# ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

# ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

## ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

Διπλωματική Εργασία που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική.

Πειραιάς

Σεπτέμβριος 2019

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ΄ αρίθμ. ..... συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική.

Τα μέλη της Επιτροπής ήταν:

-Καθηγητής Μ. Κούτρας (Επιβλέπων)

-Αναπληρωτής Καθηγητής Ε. Κοφίδης

-Επίκουρος Καθηγητής Ν. Πελέκης

Η έγκριση της Διπλωματικής Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

# Προχωρημένες Τεχνικές Οπτικοποίησης σε Δεδομένα Μεγάλων Διαστάσεων

ΠΑΝΑΓΙΩΤΙΔΗΣ ΠΑΝΑΓΙΩΤΗΣ
ΜΕΣ 17022

Σεπτέμβριος 2019

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS
AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

# Advanced visualization techniques for high dimensional data

Panagiotidis Panagiotis

**MSc Dissertation**
**submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics**
**Piraeus,**
**September 2019**

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Μάρκο Κούτρα για την ουσιαστική βοήθεια και υποστήριξη που μου παρείχε ώστε να ολοκληρωθεί η εργασία αυτή, αλλά και οι σπουδές μου στο πρόγραμμα. Επίσης θα ήθελα να ευχαριστήσω ολόκληρο το διδακτικό προσωπικό του Τμήματος για την γνώση και εμπειρία που αποκόμισα συμμετέχοντας στο πρόγραμμα.

# Περιεχόμενα


1 Περίληψη 5

2 **Abstract** 6

2.1 Οπτικοποίηση Δεδομένων Μεγάλης Διάστασης . . . . . 7

2.2 Βασικός Σκοπός της Οπτικοποίησης . . . . . 7

2.3 Βασικά Προβλήματα της Οπτικοποίησης . . . . . 10

2.4 Ιστορική Αναδρομή . . . . . 16

3 Τεχνικές Μείωσης Διαστάσεων 19

3.1 Εισαγωγή . . . . . 19

3.2 Ορολογία . . . . . 20

3.3 Ανάλυση Κύριων Συνιστωσών . . . . . 25

3.4 Τοπικές Γραμμικές Συντεταγμένες . . . . . 27

3.5 Χάρτες του Sammon . . . . . 27

3.6 Χάρτες Αυτό-Οργάνωσης . . . . . 28

3.7 Πολυδιάστατη Κλιμάκωση . . . . . 30

3.8 ISOMAP . . . . . 32

3.9 Ιδιοχάρτες του Laplace . . . . . 33

3.10 Οπτικοποίηση Δεδομένων Χρησιμοποιώντας την t-SNE . . . . . 35

4 Τεχνικές Οπτικοποίησης 38

4.1 Εισαγωγή . . . . . 38

4.2 Γεωμετρικές Τεχνικές . . . . . 38

4.2.1 Πίνακες Διαγραμμάτων Διασποράς . . . . . 38

4.2.2 Πίνακας Προβολών . . . . . 39

4.2.3 Brusing . . . . . 41

4.2.4 Υπερθηκογράμματα . . . . . 42

4.2.5 Παράλληλες Συντεταγμένες . . . . . 43

4.2.6 Οι Καμπύλες του Andrews . . . . . 44

4.2.7 Προβολή στον Κύκλο (Circle Projection) . . . . . 45

4.2.8 Κυκλικοί Τομείς . . . . . 49

4.2.9 Τοπογραφικά Διαγράμματα . . . . . 53

4.2.10 Εξερευνητική Οπτικοποίηση . . . . . 54

4.2.11 RadViz . . . . . 59

4.2.12 Απεικονίσεις Βασισμένες σε Άξονες με Ακτινικές Διαστάσεις . . . . . 61

4.2.13 Φυσαλίδες . . . . . 70
4.2.14 Εικονική Πραγματικότητα - Σχηματισμοί Εδάφους . . . . . 70
4.3 Ιεραρχικές Τεχνικές . . . . . 72
4.3.1 Δενδρογράμματα . . . . . 72
4.3.2 Θερμοχάρτες . . . . . 73
4.3.3 Δακτυλίδια . . . . . 73
4.4 Εικονογραφικές Τεχνικές . . . . . 74
4.4.1 Πρόσωπα του Chernoff . . . . . 74
4.4.2 Εικονίδια Αστέρων . . . . . 76
4.4.3 Εικονίδια Χρώματος . . . . . 76
4.4.4 Ένα μοντέλο απεικόνισης με την χρήση κινουμένων σχεδίων . . . . . 77

5 Έφαρμογή σε Πραγματικά Δεδομένα 79
5.1 Περιγραφή των Συνόλου Δεδομένων . . . . . 79
5.2 Περιγραφική Στατιστική . . . . . 79
5.2.1 Κατεψυγμένα Ψάρια . . . . . 79
5.2.2 Αναψυκτικά . . . . . 81
5.2.3 Μαρμελάδες . . . . . 84
5.3 Οπτικοποίηση των Συνόλων Δεδομένων . . . . . 86
5.3.1 Εφαρμογή στο σύνολο δεδομένων Κατεψυγμένα Ψάρια . . . . . 86
5.3.2 Εφαρμογή στο σύνολο δεδομένων Αναψυκτικά . . . . . 90
5.3.3 Εφαρμογή στο σύνολο δεδομένων Μαρμελάδες . . . . . 95
5.3.4 Παρουσίαση Αποτελεσμάτων . . . . . 100

# 1 Περίληψη

Στην εποχή των μεγάλων δεδομένων, που ζούμε, υπάρχει η ανάγκη επεξεργασίας τεράστιων ποσοτήτων δεδομένων μεγάλης διάστασης. Ωστόσο, η μεγάλη διάσταση των δεδομένων, η περιορισμένη αντίληψη του ανθρώπινου νου και η δυσκολία της απεικόνισής τους στο επίπεδο ή στον χώρο αποτελούν μια μεγάλη πρόκληση για την ανάπτυξη αποτελεσματικών μεθόδων οπτικοποίησης. Οι κλασικές μέθοδοι οπτικοποίησης που έχουν εισαχθεί λειτουργούν ικανοποιητικά για δεδομένα χαμηλών διαστάσεων, αλλά αποτυγχάνουν για τα δεδομένα υψηλής διάστασης. Οι κλασικές μέθοδοι οπτικοποίησης είναι πολύ απλουστευτικές, δεν βοηθούν να αποκτήσουμε βαθύτερη εικόνα των δεδομένων ή διαπιστώνεται ότι είναι υπερβολικά δυσκίνητες και υπολογιστικά ανεπαρκείς.

Στην εργασία αυτή χρησιμοποιούνται γνωστά σύνολα δεδομένων με σκοπό να γίνεται μια σύγκριση των κλασικών μεθόδων οπτικοποίησης και επιπλέον να προταθούν νέες μέθοδοι οι οποίες είναι αποτελεσματικές για την οπτικοποίηση δεδομένων μεγάλης διάστασης. Επιπλέον, θα γίνει εφαρμογή των πιο αποτελεσματικών μεθόδων σε πραγματικά δεδομένα καταναλωτικών αγαθών ταχείας διακίνησης (Fast Moving Consumer Goods) και θα καταλήξουμε σε συμπεράσματα Η ανάπτυξη της τεχνολογίας με εργαλεία για την επεξεργασία των μεγάλων δεδομένων μπορεί να συμβάλει τα μέγιστα στην διαδικασία αυτήν.

# 2 Abstract

In the age of big data we live in, there is a need to process huge amounts of large-scale data. However, the large dimension of data, the limited perception of the human mind, and the difficulty of visualizing them at the level or space are a great challenge for the development of effective visualization methods. The standard visualization methods that have been introduced work well for low-dimensional data but fail for high-dimensional data. Classical visualization methods are very simplistic, do not help to obtain a deeper insights of the data or are found to be overly cumbersome and computationally inadequate.

This work uses well-known datasets to compare classic visualization methods and to propose new methods that are effective for visualizing large-dimensional data. Additionally, the most effective methods will be applied to real-world data on Fast Moving Consumer Goods and we will come to conclusions Developing technology with tools for processing large data can give a great push to this process.

## 2.1 Οπτικοποίηση Δεδομένων Μεγάλης Διάστασης

Τα δεδομένα μεγάλης διάστασης είναι δύσκολο να παρουσιαστούν με μια απλή απεικόνιση [12]. Είναι σημαντική πρόκληση η απεικόνιση να καταφέρει να βοηθήσει τους χρήστες να καταλάβουν διαισθητικά τα δεδομένα. Επιπλέον, το πώς θα παρουσιαστούν οι διαστάσεις σε ένα γράφημα έχει αντίκτυπο στην απεικόνιση επιτρέποντας εξαγωγή διαφορετικών κάθε φορά συμπερασμάτων [4]. Έτσι, θα πρέπει να εξετάσουμε το πώς θα παρουσιάζουμε τις διαστάσεις χωρίς μέχρι τώρα να υπάρχει κάποιος συγκεκριμένος τρόπος. Η απεικόνιση μπορεί να προσφέρει μια ποιοτική επισκόπηση των δεδομένων μεγάλης διάστασης ώστε οι χρήστες να μπορούν να διακρίνουν την δομή, τις τάσεις, τις συσχετίσεις και τα πρότυπα πιο αποτελεσματικά. Λόγω της μεγάλης διάστασης των δεδομένων δεν είναι εφικτό να παρουσιάσουμε τις λεπτομέρειες του κάθε χαρακτηριστικού [12]. Αυτή η κατάσταση ίσως δεν είναι ιδανική όταν θέλουμε να κάνουνε ποιοτική ανάλυση. Στα δεδομένα μεγάλης διάσταση υπάρχει πάντα ένα δίλημμα ανάμεσα στην ακριβή απεικόνιση και την απλή απεικόνιση.

Ο τελικός σκοπός της απεικόνισης είναι να αποκτήσουμε μια εικόνα της δομής των δεδομένων, αλλά και να δούμε πώς τα διάφορα χαρακτηριστικά συσχετίζονται μεταξύ τους. Στις περισσότερες περιπτώσεις ορισμένες συσχετίσεις δεν έχουν ανακαλυφθεί πριν την οπτικοποίηση των δεδομένων, αυτό είναι ακριβώς που θέλουμε να πετύχουμε με την οπτικοποίηση [9]. Δηλαδή, δεν γνωρίζουμε εκ των προτέρων ποια είναι η χρήσιμη πληροφορία που κρύβουν τα δεδομένα και ευελπιστούμε αυτό να το καταφέρουμε με την απεικόνιση. Έτσι, αν δεν ξέρουμε τίποτα για τα πρότυπα ή τις συσχετίσεις που θα παρουσιαστούν με την απεικόνιση των δεδομένων, είναι πολύ δύσκολο να αξιολογήσουμε την αποτελεσματικότητα μίας συγκεκριμένης τεχνικής, μπορούμε όμως να δούμε ποιες επιπλέον πληροφορίες για τα δεδομένα μας παρέχει η κάθε τεχνική.

## 2.2 Βασικός Σκοπός της Οπτικοποίησης

Ο αντικειμενικός σκοπός της οπτικοποίησης είναι να βρούμε πρότυπα και σχέσεις ανάμεσα στις μεταβλητές. Οι διάφορες ιδιότητες και τα χαρακτηριστικά των δεδομένων μπορεί να αλλάξουν τον τρόπο που πραγματοποιούμε την οπτικοποίηση, αλλά όχι τον στόχο αυτής [4].Τα παραδοσιακά διαγράμματα διασποράς και τα διαγράμματα που χρησιμοποιούν ευθείες είναι οι πλέον ευρέως χρησιμοποιούμενες τεχνικές για την οπτικοποίηση δεδομένων με μικρό αριθμό μεταβλητών. Αυτή η τεχνική μπορεί να ενισχυθεί τοποθετώντας επιπλέον διαγράμματα στην ίδια παρουσίαση, έτσι ώστε να μπορούν να παρουσιαστούν επιπλέον μεταβλητές [58]. Μπορούμε επίσης, να απεικονίσουμε τις μεταβλητές με διάφορα γραφικά πρότυπα με διαφορετικό χρώμα, σχήμα, μέγεθος και θέση (Βλέπε κεφάλαιο 3). Η απεικόνιση όλων των

διαστάσεων δημιουργεί κάποιου είδους πρότυπα τα οποία παρέχουν σημαντικές πληροφορίες για την επιστημονική έρευνα. Οι περισσότερες από τις τεχνικές οπτικοποίησης υποθέτουν έναν Ευκλείδειο χώρο [59]. Όμως, η συγκεκριμένη επιλογή δεν είναι πάντα η καλύτερη για την οπτικοποίηση των δεδομένων. Στην εργασία αυτή δίνονται κάποιοι εναλλακτικοί τρόποι. Μια ισχυρή μέθοδος οπτικοποίησης είναι η απεικόνιση των δεδομένων παράλληλα με μια μεταβλητή που εκφράζει τον χρόνο. Αυτή η προσέγγιση με την χρήση κινούμενων σχεδίων, αναλύεται στο Κεφάλαιο 3 [37].

Η γνωστική ψυχολογία θέτει τα θεμέλια για την ανθρώπινη αντίληψη, ειδικότερα, η θεωρία του **Gestalt** ασχολείται με τον τρόπο που το ανθρώπινο μυαλό συνδέει τα στοιχεία και καθορίζει αν αυτά ανήκουν σε ένα σύνολο ή λειτουργούν ανεξάρτητα [12]. Το έργο [12] εκτιμά την ικανότητα ενός ατόμου να αναγνωρίσει μια χρωματισμένη λέξη, ενώ οι **Lamers** και **Roelof** πρόσθεσαν μια ακόμα μεταβλητή, τον χώρο. Ο **Norman** εξήγησε την ανάγκη της οπτικοποίησης [64]. Ο ίδιος λέει ότι η δύναμη του μυαλού είναι εξαιρετικά υπερτιμημένη. Χωρίς εξωτερικές ενισχύσεις, η μνήμη, η σκέψη και η συλλογιστική είναι περιορισμένες. Αλλά η ανθρώπινη νοημοσύνη είναι εξαιρετικά ευέλικτη και προσαρμοστική, εξαιρετική στην επινόηση διαδικασιών και αντικειμένων που ξεπερνούν τα δικά της όρια [62]. Η πραγματική δύναμη προέρχεται από την εκπόνηση εξωτερικών βοηθημάτων ώστε να ενισχύσουν τις γνωστικές ικανότητες. Με την εύρεση εξωτερικών βοηθημάτων βρίσκουμε πράγματα που μας κάνουν ακόμα πιο έξυπνους. Με άλλα λόγια, ενώ το ανθρώπινο μυαλό είναι ικανό για σύνθετη ανάλυση, αυτό που υπερέχει είναι οι πληροφορίες που χρειάζονται ανάλυση να μετασχηματίζονται σε μορφή οικεία με τις γνωστικές ικανότητές μας, όπως αυτές που παρέχονται μέσω της απεικόνισης των δεδομένων. Η παρουσία δεδομένων μεγάλης διάστασης μπορεί να εμποδίσει την ικανότητα του χρήστη να ξεχωρίσει κάποια πρότυπα στα δεδομένα. Για να μπορέσουν να λύσουμε αυτό το πρόβλημα οι **Ji Soo Yi** και **Youn ah Kang** [41] προσφέρουν ένα νέο πλαίσιο για την καλύτερη μέτρηση της αξιολόγησης της αποτελεσματικότητας των υπαρχόντων τεχνικών αλληλεπίδρασης. Στην μελέτη τους, αποκαλύπτουν επτά κατηγορίες αλληλεπίδρασης: την επιλογή, την εξερεύνηση, την αναδιαμόρφωση, την κωδικοποίηση, την αφαίρεση/επεξεργασία, το φιλτράρισμα και την σύνδεση. Αυτές οι επτά κατηγορίες παρέχουν μια βάση, με την οποία μπορεί κανείς να κατανοήσει τους πιθανούς τρόπους που εμπλέκεται ένας χρήστης με τις διαδραστικές απεικονίσεις. Η κατανόηση αυτών των εννοιών θα επιτρέψει σε κάποιον να καταλάβει την αλληλεπίδραση σε χαμηλότερο επίπεδο και να καθιερώσει ένα κοινό λεξιλόγιο για να συζητήσει τον τρόπο που η αλληλεπίδραση ενισχύει την οπτικοποίηση δεδομένων. Η επιλογή επιτρέπει στους χρήστες να παρακολουθούν τα αντικείμενα ενδιαφέροντος και να εκτελούν πρόσθετες ενέργειες συμπεριλαμβανομένης της διαγραφής, της μετάφρασης και της εμφάνισης ενός επιλεγμένου στοιχείου [61]. Η εξερεύνηση επιτρέπει

σε έναν χρήστη να ανακαλύψει περισσότερες πληροφορίες που ενδεχομένως καλύπτονται από τους περιορισμούς της οθόνης. Η αναδιαμόρφωση μετασχηματίζει τα δεδομένα για να αποκαλύψει μια διαφορετική οπτική γωνία τους συγκρίνοντας τα δεδομένα δίπλα-δίπλα. Η κωδικοποίηση δεδομένων περιλαμβάνει την 'μετάφραση' των δεδομένων σε μια διαφορετική απεικόνιση, π.χ. οι τοπογραφικοί χάρτες χρησιμοποιούν μπλε για να αντιπροσωπεύουν το νερό και πράσινο για να αντιπροσωπεύουν περιοχές με βλάστηση. Η αφαίρεση απλοποιεί την αντιπροσώπευση δεδομένων ενώ η επεξεργασία αποκαλύπτει συγκεκριμένα χαρακτηριστικά ενός υποσυνόλου δεδομένων. Το φιλτράρισμα εμφανίζει δεδομένα με βάση τα φίλτρα που καθορίζονται από το χρήστη. Άλλα πλεονέκτημα του φιλτραρίσματος είναι ότι επιτρέπει ελαφρές οπτικές τροποποιήσεις στα 'φιλτραρισμένα' στοιχεία, όπως για τα στοιχεία που βρίσκονται κοντά ή διαφορετική απόχρωση/χρώμα για να απεικονισθεί μια χαλαρή σχέση [39]. Η σύνδεση υπογραμμίζει τις σχέσεις μεταξύ των δεδομένων, παρουσιάζοντας διαφορετικούς τρόπους με τους οποίους συσχετίζονται. Έχοντας πλέον καθιερώσει ένα κοινό λεξιλόγιο, μπορεί κανείς να ξεκινήσει την συζήτηση για τις επιπτώσεις της αλληλεπίδρασης στις απεικονίσεις. Η κάθε τύπου αλληλεπίδραση αυξάνει την αποτελεσματικότητα της ανάλυσης δεδομένων και επιτρέπει στους χρήστες να προβάλουν γρήγορα τα δεδομένα και επιλεκτικά ανάλογα με τον σκοπό της ανάλυσης. Ο Jared Schiffman [38] ισχυρίζεται ότι η απουσία ενός μέσου αλληλεπίδρασης με την απεικόνιση είναι αυτό που κάνει την απεικόνιση ελλιπή. Καθώς, ο κόσμος προσπαθεί να απεικονίσει πιο περίπλοκα δεδομένα, η αλληλεπίδραση υπογραμμίζει την κατευθυνόμενη από τον χρήστη χειραγώγηση των δεδομένων, ένα χαρακτηριστικό που είναι απαραίτητο για την αποκάλυψη της κρυμμένης πληροφορίας σε απεικονίσεις δεδομένων μεγάλης διάστασης.

## 2.3 Βασικά Προβλήματα της Οπτικοποίησης

Αυτά τα προβλήματα δεν επιβάλλονται απαραίτητα από τεχνικά εμπόδια. Ο **Bill Hibbard** [22] και ο **Chris Johnson** έχουν επισημάνει τα παρακάτω προβλήματα. Μετέπειτα και ο **Steve Eick** ασχολήθηκε με αυτό το θέμα. Υπάρχουν πολλά προβλήματα που εμποδίζουν μια πιο αποφασιστική πρόοδο της οπτικοποίησης, έτσι ώστε να καθιερωθεί στην θέση που της αξίζει στην διαδικασία μετάδοσης της πληροφορίας. Η οπτικοποίηση οδηγεί σε λάθη. Είναι αλήθεια ότι μια εσφαλμένη χρήση της οπτικοποίησης μπορεί να μας οδηγήσει σε σφάλματα με διάφορους τρόπους. Μερικές φορές επειδή η απεικόνιση μας δείχνει μια κατάσταση η οποία στην πραγματικότητα δεν υπάρχει. Αυτή είναι μια περίπτωση των γεωμετρικών λαθών, τα οποία μπορούμε να δούμε στο βιβλίο της **W.W.Rouse Ball** [33]. Συχνά η διαίσθησή μας, μας οδηγεί σε λάθος συμπεράσματα, επειδή το σχήμα προσεγγίζει αυτό που συμβαίνει πραγματικά. Σε ορισμένες άλλες περιπτώσεις, η οπτική κατάσταση μας παραπλανά ώστε να δεχθούμε ορισμένες συσχετίσεις που εμφανίζονται, χωρίς να θεωρούμε ότι πρέπει να τις δικαιολογήσουμε πιο αυστηρά. Αλλά η πιθανότητα ότι η απεικόνιση μπορεί να οδηγήσει σε λάθος συμπεράσματα δεν μπορεί να αποτελέσει επιχείρημα εναντίον της. Ακόμα και οι πιο διαδεδομένες τεχνικές είναι επιρρεπής σε σφάλματα, όπως σφάλματα στους υπολογισμούς, στην αιτιολόγηση και άλλα. Αυτό το γεγονός πρέπει να θεωρείται φυσιολογικό. Μια απεικόνιση για παράδειγμα, με καλά προσδιορισμένους κανόνες κωδικοποίησης και αποκωδικοποίησης είναι αυτό που ο κάθε χρήστης πρέπει να έχει στο μυαλό του. Η οπτικοποίηση είναι μια πνευματική διαδικασία που είναι άμεση και αβίαστη, αλλά μόνο για εκείνον που είναι επαρκώς προετοιμασμένος να την εκτελέσει αποτελεσματικά [39]. Αυτή η προετοιμασία υποδηλώνει μια εξοικείωση με το έργο της αποκωδικοποίησης της απεικόνισης. Όταν δεν υπάρχει αυτή η προετοιμασία αυτό που για άλλους μπορεί να είναι μια ευχάριστη και αβίαστη διαδικασία μπορεί να γίνει μια επίπονη και δυσνόητη διαδικασία. Είναι αλήθεια ότι μια εικόνα είναι χίλιες λέξεις, αλλά αυτό προϋποθέτει ότι η εικόνα είναι κατανοητή [24]. Διαφορετικά, δεν μας λέει τίποτα. Ο οδικός χάρτης, για παράδειγμα, δεν είναι η πραγματικότητα του τι αντιπροσωπεύει. Είναι απλά ένα σύνολο συμβόλων που πρέπει να μάθουμε να ερμηνεύουμε. Η σωστή απόδοση μίας απεικόνισης απαιτεί μια προηγούμενη προετοιμασία, μια εκπαίδευση που πολλοί αγνοούν. Οι ερευνητές, αν και χρησιμοποιούν συνεχώς την οπτικοποίηση φαίνεται να 'ντρέπονται' για την χρήση της. Η οπτικοποίηση είναι μια δυναμική διαδικασία. Πρέπει να προσπαθήσουμε να διδάξουμε ρητά την διαδικασία της οπτικοποίησης και να δώσουμε ιδιαίτερη προσοχή στους διάφορους τύπους απεικόνισης. Επιπλέον, είναι αναγκαίο να προσπαθήσουμε να έχουμε επίγνωση της διαδικασίας κωδικοποίησης και αποκωδικοποίησης που συνεπάγεται η πρακτική της απεικόνισης και η προσπάθεια να αποσαφηνιστεί. Θα πρέπει να έχουμε την οπτικοποίηση σε μεγάλη εκτίμηση όχι μόνο στη συχνή χρήση της, αλλά και για τους διαφορετικούς τρόπους που περιλαμβάνει

[25].

Η πλειονότητα των προσεγγίσεων που έχουν εισαχθεί για την οπτικοποίηση λειτουργούν καλά με σύνολα δεδομένων χαμηλών διαστάσεων, αλλά αποτυγχάνουν σε πολύπλοκα δεδομένα υψηλής διάστασης. Στην παρούσα εργασία θα γίνει συστηματική παρουσίαση των κύριων τεχνικών γραφικής παράστασης πολυδιάστατων δεδομένων, καθώς και συγκριτική παρουσίαση των πλεονεκτημάτων και των μειονεκτημάτων της καθεμιάς. Ο σκοπός της παρούσας εργασίας είναι να βρούμε ποιες είναι πιο αποτελεσματικές τεχνικές οπτικοποίησης για δεδομένα υψηλής διάστασης. Τέλος, θα γίνει εφαρμογή αυτών των τεχνικών σε πραγματικά δεδομένα υψηλής διάστασης από μια βιομηχανία Fast Moving Consumer Goods (FMCG).

Στην συνέχεια θα μιλήσουμε για τα βασικότερα προβλήματα της οπτικοποίησης τα οποία είναι η χρηστικότητα, η κατανόηση των στοιχειωδών αντιληπτικών-γνωστικών εργασιών, η κατάρτιση, τα ουσιαστικά μέτρα ποιότητας, η επεκτασιμότητα, η αισθητική, τα προβλήματα του λογισμικού οπτικοποίησης, η αντίληψη, οι διαστάσεις, οι δεξιότητες των αναγνωστών, οι δεξιότητες παρουσίασης και το γνωστικό στυλ.

1. Οι μελέτες χρηστικότητας πρέπει να εξετάζουν αν οι χρήστες μπορούν να αναγνωρίσουν τα πρότυπα που επιδιώκονται. Επειδή αυτό συνεπάγεται αλληλένδετα αντιληπτικά-γνωστικά καθήκοντα, οι υφιστάμενες μεθοδολογίες για εμπειρικές μελέτες ενδέχεται να μην είναι άμεσα εφαρμόσιμες. Το ζήτημα της χρηστικότητας είναι ζωτικής σημασίας για όλους. Αν και η ανάπτυξη της οπτικοποίησης επιταχύνεται, η αύξηση των μελετών της χρηστικότητας και οι εμπειρικές αξιολογήσεις γίνονται με σχετικά αργό ρυθμό. Επιπλέον, τα ζητήματα χρηστικότητας εξακολουθούν να αντιμετωπίζονται με έναν συγκεκριμένο τρόπο [26]. Η πολυπλοκότητα της υποκείμενης αναλυτικής διαδικασίας που εμπλέκεται στα περισσότερα συστήματα οπτικοποίησης αποτελεί μείζον εμπόδιο. Οι τελικοί χρήστες δεν μπορούν να δουν πώς τα ακατέργαστα δεδομένα τους μετατρέπονται 'μαγικά' (για αυτούς) σε πολύχρωμες εικόνες. Η πρώτη συλλογή εμπειρικών μελετών έγινε το 1991 [7]. Αν και ο αριθμός των εμπειρικών μελετών των συστημάτων απεικόνισης αυξάνεται, οι σχεδιαστές και οι χρήστες πρέπει να βρουν εμπειρικές ενδείξεις τόσο γενικές όσο και αρκετά συγκεκριμένες για να ενημερωθούν για τις διαδικασίες λήψης αποφάσεων. Οι εμπειρικές μελέτες τείνουν να χρησιμοποιούν συστήματα ανοιχτού κώδικα και ελεύθερα διαθέσιμα [29]. Μια παρατεταμένη έλλειψη συστημάτων χαμηλού κόστους, έτοιμων προς χρήση και αναδιαμορφώσιμων συστημάτων απεικόνισης θα έχει αρνητικές επιπτώσεις στην καλλιέργεια της κρίσης του πληθυσμού των χρηστών. Ένα ισορροπημένο χαρτοφυλάκιο συστημάτων γενικής χρήσης και πλήρους λειτουργικής απεικόνισης πληροφοριών είναι βασικό για τις προοπτικές που βασίζονται σε χρήστες. Χρειαζόμαστε νέες μεθοδολογίες αξιολόγησης. Η πλειοψηφία των υφιστάμενων μελετών χρηστικότητας βασίστηκε σε μεθοδολογίες που προϋπήρχαν των τε-

χνικών απεικόνισης. Μπορεί όμως να υπάρχει ακόμη πιο σοβαρός λόγος για την έλλειψη μελετών χρηστικότητας. Η οπτικοποίηση είναι ένα εργαλείο οπτικής εξερεύνησης που επιτρέπει στο χρήστη να αλληλεπιδρά με το οπτικοποιημένο περιεχόμενο και να κατανοεί το νόημά του. Η διαδικασία εξερεύνησης είναι συχνά διερευνητική. Για παράδειγμα, οι χρήστες μπορούν να αλληλεπιδρούν με πολλούς πιθανούς τρόπους με την απεικόνιση και να ερμηνεύουν αυτό που βλέπουν [30].

2. Εργασίες όπως η περιήγηση και η αναζήτηση, απαιτούν ένα επίπεδο γνωστικών δραστηριοτήτων υψηλότερο από εκείνο της αναγνώρισης και αποκωδικοποίησης των οπτικοποιημένων αντικειμένων. Υπό αυτή την έννοια, υπάρχει αναντιστοιχία μεταξύ των χρηστών και της αξιολόγησης της χρησιμότητας των στοιχείων της απεικόνισης. Μελέτες στοιχειωδών αντιληπτικών-γνωστικών καθηκόντων εμφανίζονται στην βιβλιογραφία της ψυχολογίας και της στατιστικής, συμπεριλαμβανομένης της μελέτης των Cleveland-McGill [4] και του έργου του Treisman σε αντιληπτικά καθήκοντα. Στο πλαίσιο της απεικόνισης των πληροφοριών, οι ερευνητές έχουν κάνει σημαντική δουλειά, κυρίως μέσω του έργου του Ware στο χαρακτηρισμό της κίνησης και της αντίληψης του βάθους στην απεικόνιση, αλλά έχουμε όμως πολλά ακόμα να επιτύχουμε [11].

3. Το πρόβλημα της εκπαίδευσης είναι το τρίτο πρόβλημα που αντιμετωπίζουμε, το οποίο έχει επίκεντρο τον χρήστη. Αντιμετωπίζουμε αυτήν την πρόκληση εσωτερικά και εξωτερικά. Η εσωτερική πτυχή της πρόκλησης αυτής αναφέρεται στην ανάγκη για ερευνητές και επαγγελματίες στον τομέα της οπτικοποίησης για να μαθαίνουν και να μοιράζονται διάφορες αρχές και δεξιότητες της οπτικής επικοινωνίας. Η γλώσσα της απεικόνισης πρέπει να καταστεί κατανοητή στους πιθανούς χρήστες της [45]. Η τακτική αναθεώρηση των υπαρχουσών αρχών υπό το φως των νέων συστημάτων, πρέπει να είναι προϋπόθεση. Η εξωτερική πλευρά της πρόκλησης αναφέρεται στην ανάγκη οι δυνητικοί χρήστες εκτός του άμεσου οπτικού αποτελέσματος να δουν την αξία της απεικόνισης των πληροφοριών και πώς θα μπορούσε να συμβάλει στην εργασία τους με καινοτόμο τρόπο. Χρειαζόμαστε συναρπαστικά παραδείγματα επίδειξης για να αυξήσουμε τη συνειδητοποίηση του δυναμικού της οπτικοποίησης. Σε αυτήν την κατεύθυνση θα συμβάλει και η εισαγωγή αξιόπιστων τεχνικών αξιολόγησης των τεχνικών οπτικοποίησης.

4. Η αξιολόγηση των τεχνικών οπτικοποίησης είναι ένα ιδιαίτερα δύσκολο πρόβλημα, διότι οι ουσιαστικές μετρήσεις ποιότητας θα απαντήσουν σε βασικά ερωτήματα όπως, σε ποιο βαθμό ο σχεδιασμός της απεικόνισης αντιπροσωπεύει τα υποκείμενα δεδομένα πιστά και αποτελεσματικά και σε ποιο βαθμό διατηρούνται οι εγγενείς ιδιότητες του υποκείμενου

φαινομένου. Αυτά τα μέτρα θα μας βοηθήσουν επίσης και στην επίλυση ενός μεγάλου προβλήματος, το οποίο είναι η επιλογή τεχνικής οπτικοποίησης. Είναι ζωτικής σημασίας για το πεδίο της οπτικοποίησης να καθοριστούν ουσιαστικά μέτρα ποιότητας. Μέχρι πρόσφατα, η έλλειψη μετρήσιμων ποιοτικών τεχνικών δεν προκαλούσε μεγάλη ανησυχία [48]. Ωστόσο, η έλλειψη μετρήσιμων μετρήσεων ποιότητας και σημείων αναφοράς θα υπονομεύσει την πρόοδο της οπτικοποίησης, ιδίως την αξιολόγηση και την επιλογή της. Μια ουσιαστική μέτρηση ποιότητας θα απλοποιήσει τρομακτικά την ανάπτυξη και την αξιολόγηση διαφόρων αλγορίθμων. Η παροχή τέτοιων μέτρων ποιότητας θα επιτρέψει τη διεξαγωγή μελετών χρηστικότητας για την αξιολόγηση της συνέπειας μεταξύ της βέλτιστης λύσης που βασίζεται στις αξιολογήσεις των χρηστών και της βέλτιστης λύσης που βασίζεται στα μέτρα αυτά.

5. Ένα σχετικά πρόσφατο ερευνητικό ενδιαφέρον επικεντρώνεται στην απεικόνιση των ροών δεδομένων [38]. Η πρόκληση της απεικόνισης ροών δεδομένων οφείλεται στο ρυθμό άφιξης της ροής δεδομένων και στην επείγουσα ανάγκη να κατανοηθεί το περιεχόμενό της.

6. Ο σκοπός της απεικόνισης είναι οι πληροφορίες που μας παρέχει σχετικά με τα δεδομένα, όχι μόνο οι όμορφες εικόνες. Αλλά τι κάνει μια εικόνα όμορφη; Τι μπορούμε να μάθουμε από το να κάνουμε μια εικόνα όμορφη και να ενισχύσουμε την αναπαράσταση των πληροφοριών; Από τα παραπάνω καταλαβαίνουμε ότι είναι σημαντικό να κατανοήσουμε πώς αλληλεπιδρούν οι πληροφορίες με την αισθητική. Υπάρχει έλλειψη ολιστικών εμπειρικών μελετών για να χαρακτηρίσουμε ποιες οπτικές ιδιότητες κάνουν τους χρήστες να πιστεύουν ότι ένα γράφημα είναι όμορφο ή οπτικά ελκυστικό [40]. Η βελτίωση της αισθητικής στην απεικόνιση, με σκοπό την εξαγωγή της πληροφορίας παραμένει μια πρόκληση.

7. Όταν δημιουργούμε οπτικοποιήσεις κάνουμε τη φυσική υπόθεση ότι αυτές οι εικόνες που χρησιμοποιήσαμε μπορούν να βοηθήσουν τον άνθρωπο να μάθει πράγματα. Είναι διαισθητικά προφανές ότι οι άνθρωποι μπορούν να αντλήσουν σημαντικές πληροφορίες από τις απεικονίσεις. Είναι επίσης λογικό να πούμε ότι είναι πιο κατανοητά ορισμένα είδη πληροφοριών από κάποιες απεικονίσεις παρά από άλλες. Δεν είναι καθόλου σαφές, ωστόσο, τι κάνει ο εγκέφαλος με το φως που πέφτει στους αμφιβληστροειδείς μας για να δημιουργηθεί αυτή η πληροφορία. Αν το γνωρίζαμε αυτό, τόσο η αιτιολόγηση όσο και η έρευνα για την οπτικοποίηση θα ήταν πιο συγκεκριμένη. Τα προβλήματα της αντίληψης ήταν πάντα από τα κεντρικά ερευνητικά θέματα της γνωστικής ψυχολογίας και αυτή η έρευνα έχει οδηγήσει σε κάποιες λογικά καθιερωμένες αρχές. Οι ειδικοί για τους ανθρώπινους παράγοντες όπως ο Wickens έχουν επισημάνει χρόνια πριν ότι αυτή η έρευνα είναι σχετική με την απεικόνιση των πληροφοριών και υπάρχουν μια πλειάδα χρήσιμων αποτελεσμάτων που είναι σχετικές

με το λογισμικό της απεικόνισης. Κάποιες από τις θεωρίες που χρησιμοποιούμε τώρα είναι 100 χρόνων και αξίζει να συζητηθεί κατά πόσο οι πιο πρόσφατες γνωστικές θεωρίες της αντίληψης έχουν ενσωματωθεί με τα λογισμικά οπτικοποίησης.

Ο δέκατος ένατος αιώνας αποτέλεσε το θεμέλιο της πειραματικής ψυχολογίας, αντιμετωπίζοντας τέτοιες προκλήσεις, όπως το πόσο διαφορετικά πρέπει να είναι δύο επίπεδα φωτεινότητας για να τα αντιληφθούμε ως διαφορετικά. Το ανθρώπινο οπτικό σύστημα έχει κάποια θεμελιώδη όρια, έτσι ώστε οι ειδικοί να μπορούν να παράγουν εμπειρικούς πίνακες για να μας καθοδηγήσουν σε τέτοια θέματα όπως η επιλογή των χρωμάτων ή το μέγεθος των εικονοστοιχείων στην απεικόνιση στην οθόνη του υπολογιστή. Αυτές οι πρώιμες πειραματικές παρατηρήσεις βασίζονται πλέον σε σταθερό θεωρητικό υπόβαθρο που προέρχεται από την έρευνα [48]. Ως αποτέλεσμα, μπορούμε να συζητήσουμε την αντίληψη του χρώματος ή την ανίχνευση των χαρακτηριστικών μέσω των χαρακτηριστικών απόκρισης του αμφιβληστροειδή και των οπτικών νεύρων. Αυτές οι διαδικασίες χαμηλού επιπέδου ενσωματώθηκαν από τον **Marr** σε μια θεώρηση της αντίληψης της απεικόνισης, με τρόπο που είναι πολύ προσβάσιμες στους επιστήμονες των υπολογιστών [42]. Η θεωρία του **Marr** επικεντρώνεται στην αντίληψη των τρισδιάστατων πραγμάτων, αλλά αυτή η περιγραφή του τρόπου που η τρισδιάστατη αυτή πληροφορία διαχωρίζεται από δισδιάστατα οπτικά πεδία μπορεί να εφαρμοστεί για να περιγράψει, για παράδειγμα, το τρόπο που η υφή της οθόνης του υπολογιστή συμβάλλει στην ερμηνεία των αντικειμένων που απεικονίζονται. Η θεωρία του **Ullman** [42] για την ερμηνεία των συνόρων είναι επίσης διαθέσιμη για την απεικόνιση στο επίπεδο.

Η θεωρία του **Marr** αντιμετωπίζει τις γνωστικές λειτουργίες σαν να ήταν μια σειρά φίλτρων που λειτουργούν με τα ¨ακατέργαστα δεδομένα' που φτάνουν στα μάτια μας, που τελικά μετατρέπονται σε πληροφορίες. Η κατάσταση είναι μάλλον πιο περίπλοκη, ωστόσο ο πραγματικός κόσμος δεν φτάνει στα οπτικά μας νεύρα σαν να ήταν φωτογραφία. Η θεωρία του **Gibson** της οικολογικής οπτικής, που αργότερα εκπονήθηκε από τον **Neisser** [20], περιγράφει τον τρόπο με τον οποίο στρέφουμε την προσοχή μας στο περιβάλλον γύρω μας, δημιουργώντας έναν γνωστικό χάρτη τον οποίο χρησιμοποιούμε για να αλληλεπιδράσουμε με τον κόσμο. Οι θεωρητικοί της **Gestalt** από τις αρχές του 20ου αιώνα έδωσαν μια περιγραφή της αντιληπτικής θεωρίας που έχει αμφισβητηθεί και είναι πολύ σχετική με την απεικόνιση. Περιέγραψαν τον τρόπο με τον οποίο μετασχηματίζουμε την αντίληψή μας ώστε αυτή να καθίσταται ενιαία και συνεκτική [40]. Για να γίνει αυτό και στην οπτικοποίηση πρέπει να διατυπωθεί αυτή η θεωρία με τρόπο ώστε να οργανωθούν κανόνες που θα εφαρμόζουμε κατά την αντιληπτική διαδικασία. Αυτοί οι κανόνες μπορεί να χρησιμοποιηθούν για να εξηγήσουν γιατί κάποιες απεικονίσεις μπορούν να δουλέψουν καλύτερα από άλλες. Μας εξηγούν τον τρόπο που θα πρέπει να ερμηνεύουμε τις συνεχείς γραμμές, το κλείσιμο των ορίων, την εγγύτητα κ.τ.λ..

