UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

Dottorato di Ricerca in Informatica

XII Ciclo - Nuova Serie

Tesi di Dottorato in

3D data visualization techniques and applications for visual multidimensional data mining

Fabrizio Torre

Ph.D. Program Chair

Ch.mo Prof. Giuseppe Persiano

Supervisors

Ch.mo Prof. Gennaro Costagliola

Dott. Vittorio Fuccella

Anno Accademico 2013/2014

To my children:
Sara and Luca in the hope of being able to
lovingli preserve and grow...
the little Angels who surely watch over me from the sky...

Abstract

Despite modern technology provide new tools to measure the world around us, we are quickly generating massive amounts of high-dimensional, spatial-temporal data. In this work, I deal with two types of datasets: one in which the spatial characteristics are relatively dynamic and the data are sampled at different periods of time, and the other where many dimensions prevail, although the spatial characteristics are relatively static.

The first dataset refers to a peculiar aspect of uncertainty arising from contractual relationships that regulate a project execution: the dispute management. In recent years there has been a growth in size and complexity of the projects managed by public or private organizations. This leads to increased probability of project failures, frequently due to the difficulty and the ability to achieve the objectives such as on-time delivery, cost containment, expected quality achievement. In particular, one of the most common causes of project failure is the very high degree of uncertainty that affects the expected performance of the project, especially when different stakeholders with divergent aims and goals are involved in the project.

The second data set refers to a novel display of Digital Libraries (DL). DL allow to easily share the contribution of each individual researcher in the scientific community. Unfortunately, the webpage-based paradigm is a limit to the inquiry of current digital libraries. In recent years, in order to allow the user to quickly collect a set of documents judged as useful for his/her research, several visual approaches have been proposed.

This thesis deals with two main issues. The first regards the analysis of the possibilities offered by a 3D novel visualization technique

3DRC to represent and analyze the problem of diverging stakeholders views during a project execution and to addresses the prevention and proactive handling of the potential controversies among project stakeholders. The approach is based on the 3D radar charts, to allow easier and more immediate analysis and management of the project views giving a contribution in reducing the project uncertainty and, consequently, the risk of project failure. A prototype software tool implementing the 3DRC visualization technique has been developed to support the 3DRC approach validation with real data. The second issue regards the enhancement of CyBiS, a novel 3D analytical interface for a Bibliographic Visualization Tool with the objective of improving scientific paper search.

The benefits of using 3D techniques combined with techniques for multidimensional visual data mining will be also illustrated, as well as indications on how to overcome the drawbacks that afflict 3d visual representations of data.

Acknowledgement

Thank you! Ten times thank you!

In this brief but intense experience there are few people with whom I interacted, and whom I owe a "Thank you", no matter whether small or big but still a "Thank you!"

To Prof. Gennaro Costagliola. For his friendship, availability, and especially for his infinite patience. For giving me the opportunity to work in the field of visualization, for his expert guidance and for his encouragement and support at all levels.

To my parents, who grew me lovingly; always urging me to do more and better, without ever invading my freedom.

To Antonietta, with whom I shared the joys and sufferings during these years and who has given me the two most precious jewels.

To the President of C.S.I. Prof. Gugliemo Tamburrini who allowed the start of this unforgettable experience.

To my Technical Director Antonella who supported me before and during the doctorate, and even after my work return.

To Mattia with whom I shared this entire experience from the examinations at nights in the hotel and for the friendship shown to me.

To Vittorio for the friendship and the support provided in the writing of the thesis.

To Prof. Giancarlo Nota and Dott.sa Rossella Aiello with whom we have given life to 3DRC.

To Prof. Giuseppe Persiano for his efforts in organizing this doctorate.

And last but obviously not least, the Good God, who always accompanied me in life as well as in my studies.

Contents

Chapter 1

Introduction

> Yahweh spoke to Moses: ≪Take a census of the whole community of Israelites by clans and families, taking a count of the names of all the males, head by head. You and Aaron will register all those in Israel, twenty years of age and over, fit to bear arms, company by company.≫
>
> — Numbers 1: 2-3, *The Holy Bible*

Thanks to the progress made in hardware technology, modern information systems are capable of storing large amounts of data. For example, according to a survey conducted in 2003 by researchers from the University of Berkeley, about 1-2 Exabyte (= 1 million terabytes) of unique information were generated each year, of which a large part is available in digital format. These data, including simple tasks of daily life such as paying by credit card, using the phone, accessing e-mail servers, are generally recorded on business databases via sensors or monitoring systems [1, 2, 3].

This continuous recording tends to generate an information overload. Exploring and analyzing the vast volumes of data becomes increasingly complex and difficult. Besides, due to the number and heterogeneity of the recorded parameters we are forced to interact with multidimensional data with high dimensionality.

What encourages us to collect such data is the idea that they are a potential source of valuable information not to be missed. Unfortunately, the task of finding information hidden in such a big amount of data is very difficult. In this context, data analysis is the only way to build knowledge from them in order to successively take important decisions. In fact, this precious information can not be used until it remains locked in its "container", and the time, the relationships and the rules that govern it are not shown in a "readable" and understandable way.

In addition, many of the existing data management systems can only view "rather small portions" of data. Besides, if the data is presented textually, the amount that one can see is in the range of about 100 items, which is of no help when we have to interact with collections containing millions of data items.

Information visualization or *InfoVis* (**IV**) and *Visual Data Mining* (**VDM**) can be a valuable help to manage big amounts of data.

1.1 Introduction to Data Mining

Data Mining (**DM**) [2, 4, 5, 6] is a generic term that covers a variety of techniques to extract useful information from (large) amounts of data, without necessarily having knowledge of what will be discovered, and the scientific, industrial or operating use of this knowledge. The term "useful information" refers to the discovery of patterns and relationships in the data that were previously unknown or even unexpected. Data mining is also sometimes called *Knowledge Discovery in Databases* (**KDD**) [6].

A common mistake is to identify DM with the use of intelligent technologies that can detect, magically, models and solutions to business problems. Furthermore, this idea is false even in the presence of techniques, more and more frequent, that allow to automate the use of traditional statistical applications.

Data Mining is an interactive and iterative process. In fact, it requires the cooperation of business experts and advanced technologies in order to identify relationships and hidden features in the data. It is through this process that the discovery of an apparently useless model can often be transformed into a valuable source of information. Many of the techniques used in DM are also referred to as "machine learning" or "modelling" (see Sec. 2.3).

In the past, various methods for data mining and process models have been developed, but only in 2000 the important goal of establishing the *Cross-Industry Standard Process for Data Mining* (**CRISP-DM**) (Chapman et al. [7]) has been achieved.

The CRISP-DM methodology breaks the data mining process into six main phases [8]:

1. *Business Understanding* - identify project objectives,
2. *Data Understanding* - collect and review data,
3. *Data Preparation* - select and cleanse data,
4. *Modelling* - manipulate data and draw conclusions,
5. *Evaluation* - evaluate model and conclusions,
6. *Deployment* - apply conclusions to business.

The six phases fit together in a cyclical process and cover the entire data mining process, including how to incorporate data mining in the largest business practices. These phases are listed in detail in the diagram in figure 1.1.

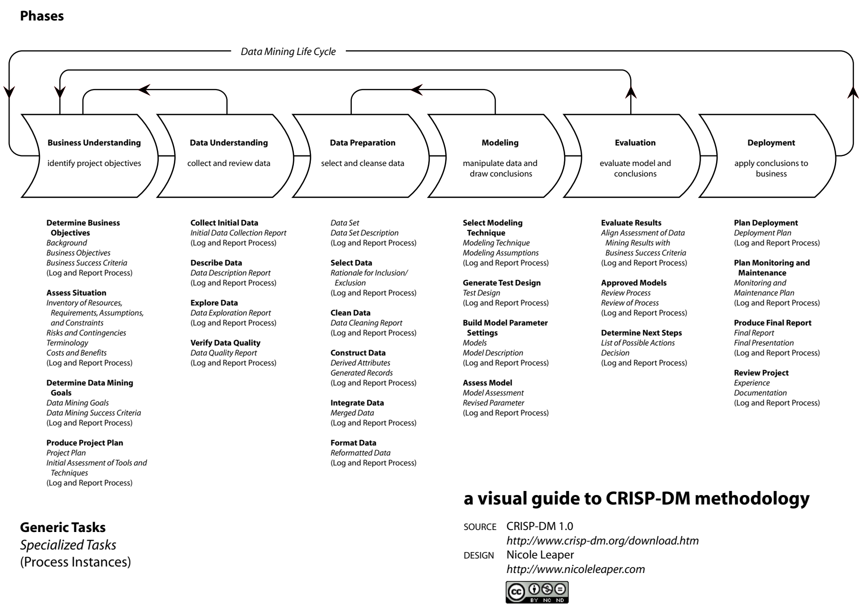


Figure 1.1: a visual guide to CRISP-DM methodology (image taken from [9])

1.2 Information Visualization

The field of information visualization has emerged "from research in human-computer interaction, computer science, graphics, visual design, psychology, and business methods. It is increasingly applied as a critical component in scientific research, digital libraries, data mining, financial data analysis, market studies, manufacturing production control, and drug discovery". [10]

In the literature there are several definitions for "IV". Here are some examples ordered by year:

- "A method of presenting data or information in non-traditional, interactive graphical forms. By using 2-D or 3-D color graphics and animation, these visualizations can show the structure of information, allow one to navigate through it, and modify it with graphical interactions." - 1998 [11]

- "The use of computer-supported, interactive, visual representations of abstract data to amplify cognition." 1999 [12]

- "Information visualization is a set of technologies that use visual computing to amplify human cognition with abstract information." 2002 [13]

- "Visual representations of the semantics, or meaning, of information. In contrast to scientific visualization, information visualization typically deals with nonnumeric, nonspatial, and high-dimensional data." 2005 [14]

- "Information visualization (InfoVis) is the communication of abstract data through the use of interactive visual interfaces." 2006 [15]

A common misconception is to confuse IV with scientific visualization or SciVis [16]. In fact, InfoVis implies that the spatial representation is chosen, while SciVis implies that the spatial representation is given.

In conclusion thanks to the use of IV we can reduce large data sets to a more readable graphical form compared to a direct analysis of the numbers. In this way human perception can detect patterns revealing the underlying structures of the data [17].

1.3 From Data Mining to Visual Data Mining

Visual data mining [18, 2] is a new approach to data mining by presenting data in visual form (or graphics). The basic idea is to use, through IV methods, the latest technologies to combine the phenomenal abilities of human intuition and the ability of the mind to visually distinguish patterns and trends in a large amount of data with traditional data mining techniques. This, to allow the data miner to “enter” in the data, get insight, draw conclusions and interact directly with them. For these reasons information visualization techniques have been considered from the outset as a valid alternative to the analytical methods based on statistical and artificial intelligence techniques.

VDM techniques have proven to be of great help in exploratory data analysis showing a particularly high potential when exploring large databases, in the case of vague knowledge about data and exploration goals, and in the presence of highly inhomogeneous and noisy data.

In [19], qualifying characteristics for VDM systems are identified as follows:

- **Simplicity** - they should be intuitive and easy to use, while providing effective and powerful techniques.
- **User autonomy** - they should guide and assist the data miner through the data mining process to draw conclusions, but also keep it under control.
- **Reliability** - They should provide estimated error or accuracy of the information provided for each stage of the extraction process.
- **Reusability** - A VDM system must be adaptable to various environments and scenarios in order to reduce the efforts of customization and improve the portability between systems.
- **Availability** - Since the search for new knowledge can not be planned, access to VDM systems should be generally and widely

available. This implies the availability of portable and distributed solutions.

- **Security** - Such systems should provide an appropriate level of security to prevent unauthorized access in order to protect data, the newly discovered knowledge, and the identity of users.

1.4 2D vs 3D Representations

The real world is three-dimensional when considering the spatial dimensions and it becomes even "four-dimensional" when we include the time. With the advent of PC graphical interfaces most of the methods of graphical visualization that required a transposition of data on sheets of paper have become obsolete. Moreover we are now also able to represent three-dimensional space (3D) on the screen of a PC. In the last decade there has been great ferment about the creation of new techniques for 3D data visualization.

Since there are obvious advantages in the use of traditional techniques (2D), some considerations are required [20]:

- the science of cognition indicates that the most powerful pattern-finding mechanisms of the brain work in 2D and not 3D;
- there is already a robust and effective methodology about the design of diagrams and the data representation in two dimensions;
- 2D diagrams can be easily included in books and reports;
- it is easy to come across an abundance of ill-conceived 3D design.

In light of the above considerations, why it is worth to study three-dimensional approach?

A quite convincing reason for an interest in 3D space perception and hence searching for new methodologies for the 3D data representation is the explosive growth of studies on 3D computer graphics. Today there are several software libraries that allow programmers to rapidly develop interactive three-dimensional virtual spaces.

Moreover a 3D representation has several advantages [21, 22, 23, 24, 25]:

- *Shape constancy*: it is the property that allows us to recognize from several points of view, irrespective of the particular 2D image projected onto the retina, the 3D shape of an object (or part of it);
- *Shape Perception*: unfamiliar shapes stand out immediately. This is done "unconsciously" and requires minimal cognitive effort;
- The use of 3D shapes enable us to encode more information than other visual properties such as scale, weight, and color;
- Users can take advantage of 3D displays to visualize large sets of hierarchical data;
- It possible to show more objects in a single screen thanks the perspective nature of 3D representations, objects shrink along the dimension of depth;
- If more information is visible at the same time, users gain a global view of the data structure;

Unfortunately there are also disadvantages that must be taken into account when designing 3D visual interfaces:

- The problem of occlusion: in a 3D scene, foreground objects necessarily block the view of the background objects.
- A lower data/pixel ratio than 2D visualizations: this significantly limits the quantity of visible data.

Since "In medio stat virtus" the best choice, used in this thesis, is to employ both approaches. It is likely that a system that allows the user to easily switch from a 2D to a 3D display, and vice versa, would be beneficial for the user. In fact, this would enable us to extend the plethora of tools at our disposal for understanding data sets.

In developing both the new version of CyBiS [26] and 3DRC [27] a thorough study has been made on the techniques and to maximize the benefits and reduce the problems regarding the usage of a 3D representation, and to create a visual representation of the semantics that is not ambiguous.

1.5 Objectives of the thesis

In literature there are a great number of information visualization techniques developed over the last decade to support the exploration of large data sets, and this thesis wants to be a small contribution in this direction.

The thesis focuses on how to apply Information Visualization and the use of 3D data visualization techniques to overcome the drawbacks of two specific domains: *Bibliographic Visualization Tools* and *Evaluation of Stakeholders Project Views*. In the following I will describe the drawbacks and the proposed solution for each of the two domains.

1.5.1 Bibliographic Visualization Tools drawbacks

Description: one of the key tasks of scientific research is the study and management of existing work in a given field of inquiry [28]. Digital Libraries (DLs) are generally referred to as "collections of information that are both digitized and organized" [29]. Digital Libraries are content-rich, multimedia, multilingual collections of documents that are distributed and accessed worldwide. The traditional functions of search and navigation of bibliographic databases, digital libraries, become day by day more and less appropriate to assist users in efficiently managing the growing number of scientific publications. Unfortunately, the webpage-based paradigm is a limit to the inquiry of current digital libraries. In fact, it is well known how difficult it is to retrieve relevant literature through traditional textual search tools. The limited number of shown items, the lack of adequate means for evaluating paper chronology, influence on future work, relevance to given terms of the query are some of the main drawbacks.

Proposed solution: a proposed solution for some of these aspects is CyBiS [26], which stands for Cylindrical Biplot System, an analytical 3D interface developed for a Bibliographic Visualization Tool with the objective of improving the search for scientific articles. Unfortunately CyBiS has some limitations regarding the choice of the data to be displayed and its corresponding 3D presentation, and presents the problem of occlusion. For these reasons, and to continue on the work undertaken on the previous version, it was necessary the realization of CyBiS2 in order to extend and emphasize strengths of CyBiS and to improve it by also adding new and useful features and operations.

1.5.2 Evaluation of Stakeholders Project Views drawbacks

Description: in order to be effective, organizations need to deal with uncertainty especially when they are involved in the realization of projects [30]. Moreover, the growth in size and complexity of the projects managed by public or private organizations leads to an increase in the degree of uncertainty. This is reflected in an increasing risk of disputes, which in turn have the potential to produce additional uncertainty and therefore may seriously threaten the success of the project. Among the most common causes of failure there are the difficulties in achieving targets, such as on-time delivery, cost containment, quality objectives planned and the lack of involvement in the project of the various stakeholders (especially when the points of view are not concordant).

Proposed solution : starting from the fundamental definition of *view* that a contractor has at a given time on the progress of a project and the related concept of *Consonance*, *Project*, *Delta* and *Gap*, the 3DRC representation is proposed for the management of potential disputes between parties that should help the reconciliation of diverging views. The 3DRC representation has been developed to allow easier and more immediate analysis and management of the project views giving a contribution in reducing the project uncertainty and, consequently, the risk of project failure.

1.5.3 The goals

The work done on CyBiS has been very useful for the realization of 3DRC, in fact some of the ideas used to improve CyBiS were of great help to the design of 3DRC. For example, the 3D information visualization combined with the salient radar-chart features taken from the metaphor of a cylindrical three-dimensional space, present in CyBiS, were the sparks of inspiration for the design and subsequent implementation of 3DRC.

In light of the foregoing my thesis aims to achieve a number of objectives:

1. describe some of the design choices that allow to remedy some of the disadvantages of 3D interfaces;

2. apply the techniques of IV and VDM to the context of project management in order to provide new tools for dealing with uncertainty in the management of complex projects;

3. present a new visualization technique, called 3DRC that addresses the prevention and proactive handling of the controversies among potential project stakeholders;

4. show, through case studies, how the salient features of a 3D representation are able to:

   (a) in the case of CyBiS: find the relevance of documents with respect to the search query or affine terms; show the influence given by the number of citations per year; highlight the correlations among documents, such as references and citations.

   (b) in the case of 3DCR: increase the analysis capability of project managers and other stakeholders; highlight the points of divergence or convergence between two or more stakeholders; compare multiple trends over time; understand how different stakeholders interact with each other across the time on the same activity; compare different activities and eventually discover some hidden correlations between them.

To support experiments towards these goals and to display the 3DRC features the 3DRC Tool was developed.

1.6 Outline of the thesis

The thesis is organized as follows.

Chapter 2 provides an introduction to the principles of Data Visualization. It presents a detailed taxonomy of visualization techniques. Furthermore, the principles of multidimensional data visualization techniques and how they can be used for Data Mining are illustrated.

Chapter 3 presents CyBys2, its previous version CyBiS, related works and the improvements provided by the new interface.

Chapter 4 describes the 3DRC visualization technique and its application domain. At first the definitions of *Consonance*, *Project*, *User view*, *Delta* and *Gap* and then the 3DRC specific properties and operations are provided.

Chapter 5 explains the architecture of the 3DRC Tool together with the user operations needed to interact with 3DRC. Furthermore a case study is provided in order to validate the model.

Finally Chapter 6 contains the final discussion, concluding remarks and future directions.

Chapter 2

Data Visualization in Data Mining

> All nations will be assembled before him and he will separate people one from another as the shepherd separates sheep from goats. He will place the sheep on his right hand and the goats on his left.
>
> — Matthew 25: 32-33, *The Holy Bible*

In order for data mining to be effective, it is important to combine human creativity and pattern recognition abilities with the enormous storage capacity and computational power of today's computers.

The basic idea of data visualization is to present data in some visual human understandable form. Visualization is important especially when the exploration objectives are still uncertain at the beginning of the data analysis process.

According to the Visual lnformation-Seeking Mantra, visual data exploration usually follows a three step process: "overview first, zoom and filter, and then details-on-demand" [31].

Overview: gains an overview of the entire collection. An overview of the data helps DM experts to identify interesting patterns. This corresponds to the data understanding phase of CRISP-DM (CRoss Industry Standard Process for Data Mining) methodology.

Zoom: users typically have an interest in some portion of a collection, and they need tools to enable them to control the zoom focus and the zoom factor [31]. For this reason zooming is one of the basic interaction techniques of information visualizations. Since the maximum amount of information can be limited by the resolution and color depth of a display, zooming is a crucial technique to overcome this limitation [32].

Filter: filtering is used to limit the amount of displayed information through filter criteria [32] and allows users to control the contents of the display. As a consequence, users can eliminate unwanted items and quickly focus on their interests [31].

Details-on-demand: as described in [33], "the details-on-demand technique provides additional information on a point-by-point basis, without requiring a change of view. This can be useful for relating the detailed information to the rest of the data set or for quickly solving particular tasks, such as identifying a specific data element amongst many, or relating attributes of two or more data points. Providing these details by a simple action, such as a mouse-over or selection (the *on-demand* feature) allows this information to be revealed without changing the representational context in which the data artifact is situated" [33].

This chapter is divided into four sections. The first section describes the classification of visualization techniques based on Daniel Keim's article [2]. The second section presents the basic graphs used in statistics mainly for exploratory analysis. The third section shows the visualization of trees and rules, which is useful for visualizing association rules and decision trees in the Machine Learning domain. The last section shows basic types of multidimensional graphs. Most of these visualization techniques are tested on the following data sets: the Iris Flower data set, which can be found at the UCI machine learning repository [34]; the total number of recorded tornadoes in 2000, arranged alphabetically by State and including the District of Columbia [35]; the Titanic data set [36].

2.1 Taxonomy of visualization techniques

The basic well known techniques are x-y plots, line plots and histograms. These techniques are useful for relatively small and low-dimensional data sets. In the last two decades, a large number of novel information visualization techniques have been developed, allowing visualization of extensive and even multidimensional data. These techniques can be classified based on three criteria [2]: the type of data to be visualized, the visualization technique and the interaction and distortion technique used, see Fig. 2.1.

