

SPACE OPTIMISATION IN MULTI-DIMENSIONAL DATA VISUALISATION

by

Wen Bo Wang

Supervisor: A/Prof. Mao Lin Huang
Co-Supervisor: Dr. Quang Vinh Nguyen

**Faculty of Engineering and Information
Technology
University of Technology, Sydney**

18 April 2016

*A thesis submitted in fulfilment for
the degree of Doctor of Philosophy*

in

*School of Software
And
The Global Big Data Technologies
Centre (GBDTC)*

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

UNIVERSITY OF TECHNOLOGY SYDNEY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

SIGNATURE OF STUDENT

ACKNOWLEDGEMENTS

How time flies. It has been nearly four years since I first set foot on the University of Technology campus in Sydney, a place that I had been dreaming for years to go. In the precious semesters that followed, I acquired a great deal of knowledge. Indeed, I came to view myself as a tiny boat sailing into the light, away from storms and loaded to full capacity with information, skills, achievements, expectations and dreams. At the tranquil and solemn campus, there is a profound academic atmosphere which encourages rigorous study aspirations. My experience at the University has left me with many happy memories and a strong sense of gratitude in my heart.

Firstly, I would like to gratefully acknowledge A/Prof. Mao Lin Huang who is an expert in the field of information visualisation, and has earned a great reputation at home and abroad for his numerous contributions to scientific research and practical achievements.. This thesis directly grew out of a series of valuable discussions that I had with him.

Secondly, I would like to thank Dr. Quang Vinh Nguyen for guiding me in the various steps towards completing the thesis including research and writing. What deeply impressed me is not just the toil he underwent in numerous readings and modifications, but the wise and noble academic vision he possessed, which will benefit me a lot throughout my life. He brought me closer to the reality I had initially perceived, and eventually enabled me to grasp its rich complexity.

Furthermore, I owe sincere thanks to the members of the Visualisation Team and fellow researchers and the professors of The Global Big Data Technologies Centre (GBDTC). Their scholarly attainments, instructions, sense of humor, and patient help benefited me in so many ways.

This research also benefited tremendously from the assistance of many other researchers and staff within the University of Technology. In addition, I would also like thank the participants in the usability study for their cooperation and valuable feedback. Their assistance was as a key factor in providing the necessary

guarantees on the accuracy, objectivity, and precision of the datum I collected.

Last but not least, I owe my deepest gratitude to my brothers and sisters. They have given me so much support and their prayers have enabled me to go through the whole process with joy and peace.

Table of Contents

Table of Contents	v
Figure List	ix
Table List	xv
Equation List	xvi
Abstract	xvii
Chapter 1 Introduction	1
1.1 Multi-dimensional Data Visualisation (MDV).....	3
1.1.1 Basis of Data Properties	7
1.1.2 Multiple Views.....	14
1.1.3 Scatterplot matrix	16
1.1.4 Parallel Coordinate Plots	23
1.1.5 Other Visualisation Techniques.....	28
1.2 Visualisation Applications of MDV	32
1.2.1 Multi-dimensional Visualisation in Social Science	33
1.2.2 Multi-dimensional Visualisation in Geography.....	34
1.2.3 Multi-dimensional Visualisation in Aerospace.....	35
1.2.4 Multi-dimensional Visualisation in Medicine	36
1.3 Visualisation Optimisation Methodologies in Visualisation.....	38
1.3.1 Modelling Optimisation	39
1.3.2 Geometrical Optimisation	41
1.3.3 Aesthetic Optimisation.....	44
1.3.4 Functional Optimisation.....	46
1.3.5 Applicability Optimisation	47

1.4	Research Challenges	48
1.5	Research Objectives	50
1.6	Contributions	51
1.7	Thesis Organisation	53
Chapter 2 Visualisation Pipelines		56
2.1	Defining Terminology	57
2.2	Data Visualisation Pipelines	60
2.2.1	Data Formatting	61
2.2.2	Data Preprocessing	63
2.2.3	Visual Mapping	66
2.2.4	Human Perception	69
2.3	Visualisation Principles	70
2.4	Summary	70
Chapter 3 Optimisation of Scatterplot Matrix		72
3.1	Framework	73
3.1.1	Idea Evolution	74
3.1.2	Space Optimisation Process	76
3.2	Algorithms	79
3.2.1	Technical Specification	80
3.2.2	Implementation Algorithms	88
3.3	Interaction Mechanism	93
3.3.1	Introduction	94
3.3.2	Interaction Method	95
3.4	Evaluation	99
3.4.1	Performance Evaluation	100
3.4.2	User Studies	102
3.4.3	Supplementary Views	104

3.5	Summary	106
Chapter 4 Interactions with Scatterplot Matrix		
Visualisation		
107		
4.1	Dimensionality Reduction	111
4.1.1	Rough Set Theory	114
4.1.2	Variable Precision Rough Set	116
4.1.3	Feature Ranking	119
4.1.4	K-Means for Data Discretisation	120
4.2	Interactive Exploration	121
4.2.1	Point to Region Interaction	122
4.2.2	Linear Approximation of Decision Trend	124
4.2.3	Augmenting Class Coverage	127
4.3	Evaluation.....	129
4.3.1	Case Studies	130
4.3.2	Usability Study	138
4.4	Summary	141
Chapter 5 MDV in Forensic Visualisation		
142		
5.1	What is Forensic investigation?	144
5.1.1	Digital Evidence	145
5.1.2	Standards and Protocols	146
5.2	MDV Improves Forensic Investigation Models	147
5.2.1	Forensic Investigation Pipeline	148
5.2.2	Visualisation based Forensics' Model	149
5.2.3	A Case Study	152
5.2.4	Summary	155
5.3	MDV Assists Visualisation Hard Disk Drives' (HDDs') Investigation ..	156

5.3.1	Hard Disk Drive	157
5.3.2	Parallel Coordinates' on HDDs	161
5.4	MDV Works in Criminal Relationship Detection	165
5.4.1	Related Works	166
5.4.2	Self-Organizing Map	168
5.4.3	Our Approach.....	174
5.4.4	Experiment and results	179
5.4.5	Conclusion.....	185
5.5	MDV Helps in Crime Analysis	186
5.5.1	Tree Visualisation	187
5.5.2	DOITree Visualisation on Enron Email Dataset	189
Chapter 6 Conclusion and Future Work.....		192
6.1	Reflections on thesis questions	193
6.2	Answers to thesis questions	195
6.3	Future Work	197
6.3.1	Technological Optimisation.....	198
6.3.2	Systematic Scatterplot matrix Evaluation Guidance.....	199
6.3.3	Cooperate with Other Domains	200
Publication List		201
Reference		203

Figure List

FIGURE 1-1 THE CLASSIFICATION OF MULTI-DIMENSIONAL VISUALISATION 1

FIGURE 1-2 DATA VISUALISATION IN SOCIAL MEDIA - VISUALIZING ONLINE CONVERSATIONS; LEFT: A GROUP WITH A SINGLE DOMINANT MEMBER. RIGHT: A GROUP WITH MANY MEMBERS AT DIFFERENT LEVELS OF PARTICIPATION (DONATH 2002). 4

FIGURE 1-3 DATA VISUALISATION IN BIOCHEMISTRY: VISUALISATION OF STRUCTURAL DYNAMICS OF THE RIBOSOME BY DIFFERENT APPROACHES; A) PRESENT CRYO-EM MAP COLOURED ACCORDING TO LOCAL RESOLUTION AS DETERMINED BY RESMAP 2; B) PRESENT CRYO-EM MAP COLOURED ACCORDING TO THE B FACTORS OBTAINED FROM THE PSEUDO-CRYSTALLOGRAPHIC ATOMIC MODEL REFINEMENT (FISCHER ET AL. 2015)..... 4

FIGURE 1-4 DATA VISUALISATION IN NETWORK SECURITY: EIGHT VERTICAL AXES REPRESENTS A 2.5 CLASS B IP RANGE(SHIRAVI, SHIRAVI & GHORBANI 2012)..... 5

FIGURE 1-5 DATA VISUALISATION IN SCIENTIFIC PROJECT TRIDENT: IT TRANSFORMS RAW DATA FROM SENSORS INTO VISUALISATIONS AND DATA PRODUCTS.(BARGA ET AL. 2008)..... 5

FIGURE 1-6 DISPLAY THE NEWSPAPER ENDORSEMENTS IN THE US PRESIDENTIAL ELECTION IN 2004. SOURCE FROM: [HTTPS://FLOWINGDATA.COM/2008/10/29/MAP-SHOWS-NEWSPAPER-ENDORSEMENTS-IN-US-PRESIDENTIAL-ELECTION](https://flowingdata.com/2008/10/29/map-shows-newspaper-endorsements-in-us-presidential-election/) / 8

FIGURE 1-7 ORDINAL DATA VISUALISATION: A GRAPH DESCRIPTION OF THE USERS’ PREFERENCE. SOURCE FROM: 8

FIGURE 1-8 QUANTITATIVE DATA VISUALISATION: A) VISUALLY MONTHLY HEALTH DATA BY MEANS OF PENCIL ICONS ON A MAP (TOMINSKI, SCHULZE-WOLLGAST & SCHUMANN 2005); B) AN ANIMATED FLOW VISUALISATION CREATED FROM STREAMLINE IMAGES (VAN WIJK 2002)..... 9

FIGURE 1-9 A VISUALISATION SAMPLE FOR LINEAR STRUCTURE: IT SHOWS GLOBAL AVERAGE TEMPERATURE FROM FIVE DATASETS SINCE THE START OF THE SATELLITE TEMPERATURE DATA ERA IN 1979 THROUGH 2009. SOURCE FROM: [HTTP://APPINSYS.COM/GLOBALWARMING/LINEARTRENDS.HTM](http://appinsys.com/globalwarming/lineartrends.htm)..... 10

FIGURE 1-10 DATA TREE STRUCTURE. 11

FIGURE 1-11 AN EXAMPLE OF A SPACE OPTIMISED TREE VISUALISATION, THE DATASET CONTAINS OF 11

FIGURE 1-12 AN EXAMPLE FOR NETWORK STRUCTURE. DETECTING CRIMINAL ORGANIZATIONS THROUGH VISUALISATION

(ROBERTSON, MACKINLAY & CARD 1991).	12
FIGURE 1-13 MULTIPLE VIEWS OF THE 6D SAMPLE DATASET SHOWN IN TABLE 1	15
FIGURE 1-14 A SCATTERPLOT MATRIX VISUALISATION SAMPLE.....	16
FIGURE 1-15 SCATTERPLOT MATRIX VARIETIES – INFORMATION ENRICHMENT: THE UPPER-LEFT IS USING HISTOGRAM TO REPLACE THE PLOTS IN THE DIAGONAL; THE UPRIGHT IS USING DATA TREND TO REPLACE THE PLOTS IN THE DIAGONAL, AND USES DIFFERENT VISUAL MARKS TO EXPLAIN THE WHOLE DATASET; THE DOWN-LEFT IS A VISUALISATION METHOD CALLED HYPERSLICE (VAN WIJK & VAN LIERE 1993), WHICH IS A MATRIX OF PLOTS WHERE “SLICES” OF MULTIVARIATE FUNCTION ARE DISPLAYED AT A CERTAIN FOCAL POINT OF INTEREST; THE DOWN-RIGHT USES VARIABLE NAME TO REPLACE THE DIAGONAL PLOTS, AND USES COLOUR TO DISPLAY THE DATASET’S CATEGORY.	17
FIGURE 1-16 BRUSHED SCATTERPLOTS.....	18
FIGURE 1-17 GENERALIZED SCATTERPLOT MATRIX	19
FIGURE 1-18 POLYGON SCATTERPLOTS.....	19
FIGURE 1-19 DIMENSION REORDERING SCATTERPLOT MATRIX	20
FIGURE 1-20 3D SCATTERPLOT MATRIX SHOWING THE 8D “OLIVE OIL” DATASET. (SANFTMANN & WEISKOPF 2012)	21
FIGURE 1-21 GEOMETRY THEORY OF PARALLEL COORDINATE IN 2D	23
FIGURE 1-22 PARALLEL COORDINATES PLOT FOR THE CAR DATASETS	25
FIGURE 1-23 CAR DATASET VISUALISED BY ARC-BASED PARALLEL COORDINATES GEOMETRY (HUANG, LU & ZHANG 2015).....	26
FIGURE 1-24 CARS DATASET VISUALISATION IN PARALLEL COORDINATES WITH A NEW AXES RE-ORDERING METHOD (LU, HUANG & HUANG 2012).....	26
FIGURE 1-25 FORBES 94, A DATASET WITH 5 VARIABLES VISUALISED IN PARALLEL COORDINATE VISUALISATION: AFTER CLUTTER REDUCTION (LU ET AL. 2010).	27
FIGURE 1-26 EXAMPLE OF GRAPHICAL LOGICAL DIAGRAMS(TAYLOR ET AL. 2006).....	29
FIGURE 1-27 THE THEORY OF THE PIXEL-ORIENTED VISUALISATION TECHNIQUE(KEIM 2000).....	30
FIGURE 1-28 A VISUALISATION SAMPLE OF CIRCLE SEGMENTS TECHNIQUE (ANKERST 2001).	30
FIGURE 1-29 CHERNOFF FACES VISUALISATION FOR LAWYERS’ RATINGS OF TWELVE JUDGES.....	31
FIGURE 1-30 A VISUALISATION APPLICATION IN SOCIAL SCIENCE (GORBAN & ZINOVYEV 2010)	33
FIGURE 1-31 AN EXAMPLE OF MULTI-DIMENSIONAL VISUALISATION IN GEOGRAPHY – 3D TERRAIN PROFILE	34

FIGURE 1-32 AN EXAMPLE OF MULTI-DIMENSIONAL DATA VISUALISATION IN AEROSPACE (HURTER, TISSOIRES & CONVERSY 2010).....	35
FIGURE 1-33 THE EXAMPLES OF MULTI-DIMENSIONAL DATA VISUALISATION IN GENE ANALYSIS (PAVER); (A) LEVEL REPRESENTATION AND (B) ANALYSIS OF THE GLOBAL MRNA OF A BACILLUS SUBTILIS 168 GLUCOSE STARVATION EXPERIMENTANALYSE (OTTO ET AL): SOURCE FROM: HTTP://WWW.DECODON.COM/PAVER-BENEFITS.HTML	37
FIGURE 1-34 THE RATIONALE OF THE ARC COORDINATES PLANE.....	42
FIGURE 1-35 (UP) ORIGINAL PLOT (HAUCK ET AL.); (DOWN) AFTER CLUTTER REDUCTION. DATA RESULTS USED WITH KIND PERMISSION OF THE AUTHOR LIANG FU LV.	43
FIGURE 1-36 CIRCULAR TREEMAPS EXAMPLE, THE RED CIRCLES REPRESENT NEWER FILES WHILE SOFT YELLOW CIRCLES REPRESENT OLDER FILES. SOURCE FROM: HTTP://LIP.SOURCEFORGE.NET/CTREEMAP.HTML	44
FIGURE 1-37 AN EXAMPLE OF COMPUTER SCREEN THAT ACHIEVES THE MAXIMIZATION (100%) OF SPACE UTILISATION AND THE MINIMIZATION (0%) OF THE OVERLAP AMONG TWO SESSION DISPLAYS BY USING THE NEW TANGRAM TREEMAPS (LIANG.J 2015)	45
FIGURE 1-38 DATA CLUSTERING IN PARALLEL COORDINATE. CASE STUDY WITH WAGES DATASET OBTAINED FROM HTTP://WWW.NBER.ORG/CPS/	46
FIGURE 1-39 CHERNOFF FACES DISPLAY LIFE IN LOS ANGELES. SOURCE FROM:	47
FIGURE 1-40 THESIS STRUCTURE DESCRIPTION.....	53
FIGURE 2-1 VISUALISATION PROCESS.....	60
FIGURE 2-2 CROSSFILTER VISUALISATION TOOL DISPLAYS THE AIRLINE ON-TIME PERFORMANCE. SOURCE FROM: HTTP://SQUARE.GITHUB.IO/CROSSFILTER/	63
FIGURE 2-3 AN EXAMPLE OF A CLUSTERING HEAT MAP. THE ROWS ARE THE HIERARCHICALLY CLUSTERED GENES, WHILE THE COLUMNS ARE THE TISSUES WITH DENDROGRAMS. THE RED IN THE HEAT MAP PRESENTS UPREGULATION WHILE BLUE PRESENTS DOWNREGULATION. CITED FROM: HTTP://ALTANALYSE.BLOGSPOT.COM.AU/2012/06/HIERARCHICAL-CLUSTERING-HEATMAPS-IN.HTML	64
FIGURE 2-4 PCA OF A MULTIVARIATE GAUSSIAN DISTRIBUTION CENTERED AT (1,3) WITH A STANDARD DEVIATION OF 3 IN ROUGHLY THE (0.878, 0.478) DIRECTION AND OF 1 IN THE ORTHOGONAL DIRECTION. THE VECTORS SHOWN ARE THE EIGENVECTORS OF THE COVARIANCE MATRIX SCALED BY THE SQUARE ROOT OF THE CORRESPONDING EIGENVALUE, AND	

SHIFTED SO THEIR TAILS ARE AT THE MEAN, SOURCE FROM:

[HTTPS://EN.WIKIPEDIA.ORG/WIKI/PRINCIPAL_COMPONENT_ANALYSIS](https://en.wikipedia.org/wiki/Principal_Component_Analysis)..... 65

FIGURE 3-1 PLOT DRAWING-POSITION OF EACH VERTEX..... 77

FIGURE 3-2 WINE DATASET WITH FOUR ATTRIBUTES IS SHOWN IN DIFFERENT SIZE OF DISPLAY SCREEN (UP-LEFT) IS THE FULL SCREEN DISPLAY; (UP-RIGHT) IS THE 50% SCREEN DISPLAY;(DOWN-LEFT)IS THE 25% SCREEN DISPLAY; (DOWN-RIGHT) IS THE 12.5% SCREEN DISPLAY..... 77

FIGURE 3-3 TRADITIONAL SCATTERPLOT MATRIX LAYOUT..... 81

FIGURE 3-4 THE LAYOUT OF A MATRIX WITH PLOTS REDUCTION..... 82

FIGURE 3-5 A EXAMPLE OF A DRAW 84

FIGURE 3-6 A) THE EXAMPLE OF MOVING A DRAW; B) THE PROCESSES OF MOVING A DRAW 86

FIGURE 3-7 A) THE SAMPLE OF ROTATING A DRAW (B) THE PROCESSES OF ROTATING $\pi/4$ OF A DRAW 87

FIGURE 3-8 THE REPOSITION METHOD FOR SCATTERPLOT MATRIX 88

FIGURE 3-9 THE ALGORITHM OF LAYOUT OPTIMISATION APPROACH..... 89

FIGURE 3-10 THE ALGORITHM HIGHLIGHTING THE APPROACH TO REFLECTING THE VARIABLE RELATIONSHIPS..... 90

FIGURE 3-11 (A)(B): THE SAMPLE OF USING COLOUR PROPERTIES TO DISCOVER PAIRWISE VARIABLE RELATIONSHIPS. DATASET DESCRIPTION: THIS DATASET OBTAINED FROM UC IRVINE MACHINE LEARNING REPOSITORY, IT CONTAINS 12 VARIABLES, INCLUDING FIXED ACIDITY, VOLATILE ACIDITY, CITRIC ACID, RESIDUAL SUGAR, CHLORIDES, FREE SULFUR DIOXIDE, TOTAL SULFUR DIOXIDE, DENSITY, PH, SULPHATES, ALCOHOL, QUALITY; WEBSITES: 91

FIGURE 3-12 THE INTERACTION METHOD – USER QUERY 96

FIGURE 3-13 THE USER QUERY RESULT WITH COLOUR PROPERTY 96

FIGURE 3-14 THE USER QUERY RESULT WITH SHAPE PROPERTY..... 97

FIGURE 3-15 (A) (B) (C): EXAMPLES OF SHAPE PROPERTY WORKING ON THE LAYOUT OF THE DATASETS: A) REPRESENTS THE DATA POINTS CONNECTED INTO DIFFERENT SHAPES, AND THIS HELPS TO DISTINGUISH THE OVERALL DISTRIBUTION FOR EACH PLOTS; B) THE GREEN LINE IS THE DIAGONAL LINE OF EACH PLOT, AND FROM THE GAP BETWEEN THE OTHER LINES AND DIAGONAL LINE, IT PROVIDES ANOTHER OVERVIEW OF THE DATASET DISTRIBUTION WITHOUT ANY POINTS DISPLAYED; C) WE USE THIS VIEW TO DETECT THE DENSITY OF DATA POINTS DISTRIBUTION. IT IS MORE CONVENIENT TO FIND THE CENTRE POINT IN A CIRCLE SHAPE COMPARED WITH DRAWING IN A PLOT (RECTANGULAR), BUT AN IMPORTANT FACTOR IS

THAT THE CIRCLE SHOULD BE TANGENTIAL WITH EDGES OF EACH RECTANGULAR (PLOTS).....	98
FIGURE 3-16 (A) THE SAVING NUMBER OF SCATTERPLOTS B): THE SAVING SPACE RATE OF OUR NEW SCATTERPLOT MATRIX.....	101
FIGURE 3-17 THE RESULTS OF USER PREFERENCE	102
FIGURE 3-18 AN OVERVIEW COMPARISON OF PLOT ARRANGEMENTS IN THE NEW LAYOUT AND THE ORIGINAL LAYOUT	104
FIGURE 3-19 DIFFERENTIATION IN COLOUR; EXPERIMENT WITH ABOLONE DATASET	105
FIGURE 3-20 DIFFERENTIATION IN COLOUR AND SHAPE; EXPERIMENT WITH ABOLONE DATASET	105
FIGURE 4-1 A ILLUSTRATION OF SCATTERPLOT MATRIX VISUALISATION	122
FIGURE 4-2: INTERACTION (MOUSE CLICK) BY USING POINT-TO-REGION CONCEPT: THAT IS, A POINT CLICK CAUSES AN ENTIRE CONVEX HULL (A CLASS) TO BE HIGHLIGHTED.....	123
FIGURE 4-3 (A) A CLASSIC SCATTERPLOTS VISUALISATION. (B) ADDING THE DECISION FLOW WHERE PLOTS WERE AUGMENTED WITH RESPECT TO THE DECISION VARIABLE.....	125
FIGURE 4-4 VISUALISATION OF THE ENTIRE WINE DATASET USING A) PARALLEL COORDINATE AND B).....	131
FIGURE 4-5: RESULTS OBTAINED FROM THE CASE STUDY WITH WINE DATA. THE UPPER DIAGONAL	133
FIGURE 4-6: CASE STUDY WITH WINE DATASET OBTAINED FROM [31]. BOXES AT “WITHOUT-TREND” AREA (AREA ABOVE THE DIAGONAL LINE) ARE SCATTERPLOTS OF EACH PAIR OF ATTRIBUTES WHILE BOXES AT “WITH-TREND” AREA (AREA BELOW THE DIAGONAL LINE) REPRESENT THE SAME VALUES AND WITH CHANGING TRENDS.....	134
FIGURE 4-7: CASE STUDY WITH CAR DATASET. WE SELECTED MILEAGE PER GALLON (MPG) AS THE DECISION AND THE DATASET HAS BEEN REDUCED TO 4 ATTRIBUTES NAMEDLY ACCELERATION, DISPLACEMENT, CYLINDERS AND HORSEPOWER.....	136
FIGURE 4-8: CASE STUDY WITH WAGES DATASET. THIS FIGURE SHOWS A BOX AT “WITH-TREND” AREA FOR CORRELATION BETWEEN <i>EXPERIENCES</i> , <i>WAGE</i> (Y-AXIS) AND <i>AGE</i> (X-AXIS).....	137
FIGURE 4-9: ACCURACY OF PARALLEL COORDINATE AND SCATTERPLOT MATRIX VISUALISATIONS CORRESPONDING TO FIVE QUESTIONS (WITH 95% CONFIDENCE INTERVALS).....	139
FIGURE 5-1: A VISUALISATION TECHNIQUE BASED FORENSIC INVESTIGATION MODEL	149
FIGURE 5-2 NEW MODEL FOR HDD VISUALISATION.....	152
FIGURE 5-3: DATA STORAGE CATEGORY.....	157
FIGURE 5-4: THE VISUALISATION OF ORIGINAL METADATA BY PARALLEL COORDINATE	163
FIGURE 5-5 USE PARALLEL COORDINATE TO IDENTIFY THE SUSPICIOUS FILE IN HDD.....	163

FIGURE 5-6: THE THEORY OF THE SELF ORGANIZING MAP (VESANTO & ALHONIEMI 2000)..... 169

FIGURE 5-7: THE DISTANCE BETWEEN TWO NEURONS..... 171

FIGURE 5-8: AN OVERVIEW OF THE WHOLE 16 ITEMS' FEATURE VALUE AND THE CIRCLED AREAS ARE THE SUSPICIOUS DATA. 180

FIGURE 5-9: THE DIFFERENCE BETWEEN ITEM 1, ITEM 4 AND ITEM 16. 181

FIGURE 5-10: THE WEIGHT PLANES OF ALL 16 ITEMS BY SOM 182

FIGURE 5-11: THE CLASSIFICATION HITS AFTER USING SOM OF 16 ITEMS 183

FIGURE 5-12: THE VISUALISATION OF CLASSIFICATION OF ALL FEATURES..... 183

FIGURE 5-13: EXAMPLES OF THE EMAILING ACTIVITIES FOR EIGHT EMPLOYEES IN ENRON. A) CHECK THE ACTIVITIES OF
BAUGHMAN-D. B) USING COLOUR IN DOITREE 190

Table List

TABLE 1-1	A 6D SAMPLE DATASET	14
TABLE 2-1	CANONICAL DATA MODEL - WHITE WINE QUALITY DATASET	61
TABLE 2-2	VISUAL MARKS.....	66
TABLE 2-3	VISUAL PROPERTIES OF MARKS	67
TABLE 5-1	NAME AND TYPE	153
TABLE 5-2	NAME AND REMARK.....	153
TABLE 5-3	TIME ATTRIBUTES.....	159
TABLE 5-4	DISK INVESTIGATOR.....	162

Equation List

EQUATION 1	THE NUMBER OF PLOTS.....	83
EQUATION 2	THE NUMBER OF PLOTS PER ROW	83
EQUATION 3	THE NUMBER OF PLOTS PER COLUMN	83
EQUATION 4	THE RELATIONSHIP BETWEEN THE NUMBER OF SCATTERPLOTS AND THE NUMBER OF VARIABLES	100
EQUATION 5:	THE INDISCERNIBILITY RELATION AMONG OBJECTS	114
EQUATION 6:	THE β POSITION REGION IN VPRS MODEL.....	116
EQUATION 7:	THE QUALITY OF CLASSIFICATION IN VPRS MODEL.....	117
EQUATION 8:	CALCULATE THE RANKING COEFFICIENT FOR EACH CONDITIONAL ATTRIBUTE	119
EQUATION 9:	THE CALCULATION OF THE SLOPE TO MEASURE THE CHANGE	126
EQUATION 10:	THE CALCULATION OF THE INTERCEPT.....	126
EQUATION 11:	INTERPOLATE THE BEST FITTING LINE AT POINT (X_0, Y_0)	126
EQUATION 12:	CALCULATE THE ACCURACY OF A RULE.....	127
EQUATION 13:	CALCULATE THE COVERAGE OF A RULE.....	128
EQUATION 14	CALCULATION OF \hat{X}	169
EQUATION 15	CALCULATION OF \hat{W}_j	169
EQUATION 16	EUCLIDEAN DISTANCE CALCULATION.....	170
EQUATION 17:	MEASURE THE DISTANCE BETWEEN TWO OBJECTS.....	171
EQUATION 18:	CALCULATE THE DISTANCES BETWEEN EVERY TWO OBJECTS.....	172
EQUATION 19:	CALCULATE THE MINIMUM DISTANCE AMONG ALL DISTANCES	172
EQUATION 20:	CERTIFY THAT THE WINNER NEURON IS THE LARGEST SCALAR PRODUCT	172
EQUATION 21:	RECOGNISE THE WEIGHT.....	173

Abstract

Multi-dimensional Data Visualisation (MDV) is the technique to generate visual presentations of datasets with more than three features (or attributes). These graphic representations of data and associated data features can facilitate human comprehension, extraction of implicit patterns and discovery of the relationships among numerous data items for visual data analysis.

Although many optimisation methods have been proposed in the past to improve the visual data processing, not many have been applied to MDV. In particular, little research work has been done in the field of display space optimisation. This thesis focuses on the optimisation of two popular Multi-dimensional visualisation methods: 1) scatterplot matrix and 2) parallel coordinate plots, visualisation by using unique approaches to achieve display space optimisation and interaction.

The first contribution of the thesis is proposing a new visualisation approach named the *Spaced Optimised Scatterplot Matrices* that achieves the maximization of the display space utilisation through position transferring. Breaking through the limitation of discovering the pairwise variable relationships, the new method is able to explore the influences of a single variable towards others. In addition, our algorithms improve the efficiency of interactive Multi-dimensional data visualisation significantly, through the reduction of the computational cost.

The second contribution of the thesis is to improve the parallel coordinate plots and apply it to the computer forensic investigation. As we are living in a big data era, it is much harder for the researchers to provide accurate evidence for victims within a certain time frame. Our research shows that visualisation techniques can improve the working efficiency of investigations in certain cases.

To conclude, we propose a concept of a space optimised Scatterplot Matrix(SPM) visualisation technique considering the shortcomings of the existed SPM and parallel coordinates in Multi-dimensional visualisation research area. In the meantime, to demonstrate the necessity of our research methodologies, we apply them into computer forensics, which is an area needed analyzing abilities with higher accuracies and efficiencies. By the tests on using Parallel Coordinates and DOITrees, the forensic specialists can easily discover the necessary information in different cases. In the future, we plan to improve our space optimised scatterplot matrix technique from technological optimisation and the broaden application aspects. For example, dealing with the visualisation coolusion problem, non-trivial computation time issue, etc; We will also do the investigations to enlarge the development and availability of Multi-dimensional visualisation techniques.

Chapter 1 Introduction

THIS CHAPTER AIMS to show that one of the biggest challenges in Multi-dimensional data visualisation is to find proper representations of multivariate data that can display the complex structure clearly and effectively (Theus 2008). Various graphics such as the scatterplot matrix, parallel coordinate plots and star plots have been developed during the last two decades. Some of these techniques will be introduced in the following sections. We will especially concentrate on their strengths, weaknesses, and the domains which they are working in. Figure 1-1 is a general scope of Multi-dimensional visualisation.

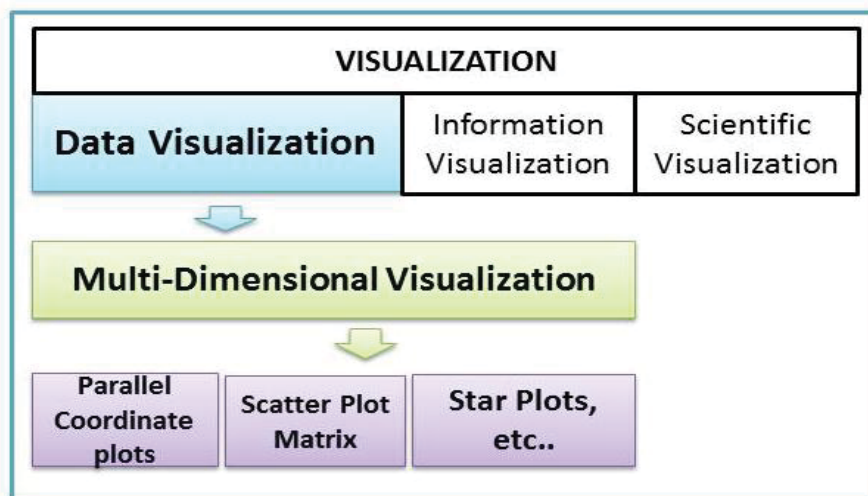


Figure 1-1 the classification of Multi-dimensional Visualisation

Following the main purpose of Chapter 1, Section 1.1 opens the research domain and directs the users' attention to Multi-dimensional data visualisation. Following this, Section 1.2 describes the domains that have benefited from visualisation techniques. Section 1.3 moves to a summary of the methodologies

that can be used for improving visualisation techniques; Section 1.4 addresses the research challenges of this thesis research. To meet the challenges, Section 1.5 briefly presents the research objectives, followed by an overview of our contributions in section 1.6. Finally, the thesis organization is given in Section 1.7.

1.1 Multi-dimensional Data Visualisation (MDV)

Data Visualisation is a specific domain to visually represent the information contained only numbers and letters, and the main goal is to improve the communication between user and data. The definition of data visualisation is:

“The use of computer-supported, interactive, visual representations of data to amplify cognition (Card, Mackinlay & Shneiderman 2009)”

Data Visualisation is slightly different from other visualisation related domains, and it is necessary to differentiate them as this might provide clearer guidance for researchers in each domain. Take *information visualisation* and *science visualisation* as an example. The term *information visualisation* (Keim et al. 2010) is generally applied to the visual representation of large-scale collections of non-numerical information, such as the files and lines of code in software systems, library and bibliographic databases, networks of relations on the internet, and so forth. While *scientific visualisation* is primarily concerned with the visualisation of 3-D phenomena (architectural, meteorological, medical, biological, etc.), where the emphasis is on realistic renderings of volumes, surfaces, illumination sources, and so forth, perhaps with a dynamic (time) component.

As we mentioned above, *data visualisation* (Yi et al. 2008) is the way to gain insight and clear results in a graphic format from massive and complex datasets. It extracts the essential information that can produce effective actions in real-time situations. It has been successfully applied in a variety of fields

including: *social media, biochemistry, network security, environmental science* (Figures 1-2, 1-3, 1-4, 1-5), and many other disciplines.

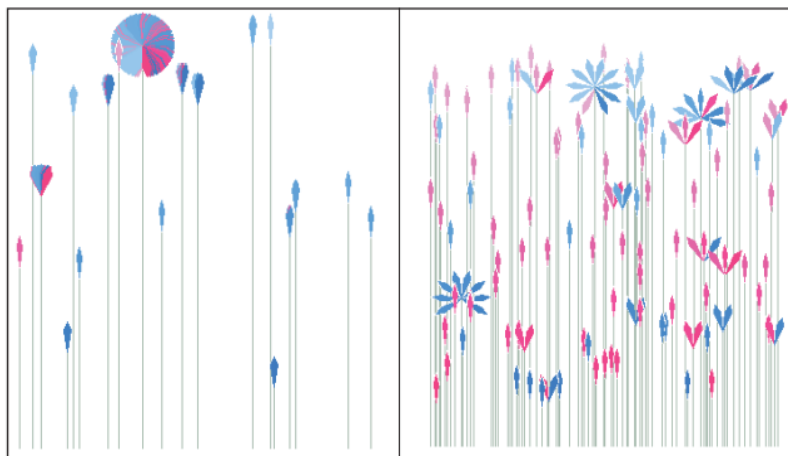


Figure 1-2 Data Visualisation in Social Media - Visualizing online Conversations; Left: a group with a single dominant member. Right: a group with many members at different levels of participation (Donath 2002).

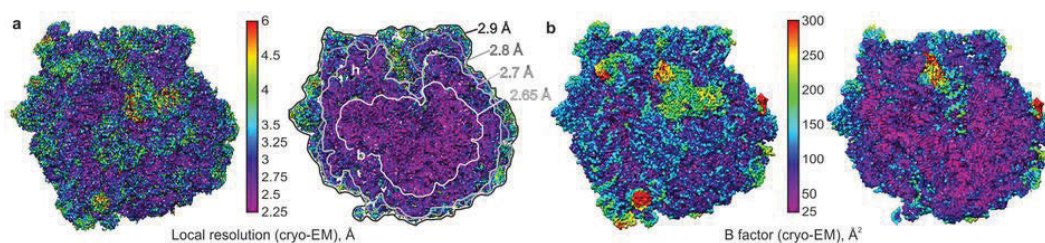


Figure 1-3 Data Visualisation in Biochemistry: Visualisation of structural dynamics of the ribosome by different approaches; a) Present cryo-EM map coloured according to local resolution as determined by Resmap 2; b) Present cryo-EM map coloured according to the B factors obtained from the pseudo-crystallographic atomic model refinement (Fischer et al. 2015).

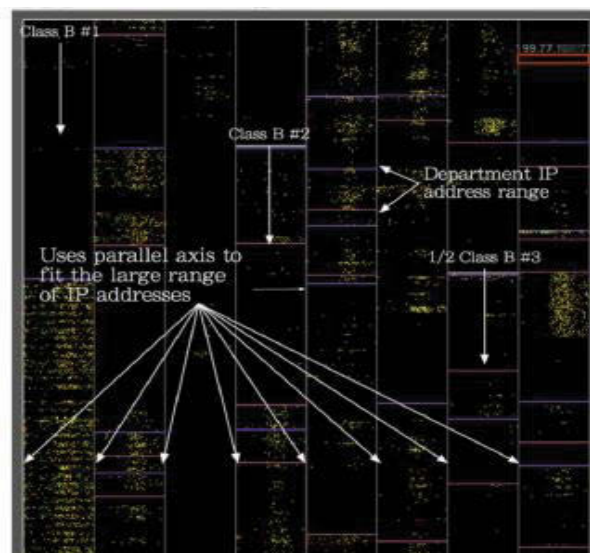


Figure 1-4 Data Visualisation in Network Security: eight vertical axes represents a 2.5 Class B IP range(Shiravi, Shiravi & Ghorbani 2012).

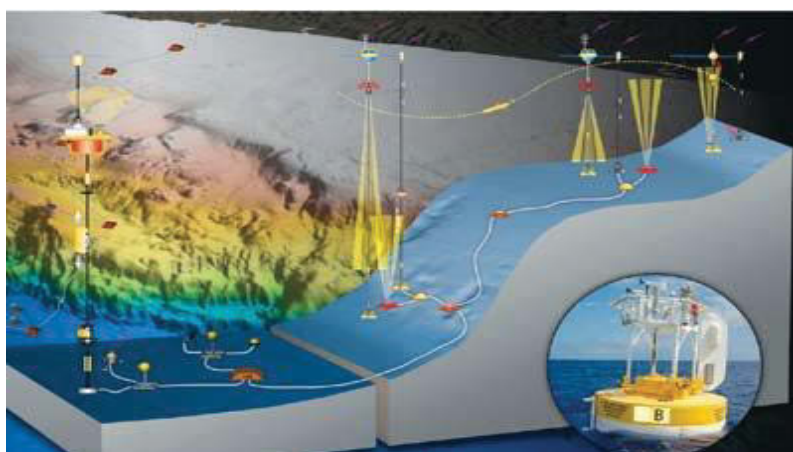


Figure 1-5 Data Visualisation in Scientific Project Trident: it transforms raw data from sensors into visualisations and data products.(Barga et al. 2008).

Specifically, the visualisation of high dimensional data has become an important branch in data visualization. One main reason is the data amounts getting larger and larger today, and it is much easier for human to use their natural well-developed ability to explore and detect information in visual patterns.

During the past two decades, many techniques have been developed for

Multi-dimensional data display. This section will outline and provide samples of some of the proposed approaches. We will introduce them from the simplest method such as multiple views to the geometric-based methods, for example, the scatterplot matrix, parallel coordinate plots and so forth.

To create proper visual representations for different datasets, it is necessary to get to know the data beforehand in Section 1.1.1. Understanding how the original data are transformed into graphics is also essential, as described in Section 1.1.2; then the most popular Multi-dimensional data visualisation techniques will be introduced in Section 1.1.3, Section 1.1.4 and Section 1.1.5 respectively.

1.1.1 Basis of Data Properties

Real world data is typically complicated, noisy, and is always large in volume. This section is about being familiar with the data. We will introduce the following properties of data: the features (types) of data; the structures of data; and the variables of data.

1.1.1.1 Features of Data

A feature is representing a characteristic of a data object. The type of a feature is according to the possible values in terms of three categories: **Nominal data** which has no inherent order and contains the symbols or name of things such as { *Sydney, Perth, Melbourne...* } { *Diabetes, Lung Cancer, Leukemia, Heart Disease...* } etc.. **Ordinal data** has some rankings among them but no measurable intervals. For example, { *first, second, third,...* } { *Small, Medium, Large* } { *Jan, Feb, Mar, Apr,...* } etc.. **Numerical data** refers to the data that can be measured and usually contains Interval data. Such as { *Latitude/Longitude, Time (event), etc.* }, and Ratio data, { *Length, Count, Time (duration), etc.* }.

Here are various displays for different types of data. Figure 1-6 is an example of the visualisation of **nominal data**. The map shows newspaper endorsements in the US Presidential Election. Data used in this map is from the Editor & Publisher's list. The divided areas with light blue and red colours represent states in the US; circles with light purple are newspapers that endorse Obama and John Kerry; While the light red ones show McCain/Bush endorsements; and the circle size is the newspaper's circulation.

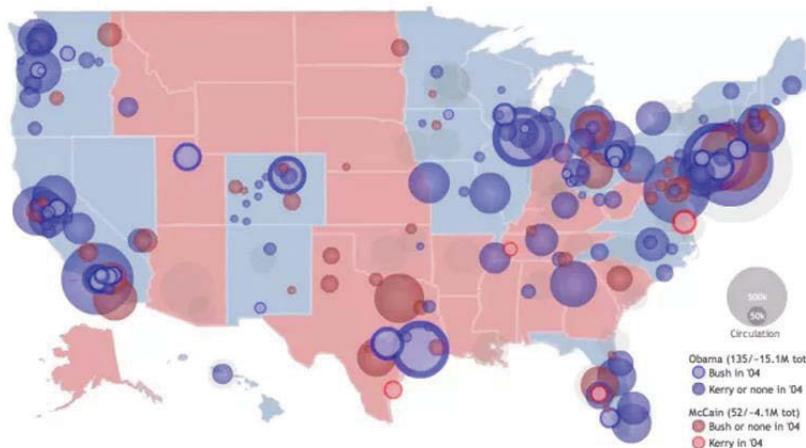


Figure 1-6 Display the newspaper endorsements in the US Presidential Election in 2004. Source from:

<https://flowingdata.com/2008/10/29/map-shows-newspaper-endorsements-in-us-presidential-election/>

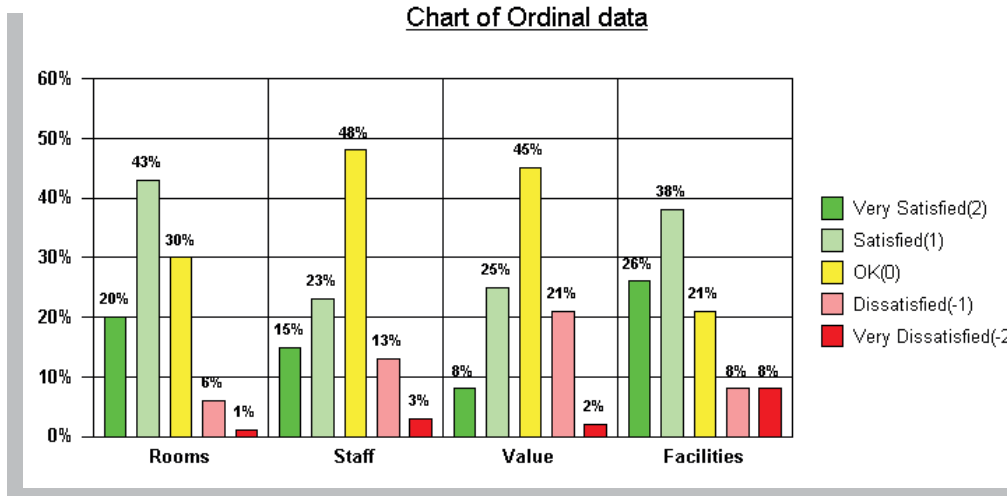


Figure 1-7 Ordinal Data Visualisation: A graph description of the users' preference. Source from:

<http://www.snapsurveys.com/blog/wp-content/uploads/2011/08/chart-of-ordinal-data.bmp>

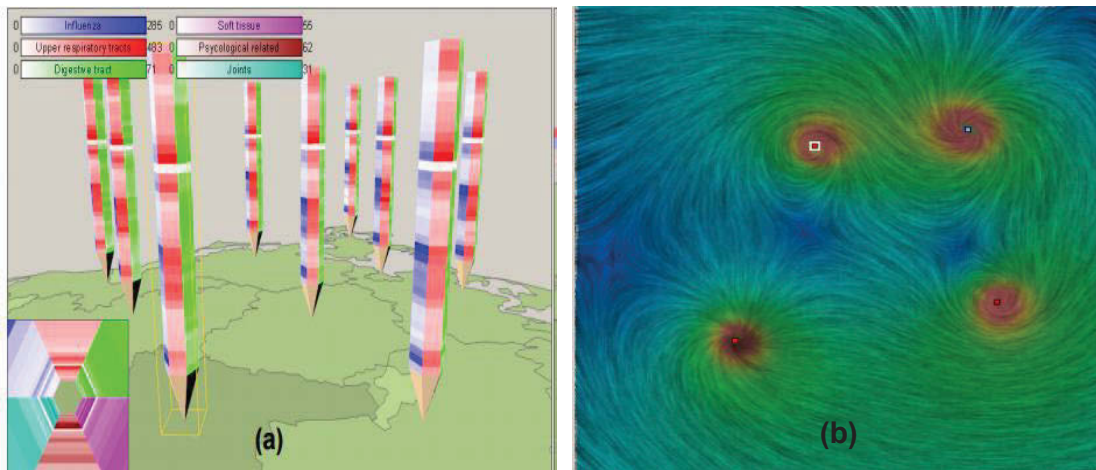


Figure 1-8 Quantitative Data Visualisation: a) Visually monthly health data by means of Pencil icons on a map (Tominski, Schulze-Wollgast & Schumann 2005); b) an animated flow visualisation created from streamline images (Van Wijk 2002).