8. Οι οπτικοποιήσεις του λογισμικού είναι ένα πλαίσιο επικοινωνίας μεταξύ του σχεδιαστή και του χρήστη της οπτικοποίησης. Ποια είναι η διαφορά μεταξύ του κειμένου του πηγαίου κώδικα για ένα λογισμικό και μίας απεικόνισης του λογισμικού; Πολλοί επιστήμονες επισημαίνουν μια θεμελιώδη διαφορά στο περιεχόμενο μεταξύ των δύο [49]. Το κείμενο είναι γραμμικό και μίας διάστασης, ενώ οι εικόνες είναι δύο διαστάσεων. Στις κινούμενες εικόνες προστίθεται ακόμα μια διάσταση η οποία είναι ο χρόνος. Στην περίπτωση της εικονικής πραγματικότητας προσθέτονται παραπάνω διαστάσεις. Οι επιστήμονες ισχυρίζονται ότι το πληροφοριακό περιεχόμενο αυξάνει εκθετικά τις χρησιμοποιούμενες διαστάσεις και αυτό είναι αληθές. Άρα εγείρεται το ερώτημα αν αυτός ο όγκος πληροφοριών μπορεί να αναπαρασταθεί από ένα λογισμικό. Το πρώτο εδώ που πρέπει να σημειωθεί είναι ότι ακόμη και το κείμενο δεν είναι πραγματικά μίας διάστασης. Ένας μεταγλωττιστής μπορεί να διαβάσει έναν πηγαίο κώδικα χρησιμοποιώντας μια γραμμική συνάρτηση, αλλά οι προγραμματιστές δεν γράφουν με αυτόν τον τρόπο, γράφουν λίγο λίγο πηγαίνοντας πίσω και βλέποντας τι έχουν κάνει και τον τροποποιούν μέχρι να φτάσουν στον σωστό κώδικα [11]. Αφού ολοκληρωθεί αυτή η δουλειά, οι αναγνώστες λαμβάνουν υπ όψιν, τις εσοχές, τα κατακόρυφα και οριζόντια σχέδια, την κατανομή του λευκού χώρου και άλλα. Σύμφωνα με τον Bertin [62], η τέχνη της παρουσίασης των πληροφοριών βρίσκεται σε μεγάλο βαθμό στο να αναθέσουμε στις διαθέσιμες διαστάσεις τις ανεξάρτητες κατηγορίες των πληροφοριών που πρόκειται να παρουσιαστούν. Ο Bertin θεωρεί ότι σε μια απεικόνιση μπορούν να χρησιμοποιηθούν μόνο οκτώ μεταβλητές. Ορισμένοι χρησιμοποιούν επίσης διαφορετικά χρώματα ή πυκνότητα για να αυξήσουν τις διαστάσεις που μπορούν να απεικονισθούν. Οι τεχνολογίες απεικόνισης μπορούν να ενσωματώσουν περισσότερες πληροφορίες παρουσιάζοντας την τρίτη ή την τέταρτη διάσταση, αλλά η κάθε μια έρχεται με αρκετό κόστος, έτσι ώστε να πρέπει να είμαστε πολύ προσεκτικοί για το πώς κατανέμουμε τις μεταβλητές στις διαστάσεις. Το κόστος αυτό είναι τόσο υπολογιστικό όσο και αντιληπτικό, η τρίτη διάσταση της αντίληψης είναι περιοριστική στο ότι εμείς μπορούμε να δούμε μόνο μια πλευρά των τρισδιάστατων αντικειμένων και η διάσταση του χρόνου ακόμα περισσότερο επειδή η αντίληψη των χρονικών γεγονότων είναι αναγκαστικά παροδική. Πόσο ενοχλητικό είναι ότι τα λογισμικά είναι πολύπλοκα και ότι υπάρχουν εκατοντάδες ανεξάρτητες μεταβλητές που πρέπει να ξεγελάσουμε στο κεφάλι μας και να μπορέσουμε να τις εκχωρήσουμε σε μια από τις διαθέσιμες διαστάσεις ώστε για να γίνουν ορατές. Ο σχεδιασμός της απεικόνισης είναι η προσπάθεια να προσαρμόσουμε πολλές μεταβλητές σε μερικές διαστάσεις. Υπάρχουν πολύ περισσότερες μεταβλητές από τις διαστάσεις, οπότε πρέπει να γίνει μια μείωση των διαστάσεων πριν την απεικόνιση. Θα πρέπει να αξιοποιήσουμε όλες τις διαστάσεις της ανθρώπινης αντίληψης και να χρησιμοποιηθούν από κοινού με την τεχνολογία. Οι πληροφορίες που είναι διαθέσιμες σε μια κατάσταση τελικά

εξαρτώνται από το πώς η κατάσταση αποκωδικοποιείται από τον παρατηρητή [30].

9. Οδηγεί η απεικόνιση σε αποτελεσματική ερμηνεία; Η οπτικοποίηση δεν εγγυάται την αποτελεσματική ερμηνεία, ειδικά για τους αρχάριους. Η ερμηνεία των απεικονίσεων είναι μια ικανότητα που κατακτάται [22]. Ο Arnheim κάνει αυτήν την συζήτηση στο θέμα των πινάκων ζωγραφικής, το οποίο είναι πολύ πιο δύσκολο όταν μιλάμε για λογισμικά απεικόνισης [51]. Οι αρχάριοι αναγνώστες συνήθως στερούνται στρατηγικών αναζήτησης και την προσοχή τους την αποσπούν τα επιφανειακά χαρακτηριστικά. Οι ειδικοί είναι πιο εξοικειωκένοι με τους τρόπους με τους οποίους οι βασικές δομές έρχονται στην επιφάνεια και μπορούν καλύτερα να αναγνωρίσουν τι είναι σημαντικό και τι ασήμαντο. Η οπτικοποίηση εξαρτάται κατά ένα μέρος από το μάτι του θεατή. Οι ειδικοί διακρίνονται από την αποκτηθείσα ικανότητά τους να βλέπουν ποιες πληροφορίες είναι σημαντικές, δηλαδή σχετίζονται με το θέμα και επιλέγουν τι πρέπει να δουν και τι να αγνοήσουν.

10. Η ελπίδα στην οπτικοποίηση είναι ότι η δεξιοτεχνία στην παρουσίαση μπορεί να αντισταθμίσει την εμπειρία του αναγνώστη. Αυτό συνεπάγεται τον εντοπισμό και την καταγραφή του τι είναι πιο σημαντικό και την καθοδήγηση του αναγνώστη με τον συνδυασμό των αντιληπτικών στοιχείων με αυτές τις προσεκτικά επιλεγμένες πληροφορίες. Αυτό σημαίνει την παροχή στήριξης στον αναγνώστη για την απόκτηση των δεξιοτήτων.

11. Ένα αξίωμα στην ψυχολογία είναι ότι τα άτομα διαφέρουν. Τα αποτελέσματα μπορεί να διαφέρουν ανάμεσα σε διαφορετικούς χρήστες, όπως ανάμεσα στις διάφορες οπτικοποιήσεις. Ξέροντας τι να περιμένουμε, που να κοιτάξουμε και τι να δούμε, μπορούμε πιο εύκολα να εξάγουμε χρήσιμες πληροφορίες. Προφανώς, η αυξανόμενη τεχνολογία αντικατοπτρίζει τις αλλαγές της αντιληπτικής στρατηγικής. Ως εκ τούτου, η απεικόνιση είναι έντονα ευάλωτη στις αδυναμίες των επιμέρους αντιληπτικών και ερμηνευτικών ικανοτήτων του κάθε χρήστη.

## 2.4 Ιστορική Αναδρομή

Τα τελευταία χρόνια η ανάπτυξη της οπτικοποίησης των δεδομένων μεγάλης διάστασης μπορεί να χωριστεί σε τέσσερα στάδια, τα οποία περιλαμβάνουν τις χρονολογικές περιόδους: πριν το 1976, από το 1977 έως το 1985, από το 1986 έως το 1991 και από το 1992 και μετέπειτα.

1. Το Στάδιο της Αναζήτησης

Το στάδιο αυτό αφορούσε την γραφική αναπαράσταση δεδομένων ενός ή δύο

χαρακτηριστικών, για την ανάλυση δεδομένων με κύριο εκφραστή της τον Tukey [12]. Οι επιστήμονες από το 1782 είχαν μελετήσει την οπτικοποίηση δεδομένων μεγάλης διάστασης, όταν ο Crome χρησιμοποίησε σημειο-σύμβολα για να δείξει την γεωγραφική κατανομή 56 προϊόντων στην Ευρώπη. Το 1950, ο Gibson ξεκίνησε την έρευνα για την υφή της οπτικής αντίληψης [23]. Αργότερα, οι Picket και White πρότειναν την απεικόνιση των δεδομένων χρησιμοποιώντας γραφικά αντικείμενα τα οποία θα αποτελούνται από γραμμές [42]. Αυτή η τεχνική διερευνήθηκε περαιτέρω από τον Picket και υλοποιήθηκε με υπολογιστή. Ο Chernoff το 1973 παρουσίασε τις συστοιχίες από πρόσωπα με την μορφή cartoon για δεδομένα μεγάλης διάστασης [21]. Σε αυτήν την πολύ γνωστή τεχνική οι μεταβλητές απεικονίζονται στο σχήμα προσώπων κινουμένων σχεδίων και στα χαρακτηριστικά τους όπως η μύτη, τα μάτια και το στόμα. Τα πρόσωπα αυτά στην συνέχεια τοποθετούνται στο δισδιάστο επίπεδο. Η έρευνα στο στάδιο αυτό χαρακτηρίζεται από την ενασχόλησή της με δεδομένα με λίγες παρατηρήσεις και εργαλεία όπως το χαρτί και οι μαρκαδόροι διαφόρων χρωμάτων. Οι στατιστικοί ήταν η κυριότερη ομάδα ερευνητών που καταπιανόταν με αυτά τα θέματα την συγκεκριμένη εποχή. Η απεικόνιση χρησιμοποιήθηκε για να αναδείξει τα βασικά χαρακτηριστικά των δεδομένων, προτείνοντας στατιστικές μεθόδους ανάλυσής τους και παρουσίασης των συμπερασμάτων.

2. Το Στάδιο της Αφύπνισης

Στο στάδιο αυτό κυριάρχησε η ανάλυση δεδομένων με τη λογική του Tukey. Η διερευνητική ανάλυση δεδομένων του σηματοδότησε μια νέα εποχή στην επιστημονική απεικόνιση των δεδομένων. Η διερευνητική ανάλυση δεδομένων είναι κάτι παραπάνω από ένα εργαλείο, είναι ένας τρόπος σκέψης. Διδάσκει τους ανθρώπους πώς να αποκωδικοποιούν τις οπτικές πληροφορίες από τα δεδομένα. Όταν ο επιστήμονας απέκτησε τον προσωπικό του υπολογιστή, η ανάλυση του Tukey έγινε το πιο ισχυρό εργαλείο στα χέρια του. Τώρα οι επιστήμονες μπορούν να οπτικοποιήσουν δεδομένα πέρα από τις δύο διαστάσεις. Ο τεράστιος όγκος υπολογισμών, που μέχρι τώρα γινόταν με το χέρι, μπορούσε τώρα να γίνει σε πραγματικό χρόνο. Οι στατιστικοί μπορούν να απεικονίζουν τα δεδομένα σε κάθε στάδιο της ανάλυσης αντί να περιμένουν να απεικονίσουν τα τελικά συμπεράσματα. Η διαθεσιμότητα λογισμικού για υπολογιστές όπως αυτό για υψηλής ευκρίνειας χρώματα έδωσε στην έρευνα της οπτικοποίησης νέες δυνατότητες. Κατά το στάδιο αυτό μελετήθηκαν δεδομένα δύο ή τριών διαστάσεων.

Ακόμα αυτή την περίοδο τα δεδομένα μεγάλης διάστασης χαίρουν μεγαλύτερης προσοχής. Ο **Asimov** παρουσίασε την **Grand Tour Technique** για την απεικόνιση δεδομένων μεγάλης διάστασης στις δύο διαστάσεις [3]. Οι δορυφόροι που τέθηκαν σε τροχιά γύρω από την γη στέλνουν συνεχώς δεδομένα και έτσι έχει φτάσει σε εμάς ένας τεράστιος όγκος δεδομένων μεγάλης διάστασης.

3. Το Στάδιο της Ανακάλυψης

Στο στάδιο αυτό οι επιστήμονες άρχισαν να ψάχνουν γραφικές μεθόδους οπτικοποίησης μέσα από ένα διαφορετικό πρίσμα. Αν και ακόμα οι γραφικές μέθοδοι ήταν σε δύο διαστάσεις, οι επιστήμονες ήταν σε θέση να κωδικοποιήσουν δεδομένα με πολλά χαρακτηριστικά με διαγράμματα στις δύο διαστάσεις τα οποία ήταν εξαιρετικά αποτελεσματικά. Ακόμα στο στάδιο αυτό αναγνωρίστηκε η αξία της οπτικοποίησης. Το 1987 έγινε ξεκάθαρη η ανάγκη για απεικόνιση στις δύο και στις τρεις διαστάσεις [57]. Η απεικόνιση στις δύο διαστάσεις δεδομένων μεγάλης διάστασης περιλαμβάνεται ως βραχυπρόθεσμος στόχος της έρευνας. Αφού έχει οριστεί ο στόχος, οι επιστήμονες αρχίζουν να επιδίδονται με ζήλο στην παρουσίαση και στην απεικόνιση μεγάλης διάστασης δεδομένων. Η περιορισμένη διαθεσιμότητα λογισμικού υψηλής ανάλυσης κατά την διάρκεια του προηγούμενου σταδίου κατακτήθηκε σταδιακά. Η πλειοψηφία των ερευνών απομακρύνθηκε από την ανάπτυξη των εργαλείων διερευνητικής ανάλυσης των δεδομένων, η οποία βασιζόταν κατά κύριο λόγο σε στατιστικά μέτρα και στράφηκε προς την απεικόνιση δεδομένων μεγάλης διάστασης τα οποία απαιτούν ταχύτητα υπολογισμών. Μερικές από τις τεχνικές που αναπτύχθηκαν σε αυτό το στάδιο είναι: **Grand Tour Methods**, **Worlds Within Worlds**, **Dimension Stacking**, **Hierarchical Axis**, **Hyperbox** και μια σειρά άλλων τεχνικών [12]. Κάποιες από αυτές τις τεχνικές προσπαθούν να οπτικοποιήσουν όλες τις διαστάσεις και όλα τα χαρακτηριστικά με μια απεικόνιση, ενώ άλλες σκοπεύουν στην πολλαπλή απεικόνιση στην οποία ο χρήστης μπορεί να αλληλεπιδράσει με την εικόνα πηγαίνοντας το ποντίκι στο μέρος της εικόνας που τον ενδιαφέρει. Η εικονική πραγματικότητα ξεκινάει να εμφανίζεται στην βιβλιογραφία.

4. Το Στάδιο της Εκπόνησης

Στο στάδιο αυτό έχουμε αλματώδη ανάπτυξη της οπτικοποίησης των δεδομένων μεγάλης διάστασης και εισάγεται μια πλειάδα τεχνικών οπτικοποίησης. Το 1990

και το 1991, δημοσιεύτηκαν τουλάχιστον δεκατέσσερις εργασίες σχετικά με την οπτικοποίηση στο συνέδριο οπτικοποίησης IEEE [8]. Μέχρι τότε είχαν δημοσιευτεί σε τρία συνέδρια οπτικοποίησης μόλις τέσσερις εργασίες. Σε αυτό το στάδιο είχε επέλθει μια περίοδος περιορισμένης ανάπτυξης των τεχνικών οπτικοποίησης. Κάποια από τα πιο πρόσφατα ανεπτυγμένα εργαλεία είχαν αναπτυχθεί σε προηγούμενα στάδια. Για παράδειγμα, η μέθοδος HyperSlice είναι μια προσπάθεια συνδυασμού των διαγραμμάτων διασποράς με την τεχνική του brushing [38]. Η τεχνική του AutoVisual είναι μια επέκταση της τεχνικής Worlds Within Worlds η οποία βασίζεται στην τεχνική της αλληλεπίδρασης. Η τεχνική XmdvTool εμπεριέχει τέσσερα εργαλεία οπτικοποίησης: διαστρωμάτωση διαστάσεων, διαγράμματα διασποράς, Glyphs και Parallel Coordinates σε ένα σύστημα με την τεχνική του Brushing [12].

Η έρευνα για την οπτικοποίηση διαμοιράζεται ανάμεσα σε πολλές επιστήμες. Έχουν ήδη γίνει προσπάθειες συνδυασμού ήχου και εικόνας. Έχει επίσης προταθεί το σενάριο βασισμένο στην τεχνική Rule-Based Queue [62] από την επιστήμη της πληροφορικής. Ένα από τα τελευταία ερευνητικά θέματα της οπτικοποίησης είναι το θέμα της αξιολόγησης των τεχνικών. Μένει να δούμε αν οι υπάρχουσες τεχνικές οπτικοποίησης μπορούν να οδηγήσουν σε καλύτερη απεικόνιση ενός προβλήματος και καλύτερη κατανόηση της υποκειμενικής επιστήμης. Η συζήτηση για την οπτικοποίηση δεδομένων μεγάλης διάσταση απέχει πολύ από την ολοκλήρωση.

# 3 Τεχνικές Μείωσης Διαστάσεων

## 3.1 Εισαγωγή

Τα δεδομένα στον πραγματικό κόσμο είναι δεδομένα μεγάλης διάστασης. Συνήθως για να αντιμετωπιστούν επαρκώς, είναι αναγκαίο να γίνει μείωση των διαστάσεών τους. Η μείωση των διαστάσεων αποσκοπεί σε μια αναπαράσταση των δεδομένων, σε ένα χώρο μειωμένης διάστασης χωρίς να χαθεί ουσιαστική πληροφορία. Ιδανικά η αναπαράσταση στις μειωμένες διαστάσεις πρέπει να έχει την πραγματική διάσταση των δεδομένων. Η πραγματική διάσταση είναι αυτή η οποία κρατάει των μικρότερο αριθμό παραμέτρων ώστε να ληφθούν υπόψιν οι παρατηρούμενες ιδιότητες των δεδομένων. Η μείωση των διαστάσεων είναι πολύ σημαντική, καθώς μετριάζει την 'κατάρα' των διαστάσεων και άλλες ανεπιθύμητες ιδιότητες των χώρων μεγάλης διάστασης [15]. Έτσι η μείωση των διαστάσεων διευκολύνει μεταξύ άλλων, την οπτικοποίηση, την ομαδοποίηση και την συμπύκνωση δεδομένων μεγάλης διάστασης. Παρα-

δοσιακά η μείωση διαστάσεων πραγματοποιείται με την χρήση γραμμικών τεχνικών όπως η ανάλυση κύριων συνιστωσών και η κλασική κλιμάκωση [16]. Ωστόσο, αυτές οι γραμμικές τεχνικές δεν μπορούν χειριστούν επαρκώς πολύπλοκα μη γραμμικά δεδομένα.

Την τελευταία δεκαετία έχει προταθεί ένας αριθμός μη γραμμικών τεχνικών για την μείωση των διαστάσεων. Σε αντίθεση με τις παραδοσιακές τεχνικές, οι μη γραμμικές τεχνικές έχουν την δυνατότητα να καταφέρνουν να χειριστούν πολύπλοκα μη γραμμικά δεδομένα. Ειδικά για τα δεδομένα του πραγματικού κόσμου, οι μη γραμμικές τεχνικές μείωσης διαστάσεων μπορούν να προσφέρουν ένα πλεονέκτημα επειδή τα δεδομένα στον πραγματικό κόσμο είναι πιθανό να είναι μη γραμμικά. Προηγούμενες έρευνες έχουν δείξει ότι οι μη γραμμικές τεχνικές είναι περισσότερο αποτελεσματικές από τις γραμμικές για πολύπλοκα δεδομένα. Για παράδειγμα, για το σύνολο δεδομένων Swiss roll (http://people.cs.uchicago.edu/ dinoj/manifold/swissroll.html) που δημιουργήθηκε για να αξιολογήσει τις τεχνικές μείωσης διάστασης και περιλαμβάνει δεδομένα που βρίσκονται σε μια σπειροειδή πολλαπλή διάσταση που είναι ενσωματωμένη μέσα σε έναν τρισδιάστατο χώρο, ένας μεγάλος αριθμός μη γραμμικών τεχνικών μπορεί να μειώσει τέλεια τις διαστάσεις, αντίθετα οι γραμμικές τεχνικές αποτυγχάνουν [15]. Σε αντίθεση με την επιτυχία των μη γραμμικών τεχνικών σε τεχνητά κατασκευασμένα σύνολα δεδομένων, η επιτυχία τους σε σύνολα δεδομένων του πραγματικού κόσμου δεν είναι ανάλογη. Πέρα από αυτήν την παρατήρηση δεν είναι ξεκάθαρο σε ποιο βαθμό οι επιδόσεις των διάφορων τεχνικών μείωσης της διάστασης διαφέρουν για τα τεχνητά και τα δεδομένα του πραγματικού κόσμου. Δεν έχει γίνει μέχρι σήμερα κάποια συστηματική μελέτη για την αξιολόγηση των τεχνικών μείωσης διάστασης. Σε αυτό το κεφάλαιο θα παρουσιαστεί η γενική ιδέα της μείωσης της διάστασης, καθώς και οι βασικές τεχνικές μείωσης της διάστασης.

## 3.2 Ορολογία

Δυστυχώς, η βιβλιογραφία για τα δεδομένα μεγάλης διάστασης πάσχει, αφού είναι ακατάλληλα ορισμένη και με ασυνεπή ορολογία. Ο όρος διαστατικότητα είναι υπερφορτωμένος. Οι μαθηματικοί θεωρούν την διάσταση ως τον αριθμό των ανεξάρτητων μεταβλητών σε μια εξίσωση. Από την άλλη μεριά, οι μηχανικοί θεωρούν τη διάσταση ως την μέτρηση ενός μεγέθους (ύψους, πλάτους, μήκους). Ακόμα και ο όρος πολλαπλό συχνά αντικαθίσταται με το πρόθεμα υπέρ. Στην βιβλιογραφία των στατιστικών το πολλαπλό σημαίνει δύο ή περισσότερα, υποδεικνύοντας έναν φυσικό διαχωρισμό ανάμεσα στην μια και στις δύο διαστάσεις. Για τον διαχωρισμό μεταξύ τριών ή τεσσάρων (και περισσότερων) διαστάσεων, χρησιμοποιείται το πρόθεμα υπέρ [4]. Χρησιμοποιούμε τον όρο πολλαπλό για να αναφερθούμε στις δύο ή

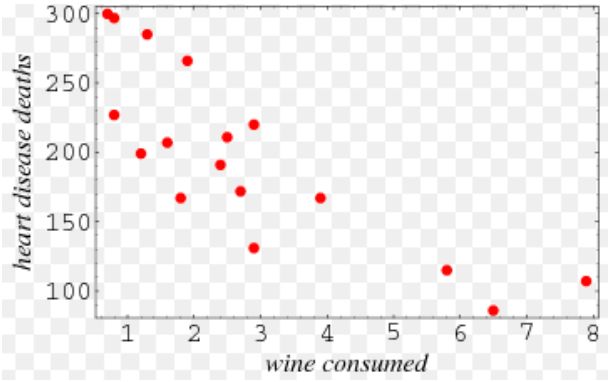
περισσότερες διαστάσεις. Επιπλέον, οι όροι Πολυδιάστατος και Πολυμεταβλητός χρησιμοποιούνται συχνά αόριστα. Ακριβολογώντας, με τον όρο πολυδιάστατο αναφερόμαστε στην διάσταση των ανεξάρτητων μεταβλητών, ενώ με τον όρο πολυμεταβλητό αναφερόμαστε στην διάσταση των εξαρτημένων μεταβλητών. Ο πιο κατάλληλος όρος για την οπτικοποίηση πολυμεταβλητών δεδομένων είναι η πολυδιάστατη πολυμεταβλητή οπτικοποίηση. Παρ΄ όλα αυτά, ένα σύνολο πολυμεταβητών δεδομένων με μεγάλη διάσταση μπορεί να θεωρηθεί πολυδιάστατο, μιας και η σχέση μεταξύ των μεταβλητών δεν είναι γνωστή εκ των προτέρων. Επομένως, η διαστατικότητα εμπεριέχει κοινή χρήση. Για λόγους ευκολίας, ο όρος χαρακτηριστικό υποδηλώνει τις ανεξάρτητες διαστάσεις και τις εξαρτημένες μεταβλητές. Ο **Beddow** [49] επισήμανε την διαφορά ανάμεσα στα πολυδιάστατα δεδομένα και τα πολυδιάστατα αντικείμενα. Τα πολυδιάστατα αντικείμενα είναι χωρικά αντικείμενα και στόχος μας είναι να κατανοήσουμε την γεωμετρία τους. Η πιο κοινή μορφή είναι οι εικόνες δύο διαστάσεων και οι τρισδιάστατοι όγκοι. Μπορούν να περιγραφούν καλύτερα ως $n$ διάστασης ευκλείδειοι χώροι. Τα πολυδιάστατα δεδομένα από την άλλη μεριά αναφέρονται στην σχέση ανάμεσα στις πολλαπλές παραμέτρους. Μαθηματικά αυτές οι παράμετροι μπορούν να διαχωριστούν σε δύο κατηγορίες: ανεξάρτητες και εξαρτημένες. Μερικοί στατιστικοί προτιμούν τους όρους παράγοντες και απόκριση [4]. Μια μεταβλητή λέμε ότι είναι εξαρτημένη όταν είναι συνάρτηση μίας άλλης μεταβλητής. Η σχέση μίας ανεξάρτητης μεταβλητής $x$ και μίας εξαρτημένης μεταβλητής $y$ μπορεί να περιγραφεί από την μαθηματική ισότητα $f(x) = y$. Μπορούμε να δεχθούμε την σύμβαση ότι ο όρος πολυδιαστατικός αναφέρεται στην διαστατικότητα των ανεξάρτητων μεταβλητών, ενώ ο όρος πολυπαραγοντικός αναφέρεται στην διαστατικότητα των εξαρτημένων μεταβλητών. Αυτός είναι ο πιο δημοφιλής τρόπος για να περιγράψουμε την διαστατικότητα στην διεθνή βιβλιογραφία της επιστημονικής οπτικοποίησης [4].

Ο **Beddow** υποστηρίζει ότι οι αναλυτικές μέθοδοι που χρησιμοποιούνται για να εξερευνήσουμε τους $n$-διάστατους ευκλείδιους χώρους $R^n$ δεν είναι γενικά κατάλληλοι για την πολυπαραγοντική ανάλυση. Στην έρευνα της οπτικοποίησης δεδομένων μεγάλης διάστασης, δεν δίνουμε μεγάλη έμφαση στους αυστηρούς ορισμούς των μαθηματικών για τις ανεξάρτητες και τις εξαρτημένες μεταβλητές, αλλά στρεφόμαστε προς τον ευρύτερο ορισμό των πολλαπλών μεταβλητών ή των παραγόντων. Αυτό δεν συμβαίνει μόνο στην έρευνα της επιστημονικής οπτικοποίησης, αλλά και στην στατιστική επιστήμη. Τα εργαλεία είναι διαφορετικά αλλά ο στόχος είναι κοινός, να βρεθούν οι κρυμμένες συσχετίσεις μεταξύ των μεταβλητών.

Γενικά, τα ακατέργαστα επιστημονικά δεδομένα μπορούν να ταξινομηθούν ιεραρχικά σε δεδομένα διαφόρων τύπων. Τα πιο γενικά και αυτά που βρίσκονται χαμηλότερα στην ιεραρχία είναι τα κατηγορικά δεδομένα, των οποίων οι τιμές δεν μπορούν να τοποθετηθούν με κάποια

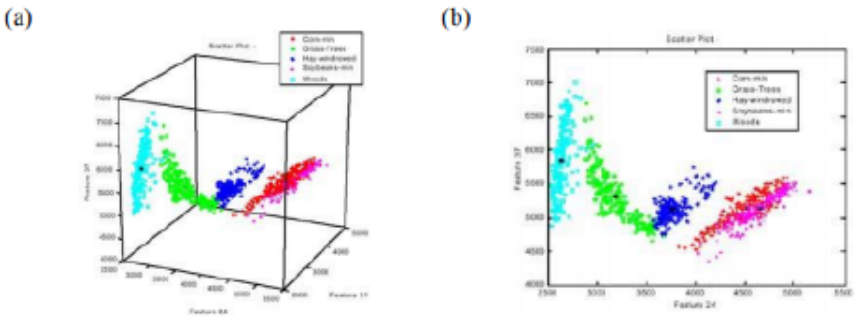
σειρά. Για παράδειγμα τα ονόματα των πόλεων είναι κατηγορικά δεδομένα. Ο επόμενως τύπος στην ιεραρχία είναι τα διατακτικά δεδομένα, για τα οποία οι τιμές τους διατάσσονται, αλλά για αυτήν την κατηγορία των δεδομένων δεν έχει νόημα η απόσταση μεταξύ των κατηγοριών τους. Σε αυτήν την κατηγορία θα μπορούσαν να μπουν τα εφτά χρώματα του ουράνιου τόξου. Στο υψηλότερο σκαλί της ιεραρχίας είναι τα αριθμητικά δεδομένα, για τα οποία για κάθε δύο τιμές έχει ερμηνεία η απόστασή τους. Ο χρόνος, η θερμοκρασία και οι αποστάσεις ανήκουν σε αυτήν την κατηγορία δεδομένων.

Η διάσταση των δεδομένων αναφέρεται στον αριθμό των χαρακτηριστικών που πρέπει να οπτικοποιηθούν [15]. Τα μονοδιάστατα δεδομένα τα οποία έχουν μόνο ένα χαρακτηριστικό μπορούν να απεικονιστούν αποτελεσματικά από τεχνικές όπως οι πίνακες και τα ιστογράμματα. Για τα δισδιάστατα δεδομένα χρησιμοποιούμαι συνήθως την απεικόνιση στο επίπεδο με τις καρτεσιανές συντεταγμένες. Μια συνηθισμένη τεχνική είναι να απεικονίσουμε την μια μεταβλητή σε σχέση με την άλλη με ένα διάγραμμα διασποράς (Σχήμα 1).

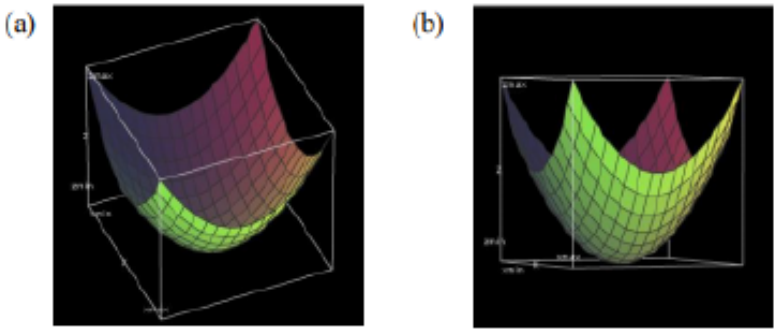



Σχήμα 1: Διάγραμμα διασποράς που δείχνει τους θανάτους ατόμων που πάσχουν από χρόνιες καρδιακές παθήσεις σε σχέση με την κατανάλωση κρασιού [68].

Τεχνικά, όταν μιλάμε για δεδομένα μεγάλης διάστασης εννοούμε δεδομένα με διάσταση πάνω από τρεις. Δημιουργούνται προβλήματα όταν θέλουμε να απεικονίσουμε στο επίπεδο δεδομένα που αναφέρονται σε τρεις διαστάσεις [16], αφού μας είναι δύσκολο να συγκρίνουμε δύο σημεία κατά μήκος του ίδιου άξονα (Σχήμα 2). Στις τρισδιάστατες επιφάνειες συναντάμε το πρόβλημα που αναφέραμε πριν, για παράδειγμα για να γίνει ορατή η ελάχιστη τιμή (Σχήμα 3) θα πρέπει να αλλάξουμε την οπτική γωνία παρατήρησης. Προφανώς το πρόβλημα οξύνεται όταν αυξάνονται οι διαστάσεις, η κατάλληλη αλληλεπίδραση θα πρέπει να είναι σε θέση να αντιμετωπίσει το συγκεκριμένο πρόβλημα.




Σχήμα 2: Απεικόνιση τρισδιάστατων δεδομένων (a) και απεικόνιση των ίδιων δεδομένων στο επίπεδο (b) [68].




Σχήμα 3: Παρατηρούμε την αλλαγή οπτικής από την εικόνα (a) στην εικόνα (b) [68].

Δεν υπάρχει σαφής διαχωρισμός μεταξύ της χαμηλής και της υψηλής διαστατικότητας. Τα δεδομένα μεγάλης διάστασης ορίζονται αυθαίρετα, αλλά συνήθως εννοούμε δεδομένα με διάσταση μεγαλύτερη του τρία. Είναι σημαντικό να αναφέρουμε ότι η γεωμετρική απεικόνιση είναι αναποτελεσματική στο να μεταφέρει την πληροφορία στον άνθρωπο. Αυτό οφείλεται στην αντιληπτική ικανότητα του ανθρώπου ανάμεσα στην χαμηλή και υψηλή διαστατικότητα.

Παρακάτω ορίζουμε το πρόβλημα της (μη γραμμικής) μείωσης διαστάσεων ας υποθέσουμε ότι έχουμε ένα σύνολο δεδομένων σε έναν $nxD$ πίνακα X, ο οποίος αποτελείται από $n$ διανύσματα $x_i$ $(i \in 1,2,.....,n)$ με διάσταση $D$ και επιπλέον ότι η πραγματική διάσταση των δεδομένων είναι $d$ (όπου $d < D$ και συχνά $d << D$) [15]. Πραγματική διάσταση σημαίνει ότι τα σημεία των δεδομένων X, βρίσκονται πάνω ή κοντά σε έναν χώρο διάστασης $d$, ο οποίος βρίσκεται σε έναν χώρο διάστασης $D$. Να σημειώσουμε ότι δεν κάναμε καμία υπόθεση για την δομή αυτού του χώρου (ίσως αυτός ο χώρος να αποτελείται από έναν αριθμό πολλαπλών χώρων οι οποίοι δεν είναι συνδεδεμένοι μεταξύ τους). Οι τεχνικές μείωσης διαστάσεων μεταφέρουν το σύνολο δεδομένων X σε ένα νέο σύνολο δεδομένων $\Upsilon$ με διάσταση

$d$, διατηρώντας την γεωμετρία των δεδομένων όσο το δυνατόν περισσότερο. Γενικά, δεν μας είναι γνωστή η γεωμετρία των πολλαπλών επιπέδων, ούτε η πραγματική διάσταση των δεδομένων αλλά ούτε η αρχική διάσταση των δεδομένων X. Άρα, η μείωση των διαστάσεων είναι ένα δύσκολο πρόβλημα που μπορεί να επιλυθεί μόνο αν υποθέσουμε ορισμένες ιδιότητες των δεδομένων (όπως η πραγματική τους διάσταση). Σε αυτήν την εργασία, θα ορίζουμε ένα σημείο μεγάλης διάστασης ως $x_i$, όπου $x_i$ είναι η i-οστή γραμμή του $D$ διάστασης πίνακα X. Το μειωμένης διάστασης αντίστοιχο σημείο του $x_i$ ορίζεται ως $y_i$, όπου $y_i$ είναι η i-οστή γραμμή του $d$ διάστασης πίνακα $\Upsilon$. Στο υπόλοιπο αυτής της εργασίας θα χρησιμοποιούμε τους παραπάνω ορισμούς.

## 3.3 Ανάλυση Κύριων Συνιστωσών

Η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis (PCA)) είναι μια γραμμική τεχνική για μείωση των διαστάσεων [16]. Παρόλο που υπάρχουν διάφορες τεχνικές που κάνουν το ίδιο, η Ανάλυση Κύριων Συνιστωσών είναι η πιο δημοφιλής από τις γραμμικές αυτές τεχνικές. Η τεχνική αυτή κατασκευάζει μια χαμηλής διάστασης αναπαράσταση των δεδομένων η οποία περιγράφει όσον το δυνατόν μεγαλύτερο μέρος της διακύμανσης των δεδομένων. Αυτό επιτυγχάνεται με την εύρεση μιας γραμμικής βάσης μειωμένης διάστασης για τα δεδομένα, η οποία καταφέρνει να κρατήσει το μεγαλύτερο ποσό της διακύμανσης των αρχικών δεδομένων.

Από μαθηματικής άποψης, η Ανάλυση Κύριων Συνιστωσών είναι μια γραμμική απεικόνιση των δεδομένων, αφού έχει γίνει μείωση την διαστάσεων. Έστω $N = x_1, x_2, ....., x_n$ τα αρχικά μας δεδομένα. Τότε έχουμε

$$SS = \sum_{r=1}^{n} \sum_{i=r+1}^{n} d^2(x_r, x_i)$$

με

$$d^2(x_r, x_i) = (x_r - x_i)^T (x_r - x_i)$$

η Ευκλείδεια απόσταση. Επιπλέον, ισχύει ότι

$$SS = nDis(N)$$

με

$$Dis(N) = \sum_{i=1}^{n} d^2(x_i, \bar{x})$$

Η ποσότητα αυτή θα λέγεται διασπορά του συνόλου σημείων $N$ και υπολογίζεται από τον τύπο

$$Dis(N) = tr(Z^T Z).$$

Για να βρούμε τον καλύτερο γραμμικό συνδυασμό, θα πρέπει να αναζητήσουμε το διάνυσμα των συντελεστών $\alpha = (\alpha_1, \alpha_2, ...., \alpha_p)^T$ έτσι ώστε να μεγιστοποιηθεί η ποσότητα

$Dis_\alpha(Y) = Dis(Y) = \alpha^T Z^T Z\alpha.$

Όμως για κάθε k

$Dis_{k\alpha} = (k\alpha)^T Z^T Z(k\alpha) = k^2(\alpha^T Z^T Z\alpha) = k^2 Dis_\alpha(N)$
οπότε θέτουμε τον περιορισμό

$||\alpha|| = 1 \Leftrightarrow \alpha^T\alpha = 1 \Leftrightarrow \sum_{i=1}^{p} \alpha_i^2 = 1.$

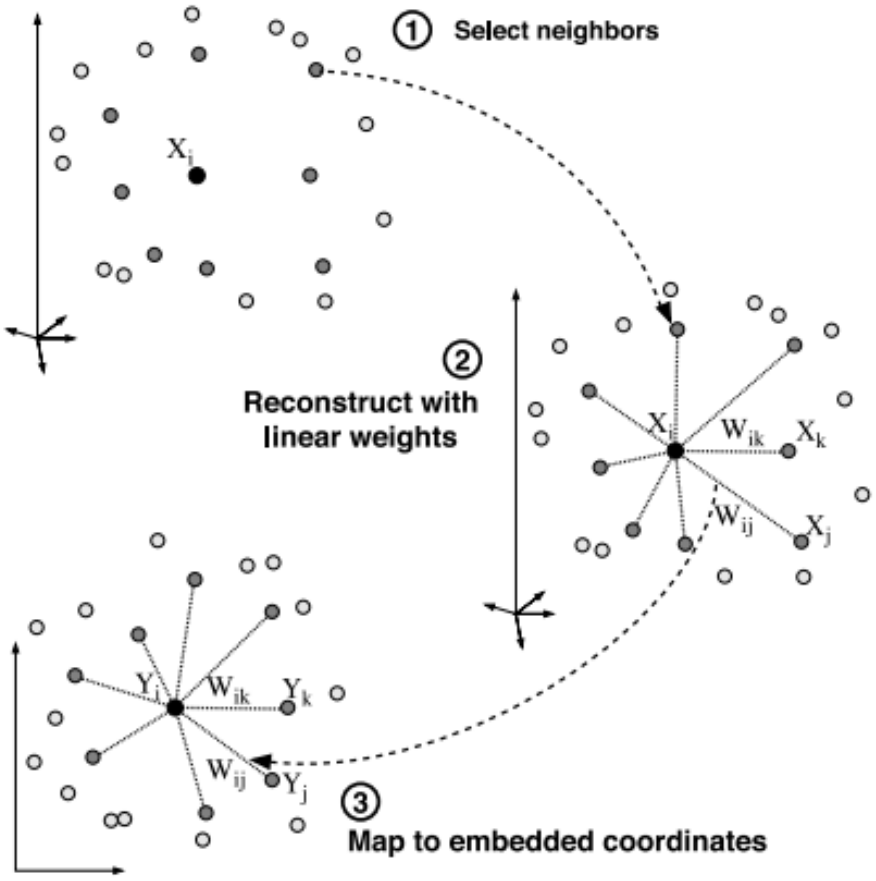
Όσον αφορά την οπτικοποίηση, μειώνοντας τον αρχικό αριθμό των μεταβλητών μπορούμε να αναπαραστήσουμε ευκολότερα τα δεδομένα και να ανιχνεύσουμε ομαδοποιήσεις ή ακραίες παρατηρήσεις. Η μεθοδολογία εύρεσης και επιλογής των κύριων συνιστωσών έχει ως εξής: έστω ότι έχουμε k μεταβλητές $x = (x_1, x_2, ..., x_k)$ από τις οποίες θέλουμε να δημιουργήσουμε τις κύριες συνιστώσες $y = (y_1, y_2, ..., y_k)$. Οι τελευταίες είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών, δηλαδή:

$y_1 = \alpha_{11}x_1 + \alpha_{12}x_2 + ..... + \alpha_{1k}x_k$
$y_2 = \alpha_{21}x_1 + \alpha_{22}x_2 + ..... + \alpha_{2k}x_k$
.....
$y_k = \alpha_{k1}x_1 + \alpha_{k2}x_2 + ..... + \alpha_{kk}x_k$
ή με μορφή πινάκων $y = \alpha x$, όπου α είναι τετραγωνικός διάστασης kxk. Η προταθείσα από τον Pearson μεθοδολογία αφορά τον εντοπισμό των στοιχείων του τετραγωνικού πίνακα α έτσι ώστε οι κύριες συνιστώσες να είναι σε φθίνουσα σειρά ως προς την διακύμανση, δηλαδή η πρώτη συνιστώσα να έχει τη μεγαλύτερη διακύμανση, η δεύτερη να έχει την δεύτερη μεγαλύτερη διακύμανση κοκ. Αποδεικνύεται ότι αν Σ είναι ο θετικά ορισμένος πίνακας διακύμανσης του διανύσματος $x$ και $(\lambda_1, \varepsilon_1), ....(\lambda_k, \varepsilon_k)$ τα ζεύγη ιδιοτιμών και ιδιοδιανυσμάτων του Σ με $\lambda_1 \geq ..... \geq \lambda_k$, τότε οι γραμμικοί συνδυασμοί $y_1 = \varepsilon_1' x, ....., y_k = \varepsilon_k' x$ είναι οι κύριες συνιστώσες, ασυσχέτιστες μεταξύ τους και με συνολική διασπορά ίση με την συνολική διασπορά των αρχικών μεταβλητών.