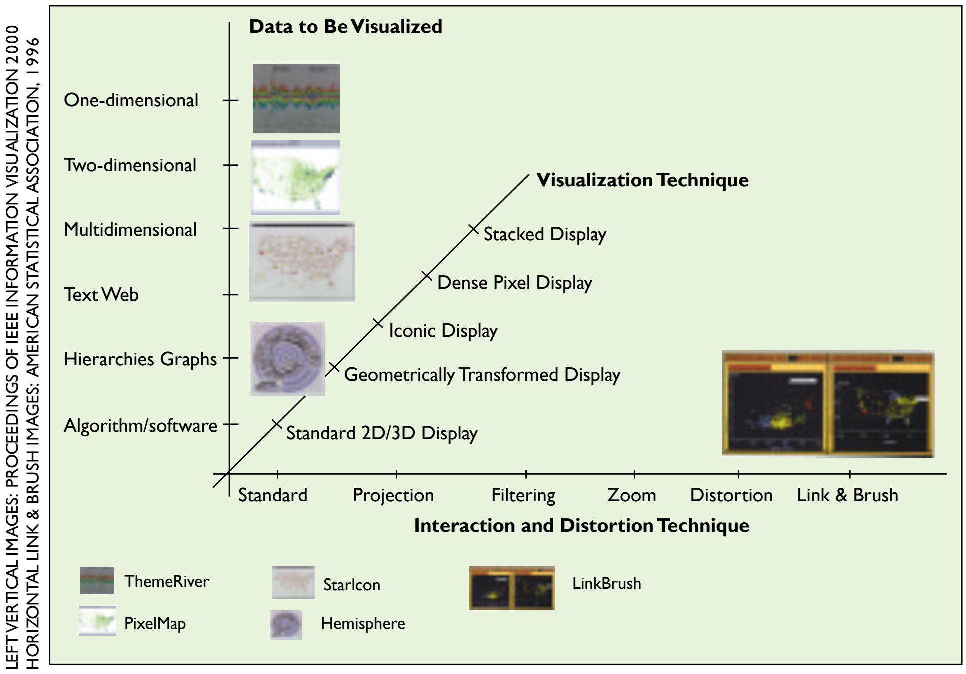


Figure 2.1: Classification of information visualization techniques (image taken from [2])

2.1.1 Data types to be visualized

The data usually consists of a large number of records each consisting of a list of values denominated data mining attributes or visualization dimensions. We call the number of variables the dimensionality of the data set. Data sets may be one-dimensional, such as temporal data; two-dimensional, such as geographical maps; multidimensional, such as tables from relational database. In this thesis we will focus mainly on these data types. Other data types are text/hypertext and hierarchies/graphs. The text data type cannot be easily described directly by numbers and therefore text has to be firstly transformed into describing vectors, for example word counts. Graphs types are widely used to represent relations between data, not only data alone. The last types are about algorithms and software, and their goal is to support software development by highlighting the structure of a program, the flow of information, etc.

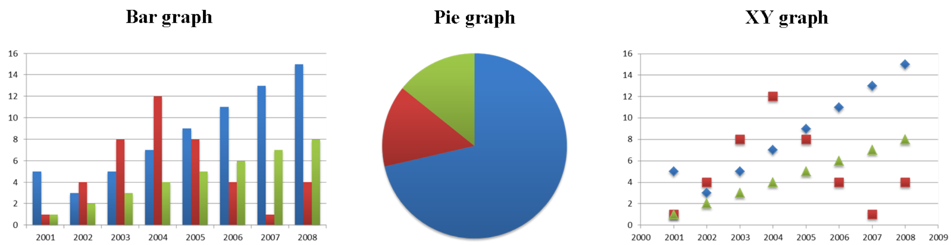


Figure 2.2: Illustration of Bar graph, Pie graph, X-Y plots

2.1.2 Visualization Techniques

The standard 2D/3D techniques

Fig. 2.2 shows some of the basic standard 2D/3D techniques: bar graph, pie graph, $x - y$ or $x - y - z$ plots and other business diagrams. A bar graph is usually used to compare the values in several time intervals, for example the progress of traffic accidents for the last 10 years. A pie graph is ideal for the visualization of a rate of value in data. The whole pie is 100%. This type of graph is used for example to visualize the rate of some type of product in the market basket. There exists a rectangle variant of this graph. A $x - y$ or $x - y - z$ plots is a common 2D/3D graph, it shows the relation between two/three attributes. This type of graph exists in two common variants, line graph for continuous variable and point graph for discrete variable.

Geometrically-Transformed Displays

The basis of geometric transformations is the mapping of one coordinate system onto another. In particular, geometrically-transformed displays techniques aim at finding transformations of multidimensional data sets, which geometrically project data from a n-dimensional space to a 2D/3D space. The typical representation for this group is the Parallel Coordinates visualization technique [37]. Some of these visualization techniques will be described more precisely in section 2.4.

Iconic Displays

The basic idea of iconic displays techniques is to map the attribute values of a multidimensional data item to the features of an icon. The Chernoff faces [38] (see Fig. 2.3) is one of the most famous application, where data dimensions are mapped to facial features such as the shape of the head and the width of the nose.

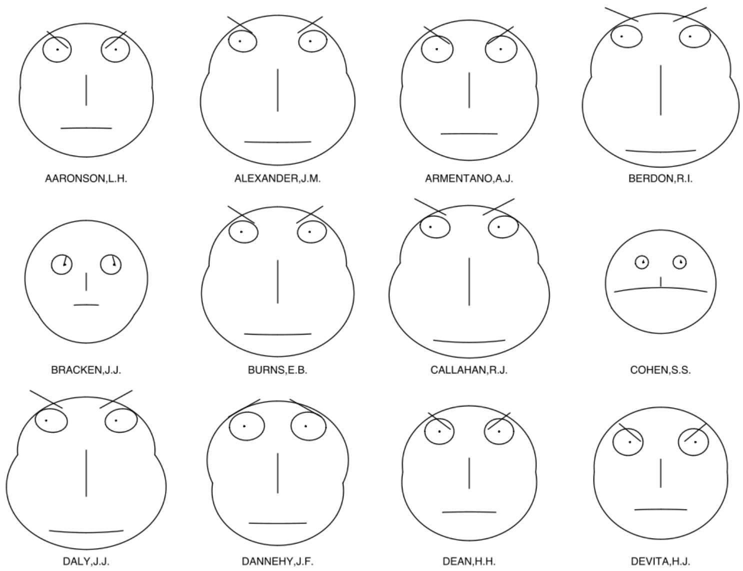


Figure 2.3: Illustration example of Chernoff Faces

Dense Pixel Displays

In dense pixel techniques we map each dimension value to a coloured pixel and group these pixels belonging to each dimension into adjacent areas. These techniques allow to visualize up to about 1.000.000 data values, the largest amount of data possible on current displays, since, in general, dense pixel displays use one pixel per data value. The main question is how to arrange the pixels on the screen. To do this, the recursive pattern and the circle segments techniques [39] are used. An example of the circle segments technique application is illustrated in Fig. 2.4, where each segment shows all the values of one attribute and each circle represents the record.

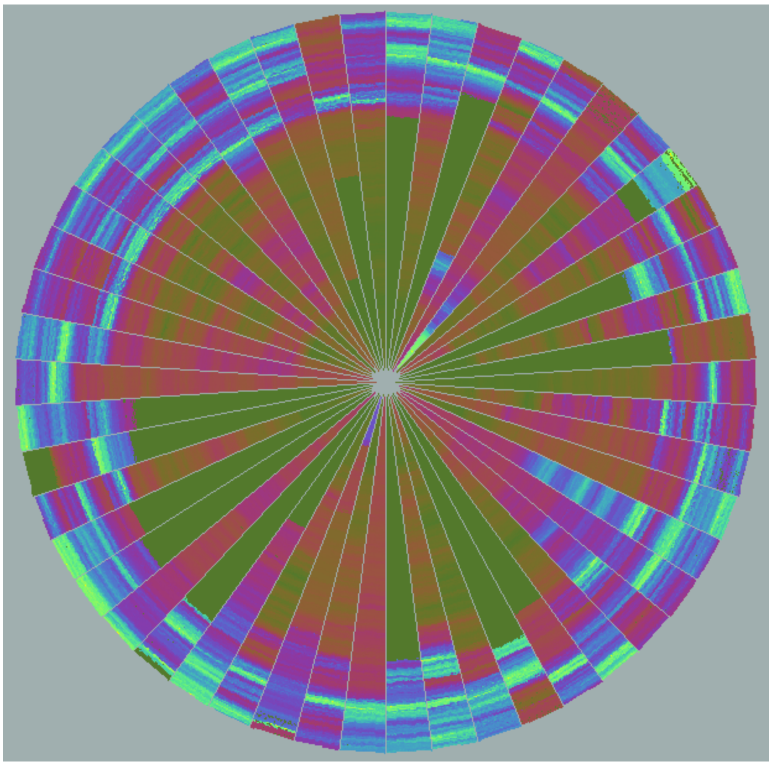


Figure 2.4: Dense Pixel Displays: Circle Segments Technique. (Image taken from [40])

Stacked Displays

Stacked display techniques are designed to present data partitioned in a hierarchical structure. The basic idea is to insert one coordinate system inside another coordinate system, i.e. two attributes form the outer coordinate system, two other attributes are embedded into the outer coordinate system, and so on. A contraindication of this technique is that we have to select the coordinate system carefully since the utility of the resulting visualization largely depends on the data distribution of the outer coordinates. Mosaic Plots [41] and Treemap [42] (see Fig. 2.5) are typical examples of use of these techniques. Mosaic Plots will be described in section 2.3.1.

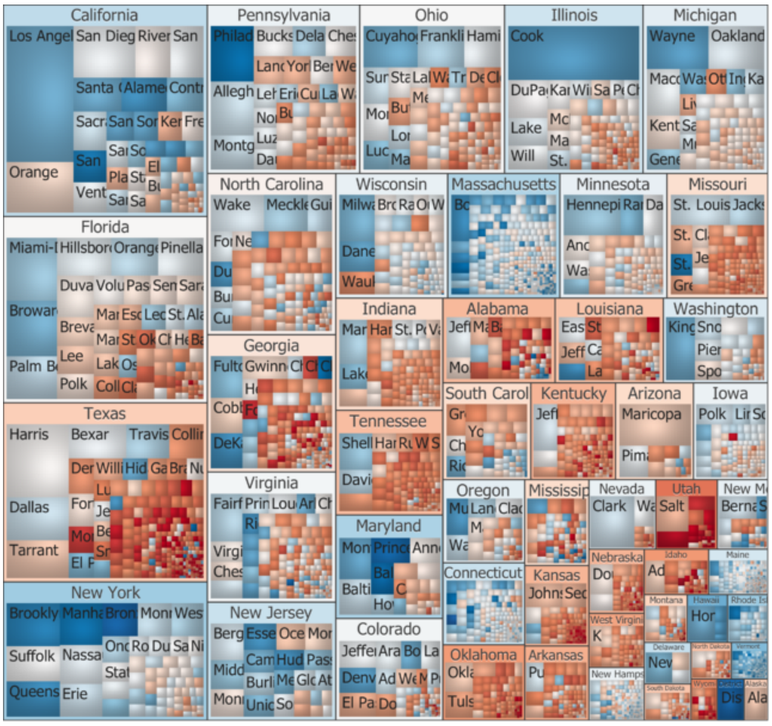


Figure 2.5: Treemap of votes by county, state and locally predominant recipient in the US Presidential Elections of 2012. (Image taken from [43])

2.1.3 Interaction and Distortion Techniques

The interaction and cooperation of different types of graphical techniques are important properties for the data analyst in visualization for an effective data exploration. Interaction techniques enable to directly interact with the visualizations, dynamically change the visualizations and facilitate the combination of multiple independent visualizations. The basic interaction techniques are dynamic projections, interactive filtering and zooming. Distortion techniques help in focusing on details while preserving data overview.

Dynamic Projections

Dynamic projections allows to automatically and dynamically change the projections in order to explore a multidimensional data set. The main difference between dynamic and interactive, where the change is made by user, is the automatic change of the projection. The sequence of projections can be random, manual, precomputed or data driven.

Interactive Filtering

Filtering facilitates the user to focus on interesting subsets of the data set by a direct selection of the subset (browsing) or by a specification of properties of the subset (querying). An example of this method for an interactive filtering is Magic Lenses [44]. The basic idea of Magic Lenses is to use a tool like a magnifying glass to support data filtering directly in the visualization. The data under the glass is processed by the filter and the result is displayed differently than the rest.

Interactive Zooming

Zooming, a well-known technique, allows to present the data in a highly compressed form that provides both overview of the data and in more detailed form depending on different resolutions. Zooming does not only mean enlarging the data object view but it also means that the data representation automatically changes to present more details on higher zoom levels. For example, the objects may be represented as single pixels, as icons or as labelled objects depending on the zoom level: low, intermediate or high resolution.

Interactive Distortion

Interactive distortion techniques assist the data exploration process by preserving an overview of the data during drill-down operations. Unlike the interactive zooming, that shows all the data in one detailed level in a single image, the idea is to show portions of the data with a lower level of detail while other parts are shown with a higher level of detail. Popular distortion techniques are hyperbolic and spherical distortion.

Interactive Linking and Brushing

Since the various techniques previously analysed have some strength and some weaknesses, the idea of linking and brushing is to combine different visualization methods to overcome the individual shortcomings.

Interactive changes made in one visualization are automatically reflected in the other, for example colouring points in one visualization is reflected to other.

2.2 Visualization in Statistics

In this section we will discuss the main graphical methods used in statistics for the exploratory analysis. Usually this analysis is the first step to take when dealing with statistical data. The goal is to find or check, for each attribute, the basic statistical characteristics such as: density, "gaps" of data, or outliers (i.e., observations that do not fit the general nature of the data), etc. To exemplify statistics graphical methods we will use the total number of tornadoes recorded in the USA grouped by State from 2000 [35], as shown in Table 2.1.

2.2.1 Dispersal graphs

Dispersal graphs (also called dot plots) are an example of one kind of graphical summary. Their construction is very simple, we draw only the data values along an axis through graphical elements such as points, thus representing the position of each data value compared to all the others. Figure 2.6 shows a dispersion graph of the tornado data from Table 2.1. The x-axis keep track of the number of tornadoes and each point corresponds to a State.

Figure 2.6: A dispersal graph of the tornado data from Table 2.1. (Image taken from [45])

2.2.2 Histograms

Histograms are another kind of graph for the evaluation of the probability density of an attribute. They are a variant of the bar graph and are produced by splitting the data range into bins of equal size.

State	**Number of Tornadoes**	**State**	**Number of Tornadoes**
Alabama	44	Montana	10
Alaska	0	Nebraska	60
Arizona	0	Nevada	2
Arkansas	37	New Hampshire	0
California	9	New Jersey	0
Colorado	60	New Mexico	5
Connecticut	1	New York	5
Delaware	0	North Carolina	23
District of Columbia	0	North Dakota	28
Florida	77	Ohio	25
Georgia	28	Oklahoma	44
Hawaii	0	Oregon	3
Idaho	13	Pennsylvania	5
Illinois	55	Rhode Island	1
Indiana	13	South Carolina	20
Iowa	45	South Dakota	18
Kansas	59	Tennessee	27
Kentucky	23	Texas	147
Louisiana	43	Utah	3
Maine	2	Vermont	0
Maryland	8	Virginia	11
Massachusetts	1	Washington	3
Michigan	4	West Virginia	4
Minnesota	32	Wisconsin	18
Mississippi	27	Wyoming	5
Missouri	28		

Table 2.1: The total number of recorded tornadoes in 2000, arranged alphabetically by State (Source: National Oceanic and Atmospheric Administration's National Climatic Data Center).

They combine data into classes or groups as a manner to generalize the details of a data set while at the same time illustrating the data's overall pattern. If the attribute has a categorical type, the x-axis represents the categories while the y-axis shows their frequencies. The tabulated frequencies are plot as adjacent rectangles, called bins, erected over discrete intervals. The height of a rectangle is equal to the frequency divided by the width of the interval. The area is equal to the frequency of the observations in the interval. The total area of an histogram is equal to the number of data. As with dispersal graphs, histograms can reveal gaps where no data values exist Fig. 2.7 (the 100-119 class).

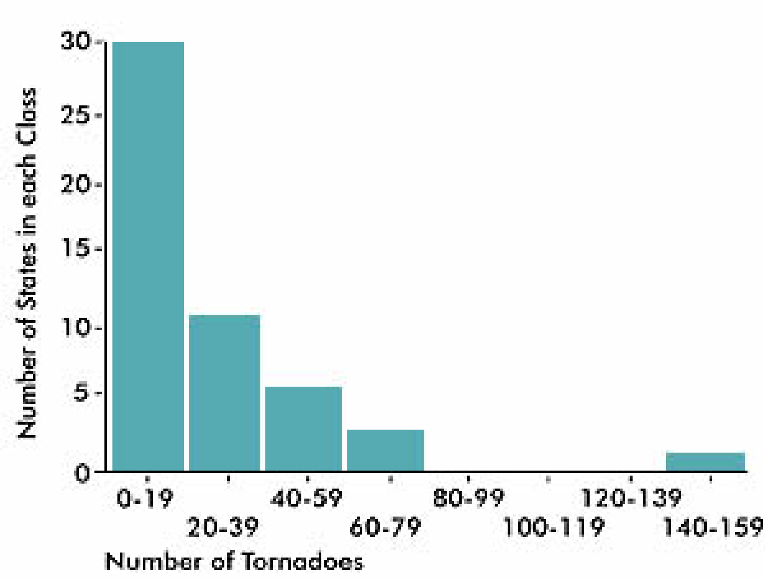


Figure 2.7: A histogram showing the tornado data from Table 2.1. (Image taken from [45])

2.2.3 Box plots

In statistics the box plots (Figures 2.8, 2.9), also called box-and-whisker plot, are a graphical representation used to describe the distribution of a sample, through simple measures of dispersion and position, and information about outliers usually plotted as individual points. Box plots are also an excellent tool for detecting and illustrating location and variation changes between different groups of data. For drawing this graph we need a 5-number summary of the data: $Min\ X, Max\ X, Q_1$ (lower quartile), Q_3 (upper quartile) and m (median).

Let X $= \{x_1, x_2, \cdots, x_n\}$ be a data set. The steps to create a box-and-whisker plot (see Fig. 2.8) are as follows:

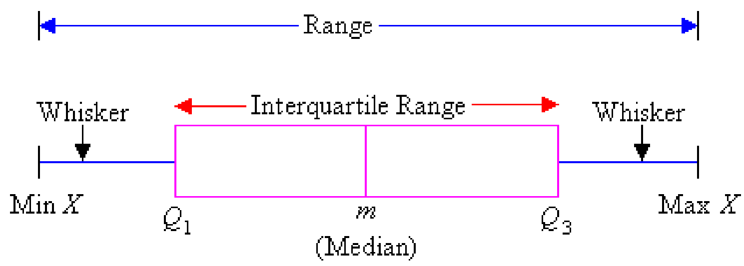


Figure 2.8

1. Arrange the data from the smallest value to the largest value: $x_{p(1)} \leq x_{p(2)} \leq \cdots \leq x_{p(n)}$, with p the corresponding permutation function;

2. Find $Min\ X$: the least value and $Max\ X$: the greatest value;

3. Find the median m of the data (second quartile or Q_2): $Q_2 = x_{p((n+1)/2)}$, if $(n+1)/2$ is a whole number. If it is not, the median is $Q_2 = \frac{x_{p(n/2)} + x_{p(n/2+1)}}{2}$;

4. Find the median of the lower half of the data (lower quartile or Q_1);

5. Find the median of the upper half of the data (upper quartile or Q_3);

6. Draw a number line and mark the points for the values found in steps 2-5 above the line;

7. Draw a box from Q_1 to m and from m to Q_3 marks. The length of the box is important, but the width can be any convenient size;

8. Draw a whisker from Q_3 mark to the maximum value. Then draw a whisker from Q_1 mark to the minimum value;

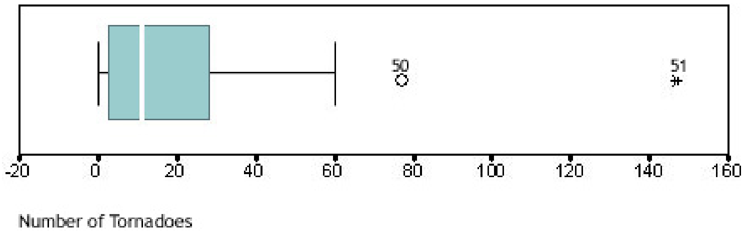


Figure 2.9: A box-plot showing the tornado data from Table 2.1. (Image taken from [45])

2.3 Visualization in the Machine Learning domain

In 1959, Arthur Samuel defined machine learning as the "Field of study that gives computers the ability to learn without being explicitly programmed" [46].

Machine learning is one of the key areas of artificial intelligence and deals with the realization of systems and algorithms that are based on observations as data for the synthesis of new knowledge. Learning can take place by capturing features of interest from sensors, data structures or examples to analyse and evaluate the relationships between the observed variables. One of the main goals of researchers in this field is to automatically learn to recognize complex patterns and make intelligent decisions based on data; the difficulty lies from the fact that the set of all possible behaviours given from all possible inputs is too large to be covered by sets of observed examples (training data). For this reason it is necessary the use of techniques to generalize the cited examples, in order to be able to produce a useful yield for new cases.

Data visualization in machine learning is usually applied to two main activities:

- model description: "*decision tree*" and "*association rules*" techniques;
- model testing: "*learning curve*" and "*Receiver Operating Characteristic (ROC) curve*" techniques.

All images in this section show data from the Titanic data set that can be downloaded from [36]. The survival status of individual passengers on the Titanic is summarized in table 2.2.