Figure 1-7 is an expression of the **ordinal data** where the colour represents the users' preference toward the evaluation standard. Figures 1-8 (a) (b) describe the visualisation on the **numerical data**. Particularly, Figure 1-8 a) is representing cases of six diseases by Pencil icons, and the face of the pencil will be changed if its position is moved; Figure 1-8 b) is generated by a user interface flow modeling and visualisation system. The user can interactively add and remove flow elements and change their properties.

1.1.1.2 Structures of data

We divide the structure of data into three categories in the aspect of visualisation: *Linear structure*; *Tree structure* and *Network structure*.

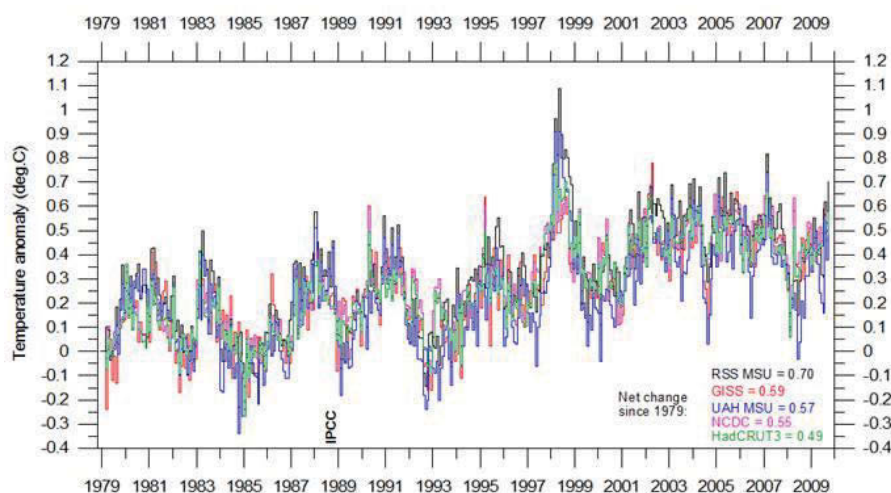


Figure 1-9 A Visualisation sample for Linear Structure: it shows global average temperature from five datasets since the start of the satellite temperature data era in 1979 through 2009. Source from: <http://appinsys.com/globalwarming/LinearTrends.htm>

As for data with a **linear structure**, the data elements have a linear relationship followed by one other, refer to Figure 1-9. **Tree structure** is a dataset with hierarchical relationship among objects, and the relationship is usually described by lines. Figure 1-10 is a simple illustration of **hierarchical structure**. The last category is **network data structure**. Rather than having a linear relationship in the objects, non-linear correlations can be discovered through the entire dataset. Finally, we give two examples of the visualization results with different data structures, see Figure 1-11 and Figure 1-12. Figure 1-11 is a data representation sample, named SO-Tree (Nguyen & Huang 2003a). This Tree Visualization was designed especially for visualizing and manipulating very large hierarchies. While Figure 1-12 explains a clear network structure visualization in computer forensic. By this result, the forensic investigators can deeply discover complicated relationships within criminal organizations.

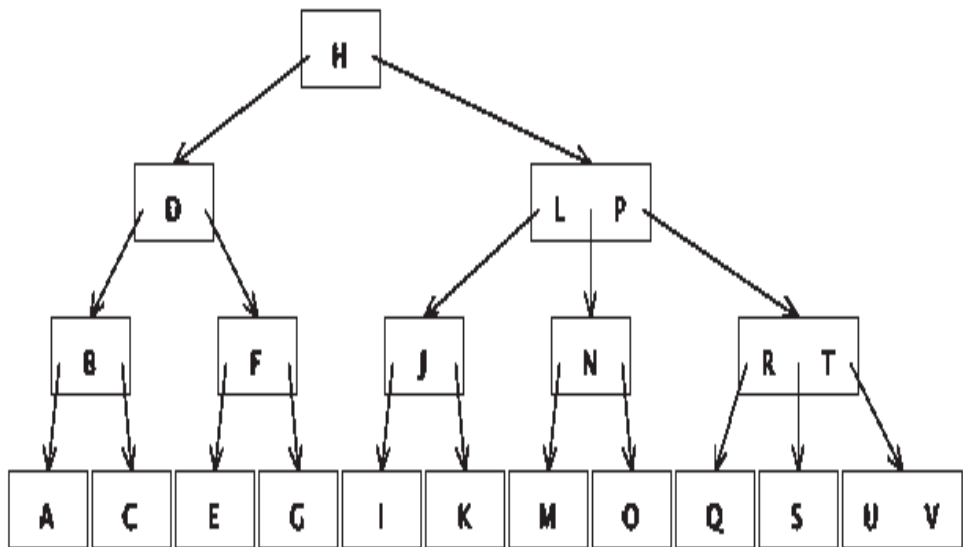


Figure 1-10 Data tree structure.

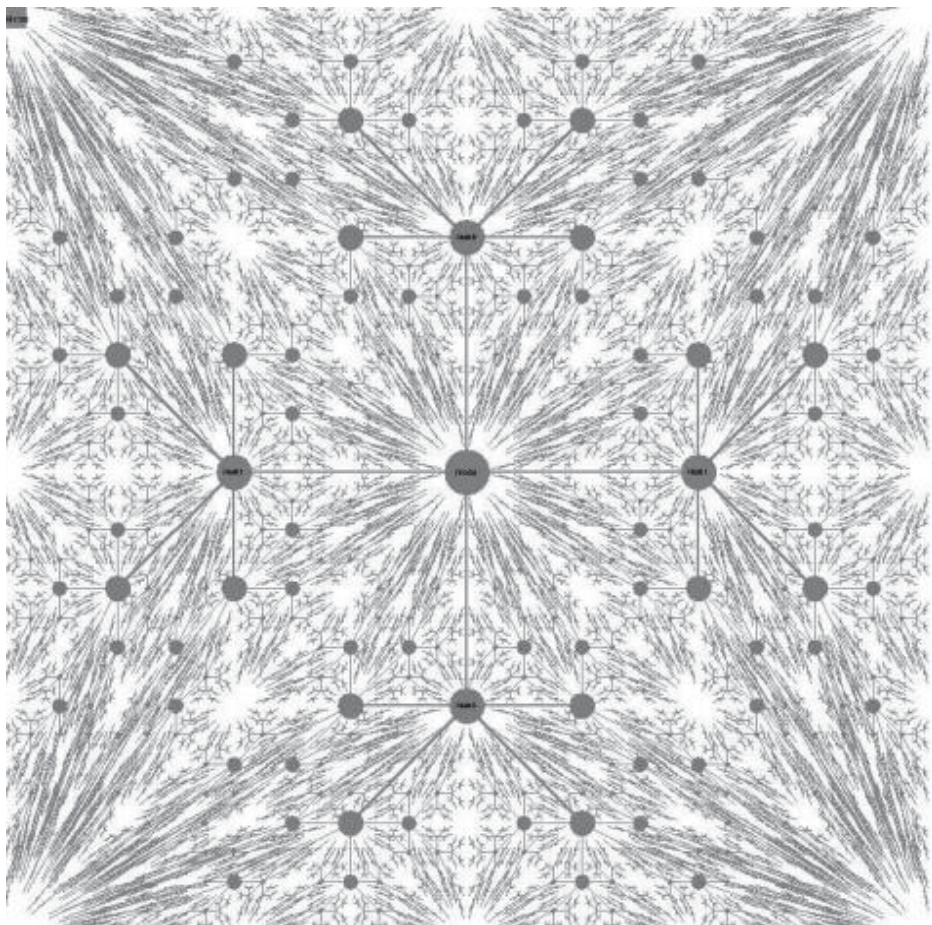


Figure 1-11 An example of a space optimised tree visualisation, the dataset contains of approximately 22000 nodes, and the running time of the method is 5 minutes 10 seconds

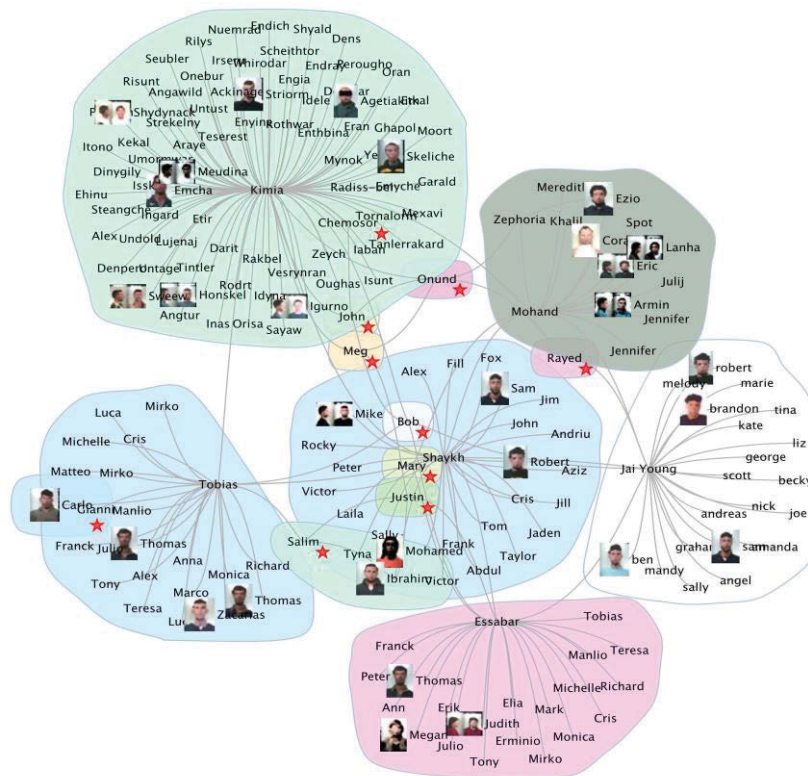


Figure 1-12 An example for Network Structure. Detecting criminal organizations through visualisation (Robertson, Mackinlay & Card 1991).

1.1.1.3 Encoding of Data

Each data consists of one or more quantitative variables for each individual. In data visualization, considering the number of variables, a dataset can be divided into four categories: *Univariate data*, *bivariate data*, *trivariate data* and *hypervariate (Multi-dimensional) data*.

Univariate dataset can be a single number or a collection of numbers. **Bivariate data** represents the data with two variables such as scatterplot visualisation. Data with three variables is called **trivariate data**. For example, the 3D scatterplot. The **hypervariate data**, also named **Multi-dimensional data**, always contains more than three variables. Significantly, dealing with the

Multi-dimensionality of data has been challenging for researchers for many years owing to the difficulty in comprehending more dimensions as well as the computational overhead.

1.1.2 Multiple Views

Multiple Views is the simplest method to display multivariate datasets. It requires one single view for each variable, and the number of the final views is equal to the number of variables per dataset. Take Table 1-1 as an example, it is a six dimension dataset, which can be simply represented by Figure 1-13. However, this approach is not satisfied when the number of variables is increased to a certain domain. One reason is that the information might be blurred, another reason is the relationships among each variable cannot be identified through this method as well.

Table 1-1 A 6D sample dataset

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	Sulfur Dioxide
7	0.27	0.36	20.7	0.045	45
6.3	0.3	0.34	1.6	0.049	14
8.1	0.28	0.4	6.9	0.05	30
7.2	0.23	0.32	8.5	0.058	47
7.2	0.23	0.32	8.5	0.058	47
8.1	0.28	0.4	6.9	0.05	30
6.2	0.32	0.16	7	0.045	30
7	0.27	0.36	20.7	0.045	45
6.3	0.3	0.34	1.6	0.049	14

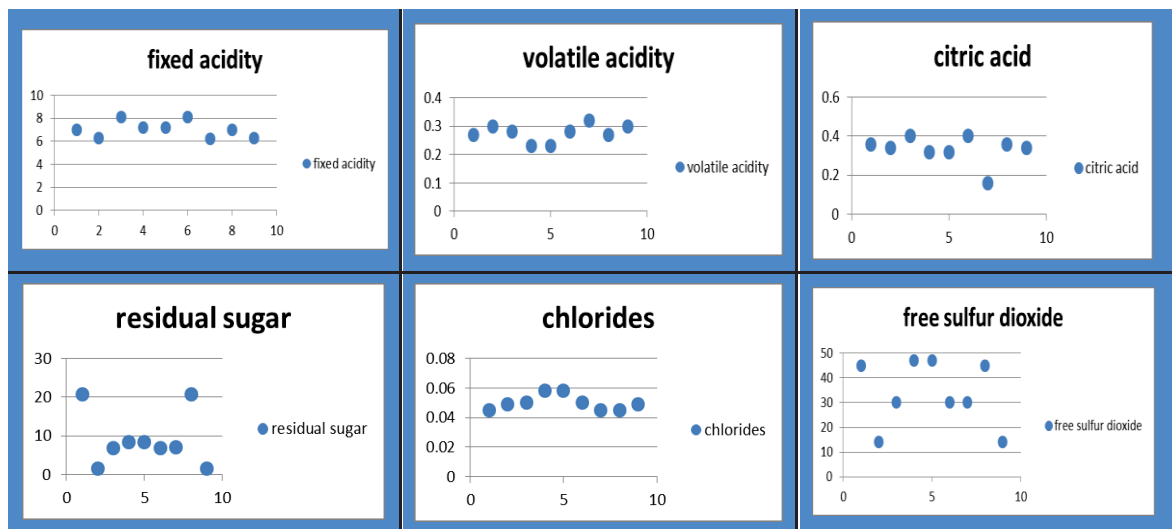


Figure 1-13 Multiple views of the 6D sample dataset shown in Table 1

1.1.3 Scatterplot matrix

Scatterplot matrix (Carr et al. 1987; Oja, Sirkiä & Eriksson 2006; Pham et al. 2014) is a frequently applied Multi-dimensional visualisation method to explore the visual presentations of multivariate data. It can roughly display the linear correlations between multiple variables. Specifically, the scatterplot matrix can be helpful in pinpointing specific variables in a large dataset that might have similar correlations to the dataset.

The basic concept of the scatterplot matrix is described below. Give a set of n variables $\{X_1, X_2, \dots, X_{n-1}, X_n\}$, the scatterplot matrix contains all the pairwise scatterplots of the variables on a single panel in a matrix format. That is, if there are n variables, the scatterplot matrix will have n rows and n columns. The i^{th} row and the j^{th} Column of this matrix is a plot of $X_i * X_j$. This technique has been increasingly common in graphical tools. Figure 1-14 is a display of the scatterplot matrix.

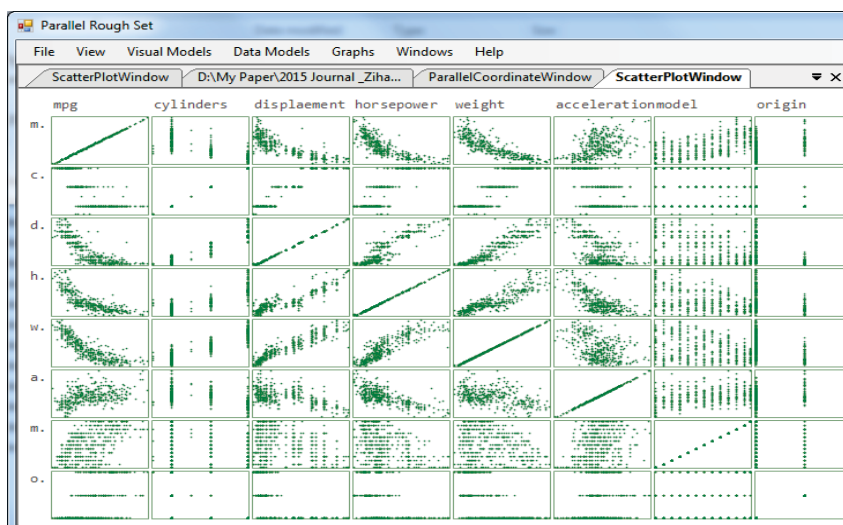


Figure 1-14 A Scatterplot matrix Visualisation Sample

As this technique has been in use many years ago (Andrews 1972; Cleveland & McGill 1985), several variations on the representation of scatterplots have been proposed. In the first variation, researchers have reused these plots to display more information and then enriched the matrix for viewers; all related varieties are shown in Figure 1-15.

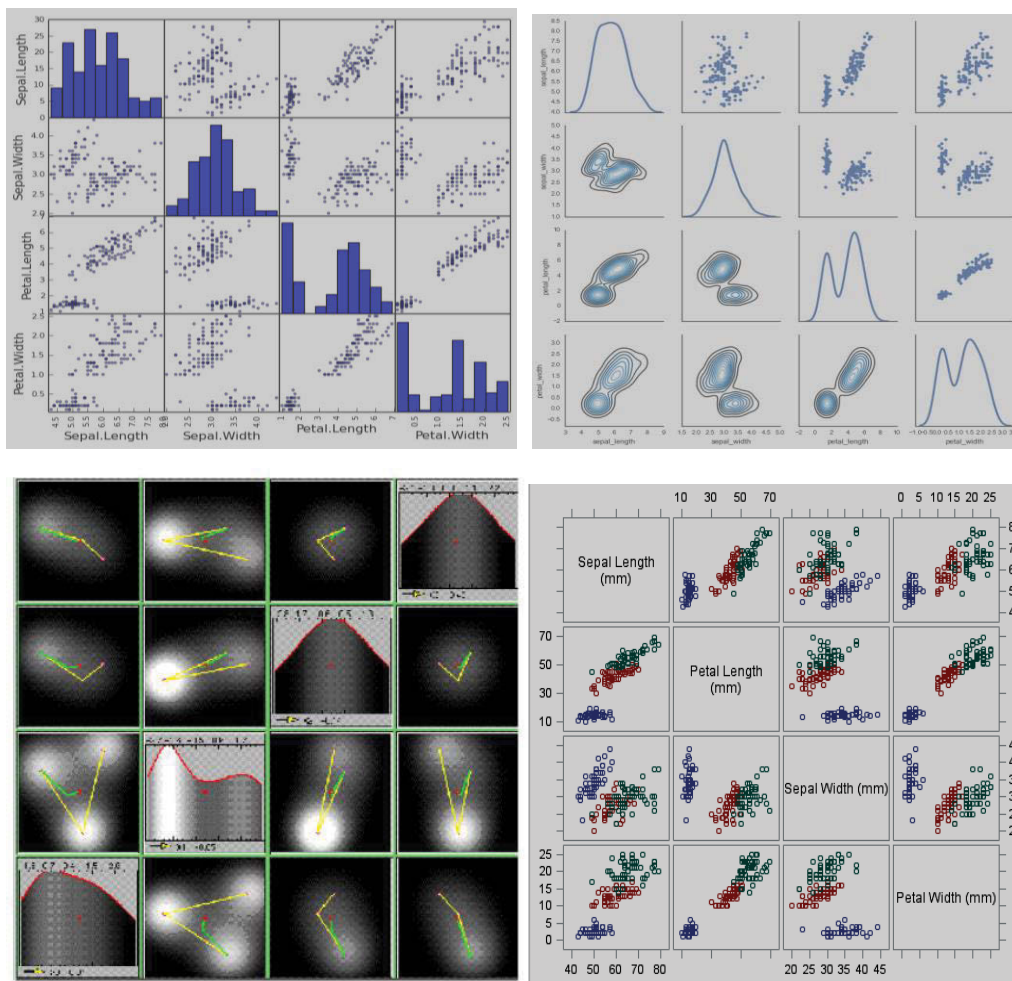


Figure 1-15 Scatterplot matrix Varieties – information enrichment: the upper-left is using histogram to replace the plots in the diagonal; the upright is using data trend to replace the plots in the diagonal, and uses different visual marks to explain the whole dataset; the down-left is a visualisation method called hyperslice (Van Wijk & Van Liere 1993), which is a matrix of plots where “slices” of multivariate function are displayed at a certain focal point of interest; the down-right uses variable name to replace the diagonal plots, and uses colour to display the dataset’s category.

Secondly, the variation is considering the limitation on plots' scalability. The technique is called brushing (Becker & Cleveland 1987), which is a dynamic graphical method to interact with each scatterplot in real time by a screen input device, see Figure 1-16. In detail, when the mouse is brushing over a certain scatterplot, the related data appears simultaneously on all the other scatterplots.

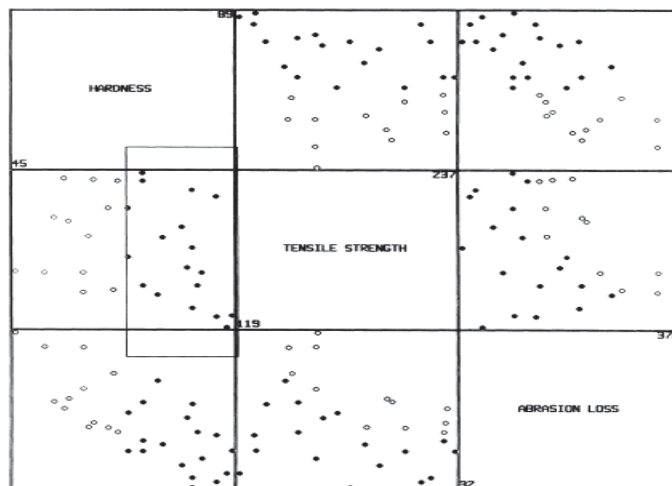


Figure 1-16 Brushed Scatterplots

The third variety is proposed to solve the problem of overlapping, which directly affects the value display. This method is called generalized scatterplots (Im, McGuffin & Leung 2013), it allows an overlap-free representation of large datasets to fit entirely into the display, See Figure 1-17. This representation provides the user with many different views for revealing patterns and relationships within the data.

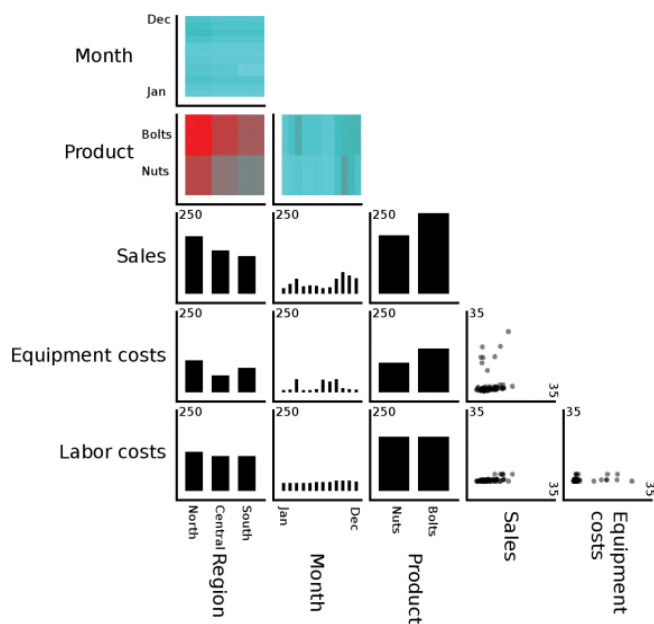


Figure 1-17 Generalized Scatterplot matrix

Polygon Scatterplots (Cleveland 1988) is another variety of the scatterplot matrix, refer to Figure 1-1. The idea is to transfer the different scatterplots to separating views with polygon. This dimensional transformation can extend the ability of visualisation. For example, if the dataset is four dimensions, it is to take a scheme for three dimensions and apply it to every choice of three out of four coordinates.

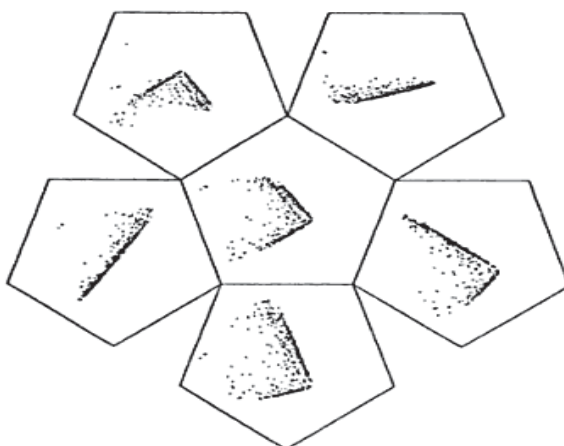


Figure 1-18 Polygon Scatterplots

The clutter problem in multivariate data visualisation is another challenge for researchers. A dimension reordering method (Albuquerque et al. 2009) was proposed in the scatterplot matrix to deal with the crowded and disordered visual entities that obscure the structure in visual displays. The dimension reordering method is the concept of calculating the similarity between dimensions, discussing several similarity measures and proposing a method to arrange dimensions. An example is given in Figure 1-19.

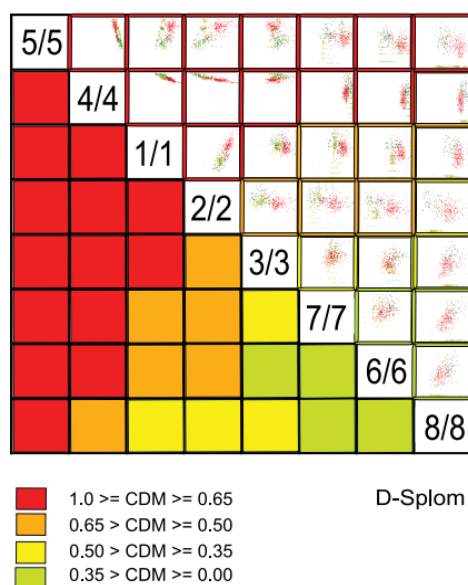


Figure 1-19 Dimension Reordering Scatterplot matrix

The fifth variety was motivated by the navigation method. Harald Sanftmann et al. (2012) extended the approaches to 3D axes by swapping one or two axes during transitions. Figure 1-20 displays the 8D oil dataset in 3D scatterplot views, where the third dimension of the data is mapped to the y -axis and all 2D projections of the 3D scatterplot matrix that preserve the y -axis mapping are projected to the back face perpendicular to the

y - axis of the cube.

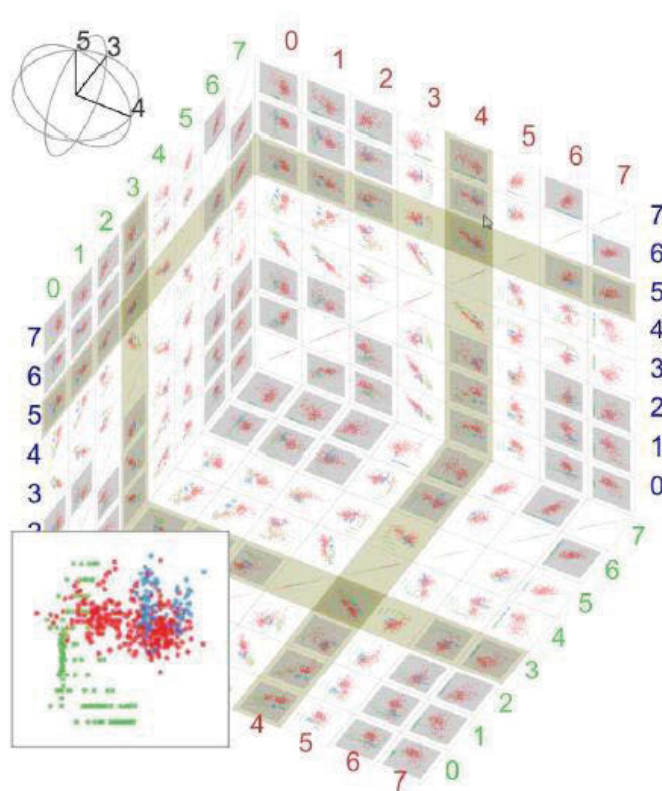


Figure 1-20 3D scatterplot matrix showing the 8D “olive oil” dataset. (Sanftmann & Weiskopf 2012)

Other main varieties of the scatterplot matrix are briefly described here. N-vision (Feiner & Beshers 1990) , a method which is similar with hyperslice, the matrix panel accommodates interactive exploration of a multivariate function. Prosection Matrix (Spence et al. 1995) is a method to support similar tasks where a set of parameter values must be chosen to lead to acceptable artifact performance, which is more suitable for data mining. Hyperbox (Alpern & Carter 1991) uses a parallelogram for every pair of dimensions, it displays simultaneously all pairwise relationships of Multi-dimensional datasets.

In summary, either the traditional scatterplot matrix or its varieties are all widely served in different domains. They have some common advantages:

- they can be easily combined with other visualisation and interaction mechanisms, like linking & brushing (Becker & Cleveland 1987).
- they are simple to evaluate, e.g., w.r.t. bivariate correlations, classifications, clusters, or trends.
- the experienced user is able to form hypotheses about multivariate relations between different dimensions of the underlying dataset.
- they are simple to implement, intuitively interpretable and also appropriate for inexperienced users.

The limitations of the scatterplot matrix cannot be ignored. One of the main limitations is the scalability. Although it can explain all of the pairwise correlated information, the display space restricts the number of dataset to at most 20 dimensions or less than 1000 records (testing by my own experiments), otherwise the display gets blurred and problems of navigation will also be experienced. For example, when we scroll up and down on the graphs, difficulties always appear in grasping target items.

1.1.4 Parallel Coordinate Plots

Parallel coordinate plots, is one of the most popular Multi-dimensional data visualisation techniques. As a geometrical-based approach, it represents the dimensions of datasets in terms of different axes.

In Section 1.1.4.1, we will introduce the fundamental theory of parallel coordinate plots, and then some improvements with samples will be given in Section 1.1.4.2.

1.1.4.1 Geometry Theory

Points on the plane are represented by lines and two points can determine a line. Take a point in two dimensional as an example: There are four points $\{a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4\}$ on X_1 and X_2 respectively. These lines can be mapped onto a line with different points. In this case, the point p is not just a segment, but a whole line: $l_p \rightarrow \infty$. Therefore, the transforming processes can be described as below in Figure 1-21.

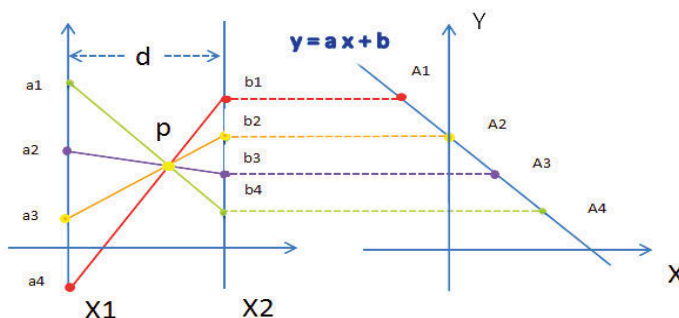


Figure 1-21 Geometry Theory of Parallel Coordinate in 2D

Generally, a N - dimensional line l can be described by the $N-1$ linear Equations:

$$l : \begin{cases} l_{1,2} : x_2 = a_2 x_1 + b_2 \\ l_{2,3} : x_3 = a_3 x_2 + b_3 \\ l_{3,4} : x_4 = a_4 x_3 + b_4 \\ \dots\dots \\ l_{i-1,i} : x_i = a_i x_{i-1} + b_i \\ l_{N-1,N} : x_N = a_N x_{N-1} + b_N \end{cases}$$

In the $X_{i-1}X_i$ plane, the relation labelled $l_{i-1,i}, i=2, \dots, N$ can be represented as a sets of points $\{P_{1,2}, P_{2,3}, P_{3,4}, \dots, P_{i-1,i}, P_{N-1,N}\}$, where (P_1, P_2) satisfies the relation $l_{1,2}$, and successive $(P_2, P_3)(P_3, P_4) \dots (P_{i-1}, P_i)$ satisfies the relations $l_{2,3}, l_{3,4}, \dots, l_{i-1,i}$ respectively. So a line $l \subset R^N$, as represented by $N-1$ points P_i .

Specifically, P_i can be computed by the following formula:

$$l : \begin{cases} p_{1,2} : x_2 = \frac{1 + 0 \times (1 - a_2)}{1 - a_2}, \frac{b_2}{1 - a_2} \\ p_{2,3} : x_3 = \frac{1 + 1 \times (1 - a_3)}{1 - a_3}, \frac{b_3}{1 - a_3} \\ p_{3,4} : x_4 = \frac{1 + 2 \times (1 - a_4)}{1 - a_4}, \frac{b_4}{1 - a_4} \\ \dots\dots \\ p_{i-1,i} : x_i = \frac{1 + (i - 2) \times (1 - a_i)}{1 - a_i}, \frac{b_i}{1 - a_i} \\ p_{N-1,N} : x_N = \frac{1 + (N - 2) \times (1 - a_N)}{1 - a_N}, \frac{b_N}{1 - a_N} \end{cases}$$

Therefore, to visualize, a parallel coordinate plots is displayed by N parallel axes, typically vertical and equally spaced. A point is represented as a polyline

with vertices on the parallel axes, and the position of the vertices on the the i^{th} axis corresponds to the i^{th} coordinate of the point. Figure 1-22 is a representation of car datasets by parallel coordinate plots.

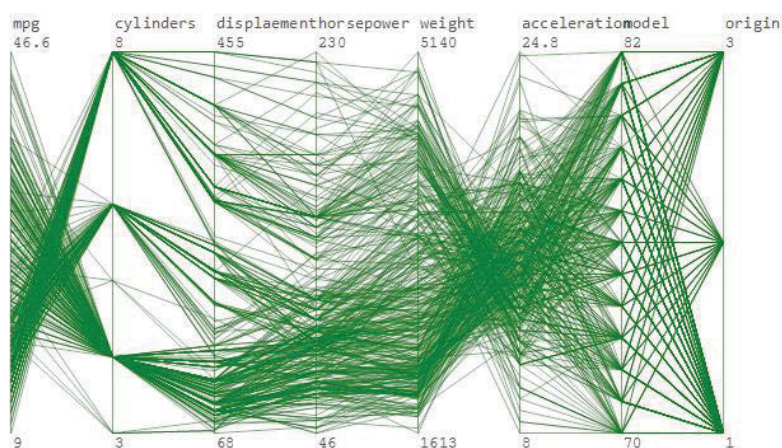


Figure 1-22 Parallel Coordinates plot for the Car Datasets.

1.1.4.2 Improvements and Examples

Since parallel coordinate plots was firstly introduced and suggested as a tool for high dimensional data analysis approximately 30 years ago, many improvements have been made to the algorithm itself. Such as arc-based parallel coordinates, which better preserves the geometric structures of data and can visualise many more data items in the same screen space (Huang, Lu & Zhang 2015). Another method is axes re-ordering. In parallel coordinate, it addresses the problems of visual clutter and computational complexity (Lu, Huang & Huang 2012). There is also a vertices optimisation in parallel coordinate plots, which deals with the representation of uncertainties in datasets (Lu et al. 2010). Samples are shown in Figures 1-23, 1-24 and 1-25 respectively.

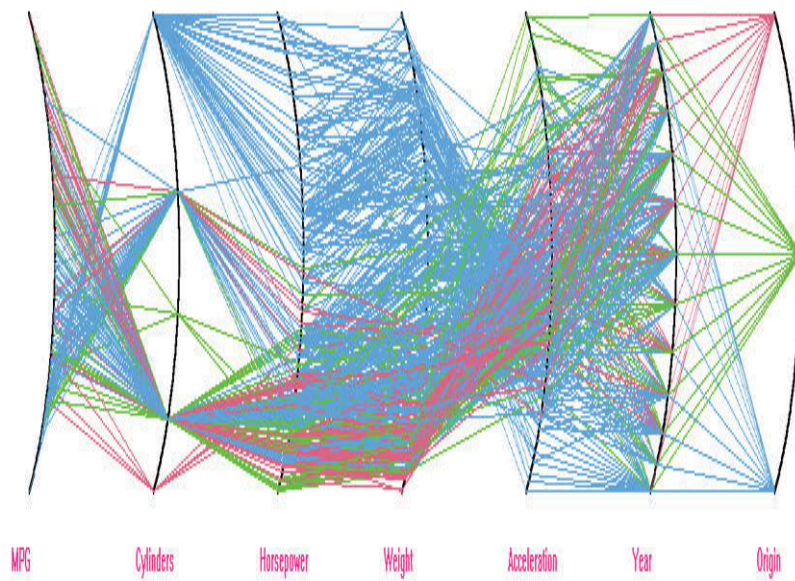


Figure 1-23 Car Dataset visualised by Arc-based parallel coordinates geometry (Huang, Lu & Zhang 2015)

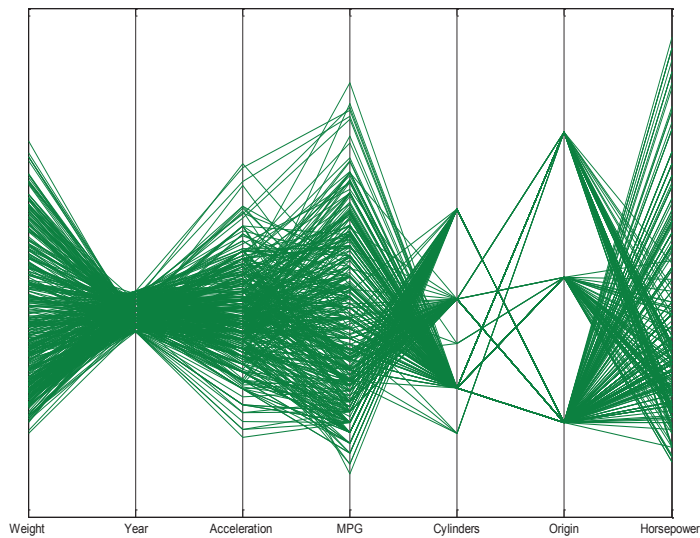


Figure 1-24 Cars dataset visualisation in parallel coordinates with a new axes re-ordering method (Lu, Huang & Huang 2012)

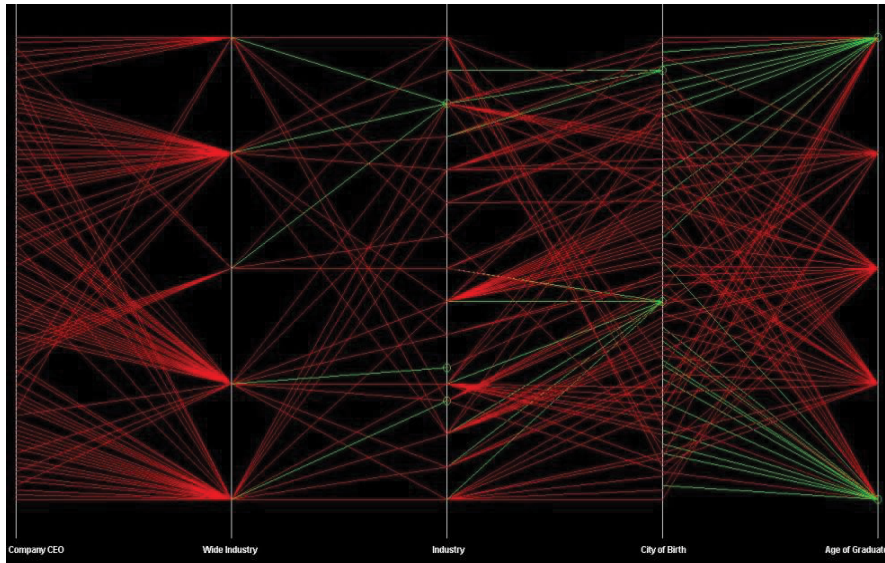


Figure 1-25 Forbes 94, a dataset with 5 variables visualised in parallel coordinate visualisation: after clutter reduction (Lu et al. 2010).

1.1.5 Other Visualisation Techniques

1.1.5.1 General Logic Diagrams

General Logic Diagrams (Ward, Grinstein & Keim 2010) also named dimensional stacking is a technique proposed by LeBalnc et al (1990). This technique provides a method for transforming and displaying Multi-dimensional data into two or three dimensions in order to provide a simple and clear visualisation. Specifically, if there is an N dimensional dataset, the dataset may be partitioned into two dimensional subspaces which are embedded into each other. The steps are shown below and will be repeated until all the attributes have been assigned.

- Calculate the stack cardinalities of each dimension with the same orientation.
- Divide the X coordinate by the number of stack cardinality.
- Continue dividing the remainders according to their stack cardinalities and until the last dimension is reached.
- Repeat the process.

One of the main advantages of the dimensional stacking visualisations is that it makes the high dimensional data relatively easy to read. For example in Figure 1-26, we can overview the pattern of the whole dataset from the image.

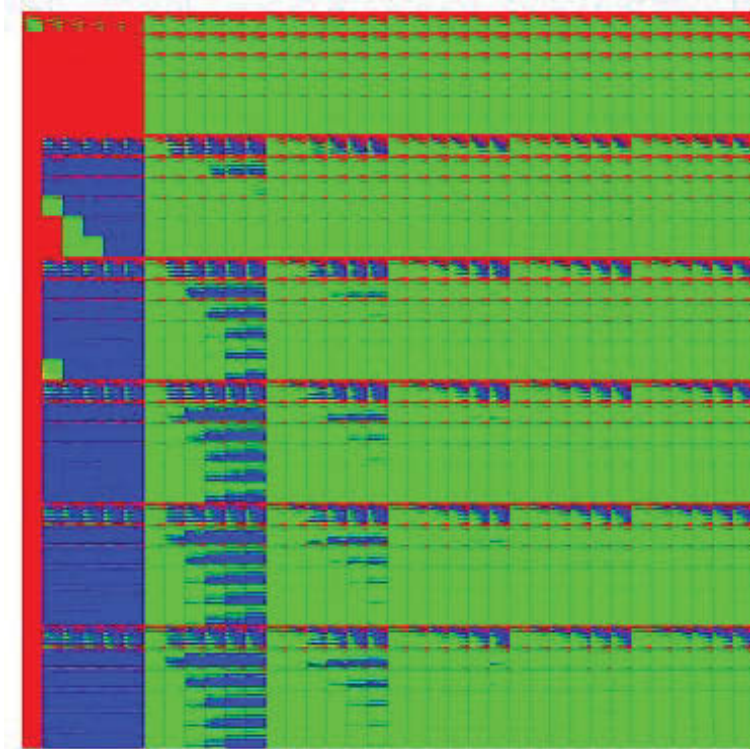


Figure 1-26 Example of Graphical Logical Diagrams(Taylor et al. 2006)

1.1.5.2 Pixel-oriented Visualization

The basic idea of pixel-oriented techniques (Keim 2000) is to map each data value to a coloured pixel. While the number of the dimensions is the number of sub-windows on the screen. For example, if a dataset has N dimensions, the pixel-oriented visualisation will create N windows, and one window for one dimension. Therefore a data record is mapped to N pixels at the corresponding positions, the values are represented by the different colours of the pixels.

We Take one of the pixel-oriented visualisation techniques – the circle segment technique” as an example. See Figure1-27, which explains the fundamental idea of this technique. This Figure displays all the data dimensions as segments

of a circle. In this sample, the testing dataset has 8 dimensions. In detail, the drawing starts from the center of the circle and it draws dimension by dimension. During the drawing, whenever the 'draw_line' encounters border lines, the 'draw_line' is moved along the orthodox line and it changes the direction at the same time. This process is repeated until the remaining dimensions are drawn. Figure 1-28 is a sample of pixel-oriented visualisation, which provides a good overview of very large amounts of high-dimensional datasets.

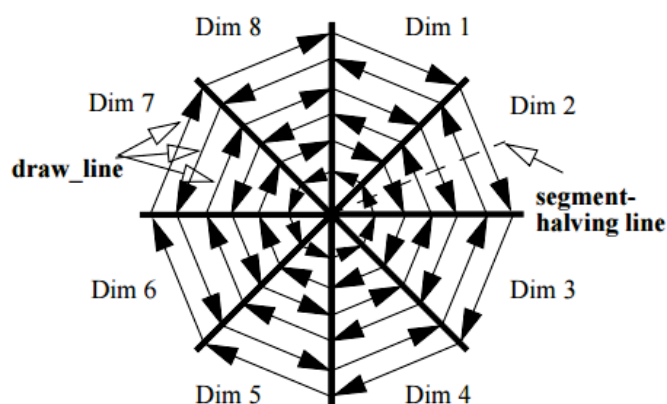


Figure 1-27 The Theory of the Pixel-Oriented Visualisation Technique (Keim 2000)

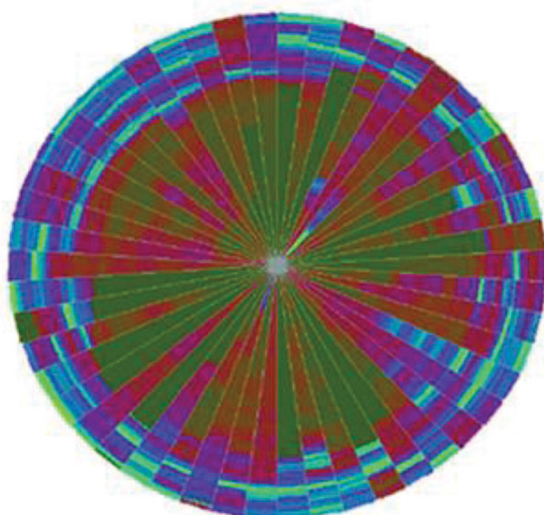


Figure 1-28 A visualisation sample of Circle Segments Technique (Ankerst 2001).