Η Ανάλυση Κύριων Συνιστωσών έχει δοκιμαστεί επιτυχημένα σε μια πλειάδα εφαρμογών, όπως η αναγνώριση προσώπου και η ομαδοποίηση κερμάτων. Αυτή η τεχνική έχει δύο βασικά μειονεκτήματα [20]. Πρώτων, στην Ανάλυση Κύριων Συνιστωσών, το μέγεθος του πίνακα συνδιακυμάνσεων είναι ανάλογο με την διάσταση των παρατηρήσεων των δεδομένων. Ως αποτέλεσμα αυτού, ο υπολογισμός των ιδιοδιανυσμάτων να είναι κάποιες φορές ακατόρθωτος για δεδομένα πολύ μεγάλης διάστασης.

## 3.4 Τοπικές Γραμμικές Συντεταγμένες

Η μέθοδος των τοπικών γραμμικών συντεταγμένων (Locally Linear Coordination (LLC)) είναι μια γραμμική τεχνική για την μείωση των διαστάσεων [12]. Αρχικά, βρίσκει τους k κοντινότερους γείτονες για κάθε σημείο. Στην συνέχεια προσεγγίζει κάθε διάνυσμα των δεδομένων ως ένα γραμμικό άθροισμα βαρών αυτών των k κοντινότερων γειτόνων. Τέλος υπολογίζει τα βάρη τα οποία κατασκευάζουν καλύτερα τα διανύσματα των δεδομένων από τους γείτονές τους και με αυτόν τον τρόπο παράγονται τα μειωμένης διάστασης διανύσματα.




Σχήμα 4: Απεικόνιση δεδομένων με την LLE

## 3.5 Χάρτες του **Sammon**

Οι χάρτες του Sammon είναι ένα χρήσιμο εργαλείο για την αναγνώριση προτύπων [31]. Είναι ένας αλγόριθμος για την ανεύρεση μίας απεικόνισης των δεδομένων διάστασης $d$ μέσα σε έναν μη γραμμικό χώρο διάστασης $m$ (όπου $d > m$), διατηρώντας όσο καλύτερα γίνε-

ται τις αποστάσεις ανάμεσα στα πρότυπα. Ο αλγόριθμος αυτός συχνά χρησιμοποιείται για την απεικόνιση δεδομένων μεγάλης διάστασης στο δισδιάστατο επίπεδο ή στον τρισδιάστατο χώρο. Ωστόσο, μπορεί να χρησιμοποιηθεί για να απεικονίσει τα δεδομένα σε οποιοδήποτε μειωμένης διάστασης χώρο, η απεικόνιση δεν είναι απαραίτητο να είναι στις δύο διαστάσεις.

Μια αδυναμία των χαρτών του Sammon, σε αντίθεση με την Ανάλυση Κύριων Συνιστωσών, είναι ότι δεν παράγει μια μαθηματική ή αλγοριθμική διαδικασία χαρτογράφησης των προηγούμενων σημείων των δεδομένων. Έτσι, όταν πρέπει να χαρτογραφηθεί ένα καινούργιο σημείο, θα πρέπει να επαναληφθεί η διαδικασία της χαρτογράφησης. Κατά την διαδικασία της χαρτογάφησης του Sammon μπορεί να χρησιμοποιηθεί η βελτιστοποίηση μίας πλειάδας κριτηρίων, όπως ο διαχωρισμός των ομάδων ή η ποσότητα της διακύμανσης που ερμηνεύει η χαρτογράφηση. Η χαρτογράφηση του Sammon προσπαθεί να διατηρήσει τις αποστάσεις στον χώρο μειωμένης διάστασης. Αυτό το καταφέρνει ελαχιστοποιώντας το σφάλμα, χρησιμοποιώντας ένα κριτήριο που δίνει ποινή στις διαφορές των αποστάσεων των σημείων στον πραγματικό χώρο και στον χώρο απεικόνισης. Αν ορίσουμε την απόσταση ανάμεσα σε δύο σημεία $x_i$ και $x_j$ με $i \neq j$, ως $d_{ij}$ στον πραγματικό χώρο και την απόσταση ανάμεσα σε δύο σημεία $y_i$ και $y_j$ με $i \neq j$, ως $D_{ij}$ στον χώρο απεικόνισης, τότε η μετρική του Sammon ορίζεται ως

$$E = \frac{1}{\sum_{i=1}^{n}\sum_{j=i+1}^{n-1} d_{ij}} \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{(d_{ij}-D_{ij})^2}{d_{ij}} \quad (1).$$

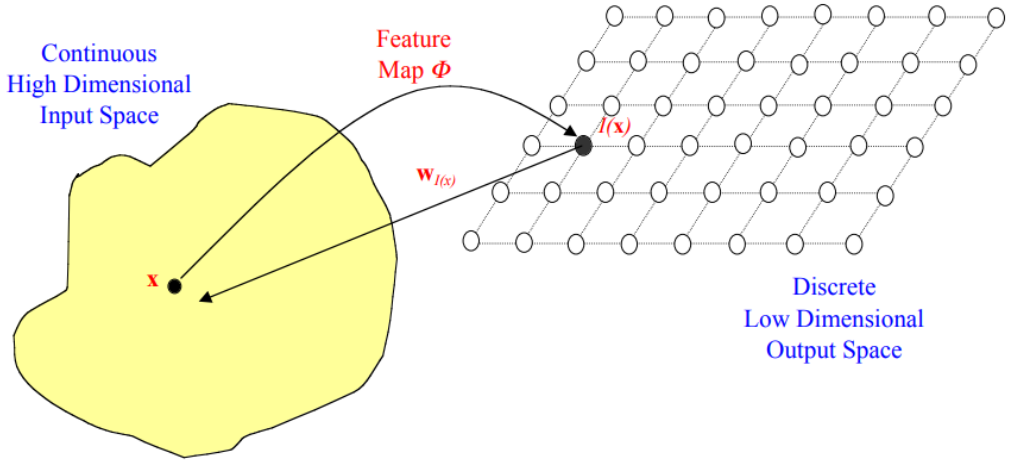
Να σημειώσουμε ότι όλες οι αποστάσεις έχουν το ίδιο βάρος καθώς διαιρούνται με βάση την απόσταση των αρχικών δεδομένων, $d_{ij}$.

## 3.6 Χάρτες Αυτό-Οργάνωσης

Μέχρι τώρα έχουμε δει δίκτυα με επιβλεπόμενες τεχνικές εκπαίδευσης, σε αυτές υπάρχει μια μεταβλητή στόχος για κάθε είσοδο δεδομένων και το δίκτυο μαθαίνει να παράγει τις απαιτούμενες εξόδους. Τώρα θα στραφούμε στην μη επιβλεπόμενη εκπαίδευση, σε αυτήν τα νευρωνικά δίκτυα μαθαίνουν από τις ομαδοποιήσεις των δεδομένων εκπαίδευσης χωρίς εξωτερική βοήθεια [32].

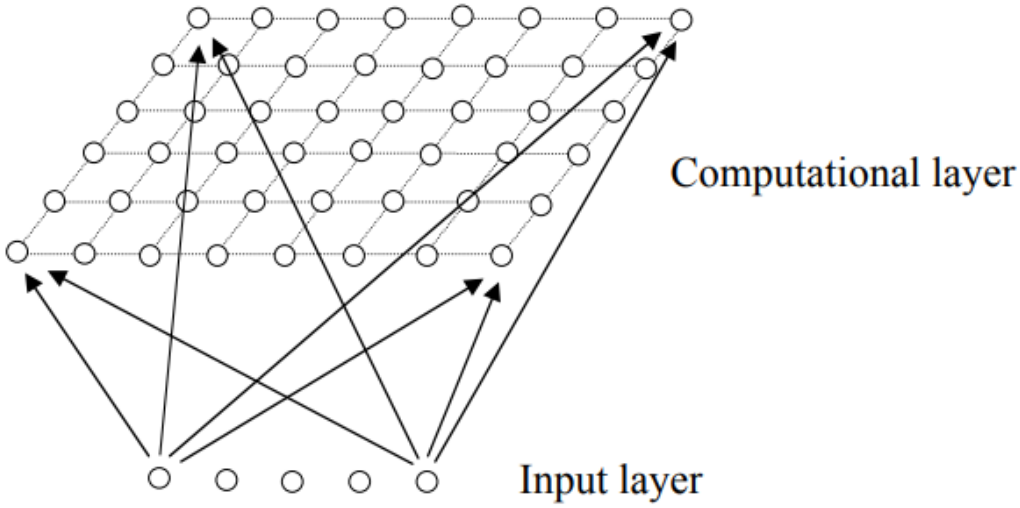
Ο βασικός σκοπός των Self Organizing Map (SOM) είναι να μετατρέψουν ένα εισερχόμενο πρότυπο οποιασδήποτε διάστασης σε απεικόνιση μίας ή δύο διαστάσεων. Επομένως, δημιουργούμε τους (SOM) τοποθετώντας τους νευρώνες στους κόμβους ενός πλέγματος μίας ή δύο διαστάσεων. Χάρτες υψηλότερης διάστασης είναι δυνατόν να δημιουργηθούν,

αλλά δεν είναι σύνηθες. Μπορούμε αυτό να το δούμε ως μια γενίκευση της Ανάλυσης των Κύριων Συνιστωσών. Στο Σχήμα (5) μπορούμε να δούμε την απεικόνιση των σημείων $x$ από τον αρχικό χώρο τους στα σημεία $I(x)$ στον χώρο μειωμένης διάστασης.




Σχήμα 5: Απεικόνιση με την τεχνική SOM του $x$ στο σημείο $I(x)$ του χώρου μειωμένης διάστασης [32].

Θα μπορούσαμε να επικεντρωθούμε σε ένα συγκεκριμένο είδος χαρτών SOM, οι οποίοι είναι γνωστοί ως δίκτυα Kohonen. Κάθε νευρώνας είναι πλήρως συνδεδεμένος με όλους τους πηγαίους κόμβους του στρώματος εισόδου (Σχήμα 6).




Σχήμα 6: Στο κάτω μέρος της εικόνας μπορούμε να δούμε τα δεδομένα εισόδου και πάνω πώς διατάσσονται από τον αλγόριθμο του Kohonen [32].

Στην απεικόνιση στην μια διάσταση θα έχει απλώς μια γραμμή (ή μια στήλη) στο πλέγμα υπολογισμού. Οι χάρτες SOM αποτελούνται από τέσσερα βασικά συστατικά. Τα βάρη σύνδεσης αρχικοποιούνται με μικρές τυχαίες τιμές. Στην συνέχεια για κάθε πρότυπο εισόδου, οι νευρώνες υπολογίζουν τις σχετικές τους τιμές χρησιμοποιώντας μια διαχωριστική συνάρτηση που παρέχει την βάση του ανταγωνισμού. Ο νευρώνας που παίρνει την μικρότερη τιμή με βάση την διαχωριστική συνάρτηση ανακηρύσσεται ο νικητής. Τέλος, οι υπάρχοντες νευρώνες μειώνουν τις ατομικές τους τιμές για την διαχωριστική συνάρτηση σε σχέση με το πρότυπο εισόδου μέσω κατάλληλης προσαρμογής των σχετικών βαρών σύνδεσης, τέτοια ώστε η ανταπόκριση του νικητή νευρώνα στην εφαρμογή μεταγενέστερα ενός παρόμοιου προτύπου εισόδου να είναι ενισχυμένη [32].

Αν ο εισερχόμενος χώρος είναι διάστασης $D$ (για παράδειγμα υπάρχουν $D$ μονάδες εισόδου) μπορούμε να γράψουμε τις μονάδες εισόδου ως $x = x_i : i = 1,...,D$ και τα βάρη εισόδου σύνδεσης της μονάδας $i$ με τους νευρώνες $j$ στο πλέγμα υπολογισμού μπορούν να γραφούν ως $w_j = w_{ji} : j = 1,...N; i = 1,...,D$, όπου $N$ είναι ο αριθμός των νευρώνων. Μπορούμε να ορίσουμε την διαχωριστική συνάρτηση ως:

$$d_j(x) = \sqrt{\sum_{i=1}^{D}(x_i - w_{ji})^2}.$$

Με άλλα λόγια, ο νευρώνας του οποίου το διάνυσμα βάρους είναι πιο κοντά στο διάνυσμα εισόδου (δηλαδή το πιο παρόμοιο με αυτό) κηρύσσεται νικητής. Με αυτόν τον τρόπο ο εισερχόμενος χώρος μπορεί να απεικονισθεί στον διακεκριμένο χώρο των νευρώνων από μια απλή διαδικασία ανταγωνισμού μεταξύ των νευρώνων.

## 3.7 Πολυδιάστατη Κλιμάκωση

Η Πολυδιάστατη Κλιμάκωση (Multidimesional Scaling (MDS)) αναφέρεται σε μια ομάδα μεθόδων οι οποίες χρησιμοποιούνται ευρέως για την μείωση των διαστάσεων. Ας ορίσουμε την απόσταση μεταξύ του $i$-οστού και του $j$-οστού αντικειμένου ως $d_{ij}$. Αν τα αντικείμενα ορίζονται από τα πολυδιάστατα σημεία $x_i = (x_{i1}, x_{i2}, ....., x_{in}), i = 1, ....., m$, η απόσταση μπορεί να οριστεί ως $d_{ij} = d(x_i, x_j)$, διάφορες αποστάσεις μπορούν να χρησιμοποιηθούν. Η απόσταση μεταξύ του $i$-οστού και του $j$-οστού αντικειμένου $y_i$ και $y_j$ στον χώρο μειωμένης διάστασης, ορίζεται ως $D_{ij}$. Ο σκοπός της Πολυδιάστατης Κλιμάκωσης είναι να βρεθούν σημεία μειωμένης διάστασης $Y_i = (y_{i1}, y_{i2}, ....., y_{iD})$, έτσι ώστε οι μεταξύ τους αποστάσεις να είναι όσο κοντά γίνεται στις πραγματικές. Η αντικειμενική συνάρτηση ελαχίστων τετρα-

γώνων που πρέπει να ελαχιστοποιηθεί είναι

$$\sigma(y) = \sum_{i<j} w_{ij}(D(y_i, y_j) - d\,ji)^2 \quad (1)$$

όπου $y = (y_1, y_2, ....., y_m)$ και τα $w_{ij}$ είναι τα βάρη. Η κανονικοποιημένη συνάρτηση ορίζεται ως

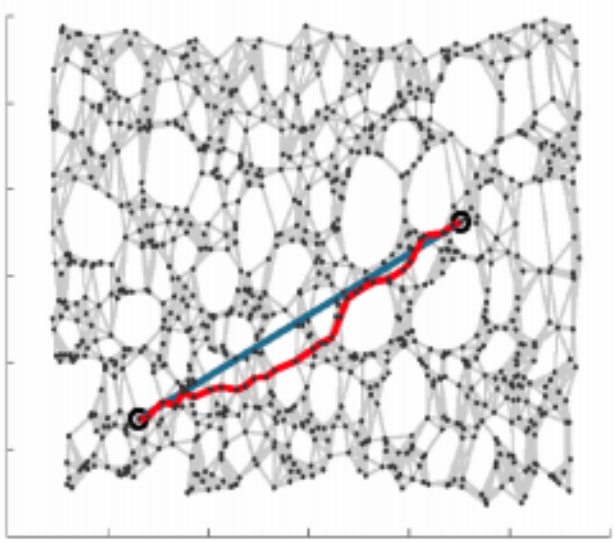
$$\sigma(y) = \frac{\sum_{i<j} w_{ij}(D(y_i, y_j) - d\,ji)^2}{\sum_{i<j} w_{ij} d\,ji^2} (2)$$

Η κανονικοποιημένη συνάρτηση δίνει μια ξεκάθαρη εξήγηση της ποιότητας της οπτικοποίησης που εξαρτάται λιγότερο από τον αριθμό των αντικειμένων $m$ και την κλίμακα των αποστάσεων. Το σχετικό σφάλμα ορίζεται από την σχέση

$$I(y) = \sqrt{\sigma(y)} = \sqrt{\frac{\sum_{i<j} w_{ij}(D(y_i, y_j) - d\,ji)^2}{\sum_{i<j} w_{ij} d\,ji^2}} (3).$$

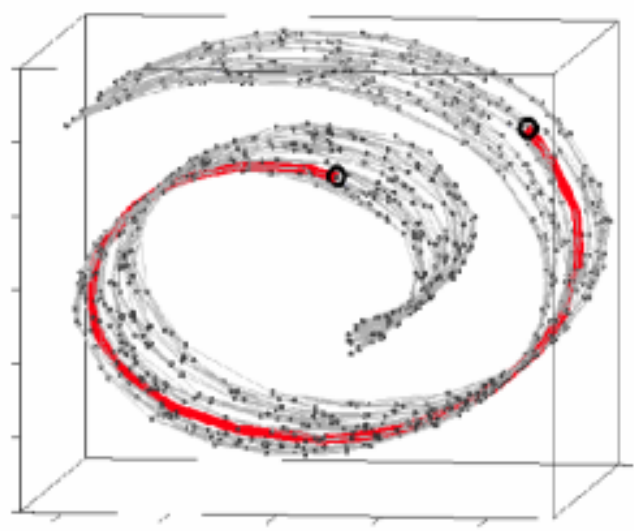
## 3.8 **ISOMAP**

Η ISOMAP διατηρεί την εγγενή γεωμετρία των δεδομένων χρησιμοποιώντας τις γεωδαιτικές αποστάσεις. Η γεωδαισιανή απόσταση είναι το μήκος της μικρότερης καμπύλης που ενώνει δύο απομακρυσμένα σημεία μίας υπερεπιφάνειας (Σχήμα 7). Συγκεκριμένα, για γειτονικά σημεία η γεωδαισιανή τους απόσταση βρίσκεται πολύ κοντά με την τιμή της ευκλείδειας απόστασής τους.



Σχήμα 7: Η γεωδαιτική απόσταση είναι με μπλε χρώμα [68].

Για απομακρυσμένα σημεία, η απόσταση προσδιορίζεται από μια ακολουθία μικρών βημάτων ματαξύ γειτονικών σημείων. Δημιουργείται από την ένωση των ακμών γειτονικών σημείων (Σχήμα 8).



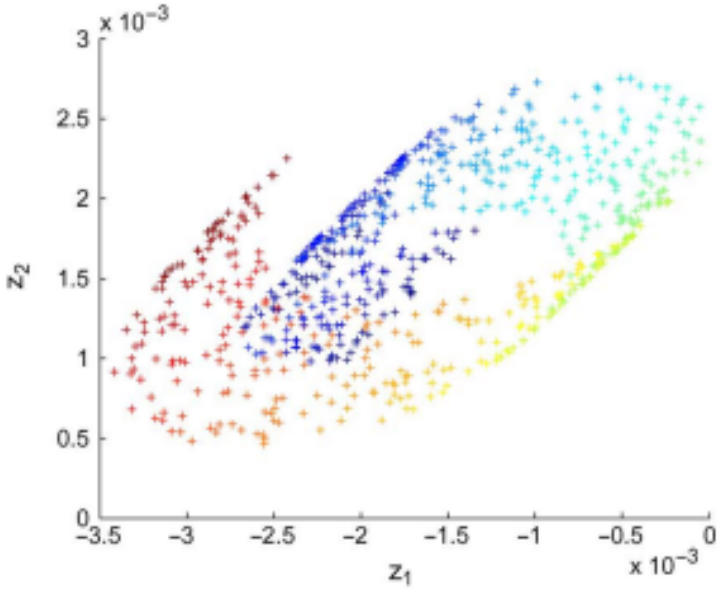
Σχήμα 8: Γεωδαιτική απόσταση δύο σημείων με κόκκινο χρώμα [68].

Ο αλγόριθμος που χρησιμοποιεί η ISOMAP περιγράφεται από τα παρακάτω βήματα. Αρχικά, γίνεται προσδιορισμός των γειτόνων κάθε σημείου $z_i$, βρίσκοντας όλα τα σημεία $z_j$ σε απόσταση ακτίνας ε. Στην συνέχεια κατασκευάζεται ο γράφος γειτονίας $G$ ενώνοντας κάθε σημείο με την ευκλείδεια ακμή $d_x(i,j)$ με τα γειτονικά του σημεία. Αφού ενωθούν όλα τα σημεία δημιουργείται ο πίνακας αποστάσεων $D_x = d_x(i,j)$. Προχωράμε με υπολογισμό των αποστάσεων πάνω στον γράφο $G$ και εφαρμογή του κλασικού αλγορίθμου MDS ( Multidimensional Scaling). Χρησιμοποιείται ο αλγόριθμος Dijkstra για τον υπολογισμό των γεωδαιτικών αποστάσεων και υπολογίζεται ο πίνακας γεωδαιτικών αποστάσεων $D_G = d_G(i,j)$. Στην συνέχεια γίνεται εφαρμογή του αλγορίθμου MDS στον πίνακα $D_G$. Τέλος, λύνοντας το πρόβλημα των ιδιοτιμών στον προηγούμενο πίνακα βρίσκουμε την ενσωμάτωση των σημείων στον ελαττωμένο χώρο. Ο ISOMAP είναι μη γραμμικός και μη επαναληπτικός αλγόριθμος. Ένα από τα μεγαλύτερα μειωνεκτήματά του είναι ότι για μικρό αριθμό δειγμάτων οδηγεί σε μη ακριβή υπολογισμό της γεωδαιτικής απόστασης και ότι η μεγάλη καμπυλότητα της επιφάνειας απαιτεί μεγάλο αριθμό γειτόνων για τον εντοπισμό της.

## 3.9 Ιδιοχάρτες του **Laplace**

Οι ιδιοχάρτες του Laplace (Laplacian Eigenmaps) βρίσκουν την αναπαράσταση της μειωμένης διάστασης που διατηρεί πιστά τοπικές δομές στον χώρο των χαρακτηριστικών. Έχουμε τρία βήματα σε αυτήν την διαδικασία. Αρχικά, κατασκευάζουμε το γράφημα γειτονίας μεταξύ $\varepsilon$-γειτονιών ή των $n$ κοντινότερων γειτόνων [12]. Επιλέγονται τα βάρη των άκρων να

είναι 1 εάν είναι συνδεδεμένοι ή αλλιώς να είναι μηδέν. Στην συνέχεια μπορούμε να πάρουμε τον πίνακα βαρών $W$. Τέλος, λύνουμε το γενικευμένο πρόβλημα των ιδιοδιανυσμάτων: $Lf = \lambda Df$, όπου ο $D$ είναι ένας διαγώνιος πίνακας βαρών που έχει ως εισόδους το άθροισμα των γραμμών του $W$ ($D_{ii} = \sum_j W_{ij}$) και $L = D - W$ είναι ο πίνακας του Laplace. Αφήνουμε εκτός το ιδιοδιάνυσμα $f_0$ και χρησιμοποιούμε τα επόμενα $d$ ιδιοδιανύσματα για ενσωμάτωση στον $d$ διάστατο ευκλείδιο χώρο, $x_i \to (f_1(i), ..., f_d(i))$.




Σχήμα 9: Μια απεικόνιση του συνόλου δεδομένων Swiss roll με την τεχνική των Ιδιοχαρτών του Laplace [9].

## 3.10 Οπτικοποίηση Δεδομένων Χρησιμοποιώντας την **t-SNE**

Θα παρουσιάσουμε την τεχνική t-SNE η οποία απεικονίζει δεδομένα μεγάλης διάστασης, δίνοντας σε κάθε σημείο των δεδομένων μια θέση στον δισδιάστατο ή τρισδιάτοτο χώρο. Η τεχνική αυτή παράγει απεικονίσεις μειώνοντας την τάση να συγκεντρώνονται τα σημεία στο κέντρο της απεικόνισης [69]. Η απεικόνιση που παράγεται από την τεχνική t-SNE είναι σημαντικά καλύτερη από αυτήν που παράγουν οι άλλες τεχνικές για πολλά σύνολα δεδομένων, στο Κεφάλαιο 4 θα γίνει σύγκριση κάποιων τεχνικών με την t-SNE. Η τεχνική t-SNE ξεκινάει με την μετατροπή των μεγάλων Ευκλείδειων αποστάσεων μεταξύ των σημείων των δεδομένων σε υπό συνθήκη πιθανότητες, οι οποίες αντιπροσωπεύουν ομοιότητες. Η ομοιότητα του σημείου $x_j$ με το σημείο $x_i$ είναι η υπό συνθήκη πιθανότητα, $p_{j|i}$, όπου το $x_i$ θα διάλεγε το $x_j$ ως γείτονά του αν οι γείτονες επιλεχθούν ανάλογα με την πυκνότητα πιθανότητας μίας Gaussian κατανομής η οποία έχει μέση τιμή το ημιάθροισμα των συντεταγμένων του σημείου $x_i$. Για τα διπλανά σημεία των δεδομένων, η $p_{j|i}$ είναι σχετικά υψηλή, ενώ για τα σημεία τα οποία βρίσκονται μακριά, η $p_{j|i}$ θα είναι σχεδόν μηδέν. Η υπό συνθήκη πιθανότητα $p_{j|i}$ δίνεται από τον τύπο

$$p_{j|i} = \frac{exp(\frac{-||x_i - x_j||^2}{2\sigma_i^2})}{\sum_{k \neq i} exp(\frac{-||x_i - x_j||^2}{2\sigma_i^2})}$$

όπου το $\sigma_i$ είναι διακύμανση της κατανομής με μέση τιμή το ημιάθροισμα των συντεταγμένων του σημείου $x_i$. Επειδή ενδιαφερόμαστε μόνο για την μοντελοποίηση ανα ζεύγη, θέτουμε την τιμή της $p_{i|i}$ να είναι μηδέν. Για τα μειωμένης διάστασης αντίγραφα $y_i$ και $y_j$ των σημείων των δεδομένων $x_i$ και $x_j$, είναι πιθανό να υπολογιστεί μια παρόμοια υπό συνθήκη πιθανότητα, την οποία ορίζουμε ως $q_{j|i}$. Θέτουμε την διακύμανση της Gaussian κατανομής $\frac{1}{\sqrt{2}}$ για τα μειωμένης διάστασης αντίγραφα $y_i$ και $y_j$. Θέτοντας μια άλλη τιμή, από την πραγματική, για την διακύμανση των αντιγράφων έχουμε τα δεδομένα με διαφορετική κλίμακα τίποτα παραπάνω, στην απεικόνιση των δεδομένων δεν αλλάζει τίποτα. Ως εκ τούτου, μοντελοποιούμε την ομοιότητα του απεικονισμένου σημείου $y_j$ με το σημείο $y_i$ με την σχέση

$$q_{j|i} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k \neq i} exp(-||y_i - y_k||^2)}$$

Επειδή, πάλι ενδιαφερόμαστε να μοντελοποιήσουμε ζευγαρωτές ομοιότητες, θέτουμε $q_{i|i} = 0$. Αν τα απεικονισμένα σημεία $y_j$ και $y_i$ μοντελοποιούν σωστά την ομοιότητα ανάμεσα στα σημεία των δεδομένων υψηλής διάστασης $x_i$ και $x_j$, τότε οι υπό συνθήκη πιθανότητες $p_{j|i}$ και $q_{j|i}$ πρέπει να είναι ίσες. Βάση αυτής της παρατήρησης η τεχνική t-SNE στοχεύει στο να βρει μια μειωμένης διάστασης απεικόνιση, η οποία ελαχιστοποιεί την αναντιστοιχεία μεταξύ $p_{j|i}$

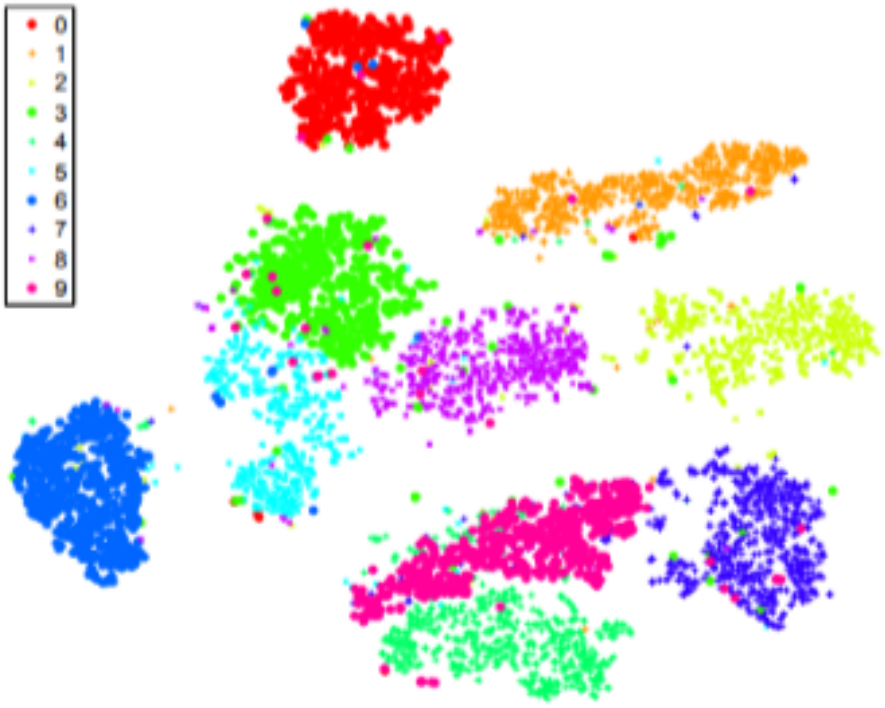
και $q_{j|i}$. Η t-SNE ελαχιστοποιεί το άθροισμα της απόκλισης των Kullback-Leibler για όλα τα σημεία των δεδομένων. Η συνάρτηση του σφάλματος $C$ δίνεται από την σχέση

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} log \frac{p_{j|i}}{q_{j|i}} (2.13.1)$$

Κάθε συγκεκριμένη τιμή της $\sigma_i$ υποδεικνύει μια κατανομή πιθανότητας, $P_i$ για όλα τα σημεία των δεδομένων. Το $Q_i$ αντιπροσωπεύει την κατανομή πιθανότητας για όλα τα σημεία δεδομένων, στην μειωμένη διάσταση, δοθέντος του σημείου $y_i$. Επειδή η συνάρτηση της απόκλισης C δεν είναι συμμετρική δίνει μεγάλο σφάλμα για την χρήση πολύ διάσπαρτων σημείων για να απεικονίσουν διπλανά σημεία των δεδομένων, αλλά υπάρχει μικρό σφάλμα στην χρήση διπλανών σημείων για να απεικονίσουν πολύ διαχωρισμένα σημεία των δεδομένων. Με άλλα λόγια, η συνάρτηση σφάλματος της t-SNE επικεντρώνεται στην διατήρηση της τοπικής δομής των δεδομένων στην απεικόνιση. Η παράμετρος που πρέπει να επιλεχθεί από τον χρήστη είναι η διακύμανση του $\sigma_i$, αν ο χρήστης επιλέξει μεγάλη τιμή για την διακύμανση τότε ως γείτονες του σημείου $x_i$ θα θεωρηθούν και σημεία τα οποία βρίσκονται μακριά από αυτό. Αντίθετα, αν ο χρήστης επιλέξει μικρή τιμή για την διακύμανση τότε μόνο σημεία που είναι κοντά στο $x_i$ θα θεωρηθούν ως γειτονικά [17]. Η απόδοση του t-SNE είναι αρκετά ισχυρή για αλλαγές στην διακύμανση και συχνά χρησιμοποιούνται οι τιμές μεταξύ 5 και 50.

Ας δούμε τώρα την οπτικοποίηση ενός γνωστού συνόλου δεδομένων με την t-SNE, του MNIST. Το σύνολο δεδομένων MNIST αποτελείται από 60000 εικόνες ψηφίων γραμμένα με το χέρι τα οποία, αφού κεντρικοποιηθούν σε εικόνες $28x28 pixel$ και μετασχηματιστούν σε κλίμακες του γκρι εισάγονται σε έναν πίνακα. (http://yann.lecun.com/exdb/mnist/index.html). Για το πείραμά μας θα επιλέξουμε τυχαία 6000 εικόνες για υπολογιστικούς λόγους. Οι ψηφιακές εικόνες έχουν $28x28 = 784$ εικονοστοιχεία (διαστάσεις).

Στο Σχήμα 13, η t-SNE παράγει μια απεικόνιση στην οποία ο διαχωρισμός μεταξύ των κατηγοριών είναι σχεδόν τέλειος. Η λεπτομερής απεικόνιση της μεθόδου t-SNE αποκαλύπτει το μεγαλύτερο μέρος της δομής των δεδομένων. Η απεικόνιση που παράγεται από την τεχνική t-SNE περιέχει λίγα σημεία που ομαδοποιούνται λάθος, αλλά σε γενικές γραμμές είναι μια καλή απεικόνιση.




Σχήμα 10: Οπτικοποίηση του συνόλου δεδομένων **MNIST** με την $t-SNE$ [17].

# 4 Τεχνικές Οπτικοποίησης

## 4.1 Εισαγωγή

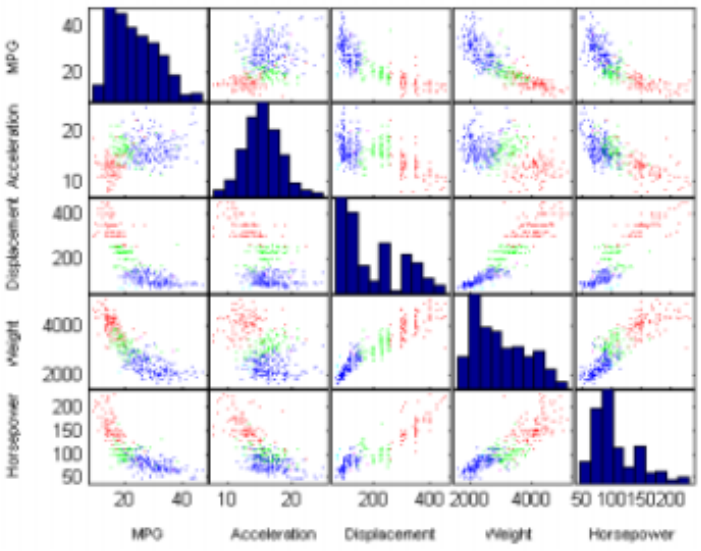
Ο **Keim** και **Kriegel** το 1996 χώρισαν τις τεχνικές οπτικοποίησης των πολυδιάστατων δεδομένων σε έξι κατηγορίες: γεωμετρικές τεχνικές, τεχνικές εικονογραφημάτων ή εικονογραφικές τεχνικές, ιεραρχικές τεχνικές, τεχνικές εικονοστοιχείων και τις υβριδικές τεχνικές [42]. Θα υιοθετήσουμε αυτήν την ταξινόμηση και θα την προσαρμόσουμε στις τεχνικές οπτικοποίησης πολυδιάστατων δεδομένων, οι οποίες ταξινομούνται σε τέσσερις ευρείες κατηγορίες σύμφωνα με την γενική αντίληψη. Οι κατηγορίες αυτές είναι: οι γεωμετρικές προβολές, τεχνικές που βασίζονται στα εικονοστοιχεία, οι ιεραρχικές και οι εικονογραφικές. Σε αυτό το κεφάλαιο θα παρουσιάσουμε τις βασικές τεχνικές για κάθε μια από τις κατηγορίες της.

## 4.2 Γεωμετρικές Τεχνικές

Οι τεχνικές αυτής της κατηγορίας αποδίδουν καλύτερα στην ανίχνευση ακραίων τιμών και των συσχετίσεων μεταξύ των διαστάσεων. Ακόμα με αυτές μπορούμε να αντιμετωπίσουμε μεγάλο όγκο δεδομένων όταν χρησιμοποιηθούν οι κατάλληλες διαδραστικές τεχνικές. Η αδυναμία των τεχνικών αυτής της κατηγορίας είναι ότι ενώ τα χαρακτηριστικά αντιμετωπίζονται ισάξια. Ένα επιπλέον μειονέκτημα των τεχνικών αυτών είναι ότι η συσσώρευση των δεδομένων σε έναν περιορισμένο χώρο είναι κάτι που προκαλεί μια χαοτική απεικόνιση, ιδιαίτερα όταν τα δεδομένα είναι μεγάλης διάστασης ή μεγάλου όγκου. Ας δούμε όμως τις επικρατέστερες από αυτές τις τεχνικές.

### 4.2.1 Πίνακες Διαγραμμάτων Διασποράς

Οι πίνακες διαγραμμάτων διασποράς (**Scatterplot Matrix**) χρησιμοποιούνται για κατηγορικά δεδομένα στα οποία προβάλλονται δύο χαρακτηριστικά στους άξονες **x-y** των καρτεσιανών συντεταγμένων [57]. Οι πίνακες διαγραμμάτων διασποράς είναι μια επέκταση των διαγραμμάτων διασποράς για δεδομένα μεγάλης διάστασης, όπου μια συλλογή από διαγράμματα διασποράς είναι ταξινομημένα στο επίπεδο ώστε να είναι ταυτόχρονα ορατές οι συσχετίσεις μεταξύ των διάφορων χαρακτηριστικών (Σχήμα 11).