Passenger category	Number aboard	Number saved	Number lost	Percentage saved	Percentage lost
Women, Third Class	165	76	89	46%	54%
Women, Second Class	93	80	13	86%	14%
Women, First Class	144	140	4	97%	3%
Women, Crew	23	20	3	87%	13%
Men, Third Class	462	75	387	16%	84%
Men, Second Class	168	14	154	8%	92%
Men, First Class	175	57	118	33%	67%
Men, Crew	885	192	693	22%	78%
Children, Third Class	79	27	52	34%	66%
Children, Second Class	24	24	0	100%	0%
Children, First Class	6	5	1	83.4%	16.6%
Total	**2224**	**710**	**1514**	**0,32**	**0,68**

Table 2.2: The contingency table of Titanic data set

2.3.1 Visualization in model description

Decision trees

In machine learning a decision tree is a predictive model, where each internal node denotes a test on an attribute; each arc to a child node represents a possible value for that property (the outcome of a test); each leaf (or terminal node) represents the predicted value of the target variable given the values of the input variables that in the tree are represented by the path from the root to the leaf.

Normally a decision tree is constructed using learning techniques to the data from the initial (data set), which can be divided into two subsets: the "training set" on which the structure of the tree is created and the "test set" that is used to test the accuracy of the predictive model created. Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value.

Instances are classified through the navigation of the nodes from the root of the tree to a leaf, along the route according to the outcome of the tests.

We can visualize all types of trees in very simple manner. The first step is to draw the root of the tree, this node forms the layer 0. The second step is to draw all edges from this root to new nodes, which form a new layer. We have to repeat this drawing, until we draw whole tree. The height of the tree is equal to the number of layers. An example of decision tree is shown in Fig. 2.10.

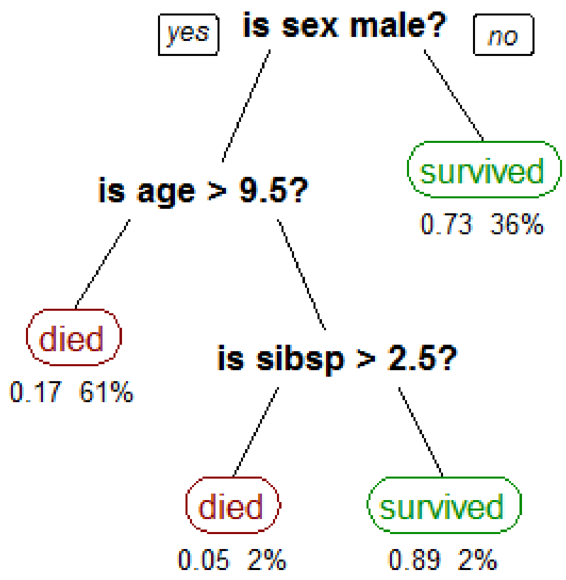


Figure 2.10: A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf. (Image taken from [47])

Visualization of association rules

The visualization of rules is an important step for subgroup discovery. The most used techniques include *Pie charts* and *Mosaic plots*.

Pie charts One of the most common way of visualizing parts of a whole are slices of pie charts. According to [48] a rule can be easily visualized through the use of two pie charts. This approach consists of a two-level pie for each subgroup. The distribution of the entire population, in terms of the property of interest of the entire example set, is represented by the base pie. The inner pie shows the division in some subgroup, which is characteristic for the rule. An example is shown in Fig. 2.11. The base pies show in both cases the relation between survived (dark blue) and non-survived (light blue) adult passengers of the Titanic. The inner pies show two subgroups, in Fig. 2.11.a the inner pie shows only data related to females, in Fig. 2.11.b the inner pie shows only data related to the crew.

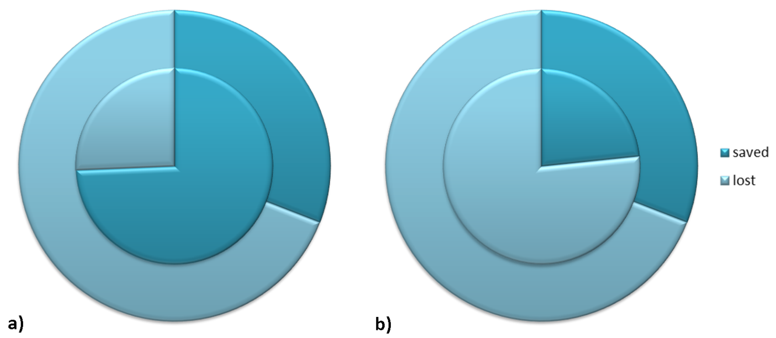


Figure 2.11: The visualization of the Titanic data set comparing the whole group of survived and non-survived adult passengers of the Titanic and the female (a) and crew (b) subgroups

Unluckily the pie graph can visualize only one selected subgroup at a time, so if we need to display multiple subgroups simultaneously we can use the *mosaic plot* tecnique.

Mosaic plots Through a mosaic plot [49, 50] it is possible to examine the relationship between two categorical, nominal or ordinal variables. To build a mosaic plot we must proceed according to the following steps: 1) draw a square with length one; 2) divide the square into horizontal stripes whose widths are proportional to the probabilities associated to the first categorical variable; 3) split each stripe vertically into stripes whose widths are proportional to the conditional probabilities of the second categorical variable. Additional splits can be made if the use of a third, fourth variable, etc. is necessary.

Fig. 2.12 shows the mosaic plot of the Titanic data set for two variables. It is possible to compare the relation between females and males and between survivors and non-survivors in one picture. Fig. 2.13 shows the mosaic plot of Titanic data set for three variables: Class, Gender and Survived.

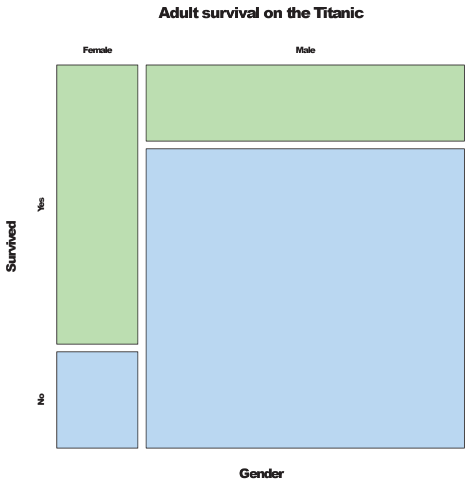


Figure 2.12: Mosaic plot of the Titanic data set for two variables: Gender and Survived

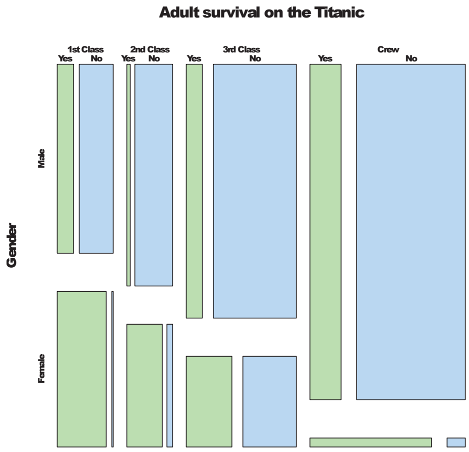


Figure 2.13: Mosaic plot of the Titanic data set for three variables: Class, Gender and Survived

2.3.2 Visualization in model testing

Visualization in model testing is mainly used in the visualization of test results. Among the most used visualization there are the *learning curve* and the *receiver operating characteristic (ROC) curve*.

Learning curve

In computer science, the term learning curve (Fig. 2.14) indicates the ratio of time required for learning (vertical axis) and amount of information correctly learned or experience (horizontal axis) and is particularly used in e-learning and in relation to software. The machine learning curve can be helpful in many aims including: choosing model parameters during design, comparing different algorithms, determining the quantity of data used for training and adjusting optimization to enhance convergence. Usually what we expect from the analysis of the learning curve is that when a machine learning algorithm has more examples for learning, the results will be more accurate.

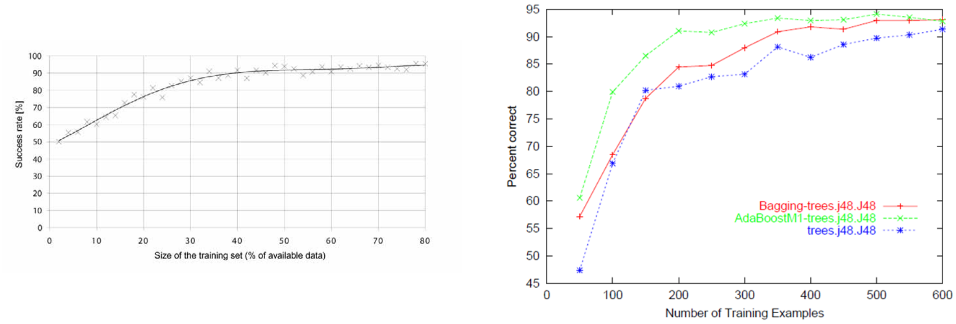


Figure 2.14: Examples of learning curves

ROC curve

The receiver operating characteristic (ROC) curve has long been used in signal detection theory. The machine learning community most often uses the ROC AUC (area under the curve) statistic for model comparison in order to select the best model.

The first step is to define an experiment from **P** positive instances and **N** negative instances for some condition and the relative confusion matrix, see table 2.3, for each model.

	Actual positives	Actual negatives
Predicted positives	True positive TP	False positive FP
Predicted negatives	False negative FN	False positive TN

Table 2.3: A confusion matrix

Where $\mathbf{P} = TP + FN$ and $\mathbf{N} = TN + FP$

The second step is to draw an ROC curve where only the true positive rate (TPR) and false positive rate (FPR) are needed, with: $TPR = \frac{TP}{P} = \frac{TP}{TP+FN}$ and $FPR = \frac{FP}{N} = \frac{FP}{TN+FP}$.

An ROC space is defined by FPR and TPR as x and y axes respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Since TPR is equivalent to sensitivity and FPR is equal to $1 - specifity$ (with $specifity = SPC = \frac{TN}{N} = \frac{TN}{TN+FP}$), each prediction result or instance of a confusion matrix represents one point in the ROC space, see the Fig. 2.15.

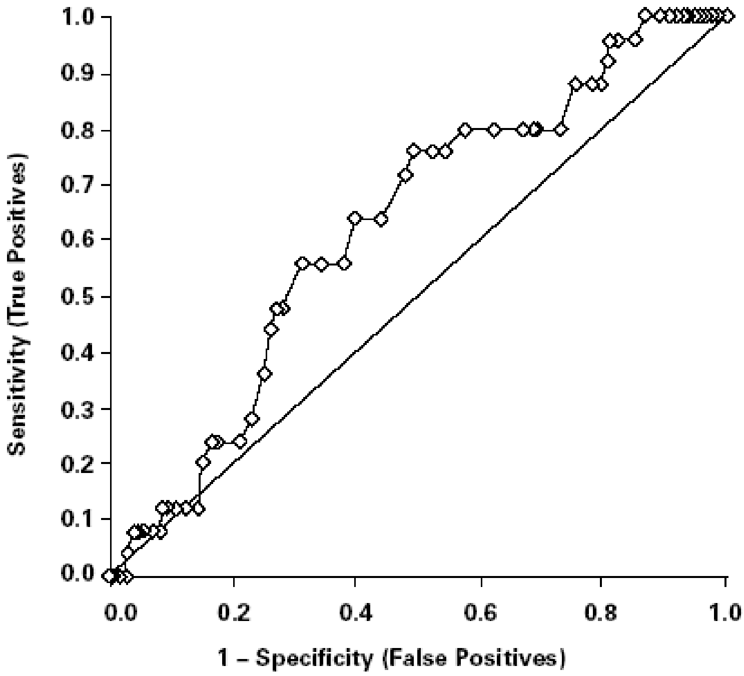


Figure 2.15: Illustration example of ROC curve

The best possible prediction method (perfect classification) would yield a point in the upper left corner or coordinate $(0, 1)$ of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives).

A model is better than another when the area under the curve for the model is larger than the one for the other model.

The diagonal line $y = x$ represents the strategy of randomly guessing a class. The diagonal divides the ROC space. Points above the diagonal represent good classification results (better than random), points below the line represent poor results (worse than random).

2.4 Multidimensional data visualization

In this section, we will examine in detail some of the techniques for the visualization of multidimensional data. We will illustrate the methods that allow the data projection from n-dimensional space to 2 or 3-dimensional space.

2.4.1 Scatter Plot

A scatter plot is useful to illustrate the relation or association between two or three variables. The positions of data points represent the corresponding dimension values. We can display directly two or three dimensions, while additional dimensions can be mapped to the colour, size or shape of the plotting symbol. For example in a 2d scatter plot each point has the value of one variable determining the position on the horizontal x-axis and the value of the other variable determining the position on the vertical y axis. The axis range varies always between the minimum and the maximum of the attribute values.

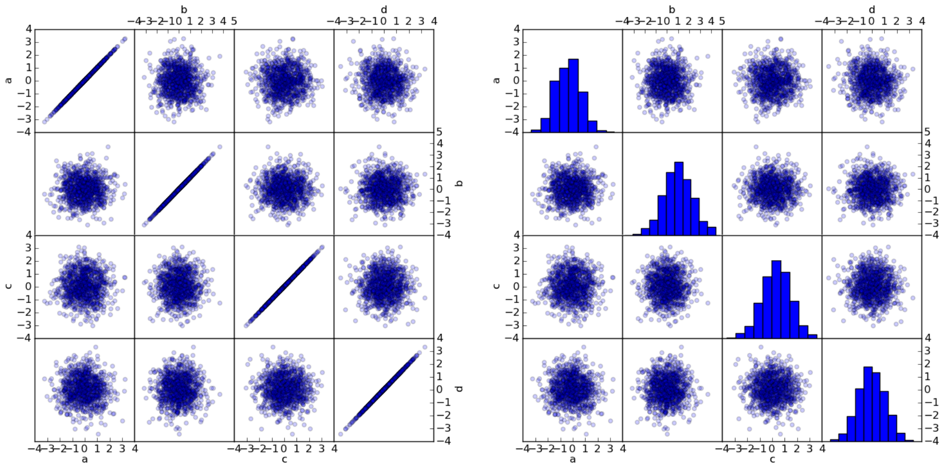
Scatter Plot matrix

Scatter plot matrices are a great way to visualize multivariate data and to roughly determine if there is a linear correlation between multiple variables. Given a dataset with N dimensions, the matrix consists of N^2 scatter plots arranged in N rows and N columns. So, each scatter plot, identified by the row i and column j in the matrix, is based on the dimensions i and j.

Since the diagonal scatter plots use the same dimensions for the x-axis and y-axis, the data points form a straight line of dots in these scatter plots (see Fig. 2.16a) showing the distribution of data points in one dimension, although its usefulness is very limited. It is also possible, to overcome this weakness, to utilize an histogram to obtain a graphical representation of the distribution of data for that specific dimension or attribute as depicted in see Fig. 2.16b.

3D Scatter Plot

3D scatter plots are used to draw data points on three axes in order to show the relationship between three variables. Each row in the data table is represented by a sketch whose position depends on its values in the columns set on the x, y, and z axes; again, additional dimensions can be mapped to the colour, size or shape of the plotting symbol. They are useful when we want to relate two attributes with respect to their temporal trend, Fig. 2.17.



(a) (b)

Figure 2.16: Scatter plot matrices

2.4.2 Multiple Line Graph and Survey plots

Multiple Line Graph is a multidimensional variant of Bar graph. The x-axis typically represents the ordering of records in the data set or time and the y-axis represents the values of attributes, see Fig. 2.18a. The graph's colour may depict the class of record. This method can be useful for finding constant attributes or attributes exhibiting linear behaviour. Survey plot is only a similar variant of Multiple Line Graph, in this case a bar of the record is centered around the x-axis, see Fig. 2.18b. This graph allows to see correlations between any two attributes especially when the data is sorted. The graph colour can again show the class or some other information about the data.

2.4.3 Parallel Coordinates

The parallel coordinate systems (see Fig. 2.19) are commonly used to visualize n-dimensional spaces and analyze multivariate data. To show a set of points in a space of n dimensions, n parallel lines are drawn, usually vertical lines, and placed at equal distance from each other. Vertical lines represent all the considered attributes. A point in n-dimensional space is represented as a poly line with vertices on the parallel axes. The position of the vertex on the i-th axis corresponds to

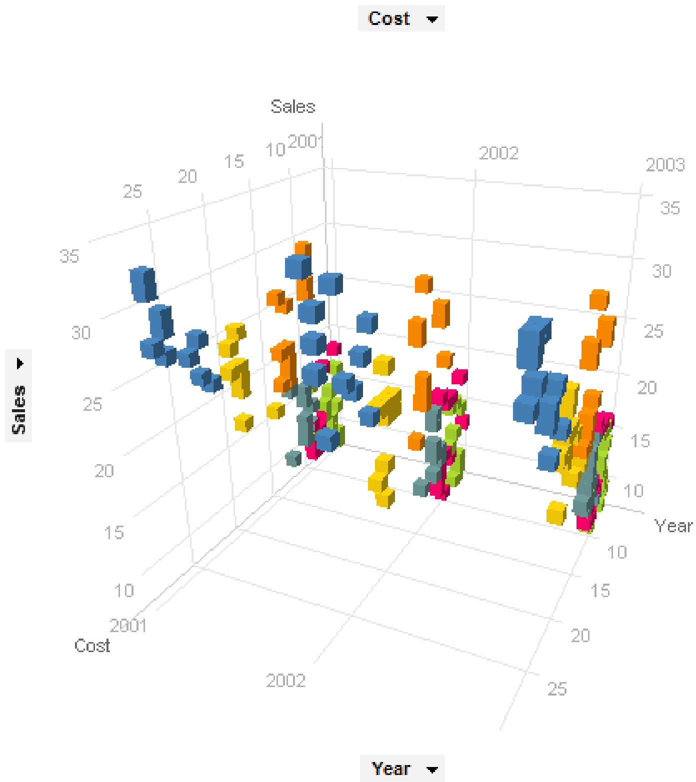
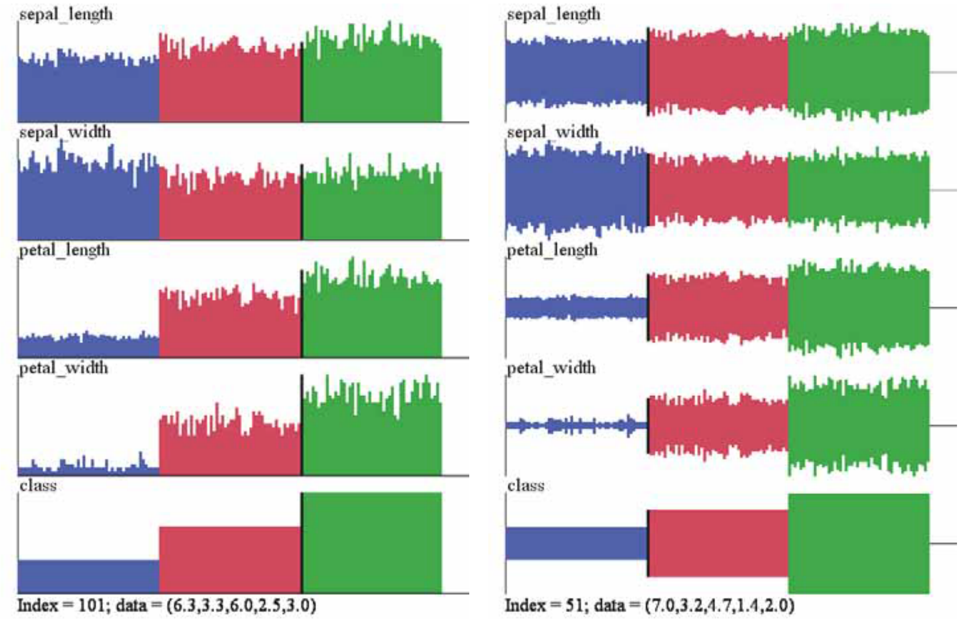


Figure 2.17: A 3D scatter plot where sales, cost, and year are plotted against each other for a number of different products (colored by product)

the i-th coordinate of the point. The maximum and minimum values of each dimension are scaled to the upper and lower point of these vertical lines.

3D Parallel Coordinates

Although parallel coordinates are a powerful method for visualizing multidimensional data they have a weakness: when applied to large data sets, they become cluttered and difficult to read. To overcome these difficulties, researchers have found different solutions. For example Fanea et al. [51] combine parallel coordinates and star glyphs into a 3D representation (see Fig. 2.20a, this technique will be further explained in 4.3.3); Streit et al. [52], as in Fig. 2.20b, assign a color vector, running from dark blue to dark red, to the plot and create an isosurface where



(a) (b)

Figure 2.18: a) Multiple Line Graph and b) Survey plots of Iris Flower data set, created by SumatraTT

the x/y-plane represents the 2D version of the parallel coordinates, while the z-dimension represents the density of the measured data events; Johansson et al. [53] propose a clustered 3D multi-relational parallel coordinates technique (Fig. 2.20c).

2.4.4 Andrews' Curves

In data visualization, an Andrews plot or Andrews curve is a visualization method which plots each N-dimensional point as a curved line using the function:

$$f(t) = \frac{x_1}{\sqrt{2}} + x_2 sin(t) + x_3 cos(t) + x_4 sin(2t) + x_5 cos(2t) + \cdots,$$

where $x \in \{x_1, \ldots, x_n\}$ are values of individual attributes. This function is plotted for $-\pi < t < \pi$, see Fig. 2.21. Thus each data point may be viewed as a line between $-\pi$ and π. This method is similar to a Fourier transformation of a data point. The drawback of this method is given by the high computation time when displaying each n-dimensional point for large data sets.

