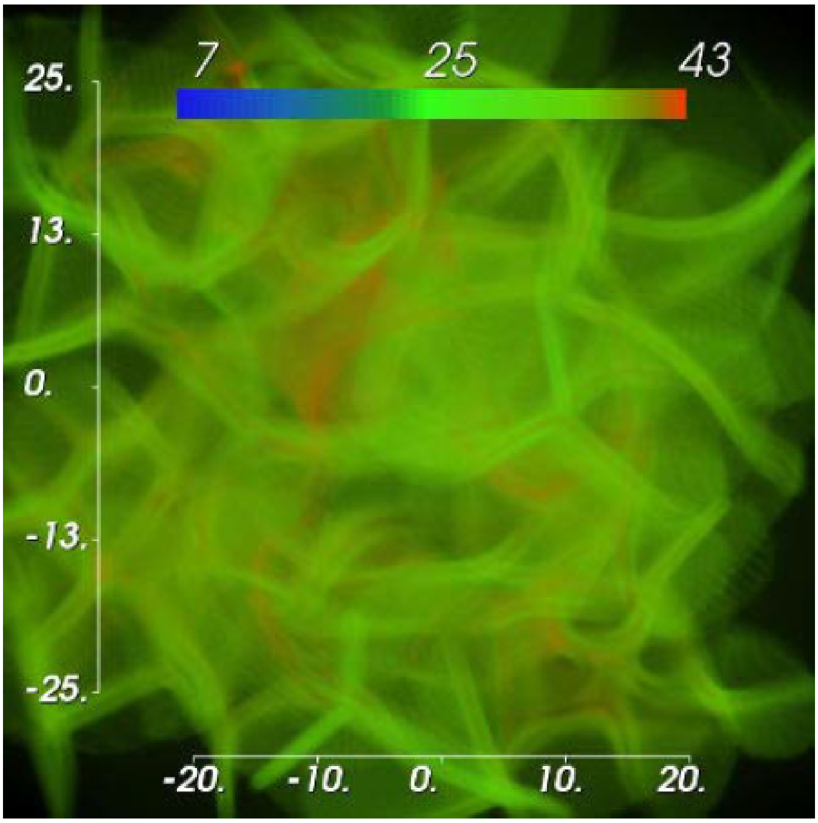


Figure 4.2: Volume rendering visualizing the orientation of fibers with color, taken from [CIBP17].

.

Since orientation is one of the most important characteristics for material engineers to base their considerations upon, Weissenböck et al. [WAS+18] focused on the visualization of this attribute. In this work the comparison of two different datasets, one real-world and one simulated, were discussed based on their orientation-tensor information. To measure the resemblance of this characteristic, three different similarity measures were computed: the degree of orientation, cosine similarity, and tensor similarity. The computed measures were visualized by overlaying a heatmap with a sequential color scheme on the XCT dataset. For a more detailed investigation, superquadric tensor glyphs were used. This explicit encoding was calculated by dividing the XCT dataset into individual cells. For each of these cells a tensor is generated by firstly calculating the fiber orientation for each fiber and then averaging the values for all fibers located in the specific cell. The comparison of the superquadric tensors was facilitated by superimposing them on their specific cells. With this tool the identification of regions with similar fiber orientation is simplified, as can be seen in Figure 4.3.

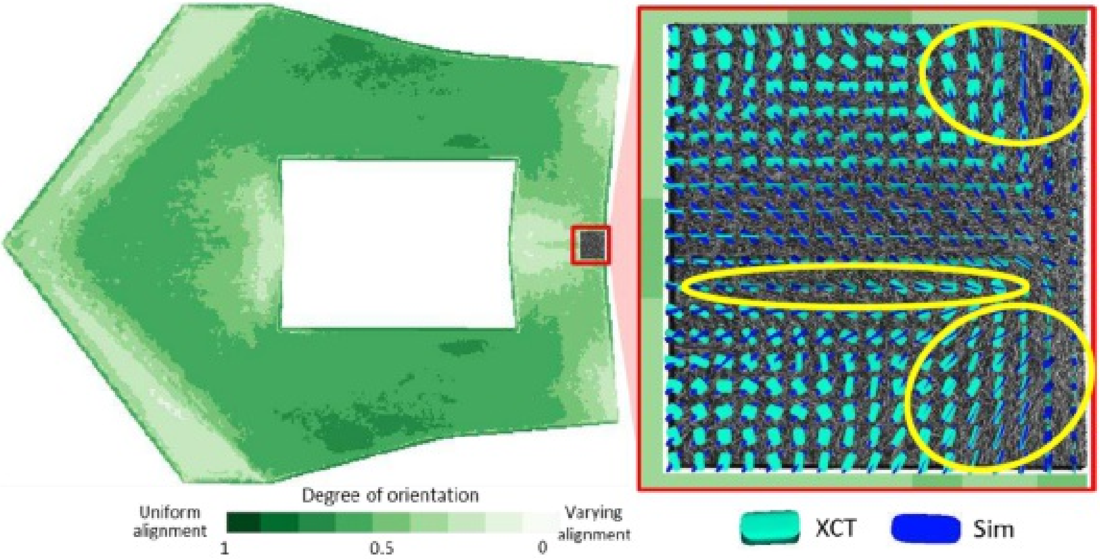


Figure 4.3: Visualizations developed by [WAS+18]: (left) heatmap visualizing the fiber orientation correlations, (right) superquadric tensor glyphs showing the orientation correlations in detail.

In summary, there are several promising approaches for visualizing single material samples in detail. However, there are still limitations. For example, when visualizing multivariate attributes of the individual structures of a specimen, currently the representation is still relying on basic charts such as the parallel coordinate plot or the scatterplot matrix, as used in FiberScout. As can be seen in Figure 4.1, these charts are only partially suitable, since they suffer from severe cluttering when visualizing multivariate attributes generated from segmented and quantified XCT datasets.

One solution for this problem is reducing the complexity of the data by focusing on one or two characteristics and neglecting the remaining attributes. In FiberScout, for example, additional charts were introduced to visualize specific characteristics like length and orientation distribution. In the work of Chiverton et al. [CIBP17], Bhattacharya et al. [BWW+17] and Weissenböck et al. [WAS+18] the specimens were only observed according to one or two attributes. The biggest disadvantage of this approach is that a lot of information is just ignored and therefore simply lost.

Furthermore, although all these works make a valuable contribution to the visualization of materials, they do not include the real-world workflow of material scientists. The comparison of many materials is essential to be able to make correct and universal statements about the properties and future behavior of the specimens. Therefore even more advanced visualization techniques are needed that manage the examination of several specimens at once and actively support comparison tasks. Some works fulfilling this specification will be described in the next section.

4.2 Ensemble Visualization Techniques

Reh et al. [RAK$^+$15] proposed a visual analysis framework to evaluate a series of in situ XCT scans, which cover the evolution of an ongoing process. To investigate the change of the internal structures of a component over time, individual features were extracted and tracked throughout the process. Different visualization methods such as a fuzzy tracking graph, a volume player, and an event explorer, were introduced to track changes in the data. During this tracking procedure, each feature is assigned to a corresponding event, i.e., creation, continuation, split, merge, or dissipation. From this information an uncertainty tracking is computed between adjacent time steps. For visualizing the uncertainty tracking a fuzzy tracking graph was developed, where tracked features were colored along time steps. The volume player is a spatial visualization that displays the events. By volume blending two subsequent time steps, changes in the data series can be observed. Consisting of different scatterplots visualizing snapshots of individual time steps, the event explorer can be used to get an overview of the distribution of the events in the data. The different views are linked together enabling a detailed examination of several time steps of one material. Figure 4.4 gives an overview of the different visualizations developed by Reh et al. [RAK$^+$15].

Amirkhanov et al. [AAS$^+$16] presented a visual analysis system, which enables material scientists to follow the creation and development of different defect types in a series of XCT scans of fiber reinforced polymers. An automatic approach was designed to extract and classify four different defect types: matrix fracture, fiber/matrix debonding, fiber pull-out, and fiber fracture. For visual analysis a framework, consisting of multiple coordinated views, was implemented. While in one view the defects are highlighted in the original CT image, another view, called the Defect Density Map, provides an overview of the defect areas and the distribution in 2D and in 3D. Furthermore, the surface of the defects can be estimated according to the material's fracture location, which is displayed as a 3D surface. For detailed exploration a 3D Magic Lens was developed, which enables an interactive exploration by allowing to compare a region of interest of one specimen with two different time steps by using a side-by-side detailed view. Figure 4.5 shows the volume rendering of a defect development process.

In the recent work of Weissenböck et al. [WFG$^+$19] the main focus laid on the comparison

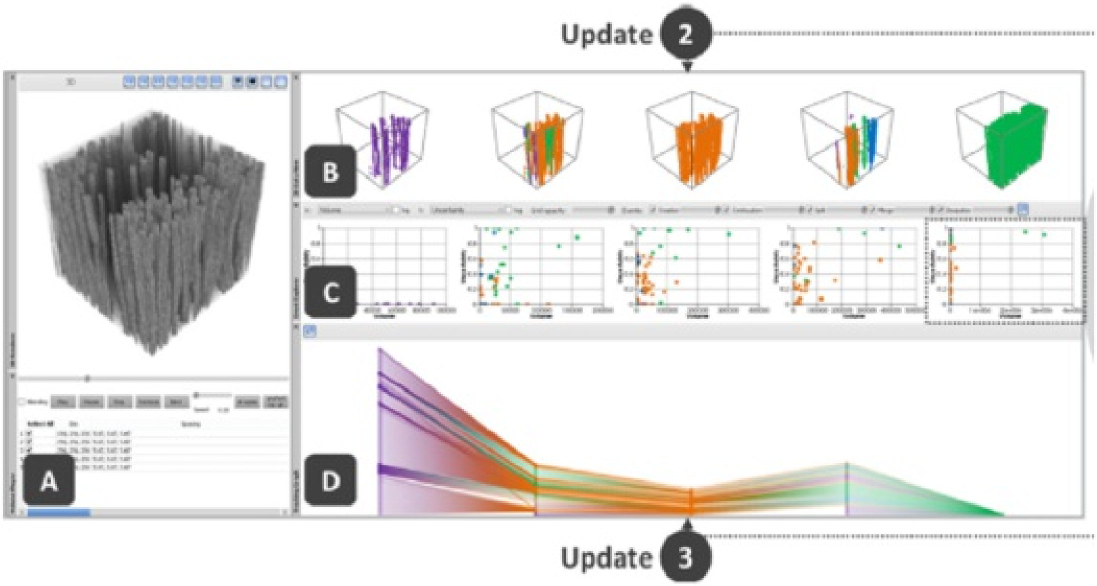


Figure 4.4: System Overview of the framework developed by [RAK+15], showing a (A) VolumePlayer, (B) 3D DataView, (C) EventExplorer and (D) FuzzyTrackingGraph.

of many spatial XCT scans of specimens. All members of an ensemble of volumes were transformed by linearizing each volume along a Hilbert space-filling curve into a 1D Hilbert line plot. As an overview, first a heatmap of the intensity frequencies was presented to the user. When zooming in, the frequencies were replaced by superimposed 1D Hilbert line plots, where one line represents the evolution of the intensities of one volume. With this visualization framework the visual analysis of many different regions is possible in parallel. Interesting spatial regions can be identified quickly based on high local intensity variations. Further the investigation of repeating patterns is facilitated, whereby the most suitable volumes can be identified. Figure 4.6 shows the heatmap visualization, while Figure 4.7 displays the non-linear scaled Hilbert lines plots and the functional box plot representation.

As can be seen, there are already some relevant works bridging the gap between material scientist's workflow and comparative ensemble visualization. However, there is still much potential for improvement. On the one hand, two of the three works focus on analysing different time steps of one specimen. Since both works abstract the development of intensity values of different points in time to higher order events, as defect creation or development, the comparison of additional details, like individual characteristics, is not possible. Further the proposed works only enable a detailed investigation of one specimen at a time. When the development of several materials should be analyzed, material engineers must again return to the inefficient method of comparing the differences manually between equal time steps. Only in the latest work of Weissenböck et al. [WFG+19] various independent specimens can be compared. All three works have in common the basis of

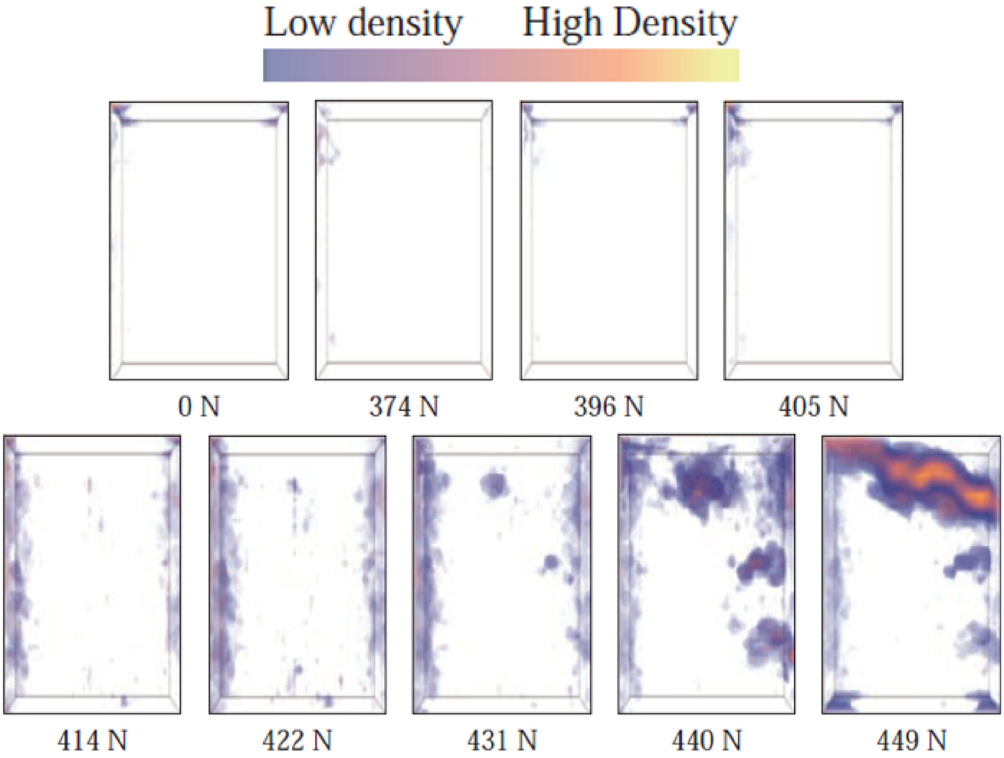


Figure 4.5: Direct volume rendering of a defect development process from Amirkhanov et al. [AAS+16].

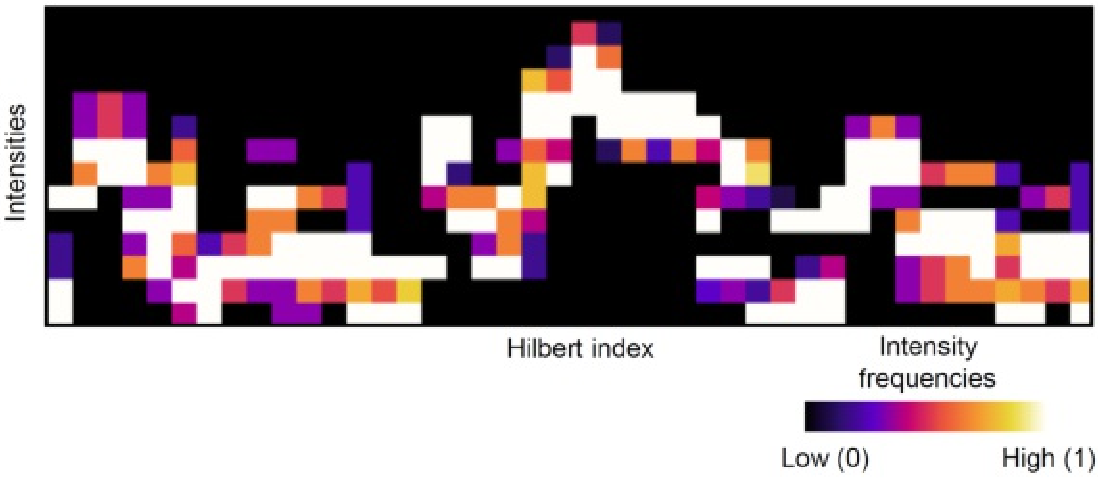
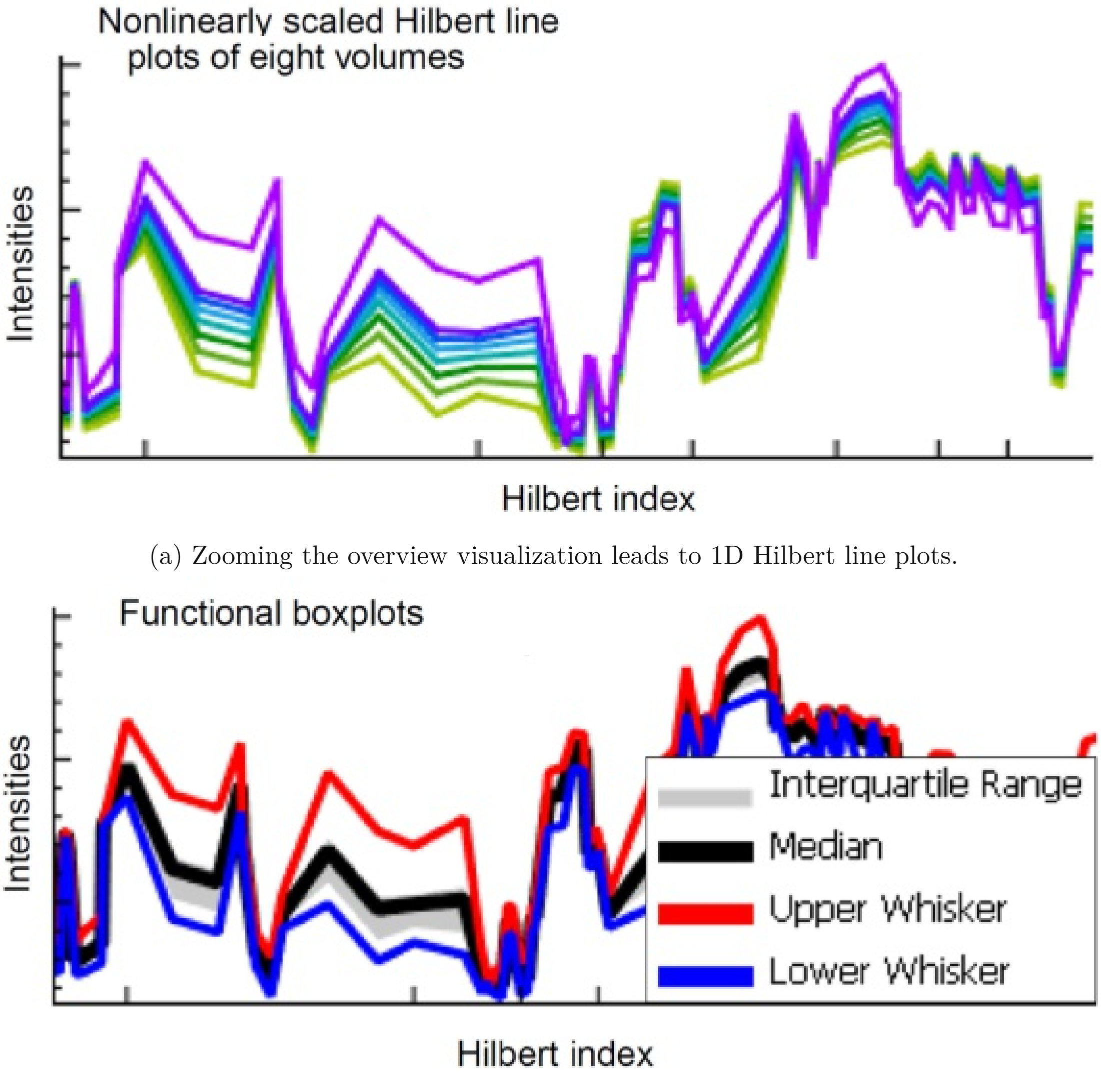


Figure 4.6: Overview visualization, developed by Weissenböck et al. [WFG+19], showing a scaled histogram map, that encodes intensity frequencies.



(a) Zooming the overview visualization leads to 1D Hilbert line plots.

(b) Functional box plots can be computed based on the individual Hilbert line plots.

Figure 4.7: Detailed representations of various volumes, visualized by the work of Weissenböck et al. [WFG+19].

their data. All focus on visualizing the spatial volume, but so far no one has yet dealt with the comparative visualization of the secondary multivariate data of several specimens. There is a high potential of gaining new insights and discovering further knowledge. Therefore our work focuses on the visualization of various multivariate ensemble datasets.

Ensemble visualization Techniques are the modification and adaptation of traditional visualization methods in combination with advanced interaction methods, to overcome the problems of the extra member dimension and multi-facetedness of ensemble data. With these methods, it is possible to understand ensemble data without having to examine each ensemble member individually [Cro18]. In the literature on ensemble visualization we see that analysis frameworks tailor their techniques to the types of domain data to be analyzed. The type of datasets represented most in ensemble visualization frameworks are climate data, since the generation of various models is crucial for climate research and is becoming increasingly popular. Therefore, the creation of ensemble visualization systems that work with spatial and spatio-temporal data has so far been investigated in great detail.

Kappe et al. [KBL19] analysed climate simulation ensembles by identifying groups of datasets with similar parameters. Their analysis system provides three different visualizations: a clustering timeline, a choropleth map, and a heatmap with a bar chart. Together they give a precise overview of data similarity and its variations over time. First the time-dependent 2D scalar field data is aggregated with a k-means clustering algorithm. The result is then visualized in a clustering timeline, giving a concise summary over the temporal occurrences of the clusters. To validate the computed clusters, a heatmap shows the similarity between the individual time steps. To combine the abstracted clustering timeline data with the original input data, filled contour maps are used. For all ensemble members and all time steps, this geospatial visualization can be displayed, visualizing binned scalar values of the unprocessed input datasets. Figure 4.8 shows the cluster-based climate ensemble analysis tool.

The analysis of time-dependent cloud ensemble datasets is the focus of the paper of Kumpf et al. [KSW19]. For aggregation first a t-Distributed Stochastic Neighbor Embedding (t-SNE) in the 2D space is used. To obtain the final clustering, various t-SNE projections are aggregated by a k-means clustering. The similarity between the final clusters is measured by comparing the distribution of their parameters. Pairwise comparison of clusters is then visualized using a cumulative distribution function (CDF), which consists of two lines, each representing a cluster. The area between the two lines defines the distance between the clusters. Further the Kuhn-Munkres-algorithm is applied to provide a one-to-one or a one-to-many matching of different clusters. Two or several matching clusters are identified automatically and visualized in a scatterplot matrix. For directly visualizing the high-dimensional data points, parallel coordinate plots are used. To further ease the comparison of the value distributions, histograms are added on the coordinate axes of the parallel coordinates for each displayed cluster. Figure 4.9 shows the 2D representation of the different clusters, and the CDF visualization of two clusters.

One of the few ensemble visualization frameworks, enabling the comparison of domain-independent volume ensemble datasets, was designed by Demir et al. [DDW14]. Datasets visualized with this system were climate data, as well as fluid simulation data. The aim of this work was to guide the analysis process towards relevant characteristics by visualizing statistical properties of 3D ensemble fields. By combining volume and information visualization techniques they seek to effectively uncover uncertainties, correlations, and trends. As a first aggregation step, the 3D voxel data is linearized along a space-filling Hilbert curve. The aggregated information and statistical values derived from it are

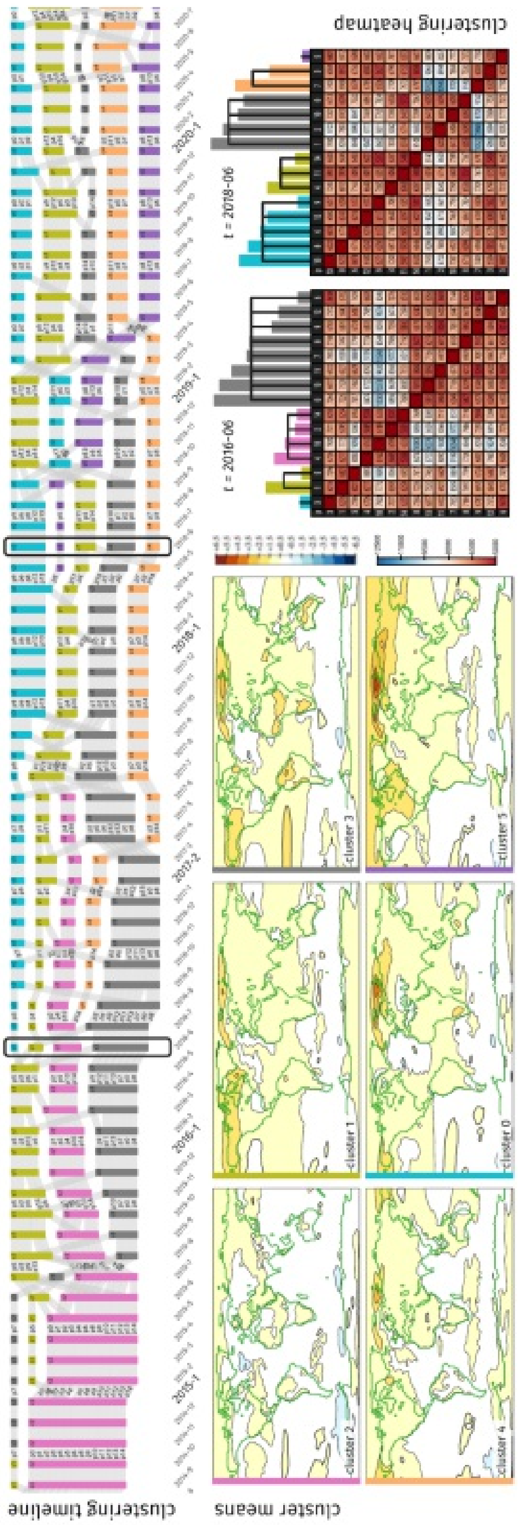


Figure 4.8: System overview taken from [KBL19]: (top) The clustering timeline gives an overview of the different clusters and their temporal occurrences. (bottom left) the contour maps visualize the original input data using discrete color-mappings. (bottom right) The heatmap shows the pairwise distances between individual time steps.

visualized in multi-charts. These charts are a combination of bar and line charts in one plot area. While the bar charts represent histograms and probability densities, line charts are overlaid to allow the viewers to compare the individual members against the ensemble. By calculating the similarity between the histograms, clustering is performed. Furthermore, they are put into spatial context by visualizing selected distributions and regions via volume rendering. Additionally, they provide a correlation computation, allowing users to draw conclusions about the connections between various values of different spatial positions. Figure 4.10 shows the overall structure of the analysis framework.

MotionRugs is a visualization framework developed by Buchmüller et al. [BJC+19] aiming to get insights into the movement patterns of groups, like swarms of fish or birds. This work solves the challenges of identifying local or time-dependent patterns in this spatio-temporal movement data. For aggregation each 2D time frame of the dataset is linearized using a Hilbert Curve, a space-filling curve approach, or a tree representation, called R-Tree. Depending on the chosen linearization method, the identified entities are arranged in a certain 1D order. Each entity in the 1D sequence is then visually encoded with a rectangle colored by its attribute value, for example speed. Finally, the generated sequences are depicted vertically in the visualization, to make a comparison between the different time steps possible. Figure 4.11 provides an overview of the visualization and its interpretation.

Although the works described before all deal with the same type of data, we can see, their solutions are highly adapted to the specific domains, for which the ensemble visualization frameworks were developed. Nevertheless they all follow the visualization pipeline described in Section 3.2. All works use some sort of aggregation technique, either clustering via k-means or linearizing the data points using a space-filling curve. Further the analysis frameworks almost always consist of at least one overview visualization and one to several detail visualizations. For composition, these frameworks mostly use juxtaposition and superimposition to combine the different members in one plotting area. Although all these systems combine very interesting visualizations, these techniques cannot be used for the data discussed in this thesis. While all these systems focus on the visualization of spatio-temporal data, the multivariate ensemble data, discussed in this thesis, is not specifically location-dependent, nor does it always have a temporal component.