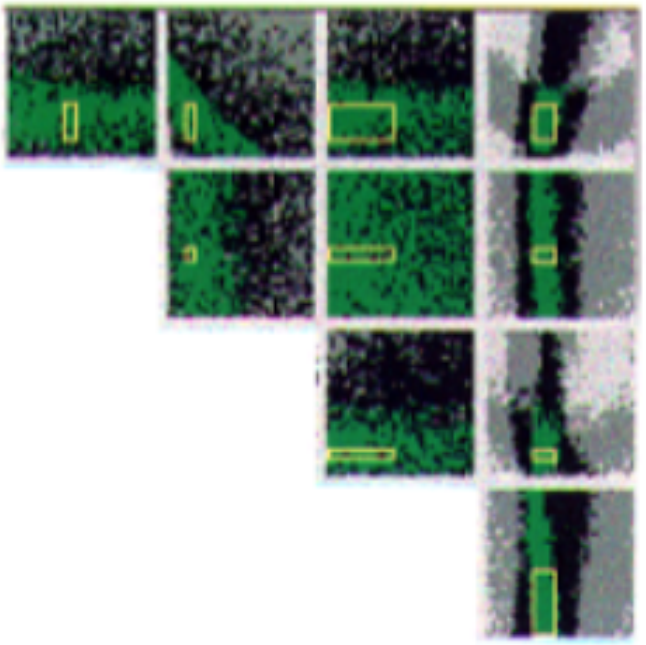



Σχήμα 11: Πίνακας διαγραμμάτων διασποράς για δεδομένα διάστασης 5, 400 αυτοκινήτων του συνόλου δεδομένων mtcars [66].

Ένας βασικός περιορισμός των πινάκων των διαγραμάτων διασποράς είναι ότι η απεικόνιση γίνεται χαοτική όταν ο αριθμός των σημείων, τα οποία αντιπροσωπεύουν τα δεδομένα, είναι πάρα πολλά. Αυτό το πρόβλημα μπορεί να λυθεί χρησιμοποιώντας την τεχνική του **brushing**, η οποία είναι μια τεχνική που συνδιάζει την διαδραστική απεικόνιση κατά την οποία ο χρήστης έχει την δυνατότητα επιλέγοντας κάποιο σημείο του διαγράμματος διασποράς να δει αυτόματα τις αλλαγές που συντελούνται και στις άλλες απεικονίσεις.

### 4.2.2 Πίνακας Προβολών

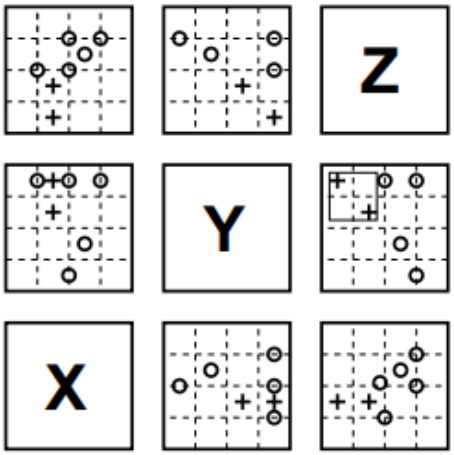
Οι Πίνακες Προβολών (**Projection Matrix**) προτάθηκαν από τους **Tweedie** και **Spence** οι οποίοι μπορούν να χρησιμοποιηθούν και για δεδομένα μεγάλης διάστασης [30]. Ο Πίνακας Προβολών είναι ένας πίνακας που περιγράφει τις συσχέτισεις μεταξύ των μεταβλητών, όπου τα σημεία κάθε διάστασης είναι χρωματισμένα με άλλο χρώμα (Σχήμα 12).



Σχήμα 12: Ένας πίνακας προβολλών για το σύνολο δεδομένων mtcars, που μας δείχνει την συσχέτιση κάθε μεταβλητής με τις υπόλοιπες, χρησιμοποιώντας διαφορετικό χρώμα για κάθε μεταβλητή [68].

### 4.2.3 **Brusing**

Η τεχνική του **Brusing** παρουσιάστηκε το 1987 [53]. Συμπεριλαμβάνεται ως μια από τις τεχνικές άμεσου χειρισμού των δεδομένων [3]. Υπάρχουν δύο είδη **Brusing** για τα διαγράμματα διασποράς, η επισήμανση και η ενισχυμένη σύνδεση. Η επισήμανση περιέχει ένα διαδραστικό **Brusing** (π.χ. ένα σημείο επισήμανσης του ποντικιού) το οποίο παίρνει και διαφορετική επισήμανση ανάλογα με την μεταβλητή που απεικονίζει. Στην ενισχυμένη σύνδεση, το **Brusing** είναι ένα ρυθμιζόμενο ορθογώνιο. Χρησιμοποιείται για την κάλυψη ενός συνόλου σημείων, τα οποία βρίσκονται σε μια περιοχή. Η παρακάτω εικόνα (Σχήμα 14) δείχνει ένα ορθογώνιο **Brusing** στο **panel** (2,3), όπου τα δεδομένα μέσα στο ορθογώνιο απεικονίζονται με το σύμβολο + αντί του 0.

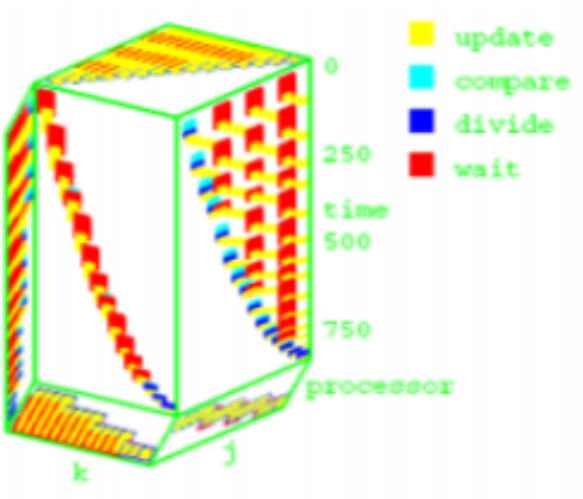



Σχήμα 13: Η τεχνική της ενισχυμένης σύνδεσης [12].

Οι ίδιες αλλαγές εφαρμόζονται και στα αντίστοιχα σημεία των δεδομένων στα άλλα **panels**. Κοιτώντας σε διαφορετικά **panels** συγκρίνοντας την οριζόντια και κάθετη έκταση του **Brusing**, έχουμε μια ενισχυμένη τεχνική σύνδεσης η οποία είναι ένα ισχυρό εργαλείο απευθείας χειρισμού της απεικόνισης. Αυτό δείχνει ότι η επίδραση του **Brusing** είναι πιο αποτελεσματική σε μια δυναμικά διαδραστική απεικόνιση. Γενικά το **Brusing** μπορεί να χρησιμοποιηθεί με μια πλειάδα τεχνικών οπτικοποίησης.

### 4.2.4 Υπερθηκογράμματα

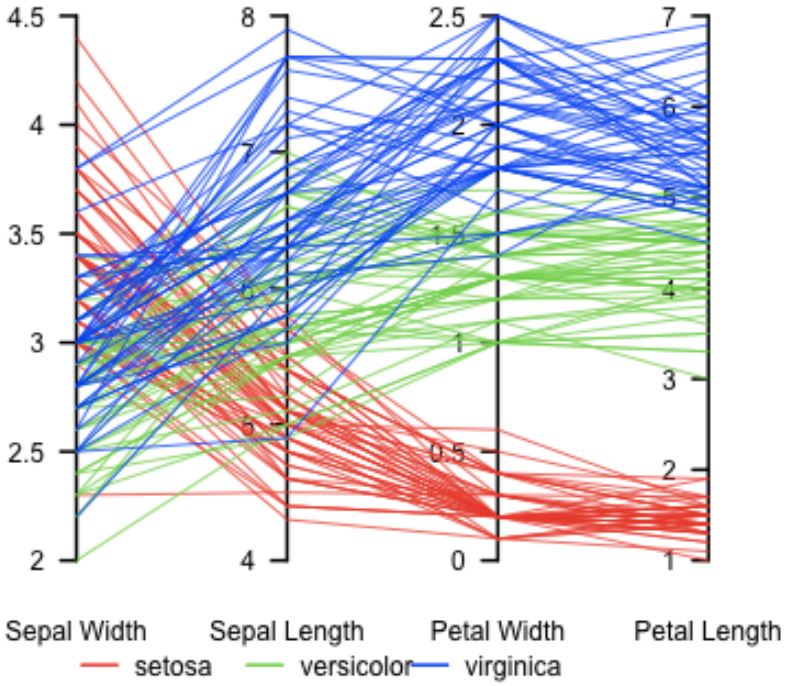
Τα Υπερθηκογράμματα (**Hyperboxes**) είναι κύβοι $n$-διάστασης οι οποίοι απεικονίζονται στις δύο διαστάσεις [1]. Το κάθε χαρακτηριστικό των δεδομένων απεικονίζεται σε μια από τις $n$ διάστασης του κύβου. Τα υπερθηκογράμματα είναι ένα ισχυρό εργαλείο αφού μπορούμε να δούμε την συσχέτιση κάθε χαρακτηριστικού με τα άλλα χαρακτηριστικά που έχουν τα δεδομένα μας. Το μειονέκτημα τους είναι ότι ο προσανατολισμός και το μήκος τον πλευρών τους είναι τυχαία και έτσι μπορεί να εξαχθούν λάθος πληροφορείες (Σχήμα 14).




Σχήμα 14: Υπερθηκόγραμμα, σε κάθε πλευρά έχουμε την συσχέτιση δύο μεταβλητών. [1].

### 4.2.5 Παράλληλες Συντεταγμένες

Οι Παράλληλες Συντεταγμένες (**Parallel Cootdinates**) είναι μια πολύ γνωστή μέθοδος οπτικοποίησης δεδομένων μεγάλης διάστασης στην οποία τα διάφορα χαρακτηριστικά απεικονίζονται ως παράλληλες ευθείες, το πλήθος των οποίον είναι ίσο με την διάσταση των δεδομένων που μελετάμε [9]. Μπορούν να χρησιμοποιηθούν για να μελετήσουμε την συσχέτιση ανάμεσα σε διάφορα χαρακτηριστικά χρησιμοποιώντας τα σημεία που τέμνονται. Επίσης μπορούν να χρησιμοποιηθούν για να δούμε την κατανομή που ακολουθεί το κάθε χαρακτηριστικό. Μειονέκτημα αυτής της τεχνικής είναι το χάος που δημιουργείται στην απεικόνιση όταν έχουμε δεδομένα πολύ υψηλής διάστασης. Σε αυτήν την περίπτωση μπορούμε να χρησιμοποιήσουμε την τεχνική του **brushing** για να βοηθηθεί η ερμηνεία της απεικόνισης.



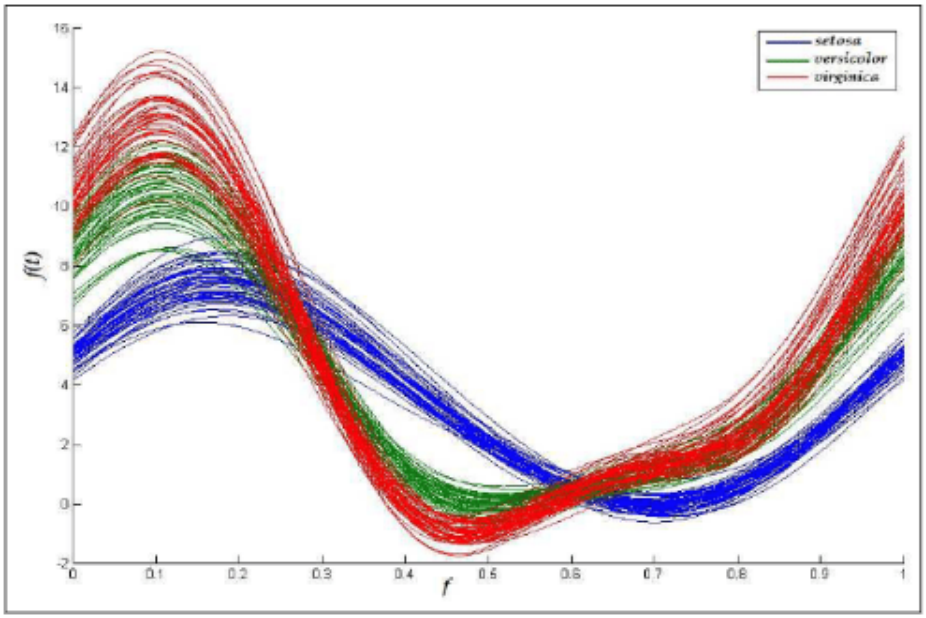

Σχήμα 15: Παράλληλες Συντεταγμένες για το σύνολο δεδομένων **Iris** [65].

### 4.2.6 Οι Καμπύλες του **Andrews**

Οι Καμπύλες του Andrews (Andrews Curves) είναι μια μέθοδος οπτικοποίησης δεδομένων μεγάλης διάστασης, η οποία απεικονίζει κάθε παρατήρηση χρησιμοποιώντας μετασχηματισμούς Fourier [2]. Κάθε πολυδιάστατο διάνυσμα $y = (y_1, y_2, ...., y_k)^T$ με k μεταβλητές μπορεί να αναπαρασταθεί στο επίπεδο με την μορφή μίας καμπύλης η οποία ορίζεται από την ακόλουθη συνάρτηση:

$$f_y(t) = y_1/\sqrt{2} + y_2 sin(t) + y_3 cos(t) + y_4 sin(2t) + y_5 cos(2t) + ...$$

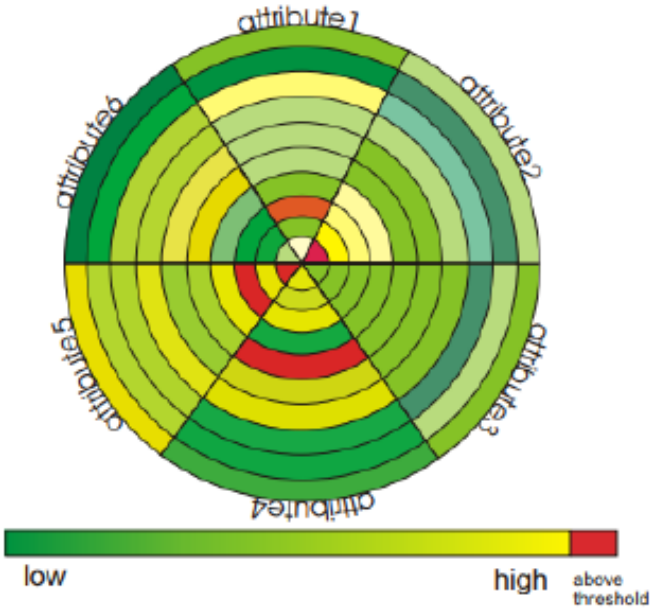
με -π $\leq t \leq$ π. Με αυτήν την μέθοδο μπορούν να εντοπιστούν ομαδοποιήσεις των δεδομένων αλλά και ακραίες τιμές. Αυτή η μέθοδος οπτικοποίησης έχει κάποια σημαντικά μειονεκτήματα. Η ανάγνωση του γραφήματος δυσκολεύει όταν έχουμε μεγάλο αριθμό παρατηρήσεων. Ακόμα η ομαδοποίηση των δεδομένων είναι υποκειμενική και ο κάθε χρήστης μπορεί να καταλήγει στην δική του ομαδοποίηση. Τέλος, έχει μεγάλη σημασία η σειρά η οποία μπαίνουν στην συνάρτηση οι μεταβλητές, αφού μεταβλητές οι οποίες εισέρχονται πρώτες έχουν μεγαλύτερη βαρύτητα από αυτές που έπονται.Άρα θα πρέπει πριν χρησιμοποιηθεί η μέθοδος να γίνει κάποια αξιολόγηση των μεταβλητών.



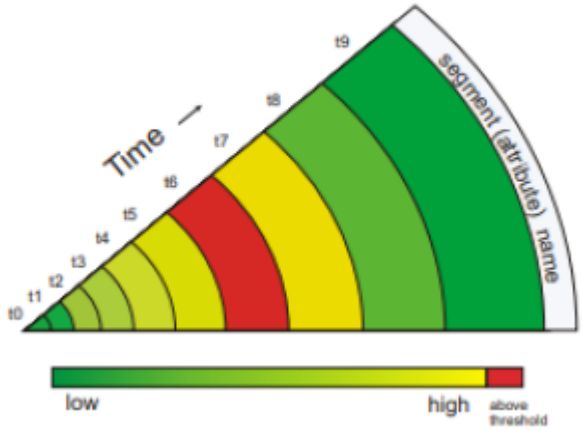

Σχήμα 16: Καμπύλες του Andrew για το σύνολο δεδομένων Iris.

### 4.2.7 Προβολή στον Κύκλο (**Circle Projection**)

Η βασική ιδέα είναι να διαιρέσουμε τον κύκλο σε έναν αριθμό τμημάτων, ανάλογα με τον αριθμό των διαστάσεων του συνόλου των δεδομένων. Στην συνέχεια, κάθε τμήμα χωρίζεται σε υποτμήματα με σκοπό την κατανομή των αλλαγών των δεδομένων που εξαρτώνται από τον χρόνο. Η παρακάτω εικόνα (Σχήμα 17) δείχνει την βασική ιδέα της απεικόνισης της κυκλικής προβολής. Ο κύκλος στην εικόνα (Σχήμα 17) χωρίζεται σε 6 τομείς ανάλογα με την επιθυμία του κάθε χρήστη. Στην συνέχεια, καθένα από τα τμήματα χρωματίζεται με κάποια απόχρωση για να απεικονισθούν και να συγκριθούν οι αλλαγές των τιμών των χαρακτηριστικών σύμφωνα με τον χρόνο. Το χρώμα κάθε υποπεριοχής παρουσιάζει την συνολική τιμή ενός χαρακτηριστικού σε ένα συγκεκριμένο χρονικό σημείο. Το παράδειγμα δείχνει πολύ καθαρά, ότι ένας χρήστης μπορεί εύκολα να συγκρίνει κάθε στιγμή ένα τμήμα του κύκλου με κάθε αντίστοιχο γειτονικό τμήμα. Αυτό είναι πολύ χρήσιμο για την απεικόνιση των συσχετίσεων στα δεδομένα και την σύγκριση διαφορετικών χαρακτηριστικών σε ένα συγκεκριμένο χρονικό σημείο. Στην παρακάτω εικόνα (Σχήμα 17) μπορούμε να δούμε ένα παράδειγμα δομής ενός τμήματος κυκλικής προβολής. Κάθε τμήμα αντιπροσωπεύει το μέγεθος ενός χαρακτηριστικού σε μια συγκεκριμένη χρονική στιγμή. Ο αριθμός των χρονικών στιγμών $n$ και η μονάδα μέτρησης του χρόνου εξαρτάται από το σενάριο εφαρμογής και μερικές φορές από τις απαιτήσεις των χρηστών. Για παράδειγμα, ένας τομέας χωρίζεται σε 24 τμήματα, ένα για κάθε ώρα και υποθέτουμε ότι ο αναλυτής θα διαλέξει μια συγκεκριμένη ώρα. Διάφορες κατατάξεις των τμημάτων του κύκλου είναι δυνατές. Ο σκοπός της κυκλικής προβολής είναι να δείξει από την μια πλευρά τα παρελθοντικά και τα παρόντα χρονικά συμβάντα την ίδια στιγμή, αλλά από την άλλη πλευρά τα παρόντα χρονικά συμβάντα είναι πιο σημαντικά. Η τεχνική της κυκλικής προβολής παρέχει παραπάνω χώρο για τα χρονικά διαστήματα των παρόντων χρονικών συμβάντων (Σχήμα 19). Σημείωση, ότι στην παρακάτω εικόνα (Σχήμα 20) δείχνονται επίσης διαφορετικές διατάξεις για τα παρόντα και τα παρελθόντα γεγονότα.




Σχήμα 17: Κυκλική προβολή που δείχνει την εξέλιξη των χαρακτηριστικών στον χρόνο [45].

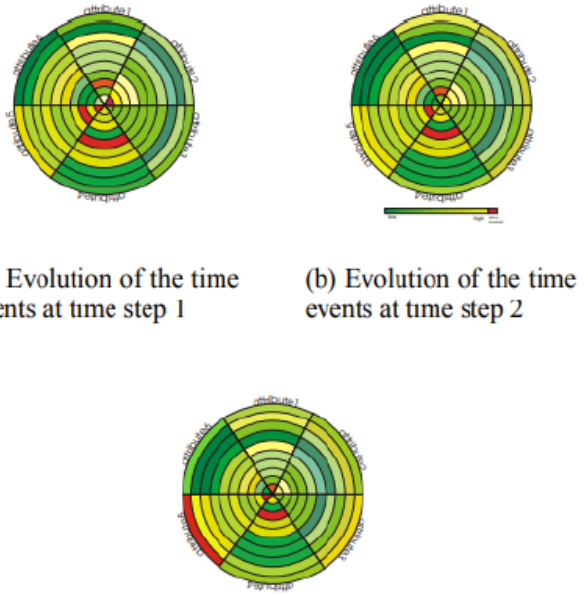



Σχήμα 18: Η εξέλιξη των γεγονότων στον χρόνο από το κέντρο στην άκρη [45].

Η προβολή στον κύκλο υποστηρίζει τις ακόλουθες δυνατότητες. Ο χρήστης μπορεί να τροποποιήσει την απεικόνιση των δεδομένων στην οθόνη, η διεπαφή προβολής κύκλου διαθέτει χειροκίνητες και αυτοματοποιημένες τεχνικές πλοήγησης. Η επιλογή αυτή παρέχει στους χρήστες τη δυνατότητα απομόνωσης ενός υποσυνόλου των δεδομένων για λειτουργίες όπως η επισήμανση, το φιλτράρισμα και η ποιοτική ανάλυση. Η διάταξη των τμημάτων δια-

δραματίζει σημαντικό ρόλο στην εξεύρεση συσχετίσεων σε σύνολα δεδομένων που μπορεί να σχετίζονται με τον χρόνο. Από αντιληπτική άποψη, είναι ευκολότερο να συγκρίνουμε τμήματα τα οποία είναι το ένα δίπλα στο άλλο. Το μάτι του χρήστη μπορεί να συγκρίνει απευθείας τις γειτονικές περιοχές, καθώς όλα τα τμήματα που περιέχονται στον ίδιο κύκλο έχουν τον ίδιο αριθμό και το ίδιο μήκος υποτομέων. Επομένως, ο στόχος είναι να βρεθεί μια κυκλική διάταξη που τοποθετεί παρόμοια τμήματα το ένα δίπλα στο άλλο. Για να αποκτήσουμε μια τέτοια χρήσιμη διάταξη, πρέπει να οριστεί η ομοιότητα ανάμεσα στους φορείς. Σημειώστε ότι είναι πολύ πιθανό, η σειρά των τμημάτων να αλλάζει από την στιγμή που εισέρχονται νέα δεδομένα. Εκτός από τον υπολογισμό μίας σειράς τμημάτων που βασίζεται στην ομοιότητα, ο χρήστης έχει την δυνατότητα να αλλάξει την θέση των τμημάτων χειροκίνητα, για να μπορούν να συγκριθούν οποιαδήποτε τμήματα της προβολής του κύκλου.

Για να αναλύσουμε τα δεδομένα που σχετίζονται με τον χρόνο, η κυκλική προβολή οπτικοποιεί τις αλλαγές των χαρακτηριστικών στο χρόνο. Η παρακάτω εικόνα (Σχήμα 19) δείχνει ένα παράδειγμα με τρία διαφορετικά σημεία του χρόνου. Στο παράδειγμα τα χρονικά τμήματα μετατοπίζονται στο κέντρο του κύκλου και νέα χαρακτηριστικά εισάγονται στην άκρη του κύκλου. Η οπτικοποίηση πρέπει να ανανεώνεται κάθε φορά που υπάρχουν νέα δεδομένα.



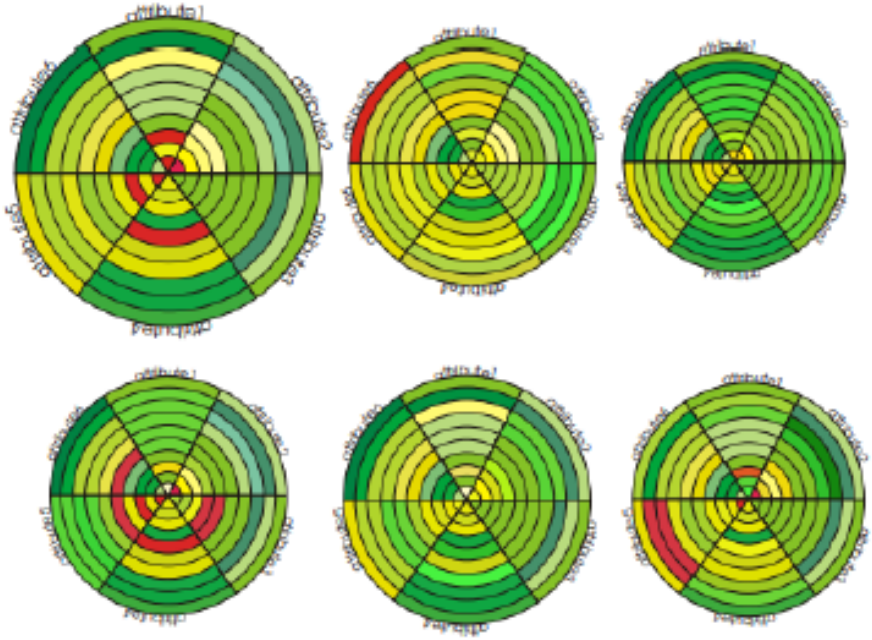
(a) Evolution of the time events at time step 1

(b) Evolution of the time events at time step 2

(c) Evolution of the time events at time step 3

Σχήμα 19: Κυκλική προβολή, συνεχής ροής δεδομένων στον χρόνο [45].

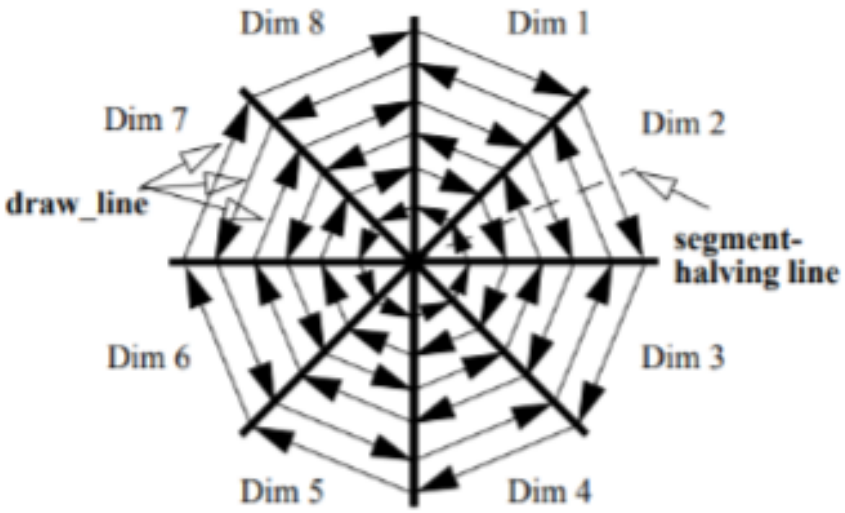
Μια επάκταση της κυκλική προβολής είναι η πολυμεταβλητή κυκλική προβολή μέσω της οποίας μπορεί να γίνει κατηγοριοποίηση των δεδομένων σε πολλαπλές προβολές κύκλων. Αυτό καθιστά δυνατή την παρακολούθηση ομάδων παρόμοιων χαρακτηριστικών. Μια τέτοια τεχνική είναι χρήσιμη για τους αναλυτές δεδομένων που αναζητούν τοπικά πρότυπα σε διαφορετικές υποκατηγορίες χαρακτηριστικών. Η τεχνική αυτή δείχνεται στο Σχήμα 20. Ως χαρακτηριστικό γνώρισμα, η διάμετρος κάθε κύκλου αντιπροσωπεύει το βάρος (π.χ. τη σημασία) της ομάδας των χαρακτηριστικών. Οι κύκλοι με υψηλότερα βάρη έχουν μεγαλύτερη διάμετρο και περισσότερο χώρο στην οθόνη ενώ κύκλοι με χαμηλότερα βάρη συρρικνώνονται.



Σχήμα 20: Συνδιασμός πολλαπλών κύκλων προβολών που δείχνει την εξέλιξη των πολλαπλών χαρακτηριστικών στον χρόνο [45].

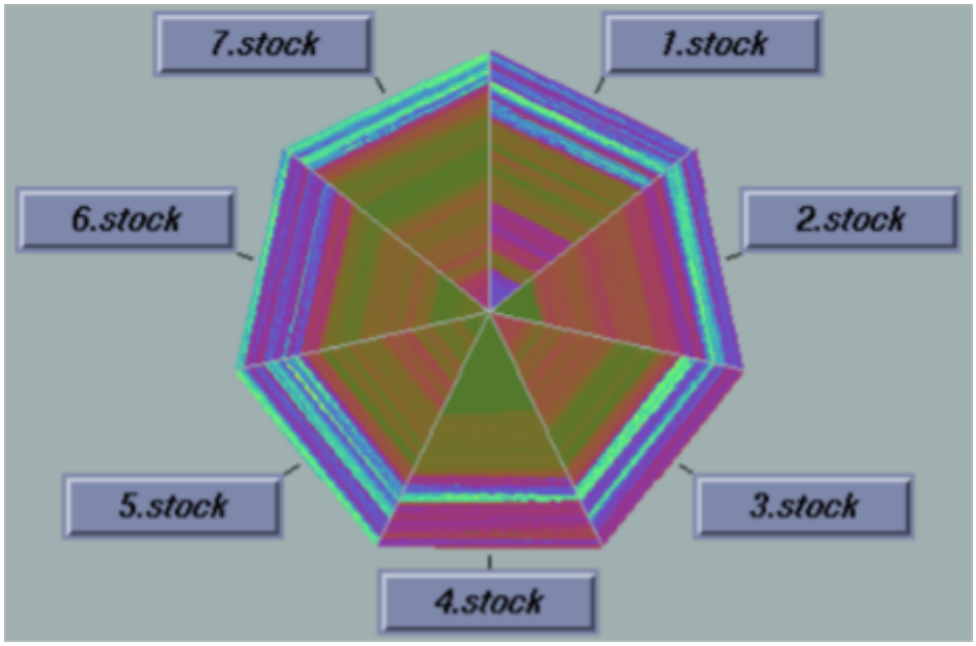
### 4.2.8 Κυκλικοί Τομείς

Οι Κυκλικοί Τομείς (**Circle Segments**) είναι μια τεχνική για οπτικοποίηση δεδομένων μεγάλης διάστασης. Η τεχνική αυτή χρησιμοποιεί έναν κυκλικό τμήμα για κάθε μεταβλητή των δεδομένων [45]. Η βασική ιδέα αυτής της τεχνικής είναι να απεικονίζει τις διαστάσεις ως τμήματα του κύκλου. Αν τα δεδομένα αποτελούνται από $k$ διαστάσεις, ο κύκλος χωρίζεται σε $k$ τμήματα που το κάθε ένα αντιπροσωπεύει μια διάσταση των δεδομένων. Μέσα στα τμήματα, οι τιμές των δεδομένων που ανήκουν σε μια διάσταση είναι διατεταγμένες από το κέντρο του κύκλου και προς τα έξω. Τα πρώτα αποτελέσματα δείχνουν ότι η τεχνική των τμημάτων του κύκλου είναι μια πολύ ισχυρή μέθοδος για την οπτικοποίηση μεγάλου όγκου δεδομένων, παρέχοντας πιο εκφραστικές απεικονίσεις από άλλες γνωστές τεχνικές. Τα στοιχεία των δεδομένων μέσα σε ένα τμήμα είναι διατεταγμένα με **draw lines**, οι οποίες είναι γραμμές που δείχνουν την κατεύθυνση της μεταβλητής σε κάθε διάσταση (Σχήμα 21).



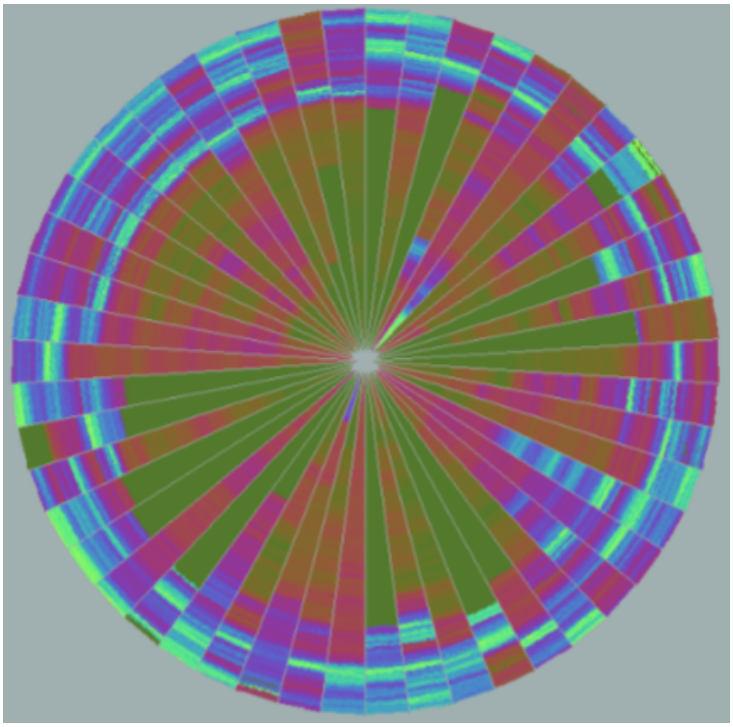

Σχήμα 21: Η τεχνική των κυκλικών τομέων για δεδομένα διάστασης οκτώ [45].

Αυτή η τεχνική απαιτεί το σύνολο δεδομένων να αποτελείται από τουλάχιστον τρεις διαστάσεις. Ένα επιλπλέον χαρακτηριστικό αυτής της τεχνικής είναι ότι μπορεί να γίνει αντιστοίχιση των διαστάσεων στα τμήματα του κύκλου από τον χρήστη. Αυτή η δυνατότητα είναι πολύ σημαντική επειδή μπορεί ο χρήστης να αλλάζει την σειρά των διαστάσεων και τον βοηθά να τις συγκρίνει αλλά και να τις ομαδοποιεί.
Ας δούμε τώρα δύο παραδείγματα. Στην παρακάτω εικόνα (Σχήμα 22) χρησιμοποιούμε την τεχνική του κυκλικού τομέα για να απεικονίσουμε δεδομένα μετοχών δέκα χρόνων (5328 εγγραφές) του χρηματιστηρίου της Φραγκφούρτης. Όταν χρησιμοποιούμε αυτήν την τεχνική στα ίδια δεδομένα, τα παλιότερα δεδομένα είναι στην μέση του κύκλου και τα πιο πρόσφατα στα εξωτερικά τμήματα του κύκλου. Οι υψηλές τιμές των δεδομένων χρωματίζονται με φωτείνες αποχρώσεις ενώ οι μη υψηλές με πιο σκοτεινές αποχρώσεις, έτσι ο χρήστης μπορεί να έχει μια διαίσθηση των δεδομένων. Μπορούμε ξεκάθαρα να δούμε ότι η πέμπτη, η έκτη και η έβδομη μετοχή ότι έχουν υψηλότερες τιμές στο τέλος της εποχής της απεικόνισης. Επιπλέον, με αυτήν την τεχνική ο χρήστης είναι σε θέση να παρακολουθήσει τις τιμές των μετοχών και να ανιχνευθούν διάφορες τάσεις μεταξύ των διαστάσεων. Για παράδειγμα η πρώτη και η δεύτερη μετοχή δείχνουν μια παρόμοια συμπεριφορά στα τελευταία χρονικά διαστήματα. Παρόμοια, στο πρώτο μισό της χρονικής περιόδου η τρίτη και τέταρτη μετοχή έχουν υψηλή τιμή και την ίδια στιγμή η τιμή της τρίτης μετοχής μένει σε υψηλή τιμή για λίγο ακόμα.




Σχήμα 22: Οπτικοποίηση δεδομένων διάστασης εφτά [45].

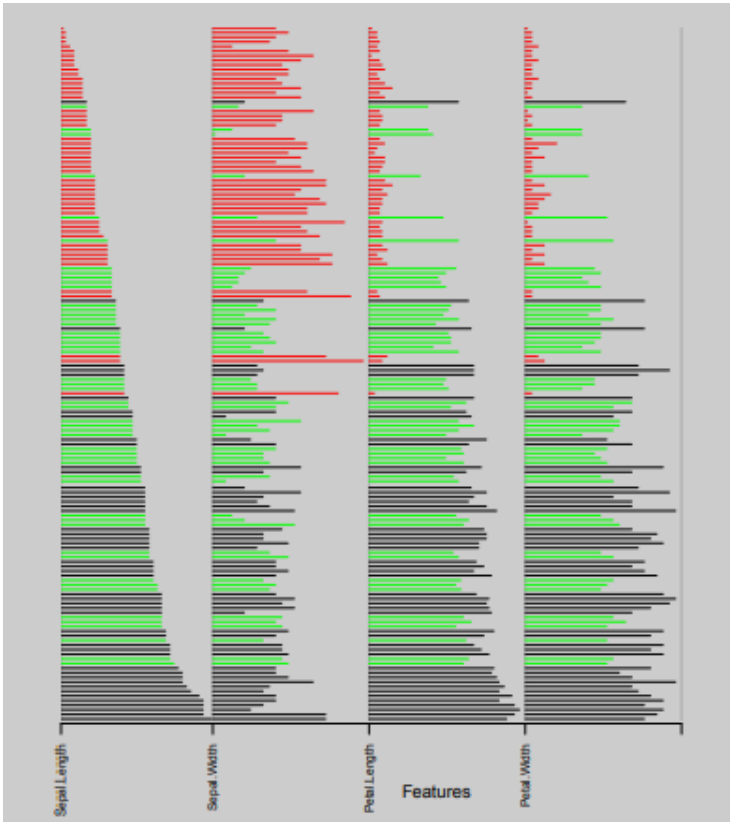
Στο Σχήμα 23 μπορούμε να δούμε πως δουλεύει η τεχνική των κυκλικών τομέων για τις τιμές 50 μετοχών του χρηματηστηρίου της Φραγκφούρτης, αντιπροσωπεύοντας 265000 τιμές δεδομένων. Λόγω, του υψηλού βαθμού επικάλυψης τα γραφήματα γραμμής δεν είναι κατάλληλα για την οπτικοποίηση τόσο πολλών διαστάσεων και για αυτό χρησιμοποιούμε τις τεχνικές σπειροειδούς και αναδρομικού προτύπου. Το κύριο πλεονέκτημα των νέων αυτών τεχνικών είναι η συνολική αναπαράσταση των δεδομένων και έτσι γίνονται καλύτερα κατανοητά.



Σχήμα 23: Απεικονίζονται 265000 τιμές δεδομένων διάστασης 50 με την τεχνική των κυκλικών τομέων [45].

### 4.2.9 Τοπογραφικά Διαγράμματα

Τα τοπογραφικά διαγράμματα (**Survey Plots**) είναι μια απλή τεχνική που μας βοηθάει να ανακαλύψουμε συσχετίσεις μεταξύ δύο μεταβλητών [34]. Μια απλή παραλλαγή αυτού επεκτείνει μια γραμμή από ένα κεντρικό σημείο, όπου το μήκος της γραμμής αντιστοιχεί στην τιμή της μεταβλητής στην κάθε διάσταση. Αυτή η συγκεκριμένη απεικόνιση των δεδομένων των $n$ διαστάσεων επιτρέπει να δούμε τους συσχετισμούς μεταξύ οποιωνδήποτε δύο μεταβλητών ειδικά όταν τα δεδομένα ταξινομούνται σύμφωνα με μια συγκεκριμένη διάσταση. Η χρήση χρώματος για διαφορετικές ομαδοποιήσεις μπορεί να βοηθήσει να καθοριστούν καλύτερα οι συντεταγμένες για να την ομαδοποίηση των δεδομένων.




Σχήμα 24: Ένα **Survey Plot** για το σύνολο δεδομένων **Iris** [4].

### 4.2.10 Εξερευνητική Οπτικοποίηση

Η Εξερευνητική Οπτικοποίηση (*Viz3D*) είναι μια τεχνική οπτικοποίησης η οποία παράγει αναπαραστάσεις δεδομένων μεγάλης διάστασης στον τρισδιάστατο χώρο, την οποία μπορούν να χειριστούν διαδραστικά οι χρήστες για να αντιμετωπίσουν την υπερβολική συσσώρευση [56]. Η τεχνική των τρισδιάστατων προβολών παρέχει πληροφόρηση για τα δεδομένα, δίνοντας στον χρήστη μεγαλύτερο έλεγχο της οπτικής αναπαράστασης χρησιμοποιώντας μια ακόμη διάσταση, η οπτική συσσώρευση εξακολουθεί να αποτελεί πρόβλημα για τον χειρισμό δεδομένων μεγάλης διάστασης.