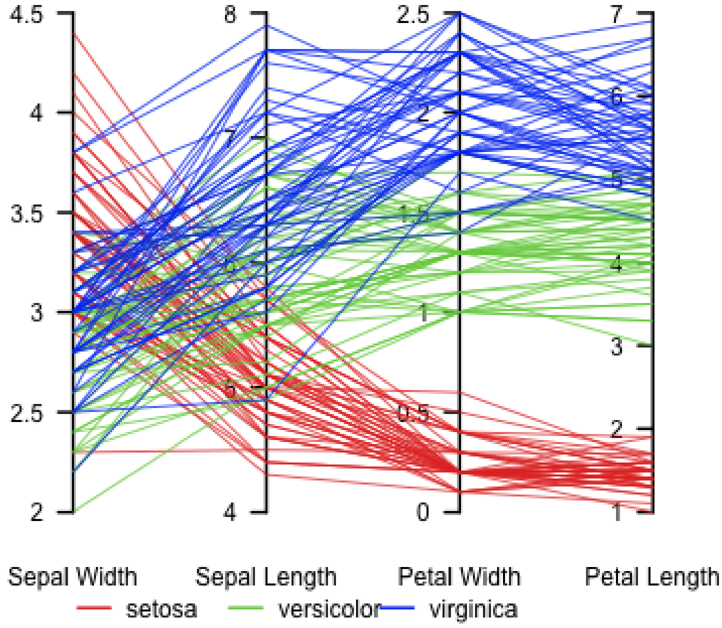


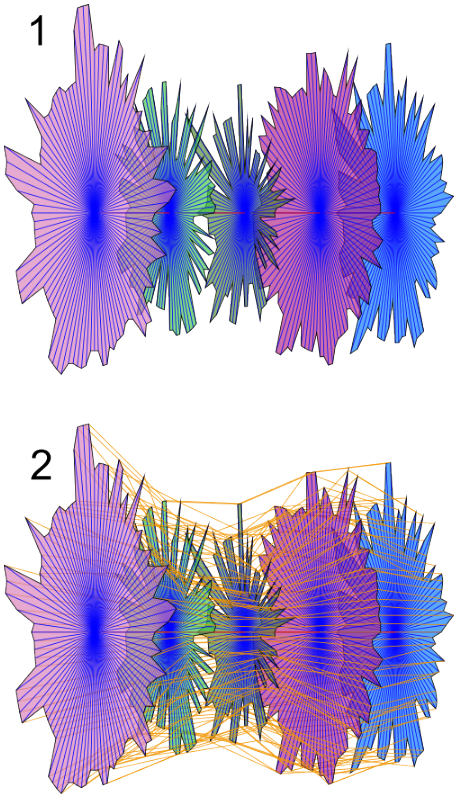
Figure 2.19: Parallel coordinate plot of Fisher Iris data

2.4.5 Using Principal Component Analysis for Visualization

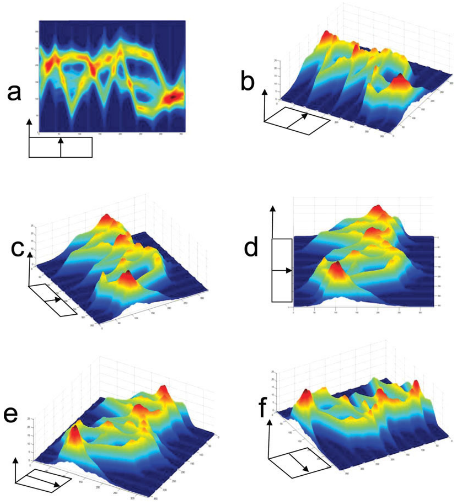
Principal component analysis (PCA) [55, 56], Pearson (1901) and Hotelling (1933), is a quite old statistical technique of exploration analysis but still one of the most used multivariate techniques today. PCA is a way of identifying patterns and expressing the data in a manner to emphasize their similarities and differences. In PCA we use orthogonal transformations to convert a set of observations of possibly correlated variables $x_1, \cdots, x_n$ into a new set $y_1, \cdots, y_m$ (where m is less than or equal to the number of original variables n) of values of linearly uncorrelated variables called principal components. This new set of attributes describes data better. Correlation between attributes is an important precondition of this transformation, if there is no correlation the new set has same size of the original.

Main idea:

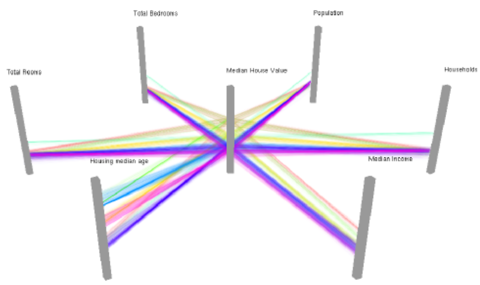
- Start with variables $x_1, \cdots, x_n$
- Find a rotation of these variables, say $y_1, \cdots, y_m$ (called principal



(a)



(b)



(c)

Figure 2.20: Examples of 3D Parallel Coordinates

components) where m is from the interval $[1, n]$, so that:

- $y_1, \cdots, y_m$ are uncorrelated. Idea: They measure different dimensions of the data.
- $Var(y_1) \geq Var(y_2) \geq \cdots \geq Var(y_m)$. Idea: y_1 is more important than y_2, etc.

For example, see Fig. 2.22, we can use the first 2 components to

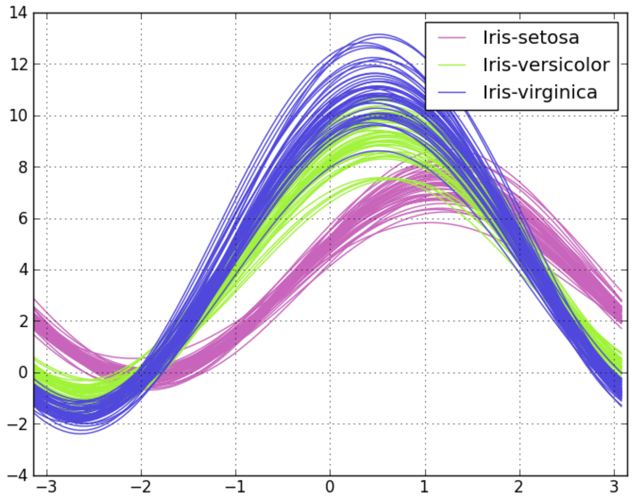


Figure 2.21: Andrews' curves of Iris Flower data set

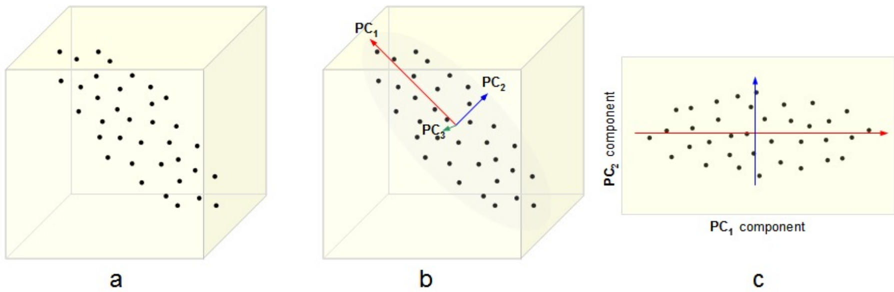


Figure 2.22: An illustration of PCA. **a)** A data set given as 3-dimensional points. **b)** The three orthogonal Principal Components (PCs) for the data, ordered by variance. **c)** The projection of the data set into the first two PCs, discarding the third one. (image taken from [54])

visualize data from 3D to a 2D space.

2.4.6 Kohonen neural network

Kohonen neural network, or self-organizing map (**SOM**), [57] is a type of artificial neural networks with two layers of nodes (or neurons) that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional) and discredited representation of the input space of the training samples. The n dimensions are used as n-input nodes to the net since the input layer has a node for each input. The output nodes of the net are arranged in a hexagonal or rectangular two-dimensional grid. All input nodes are connected to all output nodes, and these connections are associated with weights or intensity. Initially, all weights are random. When a unit wins a record, its weights (and those of other nearby units, collectively called proximity) are adjusted to better match the pattern of predictor values for that record. All input records are displayed and the weights are updated accordingly. This process is repeated several times to obtain extremely small variations. As the training progresses, the weights on the grid units are adjusted in order to form a "map" of the two-dimensional clusters, hence the term self-organizing map. At the end of the learning of the network, records similar should be close on the output map, while the very different record will be at a considerable distance. An example of Kohonen Maps is shown in the Fig. 2.23.

2.4.7 RadViz

RadViz (Radial Coordinate visualization) [59, 60] is a useful method, which uses physical Hooke law for visualization, to map in nonlinear way a set of n-dimensional points onto two dimensional space (see Fig. 2.24a) and to point to existence of natural clusters in some data sets (see Fig. 2.24b).

We need to place n points around a circle in the plain to draw a RadViz plot, this points are called dimensional anchors. In RadViz the relational value of a data record is represented through spring constants, where each line is associated with one attribute. One end of each spring is attached to each dimensional anchor, the other one to the data point. Each data point is displayed at the point, where the sum of all spring forces equals zero according to Hooke's law ($F = Kx$): i.e. where the force is proportional to the distance x to the anchor point and K is the value of the variable for the data point for each spring. The values of each dimension are normalized to $[0, 1]$ range. Data points with approximately equal or similar dimensional values are plotted closer to the center.

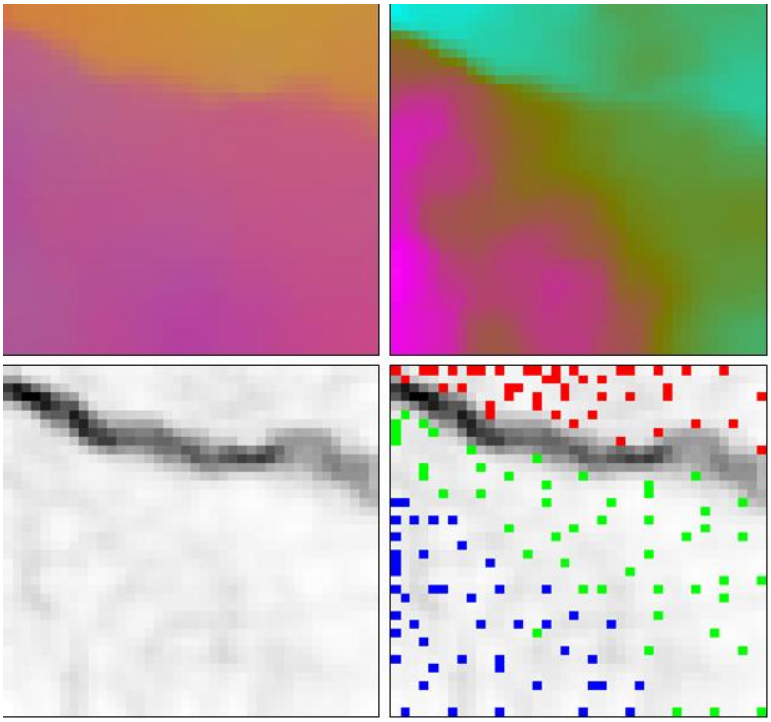
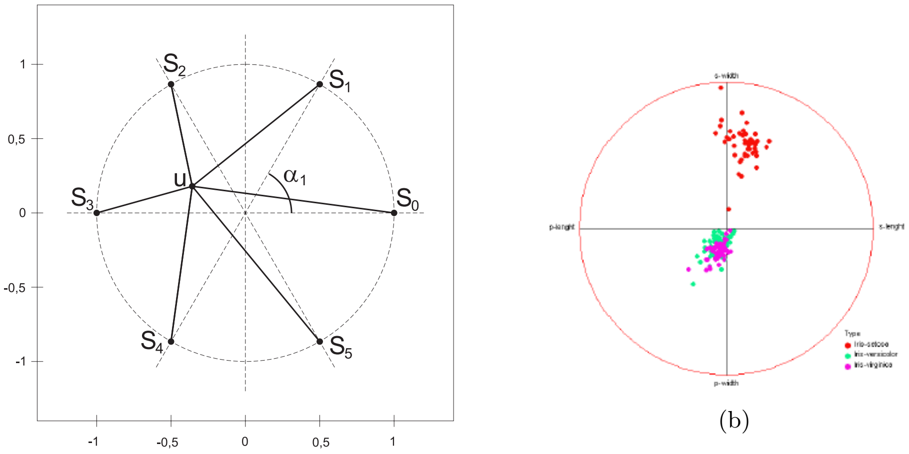


Figure 2.23: Self organizing map of Fisher's Iris flower data set: Neurons (40 x 40 square grid) are trained for 250 iterations with a learning rate of 0.1 using the normalized Iris flower data set which has four dimensional data vectors. A colour image formed by the first three dimensions of the four dimensional SOM weight vectors (top left), the pseudo-colour image of the magnitude of the SOM weight vectors (top right), the U-Matrix (Euclidean distance between weight vectors of neighboring cells) of the SOM (bottom left) and the overlay of data points (red: I. setosa, green: I. versicolor and blue: I. verginica) on the U-Matrix based on the minimum Euclidean distance between data vectors and SOM weight vectors (bottom right). (image taken from [58])

2.4.8 Star Coordinates

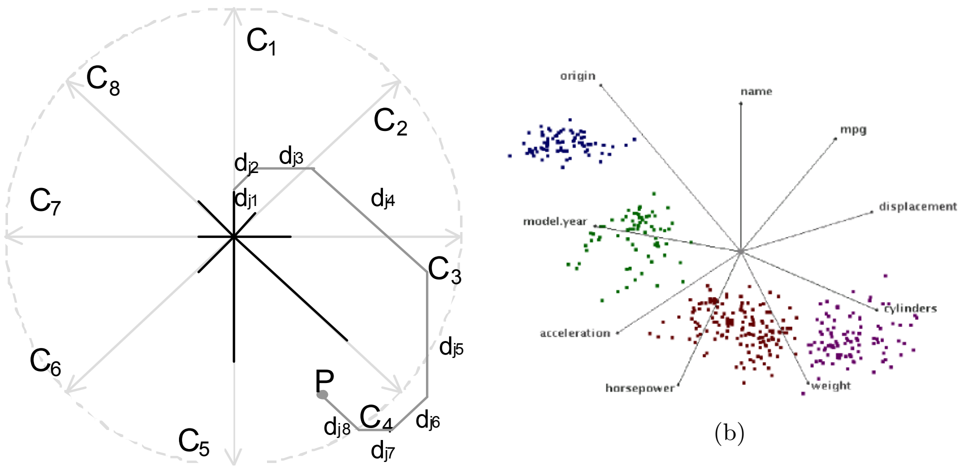
Star coordinates [61] are a multi-dimensional visualization technique with uniform treatment of dimensions. Suppose to have eight attributes: to create this plot (Fig. 2.25a) we have to arrange the coordinate axes, one for each attribute labelled C_1 through C_8, on a circle on a two-dimensional plane with equal (initially) angles between the axes with an origin at the center of the circle. Initially the length of all axes is the same. The d-labelled vectors are computed for each attribute



(a) (b)

Figure 2.24: RadViz: **a)** calculation of data point for an 6-dimensional data set, **b)** visualization of Iris Flower data set

based on the orientation of the axes and the mapping of values. Data points are scaled to the length of the axis, where the smallest value in the set for an attribute is mapped to the origin and the largest is mapped to the end of the attributes axis. In this way, no data values are "out of range". By computing the sum of the d vectors, star coordinates determines the location of star P. An example is shown in Fig. 2.25b



(a) (b)

Figure 2.25: Star coordinates: **a)** Calculation of data point for an 8-dimensional data set [61], **b)** Two-dimensional star coordinates performing cluster analysis on car specification, revealing four clusters [62]

Star Coordinates into three dimensions

Three-dimensional star coordinates [62] extends Kandogan's intuitive and easy-to-use transformations. In this new approach the axes may extend into three cartesian dimensions instead of just two. Now it is possible to use three-dimensional interactions and to maintain the two-dimensional display, i.e. the system rotation is introduced as a powerful new transformation. The users, see Fig. 2.26 have more space to exploit because stars are distributed in a volume instead of a plane, while the addition of the depth allow users to include more meaningful variables simultaneously in an analysis.

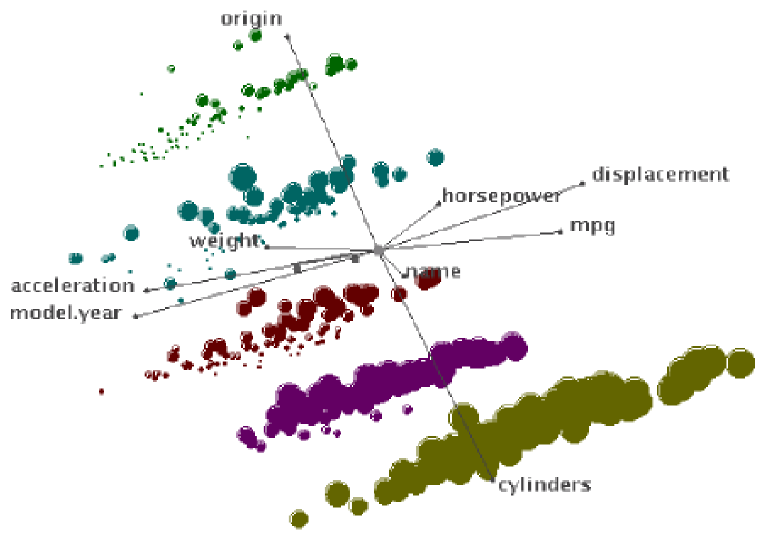


Figure 2.26: Three-dimensional star coordinates performing cluster analysis on car specification, revealing five clusters. (image taken from [62])

2.4.9 2D and 3D Radar Chart

A 2D radar chart (also known as Kiviat diagram) is a two-dimensional graphical representation of multivariate data with one spoke for each variable, and a line is drawn connecting the data values for each spoke. Radar charts are useful to find the dominant variables for a given observation, to discover which observations are most similar, to point out outliers.

3D radar charts extend 2D radar charts in order to better visualize time-dependent data. Some examples include: Kiviat Tube, 3D Parallel coordinates, Wakame (see Fig. 2.27), etc. A full survey of these techniques is proposed in section 4.3.

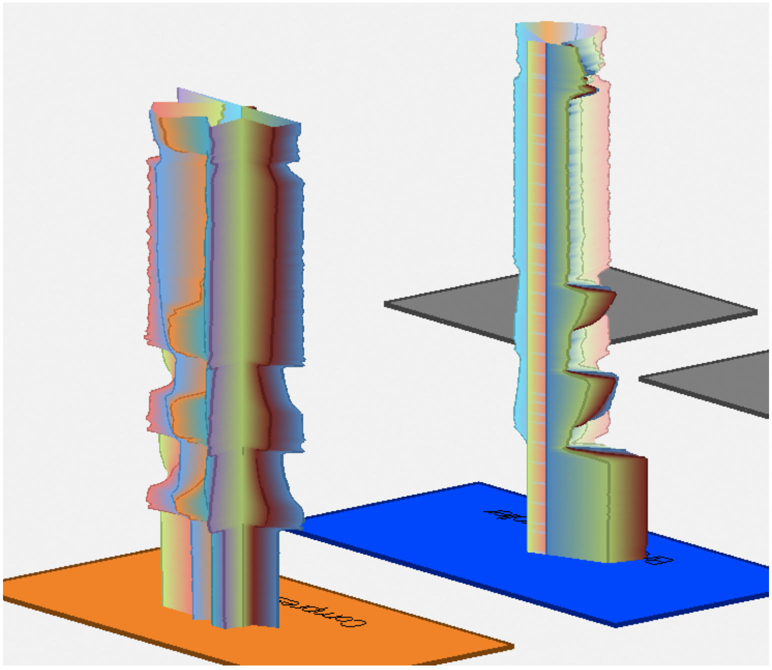


Figure 2.27: A Wakame (image taken from [63])

Chapter 3

CyBiS 2

> There was much else that Jesus did; if it were written down in detail, I do not suppose the world itself would hold all the books that would be written.
>
> — John 21: 25, *The Holy Bible*

This chapter briefly describes the work on CyBiS2, an enhanced version of the tool CyBiS [64], acronym of Cylindrical Biplot System, which is a new interface for the representation of the salient data of searched documents in a single view. This experience has been of great help to move my first steps in the field of IV. Some of the ideas derived from both the study of CyBiS and its further development were of use both to define the 3DRC technique and to realize the related tool.

3.1 Introduction

Even before the advent of the Web, it was possible to access digital libraries through cumbersome proprietary textual interfaces [65]. Nowadays DLs are equipped with a special search engine (for example: http://ieeexplore.ieee.org/, http://dl.acm.org/) where users can retrieve scientific documents.

Unfortunately, textual views can be considered ineffective from different points of view [66, 67]. For example, it is difficult to straightforwardly present users with search results grouped by specific topics related to their research themes. Other drawbacks are the limited number of shown items, the lack of adequate means for evaluating paper chronology, influence on future work, relevance to given terms of the query. To overcome the above drawbacks Information Visualization (IV) is often used. In literature, several visual interfaces for Bibliographic Visualization Tools (BVTs) have already been proposed and actually there are no standards in this field and there is not even a well established terminology. Many visual tools described in the literature provide different features in order to evaluate if a document meets his/her needs like its relevance with respect to the query terms, influence over the field, "age" and relation with other documents. Nevertheless, it is rare that all of them are shown in a single view at the same time: e.g., some systems use different views to show relevance to terms and citation relations.

The rest of the chapter is organized as follows: the next section contains a brief survey on BVTs; section 3.3 presents the revised design and the improvements of the new interface; some final remarks and a brief discussion on possible future work conclude the chapter.

3.2 Related Work

In [28], the authors identified the main objectives of BVTs. BVTs can also be useful for activities related to paper search: finding authors or topics and tasks that give an overview of the research in an area, such as the study of citation network, chronology or collaboration. Following the model presented in [68], we classify the interfaces of BVTs depending on the way in which they present the results of a search query:

- *analytical*: an individual document is represented by an information item;
- *holistic*: groups of documents sharing the same value of a categorization attribute are represented by information items;
- *hybrid*: aggregated and detailed aspects of a search result are presented simultaneously.

A graph representation is one of the most commonly used by analytical interfaces. For example BIVTECI [66] uses a simple interface based on graphs visualization. Citespace II [69] uses a more complex graph, in which one of the several characteristics is that information related to the change of influence of the paper over time are indicated by the color of the node. In [68] and [67] the authors present simple analytical interfaces based on vector space.