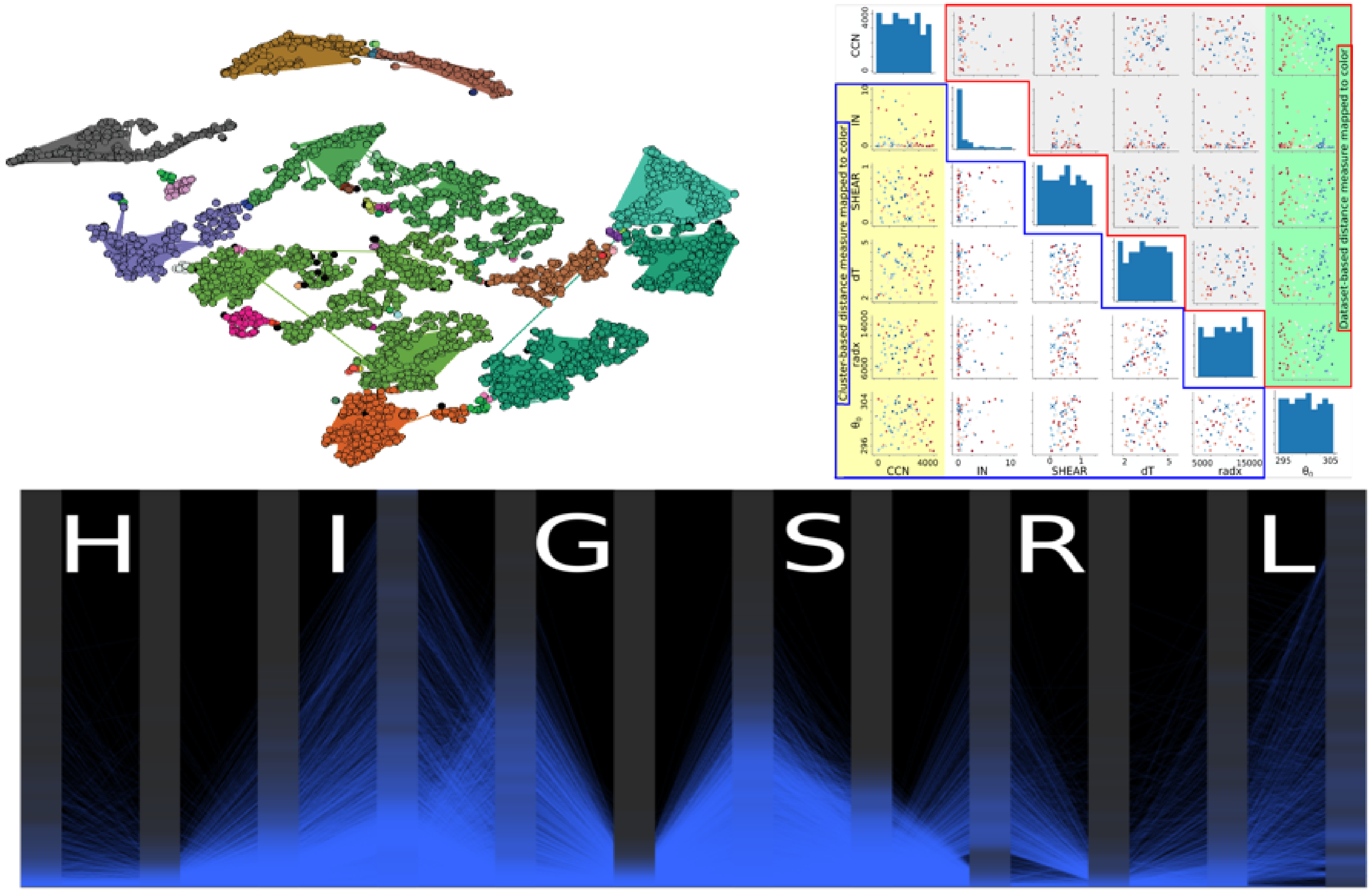
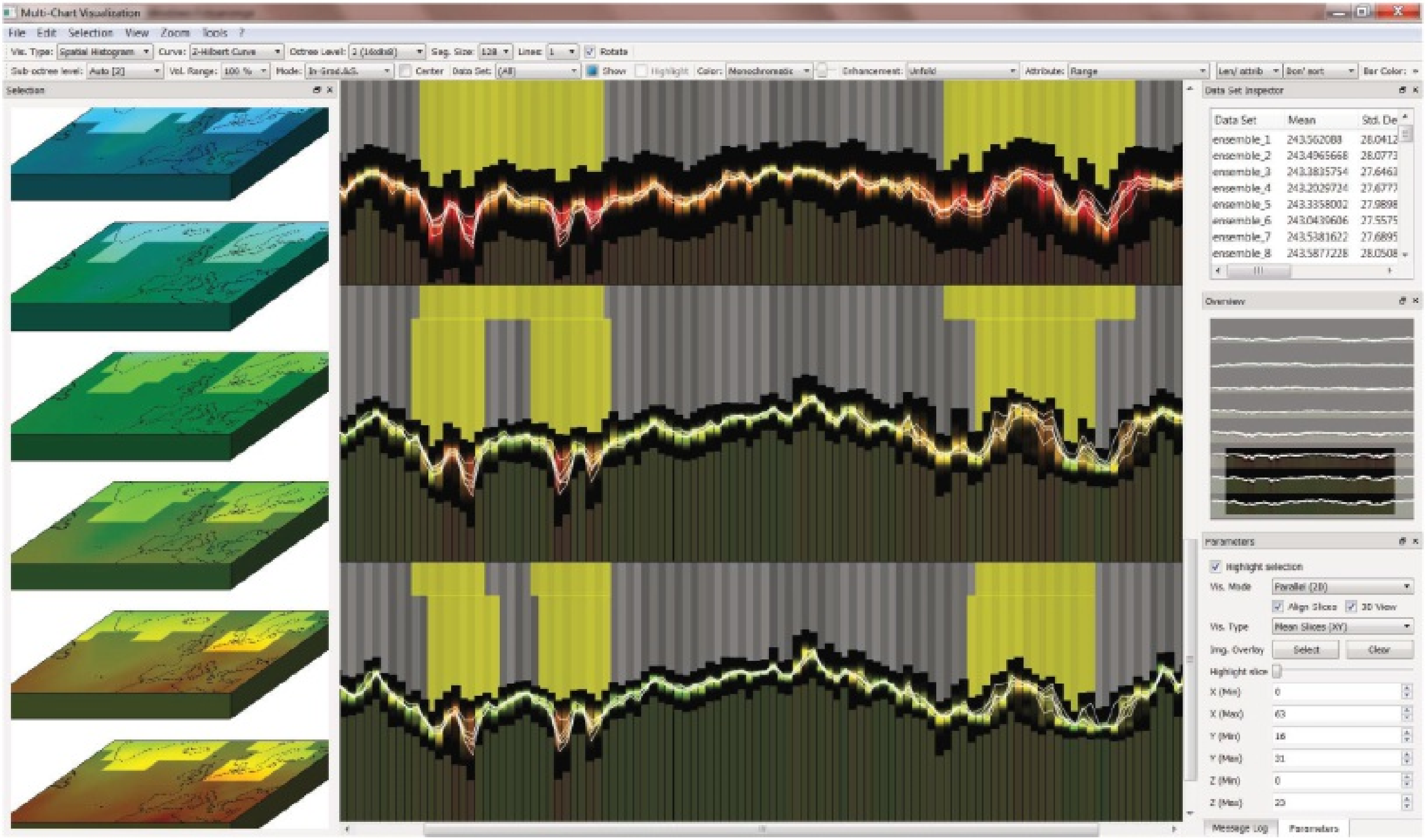


Figure 4.9: System overview taken from [KSW19]: (top left) The scatterplot visualization of the clustered voxel data determined by applying multiple k-means operations on t-SNE projections. (top right) The scatterplot matrix shows the automatic matching of the clusterings. (bottom) parallel coordinates plot shows the per-voxel parameters.


Multi-Chart Visualization
File Edit Selection View Zoom Tools ?
Selection
Overview
Message Log
Parameters
Select
Clear

Vis. Type: Spatial Histogram Curve: 2-Hilbert Curve Octree Level: 3 (16x8x8) Seg. Size: 128 Lines: 1 ☑ Rotate

Sub octree level: Auto [2] Vol. Range: 100 % Mode: In-Grad.&S. □ Center Data Set: (All) ☑ Show □ Highlight Color: Monochromatic Enhancement: Unfold Attribute: Range Len/ attrib Bon/ sort Bar Color:

Data Set Inspector

Data Set	Mean	Std. De
ensemble_1	243.562088	28.0412
ensemble_2	243.4965668	28.0773
ensemble_3	243.3835754	27.6463
ensemble_4	243.2029724	27.6777
ensemble_5	243.3358002	27.9898
ensemble_6	243.0430606	27.5575
ensemble_7	243.5381622	27.6895
ensemble_8	243.5877228	28.0508

Parameters

☑ Highlight selection
Vis. Mode Parallel (2D)
☑ Align Slices ☑ 3D View
Vis. Type Mean Slices (XY)
Img. Overlay
Highlight slice
X (Min) 0
X (Max) 63
Y (Min) 16
Y (Max) 31
Z (Min) 0
Z (Max) 20

Figure 4.10: System overview taken from [DDW14]: The multi-charts visualization tool consists of a combination of a bar chart and a line chart displaying statistical information. The multi-charts are linked with the volume visualization, to enable detailed analysis of the weather forecast data.

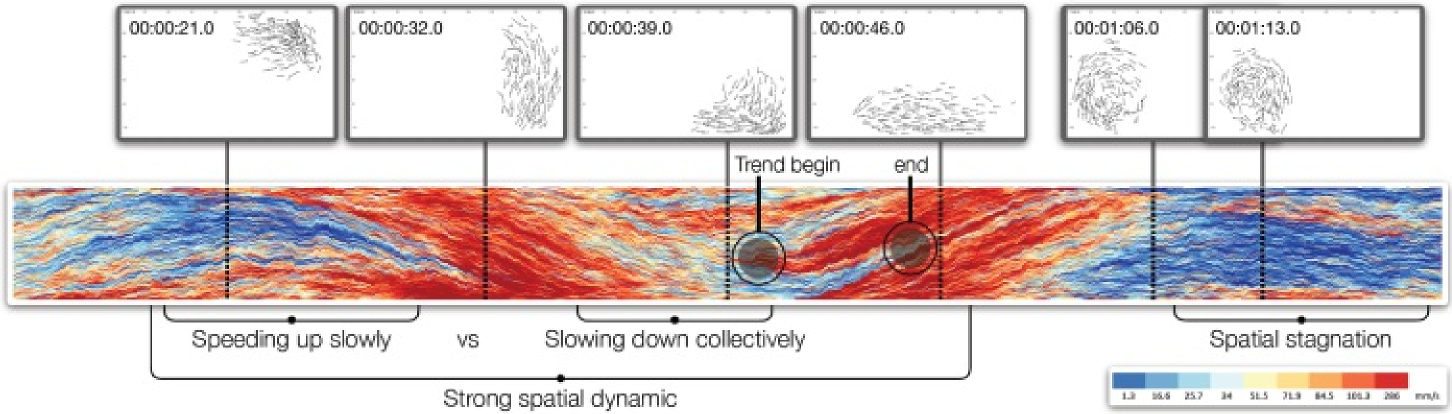


Figure 4.11: System overview taken from [BJC+19]: Each vertical line of the overview visualization encodes a 2D frame of motion data, thus facilitating the exploration of collective patterns.

Since most research in the field of ensemble visualization is done on spatial or spatio-temporal data, the visualization of multivariate ensemble data has not been in focus so far. Therefore, there are only a few papers dealing with visualization techniques for non-spatial data [WHLS19]. Most papers, visualizing multivariate data alone, or data related to the spatial domain, use parallel coordinate plots [MGKH09, ZMM12, HTWL18, KSW19], scatterplot matrices [KSW19], or simple heat maps [JFSK16, KBL19] to give an overview of data similarity. These plots are often chosen because they do not require major modifications to represent multivariate ensemble datasets [WHLS19, TAES09]. In principle, these techniques are designed to enable the identification of patterns or trends in many objects and their characteristics. However, their readability decreases with the amount of data, since clutter makes it difficult to identify trends or individual objects.

To visualize ensemble datasets based on high-dimensional multivariate data, considering some works focusing on the visualization of large amounts of multivariate data is essential. The field of Genome Data Visualization faces the same challenges and strives for the same goals as the ensemble visualization developed in this thesis. In this research area the goal is to find similarities and outliers in very complex datasets. Although genome data has not yet been classified as ensemble datasets, we give arguments for doing so. Based on the five dimensions of ensemble datasets, we can show that genome data can also be described in this form. First, there exist various different types of genome data depending whether the genome inspected originates from viruses, bacteria, etc. (ensemble dimension). For each species and within each species, the structure of the genomes can vary considerably (member dimension). Furthermore, genomes can evolve and change over time (time). In their structure, genomes consist of tens of thousands of DNA sequences (genes) that are connected to each other (location dimension). The individual sequences in turn consist of various combinations of different base pairs (variable dimension). As shown above, genome datasets are also very large and complex. Since the analysis and interpretation of these multilayered data are routine in the fields of biology and medicine, there are some works focusing on their visualization [ADG11, CBS+15].

Albers et al. [ADG11] aim to visualize up to 100 different genomes according to their underlying DNA sequences in their tool called SequenceSurveyor. They focus on providing a concise overview, to allow the viewers to compare sequences more quickly, and facilitate the task of finding matching regions by using alignment visualization. The overview representation consists of horizontal stripes, where each row shows the sequences of one genome. The visual encoding of the different attributes of the sequences, is defined by the user and can be changed at any time. There are two different mappings possible, to position as well as to color. By assigning different combinations of attributes to position or color, different patterns can be made visible. Since the size of the sequences exceeds the number of available pixels on a screen, the sequences have to be aggregated to be displayed. For aggregation, the sequences are grouped in blocks. The color of the blocks is determined by four different aggregation methods: aggregation, robust aggregation, event stripping, and color weaving. While averaging reveals high-level trends, robust averaging has the advantage of reducing the influence of outliers compared to the former.

Outliers are highlighted when using event stripping. The most detailed information is provided by the aggregation method color weaving, which visualizes the distribution of the different genes in the blocks. Figure 4.12 shows the alignment overview visualization.

LayerCake is the analysis tool developed by Corell et al. [CBS$^+$15]. This framework is intended to enable the rapid exploration of sequence variations in viral populations. Each row in the overview visualization represents a sample of a viral population. The various areas of the samples are color coded according to the deviation from a previously chosen reference sample. Since it is not possible to visualize all sequences of the samples simultaneously, LayerCake automatically aggregates regions of genomes into discrete bins. The color of the block results from averaging the different variations contained in the block. With this technique it is possible to give a concise overview of thousands of sequences, facilitating the exploration of similarities and differences of variability in viral genomes. By using averaging as an aggregation method, details such as sporadically occurring locations of high variation can be erased. LayerCake supports recovering details, by providing focus + context lenses and event stripping. When clicking on a bin, the user is able to enlarge its content in a detailed view, while the rest of the line is reduced in size. When using the method event stripping, the user has to select a specific threshold. Bins containing the specified range are highlighted by adding a dark red bar inside the bin. This technique facilitates the identification of outliers in the sequence data. Figure 4.13 shows the overview visualization of LayerCake. As can be seen, the identification of patterns of variability is easily possible.

Both works are based on alignment visualizations, which focus on the comparison of similarities of different sequences. Since the main goal is to provide an overview of many different genomes, it is necessary to abstract the very complex data. Therefore mainly statistical methods are used. In the design phase of both works, principles of human perception were considered. For this reason, visualizations are easy to read, similar regions can be found at a glance due to their visual encoding, and there is no clutter restricting readability. All these advantages speak for using different concepts of alignment visualizations also for the comparison of complex material datasets.

Since genome data consist of concatenated sequences, a fixed order is naturally given. Features such as fibers and pores do not have a natural ordering, which is why the visualizations presented here cannot simply be applied to these datasets. But, as was shown in these works, alignment visualizations are a good alternative to conventional visualizations, such as parallel coordinate plots or scatterplot matrices. Since they naturally support the detection of collective patterns and trends, the design of our analysis framework is based on this type of visualization.

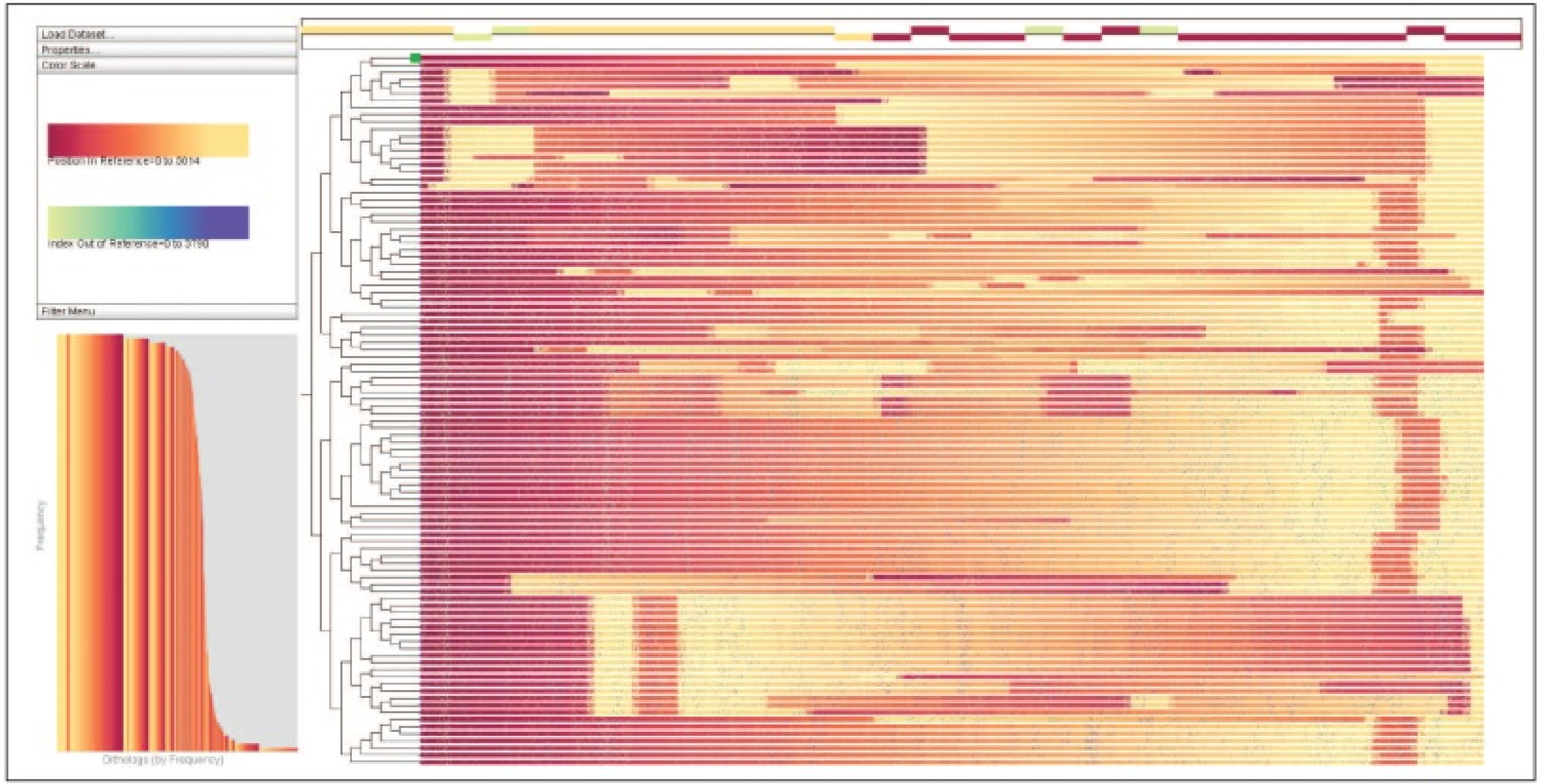


Figure 4.12: System overview taken from [ADG11]: Each row displays one genome according to all its sequences. The position of the individual genes is encoded using color. The first genome, used as reference, defines the mapping of individual genes to the color scheme.

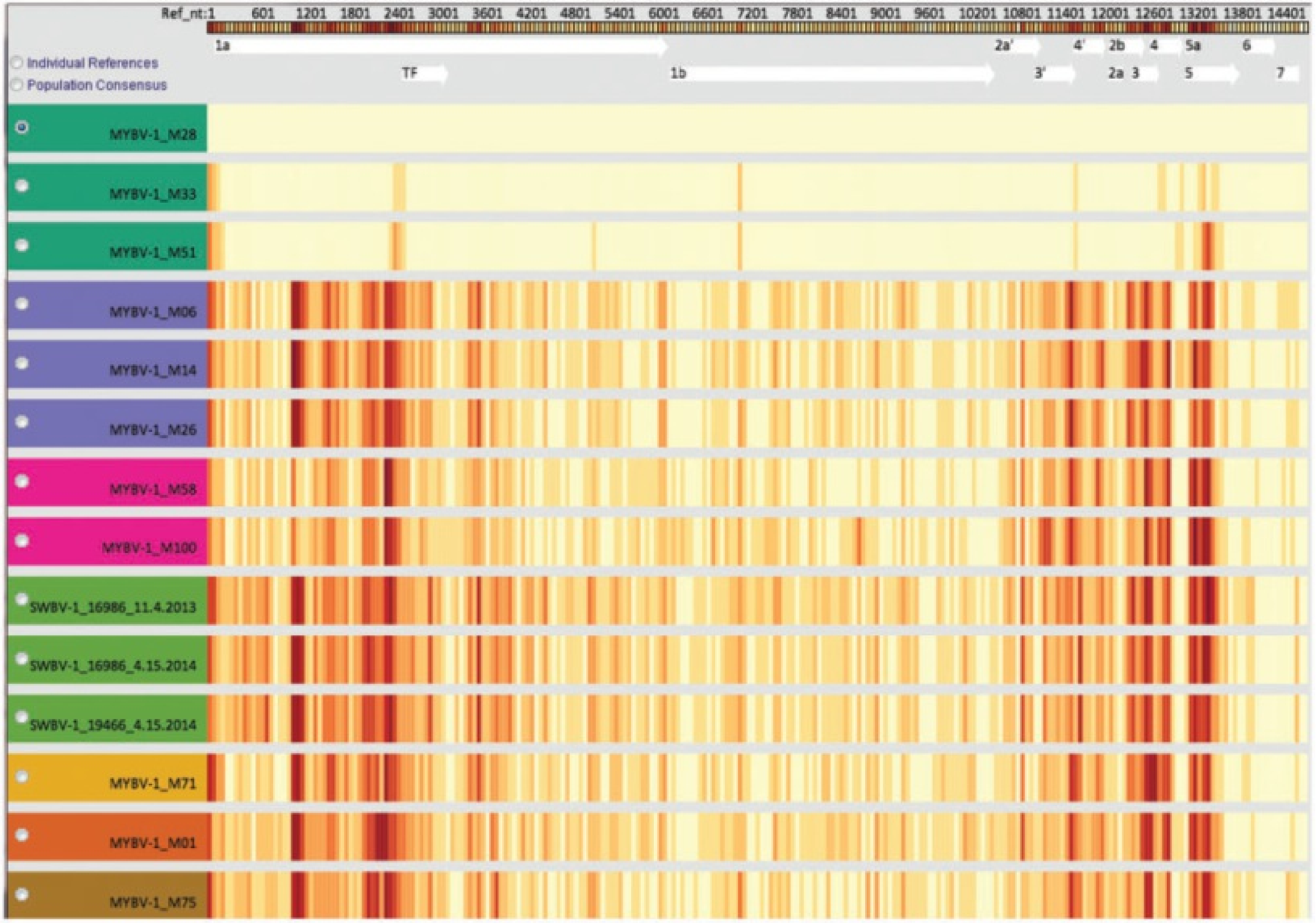


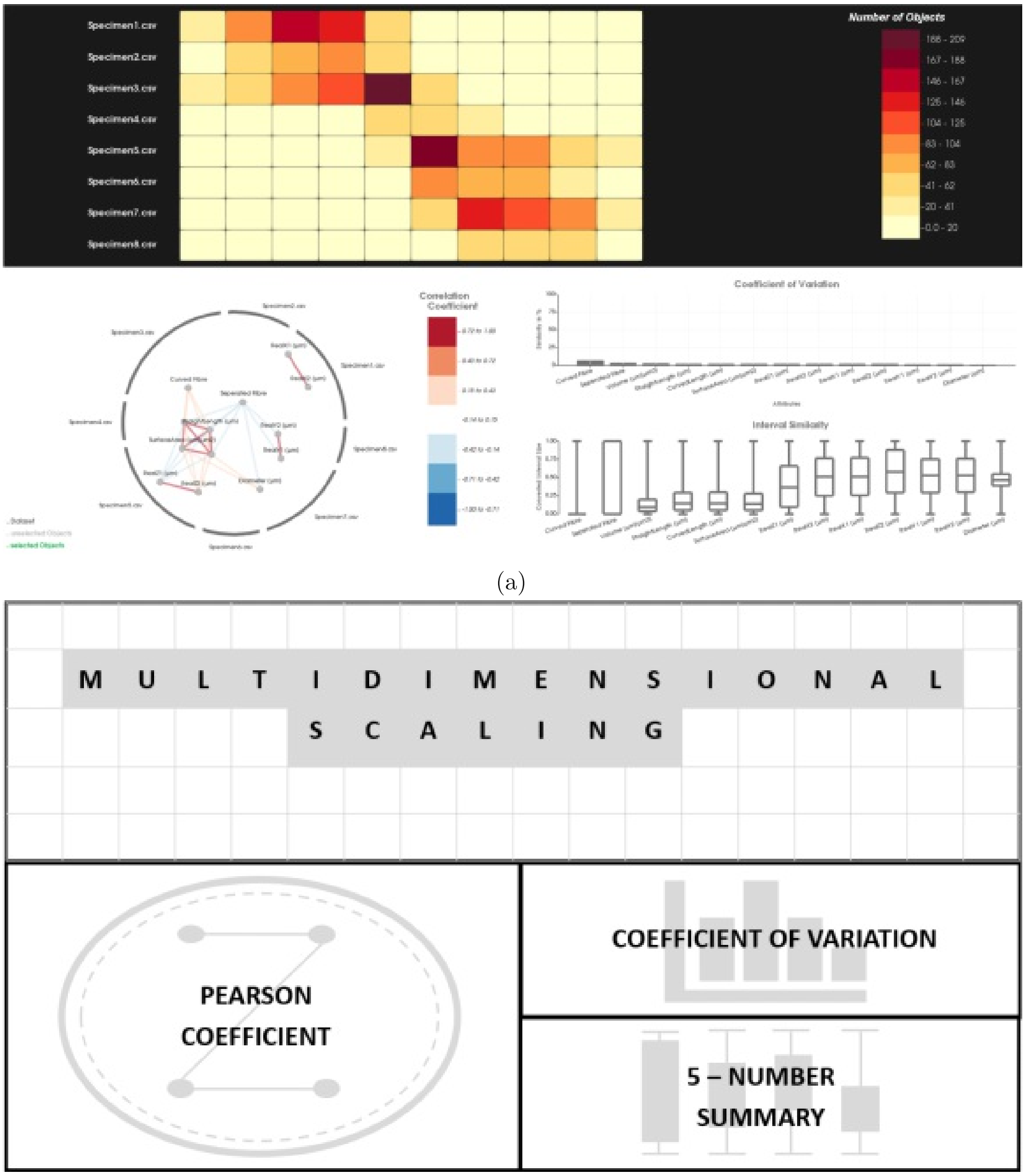
Figure 4.13: System overview taken from [CBS+15]: The overview visualization facilitates the identification of similar subgroups by color-coding the sequence variations in the different samples.

CHAPTER 5

Methodology

To visualize several material samples for comparison, our analysis framework is based on an overview visualization and three detailed visualizations. Since most tools in material science focus on a detailed exploration of one or only several materials at once, we wanted to provide a framework where a huge number of datasets can be compared at one glance. We fulfilled this goal by creating an overview visualization called Histogram-Table. To summarize the huge amount of data we use a dimension reduction method, called Multidimensional Scaling (MDS), to map each individual object in the datasets to a specific position on a 1D line. A dataset is visualized using a horizontal stripe. Each additional dataset is drawn underneath the previous one, in a top to bottom style. For a more detailed exploration we provide two different zooming techniques. To facilitate the comparison task and make a fast detection of similar datasets or regions possible, we implemented an automatic ranking technique based on a histogram distance metric. To identify in which attribute the objects are similar, a statistic parameter, called coefficient of variation, is calculated for every attribute and visualized using a bar chart. In order to explore to which extent the various attributes are similar, a box plot is drawn showing the intervals of the characteristics. Finally, users are also able to make assumptions about the correlation of the individual attributes, by exploring the last detail visualization, called correlation map. Figure 5.1 shows the interface of our analysis framework.

To give a detailed explanation of our visualization approach, first we explain our design considerations. Then we discuss the Histogram-Table in more detail. Finally the computation and depiction of the individual detail visualizations are explained.



(a)

(b)

Figure 5.1: (a) Interface of our analysis framework consisting of an overview visualization and three detail visualizations. (b) Mathematical calculation whose results are displayed in the corresponding visualizations

5.1 System Design

Until now material scientists mostly depend on the sequential comparison of datasets to find similarities and differences. The number of specimens to be manually explored is restricted to a limited number of samples, since otherwise this cognitively demanding task requires even more time, leading to a higher probability of errors. To reduce complexity and allow the users to explore several specimens at a time more efficiently, we first summarized the detailed analysis tasks most important to material scientists when exploring different samples. Secondly, we organized all available information about required tasks and dataset structure, according to the four considerations from Gleicher et al. [Gle18], which help in establishing design solutions for scenarios involving comparison. As stated by Albers et al. [ADG11], including the knowledge of perceptual science is a key advantage when designing visualizations for large amounts of huge data. Therefore, we give a brief overview of the perceptual principles that have been studied and on which we tried to rely as much as possible.

5.1.1 Analysis Tasks

Based on discussions and interviews about their workflow with material experts, the following analysis tasks were determined:

- **Comparing Specimens** - Since minor modifications in the generation process of materials, can have a huge impact on their inner structures, rapidly comparing different samples of composites of the same type would facilitate the work of material scientists. Furthermore, identifying samples of interest from a large amount of data and then having to analyse only the selected ones, could save both working time and effort. Also for educational purposes, a rough overview of different materials could help to memorize figuratively the wide range of variations in their inner structure.

- **Finding Patterns** - Finding patterns in material datasets can help to make more accurate statements or predictions about the behavior of the samples. The comparative analysis of several adjacent areas in one material can also support to predict its mechanical properties.

- **Comparing Individual Objects** - Evaluating groups or individual objects, as fibers and pores, based on their attributes can be useful for exploring specific regions in more detail. Thereby the detailed analysis should enable the examination of all attributes of the objects, but also of individual attributes, which the user is currently most interested in.

- **Identifying Outliers** - By determining similar or different regions inside one specimen, irregularities can be detected. Also the fast detection of outliers in a lot of different specimens can help to focus on different tasks. For example, the user can decide to examine just the similar samples more closely. The interest could also fall on one or more outliers, which can then be investigated in more detail.

5.1.2 Visual Design Considerations

In our work, comparison is the most important task, even influencing other analysis tasks. The visual concepts used in our visualization system are based on the design framework developed by Gleicher et al. [Gle18]. The authors give a broad overview of what needs to be considered when designing a comparative framework and base their recommendations on an extensive literature study. For this reason, we structure our information according to the four defined categories: Comparative Elements, Comparative Challenges, Scalability Strategies, and Comparative Designs.

C1. **The Comparative Elements**, the set of items that should be compared in our final visualization framework, are two-layered. First, a comparison between different specimens should be possible. Each specimen consists of thousands of objects, each described by 10 to hundreds of attributes. Second, a comparison among groups of objects should be possible. **Therefore our visualization tool has to be adaptable to different levels of detail.**

C2. **Comparative Challenges**, determining the degree of difficulty of a comparison, are defined by three categories: the number of items to compare, the complexity of the items themselves, and the complexity of the relationships between the items. Since in this work not only several different datasets, but also groups with varying numbers of objects, should be compared, the number of the comparative elements is quite high, lying in the range of three to hundreds of dataset and thousands of objects. Since the individual objects are described by a large number of characteristics, usually in the range of ten to hundreds, the items themselves are complex as well. Solely the relationships among the elements can be characterized as simple, since each object has the same number and types of attributes, where the variation can be depicted by calculating the difference. **As a result, our visualization framework has to be able to deal with a huge amount of data, providing an abstraction by summarizing the data in an overview visualization. Since exact characteristics are necessary for domain scientists to analyse and interpret the specimens' properties, details cannot be omitted. Therefore additional detail visualizations are necessary to display important supplementary information about the detailed specifications.**

C3. A **Comparative Strategy** is required to solve the scalability problems imposed through the comparative challenges. To solve comparison challenges, Gleicher et al. categorized three different approaches: scanning sequentially, selecting a subset, and summarization. While the user is able to explore each item serially in the first approach, selecting a subset can help identifying patterns or commodities. Summarization can constitute an advantage when an overview or context-specific properties have to be analysed. **Since we wanted material scientists being able to explore their data on a high-level as well as individually, we combined all three concepts in our tool.**

C4. Comparative Design, describes the visual design used for comparative visualizations. As described by several prior papers [Kim2017, Gleicher2018, Wang2019] there are three basic designs: juxtaposition, superimposition, explicit encoding. While in the first strategy the items are placed in different spaces, next to each other, in the superimposition design the items are place in the same space, on top of each other. In contrast to those, explicit encoding visualizes the relationships between the items. **Since the comparative design is depending on the task at hand, in our tool all three concepts were used in different combinations.**

5.1.3 Perceptual Phenomena

Human perception plays a major role in the field of visualization, since the correct use of graphic features exploits the low-level visual system of human beings and allows them to be recognized without effort. When visualizing huge amounts of information, the human visual system is easily overwhelmed. Therefore understanding perception can significantly improve material science visualization by exploiting perceptual mechanisms [ADG11, FP02]. Perceptual phenomena, on which we focused during the development of our framework are: Pre-Attentive Phenomena, Visual Search, Visual Clutter, and Summarization.