1.1.5.3 Glyphs Visualization

Chernoff faces (Chernoff 1973; Keim 2000) is one of the visualisation techniques in Glyphs, which uses faces to graphically represent points in a high dimensional space. This technique displays all the data in the shape of a human face. The data value decides the size, shape or placement of different parts of the face, such as the eye, nose, mouth, etc. Figure 1-29 shows Chernoff faces for lawyers' ratings of twelve judges.

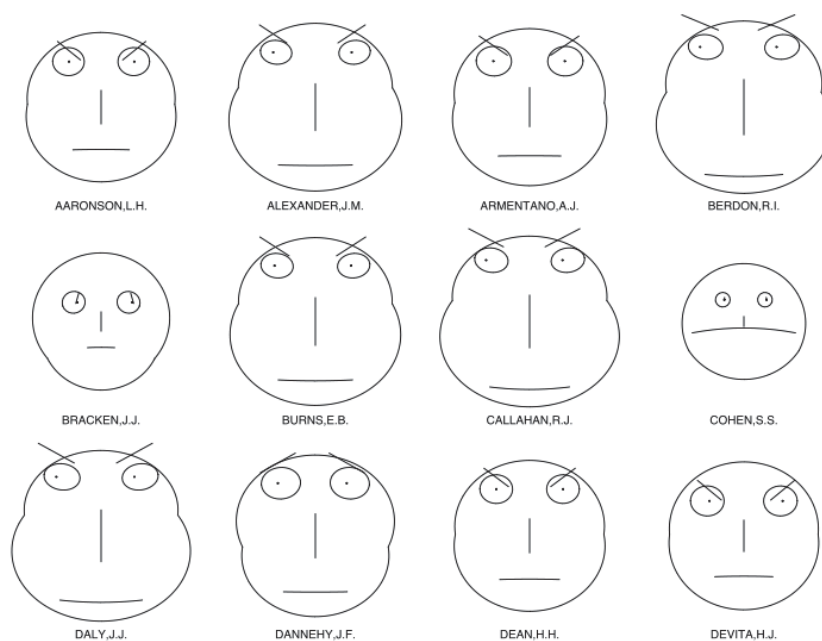


Figure 1-29 Chernoff faces Visualisation for lawyers' ratings of twelve judges.

Source from : https://en.wikipedia.org/wiki/Chernoff_face

1.2 Visualisation Applications of MDV

Multi-dimensional (Multivariate) Visualisation normally refers to the visualisation of datasets that have more than three variables. Although the challenges of high dimensionality in visualisation have not been solved yet, many research domains have benefited from visualisation techniques since the time when the data visualisation was first proposed. This section highlights the recent developments of classic applications in Multi-dimensional data visualisation. Throughout the section, various applications of Multi-dimensional data visualisation are presented, including their use in social science, geography, aerospace and medicine. Generally, this section reveals the possibilities and advantages of the visual analysis.

1.2.1 Multi-dimensional Visualisation in Social Science

An application of Multi-dimensional social data visualisation is presented here. The example simply interprets how visualisation works in the domain of economics. Figure 1-30 is an investigation of economies of countries around. In this application, Zinovyev et al (2010) used the idea of scatterplots to visualise simultaneously the values of GDP, population growth rates, employment rate by three indicators, colour, shape size, and text size. By this tool, it is much easier and clearer for viewers to observe the differences and similarities of the changes in each element. For example, the GDP growth in Asian countries such as China, India, and Indonesia are similar. They all grow around \$2.8k per capita. While in the western countries, like Germany and Spain, their GDP growth is higher, being between \$30k and \$40k per capita.

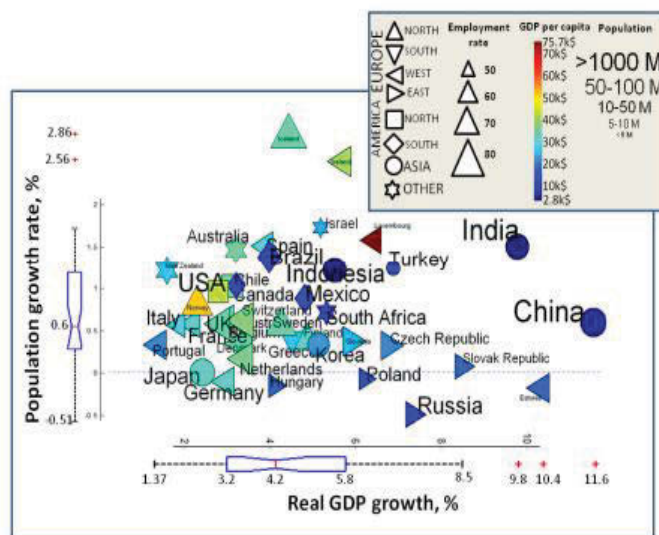


Figure 1-30 A visualisation application in social science (Gorban & Zinovyev 2010)

1.2.2 Multi-dimensional Visualisation in Geography

This Section introduces how visualisation techniques help users to understand the increasing volume of geospatial data. Particularly, it is more convenient to explore the oceanographic ecosystems. Take ViNeu (Kreuseler 2000) as an example, which is a system for visual analysis of complicated environmental phenomena. In Figure 1-31, ViNeu displays a 3D terrain profile with measuring points. In detail, each depth plot above the terrain represents one point data in time at their location. It uses colour to distinguish between elements, and after the data has been measured, the temperature, salinity, and relative oxygen concentration can all be visualised. Accordingly, the difference among different elements is also easy to obtain.

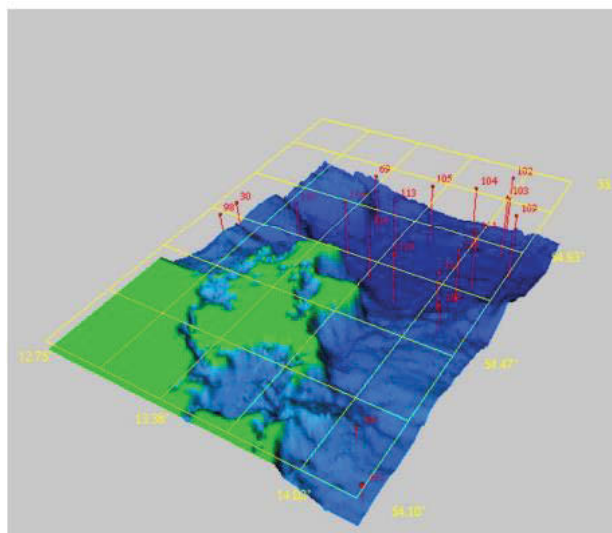


Figure 1-31 An example of Multi-dimensional visualisation in Geography – 3D terrain profile where measuring points and coordinate grid measuring points are numbered and can be selected (Kreuseler, Lopez & Schumann 2000)

1.2.3 Multi-dimensional Visualisation in Aerospace

Multi-dimensional Visualisation also plays an important role in the Air Traffic Control field. As the air traffic stakeholders always need up to the minute air traffic information in order to avoid flight clashes and to make sure of the security of every trajectory, so the visualisation technique is a good tool to furnish this required information. Figure 1-32 is one-day's record of traffic over France by Tool – FromDaty (Hurter, Tissoires & Conversy 2009). It represents a trajectory visualisation technique to deal with the problems of multiple trails. Specifically, the basic idea of FromDaty is from Scatterplots, interaction, and rapid visual design. The colour from green to blue represents the different altitude of each aircraft; the green colour represents the lowest altitude, while the blue colour shows the highest altitude. This tool helps the aircraft stakeholder to query iteratively by simple views and extract large amounts of trajectory data visually.

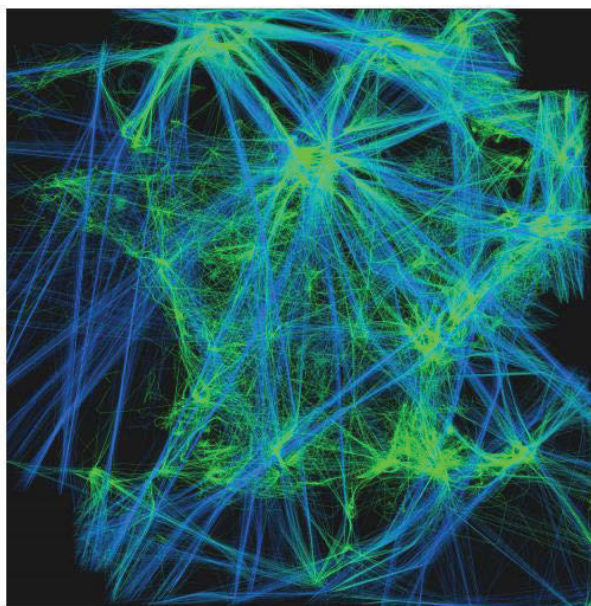
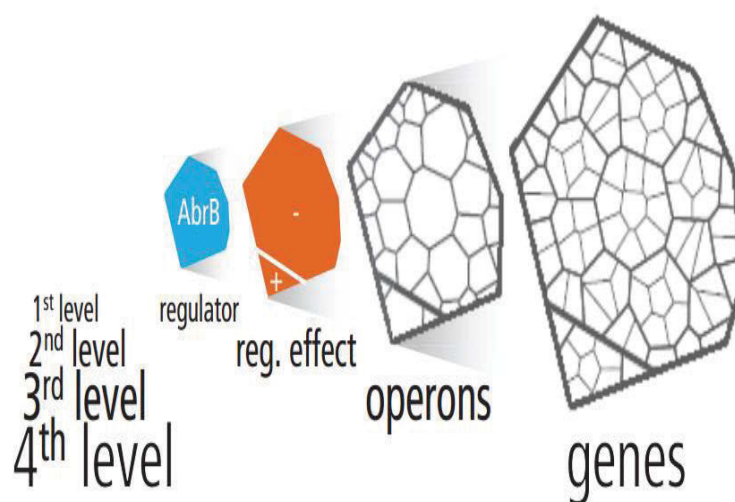


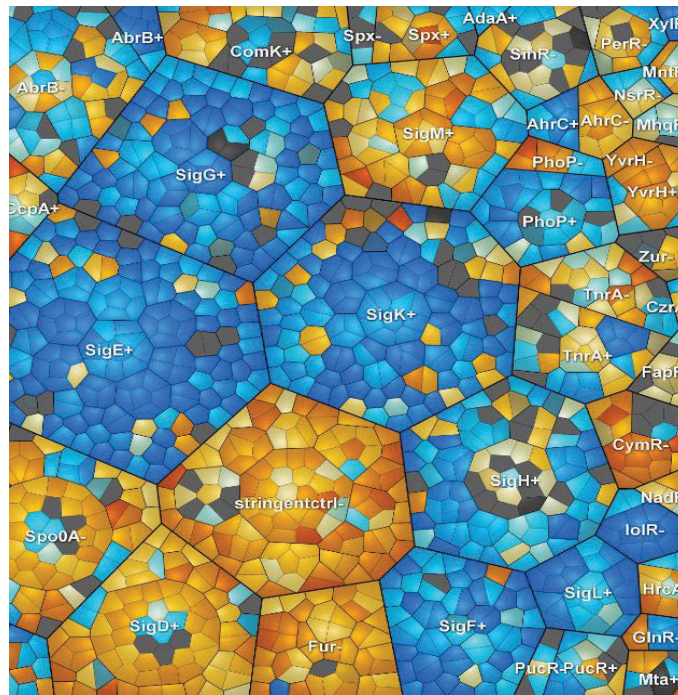
Figure 1-32 An example of Multi-dimensional data visualisation in aerospace (Hurter, Tissoires & Conversy 2010).

1.2.4 Multi-dimensional Visualisation in Medicine

High dimensional data visualisation of Gene ontology data (Baehrecke et al. 2004; Khatri & Drăghici 2005) has many modern applications. In this investigation, Voronoi Treemaps (Balzer, Deussen & Lewerentz 2005; Horn, Tobiasz & Shen 2009) are chosen to visualise the objects under consideration of these being the mRNA of a *Bacillus subtilis* 168 glucose starvation (Figure 1-33(b)). As the Voronoi Treemap is a hierarchical structure visualisation technique, it allows for creating within areas of arbitrary shapes, such as triangles and circles. It also offers low aspect ratios, better interpretability of hierarchical structures, and flexible adaptability regarding the enclosing shape. Therefore, in gene analysis, the various shapes can provide improved visual perception of the hierarchical classification of bacteria, See Figures 1-33(a) (b).



(a)



(b)

Figure 1-33 The examples of Multi-dimensional data visualisation in gene analysis (Paver); (a) level representation and (b) analysis of the Global mRNA of a *Bacillus subtilis* 168 glucose starvation experimentanalyse (Otto et al): Source From: <http://www.decodon.com/paver-benefits.html>

1.3 Visualisation Optimisation Methodologies in Visualisation

Optimisation in visualisation employs mathematical methodologies and graphic design approaches to help provide useful visual representations and deeper insight knowledge. Usability and utility are the parameters to consider in terms of the optimisation of visualisation techniques. In this section, we divide different optimisation methods into different groups, and in the following sections, we introduce them. Firstly, in Section 1.3.1 modelling optimisation will be explained, followed by geometrical optimisation in Section 1.3.2; then Section 1.3.3 introduces aesthetic optimisation; finally, functional optimisation and applicability optimisation will be described in Sections 1.3.4 and 1.3.5 respectively.

1.3.1 Modelling Optimisation

Visualisation Model also called visualisation workflow normally contains statistical analysis, visualisation, human-computer interaction, human–human communication (Chung et al. 2015). A proper model could improve the cost-benefit ratio of a visualisation technique, such as results accuracy, analysing speed, saving human resources, reducing computing consumption, and so forth. Therefore, discovering a productive model has become more and more popular and essential for managing today’s big data.

Traditionally, it is important to optimise a visualisation model from an information-theoretical perspective; it is also desirable to improve the workflow mathematically. In either process, three main components, analysis, visualisation and interaction, are considered. Some researchers break down the different components into steps (data transformation or data mapping etc.) to improve the visualisation working efficiencies. For example, Upson et al. proposed one of the earliest abstractions of a workflow, which mentioned collecting data, data filtering, data mapping and rendering, and result output. Later, interaction and cognition were brought into the visualisation pipeline by van Wijk et al (2008); While other researchers tend to implement the optimisation by designing methods. For example, Abram et al (1995) pointed out the problem of data flow execution in the data visualisation. They proposed an extended new model to enhance the implementation of the execution model in the data explorer in response to user requirements for data analysis and visualisation. Later in 2000, Munzner et al (2009) proposed a visualisation model for a taxonomy of visualisation techniques, comprising data, visualisation,

and visual mapping transformation. This has greatly helped implementers to understand the space of design and how information visualisation techniques can be applied. Compared with the above approaches, developing theories of visualisation needs more effort. For example, the application of information-theoretic for visualisation (Chen & Jaenicke 2010) , which explains some phenomena and events in visualisation, such as visual mapping, context + detail, etc.; Chen et al (2010) also gave another mathematical explanation as to be optimised in successful visualisation processes.

In summary, the optimisation techniques on visualisation are increasingly more important in the domain of Multi-dimensional data analysis. And the current practices in different representations or visualisations with various optimisation models have brought benefits in terms of working efficiency and processing ability from the point of view of a visualisation perspective.

1.3.2 Geometrical Optimisation

The visualisation optimisation can also be considered from geometrical property perspective. In this section, we use parallel coordinate plot as a sample to display how this idea works in the domain of visualisation.

Although parallel coordinate plot is a popular used visualisation technique, the visual clutter problem is always a difficult issue. Liang fu Lv etc. (2015) proposed an arc-based parallel coordinates visualisation method, termed arc coordinate plots (ACP), which reduces the clutter problem. The approach he mentions is based on geometry computation on axes drawing. It is easy to find that the axes in the traditional plot (PCP) are constructed by vertical straight lines, and if we suppose there are two points which are start and end points, it is longer to draw an arc between them rather than drawing a line. So more items could be displayed on parallel arranged arcs compared to the parallel lines. This benefit brings better geometric structure to some circular datasets. Figure 1-34 provides the proof of this theory. Based on the theory, we find the following equation:

$$|O_1M| = |O_1M'| = |x_0| = \frac{\sqrt{2}}{2}$$

While

$$|MM'| = 2T = 1.$$

Hence, $\Delta O_1MM'$ is a right angled isosceles triangle. The length of arc MM' equals to one quarter of the perimeter of a circle exactly, i.e.

$\frac{\pi x_0}{2} = \frac{\sqrt{2}\pi}{4}$. Therefore, the proof of extension rate is $\frac{\sqrt{2}\pi}{4}$. Two figures in Figures

1-35 (up) (down) explain the geometrically optimised parallel coordinates

compared to the original parallel coordinates.

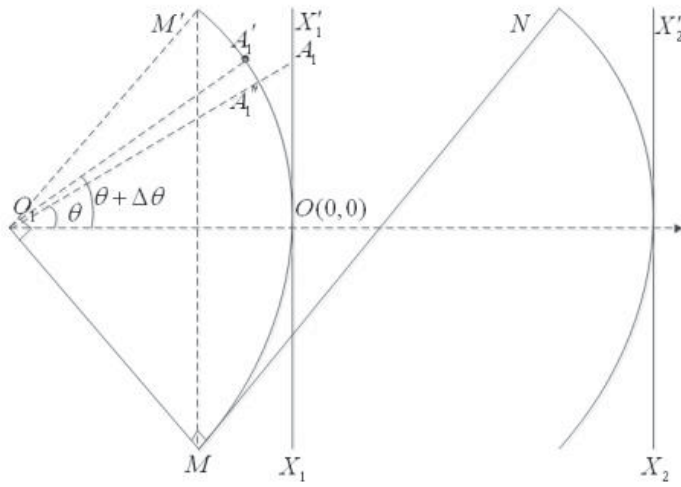
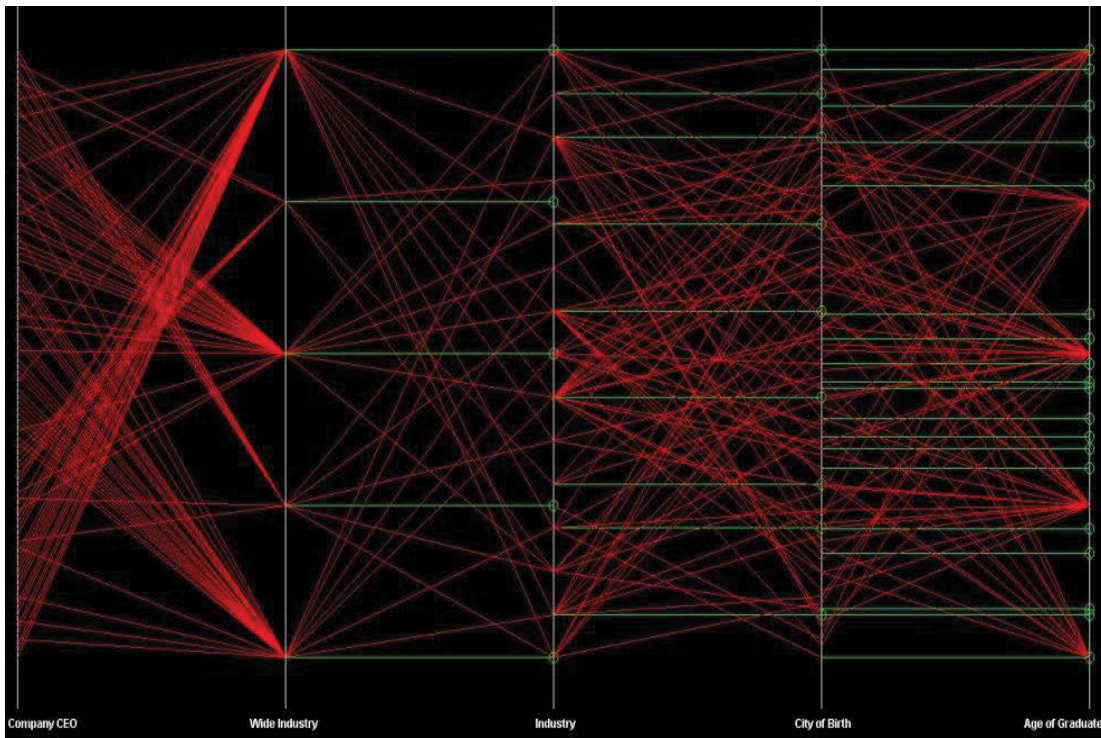


Figure 1-34 The rationale of the arc coordinates plane.



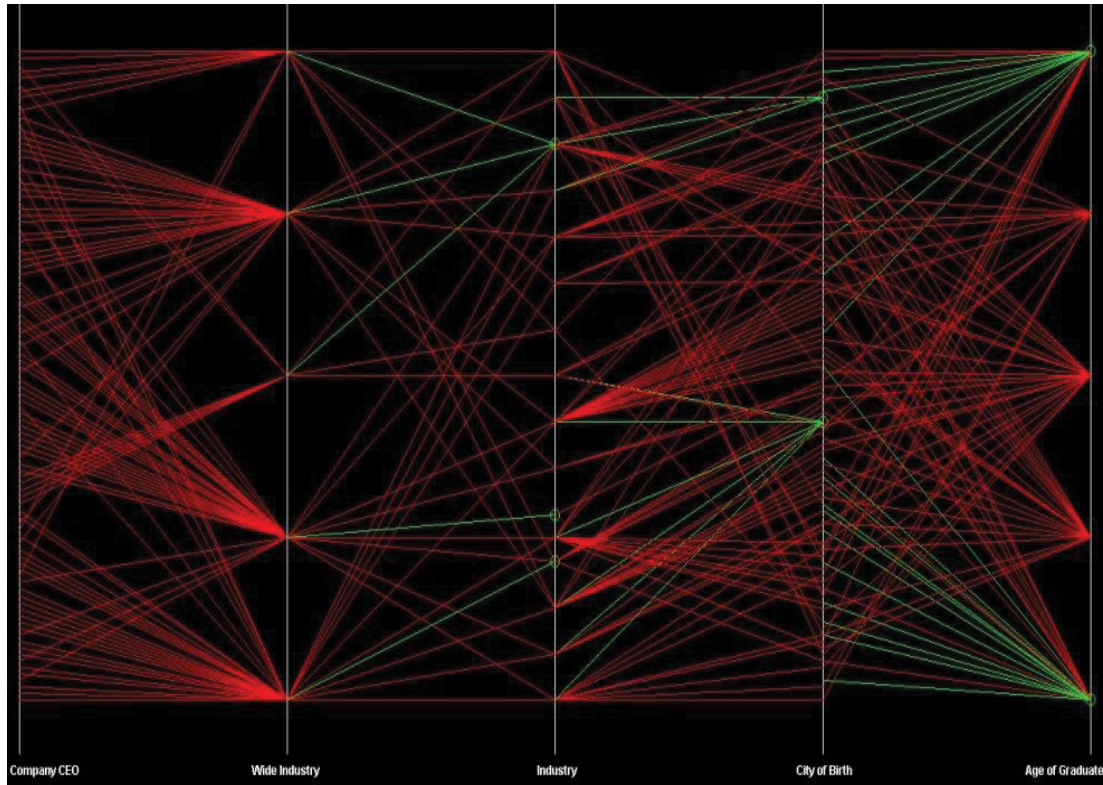


Figure 1-35 (Up) Original Plot (Hauck et al.); (Down) after clutter reduction. Data results used with kind permission of the author Liang Fu Lv.

1.3.3 Aesthetic Optimisation

In this section, we discuss the optimisation of aesthetic criteria for visualisation. It is necessary for a representation to be nicely perceivable. Normally, the criteria can be affected by the position of the nodes or edges, the colour of the nodes or edges, the shape of the nodes or edges, the aspect ratio aesthetics rule, the number of edge crossings and the path length between nodes or the angle between edges. Therefore, researchers who are interest in the optimisation of representation from the aesthetic point of view basically obtain their idea from the above aspects.

Take Circular Treemaps (Al-Awami et al. 2016) , Figure 1-36, as an example. It is presented to give the visualisation an unusual, creative look. Although using nested circles as a mode of display has some shortcomings such as space wasting and slower drawing speed, the visually attractive layout is still engaging for many applications.

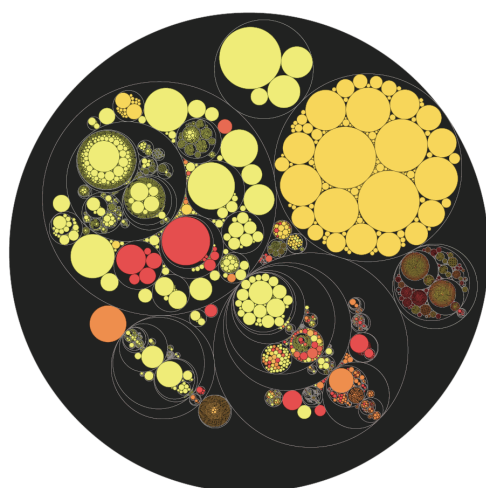


Figure 1-36 Circular Treemaps example, the red circles represent newer files while soft yellow circles represent older files. Source from: <http://lip.sourceforge.net/ctreemap.html>

Another good sample of aesthetic optimisation of visualisation techniques is Tangram Treemap (Liang.J 2015) . This approach considered the maximization of space utilisation of the entire computer screen that is commonly shared by multiple sessions . It uses alternative Treemaps to partition the hierarchical data structures in a variety of shapes to achieve the layout variability in enclosure data visualisation, see Figure 1-37.

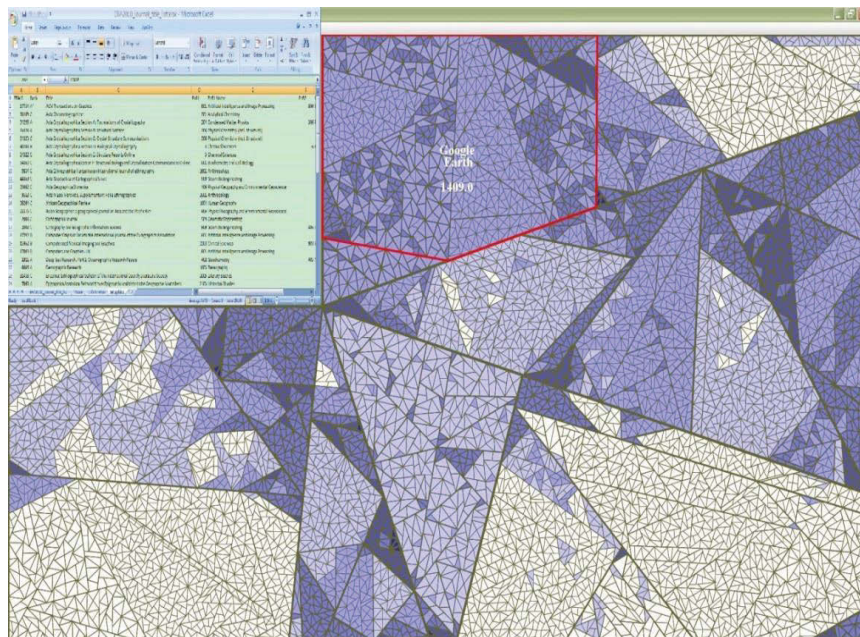


Figure 1-37 An example of computer screen that achieves the maximization (100%) of space utilisation and the minimization (0%) of the overlap among two session displays by using the new Tangram Treemaps (Liang.J 2015) .

1.3.4 Functional Optimisation

Enhancing the functionality of visualisation techniques is also an ideal way to implement optimisation. During the past years, many data mining methodologies have been introduced into visualisation. For example, suppose a large volume of dataset needs to be analysed, and the configurations of computing machines (normally use computer) have limitations. Using data reduction methodologies, such as Principal Components Analysis (PCA), Attribute Subset Selection (ASS), this problem can be solved. The resulting visualisation will also be clear with a higher accuracy. Other functions, for example, outlier detection, classification, density detection, are all popularly used in visualisation optimisation to propose a better result. Figure 1-38 is a sample of using data clustering to improve multi-dimensional data visualisation techniques (Huang, Huang & Zhang 2016).

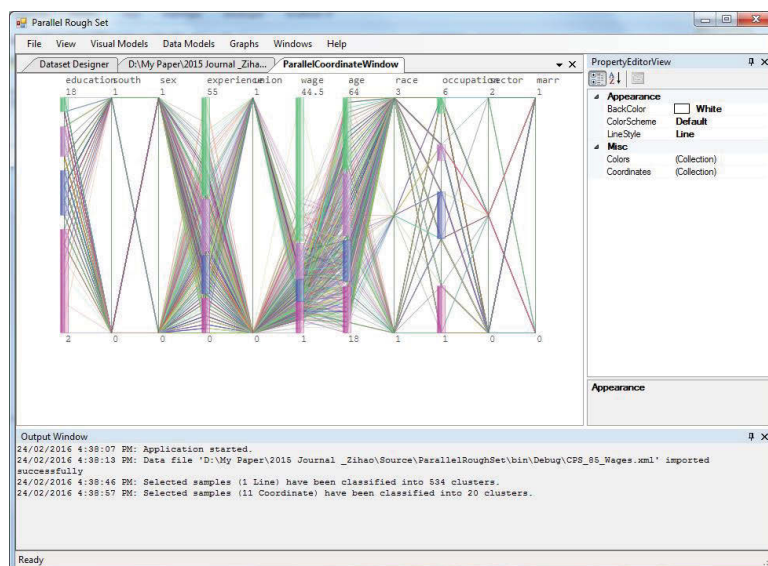


Figure 1-38 Data Clustering in Parallel Coordinate. Case study with Wages dataset obtained from <http://www.nber.org/cps/>.

1.3.5 Applicability Optimisation

The last visualisation optimisation method is applicability extension. It means to apply the visual techniques into a new domain that can solve the practicable problems. For example, Figure 1-39 interprets Chernoff Faces (Chernoff 1973) which is used in the detection of the life in Los Angeles.

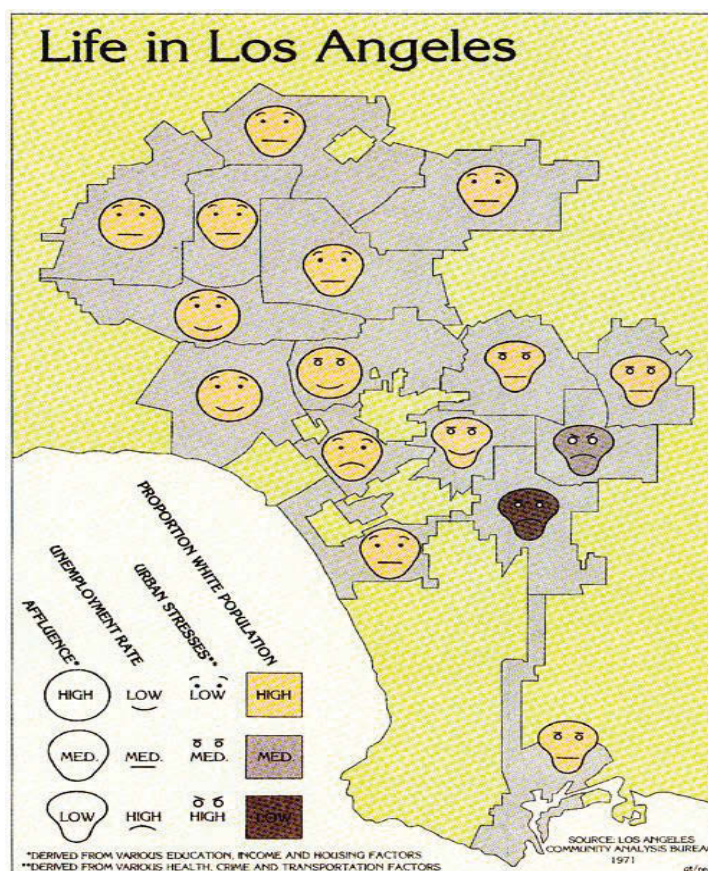


Figure 1-39 Chernoff faces display life in Los Angeles. Source from:

<https://cartastrophe.wordpress.com/2010/06/16/on-the-abuse-of-chernoff-faces/>

1.4 Research Challenges

In today's data era, our daily lives can not be separated with digital devices, such as working by computers and communicating through phones, etc.. As a result, large volumes of data is generated each day. Therefore, a good way to improve our daily livings and society developments is discovering useful information from the generated data.

MDV methodology is one of the important ways to analyze data. Although many MDV algorithms have been proposed during the past years, research challenges are still existed. In our thesis, we focus on the improvement of optimisation methodologies on MDV.

Solving multi-dimensional data visualisation problems by optimisation techniques involves more than one perspective. Through the approaches described above, a visualisation technique can be improved from different views. Rather than propose or design a novel technique, in our work, we focus on the technical improvements of the scatterplot matrix and the application extension of parallel coordinate plots in forensics.

We conclude the research challenges below.

- **Research Challenge 1**

While most of scatterplot matrix varieties have achieved the representation of an ordered collection of bivariate graphs, they do not consider the maximization of space utilisation of the entire computer screen. Specifically, some plots are duplicated, so the space of these

plots on the screen is wasted. Therefore, recalculating the plots' positions where a balance between information ordered clarity and maximum space utilisation should be considered.

- **Research Challenge 2**

Scatterplot matrix is one of the major multi-dimensional visualisation techniques. It is a great way to roughly determine whether a linear correlation exists between multiple variables. However, it is a theoretically impossible problem to enable direct interactions with scatter points.

- **Research Challenge 3**

Forensic investigation is the application of science to criminal and civil laws. The computer related forensic investigators collect, preserve and analyse the data to obtain useful evidence in the real court for protecting human rights. However, the special requirements of time and accuracy on forensic evidence are always challenges for investigators.

1.5 Research Objectives

The overall objective of this research is to optimise the multi-dimensional visualisation techniques which address all of the challenges in Section 1.4. More specifically the research objectives are described below:

- **Research Objective 1**

To investigate space optimised visualisation techniques that can maximize the overall utilisation of computer screens without losing the clarity of displaying variable relationships in order.

- **Research Objective 2**

To conduct experimental and user-centered evaluation of techniques produced in research objectives 1.

- **Research Objective 3**

To conduct an experimental evaluation and usability study of techniques in regard to a scatterplot matrix with decision trend analysis and interactive data exploration.

- **Research Objective 4**

To investigate suitable visualisation optimisation methodologies that can be applied to the investigation of computer forensics to improve the working efficiency of forensic investigators or researchers.

1.6 Contributions

To date, most of the existing scatterplot matrix varieties only concern the content of data information that can be displayed in each plot. They use data mining approaches (clustering or density detection) to obtain more information from the original dataset; They also add interactive techniques to achieve the human-computer interaction and make it user friendly. However, there is a space utilising problem in the ordered visualisation techniques that the duplicated plots are reduced the original order needs to be remained. Specifically, while the number of plots are increased, the lower the space utilisation rate of the computer screen.

Therefore, in regard to scatterplot matrix of our thesis, the major significance is that it is the first time to address the issue of optimizing the display space utilisation. The major contribution to this challenge is also providing an appropriate solution to this issue for the first time. In addition, this proposed space optimised scatterplot matrix has been evaluated with scientific experiments and user studies and tested in case studies.

Furthermore, considering the existing interactive functions in a scatterplot matrix, for example, brushing. No research has discussed interactively displaying data points and their classifications, which is a theoretically impossible problem. So in our thesis, another contribution to scatterplot matrix is conducting the experimental evaluation and usability study to evaluate an enabled data trend analysis with interactive scatterplot matrix.

Lastly, our thesis also discusses the visualisation optimisation from the application perspective. We found that the visualisation techniques could bring many benefits to forensic investigations, either for modelling investigation process, or visualising the digital evidence. Particularly, we use multi-dimensional data visualisation techniques to dealing with the time zone problems in the display the hard disk drives.

1.7 Thesis Organisation

In this thesis, we discussed the methodologies of optimising multi-dimensional visualisation techniques. Firstly, we propose a space optimised scatterplot matrix and evaluate an interactive scatterplot matrix by pilot studies, in which the main idea is to compare with parallel coordinates by answering questions. Then calculating the mean value of answer accuracy. What's more, we consider the optimisation method of technique application in multi-dimensional visualisation. We focus on a new research domain, computer forensics. Through our experiments, it is also clear that visualisation can greatly help forensic investigations. Figure 1-40 is a description of the thesis structure.

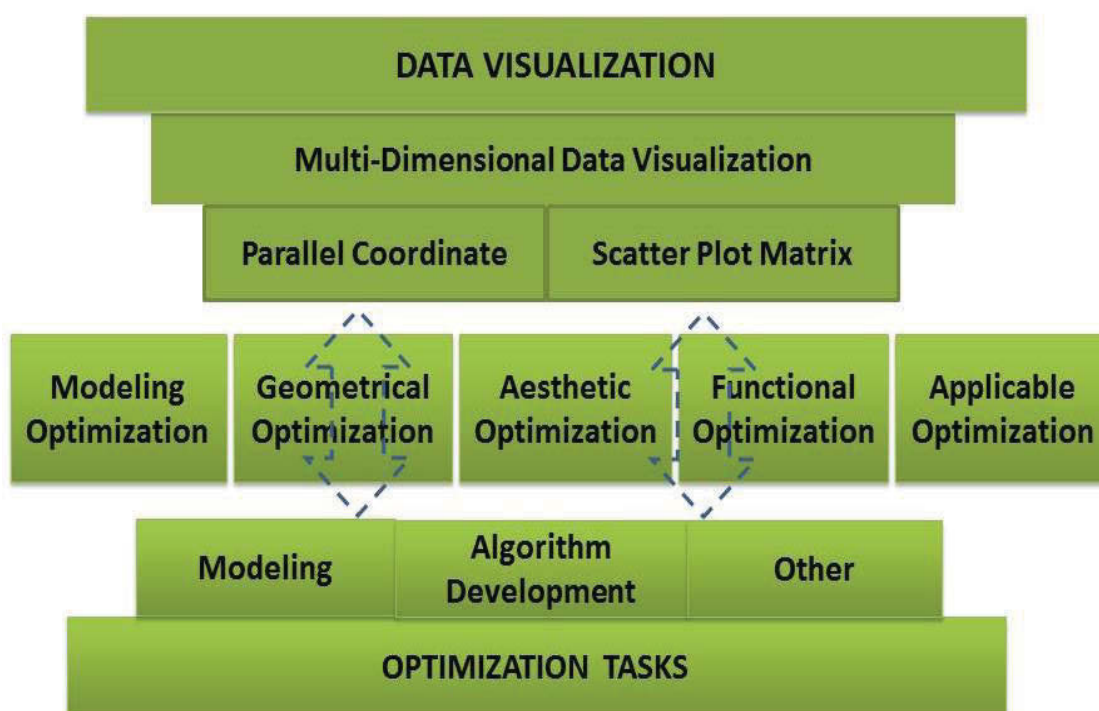


Figure 1-40 Thesis structure description

Specifically, the thesis is organised as below: Following the introduction and

background introduced in Chapter 1. Chapter 2 explains the pipelines of visualisation process. Chapter 3 introduces our new space optimised scatterplot matrix visualisation technique. In detail, Section 3.1 shows the framework. Section 3.2 describes the techniques and algorithms required in the implementation. In Section 3.3, we display the interaction mechanism which enhanced the usability of this application. Followed by Section 3.4 to evaluate this technique from a technical perspective and also the usability studies, and the last Section 3.5, a brief conclusion will be given.

Chapter 4 explains an interactive exploration in scatterplot matrix visualisation. We mainly work on the evaluation of the interactive exploratory scatterplot matrix by comparing with parallel coordinates.

Chapter 5 examines how important the visualisation technique is in computer forensics. We use different cases to certify the effectiveness of parallel coordinates in forensic investigations. Particularly, we propose visualisation involved models for investigators to get the results effectively and efficiently. we also use parallel coordinates on the presentation of hard disk drives, and then explore criminal relationships through multi-dimensional visualisation approaches.

Finally, the work is concluded in Chapter 6. We recaps the research strength and connects our findings with research challenges and original contributions in Section 6.1 and 6.2. Section 6.3 discusses the limitations of the research, which opens up opportunities for future work. The areas for further development and research include: technical improvements, alignment with Industry, scatterplot matrix Design Guidelines and Systematic Application Evaluation

principles. In the meanwhile, a summary of the significance of the research in the field of data visualisation will be given.

Chapter 2 Visualisation Pipelines

THIS CHAPTER AIMS to understand the basics of visualisation from different expectations.

In this chapter, firstly, we define our terminology to help users have a clear understanding of the words which are overused in the visualisation literature.

Secondly, we introduce the pipeline of visualisation to help users have a general view of how visualisation can be implemented.

Thirdly, we specify the main principles of a good visualisation.

2.1 Defining Terminology

It is essential to differentiate the words such as “ Raw Data ” “ Variables (attributes) ” “ Multi-dimensional (high-dimensional) ” and “ Multivariate Data ” which are frequently discussed in the visualisation literature.

The term **raw data** can be applied to idiosyncratic formats. It has many forms, such as spread sheets, or texts. Usually it is unstructured, so normally this data needs to be transformed into a **data table** where data is more structured and thus easier to map to visual forms. Then the term **attribute** also named as **variable** represents a feature of a data object. An object of data is composed by **attributes**, and if a data object contains more than one attribute it is then called a **multivariate data** object.

There are slight differences among multivariate datasets, multi-dimensional datasets, and multivariable datasets. As for a **multivariate dataset**, it is composed of dependent variables, which might be correlated to each other to varying degrees. **Multi-dimensional dataset** is a dataset that has many independent variables clearly identified, and one or more dependent variables related to them. **Multivariable dataset** can be either multivariate or multi-dimensional, it is also named as **high-dimensional dataset** when the dataset contains more than three attributes, either dependent or independent. In this thesis, we are concerned with the visualisation optimisation of **multivariable dataset (high-dimensional dataset)**.

Some other terms are described below:

A Visualisation Method (V_m): could be a modeling scheme, a layout algorithm, or a viewing technique, etc.

A set of Properties: is a set of technical features of V_m , these technical features could be high interactive speed, high efficiency of space utilisation, low computational complexity, different physical references and so on.

Data Space (D): the original dataset.

A Graph $G(V, E)$: is defined as a pair (V, E) consisting of a finite set V of vertices and a finite set E of edges, where $E = \{(u, v) \mid u, v \in V\}$

A Drawing $D(G_i)$: is a geometric drawing of graph G_i . It consists of a position for each vertex $v \in V_i$.

Time Consumption (TC): is the running time spent in drawing in a graph from its initiate state to the end state.

Visual Metaphor: is a visual representation of a dataset by means of visual attributes, which is well-known to users, for example, colour, size, shape, etc.

Vertex Position Variance (VPV): indicates the position change of a vertex from its starting position to its end position in the drawing.

A Shape (SP): is defined by an x- and y-coordinate, and is defined by a group of vertices.

2.2 Data Visualisation Pipelines

The visualisation is the process of mapping data to a visual form which can be easily perceived by humans. We summarise a list of four different steps for visualisation processing, See Figure 2-1. Generally, It contains data formatting, data mining, data visual mapping and human perception. In the following sections, we will explain these activities briefly.

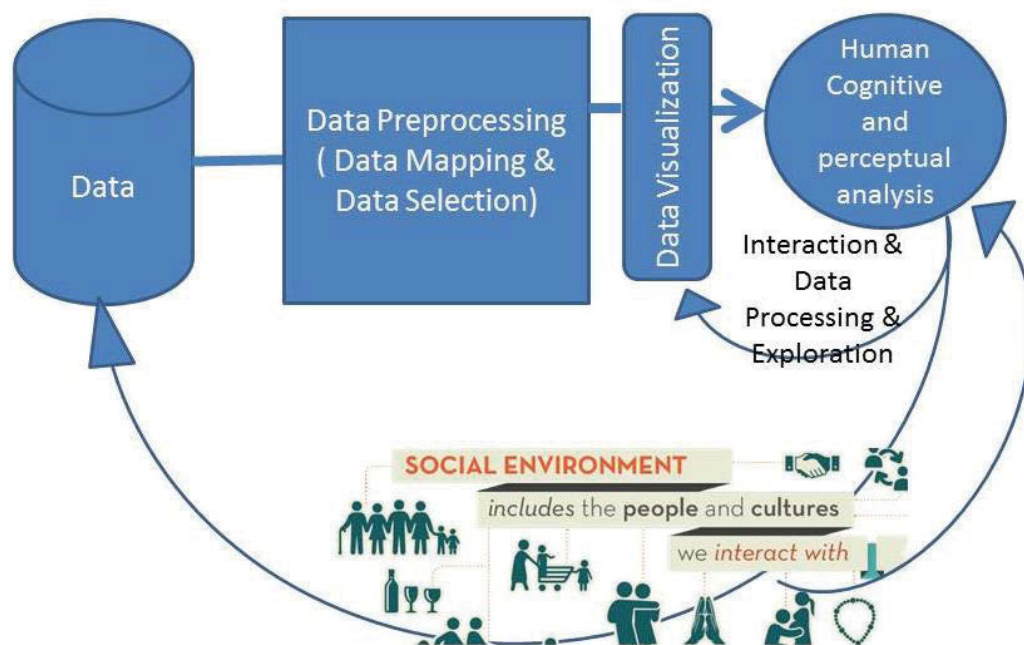


Figure 2-1 Visualisation Process

2.2.1 Data Formatting

Real data comes in many forms, from spreadsheets to the texts. In visualisation, data requires structure and a data model. The usual strategy is to transform the real data into a set of relations that are more structured, and a *Data Table* will be generated. Table 2-1 is a canonical data model about white wine quality. The first row is the attribute of the dataset, also named dimension, variable, field. Attributes can either be dependent on or independent from each other. The rest of rows are data values, and the types of these values have been mentioned in Section 1.1.1.1- numerical, ordinal and nominal.