Pure holistic interfaces are not very common, since hybrid interfaces may allow the same kind of analysis and also show the individual documents. A graph-based interface in which each node represents a term is used by the Interactive Concept Map of the CiteWiz system [28]. Another example of a pure holistic interface is presented in the Periscope system [68].

The hybrid interfaces use an attribute suitable for clustering group nodes. For this purpose, they generally utilize geometric figures incorporated in other figures. Real world metaphor are often used to represent a group and its articles. Among the different instances one can cite a plant and its flowers in Knowledge Garden [70], a library and its books in the system Docu-World [71], a land and its mountains VxInsight system [72]. An index card metaphor is used by INVISQUE [73] to display library content, organized in a way that visually integrates attributes such citations and publication date. In [74] the body of a butterfly is an article, while the two groups of its references and citers are associated to its left and right wing, respectively. In [75], the authors use a graph view where the articles are represented by small squares grouped in rectangular clusters. In a view of the Periscope system [68], coaxial cylinders located along the vertical axis represent a group of documents from the same domain (org, com, etc.), while the bricks embedded in the cylinders represent individual documents. Bricks of the same color represent documents coming from the same website.

3.3 The revised design

To develop the new design of CyBiS2 I chose to continue on the work undertaken on the previous version, trying to extend and emphasize strengths and improve some aspects by developing new and useful features and operations.

3.3.1 The previous CyBiS

The previous CyBiS approach showed document items as spheres embedded in a 3D cylinder placed horizontally on the screen (see Fig. 3.1). The relevance of a document to terms was given by the height and depth of the item in the cylinder; the horizontal position, instead, represented the document's publication year; the influence of the document was given by the color gradation of the item; citations were shown on-demand and were represented through connections. With this tool, the user was not compelled to switch between different visualizations thus losing the focus on the current context.

Several operations were available in order to refine search. Firstly, the cylinder could rotate on the horizontal axis, in order to bring in the foreground the documents more related to a given term. It was also possible to zoom, which is equivalent to cut the cylinder along its horizontal axis, in order to select a time slice or to filter, for selecting a subset of items relevant to given terms.

3.3.2 The improvements of the new interface

First, improvements have been made to the pre-existing design, aimed at improving the user's perception of the characteristics of a document, particularly its relevance to the term. The revised design has been defined based on the advice of some pilot users, who have been subjected to the vision of mock-ups of different versions of the interface in order to establish the best possible solution.

In the CyBiS approach the relevance of a document with respect to the term is represented through the use of a biplot shown in perspective. In the previous tool CyBiS, the relationship between the vectors representing the term and the keyword was represented by an association of colors. In the new design, the relationship is direct, with the keywords listed in correspondence of the vectors at the base of the cylinder. The procedure for the selection of the most recognized term to display at

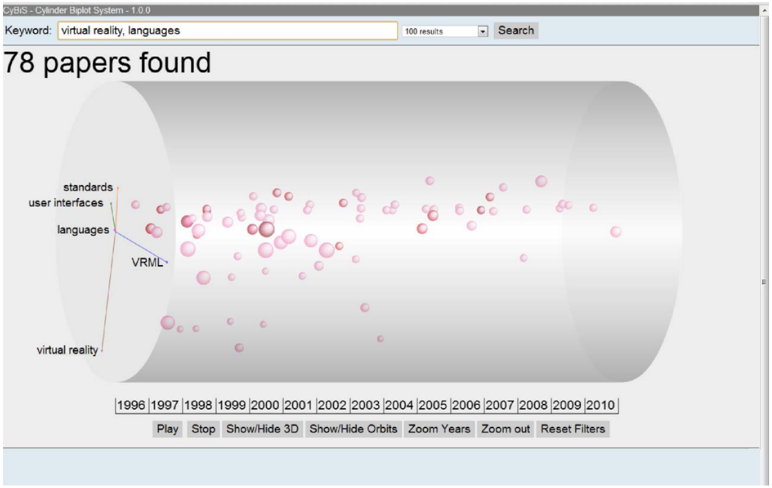


Figure 3.1: The old CyBiS interface.

the base of the cylinder is based on calculating the average frequency of the term within the selected subset and the resulting value TF-IDF (Salton & Buckley, 1988) [76].

Moreover, most of the time the list of the five most significant terms did not contain the search terms entered by the user. Being the items distributed in the cylinder with respect to such terms, this aspect tended to create confusion and frustration to the end user. In CyBiS2 search terms are entered by default, sorted by relevance, on the top of the list of the most significant terms. In any case, since the possibility to display the items with respect to the list of the five most significant terms results to be interesting, the new tool allows to manage both types of research.

Furthermore, the items corresponding to documents were originally represented by spheres. With this design it was difficult to assess the depth, whose representation was left solely to the size of the sphere. Each item document, unlike the previous interface, it is now represented by a mini "post-it", representing a section of the cylinder's surface, located inside the cylinder with the specific coordinates, which assumes different shape and curvature depending on the angle of rotation of the cylinder and its distance from the axis. This choice was dictated by the desire to offer the user a greater understanding of the depth assumed by each item in the space of the cylinder. To emphasize the three-dimensional appearance of the display, the items that represent documents are scaled respect to their relative position in the cylinder: those ones closer to the point of view have larger size, while those more distant are noticeably smaller. The new interface is shown in figure 3.2.

The toolbox in the “View Results” has been modified to simplify and rationalize the content. In CyBiS2 the buttons are “On/Off” and grouped by type.

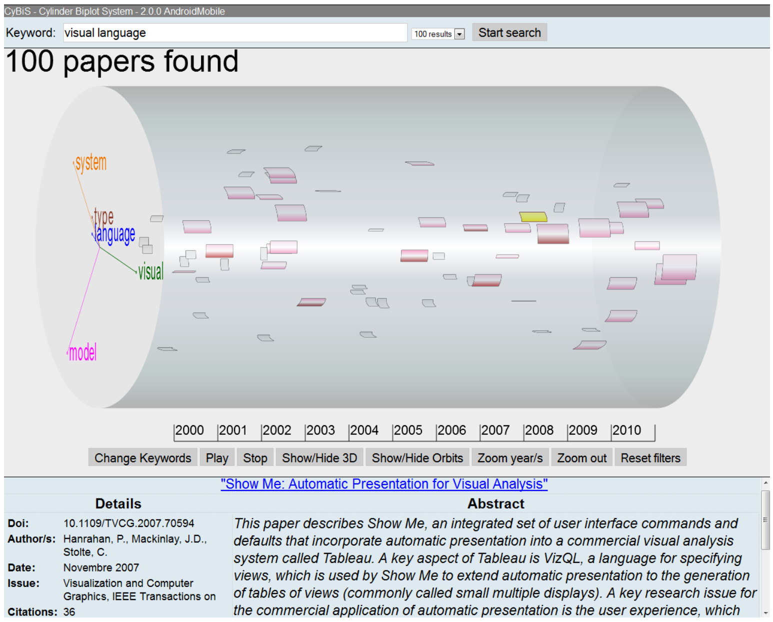


Figure 3.2: The new CyBiS interface initialized with a query composed of the terms *visual language*.

3.3.3 The problem of occlusion

To finally resolve the problem of occlusion (the most important item, being attracted towards the center of the cylinder, tend to be hidden from those who are outward) two new options were added for displaying / navigation. In CyBiS2 it is possible to zoom in and navigate inside the cylinder directly through the use of a pitch-to-zoom function, using touch gestures on a touchscreen, or via the “zoom in” / “zoom out” and use the mouse or the fingers to scroll the area of the cylinder.

In addition, the new interface has been equipped with the new display function “sheet of onion”. The user may decide to group, either manually or automatically, the items in concentric cylinders according to their distance from the center. Once this is done it becomes possible to hide or delete sections of the cylinder to better focus on a particular distribution of the retrieved items.

Due to the "graduated transparency",the system allows the user to set the transparency of the items according to their distance from the center. Those that are located closer to the surface of the cylinder will be more transparent than those that are more attracted toward the center. The system also allows to set the degree of transparency both automatically and manually according to specific user needs.

3.4 Discussion and Future Work

Future work will focus on evaluating the effectiveness and efficiency of the interface through studies focused on users. In addition to relying upon any opinion expressed by users, we aim to obtain quantitative data for evaluating the performance of the interface and compare them with those resulting from the use of other systems.

Quantitative data can be extracted both by the speed with which users complete tasks with each interface and by measures related to the relevance of the tasks completed compared to the result that was expected after their completion.

Chapter 4

3DRC

> The priests, the sons of Levi, will then step forward, these being the men whom Yahweh your God has chosen to serve him and to bless in Yahweh's name, and it being their business to settle all cases of dispute or of violence.
>
> — Deuteronomy 21: 5, *The Holy Bible*

4.1 Introduction

Nowadays the management of projects for public or private organizations is becoming increasingly difficult due to their growth in size and complexity. Statistical analysis [77, 78] highlight how projects fail at an alarming rate because of the difficulty and the ability to achieve the objectives. Among the most frequent causes there is the very high degree of *uncertainty* that affects the expected performance of the project. If the projects become more complex for a number of reasons ranging from the size of the technical complexity, or environmental conflicts or political constraints, then the usual project management methodologies need to be adapted as well as managers must be able to analyze situations from different perspectives and work using a larger number of models. The lack of involvement in the project of the various stakeholders also contributes to increase the uncertainty [79, 80, 81] about their satisfaction degree, especially when the points of view of involved stakeholders are not consistent.

Providing new tools for visualization and analysis would be of great help to reduce project failures. In this context, an aid may come from the use of Information visualization (IV), whose usefulness has been demonstrated in many areas [26, 82, 83]. Graphical representations such as Gantt graphs, network diagrams, S-curve histograms or those considered by Earned Value Analysis [84, 85] to measure project performance in areas such as cost, schedule and scope are very frequently applied in Project Management. However, most of all techniques are bidimensional while 3D representations of project variables and management techniques would increase the analysis capability of project managers and other stakeholders. The goal of this thesis is to introduce a new 3D visualization to represent and analyze the problem of diverging stakeholders views during a project execution.

The chapter is structured as follows: the first section introduces the "pursuit of consonance" as a reducing factor of the project risks together with the some basic definitions about the concept of stakeholders views and gap; the second section summarizes the existing techniques in literature for creating three-dimensional Kiviat diagrams. The new 3DRC approach will be presented in the third section while the fourth shows some operations to handle 3DRC. The next section presents a case study and, finally, we draw the conclusions and discuss future work.

4.2 The Pursuit of Project Stakeholders Consonance

According to the Viable System Approach [86, 87], *consonance* is considered as a potential compatibility between systems, that facilitates their connection. The virtuous interaction (harmonizing) between the two drivers becomes the *resonance*. While consonance concerns structural concepts, resonance is a systemic concept. The consonance represents a situation aiming at reaching an harmony or agreement among two or more systems, but they effectively become resonant when an harmonic interaction between components exists. In [88] the authors discuss how an agreement on a contractual basis for a project realization creates a situation of consonance among the contractors that appears to be the minimum necessary to enable the resonance. The stakeholders will to contract indicates that exists a relationship of trust and harmony that comes from sharing the same design goals. When strategy, mutual commitments, values and goals are shared, a cooperative behavior emerges from the system composed of the stakeholders. If some conflict of interest arises, each stakeholder (and, in general, each system) must identify the best possible way to cooperate, possibly changing its behavior to reach the convergence of individual perspectives. However, it is not always possible to predict in advance and manage the risks of adverse events to the project realization; in addition, due to the conflicting and opportunistic roles that stakeholders hold, positions and different views on one or more project activities often emerge. In these cases, the balance achieved in the degree of initial consonance is disturbed and some corrective actions need to be provided by one or both parties, to try to restore it to the initial levels or at the levels considered as satisfactory for the project continuation. Some negotiation cycles are necessary before reaching an agreement on potential controversies generated during the project. A feedback model formalizing this situation, is discussed in [89, 88] where contract management together with dispute management act as a regulatory instruments to maintain adequate levels of consonance and resonance among systems cooperating for the achievement of a common goal. In the following we provide some definitions [30, 89], on which the 3DRC visualization technique is based.

4.2.1 Definition of Project

A project can be seen as a set of sequential and/or parallel activities. We define a project P as a set of quadruple:

$$P = \{[a_1, c_1, s_1, q_1], [a_2, c_2, s_2, q_2], \cdots, [a_m, c_m, s_m, q_m]\} \tag{4.1}$$

where $a_1, \cdots, a_m$ are the project activities and:

- $c_1, \cdots, c_m$ are the associated planned costs,
- $s_1, \cdots, s_m$ are the scheduled execution times where each $s_i =$

$[t_{start}, t_{end}, state]$ $i = 1, \ldots, m$ with the attribute *state* that, at a given time, can assume one of the values contained in the collection {`started`, `not started`, `in progress`, `suspended`, `completed`}.

- $q_1, \cdots, q_m$ define the quality constraints specified in relation to the output design.

4.2.2 Definition of View

We now introduce the concept of view associated with each contractor at a certain time instant. Assume that two contractors A and B have signed a contract at time t_0. At a generic time $t > t_0$ each contractor $x \in \{A, B\}$ has the following point of view with respect to the execution of the project P:

$View_x(P, t) = \{[a_i, c_i^t, s_i^t, q_i^t]\}\ i = 1, \ldots, m$

where:

- each c_i^t represents the cost assigned by x at time t with respect to the activity a_i
- each $s_i^t = [t_{start}, t, state]$ is the time observed for implementing the activity a_i until t
- each q_i^t indicates the degree of quality observed by x on activity a_i at the observation time.

At time $t_0, View_A = View_B$ because the parties have reached a contractual agreement; therefore, a view represents the "*planned values*" for project variables specified at time t_0. Over time, a view assumes the meaning of "*observed value*" derived from the perception that a contractor has with respect to the execution of the project activities. Since the point of view is a subjective value, it happens frequently that the views of the parts tend to vary over time. The idea of divergence on the values of fundamental project variables is formalized below by the function Δ.

4.2.3 Definition of Gap

Supposing that two observers A and B observe an activity a_i at a generic time instant t_k, the *gap* between the two views is represented by the Δ_i function as follows:

$$\Delta_i(View_A, View_B, t_k) = [a_i, \Delta_i^c, \Delta_i^s, \Delta_i^q] \tag{4.2}$$

where $\Delta_i^c, \Delta_i^s, \Delta_i^q$ are the projections of the distance between the views $View_A, View_B$ on the dimensions of cost, time and quality of the activity a_i, respectively.

If the observers are interested in a subset of activities $X = \{a_1, a_2, \ldots, a_l\}$, $l <= m$

$$\Delta_X(View_A, View_B, t_k) = \{[a_j, \Delta_j^c, \Delta_j^s, \Delta_j^q]\}\ j = 1, \ldots, l \tag{4.3}$$

if $k = m$, A and B observe the whole project. More details about the computation of Δ on the three dimensions can be found in [30, 88].

The *total gap* between two views of the subset X according a dimension d (time, cost or quality) is calculated as the sum of the gap on all activities belonging to X:

$$T\Delta_X^d(t_k) = \sum_{j=1}^{l} \Delta_j^d. \tag{4.4}$$

As above, if X contains all the activities of the project P, $T\Delta$ represents the total gap of the P.

The analysis of the $T\Delta_d$ over time also gives us an idea of how the project consonance level between A and B evolves with respect to the considered dimension. If the views diverge at the time of the observation, the consonance between contractors tends to decrease. When the points of view are irreconcilable, a dispute arises slowing down or possibly stopping the project. It may then be necessary to embark on a series of informal negotiations between the contractors to re-establish an acceptable level of consonance or, in the case of irreconcilable disagreement between the parties, the initiation of a formal process of dispute resolution governed by a mediator.

4.3 Background

A radar chart (also known as Kiviat diagram) is a two-dimensional graphical representation of multivariate data with one spoke for each variable. The sequence of spokes, or radii, is equi-angular and starts from the same central point which represents zero. The angle of the axes and relative position is typically uninformative. A radar chart requires at least three categories; the length of each spoke is proportional to the size of the variable relative to the maximum size of the variable across all data points. The data values for each spoke are connected by a line giving the plot a star-like appearance. Each star represents a single observation. A radar chart represents multiple observations through multi-plot format with many stars on each page and each star representing one observation. This plot is useful to answer to the following questions:

- Which observations are most similar?
- What variables are dominant for a given observation?
- Are there outliers?

In the literature several researches on 3D radar-chart construction can be found.

4.3.1 Yuhua and Kerren 3D Kiviat diagram

Among the most interesting we cite Yuhua and Kerren [90, 91] who proposed a 3D Kiviat diagram in which the same axes are folded up and down in 3D (Figure 4.1), with the result of allowing the observer to better perceive the data-dependent time.

4.3.2 Kiviat tubes

The Kiviat tubes, as suggested by Hackstadt and Malony in [92] (Figure 4.2a), are achieved by rendering the surrounding surface obtained by displaying Kiviat diagrams placed along a time axis. In the first version the surface connecting all the Kiviat diagrams emphasizes the shape of the tube but hides the information about individual data elements. This

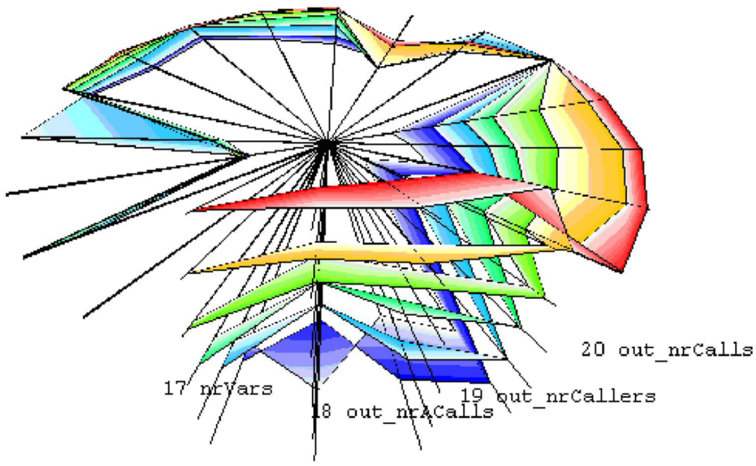


Figure 4.1: a 3d radar-chart according to Yuhua (image taken from [90])

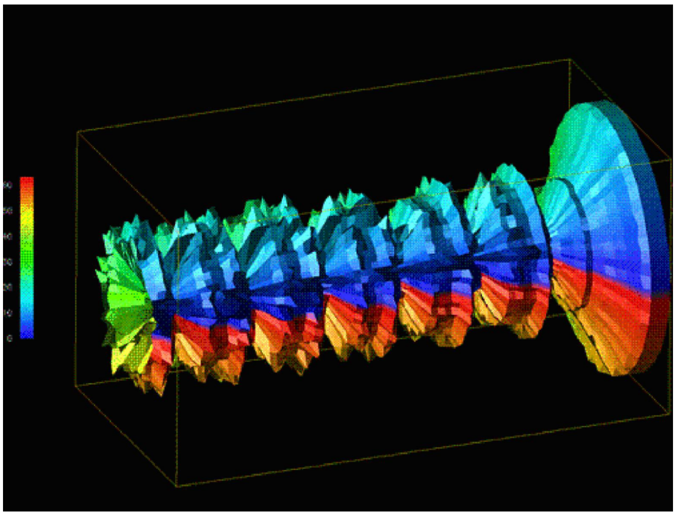
inconvenience has been solved thanks to transparencies and composition techniques that allow an improved display which lets scroll a slice of a 2D Kiviat diagram through a partially transparent Kiviat tube.

4.3.3 3D Parallel coordinates

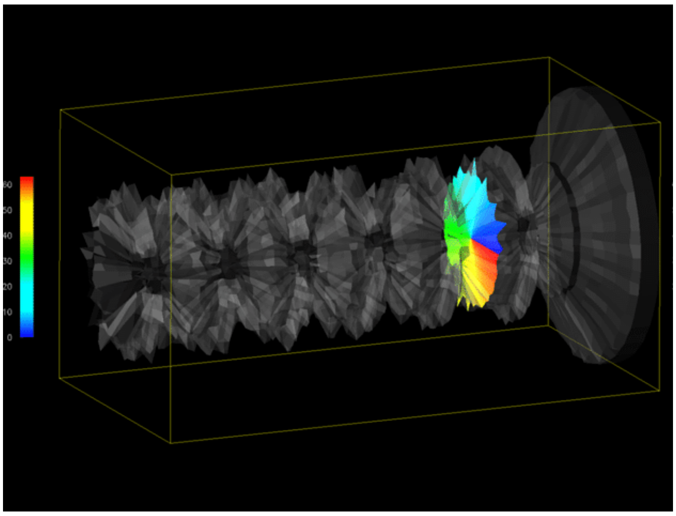
Fanea et al. [51] extended the 2D Parallel coordinates in 3D Parallel coordinates, by combining multiple views of 2D parallel coordinates in a single view, as shown in Figure 4.3. This was achieved by placing the matrix diagrams of 2D Kiviat in a cylinder aligned along the time axis and connecting the vertices with polylines. The result of this approach is similar to 3D Kiviat tubes; unfortunately, the amount of data sets that can be displayed is still limited because when too much data must be handled, their recognition and comparison can be very difficult.

4.3.4 3D Kiviat tube

Tominski et al. [93] presented some approaches to achieve an interactive visualization based on the axes that can be used to explore and analyze multivariate time series data. The main idea is to arrange the axes to form a circle around a commonly shared time axis and render, for each time step, a surface representing a Kiviat diagram, see Figure 4.4.