- **Pre-Attentive Phenomena** summarize visual properties that can be identified fast and without effort by the viewer. The use of these pre-attentive features allows the viewers to detect important points among a large number of elements by making them "pop-out". An example where this phenomenon occurs is when finding a few blue points in a set of red ones. However, the number of pre-attentive features is very limited and the combination of several of them is also problematic, since interference can eliminate the effect. Therefore in our work the only pre-attentive graphical feature we use is color. Our encoding emphasizes pre-attentive pattern finding and summarization, since large fields of color symbolize the number of similar objects in the datasets or specific regions.

- **Visual Search** is a common task in comparison, since it occurs when viewers cannot detect targets pre-attentively, but have to look carefully over the whole scene while searching for them. Search tasks can be quite time consuming and demanding without visual aid. Designing tools that incorporate perceptual mechanisms supporting the visual scans are necessary. To efficiently support users in finding similar fiber or pore regions, visual search mechanisms are the key to a guided search. In our visualization tool we map each dataset to a horizontal line, allowing the user to scan the inside of one dataset by horizontal reading, and to search across multiple datasets by vertical reading. In addition, an automatic ranking helps to identify the datasets that are most similar. The more similar the datasets are, the closer they are positioned to each other, minimizing the time needed to search the entire screen.

- **Visual Clutter** prevents from fast identification of single objects or their connections, because the number of elements, the visual encoding, or the layout prevents a direct identification. This slows down the search process and increases the cognitive load of the user. To avoid visual clutter, our framework uses diverse aggregation techniques and a color-field visualization, limiting the displayed information. Clutter can emerge in our prototype only on the most detailed level view, when each individual object is shown.

- **Summarization** describes the aggregation of whole regions or groups by using statistical methods. The result is representative overview information, whereby the context of the information is preserved, while irrelevant details are omitted. Summarization is currently not very common in the visualization of material data because of the danger of losing valuable detail information. However, to be able to compare different materials efficiently, the large amount of data has to be reduced. In our framework the given objects are summarized according to all their attributes to a specific position on a 1D line. Therefore the graphical feature, encoding the similarity of individual objects, is position.

5.2 Overview Visualization - Histogram-Table

The primary goal in the design of our framework was enabling the recognition of patterns and trends of several specimens. Our framework follows the visualization pipeline from Section 3.3 as follows: First an aggregation is applied to the data. Next, the visual encodings are assigned, which are then merged by a composition. In our work the goal is to compare as many objects from as many datasets as possible. To allow the visualization to be highly scalable, the given information must be minimized as much as possible to make a visual summary of the whole data possible. To ensure serial scanning of the datasets or objects, we decided to map every single object into 1D space. The points, originating from the same dataset, are mapped together on a horizontal line. This way a space-filling visualization can be created, which allows the viewers to compare an unlimited number of datasets.

5.2.1 Aggregation with Multidimensional Scaling (MDS)

For aggregation, the dimension reduction technique Multidimensional Scaling was applied. The overall idea of MDS is to map the original high-dimensional data into lower dimensional data. We consider s objects defined in an m-dimensional space M. The goal is to map each of the given s objects into a lower r-dimensional space R, where $r \ll m$. In the original space a proximity measure δ_{ij} describes the correct distance between the pairwise objects (i, j). The computation of this measure results in a matrix

$C = (\delta_{ij})$, where the dimension of C is defined as $dim\,(C) = s \times s$. The MDS results in a configuration matrix X, with the dimension $dim\,(X) = s \times r$, that contains the newly calculated positions of objects in r-dimensional space. To compute the configuration matrix X, the MDS tries to replicate the proximities of the m-dimensional space in the r-dimensional space. To obtain a result for configuration X, the following function is defined:

$$\sigma(X) = \sum_{i<j\leq n} w_{ij}\left(\delta_{ij} - d_{ij}(X)\right)^2 \tag{5.1}$$

The function $\sigma(X)$ describes a loss or stress function that measures the squared differences between the ideal (m-dimensional) distances and the actual distances in r-dimensional space. The term δ_{ij} represents the ideal distance, here called proximity, between the objects in the original m-dimensional space. The term $d_{ij}(X)$ measures the distances between object i and j in the embedding space R. $w_{ij} \geq 0$ describes a weight for the measurement between the pair of objects (i, j). The better the configuration X minimizes the function $\sigma(X)$, the better the points in the new space correspond to the points in the original space.

Input Data for MDS

To obtain points in a lower dimensional space, the MDS requires similarity or dissimilarity information as input. The proximity, denoted as δ_{ij}, describes the closeness information of the points in the input data. For MDS, there exists a multitude of different proximity measures. In general, proximity information is organized in a proximity matrix and different metrics primarily vary in the way they were collected, either directly or indirectly [SNHS18].

- **Direct proximity measures** are based on qualitative judgements or quantitative assessments that directly indicate the condition of the point. So the similarities or dissimilarities of this measure are directly encoded. Examples for direct proximities are subjective rankings, opinions or perceptions, expressed by users. For this reason, variables used in this type of measure are generally dichotomous (binary). Here the similarity is often expressed with presence or absence of the defined phenomena or opinion.

- **Indirect proximity measures** are the result of computations from certain available types of information. In most cases this information is available as a matrix, containing similarities in form of variables. It should be noted that the available variables can be quantitative, qualitative, or mixed.

One major disadvantage when using MDS for similarity computation, is that the generation of proximity information, i.e., the creation of the proximity matrix, is often quite complex and time consuming. For simplification, certain assumptions can be made to ease this process:

1. $\delta_{ij} > 0$, describes non-negativity
2. $\delta_{ii} = 0$, describes identity, so the point is always similar to itself
3. $\delta_{ij} = \delta_{ji}$, describes symmetry, stating that the distance from i to j is the same as from j to i

There are some authors, who disagree with the above conditions [SNHS18]. If $\delta_{ij} = 0$ holds for two objects i and j, it may happen that $\delta_{ik} \neq \delta_{jk}$ for a third point k. To ensure that the above assumptions hold, it is necessary to check whether the similarity information satisfies the definiteness property or the triangle inequality, so either i and j coincide for $\delta_{ij} = 0$ or $\delta_{ij} \leq \delta_{ik} + \delta_{jk}$. If these requirements are fulfilled, well-known distance metrics like the Minkowski family of distances can be applied as proximity metrics. For two objects $i, j \in M$ the Minkowski distance of order p, can be defined as follows if we consider m quantitative variables Y_k with observations y_{ik}:

$$\delta_{ij} = \left(\sum_{k=1}^{m} |y_{ik} - y_{jk}|^p \right)^{\frac{1}{p}} \tag{5.2}$$

For $p = 1$ the Minkwoski metric corresponds to the Manhatten Distance, for $p = 2$ to the Euclidean distance and for $p = \infty$ to the Lagrange and Chebyshev models. Furthermore, Table 5.1 shows various popular proximity metrics, where δ_{ij} describes the dissimilarity and ρ_{ij} describes the similarity between two objects i and j. The first four entries display four different Minkowski distances, that directly integrate dimensional differences. In contrast to the Minkowski family, the models from row 5 to 10 in Table 5.1 control the dispersion of the variables. The Canberra Distance is used if absolute differences should be corrected along each dimension. In addition, if y_{ik} have negative values, then δ_{ij} reaches an asymptote of infinity. Consequently, the Canberra Distance should only be used if solely positive values are involved. The Bay-Curtis distance tries to correct the sum of absolute differences, and is mainly used in the domain of environmental research. Both, this metric and the chord distance perform well with positive values of y_{ik}. The resulting values of the correlation coefficient lie in the range of -1 and 1. In contrast to the other metrics, this distance describes the linear similarity between the individual objects i and j according to the Pearson Correlation. It works especially well for very large data with many dimensions m.

In our work we focused on the distances based on angles, defined in row 9 and 10 in Table 5.1. The Angular Separation, also known as cosine similarity, measures the similarity of two non-zero vectors. This measure is defined to be equal to the cosine of the angle between them, which is also equal to the inner product of the normalized vectors. Angular Separation is a measure of orientation of the vectors, since unit vectors are maximally

No.	Model	Formula
1	Manhattan	$\delta_{ij} = (\sum_{k=1}^{m} \lvert y_{ik} - y_{jk} \rvert)$
2	Euclidean	$\delta_{ij} = \left(\sum_{k=1}^{m} \lvert y_{ik} - y_{jk} \rvert^2\right)^{\frac{1}{2}}$
3	Lagrange and Chebyshev	$\delta_{ij} = \max_{k=1}^{m} \lvert y_{ik} - y_{jk} \rvert$
4	Minkowski	$\delta_{ij} = (\sum_{k=1}^{m} \lvert y_{ik} - y_{jk} \rvert^{p})^{\frac{1}{p}}$ with $p \geq 1$
5	Canberra	$\delta_{ij} = \sum_{k=1}^{m} \frac{\lvert y_{ik} - y_{jk} \rvert}{\lvert y_{ik} + y_{jk} \rvert}$
6	Bray-Curtis	$\delta_{ij} = \frac{\sum_{k=1}^{m} \lvert y_{ik} - y_{jk} \rvert}{\sum_{k=1}^{m} (y_{ik} + y_{jk})}$
7	Chord	$\delta_{ij} = \left(\sum_{k=1}^{m} \left\lvert y_{ik}^{\frac{1}{2}} - y_{jk}^{\frac{1}{2}} \right\rvert^2\right)^{\frac{1}{2}}$
8	Correlation Coefficient	$\rho_{ij} = \frac{\sum_{k=1}^{m} (y_{ik} - \bar{y}_i)(y_{jk} - \bar{y}_j)}{(\sum_{k=1}^{m} (y_{ik} - \overline{y_i})^2)^{\frac{1}{2}} \left(\sum_{k=1}^{m} (y_{jk} - \overline{y_j})^2\right)^{\frac{1}{2}}}$
9	Angular Separation	$\rho_{ij} = \frac{\sum_{k=1}^{m} y_{ik} y_{jk}}{(\sum_{k=1}^{m} y_{ik}^2)^{\frac{1}{2}} (\sum_{k=1}^{m} y_{jk}^2)^{\frac{1}{2}}}$
10	Angular Distance	$\delta_{ij} = 1 - \arccos\left(\frac{\sum_{k=1}^{m} y_{ik} y_{jk}}{\sqrt{\sum_{k=1}^{m} y_{ik}^2 \cdot \sum_{k=1}^{m} y_{jk}^2}}\right) \cdot (1/\pi)$

Table 5.1: Proximity Metrics

similar if they are parallel, while they are maximally dissimilar if they are orthogonal to each other. The resulting similarity values have a range of $[-1, 1]$. Values close to -1 mean the compared vectors are exactly opposite, while 0 indicates decorrelation and 1 denotes equality. The Angular Separation is very similar to the Pearson correlation, because it can be transferred to the Pearson correlation by a simple subtraction of the vector mean values from the vectors.

In contrast to Angular Separation, which expresses the similarity ρ_{ij} between two objects i and j, we consider the dissimilarity δ_{ij} between i and j. In our work, we are only interested in the ordering of the dissimilarities within the set of input vectors. Since the order of the resulting dissimilarities is always the same, regardless which of the angular functions was chosen, we neglected the term $(1/\pi)$ and we do not subtract the arccos-term from 1. To assign different importance to the individual attributes y_{ik} and y_{jk} of objects i and j, we add a weighting described by α_k. With this weighting, the users are able to determine the similarity of the objects based on the attributes that are of interest to them. The resulting proximity calculation, used in our implementation, has the following structure:

$$\delta_{ij} = \arccos\left(\frac{\sum_{k=1}^{m} \alpha_k^2 y_{ik} y_{jk}}{\sqrt{\sum_{k=1}^{m} \alpha_k^2 y_{ik}^2 \cdot \sum_{k=1}^{m} \alpha_k^2 y_{jk}^2}}\right) \tag{5.3}$$

Before the MDS is applied to the data in our approach, each value of each column of the initial matrix is divided by the maximum value of its column. This ensures that all values of a column lie in the range of $[0, 1]$ and prevents saturation effects of the numerical values. Each row vector of the matrix represents an item, i.e. a fiber or pore, the individual columns represent the individual attributes. In general, the resulting proximity matrix $C = (\delta_{ij})$ is symmetric, non-negative, and is hollow, thus has a diagonal filled with zeros [dLM09]. C can now be used as input proximity matrix for the MDS calculation.

Loss Function for MDS

To map high-dimensional data points to a lower dimension, the MDS positions all points in a low-dimensional Euclidean space, in such a way that the distances between the points approximate the given dissimilarities. To get the best result, the MDS tries to find the best configuration matrix X such that the Euclidean distance $d_{ij}(X)$ approximates the proximity δ_{ij}:

$$d_{ij}(X) \approx \delta_{ij}$$

This equation can be represented as an optimization problem by the stress function shown in Equation 5.1. The authors Leeuw et al. [dLM09] developed an efficient algorithm, named SMACOF (Scaling by Majorizing a Complicated Function), that minimizes the stress function by iterative majorization. In each iteration the stress $\sigma(X)$ is calculated and the algorithm can either be stopped if the difference between the stress value of the previous iteration step and the stress value of the current iteration is quite small, lower than a specified threshold, or if a certain iteration limit is reached. Majorization guarantees a reduction or at least preservation of the stress value for each iteration step and provides a linear convergence rate. A disadvantage of this method is the possibility to get stuck in local minima. The smaller the dimension of the final low-dimensional space is specified, the more likely the algorithm will return a local minimum as result [dLM09]. In this work we will not explain the mathematical details of SMACOF, but introduce the fundamental equation, called Guttman transform:

$$X = V'B(Z)Z \tag{5.4}$$

where V' is the Pseudo inverse, also called Moore-Penrose inverse. The Guttman transform is obtained by setting the derivative of $\sigma(X)$ equal to zero, that is, $\nabla\sigma(X) = 0$, where ∇ symbolizes the derivation. This equation can also be written as $VX = B(Z)Z$.

$$V = (v_{ij}) \tag{5.5}$$

$$v_{ij} = \begin{cases} -w_{ij} & \text{if } i \neq j \\ \sum_{i \neq j} w_{ij} & \text{if } i = j \end{cases} \tag{5.6}$$

$$B(z) = (b_{ij}) \tag{5.7}$$

$$b_{ij} = \begin{cases} -w_{ij}\delta_{ij}/d_{ij}(Z) & \text{if } i \neq j \\ 0 & \text{if } d_{ij}(Z) = 0, i \neq j \\ -\sum_{i \neq j} b_{ij} & \text{if } i = j \end{cases} \tag{5.8}$$

If the weights are equal for all distances, i. e., $w_{ij} = 1$, then

$$V = s\left(I - \frac{ee^t}{s}\right) \tag{5.9}$$

$$V' = 1/s\left(I - \frac{ee^t}{s}\right) \tag{5.10}$$

where $e = (1, ..., 1)^t$ is a unit vector whose length is r, as the resulting lower dimensional space R, and t denotes the transpose of the vector. I is the identity matrix, and s is the number of objects. In our work we use Equation 5.10 for V'. The iteration of the SMACOF algorithm is processed by substituting $X^{[k-1]}$ into Z of Equation 5.4, resulting in:

$$X^{[k]} = V'B(X^{[k-1]})X^{[k-1]} \tag{5.11}$$

where $X^{[k]}$ is the configuration matrix after k iterations. Based upon the previous definitions, the SMACOF algorithm now executes the following steps. First $X^{[0]}$ is initialized with a random or predefined configuration matrix. In each iteration k a new configuration $X^{[k]}$ is calculated by computing the Guttman transform of Equation 5.11. Then the difference $\triangle\sigma(X^{[k]})$ between the stress of the old configuration and the stress of the new configuration is calculated, that is, $\triangle\sigma(X^{[k]}) = \sigma^{[k-1]} - \sigma^{[k]}$. If $\triangle\sigma(X^{[k]}) < \epsilon$, where ϵ is a relatively small threshold constant, the computation is finished and the configuration matrix $X^{[k]}$ is taken as result. The iterative process performed by the SMACOF algorithm is summarized by the pseudo code example in Algorithm 5.1. Since in our work the information regarding each object should be compressed as much as possible, we chose the one-dimensional space as the result space, thus for our case $r = 1$. For a more detailed description of the derivation of the SMACOF algorithm, we refer to Leeuw et al. [dLM09] and Bae et al. [Bae08].

Output of MDS

The resulting output matrix X now stores the information about the relationship between all objects, where similar elements are located in close proximity to each other, while dissimilar elements are positioned far apart. The coordinates produced by the MDS are

Algorithm 5.1: SMACOF

```
Input: A random configuration matrix $\mathbf{X}^{[\mathbf{0}]}$ and a threshold $\epsilon$
Output: A configuration matrix $\mathbf{X}$
$\mathbf{Z} \leftarrow \mathbf{X}^{[\mathbf{0}]}$;
$k \leftarrow 0$;
Compute $\sigma^{[0]} = \sigma(\mathbf{X}^{[\mathbf{0}]})$;
while $k = 0$ or $\left(\triangle\sigma(\mathbf{X}^{[\mathbf{k}]}) > \epsilon \textit{ and } k \leq \textit{maximum iterations}\right)$ do
    $k \leftarrow k + 1$;
    Compute the Guttman transform $\mathbf{X}^{[\mathbf{k}]}$ given in Equation 5.11;
    Compute $\sigma^{[k]} = \sigma(\mathbf{X}^{[\mathbf{k}]})$;
    $\mathbf{Z} \leftarrow \mathbf{X}^{[\mathbf{k}]}$;
end
$\mathbf{X} \leftarrow \mathbf{X}^{[\mathbf{k}]}$;
return $\mathbf{X}$;
```

not unique, since the output configuration is not explicit in terms of reflection, rotation, or dilation. For this reason a comparison of several configurations cannot be performed without considering the ambiguity caused by these three transformations [SNHS18]. In our approach the MDS generates a configuration matrix X of size $dim\,(X) = s \times 1$, in which for each object i of all used datasets a certain position was computed.

5.2.2 Visualization

The overview visualization, developed in our work, uses the output matrix of the MDS and visualizes each object according to its computed position. All objects are ordered according to their affiliation to a certain dataset. The objects of a dataset are mapped to a common horizontal number line. Each object is represented as a circle according to its calculated position on the respective number line. The MDS maps similar objects close together and dissimilar objects farther apart. Therefore points that are located closer to each other represent similar elements, while points positioned further apart are more dissimilar. The numerical values, calculated by the MDS, cannot be interpreted on their own, since merely the differences between them are meaningful. Therefore we have omitted concrete axis labels in the visualization to not distract the viewer during the analysis. Figure 5.2 visualizes the mapping of the objects to their positions on the appropriate number lines, resulting in a point visualization to provide an overview.

If we look at dataset *D1* in Figure 5.3, we can see that many points on the horizontal line are located further to the left, when viewed from the center. Only a very small percentage of the points are further to the right. These elements are outliers in this dataset, since their positions do not align with the positions of the majority of the points. Since the MDS considers the distances from all objects from all datasets, different datasets can be compared with each other. Points that are vertically below each other are also similar, because they have the same horizontal position. For example, if we compare the horizontal line of dataset *D1* with that of dataset *D2*, we see that both datasets are rather different. While in *D1* the majority of the elements is more to the left, the elements of *D2* are positioned from the center to the far right. The rightmost positioned elements of *D1* and the objects of *D2*, positioned underneath, are quite similar, marked in light green in Figure 5.3. The outermost right elements of *D2* and the outermost left ones of *D1* are most dissimilar, highlighted in light red in Figure 5.3. Since there are thousands of objects to be visualized, mapping each element to one specific position leads to clutter in the visualization, as indicated by a dark red marking in Figure 5.2.

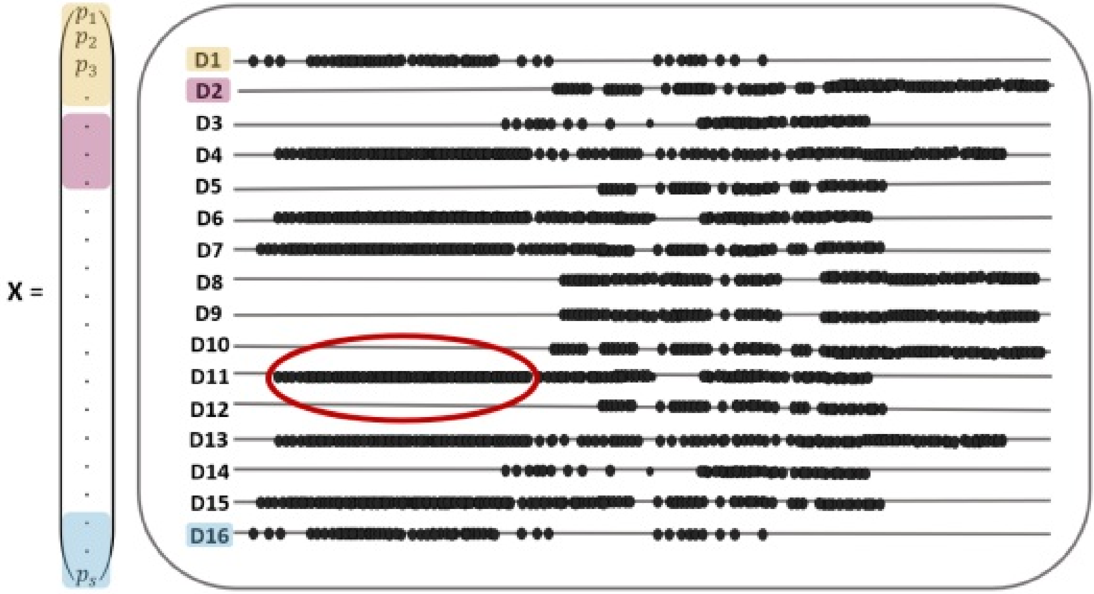


Figure 5.2: Illustration that shows how the individual objects are visualized on their respective number lines. For each dataset D_j, $\forall j = 1, ..., 16$, a number line is displayed. Each object i, $\forall i = 1, ..., s$, is represented by a position p_i. According to its position, each object is depicted on the horizontal number line. If too many objects are mapped to similar positions, clutter occurs, which is marked in dark red.

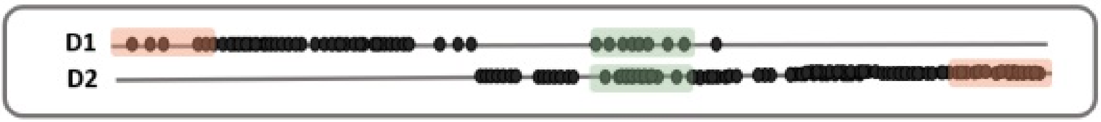


Figure 5.3: Closer examination of dataset *D1* and dataset *D2*: The light red highlighting shows the objects of the two datasets that are most dissimilar to each other. The centrally positioned elements of *D1* and *D2* that lie vertically above each other, highlighted with light green, are very similar.

To avoid clutter, the number lines are organized into regions and the number of elements inside a region is color coded. Through dividing the number lines into adjacent bins, a summarization visualization is created that avoids visual clutter and allows users to detect similar objects and datasets by using color as pre-attentive feature. This results in the overview visualization shown in Figure 5.4, called Histogram-Table. After binning has been performed, the visualization is still the same: Regions that are close to each other are similar, while regions that are far apart are dissimilar.

A sequential scheme from yellow to red was chosen as the color palette. It was selected based on experiments with different schemes. The chosen scheme was best suited to identify different adjacent regions based on visual perception. Using this color scheme facilitates the visual search for similar patterns across the different datasets. The visualization of the same patterns across different datasets, reveals the same underlying quantity of objects. Datasets with objects that have similar properties and are similar in number can be found quickly and without mental effort. In Figure 5.4 we can see at a first glance that the datasets *Specimen1* to *Specimen3* have similar distributions of objects, since most of the more dark yellow to redder regions are located on the left side. In contrast, the datasets *Specimen5* to *Specimen8* have the distributions of their objects on the right side. From these two observations we can immediately conclude that *Specimen1* to *Specimen3* are similar and differ strongly from *Specimen5* to *Specimen8*.

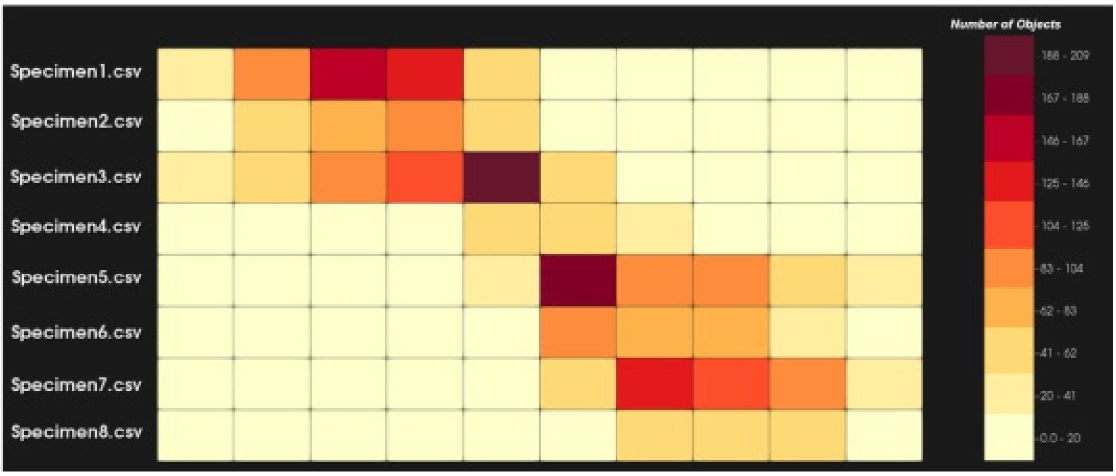


Figure 5.4: Overview Visualization - Histogram-Table: This visualization gives an overview over eight datasets. The individual regions are color coded according to the number of objects inside. The redder the color, the more objects are located inside a region, the brighter yellow the region is colored, the less to no elements are contained.

5.2.3 Linear Zoom

In the overview visualization, the horizontal number line is very roughly partitioned into ten individual areas. Even though this subdivision provides a compact overview of the object distribution, important details are lost. A finer subdivision would produce a more precise representation of the distribution. This would make it possible to identify smaller collections of objects that were previously aggregated by the rough classification. For this reason, three different zoom levels were integrated to obtain a more detailed representation. Since the comparison between the different zoom levels should be possible, the color palette was kept for each level. Figures 5.5, 5.6, and 5.7 show the individual zoom levels. The higher the zoom level, the more precisely the exact positions, where the objects are located, are displayed. By scrolling in or out with the mouse wheel, the different zoom levels are activated or deactivated.

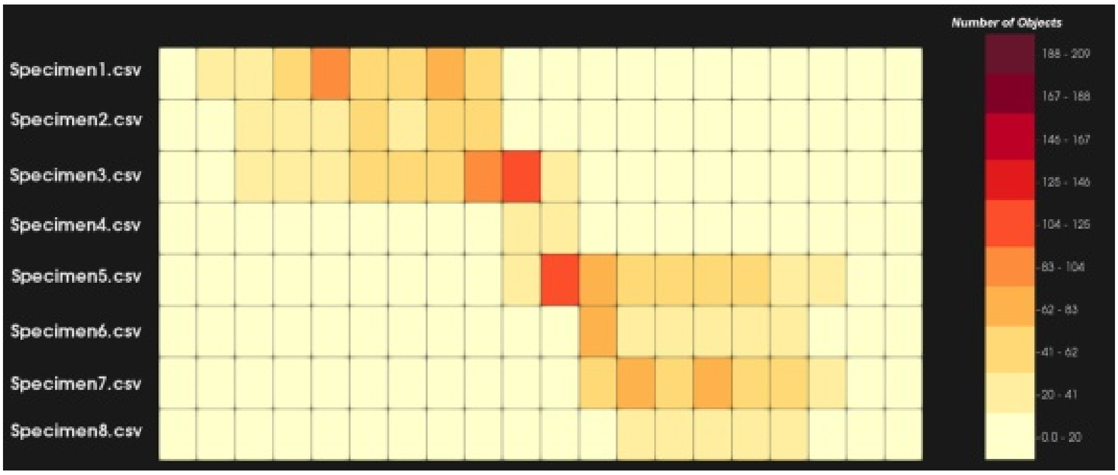


Figure 5.5: Zoom level 1: 20 regions per horizontal line

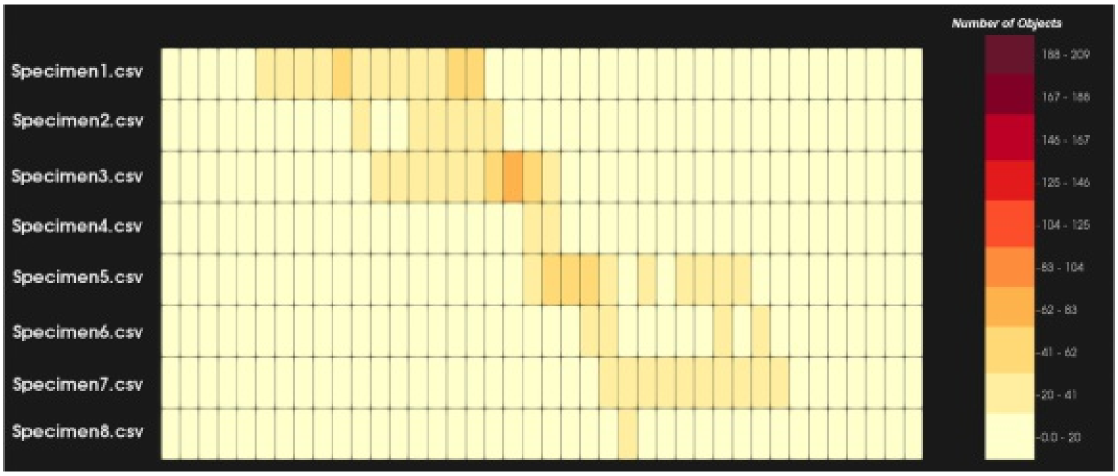


Figure 5.6: Zoom level 2: 40 regions per horizontal line

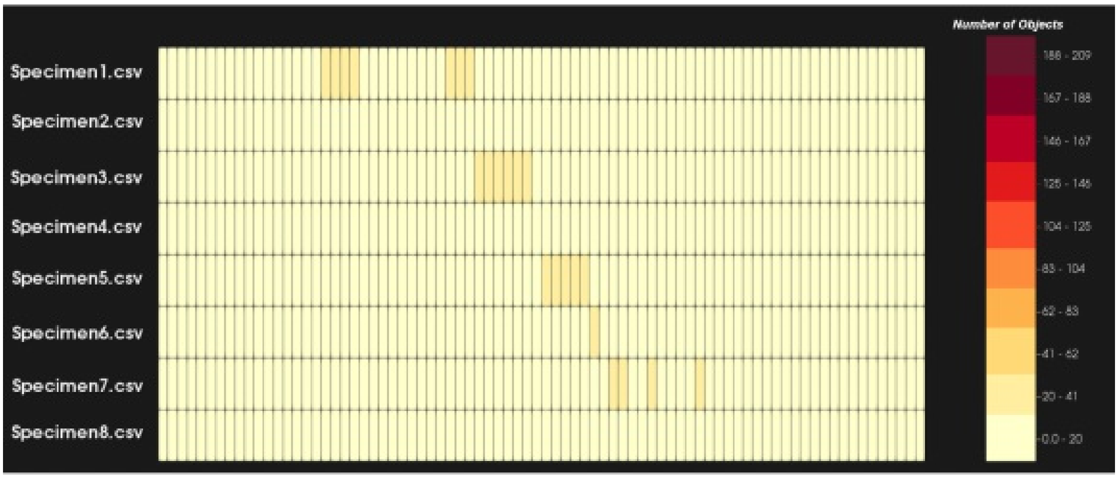


Figure 5.7: Zoom level 3: 80 regions per horizontal line

5.2.4 Non-Linear Zoom

In linear zoom mode all datasets are enlarged or reduced simultaneously. To be able to inspect interesting areas individually, a non-linear zoom mode was developed. Figure 5.8 shows the non-linear zooming applied on various regions. The user can select any region at any zoom level and view the individual objects located in that region in a separate area. It is possible to select single regions or several regions at once in the same dataset, as well as for several datasets. Thus, individual regions can be compared with each other, but also whole datasets can be compared in detail. Inside the newly created zooming areas, the selected objects are positioned in relation to the two selected objects with the smallest and largest position, i.e., similarity value. On the left border of the zooming area, the object with the smallest position is drawn, while on the right border the object with the highest position is visualized. Inside the zooming area the distribution of the selected objects is displayed according to the selected minimum and maximum position. Since we have to deal with thousands of objects, the subdivision into individual regions is also applied to the zooming area.