Εδώ, προτείνουμε μια τεχνική απεικόνισης δεδομένων μεγάλης διάστασης, η οποία προβάλλει τα δεδομένα στον τρισδιάστατο χώρο, παρέχοντας μια απεικόνιση η οποία μπορεί να είναι διαδραστική και επιτρέπει στους χρήστες να διαχειριστούν την συσσώρευση των σημείων. Αυτή η τεχνική που ονομάζεται *Viz3D* μπορεί να οπτικοποιεί συστοιχίες που υπάρχουν στα δεδομένα. Η *Viz3D* είναι μια τεχνική που διατηρεί την ικανότητα να αποκαλύπτει ομάδες, ενώ αντιμετωπίζει την συσσώρευση των παρατηρήσεων πιο αποτελεσματικά [56]. Τα δεδομένα προβάλλονται στην επιφάνεια και στο εσωτερικό ενός τρισδιάστατου κύκλου (Σχήμα 25), δημιουργώντας μια αναπαράσταση που φιλοξενεί έναν μεγάλο αριθμό σημείων και προσφέρει μια κατάλληλη αναπαράσταση για την παρουσίαση δεδομένων μεγάλων διαστάσεων. Πολλές καταστάσεις συσσώρευσης μπορούν να αντιμετωπιστούν με αυτήν την τεχνική.

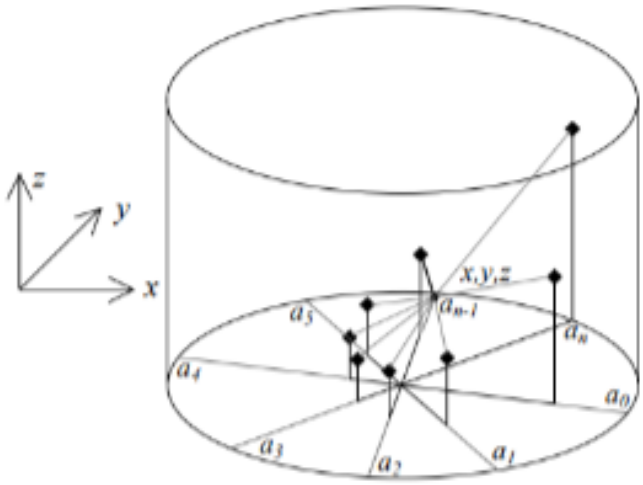
Υποθέτοντας ένα σύνολο δεδομένων που είναι αποθηκευμένο σε έναν πίνακα $D_{mxn}$, μια αναπαράσταση *Viz3D* επιτυγχάνεται απεικονίζοντας τις $n$ διαστάσεις των $m$ εγγραφών στις συντεταγμένες των τριών διαστάσεων $(x_i, y_i, z_i)$ σύμφωνα με τους τύπους:

$$x_i = x_c + \frac{1}{n}\sum_{j=0}^{n-1} \frac{D_{i,j} - min_j D_{i,j}}{max_j D_{i,j} - min_j D_{i,j}} cos(\frac{2\pi j}{n})$$

$$y_i = y_c + \frac{1}{n}\sum_{j=0}^{n-1} \frac{D_{i,j} - min_j D_{i,j}}{max_j D_{i,j} - min_j D_{i,j}} sin(\frac{2\pi j}{n})$$

$$z_i = z_c + \frac{1}{n}\sum_{j=0}^{n-1} \frac{D_{i,j} - min_j D_{i,j}}{max_j D_{i,j} - min_j D_{i,j}}$$

για $i = 0, ...., m-1$ και $j = 0, ...., n-1$. Στην παρακάτω εικόνα μπορούμε να δουμε τις προβολές $n$ διάστασης εγγραφών με χαρακτηριστικά $(\alpha_0, ...., \alpha_{n-1})$. Οι διάφορες διατάξεις των ακτινικών αξόνων παράγουν διαφορετικές οπτικοποιήσεις. Επομένως, η τοποθέτηση των αξόνων μπορεί να χρησιμοποιηθεί σε περιπτώσεις όπου έχουμε επικάλυψη σημείων.

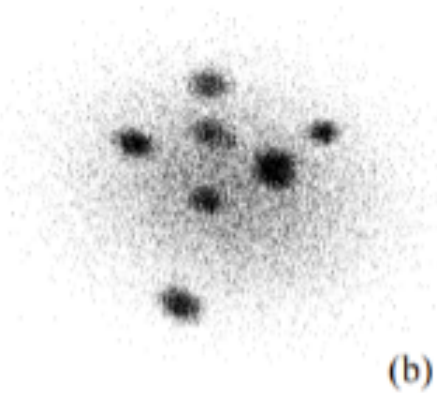



Σχήμα 25: *Viz3D* προβολή [46].

Αν και η *3D* προβολή που εκτελείται στην τεχνική *Viz3D* μπορεί να φιλοξενήσει περισσότερα γραφικά στοιχεία, ωστόσο η οπτική συσσώρευση εξακολουθεί να υπάρχει όταν έχουμε δεδομένα μεγάλης διάστασης. Αυτό το γεγονός απεικονίζεται στην παρακάτω εικόνα (Σχήμα 26), που δείχνει μια *Viz3D* απεικόνιση ενός τεχνητού συνόλου δεδομένων του *Sint3.data* (διαθέσιμο στην ιστοσελίδα $http://lib.stat.cmu.edu/datasets/$), το οποίο περιέχει 38850 εγγραφές δεδομένων 20 χαρακτηριστικών. Τα δεδομένα αποτελούνται από 8 ομάδες όπως περιγράφεται στην παρακάτω εικόνα (Σχήμα 26). Ένα σύνολο 14831 εγγραφών τα οποία ανήκουν σε ομάδες, οι υπόλοιπες 24019 εγγραφές είναι θόρυβος.

| Cluster# | Records |
|---|---|
| 1 | 2,272 |
| 2 | 1,521 |
| 3 | 2,272 |
| 4 | 1,518 |
| 5 | 2,271 |
| 6 | 1,352 |
| 7 | 2,272 |
| 8 | 1,353 |

(a)



(b)

Σχήμα 26: (*a*) Ο αριθμός των εγγραφών σε κάθε ομάδα του συνόλου, (*b*) Απεικόνιση με την τεχνική *Viz3D* [46].

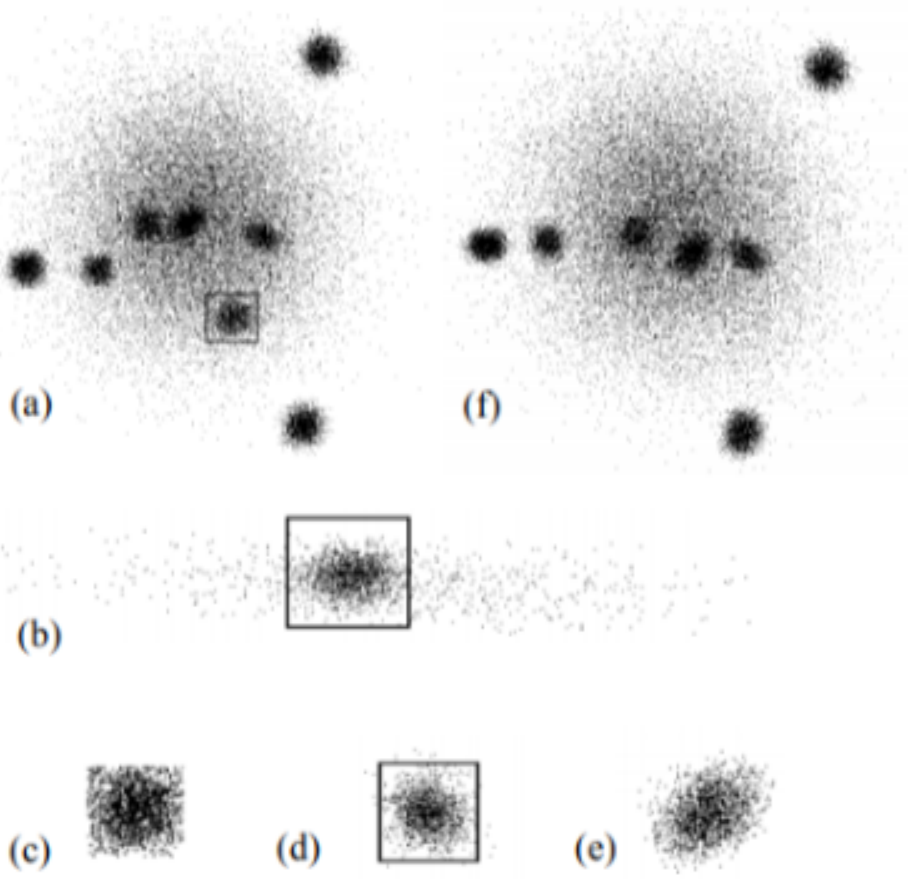
Σε αυτήν την περίπτωση, η αλληλεπίδραση με την αναπαράσταση στις τρεις διαστάσεις δίνει την δυνατότητα να εντοπιστούν οι 7 από τις 8 ομάδες. Για την παραγωγή πιο αποτελε-

σματικής απεικόνισης έχουν προταθεί δύο στρατηγικές.

Χρησιμοποιείται η πυκνότητα των δεδομένων στην οπτικοποίηση αντί να αντιμετωπίζονται ως μεμονωμένες εγγραφές. Η τεχνική αυτή παρέχει την βάση για μια προσέγγιση της οπτικής ομαδοποίησης με γνώμονα τον χρήστη, η οποία εφαρμόζεται σε δεδομένα μεγάλης διάστασης και είναι πολύ απλή, διαισθητική και αποτελεσματική. Στην πρώτη, ο χρήστης αλληλεπιδρά άμεσα με την *Viz3D* απεικόνιση οριοθετώντας οπτικά τις ομάδες των δεδομένων. Στην δεύτερη απεικόνιση ο χρήστης αλληλεπιδρά με μια εμπλουτισμένη πυκνότητα της απεικόνισης *Viz3D* για να αναγνωρίσει ομάδες σε δεδομένα που έχουν πολύ θόρυβο, τα οποία παράγουν πολύ ακατάστατες οπτικοποιήσεις. Οι προσεγγίσεις οπτικής ομαδοποίησης επιτρέπουν στους χρήστες να αλληλεπιδράσουν με τις οπτικές αναπαραστάσεις για να οδηγηθούν στην αναγνώριση ομάδων. Σε αυτήν την προσέγγιση ο χρήστης αλληλεπιδρά με την *Viz3D* απεικόνιση, χρησιμοποιώντας το ποντίκι του υπολογιστή ώστε να οριοθετήσει περιοχές οι οποίες πιθανά περιέχουν μια ομάδα. Μια διαδραστική διαδικασία συνδέεται με τρόπο ώστε, μόλις μια ομάδα γίνει οπτικά ορατή, οι εγγραφές που περιέχονται σε αυτήν ταυτοποιούνται κατάλληλα και ο χρήστης μπορεί να προχωρήσει στην αναγνώριση άλλης ομάδας. Η διαδικασία αυτή φαίνεται στην παρακάτω εικόνα (Σχήμα 27) και εφρμόζεται στα δεδομένα Sint3.

Η παρακάτω εικόνα (Σχήμα 27 (*a*)) δείχνει μια απεικόνιση στην οποία ο χρήστης έχει οριοθετήσει μια ορθογώνια περιοχή η οποία περιέχει μια προβαλλόμενη ομάδα. Στην επόμενη εικόνα δείχνονται μόνο τα στοιχεία που έχουν επιλεγεί. Επειδή η επιλογή των στοιχείων αυτών είναι στο δισδιάστατο επίπεδο, όσο ο χρήστης περιστρέφει την εικόνα παρατηρεί και διαφορετικές προβαλλόμενες όψεις, δείχνοντας ότι οι εγγραφές δεν ανήκουν ξεκάθαρα στην ομάδα, όπως στην παρακάτω εικόνα (Σχήμα 27 (*b*)). Ο χρήστης μπορεί να επιλέξει ξανά τις περιοχές σε αυτές τις νέες προβολές, βελτιώνοντας την επιλογή των εγγραφών στην ομάδα (Σχήμα 27 (*c*) και (*d*)), μέχρι να έχουμε μια ακριβή οριοθέτηση της ομάδας βάσει της αντίληψης του χρήστη. Αυτό φαίνεται στην παρακάτω εικόνα (Σχήμα 27 (*e*)), η οποία δείχνει τα αποτελέσματα της επιλογής της περιοχής στην εικόνα (Σχήμα 27 (*d*)). Αφού μια ομάδα έχει οριστεί πλήρως, οι εγγραφές της επισημαίνονται με ένα διακριτικό, π.χ. ομάδα 1. Η εικόνα (Σχήμα 27 (*f*)) δείχνει τα δεδομένα Sint3 χωρίς οι εγγραφές να έχουν ανατεθεί στην ομάδα 1. Η συσσώρευση μειώνεται βαθμιαία κατά την διάρκεια της διαδικασίας αυτής, απλοποιώντας το έργο της αναγνώρισης νέων ομάδων. Έτσι, η διαδικασία αυτή μπορεί να χρησιμοποιηθεί σε πολύ πυκνές και θορυβώδης απεικονίσεις. Ωστόσο, σε συνθήκες πολύ μεγάλης συσσώρευσης και με πολύ θορυβώδη δεδομένα, η οριοθέτηση μπορεί να γίνει δύσκολη και κουραστική. Εκθέτοντας πληροφορίες για την πυκνότητα των δεδομένων, παρά για τα ακατέργαστα δεδομένα, παρέχεται μια πιο κατάλληλη αναπαράσταση για την οπτικοποίηση

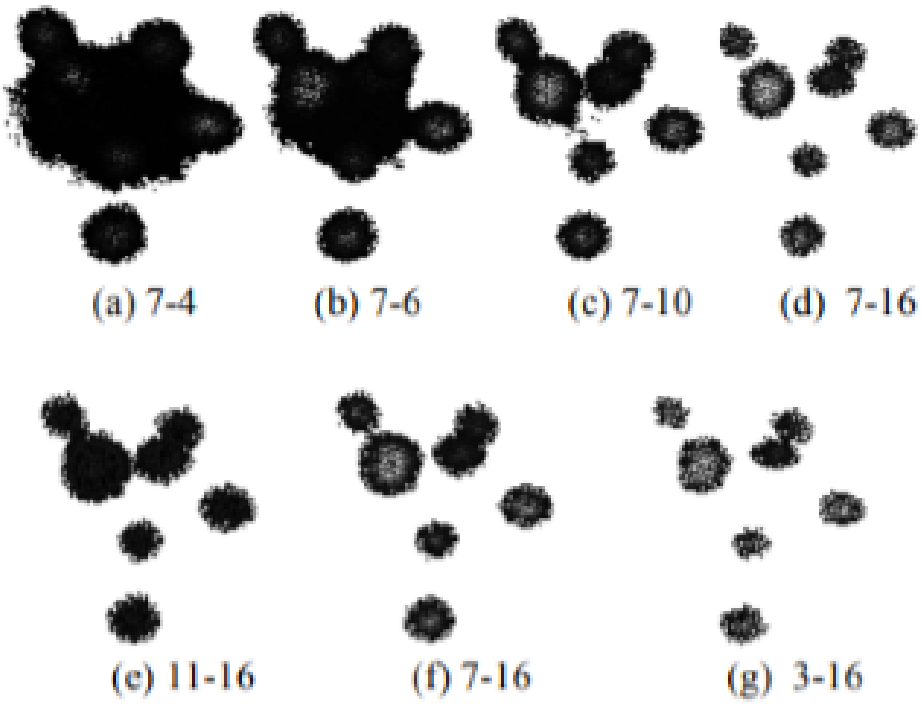
των ομάδων. Οι οπτικοποιήσεις της πυκνότητας των δεδομένων 'φιλτράρουν' τις πληροφορίες που παράγει η οπτικοποίηση, έτσι ώστε να δημιουργηθούν απεικονίσεις οι οποίες είναι πιο αποτελεσματικές για την ανακάλυψη μοτίβων.



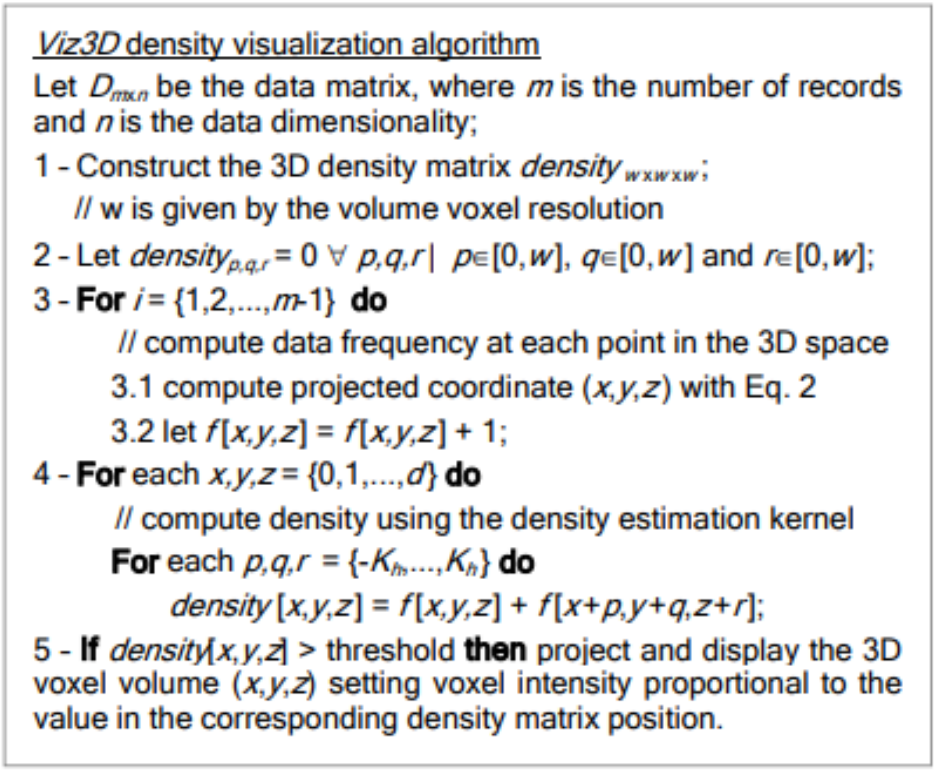

Σχήμα 27: Διαδραστική ομαδοποίηση με την *Viz3D* [56].

Ο υπολογισμός της πυκνότητας στον αρχικό χώρο των $n$ διαστάσεων είναι απαγορευτικός για μεγάλες τιμές του $n$. Υπολογίζουμε την πυκνότητα των δεδομένων στον τρισδιάστατο χώρο. Ο αλγόριθμος δείχνεται στο Σχήμα 33. Η *Viz3D* προβολή παράγει τις προβαλλόμενες συντεταγμένες $(x, y, z)$ των $n$ διάστατων εγγραφών. Η πυκνότητα των δεδομένων σε κάθε σημείο του τρισδιάστατου διακριτού χώρου υπολογίζεται και αποθηκεύεται στον πίνακα της πυκνότητας. Η προκύπτουσα απεικόνιση είναι αυτή που δημιουργήθηκε από τον πίνακα πυκνότητας. Κάθε στοιχείο του πίνακα πυκνότητας $(i, j, k)$ προβάλλεται με εικονοστοιχεία τα οποία ορίζουν ένα σημείο στον τρισδιάστατο χώρο με συντεταγμένες $(x, y, z)$, των οποίων η τιμή είναι ανάλογη προς την τιμή της πυκνότητας στην αντίστοιχη θέση του πίνακα. Υποθέτοντας μια μονόχρωμη απεικόνιση, η απεικόνιση των τρισδιάστατων σημείων των δεδομένων με μεγαλύτερη πυκνότητα θα προβάλλεται σε εικονοστοιχεία αποτυπομένα με υψηλότερες εντάσεις, επιτρέποντας έτσι μια άμεση οπτική αναγνώριση των περιοχών των

ομάδων των δεδομένων. Αντιστρέφοντας την απεικόνιση έτσι ώστε οι μικρότερες πυκνότητες να απεικονίζονται με εικονοστοιχεία υψηλότερης έντασης, επιτρέποντας την ταυτοποίηση των έκτροπων τιμών. Σε μια διερευνητική διαδικασία ένας χρήστης ελέγχει δύο παραμέτρους, το μήκος της διαμέτρου του πυρήνα ($K_w$) και το χαμηλότερο όριο της πυκνότητας του προβαλλόμενου τρισδιάστατου σημείου. Το μεταβαλλόμενο πλάτος του πυρήνα επιτρέπει τον έλεγχο του επιπέδου της διαδικασίας της ομαδοποίησης, ένα ελάχιστο πλάτος αντιστοιχεί στην εξέταση κάθε τρισδιάστατου σημείου ως ομάδα. Η παρακάτω εικόνα (Σχήμα 28) δείχνει διάφορες πυκνοτικές οπτικοποιήσεις των δεδομένων Sint3. Οι δύο αριθμοί που εμφανίζονται κάτω από κάθε εικόνα δείχνουν το πλάτος του πυρήνα (αριστερά) και το κατώφλι της πυκνότητας (δεξιά).




Σχήμα 28: Πυκνοτικές οπτικοποιήσεις των δεδομένωνSint3, αυξάνοντας τα κατώφλια [56].

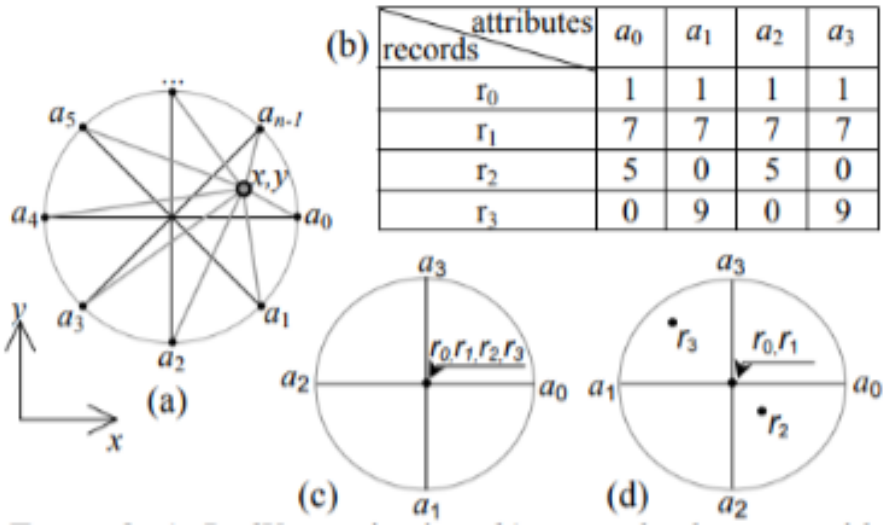



Σχήμα 29: Αλγόριθμος Viz3D [56].

Η σειρά των βημάτων από το $(a)$ στο $(d)$ στην παραπάνω εικόνα (Σχήμα 28) δείχνει την επίδραση της αύξησης του κατωφλιού της πυκνότητας: μια αρχική τιμή αυξάνεται σταδιακά, ώστε σε κάθε βήμα να εμφανίζονται οι περιοχές στις οποίες αυξάνεται η πυκνότητα των δεδομένων, το οποίο έχει ως αποτέλεσμα να εμφανίζονται οι ομάδες των δεδομένων. Η σειρά των βημάτων από το $(e)$ στο $(g)$ (Σχήμα 28) δείχνει την επίδραση της μείωσης του πλάτους του πυρήνα, $K_d$.

### 4.2.11 **RadViz**

Η τεχνική *RadViz* απεικονίζει τα δεδομένα μεγάλης διάστασης στις δύο διαστάσεις, τοποθετώντας κάθε στοιχείο των δεδομένων σε έναν δισδιάστατο κύκλο (Σχήμα 30) [56]. Σε αυτήν την απεικόνιση, οι $n$ άξονες ξεκινούν από το κέντρο του κύκλου και τερματίζονται στην περίμετρό του, όπως φαίνεται στην εικόνα (Σχήμα 30).




Σχήμα 30: Στην $(a)$ βλέπουμε την τεχνική των προβολλών *RadViz*, στην $(b)$ βλέπουμε ένα σύνολο δεδομένων με εγγραφές τεσσάρων χαρακτηριστικών, στην $(c)$ βλέπουμε την απεικόνιση αυτών το εγγραφών με την *RadViz* και στην $(d)$ βλέπουμε μια απεικόνιση των ίδιων δεδομένων με αλλαγή στην θέση των αξόνων, που αντιμετωπίζει την συσσώρευση [46].

Κάθε άξονας αντιπροσωπεύει ένα χαρακτηριστικό των δεδομένων. Η θέση των δεικτών ορίζεται σταθμίζοντας τις τιμές των χαρακτηριστικών των δεδομένων. Η προκύπτουσα προβολή αποτελεί ένα μη γραμμικό μετασχηματισμό που διατηρεί μερικές συμμετρίες των δεδομένων. Σε αυτήν την τεχνική οι εγγραφές πιέζονται προς τον άξονα απεικόνισης του χαρακτηριστικού με την μεγαλύτερη τιμή. Οι εγγραφές που έχουν ίσες τιμές προβάλλονται στον κέντρο του κύκλου. Οι εγγραφές οι οποίες έχουν ίσες τιμές προβάλλονται στο κέντρο του κύκλου. Σημαντικό πλεονέκτημα είναι η χαμηλή πολυπλοκότητα και ότι οι όμοιες παρατηρήσεις στον χώρο διάστασης $D$ προβάλλονται κοντά στον χώρο των δύο διαστάσεων, ευνοώντας την αναγνώριση των ομάδων.

Μια επέκταση της τεχνικής *RadViz* για τον τρισδιάστατο χώρο είναι η *RadViz3D*. Μια πιθανή προσέγγιση που μπορεί να αποτρέψει να αποτρέψει την συσσώρευση των χαρακτηριστικών που σχετίζονται υψηλά στο κέντρο του κύκλου είναι να τα τοποθετήσουμε σε αντίθετες κατευθύνσεις στο ακτινικό σύστημα. Ακολουθούμε την πρόταση των Ankerst et al. παράγοντας μια σειρά των αξόνων βάση των συσχετισμών των χαρακτηριστικών τους. Με μια διάταξη αξόνων η οποία κρατάει κοντά τις υψηλές συσχετίσεις χαρακτηριστικών μπορεί να επιτευχθεί η εξαγωγή πληροφοριών σχετικά με την ομοιότητα των χαρακτηριστικών. Η προσέγγισή μας συνιστά την εξαγωγή ενός πίνακα ομοιοτήτων $S_{nxn}$ από τον πίνακα $D_{mxn}$.

Ο πίνακας $S$ περιέχει μέτρα ομοιότητας μεταξύ όλων των ζευγών δεδομένων, όπου το $s_{ij}$ δίνει ένα μέτρο ομοιότητας ανάμεσα στα χαρακτηριστικά $i$ και $j$ που υπολογίζονται από τον τύπο

$$s_{i,j} = 1 - \frac{1}{m}\sum_{k=1}^{m} \mid \frac{D_{k,i} - Min(D_i}{Max(D_i) - Min(D_i)} - \frac{D_{k,j} - Min(D_j)}{Max(D_j) - Min(D_j)} \mid$$

όπου το $Min(D_i)$ και $Max(D_i)$ είναι η μικρότερη και η μεγαλύτερη τιμή της στήλης $i$. Μεγάλες τιμές του $s_{ij}$ υποδεικνύουν μεγάλη ομοιότητα των χαρακτηριστικών $i$ και $j$. Αυτό το μέτρο εξαρτάται από την κλίμακα των δεδομένων, για αυτόν τον λόγο χρειάζεται πρώτα να γίνει κανονικοποίηση των δεδομένων. Αξίζει εδώ να σημειώσουμε ότι η ισότητα (3) δίνει ένα πολύ απλοϊκό μέτρο ομοιότητας. Μπορούν να χρησιμοποιηθούν και άλλα μέτρα ομοιότητας. Βασιζόμενοι στην εξίσωση (3) οι άξονες είναι διατεταγμένοι με βάση την ομοιότητα των χαρακτηριστικών τους, έτσι ώστε τα χαρακτηριστικά με υψηλή συσχέτιση να τοποθετούνται (άξονες) το ένα δίπλα στο άλλο. Ένας χρήστης μπορεί να αφαιρέσει διαδραστικά ένα από τους δύο άξονες οι οποίοι έχουν μεγάλη συσχέτιση. Παρέχεται μια αυτοματοποιημένη διαδικασία για την κατάργηση των εξαιρετικά συσχετισμένων χαρακτηριστικών, όπου οι δυνητικοί υποψήφιοι για εξάλειψή τους καθορίζονται με βάση την υψηλότερη τιμή του $s_{ij}$. Η διαδικασία αυτή μπορεί να επαναληφθεί μέχρι ότου να επιτευχθεί ο απαιτούμενος αριθμός χαρακτηριστικών. Αυτή την διαδικασία μπορούμε να την δούμε και ως μια διαδικασία προς τα πίσω επιλογής χαρακτηριστικών, η οποία ξεκινάει με ένα σύνολο χαρακτηριστικών και βαθμιαία εξαλείφει αυτά τα οποία είναι υψηλά συσχετισμένα.

#### 4.2.12 Απεικονίσεις Βασισμένες σε Άξονες με Ακτινικές Διαστάσεις

Αυτές οι τεχνικές χρησιμοποιούνται για την οπτικοποίηση δεδομένων μεγάλης διάστασης. Θα εισάγουμε παρακάτω δύο νέες ακτινικές τεχνικές, την **TimeWheel** και την **MultiComb**, ως πολλά υποσχόμενες τεχνικές για την ανάλυση δεδομένων μεγάλης διάστασης [11]. Αυτές είναι μέρος μίας διαδραστικής διαδικασίας η οποία λέγεται **VisAxes**, η οποία απεικονίζει τέτοια σύνολα δεδομένων με διαφερετική ακτινική διάταξη στην απεικόνιση και παρέχει υποστήριξη για μια ποικιλία λειτουργιών πλοήγησης.

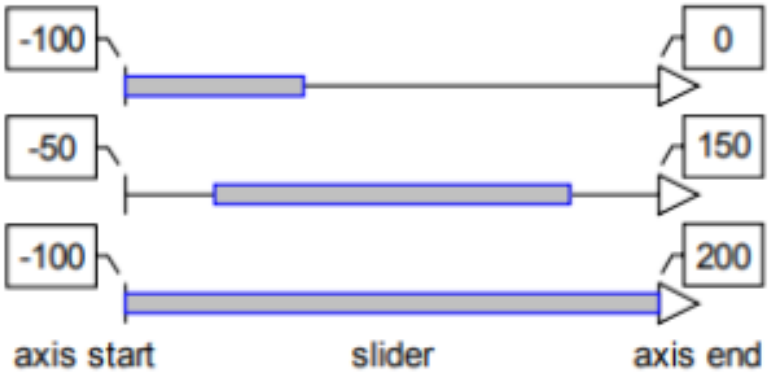
Όταν οριμένοι από τους άξονες αντιπροσωπεύουν ιεραρχικά οργανωμένα δεδομένα, είναι φυσικό να παρέχουν μηχανισμούς ώστε να μας κάνουν να χρησιμοποιήσουμε την ιεραρχία στην διαδικασία της πλοήγησης. Αυτό που αξίζει να σημειωθεί είναι ένα σημαντικό χαρακτηριστικό των πιο πρόσφατων μεθόδων που στηρίζονται στους άξονες, δεδομένου ότι στην

αρχική τους μορφή επικεντρώνονται μόνο σε ένα επίπεδο λεπτομερώς. Αυτή η δυνατότητα γίνεται πολύ σημαντική όταν αντιμετωπίζουμε δεδομένα που σχετίζονται με τον χρόνο. Δεδομένου ότι σε πολλά πραγματικά σύνολα οι άξονες δεν είναι ίσοι, δηλαδή γίνεται μια διάκριση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Αυτή η διάκριση πρέπει να εκφραστεί σε μια απεικόνιση που βασίζεται στους άξονες. Για παράδειγμα, στην περίπτωση που τα δεδομένα μας βασίζονται στον χρόνο, ο χρόνος παίζει αυτόν τον ρόλο. Ωστόσο, σε ορισμένες περιπτωσεις αυτός ο ειδικός άξονας δεν παίζει ρόλο. Ένα επιθυμητό χαρακτηριστικό των οπτικών αναπαραστάσεων που βασίζονται σε άξονες είναι η ανίχνευση συσχετίσεων μεταξύ γειτονικών αξόνων. Ωστόσο, είναι δύσκολο να ανιχνευθούν συσχετισμοί μεταξύ των αξόνων μακριά από την ενσωμάτωση. Η εξερεύνηση δεδομένων μεγάλης διάστασης που στηρίζεται σε τεχνικές αξόνων απαιτεί μια ολοκληρωμένη λύση. Επιπλέον, οι ακτινικές ρυθμίσεις μπορούν να χρησιμοποιηθούν σε μια ποικιλία περιπτώσεων όπως ο συνδιασμός οπτικοποιήσεων που στηρίζονται σε άξονες για να απεικονιστεί ένα σύνολο δεδομένων.

Έχουν προταθεί πολλές μέθοδοι βασισμένοι στους άξονες για την οπτικοποίηση δεδομένων μεγάλης διάστασης. Στόχος μας εδώ είναι η ανάπτυξη του ευέλικτου πλαισίου VisAxes για την υποστήριξη της δημιουργικότητας και την αξιολόγηση των αξόνων κάτω από το πρίσμα των ακτινικών ρυθμίσεων. Στις απεικονίσεις που βασίζονται σε άξονες, κάθε άξονας συνδέεται με μια δεδομένη μεταβλητή. Ο σχεδιασμός και η κλίμακα ενός άξονα εξαρτάται έντονα από τον τύπο των δεδομένων. Αν μια μεταβλητή έχει ένα μεγάλο εύρος τιμών κάποιες είναι απαραίτητο να ομαδοποιηθούν. Αν οι τιμές των μεταβλητών είναι ιεραρχικά δομημένες είναι φυσικό να γίνει κάποια χρήση της ιεραρχίας κατά την περιήγηση των δεδομένων. Ένα άλλο επιθυμητό χαρακτηριστικό είναι να έχουν την δυνατότητα οι χρήστες να απεικονίσουν ένα ορισμένο σύνολο τιμών. Για να ανταποκριθεί στις απαιτήσεις αυτές και να παρέχει ένα ευέλικτο και υψηλό βαθμό αλληλεπίδρασης που σχετίζονται με δεδομένα τα οποία έχουν μεγάλο αριθμό παρατηρήσεων, σχεδιάζουμε τρεις τύπους διαδραστικών αξόνων, οι οποίοι είναι οι άξονες κύλισης, οι ιεραρχικοί άξονες και οι άξονες εστίασης.

Η κύρια χρήση του άξονα κύλισης είναι για μεταβλητές οι οποίες έχουν μεγάλο αριθμό τιμών. Σε αυτήν την περίπτωση είναι πιο αποτελεσματικό να εμφανίζονται μόνο οι τιμές σε ένα τμήμα ενδιαφέροντος. Έτσι, η ιδέα είναι να σχεδιάσουμε έναν ολισθητή σε έναν άξονα (Σχήμα 31). Ο ολισθητής μπορεί να κινηθεί διαδραστικά στον άξονα. Με αυτόν τον τρόπο, ένας χρήστης μπορεί να επιλέξει το τμήμα ενδιαφέροντος εντός της περιοχής της μεταβλητής, η οποία στην συνέχεια αντιστοιχίζεται στον άξονα της μεταβλητής. Ο ολισθητήρας μπορεί να στενεύσει ή να διευρυνθεί διαδραστικά στο πλάτος της περιοχής ενδιαφέροντος. Επειδή, το εύρος των τιμών είναι μικρό για έναν στενό ολισθητήρα και

μεγάλο για έναν μακρύ ολισθητήρα, ένας χρήστης μπορεί να χρησιμοποιήσει ένα κυλιόμενο ολισθητήρα για να μεγενθύνει τις μεταβλητές ή να έχει μια συνολική εικόνα της περιοχής τιμών της μεταβλητής.



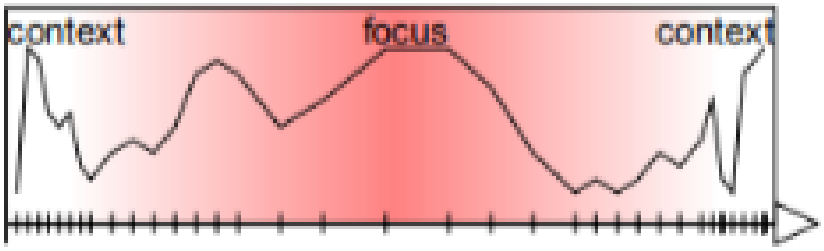

Σχήμα 31: Διάφοροι άξονες κύλισης για μια μεταβλητή με μικρότερη τιμή το -100 και μεγαλύτερη την 200. Το πλάτος και η θέση καθορίζει την κλιμάκωση των αξόνων που ανταποκρίνεται στις απεικονισθέντες τιμές [67].

Ο δεύτερος τύπος αξόνων είναι οι ιεραρχικοί άξονες (Σχήμα 32). Χρησιμοποιούνται στην περίπτωση που έχουμε δεδομένα ιεραρχικής δομής. Εάν ένας χρήστης επιλέξει ένα τμήμα, αυτό διαχωρίζεται σύμφωνα με τον αριθμό των κόμβων. Μια άλλη διαδραστική επιλογή μπορεί να χρησιμοποιηθεί είτε για να ανοίξει περισσότερα τμήματα, είτε να ανατεθεί σε ένα μόνο τμήμα. Δεδομένου ότι η απόφαση για το άνοιγμα περισσότερων τμημάτων σημαίνει ότι ο χρήστης αναζητά πιο λεπτομερή πληροφορία σχετικά με το τμήμα. Επιπλέον, η κινούμενη εικόνα χρησιμοποιείται για την απεικόνιση του ανοίγματος ή του κλεισίματος του κόμβου προκειμένου να αποφευχθεί η οπτική ασυνέχεια.


Σχήμα 32: Ένας ιεραρχικός άξονας του χρόνου μετά από κάποια βήματα αλληλεπίδρασης. Μπλε, πράσινα και κόκκινα πλαίσια υποδεικνύουν τα οπτικοποιημένα τμήματα [67].

Ο τρίτος τύπος αξόνων είναι ο άξονας εστίασης. Είναι χρήσιμος όταν απαιτείται απεικόνι-

ση ολόκληρης της περιοχής της μεταβλητής. Ωστόσο, ο άξονας κύλισης είναι κλιμακωτός ομοιόμαρφα, ενώ η κλιμάκωση εντός αυτού του άξονα είναι μη ομοιόμορφη (Σχήμα 33). Στην περίπτωση αυτή χρησιμοποιούμε μια από τις γνωστές λειτουργίες του μετασχηματισμού μεγένθυνσης (Σχήμα 33) στην διαδικασία της απεικόνισης. Με αυτό τον τρόπο παρέχεται μια πιο λεπτομερής προβολή (εστίαση) χωρίς να χάσουμε την συνολική προβολή που παρέχεται ως πλαίσιο. Ένας χρήστης μπορεί να αλλάξει διαδραστικά την εστίαση εντός των ορίων των τιμών μίας μεταβλητής.

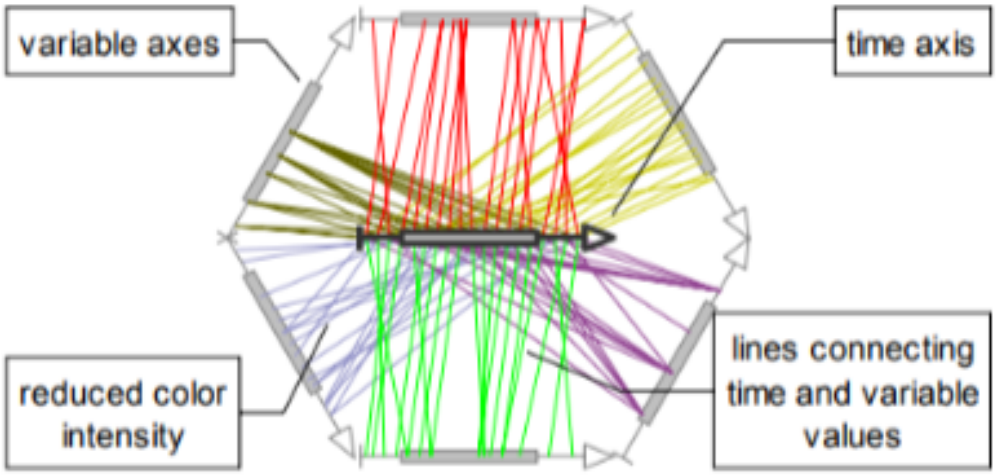



Σχήμα 33: Ένα μη ομοιόμορφος κλιμακωτός άξονας που συνδιάζεται με ένα διάγραμμα μίας μόνο μεταβλητής [67].

