ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

Γραφικές Μέθοδοι Παρουσίασης Πολυδιάστατων Δεδομένων

Κωνσταντίνος Ρ. Ράτσης

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

Μάιος 2009

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' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Καθηγητής Μ. Κούτρας (Επιβλέπων)
- Επίκουρη Καθηγήτρια Μ. Κατέρη
- Λέκτορας Γ. Τζαβελάς

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

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

Graphical Methods for Multidimensional Data Visualization

Konstantinos R. Ratsis

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfillment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece

May 2009

ΠΕΡΙΛΗΨΗ

Στην εποχή μας έχει δημιουργηθεί η ανάγκη για αξιοποίηση και μελέτη ολοένα και μεγαλύτερου όγκου δεδομένων με πολύπλοκη δομή και πολλαπλά χαρακτηριστικά. Με την εξέλιξη των σύγχρονων ηλεκτρονικών υπολογιστών και τις μεγάλες δυνατότητες που παρέχουν αυτοί τόσο στην επεξεργασία όσο και στα γραφικά, κατέστη εφικτό να αναπτυχθούν αποτελεσματικά εργαλεία για τη μελέτη τεράστιων όγκων δεδομένων με στόχο την κατανόηση της υφής τους και τον εντοπισμό των σχέσεων που πιθανό να υπάρχουν μεταξύ των ατόμων του πληθυσμού στους οποίους αναφέρονται ή μεταξύ των διαφόρων υπό μελέτη χαρακτηριστικών.

Ένα από τα σημαντικότερα ζητήματα που προκύπτουν κατά την εξέταση συνόλων δεδομένων με πολλές μεταβλητές είναι οι δυνατότητες οπτικοποίησής τους. Στη διεθνή βιβλιογραφία έχουν προταθεί αρκετές μέθοδοι και τεχνικές που εξυπηρετούν τη γραφική αναπαράσταση πολυδιάστατων δεδομένων, ιδιαίτερα κατά το χρονικό διάστημα 1970-1990.

Στην παρούσα διπλωματική εργασία επιχειρείται μια συστηματική παρουσίαση των σημαντικότερων μεθόδων οπτικοποίησης πολυδιάστατων δεδομένων καθώς και ο εντοπισμός των κυριοτέρων πλεονεκτημάτων και μειονεκτημάτων αυτών. Τέλος αναφέρονται τα διαθέσιμα υπολογιστικά μέσα (λογισμικά) με τα οποία υλοποιούνται οι τεχνικές αυτές αλλά και χαρακτηριστικά παραδείγματα ώστε να γίνει πιο κατανοητή η εφαρμογή τους.

ABSTRACT

In our era, the need for studying and making beneficial use of increasingly larger amounts of data with complex structure and multiple characteristics (variables) under study is very frequently encountered. The rapid development of the contemporary computers and the flexible features they are offering nowadays for data processing and graphing, made it possible powerful software to be set up for the statistical analysis of enormous data sets; as a result understanding the data structure and revealing the possible interrelations between the population members and the characteristics under analysis, can be easily achieved.

One of the major needs that arise when considering data sets with many variables, is the use of appropriate visualization procedures. In the international literature, several methods and techniques have been proposed to gain (usually two dimensional) graphic representation of multidimensional data, particularly during the period 1970-1990.

The present MSc thesis offers a systematic presentation of multidimensional data visualization methods and reports the main advantages and disadvantages of each of them. Finally the available software for implementing these techniques is referenced and some illustrative examples are worked out so that the reader gets a better understanding of the way the method is applied in a specific dataset.

Στη γυναίκα μου
στο γιο μου
& στους γονείς μου

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 Προγράμματος Μεταπτυχιακών Σπουδών «Εφαρμοσμένη Στατιστική» του Πανεπιστημίου Πειραιώς για τη γνώση και την εμπειρία που αποκόμισα συμμετέχοντας στο πρόγραμμα.