As can be seen in Figure 5.8, in the first dataset *Specimen3* three adjacent regions were selected, visualized with a green outline. The zooming area is visualized as a continuous line, where the selected objects are distributed. We can see that the distribution is quite homogeneous in the first two selected region, while in the rightmost area a cluster with a higher number of objects can be detected. In dataset *Specimen1* the same regions were selected, so the selected regions of the datasets can now be compared with each other. Although the objects from these regions are similar, a detailed analysis can reveal that they are not exactly the same. For example, the objects in dataset *Specimen3* are located further to the right, while the objects in dataset *Specimen1* form a cluster to the left. Finally, the same regions were selected again in dataset *Specimen6*. Here the zooming area is empty, since there are no objects in the selected regions.

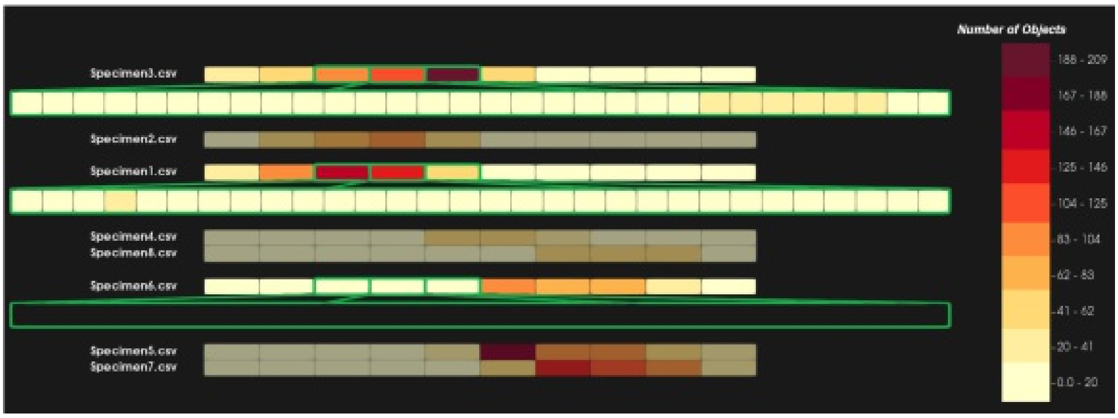


Figure 5.8: Non-linear zoom mode: Various regions of three different datasets were selected for a detailed comparison. For each dataset a new zooming area is visualized containing only the objects of selected regions.

As in the linear zoom mode, it is possible to adjust the region size with the mouse wheel. In the non-linear zoom mode, we provide four different zoom levels. In the first three zoom levels the size of the bins is changed, while in the fourth level, called the point level, the underlying point visualization is shown. Since binning causes blockiness and positional inaccuracy, these introduced errors should be compensated by the point visualization [ADG11]. Figure 5.9 shows the zoom mode on point level. We can see that the selected region in red from dataset *Specimen3* contains most of the objects, but they are distributed within the region in a uniform way. The yellow region from dataset *Specimen1*, that is located vertically below the selected, red region, on the other hand, contains much fewer objects, which are also much less overlapping. To facilitate a direct comparison of datasets that are located far away from each other, it is also possible for users to manually position the datasets one below the other.

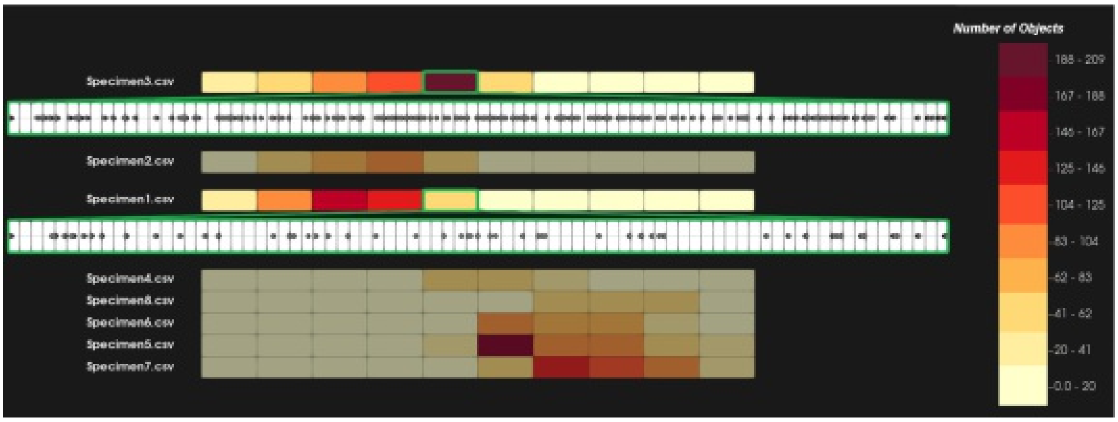


Figure 5.9: Non-linear zoom mode showing the point representation. Each object is visualized as a circle.

5.2.5 Similarity Ranking

The main goal of our visualization was to support users in identifying similar and dissimilar datasets without having to look carefully over the whole ensemble. If there are many very similar datasets, it is difficult to visually determine which of the datasets are most similar. For this reason, an automatic calculation of the most similar datasets has been incorporated into the system. Since the Histogram-Table represents histograms, we chose the Chi-squared distance as comparison metric between the individual datasets.

The overview visualization can be interpreted as a histogram, since the overview visualization is a depiction of several histograms, each shown as a colored bar. A histogram is constructed by binning a range of values. In our case, for each dataset the positions of the objects are clustered according to a predefined number of regions or bins. The entire value range is thus divided into a series of intervals. Then the number of values in each bin, called frequency, is counted. Often the bins are equal in size, and are visualized by rectangles, that are built with a height proportional to the number of cases in each respective bin. Also in our work the bins are of the same size, but instead of drawing a rectangle for each bin, we divide the different frequencies into discrete intervals and encode them in color. So we get a space-saving visualization of a histogram.

Since in the Histogram-Table each row represents a histogram, a histogram comparison metric can be applied to determine the similarity of the individual datasets. There are many different statistical methods to calculate the similarity between histograms. A very good overview is given by Cha et al. [Cha08] and Meshgi et al. [MI15]. One of the most popular metrics is the Chi-squared measure χ^2, which is defined as follows:

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 \,/\, (O_i + E_i)\,,$$

where O_i is the observed frequency for bin i and E_i is the expected frequency for bin i. In our work, the expected frequency E_i is represented by the dataset, which is used as reference for the comparison. The other datasets, which will be ordered according to the reference one, are represented by O_i. The Chi-squared measure is an unbound measure, meaning that it does not have a fixed maximum. The smallest value delivered by this measure is zero. The smaller the value of the metric, the more similar the expected and observed histograms are. Chi-squared measure is a kind of weighted Euclidean distance. As stated by Naik et al. [NPJ09] the chi-squared measure does not deliver the most accurate results when comparing a large group of datasets with very different types of histograms, but creates very accurate results when comparing sets of very similar multimodal histograms. Since our analysis tool should help to determine the similarity of materials, which can also have a very similar distribution of attributes, we have chosen the Chi-square metric. The automatic sorting by similarity is especially useful in cases where many very similar datasets are inspected. Figure 5.10 shows the ensemble before and after the ranking was performed, according to a dataset selected by the user. All other datasets are ordered according to their similarity to the chosen reference dataset. So, the closer the other datasets are positioned to the reference one, the more similar they are to it.

The datasets can be ranked not only by all their bins, but also by individual ones or groups. The user can select one or more bins from a dataset, according to which the other datasets should be ordered. Figure 5.10c shows the ordering of all datasets according to one individual bin outlined in green. This makes it possible, for example, to sort the data records more specifically according to a certain type of objects.

5.2.6 Ordering

Up to now, the user can recognize collective patterns and trends in the Histogram-Table, filter out the most similar datasets, and set varying zoom levels through different interactions. Since the individual datasets can consist of a different number of objects, a functionality must be provided to study the datasets also by the number of their contained elements. To make this possible, an additional visualization was implemented, which visualizes the datasets sorted by the total number of their objects. Figure 5.11 shows the three different orderings that are provided to the viewer to explore the varying number of objects: sorted in ascending order, sorted in descending order, and sorted in the order the datasets had when the user had selected and loaded them into the application. The visualizations consist of bars that represent the individual number of objects per dataset. On the right, the total number of objects in each dataset can be read.

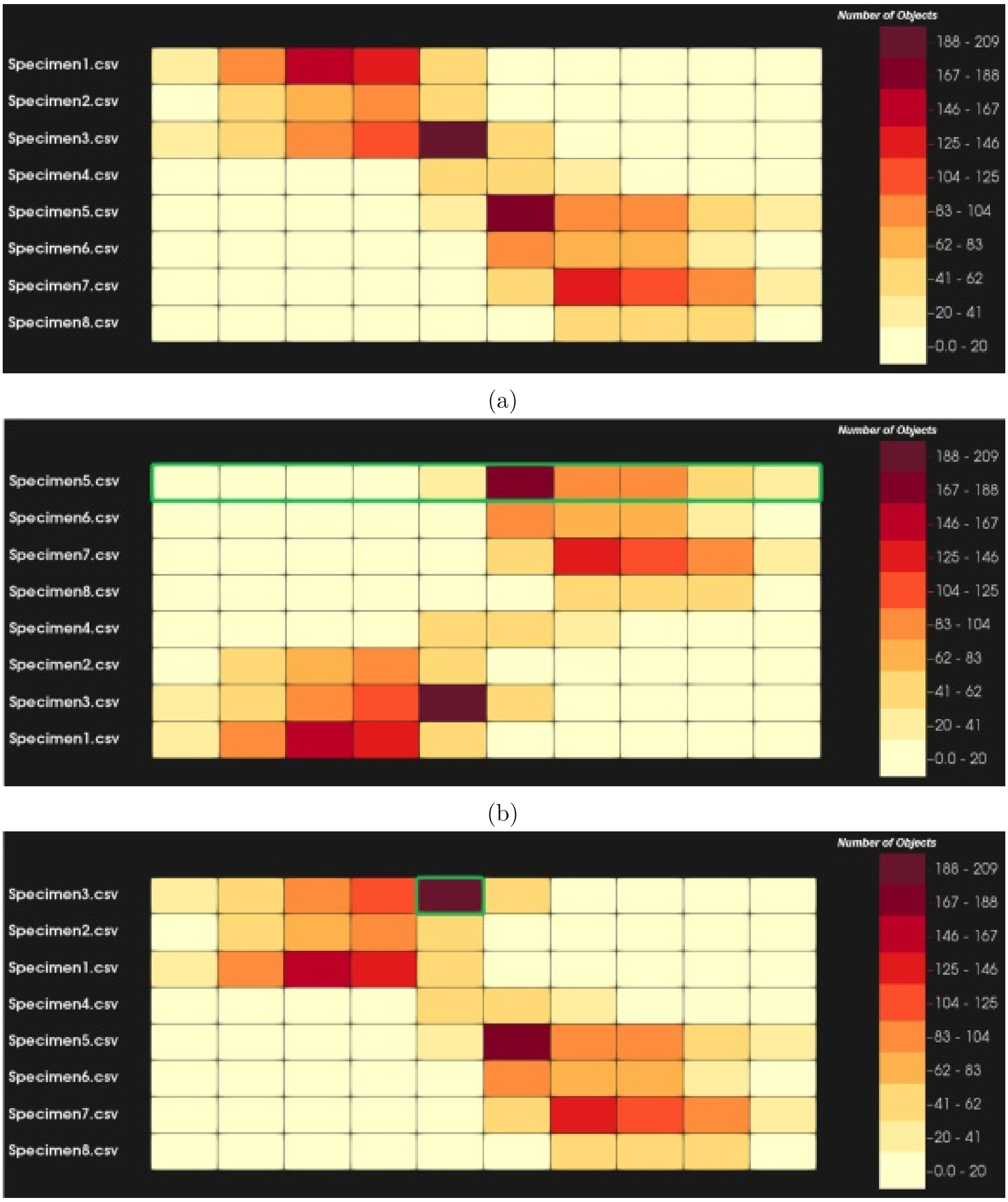
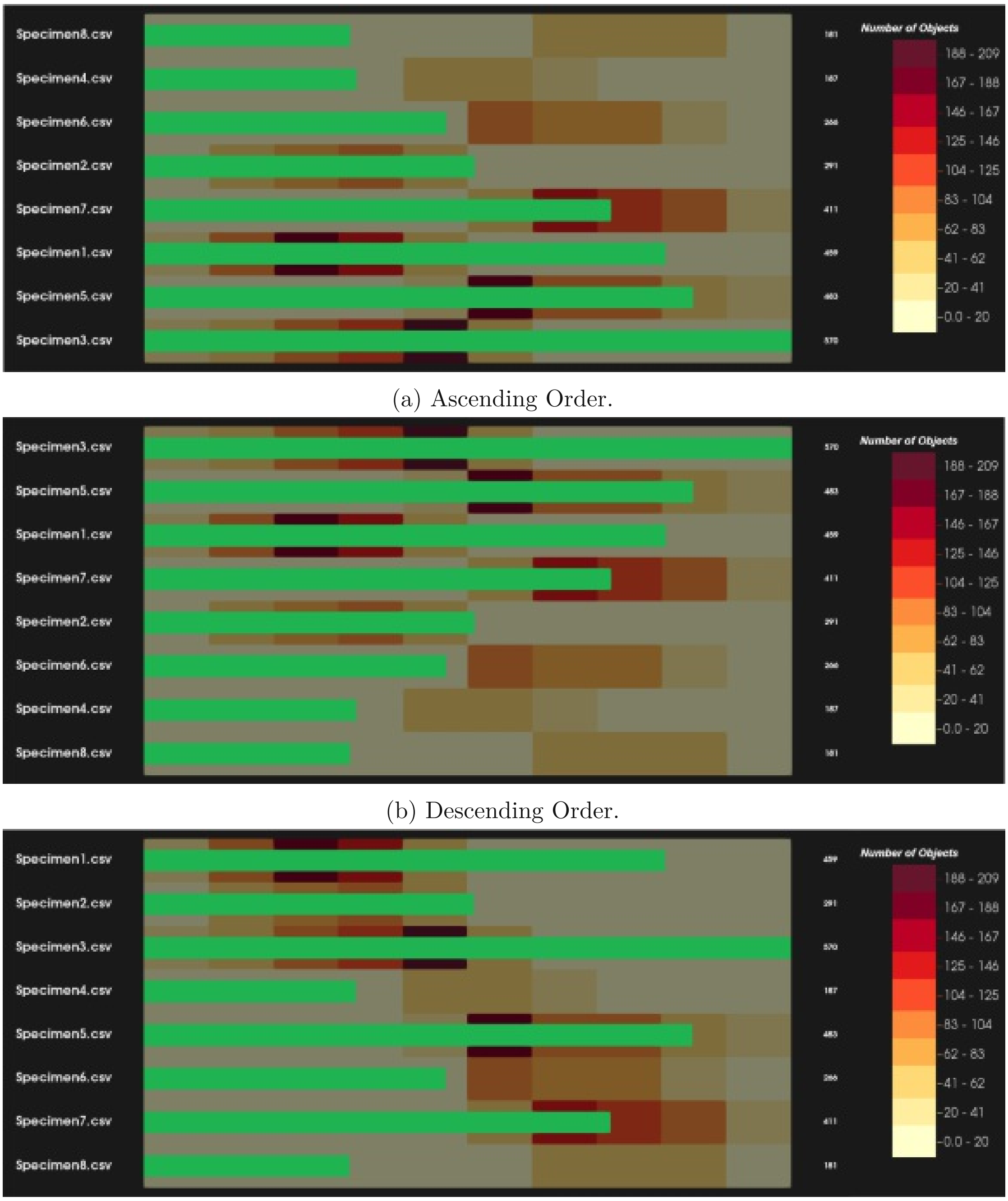


Figure 5.10: The Histogram-Table (a) before and (b) after the automatic ranking was performed. The row outlined in green represents the dataset that was selected as reference by the user. (c) demonstrates the ranking based on one individual bin (green).



(a) Ascending Order.

(b) Descending Order.

(c) Ordering according to the initial order the datasets had when the user loaded them into the application.

Figure 5.11: The different orderings of the datasets, based on the number of their contained elements.

5.3 Detail Visualizations

Since the Histogram-Table visualization and the aggregation required for it, abstracts the data very strongly, essential details can no longer be determined. For example, material scientists need knowledge about the individual attributes of fibers and pores in order to draw conclusions about the properties of the material. Since this detailed information is lost in the overview visualization, it must be made visible again by additional visualizations. Especially if the user has inspected similar datasets or regions, he or she needs answers to questions like:

- In which attribute or attributes are the objects in the regions or datasets similar?
- To what extent are the elements similar in a particular attribute?
- If the elements are similar in one attribute, can the similarity in other attributes be inferred?

A concrete example would be if the material scientists now want to examine the properties, of a specific region containing a large number of fibers, more closely. Above all, he is interested in why these fibers are all located in the same region, i.e., more specifically, what commonalities they share. If, for instance, the fibers all behave similarly in their length attribute, the follow-up question arises as to whether all fibers are equally long or equally short. Since the comparison of many materials provides a lot of data about their attributes, a further question can occur whether one attribute influences another one. For example, long fibers could always have increased volumes. Based on these considerations, we added three different detailed visual metaphors to the analysis framework:

1. A **bar chart** visualizing the most similar attributes computed based on a statistical measure
2. A **box plot** displaying the interval range in which the objects' attributes lie
3. A **correlation map** showing the relation between the individual attributes.

Since the various visualizations complement each other, the combination of all of them forms a framework for analysis that both offers an overview and enables a more detailed investigation.

5.3.1 Bar Chart

The bar chart visualizes all attributes given for each individual object. The height of the bars represents the individual similarity values computed based on the empirical coefficient of variation. The empirical coefficient of variation is a measure of descriptive statistics. It is a relative measure of variation and can be used to determine a series of measurements. The empirical coefficient of variation c_v is defined as follows:

$$c_\text{v} = \frac{1}{\sqrt{s}} \cdot \frac{\sigma}{\mu}, \tag{5.12}$$

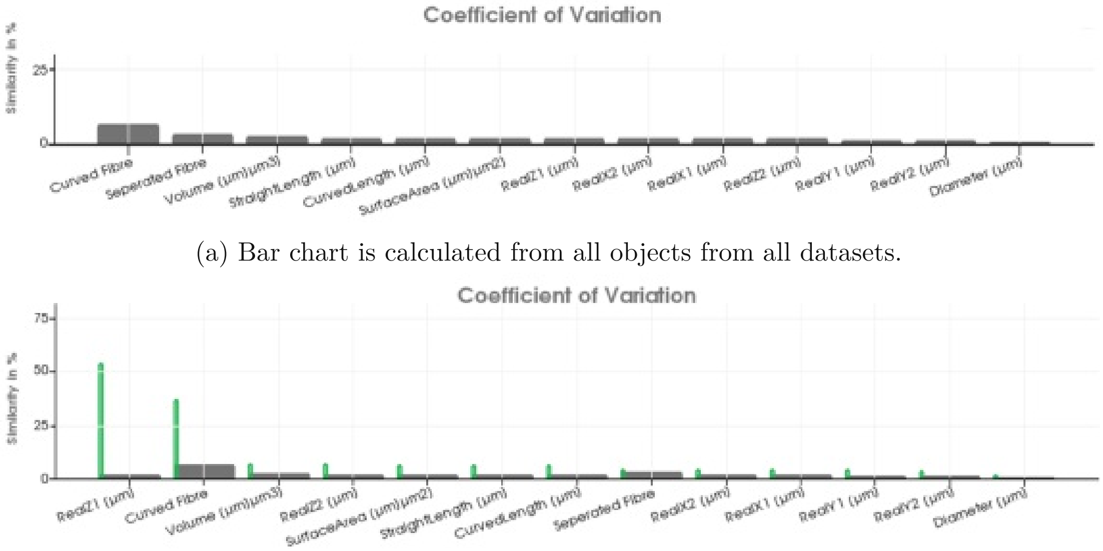
where μ defines the mean value and σ describes the empirical standard deviation [Koh06]. In general, the empirical coefficient of variation shows the degree of variability in relation to the mean value of the measurement and the resulting interval is in the range of $[0, 1]$. The more similar the values are, the closer the result of the coefficient of variation is to 0, the more dissimilar they are, the closer the result is to 1. To ease the readability and eliminate the need for prior knowledge, we mapped the resulting interval of the coefficient of variation to a percentage in the range of $[0, 100]$, where 0 means there is no similarity and 100 implies equality.

Figure 5.12a shows the ranking of the 13 different attributes according to their similarity. All fibers from all datasets were considered for this calculation. As we can see, the attribute with the least dispersion of its values is the attribute *Curved Fibre* with about 5% to 10% similarity. The other attributes follow in descending order. In general, it can be said that all fibers from all datasets are not very similar overall.

If a selection of certain regions was made in the Histogram-Table visualization, the bar chart will also be adjusted based on the selected regions. The coefficient of variation is recalculated based only on the objects contained in the selected bins. The various results are then visualized with green bars that are superimposed on the original grey bars. So, a comparison between the similarity of the selected objects and all the objects can be made. In addition, the attributes are re-sorted depending on which attributes the selected objects are most similar. Furthermore, the width of the green bars is variable, depending on the ratio between the number of selected objects and the number of all objects from all datasets. The fewer objects are selected compared to the original number of elements, the thinner the bar is drawn. In Figure 5.12b we can see that the attributes, in which the selected objects are most similar, is the characteristic *RealZ1* with more than 50%, followed by the attribute *Curved Fibre* with approximately 35%.

5.3.2 Box Plot

Now that the user has been able to determine in the bar chart, in which attributes the fibers are similar, the question arises to what extent they are similar. In order to visualize the dispersion and skewness of the individual attributes, a box plot is drawn for each characteristic. For this reason, the five-number summary is calculated based on all objects from all datasets for each attribute. This statistical summary consists of the minimum, the median, the first and third quartiles, and the maximum. The minimum and the maximum define the lowest and largest data point excluding any outliers. The median is the middle value of the dataset. The lower quartile is the median of the lower half of the dataset, while the third quartile is specified as the median of the upper half of



(a) Bar chart is calculated from all objects from all datasets.

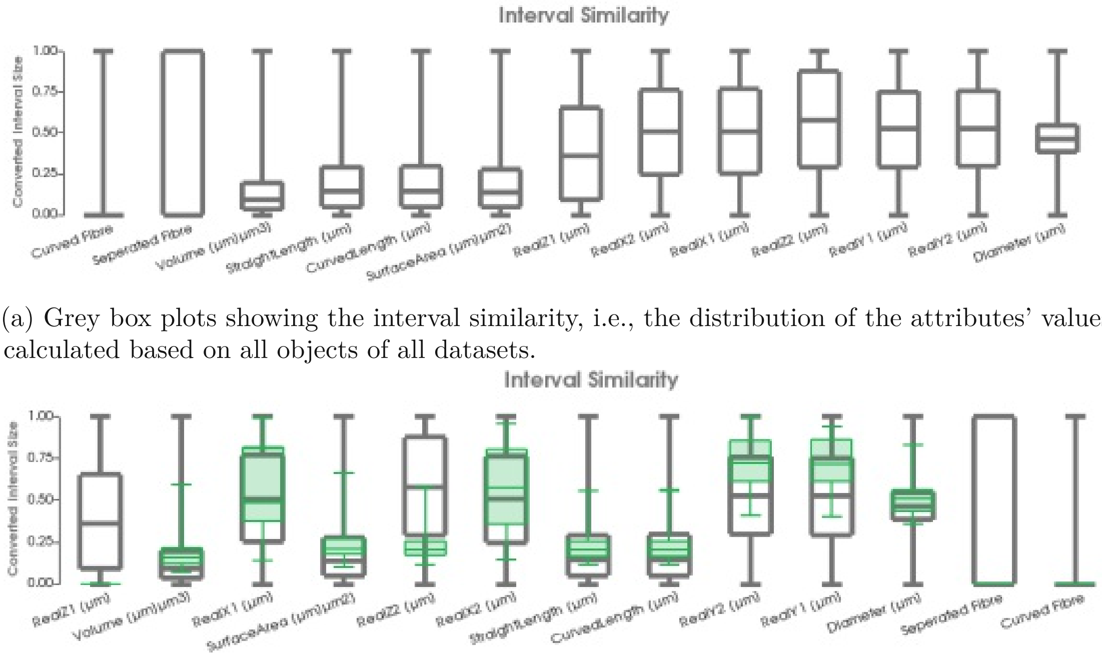
(b) Bar chart recalculated according to the selected objects. The green bars represent the similarity of the selected elements, while the grey bars show the similarity based on all objects from all datasets. The width of the green bars is based on the ratio of the amount of selected items and the overall number of objects.

Figure 5.12: Bar chart shows the similarity for each available attribute computed by the coefficient of variation. The bars are arranged in descending order from most similar to most dissimilar attribute.

the data. Each box plot consists of a box and two lines, called whiskers, which extend the rectangle. The bottom line represents the minimum and the top line represents the maximum. The box is drawn from the 25%-quartile to the 75%-quartile. The horizontal line in the box represents the value of the median. Since the different attributes can have very diverse values, the minimum is mapped to the value 0 and the maximum to the value 1. This ensures that all box plots can be displayed side by side in a diagram. The ordering of the box plots is bound to the similarity ranking of the bar chart. Therefore the attributes are arranged in the same order as the attributes in the bar chart.

Figure 5.13a shows the box plots for the individual attributes. If we look at attribute *Volume*, we can see that the first and third quartiles are very close together. Since this means that the variance of the values is rather small, we can conclude that this attribute has very similar values. In contrast, if we look at attribute *RealZ1*, we see that the box, i.e., the interquartile range, is very large, which means that there is a large variation, so it is not very similar in its values. If we go back to attribute *Volume*, we can also observe that the box is displayed almost at the bottom on the vertical axis. From this we can deduce that most of the values are rather small compared to the maximum value this attribute can assume. A similar statement can be made for the attribute *Diameter*. Here the box is very small and located in the middle. This reveals that the majority of the objects has an average value with a few smaller and larger outliers.

Like the bar chart, the box plot is also recalculated based on the selected regions. For each attribute additional green box plots are drawn, for which the calculation is based only on the selected objects. To perform a comparison with the grey reference box plots, which are computed using all objects from all datasets, the green box plots are superimposed on the grey ones. Furthermore, the ordering of the box plots is changed using the calculated similarity values based on the coefficient of variation as depicted in the bar chart. Figure 5.13b shows box plots for some selected elements. As we can see, the selected objects are most similar in the attribute *RealZ1*. Here the box is particularly small and also positioned very low. Therefore, we can conclude that the values of the selected elements are very small, compared to the numerical values of all objects in all datasets. If we look at the second attribute *Volume*, we can observe that the selected objects have an increased volume compared to all elements, because the lower quartile of the selected elements is above the median of the original objects. Furthermore, we can conclude that the currently selected items do not contain the objects with the highest or lowest numerical value. This can be determined by the fact that the whiskers of the green box do not reach the whiskers of the grey box.



(a) Grey box plots showing the interval similarity, i.e., the distribution of the attributes' values, calculated based on all objects of all datasets.

(b) Interval similarity calculated based on the selected objects superimposed in green over the grey reference boxplots.

Figure 5.13: Box plot visualizing the distribution of each individual attribute.

5.3.3 Correlation Map

On the basis of the two detail visualization techniques it can now be determined in which attributes the objects resemble each other and to what extent they are similar. Now it can be clarified whether one attribute influences another one, or in other words: Whether it can be shown that if in the ensemble many objects are similar in one attribute, i.e., have similar values, that they are also similar in another attribute. To detect relations between attributes, correlation analysis can be used. This analysis determines whether a relationship between pairs of variables exists and describes how strong it is. Correlation calculates a statistical relationship, where often the linear correlation of two variables is of interest. Also in our work we are interested in the linear correlation between individual attribute pairs. A standard method in correlation analysis is the correlation coefficient, also called Pearson product-moment correlation coefficient. For a given attribute pair (a, b), $\forall a, b = 1, ..., m$ and $a \neq b$, each consisting of s values given by the s objects, the pairwise correlation coefficient r_{ab} between the two attributes a and b is defined as follows:

$$r_{ab} = \frac{\sum_{k=1}^{s} (a_k - \bar{a}) \left(b_k - \bar{b}\right)}{\left(\sum_{k=1}^{s} (a_k - \bar{a})^2\right)^{\frac{1}{2}} \left(\sum_{k=1}^{s} \left(b_k - \bar{b}\right)^2\right)^{\frac{1}{2}}} \tag{5.13}$$

The correlation coefficient can assume values in the interval $[-1, 1]$. The value 1 means that the two variables have a perfect positive relation. If one of the two variables increases, the value of the other variable will also increase. The result -1 shows a perfect negative correlation, which means that the more one variable decreases in its value, the greater the value of the other variable becomes. If the correlation coefficient is zero, no linear relationship between the two variables can be calculated. To interpret the results of the correlation coefficient correctly, the following properties of this measure have to be considered:

- **Existence:** Even if the correlation coefficient assumes the value 0 for a pair of attributes (a, b), this does not mean that there is no relation at all, but only that in this case no linear relation can be determined. If we assume, for example, that the attribute a has a standard normal distribution and $b = a^2$, thus depends completely on a. This would result in a complete dependence between both variables, but their correlation would be 0 according to the correlation coefficient.