Οι προτεινόμενοι άξονες σχεδιασμού μπορούν να χρησιμοποιηθούν για διαφορετικούς τύπους δεδομένων και διαφορετικά δομημένες μεταβλητές. Κατά την απεικόνιση διακριτά και συνεχή πεδία και των τριών τύπων αξόνων εφαρμόζονται.

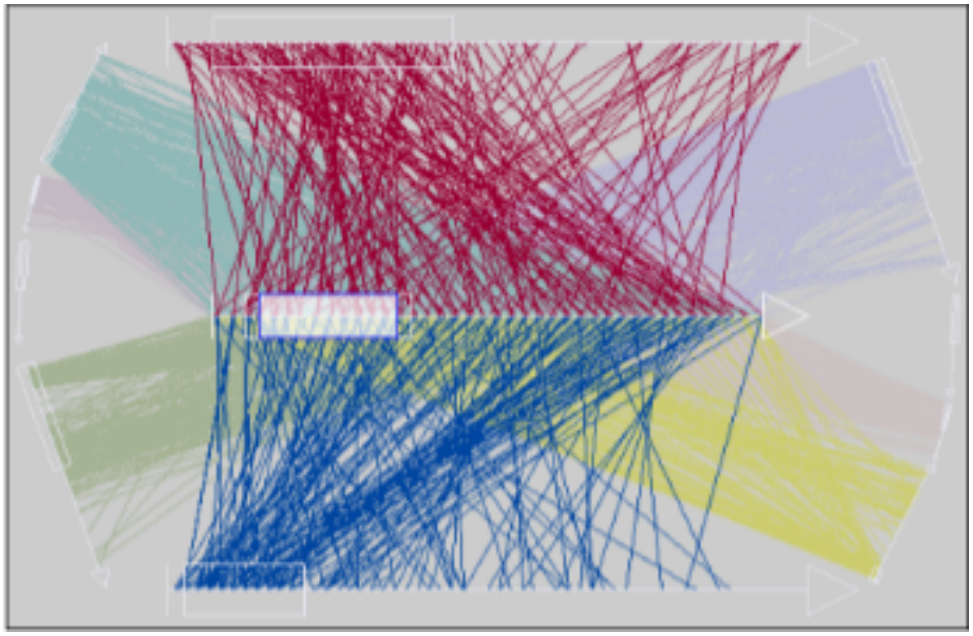
Η διευθέτηση των αξόνων είναι μια μη τετριμένη εργασία. Ως εκ τούτου, μερικά από τα χαρακτηριστικά ενός συνόλου δεδομένων πρέπει να εξεταστούν βάση αυτού του πλαισίου. Ειδικότερα, πρέπει να γίνει διάκριση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Παρακάτω, παρουσιάζονται δύο ακτινικές διατάξεις. Κάθε άξονας συνδέεται με ένα συγκεκριμένο χρώμα.

Η πρώτη τεχνική η TimeWheel είναι να παρουσιάσουμε τον άξονα αναφοράς (εδώ χρόνος) στο κέντρο της οθόνης και να ρυθμιστούν κυκλικά οι εξαρτημένοι άξονες γύρω από αυτόν (Σχήμα 34). Για κάθε χρονική τιμή ένα έγχρωμο ευθύγραμμο τμήμα ζωγραφίζεται για κάθε μεταβλητή στην απεικόνιση.



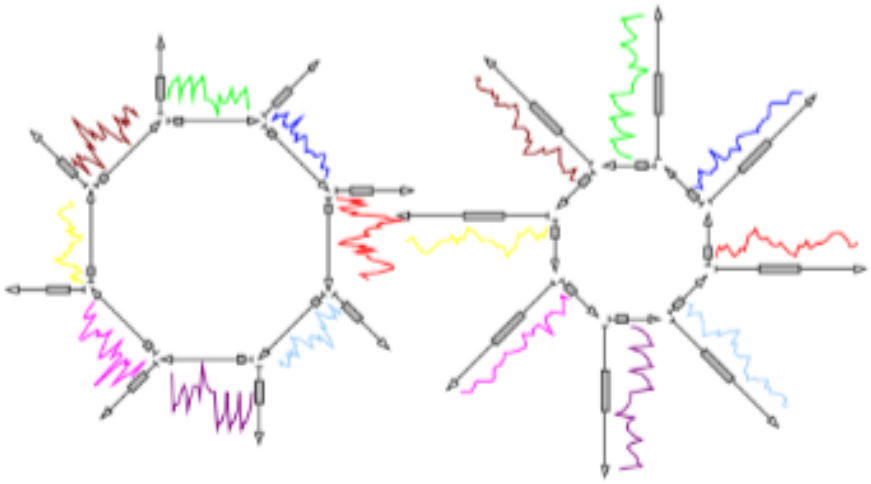

Σχήμα 34: Ένας TimeWheell. Έξι άξονες μεταβλητών τοποθετούνται κυκλικά γύρω από έναν άξονα του χρόνου [67].

Οι σχέσεις μεταξύ των χρονικών τιμών και των τιμών μίας μεταβλητής μπορούν να διερευνηθούν με τον πιο αποτελεσματικό τρόπο, εάν ο άξονας των μεταβλητών είναι τοποθετημένος παράλληλα προς τον άξονα του χρόνου. Η διαδραστική περιστροφή της τεχνικής του TimeWheel παρέχεται έτσι ώστε ένας χρήστης να μπορεί να μετακινήσει τους άξονες ενδιαφέροντος σε τέτοια θέση. Όταν ένας άξονας είναι κάθετος στον άξονα του χρόνου είναι δύσκολη η οπτική του ανάλυση. Για να μετριάσουμε αυτήν την δυσκολία χρησιμοποιούμε την χρωματική εξασθένηση για την απόκρυψη γραμμών μεταξύ αυτών των αξόνων (Σχήμα 35). Επομένως, το χρώμα υπολογίζεται σε σχέση με την γωνία που σχηματίζεται από τους άξονες που συνδέει η γραμμή του χρόνου. Επιπλέον, εμείς ρυθμίζουμε τα μήκη των αξόνων, οι οποίοι αντιπροσωπεύουν τις μεταβλητές, σύμφωνα με τις γωνίες τους με τον άξονα του χρόνου. Με αυτόν τον τρόπο διαθέτουμε περισσότερο χώρο για τους άξονες που μας ενδιαφέρουν (οι άξονες είναι παράλληλοι στους χρονικούς άξονες), αφού οι άλλοι άξονες αντιπροσωπεύονται σε μικρότερο βαθμό λεπτομέρειας μειώνοντας τα μήκη τους. Μπορεί να χρησιμοποιηθεί η χρήση διαφορετικών μηκών αξόνων. Χάρη τόσο στο ξεθώριασμα όσο και στη ρύθμιση του μήκους αποφεύγουμε υπερπλήρωση της οθόνης και μείωση της συσσώρευσης. Οι χρήστες οι οποίοι είναι οικείοι με τις παράλληλες συντεταγμένες θα δουν τον TimeWheel ως μια ενδιαφέρουσα εναλλακτική λύση για δεδομένα μεγάλης διάστασης που σχετίζονται με τον χρόνο.



Σχήμα 35: Ένας TimeWheel. Το μήκος των κυκλικών αξόνων και το ξεθώριασμα του χρώματος υπολογίζεται σύμφωνα με την γωνία που σχηματίστηκε από τον κάθε άξονα με τον κεντρικό άξονα ως άξονα αναφοράς [67].

Μια άλλη τεχνική αξόνων είναι η MultiComb. Η βασική ιδέα είναι να κάνει χρήση των διαγραμμάτων για την οπτικοποίηση των δεδομένων μεγάλης διάστασης. Ως εκ τούτου, διαφορετικού τύπου διαγράμματα, διατάσσονται κυκλικά στην οθόνη (Σχήμα 36). Υπάρχουν δύο πιθανοί τρόποι για την τοποθέτηση των διαγραμμάτων. Στην μια περίπτωση, τα διαγράμματα διατάσσονται κυκλικά στην οθόνη (στην αριστερή μεριά στο Σχήμα 36). Στην δεύτερη περίπτωση, τα διαγράμματα διατάσσονται από το κέντρο και προς τα έξω της απεικόνισης (στην δεξιά πλευρά στο Σχήμα 36). Για να αποφευχθούν αλληλεπικαλυπτόμενα διαγράμματα, οι άξονες δεν ξεκινούν από το κέντρο της απεικόνισης. Έτσι, η κεντρική περιοχή μπορεί να χρησιμοποιηθεί για να αντιπροσωπεύσει πρόσθετες πληροφορίες προκειμένου να βελτιωθεί η τεχνική του MultiComb. Η MultiComb μπορεί να περιστρέφεται διαδραστικά, όπως η TimeWheel.



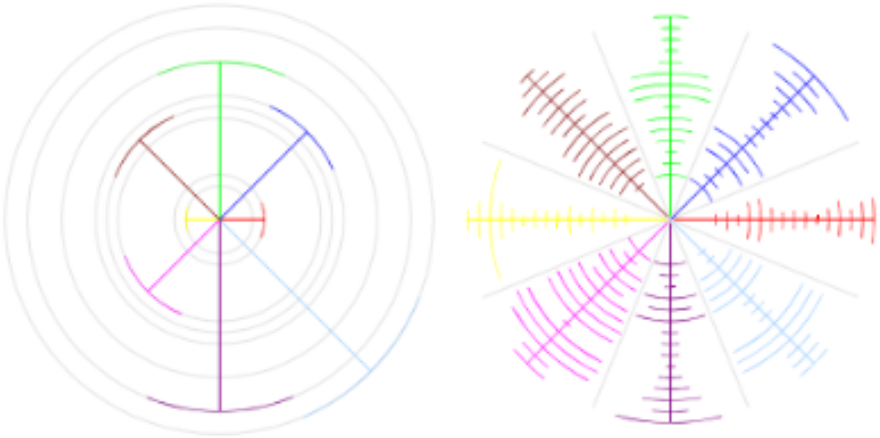
Σχήμα 36: Η τεχνική **MultiComb**. Αριστερά: Οι άξονες των εξαρτημένων αξόνων εκτείνονται εκτός του κέντρου. Δεξιά: Ο άξονας αναφοράς εκτείνεται προς τα έξω [67].

Για να βοηθηθεί η εξερεύνηση και η ανάλυση των δεδομένων που εξαρτώνται από τον χρόνο, θα προσθέσουμε στατιστικά στοιχεία στους δείκτες. Θα δούμε πως χρησιμοποιείται ένας άξονας ολίσθησης για την αναπαράσταση του χρόνο που τρέχει στο παρόν, μπορούν να ερμηνευτούν ως παρόντες και οι τιμές πριν ή μετά από αυτό το διάστημα ως παρελθόν και μέλλον αντίστοιχα. Σύμφωνα με αυτό, χρησιμοποιούμε στους άξονες δείκτες διαφορετικού σχήματος (βλέπε 37) (π.χ. ένας κύκλος για το παρόν, και ένα δεξί/αριστερό βέλος για το παρελθόν και το μέλλον). Κάθε δείκτης συνδέεται με ένα κατάλληλο στατιστικό (π.χ., ελάχιστη, μέγιστη ή μέση τιμή), το οποίο καθορίζει την πραγματική θέση του δείκτη σε έναν άξονα.

Σχήμα 37: Ένα παράδειγμα στατιστικών δεικτών σε έναν κυλινδρικό άξονα [67].

Όπως αναφέρθηκε παραπάνω για την **MultiComb**, η κεντρική περιοχή της απεικόνισης μπορεί να χρησιμοποιηθεί για την παρουσία πρόσθετων πληροφοριών. Ως εκ τούτου, χρησιμοποιούμε ένα εικονοστοιχείο στην μορφή ακίδας για να επιτρέψει την εύκολη σύγκριση των τιμών και μια συγκεντρωτική προβολή των παρελθοντικών τιμών. Ένας χρήστης τώρα επιλέγει μια τιμή από τον άξονα αναφοράς. Το μήκος κάθε ακίδας τότε υπολογίζεται σύμφωνα με την τιμή της εξαρτώμενης μεταβλητής που συνδέεται με την ακίδα στο επιλεγμένο σημείο

αναφοράς. Για να γίνει ευκολότερη η σύγκριση των εικονοστοιχείων ακίδων σχεδιάζεται ένα έγχρωμο κυκλικό τόξο με ακτίνα ανάλογη προς το μήκος της ακίδας.

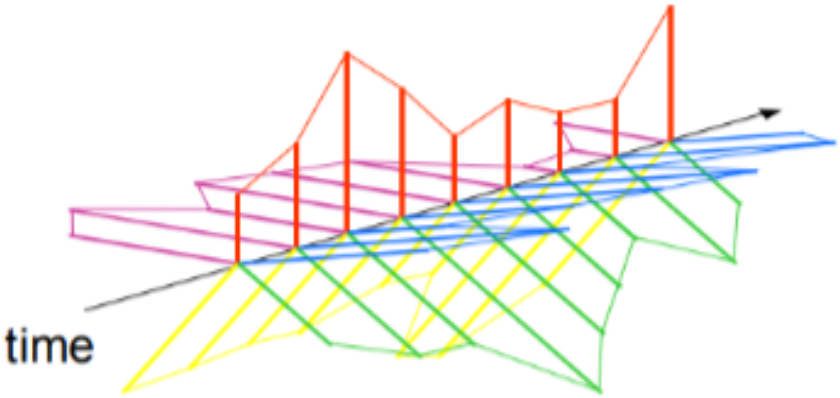


Σχήμα 38: Ένα εικονοστοιχείο ακίδα (αριστερά) και μια συνολική εικόνα τον παρελθόντων τιμών (δεξιά) [67].

Η προηγούμενη απεικόνιση μπορεί να συνδιαστεί με την τεχνική του άξονα κύλισης. Για να δημιουργήσουμε μια συνολική προβολή για κάθε άξονα μεταβλητής που χρησιμοποιείται για την οπτικοποιήση ενός επιπρόσθετου άξονα που δημιουργείται από το κέντρο της απεικόνισης από τους συνδεόμενους άξονες (Σχήμα 38 δεξιά). Σε κάθε επιπρόσθετο άξονα ένας μικρός αριθμός κυκλικών τόξων μπορεί να χρησιμοποιηθεί για να αντιπροσωπευθούν οι κεντρικές τιμές. Επομένως, για κάθε κυκλικό τόξο συγκεντρώνεται ένας συγκεκριμένος αριθμός τιμών των δεδομένων και οι τιμές απεικονίζονται στην γωνία του τόξου. Παρέχοντας μια συγκεντρωτική προβολή, κατά την περιήγηση ο χρήστης μπορεί να πάρει μια ιδέα των τιμών των μεταβλητών για τα προηγούμενα διαδραστικά βήματα. Λαμβάνοντας υπόψιν τους δείκτες, το εικονογράφημα ακίδα και την συγκεντρωτική προβολή υποστηρίζεται ένας χρήστης κατά την διαδικασία της εξερεύνησης. Γνωρίζοντας το απεικονιζόμενο διάστημα, το επιλεγμένο σημείο των δεδομένων και τα σχετικά στατιστικά ένας χρήστης μπορεί να πάρει μια ιδέα των συσχετίσεων ανάμεσα στις παρόντες, τις παρελθόντες και τις μελλοντικές τιμές των δεδομένων.

Μια ακόμα επέκταση αυτών των τεχνικών είναι να τις τοποθετήσουμε στον τρισδιάστατο χώρο (Σχήμα 39). Αυτό ανοίγει πολλές προοπτικές για το σχεδιασμό νέων διατάξεων. Αυτή η τεχνική ονομάζεται τρισδιάστατη **MultiComb**. Δημιουργείται τοποθετώντας διάφορα εικονοστοιχεία σε σχήμα αστεριού για κάθε εγγραφή των δεδομένων που συνδέονται με το χρόνο. Αυτή η τεχνική μπορεί να χρησιμοποιηθεί σε όλες τις προηγούμενες απεικονίσεις και

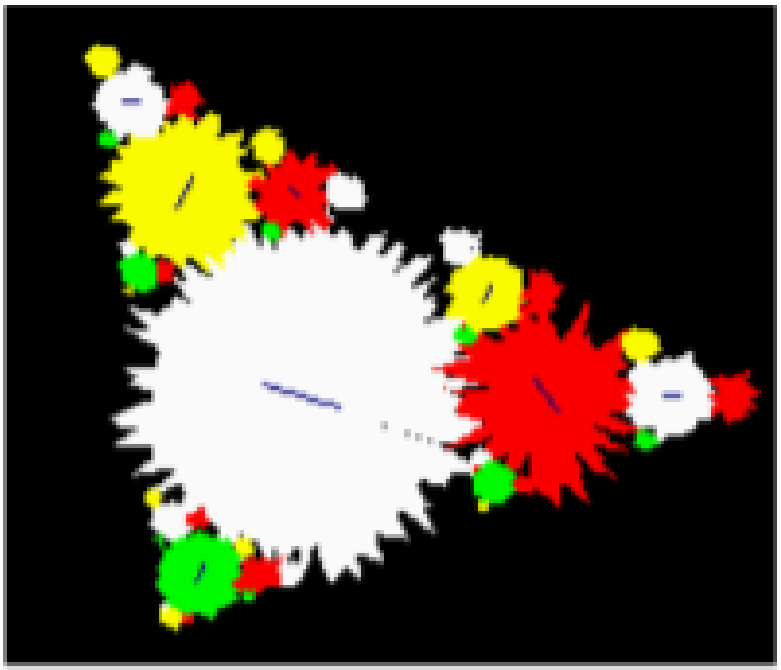
υπόκειται στις ίδιες διαδικασίες που ισχύουν για άλλα μοτίβα που βασίζονται σε άξονες.




Σχήμα 39: Ένα σχέδιο της 3-D MultiComb [67].

### 4.2.13 Φυσαλίδες

Οι Φυσαλίδες (**Foam Bubbles**) είναι μια τεχνική για οπτικοποίηση δεδομένων μεγάλης διάστασης, η οποία παρουσιάζει τις συσχετίσεις μεταξύ των διαστάσεων αναδρομικά (Σχήμα 40). Διαλέγουμε αρχικά μια διάσταση η οποία αναπαρίσταται ως μια έγχρωμη φυσαλίδα. Οι άλλες διαστάσεις αναπαρίστανται ως μικρότερες φυσαλίδες πηγαίνοντας δεξιόστροφα γύρω από την κύρια φυσαλίδα. Το μέγεθος κάθε φυσαλίδας αντιπροσωπεύει την συσχέτιση της πρώτης διάστασης με την κάθε διάσταση. Αυτή η διαδικασία επαναλαμβάνεται με μικρότερες φυσαλίδες γύρω από κάθε μία από τις άλλες φυσαλίδες παρουσιάζοντας τις ζευγαρωτές τους συσχετίσεις. Οι διαστάσεις που είναι υψηλά συσχετισμένες θα έχουν μεγαλύτερες φυσαλίδες. Άλλα στατιστικά μέτρα όπως η διακύμανση, η κύρτωση και η λοξότητα μπορούν να απεικονισθούν. Ένα παράδειγμα μπορούμε να δούμε στην παρακάτω εικόνα.

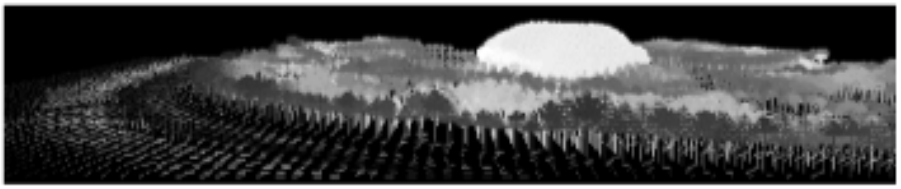


Σχήμα 40: Η τεχνική των φυσαλίδων για την απεικόνιση των δεδομένων **Iris** [65].

### 4.2.14 Εικονική Πραγματικότητα - Σχηματισμοί Εδάφους

Ένας αριθμός από τεχνικές εικονικής πραγματικότητας έχει εφαρμοστεί για να απεικονισθούν δεδομένα μεγάλης διάστασης και συνήθως ονομάζονται τεχνικές **Benediktine**. Αυτές οι απεικονίσεις έχουν το ίδιο είδος προβλημάτων με αυτές που απεικονίζουν δεδομένα μεγάλης διάστασης στις δύο διαστάσεις [62]. Ο **Benediktine** αναφέρθηκε στα προβλήματα απεικόνισης

με τους όρους των εξωτερικών διαστάσεων (τρείς χωρικές συντεταγμένες την φορά) και των εσωτερικών διαστάσεων όπως το σχήμα και το χρώμα. Το πρόβλημα της απεικόνισης είναι μεγαλύτερο για τις τρεις διαστάσεις, όπου δεν υπάρχουν προφανείς εξωτερικές διαστάσεις των δεδομένων. Η διάδραση είναι απαραίτητη σε κάθε απεικόνιση με την τεχνική της εικονικής πραγματικότητας και αυτό απαιτεί γνώση από την πλευρά των χρηστών. Η τεχνική **SGI** του **Mineset** που χρησιμοποιεί τριών διαστάσεων διαγράμματα διασποράς κινουμένων εικονοστοιχείων είναι ένα από τα καλύτερα παραδείγματα και μας επιτρέπει να πάρουμε πληροφορίες για τα δεδομένα. Η τεχνική **Natural Paradigm** είναI μια άλλη τεχνική εικονικής πραγματικότητας, η οποία προσπαθεί να εκμεταλλευτεί το ανθρώπινο οπτικό σύστημα για να απεικονίσει τα δεδομένα με φυσικές σκηνές. Αυτές οι φυσικές σκηνές αναφέρονται σε σκηνές που μπορούμε να δούμε στην φύση, όπως δέντρα και βουνά. Το ανθρώπινο οπτικό σύστημα αναπτύσσεται στην φύση και είναι εξοικειωμένο στο να αναγνωρίζει χαρακτηριστικά που υπάρχουν στο φυσικό περιβάλλον. Αυτήν την τεχνική μπορούμε να δούμε στο παρακάτω Σχήμα.



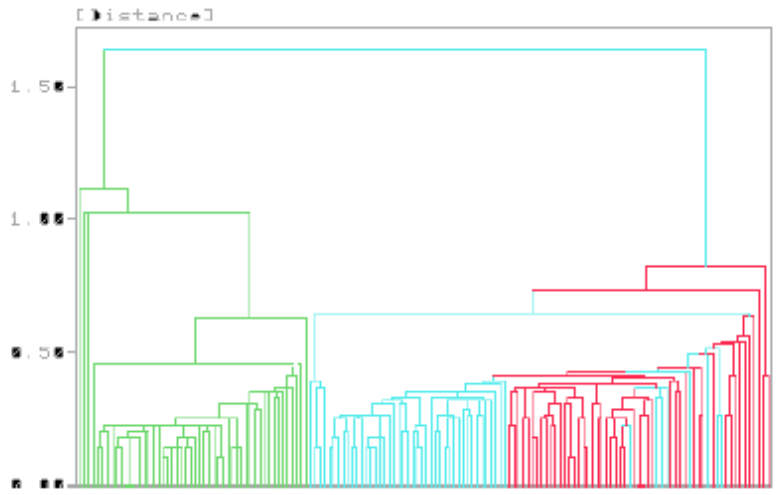
Σχήμα 41: Ένα παράδειγμα εικονογραφημάτων με φυσικά τοπία [65].

## 4.3 Ιεραρχικές Τεχνικές

Οι ιεραρχικές τεχνικές υποδιαιρούν το χώρο των δεδομένων σε υποχώρους χρησιμοποιώντας μια ιεραρχική δομή. Τα χαρακτηριστικά μπορούν να αντιμετωπιστούν με διαφορετικούς τρόπους, κάθε διαφορετική απεικόνιση παράγει και άλλη οπτική των δεδομένων. Συνεπώς, η ερμηνεία της κάθε οπτικής απαιτεί κατάρτιση. Οι τεχνικές αυτές χρησιμοποιούνται για ιεραρχικά δεδομένα, αλλά και όταν κάποια από τα χαρακτηριστικά είναι πιο ενδιαφέροντα για τους χρήστες.

### 4.3.1 Δενδρογράμματα

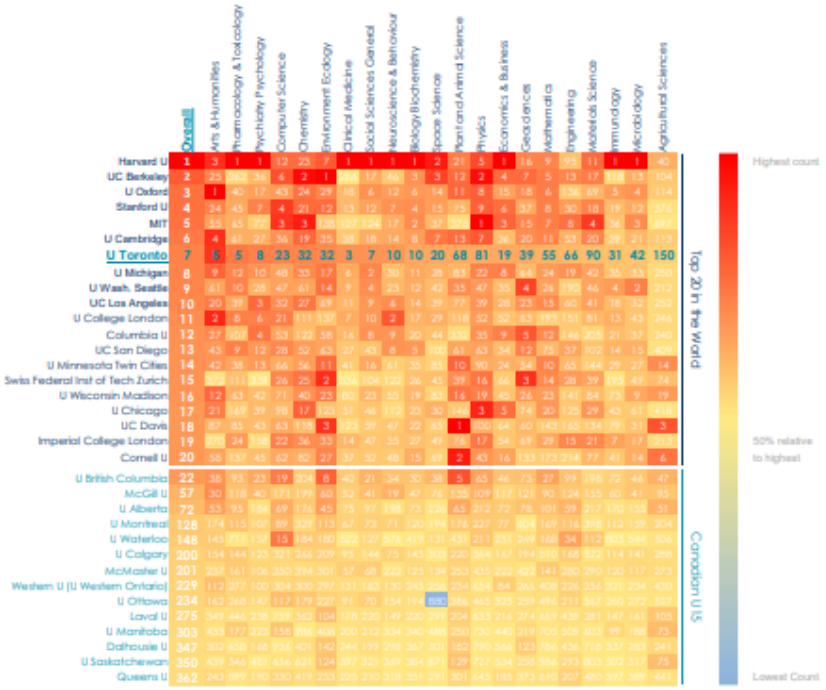
Τα δενδροδιαγράμματα χρησιμοποιούν έναν ιεραρχικό διαχωρισμό της οθόνης, που εξαρτάται από τις τιμές των χαρακτηριστικών (Σχήμα 42) [7]. Όταν δύο παρατηρήσεις ομαδοποιούνται στο διάγραμμα, χρησιμοποιείται μια οριζόντια και δύο κάθετες γραμμές για να δηλώσουν το σχηματισμό της αντίστοιχης συστάδας. Το χρώμα της κάθε περιοχή μπορεί να παραστήσει ένα επιπλέον χαρακτηριστικό. Τα δενδρογράμματα χρησιμοποιούνται για την οπτικοποίηση πολυδιάστατων δεδομένων με ιεραρχική δομή. Ένα μειονέκτημα των δενδροδιαγραμμάτων είναι ότι για μεγάλα σύνολα δεδομένων δεν είναι αποτελεσματικά, καθώς απαιτούν πολύ μεγάλη επιφάνεια οπτικοποίησης.




Σχήμα 42: Δενδροδιάγραμμα (hierarhical clustering) του συνόλου δεδομένων Iris [65].

### 4.3.2 Θερμοχάρτες

Οι θερμοχάρτες (**Heatmaps**) χρησιμοποιούν τα χρώματα για να απεικονίσουν τα πολυδιάστατα δεδομένα μεταβάλλοντας τις αποχρώσεις των χρωμάτων (Σχήμα 43) [7]. Κάθε στήλη αντιστοιχεί και σε μία μεταβλητή και οι γραμμές αντιπροσωπεύουν τις πολυδιάστατες παρατηρήσεις. Μπορούν να χρησιμοποιηθούν για να δείξουν την πυκνότητα του πληθυσμού μίας χώρας χρησιμοποιώντας τα φάσματα των χρωμάτων. Τα χρώματα χρησιμοποιούνται για να απεικονίσουν την συχνότητα κάθε πολυδιάστατης παρατήρησης, αυτός είναι ο λόγος που είναι εύκολη η ερμηνεία τους. Για παράδειγμα όταν ένας πίνακας συσχετίσεων αντικαθίσταται από έναν χάρτη θερμότητας μπορούμε να μεταβάλλουμε το χρώμα του φόντου από το λευκό στο κόκκινο για να αντιστοιχίσουμε τις συσχετίσεις από το 0 στο 1 για παράδειγμα.

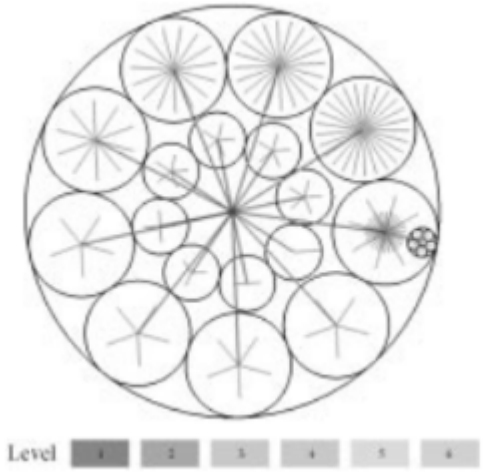



Σχήμα 43: Ένας θερμοχάρτης, ο οποίος δείχνει των αριθμό των δημοσιεύσεων των μεγαλύτερων πανεπιστημίων, όσο πιο κόκκινο γίνεται το χρώμα τόσο περισσότερες δημοσιεύσεις έχει το πανεπιστήμιο στην συγκεκριμένη ερευνητική περιοχή [61].

### 4.3.3 Δακτυλίδια

Τα Δακτυλίδια (**Rings**) είναι μια τεχνική οπτικοποίησης μεγάλων συνόλων, μεγάλης διάστασης δεδομένων με ιεραρχική δομή (Σχήμα 44) [10]. Το μεγάλο πλεονέκτημα της μεθόδου αυτής είναι ότι χρησιμοποιεί αποτελεσματικά τον περιορισμένο χώρο απεικόνισης προβάλλοντας περισσότερους κόμβους μαζί. Η μέθοδος μας δίνει και την δυνατότητα δια-

δραστικής πλοήγησης στο γράφημα παρέχοντας μας επιπλέον πληροφόρηση για τα δεδομένα χωρίς να επηρεάζεται η σαφήνεια της περιοχής που μας ενδιαφέρει.



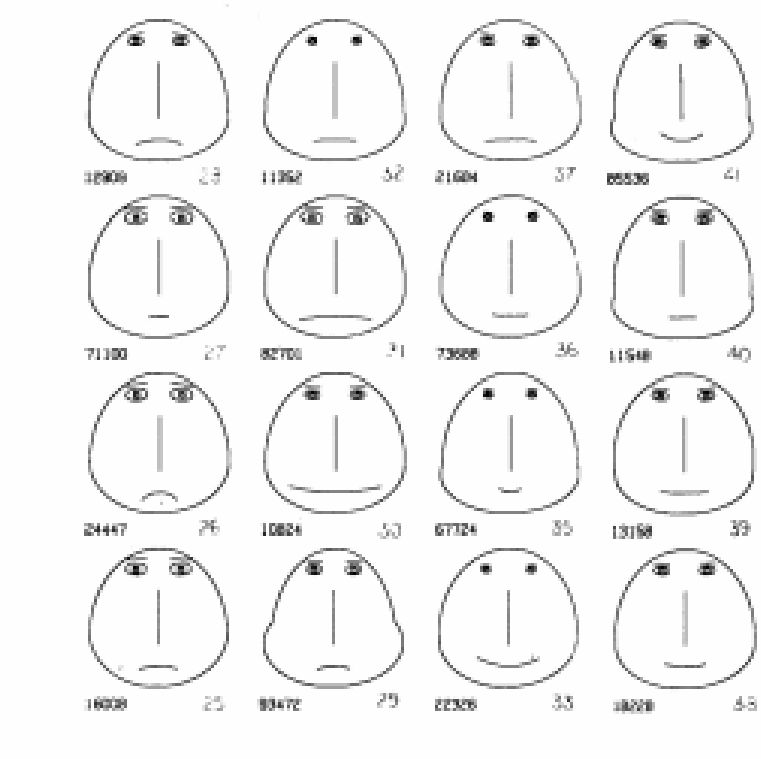

Σχήμα 44: Η τεχνική των δακτυλιδιών [10].

## 4.4 Εικονογραφικές Τεχνικές

Οι Εικονογραφικές Τεχνικές τοποθετούν κάθε πολυδιάστατη παρατήρηση σε ένα εικονίδιο. Τα γεωμετρικά χαρακτηριστικά, οι χρωματισμοί ή η θέση των εικονιδίων χρησιμοποιούνται για να οπτικοποιηθούν οι τιμές των μεταβλητών των δεδομένων. Με αυτόν τον τρόπο μπορούμε να παρατηρήσουμε εικονίδια που μοιάζουν μεταξύ τους και άλλα που δεν μοιάζουν και να προβούμε σε ομαδοποιήσεις. Αυτές οι τεχνικές οπτικοποίησης είναι οικίες για το ανθρώπινο μάτι. Οι τεχνικές αυτές είναι αποτελεσματικές για σύνολα δεδομένων με περιορισμένο σύνολο μεταβλητών.

### 4.4.1 Πρόσωπα του **Chernoff**

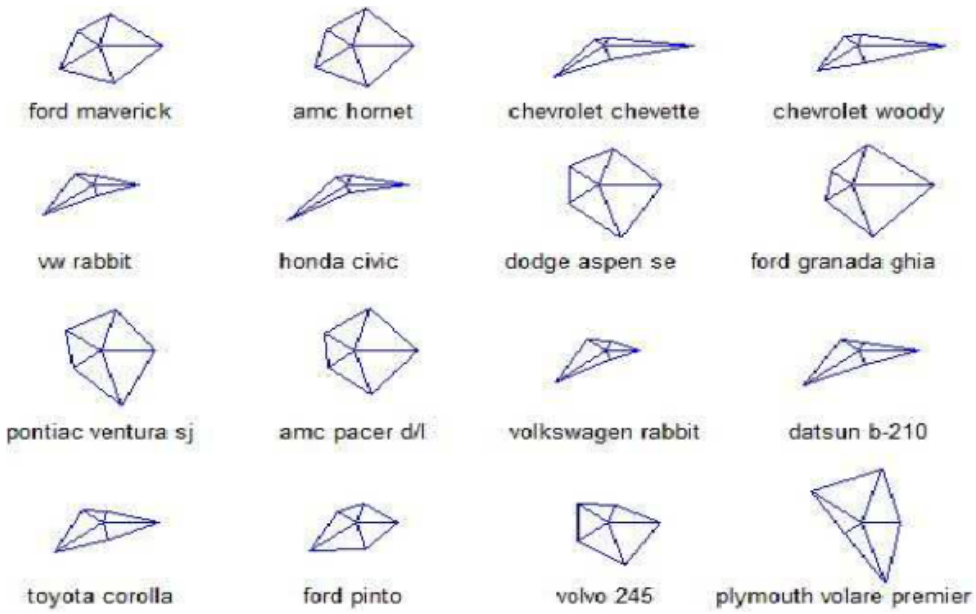
Τα Πρόσωπα του Chernoff είναι ίσως η πιο γνωστή τεχνική απεικόνισης από τις εικονογραφικές τεχνικές [21]. Ο Herman Chernoff εισήγαγε τα πρόσωπα του Chernoff με τα οποία μπορούμε να παρουσιάσουμε πολυδιάστατα δεδομένα, διάστασης μέχρι 8, μέσω εικονιδίων τα οποία μοιάζουν με ανθρώπινα πρόσωπα (Σχήμα 45). Σε κάθε χαρακτηριστικό του προσώπου αντιστοιχίζουμε μια μεταβλητή (μέγεθος ματιών, μέγεθος μύτης, σχήμα προσώπου κ.α). Αυτή η τεχνική είναι εξαιρετικά χρήσιμη για να ανακαλύψουμε τις συσχετίσεις μεταξύ των μεταβλητών, αλλά και ακραίες παρατησήσεις. Από την άλλη μεριά είναι δύσκολο να συγκρίνουμε τα διαφορετικά πρόσωπα, αυτός είναι και ο λόγος που πολλοί κατά την χρήση τους επικεντρώνονται μόνο σε κάποια χαρακτηριστικά που τους ενδιαφέρουν.



Σχήμα 45: Πρόσωπα του Chernoff [21].

### 4.4.2 Εικονίδια Αστέρων

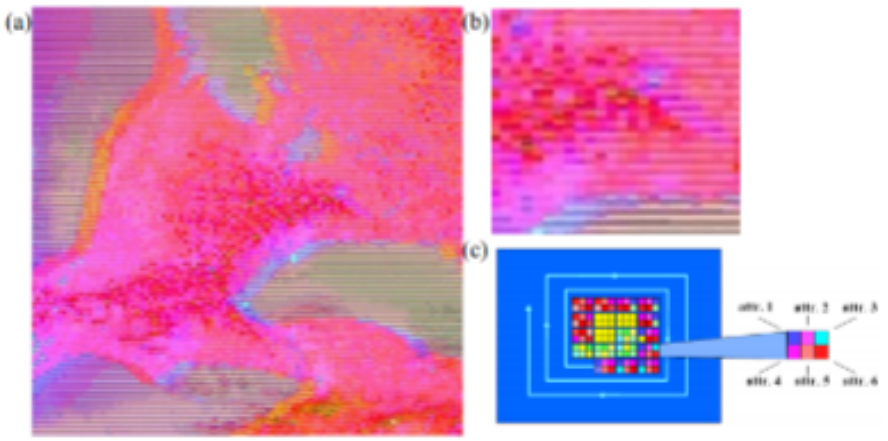
Υπάρχουν πολλές παραλλαγές στην οικογένεια των εικονιδίων, μια από αυτές είναι και τα Εικονίδια Αστέρων (Star Plots), τα οποία παρουσιάστικαν από τον Chamber [61]. Είναι μια ακόμα τεχνική οπτικοποίησης πολυδιάστατων δεδομένων. Η κάθε πολυμεταβλητή παρατήρηση απεικονίζεται με την μορφή ενός αστεριού, το οποίο έχει τόσες ακτίνες όσες είναι και οι μεταβλητές (Σχήμα 46). Οι ακτίνες ισαπέχουν μεταξύ τους και το μήκος κάθε μίας υποδηλώνει την τιμή της αντίστοιχης μεταβλητής. Τέλος, οι ακτίνες ενώνονται μεταξύ τους σχηματίζοντας ένα πολύγωνο. Για μεγάλο πλήθος παρατηρήσεων δεν παρέχουν ουσιαστική πληροφόρηση, καθώς δεν υπάρχει αρκετός χώρος για να απεικονισθούν και έτσι δημιουργείται μια χαοτική εικόνα.




Σχήμα 46: Διαγράμματα Αστέρων για το σύνολο δεδομένων mtcars [54].

### 4.4.3 Εικονίδια Χρώματος

Η τεχνική των Εικονιδίων Χρώματος (Color Icon) είναι ένας συνδιασμός σπειροειδών αξόνων που βασίζονται σε εικονίδια (Σχήμα 47) [54]. Τα εικονοστοιχεία αντικαθίστανται από συστοιχίες πεδίων χρωμάτων που αντιπροσωπεύουν τις τιμές των χαρακτηριστικών. Για δεδομένα διάστασης n θα χρειαστούν n χρωματισμένα εικονοστοιχεία. Το χρώμα, το σχήμα, το μέγεθος, τα όρια και οι υποδιαιρέσεις της κάθε περιοχής μπορούν να χρησιμοποιηθούν για να απεικονισθούν τα διάφορα χαρακτηριστικά των πολυδιάστατων δεδομένων.