In addition, some simple manipulations of the data table are necessary to understand in the working flow of data visualisation, which contains data selection, data projection, data aggression, join table rows/columns, transpose table rows and columns and sorting.

Table 2-1 Canonical data model - white wine quality dataset

<u>fixed acidity</u>	<u>volatile acidity</u>	<u>citric acid</u>	<u>residual sugar</u>	<u>chlorides</u>	<u>free sulfur dioxide</u>	<u>total sulfur dioxide</u>	<u>Density</u>	<u>pH</u>
7	0.27	0.36	20.7	0.045	45	170	1.001	3
6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3
8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19
8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26
6.2	0.32	0.16	7	0.045	30	136	0.9949	3.18
7	0.27	0.36	20.7	0.045	45	170	1.001	3
6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3

Space Optimisation in Multi-dimensional Data Visualisation

8.1	0.22	0.43	1.5	0.044	28	129	0.9938	3.22
8.1	0.27	0.41	1.45	0.033	11	63	0.9908	2.99
8.6	0.23	0.4	4.2	0.035	17	109	0.9947	3.14
7.9	0.18	0.37	1.2	0.04	16	75	0.992	3.18
6.6	0.16	0.4	1.5	0.044	48	143	0.9912	3.54
8.3	0.42	0.62	19.25	0.04	41	172	1.0002	2.98
6.6	0.17	0.38	1.5	0.032	28	112	0.9914	3.25
6.3	0.48	0.04	1.1	0.046	30	99	0.9928	3.24
6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33
7.4	0.34	0.42	1.1	0.033	17	171	0.9917	3.12
6.5	0.31	0.14	7.5	0.044	34	133	0.9955	3.22
6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33

2.2.2 Data Preprocessing

Up to now, the data volume is one of the main challenges in data visualisation. When the amount of data is getting numerous, it is more difficult to discover the essential information among them, and the time cost is always out of expectation. Here some data mining methodologies could provide approaches to simplify the complicated dataset. For example, data reduction, data clustering or dimensionality reduction.

In data reduction, people can choose a sampling approach or filtering approach. The former method cannot show every element, only the selected subset, but it is efficient for large datasets. While the latter one defines criteria to remove data from original datasets. Figure 2-2 is a CrossFilter visualisation tool implemented first by the filtering method.

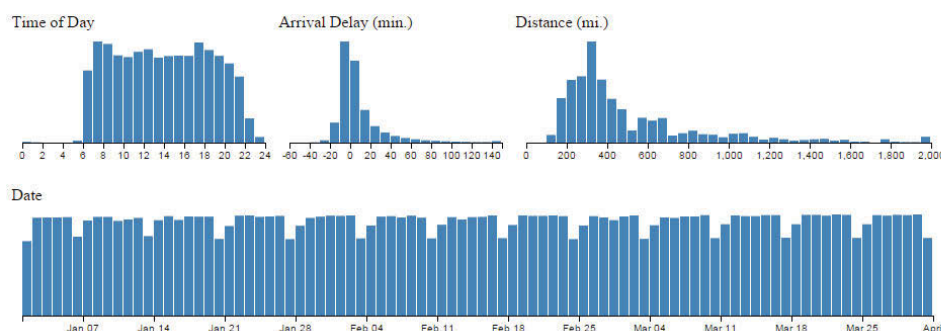


Figure 2-2 CrossFilter visualisation tool displays the airline on-time performance.
Source from: <http://square.github.io/crossfilter/>

Clustering is to classify similar items into groups. The methodologies are based on similarity measures, such as Euclidean distance $d(p,q) = \sqrt{\sum_{i=1}^n ((q_i - p_i)^2)}$, Pearson

correlation $\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$, etc.. By these methods to calculate the distance between two items and then divide them by partition algorithms. For example, *k* - means, Affinity propagation, Bi-clustering and Fuzzy clustering, etc.. Specifically, in data visualisation, the clustering method helps to order data, brush data, and also to aggregate data. Figure 2-3 is an example of using clustering in visualisation to get a data aggregation.

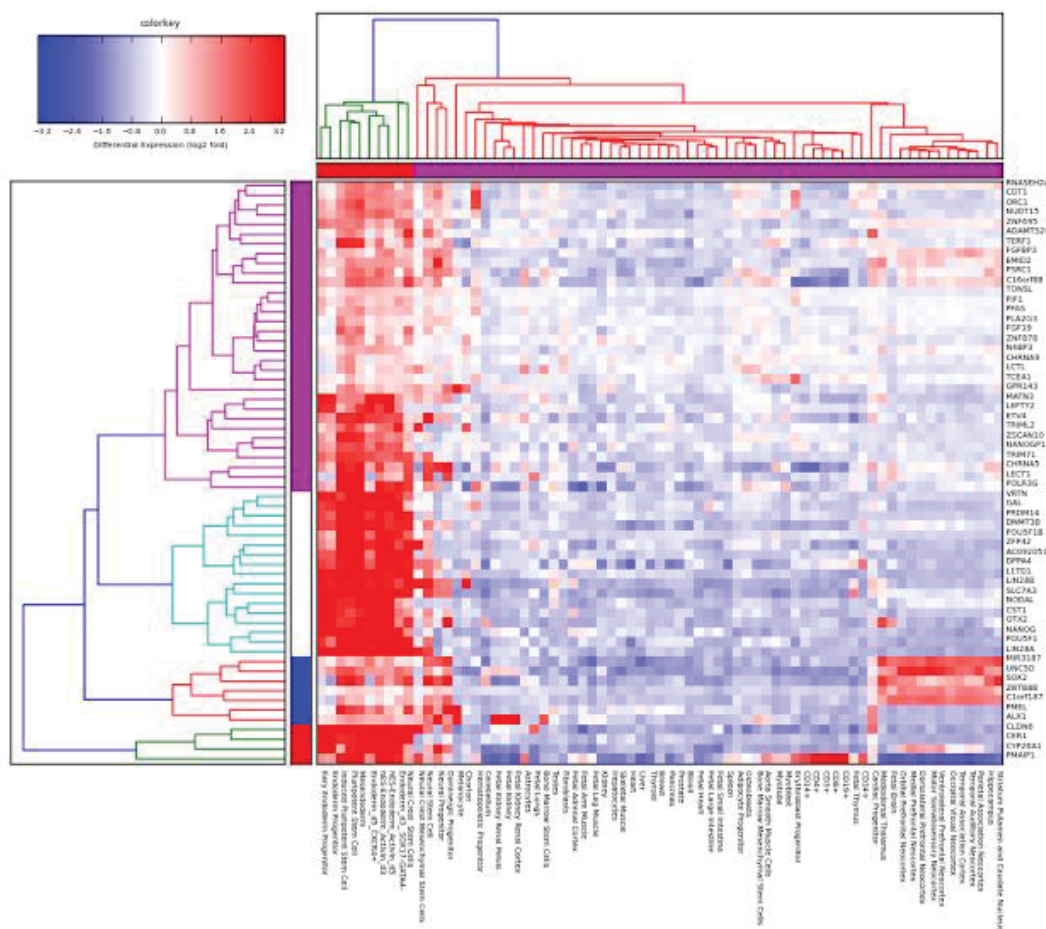


Figure 2-3 An example of a clustering heat map. The rows are the hierarchically clustered genes, while the columns are the tissues with dendrograms. The red in the heat map presents upregulation while blue presents downregulation. Cited from: <http://altanalyse.blogspot.com.au/2012/06/hierarchical-clustering-heatmaps-in.html>.

Many techniques in reducing high dimensional to lower dimensional space have been proposed during the past decades. Take principal components analysis (PCA) as an example, this statistical procedure uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. Supposing that the data vectors to be processed have N dimensions, PCA aims to find k vectors that are the best to be used to describe the N dimensional data, where $n \geq k$. PCA computes k vectors as the principal components, and these components will be sorted in order to be used as the new axes. The original data can be displayed on the new axes, and this should be a good approximation of the original data because the weaker components of data have been eliminated, see Figure 2-4.

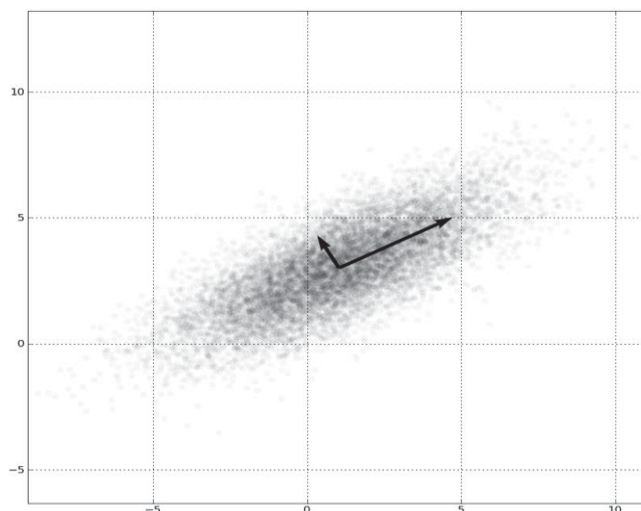


Figure 2-4 PCA of a multivariate Gaussian distribution centered at (1,3) with a standard deviation of 3 in roughly the (0.878, 0.478) direction and of 1 in the orthogonal direction. The vectors shown are the eigenvectors of the covariance matrix scaled by the square root of the corresponding eigenvalue, and shifted so their tails are at the mean, source from: https://en.wikipedia.org/wiki/Principal_component_analysis

2.2.3 Visual Mapping

Data mapping represents how to transfer data into visual form, which means to visualise the mathematical relations based on graphical properties. Normally, there are two steps to map data. The first step is to map data items as visual marks: points, lines, areas, volume glyphs, see Table 2-2. Followed by the second step of mapping data attributes as visual properties of marks, see Table 2-3.

Table 2-2 Visual Marks









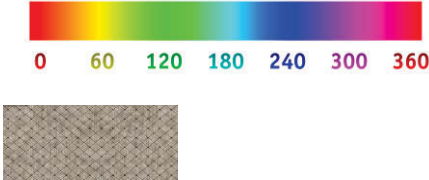


Name	Graphs
Points	
Lines	
Areas	
Volumes	
Glyphs	

Table 2-3 Visual Properties of Marks

Name	Representations
Position	
Length, Area, Volume	
Orientation, angle , slope	
Colour , Texture	
Shape	
Animation, Blink, Motion	

From a Geometric perspective, understanding data mapping processes are essential in implementing the visual marks and their properties. Suppose a dataset contains n data objects, and they are with m varieties, then the data relations can be explained as a set of tuples,

$$\{ \langle D_{11}, D_{12}, \dots, D_{1m} \rangle \langle D_{21}, D_{22}, \dots, D_{2m} \rangle \dots \langle D_{j1}, D_{j2}, \dots, D_{jm} \rangle \dots \langle D_{n1}, D_{n2}, \dots, D_{nm} \rangle \}$$

Where:

D_{ij} - represents a Data space;

After data mapping, the tuples can be displayed by a new form:

$$\{ \langle V_{11}, V_{12}, \dots, V_{1k} \rangle \times \langle V_{21}, V_{22}, \dots, V_{2k} \rangle \dots \langle V_{i1}, V_{i2}, \dots, V_{ik} \rangle \dots \langle V_{t1}, V_{t2}, \dots, V_{tk} \rangle \},$$

Where:

V - is the visual space,

t - represent the number of components in V , and it determines the number of graphical properties of the visual mark (refer to Table 2-4),

k - is the number of spatial coordinates used to locate a visual mark, such as in 2D or 3D environment to visualise visual marks. Specifically, in a 2D environment, if the aim is to visualise a high dimensional dataset, we can also display data with a Z coordinate, and we need to further transfer the above visual form into $\{A_1, A_2, \dots, A_i, \dots, A_k\}$, which means to take a whole row or column of data as one object, and then display them in a graphic form.

To conclude, a good mapping always produces a visual representation with higher accuracy. This contributes to describing information if there is an accurate relationship between data objects and visual objects.

2.2.4 Human Perception

Visualisation technology always occurs with assistance by humans. Term as cognitive support (Tory & Möller 2004), which means Human perception and cognition is an important factor in visualisation.

Generally, the visualisation interface is able to quickly recognise and easily interpret visualisation. All the variables should be visible so that a viewer can perceive information accurately, including: the overview of data objects, the relationship among data objects (either explicit or implicit), and the metaphors which are present to help the user understand how to explore through display.

2.3 Visualisation Principles

It is also essential to evaluate visualisation techniques to verify the effectiveness of techniques in different tasks. The early work in data visualisation was concerned with defining the quality of graphical displays and developing theories of design of graphical displays (Bertin,1981). Later, Tufte (1983) proposed the principles of excellence and integrity of data graphics. According to these principles, good displays should have the following factors:

- Display the whole data without losing information
- Lead the viewer to think about the substance of the data, rather than about the technical approaches or graphical design methodologies
- The distortion of data explanation
- Present Big Data (large volume, variety, velocity)
- Discovery relationships among datasets
- Avoid unreasonable purpose: representation, decoration

In summary, a good graphic is not only about displaying the whole data, but considering many aspects, such as data relationships or data scalability, ect.. and these elements makes giving good displays much harder.

2.4 Summary

Data Visualisation has been proposed many years ago, with the data is gradually complicated, the visualisation techniques had more challenges than before. Therefore, to improve the visual methodologies and grasp more

understanding on data visual results, it is important to know how data visualisation work and what is the standards of the visualisation.

This chapter gives a clear description of the above questions, and these would greatly help us to do the further study on multi-dimensional data visualisation.

Chapter 3 Optimisation of Scatterplot Matrix

THIS CHAPTER AIMS to present a space optimised scatterplot matrix with a clear display of pairwise variable relationships.

The basic concept of the scatterplot matrix is simple. Give a set of n variables $\{X_1, X_2, \dots, X_{n-1}, X_n\}$, scatterplot matrix contains all the pairwise scatterplots of the variables on a single panel in a matrix format. That is, if there are n variables, the scatterplot matrix will have n rows and n columns and the j^{th} row and the i^{th} Column of this matrix is a plot of $X_i * X_j$. It has been increasingly commonly used amongst other graphical tools.

Although a scatterplot matrix can explain all of the pairwise correlated information, its limitations are also clear, that is the issue of space utilisation. A scatterplot matrix will take up much more space while the number of data attributes is increased. The problem of navigation will also appear: While the volume of a dataset is getting larger, it is hard to display the information when the users scroll up and down.

In reponse to the above problems in a traditional scatterplot matrix, we present a space optimised scatterplot matrix with a clear display of pairwise variable relationship.

3.1 Framework

In this section we describe the proposed space optimised scatterplot matrix in an overall view, including the evolution of ideas and working processes.

3.1.1 Idea Evolution

Prior work in scatterplot matrix visualisation hasn't addressed the concerns of the adaptability to the space utilisation in the design of display containers based on the size of the display screen. This limitation might affect the clarity of information to be visualised on the screen at a time, especially with a large volume of dataset. The methodology proposed here provides better supporting multivariate data to be visualised in a scatterplot matrix with a 100% space utilisation of the various size of the display screen.

In scatterplot matrix design, all the plots could be simulated by multiple pointed rectangles with finite width and height. Hence, in order to be employed in an unstable size of display container, we aim to invent a new plot (rectangle) drawing method.

In the context of the graph drawing approach, the new approach should utilise the **Processing** draws directed graphs. This approach is able to arrange plots in the following ways: The number of the containments per row; the number of the containments per column; the width of the containments; the height of the containments. Generally, our new method offers flexibility and clarity in the above four ways to represent multi-dimensional data. With modification of the algorithms, the new approach should be capable of producing space optimised layouts for various high dimensional dataset as well.

The evolution of the research approach is abstracted in the following stages. In the first stage, we relax the size of the display screen constraint completely and add freedom for rectangular drawing. The new process creates visualisation mainly with rectangular fixing vertically. In order to increase visibility, we add a

design metaphor - colour to differentiate the plots. The layout of scatterplot matrix is improved with the arrangement and metaphor combined resolution.

In the second stage, we maintain the number of the original plots, by a mathematics ranking algorithm. This stage creates the traditional layout of a scatterplot matrix.

In the third stage, we reduce the duplicated plots, and relocate the positions of the left plots in the condition of minimally affecting the relationship between the variables which are originally in the same row or column, and this step is to keep our technique being balanced between the space utilisation and information clarity.

In order to fit into user tasks and scenarios', we employ these approaches in two environments mentioned below: Containment control (the size of the display screen); Container control (the number or size of the rectangular in each row and column). So that application designers can utilise it with other visualisation methods for diverse domain purpose and users can effectively and clearly discover the information over this new scatterplot matrix visualisation layout.

3.1.2 Space Optimisation Process

The principle of a scatterplot matrix is to roughly determine if the views have a linear correlation between multiple variables in two-dimensional space. The efficiency of the proposed method is based on the balance of a suitable view with clear variable correlation with the scatterplot matrix paradigm. Particularly, it inherits the rectangular drawing principals from processing approach to ensure the maximum utilisation of geometrical space for displaying plots, while they can also discover the relationships with each other by highlighting correlations with different colours. The series of procedures for constructing such visualisation is described as below:

Step 1 - Number Calculation (Row & Column): This step calculates the number of plots per row and per column. Particularly, the number of rectangles in a row or a column depends on both the size of the display screen and the number of the data variables (dimensions).

Step 2 - Space Partitioning: In a constrained space, (the maximum is the size of the display screen, the minimum is size 0), Step 2 needs to repartition the entire display space into a set of rectangular called plots when the size of the screen is resettled. After partitioning, each rectangle is then drawn in an orderly manner.

Step 3 - Vertex Calculation: This step calculates the position values for every rectangle in the scatterplot matrix. The position of a plot is defined as four vertex value pairs $P_1 \langle V_1, V_1 \rangle; P_2 \langle V_2, V_2 \rangle; P_3 \langle V_3, V_3 \rangle; P_4 \langle V_4, V_4 \rangle$, see Figure3-1, associated with the location of a plot.

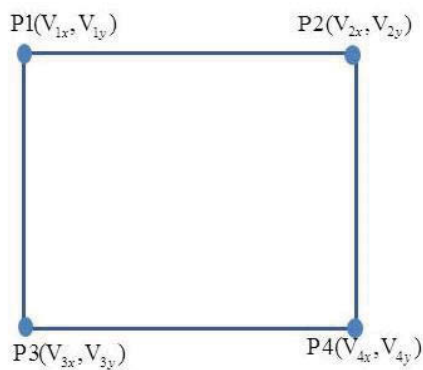


Figure 3-1 Plot Drawing-position of each vertex.

row and first column, Step4 computes the vertex of all plots and their widths and heights. Each plot is bounded by the flexible size of the display screen. Figure 3-2 contains examples to explain the flexibility of the technique representation on various screen size.

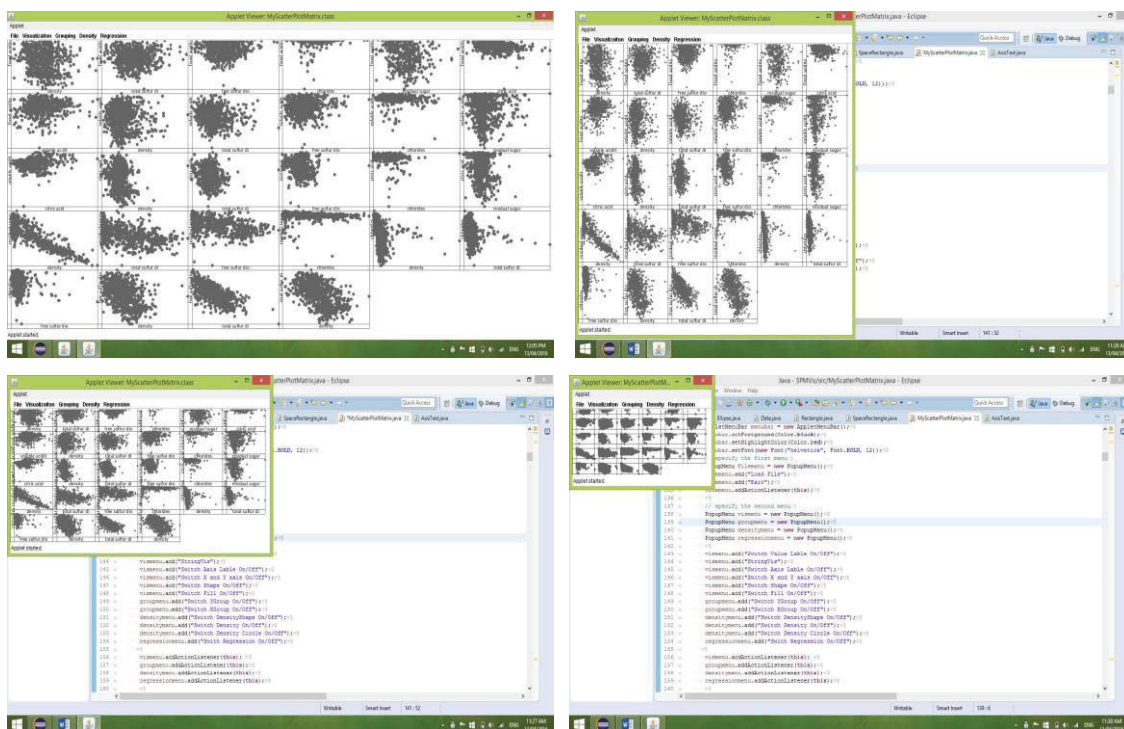


Figure 3-2 Wine dataset with four attributes is shown in different size of display screen (up-left) is the full screen display; (up-right) is the 50% screen display;(down-left)is the 25% screen display; (down-right) is the 12.5% screen display.

Step 4 – Point Positioning: Step 4 assigns the graphical layout, followed by \ positioning the data into their plots accordingly. In this step, we use visual cues (colour, shape and size) to differentiate alternative value types. To interact with visual representation, Step 4 employs the interaction scheme along with animation at this stage as well.

3.2 Algorithms

This section explains the methodology. Specifically, Section 3.2.1 illustrates the technical specifications, while Section 3.2.2 discusses the space optimised scatterplot matrix in detail.

3.2.1 Technical Specification

To streamline the graphics notation and convention in technical sections, all the symbols and notation are defined in this section.

In the geometry space, R^n represents an n -dimensional plane in Euclidean geometry, therefore, R^2 is a two-dimensional plane.

Generally, in the graphics with a two-dimensional geometry space, N indicates a node. A subset of nodes are represented by $N = \{n_1, n_2, \dots, n_m\}$, where m indicates the number of data points.

Specifically, in graphics design of scatterplot matrix, the data presented in each plot are different. In this thesis, for example, each plot can be seen as a small space. The data in a matrix is described as:

$$N = \left\{ n_1 (P_{1x}, P_{2x}, \dots, P_{ix}) n_2 (P_{3x}, P_{4x}, \dots, P_{jx}) \dots n_m (P_{px}, P_{tx}, \dots, P_{kx}) \right\}$$

OR

$$N = \left\{ n_1 (P_{1y}, P_{2y}, \dots, P_{iy}) n_2 (P_{3y}, P_{4y}, \dots, P_{jy}) \dots n_m (P_{py}, P_{ty}, \dots, P_{ky}) \right\}$$

Where:

P_{ix} / P_{iy} - is the x-axis / y-axis in the i^{th} plot;

m - is the number of notes.

Given a dataset with k variables (dimensions), the traditional scatterplot matrix is a $k \times k$ plot matrix, with k rows and k columns. Figure 3-3 is the traditional layout of the scatterplot matrix, where the viewer can discover the relationship among the neighbour variables.

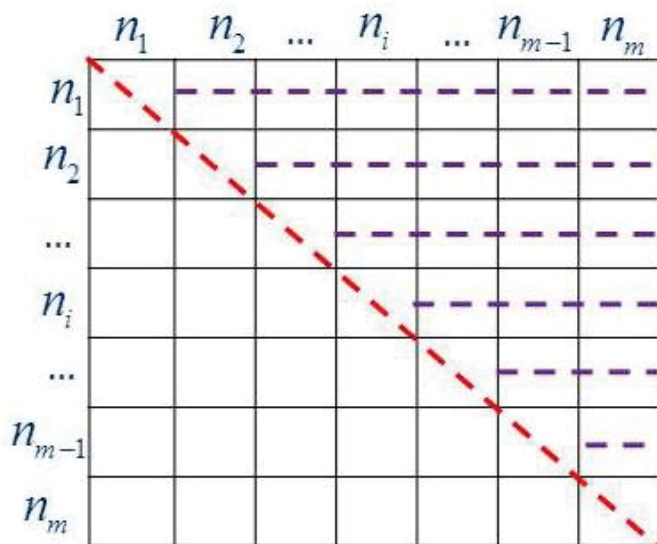


Figure 3-3 Traditional scatterplot matrix layout.

From Figure 3-3, it is obvious that the layout of a traditional scatterplot matrix is space wasting. In addition, two types of plots in the matrix reduces the space utilisation rate: One is the plots in which their X axis and Y axis represents the same variables, e.g. $(n_1, n_1)(n_2, n_2)(n_i, n_i) \dots (n_m, n_m)$; the other category of plots is symmetric plots, which are the array pairs in the following format: $\{(n_1, n_2), (n_2, n_1)\}; \{(n_3, n_4), (n_4, n_3)\}; \dots; \{(n_i, n_j), (n_j, n_i)\}$. In this case, only half of the plots need to be remained, so the plot either $(n_1, n_2), (n_3, n_4) \dots (n_i, n_j)$ OR $(n_2, n_1), (n_4, n_3) \dots (n_j, n_i)$ normally can be deleted, and the layout of matrix with plots reduction is in Figure 3-4.

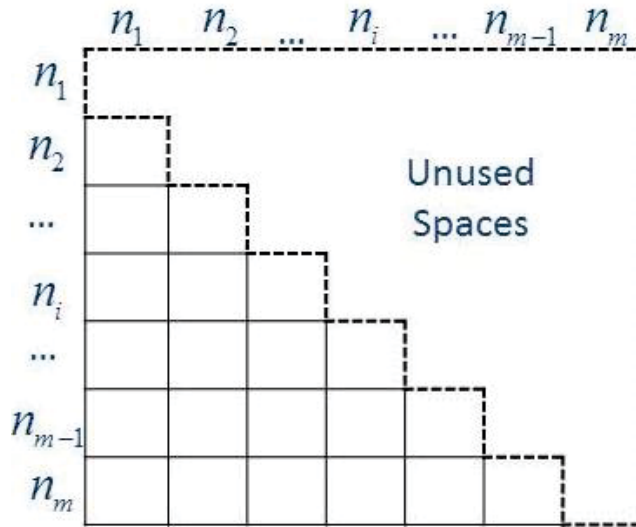


Figure 3-4 the layout of a matrix with plots reduction

In the graphics, a plot is also a rectangular. A rectangular is composed by four vertices and four edges, presented in a set of

$$\{V_{11}, V_{12}, V_{13}, V_{14}\} \{V_{21}, V_{22}, V_{23}, V_{24}\} \dots \{V_{i1}, V_{i2}, V_{i3}, V_{i4}\} \dots \{V_{t1}, V_{t2}, V_{t3}, V_{t4}\}$$

And

$$\{E_{11}, E_{12}, E_{13}, E_{14}\} \{E_{21}, E_{22}, E_{23}, E_{24}\} \dots \{E_{i1}, E_{i2}, E_{i3}, E_{i4}\} \dots \{E_{t1}, E_{t2}, E_{t3}, E_{t4}\}$$

Where:

t - represents the number of plots.

After the introduction of the technical convention, we now review the mathematical theory in our technique.

Suppose the dataset contains n variables, we might have a $n*n$ matrix and after reducing the duplicated plots, the total number of the plots N in the matrix would be in Equation 1.

$$N = n * (n - 1) / 2$$

Equation 1 The number of plots

Considering the balance between space utilisation and data relationship clarity, we position the plots as follows.

The number of row in a matrix is defined as NR , see Equation 2:

$$NR = Roundup\sqrt{N}$$

Equation 2 The number of plots per row

The number of column in a matrix is defined as NC , see Equation3:

$$NC = ceiling\left(\frac{NR}{N}\right)$$

Equation 3 The number of plots per column

Then, we start to draw the matrix by the principal of processing.org. See Figure 3-5. The whole display screen is an axis. The start point is the upper left corner point $(0,0)$, and other points can be described as $p(x_i, y_i)$. A line is composed

by two points, and it is defined as $line(x_i, y_i, x_j, y_j)$, while (x_i, y_i) is the start point of the line, (x_j, y_j) is the end point. A rectangular is defined as $rect(x, y, width, height)$, where the value of x and y is the value of the upper left vertex of the rectangular, the $width$ is the width of the rectangular and it can be calculated by the x axis value of the upper-left (bottom-left) vertex and upper-right (bottom -right) vertex, which equals to

$$width = P_{x_{upper-right}} - P_{x_{upper-left}} \text{ OR } width = P_{x_{bottom-right}} - P_{x_{bottom-left}} ;$$

The height of the rectangle is $height$, and it can be calculated by

$$height = P_{y_{upper-right}} - P_{y_{bottom-right}} \text{ OR } height = P_{y_{upper-left}} - P_{y_{bottom-left}} .$$

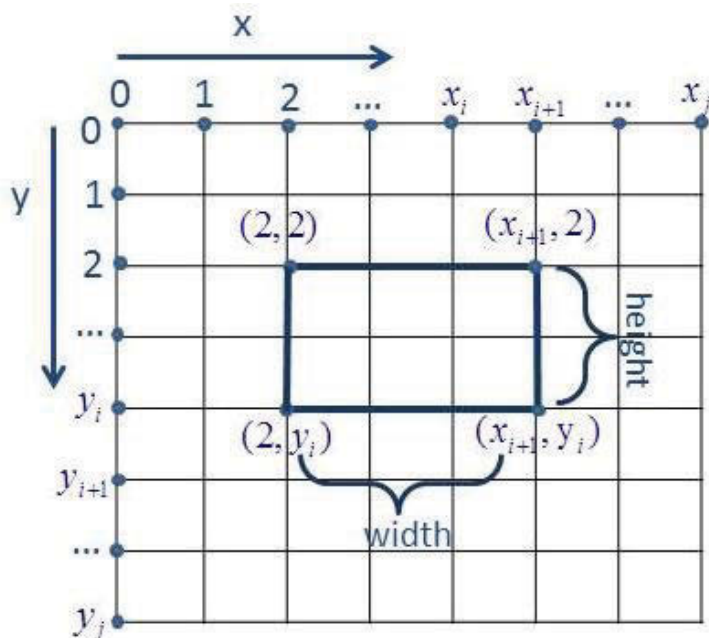
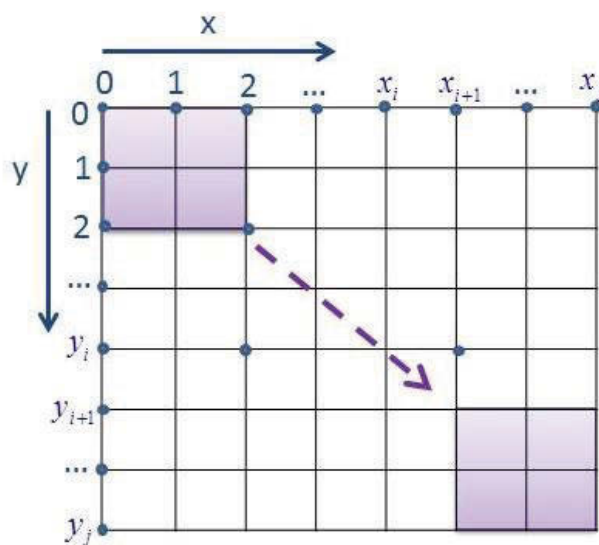


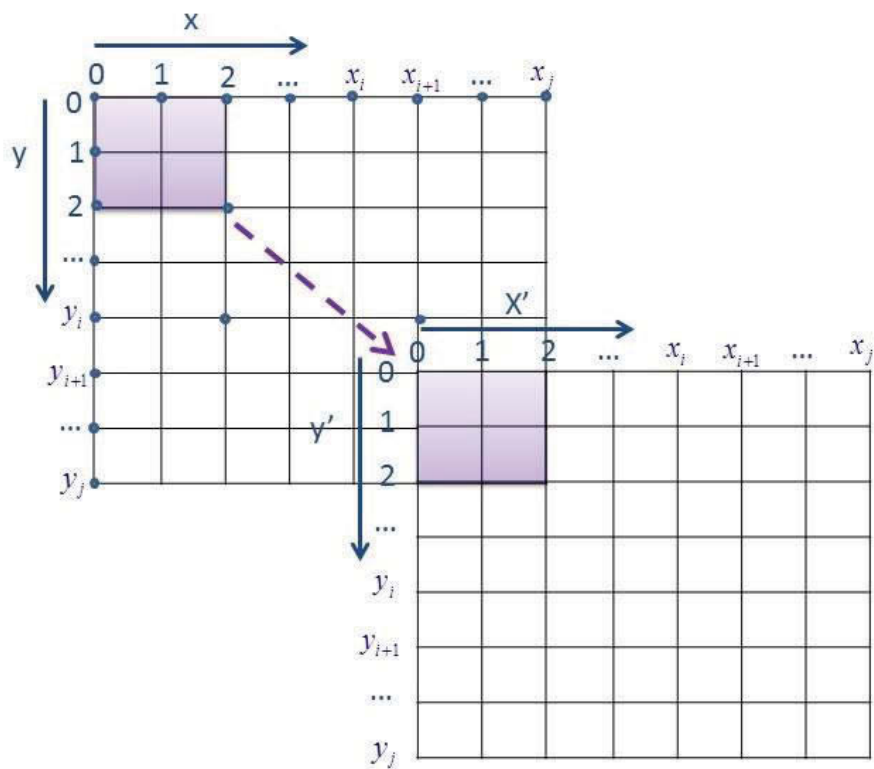
Figure 3-5 A Example of a draw

Particularly, it is important to understand how the shape is moving and rotating in the preceding diagram, see Figures 3-6 (a) (b) and Figures 3-7 (a) (b). In order to move a draw, the coordinate system should be moved to a new position, and then redraw the square on the same point. Similarly, it takes two steps to rotate a draw, the first step (S1) is to transfer the start point (upper left point) of the axis to a new position, where the draw is to be placed; Step2 (S2) is to rotate the axis 45 degrees, then draw the square based on the original coordinates.

Generally, an essential principal of drawing is to change the coordinate system through translation or rotation, and then draw with the original. Never change the position of the draw.

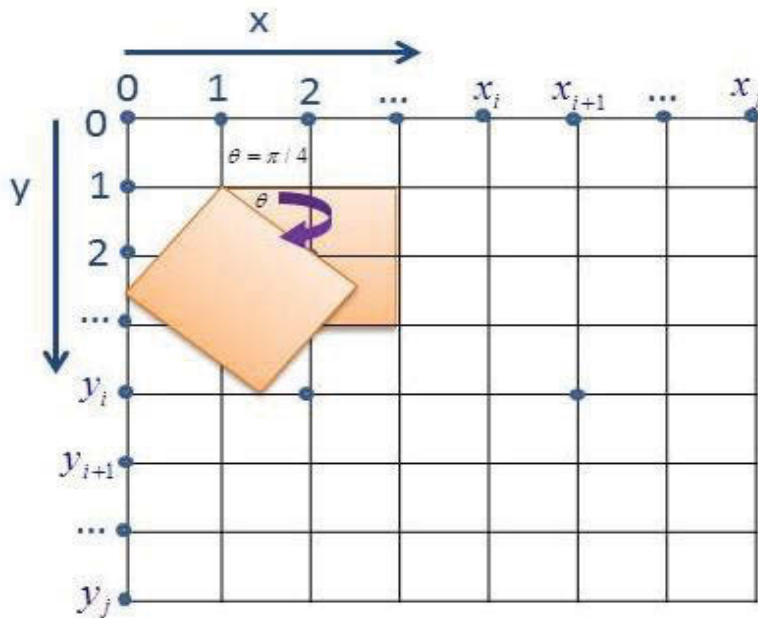


(a)

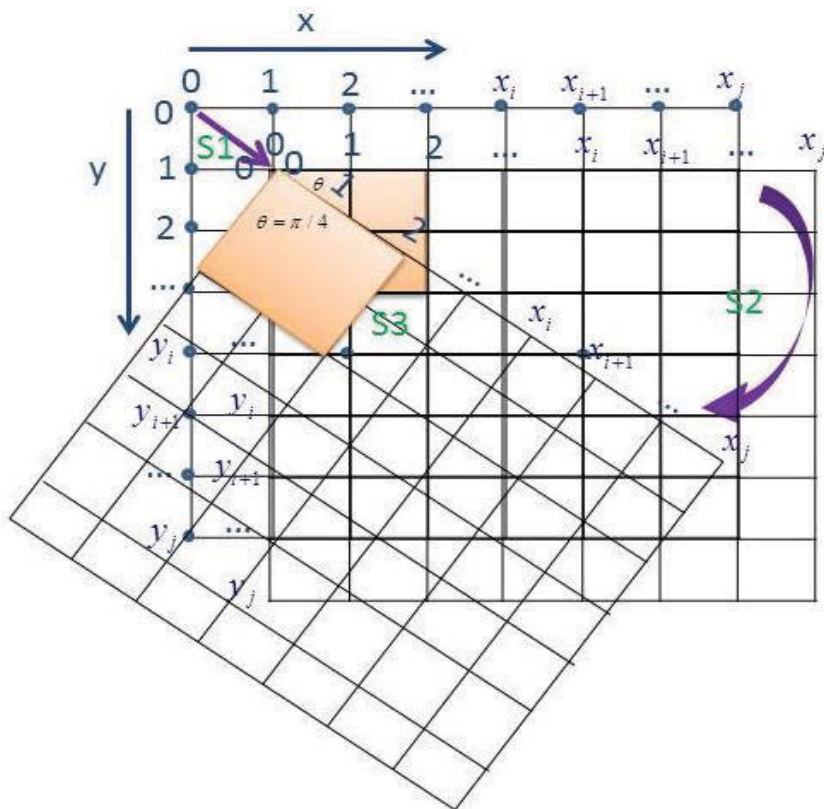


(b)

Figure 3-6 a) The Example of moving a draw; b) the processes of moving a draw



(a)



(b)

Figure 3-7 a) The sample of rotating a draw (b) the processes of rotating $\pi/4$ of a draw

3.2.2 Implementation Algorithms

The space optimised scatterplot matrix is based on the layout arrangement of the plots reduction matrix described in Section 3.2.1. The principle of repositioning the plots is to retain the visual ability of neighbour variables and to fully utilise the display space as well. To achieve optimisation goal, we use fill – in method and colour mapping approach. In the first stage, we calculate the number of plots per row and per column that ensure the plots expressed the same variables as being continuous, see Figure3-8; in the second stage, we map the data into the relative plots; in the third stage, we use a colour metaphor to visualise one more data dimension and by this approach, the viewer can distinguish the variables both hierarchically and horizontally.

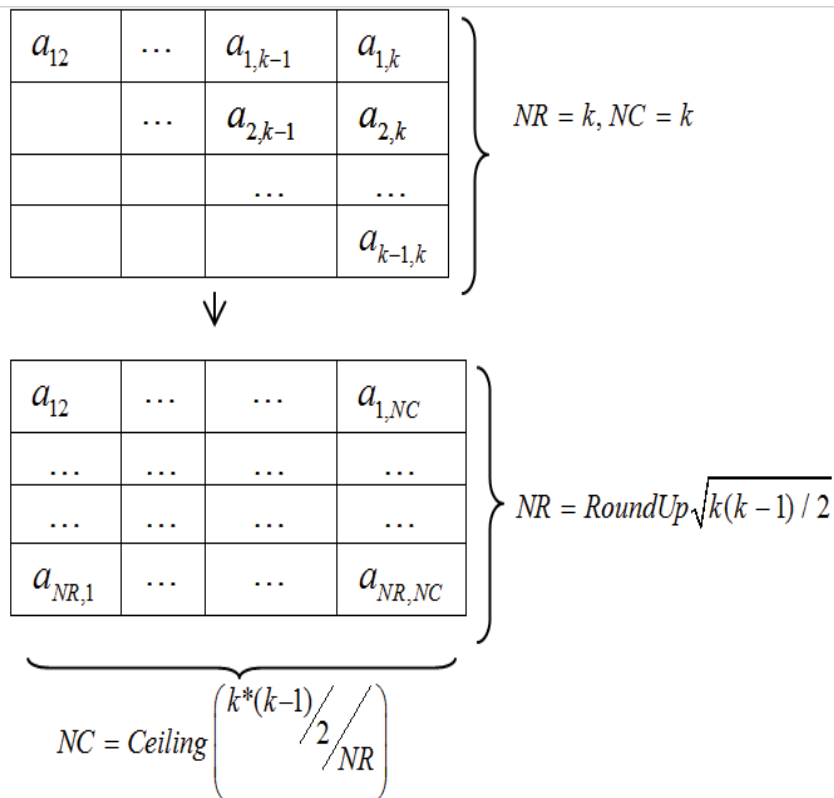


Figure 3-8 The reposition method for scatterplot matrix

Where,

K - the number of variables,

N - the total number of scatterplots

NR – the number of plots per row.

NC – the number of plots per column

The algorithm for rearranging the plots is outlined in Figure 3-9:

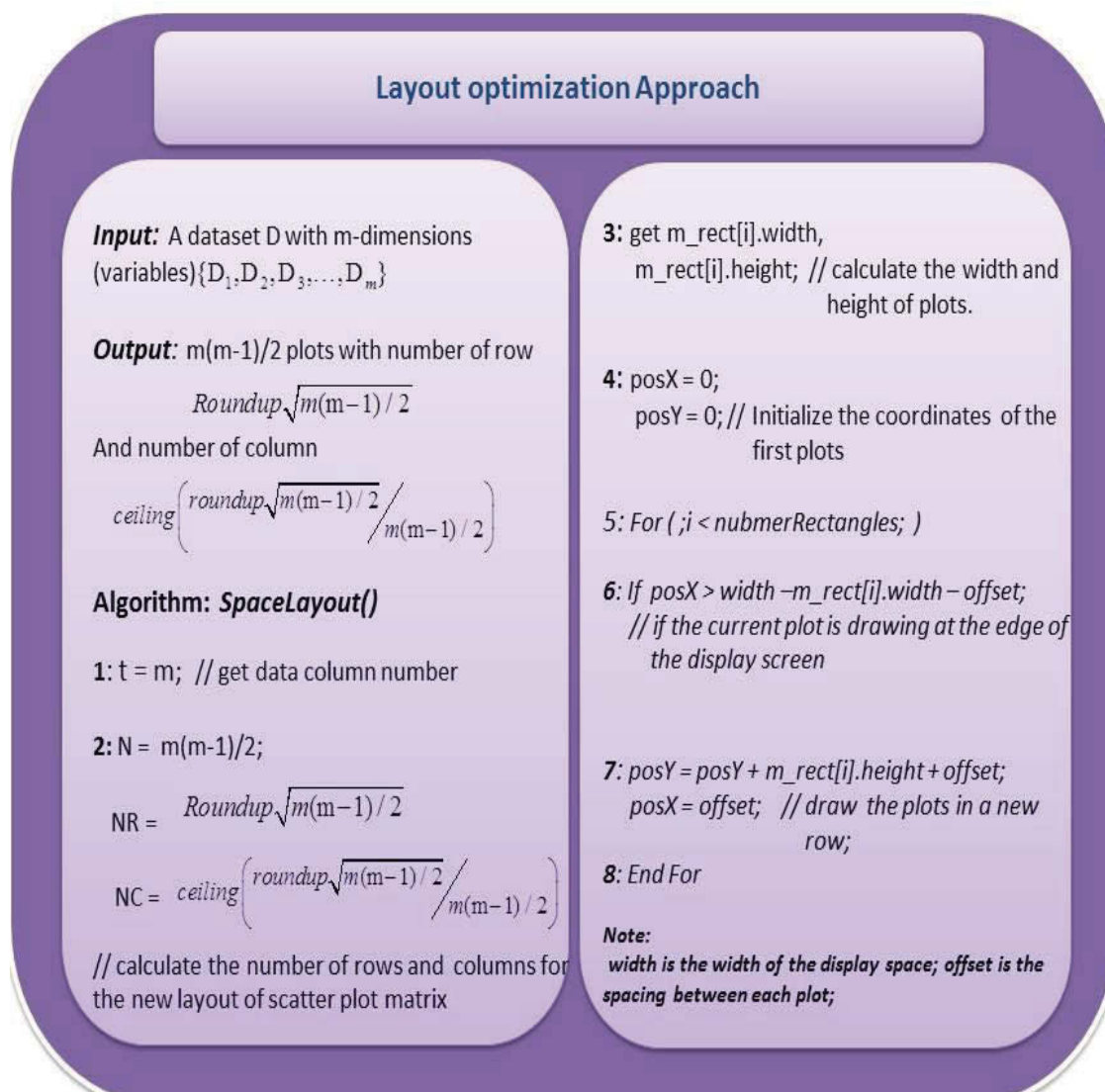


Figure 3-9 The algorithm of Layout Optimisation Approach

It is obvious that after the rearrangement of the position for each plot, the plots

could not be in a linear order. The main reason is that some of the plots had been moved up in order to keep a higher rate of space utilisation. Therefore, to retain the benefits of traditional scatterplot matrix, having a linear correlation between multiple variables, we introduce a third parameter, colour, to remedy the disadvantages in our proposed space optimised scatterplot matrix (Also an interaction mechanism).