- **Subpopulation Size:** The correlation coefficient can be significantly influenced by the number of the chosen subset of the objects. A limitation of the variability of the data occurs, if a attribute has similar values. This happens, for example, if the subpopulation does not contain all or only a very limited range of possible values of an attribute. Thus the subpopulation does not sufficiently represent the whole interval range of the values of an attribute. An unbalanced selection of the subpopulation can lead to a wrong correlation being calculated. Furthermore, correlations for very small subpopulations are not very meaningful, since high correlations are easier to obtain in smaller subpopulations than in larger ones.

- **Robustness:** The correlation coefficient is not robust against outliers. This means that outliers can both artificially increase and decrease the correlation coefficient.

To allow a correct interpretation of the correlation coefficient, the above mentioned properties should be kept in mind. Furthermore, for easier interpretation, the rules of thumb of Cohen [Coh13] can be taken into account, which state that if $|r_{ab}| = 0.1$ the effect is small, at $|r_{ab}| = 0.3$ a medium correlation can be found and with $|r_{ab}| \geq 0.5$ a high correlation can be observed.

Dealing with datasets consisting of a large number of attributes, it becomes difficult to interpret the many resulting correlation relationships, since the information to be examined grows very quickly. Given m attributes, the correlation computation will compute $O\left(m^2\right)$ correlation pairs. Therefore, an effective visual interface is required to enable analysts to get a quick overview. Traditionally, visualization methods such as scatterplots, parallel coordinate plots, or correlation matrices are used to display the correlation of many attributes. Scatterplots do not visualize the correlation coefficients themselves, but the underlying data points. This visualization shows the relationship between two attributes measured for the same objects. The values of one attribute are drawn on the horizontal axis, while the values of the other attribute are displayed on the vertical axis. Each individual object is thus represented as a point, whose x- and y-position are determined by the values of its two attributes. The overall pattern formed by the points describes the correlation of the attributes. In the parallel coordinates plot, each attribute corresponds to an axis, and the m axes are organizes as uniformly spaced vertical lines. An object, consisting of m attributes, manifests itself as a set of points, one on each axis, connected by a polyline. The pattern of the polylines determines the type of correlation. Both types of visualizations often suffer from clutter, because either points or different lines overlap. An alternative approach is the use of a correlation matrix visualized as a heatmap. In this depiction the correlation coefficients of each attribute pair are encoded using a color scale [SS04]. Although the perception of the correlations is improved by encoding the correlation coefficients with color and brightness, the authors Bertin et al. [Ber83] found that position and size are preferable for interpreting quantitative information.

For these reasons we followed the approach of Zhang et al. [ZMM12, ZMZM15] and decided to display the correlation information as a graph, called correlation map. The correlation map is a complete graph G consisting of a set of vertices V and a set of edges E. The set $V = \{v_1, v_2, ..., v_m\}$ contains a node for each of the m attributes. Since G is a complete graph, the set of edges is defined as $E = \{\{v_a, v_b\} : 1 \leq a < b \leq m\}$, where every edge $\{v_a, v_b\}$ describes the correlation coefficient r_{ab} between the attribute pair (a, b).

The vertices in the graph are positioned according to a force-directed layout. Attributes, which have a strong correlation, are positioned closer to each other, while attributes, which have no linear correlation, are positioned far away from each other. For computing the layout, the algorithm of Fruchterman and Reingold [Kob13] was applied, which treats vertices in the graph as atomic particles exerting attractive and repulsive forces on each other. In our approach the attractive forces were weighted by the result of the correlation coefficients to fulfill the requirement to position the vertices, i.e., attributes, with stronger correlations closer to each other. The following formulas were implemented for the attractive force $f_a(D)$ and the repulsive force $f_r(D)$:

$$f_a(D) = \left(D^2/D_{opt}\right) \cdot |r_{ab}|, \tag{5.14}$$

$$f_r(D) = -D_{opt}^2/D, \tag{5.15}$$

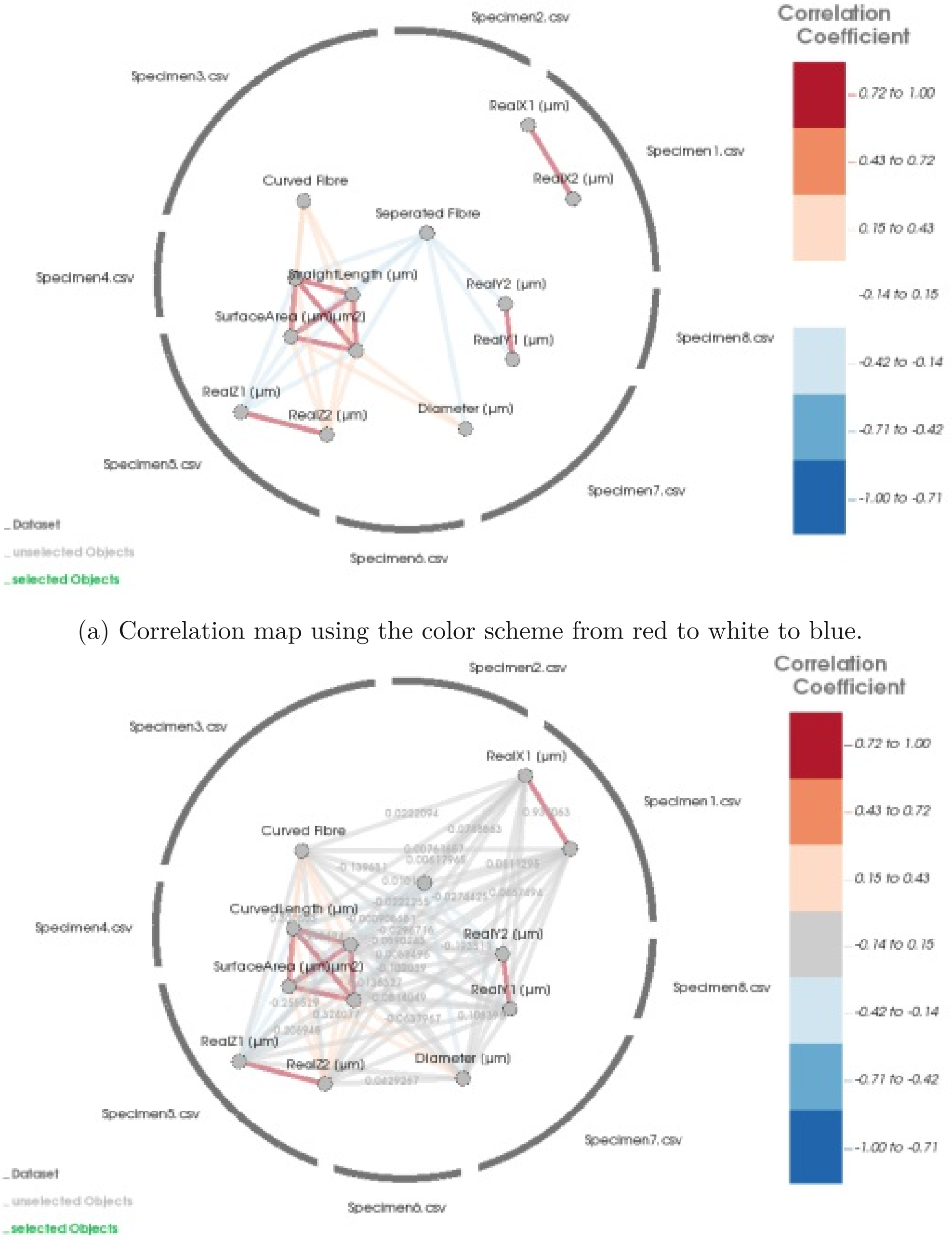
where D describes the distance between two vertices and the optimal distance D_{opt} between vertices is defined as

$$D_{opt} = Z\sqrt{\frac{area}{number\ of\ vertices}} \tag{5.16}$$

where Z is a constant, *area* is the available width and length of the drawing area, and *number of vertices* is the total number of vertices. In our approach the vertices' positions are initialized randomly, then the layout-algorithm is applied. The edges are color coded according to the type of correlation. For coloring, two different diverging color schemes where chosen with seven discrete color steps. The first color scheme runs from red, i.e., positive correlation, to white, i.e., no correlation, to blue, i.e., negative correlation. The second one contains the same colors except that the color white has been replaced by grey, i.e., no correlation. Figure 5.14a shows a correlation map, where the correlation coefficients were calculated between all attribute pairs. As we can see, when the first color scheme is applied, the correlation results with values close to 0 are invisible. This makes it possible to get a good overview of the strong correlations between the attributes without severe clutter. Furthermore, through the positioning of the nodes we can see which attributes do not depend on other characteristics. For example, the attributes *RealX1* and *RealX2* are both very far away from the other attributes, indicating that they do not have a linear relationship with the other characteristics. Figure 5.14b shows the same result using the second color scheme. Here we can see that correlations were calculated between all attribute pairs, but for many of them the result was close to zero. Since this color mapping produces a cluttered view because of the many overlapping edges, that do not represent important correlation information, we provided the color scheme used in Figure 5.14a. However, to allow the user to inspect the edges that have no correlation more closely, the user can switch between these two color schemes.

The grey circular segments, positioned around the correlation map, represent the individual datasets. The length of the circular segments indicates the ratio of the number of objects between the different datasets. As we can see in Figure 5.14, *Specimen4* and *Specimen8* contain the fewest objects, since they are represented by the shortest segments. The representation of the ratio between the number of objects, contained in the different datasets, should show which attributes of which objects were used to calculate the correlation map. For example, if one dataset has significantly more objects than another one, we could conclude that the smaller one is underrepresented for the correlation calculation. These segments help the user to select a suitable subpopulation of the datasets. The distance of the vertices to the circular segments does not contain any information.

Like the bar chart and the box plot, the correlation map is also recalculated after selecting certain regions in the Histogram-Table. Figure 5.15 shows the re-calculated correlation map after various bins of two different datasets were selected by the user. The correlation calculation is now only performed on the basis of the objects, which are contained in the selected bins. The circular segments were updated to indicate that only two datasets were included in the calculation. The dark-grey, external segments show the ratio between the two selected datasets, based on all their objects, while the green inner segments represent the number of selected objects per dataset. Only the elements represented by these green segments are used to calculate the correlation coefficients. The light-grey segments show the ratio of non-selected objects inside each dataset with respect to the number of selected objects, visualized by the green segments.



(a) Correlation map using the color scheme from red to white to blue.

(b) Correlation map using the color scheme from red to grey to blue.

Figure 5.14: Correlation map visualizing the linear correlation between the individual attributes. The positioning of the vertices represents the strength of the correlations, while the edges encode the type of correlation with color. The length of the circular segments visualize the ratio of the elements in each dataset. The longer the segments are, the more objects are stored in the respective dataset.

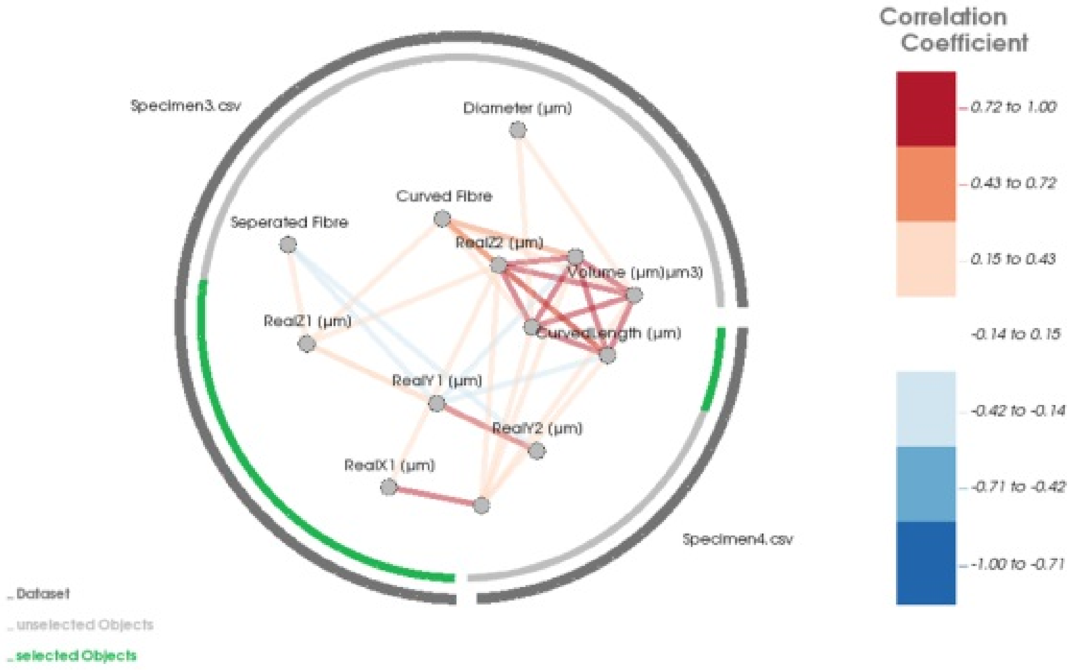


Figure 5.15: Correlation map after some bins of two different datasets were selected. The correlation map is recalculated based only on the chosen objects. The dark-grey, external segments visualize the ratio between the different datasets based on the total number of their objects. The green segments show the ratio of the selected objects with respect to the dark-grey, external segments, showing all objects inside their dataset. The light-grey lines represent the ratio of non-selected objects with respect to the selected objects inside each dataset.

5.4 Implementation

Our application is based on the programming language C++. It was developed as module for the application open_iA [FWS+19]. Open_iA is an open source application for the analysis of volume data based on image processing and visualization. Its main focus is on industrial computed tomography. As graphical user interface, the cross-platform framework Qt is used, which provides an easy to use and intuitive surface. Open_iA allows slice-by-slice navigation in the 2D views, 3D navigation and arbitrary slice planes in the 3D views and furthermore the definition of user-specific views for individual visualizations. For visualization and image processing purposes, the toolkits ITK [SNC03] and VTK [SML06] are used in addition to in-house developments, which extend open_iA to a comprehensive tool for 3D visualization and CT data analysis. open_iA is easily extensible and provides various filters and different tools for an accurate visualization of material properties such as fibers and pores.

CHAPTER 6

Discussion and Evaluation

In this chapter we will go into more detail about the practical applicability of our analysis framework. We will demonstrate the usefulness of our analysis framework by means of three different usage scenarios. To test the readability and cognitive difficulty of the applied visual encodings in our visualization framework, we conducted a qualitative user study. For this purpose, we formulated open tasks that the material engineers should execute. In addition to checking the understanding of the visualization techniques, the material experts were also interviewed about the usefulness of our framework in their daily activities. Finally, we will describe the benefits and limitations of our tool on the basis of the evaluation we have conducted.

6.1 Description of Data Used in This Work

The datasets used in our work are various CSV files consisting of hundreds to thousands of rows, representing the individual objects. Each object is described through a varying number of attributes. In the following we will describe three different use cases based on three different ensembles. The first two ensembles come from the material science, while the last ensemble represents a general dataset usually visualized in the information visualization domain.

The first ensemble consists of six individual datasets, which describe two different materials based on the contained pores. One dataset describes a small material scan, while the other five datasets each describe a region of a larger material scan, which was divided into five adjacent regions. The performed cutting is illustrated in Figure 6.1. The aim is to compare the smaller material scan with the larger one and to determine whether the two materials are similar and if so, in which region. The pores are described by 23 different features, including attributes such as different orientation specifications, flatness, roundness, length, elongation, etc. The smaller material scan contains 1722 pores, while each region of the big material scan contains around 1500 pores.

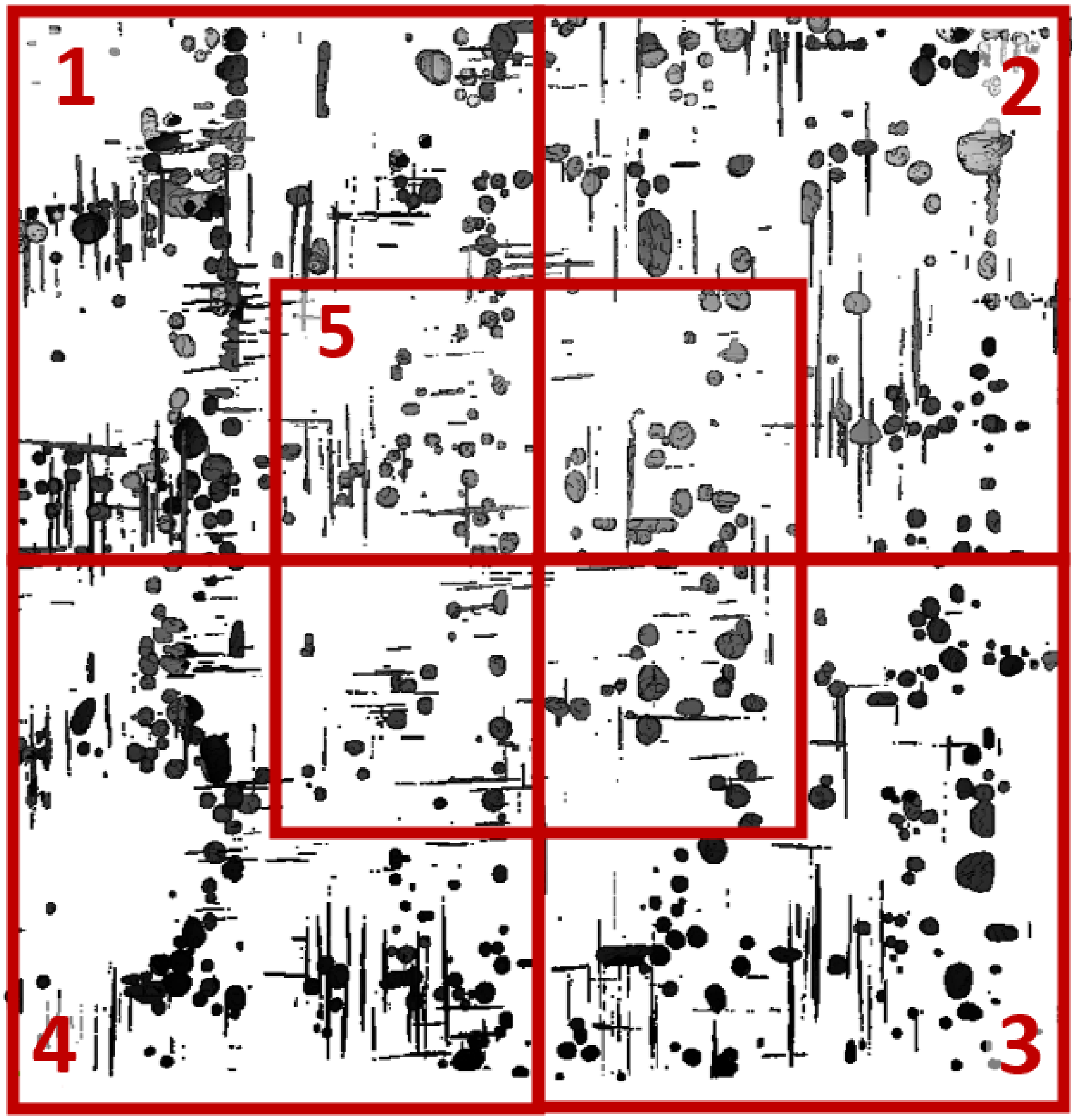


Figure 6.1: Regions, into which the larger sample was divided. All these will be compared with a smaller material scan.

The second ensemble describes a fiber material that has been modified by an in situ test. The first time the material has been scanned after it had been exposed to a shearing force for 10 minutes. Then, it was scanned a second time after a total of 60 minutes of shear force application. In this ensemble, the aim is to compare the two different time steps. To enable the user to make a detailed analysis, the material was divided into the same four areas in each scan, as shown in Figure 6.1. The individual fibers are described by 13 different characteristics like their start and end positions in the different dimensions, length, diameter, area, volume, etc. Each of the individual regions contains about 2.500 fibers, so a total of about 20.000 fibers are compared in this ensemble.

To demonstrate the generalizability of our analysis tool, we also show the applicability to a time-varying spatial dataset, as provided by ILOSTAT [Org20] on the platform Gapminder [Gap20]. This ensemble consists of six different datasets that describe the unemployment of men and women in different age categories. Three datasets describe the unemployment of women aged 15-24, 25-54 and 55-64, while the other three datasets describe the unemployment of men in the same age groups. These CSV files are structured as follows: For about 165 countries of the world, the percentage of unemployment is plotted for the time period of 29 years. This dataset can be interpreted in such a way that the individual objects are represented by the different countries, while the past years reflect the individual attributes. By comparing these datasets, it will be shown that unemployment behaves similarly within the same gender, while a clear difference between the sexes can be identified.

6.2 Usage Scenarios

In the first scenario, two different specimens are compared based on their pore properties. The second scenario describes a fiber material on which a force has been applied for a certain period of time. The third scenario relies on datasets that do not originate from material science. To show the generalizability of our tool, six different spatio-temporal datasets are compared. Based on the different user scenarios we will show how to read the different visualizations and what conclusions can be derived from them.

6.2.1 Analysis of Porous Materials

In this scenario the ensemble consists of two different material scans. The larger scan has been partitioned into five areas, as can be seen in Figure 6.1. In the Histogram-Table in Figure 6.2 we can explore the dataset *SmallPores* and the datasets with regions 1 to 5, describing the various areas of the larger material scan. At first glance, there are no major differences between the individual regions and the *SmallPores* dataset. However, dataset *Region5* is probably the most dissimilar, since it reveals large clusters, i.e., very red areas, more on the right side of the Histogram-Table. Figure 6.3 demonstrates the bar chart and box plot calculated for the ensemble. Inspecting the bar chart in Figure 6.3a, we can determine that most similarity between all pores is mainly in their flatness attribute and in their volume. In all other attributes, the statistical calculation does not show much similarity. In the box plot in Figure 6.3b we can now take a closer look at the more detailed depictions of the interval range of the individual attributes. When inspecting the first few attributes, in general we can state that the pores are very flat, have a very small volume and have a very similar orientation, based on the tensor orientation *a22*. Figure 6.4 presents the correlation map calculated for this ensemble. The visualization reveals that, for example, the attribute flatness does not show a strong correlation with

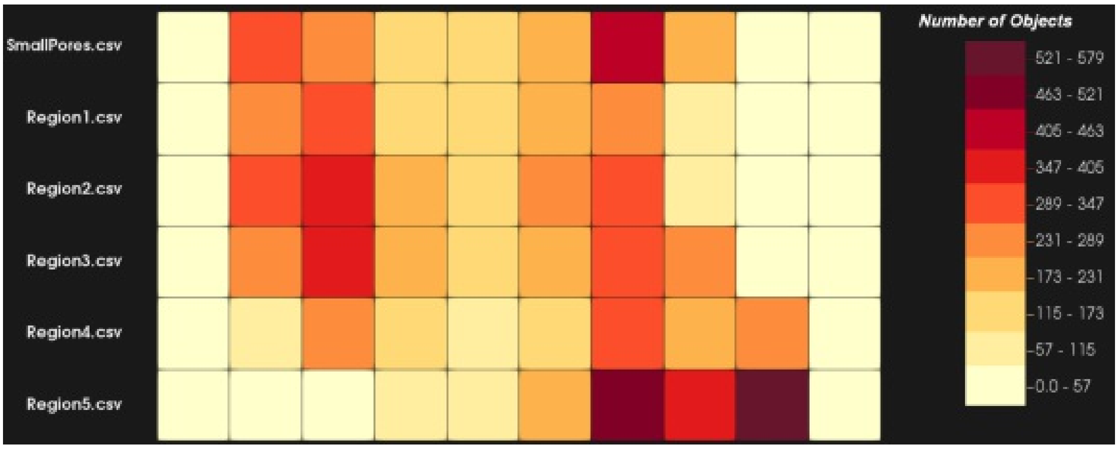


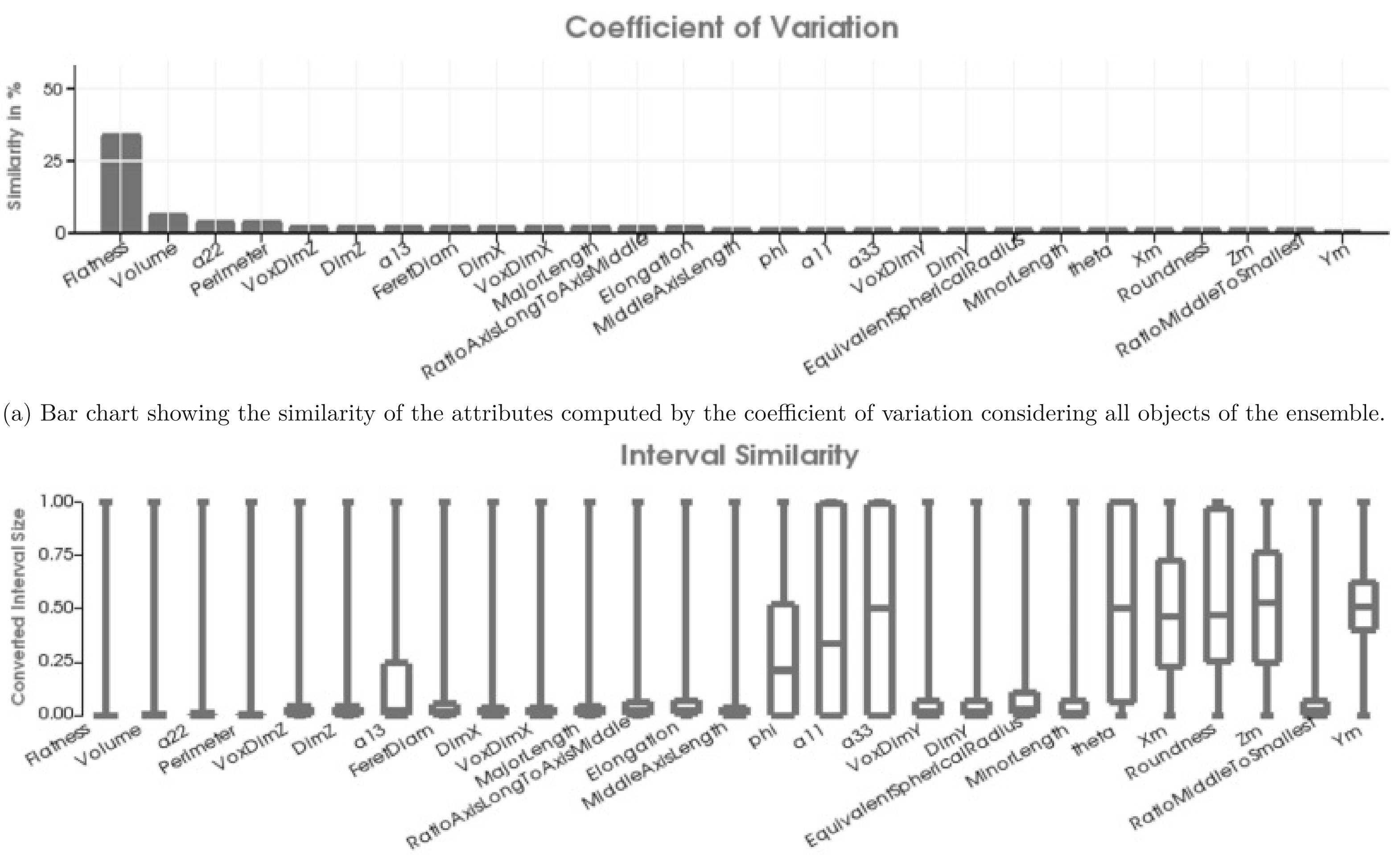
Figure 6.2: Histogram-Table displaying the *SmallPores* dataset and the five regions of the larger dataset.

any other attribute, while a more or less strong correlation was calculated between the majority of the attributes.

Although we know that *SmallPores* is the smaller volume, we wanted to determine the number of segmented pores more precisely. For this reason we sort the datasets in ascending order. Figure 6.5 shows the result of the sorting. We can see that *SmallPores* does not contain the smallest number of pores with its 1722 objects. *Region1* and *Region4* even contain smaller numbers of pores, i.e., 1437 and 1597 respectively. The area with the most pores is *Region5*, with 1917 found elements. From these observations we can conclude that although *SmallPores* covers a smaller volume, it has a higher pore density compared to individual areas of the larger volume.

Our main goal is to find the most similar region to the volume of *SmallPores*. Since the individual regions all have a very similar pore distribution, it is difficult to tell at a glance which one is the most similar. For this reason we perform the automatic classification according to the most similar datasets. Figure 6.6a shows the result of the similarity ranking. The most similar dataset is *Region2*, while the second most similar dataset is *Region1*. If we zoom in, as displayed in Figure 6.6b, the bin size is reduced, and we can explore the distribution of the objects in more detail. At this zoom level, the subtle differences in the distributions of the datasets *SmallPores* and *Region2* become visible.

Now we want to explore why *Region2* and the reference dataset *SmallPores* are similar. This is why we select the bins with the most pores. Figure 6.7 displays the updated Histogram-Table. All bins framed in green have been selected and the distribution of their objects is drawn in more detail via the non-linear zoom. All in all, the distribution of the objects is quite homogeneous in both datasets. The most significant difference between the datasets is visible on the right side of the *SmallPores* dataset. There we can locate many similar objects, while in the bins below, in dataset *Region2*, there is no increased number of objects visible. Figure 6.10 represents the recomputed bar chart



(a) Bar chart showing the similarity of the attributes computed by the coefficient of variation considering all objects of the ensemble.

(b) Box plot displaying the interval similarity, i.e., the distribution of the attributes' values, computed based on all objects of the ensemble.

Figure 6.3: Bar chart and box plot visualized for the analysis of two porous material scans.