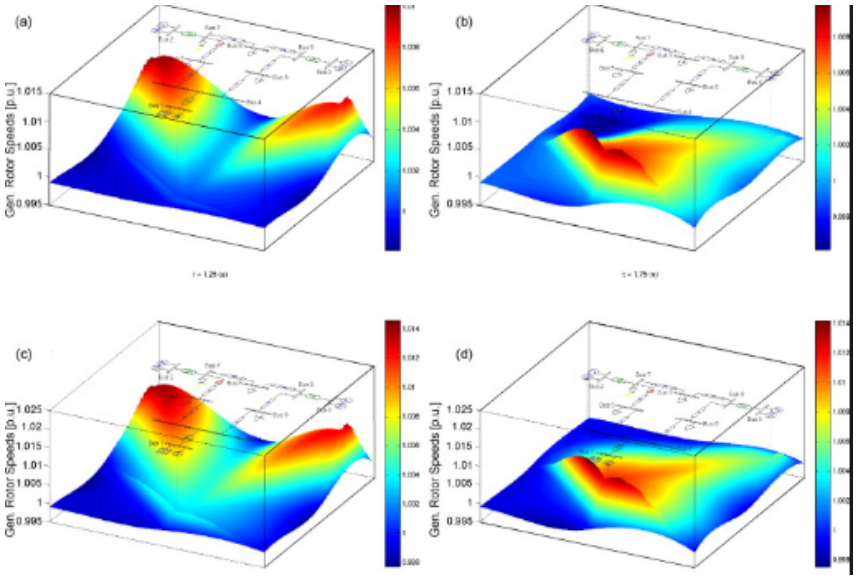

Σχήμα 47: (*a*) Απεικόνιση δεδομένων διάστασης 5 με εικονίδια χρώματος, (*b*) τμήμα της (*a*) τα δεδομένα μεγεθυμένα, (*c*) σχεδιάγραμμα της τεχνικής των εικονιδίων χρώματος [35].

### 4.4.4 Ένα μοντέλο απεικόνισης με την χρήση κινουμένων σχεδίων

Σήμερα στην εποχή των υπερυπολογιστών που ζούμε μπορούν να παραχθούν εικονίδια σε πραγματικό χρόνο. Για δεδομένα με $n$ μεταβλητές, υπάρχουν $n!$ δυνατές απεικονίσεις των μεταβλητών οι οποίες μπορούν να αντικαταστήσουν τις παραμέτρους των εικονιδίων. Παρέχεται ένα σύνολο δυνατοτήτων στις οποίες συμπεριλαμβάνεται διαδραστική ανάλυση σε πραγματικό χρόνο. Σε αυτές ακόμα συμπεριλαμβάνεται η δυνατότητα οριοθέτησης του μεγέθους του κάθε εικονιδίου, τον προσανατολισμό των εικονιδίων ως προς τους άξονες αναφοράς και τον μέγιστο αριθμό μετατόπισης των εικινιδίων. Οι χρήστες μπορούν να προσαρμόσουν αυτές τις παραμέτρους ελέγχου μέχρι να εντοπιστούν ενδιαφέροντα πρότυπα. Αυτό το είδος συνεχούς παραμετρικής αλληλεπίδρασης μπορεί να ληφθεί υπόψη ως η τέλεια φυσική αναπαράστση των δεδομένων.

Ο **Bragatto et al.** περιγράφει ένα μοντέλο που έχει σχεδιαστεί για να ζωντανέψει σε κλιμακωτά κινούμενα επίπεδα (**A Scalar Visuallization Animation Model**) τα δεδομένα μεγάλης διάστασης (Σχήμα 48). Τα δεδομένα που μπορεί να υποστηρίξει αυτή η μέθοδος είναι μεγάλης διάστασης ογκομετρικά δεδομένα. Τρεις από τις πιθανές τεχνικές κίνησης που συναντάμε εδώ είναι μια ακολουθία τρισδιάστων κινούμενων όγκων οι οποίοι κινούνται σύμφωνα με τον χρόνο, ένας στατικός όγκος αν ο χρόνος δεν συμπεριλαμβάνεται ή μια σειρά επιπέδων εάν η μια διάσταση δεν χρησιμοποιείται. Τα δεδομένα μπορεί να είναι χωρικά (δηλαδή ένα υποσύνολο του χώρου τριών διαστάσεων) ή χρονικά (δηλαδή ένα συγκεκριμένο χρονικό διάστημα). Τα δεδομένα απεικονίζονται σε τρισδιάστατα σκιασμένα πλέγματα που χαρακτηρίζονται από το χρώμα και την υφή για κάθε σημείο του πλέγματος. Μια κινούμενη εικόνα αποτελείται

από μια ή περισσότερες σκηνές. Υπάρχει ακόμα ένα 'ρολόϊ' του συστήματος για να ρυθμίζει την ταχύτητα της κίνησης της εικόνας.




Σχήμα 48: Ένα μοντέλο κλιμακωτής απεικόνισης κινουμένων σχεδίων [12].

# 5 Έφαρμογή σε Πραγματικά Δεδομένα

## 5.1 Περιγραφή των Συνόλου Δεδομένων

Το πρώτο σύνολο δεδομένων με το οποίο θα ασχοληθούμε περιέχει 392 προϊόντα και 258 μεταβλητές για τα προϊόντα αυτά, δηλαδή έχουμε έναν πίνακα $392x258$. Τα προϊόντα τα οποία αναφερόμαστε στο πρώτο σύνολο δεδομένων είναι κατεψυγμένα ψάρια. Για κάθε προϊόν τα χαρακτηριστικά μεταξύ άλλων είναι η μάρκα, πόσα τεμάχια έχει η συσκευασία και άλλα (Brand, Segment, Forme, κ.λ.π.).

Το δεύτερο σύνολο δεδομένων με το οποίο θα ασχοληθούμε περιέχει 332 προϊόντα και 259 μεταβλητές για τα προϊόντα αυτά, δηλαδή έχουμε έναν πίνακα $332x259$. Τα προϊόντα τα οποία αναφερόμαστε στο δεύτερο σύνολο δεδομένων είναι αναψυκτικά. Για κάθε προϊόν τα χαρακτηριστικά μεταξύ άλλων είναι η μάρκα, η γεύση, η ποσότητα και άλλα (Brand, Flavour, Size, κ.λ.π.)

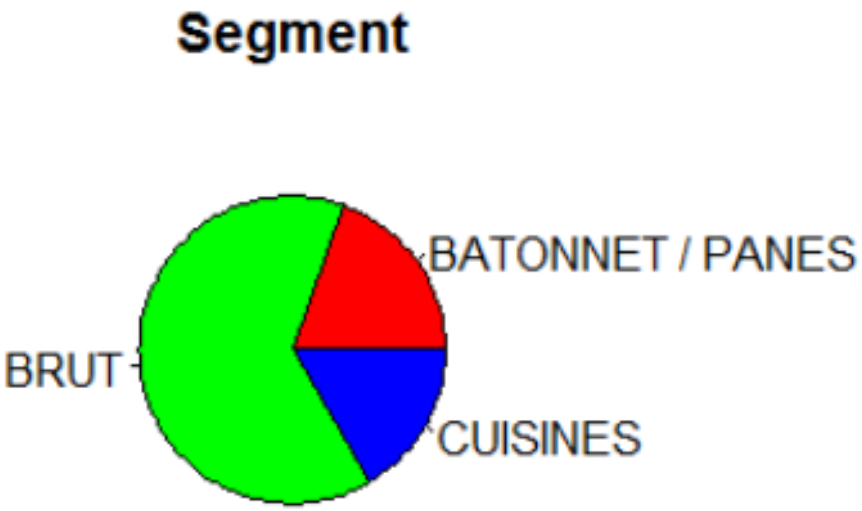
Το τρίτο σύνολο δεδομένων με το οποίο θα ασχοληθούμε περιέχει 490 προϊόντα και 7133 μεταβλητές για τα προϊόντα αυτά, δηλαδή έχουμε έναν πίνακα $490x7133$. Τα προϊόντα τα οποία αναφερόμαστε στο τρίτο σύνολο δεδομένων είναι μαρμελάδες. Για κάθε προϊόν τα χαρακτηριστικά μεταξύ άλλων είναι η μάρκα, η γεύση, η συνταγή και άλλα (Brand, Flavour, Recipe, κ.λ.π.)

Σε όλα τα σύνολα δεδομένων θα εφαρμόσουμε τις ίδιες τεχνικές οπτικοποίησης στις οποίες θα αναφερθούμε στην συνέχεια.

## 5.2 Περιγραφική Στατιστική

### 5.2.1 Κατεψυγμένα Ψάρια

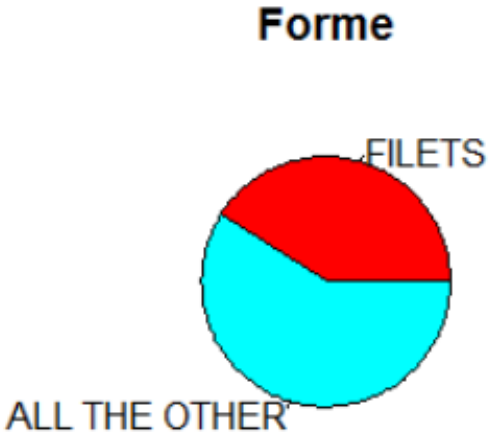
Για το σύνολο δεδομένων με τα κατεψυγμένα ψάρια για τον τρόπο τεμαχισμού (Segment) των ψαριών, 250 ολόκληρα ψάρια (Brut), 77 σε κροκέτες (Batonnet/Panes) και 65 έτοιμα για μαγείρεμα (Cuisines).




Σχήμα 49: Τρόπος Τεμαχισμού

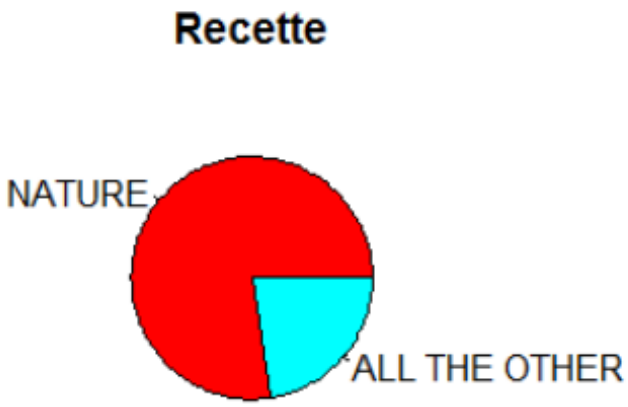
Στα δεδομένα μας έχουμε προϊόντα και από τις 24 μάρκες, με το ποσοστό της μάρκας με τα περισσότερα προϊόντα να ανέρχεται στο 34,4%, ενώ της αμέσως επόμενης στο 14,54%. Έχουμε όλα τα είδη (Forme) ψαριού με την μεγαλύτερη κατηγορία να είναι ένα είδος μπακαλιάρου με ποσοστό 26% και έπεται επίσης ένα άλλο είδος μπακαλιάρου με ποσοστό 16,8%. Οι υπόλοιπες κατηγορίες αντιπροσωπεύονται με πολύ μικρότερα ποσοστά.

Για το σχήμα του ψαριού έχουμε ότι το μεγαλύτερο ποσοστό τους είναι σε φιλέτα το 40,8% (Filets).



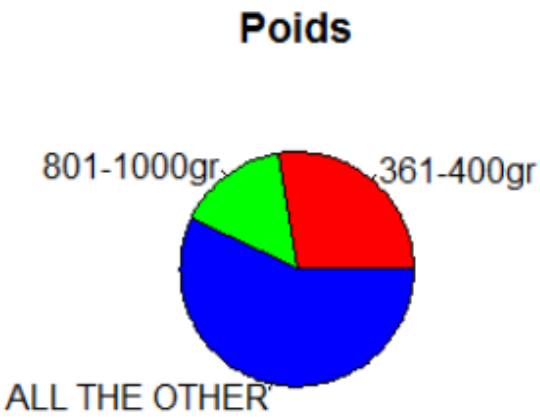

Σχήμα 50: Σχήμα Ψαριού

Η συντριπτική πλειοψηφία των συνταγών στα δεδομένα το 77,55%, μαγειρεύονται με την κλασική συνταγή και οι υπόλοιπες συνταγές αντιπροσωπεύονται με πολύ μικρά ποσοστά.




Σχήμα 51: Τρόπος Μαγειρέματος

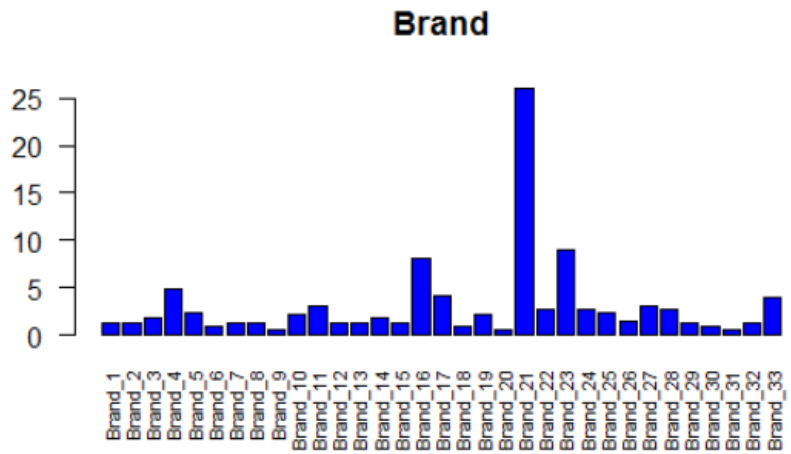
Για την ποσότητα της συσκευασίας των προϊόντων έχουμε ότι το 27,8% είναι σε συσκευασίες των $361-400gr$ και το 15,27% σε συσκευασίες των $801-1000gr$.




Σχήμα 52: Ποσότητα Συσκευασίας

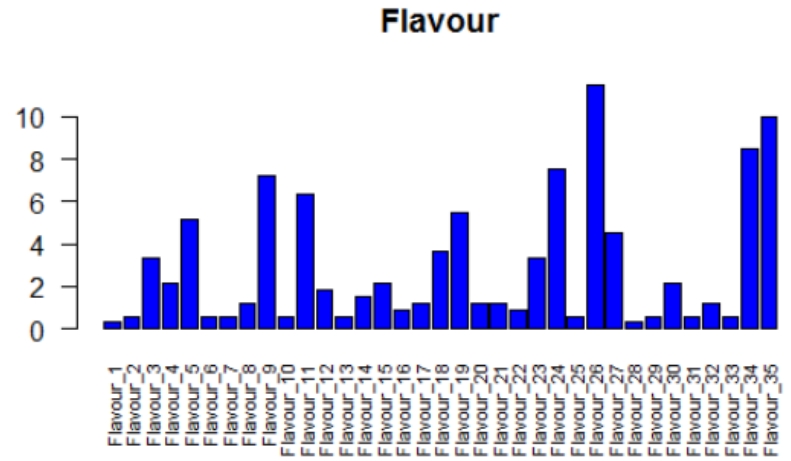
### 5.2.2 Αναψυκτικά

Στα δεδομένα μας έχουμε τριάντα τρεις μάρκες αναψυκτικών. Το 26% των προϊόντων ανήκει σε κάποια άλλη μάρκα από αυτές που έχουμε στα δεδομένα μας και ακολουθούν με χαμηλότερα ποσοστά οι πιο γνωστές μάρκες αναψυκτικών.




Σχήμα 53: Μάρκες Αναψυκτικών

Για τις γεύσεις των αναψυκτικών έχουμε ότι το 10% είναι χωρίς γεύση, το 7,5% πορτοκάλι, το 7% λεμόνι και οι υπόλοιπες γεύσεις σε μικρότερα ποσοστά.

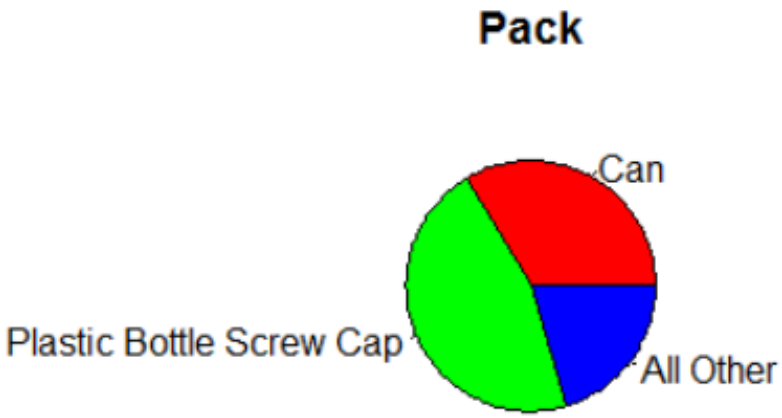



Σχήμα 54: Γεύσεις Αναψυκτικών

Για την ποσότητα έχουμε ότι 17,7% των προϊόντων είναι σε συσκευασίες των 330ml, το 52% σε συσκευασίες των 500ml και το 30,3% σε άλλου μεγέθους συσκευασίες.

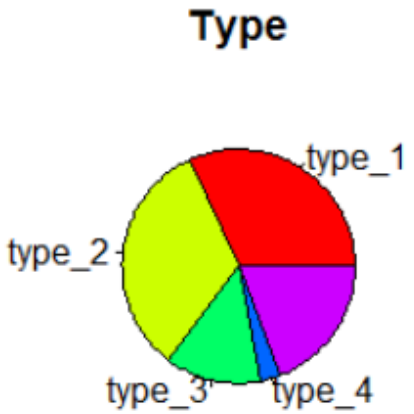

Σχήμα 55: Ποσότητα Συσκευασίας

Για την συσκευασία έχουμε ότι 45,8% των προϊόντων είναι συσκευασμένα σε Plastic Bottle Screw Cap και το 33,4% σε Can.




Σχήμα 56: Συσκευασία

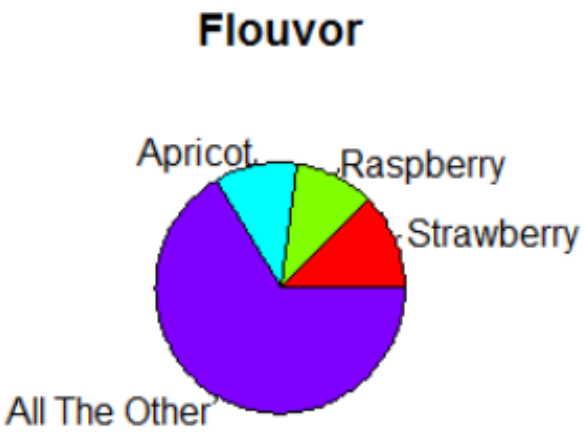
Για τον τύπο του αναψυκτικού έχουμε τέσσερα είδη αναψυκτικών (κανονικά, χωρίς ζάχαρη, light, διαίτης κ.λ.π.). Τα οποία μπορούμε να δούμε παρακάτω πως κατανέμονται στα δεδομένα.




Σχήμα 57: Τύπος Αναψυκτικών

### 5.2.3 Μαρμελάδες

Όσο αφορά τώρα το σύνολο δεδομένων με τις μαρμελάδες. Έχουμε ότι 12,65% έχουν γεύση φράουλα, 10,2% έχουν γεύση βατόμουρο, 10,6% γεύση βερίκοκο και οι υπόλοιπες γεύσεις αντιπροσωπεύονται με πολύ μικρά ποσοστά.




Σχήμα 58: Γεύσεις Μαρμελάδων

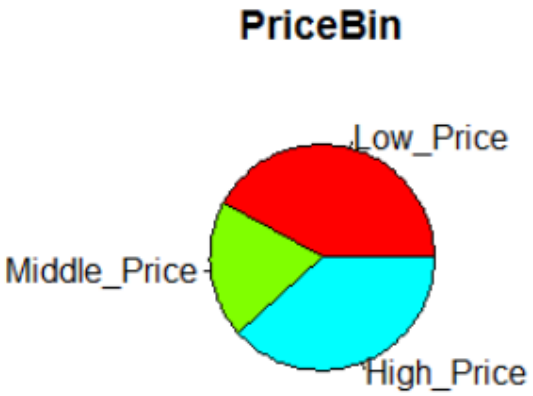
Για τις μάρκες των μαρμελάδων έχουμε ότι το 17,55% των προϊόντων ανήκει στην μάρκα 1, το 18,98% στην μάρκα 2, το 9,6% στην μάρκα 3, το 9,2% στην μάρκα 4 και έπονται άλλες 19 μάρκες με πολύ μικρότερα ποσοστά.


Σχήμα 59: Μάρκες Μαρμελάδων

Για την περιεκτικότητα των μαρμελάδων έχουμε ότι περίπου το 75% από αυτές είναι μαρμελάδες με ολόκληρα κομμάτια φρούτων, ενώ περίπου το 15% είναι μαρμελάδες με φρούτα οι οποίες περιέχουν μικρά κομμάτια φρούτων, σε μορφή ζελέ είναι περίπου το 7,5% και σε κρεμώδη μορφή περίπου το 2,5%.


Σχήμα 60: Περιεκτικότητα Μαρμελάδων

Επιπλέον, περίπου το 42% των προϊόντων ανήκουν στο χαμηλό εύρος τιμών, το 19,6% στο μεσαίο εύρος τιμών και το 38,4% στο υψηλό εύρος τιμών.




Σχήμα 61: Εύρος Τιμών Μαρμελάδων

## 5.3 Οπτικοποίηση των Συνόλων Δεδομένων

### 5.3.1 Εφαρμογή στο σύνολο δεδομένων Κατεψυγμένα Ψάρια

Στην ανάλυση μας θα χρησιμοποιήσουμε τις παρακάτω τεχνικές οπτικοποίησης: **chernoff faces**, **andrews plots** και **radial plots**. Ας δούμε τώρα κάποιες απεικονίσεις του συνόλου δεδομένων με τις παραπάνω τεχνικές.

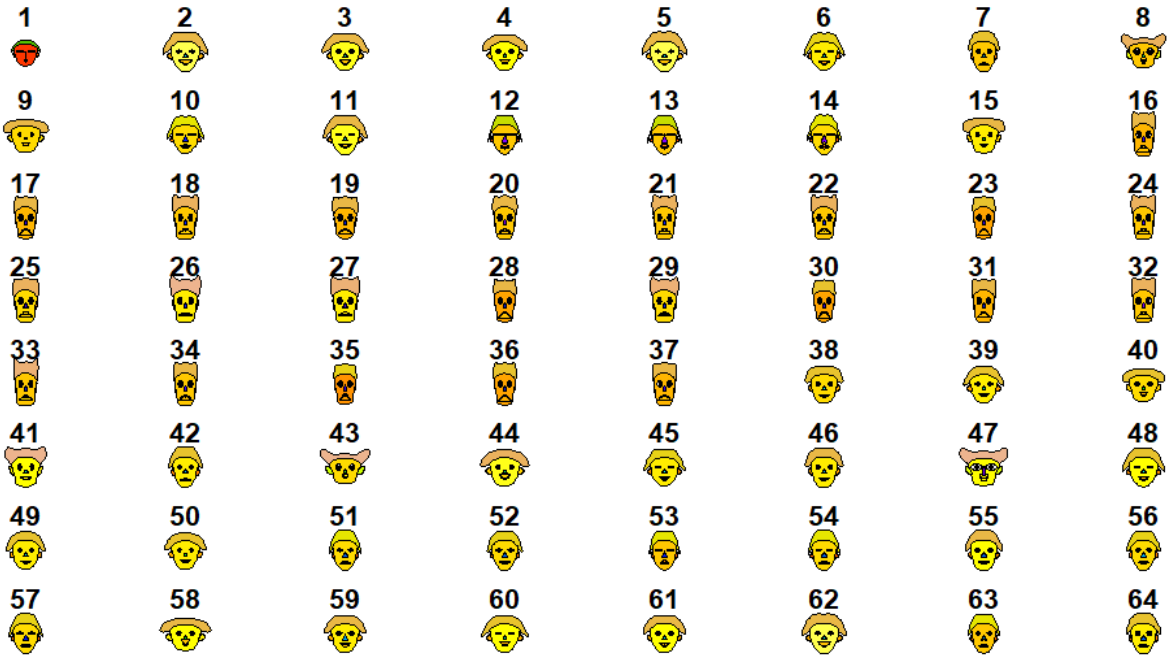
Θα εφαρμόσω πρώτα στα δεδομένα μου, τα οποία αποτελούνται από 392 εγγραφές και 259 μεταβλητές, την **PCA** για να μειώσω την διάσταση των δεδομένων μου. Από το Σχήμα 62 μπορούμε να δούμε ότι η πρώτη μεταβλητή εξηγεί το 28,743 της μεταβλητότητας των δεδομένων και μέχρι και την τέταρτη μεταβλητή εξηγείται το 44,315 της μεταβλητότητας. Οπότε θα απεικονίσω τα δεδομένα μου στις 4 διαστάσεις, επειδή μετά την τέταρτη διάσταση κάθε επόμενη διάσταση εξηγεί πολύ μικρό μέρος της μεταβλητότητας των δεδομένων.

Eigenvalues

| | Dim.1 | Dim.2 | Dim.3 | Dim.4 | Dim.5 | Dim.6 | Dim.7 | Dim.8 |
|---|---|---|---|---|---|---|---|---|
| Variance | 77.032 | 16.025 | 14.306 | 11.402 | 9.120 | 7.634 | 7.062 | 6.462 |
| % of var. | 28.743 | 5.980 | 5.338 | 4.254 | 3.403 | 2.848 | 2.635 | 2.411 |
| Cumulative % of var. | 28.743 | 34.723 | 40.061 | 44.315 | 47.718 | 50.567 | 53.202 | 55.613 |
| | Dim.9 | Dim.10 | Dim.11 | Dim.12 | Dim.13 | Dim.14 | Dim.15 | Dim.16 |
| Variance | 5.202 | 4.321 | 4.009 | 3.403 | 2.918 | 2.786 | 2.639 | 2.501 |
| % of var. | 1.941 | 1.612 | 1.496 | 1.270 | 1.089 | 1.039 | 0.985 | 0.933 |
| Cumulative % of var. | 57.554 | 59.166 | 60.662 | 61.932 | 63.021 | 64.060 | 65.045 | 65.978 |
| | Dim.17 | Dim.18 | Dim.19 | Dim.20 | Dim.21 | Dim.22 | Dim.23 | Dim.24 |
| Variance | 2.336 | 2.153 | 2.130 | 2.004 | 1.898 | 1.817 | 1.809 | 1.683 |
| % of var. | 0.872 | 0.804 | 0.795 | 0.748 | 0.708 | 0.678 | 0.675 | 0.628 |
| Cumulative % of var. | 66.850 | 67.653 | 68.448 | 69.196 | 69.904 | 70.582 | 71.257 | 71.885 |

Σχήμα 62: Πίνακας της PCA που μας δίνει το ποσοστό της μεταβλητότητας που εξηγεί η κάθε μεταβλητή.

Στην συνέχεια θα χρησιμοποιήσουμε της συντεταγμένες για τις 4 διαστάσεις που μας δίνει η PCA και σε αυτές θα εφαρμόσουμε τις τεχνικές οπτικοποίησης που αναφέραμε παραπάνω. Θα ξεκινήσουμε εφαρμόζοντας την τεχνική των chernoff faces.

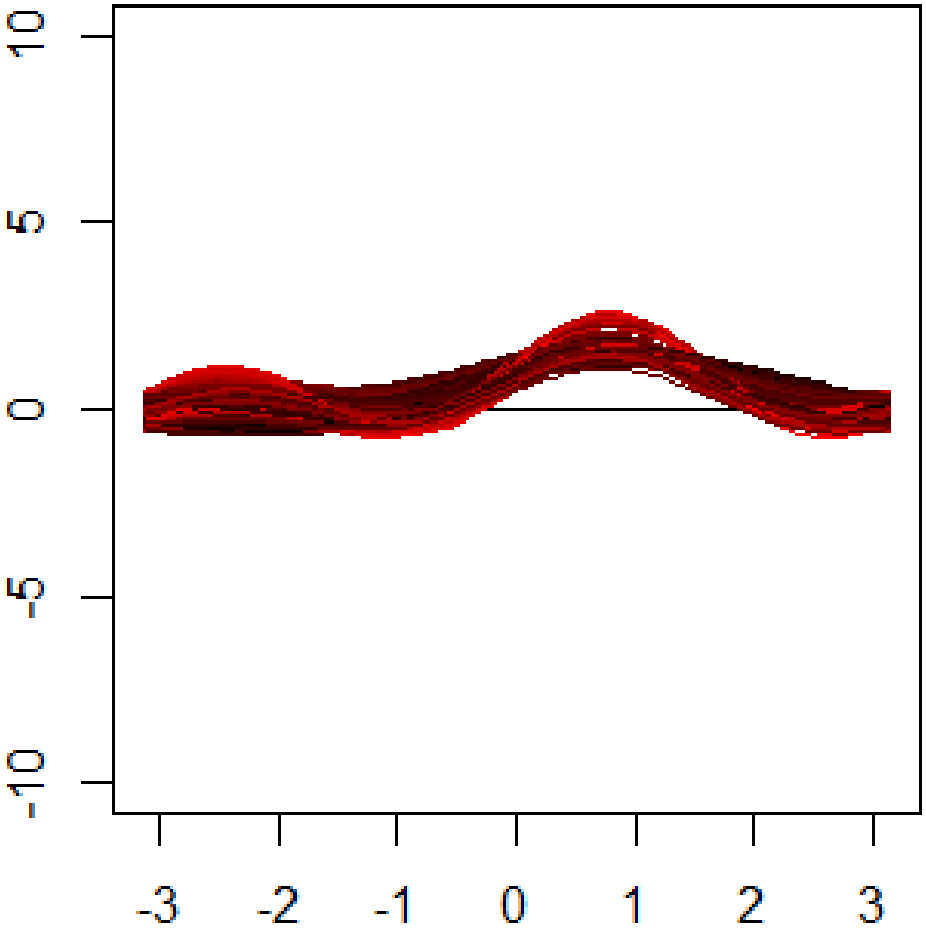



Σχήμα 63: Chernoff faces

Στο Σχήμα 63 παρατηρούμε τα chernoff faces παρατηρούμε τις 63 πρώτες εγγραφές των δεδομένων να αναπαρίστανται η κάθε μια στις τέσσερις διαστάσεις, το μήκος του προσώπου και το μήκος των μαλλιών αναπαριστά την πρώτη διάσταση, το πλάτος του προσώπου και το πλάτος των μαλλιών την δεύτερη διάσταση, η κατασκευή του προσώπου και το είδος του κουρέματος την τρίτη διάσταση και τέλος η τέταρτη διάσταση αναπαρίσταται από το πλάτος

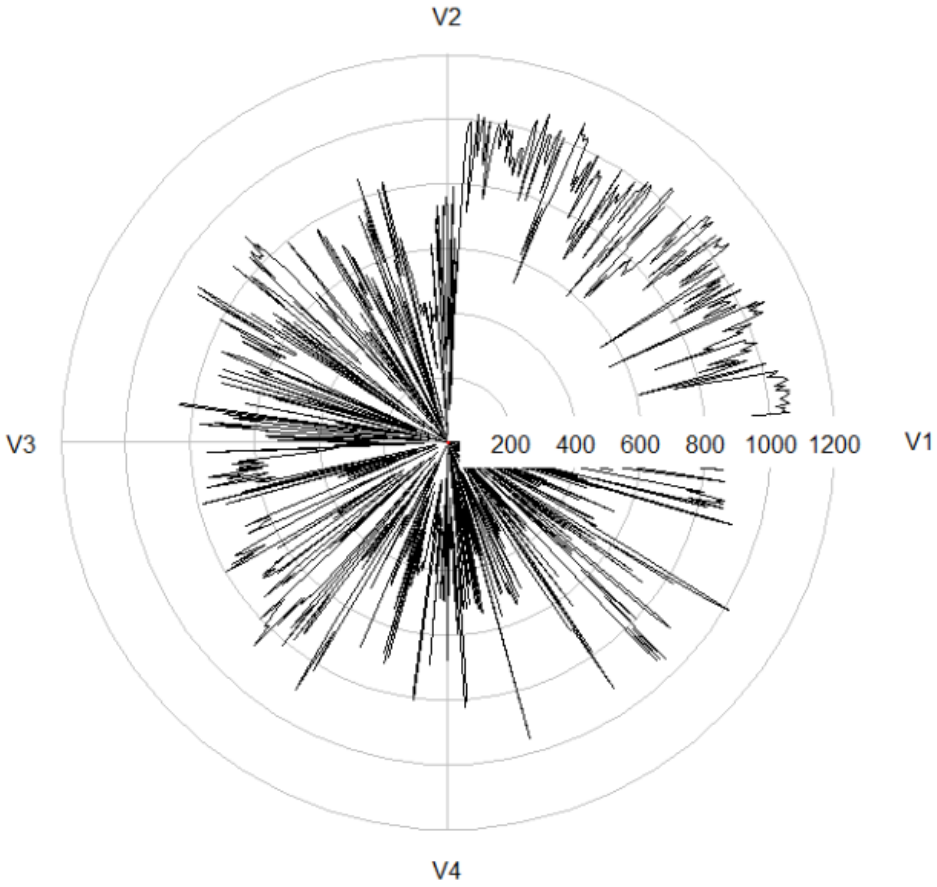
της μύτης και το πλάτος των ματιών. Στο Σχήμα 63 μπορούμε να παρατηρήσουμε ότι τα πρόσωπα 2 έως και 5 μοιάζουν αρκετά μεταξύ τους και αυτό γιατί είναι προϊόντα της ίδιας εταιρείας και μίας άλλης εταιρείας είναι τα προϊόντα 16 έως 25 τα οποία και αυτά μοιάζουν.

Συνεχίζουμε με τα andrews plots. Στο Σχήμα 64 μπορούμε να δούμε τις καμπύλες του andrews για τα δεδομένα μας, παρατηρούμε ότι δεν δίνουν καμιά ουσιαστική πληροφορία.



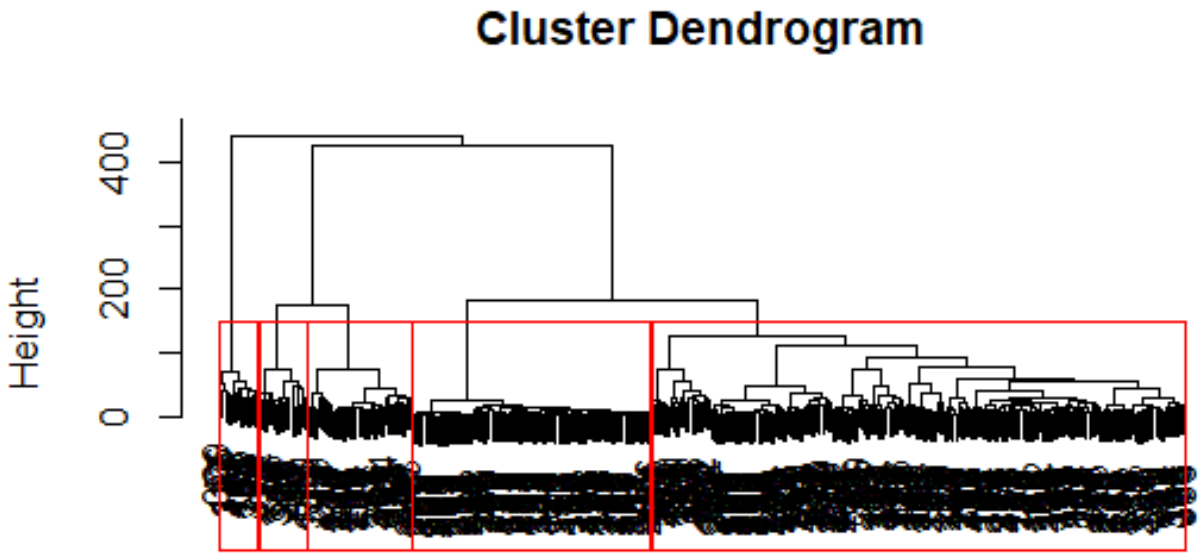

Σχήμα 64: andrews plot

Συνεχίζουμε με τα radial plots. Σε αυτήν την τεχνική απεικονίζονται τα δεδομένα σε κυκλικούς, σε κάθε κυκλικό τομέα αντιστοιχούν και δύο διαφορετικές διαστάσεις. Στο Σχήμα 65 μπορούμε να δούμε ένα radial plot για τα δεδομένα μας από την αναπαράστασή τους στην πρώτη και στην δεύτερη διάσταση παρατηρούμε ότι εκεί παίρνουν τις μεγαλύτερες τιμές τους. Ενώ, από την αναπαράστασή τους στις υπόλοιπες διαστάσεις, παρατηρούμε ότι παίρνουν σχεδόν ίδιες τιμές.




Σχήμα 65: radial plot

Τέλος, θα χρησιμοποιήσουμε ένα δενδρόγραμμα. Στο σχήμα 66 βλέπουμε ένα δενδρόγραμμα για τα δεδομένα μας. Όπως, παρατηρούμε τα δεδομένα μας χωρίζονται σε 5 ομάδες ανάλογα με την τιμή του προϊόντος. Η δεξιά ομάδα περιέχει τα πιο φθηνά προϊόντα και πηγαίνοντας προς τα αριστερά έχουμε τα πιο ακριβά προϊόντα μέχρι την ομάδα στην οποία ανήκουν τα πιο ακριβά προϊόντα, η οποία έχει και τα λιγότερα προϊόντα σε σχέση με τις άλλες.




Σχήμα 66: Δενδρόγραμμα

Μια γενική παρατήρηση που μπορούμε να κάνουμε είναι ότι η τεχνική του δενδρογράμματος μας δίνει καλύτερη πληροφόρηση για το πώς απεικονίζονται οι παρατηρήσεις μας.

### 5.3.2 Εφαρμογή στο σύνολο δεδομένων Αναψυκτικά

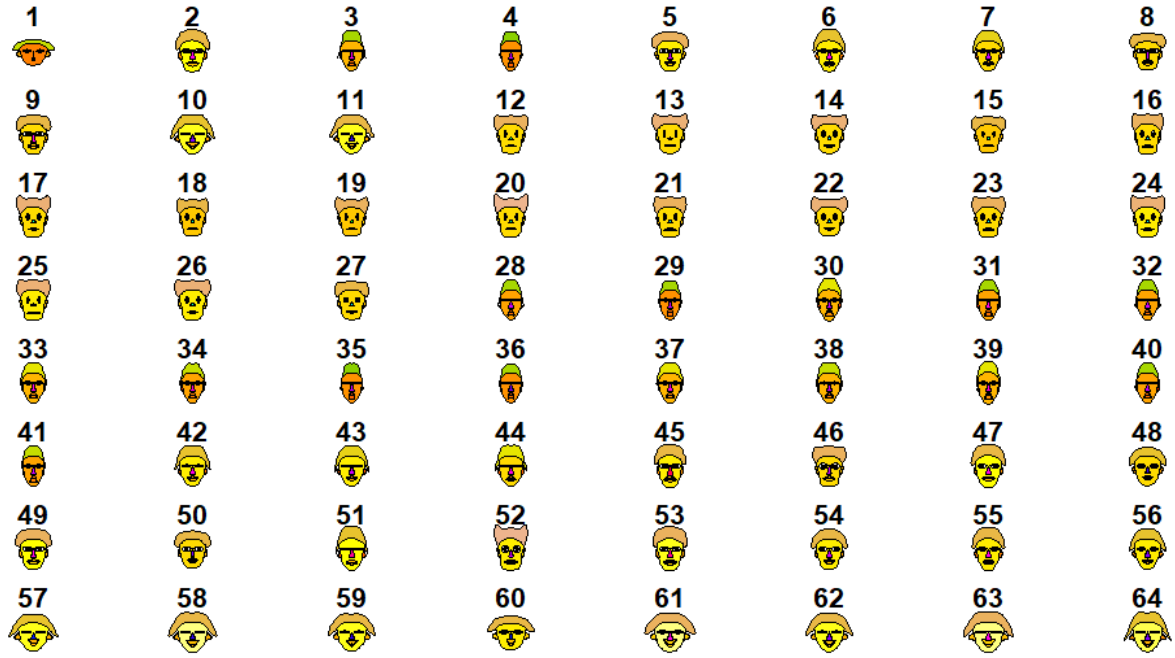
Θα εφαρμόσω πρώτα στα δεδομένα μου, τα οποία αποτελούνται από 392 εγγραφές και 268 μεταβλητές, την PCA για να μειώσω την διάσταση των δεδομένων μου. Από το Σχήμα 66 μπορούμε να δούμε ότι η πρώτη μεταβλητή εξηγεί το 25,47 της μεταβλητότητας των δεδομένων και μέχρι και την τέταρτη μεταβλητή εξηγείται το 40,236 της μεταβλητότητας. Οπότε θα απεικονίσω τα δεδομένα μου στις 4 διαστάσεις, επειδή μετά την τέταρτη διάσταση κάθε επόμενη διάσταση εξηγεί πολύ μικρό μέρος της μεταβλητότητας των δεδομένων.