The main idea is give each plot a colour to distinguish them based on their name of x -axis or y -axis. The approach summarises the existing capability of a matrix to discover the variable relationships easily is described in Figure 3-10.

Highlighting Approach - Variable Relationships

Input: Dataset D with m variables.
Output: plots classifications with different colors.

1. `getRowName()/ getColumnName();` // get the name of X-axis or Y-axis for each plot.
2. `For (; k < l ;)`
3. `If (m_rectangles[k].rowName == m_rectangles[i].rowName)`
4. `m_rectangles[k].color = m_rectangles[i].color;`
5. `break;`
- 7: `End For;` // if the axis's name of plots are the same, they are given the same color;

Note: we use `Random()` and `HSBColor` in the plots coloring design.

Figure 3-10 The algorithm highlighting the approach to reflecting the variable relationships.

Figures 3-11 (a) and (b) are samples of the comparison between remedying the disadvantages of the space utilisation by colour property and the traditional scatterplot matrix. Specifically, we use wine dataset (<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>) to do the experiment. Firstly, choose 8 variables from the dataset accordingly, the scatterplot matrix can be divided into 8 classifications. Including: *fixed acidity*, *volatile acidity*, *citric acid*, *residual sugar*, *chlorides*, *free sulfur dioxide*, *total sulfur dioxide*, *density*, *PH*, *sulphates*). Based on our classification principal, each category belongs to one type of colour. So we can find that the *fixed acidity* (the first variable in dark yellow) has a similar influence on all the other attributes, while the *PH* (the variable in pink) has a lesser influence on the *density sulphates*, but has a stronger influence on the rest of the variables.

3.3 Interaction Mechanism

A navigation mechanism can enable users to interactively adjust views to reach the clearer view of a complicated graph, and having this scheme that let users discover information. Capabilities of visualisation tools can maximize human capabilities to perceive and understand complex and dynamic data.

Up to now, there have been many interaction methods that use focus + context views (Stasko & Zhang 2000), overall + details views (Plaisant, Carr & Shneiderman 1995), fish eye views (Schaffer et al. 1996) and so forth that have been proposed in information visualisation.

The scatterplot matrix, a widely used Multi-dimensional data visualisation technique, is particularly helpful in pinpointing specific variables that might have similar correlations to the genomic or proteomic data.

Interactivity is the key to making the scatterplot matrix method more useful. An important step involved in the interactive process is user navigated visualisation. One of the most important issues involved in navigation is that the users are always able to choose their preference information. This allows users to maintain the perception of what they would like to discover from the large information spaces. This also assists users in making further decisions about their discovery, while interactively navigating through the large amounts of information.

3.3.1 Introduction

Although the optimised layout in a scatterplot matrix is efficient in terms of space utilisation and information relationship representation, the issues of “view-ability” to produce user-friendly interactive interfaces and the ability to explore information efficiently are critical. Especially visualise a multi-dimensional dataset. The reason is that in large data visualisation techniques, it is hard to discern between data when thousands of items in a dataset are displayed concurrently. Therefore, an efficient and effective interaction scheme, combined with a visualisation which provides users with important knowledge, is essential when learning from a large dataset.

The interaction mechanism should enable users to interactively discover how to adjust views to reach their destination (or goal), allowing them to obtain deeper understanding of relationships among data and variables (or different dimensions). With a navigation scheme that lets users pick their preference methods based on their accordingly needs. Capabilities of visualisation techniques can maximize human abilities to understand high dimensional data , and promote their working efficiencies in terms of saving time and computing resources.

In interaction control for space-optimised scatterplot matrix, the interaction is applied in a visualisation process that allows users to change their preferred views on the same dataset to explore information more deeply.

3.3.2 Interaction Method

Up to now, there are many interaction techniques in Multi-dimensional visualisation, such as “focus + context” (Stasko & Zhang 2000), Zooming in and out (Baker & Erik, 1995), dynamic queries (B.Shneiderman,1994), etc., which have been proposed in data visualisation, and only very few techniques have already been applied to scatterplot matrix technique.

“Dynamic Queries” (Shneiderman 1996), is one of the most popular interactive techniques used in information visualisation. It is a natural method for requesting data when the output is going to have a visual form, and could continuously update search results as users select buttons to gain the answers to simple questions. In our thesis, we use this method to implement the interaction function. Specifically, we use two properties, colour and shape, to differentiate views from the same dataset so that users can gain more information from different perspectives.

Differentiation in colour attracts the attention of users, as it is one of the most effective ways to enhance and clarify a presentation (Stone 2006). Figure 3-12 explains a simple user query interactive method in our scatterplot matrix. The user can choose their preference and get a further view. Figure 3-13 and Figure 3-14 are the updating results by clicking “votalie acidity” by colour property and shape property respectively. Figures 3-15 (a) (b) (c) are the extensions of shape property working on the layout of each plots, they help users to have a different view of the overall dataset.



Figure 3-12 The interaction method – user query

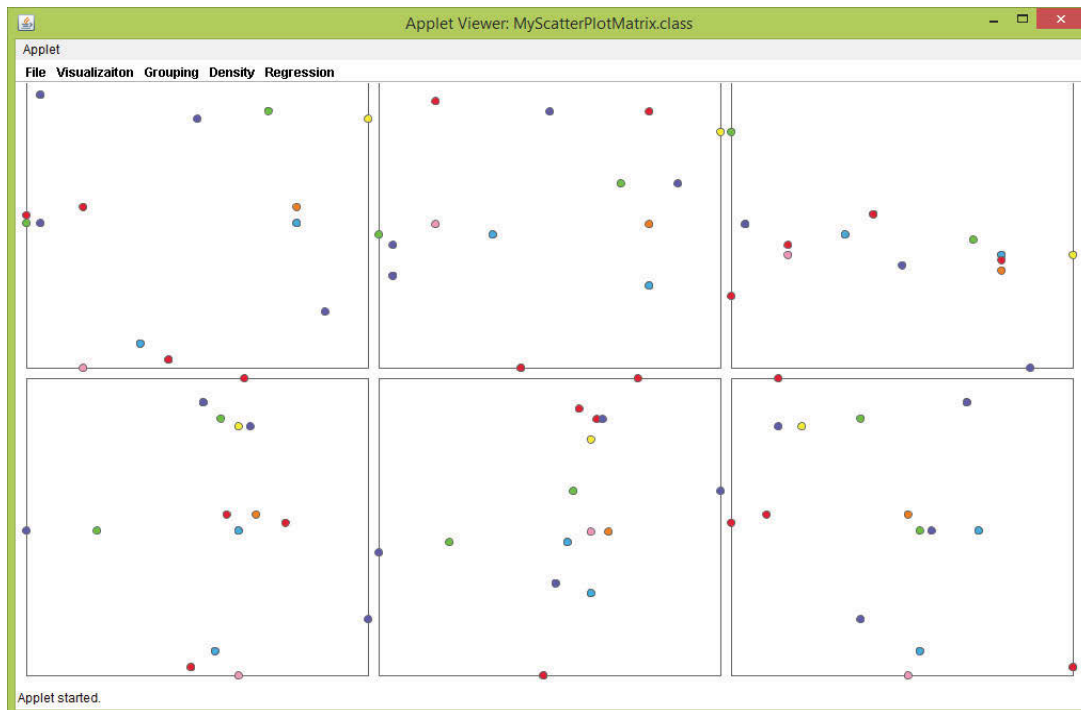


Figure 3-13 The user query result with Colour property

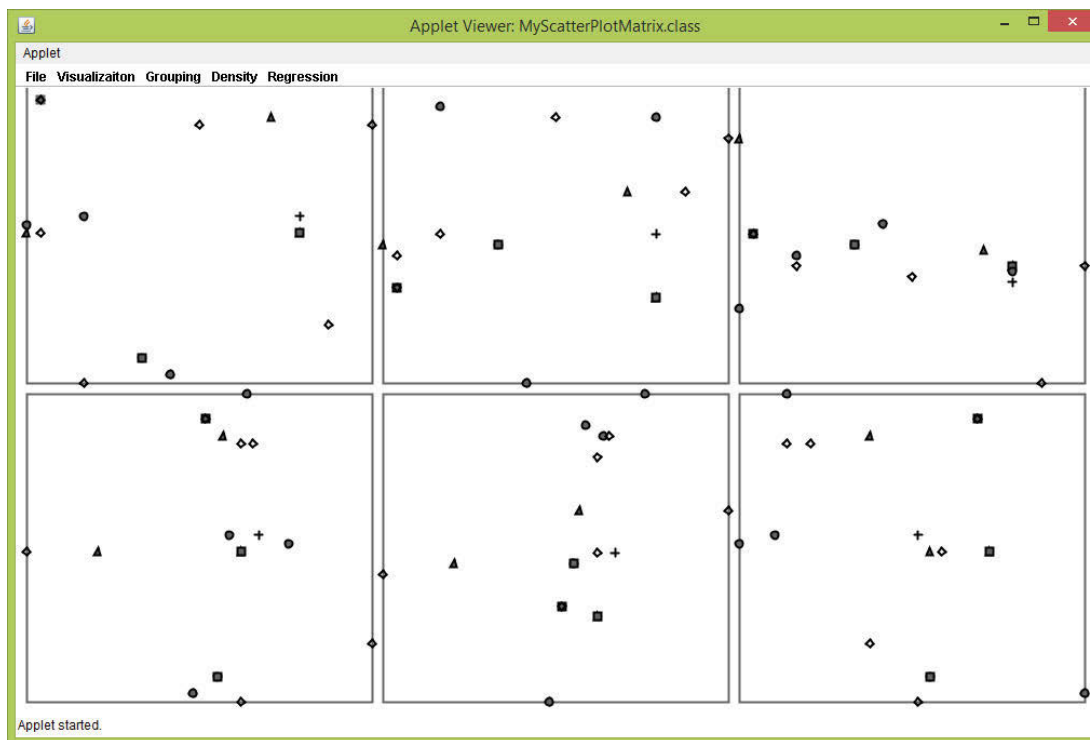
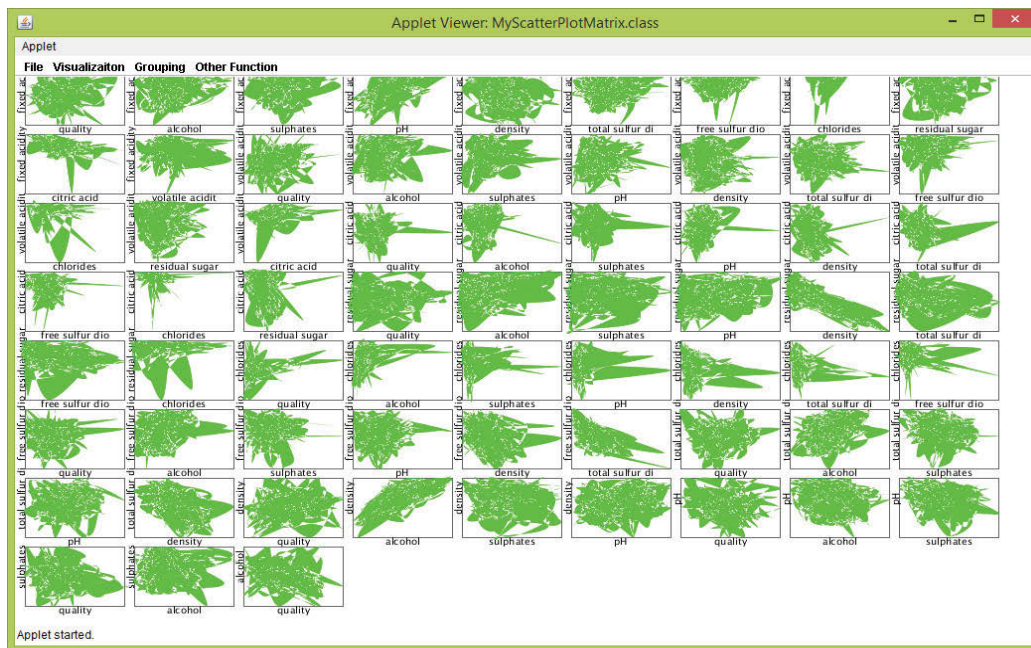


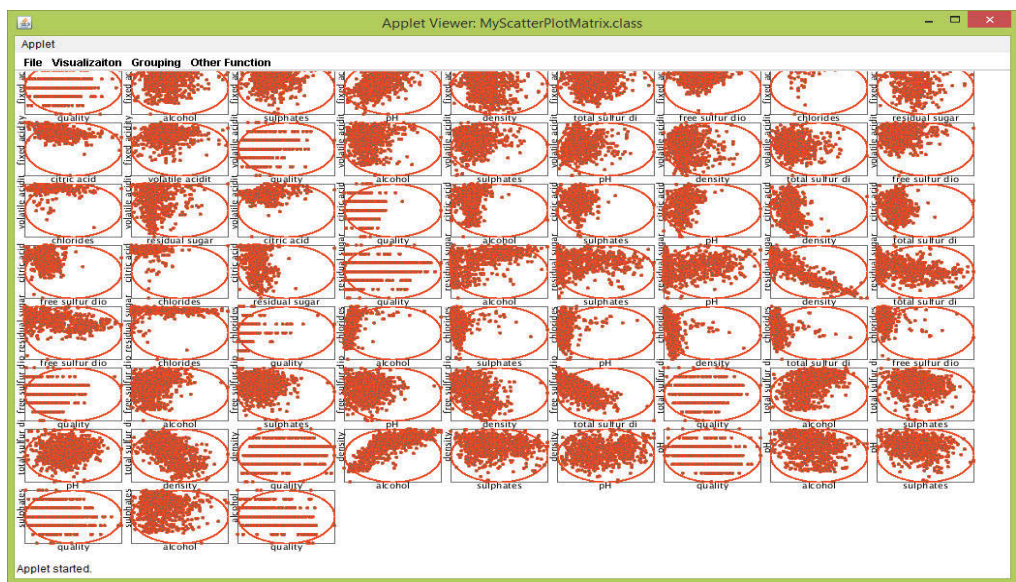
Figure 3-14 The user query result with Shape property



(a)



(b)



(c)

Figure 3-15 (a) (b) (c): Examples of Shape property working on the layout of the datasets: a) represents the data points connected into different shapes, and this helps to distinguish the overall distribution for each plots; b) the green line is the diagonal line of each plot, and from the gap between the other lines and diagonal line, it provides another overview of the dataset distribution without any points displayed; c) we use this view to detect the density of data points distribution. It is more convenient to find the centre point in a circle shape compared with drawing in a plot (rectangular), but an important factor is that the circle should be tangential with edges of each rectangular (plots).

3.4 Evaluation

This section evaluates this space optimised scatterplot matrix according to a set of design guidelines. The first objective of this research is to design a layout which meets the traditional scatterplot matrix design advantages, and also can fully utilise the display spaces. Therefore, from the technical point, we will investigate our new scatterplot matrix in two areas, space utilisation ratio of display screens, and the degree of difficulty in discovering pairwise variable relationships, in comparison with traditional scatterplot matrix techniques. In addition, to further investigate how well space optimised scatterplot matrix work in the scenario based tasks in Multi-dimensional data visual analysis process, we have conducted a user study to compare with the traditional scatterplot matrix.

3.4.1 Performance Evaluation

Our performance evaluation metric is space saving and improving the utilisation rate.

The experimental environment is described below. Firstly, Java 2.0 with Eclipse Platform was used to develop the prototype (the Java program) that implements our space optimised scatterplot matrix. The Java program was executed on a Personal Computer with the CPU: Intel Core i5-5200U, 2.7 GHz, and 8GB of ROM. Secondly, the space utilisation rate is how much the display screen has been utilised to view the results. We ran our application in the above environment with datasets which contain different number of variables (dimensions). The details of the datasets are described in Section 3.4.3 supplementary views.

Figures 3-16 (a) and (b) are the comparison figures of space utilisation. Obviously, the number of scatterplots is proportional to the number of the variables, which means that if the number of data attributes is increased in one dataset, the space of each spot is getting smaller and the number of the plots to be reduced is increased. Refer to Equation 4:

$$\begin{cases} S = \{S_1, S_2, \dots, S_N\} \\ a = N/S_i, i = \{1, 2, \dots, N\} \end{cases}$$

Equation 4 the relationship between the number of scatterplots and the number of variables

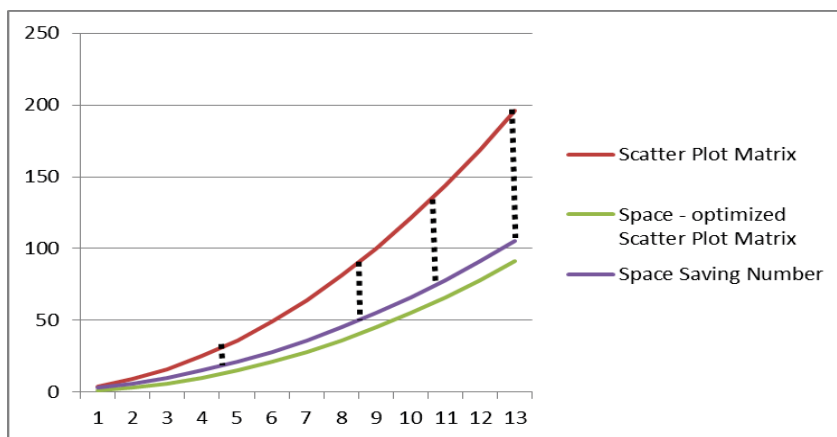
Where:

S - Size of the display screen

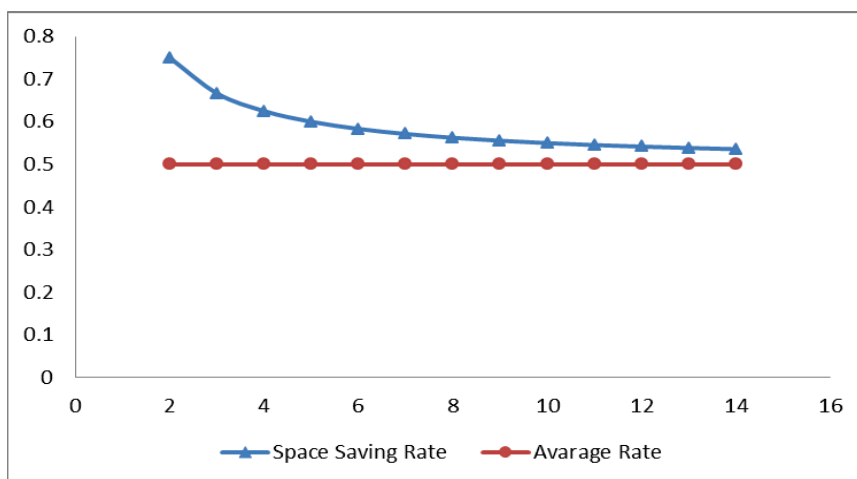
N - Number of the scatterplots

S_i - Size of the i^{th} scatterplot

Suppose a dataset has four dimensions, the plots can decrease from 16 to 6, which saves 61% of the spaces. Accordingly, if the dataset contains 8 or 12 variables, their plots will reduce from 64 or 144 to 28 or 66. Their space saving rates are 56.25% and 54.2% respectively. In Figure 3-16(b), we found that although the space-utilisation rate is decreased while the dimension is increased, the rate is always above 50%.



(a)



(b)

Figure 3-16 (a) The saving number of scatterplots b): The saving space rate of our new scatterplot matrix

3.4.2 User Studies

We conducted a usability study with users who are mostly unfamiliar with scatterplot matrix. The goal was to test whether the approach toward the design improved the user satisfaction on some challenges of normal visualisation techniques, including: Space Utilisation, Visual Clarity, and Information Navigation.

We involved 10 participants (Male and Female), who are PhD students and graduate students in different research areas and majors, including Science, Information Technology and Engineering. Four questions are designed for the evaluation, and all the participants needed to do a comparison between our space-optimised scatterplot matrix with the original scatterplot matrix by answering the questions. Additionally, they need to tell how they regard these two applications on the three aspects: space utilisation, visual clarity, and information navigation. Lastly, the participants also needed to give their overall preference for using the space-optimised scatterplot matrix and the original scatterplot matrix.

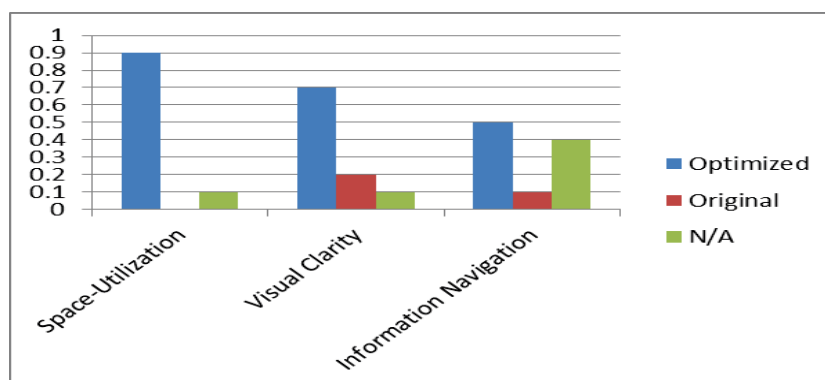


Figure 3-17 The results of user preference

From the evaluation result, the space-optimised scatterplot matrix is generally better than the original scatterplot matrix in respect to the space utilisation, visual clarity and also information navigation. The subjects showed high preference for using the space-optimised scatterplot matrix with their own dataset. Refer to Figure 3-17.

3.4.3 Supplementary Views

This section provides more view results (Figure 3-18, Figure 3-19, and Figure 3-20) from our space-optimised scatterplot matrix, which help users to understand more about its visualisation performance.

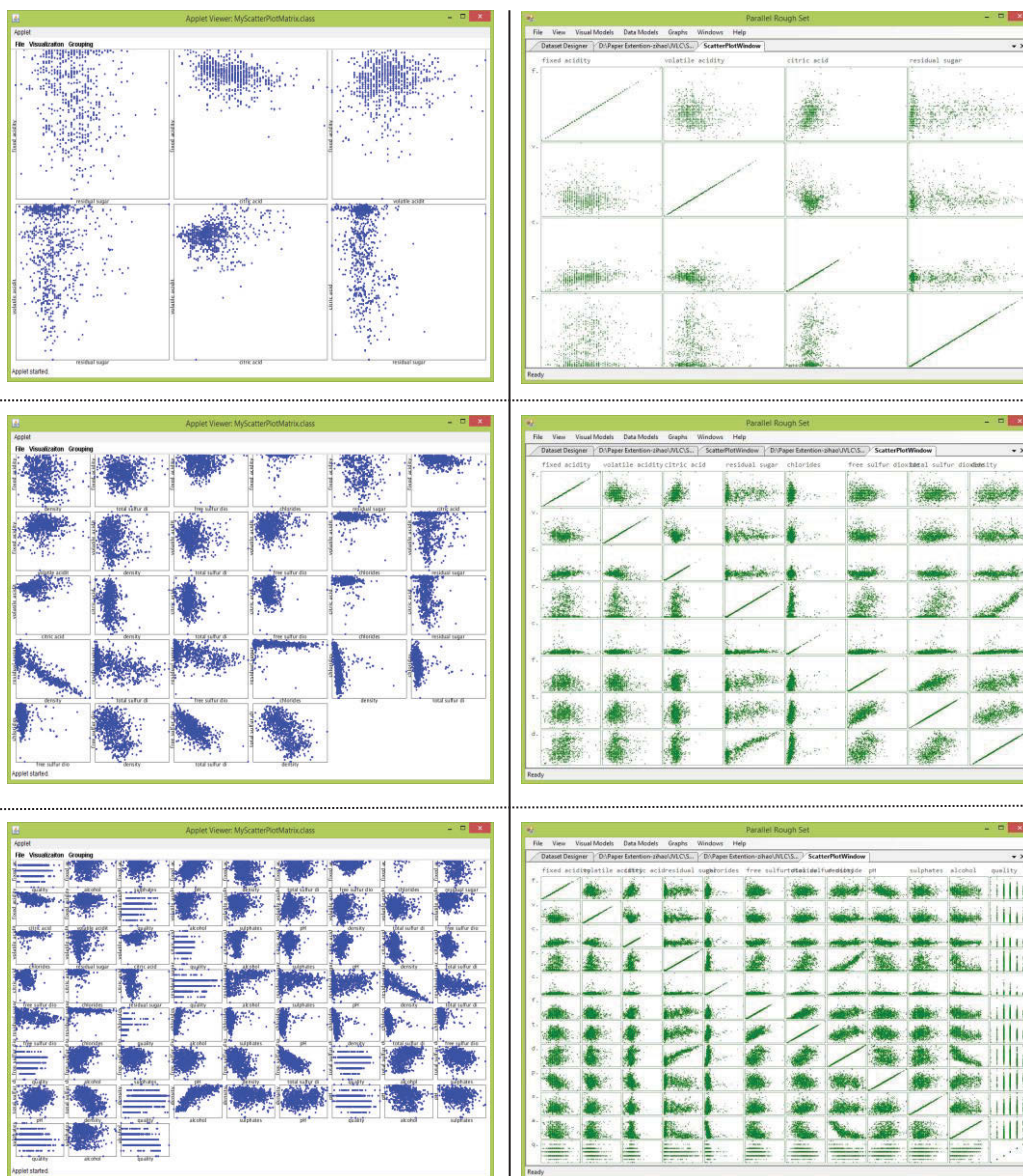


Figure 3-18 An overview comparison of plot arrangements in the new layout and the original layout

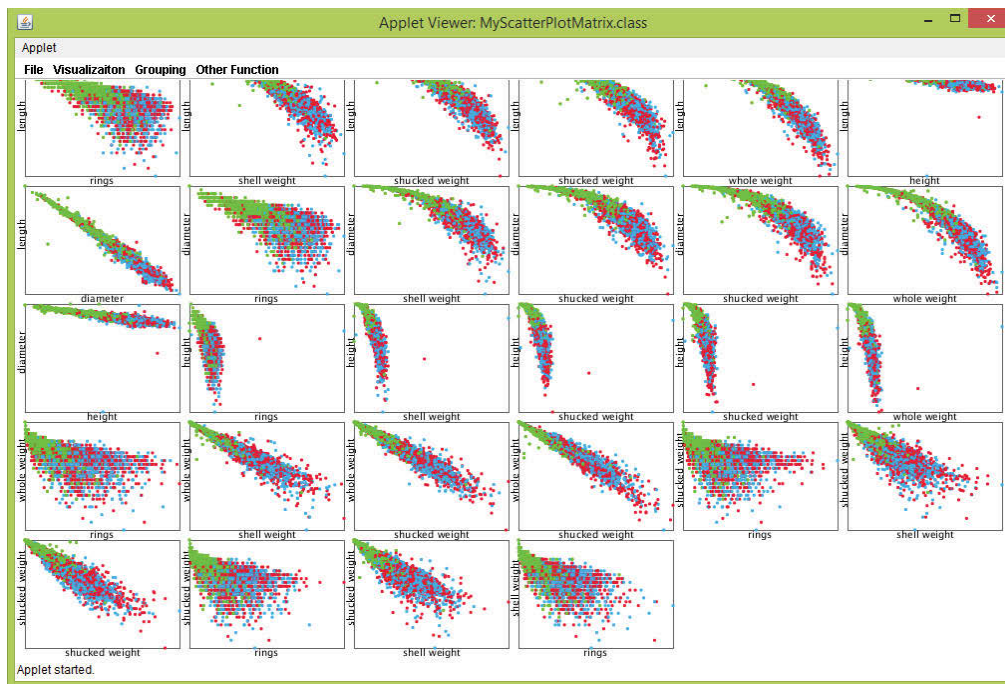


Figure 3-19 Differentiation in Colour; Experiment with abalone Dataset

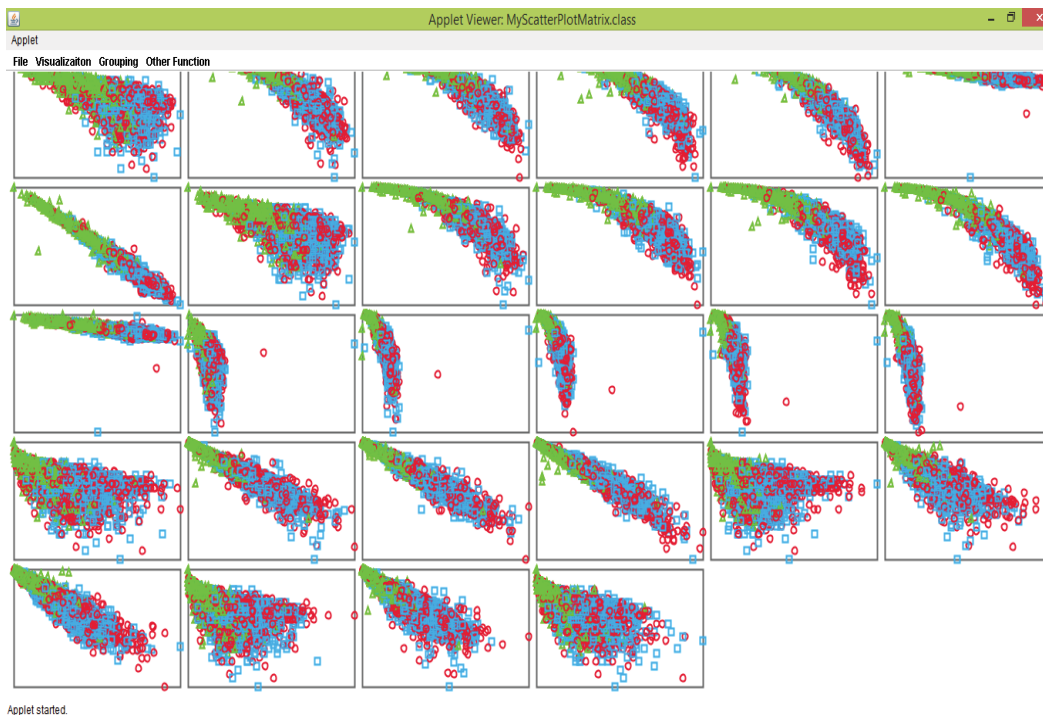


Figure 3-20 Differentiation in Colour and Shape; Experiment with abalone Dataset

3.5 Summary

A scatterplot matrix is a table of all the pairwise scatterplots of variables on a single view. In this new scatterplot matrix visualisation, we solve the space - utilisation cost problem, and also improve the clarity of navigation when scrolling up and down. This scatterplot matrix is especially designed to enhance the understanding of multi-dimensional dataset.

Space-optimised scatterplot matrix is an effective and efficient technique for visualizing Multi-dimensional dataset (Herman, Melançon & Marshould 2000). This technique optimises the space-utilisation by reducing scatterplots and rearranging the display layout that allows the space to be fully utilised, and all the relationships among each variable can be displayed with a higher clarity. What's more, we apply a colour property to differentiate the variables, and the advantage of traditional scatterplot matrix, displaying pairwise variables relationships in order, is also implemented.

In our technique, an interactive mechanism - user queries is also applied to make our tool user-friendly by colour and shape properties, which brings more benefits on our new scatterplot matrix visualisation.

Chapter 4 Interactions with Scatterplot Matrix Visualisation

THIS CHAPTER AIMS to introduce the interactive scatterplot matrix visualisation in Sections 4.1 and 4.2; then a detailed introduction is given to the evaluation methods applied to this technique in Section 4.3. Particularly, this chapter will conduct an evaluation for the interactive exploration approach from two perspectives: User Studies (Section 4.3.1) and Case Studies (Section 4.3.2).

Multi-dimensional data exploration presents a great challenge to information visualisation. Because features of data are inherently sparse in high dimensional data and the over-plotting of visual display makes it even more difficult to observe any useful patterns. However, visualisation methods for large dimensional data are not usually effective due to the density of high dimensions and the limitations of screen display. Therefore, although the interaction is still limited some of the contextual information could be lost during the navigation ,interactive zooming is still an aid for exploring and reducing the number of dimensions, such as the zooming function provided by Diesburg et al (2010).

The efficiency of knowledge discovery tends to decline while the processing cost of information interpretation tends to increase. Because some are noisy data and it is not necessary that all the dimensions need to be analysed. This phenomenon is also known as the curse of dimensionality which was first apparently coined by Bellman et al (2005) , who mentioned that data samples

would grow exponentially according to the changes of the number of dimensions because of the necessity of fitting a multivariate function for a given degree of accuracy.

Therefore, dimensionality reduction is important to be a preprocess method dealing with the large volume of dataset. Particularly, dimensionality reduction is important in many application domains for being facilitated with classification, visualisation dealing with the complexity of multi-dimensional data. It reduces the intrinsic dimensionality of the data in order to cut down the cost of time and space complexities required for subsequent computation and analytic tasks. The terms variable, feature and attribute are commonly quoted in various research fields hence we use them interchangeably.

Dimensionality reduction can be divided into feature selection and feature extraction. Feature selection is mainly to select a subset of the original variables according to selection principals. In supervised methods, the general criteria requires users to guide the selection process through choosing weighted quality metrics, therefore the selection rule would prefer the attributes weighted above the threshold. However in this case, user's expertise about quantisation would have a great influence on the effectiveness of variable selection as quantisation is typically not a trivial task. More importantly, empirical studies are the fundamental basis of applying quantisation; hence the method may work well on this dataset but might completely fail on another. On the other hand, feature extraction is a typically unsupervised technique with minimal consideration about user factors. The absence of user guidance raises the challenge of information interpretation if the result is unintuitive or not expected by the user and this is often criticised as information loss. Most techniques developed in the

past are projection based, implying that the phenomena of interest higher than second order could not be discovered. Strictly speaking, projection is orthogonal. The oversimplified pattern is not adequate to support interactive data exploration that requires iterative interaction through visualisation for the adjustment of input vectors to increase the accuracy of analytical results for decision trend analysis.

Multi-dimensional data exploration via dimensionality reduction is really a user centric task in information visualisation. Most dimensional reduction methods do not provide multiple results and make no assumption about the consideration of the user's concern. Ideally, an effective method should only require the user to guide the procedures of dimensionality reduction, in terms of specifying a centrally concerned attribute and adjusting the values of input vectors subjectively.

In the previous works, Tze-Haw Huang integrated Rough Set Theory (Ankerst) with parallel coordinates (Inselberg & Dimsdale 1991) and scatterplots (Tusher, Tibshirani & Chu 2001) for interactive feature selection. RST (Pawlak 1998) is a mathematical approach to data vagueness and uncertainty, which can be considered as discovering facts from complicated data through dimension reduction with a given dimension known as a decision specified by the user. Later in 2014, Tze-Haw Huang et al (2011) further extended his prior work with additional contributions summarised as follows:

- A feature ranking method on the results to guide the user in Multi-dimensional data analysis.
- Interactive data exploration support in scatterplot matrix for class data.

- Enhanced scatterplot matrix for decision trend analysis.

Based on his recent works, we provide more case studies to illustrate the interactive scatterplot matrix visualisation technique on different datasets, and do comparisons between parallel coordinates with Rough Set Theory and scatterplot matrix with Rough Set Theory. In addition, we carry out a pilot usability study on the visualisations.

4.1 Dimensionality Reduction

There are several techniques for visualizing multi-dimensional data, such as Parallel Coordinate (Inselberg 1985), Start Plots (Schein et al. 2002), Scatterplot matrix (Andrews 1972), Mosaic Plots (Hofmann, Siebes & Wilhelm 2000), Heat Map (Wilkinson & Friendly 2012), and Glyphs and Icons (Olcott 2006). Among them, the Parallel Coordinate and the Scatterplot matrix are considerably popular techniques for large scale datasets. Theoretically, they are capable of visualising the data with an unlimited number of dimensions nevertheless their visual efficiencies tend to decline when the number of dimensions grows.

Some developments addressed the problem by visual transformation. Guo (2010) and Artero et al (2004) used clustering to highlight the patterns of homogenous data in parallel coordinates. Peng et al (2004) applied dimension reordering to rearrange the dimension axes based on visual neighbouring similarity for clutter reduction. However, using visual transformation to enhance the visual structure still left data in high dimensional space with sparse features. Nguyen et al (2013) presented Multi-dimensional data visualisation system based on a scatterplot with flexible axis and attribute mapping. The tool also provided interaction, filtering, zooming and dynamically controlled visualisation. Although these techniques are quite effective in visualising small numbers of dimensions, dealing with high numbers of dimensions remains a challenge.

The widely accepted dimensionality reduction methods are Principal Component Analysis (PCA) (Person 1901), Multi-dimensional Scaling (MDS)

(Kruskal 1964) and Self-Organizing Map (Yasinsac et al.) (Kohonen 1990). PCA is a linear transformation method that projects the original data onto a much smaller set without original results. The selection principles are typically interested in dimensions with largest eigenvalues, known as principal components, because they explain the majority of variability. The low dimensional view that represents the high dimensional dataset is formed by rotating the principal components along the linear directions of maximum variability. MDS aims to place the data points so that the pairwise distances are preserved as effectively as possible. SOM is an unsupervised learning algorithm based on neural network model, reducing the dimensions to low-dimensional (typically 2D) layer of neurons. Locally Linear Embedding (Roweis & Saul 2000) is another popular unsupervised learning technique that computes the nearest neighbourhood of each dimension to obtain the low dimensional embedding of high dimensional data. One common drawback of these methods is that they project the dataset into extremely low dimensions that could oversimplify patterns. Projecting an information correlated dataset i.e. survey dataset, into 2D space is usually meaningless for human centric knowledge discovery.

Projection Pursuit (Huson et al.) (Friedman & Tukey 1974) is a type of statistical technique for the pursuit of the choices about possible projections in multi-dimensional data that can reveal the most details about the structure defined by a projection index. The pursuit of the possible projections globally involves a non-trivial computational intensive task (Friedman & Stuetzle 1981) . XGobi (Swayne, Buja & Hubbell 1992) is a visualisation system that integrated PP for viewing high dimensional data. The choices of possible relevance are the commonality between our work and PP. The main problem of PP is the difficulty to quantise the value of the projection index because it is possible to present

spurious interesting structures with an inappropriate projection index.

Several Visual Dimensionality Reduction (VDR) methods have been proposed by taking advantages of information visualisation at different stages. Yang et al (2002) proposed the Visual Hierarchical Dimensionality Reduction (VHDR) method by visually grouping dimensions into a hierarchy and constructing a new representation through the clusters of the hierarchy. VHDR has been integrated into XmdvTool (Ward 1994) since version 6.0. Yang et al (2003) further extended VHDR to propose a hierarchical Dimension Ordering, Spacing and Filtering Approach (DOSFA). DOSFA is similar to VHDR with additional improvements in visual structure via dimension ordering and spacing. Guo et al (2003) contributed a method that computed the entropy matrix and hierarchical clustering for low dimensional feature selection. Later, Johansson et al (2009) applied several user-defined combinations of quality metrics such as similarity, outlier and clustering to measure the importance of attributes. The attributes are selected for these weights above the threshold defined by the user. By strict definition, they are feature selection techniques using some quality metrics as a measure to determine the feature subset selection.

4.1.1 Rough Set Theory

RST was first introduced by Pawlak et al (1995; 2012) to distinguish objects into sets under the given conditions necessary to make decisions specified by *decision attributes*. In general, it seems to be of fundamental importance to many fields that require classification tasks such as feature selection, decision analysis, knowledge discovery and pattern recognition, etc.

In RST terms, a dataset is called a *decision table* which contains a finite set of data, namely the universe, denoted as U . In the *decision table*, rows of a decision table are known as *decision rules*, which give conditions to make decisions, and let $A = \{a_1, a_2, a_3, \dots, a_n\}$ represent a superset of attributes. A is further classified into two disjoint subsets $A = (C, A \cup \{D\}), C \cap D = \emptyset$ where C and D denote the condition and decision respectively. RST is unable to deal with single objects because of the impossibility of discerning some objects by the existing information, so the objects need to be grouped into a set of equivalent classes by finding their indiscernibility relation expressed as in Equation 5:

$$E(P) = \{(x, y) \in U; \forall_a \in P: a_i(x) = a_i(y)\}$$

Equation 5: The indiscernibility relation among objects

Where $P \in A$ and x, y are the objects in the universe, $a_i(x)$ is the value of an attribute a , for an object x . Equivalence classes are further classified into an approximation space where RST defines three regions of approximations namely lower approximation, upper approximation and boundary region. The

first one is where the union of all original sets are included in every set, the second is where the union of all original sets have nonempty intersection with every set, and the third is that which represents the difference between the upper and lower approximation. In our work, we only care about the lower approximation as it determines the quality of classification. Lower approximation and upper approximation are also called positive and negative regions respectively in RST terms.

4.1.2 Variable Precision Rough Set

Classic RST was designed to deal with a consistent dataset by its assumption of being not possible under a certain level of error on classification. For example, if $ab \rightarrow D$, then $cd \rightarrow D$ is considered inconsistent. This assumption of failure-free-decision-making is unrealistic in most real world datasets. To deal with inconsistency, Ziarko et al (1993) argued that probabilistic classification rules should be incorporated and hence proposed Variable Precision Rough Set (VPRS) model as an extension to RST. Beynon et al (2001) provided the detailed VPRS concept, notations and case study. The VPRS model allows the probability classification by introducing a given probability value β to deal with the restricted classification in original RST. It introduces the concept of major inclusion to tolerate the inconsistent dataset and the definition of majority is defined to lie between 0.5 and 1, which implies a less than 50% classification error.

The β position region in VPRS model is approximated as in Equation 6:

$$POS_{\rho}^{\beta} = U_{\Pr(Y|x_i) \geq \beta} \{x_i \in E(P)\} PO$$

Equation 6: The β position region in VPRS model

where $Y \in U, \Pr(Y|x_i) = |Y \cap x_i|/|x_i|$ is a conditional probability function and $E(P)$ denotes a set of equivalent classes partitioned using Equation 5. Clearly, a portion of objects with specified value β in the equivalent classes need to be classified into Y for it to be included in the β positive region. Given Equation 6, we could find the *quality of classification* that measures the percentage of

objects in conditional classes which C has approximated into the position region in decision attributes D . It is used to extract β reduct and we will explain the definition of *reduct* later. The quality of classification in VPRS model is defined as in Equation 7:

$$r^\beta(C, D) = \frac{|\bigcup_{\Pr(E(D)|x_i) \geq \beta} \{x_i \cup E(P)\}|}{|U|}$$

Equation 7: The quality of classification in VPRS model

A subset of attributes that meets the classification requirement is called a *reduct* which is sufficient to describe the original attributes without loss of classification. In the VPRS model, *reduct* is called β *reduct* or *approximate reducts* denoted as $RED^\beta(C, D)$ and according to Ziarko that a subset $P \subseteq C$ is a reduct of C with respect to D if and only if the following two criteria are satisfied:

1. $r^\beta(C, D) = r^\beta(RED^\beta(C, D), D)$ and
2. No attributes can be eliminated from $RED^\beta(C, D)$ without affecting the requirement (1).

In the first requirement, Ziarko has defined the strict satisfaction of β *reduct* as being that some attributes can only be removed if and only if its qualification of classification r^β for subset $P \subseteq C$ must not be affected by the r^β for the whole set of conditional attributes C .

Our task of dimensionality reduction is relatively computationally expensive by exhaustive approaches. Basically, we generate all the possible candidates from

the conditional attributes and test them for satisfaction of β reduct criteria. Given n conditional attributes, we start from $k=2$ until $k=n$ so there are $\binom{k}{n} = \frac{n!}{k!(n-k)!}$ combinations in the search space. There is a more efficient algorithm called *QuickReduct*, but its discussion is outside the scope here.

Algorithm 1 describes the feature selection procedures where G is a function used to satisfy the second requirement defined by Ziarko.

**Algorithm 1 Dimensionality reduction
algorithm based on VPRS model.**

Input: A dataset U with conditional C , decision D

and precision β .

Output: A set of reducts with respect to D .

1. $R \leftarrow \emptyset$
 2. for $k \in K$ do
 3. if $\gamma^\beta(C, D) = \gamma^\beta(RED^\beta(k, D), D)$ then
 4. if $G(RED^\beta(k, D))$ then
 5. $R \leftarrow k$
 6. end if
 7. end if
 8. end for
 9. return R
-

4.1.3 Feature Ranking

Typically we expect to find many *reducts* from the procedures described earlier and they have no discrepancy from RST perspective because they are all sufficient to represent the whole set of attributes without loss of classification quality. Unfortunately, it might be a legitimate concern from the user perspective as to which attribute is the most useful to start with if more than one exists. Feature ranking is commonly used in this situation that measures the correlation between classes based on ranking criteria. The correlation here refers to the linear relationship between two variables.

We applied the Spearman et al (1904) proposed rank correlation coefficient which is a non-parametric measure of statistical dependence between variables which ranks the order of data items instead of calculating the mean value. Thus, it is less susceptible to outlier or boundary items over other algorithms. Given a *reduct*, we first compute the ranking coefficient for each conditional attribute against the decision attribute, see Equation 8.

$$r = 1 - 6 \sum_{i=0}^N d_i^2 / n(n^2 - 1)$$

Equation 8: Calculate the ranking coefficient for each conditional attribute

Where d_i denotes the difference between ranks for data items and r measures the degree of linear dependency. The overall ranking weight of a *reduct* can be easily calculated by $\sum_i^N r_i$.

4.1.4 K-Means for Data Discretisation

Recall that RST gets the results in the form of classification derived from a set of objects. If the underlying numerical attributes are continuous, then there will be too many weak equivalent classes generated, remembering that a continuous data range can be theoretically unlimited.