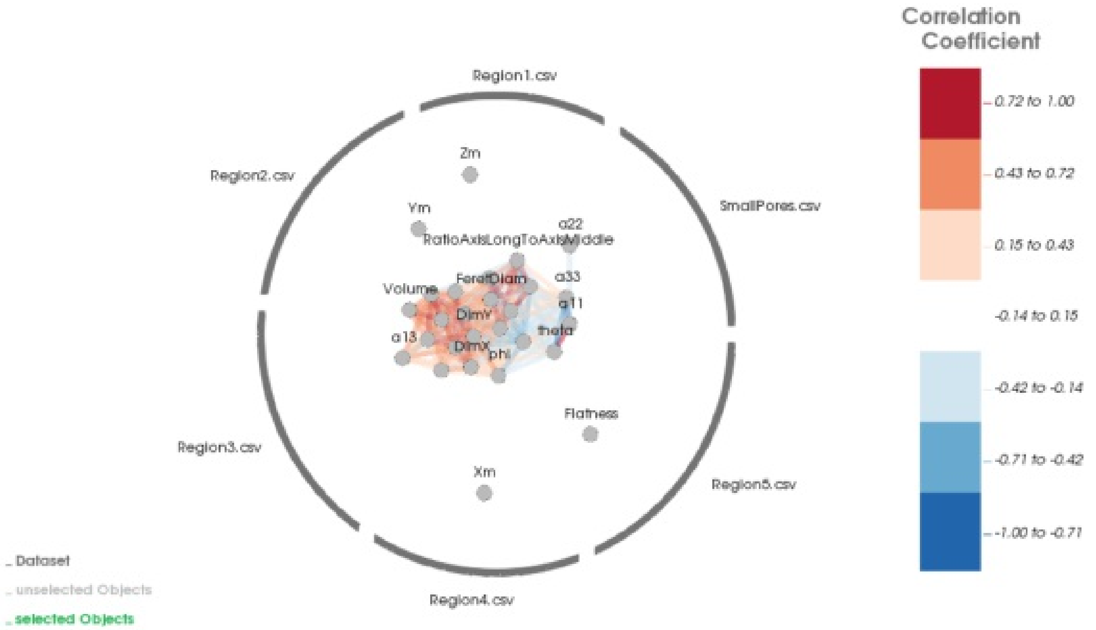


Figure 6.4: Correlation map computed based on all objects of the ensemble.

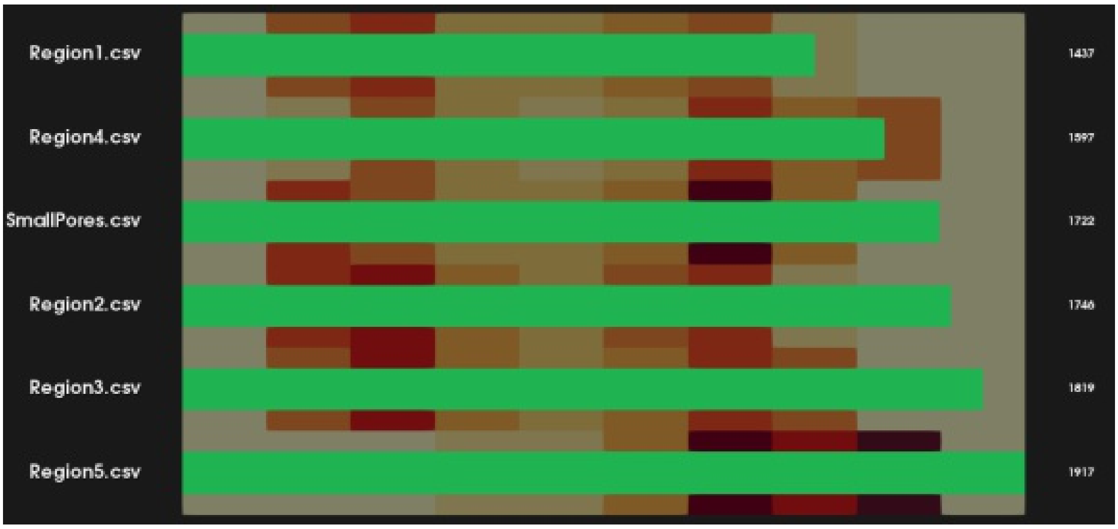
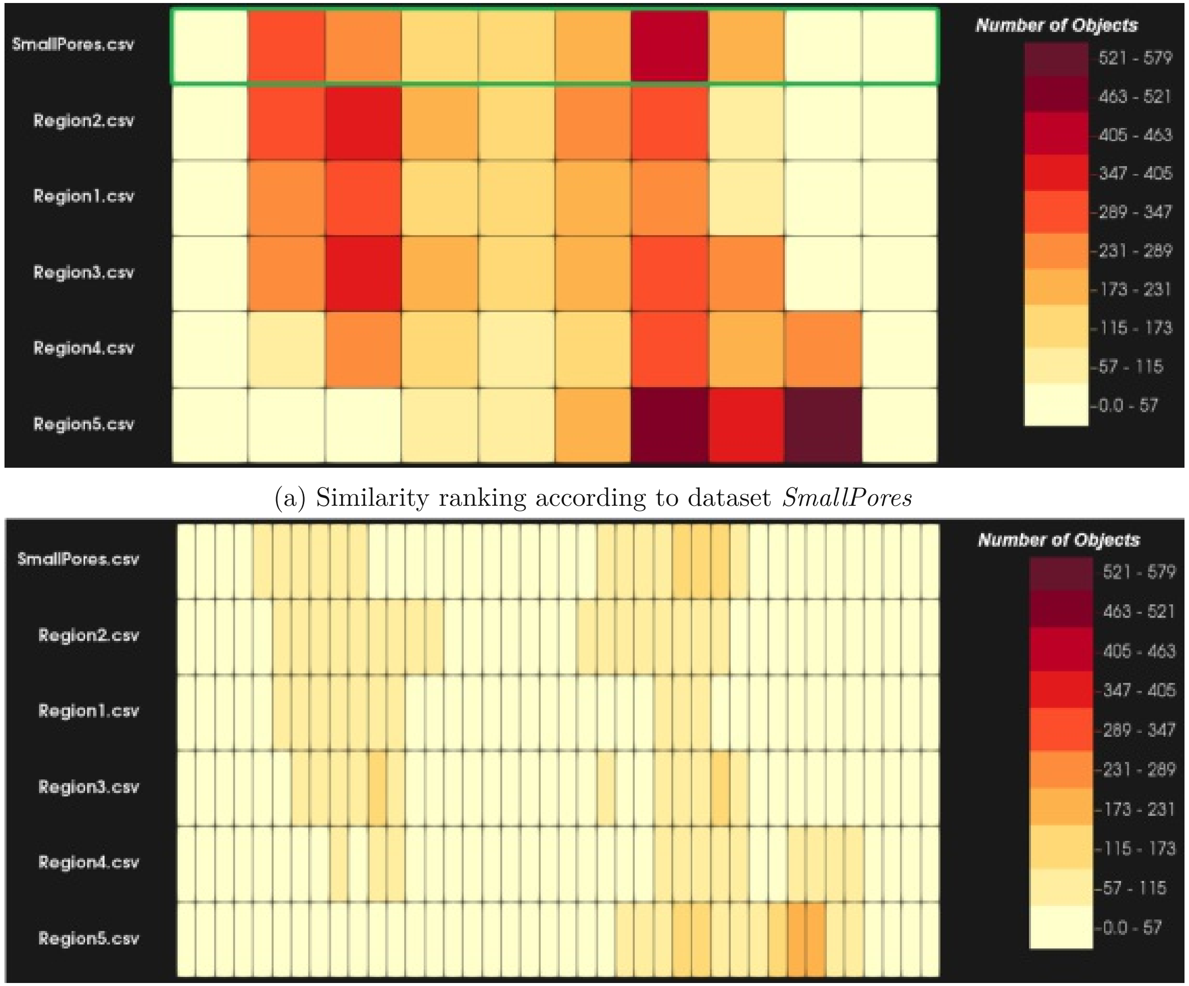


Figure 6.5: Ascending Ordering according to the contained number of pores.

and the newly computed box plot. As we can see in Figure 6.8a, the pores in these two datasets are resembling each other for 50% in the first attribute, in contrast to the overall similarity of about 35%. In addition, the green bars are about half as thick as the grey reference ones, thus we can conclude that the two datasets together contain slightly more than one third of all objects. In the box plot in Figure 6.8b, we can see that very flat pores are present in the two datasets, since the box plot of attribute *Flatness* is very small and does not contain the outliers with the higher values. The selected pores are also very similar in the attribute *Volume*. Further we can observe that the selected pores have a very small volume, since the green box plot is very small and located very far below. In summary, we can conclude that the pores located in the *SmallPores* material scan, as well as in *Region2* of the larger scan, are very flat and small.

As a final step in our exploration analysis, we want to take a closer look at the properties of the pores in *Region5*, which are the most different ones compared to those in *Region2* and in *SmallPores*. To do this, we select the three rightmost bins of *Region5*. In Figure 6.9 we can observe the representation of the individual pores and their distribution. In the bar chart, shown in Figure 6.10a, we notice that the selected pores of *Region5* are also most similar in the attributes *Flatness* and *Volume*. If we take a closer look at the interval range of the pores, selected in *Region5*, we can see that they assume much higher values in both attributes, *Flatness* and *Volume*, compared to the pores selected in the datasets *SmallPores* and *Region2*. We can distinguish the difference in the interval range of the two attributes by looking at Figure 6.8b and Figure 6.10b. In Figure 6.10b, the maximum values of the intervals of the attributes from Figure 6.8b are visualized using two red markings. Thus, we can conclude that *Region5* contains larger and rounder pores.



(a) Similarity ranking according to dataset *SmallPores*

(b) Zoom level 2 applied on the ranking of (a)

Figure 6.6: Exploration of the similarity of all datasets according to one reference dataset

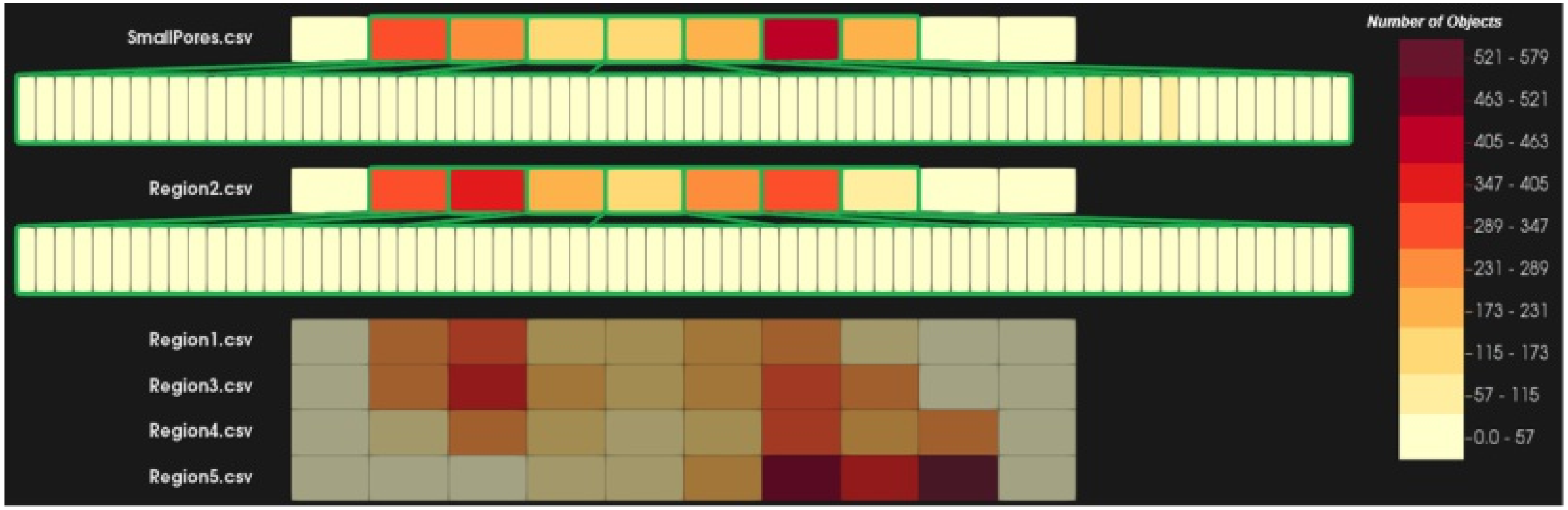
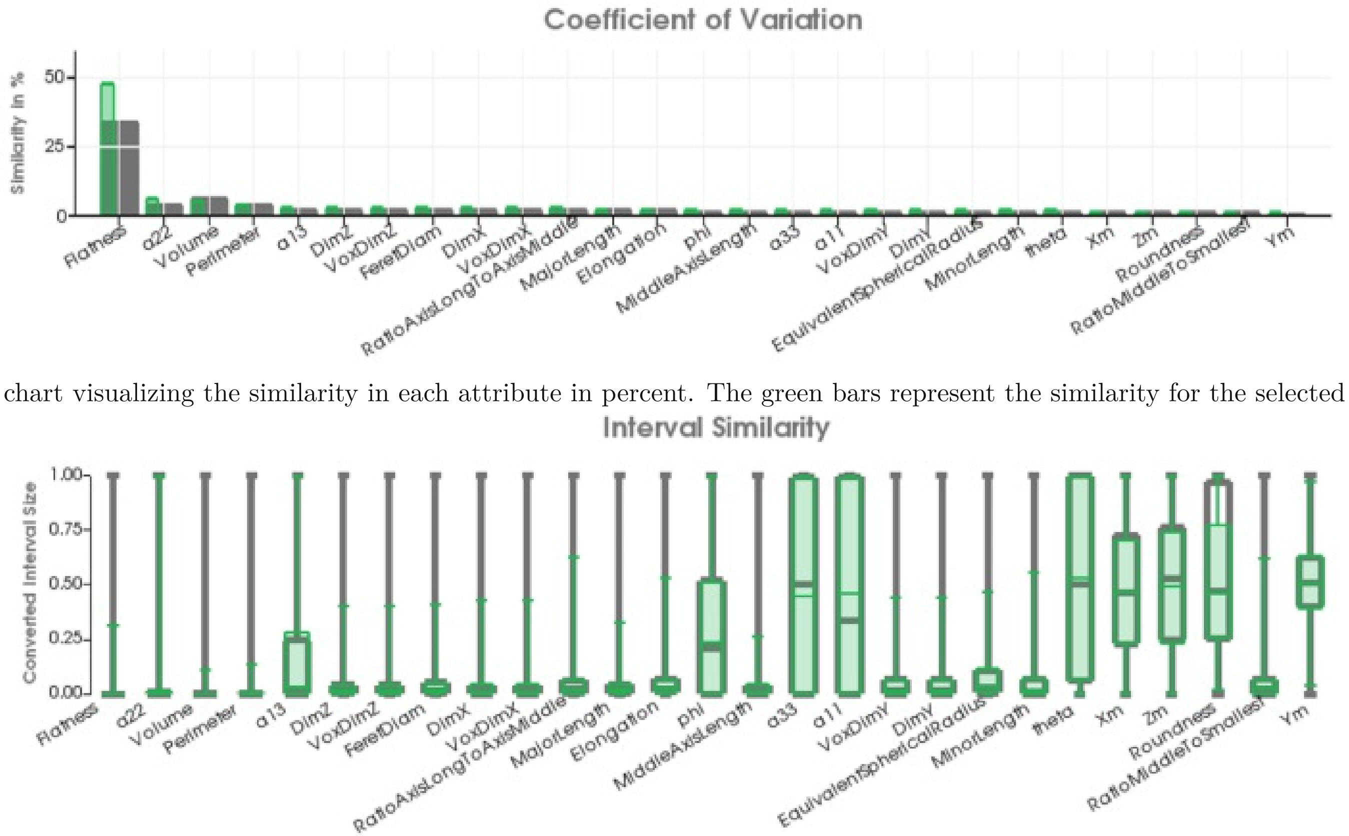


Figure 6.7: Histogram-Table showing the selected bins of two different datasets. All in all, the distribution of the objects is quite homogeneous in both datasets. The most significant difference is visible on the right side of the *SmallPores* dataset. There we can locate many similar objects, while in the bins below in dataset *Region2* there is no increased number of objects visible.



(a) Bar chart visualizing the similarity in each attribute in percent. The green bars represent the similarity for the selected pores.

(b) Box plot visualizing the interval similarity of various attributes. The superimposed green boxplots represent the selected pores.

Figure 6.8: Exploration of the similarity of all datasets according to one reference dataset

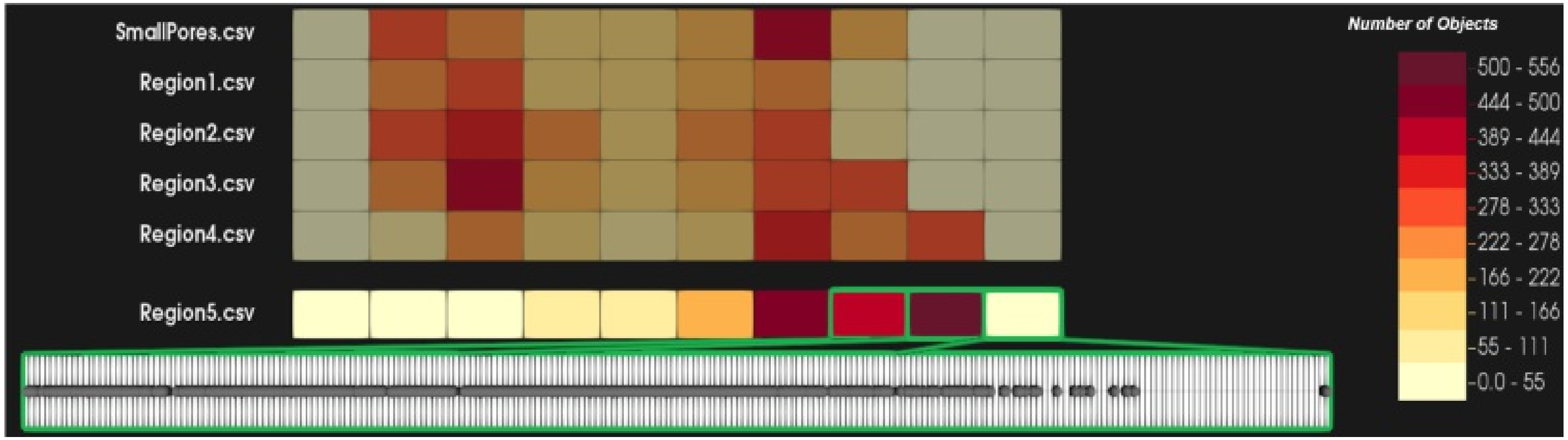
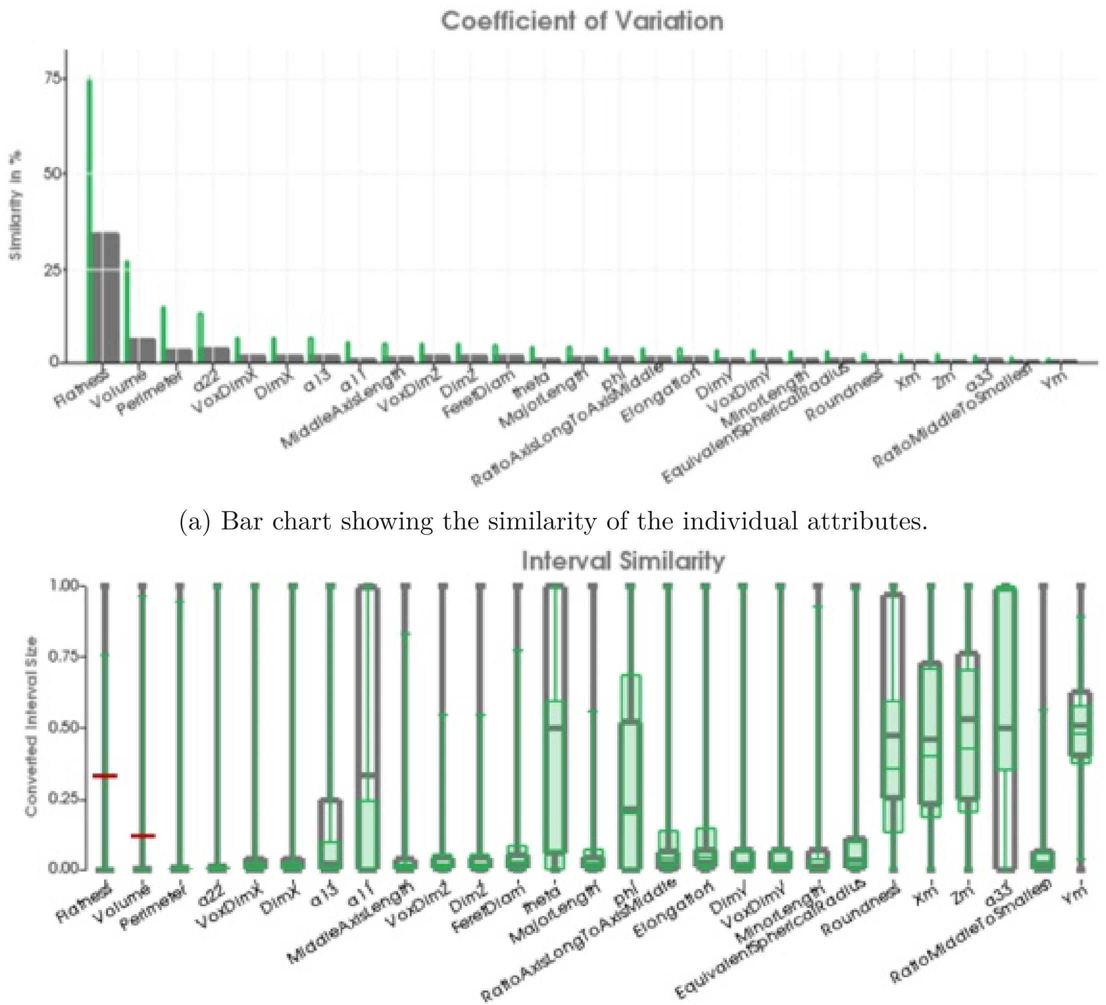


Figure 6.9: Histogram-Table visualizing the selected bins with their point representation.



(a) Bar chart showing the similarity of the individual attributes.

(b) Box plot visualizing the interval similarity of the attributes.

Figure 6.10: Detailed exploration of the pores, selected in Figure 6.9 (green), in respect to all pores of the ensemble (grey).

6.2.2 Analysis of a Fiber Material after Application of a Shear Force

In the second scenario, we show the visualization of a fiber material that was subjected to a shear force over a certain period of time. A more detailed description of the ensemble can be found in Section 6.1. At the beginning, we are mainly interested in the distribution of the individual fibers. Therefore we take a closer look at the Histogram-Table in Figure 6.11. For two different time steps, the four individual areas of the sample are visualized. The first four rows represent the four areas of the sample after a 10 minute exposure to a shear force, the last four rows in the Histogram-Table show the same four areas after 60 minutes.

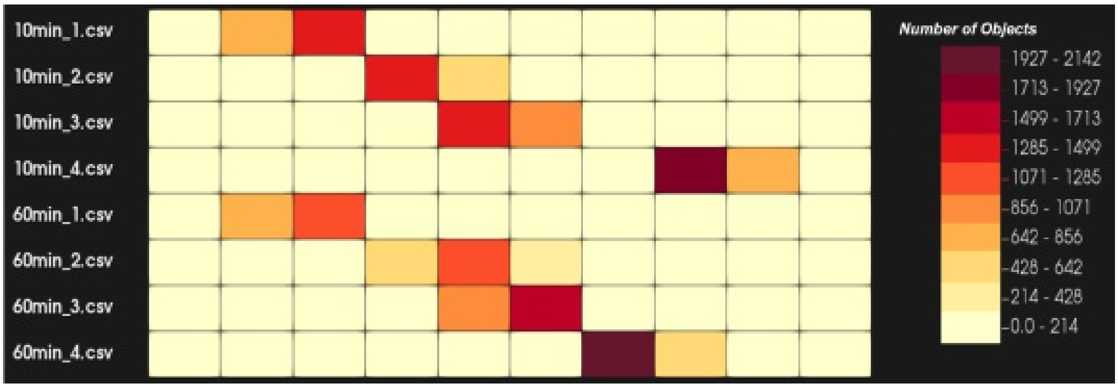


Figure 6.11: Histogram-Table shows each of the four areas of the sample in two different time steps, i.e., after 10 minutes and after 60 minutes.

At first sight, we can detect a similarity between the two time steps. Area *1* of the material scan, indicated with *_1* in the two time steps *10min_1* and *60min_1*, has changed the least in the elapsed time. The areas *2* and *3* already show somewhat larger changes, recognizable by the changing distributions of the fibers, which has moved slightly to the right. The greatest change can be found in area *4*, where a large number of fibers have moved one bin further to the left. Although slight changes in the areas can be observed, the similarity of the individual regions can still be clearly seen.

If we take a closer look at the similarities of the individual attributes, we can see that only low similarities result based on the entirety of all fibers. All bars in Figure 6.12 are very low. However, the three most similar attributes are *Curved Fibre*, *Separated Fibre*, and *Volume*. The first two are deterministic classification values and can only assume boolean values. The attribute *Volume* is a continuous variable and indicates the calculated volume of the individual fibers.

In the box plot in Figure 6.13 we can now take a closer look at the exact characteristics. Although the attributes *Curved Fibre* and *Separated Fibre* can only assume two different values, we can use the box plot to estimate which of the two values will occur more frequently. For the attribute *Curved Fibre* the interquartile range, is equal to the minimum value, which symbolizes that the majority of the fibers in this attribute have a value

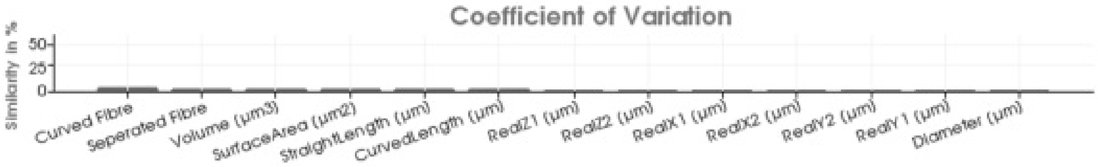


Figure 6.12: Bar chart showing the similarities of the individual attributes. In general, no high degree of similarity can be determined for all fibers, which can be seen from the small size of the bars.

of 0. In contrast, for the attribute *Separated Fibre* the interquartile range ranges from the minimum to the maximum value, which is why the whiskers of the box plot are occluded. This means that this attribute has mainly stored 1 as value. The box plots of the next four attributes show a high similarity of the characteristics, since the boxes are generally very small. Furthermore, they are also located on the lower end of the vertical axis, so we can conclude that the fibers are generally very small, with only a few bigger outliers. The next six attributes describe the start and end positions of the segmented fibers in the three different axes x, y and z. Since their box plots are quite large, one can conclude that the fibers are distributed relatively regularly in the volume. The last attribute describes the diameter of the fibers. Due to the relatively small box plot, we can assume that many fibers have a rather similar diameter. Since the box is placed in the upper third of the vertical axis, most of the fibers have a relatively large diameter, but there exist also some larger and smaller outliers.

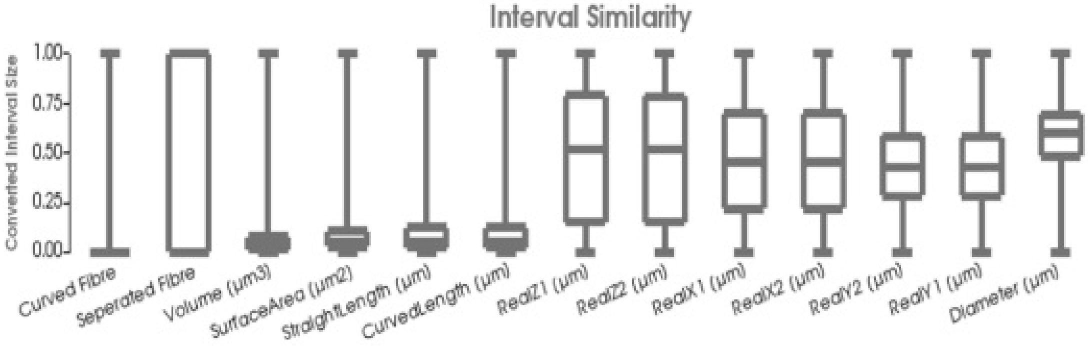


Figure 6.13: Box plot visualizing the interval range of the attributes. The attributes *Curved Fibre* and *Separated Fibre* have the most similar values. While *Curved Fibre* mainly assumes the value zero, *Separated Fibre* mostly assumes the value one. The *Volume* attribute shows a very high similarity, with the fibers shown here all having a very small volume.

As the last step of the initial exploration, we want to check whether the attributes influence each other. Therefore we inspect the correlation map in Figure 6.14. At first sight we can see that there is a positive linear correlation especially between the start and end positions of the different dimensions. Since the length of the fibers, according to the box plot, is relatively constant, the end position of the fibers is at a higher position, the higher the start position is located. Furthermore, the location of these positions shows that they are not, or only slightly, related to the other attributes. For this reason, they are positioned further apart from the other attributes such as *Diameter* or *CurvedLength*. On closer inspection, as seen in Figure 6.15, we can discover four different attributes, which all influence each other. The four attributes are *Volume*, *SurfaceArea*, *CurvedLength*, and *StraightLength*. The red edges between them symbolize that if one attribute of these four has a higher value, all others will also have a higher value.

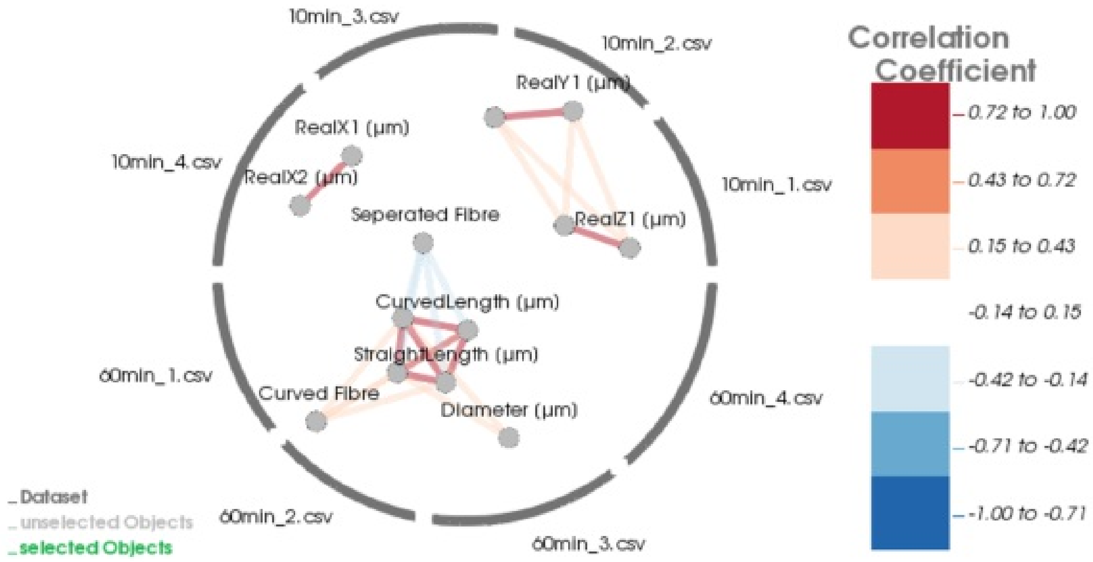


Figure 6.14: Correlation map visualizing the initial state of the fiber ensemble.