Eigenvalues

| | Dim.1 | Dim.2 | Dim.3 | Dim.4 | Dim.5 | Dim.6 | Dim.7 | Dim.8 |
|---|---|---|---|---|---|---|---|---|
| Variance | 68.261 | 16.228 | 12.227 | 11.118 | 8.423 | 7.887 | 6.726 | 4.666 |
| % of var. | 25.470 | 6.055 | 4.562 | 4.148 | 3.143 | 2.943 | 2.510 | 1.741 |
| Cumulative % of var. | 25.470 | 31.526 | 36.088 | 40.236 | 43.380 | 46.322 | 48.832 | 50.573 |
| | Dim.9 | Dim.10 | Dim.11 | Dim.12 | Dim.13 | Dim.14 | Dim.15 | Dim.16 |
| Variance | 4.228 | 3.943 | 3.697 | 3.370 | 3.196 | 2.774 | 2.748 | 2.573 |
| % of var. | 1.578 | 1.471 | 1.380 | 1.257 | 1.193 | 1.035 | 1.025 | 0.960 |
| Cumulative % of var. | 52.151 | 53.622 | 55.002 | 56.259 | 57.452 | 58.487 | 59.512 | 60.472 |
| | Dim.17 | Dim.18 | Dim.19 | Dim.20 | Dim.21 | Dim.22 | Dim.23 | Dim.24 |
| Variance | 2.533 | 2.272 | 2.247 | 2.197 | 2.040 | 1.997 | 1.953 | 1.881 |
| % of var. | 0.945 | 0.848 | 0.838 | 0.820 | 0.761 | 0.745 | 0.729 | 0.702 |
| Cumulative % of var. | 61.418 | 62.266 | 63.104 | 63.923 | 64.685 | 65.430 | 66.159 | 66.860 |
| | Dim.25 | Dim.26 | Dim.27 | Dim.28 | Dim.29 | Dim.30 | Dim.31 | Dim.32 |
| Variance | 1.790 | 1.779 | 1.729 | 1.704 | 1.642 | 1.615 | 1.610 | 1.560 |
| % of var. | 0.668 | 0.664 | 0.645 | 0.636 | 0.613 | 0.603 | 0.601 | 0.582 |
| Cumulative % of var. | 67.529 | 68.192 | 68.838 | 69.473 | 70.086 | 70.689 | 71.289 | 71.871 |
| | Dim.33 | Dim.34 | Dim.35 | Dim.36 | Dim.37 | Dim.38 | Dim.39 | Dim.40 |
| Variance | 1.533 | 1.491 | 1.445 | 1.374 | 1.347 | 1.325 | 1.264 | 1.263 |
| % of var. | 0.572 | 0.556 | 0.539 | 0.513 | 0.503 | 0.494 | 0.472 | 0.471 |
| Cumulative % of var. | 72.443 | 73.000 | 73.539 | 74.052 | 74.554 | 75.049 | 75.520 | 75.992 |
| | Dim.41 | Dim.42 | Dim.43 | Dim.44 | Dim.45 | Dim.46 | Dim.47 | Dim.48 |
| Variance | 1.232 | 1.207 | 1.184 | 1.179 | 1.148 | 1.122 | 1.106 | 1.067 |
| % of var. | 0.460 | 0.450 | 0.442 | 0.440 | 0.429 | 0.419 | 0.413 | 0.398 |
| Cumulative % of var. | 76.452 | 76.902 | 77.344 | 77.784 | 78.212 | 78.631 | 79.043 | 79.441 |

Σχήμα 67: Πίνακας της PCA που μας δίνει το ποσοστό της μεταβλητότητας που εξηγεί η κάθε μεταβλητή.

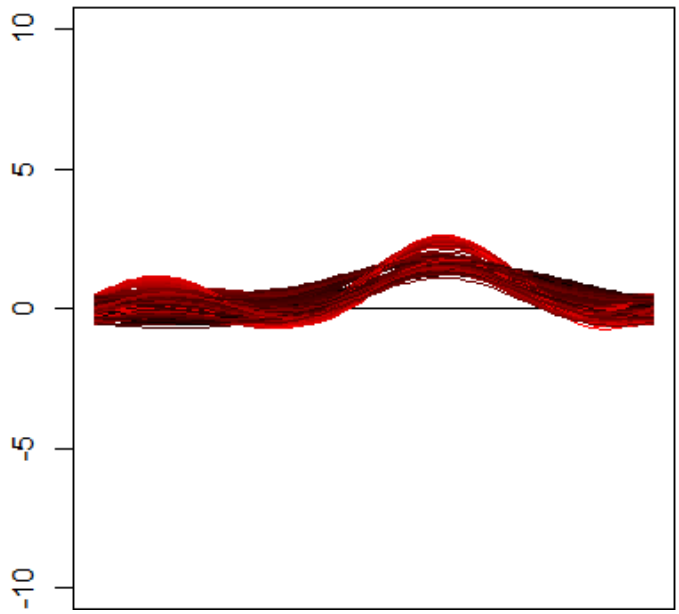
Στην συνέχεια θα χρησιμοποιήσουμε της συντεταγμένες για τις 4 διαστάσεις που μας δίνει η PCA και σε αυτές θα εφαρμόσουμε τις τεχνικές οπτικοποίησης που αναφέραμε παραπάνω. Θα ξεκινήσουμε εφαρμόζοντας την τεχνική των chernoff faces.




Σχήμα 68: Chernoff faces

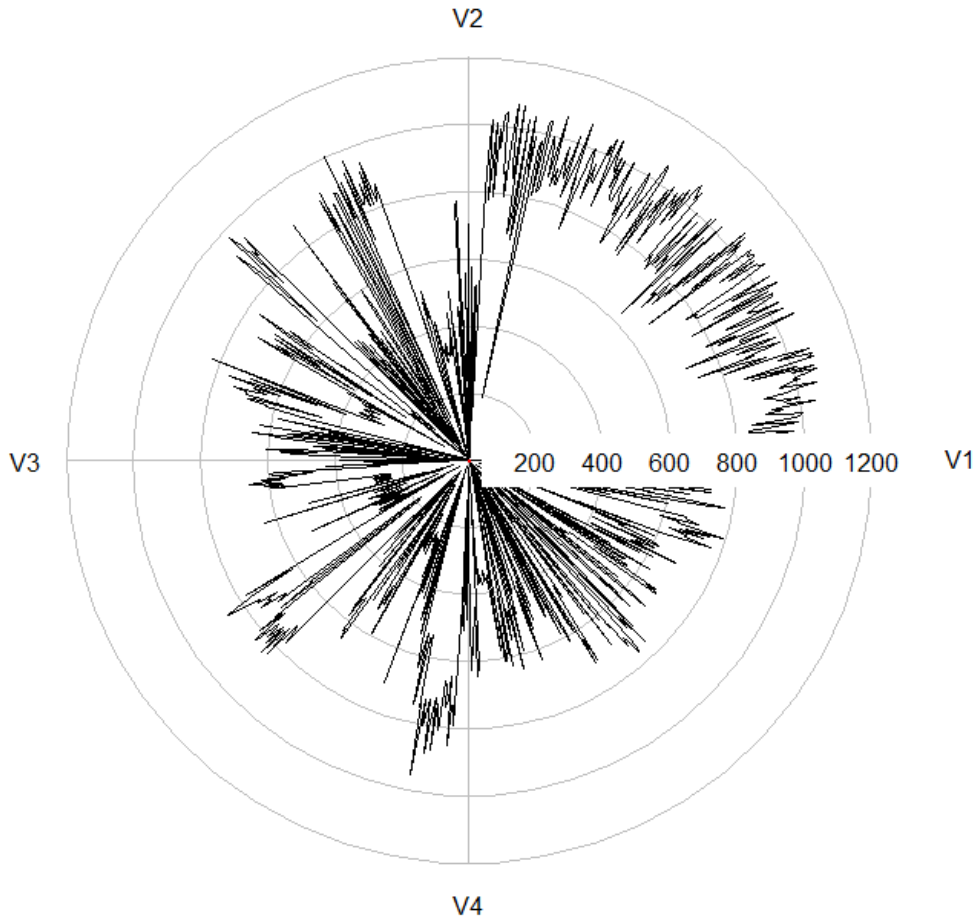
Στο Σχήμα 67 παρατηρούμε τα chernoff faces παρατηρούμε τις 63 πρώτες εγγραφές των δεδομένων να αναπαρίστανται η κάθε μια στις τέσσερις διαστάσεις, το μήκος του προσώπου και το μήκος των μαλλιών αναπαριστά την πρώτη διάσταση, το πλάτος του προσώπου και το πλάτος των μαλλιών την δεύτερη διάσταση, η κατασκευή του προσώπου και το είδος του κουρέματος την τρίτη διάσταση και τέλος η τέταρτη διάσταση αναπαρίσταται από το πλάτος της μύτης και το πλάτος των ματιών. Στο Σχήμα 63 μπορούμε να παρατηρήσουμε ότι τα πρόσωπα 17 έως και 23 μοιάζουν αρκετά μεταξύ τους και αυτό γιατί είναι προϊόντα της ίδιας εταιρείας και μίας άλλης εταιρείας είναι τα προϊόντα 33 έως 40 τα οποία και αυτά μοιάζουν.

Συνεχίζουμε με τα andrews plots. Στο Σχήμα 68 μπορούμε να δούμε τις καμπύλες του andrews για τα δεδομένα μας, παρατηρούμε ότι δεν δίνουν καμιά ουσιαστική πληροφορία.



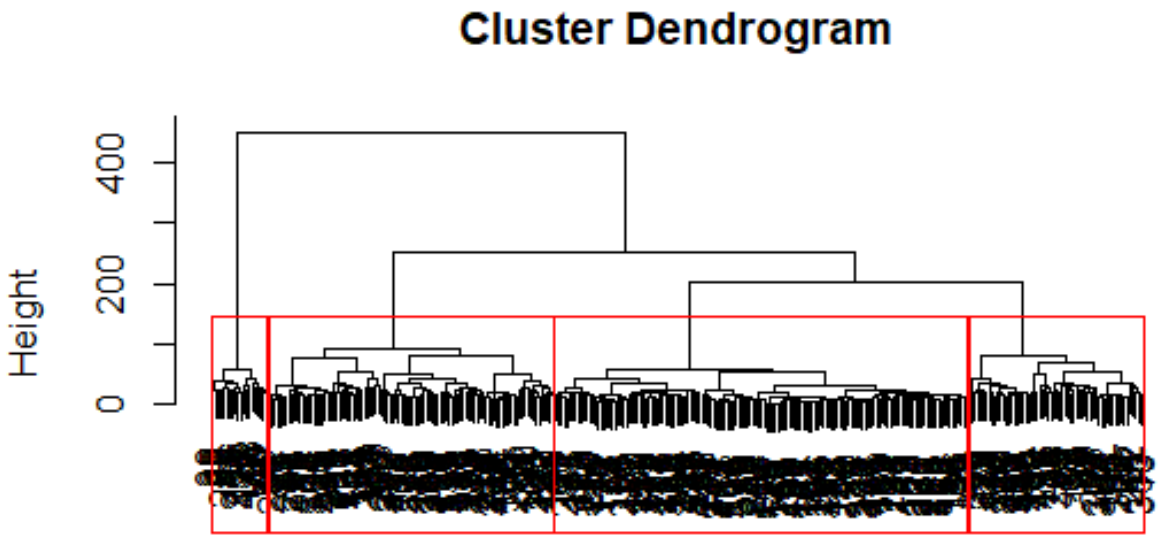

Σχήμα 69: andrews plot

Συνεχίζουμε με τα **radial plots**. Στο Σχήμα 69 μπορούμε να δούμε ένα **radial plot** για τα δεδομένα μας, παρατηρούμε ότι στην πρώτη και στην δεύτερη διάσταση οι παρατηρήσεις μας παίρνουν τις μεγαλύτερες τιμές τους, ενώ τις μικρότερες τιμές τους στην πρώτη και τέταρτη διάσταση.




Σχήμα 70: radial plot

Τέλος, θα χρησιμοποιήσουμε ένα δενδρόγραμμα. Στο σχήμα 71 βλέπουμε ένα δενδρόγραμμα για τα δεδομένα μας. Όπως, παρατηρούμε τα δεδομένα μας χωρίζονται σε 4 ομάδες ανάλογα με το είδος του αναψυκτικού. Στην δεύτερη από δεξιά ομάδα η οποία είναι και η μεγαλύτερη ανήκουν τα κανονικού τύπου αναψυκτικά, στην πρώτη από δεξιά τα αναψυκτικά τύπου cola, στην τρίτη ομάδα τα διαιτητικά αναψυκτικά και στην τελευταία η οποία είναι και η μικρότερη ανήκουν τα ενεργειακά ποτά.




Σχήμα 71: Δενδρόγραμμα

Μια γενική παρατήρηση που μπορούμε να κάνουμε είναι ότι η τεχνική του δενδρογράμματος μας δίνει καλύτερη πληροφόρηση για το πώς απεικονίζονται οι παρατηρήσεις μας.

### 5.3.3 Εφαρμογή στο σύνολο δεδομένων Μαρμελάδες

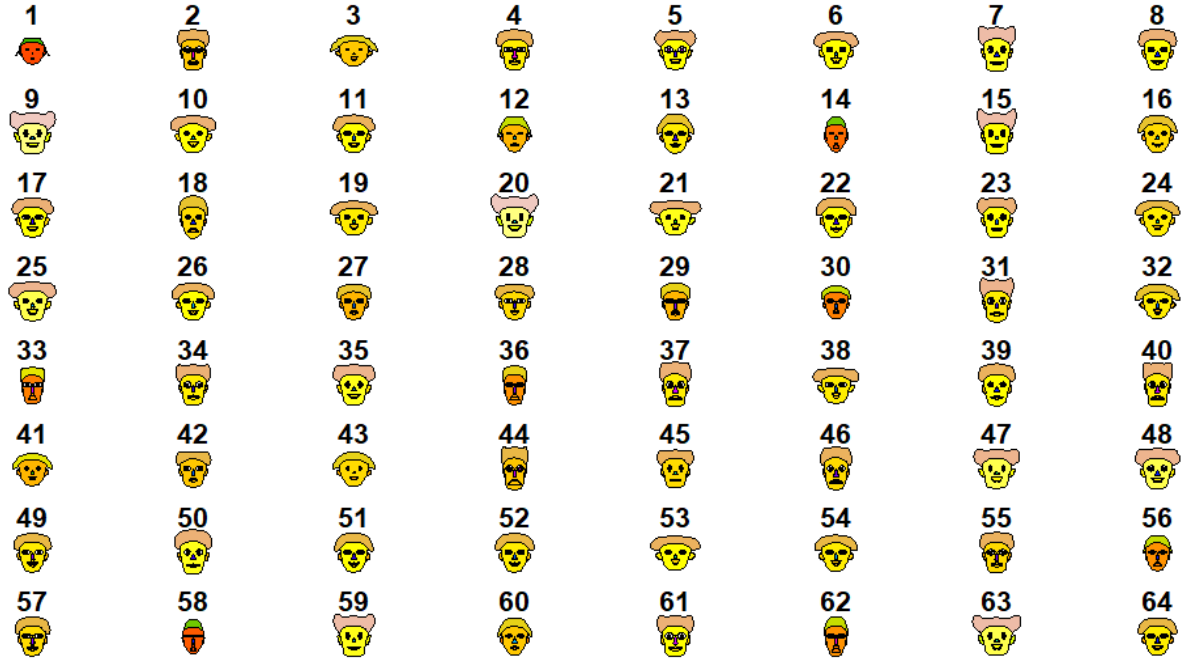
Θα εφαρμόσω πρώτα στα δεδομένα μου, τα οποία αποτελούνται από 490 εγγραφές και 7133 μεταβλητές, την PCA για να μειώσω την διάσταση των δεδομένων μου. Από το Σχήμα 70 μπορούμε να δούμε ότι η πρώτη μεταβλητή εξηγεί το 19,264 της μεταβλητότητας των δεδομένων και μέχρι και την τέταρτη μεταβλητή εξηγείται το 36,622 της μεταβλητότητας. Οπότε θα απεικονίσω τα δεδομένα μου στις 4 διαστάσεις, επειδή μετά την τέταρτη διάσταση κάθε επόμενη διάσταση εξηγεί πολύ μικρό μέρος της μεταβλητότητας των δεδομένων.

Eigenvalues

| | Dim.1 | Dim.2 | Dim.3 | Dim.4 | Dim.5 | Dim.6 | Dim.7 |
|---|---|---|---|---|---|---|---|
| Variance | 1374.328 | 566.573 | 421.768 | 249.975 | 175.249 | 145.409 | 123.250 |
| % of var. | 19.264 | 7.942 | 5.912 | 3.504 | 2.457 | 2.038 | 1.728 |
| | Dim.8 | Dim.9 | Dim.10 | Dim.11 | Dim.12 | Dim.13 | Dim.14 |
| Variance | 104.842 | 82.237 | 73.131 | 70.666 | 57.764 | 54.034 | 49.924 |
| % of var. | 1.470 | 1.153 | 1.025 | 0.991 | 0.810 | 0.757 | 0.700 |
| | Dim.15 | Dim.16 | Dim.17 | Dim.18 | Dim.19 | Dim.20 | Dim.21 |
| Variance | 49.192 | 46.189 | 43.020 | 41.420 | 38.565 | 37.980 | 37.004 |
| % of var. | 0.690 | 0.647 | 0.603 | 0.581 | 0.541 | 0.532 | 0.519 |
| | Dim.22 | Dim.23 | Dim.24 | Dim.25 | Dim.26 | Dim.27 | Dim.28 |
| Variance | 35.225 | 34.107 | 32.448 | 31.738 | 30.973 | 30.538 | 30.035 |
| % of var. | 0.494 | 0.478 | 0.455 | 0.445 | 0.434 | 0.428 | 0.421 |
| | Dim.29 | Dim.30 | Dim.31 | Dim.32 | Dim.33 | Dim.34 | Dim.35 |
| Variance | 28.830 | 28.256 | 27.467 | 26.957 | 26.760 | 26.423 | 25.939 |
| % of var. | 0.404 | 0.396 | 0.385 | 0.378 | 0.375 | 0.370 | 0.364 |
| | Dim.36 | Dim.37 | Dim.38 | Dim.39 | Dim.40 | Dim.41 | Dim.42 |
| Variance | 25.657 | 25.348 | 25.139 | 24.641 | 24.558 | 24.094 | 23.852 |
| % of var. | 0.360 | 0.355 | 0.352 | 0.345 | 0.344 | 0.338 | 0.334 |
| | Dim.43 | Dim.44 | Dim.45 | Dim.46 | Dim.47 | Dim.48 | Dim.49 |
| Variance | 23.636 | 23.325 | 22.889 | 22.582 | 22.461 | 22.140 | 22.110 |
| % of var. | 0.331 | 0.327 | 0.321 | 0.317 | 0.315 | 0.310 | 0.310 |
| | Dim.50 | Dim.51 | Dim.52 | Dim.53 | Dim.54 | Dim.55 | Dim.56 |
| Variance | 21.930 | 21.749 | 21.332 | 21.110 | 20.705 | 20.669 | 20.501 |
| % of var. | 0.307 | 0.305 | 0.299 | 0.296 | 0.290 | 0.290 | 0.287 |
| | Dim.57 | Dim.58 | Dim.59 | Dim.60 | Dim.61 | Dim.62 | Dim.63 |
| Variance | 20.431 | 20.083 | 19.922 | 19.825 | 19.610 | 19.485 | 19.235 |
| % of var. | 0.286 | 0.282 | 0.279 | 0.278 | 0.275 | 0.273 | 0.270 |
| | Dim.64 | Dim.65 | Dim.66 | Dim.67 | Dim.68 | Dim.69 | Dim.70 |
| Variance | 18.793 | 18.764 | 18.537 | 18.381 | 18.278 | 18.038 | 17.975 |
| % of var. | 0.263 | 0.263 | 0.260 | 0.258 | 0.256 | 0.253 | 0.252 |
| | Dim.71 | Dim.72 | Dim.73 | Dim.74 | Dim.75 | Dim.76 | Dim.77 |
| Variance | 17.827 | 17.636 | 17.310 | 17.179 | 17.005 | 16.951 | 16.782 |
| % of var. | 0.250 | 0.247 | 0.243 | 0.241 | 0.238 | 0.238 | 0.235 |

Σχήμα 72: Πίνακας της PCA που μας δίνει το ποσοστό της μεταβλητότητας που εξηγεί η κάθε μεταβλητή.

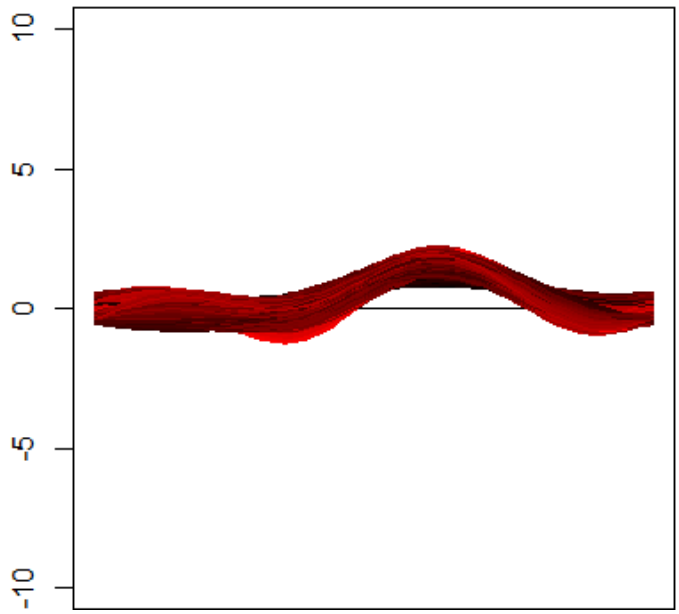
Στην συνέχεια θα χρησιμοποιήσουμε της συντεταγμένες για τις 4 διαστάσεις που μας δίνει η PCA και σε αυτές θα εφαρμόσουμε τις τεχνικές οπτικοποίησης που αναφέραμε παραπάνω. Θα ξεκινήσουμε εφαρμόζοντας την τεχνική των chernoff faces.




Σχήμα 73: Chernoff faces

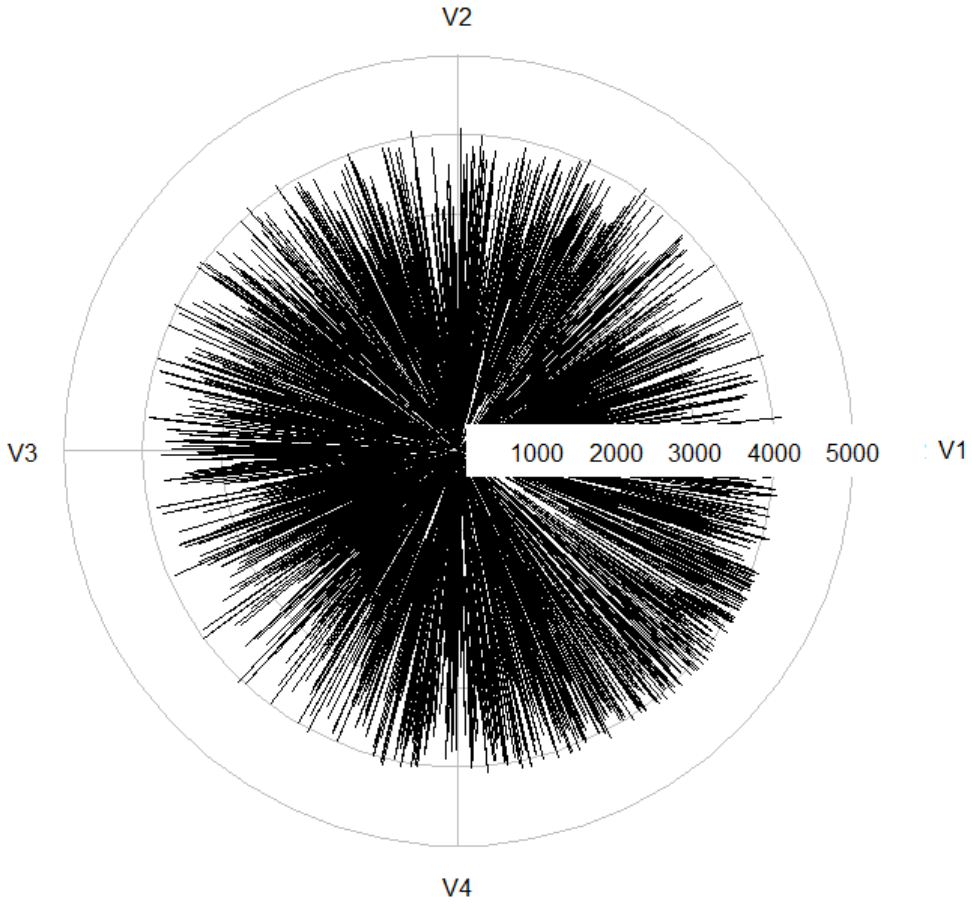
Στο Σχήμα 71 παρατηρούμε τα chernoff faces παρατηρούμε τις 63 πρώτες εγγραφές των δεδομένων να αναπαρίστανται η κάθε μια στις τέσσερις διαστάσεις, το μήκος του προσώπου και το μήκος των μαλλιών αναπαριστά την πρώτη διάσταση, το πλάτος του προσώπου και το πλάτος των μαλλιών την δεύτερη διάσταση, η κατασκευή του προσώπου και το είδος του κουρέματος την τρίτη διάσταση και τέλος η τέταρτη διάσταση αναπαρίσταται από το πλάτος της μύτης και το πλάτος των ματιών. Στο Σχήμα 71 μπορούμε να παρατηρήσουμε ότι κανένα από τα πρόσωπα δεν μοιάζει τόσο με το άλλο.

Συνεχίζουμε με τα andrews plots. Στο Σχήμα 72 μπορούμε να δούμε τις καμπύλες του andrews για τα δεδομένα μας, παρατηρούμε ότι δεν δίνουν καμιά ουσιαστική πληροφορία.



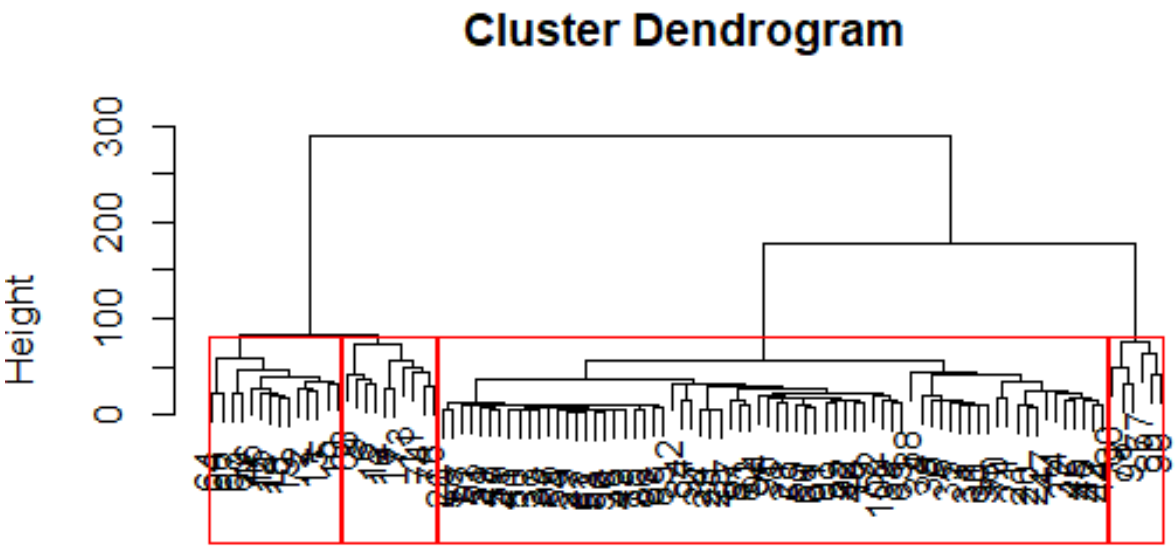

Σχήμα 74: andrews plot

Συνεχίζουμε με τα radial plots. Στο Σχήμα 73 μπορούμε να δούμε ένα radial plot για τα δεδομένα μας, παρατηρούμε ότι σε όλες τις διαστάσεις οι παρατηρήσεις μας έχουν σχεδόν ίδιες τιμές.




Σχήμα 75: radial plot

Τέλος, θα χρησιμοποιήσουμε ένα δενδρόγραμμα. Στο σχήμα 76 βλέπουμε ένα δενδρόγραμμα για τα δεδομένα μας. Όπως, παρατηρούμε τα δεδομένα μας χωρίζονται σε 4 ομάδες ανάλογα με το είδος του προϊόντος. Στην δεύτερη από δεξιά ομάδα η οποία είναι και η μεγαλύτερη ανήκουν οι μαρμελάδες που δεν είναι πηκτές και αλείφονται εύκολα, στην πρώτη από δεξιά ομάδα η οποία είναι και η μικρότερη ανήκουν τα γλυκά του κουταλιού, στην τρίτη ομάδα από δεξιά οι μαρμελάδες με κομμάτια φρούτων και στην τελευταία ανήκουν οι πηκτές μαρμελάδες.




Σχήμα 76: Δενδρόγραμμα

Μια γενική παρατήρηση που μπορούμε να κάνουμε είναι ότι η τεχνική του δενδρογράμματος μας δίνει καλύτερη πληροφόρηση για το πώς απεικονίζονται οι παρατηρήσεις μας.

### 5.3.4 Παρουσίαση Αποτελεσμάτων

Ένα γενικό αρχικό συμπέρασμα είναι ότι και στα τρία σύνολα δεδομένων οι καλύτερες απεικονίσεις δίνονται με την τεχνική των δενδρογραμμάτων, αλλά ακόμα και αυτή η τεχνική είναι απλουστευτική και αποτυγχάνει στο να μας δώσει μια ξεκάθαρη εικόνα των δεδομένων μας. Οι τεχνικές των chernoff faces και των andrews plots ενώ σε σύνολα μικρά σύνολα δεδομένων με λίγες διαστάσεις πετυχαίνουν να τα οπτικοποιήσουν αρκετά καλά, παρατηρούμε ότι για μεγάλα σύνολα δεδομένων δεν έχουμε ανάλογα αποτελέσματα. Από τα παραπάνω είναι επιτακτική η ανάγκη να αναζητήσουμε νέες πιο αποτελεσματικές τεχνικές οπτικοποίησης οι οποίες θα μας δίνουν μια πρώτη εικόνα των δεδομένων μας.

## Αναφορές

[1] Alpern, B. and Carter, L. (1991). The Hyperbox, *IEEE Visualization, Proceedings of the 2nd conference on Visualization '91,* 133-139, San Diego California.

[2] Andrews, D. F. (1972). *Plots of high dimensional data, Biometrics*, 28, 125-136.

[3] Asimov, D. (1985). The grand tour: a tool for viewing multidimensional data, *SIAM Journal on Scientific and Statistical Computing*, 6, 128-143.

[4] Cleveland, W. S. (1993). *Visualizing Data,* Hobart Press.

[5] Friendly, M. (1994). Mosaic Displays for Multi-Way Contingency Tables, *Journal of the American Statistical Association*, 89, 190-200.

[6] Huber, P. J. (1985). Projection Pursuit, *The Annals of Statistics*, 13, 435-475.

[7] Johnson, B. and Shneiderman, B. (1991). Treemaps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures, *In Proceedings of the IEEE Information Visualization '91,* 275–282.

[8] Kohonen, T. (1990). The Self-Organizing Map, *Procedding of the IEEE,* 78, 1464-1480.

[9] Hyper-dimensional data analysis using parallel coordinates, *Journal of the American Statistical Association,* 85, 664-675.

[10] Teoh, S. T. and Ma, K-L. (2002). RINGS: A technique for visualizing large hierarchies, In GD '02: *Revised Papers from the 10th International Symposium on Graph Drawing,* 268-275.

[11] Wegman, E. J. and Carr, D. B. (1993). *Statistical Graphics and Visualization,* Center for Computational Statistics, George Mason University.

[12] Wong, P. C. and Bergeron, R. D. (1994). 30 Years of Multidimensional Multivariate Visualization, *IEEE Computer Society, Washington, DC, USA,* 3-33.

[13] Y. Bengio, O. Delalleau, N. Le Roux, J.-F. Paiement P. Vincent, and M. Ouimet. Learning eigenfunctions links spectral embedding and Kernel PCA. *Neural Computation,* 16(10):2197– 2219, 2004.

[14] T. Cox and M. Cox. *Multidimensional scaling.* Chapman and Hall, London, UK, 1994.

[15] J.A. Lee and M. Verleysen. *Nonlinear dimensionality reduction.* Springer, New York, NY, USA, 2007.

[16] Μ. Κούτρας Σημειώσεις του μαθήματος Εφαρμοσμένη Πολυμεταβλητή Ανάλυση. Πανεπιστήμιο Πειραιώς, 2017-2018.

[17] L.J.P. van der Maaten and G.E. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research,* 9(No):2431–2456, 2008.

[18] L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. Dimensionality reduction: A comparative review, 2008.

[19] K.Q. Weinberger and L.K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research,* 10(Feb):207–244, 2009.

[20] C.K.I. Williams. On a connection between Kernel PCA and metric multidimensional scaling. *Machine Learning,* 46(1-3):11–19, 2002.

[21] H. Chernoff. The use of faces to represent points in k-dimensional space graphically *Journal of American Statistical Association*, 68:361–368, 1973

[22] William S. Cleveland. *Visualizing Data.* Hobart Press, Summit, New Jersey, 1993.

[23] John Keenan Taylor. *Statistical Techniques for Data Analysis.* Lewis Publishers, Boca Raton, Florida, 1990.

[24] John W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, 1977.

[25] Steven Feiner and Clifford Beshers. Visualizing n-dimensional virtual worlds with n-Vision. *Computer Graphics,* 24(2):37–38, March 1990.

[26] Steven Feiner and Clifford Beschers. Worlds within worlds: Metaphors for exploring n-dimensional virtual worlds. In *Proc. UIST '90, ACM Symposium on User Interface Software and Technology,* pages 76–83, Snowbird, UT, October 1990.

[27] A. Inselberg and B. Dimsdale, "Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry", *Proceedings of the 1st IEEE Conference on Visualization '90,* 31-375, 1990.

[28] M. C. F. de Oliveira and H. Levkowitz, "From Visual Data Exploration to Visual Data Mining: A Survey", *IEEE Transactions on Visualization and Computer Graphics,* vol.9, no.3, 378-394, 2003.

[29] C. Ware, *Information Visualization: Perception for Design,* Morgan Kaufmann Publishers, 2004.

[30] R. Spence, *Information Visualization,* Addison Wesley, ACM Press, 2000.

[31] Sammon Jr, J.W., 1969 *A nonlinear mapping for data structure analysis.* IEEE Trans. Comput. 18, 401–409.

[32] Kohonen, Teuvo [2000]. *Self-Organizing Maps.* 3rd Edition. Springer,90-93, Dec 2000.

[33] Mazza, Riccardo [2009]. *Introduction to Information Visualization.* Springer, 129-133, 1st Mar 2009.

[34] Munzner, Tamara [2014]. *Visualization Analysis and Design.* CRC Press, 135-184, 26th Nov 2014.

[35] H. Levkowitz, *"Color Icons: Merging Color and Texture Perception for Integrated Visualization of Multiple Parameters,"* presented at IEEE Visualization '91, 1991.

[36] J. H. Friedman, *"Exploratory Projection Pursuit," Journal of the American Statistical Association,* vol. 82, 249-266, 1987.

[37] C. L. Bentley and M. O. Ward, *"Animating Multidimensional Scaling to Visualize N-Dimensional Data Sets,"* presented at IEEE Information Visualization '96, San Francisco, CA, 1996.

[38] J. A. Wise, J. J. Thomas, et al, *"Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Docments,"* presented at IEEE Information Visualization'95, Atlanta, GA, 1995.

[39] C. Rohrdantz, D. Oelke, M. Krstajic, and F. Fischer, *"Realtime visualization of streaming text data: Tasks and challenges," in Workshop on Interactive Visual Text Analytics for Decision-Making at the IEEE VisWeek,* vol. 201, 2011.

[40] G. Kress and T. van Leeuwen. *Reading Images: The Grammar of Visual Design.* Routledge, 1996.

[41] Yi, Ji Soo, Rachel Melton, John Stasko and Julie Jacko [2005a]. Dust and Magnet: *Interactive Visualization for Everyday Data.* Information Interfaces Group, Georgia Tech. 2005.

[42] Keim D.A., Kriegel H.P.:Database Exploration using Multidimensional Visualization, *Computer Graphics and Applications* 40-49, Sept. 1994.

[43] HANSON A. J., CROSS R. A.: Interactive visualization methods for four dimensions. *In IEEE Conference on Visualization,* 196-203, 1993.

[44] E. Kandogan. Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. *In IEEE Information Visualization Symp.,* 2000.

[45] Y. Schler. Circle graphs: New visualization tools for text-mining *In PKDD*, 1999.

[46] J. Sharko, G. Grinstein, and K. A. Marx. Vectorized radviz and its application to multiple cluster datasets. *Visualization and Computer Graphics,* IEEE, 2008.

[47] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proc. IEEE Symposium on Information Visualization*, 2003.

[48] Carroll, J. M. , Kellogg, W. A. and Rosson, M. B. (1991) The task-artifact cycle. In J. M. Carroll (Ed. ) *Designing Interaction: Psychology at the Human-Computer Interface.* Cambridge: Cambridge University Press.

[49] Edwards, B. (1979). *Drawing on the Right Side of the Brain.* Los Angeles: J. P. Tarcher.

[50] Hyper-dimensional data analysis using parallel coordinates, *Journal of the American Statistical Association,* 85, 664-675.

[51] J.W. Sammon. A nonlinear mapping for data structure analysis. In *IEEE Trans. on Comp.,* volume C-18,401–409, May 1964.

[52] I. T. Jolliffe. *Principal Component Analysis.* Springer, 2 edition, 2002.

[53] Buja A., McDonald J. A., Michalak J., Stuetzle W.: *'Interactive Data Visualization Using Focusing and Linking'* Visualization '91, San Diego, CA, 1991, 156-163.

[54] Pickett R. M., Grinstein G. G.: *'Iconographic Displays for Visualizing Multidimensional Data',* IEEE Press, Piscataway, NJ, 1988, 514-519.

[55] Ward M. O.: *'XmdvTool M. G.: Integrating Multiple Methods for Visualizing Multivariate Data',* Proc. Visualization '94, Washington, DC, 326-336, 1994

[56] Keim D. A., Kriegel H.-P.: *'VisDB: A System for Visualizing Large Databases',* San Jose, CA, 164-345, 1995.

[57] Carr, D.B., Littlefield, R.J., Nicholson, W.L., Littlefield, J.S.: *Scatterplot Matrix Techniques for Large N.* Journal of the American Statistical Association 82(398), 424–436 (1987).

[58] Elmqvist, N., Dragicevic, P., Fekete, J.-D.: Rolling the dice: *Multidimensional visual exploration using scatterplot matrix navigation.* IEEE Trans. Vis. Comput. Graph. 14(6), 1539–1148 (2008).

[59] Inselberg, A. (1985). The plane with parallel coordinates, *The Visual Computer*, 5, 69-91.

[60] I. Friedel and A. Keller. Fast generation of randomized low-discrepancy point sets. In H. Niederreiter, K. Fang, and F. Hickernell, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2000,* 257–273. Springer, 2002.

[61] Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. (1983). *Graphical methods for data analysis,* The Wadsworth statistics/Probability series.

[62] D. Bordwell and K. Thompson. *Film Art: An Introduction* McGraw-Hill, 2003.

[63] C.Ware. *Information visualization: perception for design.* Morgan Kaufmann, San Francisco, CA, 2004.

[64] W. Wojtkowski and W. G. Wojtkowski. Storytelling: its role in information visualization. In *European Systems Science Congress,* 2002.

[65] Patric Hoffman. *TABLE VISUALIZATIONS: A FORMAL MODEL AND ITS APPLICATIONS.* NORTHEASTERN UNIVERSITY OF BOSTON, 1977.

[66] Georges Drinstein. *High-Dimensional Visualizations.* University of Massachussets Lowell, 1990.

[67] Christian Tominski. *Axes-Based Visualizations with Radial Layouts.* ACM Symposium on Applied Computing, 2004.

[68] Winnie Wing-Yi Chan. *A Survey on Multivariate Data Visualization.* Hong Kong University of Science and Techology, 2006.

[69] Laurens var der Maaten. *Visualizing Data using t-SNE.* Tilburg University, 2008.