Discretisation is a process that transfers the attributes with continuous data into their discrete counterparts. It has received significant attention as a data pre-processing technique in many data mining systems i.e. ROSETTA (Øhrn & Komorowski 1997). Equal Interval Width is the simplest discretisation method but it is vulnerable in dealing with an uneven distribution of data. In our work, the k -means clustering (Hartigan & Wong 1979) is extended to discretisation on attributes with continuous data before doing classification. It computes object similarities through distance function which generates the minimum value of the average inner-cluster separation, and so the uneven distribution of values can be well separated. Its main disadvantage is, however, that the input parameter of k clusters must be known in advance as opposed to hierarchical clustering. Nevertheless, specifying k is considered easier than defining a stopping rule for optimal clusters in hierarchical clustering.

4.2 Interactive Exploration

This section introduces the interaction mechanism in multi-dimensional visualisation. In our new visualisation technique, we use point to Region interaction and decision trend, which will be described in Section 4.2.1 and Section 4.2.2 respectively.

4.2.1 Point to Region Interaction

Identifying class patterns and their correlations, such as linear relationships, is a fundamental task in Multi-dimensional data exploration. We use scatterplot matrix to visualise the result of dimensionality reduction. A scatterplot matrix shows all the pairwise scatterplots of attributes in a single view with multiple scatterplots in a matrix format, as shown in Figure 4-1.

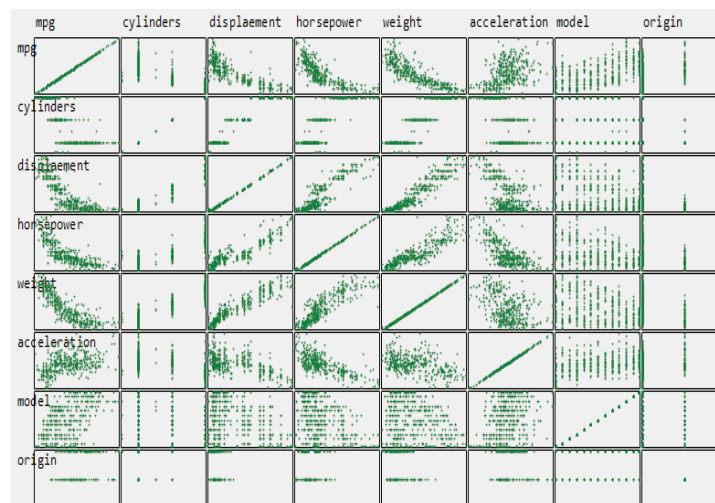


Figure 4-1 A illustration of scatterplot matrix visualisation

The motivations behind this choice are 1) it is generally more intuitive to perceive data correlation in low projection view and 2) it is less susceptible to visual clutter created by over-plotting as opposed to parallel coordinate visualisation. Interaction is an important function in our visualisation which turns the static info-graphics into a dynamic display to uncover insights by allowing users to manipulate the data transformation directly through the visual interface. In the interaction design, we allow the user to use the “focus + context” concept in interacting with scatter points directly. This interaction method can achieve

noise reduction in class selection processes. For example, when visualisation detects a point that has been clicked, the entire convex hull of a corresponding class will be highlighted and the background of the convex hull will be greyed out as illustrated in Figure 4-2. In other words, the system provides the interaction for individual data at the class level granularity, focusing on the subset with the area covered by a convex hull.

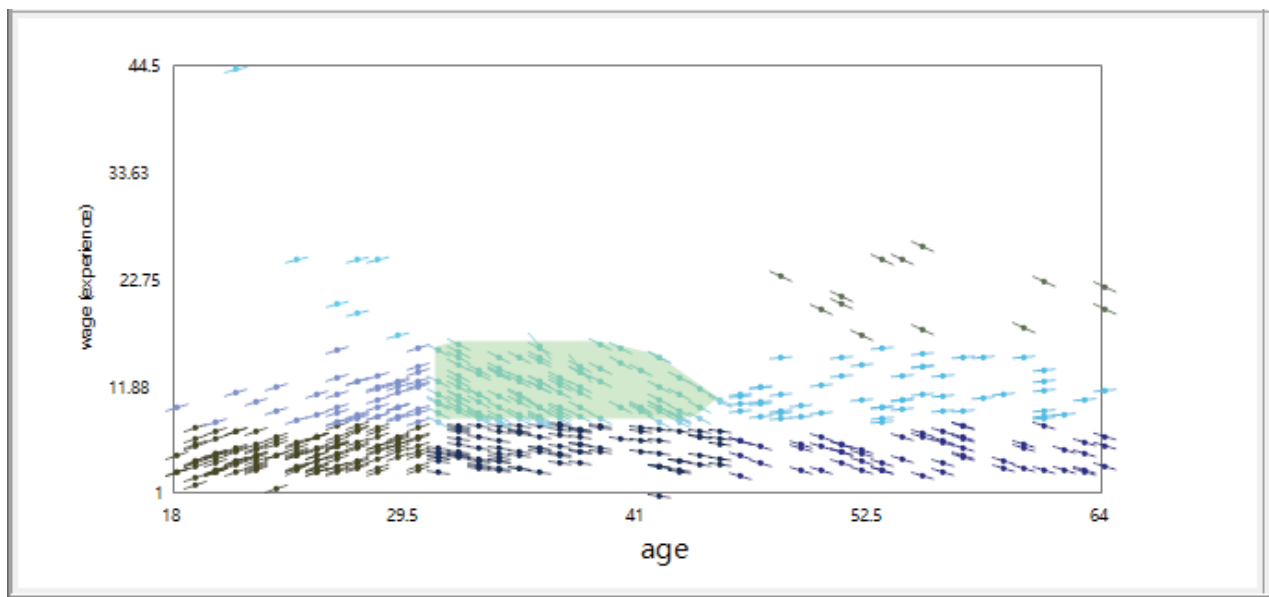


Figure 4-2: Interaction (mouse click) by using point-to-region concept: that is, a point click causes an entire convex hull (a class) to be highlighted.

4.2.2 Linear Approximation of Decision Trend

Decision attribute is the most distinct conception in RST compared to other methods. It explicitly asks users to choose a preferred attribute from a given dataset so the attributes are reduced according to it. It would be useful to the user if the data exploration task is designed with a more decision oriented approach. Since a scatterplot can only reveal a data correlation between two variables, we augment a parameter to approximate its relationship with a corresponding point in a third virtual dimension, that is, a decision attribute in this case. We acknowledge that a flow based scatterplot was previously discussed by Chan et al (2010) to study the sensitivity, but we further extended it to scatterplot matrix with interaction for class exploration by rough set model. A scatter point is positioned by its data value with a line which represents the derivative of function y specifically; the slope indicates the positive or negative correlation with respect to x or in global linear approximation and all the points reveal the same trend when there is one slope. They also computed local neighbourhood of radius to smooth the local trend around a given point. In our case, the equivalent class is already a set so we compute the local trend from the members in the class of a given point. Figures 4-3 (a) (b) provide a visual comparison between the classic and flow based scatterplot representations. Clearly, it is simple yet powerful visual augmentation that helps the user to study the decision trend as opposed to the classic metaphor which does not show the phenomena of interest as the decision trend.

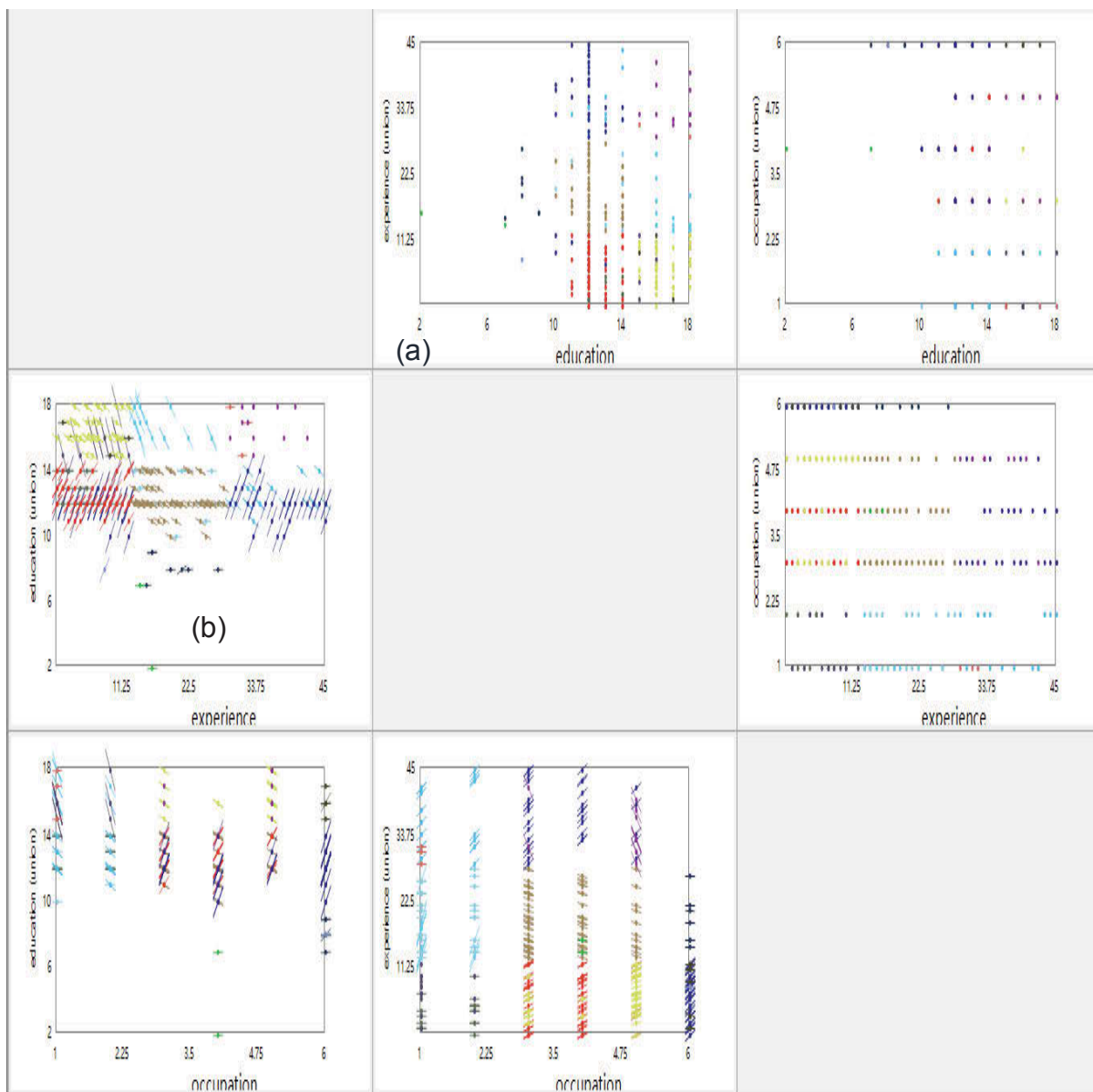


Figure 4-3 (a) A classic scatterplots visualisation. (b) Adding the decision flow where plots were augmented with respect to the decision variable.

To approximate the decision trend, we apply fewer squares in the linear regression model (Chatterjee & Hadi 2009) to the best fit line of a given point with respect to the decision attribute. In the linear regression model expressed by Equation 9 and Equation 10, there are two important coefficients that need to be solved first, where the slope that measures the change with respect to X and b_0 is the intercept. They are defined as follows:

$$b_1 = \frac{N \sum_i^N (X_i - X_0)(Y_i - Y_0) - \sum_i^N (X_i - X_0) \sum_i^N (Y_i - Y_0)}{N \sum_i^N (x_i - x_0)^2 - (\sum_i^N (X_i - X_0))^2}$$

Equation 9: The calculation of the slope to measure the change

$$b_0 = \frac{\sum_i^N (Y_i - X_0) - b_1 \sum_i^N (X_i - X_0)}{N}$$

Equation 10: The calculation of the intercept

where $x_i \in E(P)$ and $x_0 \in E(P)$, substituting b_0 and b_1 into the linear Equation 11 to interpolate the best fitting line at point (X_0, Y_0) :

$$Y_i(X_0 \pm k) = Y_0 + b_1(X_0 \pm k) + b_0$$

Equation 11: Interpolate the best fitting line at point (X_0, Y_0)

where k is the desired length. Please note that we have added the value of Y_0 because Y_i is a local linear approximation from a given point (X_0, Y_0) .

In the interactive design for decision trend analysis, we enable the user to switch the view between (X_0, Z_0) and (Y_0, Z_0) by simply clicking on the coordinate label.

4.2.3 Augmenting Class Coverage

We mentioned earlier that the data covered by a convex hull belongs to an equivalent class. It essentially represents a rule expressed as $E(P) \rightarrow D_i$ that has been learned from approximating a set with respect to a decision class using Equations 5, 6 and 7.

For example, the rule $E(P) = \{weight_{high}, accel_{low}\} \rightarrow 80\%cylinder_{high}$ means that there is eighty percent confidence that cars having more *cylinders* should have higher *weight* and lower *acceleration*. In fact, approximation regions are rule templates, a certain rule would classify the equivalent classes into positive regions, while uncertain or negative rules would classify the classes or negative regions. We are only interested in the rules that explain the phenomenon of interest. The two key elements associated with a rule are accuracy and coverage (Tsumoto 2002). Given a rule, its accuracy is defined as:

$$accuracy(E(P) \rightarrow D_i) = \frac{|E(P) \cap D_i|}{|E(P)|}$$

Equation 12: Calculate the accuracy of a rule

where $E(P)$ and D_i denote the condition and decision class respectively. The accuracy measures the strength of a rule with respect to D_i . A weak rule has accuracy of less than β and is too weak to be meaningful. Similarly, the coverage of a rule can be measured by:

$$\text{coverage}(E(P) \rightarrow D_i) = \frac{|E(P) \cap D_i|}{|D_i|}$$

Equation 13: calculate the coverage of a rule

The coverage measures the generality of a rule pointing to a certain class in D . In general, a rule with higher accuracy does not necessarily imply a lower coverage rule (Yao & Zhao 2008) and vice versa.

In the visualisation, we map the coverage to a hot-cold map with colours ranging from red to blue. For example, the background colour for the area covered by the convex hull will be close to red for higher coverage.

4.3 Evaluation

Interactive scatterplot matrix visualisation explores data using a rough set rules and decision trend analysis approach which establish target interests for users to reach their requirement explicitly.

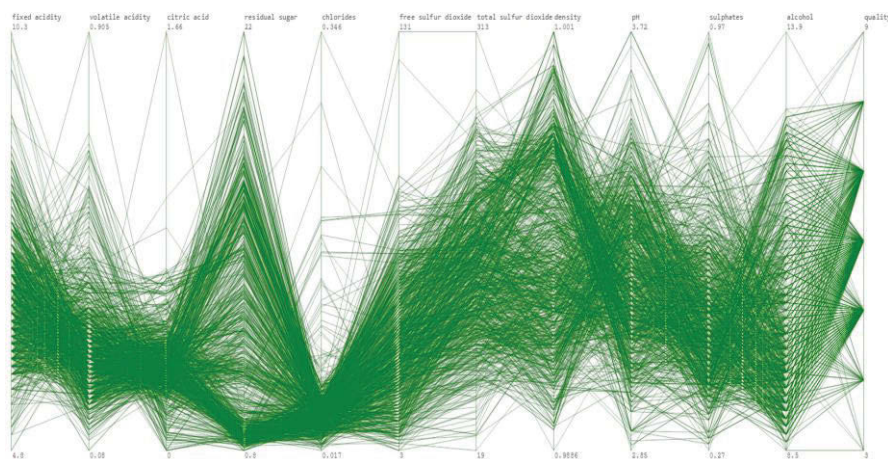
For further investigation of how well this interactive exploration technique works in the scenario based tasks during the visual analysis process, we apply our concept of decision trend interactive visual analysis to real cases (Section 4.3.1); and we also conducted a user study to compare the rough set theory working on scatterplot matrix and parallel coordinates.

4.3.1 Case Studies

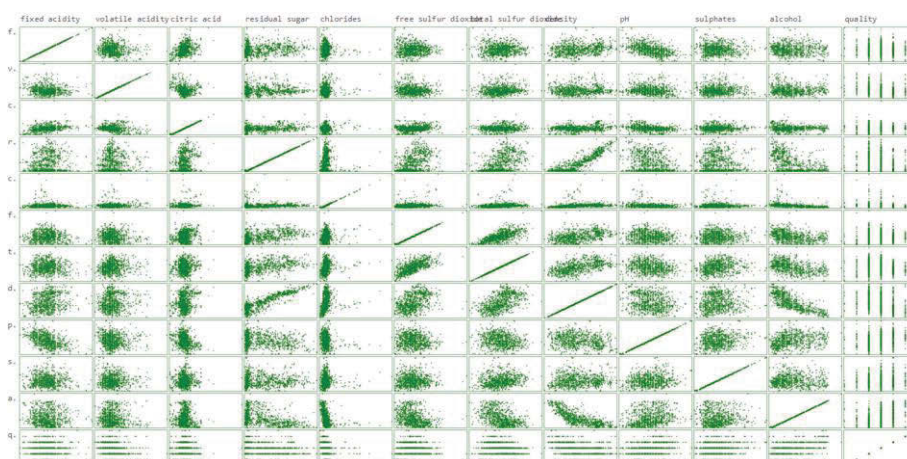
We applied our technique to three popular datasets to demonstrate its effectiveness. The case studies are presented as below.

4.3.1.1 Wine Data

We used the wine dataset obtained from Machine Learning Depository which consisted of 12 attributes with 4898 samples for modelling wine quality based on physicochemical tests. The attributes cover the sufficient information to describe the characteristics of a wine such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality (see Figures 4-4 (a) (b) for the visualisation of the entire dataset using standard parallel coordinates and the scatterplot matrix).



(a)



(b)

Figure 4-4 Visualisation of the entire wine dataset using a) parallel coordinate and b) scatter plot matrix. Dataset from: <https://archive.ics.uci.edu/ml/datasets/Wine+Quality>

Although the visualisations in Figures 4-4 (a) (b) provide contextual information about the entire dataset, the inclusion of many dimensions makes the visualisations less readable due to the density at the parallel coordinates and the size reduction of the scatterplot matrix.

In our visualisation of the dataset after using the rough set theory, the wine quality is the decision attribute (or dimension) and the rest become conditional attributes. The attributes are partitioned into three groups (or clusters) using K-means. There were five ranked feature sets obtained from VPRS procedures and each contains two conditional attributes and one decision where we selected four of them as shown in Figures 4-5 (a) (b). The points with the same colour indicate that they belong to the same class. Some outlier classes are annotated with an arrow. In the visual data exploration of the scatterplot matrix, it displays that both fixed acidity and volatile acidity have negative impact on the wine quality revealed in the trends in Figure 4-5 (a) and Figure 4-5 (b). We further identified an outlier class which has the worst impact on the wine quality

in Figure 4-5 (a). It is also interesting to note that the lower the level of free sulphur dioxide and residual sugar tend to have positive impact on the quality as displayed in Figure 4-5 (c) and Figure 4-5 (d).

Figure 4-6 illustrates another example of the wine dataset where 1) the number of clusters was set to two, because without clustering, the rough set will classify too many weak rules due to continuous variables, 2) fixed acidity was chosen as the rough set decision attribute, and 3) the acceptable classification error rate of quality was set to 80%, which interpreted that it is allowable to have up to 20% incorrectness in the final clustering. After carrying out the RST process, there were 6 feature sets generated for classifying data into classes, including: a) {fixed acidity, alcohol, quality}, b) {fixed acidity, volatile acidity, residual sugar, chlorides, alcohol, quality}, c) {fixed acidity, residual sugar, alcohol, quality}, d) {residual sugar, PH, quality}, e) {residual sugar, alcohol, quality}, f) {fixed acidity, volatile acidity, alcohol, quality}. The feature set (b) was used in our experiment after several trials in comparing the data quantity and visual quality of the classification results.

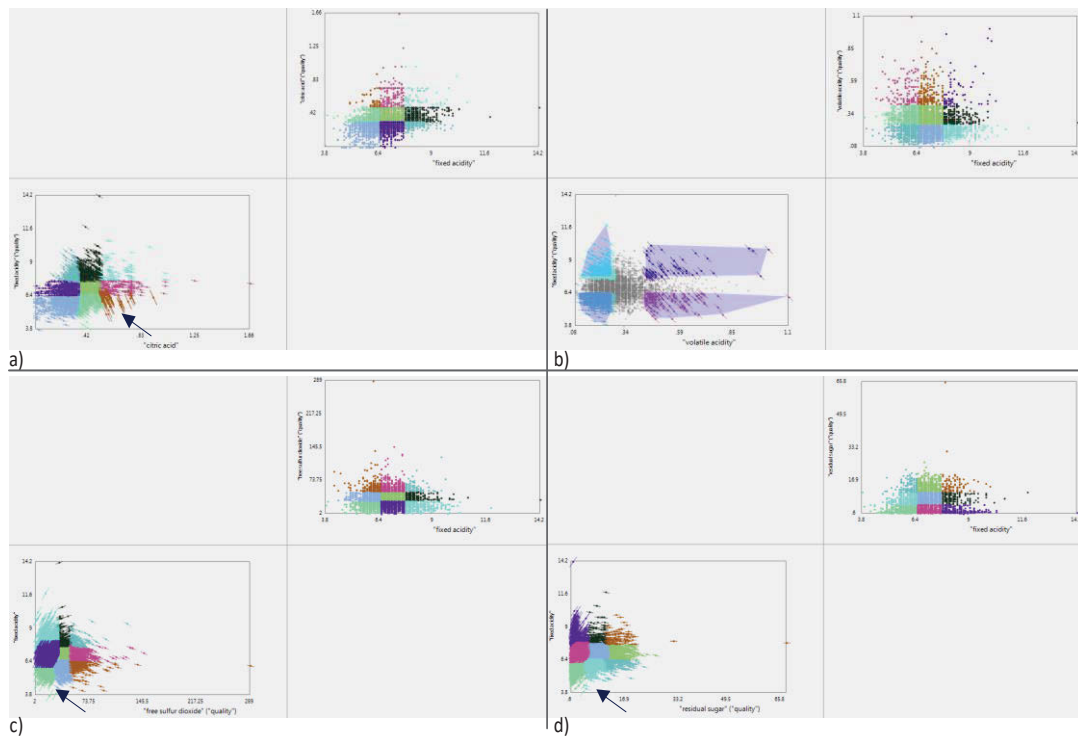


Figure 4-5: Results obtained from the case study with wine data. The upper diagonal matrix displays the classic scatterplots and the lower diagonal matrix has been augmented by the decision trend a) { citric acid, fixed acidity, quality} with (Y_0, Z_0) . b) {volatile acid, fixed acidity, quality} with (Y_0, Z_0) . c) { free sulfur dioxide, fixed acidity, quality} with (X_0, Z_0) . d) { residual sugar, fixed acidity, quality} with (X_0, Z_0) .



Figure 4-6: Case study with wine dataset obtained from [31]. Boxes at “without-trend” area (area above the diagonal line) are scatterplots of each pair of attributes while boxes at “with-trend” area (area below the diagonal line) represent the same values and with changing trends.

As seen at Figure 4-6, X -axis and Y -axis give the same wine properties including {fixed acidity, volatile acidity, residual sugar, chlorides, alcohol, quality}. Boxes at “with-trend” area (area above the diagonal line) are scatterplots of each pair of attributes while boxes at “without-trend” area (area below the diagonal line) represent both the points' value and their changing trends. For example, it is easy to discover from boxes (a) and (b) that the wine data have been divided into 6 clusters. In the box (a), the wine with higher

chlorides and lower *fixed acidity* has positive influences on the *quality*, while the highest *chlorides* have invisible influence on wine. The visualisation at boxes (a) (c) and (d) also indicated that, when the *chloride* is the same, *fixed acidity and volatile acidity and residual sugar* might have different impacts on the wine clusters. *Particularly, both volatile acidity and residual sugar* impacts the wine *quality* negatively, while *fixed acidity* impacts the wine *quality* positively.

4.3.1.2 Car data

In this case study, we used a well-known car dataset obtained from <http://lib.stat.cmu.edu/datasets/cars.data>. The dataset contains 8 attributes with 392 samples after the removal of the missing attribute data. The dataset describes the car information about its origin, model, acceleration, weight, horsepower, cylinder, mileage per gallon (mpg) and displacement. The dimensionality reduction result is described in Figure 4-7. This case study is used only for illustrative purposes to show that a feature set with more dimensions has been captured from the feature selection procedures.

Through the demonstration of the case study, we have shown the ease of use provided by the system for Multi-dimensional data exploration, visual analysis and decision making.

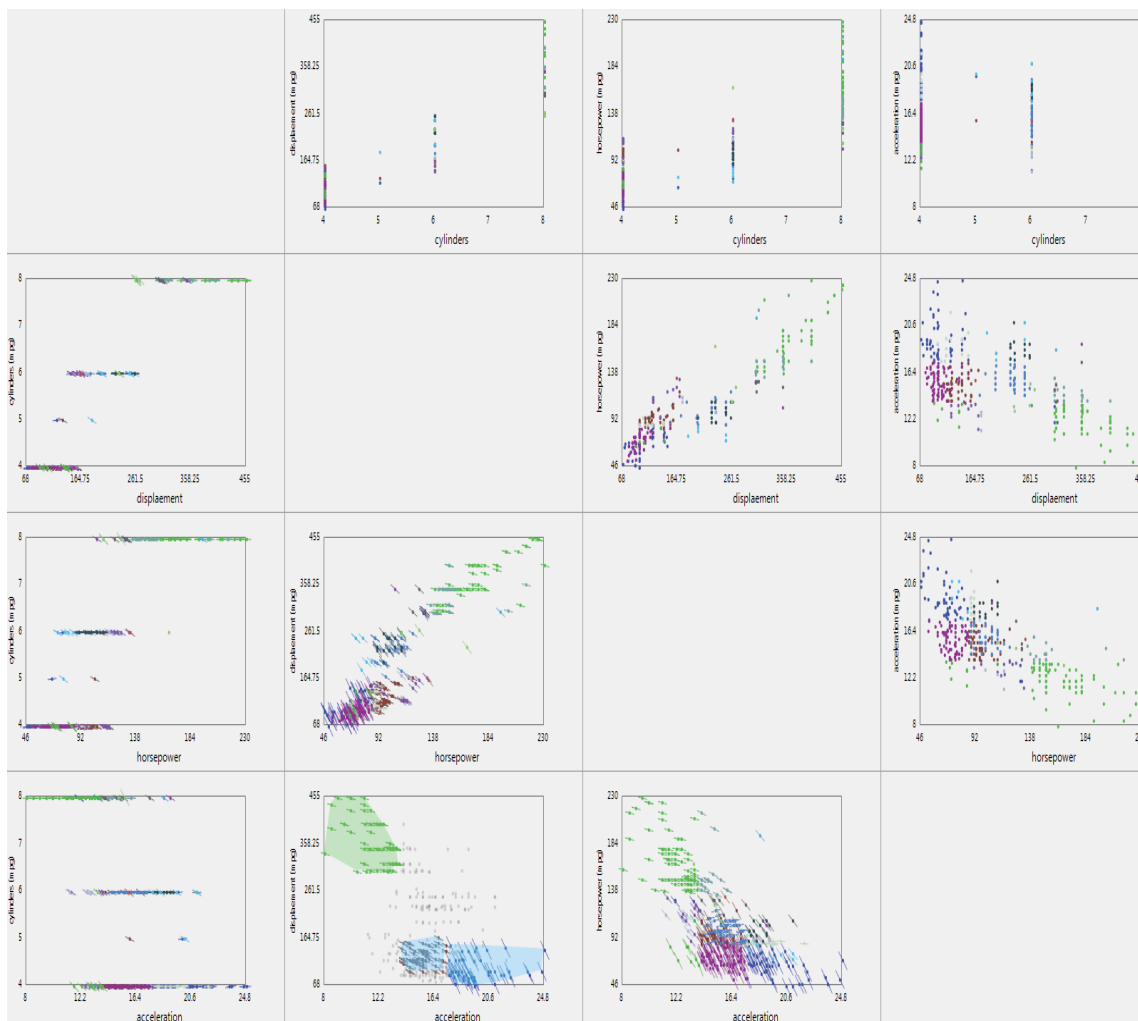


Figure 4-7: Case study with car dataset. We selected mileage per gallon (MPG) as the decision and the dataset has been reduced to 4 attributes namely acceleration, displacement, cylinders and horsepower.

4.3.1.3 Wage data

In the third case study, we choose a wage dataset as a test sample. The wage dataset collected from <http://www.nber.org/cps/> contains 534 observations on 11 variables sampled from the Current Population Survey of 1985. This dataset includes attributes including education, south, sex, experience, union, wage, age, race, occupation, sector and Marr (Marital Status).

In this case study, experience, wage and age are the features in one rule using education as the decision attribute. The visualisation at Figure 4-8 shows seven categories based on the work experience.

More specifically, there are fewer people with higher wages, and they are either at a younger age (25-29) or an older age (52-60). Although fewer young people earn high wages, they have a positive influence on the classification with respect to experience, on the contrast; people who are older with a higher wage tend to impact the classification negatively.

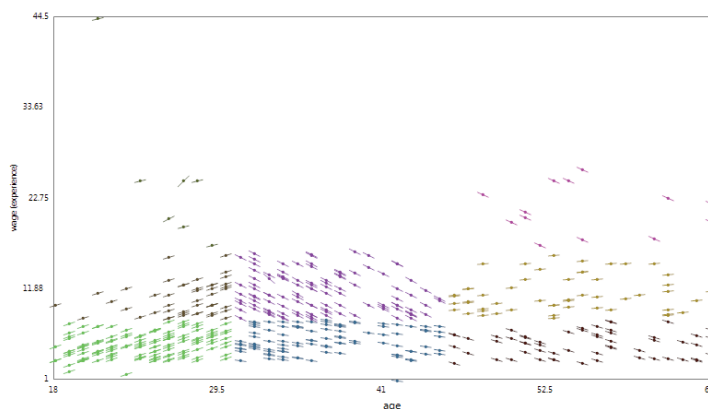


Figure 4-8: Case study with Wages dataset. This figure shows a box at “with-trend” area for correlation between *experiences*, *wage* (y-axis) and *age* (x-axis).

4.3.2 Usability Study

We conducted a pilot usability study with students from various backgrounds. The goal was to evaluate whether a scatterplot matrix is more effective than a Parallel Coordinate when using RST results, in terms of accuracy and user preference.

4.3.2.1 Methods

1) Participants: recruited to the usability study were 16 participants (5 female, 11 male), with ages ranging from 25 – 40 who were students from different backgrounds, including information technologies, sciences and business. Most of the participants indicated that they had never used Parallel Coordinate and Scatterplot Matrix before. None of them knew about Rough Set Theory. All participants were fluent English speakers and accustomed to the methodology of understanding.

2) Experimental design and tasks: two similar datasets were used in the study. Each participant performed two experiments on the two datasets using the two visualisation techniques, the Parallel Coordinate and the Scatterplot matrix. The experiments were run on a 24 inches full HD screen. All the tasks in each trial took approximately 20 minutes to complete. For dataset 1, we set clusters to 2, and applied RST in the condition of using mpg as the decision attribute; the classification error rate was set to 0.8 on horsepower. For dataset 2, we set clusters to 3, and applied RST with the condition of using education as the decision attribute; the classification error rate was set to 0.8 on experience. We

only collected the accuracy result in our study; each correct answer was marked as 1 and an incorrect answer was marked as 0.

Five questions were designed for the evaluation as follows, and listed on a multiple choice questionnaire.

Q1. After using Rough Set Theory, how many attributes do we have on the visualisation?

Q2. Which one is the decision attribute?

Q3. Is one attribute more influential than another attribute?

Q4. Are selected attributes more influential than the decision attribute?

Q5. Find the number of clusters in the visualisation.

At the end, the participants were requested to rank each visualisation technique on a 5-point Likert scale, from 1 (strongly disagree) to 5 (strongly agree).

4.3.2.2 Results

Our results are demonstrated by accuracy and participants' feedback.

1) Accuracy

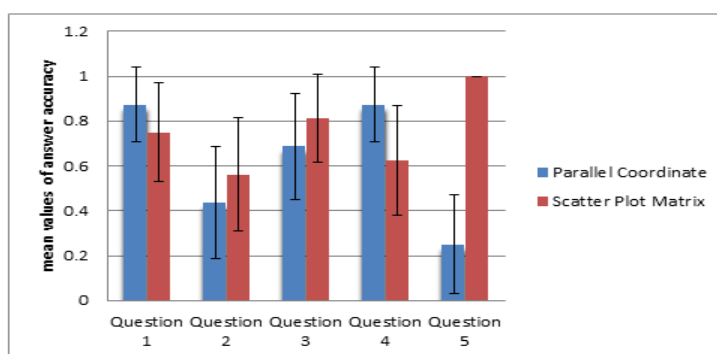


Figure 4-9: Accuracy of Parallel Coordinate and Scatterplot matrix Visualisations corresponding to five questions (with 95% confidence intervals).

Figure 4-9 shows the accuracy for the two visualisations corresponding to the five questions.

The early analysis of the results indicate that the average accuracy was significantly better in question 2 and 3 for Scatterplot matrix than for Parallel Coordinate: Question 2 (M = 0.56, SD = 0.26 versus M = 0.44, SD = 0.26), Question 3 (M = 0.81, SD = 0.16 versus M = 0.69, SD = 0.23). The average accuracy was significantly lower in question 1 and 4 for Scatterplot matrix than for Parallel Coordinate: Question 1 (M = 0.75, SD = 0.20 versus M = 0.88, SD = 0.12), Question 4 (M = 0.63, SD = 0.25 versus M = 0.88, SD = 0.12). Noticeably, all the participants could identify the number clusters in the Scatterplot matrix while they have difficulty identifying them in the Parallel Coordinate (Question 5: M = 1.00, SD = 0.00 versus M = 0.25, SD = 0.20).

2) Participants' Feedback

Participants gave subjective feedback about the Parallel Coordinate and Scatterplot Visualisations: on the five-point Likert scale, the participants evaluated their overall preference using the two techniques on the RST datasets. The result indicated that that the participants prefer the Scatterplot matrix (M = 4.19, SD = 0.16) over the Parallel Coordinate (M = 2.13, SD = 0.65).

4.4 Summary

Visual analysis is an important subject in Multi-dimensional visualisation, but it is often discussed in a standalone manner with many areas left unexplored. Thus, it is often considered as a viewing step in Multi-dimensional dataset. Dealing with high dimensional dataset is always challenging and we believe that the most effective way is through iterative visualisation and interaction on the data subset. This is because the iterative interaction process will involve human's eye-brain system in the data analytics and the eye-brain system is considered to be the most efficient system for data analysis.

We contributed the novel scatterplot matrix visualisation for Multi-dimensional data and decision trend analysis. Our solution is a more comprehensive approach with a novel interaction model that is tightly integrated with the dimensionality reduction based on RST. We highlight the decision rule based concept offered by RST because it explicitly requires the user to establish a target interest in the visual analytic task. We illustrated the visualisations in three case studies with three popular datasets including wine quality, cars and wages. Our pilot usability study indicates the higher accuracy of Scatterplot matrix visualisation over Parallel Coordinate visualisation in determining the decision attribute, in which attributes have more influence, and recognizing the clusters. The participants also preferred Scatterplot matrix to the Parallel Coordinate in the analysis task.

Chapter 5 MDV in Forensic Visualisation

THIS CHAPTER AIMS to discuss how visualisation techniques improve the efficiency of investigating computer related crimes' in the domain of computer forensics. We mainly discuss this from three perspectives: the investigation model; hard disk drive investigation; and criminal detection.

Up to now, a great deal of time and money has been wasted by investigators trying to explain complicated large volumes of data that is uncorrelated or meaningless without high levels of patience and tolerance for mistakes. Many approaches have been proposed to help the investigators during the process of exploring evidence, such as data mining, etc. In our thesis, we discuss the application of the multi-dimensional visualisation technique, which is an approach using the computer-supported, interactive, visual representations of data to amplify cognition (Card, Mackinlay & Shneiderman 1999), and a way to discover decision making and explanation which enable human's ability to visually interpret and comprehend a textual description, into forensic research.

We discovered that multi-dimensional visualisation techniques greatly aid researchers in guiding their searching for target files, in effect supporting the explanation process. In detail, Section 5.1 gives a short description about forensic investigation, then in Section 5.2, we demonstrate the pipelines of visualisation techniques working in forensic investigation. Section 5.3 delivers examples of visualizing computer hard disk drives. Section 5.4 develops the method of detecting criminal relationships with the Self organizing Map. Section

5.5 gives a short introduction to tree visualisation techniques in forensics.

5.1 What is Forensic investigation?

As the digital device has become a public and necessary tool in human's daily life, it is more often taken advantage of committing illegal activities. The forensics domain is proposed to examine the digital media in a forensically sound manner with the aim of identifying, preserving, recovering, analysing and presenting facts and opinions about the information (Teerlink & Erbacher 2006). In detail, when problems happened, forensic investigators firstly access the original data from these storages, then analyse the original data to dip into the deeper information, either constructing the scene or predicting the behaviour of the criminals, to simplify the complicated situation. However, at the end, they need to provide documentary evidence, which is the objective, unbiased truth of the matter. The evidence is prepared to present in court in adversarial and sometimes very probing proceedings, to support the results of a computer forensic examination.

However, with the development of digital storage resources, such as telephones, mobile devices, laptops, desktops, routers, firewalls, and also compact disks, floppy disks, magnetic tapes, high capacity flip, zip, and jazz disks, memory sticks, USB storage devices and so forth, there are more challenges for techniques which aid analysts (forensic investigators) in collecting and gaining digital results based on these hardware facilities.

The following Section 5.1.1 will introduce more about digital evidence and then in Section 5.1.2, a description of standards and protocols will be provided.

5.1.1 Digital Evidence

Digital evidence (Noblett, Pollitt & Presley 2000) is any probative information stored or transmitted in digital form that a party to a court case may use at trial. It reveals how a crime was committed, and provides investigative leads. It is also used as a means of disproving or supporting witnesses. In addition, different crimes result in different types of digital evidence. For example, cyber stalkers often use e-mail to harass their victims, computer crackers sometimes inadvertently leave evidence of their activities in log files, and child pornographers sometimes have digitised images stored on their computers.

Nowadays, digital data can be stored in various forms. As for a large volume of dataset, which is always vulnerable and sensitive, the process of analysis would be time-consuming. It is difficult to keep digital evidence readable and accurate due to its large volume. The frequency of computer fraud and other digital crimes are growing day by day. Unfortunately, less than two percent of the reported cases result in convictions. Therefore, it also needs ongoing efforts to develop examination standards and to provide structures for computer forensic examinations.

5.1.2 Standards and Protocols

Computer forensics will be presented in court in adversarial and sometimes very probing proceedings. To support the results of a computer forensic examination, procedures are needed to ensure that the true information exists on the computer storage media, unaltered by the examination process.

Basically, there are general forensic and procedural principles (Yasinsac et al. 2003) (Garber 2001) to be applied during the investigation process:

- (1) Actions taken to secure and collect digital evidence should not affect the integrity of that evidence.
- (2) Persons conducting an examination of digital evidence should be trained for that purpose.
- (3) Activity relating to the seizure, examination, storage, or transfer of digital evidence should be documented, preserved, and available for review.

Through all of this, the examiner should be cognisant of the need to conduct an accurate and impartial examination of the digital evidence.

5.2 MDV Improves Forensic Investigation Models

Models aim to establish clear guidelines for dealing with a complex problem. A suitable forensic investigation pipeline could provide clear guidance for specialists. In Section 5.2.1 we describe the fundamental pipelines of forensic investigation; followed by the investigation processes with visualisation in Section 5.2.2. Then in Section 5.2.3, a case study on Hard Disk Drive forensic analysis is presented. Finally, we have a short summary on the improvement of working processes for forensic analysis.

5.2.1 Forensic Investigation Pipeline

Generally, forensic investigation is composed by preparation, forensic analysis, forensic report and admissibility. Many models are the extensions of these four steps. For example, Reith et al (2004) extended the model by evidence identification, preparation, approach strategy, preservation, collection, examination, analysis, presentation, and returning evidence. Later, Carrier et al (2005) carried another model to make an interaction between physical and digital investigation, allowing the model to be implemented in real digital crime scenes. In the same year, N.L.Beebe et al (2005) introduced a multi-tier, hierarchical framework, which is beneficial for logical analysis of investigation and have more applications for concerns about user community applicability; Followed by this, Ricci S.C.Leong et al (2010) discovered a hierarchical and objectives-based model to focus on the availability of real investigation scenes. These models provided an abstract reference framework for forensic analysts. They might also help develop and apply methodologies to new technologies in forensics domain.

Although the models are extended by various ways and they all played important roles in forensic investigations during the past years, fewer researchers have considered enhancing the analysing process by visualisation techniques. So one contribution in our thesis is to introduce a visualisation technique based forensic investigation model.

5.2.2 Visualisation based Forensics' Model

This visualisation based forensic investigation model not only displays how visualisation techniques work well in the whole investigation process; it also considers the protection and accuracy of digital evidence by applying security certification before and after the analysis process, as is shown in Figure 5-1.

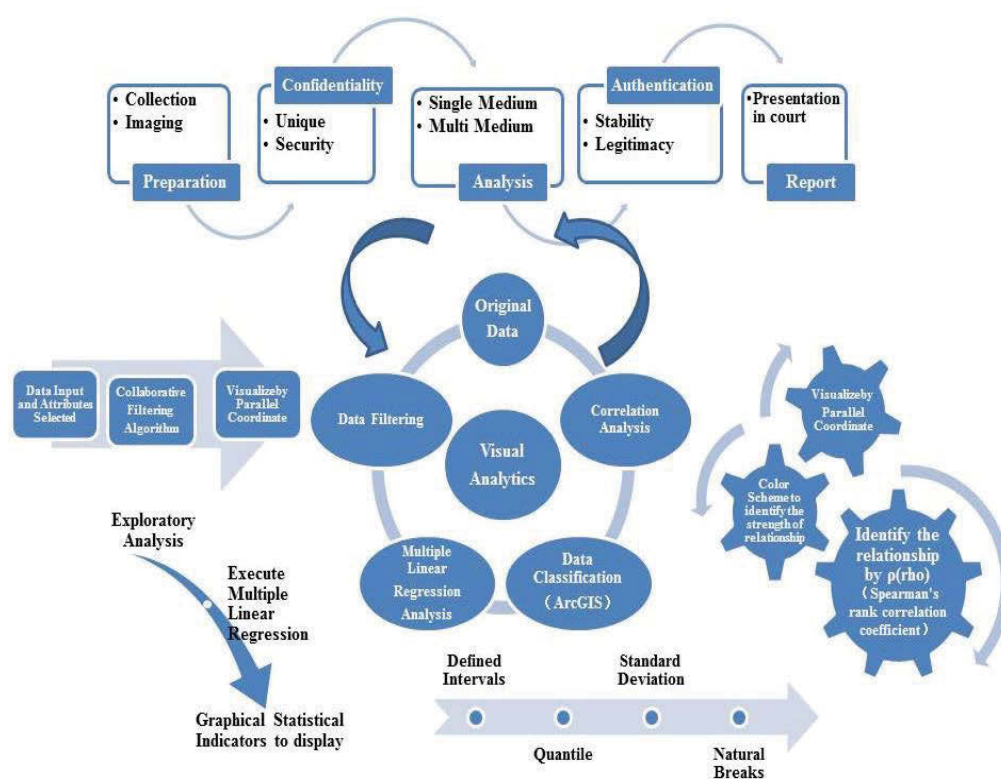


Figure 5-1: A visualisation technique based Forensic investigation Model

According to the model proposed by Baryamureeba et al (2004). We divide the investigating processes into five parts including: *evidence preparation*, *data protection*, *analysis*, *data certification*, and *report*. We will explain them accordingly as below.

Firstly, as the other models mentioned, evidence preparation is required. In this step, it needs data collection, imaging data and duplication of digital storage media. For example, researchers use a series of text-based commands to duplicate files.

The second stage is mainly for protecting the integrity of evidence. Analysts start the work with a sure knowledge of the data security. For example, they can use 'fingerprints' to keep the security of the original information.

The third stage is always considered to be the most tedious and complicated part as the investigators analyse the data and discover available results. Most researchers choose data mining approaches and apply them repetitively until they find the desirable conclusions. However, in our new model, we combine data visualisation approaches (Han, Kamber & Pei 2011) with data mining methodologies to deal with the original data instead of text searching or image searching. Specifically, this combined part can be split into 5 parts: *visualizing original data, correlation analysis, Data classification, multiple linear regression, and data filtering*.

In the first part, after we collected the metadata which are normally in numbers or texts, we visualise them by Parallel Coordinates so that the users can get an overview of the whole dataset. Through the visualisation of correlation analysis, the view is easier to discover the data relationships, and a colour scheme can also clearly reflect the strength of the relationship among data. We also used cluster methodologies to represent the data classifications of the parallel coordinate's plots.

Then we came to our fourth step: information authentication. Because of the particularity of digital evidence, mentioned in Section 5.1.1, the investigators keep the stability and legitimacy of investigating results. Followed by this step, the forensic analyst will prepare a report that will be formatted to provide an easy-to-read document including all evidence recovered throughout the investigation and analysis.

In summary, this model provides a view for understanding the process of investigations, and considers security problems of original data, processing data and also data results due to being placed in the law enforcement environment. Our model also concentrates on an important process, the data analysis. By using information visualisation techniques, the analysts can save time discovering the data relationships among other variables, and this could improve the investigating efficiency to a certain extent.