Next, we want to take a closer look at the area that changes most under the influence of shear forces. For this reason, we first perform an automatic similarity ranking regarding region *4* in the first time step. As expected, region *4* at the first time step is most similar to the same region in the second time step, as shown in Figure 6.16.

We now look at the two specific bin regions of the *10min_4* and *60min_4* datasets, which contain the most fibers. Figure 6.17 shows the selection of two bins of interest. We will first find out to which extent the fibers are similar within their respective regions. Then we can conclude how the fibers in the two regions differ and thus draw conclusions about their development.

Figure 6.17a shows the exploration of the fibers contained in the red bin in dataset *10min_4*. According to the color scheme, about 2000 fibers are positioned in this bin. The bar chart in Figure 6.18a shows that the fibers are almost 100% similar in the attribute *Curved Fiber*. In the second most similar attribute, the *Volume*, the selected fibers are only 2% similar. Compared to the original calculation, which is based on all

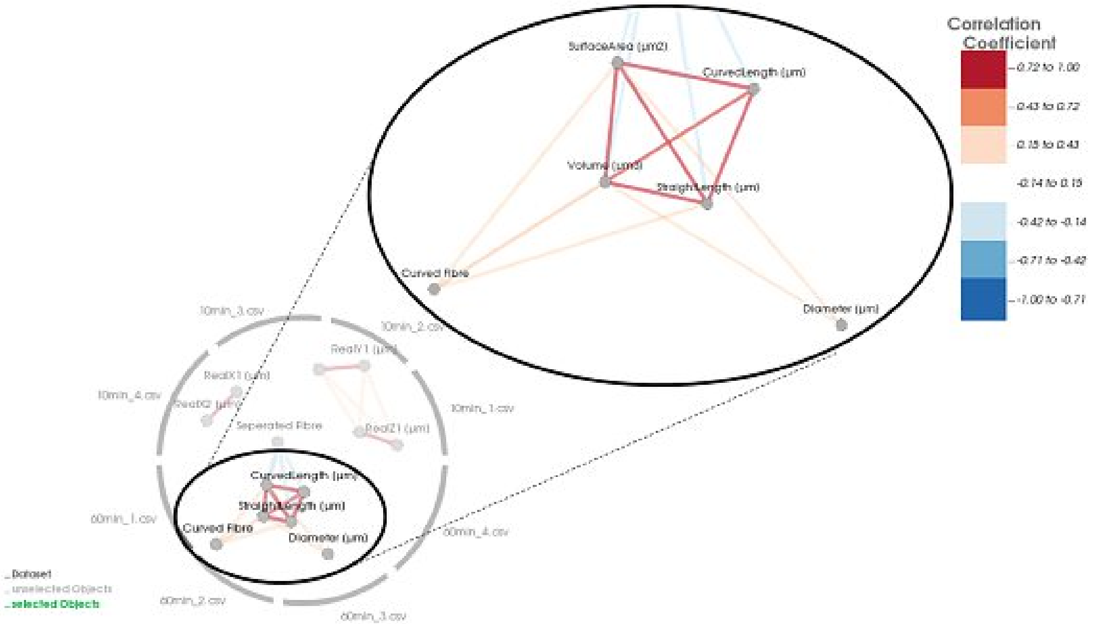


Figure 6.15: Zooming to four individual attributes, where a strong linear relationship could be computed.

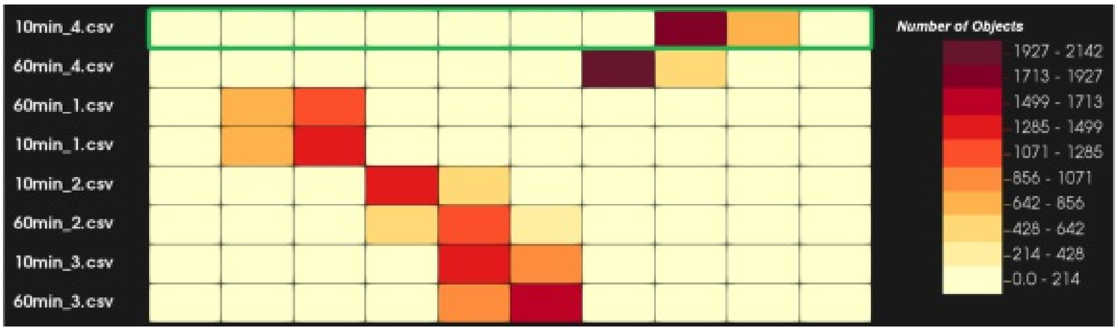
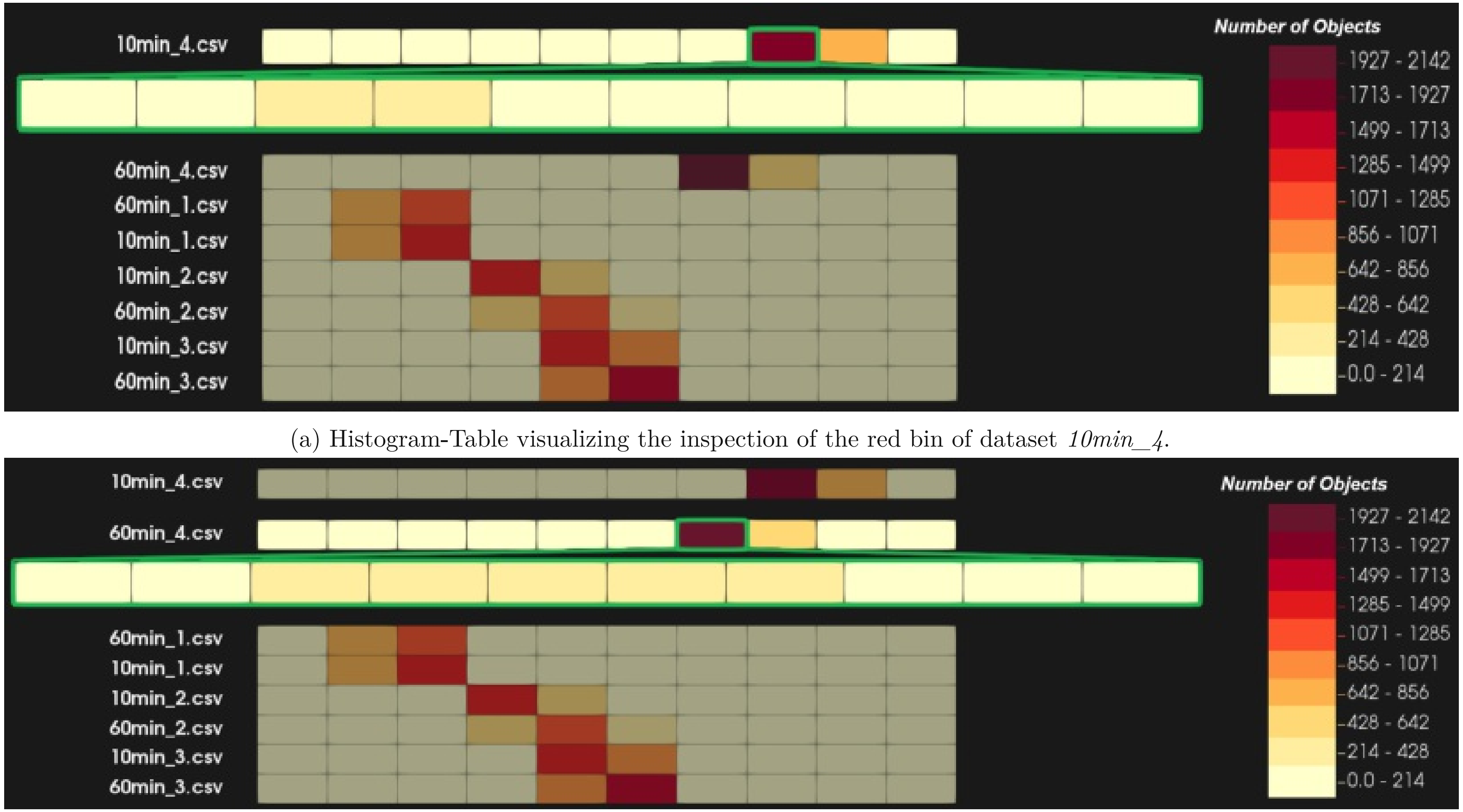


Figure 6.16: Resulting ranking based on dataset *10min_4*.

fibers, their similarity is higher. In the box plot in Figure 6.18b, we can take a closer look at the interval range. At first sight, we can see that the volume is visualized by a very narrow box plot, which is relatively far at the bottom, inside the grey reference box plot. The tool-tip shows the exact values of the volume. As we can see, the first quartile is 6213 μm^3, the median is 11459 μm^3 and the third quartile is 22625 μm^3. Interesting to observe are also the attributes *RealZ1* and *RealZ2*. In contrast to the original boxes, the green boxes are very small and lie pretty much exactly on the first quartile of the original boxes. This suggests that the fibers lying in this bin all have a very similar start and end position on the z axis, meaning that they all lie together in a similar position in the volume.

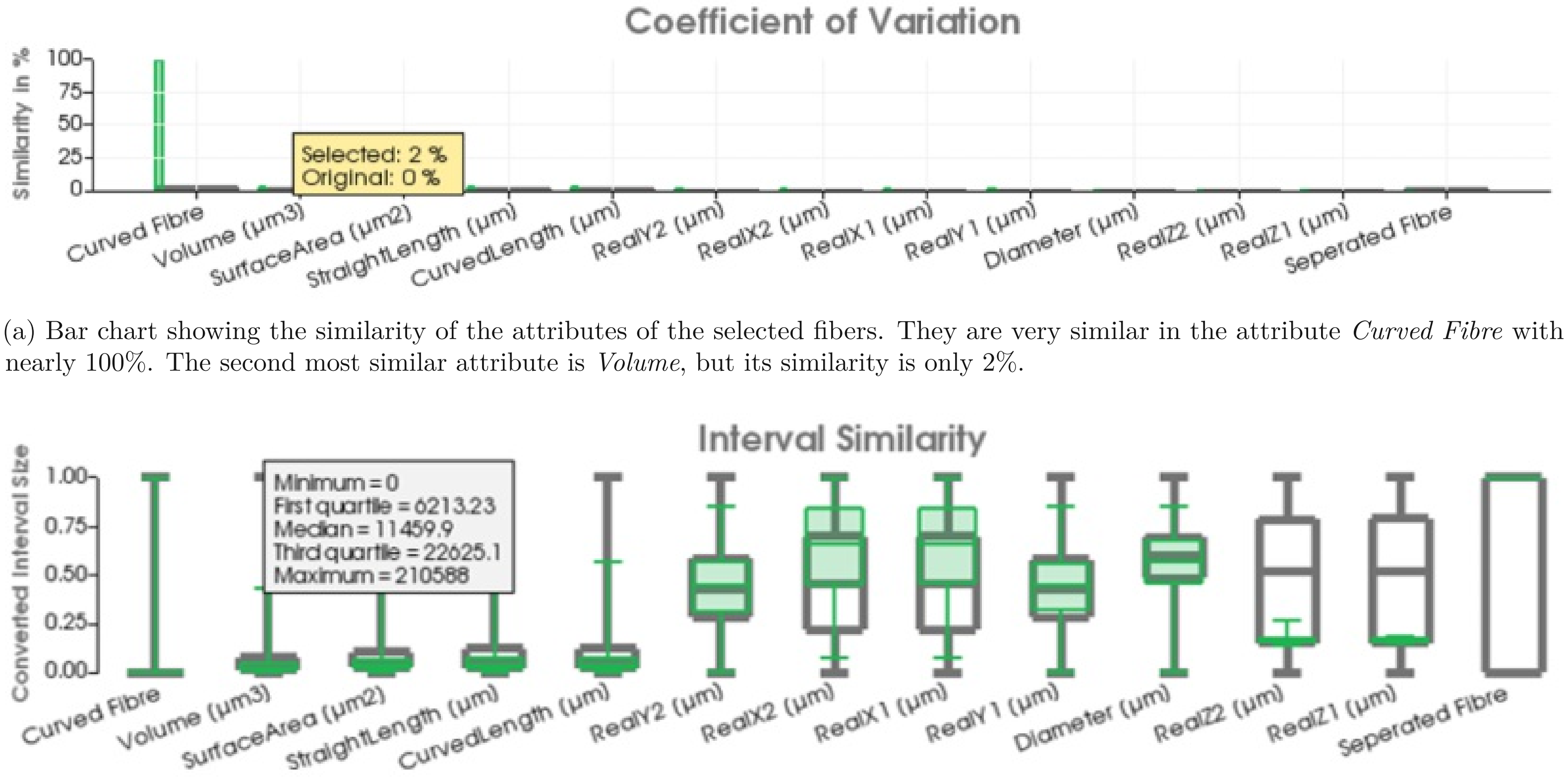
Now we consider the similarity of the red bin, containing the most objects, in area *4* in the second time step. In contrast to the red bin in dataset *10min_4*, this bin is located a bit more on the left side, thus we can observe a development of the fibers in the Histogram-Table in Figure 6.17b. If we now take a closer look at the similarities of the attributes in the bar chart in Figure 6.19a, we discover that these fibers are not so similar in their attribute *Curved Fibre*. Even though the fibers in this attribute are still the most similar, they are only about 25% resembling each other. However, the similarity in the *Volume* attribute is still the same, as the tool-tip shows with a similarity of 2%. The box plot in Figure 6.19b shows the interval similarity of the attributes. The volumes of these fibers are still very similar, and also relatively small. If we take a closer look at the exact values in the tool-tip, we can determine the following values: The first quartile corresponds to 6013 μm^3, the median to 11579 μm^3 and the third quartile to 22653 μm^3.



(a) Histogram-Table visualizing the inspection of the red bin of dataset *10min_4*.

(b) Histogram-Table visualizing the inspection of the red bin of dataset *60min_4*.

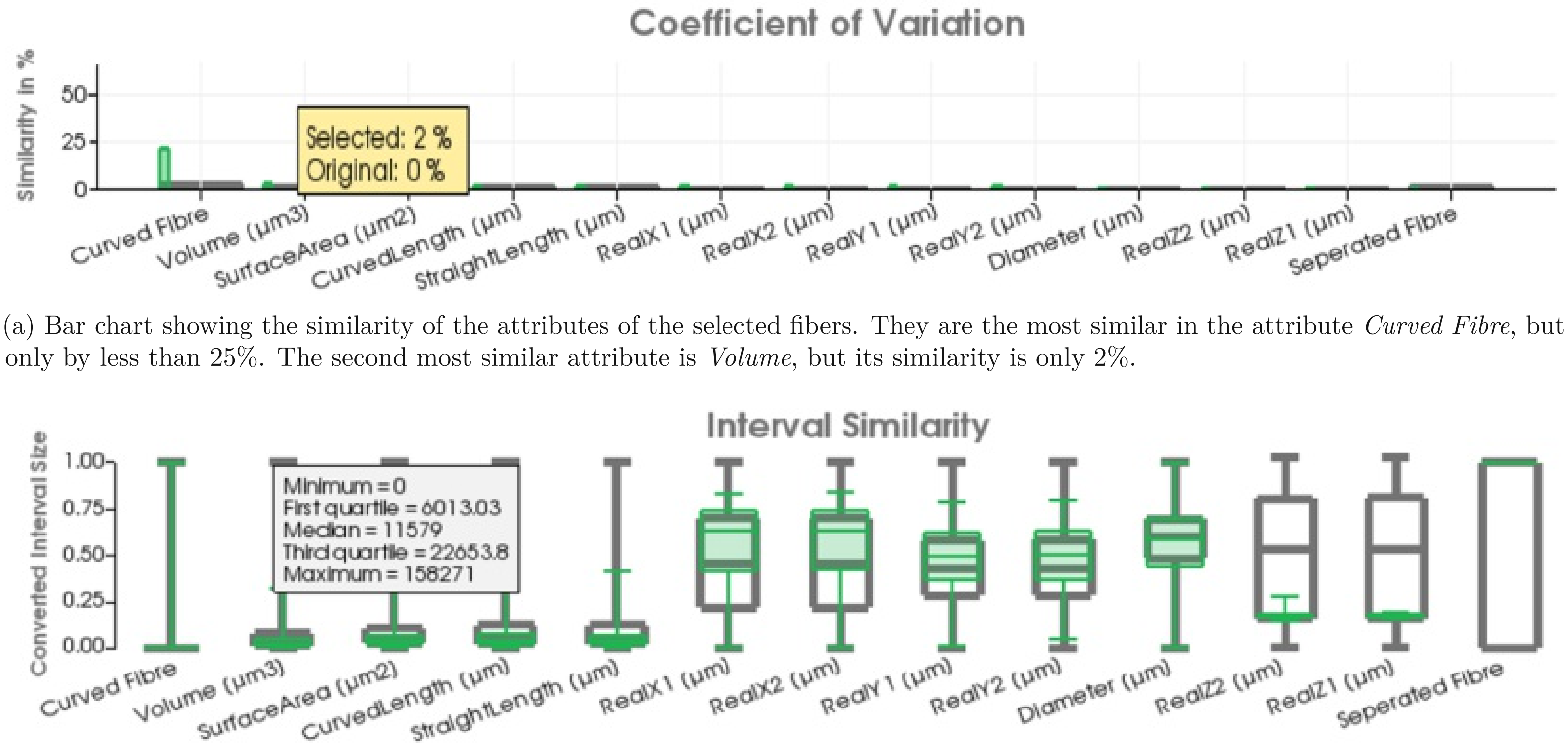
Figure 6.17: Exploration of the similarity of the fibers in area *4* in the first and second time step.



(a) Bar chart showing the similarity of the attributes of the selected fibers. They are very similar in the attribute *Curved Fibre* with nearly 100%. The second most similar attribute is *Volume*, but its similarity is only 2%.

(b) Box plot showing the interval similarity of the fibers. The *Volume*, as well as the following three attributes, are very similar. The precise values of the volume can be read in the tool-tip.

Figure 6.18: Detailed exploration of the similarity of the fibers in area *4* in the first time step.



(a) Bar chart showing the similarity of the attributes of the selected fibers. They are the most similar in the attribute *Curved Fibre*, but only by less than 25%. The second most similar attribute is *Volume*, but its similarity is only 2%.

(b) Box plot showing the interval similarity of the fibers. The *Volume*, as well as the following three attributes, are very similar. The precise values of the *Volume* can be read in the tool-tip.

Figure 6.19: Detailed exploration of the similarity of the fibers in area *4* in the second time step.

6.2.3 Analysis of Unemployment among Men and Women of different Age Classes

In the last scenario we want to show that our analysis framework can also be used to visualize generic datasets and is not necessarily bound to material datasets. Since we represent the objects with their characteristics in a very abstract way, we are not limited to the exact structure and meaning of attributes in material structures. For this reason we visualize an ensemble that describes the unemployment of men and women of different age groups. The exact structure of the ensemble is explained in Section 6.1.

Figure 6.20 shows the Histogram-Table visualizing the Unemployment Ensemble. We can immediately see that there is a clear difference between unemployment for women and men. Even if the largest clusters are located in the middle of the Histogram-Table, a distribution to the left for women and to the right for men can be recognized. However, there are also many countries in which the unemployment rates of both sexes are very similar. These are depicted by the very dark red regions in the middle. If we look more closely at unemployment within the gender group of women, we can observe a greater variability than if we compare the regions within the group of men. Figure 6.21 displays the bar chart and box plot that visualize the unemployment rate in percent for various countries over 39 years. In the bar chart in Figure 6.21a we can see that in the years *1981* and *1979*, the countries' unemployment rates were the most similar. In Figure 6.21b the box plot reveals the interval range of the unemployment rates. We can observe that the unemployment rates were very similar and had a very low value in 1980s and the 1990s. From 2000 onward, the unemployment rates of the individual countries developed differently, but increased overall. This can be observed by the growth of the interquartile ranges of the box plots.

In this scenario we want to focus on the similarity of countries describing the unemployment of men in the age group between 25 and 54. Figure 6.22 shows the selected regions of interest. The bar chart visualizes for each year, how much the unemployment rates differ between the sexes and age groups. The smaller the bars are, the greater the difference between the groups. Figure 6.23a depicts the bar chart, which shows the similarity of all countries per year. It can be seen that unemployment was more similar in the earlier years, but became more dissimilar towards the turn of the millennium. Thus, unemployment among men between the ages of 25 and 54 has evolved divergingly in the different countries. The box plot in Figure 6.23b shows that the unemployment rates increased especially in the 2000s. This can be seen from the fact that the interquartile ranges, i.e., boxes, are getting larger. The larger the interquartile ranges become, the more varying unemployment rates are recorded for a given year. Since the boxes are getting bigger in size, higher unemployment rates were measured in these years.

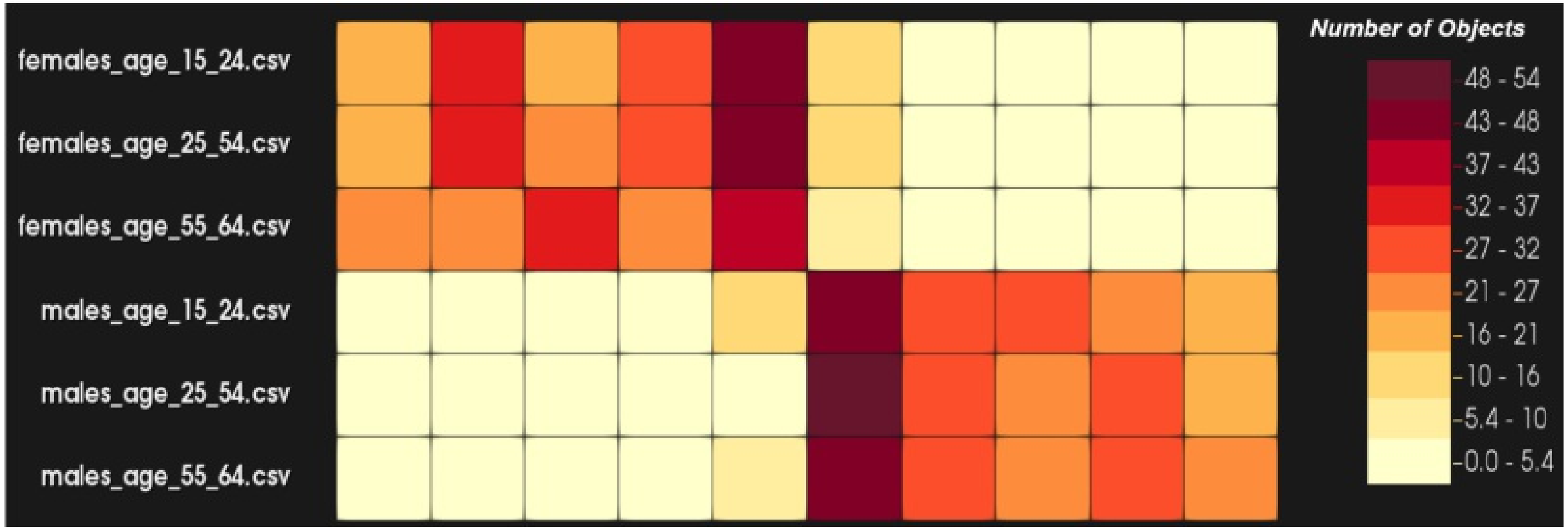
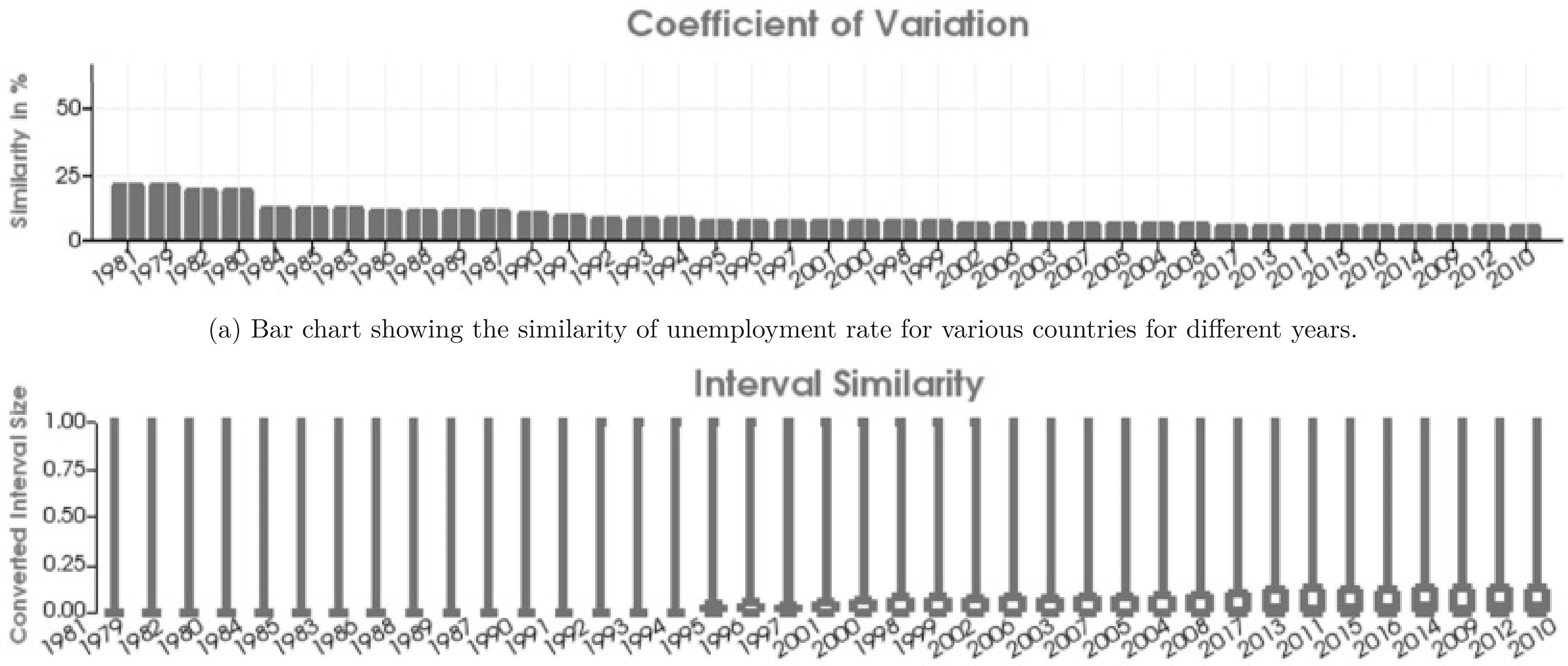


Figure 6.20: Histogram-Table visualizing the selected bins with their point representation.



(a) Bar chart showing the similarity of unemployment rate for various countries for different years.

(b) Box plot visualizing the interval similarity of the different years.

Figure 6.21: Detailed exploration of the unemployment rate among men and women.

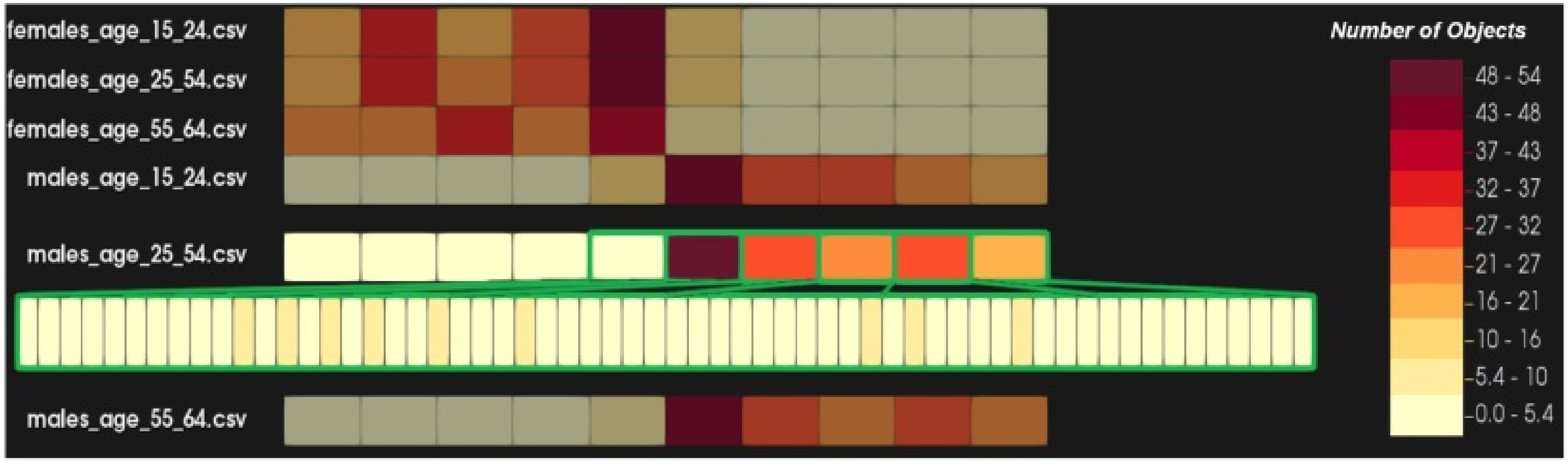
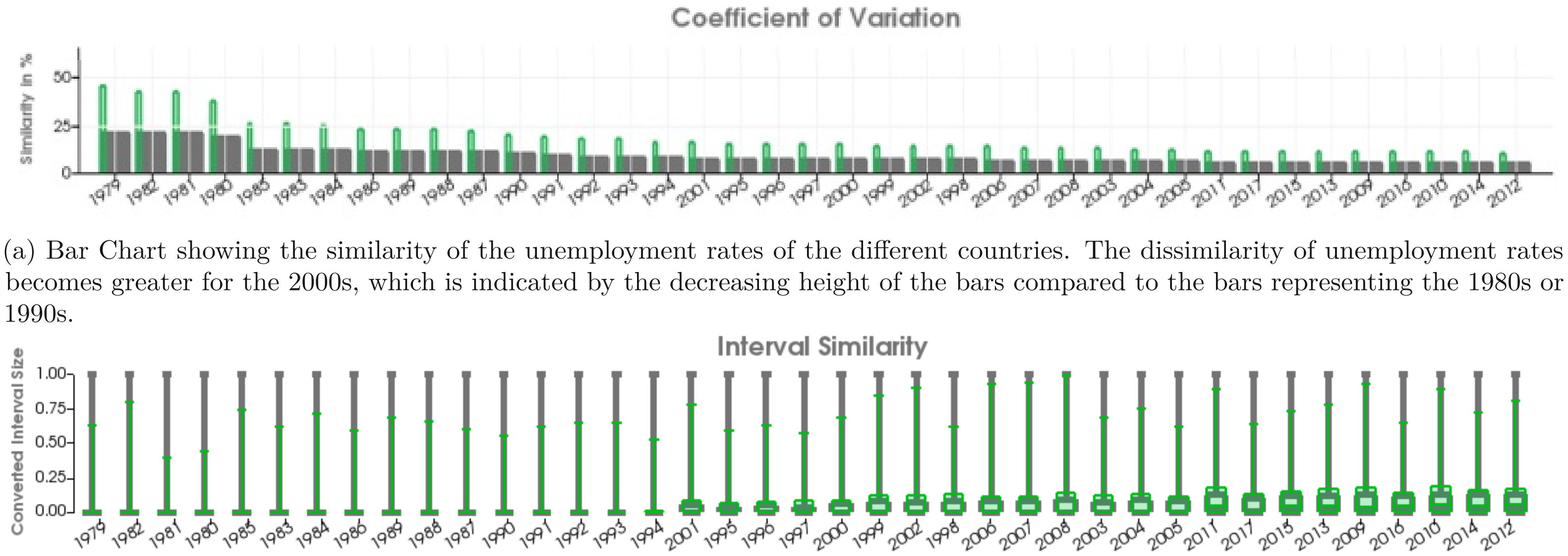


Figure 6.22: Histogram-Table showing the selection of countries in the dataset, which describes the unemployment rate of males with age 25 to 54.