(a)



(b)

Figure 4.2: A Kiviat tube in the first and second version (image taken from [92])

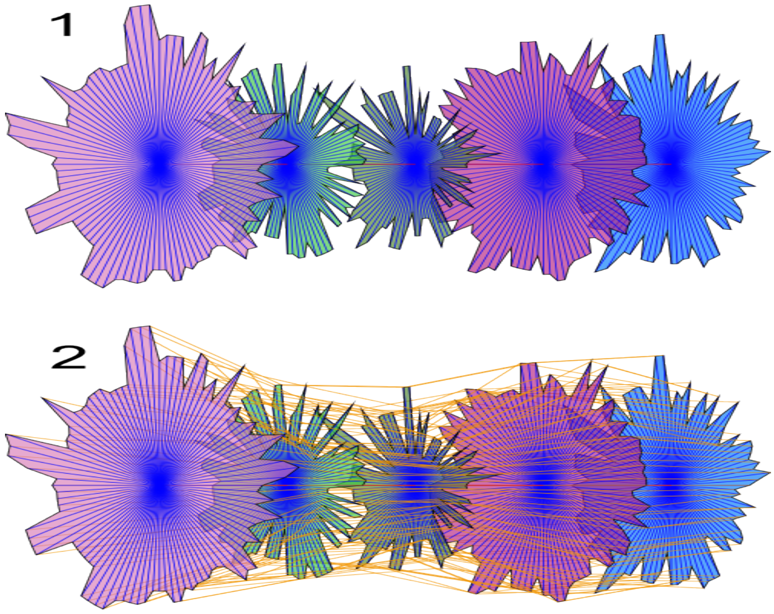


Figure 4.3: 3D integration of parallel coordinates and Star Glyphs (image taken from [51])

4.3.5 Wakame

The Wakame [63] (as in Figure 4.5) are obtained with a very similar approach to those used by Hackstadt and Malony or Tominski et al. for the construction of their Kiviat tubes. A traditional multidimensional radar chart (Figure 4.5.a)) is drawn on the bottom using a 3D representation (Figure 4.5.b)). As with other extrusion techniques, the time is mapped along the y axis, and the various sequential measures become radar-chart drawn at different heights above the floor from the bottom upwards (Figure 4.5.c)). These drawings are used to create "hollow tubes" whose shape illustrates the changes in data over time (Figure 4.5.d)). Compared to the original implementation (Figure 4.5.d)) where each Wakame is colored with a unique color, the new prototype assigns a different color to each of the vertices that correspond to the different sizes (Figure 4.5.f)). With this coloration, both the overall shape and shape changes inside each dimension are more easily comprehensible.

Furthermore, a Wakame also allows the comparison of an heterogeneous data collection emphasizing the possible correlations.

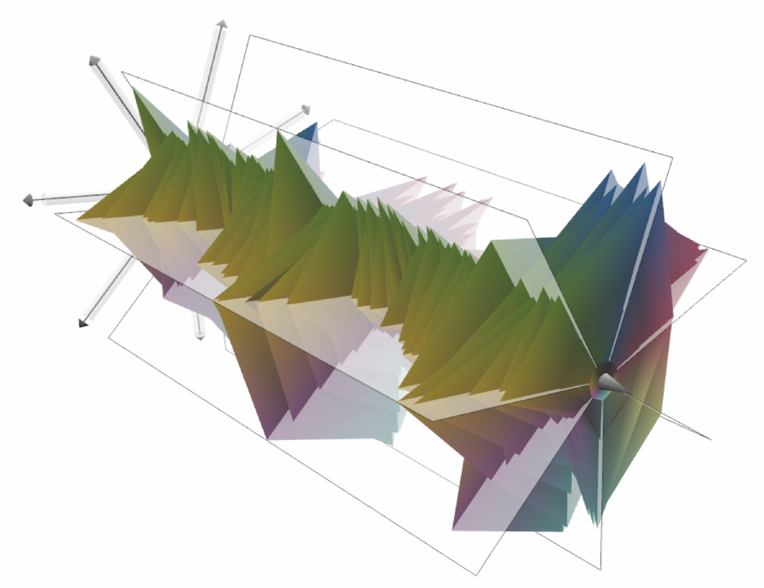


Figure 4.4: A 3D Kiviat tube displays seven attributes respect to the central axis of time (image taken from [93])

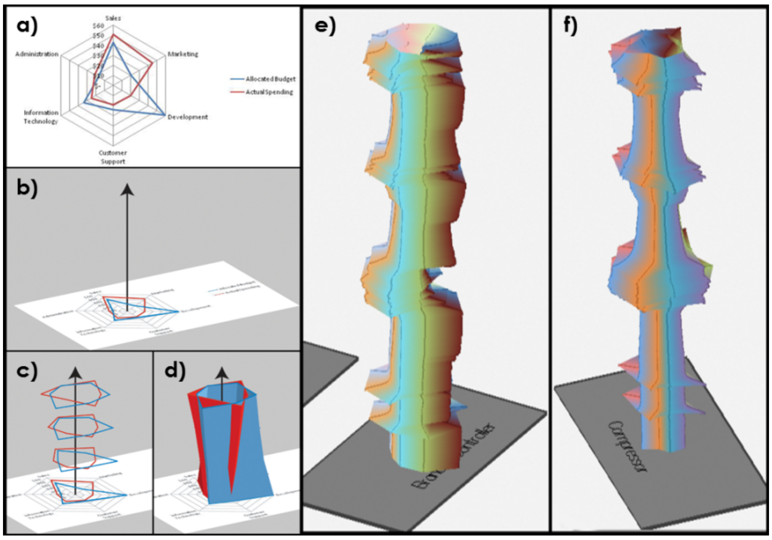


Figure 4.5: How a Wakame is built (image taken from [63])

4.4 3DRC

In this research, a 3D graphical visualization technique based on radar charts has been applied to the specific problem of the representation and analysis of divergent views and their variations in time, according to the definitions presented in the second section. For each gap $\Delta_i(View_A, View_B, t_k)$ between the views of two observers A and B at time t_k for activity i we define the *radar chart* $rc_i^{t_k}(A, B)$ as shown in Figure 4.6.a), where the three axes represent the three components $\Delta_i^c, \Delta_i^s, \Delta_i^q$. These components are then normalized. A radar chart allows us to easily detect the following information:

- the variations of a single component Δ_i^d with respect to another;
- a relative measure of the size of the gap Δ_i: the smaller the triangle area (with the vertices that tend to the center axis), the smaller Δ_i.

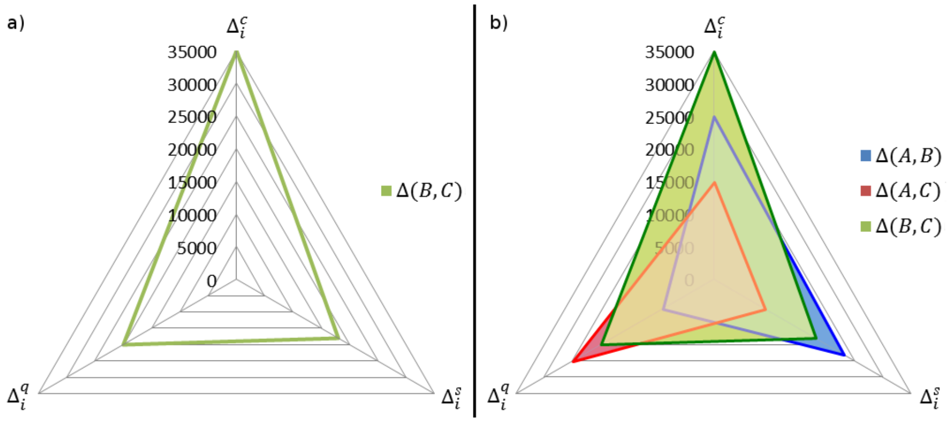


Figure 4.6: Radar chart representations of gaps between stakeholders A and B (a) and stakeholders A, B, C (b).

The *composite radar chart* $rc_i^{t_k}(A, B, C)$ shown in Figure 4.6.b) is obtained from the three radar charts $rc_i^{t_k}(A, B)$, $rc_i^{t_k}(A, C)$, $rc_i^{t_k}(B, C)$ in the following way:

$$rc_i^{t_k}(A, B, C) = rc_i^{t_k}(A, B) \uplus rc_i^{t_k}(A, C) \uplus rc_i^{t_k}(B, C) \quad (4.5)$$

where the operator $\uplus$ indicates the overlap operation on radar charts. Composite radar charts allow to effectively compare Δ_i to highlight the points of divergence or convergence between more than two stakeholders by giving an immediate identification of the differences among gaps.

To display the variation of one Δ_i with respect to time, a *3-Dimensional Radar Chart (3DRC)* is introduced as a temporal sequence of radar charts:

$$3DRC_i^{A,B} = \langle rc_i^{t_0}, rc_i^{t_1}, \ldots, rc_i^{t_k} \rangle \tag{4.6}$$

In other words, a 3DRC represents the trend of a radar chart over time (Fig. 4.7).

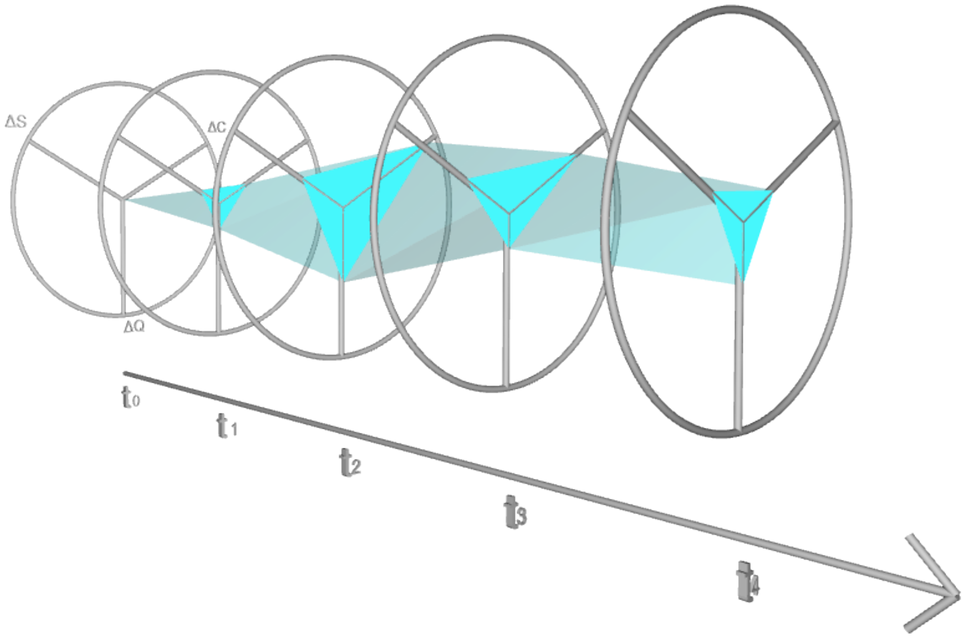


Figure 4.7: The representation of a 3DRC$_i$

In order to compare multiple trends over time, two or more 3DRCs can be overlapped as shown in fig. 4.8 where two 3DRCs (a) are overlapped to form a composite 3DRC (b). In particular, this is useful when visualizing the trends of more than two stakeholders. For example, the composite 3DRC for stakeholders A, B, and C is obtained as follows:

$$c3DRC_i^{A,B,C} = 3DRC_i^{A,B} \uplus 3DRC_i^{A,C} \uplus 3DRC_i^{B,C} \tag{4.7}$$

The definition stated above for radar charts and 3DRC on activity a_i also continues to be valid for a subset $X = a_1, a_2, \ldots, a_k$ of activities. Indeed, for stakeholders A and B, the radar chart $rc_X^{t_k}(A, B)$ gives the representation of $T\Delta_X$ where the value of each component Δ_X^d $\forall d \in \{c, s, q\}$ is obtained as a sum of all the Δ_i^d $\forall i \in \{1, \ldots, k\}$ and

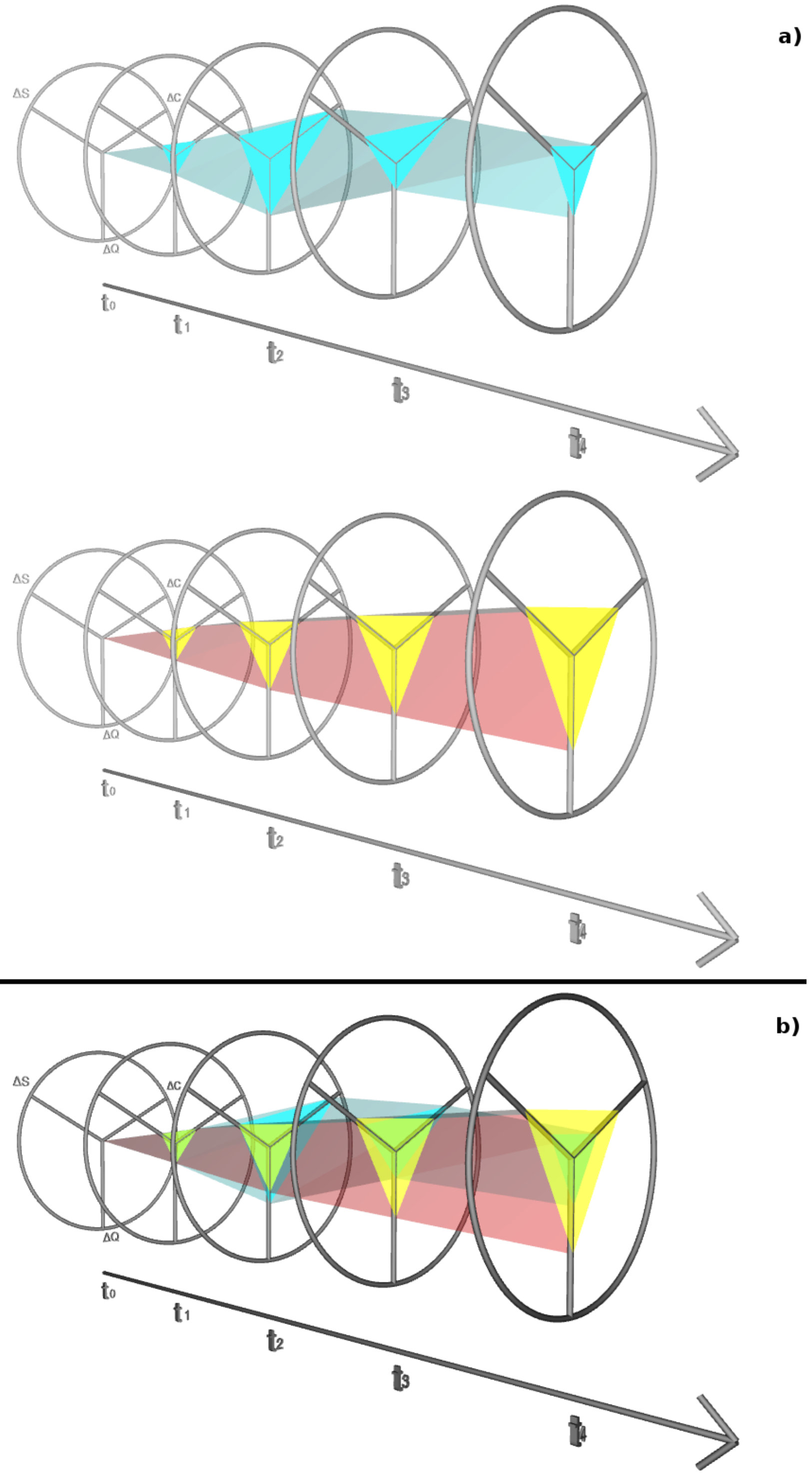


Figure 4.8: The overlapping of two 3DRCs

$\forall d \in \{c, s, q\}$; we then define the resulting 3DRC as follows:

$$3DRC_X^{A,B} = \left\langle rc_X^{t_0}, rc_X^{t_1}, \ldots, rc_X^{t_k} \right\rangle \tag{4.8}$$

If X contains all the project activities, the radar charts correspond to the representation of the total gap $T\Delta$.

The visualization algorithm places a circumference with three equally spaced radii representing the three components Δ_i^c, Δ_i^s, Δ_i^q at the origin of the t-axis representing the time; it draws a 3DRC$_i$ starting with a point at time t_0 (that means gap=0 at time of contract signing) and going on with radar-charts at discrete points of the *time*-axis.

In other words, with an approach similar to [92, 93, 63], the time is mapped along the t axis, and the various sequential measures of the function Δ are represented as radar-charts drawn on the t axis. Depending on the visual needs, the time intervals are drawn either as normalised at an equal distance or respecting the real effective distances.

As in [51], the model connects the vertices of the radar charts with polylines and assigns a single transparent color to the resulting solid obtaining something like a piece of transparent glass. In fact these drawings are used to create “filled tubes”.

When a composite 3DRC must be created, a different color is assigned to every single 3DRC to distinguish the different stakeholders views.

While in other implementations of Kiviat tubes the visual information is shown on the surface of the tube, in our case the information is represented in the volume variations that the 3DRC assumes along the time.

4.4.1 3D Shape Perception: learning by observing

From the field of perceptual psychology [22, 23] there is evidence that:

- the repeated viewing of a 3D shape is impressed rapidly in memory;
- unfamiliar shapes stand out immediately. This is done “unconsciously” and requires minimal cognitive effort.

Thanks to these properties the graphical appearance of a 3DRC is used to

1. observe all three components of the Δ_i at a glance;

2. distinguish the individual radar charts composing the 3DRC to see how the views between two stakeholders about an activity (or a subset of activities) change across the time;

3. overlap 3DRCs to understand how different stakeholders interact with each other across the time on the same activity (or subset of activities);

4. overlap two or more 3DRCs in order to compare different activities and eventually discover some hidden correlations between them.

Fig. 4.7 shows how a series of radar charts move in a certain direction over time to immediately understand the presence of some critical issues that could reveal some potential controversies. When the agreement and thus the consonance among stakeholders have a constant trend, Δ values show little variations with respect to the planned values while greater variations in one or more components draw attention to different points of view of stakeholders about the correct project execution. By observing a 3DRC we note that a "contraction" of diverging views indicates a reconciliation between those views observed through the increment of consonance among the stakeholders.

In order to augment the model with some alerting functions, two cylinders are added to a 3DRC: the internal cylinder defines a green zone representing an acceptable consonance situation while the external one delimits the maximum tolerance beyond which it is very hard to restore the minimum necessary consonance to continue the project (red zone). The intermediate area (yellow zone) indicates a divergence of points of view that requires some smooth actions by stakeholders to converge towards more tolerant consonance values. In the following, we define some alerting functions. Given:

- $CircumferenceArea_{min}$: the area of circumference of radius min, where min is the minimum tolerance degree;

- $CircumferenceArea_{max}$: the area of circumference of radius max, where max is the maximum tolerance degree;
- $Distance_{\Delta^d}$ represents the position of the radar chart vertices on the axes; it coincides with the value of Δ_i^d;

we have the following relations:

- for each radar chart in a 3DRC, if the $Distance_{\Delta^d} \leq min$ for each dimension $d \in \{c, t, q\}$ (see Fig. 4.9 at time t_1 and t_4) then the consonance is acceptable.
- If some components of the Δ_i in one or more radar charts are outside $CircumferenceArea_{min}$ but inside $CircumferenceArea_{max}$, in other words $\exists\, d \in \{c, t, q\} \mid Distance_{\Delta^d} > min$ and $Distance_{\Delta^d} < max$, then there is a partial consonance.
- If some components of the Delta function are outside the $CircumferenceArea_{max}$ then a high risk situation for the project has highlighted.

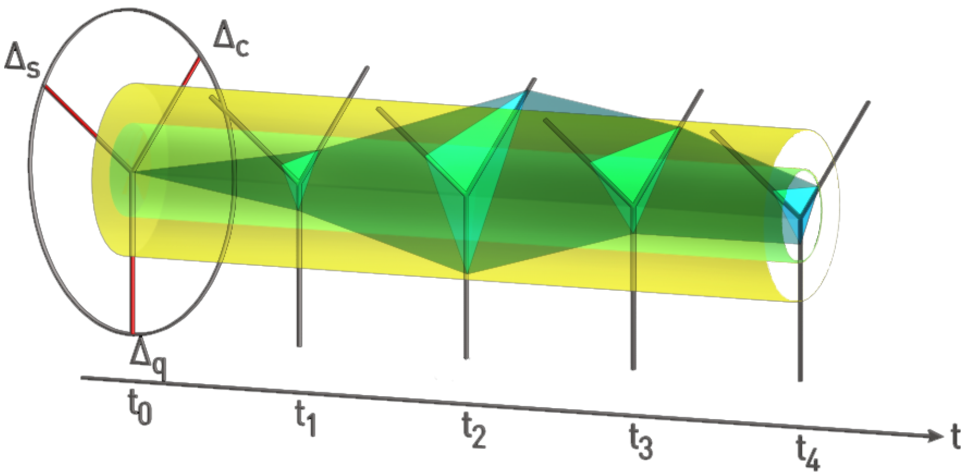


Figure 4.9: Example of monitoring not exceeded the minimum constraints of consonance

The overlapping of 3DRCs can also be executed when we consider 3DRCs representing the trend of stakeholders views about different activities. If we consider 3DRC for a set of activities X, the composite

3DRC is obtained as follows:

$$c3DRC_X^{A,B} = \biguplus_{j \in X} c3DRC_j^{A,B} \tag{4.9}$$

With this operation it is possible to analyze the possible relationships among activities and to answer questions like these:

- a high gap on an activity may have brought benefits or disadvantages on the execution of another one?
- Could some actions aimed at reducing a gap have an impact on other activities?

As an example, if there is an activity with a cost higher than the one planned, the company will probably tend to compensate this extra cost reducing the assigned resources or the quality of some other activities. By composition and comparison of 3DRCs, the analyst has a visual and more immediate perception of the cause-effect relations among activities having also a valid support for the monitoring and the proactive control of the project execution.