ΠΕΡΙΕΧΟΜΕΝΑ

Κεφάλαιο 1: Εισαγωγή

1.1 Οπτικοποίηση δεδομένων 1
1.2 Ιστορική αναδρομή 3
1.3 Ταξινόμηση μεθόδων οπτικοποίησης δεδομένων 5

Κεφάλαιο 2: Τυπικές Τεχνικές

2.1 Εισαγωγή 9
2.2 Πολλαπλά σημειογραφήματα 9
2.3 Πίνακες μεταθέσεων 12
2.4 Διαγράμματα survey 13
2.5 Πίνακες διαγραμμάτων διασποράς 14
2.6 Διαγράμματα trellis 15
2.7 Διαγράμματα profile 17

Κεφάλαιο 3: Εικονογραφικές Τεχνικές

3.1 Εισαγωγή 19
3.2 Πρόσωπα Chernoff 19
3.3 Stick figure icons 21
3.4 Κωδικοεικόνες 22
3.5 Χρωμοεικόνες 24
3.6 Γλυφογραφήματα 26
3.7 Διαγράμματα αστέρων 27
3.8 Δένδρα των Kleiner & Hartigan 29

Κεφάλαιο 4: Τεχνικές Εικονοστοιχείων

4.1 Εισαγωγή 31
4.2 Διαγράμματα κυκλικών τομέων 31
4.3 Τεχνικές σπείρας / τεχνικές αξόνων 33
4.4 Τεχνικές αναδρομικών σχηματισμών 36
4.5 Θηκογράμματα εικονοστοιχείων 38
4.6 Attribute blocks 40

Κεφάλαιο 5: Γεωμετρικές Τεχνικές

5.1 Εισαγωγή 43
5.2 Διαγράμματα παράλληλων συντεταγμένων 43
5.3 Καμπύλες του Andrews 46
5.4 Οπτικοποίηση πολικών συντεταγμένων 49
5.5 Διαγράμματα hammock 53
5.6 Διαγράμματα μωσαϊκού 55
5.7 Πτερυγογράμματα 58
5.8 Υπερθηκόγραμμα 60
5.9 Υφαντόγραμμα 61

Κεφάλαιο 6: Ιεραρχικές Τεχνικές

6.1 Εισαγωγή 65
6.2 Δενδροχάρτες 65
6.3 Στιβογράμματα 72
6.4 Διαγράμματα Venn 74
6.5 Δενδρογράμματα 75
6.6 Σμηνογράμματα 76
6.7 Θερμοχάρτες 78
6.8 RINGS 80

Κεφάλαιο 7: Μέθοδοι Μείωσης Διαστάσεων

7.1 Εισαγωγή 83
7.2 Projection Pursuit / Grand Tour method 83
7.3 Ανάλυση κύριων συνιστωσών 85
7.4 Biplots 88
7.5 Πολυδιάστατη κλιμάκωση 90
7.6 Χάρτες του Sammon 92
7.7 Χάρτες αυτοοργάνωσης 93