(a) Bar Chart showing the similarity of the unemployment rates of the different countries. The dissimilarity of unemployment rates becomes greater for the 2000s, which is indicated by the decreasing height of the bars compared to the bars representing the 1980s or 1990s.

(b) Box plot showing the interval range of different unemployment rates. Unemployment rises in several countries as the boxes get bigger over time.

Figure 6.23: Detailed exploration of the unemployment rate among men between the age of 25 and 54.

6.3 Performance

All three usage scenarios were run on the same test setup, a laptop equipped with an Intel i7-6820HK CPU with 16 GB of RAM and a 17 inch screen size. In the first two usage scenarios, in each case, a total of about 20000 objects were compared, which is why the calculation of the visualizations took about 17 minutes each. In the third scenario a total of about 1000 elements were compared, resulting in a computation time of about 10 seconds.

The bottleneck in our application is the calculation of the dimension reduction method MDS. The MDS is a very memory consuming method, since it uses a proximity matrix C as input parameter. This matrix has a dimension of $dim\left(C\right) = s \times s$, where s represents the number of all objects from all datasets [Bae08]. This input matrix can become very large, so that the possible number of objects to be compared is currently bound to the size of the memory. In our implementation we used vectors that store the data type double. In C++ each double value requires 8 bytes. Therefore $8 \cdot s \cdot s$ bytes need to fit in the RAM for a reasonably fast calculation. In our experiments the comparison of a maximum of 20000 objects was possible.

Another limitation is the calculation of the MDS. The runtime of the SMACOF algorithm is $O\left(n^3 \cdot k\right)$, where k represents the number of iterations and n the number of objects [Bae08]. The display of the Histogram-Table depends on the calculation time of the MDS and thus forms the bottleneck of our visualization framework.

These performance issues could be solved by, for instance, using a more effective implementation of MDS, as shown in the work of Bea et al. [Bae08]. Another solution would be to implement a less costly and memory intensive dimension reduction method for the abstraction of individual objects, like PCA or Locally linear embedding (LLE) [ZKL18]. When using a different method, however, one must be aware of the different calculation method and the different context of the resulting objects.

6.4 Evaluation Workflow

To check the comprehensibility and usability of our analysis framework, we conducted a qualitative user study. The main focus was on testing the readability and intuitive interpretability of the visualizations. Since our tool was mainly developed for the application in material science, only persons with background knowledge in this field were included in the study. A total of 12 people participated in the study, where six individuals would describe themselves as rather inexperienced, since they were in the middle of their studies or did not deal with this kind of data every day. The remaining participants worked with these type of data almost every day or taught students how to understand and process them. Five of the 12 participants were female.

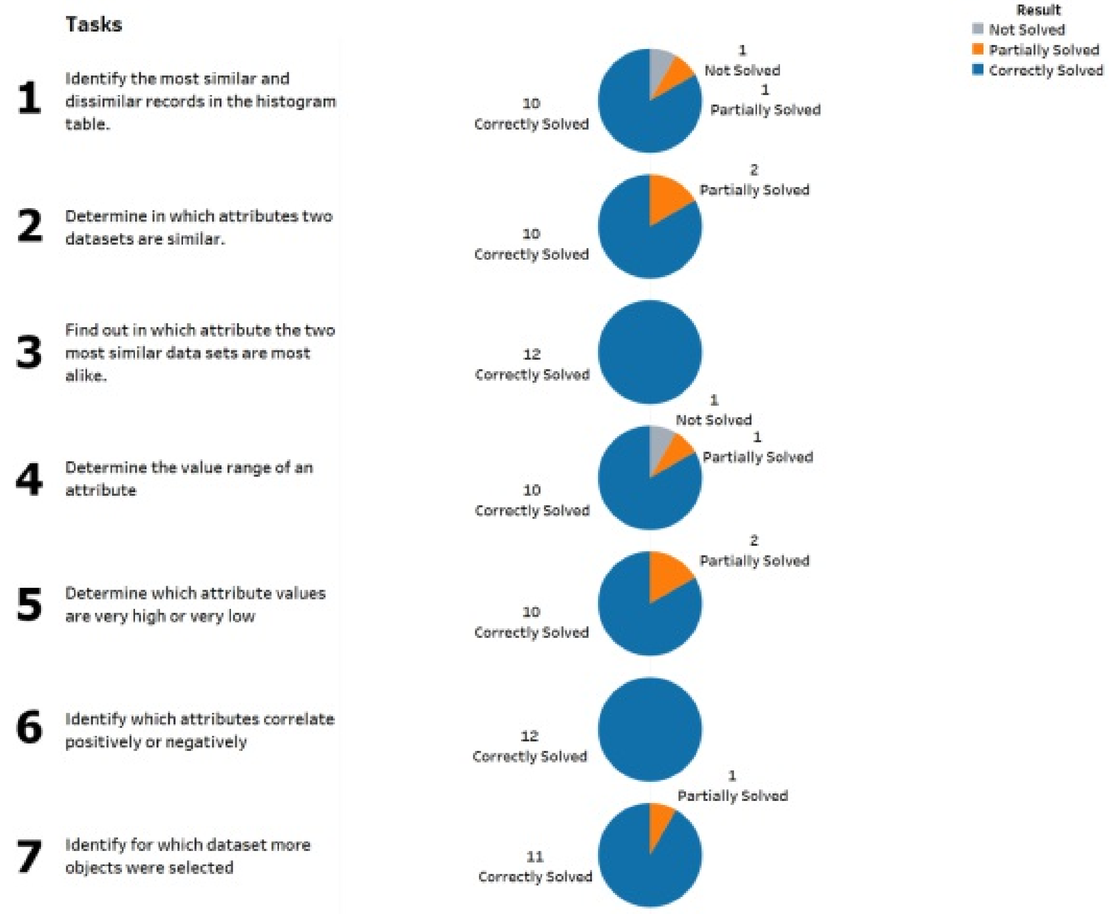
The study was conducted as follows: First, the participants were given a 15 minute introduction to the framework. A small fiber materials example was used to explain the readability of the various visualizations and the different interaction possibilities. The description of the Histogram-Table also contained a short, very general explanation of how the individual objects were mapped to points to make it easier for the participants to interpret the very abstract information. Then the attendees were presented with a new dataset, which they could explore independently for 20 minutes. Meanwhile, the participants were asked to formulate their thoughts aloud and how they would interpret the individual representations. Qualitative tasks were prepared in advance to ensure that all possible interactions were tested and that the visual encodings and their meaning for the participants were vocalized at least once. Figure 6.24 shows the formulated challenges and whether the tasks were immediately answered correctly, partially correctly, or not at all. The dataset from Section 6.2.2 was made available for the participants for free exploration during the study. None of the participants had ever worked with or seen the presented dataset before. They were verbally informed what the dataset contained and how it was generated. This information was seen as an important factor for the user study, as material experts should contribute their domain knowledge during the exploration, to be able to test the usability of the framework under realistic conditions. After the exploration phase, attendees were invited to participate in a short semi-structured interview of about 10 minutes. The following questions were of particular interest to us:

- Were you able to detect shear typical changes of the sample through our visualization techniques?
- Can you imagine using this analysis framework in your daily work routine?
- Have you already worked with Box Plots before?
- Have you ever seen a visualization similar to the Histogram-Table?
- Do you consider the interpretation of the Histogram-Table difficult?

6.5 Evaluation Results

In general, the analysis tool was very positively received. The various tasks were mostly solved independently by the participants during the exploration phase. The exact results of the tasks are shown in Figure 6.24.

The Histogram-Table was perceived by the participants as very clear and descriptive. Once explained how to interpret this visualization, they found the visual encoding obvious and intuitive. Especially the color coding was positively received, as it clearly distinguished the individual bins from each other. The positive perception of the Histogram-Table is demonstrated by the good results in Task **1**. Only two participants could not recognize at first glance which datasets were similar to each other. These users first had difficulties with the strong abstraction of the data. After another short explanation, one of the two participants understood how the individual fibers were mapped. The second person felt,



Tasks

1 Identify the most similar and dissimilar records in the histogram table.

2 Determine in which attributes two datasets are similar.

3 Find out in which attribute the two most similar data sets are most alike.

4 Determine the value range of an attribute

5 Determine which attribute values are very high or very low

6 Identify which attributes correlate positively or negatively

7 Identify for which dataset more objects were selected

Figure 6.24: Visualization of the individual tasks and how the participants solved them

as he expressed himself, that *"the abstraction was too strong and lacked exact numerical values to read the similarity"*. Furthermore, as the study progressed, it became more and more apparent that empty bins, with no objects in them, required their own color coding in the Histogram-Table. This would make it easier for viewers to create an interpretation. Furthermore, it was discussed whether a separate color table for the zoomed Histogram-Table would be useful. Some of the participants thought that it would be advantageous, to have a customized color table to better distinguish the smaller number of objects and thus not only perceive homogeneously colored regions at higher zoom levels. On the other hand, the participants found it good to be able to compare the different zoom levels by using the same color table. This enabled them to keep the overview and context between the changes.

The bar chart was immediately understood by all participants, and was judged by nearly all to be very helpful, as it was easy to identify at a glance, in which attribute the fibers were most similar. This positive feedback is also reflected in the excellent results of Task

2 and Task **3**. Only two participants solved the task merely partially. They argued that, *"they preferred to look at the box plot rather than the bar chart"*. This opinion was based on the fact that the individual bars in the bar chart were very small in size, as they represented low similarity values of about 10% to 15% and the initial axis was scaled to 100%. In their opinion, the axis should be adjusted so that the largest value also takes up all of the displayed part of the axis, even if there is a risk that small similarities are then interpreted as big ones.

Besides the Histogram-Table, the box plot was the visualization that the participants were most interested in. The Histogram-Table was used as overview and to filter interesting areas, while the box plot made the underlying details visible for the selected elements. Most of the test persons felt that the adjacency of the different attributes made a clear exploration easy. The readability of the box plot was strongly dependent on the previous knowledge of the test persons. Tasks **4** and Task **5** could be solved without problems, especially by those who had worked with box plots before. The person who found it difficult to interpret the box plot had never heard of this type of visualization before, and therefore required a more detailed explanation in addition to the general introduction. Although it was possible to inspect the exact values of the individual attributes, the representation of the interquartile ranges was often sufficient to get the necessary impression of the characteristics.

The correlation map got the best feedback regarding its visual encoding. According to the participants, the color coding of the linear correlations was clear and intuitive, while the positioning of the attributes, as stated by one person, *"felt a bit arbitrary"*. Also the change between the different color schemes to view the relations in more detail or, if necessary, to *"switch invisible"* according to one participant, was very popular. The good readability can be shown by the result of Task **6**. The circular segments, positioned around the correlation map, were also evaluated positively and could be interpreted without problems, as shown in Task **7**.

The longer the exploration lasted, the more motivated the participants became to continue using our analysis tool. One respondent noted that *"you can get a lot of information out of the tool, but it requires thinking in order to correctly interpret what you see"*. Another test person made a similar statement saying *"you have to get a feeling for the visualization first, but then you can discover many interesting things"*. Another meant that *"in the beginning it was difficult to approach the problem in such an abstract way, but with time you get used to it"*. The test persons also noticed positively, that they always saw the attributes of the selected fibers in comparison to the ensemble as a whole. Thus, as one of the participants stated, they were able to *"interpret the exact values more easily in comparison to the given ensemble"*.

When asked whether the participants could also use this analysis tool for their daily activities, almost all agreed. Especially in areas such as utility value analysis, where the homogeneity of materials plays an important role, this analysis tool could be tested. The experts can also imagine using our framework for pre-processing tasks. This way, they would not have to display the distribution of each attribute in a separate chart and compare them manually.

To the question, whether the participants had seen a similar visualization before, only three of the test persons affirmed the question. Two of the three regularly deal with different visualizations and had therefore already seen similar ones. The third person had often dealt with biological data, and therefore had seen similar visualizations before. All other participants stated that they mostly worked with graphics that could only show the distribution of one attribute. One visualization they used, but rarely, to display several attributes at once, was the star plot.

All in all, the readability and interpretability of our analysis tool was found to be very good. It turned out that a certain training period was necessary for an efficient use, but the developed visualization methods also offer new interpretation possibilities and perspectives.

CHAPTER 7

Conclusion and Future Work

We would like to conclude this thesis with a summary and an explanation of how the research questions defined in Chapter 1 were addressed. Furthermore, we briefly discuss the limitations of our project and formulate goals and tasks for future work.

7.1 Summary

The main task in our work was to develop a visualization system that would enable material scientists to compare several materials at once, based on their similarity or dissimilarity. In order to fulfill this task in the best possible way, we thoroughly researched the field of material science through literature research and personal interviews with experts in the field. The precise knowledge of their workflow and existing visualizations should help to create a balanced visualization system that provides appropriate tools. In Chapter 2 the findings from these researches were presented in detail.

As described in Chapter 3, we tried to include the broad knowledge from the field of visualization and examined the material science datasets for their structure. We found out that they represent a kind of ensemble of datasets, and thus examined the current state-of-the-art visualization concepts in this area.

Based on the analysis of the individual work steps and the defined domain-specific questions that were established in Chapter 1, we defined three main questions **R1**-**R3**, that our tool was supposed to answer. In the following we describe how our framework addresses these questions:

(R1) *Which datasets are similar or dissimilar?* - Above all, this question was the focus of our work and significantly influenced our implemented framework. It was intentionally kept so general as to take into account the existing scaling problem that the structure of the ensemble data entails. This question can mainly be answered by the Histogram-Table. By abstracting the individual objects, their distribution can be examined more closely not only within a dataset, but also among several datasets. On the one hand, this visualization can handle different degrees of complexity. In general, our visualization can process data, consisting of varying number of attributes, i.e. dimensions, like one or two, up to m dimensions. In addition, our visualization scales well, since each dataset is represented by a single line. The scalability is only limited by the height of the lines in relation to the height of the screen. Furthermore, our visualization system is able to deal with various kinds of data. Due to the high degree of abstraction, the exact structure or the meaning of the structure is not decisive to know to enable a comparison.

(R2) *In which characteristics are the datasets similar or different?* - For material experts, the detail values or at least the interval ranges of the individual attributes are essential for estimating the differences or similarities of the materials. Since important numerical details are lost due to the abstraction of the data in the Histogram-Table, detail visualizations were added to show these details separately. On the one hand, a bar chart was integrated, which visualizes the similarity of the individual attributes by means of a statistical calculation. This provides an overview of the details. On the other hand, a box plot was included, which provides deeper insight by visualizing the exact interval boundaries. This allows users to verify even more precisely in which attributes the datasets are most similar or dissimilar.

(R3) *Is there a correlation between certain characteristics?* - This last question can be answered mainly by the correlation map. Here the linear correlation between the individual attributes is depicted according to a statistical calculation. Due to representation of the characteristics as nodes and the correlations as color-coded edges, a simple interpretation of the results is possible.

With the help of a qualitative evaluation, we tried to determine whether material experts could find answers to the above questions using our analysis tool. The result of the evaluation was quite positive. Additionally, feedback from the participants helped to collect some improvements and suggestions for further development.

7.2 Limitation and Future Work

During the evaluation we could not only collect suggestions for improvement of our individual visualizations, but also discuss further potentially interesting comparison tasks with the participants. Two approaches were considered by many participants as particularly interesting and also important for an even more detailed exploration.

The participants thought that it should also be possible to order the objects in the Histogram-Table according to attributes selected by the user. This would enable a targeted search in the datasets for specific characteristics. This is partly already possible in our framework, since the weighting of the individual attributes can be determined before the calculation of the MDS and thus the resulting position of the objects can be defined according to user-chosen characteristics. If we want to filter for a specific attribute, we can set the weight of this attribute to 100% and the weights of all other attributes to 0%. Since, as mentioned above, the calculation can require a very long time depending on the number of objects, this functionality is not realized in real time in our framework. One solution could be that the MDS pre-calculates a solution for all individual attributes, so that the finished computations are available in real-time.

In our framework the focus was put on comparing selected objects with all objects of the ensemble. During the evaluation, this approach proved to be useful and valuable, but it was also found that the comparison between the selected objects themselves was equally important. Therefore, our next step will be to focus on the comparison of selected objects with each other. This should make it easier to examine the objects for their dissimilarities.

In general, our framework is better suited to visualize similarities than dissimilarities. If objects are similar, this is indicated in the bar chart by large bars, in the box plot by small boxes. Due to the small size of the box plots, the viewer is able to estimate the interval range, in which the attributes are located. For dissimilar objects, the bars in the bar chart are small, the boxes in the box plot are large. We can tell that the selected objects are dissimilar, but we cannot identify which objects have what exact numerical values. A better separation of the visualization of similarities and dissimilarities could lead to additional insights in both use cases.

These limitations and future suggestions for improvement show that there is still much to be explored in the comparative visualization of materials. This work is a further advancement in the integration of comparative visual analysis techniques to improve the visualization of diverse materials. Furthermore, this system is also interesting for use in other fields. Here, specific evaluations with experts from other domains would be necessary and fruitful. In this thesis we presented a comparative analysis framework that enables material scientists to visually investigate various materials for their similarities and dissimilarities. The goal was to enable an efficient identification of trends and patterns in the ensembles, thus enabling a faster, high quality inspection of materials.

List of Figures

Bibliography

[AAS+16] Alexander Amirkhanov, Artem Amirkhanov, Dietmar Salaberger, Johann Kastner, M. Eduard Gröller, and Christoph Heinzl. Visual analysis of defects in glass fiber reinforced polymers for 4dct Interrupted In situ Tests. *Computer Graphics Forum*, 35(3):201–210, jun 2016.

[ADG11] Danielle Albers, Colin Dewey, and Michael Gleicher. Sequence surveyor: Leveraging overview for scalable genomic alignment visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2011.

[AHT20] Shaeela Ayesha, Muhammad Kashif Hanif, and Ramzan Talib. Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion*, jan 2020.

[Bae08] Seung-Hee Bae. Parallel multidimensional scaling performance on multicore systems. In *2008 IEEE Fourth International Conference on eScience*. IEEE, dec 2008.

[Ber83] Jacques Bertin. Semiology of graphics: diagrams. *Networks, Maps*, 1983.

[BJC+19] Juri Buchmuller, Dominik Jackle, Eren Cakmak, Ulrik Brandes, and Daniel A. Keim. MotionRugs: Visualizing collective trends in space and time. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):76–86, jan 2019.

[BWW+17] Arindam Bhattacharya, Johannes Weissenböck, Rephael Wenger, Artem Amirkhanov, Johann Kastner, and Christoph Heinzl. Interactive exploration and visualization using MetaTracts extracted from carbon fiber reinforced composites. *IEEE Transactions on Visualization and Computer Graphics*, 23(8):1988–2002, aug 2017.

[CBS+15] Michael Correll, Adam L. Bailey, Alper Sarikaya, David H. O'Connor, and Michael Gleicher. LayerCake: a tool for the visual comparison of viral deep sequencing data. *Bioinformatics*, 31(21):3522–3528, jul 2015.

[CEK+19] Sang-Yeop Chung, Mohamed Abd Elrahman, Ji-Su Kim, Tong-Seok Han, Dietmar Stephan, and Pawel Sikora. Comparison of lightweight aggregate and foamed concrete with the same density level using image-based characterizations. *Construction and Building Materials*, 211:988–999, jun 2019.

[Cha08] Sung-Hyuk Cha. Taxonomy of nominal and type histogram and distance measures. *City*, 2008.

[CIBP17] John P. Chiverton, Olubisi Ige, Stephanie J. Barnett, and Tony Parry. Multiscale shannon's entropy modeling of orientation and distance in steel fiber micro-tomography data. *IEEE Transactions on Image Processing*, 26(11):5284–5297, nov 2017.

[CMK20] Angelos Chatzimparmpas, Rafael Messias Martins, and Andreas Kerren. t-visne: Interactive assessment and interpretation of t-sne projections. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[Coh13] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[Cro18] Patricia Crossno. Challenges in visual analysis of ensembles. *IEEE Computer Graphics and Applications*, 38(2):122–131, mar 2018.

[CSR^{+}20] Sang-Yeop Chung, Pawel Sikora, Teresa Rucinska, Dietmar Stephan, and Mohamed Abd Elrahman. Comparison of the pore size distributions of concretes with different air-entraining admixture dosages using 2d and 3d imaging approaches. *Materials Characterization*, 162:110182, apr 2020.

[CSS14] Horst Czichos, Birgit Skrotzki, and Franz-Georg Simon. *Das Ingenieurwissen: Werkstoffe*. Springer Berlin Heidelberg, 2014.

[DDW14] Ismail Demir, Christian Dick, and Rüdiger Westermann. Multi-charts for comparative 3d ensemble visualization. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2694–2703, dec 2014.

[dLM09] Jan de Leeuw and Patrick Mair. Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software*, 31(3), 2009.

[FHG^{+}09] Laura Fritz, Markus Hadwiger, Georg Geier, Gerhard Pittino, and M Eduard Gröller. A visual approach to efficient analysis and quantification of ductile iron and reinforced sprayed concrete. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1343–1350, nov 2009.

[FP02] J.-D. Fekete and Catherine Plaisant. Interactive information visualization of a million items. In *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.* IEEE Comput. Soc, 2002.

[FWS^{+}19] Bernhard Fröhler, Johannes Weissenböck, Marcel Schiwarth, Johann Kastner, and Christoph Heinzl. open_iA: A tool for processing and visual analysis of industrial computed tomography datasets. *Journal of Open Source Software*, 4(35):1185, mar 2019.

[Gap20] Unemployment data. https://www.gapminder.org/data/, 08 2020.

[GAW$^+$11] Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D. Hansen, and Jonathan C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, sep 2011.

[Gle18] Michael Gleicher. Considerations for visualizing comparison. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):413–423, jan 2018.

[GVTA10] Sergi Grau, Eduard Vergés, Dani Tost, and Dolors Ayala. Exploration of porous structures with illustrative visualizations. *Computers & Graphics*, 34(4):398–408, aug 2010.

[HHB16] Lihua Hao, Christopher G. Healey, and Steffen A. Bass. Effective visualization of temporal ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):787–796, jan 2016.

[HS17] Christoph Heinzl and Stefan Stappen. STAR: Visual computing in materials science. *Computer Graphics Forum*, 36(3):647–666, jun 2017.

[HTWL18] Xiangyang He, Yubo Tao, Qirui Wang, and Hai Lin. A co-analysis framework for exploring multivariate scientific data. *Visual Informatics*, 2(4):254–263, dec 2018.

[JFSK16] Dominik Jackle, Fabian Fischer, Tobias Schreck, and Daniel A. Keim. Temporal MDS plots for analysis of multivariate data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):141–150, jan 2016.

[KBL19] Christopher P. Kappe, Michael Böttinger, and Heike Leitte. Exploring variability within ensembles of decadal climate predictions. *IEEE Transactions on Visualization and Computer Graphics*, 25(3):1499–1512, mar 2019.

[Kie07] Denis Kiefel. *Quantitative Porositätscharakterisierung von CFK-Werkstoffen mit der Mikro-Computertomografie*. Dissertation, Technische Universität München, München, 2007.

[Kob13] Stephen G. Kobourov. Force-directed drawing algorithms. In Roberto Tamassia, editor, *Handbook of graph drawing and visualization*, pages 383–408. CRC press, 2013.

[Koh06] Wolfgang Kohn. *Statistik: Datenanalyse und Wahrscheinlichkeitsrechnung*. Springer-Verlag, 2006.

[KSW19] Alexander Kumpf, Josef Stumpfegger, and Rüdiger Westermann. Cluster-based Analysis of Multi-Parameter Distributions in Cloud Simulation Ensembles. In Hans-Jörg Schulz, Matthias Teschner, and Michael Wimmer, editors, *Vision, Modeling and Visualization*. The Eurographics Association, 2019.

[KW78] Joseph B. Kruskal and Myron Wish. Multidimensional scaling (quantitative applications in the social sciences). *Beverly Hills*, 1978.

[MGKH09] Kresimir Matkovic, Denis Gracanin, Borislav Klarin, and Helwig Hauser. Interactive visual analysis of complex scientific data as families of data surfaces. *IEEE Transactions on Visualization and Computer Graphics*, 2009.

[MHPK19] Julia Maurer, Christian Hannesschläger, Bernhard Plank, and Johann Kastner. Damage characterisation of short glass fibre reinforced polyamide with different fibre content by an interrupted in-situ x-ray computed tomography test. *International Symposium on Digital Industrial Radiology and Computed Tomography – DIR2019*, 2019.

[MI15] Kourosh Meshgi and Shin Ishii. Expanding histogram of colors with gridding to improve tracking accuracy. In *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*. IEEE, may 2015.

[NKUC20] Kakur Naresh, Kamran Khan, Rehan Umer, and Wesley James Cantwell. The use of x-ray computed tomography for design and process modeling of aerospace composites: A review. *Materials & Design*, 190:108553, may 2020.

[NPJ09] Nikhil Naik, Sanmay Patil, and Madhuri Joshi. A scale adaptive tracker using hybrid color histogram matching scheme. In *Second International Conference on Emerging Trends in Engineering and Technology, ICETET-09*. IEEE, 2009.

[OBJ16] Harald Obermaier, Kevin Bensema, and Kenneth I. Joy. Visual trends analysis in time-varying ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 2016.

[Org20] International Labour Organization. Employment by sex and age — ilo modelled estimates. https://ilostat.ilo.org/data, 08 2020.

[RAK$^+$15] Andreas Reh, Aleksandr Amirkhanov, Johann Kastner, Eduard Gröller, and Christoph Heinzl. Fuzzy feature tracking: Visual analysis of industrial 4d-XCT data. *Computers & Graphics*, 53:177–184, dec 2015.

[RPML19] Dipen Rajak, Durgesh Pagar, Pradeep Menezes, and Emanoil Linul. Fiber-reinforced polymer composites: Manufacturing, properties, and applications. *Polymers*, 11(10):1667, oct 2019.

[Sch16] Johanna Schmidt. *Scalable Comparative Visualization*. Dissertation, Technische Universität Wien, 2016.

[Sch18] Karlheinz Schiebold. *Zerstörende und Zerstörungsfreie Werkstoffprüfung*. Springer Berlin Heidelberg, 2018.

[Shn96] Ben Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*. IEEE Comput. Soc. Press, 1996.

[SKK$^+$11] Dietmar Salaberger, K. Arunachalam Kannappan, Johann Kastner, Jens Reussner, and Thomas Auinger. Evaluation of ct data from fibre reinforced polymers to determine fibre length distribution. *International Polymer Processing*, 26(3):283–291, 2011.

[SML06] Will Schroeder, Ken Martin, and Bill Lorensen. *The Visualization Toolkit (4th ed.)*. Kitware, 2006.

[SNC03] Will Schroeder, Lydia Ng, and Josh Cates. The ITK Software Guide. 2003. ISBN 1-930934-10-6.

[SNHS18] Nasir Saeed, Haewoon Nam, Mian Imtiaz Ul Haq, and Dost Bhatti Muhammad Saqib. A survey on multidimensional scaling. *ACM Computing Surveys*, 51(3):1–25, may 2018.

[SS04] Jinwook Seo and Ben Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. IEEE, 2004.

[SSW17] Yapeng Su, Qihui Shi, and Wei Wei. Single cell proteomics in biomedicine: High-dimensional data acquisition, visualization, and analysis. *Proteomics*, 17(3-4):1600267, feb 2017.

[TAES09] Andrada Tatu, Georgia Albuquerque, Martin Eisemann, and Jorn Schneidewind. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2009.

[TMG$^+$10] Matthias Teßmann, Stephan Mohr, Svitlana Gayetskyy, Ulf Haßler, Randolf Hanke, and Günther Greiner. Automatic determination of fiber-length distribution in composite material using 3d CT data. *EURASIP Journal on Advances in Signal Processing*, 2010(1), may 2010.

[WAL$^+$14] Johannes Weissenböck, Artem Amirkhanov, Weimin Li, Andreas Reh, Alexander Amirkhanov, Eduard Gröller, Johann Kastner, and Christoph Heinzl. FiberScout: An interactive tool for exploring and analyzing fiber reinforced polymers. In *2014 IEEE Pacific Visualization Symposium*. IEEE, mar 2014.

[WAS$^+$18] Johannes Weissenböck, Mustafa Arikan, Dietmar Salaberger, Johann Kastner, Jan De Beenhouwer, Jan Sijbers, Stefanie Rauchenzauner, Tanja Raab-Wernig, Eduard Gröller, and Christoph Heinzl. Comparative visualization of orientation tensors in fiber-reinforced polymers. In *Proceedings of the 8th*

Conference on Industrial Computed Tomography (iCT 2018), Wels, Austria, 2018.

[WDJ15] Wolfgang Weißbach, Michael Dahms, and Christoph Jaroschek. *Werkstoffkunde*. Springer Fachmedien Wiesbaden, 2015.

[WFG+19] Johannes Weissenböck, Bernhard Fröhler, Eduard Gröller, Johann Kastner, and Christoph Heinzl. Dynamic volume lines: Visual comparison of 3d volumes through space-filling curves. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):1040–1049, jan 2019.

[WHLS19] Junpeng Wang, Subhashis Hazarika, Cheng Li, and Han-Wei Shen. Visualization and visual analysis of ensemble data: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 25(9):2853–2872, sep 2019.

[ZKL18] Jelena Zubova, Olga Kurasova, and Marius Liutvinavičius. Dimensionality reduction methods: the comparison of speed and accuracy. *Information Technology And Control*, 47(1), mar 2018.

[ZMM12] Zhiyuan Zhang, Kevin T. McDonnell, and Klaus Mueller. A network-based interface for the exploration of high-dimensional data spaces. In *IEEE Pacific Visualization Symposium 2012*. IEEE, 2012.

[ZMZM15] Zhiyuan Zhang, Kevin T. McDonnell, Erez Zadok, and Klaus Mueller. Visual correlation analysis of numerical and categorical data on the correlation map. *IEEE Transactions on Visualization and Computer Graphics*, 21(2):289–303, feb 2015.