5.2.3 A Case Study

Most digital forensic tools display files in graphs. They use the file as a single object. To be more specific, a grid represents a file if you use Tree-Map, and each face is a file if you choose Chernoff Faces. In this situation, if users click on the required files, they can check the details of files, such as: the name of the file, the extension of the file, the date it was created, the date it was modified, and the logical size of the file, etc. In our experiments, we applied our visualisation based forensic investigation model into a Hard disk drive investigation. Specifically, we adjusted our model into a new format according to the scene, see Figure 5-2. By this visualisation technique, the understandability and clarity of investigation results are improved.

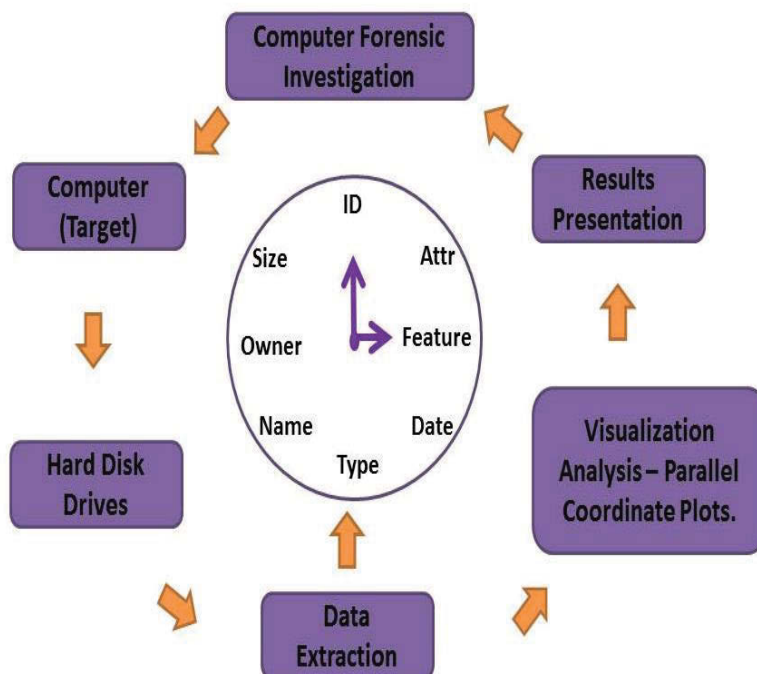


Figure 5-2 New Model for HDD Visualisation

This new model includes three parts: File Extraction, File Analysis and File Visualisation. We use Disk Investigator to extract all files in HDD, and use parallel coordinates to display all the original files then to specify the suspicious files based on the cluster function of Parallel Coordinate. Disk Investigator can display Time attributes, File Name, File Size and File Path. In our new model, we mainly use the following six attributes: File ID, File Size, File Attribute, File Path, Item Type, and Owner. All of them are stable attributes compared to the dynamic attribute-Time. Therefore, information analysed will be more accurate. Table 5-1 and Table 5-2 give more details about these six attributes.

Table 3-1: Name and Type

Name	×10
File ID	{1,2,3,4, 5,.....,N}
File Size	{1,2,3,4, 5,.....,N}
File Attr	{1,2,3,4, 5,.....,N}
Item Type	{1,2,3,4, 5,.....,N}
Owner	{1,2,3,4, 5,.....,N}
File Feature	{1,2,3,4, 5,.....,N}

Table 5-2: Name and Remark

Name	Remark
FileID	Each file has a file id to connect file attributes. For example, if the file's number is N, then the last file's ID is N.
File Size	File is divided into 7 categories according to size, they are: Unspecified (file folder), Gigantic (>128MB), Huge (16-128MB), Large (1-16MB), Medium (100KB-1MB), Small

	(10-100KB), Tiny (0-10KB)
File Attr	<p>File attributes are settings associated with computer files that grant or deny certain rights to how a user or the operating system can access that file. In Microsoft Windows, they are marked as: Null, H, DHS, HAS, R,RA, RHA,A, RSA,RH,SA,D,RD,HD,RSD,DAE, SA, Other (DI, RHDI, HDI, etc.).</p> <p>Read – Only allows a file to be read, but nothing can be written to the file.</p> <p>Archive – Tells Windows Backup to back up file.</p> <p>System – System file.</p> <p>Hidden – File will not be shown when doing a regular dir. from DOS.</p>
Item Type	For example: Adobe Acrobat, MATLAB, JPEG image, Microsoft Word, PNG image, shortcut, etc.
Owner	In our test, for example, the owner contains: SYSTEM, Administrators, FEIT\11485570)
File Feature	In our test, We use Disk Investigator to recognise File Feature, including: Trivial File, Broken File, Hided file, Deleted file and Encrypted File, Encrypted Directory, Trivial Directory.

5.2.4 Summary

The traditional digital forensics approach (Moore 2010) involves seizing a system(s)/media, transporting it to the lab, making a forensic image(s), and then searching the entire system for potential evidence. However, this is no longer appropriate in some circumstances. In cases such as child abductions or missing or exploited persons, time is of the essence. In these types of cases, investigators dealing with the suspect or crime scene need investigative leads quickly. In many cases, it has differences between life and death for the victim(s). So the need for the timely identification, analysis and interpretation of digital evidence is becoming more crucial.

Our new visualisation involved forensics' model improves the quality of investigating. (1) It attacks certain problems of data visualisation, especially when data are large in size, high in complexity, and also intractable. (2) It advances forensic developments in evidence analysis, human and computer interacting, and accurate representation in courts.

5.3 MDV Assists Visualisation Hard Disk Drives' (HDDs') Investigation

Computer forensics implies a connection between computer and crime detection. Computer-based evidence has obviously become a critical part of legal systems throughout the world. In this part, our work targets at one of the main digital devices - Hard Disk Drive. Different Multi-dimensional visualisation techniques will be utilised in the analysis of hard disk drives based forensic cases.

This section introduces how parallel geometry helps in visualizing high-dimensional metadata, especially in regard to the large capacity of HDD. Section 5.3.1 introduces hard disk drive and Section 5.3.2 describes how parallel coordinate visualisation techniques work on the analysis of hard disk drives.

5.3.1 Hard Disk Drive

Nowadays, digital data is generated exponentially due to the quick development of digital techniques. The statistic shows that 2.5 Quintillion Bytes of data are created each day. Beyer et al (2011) pointed that today's data is high volume, high velocity and/or high variety, and information assets that require new forms of processing to deal with the scalability, formatting and semantics problems, performance issue and the privacy issue as well.

But no matter how complicated the data are, they all need medium devices to store. Such as laptops, desktops, USB storage devices and so forth. Therefore, before analysing the data, it is essential to understand how and where storage devices preserve data.

Generally, a storage device is using for holding and processing data, and a storage device always contains three categories: Memory, Hard Disk Drive and Removable Media Drive, see Figure 5-3.

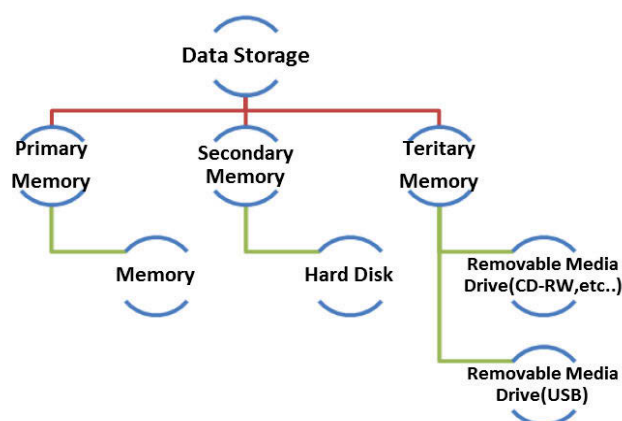


Figure 5-3: Data Storage Category

In this thesis, we concentrate on the hard disk drive related investigation. As one of the semi-portable method devices, Hard Disk Drive (HDD) is more often used for storing and retrieving digital information by rapidly rotating. A HDD can contain data even when powered off. Particularly, in today's informatics era, a hard disk drive has the ability to contain all the data generated all over the world in a day (a 2TB hard disk drive is common today). However, such large volume devices also bring a higher level of difficulties for forensic investigators to explore available digital evidence.

To analyse the data collected from a hard disk drive, the simplest way is to read the data through disk investigation tools. In our work, Disk Investigator (Casey & Stanley 2004) is the collection tool used to reach data in a hard disk drive. This tool can extract the disk data in three formats, Hex Text and Decimal. It also shows the details of files and also the disk information, such as: size, logical sectors, cluster size, free clusters, MFT size, MFT start cluster, MFT start zone cluster, MFT zone size, MFT mirror start, media descriptor, root Entries, heads, sectors per track, reserved sectors, volume label, etc.. In addition, disk investigator provides a function of special cluster viewer; the cluster details will be shown in dec, hex and text. You can also copy all the results. However, it is hard for the layperson to analyse the results as all the disk information will be shown by numbers and letters, such as Sofs: 0000 Hex: EB 58 90 4D 53 44 4F 53 Text: .X.H S D O S Decimal: 235 88 144 77 83 68 79 83; Cluster Selection F0fs: 000000000033 Hex, 51 Dec. C0fs: 0033 Hex, 51 Dec Selection Sector: 16356000.

More often, the data will be shown in a hierarchical structure with file attributes.

For example, in a Windows7 System, there are more than 200 attributes in each file. these include: *File Name, File Type, Total Editing Time (Only for .doc or .docx), Computer, Date Created, Date Accessed, Date Modified, Content Created Time, Date Last saved Time, File attributes, Path, Size, Data rate, Email, Description, Due Date, Date Sent, Date Taken, Date Visited, etc.* However, as the value of the attributes might be wrong and some attributes have no value, as they are displayed “null”. In a forensic investigation, not all of these attributes are applicable to the analysis process.

Take Time Attribute as an example; it is usually considered as an important element in different data analysis, not to mention in the forensic investigations which have a time limitation for each case and need time to navigate and track crime on most occasions. During the past years, some computer forensic researchers utilised Time as the intermediate source in their investigations, such as, they calculated the file’s access frequency by time to detect the suspicious targets. In addition, most of the forensic tools have been developed with the function of a representation with timelines.

However, some drawbacks still exists in the utilisation of time attributes in forensic investigation. The first is about the accuracy of data value displayed by systems, including the Modified Date, Data Accessed, Date Last Saved and Date Created, see Table 5-3.

Table 5-3: Time Attributes

Name	Date Modified	Date Accessed
File1	13/03/2013 3:35PM	21/03/2013 2:40PM
File2	13/03/2013 2:25PM	21/03/2013 4:40PM

File3	13/03/2013 3:34PM	21/03/2013 2:40PM
Name	Date Last Saved	Date Created
File4	2/05/2012 7:57 PM	4/02/2013 3:19 PM
File5	13/03/2013 3:34PM	21/03/2013 2:40PM
Name	Content Created	Date Created
File7	19/11/2004 2:20PM	20/03/2013 3:36PM
File8	20/03/2013 3:54PM	21/03/2013 2:40PM
File9	13/03/2013 3:36PM	21/03/2013 2:40PM
Name	Date Last Saved	Date Modified
File10	25/03/2007 9:05PM	21/03/2013 8:47AM

The second is the deviation value of a data attribute, which means that the data might be stored in a different Time Zone or with a different system time-setting. So if the analysts require time-related data, more situation should be considered. Such as: where the computer is, whether computer setting has been changed, etc. We use a small dataset (Windows 7) to illustrate the Time Problem; reading three the files in Table 5-3, we can get four conclusions: firstly, files' Modified Date might be earlier than its Accessed Date; secondly, the files' Last Saved Date might be earlier than its Created Date; thirdly, the files' Content Date might be different to the Date created; finally, the files' Last Saved Date might differ from its Modified Date. To conclude, time attribute is an important element, but it has to be read carefully as it may cause a calculation error and results with deviation.

5.3.2 Parallel Coordinates' on HDDs

The aim of applying parallel coordinate plots to a hard disk drive is to transform the complex and incomprehensible text or number into a graph. It might be easily understandable by all people, irrespective of their degree of training. In our experiment, the testing computer is configured as follows: Processor is Intel(R) Core i5-2400 CPU @ 3.10GHz; Installed memory is 4G; System type is 64-bit Operating System, Local Disk 465G.

We choose Disk Investigator (mentioned in Section 5.3.1), a public disk extraction tool, to extract data. Through Disk Investigator, we got file details with parts of their attributes: file_name, dos_name, extension, attribute, size, modified time, created time and last accessed time. The files are displayed with four different colours: black represents normal files, green represents directories and files with size 0, rose red represents deleted files, and deep red represents deleted directories and the directories with size 0, as shown in Table 5-4.

Although disk investigator would provide File size and Attr, it is difficult to record data numerically for the implementation of parallel coordinate plots. Therefore, we transferred the File Feature in the format of {0.1,0.2,0.3,0.4...}. Followed by the data transformation, we use system default file attributes to get other datasets elements, including Owner, File Feature, File Attr and Size, and they are stored as {0.1,0.2,0.3,0.4...}.

Table 5-4 Disk Investigator

Name	~\$e use of Pa	~\$per-p repare referen ce.docx	Paralle l Coordi nate	0_Osakns_1- 150
Dos Name	?\$EUS E~1	~\$PER. ~1	PARAL L ~1	?_PSAL~1
Ext	Doc	Doc		
Attr	a...h	a...h	.d...	.d...
Size	162	162	0	16,384
Modified	2013-3- -27 11:00:0 0	2013-3- 15 11:00:0 0	2012-1 2-11 11:00:0 0	2013-3-13 11:00:00

Figure 5-4 and Figure 5-5 explain the further information obtained through applying parallel coordinates. By these two figures, our first conclusion is that the user has a clear overview of all files. Secondly, each file attribute value is shown as a point, and one line represents one file. When the user clicks on any points, they can identify other attributes which belong to the same file or the files which have the same attribute. Thirdly, from this graph, you can easily recognise files in a same classification. For example, if there are 20 points together, this represents that these 20 files have the same attribute, and can be recognised as a group. Specifically, when you click on a clustering point, you can get the number of values which have the same attributes and their files' attribute will be displayed at the same time. Fourthly, this graph easily allows the user to find the anonymous files. Take Figure 5-4 as an example, there are two files clustered at one point, click on this point, it will show the value of this point is 0, which represents that the size of these two files is 0. Furthermore, the points have been connected by a line, and by checking each line, the user can get information about the file and then to find the suspicious ones. As shown in

Figure 5-5, one file is a deleted file with encrypted attributes, file ID is 18 and item type is Matlab, another is a hidden file with encrypted attributes, file ID is 20 and file extension is .doc. Through analysing the value of the attribute, File 20 and 18 are common, temporary and secure files. So they can be excluded from the anonymous files queue. Forensic investigators will also be able to prioritise the analysis of other data.

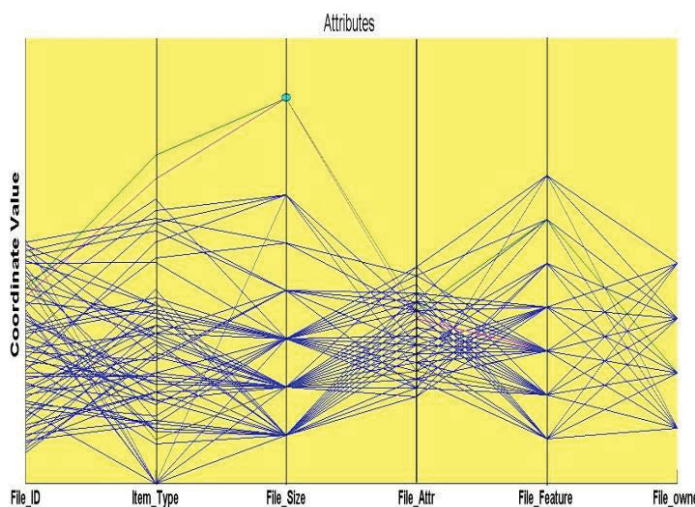


Figure 5-4: The Visualisation of Original Metadata by Parallel Coordinate

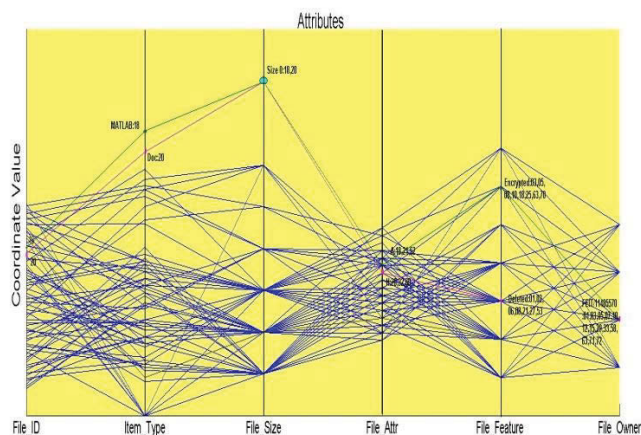


Figure 5-5 Use Parallel Coordinate to identify the suspicious file in HDD

To conclude, visualisation techniques play an important role in displaying high-dimensional metadata. Our results show that during the investigation of the

large capacity HDD, Parallel Coordinates not only aids the easy visual identification of files, it also displays the relationship among files directly. The mapping of the multivariate data on 2D space greatly helps the analysis of files' relationships with similar attributes. In addition, this method also has low representational complexity and is mathematically rigorous.

5.4 MDV Works in Criminal Relationship Detection

In forensic investigation processes, searching for the targeted criminal among many crime suspects is important. In this section, we discuss how visualisation techniques work well in the feature analysis among people and activities. In other words, we discovered that visualisation techniques in selecting features for forensic investigations not only improves the time of selection, but also deeply and obviously displays the slight changes of features in relation to criminals and also the relationships among various features and criminals. Which is an essential part of finding the target with significant differences to others. Additionally, it also predicts more active features to help investigate similar data resources in future.

In detail, Section 5.4.1 introduces the related works in forensic investigation, and Section 5.4.2 describes the fundamental techniques we apply to the self-organizing map, then we propose our approach in analysing the criminal relationships in Section 5.4.3, followed by the experiment and result representation and a short conclusion in Section 5.4.4 and Section 5.4.5 respectively.

5.4.1 Related Works

Displaying criminal data and understanding relationships by visualisation techniques (Jewitt & Van Leeuwen 2001) is necessary due to the cognitive and intelligence benefits. Some approaches in the visualisation domain had already been applied to forensic investigations during the past years. For example: Link Discovery (tool), COPLINK (tool) and hyperbolic tree view.

- Link Discovery (Horn, Birdwell & Leedy 1997) is developed by the University of Tennessee and St. Petersburg Police Department. It visualises associations such as relatives, criminals etc.. But it only deals with one target instead of different entities such as persons, addresses, and mobile numbers. It could not identify the relationships among items if they are associated with other features (e.g. if the two persons wore the same clothes).
- COPLINK (Hauck et al. 2002) is another detecting tool used in crime analysis. This tool provides two visualisation ways to view data information, a hyperbolic tree view and hierarchical list view. The connections between items will be shown between nodes. As the degree of associations are computed by co-occurrence weight which predicts the existence of a relationship. The higher a co-occurrence weight, the more likely the two items involved have a relationship of influence on each other.
- In the hyperbolic tree view (Pirolli, Card & Van Der Wege 2003), nodes are classified into centre node and other nodes, where the centre node is

the target item, and other nodes surrounding of centre node are the associated targets. So that researchers can find the global structure and details at the same time. In the hierarchical list view (Zhou & Feiner 1998), the target item will be on the first level and as others are divided into different levels hierarchically, which is easier to understand.

However, with the increase of data in volume, tools like COPLINK, are not efficient means for forensic investigators to find information from a large amount of data. Most of the data are too crowded to be displayed clearly (Mohay 2005).

5.4.2 Self-Organizing Map

It is always difficult for humans to read Multi-dimensional data figuratively, but Self-Organizing Map (Yasinsac et al.) technology can help people to visually understand the characteristics of these complicated data.

Self-Organizing Map (Yasinsac et al.) is a type of artificial neural network that is trained using unsupervised learning and the choice of visual clustering approaches to solve clustering problems. It is also regarded as an important way of data visualisation and data reduction because it gives a complete information illustration in results.

Geometrically, this approach maps high-dimensional data, producing a two dimensional representation of the input space of the training samples, and clusters data with qualitative features (Vesanto & Alhoniemi 2000). This method is implemented through two main steps: training and mapping. Training builds the map using input examples, and mapping, which automatically classifies a new input vector. Particularly, there are two main components called nodes or neurons (Hoglund, Hatonen & Sorvar 2000), and two layers (Ramadas, Ostermann & Tjaden 2003): input layer and output layer in the whole implementation process. The input layer analogs and shows the retina the outside information, while the output layer simulates responses from the cerebral cortex. When data has been input, one data will be mapped into different neurons, and then the data will be trained until features are stable, see Figure 5-6. The principle of SOM is based on biological competition principle and the details

are shown below:

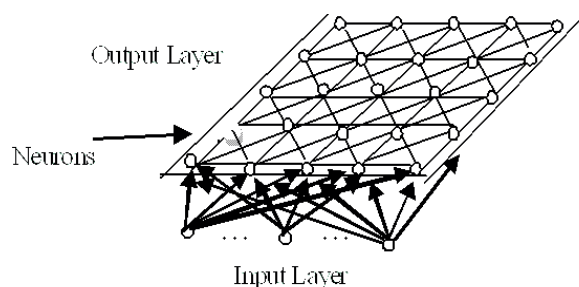


Figure 5-6: The theory of the Self Organizing Map (Vesanto & Alhoniemi 2000)

Firstly, initialise the network through setting random value to the weights between the inputting and mapping layer, then normalise the vectors $X_j, W_j (j=1,2,\dots,m)$ to get \hat{X} and \hat{W}_j by Equations 14 and 15.

$$\hat{X} = \frac{X}{\|X\|} = \left(\frac{x_1}{\sqrt{\sum_{j=1}^n x_j^2}} \dots \frac{x_n}{\sqrt{\sum_{j=1}^n x_j^2}} \right)$$

Equation 14 calculation of \hat{X}

$$\hat{W}_j = \frac{W_j}{\|W_j\|} = \left(\frac{w_{j1}}{\sqrt{\sum_{j=1}^n w_{j1}^2}} \dots \frac{w_{jn}}{\sqrt{\sum_{j=1}^n w_{jn}^2}} \right)$$

Equation 15 calculation of \hat{W}_j

to generate an initial winning field $N_{j^*}(0)$, and also initialize the training rate η .

The next step is to input data into the input layer, followed by computing the

distance between neurons weight vector of mapping layer and input by Euclidean distance measurement or Cosine Similarity. For Euclidean distance, as is expressed in Equation 16:

$$d = \|X - X_i\| = \sqrt{(X - X_i)(X - X_i)^T}$$

Equation 16 Euclidean distance calculation

In this formula, while d is smaller and X and X_i is closer, which represents that X and X_i is more similar; if $d=0$, then $X = X_i$; let $d = T(\text{constant})$ be the standard to cluster input data. For example, $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$ are the input data, compute their distances between each other. Suppose:

$$d_{12} < T, d_{24} < T, d_{14} < T, d_{38} < T, d_{78} < T, d_{67} < T, d_{56} < T, d_{35} < T$$

while

$$\begin{aligned} d_{36} > T, d_{56} > T, d_{67} > T, d_{68} > T; \\ d_{13} > T, d_{15} > T, d_{16} > T, d_{17} > T, d_{18} > T; \\ d_{23} > T, d_{25} > T, d_{26} > T, d_{27} > T, d_{28} > T; \\ d_{34} > T, d_{45} > T, d_{46} > T, d_{47} > T, d_{48} > T; \end{aligned}$$

then the data can be classified into two clusters: $A(X_1, X_2, X_4)$,

$B(X_3, X_5, X_6, X_7, X_8)$, as is shown in Figure 5-7.

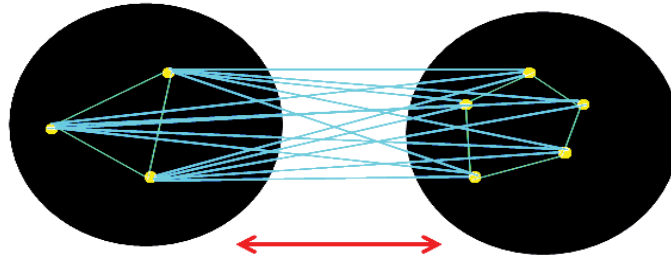


Figure 5-7: the distance between two neurons

Cosine Similarity, Equation 17 is another way to measure the distance between objects:

$$\cos \varphi = \frac{X^T X_i}{\|X\| \|X_i\|}$$

Equation 17: measure the distance between two objects

where φ is the angle between two vectors;

when φ is smaller, X and X_i is closer, which represents X and X_i is more similar; if $\varphi = 0$, then $\cos \varphi = 1, X = X_i$; let $\varphi = \varphi_0$ be the standard to cluster.

Get the minimum among all the distances, which is called Winner Neuron, and its weight is \hat{W}_{j^*} , where distance d equals to Equation 18 :

$$\begin{aligned}
 d &= \sqrt{(\hat{X} - \hat{W}_{j^*})(\hat{X} - \hat{W}_{j^*})^T} = \sqrt{\hat{X} \hat{X}^T - \hat{W}_{j^*} \hat{X}^T - \hat{X} \hat{W}_{j^*}^T + \hat{W}_{j^*} \hat{W}_{j^*}^T} \\
 &= \sqrt{\hat{X} \hat{X}^T - 2\hat{W}_{j^*} \hat{X}^T + \hat{W}_{j^*} \hat{W}_{j^*}^T} \\
 &= \sqrt{\hat{X} \hat{X}^T - 2\hat{W}_{j^*} \hat{X}^T + \hat{W}_{j^*} \hat{W}_{j^*}^T} \\
 &= \sqrt{2E - 2\hat{W}_{j^*} \hat{X}^T} = \sqrt{2(E - \hat{W}_{j^*} \hat{X}^T)}
 \end{aligned}$$

Equation 18: calculate the distances between every two objects

Therefore

$\hat{W}_{j^*} \hat{X}^T$ is the maximum value, if the minimum distance equals to Equation 19

$$\min_{j \in \{1, 2, \dots, n\}} \left\{ \left\| \hat{X} - \hat{W}_{j^*} \right\| \right\}$$

Equation 19: calculate the minimum distance among all distances

and the j^{th} neuron is the winner neuron, and the winner neuron will be the one with the largest scalar product, see Equation 20:

$$\hat{W}_{j^*} \hat{X}^T = \max_{j \in \{1, 2, \dots, m\}} \left(\hat{W}^T \hat{X} \right)$$

Equation 20: certify that the winner neuron is the largest scalar product

The last step is to reorganize *weight*, see Equation 21

$$\begin{cases} W_j(t+1) = \hat{W}_j(t) + \Delta W_j = \hat{W}_j(t) + \mu(t)(\hat{X} - \hat{W}_j), j = j^* \\ W_j(t+1) = \hat{W}_j(t), j \neq j^* \end{cases}$$

Equation 21: recognise the weight

Repeat the above training steps until training rate decays to zero or a positive decimal.

The results can be shown in five different graphs: topology graph, neighbour weight distance, weight planes, classification hits, and neuron's weight vectors.

- **Topology Graph:** Describes the layout of neurons in competitive layer.
- **Neighbour Weight Distance:** Each blue dot represents a neuron, and the red line represents a connection between neurons, the rhombus represents the distance between neurons, while the colour is between yellow and dark; when the colour is darker, the distance is larger.
- **Weight Planes:** This graph consists of a set of subplots. Each i^{th} subplot shows the weights from the i^{th} input to the layer's neurons, with the most negative connections shown as blue, zero connections as black, and the strongest positive connections as red.
- **Classification Hits:** It is a SOM layer, with each neuron showing the number of input vectors that it classifies. The relative number of vectors for each neuron is shown via the size of a blue patch.
- **Neuron's Weight Vector:** The input vectors are green dots, and show how the SOM classifies the input space through representing blue dots for each neuron's weight vector, and the red lines show the connections between neighbouring neurons.

5.4.3 Our Approach

Considering the amount of data and the requirements of crime analysis of levels of simplicity and legibility (leong 2006), lots of effort has been put into visual understanding and analysing the relationships among criminal data. In our work, we propose a method to reduce data complexity without changing the variation trend, and present all results simply and legally in a graphical manner. This visualisation based method could also reduce the analysing time, and increase the probability of identifying criminal or criminal activities. In addition, our approach is also able to help users to query a specific item or feature to see statistics during the analysis. Such as: item details in all features, the relationship among all items, the variation trends and relationships among all features, and also the relationship between item and features.

Our approach is based on the self-organizing map, and the detail is described below:

(1) Initialization

Set ransom value to the weights between inputting and mapping layer, and normalize a vector (Fei et al. 2006) $\hat{W}_j, j=1,2,\dots,m$, use the methods in Equations 14 and 15 to get \hat{X} and \hat{W}_j , and then initialise a winning field $N_{j^*}(0)$.

(2) Learning Process

After initialising data, each training data will be submitted to SOM orderly, and normally it takes several incidents of iteration. The iteration includes five parts.

- **Getting Input data:** Choose an input model from training sets, normalize it and get: $X^p, p \in \{1, 2, \dots, p\}$.
- **Finding Winner Neuron:** calculated by the flowing equation

$$\begin{aligned}
 \|\hat{X} - \hat{W}_{j^*}\| &= \min_{j \in \{1, 2, \dots\}} \{\|\hat{X} - \hat{W}_{j^*}\|\} \\
 \|\hat{X} - \hat{W}_{j^*}\| &= \sqrt{(\hat{X} - \hat{W}_{j^*})^T (\hat{X} - \hat{W}_{j^*})} \\
 &= \sqrt{\hat{X}^T \hat{X} - 2\hat{X} \hat{W}_{j^*} + \hat{W}_{j^*}^T \hat{W}_{j^*}} \\
 &= \sqrt{2(1 - \hat{W}_{j^*}^T \hat{X})}
 \end{aligned}$$

Therefore $\|\hat{X} - \hat{W}_{j^*}\| = \max_{j \in \{1, 2, \dots\}} \{\hat{W}_{j^*}^T \hat{X}\}$.

Specifically,

$$\begin{aligned}
 D(Item_i, Item_j) &= \sqrt{(Item_{i,1} - Item_{j,1})^2 + (Item_{i,2} - Item_{j,2})^2 + \dots + (Item_{i,n} - Item_{j,n})^2} \\
 &= \sqrt{\sum_{i,j=1}^n (Item_{i,k} - Item_{j,k})^2}
 \end{aligned}$$

Where:

If d is smaller, then X and X_i have more similarity.

If $d=0$, then $X = X_i$

- **Wining** : Define a wining area (Rauber, Merkl & Dittenbach 2002) $N_{j^*}(t)$, which sets j^* as the centre to search the weight adjustment area at t . When t is increased, $N_{j^*}(t)$ is decreased, and normally, $N_{j^*}(0)$ is the maximum value among all $N_{j^*}(t)$.

- **Weight Adjustment**: We use the following formula to adjust the weights.

$$W_{j^*}(t+1) = \hat{W}_{j^*}(t) + \Delta W_{j^*} = \hat{W}_{j^*}(t) + \mu(t, N)(\hat{X} - \hat{W}_{j^*})$$

Where:

t - The t^{th} iteration

N - Topological distance between the j^{th} neuron and winning neuron j^* in the adjacent area. $W_{j^*}(t)$ is the weight of the j^{th} neuron; $\mu(t, N)$ is the association between training time t and the topological distance N . If N is increasing, $\mu(t, N)$ is increasing, but while t is increasing, $\mu(t, N)$ has a decreased trend, the relationship is $\mu(t, N) = \mu(t)e^{-N}$. The result displays weight planes and classification hits. In the weight planes, the graph consists of a set of subplots. Each l^{th} subplot shows the weights from the l^{th} input to the layer's neurons, with the most negative connections shown as blue, zero connections as black, and the strongest positive connections as red. In the result of classification hits, there is a SOM layer, with each neuron showing the number of input vectors that it classifies. The relative number of vectors for each neuron is shown via the size of a blue patch.

- **Graphical Classification:** We suppose a dataset

$$U = \{U_1, U_2, \dots, U_n\}$$

Where:

n - The number of items

And another dataset stores L labels for U_i , which is defined as:

$$L = \{L_1, L_2, \dots, L_m\}. \text{ Therefore, } U_i \text{ can be marked as: } U_i = \{U_{i,1}, U_{i,2}, \dots, U_{i,m}\}.$$

Where m represents the number of the labels, while U can be transformed as:

$$\begin{bmatrix} U_{1,1} & U_{1,2} & \dots & U_{1,m} \\ U_{2,1} & U_{2,2} & \dots & U_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ U_{n,1} & U_{n,2} & \dots & U_{n,m} \end{bmatrix}$$

Where:

$$n \in N, m \in N, i \in N, j \in N$$

As $U_i \in U$, which is related to $l \in 2^{L_i}$, where: $l = \{l_1, l_2, \dots, l_m\} (m \leq n), U \notin \left(\frac{l}{L}\right)$. If

L is consisting of $|L|$ multi-label examples $(L_i, U_i) = i = 1, \dots, |L|$, two variables are generated: label density and cardinality. Label density (Andresen 2006; Maciejewski et al. 2008) is the average number of labels of the examples in L divided by $|D|$, where $|D|$ is much larger compared to the number of labels for each sample.

The formula is shown as below.

$$LD(L) = \frac{1}{|L|} \sum_{i=1}^L \frac{U_i}{D};$$

while the Label Cardinality of L (McQuaid et al. 1999) is the average number of labels of the examples in L , and the formula is shown below.

$$LC(L) = \frac{1}{|L|} \sum_{i=1}^L U_i$$

In addition, Label cardinality is used to quantify the number of alternative labels that characterise the examples of a multi-label training dataset, and is dependent of the number of labels $|D|$ in the classification. Generally, Label density takes into consideration the number of labels in the classification problem. Two datasets with the same label cardinality and with similar label density might exhibit the same properties and cause similar behaviour to the multi-label classification (Tsoumakas & Katakis 2006) (Read et al. 2011). The relations between Label density and Label Cardinality can be calculated by

$$LC(F) = |D|LD(F).$$

The above is the whole algorithm of our self-organizing map visualisation approach, and the results will be displayed in Section 5.4.4 with real dataset experiments.

5.4.4 Experiment and results

In our experiment, the dataset contains 16 items with 16383 features, and we recognise these 16 items as the suspicious people. Our aim is to visually analyse the associations between these suspicious targets, and detect the more effective features to be used for future studies among 16383 features by visualisation approaches.

Preliminary works:

Suppose there are 16 suspicious people, which are defined as $I = \{Item_1, Item_2, Item_3, \dots, Item_{14}, Item_{15}, Item_{16}\}$. Each character is marked as F_i . $F_i = \{F_{i,1}, F_{i,2}, \dots, F_{i,m-1}, F_{i,m}\}$. The characters contain, for example, eye pupil distance, nose height, lip thickness, distance between eyes and eye brows, blood concentrations, etc. Some features are of the same value (or have no value) towards each item, and they have no detectable effects in our experiments, would also affect the processing rate for future analysing. So we pre-process them by deletion. What's more, in this dataset, each person can be represented by 862 characters, which means a single item can be classified into 862 different groups.

Based on the above experiment conditions, our test contains three main steps:

a) Original data (Statistic) Visualisation

In comparison with other visual techniques (Herman, Melançon & Marshould

2000; Lamping & Rao 1994; Morse, Lewis & Olsen 2000), our tool is able to be displayed in lines. Each feature is to be used as a label for each item. We can find the clear differences among all items from the entire graph, and each point is displayed with values. See Figure 5-8

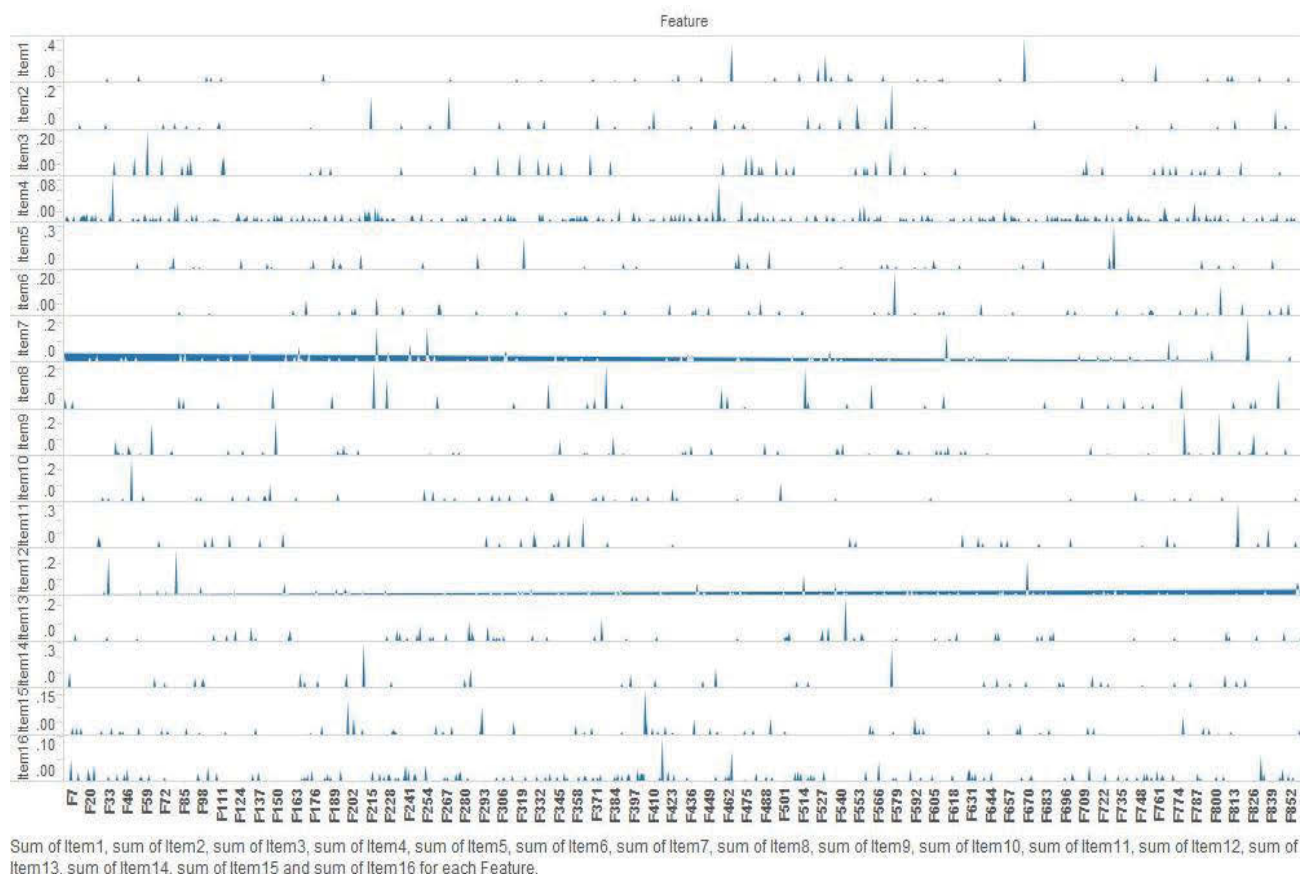


Figure 5-8: An overview of the whole 16 items' feature value and the circled areas are the suspicious data.

In Figure 5-9, the *Y*-axis (vertical) represents the 16 items (Person), and the *X*-axis (horizontal) represents the 862 features. We can discover the general changing trends of each item. For example, as the feature number is increased, the changing rate of Item 7 is decreased, which means that Item 7 has been influenced more by the first few features, while Item 12 is influenced more by the

last few features. There is no special feature that has influence on the rest of the items, as they are distributed averagely. In Figure 5-9, we can also discover the overall difference between each item. For example, we can find Item1, Item 4 and Item 16 have similar values on most of the features.

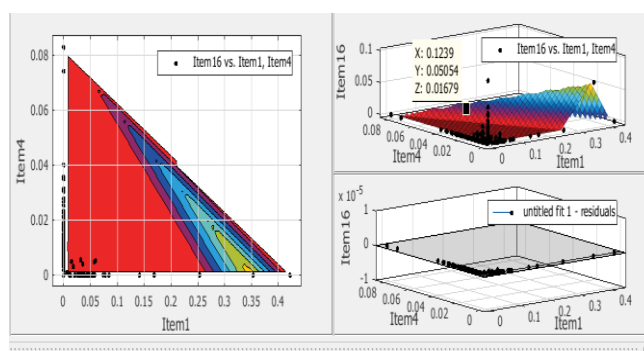


Figure 5-9: the difference between Item 1, Item 4 and Item 16.

We can also find that because of the limited space, some of the features are hidden and only part of them can be shown. In order to solve this problem, we use the self-organizing map to pre-process the data without influencing the changing trends.

b) Reducing Data by Self-Organizing Map

Set the net size equal to 30, Figure 5-10 is the visualisation result of all features through SOM (Bingham & Mannila 2001). There are some highlighted parts in each panel, and the highlighted parts reflect the influences of the features on the items. The relative classification hits are also generated, which is shown in Figure 5-11, and we find that the larger the polygon, the more relative the vectors.

c) Discover the Anonymous Targets

After using SOM to pre-process the dataset, we visualise the new dataset, see Figure 5-12.

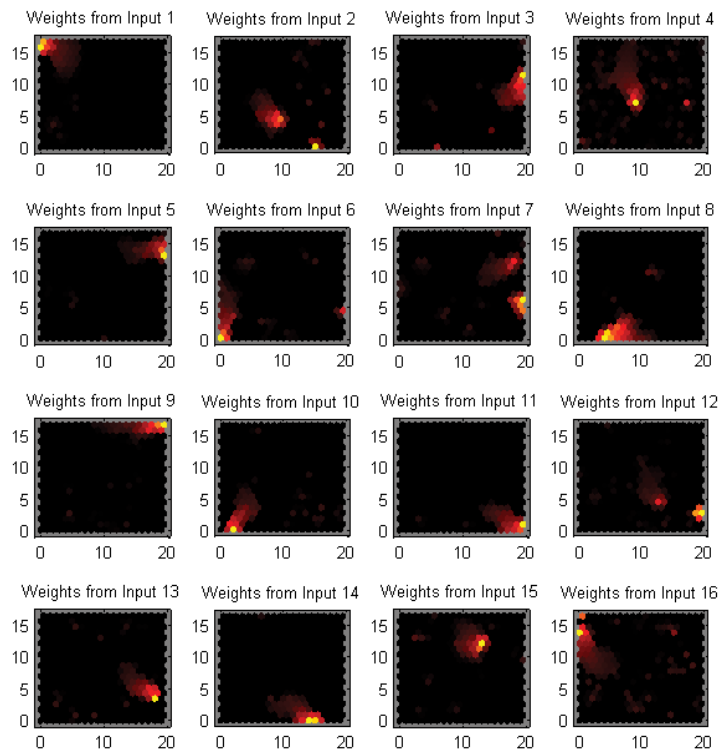


Figure 5-10: the weight planes of all 16 items by SOM

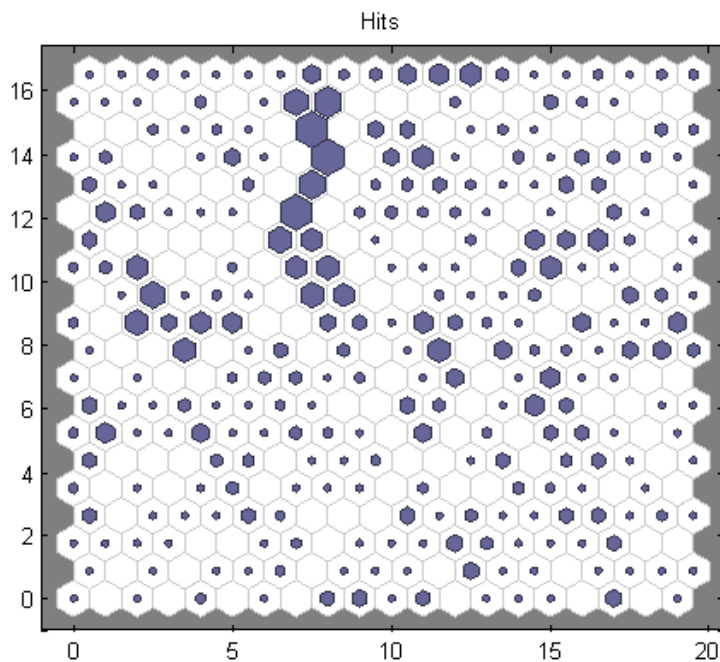


Figure 5-11: the classification hits after using SOM of 16 items

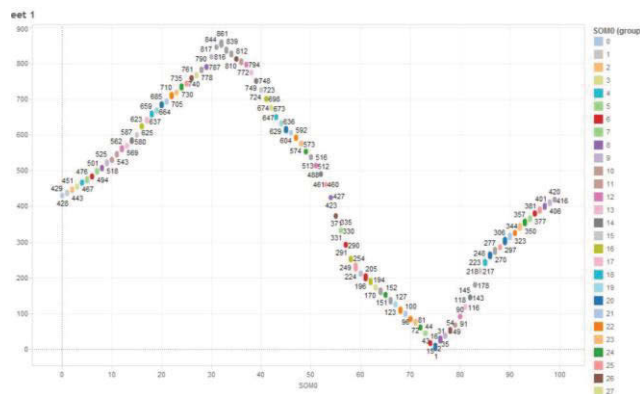


Figure 5-12: the visualisation of classification of all features

We could find the difference between each data net. These graphs describe that related features are grouped into a cluster. Although in each cluster with a generated new attribute, the graph trends are similar, and researchers could analyse the clustering group instead the whole features and we can click each

cluster to find the inner relationships.