Chapter 5

The 3DRC Tool

> Moses replied as follows, ≪But suppose they will not believe me or listen to my words, and say to me, "Yahweh has not appeared to you"?≫Yahweh then said, ≪What is that in your hand?≫≪A staff≫, he said.≪Throw it on the ground≫, said Yahweh. Moses threw it on the ground; the staff turned into a snake and Moses recoiled from it. Yahweh then said to Moses, ≪Reach out your hand and catch it by the tail≫. He reached out his hand, caught it, and in his hand it turned back into a staff. Thus they may believe that Yahweh, the God of their ancestors, the God of Abraham, the God of Isaac and the God of Jacob, has appeared to you≫.
>
> — Exodus 4: 1-5, *The Holy Bible*

This chapter provides a description of the 3DRC Tool. In order to explore and develop our 3DRC visualization approach and to develop the analytical characteristics of the user interface a prototype system capable of displaying multi-dimensional time-based data was implemented. The figures and examples used in this chapter have been elaborated based on real the data obtained by the COSMO Project, a project financed by Campania Region (in the South of Italy) with POR 2000-2006.

This chapter is structured as follows: in section 5.1, the system architecture and its main modules and their interactions are discussed. The next section presents the operations that a user may perform when interacting with the 3DRC system. Finally, the model validation is presented trough a case study.

5.1 System Architecture

The 3DRC Tool consists of four modules (see Fig.: 5.1), each assigned to a particular task:

- User Access Control Module (UACM): the module for user administration;
- Entity Definition Module (EDM): the module for the definition of projects and related entities;
- Project Population Module (PPM): the module for the population of a project;
- 3DRC Module: (3DRCM): the display module.

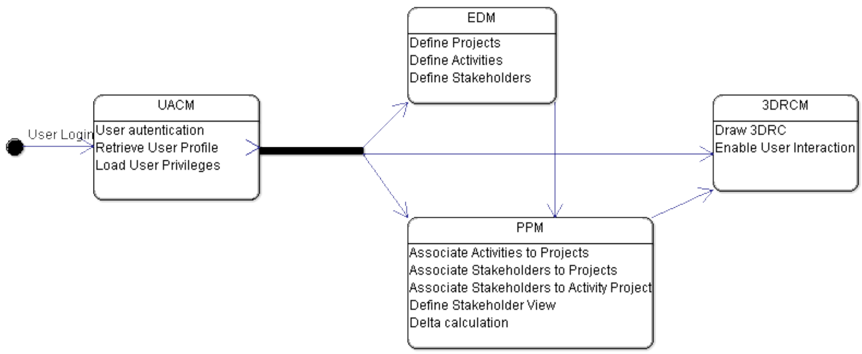


Figure 5.1: The 3DRC tool logic modules

5.1.1 UACM Module

The UACM performs some basic operations:

- definition of users who can log into the system;
- definition of criteria for the profiling of registered users and their permissions;
- access control.

This module has been designed to provide an appropriate level of security to prevent unauthorized access in order to protect data, the newly discovered knowledge, and the identity of the users.

5.1.2 EDM Module

In this module it is possible to define/modify projects, activities, and stakeholders. For each entity, the user can enter the title and a brief description. For each project, the user can also create a list of users who can view the information in read-only or in writing mode. The EDM interacts with the UACM to retrieve the set of allowed operations for each user.

In addition, this module allows to export/import, using the XML format, a project, the activity-stakeholder associations and the set of stakeholder views in/from a file.

5.1.3 PPM Module

The PPM allows to assign activities and stakeholders to individual projects. After completing the activity-stakeholder associations the user will be able to create and populate the views that stakeholders have on a specific activity project. This module also deals with the calculation of the Δ, for each activity, obtained from the information contained in the stakeholders views. The PPM also offers a preview of the 2D radar chart (and therefore of the three components Δ_c, Δ_q and Δ_t) which will be useful for the creation of 3DRC. This module interacts with the EDM for the retrieval of projects, activities and stakeholders defined by a user and with the UACM for the obtainment of the allowed operation set.

5.1.4 3DRCM Module

This module takes the data processed by the PPM to generate 3DRC. Any change in PPM triggers an update to the appearance of the 3DRC in the 3DRCM. It interacts with the UACM for the retrieval of the set of 3DRC a user can view. The 3DRCM also provides a set of operations to be able to interact with 3DRC (see Section 5.2.2).

5.2 Interacting with 3DRCs

5.2.1 The GUI: FLUID or WIMP?

The traditional desktop interfaces are based on using the conventional WIMP desktop metaphor (Windows, Icons, Menus and Pointer) with a control panel, push buttons, and checkboxes. For touch-based devices it is also available another interface, called FLUID, that eliminates the control panel and uses touch to define actions on the visualized data. Nowadays, application developers need to consider the increasingly spread of touch-based devices (tablet, smart phone, etc.). For these reasons, when dealing with these devices, they are faced with a dilemma in implementing interactions: do we simply port the WIMP-style interfaces that are so familiar on the desktop by only replacing the mouse click with a finger tap; or should we change the interaction style to a more "touch-centric" interface that affords more direct manipulation of that application? [94]

The early 3DRC Tool Prototype was initially created to run on traditional desktop systems while the current 3DRCM partially supports the use of a "touch-centric" interface.

To meet the user needs and to execute the 3DRC Tool satisfactorily on tablets, smart-phones, etc., I tried to create a GUI able to switch, as much as possible, from a WIMP to a FLUID interface depending on the device. For these reasons, I made use of libraries and frameworks that would allow the use of both classical specific interactions required for desktop systems and touch systems.

5.2.2 The operations

Figure 5.2 shows the interface of the prototype implementing the 3DRC visualization model. The users mainly interact with this interface through the mouse and keyboard or using the finger on a touch-screen. In the following we list the main available operations:

- *Space rotation*: the 3D space containing the 3DRCs can be rotated around one or more axes. The effect is the changing of the point of view for the visualized items. The operation is performed by

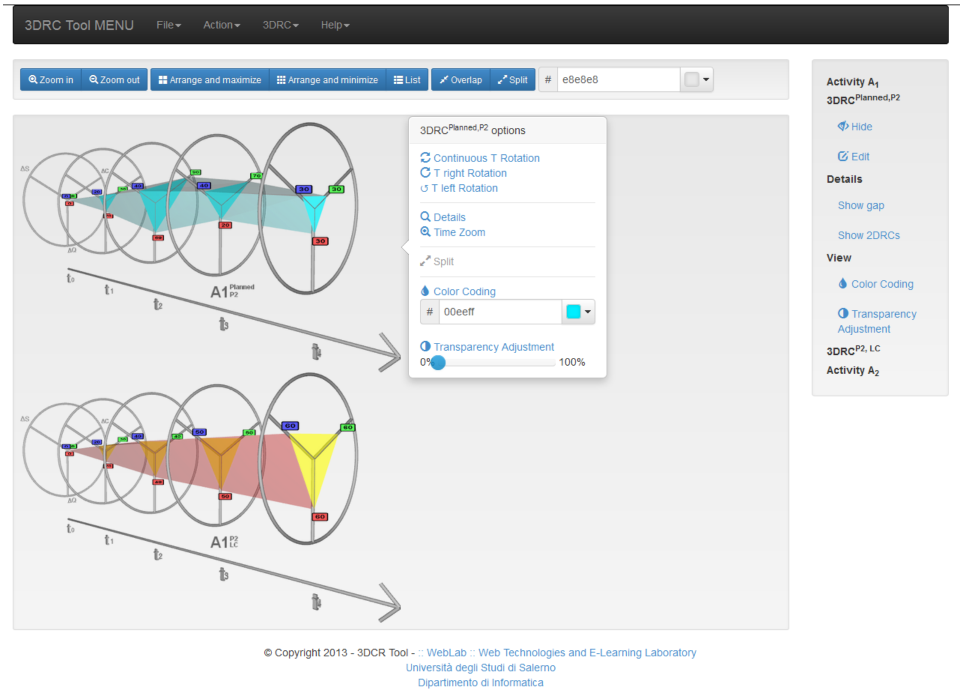


Figure 5.2: The interface of the prototype

clicking the left button and moving the mouse or by dragging a finger on the touch-screen;

- *Space shift*: the 3D space containing the 3DRCs can be moved along the x and y axes. The operation is performed by clicking the right button and moving the mouse or by dragging three fingers on the touch-screen;

- *T-rotation*: a single 3DRC can be rotated on its center around the t-axis (see Fig. 5.3). The effect is the changing of the point of view for the visualized items. This operation entails a rotation of the three Delta vectors. The operation is performed by clicking the left button on the center axis of the 3DRC and rotating the mouse wheel or by tapping with a finger on the touch-screen and use an other finger to pivot around the first; the cylinder can alternatively be animated by pressing the Play button in the toolbox associate to the 3DRC;

- *Tooltip*: when the user hovers the cursor over an item, a box con-

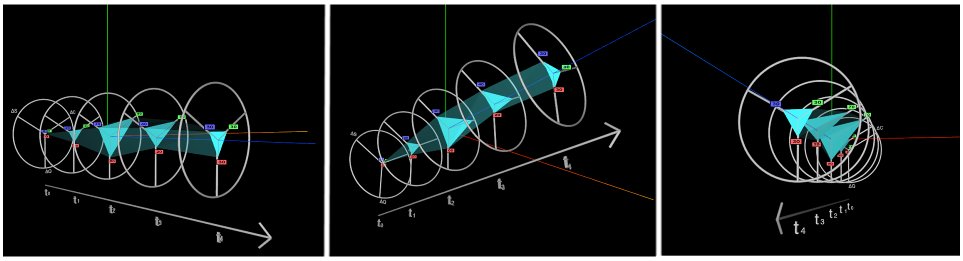


Figure 5.3: An example of T-rotation on 3DRC

taining $\Delta(View_{User_i}, View_{User_j})$, $T\Delta_d(View_{User_i}, View_{User_j})$ appears; an arrow indicates whether the trends of the gaps are increasing or decreasing with respect to a particular size;

- *Details*: further details are obtained when the user clicks over an item. The item is selected and the selection is highlighted through a yellow light surrounding the item. Details of the selected item (see Figure 5.6) are shown in the *details view* and they include, as an example, the monitoring panel for the GAP analysis of project activities or the possibility of seeing the single radar chart associated with that item;

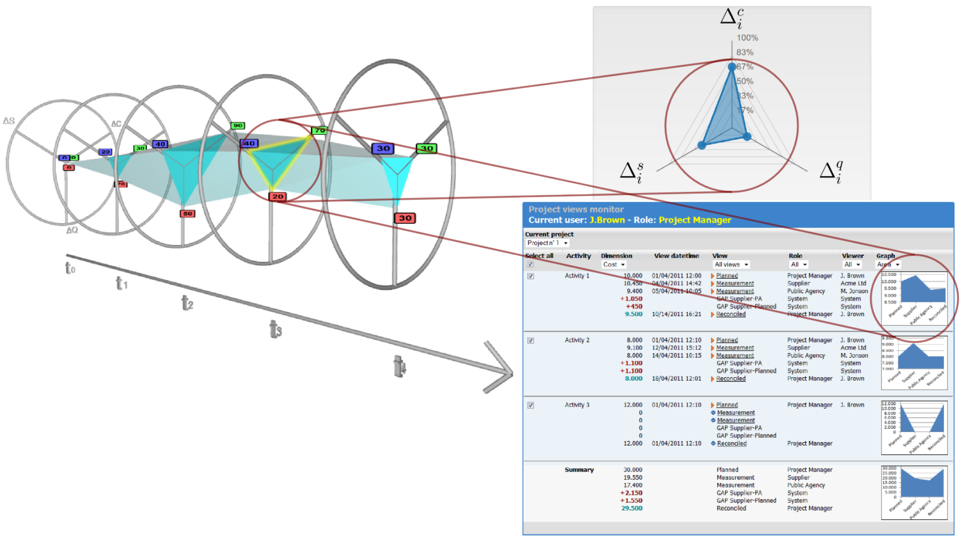


Figure 5.4: An example of Details operation on 3DRCs

- *Time zoom*: this operation is useful when the user is interested in a shorter time interval than the one currently visualized. The interval of interest can be selected by dragging the mouse over a horizontal slice of the 3DRC. The operation cuts off the time intervals complementary to the selection and stretches the selected slice over the whole result view. The subset of items included in the selection is visualized with a larger spacing, thus allowing the user to have a more detailed view. Further zoom operations are possible on the selection.

- *Zoom*: this operation allows to zoom in or zoom out the 3DRC to navigate inside of it. The operation is performed by rotating the mouse wheel or move two fingers apart/toward each other;

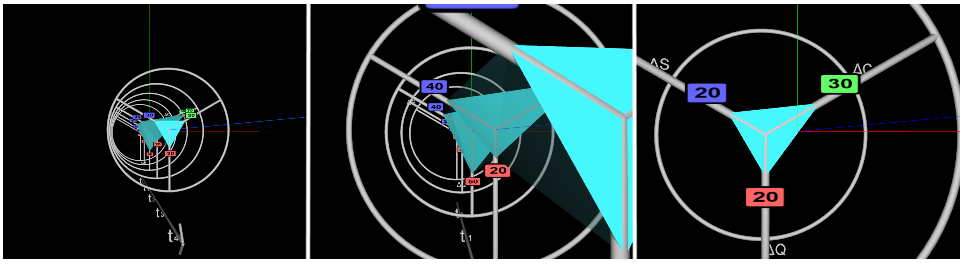


Figure 5.5: An example of Zoom operation on 3DRCs

- *Filter*: it allows to select what activities and/or stakeholders must be considered for the visualization of the 3DRCs. The operation is performed by the 3DRC contextual menu;

- *Overlap*: the users can overlap multiple 3DRCs as discussed in section 4.4. The operation is performed by selecting at least two 3DRCs and click on the overlap button or touching at the same time two 3DRCs with long presses and performing a tap;

- *Split*: given a composite 3DRC, the user can decompose it in the original 3DRCs. The operation is performed by selecting the composite 3DRC and clicking on the unoverlap button or selecting by a finger the composite 3DRC and make a double tap on the composite 3DRC;

- *Color Coding Switch*:the user can switch on/off color coding of a 3DRC. The operation is performed by the 3DRC contextual menu;

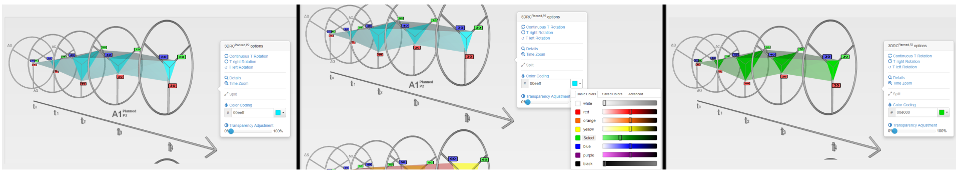


Figure 5.6: An example of the Color Coding Switch operation on 3DRCs

- *Transparency Adjustment*: the transparency degree of the selected 3DRC may be adjusted. The operation is performed by the 3DRC contextual menu;

5.3 The case study

The 3DRC approach has been validated with real data obtained by "COSMO Project - Contract & SLA Management Outlook" financed by Campania Region (in the South of Italy) with POR 2000-2006 - Measure 3.17 - whose aim was the realization of a sensor network for the monitoring of environment variables and fire prevention [88]. The project had a duration of 18 months from September 2008 and it was presented by a Virtual Enterprise composed of four companies of Campania Region, together with the University of Benevento (Department of Engineering).

The COSMO project was divided in nine work packages with a total of 39 activities. Each activity was assigned to a partner (P_i, i=1,..,4) with the coordination of the leading company (LC) that had also the role of the front-end towards the administration offices of Campania Region. Three official project milestones were provided for the reporting of the project, but also other internal milestones were established among the partners to control the execution and the exchange of information and project documentation. For brevity, only data showing a summary of views and Δ about the project activity A_1 are reported in Table 5.1. This activity was assigned to the partner P_2 and three internal milestones were defined by LC for the monitoring. The *Planned* column shows activity data as approved by Campania Region to be eligible for funding. Other two columns on the right compare P_2 and LC views respectively while the last column visualizes the official cost data recognized by Campania Region after the first Project Review Meeting.

Table 5.1: Data Summary about Activity A_1

Observation Time	Attribute	Planned	View$_{P_2}(A_1)$	View$_{LC}(A_1)$	View$_{CR}(A_1)$
	Start	1 Oct 2008	15 Oct 2008	15 Oct 2008	15 Oct 2008
	End	3 Nov 2008	24 Nov 2008	2 Dec 2008	4 Dec 2008
t_1:6 nov 2008	**State**	completed	in progress	in progress	
	Time	$\Delta_i^{dur}(Planned, P_2) = 12$	$\Delta_i^{dur}(P_2, LC)$=0		
	Cost	12390	10725	9780	
		$\Delta_i^c(Planned, P_2)$=1665	$\Delta_i^c(P_2, LC)$=945		
	Quality	D1.1:4	D1.1:4	D1.1:3	
		$\Delta_i^q(Planned, P_2) = 0$	$\Delta_i^q(P_2, LC)$=1		
t_2:27 nov 2008	**State**	completed	completed	in progress	
	Time	$\Delta_i^{dur}(Planned, P2)$=6	$\Delta_i^{dur}(P_2, LC)$=3		
	Cost	12390	12390	11890	
		$\Delta_i^c(Planned, P_2)$=0	$\Delta_i^c(P_2, LC)$=500		
	Quality	D1.2:4	D1.1: 4,D1.2: 4	D1.1:4, D1.2:2	
		$\Delta_i^q(Planned, P_2) = 0$	$\Delta_i^q(P_2, LC)$=2		
t_3:2 dec 2008	**State**	completed	completed	completed	
	Time	$\Delta_i^{dur}(Planned, P_2)$=6	$\Delta_i^{dur}(P_2, LC)$=8		
	Cost	12390	12390	12035	
		$\Delta_i^c(Planned, P_2)$=0	$\Delta_i^c(P_2, LC)$=355		
	Quality	D1:4	D1:4	D1:4	
		$\Delta_i^q(Planned, P_2) = 0$	$\Delta_i^q(P_2, LC)$=0		
t_4:15 feb 2009	**Cost**	12390	12390		11650
					$\Delta_i^c(P_2, CR)$=740

In the table, Δ^{dur} corresponds to the difference of the activity durations declared by each stakeholder ([88]), where duration dur at time t of an activity a_i is computed according to the following statement:

$$dur_i^t = \begin{cases} t_{end} - t_{start} & \text{if state=completed} \\ t - t_{start} & \text{if state = started} \\ undefined & \text{if state = not\_started} \end{cases} \tag{5.1}$$

Δ^c is obtained as the difference between declared costs while the quality dimension has been summarized with a range of numbers between 0 and 4, where each number in the range indicates the corresponding value in the set {`not complaint`, `low quality`, `sufficient quality (or complaint)`, `good quality`, `excellent`}. By data comparison at the first milestone, different perceptions about activity state can be

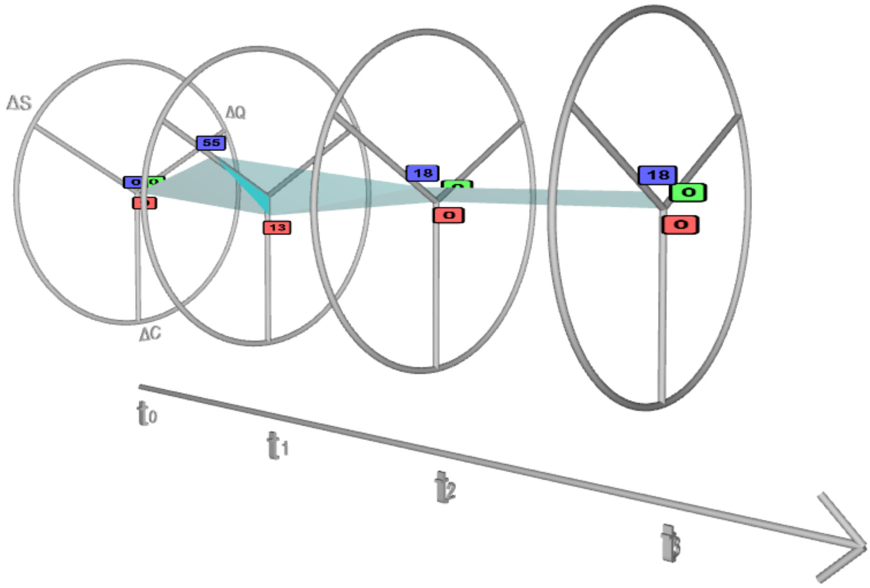


Figure 5.7: The 3DRC generated about diverging views between stakeholders: Planned and P_2

observed. With respect to the planned values (fig. 5.7 at time t_2), P_2 declares the beginning of activity as shifted of 15 days, and duration shows a gap of about 6 days while gaps on cost and quality arise because the activity is still running. Comparing the views of P_2 and LC (fig. 5.8), we observe that LC confirms the shifting of the activity start and some differences with respect to the planned values also remain in duration (time t_1); moreover, there are gaps on other two dimensions, due to the missing of some statements of expenditure and a delay in the delivery of the preliminary version of project document D1. At the second milestone t_2, after an exchange of other documents, the gap about the cost decreases while there is a greater quality gap on the second section of D1: for LC the activity is still running while P_2 declares its completion. An agreement is reached at t_3 when the activity results completed and the deliverable is accepted by LC. Finally, according to the official milestone, financial reporting made by Campania Region (RC) shows a difference in the cost because some expenditures resulted as not eligible.

By analyzing the two graphical representations, with no previous knowledge about data, an observer immediately perceives the differences

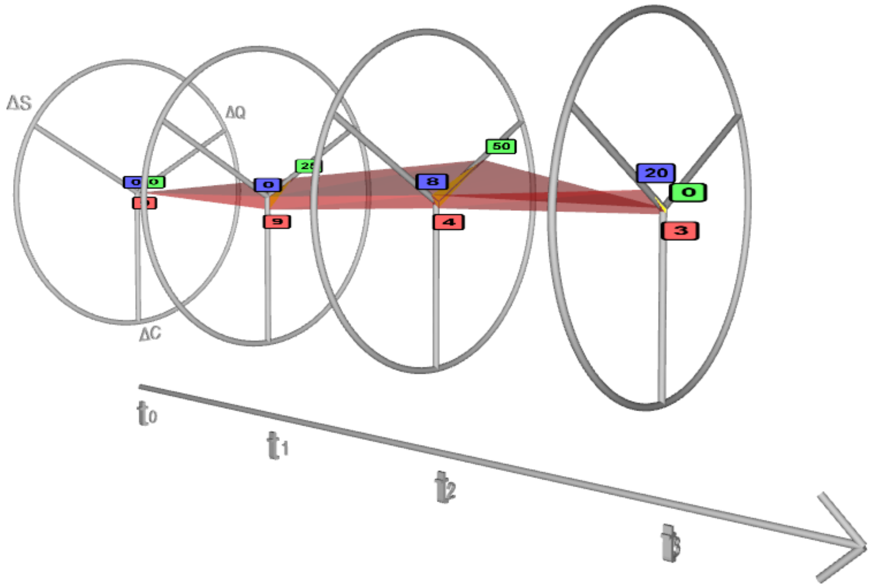


Figure 5.8: The 3DRC generated about diverging views between stakeholders: P_2 an L_C

between what P_2 declares and what the controller (in this case the leading company LC) verifies. In the case of A_1, according to the P_2 view the activity has a trend very close to the planned values (fig. 5.8) while the control action made by the leading company highlights a lowest quality than that required.

Chapter 6

Conclusions

≪It has already happened.
I am the Alpha and the Omega,
the Beginning and the End.
I will give water from the well of life free to anybody who is thirsty;
anyone who proves victorious will inherit these things;
and I will be his God and he will be my son.≫

— Revelation 21: 6-7, *The Holy Bible*

Nowadays data are generated in increasingly higher volumes. The exploration and analysis of these vast amounts of data becomes more and more an important but difficult problem. The techniques of information visualization can help solve this problem.

In this thesis, two types of datasets was presented: one in which the spatial characteristics are relatively dynamic and the data are sampled at different periods of time, and the other where many dimensions prevail, although the spatial characteristics are relatively static. One dataset referred to the dispute management, a peculiar aspect of uncertainty arising from contractual relationships that regulate a project execution. The other data set referred to a novel display of Digital Libraries.

Two main issues were also presented. The first issue regarded the enhancement of CyBiS, a novel 3D analytical interface for a Bibliographic Visualization Tool with the objective of improving scientific paper search. The second regarded the analysis of the possibilities offered by a 3D novel visualization technique 3DRC to represent and analyse the problem of diverging stakeholders views during a project execution and to addresses the prevention and proactive handling of the potential controversies among project stakeholders.

In addition, the benefits of using 3D techniques in combination with techniques for multidimensional visual data mining have been shown also by providing guidance on how to overcome the problems that plague 3D visual representations of data.

In this thesis I have presented two visualization techniques:

- The first, called CyBiS2, an enhanced version of the tool CyBiS, is a new interface for the representation of the salient data of searched documents in a single view. In fact one of the main distinguishing feature of CyBiS, rarely exploited previously, was the inclusion in a single view of the most important data of interest for a scientific paper. This feature allowed a user to evaluate documents at the same time without the need of switching among different views or changing the context.

  Future work will focus on: evaluating the effectiveness and efficiency of the interface through studies focused on users; replacing the old 3D graphic engine with the new and more performant used in the 3DRC Tool; comparing the performance of the interface with respect to that of other systems.

- The second, called 3DRC, addresses the prevention and proactive handling of the potential controversies among project stakeholders. The approach, based on the 3D Kiviat diagrams, uses radar charts to represent the points of view of stakeholders about the execution of the project activities and the trends of the project consonance. 3DRCs use radar charts to show the gap variations over time; moreover, the overlap operation provides functionalities to merge and compare different 3DRCs.

  The novelty of this work is the application of a 3D graphical visualization technique to the specific problem of the Project Management concerning the dispute management. Some strengths of 3DRC are the possibility of analyzing the relationships among activities, understanding how different stakeholders interact with each other across the time on the same activity (or subset of activities), comparing different activities and eventually discovering some hidden correlations between them. The method discussed in section 4.2 has been validated in the context of the COSMO project obtaining satisfactory results in terms of prevention or immediate handling of potential disputes. I believe that the introduction of the 3DRC technique offers additional benefits to stakeholders about real time project monitoring giving a visual and more immediate perception of the project trends not easily identifiable from tabular data.

Future work will focus on continuing to test the effectiveness of the 3DRC technique on real projects although this requires a long time. Currently, a validation phase against the SMART-TUNNEL project (involving the realization of an e-logistics system for the Italian ports) is in progress. The preliminary ideas of 3DRC, discussed at the PKMT2012 workshop [95] and at the ACM conference [27], gained a lot of interest from certified Project Managers and practitioners encouraging us to continue the research work. Other future work include the implementation of a full FLUID interface to provide an increasing usability and a full user experience on touch-systems, and the testing of new gestures designed for touch systems through the use of case studies.

Bibliography

[1] D. A. Keim, "Visual exploration of large data sets," *Communications of the ACM*, vol. 44, no. 8, pp. 38–44, 2001.

[2] D. A. Keim, "Information visualization and visual data mining," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 8, pp. 1–8, Jan. 2002.

[3] D. A. Keim, M. Sips, and M. Ankerst, *Visual data-mining techniques*. Bibliothek der Universität Konstanz, 2004.

[4] G. Linoff and M. Berry, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. IT Pro, Wiley, 2011.

[5] I. Witten, E. Frank, and M. Hall, *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems, Elsevier Science, 2011.

[6] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems, Elsevier Science, 2011.

[7] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "Crisp-dm 1.0 step-by-step data mining guide," tech. rep., The CRISP-DM consortium, August 2000.

[8] C. Shearer, "The crisp-dm model: The new blueprint for data mining," *Journal of Data Warehousing*, vol. 5, no. 4, 2000.

[9] N. Leaper, "a visual guide to crisp-dm methodology - from wikipedia, the free encyclopedia." `http://exde.wordpress.com/2009/03/13/a-visual-guide-to-crisp-dm-methodology/`.

[10] B. B. Bederson and B. Shneiderman, *The craft of information visualization: readings and reflections*. Morgan Kaufmann, 2003.

[11] "University of illinois at urbana-champaign digital libraries initiative, glossary." `http://dli.grainger.uiuc.edu/glossary.htm`.

[12] S. K. Card, J. D. Mackinlay, and B. Shneiderman, "Readings in information visualization," ch. Information Visualization, pp. 1–34, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999.

[13] S. Card, "The human-computer interaction handbook," ch. Information Visualization, pp. 544–582, Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 2003.

[14] C. Chen, "Top 10 unsolved information visualization problems," *Computer Graphics and Applications, IEEE*, vol. 25, pp. 12–16, July 2005.

[15] D. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler, "Challenges in visual data analysis," in *Information Visualization, 2006. IV 2006. Tenth International Conference on*, pp. 9–16, July 2006.

[16] T. Munzner, "Process and pitfalls in writing information visualization research papers," in *Information visualization*, pp. 134–153, Springer, 2008.

[17] S. K. Card, G. G. Robertson, and J. D. Mackinlay, "The information visualizer, an information workspace," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '91, (New York, NY, USA), pp. 181–186, ACM, 1991.

[18] D. A. Keim, W. Müller, and H. Schumann, *Visual data mining*. Bibliothek der Universität Konstanz, 1997.

[19] P. C. Wong, "Visual data mining," *Computer Graphics and Applications, IEEE*, vol. 19, pp. 20–21, Sept 1999.

[20] C. Ware, *Information Visualization: Perception for Design*. Interactive Technologies, Elsevier Science, 2012.

[21] K. Koffka, "Principles of gestalt psychology," 1935.

[22] H. von Helmholtz and J. P. Southall, "Helmholtz's treatise on physiological optics, vol 1 (trans," 1924.

[23] E. Boring, *A history of experimental psychology*. Century psychology series, The Century Co., 1929.

[24] Z. Pizlo, *3D shape: Its unique place in visual perception*. MIT Press, 2010.

[25] M. Tavanti and M. Lind, "2d vs 3d, implications on spatial memory," in *Information Visualization, 2001. INFOVIS 2001. IEEE Symposium on*, pp. 139–145, Oct 2001.

[26] G. Costagliola and V. Fuccella, "Cybis: A novel interface for searching scientific documents," in *Information Visualisation (IV), 2011 15th International Conference on*, pp. 276 –281, july 2011.

[27] R. Aiello, G. Nota, G. Costagliola, and F. Torre, "3drc: A novel technique in evaluating stakeholders project views," in *Proceedings of the 6th International Symposium on Visual Information Communication and Interaction*, VINCI '13, (New York, NY, USA), pp. 88–96, ACM, 2013.

[28] N. Elmqvist and et al., "Citewiz: A tool for the visualization of scientific citation networks," tech. rep., 2007.

[29] M. Lesk, *Practical Digital Libraries: Books, Bytes, and Bucks*. The Morgan Kaufmann series in multimedia information and systems, Morgan Kaufmann Publishers, 1997.

[30] G. Nota, R. Aiello, and R. Carvello, "Modeling the reconciliation process for stakeholder diverging project views," in *Second Proceedings: Short papers, Doctoral Symposium and Workshops of the*

PROFES 2011, pp. 128–134, ACM, 20 June 2011. in PKMT2011 Workshop.

[31] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, (Washington, DC, USA), pp. 336–, IEEE Computer Society, 1996.

[32] "Infovis:wiki @ONLINE." `http://www.infovis-wiki.net/index.php/Visual_Information-Seeking_Mantra`.

[33] B. Craft and P. Cairns, "Beyond guidelines: What can we learn from the visual information seeking mantra?," in *Proceedings of the Ninth International Conference on Information Visualisation*, IV '05, (Washington, DC, USA), pp. 110–118, IEEE Computer Society, 2005.

[34] K. Bache and M. Lichman, "UCI machine learning repository," 2013.

[35] "Noaa's national climatic data center (ncdc) @ONLINE." `http://www.ncdc.noaa.gov/`.

[36] "Titanic data @ONLINE." `http://lib.stat.cmu.edu/S/Harrell/data/descriptions/titanic.html`, Mar. 1999.

[37] A. Inselberg, "The plane with parallel coordinates," *The Visual Computer*, vol. 1, no. 2, pp. 69–91, 1985.

[38] H. Chernoff, "The use of faces to represent points in k-dimensional space graphically," *Journal of the American Statistical Association*, vol. 68, no. 342, pp. pp. 361–368, 1973.

[39] D. Keim, "Designing pixel-oriented visualization techniques: theory and applications," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 6, pp. 59–78, Jan 2000.

[40] D. A. Keim, "Visual exploration of large data sets," *Commun. ACM*, vol. 44, pp. 38–44, Aug. 2001.

[41] M. Friendly, "A brief history of the mosaic display," *Journal of Computational and Graphical Statistics*, vol. 11, no. 1, 2002.

[42] B. B. Bederson, B. Shneiderman, and M. Wattenberg, "Ordered and quantum treemaps: Making effective use of 2d space to display hierarchies," *ACM Trans. Graph.*, vol. 21, pp. 833–854, Oct. 2002.

[43] "Treemap of the us presidential elections of 2012 @ONLINE." `http://www.treemap.com/datasets/uselections/`, Nov. 2012.

[44] E. A. Bier, M. C. Stone, K. Pier, W. Buxton, and T. D. DeRose, "Toolglass and magic lenses: The see-through interface," in *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '93, (New York, NY, USA), pp. 73–80, ACM, 1993.

[45] "Understanding descriptive statistics @ONLINE." `http://nationalatlas.gov/articles/mapping/a_statistics.html`, Jan. 2013.

[46] P. Simon, *Too Big to Ignore: The Business Case for Big Data*. Wiley and SAS Business Series, Wiley, 2013.

[47] S. Milborrow, "Decision tree learning @ONLINE." `http://en.wikipedia.org/wiki/Decision_tree_learning/`, Aug. 2011.

[48] D. Gamberger, N. Lavrac, and D. Wettschereck, "Subgroup visualization: A method and application in population screening," in *In Proceedings of the International Workshop on intelligent Data Analysis in Medicine and Pharmacology, IDAMAP*, 2002.

[49] M. Friendly, "A brief history of the mosaic display," *Journal of Computational and Graphical Statistics*, vol. 11, no. 1, pp. 89–107, 2002.

[50] H. Hofmann, A. P. J. M. Siebes, and A. F. X. Wilhelm, "Visualizing association rules with interactive mosaic plots," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, (New York, NY, USA), pp. 227–235, ACM, 2000.

[51] E. Fanea, S. Carpendale, and T. Isenberg, "An interactive 3d integration of parallel coordinates and star glyphs," in *Proceedings of the IEEE Symposium on Information Visualization (InfoVis 2005, October 23–25, 2005, Minneapolis, Minnesota, USA)* (J. Stasko and M. Ward, eds.), (Los Alamitos, CA), pp. 149–156, IEEE Computer Society, 2005.

[52] M. Streit, R. C. Ecker, K. sterreicher, G. E. Steiner, H. Bischof, C. Bangert, T. Kopp, and R. Rogojanu, "3d parallel coordinate systemsa new data visualization method in the context of microscopy-based multicolor tissue cytometry," *Cytometry Part A*, vol. 69A, no. 7, pp. 601–611, 2006.

[53] J. Johansson, M. Cooper, and M. Jern, "Jern m.: 3-dimensional display for clustered multi-relational parallel coordinates," in *In Proceedings of the 9th International Conference on Information Visualization (2005*, pp. 188–193.

[54] L. Kavraki, "Dimensionality reduction methods for molecular motion @ONLINE." http://cnx.org/content/m11461/1.10/, June 2011.

[55] M. Natrella, *NIST/SEMATECH e-Handbook of Statistical Methods*. NIST/SEMATECH, July 2010.

[56] L. I. Smith, "A tutorial on principal components analysis," tech. rep., Cornell University, USA, February 26 2002.

[57] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, pp. 1464–1480, Sep 1990.

[58] "Self-organizing map @ONLINE." `http://en.wikipedia.org/wiki/Self-organizing_map`.

[59] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley, "Dna visual and analytic data mining," in *Proceedings of the 8th Conference on Visualization '97*, VIS '97, (Los Alamitos, CA, USA), pp. 437–ff., IEEE Computer Society Press, 1997.

[60] P. Hoffman, G. Grinstein, and D. Pinkney, "Dimensional anchors: A graphic primitive for multidimensional multivariate information visualizations," in *Proceedings of the 1999 Workshop on New Paradigms in Information Visualization and Manipulation in Conjunction with the Eighth ACM Internation Conference on Information and Knowledge Management*, NPIVM '99, (New York, NY, USA), pp. 9–16, ACM, 1999.

[61] E. Kandogan, "Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions," in *In Proceedings of the IEEE Information Visualization Symposium, Late Breaking Hot Topics*, pp. 9–12, 2000.

[62] N. D. Cooprider and R. P. Burton, "Extension of star coordinates into three dimensions," in *Electronic Imaging 2007*, pp. 64950Q–64950Q, International Society for Optics and Photonics, 2007.

[63] C. Forlines and K. Wittenburg, "Wakame: sense making of multidimensional spatial-temporal data," in *Proceedings of the International Conference on Advanced Visual Interfaces*, pp. 33–40, ACM, 2010.

[64] G. Costagliola and V. Fuccella, "Cybis: A novel interface for searching scientific documents," in *Information Visualisation (IV), 2011 15th International Conference on*, pp. 276 –281, july 2011.

[65] P. Bergstrom and D. Atkinson, "Augmenting the exploration of digital libraries with web-based visualizations," in *Digital Information Management, 2009. ICDIM 2009. Fourth International Conference on*, pp. 1 –7, nov. 2009.

[66] D. Modjeska, V. Tzerpos, P. Faloutsos, and M. Faloutsos, "Bivteci: a bibliographic visualization tool," in *Proceedings of the 1996 conference of the Centre for Advanced Studies on Collaborative research*, CASCON '96, pp. 28–, IBM Press, 1996.

[67] T. Nguyen and J. Zhang, "A novel visualization model for web search results," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, pp. 981–988, Sept. 2006.

[68] W. Wiza, K. Walczak, and W. Cellary, "Periscope: a system for adaptive 3d visualization of search results," in *Proceedings of the ninth international conference on 3D Web technology*, Web3D '04, (New York, NY, USA), pp. 29–40, ACM, 2004.

[69] C. Chen, "Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature," *J. Am. Soc. Inf. Sci. Technol.*, vol. 57, pp. 359–377, Feb. 2006.

[70] M. Crossley, J. Davies, A. Mcgrath, M. Rejman-Greene, and V. Candide, "The knowledge garden," 1999.

[71] K. Einsfeld, S. Agne, M. Deller, A. Ebert, B. Klein, and C. Reuschling, "Dynamic visualization and navigation of semantic virtual environments," in *Proceedings of the conference on Information Visualization*, IV '06, (Washington, DC, USA), pp. 569–574, IEEE Computer Society, 2006.

[72] G. S. Davidson, B. Hendrickson, D. K. Johnson, C. E. Meyers, and B. N. Wylie, "Knowledge mining with vxinsight: Discovery through interaction," *JOURNAL OF INTELLIGENT INFORMATION SYSTEMS*, vol. 11, pp. 259–285, 1998.

[73] B. Wong, S. Choudhury, C. Rooney, R. Chen, and K. Xu, "Invisque: Technology and methodologies for interactive information visualization and analytics in large library collections," in *Research and Advanced Technology for Digital Libraries* (S. Gradmann, F. Borri, C. Meghini, and H. Schuldt, eds.), vol. 6966 of *Lecture Notes in Computer Science*, pp. 227–235, Springer Berlin Heidelberg, 2011.

[74] J. D. Mackinlay, R. Rao, and S. K. Card, "An organic user interface for searching citation links," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '95, (New York, NY, USA), pp. 67–73, ACM Press/Addison-Wesley Publishing Co., 1995.

[75] Q. V. Nguyen, M. L. Huang, and S. Simoff, "Visualization of relational structure among scientific articles," in *Proceedings of the 9th international conference on Advances in visual information*

systems, VISUAL'07, (Berlin, Heidelberg), pp. 415–425, Springer-Verlag, 2007.

[76] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," in *INFORMATION PROCESSING AND MANAGEMENT*, pp. 513–523, 1988.

[77] T. S. Group., "Chaos summary," 2009. http://www1.standishgroup.com/newsroom/ chaos_2009.php.

[78] M. Bloch, S. Blumberg, , and J. Laartz, "Delivering large-scale it projects on time, on budget, and on value," October 2012. http://www.mckinsey.com/insights/business_technology/delivering_large-scale_it_projects_on_time_on_budget_and_on_value.

[79] A. De Meyer, C. H. Loch, and M. T. Pich, "Managing Project Uncertainty: From Variation to Chaos," *MIT Sloan Management Review*, vol. 43, no. 2, pp. 60–67, 2002.

[80] S. Ward and C. Chapman, "Stakeholders and uncertainty management in projects," *Construction Management and Economics*, vol. 26, pp. 563–577, June 2008.

[81] H. Krane, A. Rolstads, and N. Olsson, "A case study of how stakeholder management influenced project uncertainty regarding project benefits," *International Journal of Information Technology Project Management (IJITPM)*, vol. 3, pp. 21–37, June 2012.

[82] G. Costagliola and V. Fuccella, "Fine-grained analysis of web tasks through data visualization," in *ICWE'09: Proceedings of the Ninth International Conference on Web Engineering*, (San Sebastian, Spain), pp. 1–15, Jun. 24-26 2009.

[83] G. Costagliola, V. Fuccella, M. Giordano, and G. Polese, "Monitoring online tests through data visualization," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 21, no. 6, pp. 773–784, 2009. Washington, DC, USA.

[84] F. T. Anbari, "Earned value project management method and extensions," 2004.

[85] C. Budd and C. Budd, *A Practical Guide to Earned Value Project Management*. Management Concepts, 2005.

[86] G. Golinelli, *Viable Systems Approach (VSA) - Governing Business Dynamics*. CEDAM, 2010.

[87] S. Barile, *L'impresa come sistema Contributi sull'approccio Sistemico Vitale (ASV)*. Giappichelli, 2008.

[88] G. Nota and R. Aiello, "Managing uncertainty in complex projects," in *NEW Series Chaos Theory: What perspectives for the Economics* (M. Faggini, ed.), Springer Verlag, 2013.

[89] G. Nota, R. Aiello, and L. Cervelli, "Statechart modeling of contractual disputes," in *Software, Knowledge Information, Industrial Management and Applications. Int. Conf. 5th 2011 (SKIMA 2011)*, pp. 7–12, Institute of Electrical and Electronics Engineers (IEEE), 8-10 september 2011.

[90] Y. Guo, "Implementation of 3d kiviat diagrams," 2008. Bachelor's thesis, Vxj University, Sweden, 2008.

[91] A. Kerren and I. Jusufi, "Novel visual representations for software metrics using 3d and animation," *In J. Mnch, P. Liggesmeyer, editors, Software Engineering 2009 - Workshopband*, pp. 147–154. Lecture Notes in Informatics (LNI), Proceedings Vol. P-150, 2009. GI-Edition.

[92] S. Hackstadt and A. Malony, "Visualizing parallel programs and performance," *Computer Graphics and Applications, IEEE*, vol. 15, no. 4, pp. 12–14, 1995.

[93] C. Tominski, J. Abello, and H. Schumann, "Interactive poster: 3d axes-based visualizations for time series data," Citeseer, 2005.

[94] S. M. Drucker, D. Fisher, R. Sadana, J. Herron, and m. schraefel, "Touchviz: A case study comparing two interfaces for data analytics

on tablets," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, (New York, NY, USA), pp. 2301–2310, ACM, 2013.

[95] "Pmi-sic project management institute south italy chapter - pkmt 2012 workshop: Complessit nel project e knowledge management @ONLINE." http://www.pmi-sic.org/index.php?option=com_content&view=article&id=265&Itemid=133, Dec. 2012.