Overall, our experiment shows that the combination of self-organizing map and visualisation technique can improve the analysis efficiency in forensics. Particularly, we use self-organizing map to reduce and classify the original criminal record, and the visualisation techniques allow investigators to interactively select the target for better understanding of the activities or possibilities. Additionally, this approach represents the hierarchically reduced features, data distribution and relationships between features and items. The high risk features and items can also be predicted. It is highly efficient to identify the suspicious criminals among a large amount of data

5.4.5 Conclusion

This proposed graph-based feature analysis method in analysing associations between features and criminals is effective. It saves time for forensic investigators in targeting suspicious criminals, and also brings benefits to future analysis through displaying the changing trends of each feature.

5.5 MDV Helps in Crime Analysis

Visualizing Information can extract (Leung & Khan 2006; Tarjan 1983) essential information and knowledge from large and complex datasets, and can also represent them clearly and comprehensively in graphic forms for a better understanding of the patterns hidden. Therefore, using the concept of visual perception, many applications or tools have been developed that display file information in a graphical manner. These visualisation techniques reduce the time which investigators needed to analyse digital evidence, and greatly increase the probability of locating suspicious data.

In this section, we describe how visual cues support the hierarchical exploration in the forensic domain, and particularly introduce how visual cues within DOITree Visualisation analyses the Enron Email Dataset.

5.5.1 Tree Visualisation

Today, a large majority of digital information exists within a hierarchical structure, for example, the directory structure in an information management system, the pyramid selling scheme in business, and the work breakdown structure in project management and so forth. They can all be represented in a tree graphical form, and are usually called hierarchical trees.

Such structures play important roles in understanding the distribution of data across the hierarchy, and also summarising the low-level data into manageable data to reduce information overload. Consequently, developing a user-friendly visualisation technique of the whole dataset, with an ability to allow for deeper discovery at certain levels of granularity is essential for the researchers in the information exploration process.

In the early stage of researches in the presentation of hierarchical structure, tree visualisation has been categorized into two main streams (Nguyen & Huang 2002). First is the vertex-edge (node-link) diagram layout based visualisation, second is the space-filling layout based visualisation. Specifically, the vertex-edge diagram is capturing entities as vertices and relationships, by visible graphical edges, usually lines, connecting vertices from the parents to their children to present relationships among data objects. The advantage of this approach is that the user can discover the links among data directly and clearly. This technique has the simplification of the process of understanding the relational structures of information in a graph, while the space-filling method is to use enclosure to represent the tree structure rather than use connection

method.

Up to now, many techniques have been developed in tree visualisation. Such as the balloon view (Herman, Melançon & Marshould 2000), space-optimised tree (Nguyen & Huang 2003b), EncCon (Nguyen & Huang 2005), radial view (Bingham & Sudarsanam 2000), tree-maps (Shneiderman 1992) , etc. In addition, many of them had been successfully supported by various research and business enterprises during the past years. In forensic investigations, for instance, disk tree (Chi et al. 1998) can assist in explaining the file systems for deeper investigation. However, considering the large volume of datasets today, SpaceTree (Plaisant, Grosjean & Bederson 2002) and DOITrees (Heer & Card 2004) are the most acceptable techniques for large Tree visualisations. In the following Section 5.5.2, we will further discuss how an optimised DOITree visualisation works well in forensic investigation.

5.5.2 DOITree Visualisation on Enron Email Dataset

DOITree (degree-of-interest tree) (Heer & Card 2004; Nguyen, Simoff & Huang 2014) visualisation is a multi-focal tree layout algorithm, for discovering large hierarchical structures. It has the capability to provide multiple focused nodes and a dynamic rescaling of substructures to fit the variable space. However, the original DOITree is limited to provide a clear presentation of the hidden structure. So visual cues are proposed to help with DOITree (Nguyen, Simoff & Huang 2014) for the effectiveness and capability of providing multiple focus branches, and following for the visibility and readability of the hidden structures to be enhanced.

Figures 5-13 (a) (b) are the optimised DOITree descriptions of Enron Email Dataset (Shetty & Adibi 2004), which was originally made public and posted to the web by the Federal Energy Regulatory Commission during its investigation after the company's collapse. Enron Email is also a large database over 600,000 emails generated by 158 users, mostly senior management of Enron Company, and all the files are organised into folders.

By applying DOITree, users can click any item to explore further information. Specifically, the arc represents the links between data objects. The triangle is a guide for the users to overview the contents. Two attributes of the triangles, colour and size, explain the amount of data stored per object. The darker and larger the triangle, the more data items it contains. See Figure 5-13(a), the triangle in rectangle A is larger and darker than the triangles in rectangle B, which means the employees in rectangle A are closely related to each other, and

they communicate with bauchman-d under the same case. While the employees, whose email address is andrea.p.williams@pwcglobal.com and cjwilson@lcc.net, are more likely involved in a same company project.

In addition, we can use colour property in DOITree visualisation to identify the same objects (employees). In Figure 5-13(b), it is clear to discover that Matthew appeared twice, and the green colour represents the category they belong to at different levels.

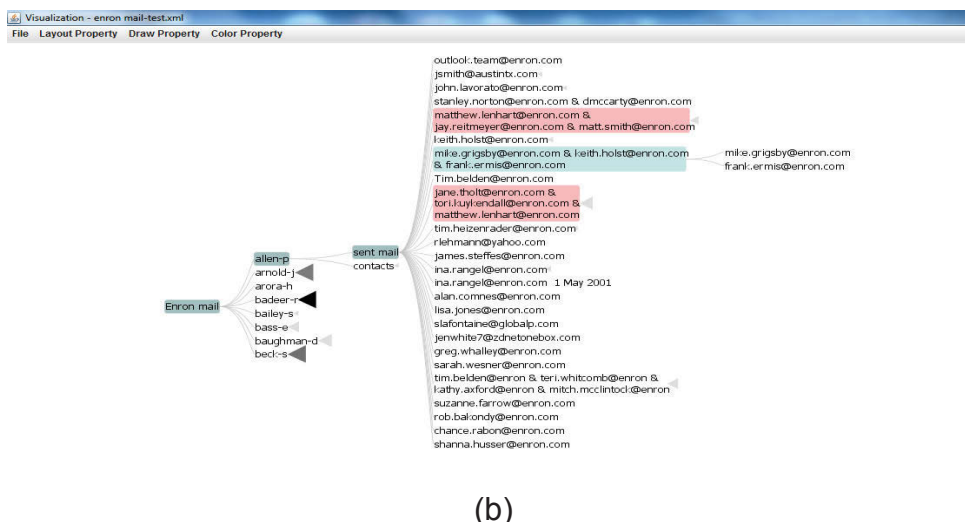
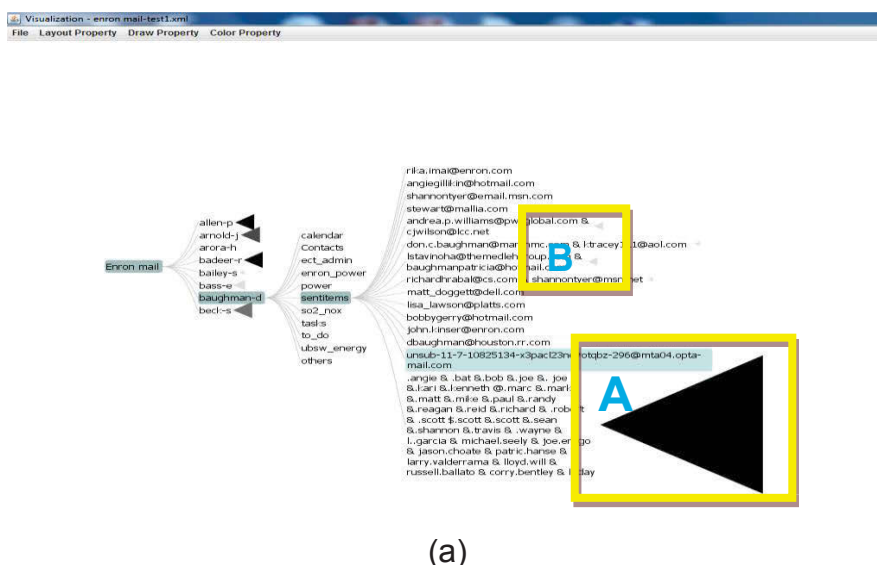


Figure 5-13: Examples of the emailing activities for eight employees in Enron. a) check the activities of bauchman-d. b) using colour in DOITree

In summary, there are many advantages of applying DOITrees to visually control the investigation processes. 1) It is a representing tree in a universally accepted classical way. 2) It provides multiple-foci views with “focus + context” interaction. 3) It utilises display space. 4) It provides smooth fading in/out animations among transitions.

Chapter 6 Conclusion and Future Work

The preceding chapters have presented, discussed and illustrated all the relevant backgrounds and technical components of my PhD research. This chapter concludes the thesis by providing a summary of key contributions made by this thesis to the discipline of Multi-dimensional data visualisation. Finally, this chapter outlines the directions for the future work.

6.1 Reflections on thesis questions

In this information age, we need timely, accurate and correct answers for almost every event. However, it is getting harder because we are in the era of “big data”, where datasets are characterised by a high volume velocity and variety. People are also highly dependent on the datasets to make right decisions and solve their problems.

However, the traditional analytic methodologies have not been sufficient to analyse and understand such large, complex and dynamic datasets. To make an improvement on the conventional approaches to multi-dimensional data visualisation, proper visualisation techniques could enable humans to merge and simplify information, discover deeper insights from complicated data, and provide available suggestions without delay. The optimised visualisation applications can enhance humans' capacity of perceiving and exploring data.

As the typical multi-dimensional data visualisation, scatterplot matrix and parallel coordinates provide the capability of displaying multivariate data. So far, they have been successfully applied to datasets from various domains, including bioinformatics (Frank et al. 2004; Saraiya, North & Duca 2005), geography (Anselin 1993) and various others (Inselberg 2009).

However, the traditional scatterplot matrix visualisation and parallel coordinate visualisation all have challenges and shortcomings as mentioned in the introduction. Firstly, in the scatterplot matrix, the analysts often need the representation of multiple plots (views) at the same time, due to the requirements of discovering features among all variables (dimensions). More

often, the visualisation tool displays all the plots when the visualisation of scatterplot matrix function is required. To avoid blurring the results and wasting screen space, visualisation researchers have to consider the utility of each plot and the place of each plot. For example, reducing the duplicated plots without shifting positions and transferring the different scatterplots to separating views with polygons do well to simplify results. However, these methodologies do not consider the display space utilisation of the computer screen, where it can always be fully filled by all necessary plots without considering its flexible transformation of the screen size, either enlarging or minimising (Research Objective 1 and 2).

Secondly, most of the existing scatterplot matrix have a common limitation: they display the differentiations among all variables, but could not reflect well the information inside each plot. Additionally, they are not handy for users to point out their desired information among a large number of choices (data values) in each plot (Research Objective 3).

Thirdly, multi-dimensional data visualisation techniques have been widely applied in many research domains and some of them have achieved successful results. However, most of the investigations in forensics have not considered adopting visualisation techniques in their analysis processes (Research Objective 4).

6.2 Answers to thesis questions

Motivated by these three challenges, we propose the concept of a space-optimised scatterplot matrix (Answer Research Objective 1 and 2). Specifically, we created a layout calculation method for scatterplot matrix with implementation algorithms. Our new scatterplot matrix manages to effectively utilise display screen space optimisation on the one hand; on the other hand, the colour mapping method in our technique offers clear recognition of variables' relationships, and this keeps one of the main contributions of the original scatterplot matrix. Generally, our method maintains a balance between space utilisation and information clarity.

In addition, we achieve an interactive exploration in scatterplot matrix (Answer Research Objective 3). Our solution is a more comprehensive approach with a point-to-region interaction model for class data, a feature ranking method, rough set theory, for dimensionality reduction the Multi-dimensional data. We also use a decision trend analysis function to guide the user for info prediction.

Finally, we enhance the analysis productivity in a specific domain, computer forensic investigation, through the application of multi-dimensional data visualisation techniques from various aspects (Answer Research Objective 4). Specifically, we carry out a new model for a timely investigation to search out the crime suspects by introducing visualisation techniques in different analytical steps. We also exploit parallel coordinates to display one of the main data storage mediums, the hard disk drive, by fixing the time zone problem, which is a key element in crime investigation. To help the investigation of digital crimes, we apply the self-organizing map to a special case, the criminal relationship

detection from our study. Under this method, specialists are much easier to discover the anonymous relationships through the visualisation results.

6.3 Future Work

For future work, we focus on three aspects: technological optimisation on scatterplot matrix in Section 6.3.1; systematic scatterplot matrix evaluation guidelines in Section 6.3.2; and collaborative Multi-dimensional data visualisation techniques with other domains in Section 6.3.3.

6.3.1 Technological Optimisation

Firstly, the space optimised scatterplot matrix technique presented in this thesis is still in the early stage of becoming a mature visualisation tool. There are still a number of technical imperfections which need to be improved. For example, overlapping data point causes visualisation occlusion when the dataset increases dramatically, not enough dynamic query functions provide enough guidance for users as well.

In addition, the current system of interactive exploration in scatterplot matrix also requires non-trivial computational time to search the solution space exhaustively. So we would like to integrate the evolutionary algorithm as suggested in (Bjorvand & Komorowski 1997) with *QuickReduct* to improve the time complexity.

6.3.2 Systematic Scatterplot matrix Evaluation Guidance

Although we conducted formal usability studies to evaluate both the space-optimised scatterplot matrix and the interactive scatterplot matrix visualisation that could prevent the shortcomings of empirical metrics. There is still a need to understand how to evaluate visual analytics systems as a whole. Up to now, no formal evaluation pipeline has been proposed. Therefore, it is appropriate to provide fundamental guidelines for visualisation evaluation.

6.3.3 Cooperate with Other Domains

As researchers in data visualisation gradually transferred to visual analytics which focuses on analytical reasoning, designing user centred visualisation interfaces should be combined with human perception analysing processes. It is predictable that even with a very well-designed visualisation interface and in a high configuration hardware environment, users still discern that the time spent on exploiting and interpreting worthwhile information by using visual analytic tools is significantly long. The research shows that the main cause of visualisation tool failing to fulfil the analysts' need is the gaps between visual analytics and analytical reasoning in applying visual analytics. The ways of reflecting and corresponding between human and data can be influenced negatively when the result is presented with an improper structure. Therefore, in the future, we will also cooperate with domain specialists and put more effort into finding out their industry requirements. We will also examine the analytic pipeline and the utilisation of visualisation applications from a user's aspect.

On the other hand, it is important to enlarge the development and availability of multi-dimensional visualisation techniques. For example, integrating them into other domains, where they are required for the discovery of knowledge from multivariate datasets. Although we found that some visualisation techniques could play an important role in forensic science, more experiments and tests need to be accomplished at a future date.

Publication List

1. Wen-bo Wang, Mao Lin Huang, Quang Vinh Nguen. A Space optimised Scatterplot matrix Visualisation. The 13th International Conference on cooperative Design, Visualisation and Engineering. 2016.
2. Wen-bo Wang, Maolin Huang, Quang Vinh Nguyen, Kang Zhang, Tze-Haw Huang. Enabling Decision Trend Analysis with Interactive Scatterplot Matrix Visualisation. Journal of Visual Languages and Computing. 2016.
3. Wen-bo Wang, Mao Lin Huang, Jinson Zhang, Wei Lai. Detecting Criminal Relationships through SOM Visual Analytics. The 19th International Conference Information Visualisation.pp:316-321,2015.
4. Wen-bo Wang, Mao Lin Huang, Jinson Zhang, Liangfu Lu, Improving performance of Forensic investigation with Parallel Coordinates Visual Analytics. The 8th International Conference on Frontier of Computer Science and Technology,(FCST 2014).pp:1839-1843.
5. Jinson Zhang, Maolin Huang, Wen-bo Wang. Big Data Density Analytics using

Parallel Coordinate Visualisation, The 13th International Symposium on Pervasive Systems, Algorithms, and Networks (I-SPAN2014).pp.1115-1120.

6. Wen-bo Wang, Mao Lin Huang, Parallel Coordinates Visualisation of Large Data Investigation on HDDs. Proceeding of the 10th international conference of Computer Graphics, imaging and visualisation, 2013.(Best Paper Award)

Reference

- Abram, G & Treinish, L. 1995, 'An extended data-flow architecture for data analysis and visualisation', *Proceedings of the 6th conference on Visualisation'95*, IEEE Computer Society, p. 263.
- Al-Awami, A.K., Beyer, J., Haehn, D., Kasthuri, N., Lichtman, J.W., Pfister, H. & Hadwiger, M. 2016, 'NeuroBlocks–Visual Tracking of Segmentation and Proofreading for Large Connectomics Projects', *Visualisation and Computer Graphics, IEEE Transactions on*, vol. 22, no. 1, pp. 738-46.
- Albuquerque, G, Eisemann, M., Lehmann, D.J., Theisel, H. & Magnor, M.A. 2009, 'Quality-Based Visualisation Matrix', *VMV*, pp. 341-50.
- Alpern, B. & Carter, L. 1991, 'The hyperbox', *Visualisation, 1991. Visualisation'91, Proceedings., IEEE Conference on*, IEEE, pp. 133-9, 418.
- Andresen, M.A. 2006, 'Crime measures and the spatial analysis of criminal activity', *British Journal of Criminology*, vol. 46, no. 2, pp. 258-85.
- Andrews, D.F. 1972, 'Plots of high-dimensional data', *Biometrics*, pp. 125-36.
- Ankerst, M. 2001, 'Visual data mining with pixel-oriented visualisation techniques', *Proceedings of the ACM SIGKDD Workshop on Visual Data Mining*.
- Anselin, L. 1993, *The Moran scatterplot as an ESDA tool to assess local instability in spatial association*, Regional Research Institute, West Virginia University Morgantown, WV.
- Artero, A.O., de Oliveira, M.C.F. & Levkowitz, H. 2004, 'Uncovering clusters in crowded parallel coordinates visualisations', *Information Visualisation, 2004. INFOVIS 2004. IEEE Symposium on*, IEEE, pp. 81-8.
- Baehrecke, E.H., Dang, N., Babaria, K. & Shneiderman, B. 2004, 'Visualisation and analysis of microarray and gene ontology data with treemaps', *BMC bioinformatics*, vol. 5, no. 1, p. 84.
- Balzer, M., Deussen, O. & Lewerentz, C. 2005, 'Voronoi treemaps for the visualisation of software metrics', *Proceedings of the 2005 ACM symposium on Software visualisation*, ACM, pp. 165-72.
- Barga, R., Jackson, J., Araujo, N., Guo, D., Gautam, N. & Simmhan, Y. 2008, 'The trident scientific workflow workbench', *eScience, 2008. eScience'08. IEEE Fourth International Conference on*, IEEE, pp. 317-8.
- Baryamureeba, V. & Tushabe, F. 2004, 'The enhanced digital investigation process model', *Proceedings of the Fourth Digital Forensic Research Workshop*, Citeseer, pp. 1-9.
- Becker, R.A. & Cleveland, W.S. 1987, 'Brushing scatterplots', *Technometrics*, vol. 29, no. 2, pp. 127-42.
- Beebe, N.L. & Clark, J.G. 2005, 'A hierarchical, objectives-based framework for the digital investigations process', *Digital Investigation*, vol. 2, no. 2, pp. 147-67.
- Beyer, M. 2011, 'Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data', *Gartner. Archived from the original on*, vol. 10.
- Beynon, M. 2001, 'Reducts within the variable precision rough sets model: a further investigation', *European journal of operational research*, vol. 134, no. 3, pp. 592-605.
- Bingham, E. & Mannila, H. 2001, 'Random projection in dimensionality reduction: applications to image

- and text data', *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 245-50.
- Bingham, J. & Sudarsanam, S. 2000, 'Visualizing large hierarchical clusters in hyperbolic space', *Bioinformatics*, vol. 16, no. 7, pp. 660-1.
- Bjorvand, A.T. & Komorowski, J. 1997, 'Practical applications of genetic algorithms for efficient reduct computation', *Wissenschaft & Technik Verlag*, vol. 4, pp. 601-6.
- Card, S., Mackinlay, J. & Shneiderman, B. 2009, 'Information visualisation', *Human-computer interaction: design issues, solutions, and applications*, vol. 181.
- Card, S.K., Mackinlay, J.D. & Shneiderman, B. 1999, *Readings in information visualisation: using vision to think*, Morgan Kaufmann.
- Carr, D.B., Littlefield, R.J., Nicholson, W. & Littlefield, J. 1987, 'Scatterplot matrix techniques for large N', *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 424-36.
- Carrier, B. 2005, *File system forensic analysis*, Addison-Wesley Professional.
- Casey, E. & Stanley, A. 2004, 'Tool review—remote forensic preservation and examination tools', *Digital Investigation*, vol. 1, no. 4, pp. 284-97.
- Chan, Y.-H., Correa, C.D. & Ma, K.-L. 2010, 'Flow-based scatterplots for sensitivity analysis', *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, IEEE, pp. 43-50.
- Chatterjee, S. & Hadi, A.S. 2009, *Sensitivity analysis in linear regression*, vol. 327, John Wiley & Sons.
- Chen, M. & Jaenicke, H. 2010, 'An information-theoretic framework for visualisation', *Visualisation and Computer Graphics, IEEE Transactions on*, vol. 16, no. 6, pp. 1206-15.
- Chernoff, H. 1973, 'The use of faces to represent points in k-dimensional space graphically', *Journal of the American Statistical Association*, vol. 68, no. 342, pp. 361-8.
- Chi, E.H., Pitkow, J., Mackinlay, J., Pirolli, P., Gossweiler, R. & Card, S.K. 1998, 'Visualizing the evolution of web ecologies', *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press/Addison-Wesley Publishing Co., pp. 400-7.
- Chung, D.H., Legg, P.A., Parry, M.L., Bown, R., Griffiths, I.W., Laramée, R.S. & Chen, M. 2015, 'Glyph sorting: Interactive visualisation for Multi-dimensional data', *Information Visualisation*, vol. 14, no. 1, pp. 76-90.
- Cleveland, W.S. 1988, *The Collected Works of John W. Tukey: Graphics 1965-1985*, vol. 5, CRC Press.
- Cleveland, W.S. & McGill, R. 1985, 'Graphical perception and graphical methods for analyzing scientific data', *Science*, vol. 229, no. 4716, pp. 828-33.
- Convertino, G., Chen, J., Yost, B., Ryu, Y.-S. & North, C. 2003, 'Exploring context switching and cognition in dual-view coordinated visualisations', *Coordinated and Multiple Views in Exploratory Visualisation, 2003. Proceedings. International Conference on*, IEEE, pp. 55-62.
- Daum, F. 2005, 'Nonlinear filters: beyond the Kalman filter', *Aerospace and Electronic Systems Magazine, IEEE*, vol. 20, no. 8, pp. 57-69.
- Diesburg, S.M. & Wang, A.-I.A. 2010, 'A survey of confidential data storage and deletion methods', *ACM Computing Surveys (CSUR)*, vol. 43, no. 1, p. 2.
- Donath, J. 2002, 'A semantic approach to visualizing online conversations', *Communications of the ACM*, vol. 45, no. 4, pp. 45-9.
- Fei, B., Eloff, J.H., Olivier, M.S. & Venter, H.S. 2006, 'The use of self-organising maps for anomalous

- behaviour detection in a digital investigation', *Forensic Science International*, vol. 162, no. 1, pp. 33-7.
- Feiner, S.K. & Beshers, C. 1990, 'Worlds within worlds: Metaphors for exploring n-dimensional virtual worlds', *Proceedings of the 3rd annual ACM SIGGRAPH symposium on User interface software and technology*, ACM, pp. 76-83.
- Fischer, N., Neumann, P., Konevega, A.L., Bock, L.V., Ficner, R., Rodnina, M.V. & Stark, H. 2015, 'Structure of the E. coli ribosome-EF-Tu complex at < 3 Å resolution by Cs-corrected cryo-EM', *Nature*, vol. 520, no. 7548, pp. 567-70.
- Frank, E., Hall, M., Trigg, L., Holmes, G & Witten, I.H. 2004, 'Data mining in bioinformatics using Weka', *Bioinformatics*, vol. 20, no. 15, pp. 2479-81.
- Friedman, J.H. & Stuetzle, W. 1981, 'Projection pursuit regression', *Journal of the American statistical Association*, vol. 76, no. 376, pp. 817-23.
- Friedman, J.H. & Tukey, J.W. 1974, 'A projection pursuit algorithm for exploratory data analysis'.
- Garber, L. 2001, 'Encase: A case study in computer-forensic technology', *IEEE Computer Magazine January*.
- Gorban, A.N. & Zinovyev, A. 2010, 'Principal manifolds and graphs in practice: from molecular biology to dynamical systems', *International journal of neural systems*, vol. 20, no. 03, pp. 219-32.
- Guo, D. 2003, 'Coordinating computational and visual approaches for interactive feature selection and multivariate clustering', *Information Visualisation*, vol. 2, no. 4, pp. 232-46.
- Guo, P., Xiao, H., Wang, Z. & Yuan, X. 2010, 'Interactive local clustering operations for high dimensional data in parallel coordinates', *Visualisation Symposium (PacificVis), 2010 IEEE Pacific*, IEEE, pp. 97-104.
- Han, J., Kamber, M. & Pei, J. 2011, *Data mining: concepts and techniques*, Elsevier.
- Hartigan, J.A. & Wong, M.A. 1979, 'Algorithm AS 136: A k-means clustering algorithm', *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100-8.
- Hauck, R.V., Atabakhsb, H., Ongvasith, P., Gupta, H. & Chen, H. 2002, 'Using Coplink to analyze criminal-justice data', *Computer*, vol. 35, no. 3, pp. 30-7.
- Heer, J. & Card, S.K. 2004, 'DOITrees revisited: scalable, space-constrained visualisation of hierarchical data', *Proceedings of the working conference on Advanced visual interfaces*, ACM, pp. 421-4.
- Herman, I., Melançon, G & Marshould, M.S. 2000, 'Graph visualisation and navigation in information visualisation: A survey', *Visualisation and Computer Graphics, IEEE Transactions on*, vol. 6, no. 1, pp. 24-43.
- Hofmann, H., Siebes, A.P. & Wilhelm, A.F. 2000, 'Visualizing association rules with interactive mosaic plots', *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 227-35.
- Hoglund, A.J., Hatonen, K. & Sorvar, A.S. 2000, 'A computer host-based user anomaly detection system using the self-organizing map', *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, vol. 5, IEEE, pp. 411-6.
- Horn, M.S., Tobiasz, M. & Shen, C. 2009, 'Visualizing biodiversity with voronoi treemaps', *Voronoi Diagrams, 2009. ISVD'09. Sixth International Symposium on*, IEEE, pp. 265-70.
- Horn, R.D., Birdwell, J.D. & Leedy, L.W. 1997, 'Link discovery tool', *Proceedings of the Counterdrug*

Technology Assessment Center's ONDCP/CTAC International Symposium Chicago, IL.

- Huang, M.L., Huang, T.-H. & Zhang, X. 2016, 'A novel virtual node approach for interactive visual analytics of big datasets in parallel coordinates', *Future Generation Computer Systems*, vol. 55, pp. 510-23.
- Huang, M.L., Lu, L.F. & Zhang, X. 2015, 'Using arced axes in parallel coordinates geometry for high dimensional BigData visual analytics in cloud computing', *Computing*, vol. 97, no. 4, pp. 425-37.
- Huang, T.-H., Huang, M.L. & Jin, J.S. 2011, 'Parallel rough set: dimensionality reduction and feature discovery of Multi-dimensional data in visualisation', *Neural Information Processing*, Springer, pp. 99-108.
- Hurter, C., Tissoires, B. & Conversy, S. 2009, 'Fromdady: Spreading aircraft trajectories across views to support iterative queries', *Visualisation and Computer Graphics, IEEE Transactions on*, vol. 15, no. 6, pp. 1017-24.
- Hurter, C., Tissoires, B. & Conversy, S. 2010, 'Accumulation as a tool for efficient visualisation of geographical and temporal data', *AGILE 2010, 13th International Conference on Geographic Information Science*, p. pp xxx.
- Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M. & Rupp, R. 2007, 'Dendroscope: An interactive viewer for large phylogenetic trees', *BMC bioinformatics*, vol. 8, no. 1, p. 1.
- Ieong, R.S. 2006, 'FORZA—Digital forensic investigation framework that incorporate legal issues', *digital investigation*, vol. 3, pp. 29-36.
- Im, J.-F., McGuffin, M.J. & Leung, R. 2013, 'GPLOM: the generalized plot matrix for visualizing Multi-dimensional multivariate data', *Visualisation and Computer Graphics, IEEE Transactions on*, vol. 19, no. 12, pp. 2606-14.
- Inselberg, A. 1985, 'The plane with parallel coordinates', *The visual computer*, vol. 1, no. 2, pp. 69-91.
- Inselberg, A. 2009, *Parallel coordinates*, Springer.
- Inselberg, A. & Dimsdale, B. 1991, 'Parallel coordinates', *Human-Machine Interactive Systems*, Springer, pp. 199-233.
- Jewitt, C. & Van Leeuwen, T. 2001, *Handbook of visual analysis*, Sage Publications.
- Johansson, S. & Johansson, J. 2009, 'Interactive dimensionality reduction through user-defined combinations of quality metrics', *Visualisation and Computer Graphics, IEEE Transactions on*, vol. 15, no. 6, pp. 993-1000.
- Keim, D.A. 2000, 'Designing pixel-oriented visualisation techniques: Theory and applications', *Visualisation and Computer Graphics, IEEE Transactions on*, vol. 6, no. 1, pp. 59-78.
- Keim, D.A., Kohlhammer, J., Ellis, G & Mansmann, F. 2010, *Mastering the information age-solving problems with visual analytics*, Florian Mansmann.
- Khan, A., Wiil, U.K. & Memon, N. 2010, 'Digital forensics and crime investigation: legal issues in prosecution at national level', *Systematic Approaches to Digital Forensic Engineering (SADFE), 2010 Fifth IEEE International Workshop on*, IEEE, pp. 133-40.
- Khatri, P. & Drăghici, S. 2005, 'Ontological analysis of gene expression data: current tools, limitations, and open problems', *Bioinformatics*, vol. 21, no. 18, pp. 3587-95.
- Kohonen, T. 1990, 'The self-organizing map', *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464-80.

- Kreuseler, M. 2000, 'Visualisation of geographically related Multi-dimensional data in virtual 3D scenes', *Computers & Geosciences*, vol. 26, no. 1, pp. 101-8.
- Kreuseler, M., Lopez, N. & Schumann, H. 2000, 'A scalable framework for information visualisation', *Information Visualisation, 2000. InfoVis 2000. IEEE Symposium on*, IEEE, pp. 27-36.
- Kruskal, J.B. 1964, 'Multi-dimensional scaling by optimizing goodness of fit to a nonmetric hypothesis', *Psychometrika*, vol. 29, no. 1, pp. 1-27.
- Lamping, J. & Rao, R. 1994, 'Laying out and visualizing large trees using a hyperbolic space', *Proceedings of the 7th annual ACM symposium on User interface software and technology*, ACM, pp. 13-4.
- LeBlanc, J., Ward, M.O. & Wittels, N. 1990, 'Exploring n-dimensional databases', *Proceedings of the 1st conference on Visualisation'90*, IEEE Computer Society Press, pp. 230-7.
- Leung, C.K.-S. & Khan, Q.I. 2006, 'DSTree: a tree structure for the mining of frequent sets from data streams', *Data Mining, 2006. ICDM'06. Sixth International Conference on*, IEEE, pp. 928-32.
- Liang, J., N.Q.V., Simoff, S. and M.L.Huang. 2015, 'Tangram treemaps - A New Technique for Visualizing Large Hierarchies.', *Journal of Visual Languages and Computing*
- Lu, L.F., Huang, M.L. & Huang, T.-H. 2012, 'A new axes re-ordering method in parallel coordinates visualisation', *Machine Learning and Applications (ICMLA), 2012 11th International Conference On*, vol. 2, IEEE, pp. 252-7.
- Lu, L.F., Zhang, J.W., Huang, M.L. & Fu, L. 2010, 'A new concentric-circle visualisation of Multi-dimensional data and its application in network security', *Journal of Visual Languages & Computing*, vol. 21, no. 4, pp. 194-208.
- Maciejewski, R., Rudolph, S., Hafen, R., Abusalah, A., Yakout, M., Ouzzani, M., Cleveland, W.S., Grannis, S.J., Wade, M. & Ebert, D.S. 2008, 'Understanding syndromic hotspots-a visual analytics approach', *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*, IEEE, pp. 35-42.
- McQuaid, M.J., Ong, T.-H., Chen, H. & Nunamaker, J.F. 1999, 'Multi-dimensional scaling for group memory visualisation', *Decision support systems*, vol. 27, no. 1, pp. 163-76.
- Mohay, G 2005, 'Technical challenges and directions for digital forensics', *Systematic Approaches to Digital Forensic Engineering, 2005. First International Workshop on*, IEEE, pp. 155-61.
- Moore, R. 2010, *Cybercrime: Investigating high-technology computer crime*, Routledge.
- Morse, E., Lewis, M. & Olsen, K.A. 2000, 'Evaluating visualisations: using a taxonomic guide', *International Journal of Human-Computer Studies*, vol. 53, no. 5, pp. 637-62.
- Munzner, T. 2009, 'A nested model for visualisation design and validation', *Visualisation and Computer Graphics, IEEE Transactions on*, vol. 15, no. 6, pp. 921-8.
- Nguyen, Q.V. & Huang, M.L. 2002, 'A space optimised tree visualisation', *Information Visualisation, 2002. INFOVIS 2002. IEEE Symposium on*, IEEE, pp. 85-92.
- Nguyen, Q.V. & Huang, M.L. 2003a, 'Improvements of space optimised tree for visualizing and manipulating very large hierarchies', *Selected papers from the 2002 Pan-Sydney workshop on Visualisation-Volume 22*, Australian Computer Society, Inc., pp. 75-81.
- Nguyen, Q.V. & Huang, M.L. 2003b, 'Space optimised tree: a connection+ enclosure approach for the visualisation of large hierarchies', *Information Visualisation*, vol. 2, no. 1, pp. 3-15.

-
- Nguyen, Q.V. & Huang, M.L. 2005, 'EncCon: an approach to constructing interactive visualisation of large hierarchical data', *Information Visualisation*, vol. 4, no. 1, pp. 1-21.
- Nguyen, Q.V., Qian, Y., Huang, M. & Zhang, J. 2013, 'TabuVis: A tool for visual analytics Multi-dimensional datasets', *Science China Information Sciences*, vol. 56, no. 5, pp. 1-12.
- Nguyen, Q.V., Simoff, S. & Huang, M.L. 2014, 'Using Visual Cues on DOITree for Visualizing Large Hierarchical Data', *Information Visualisation (IV), 2014 18th International Conference on*, IEEE, pp. 1-6.
- Noblett, M.G., Pollitt, M.M. & Presley, L.A. 2000, 'Recovering and examining computer forensic evidence', *Forensic Science Communications*, vol. 2, no. 4, pp. 1-13.
- Øhrn, A. & Komorowski, J. 1997, 'Rosetta--a rough set toolkit for analysis of data', *Proc. Third International Joint Conference on Information Sciences*, Citeseer.
- Oja, H., Sirkiä, S. & Eriksson, J. 2006, 'Scatter matrix and independent component analysis', *Austrian Journal of Statistics*, vol. 35, no. 2, pp. 175-89.
- Olcott, P.L. 2006, 'Method and system for recognizing machine generated character glyphs and icons in graphic images', Google Patents.
- Pawlak, Z. 1995, 'Vagueness and uncertainty: a rough set perspective', *Computational intelligence*, vol. 11, no. 2, pp. 227-32.
- Pawlak, Z. 1998, 'Rough set theory and its applications to data analysis', *Cybernetics & Systems*, vol. 29, no. 7, pp. 661-88.
- Pawlak, Z. 2012, *Rough sets: Theoretical aspects of reasoning about data*, vol. 9, Springer Science & Business Media.
- Peng, W., Ward, M.O. & Rundensteiner, E.A. 2004, 'Clutter reduction in Multi-dimensional data visualisation using dimension reordering', *Information Visualisation, 2004. INFOVIS 2004. IEEE Symposium on*, IEEE, pp. 89-96.
- Person, K. 1901, 'On Lines and Planes of Closest Fit to System of Points in Space. Philosophical Magazine, 2, 559-572'.
- Pham, T., Metoyer, R., Bezrukova, K. & Spell, C. 2014, 'Visualisation of cluster structure and separation in multivariate mixed data: A case study of diversity faultlines in work teams', *Computers & Graphics*, vol. 38, pp. 117-30.
- Pirolli, P., Card, S.K. & Van Der Wege, M.M. 2003, 'The effects of information scent on visual search in the hyperbolic tree browser', *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 10, no. 1, pp. 20-53.
- Plaisant, C., Carr, D. & Shneiderman, B. 1995, 'Image-browser taxonomy and guidelines for designers', *Software, IEEE*, vol. 12, no. 2, pp. 21-32.
- Plaisant, C., Grosjean, J. & Bederson, B.B. 2002, 'Spacetree: Supporting exploration in large node link tree, design evolution and empirical evaluation', *Information Visualisation, 2002. INFOVIS 2002. IEEE Symposium on*, IEEE, pp. 57-64.
- Ramadas, M., Ostermann, S. & Tjaden, B. 2003, 'Detecting anomalous network traffic with self-organizing maps', *Recent Advances in Intrusion Detection*, Springer, pp. 36-54.
- Rauber, A., Merkl, D. & Dittenbach, M. 2002, 'The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data', *Neural Networks, IEEE Transactions on*, vol. 13,

- no. 6, pp. 1331-41.
- Read, J., Pfahringer, B., Holmes, G & Frank, E. 2011, 'Classifier chains for multi-label classification', *Machine learning*, vol. 85, no. 3, pp. 333-59.
- Robertson, GG, Mackinlay, J.D. & Card, S.K. 1991, 'Cone trees: animated 3D visualisations of hierarchical information', *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, pp. 189-94.
- Rogers, M.K. & Seigfried, K. 2004, 'The future of computer forensics: a needs analysis survey', *Computers & Security*, vol. 23, no. 1, pp. 12-6.
- Roweis, S.T. & Saul, L.K. 2000, 'Nonlinear dimensionality reduction by locally linear embedding', *Science*, vol. 290, no. 5500, pp. 2323-6.
- Sanftmann, H. & Weiskopf, D. 2012, '3D scatterplot navigation', *Visualisation and Computer Graphics, IEEE Transactions on*, vol. 18, no. 11, pp. 1969-78.
- Saraiya, P., North, C. & Duca, K. 2005, 'An insight-based methodology for evaluating bioinformatics visualisations', *Visualisation and Computer Graphics, IEEE Transactions on*, vol. 11, no. 4, pp. 443-56.
- Schaffer, D., Zuo, Z., Greenberg, S., Bartram, L., Dill, J., Dubs, S. & Roseman, M. 1996, 'Navigating hierarchically clustered networks through fisheye and full-zoom methods', *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 3, no. 2, pp. 162-88.
- Schein, A.I., Popescul, A., Ungar, L.H. & Pennock, D.M. 2002, 'Methods and metrics for cold-start recommendations', *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 253-60.
- Shetty, J. & Adibi, J. 2004, 'The Enron email dataset database schema and brief statistical report', *Information sciences institute technical report, University of Southern California*, vol. 4.
- Shiravi, H., Shiravi, A. & Ghorbani, A.A. 2012, 'A survey of visualisation systems for network security', *Visualisation and Computer Graphics, IEEE Transactions on*, vol. 18, no. 8, pp. 1313-29.
- Shneiderman, B. 1992, 'Tree visualisation with tree-maps: 2-d space-filling approach', *ACM Transactions on graphics (TOG)*, vol. 11, no. 1, pp. 92-9.
- Shneiderman, B. 1996, 'The eyes have it: A task by data type taxonomy for information visualisations', *Visual Languages, 1996. Proceedings., IEEE Symposium on*, IEEE, pp. 336-43.
- Shrinivasan, Y.B. & van Wijk, J.J. 2008, 'Supporting the analytical reasoning process in information visualisation', *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM, pp. 1237-46.
- Spearman, C. 1904, 'The proof and measurement of association between two things', *The American journal of psychology*, vol. 15, no. 1, pp. 72-101.
- Spence, B., Tweedie, L., Dawkes, H. & Su, H. 1995, 'Visualisation for functional design', *Information Visualisation, 1995. Proceedings., IEEE*, pp. 4-10.
- Stasko, J. & Zhang, E. 2000, 'Focus+ context display and navigation techniques for enhancing radial, space-filling hierarchy visualisations', *Information Visualisation, 2000. InfoVis 2000. IEEE Symposium on*, IEEE, pp. 57-65.
- Stone, M. 2006, 'Choosing colours for data visualisation', *Business Intelligence Network*.
- Swayne, D.F., Buja, A. & Hubbell, N. 1992, *XGobi meets S: Integrating software for data analysis*,

DTIC Document.

- Tarjan, R.E. 1983, 'Updating a balanced search tree in $O(1)$ rotations', *Information Processing Letters*, vol. 16, no. 5, pp. 253-7.
- Taylor, A.L., Hickey, T.J., Prinz, A.A. & Marder, E. 2006, 'Structure and visualisation of high-dimensional conductance spaces', *Journal of Neurophysiology*, vol. 96, no. 2, pp. 891-905.
- Teerlink, S. & Erbacher, R.F. 2006, 'Foundations for visual forensic analysis', *Information Assurance Workshop, 2006 IEEE*, IEEE, pp. 192-9.
- Theus, M. 2008, 'High-dimensional data visualisation', *Handbook of data visualisation*, Springer, pp. 151-78.
- Tominski, C., Schulze-Wollgast, P. & Schumann, H. 2005, '3d information visualisation for time dependent data on maps', *Information Visualisation, 2005. Proceedings. Ninth International Conference on*, IEEE, pp. 175-81.
- Tory, M. & Möller, T. 2004, 'Human factors in visualisation research', *Visualisation and Computer Graphics, IEEE Transactions on*, vol. 10, no. 1, pp. 72-84.
- Tsoumakas, G & Katakis, I. 2006, 'Multi-label classification: An overview', *Dept. of Informatics, Aristotle University of Thessaloniki, Greece*.
- Tsumoto, S. 2002, 'Accuracy and coverage in rough set rule induction', *Rough Sets and Current Trends in Computing*, Springer, pp. 373-80.
- Tusher, V.G, Tibshirani, R. & Chu, G 2001, 'Significance analysis of microarrays applied to the ionizing radiation response', *Proceedings of the National Academy of Sciences*, vol. 98, no. 9, pp. 5116-21.
- Van Wijk, J.J. 2002, 'Image based flow visualisation', *ACM Transactions on Graphics (TOG)*, vol. 21, ACM, pp. 745-54.
- Van Wijk, J.J. & Van Liere, R. 1993, 'HyperSlice: visualisation of scalar functions of many variables', *Proceedings of the 4th conference on Visualisation'93*, IEEE Computer Society, pp. 119-25.
- Vesanto, J. & Alhoniemi, E. 2000, 'Clustering of the self-organizing map', *Neural Networks, IEEE Transactions on*, vol. 11, no. 3, pp. 586-600.
- Ward, M.O. 1994, 'Xmdvtool: Integrating multiple methods for visualizing multivariate data', *Proceedings of the Conference on Visualisation'94*, IEEE Computer Society Press, pp. 326-33.
- Ward, M.O., Grinstein, G & Keim, D. 2010, *Interactive data visualisation: foundations, techniques, and applications*, CRC Press.
- Wilkinson, L. & Friendly, M. 2012, 'The history of the cluster heat map', *The American Statistician*.
- Yang, J., Peng, W., Ward, M.O. & Rundensteiner, E.A. 2003, 'Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets', *Information Visualisation, 2003. INFOVIS 2003. IEEE Symposium on*, IEEE, pp. 105-12.
- Yang, J., Ward, M.O. & Rundensteiner, E.A. 2002, 'Visual hierarchical dimension reduction for exploration of high dimensional datasets'.
- Yao, Y. & Zhao, Y. 2008, 'Attribute reduction in decision-theoretic rough set models', *Information sciences*, vol. 178, no. 17, pp. 3356-73.
- Yasinsac, A., Erbacher, R.F., Marks, D.G, Pollitt, M.M. & Sommer, P.M. 2003, 'Computer forensics education', *IEEE Security & Privacy*, no. 4, pp. 15-23.

- Yi, J.S., Kang, Y.-a., Stasko, J.T. & Jacko, J.A. 2008, 'Understanding and characterizing insights: how do people gain insights using information visualisation?', *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaluation methods for Information Visualisation*, ACM, p. 4.
- Zhou, M.X. & Feiner, S.K. 1998, 'Visual task characterization for automated visual discourse synthesis', *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press/Addison-Wesley Publishing Co., pp. 392-9.
- Ziarko, W. 1993, 'Variable precision rough set model', *Journal of computer and system sciences*, vol. 46, no. 1, pp. 39-59.