Βιβλιογραφία 95

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Οπτικοποίηση Δεδομένων

Τις τελευταίες δεκαετίες έχει αυξηθεί ραγδαία η ποσότητα των δεδομένων που παράγονται, αποθηκεύονται και στη συνέχεια επεξεργάζονται. Στο γεγονός αυτό έχει συμβάλλει η έκρηξη της τεχνολογικής εξέλιξης των ηλεκτρονικών υπολογιστών, η οποία έχει φέρει στη διάθεση των επιστημόνων – ερευνητών τεράστιες βάσεις δεδομένων, ογκώδη και συνεχή ροή πληροφοριών αλλά και πανίσχυρα υπολογιστικά εργαλεία. Οι δυνατότητες που προσφέρονταν πλέον καθιστούσαν αναγκαία την παράλληλη ανάπτυξη και εφαρμογή μεθόδων συλλογής, επεξεργασίας και αξιοποίησης των διαθέσιμων δεδομένων. Και σε αυτόν τον τομέα οι ηλεκτρονικοί υπολογιστές διαδραμάτισαν καθοριστικό ρόλο καθώς επέτρεψαν τη δημιουργία νέων τεχνικών και καινοτομιών. Επιπλέον κάποιες παλαιότερες ιδέες οι οποίες ήταν αδύνατον να εφαρμοσθούν σε πρακτικό επίπεδο, μπορούσαν τώρα να υλοποιηθούν και να συμβάλλουν στη μελέτη των δεδομένων. Σε όλη αυτή την επιστημονική και τεχνολογική εξέλιξη, πρωτεύοντα ρόλο είχε η Στατιστική ως η βασική Επιστήμη συλλογής, καταγραφής, επεξεργασίας και αξιοποίησης των δεδομένων. Ο όγκος και η πολυπλοκότητα των δεδομένων απαιτούσαν νέες τεχνικές επεξεργασίας και παρουσίασής τους, ώστε να μπορέσουν οι επιστήμονες να τα αξιοποιήσουν κατάλληλα. Σήμερα στα διαθέσιμα δεδομένα εμφανίζονται πολλές αλληλεπιδράσεις, ενω παράλληλα αυτά αναφέρονται σε μεγάλο πλήθος μεταβλητών και μπορούν να παρατηρηθούν σε πολλά χρονικά σημεία, μετατρέποντας έτσι τη μελέτη τους σε πραγματική πρόκληση. Μια πρόκληση που βρήκε και συνεχίζει να βρίσκει ανταπόκριση από τους ερευνητές διαφόρων επιστημονικών κλάδων, οι οποίοι αγωνίζονται να ανακαλύψουν καινούριες μεθόδους, ακόμα και να βελτιώσουν τις ήδη υπάρχουσες, με σκοπό την όσο το δυνατόν καλύτερη αξιοποίηση των διαθέσιμων στοιχείων. Ούτως ή άλλως κανείς δε μπορεί να ισχυριστεί οτι η τεχνολογική πρόοδος δε θα συνεχιστεί τα επόμενα χρόνια, οπότε και οι ανάγκες για πολυπλοκότερη, πολυπληθέστερη και ταχύτερη επεξεργασία της πληροφορίας θα αυξάνονται.

Η φύση των σύνθετων δεδομένων που καλούμαστε να διαχειριστούμε στη σύγχρονη εποχή, έφερε στο προσκήνιο την ανάγκη για τρόπους απεικόνισής τους ώστε να γίνονται αντιληπτές οι σχέσεις και οι δομές που κρύβονται μέσα σε αυτά. Πρόκειται για την Οπτικοποίηση Δεδομένων (Data Visualization) που αποτελεί τη διαδικασία μετατροπής των αριθμητικών δεδομένων σε εικόνα με σκοπό την ερμηνεία των πληροφοριών και των συνδέσμων οι οποίοι περιέχονται σε αυτά. Η μέθοδος αυτή έρχεται να συμπληρώσει τη στατιστική μελέτη των δεδομένων, εκμεταλλευόμενη το γεγονός οτι το ανθρώπινο μάτι αντιλαμβάνεται πολύ πιο εύκολα και γρήγορα την ύπαρξη μιας πληροφορίας όταν τα δεδομένα δίνονται ως σημεία ή σχήματα στο χώρο παρά ως αριθμητικές τιμές σε έναν πίνακα. Παραδοσιακά η Οπτικοποίηση διαχωρίζεται σε δύο κύριους τομείς: την Επιστημονική Οπτικοποίηση (Scientific Visualization) και την Οπτικοποίηση Πληροφοριών (Information Visualization), ένας διαχωρισμός ο οποίος γίνεται με βάση τη φύση των μοντέλων από τα οποία προέρχονται τα δεδομένα. Σύμφωνα με τους Wong and Bergeron (1997) καθως και τους Tory and Moller (2004), στη μεν Επιστημονική Οπτικοποίηση τα δεδομένα προέρχονται από συνεχή υποδείγματα ενώ στην Οπτικοποίηση Πληροφοριών τα δεδομένα προέρχονται από διακριτά υποδείγματα. Σε αντιστοιχία με τον παραπάνω διαχωρισμό γίνεται και η διάκριση μεταξύ πολυδιάστατων (multidimensional) και πολυμεταβλητών (multivariate) δεδομένων. Πιο συγκεκριμένα, ως σύνολο πολυδιάστατων δεδομένων θεωρείται αυτό το οποίο έχει αρκετές, σαφώς καθορισμένες ανεξάρτητες μεταβλητές και μια ή περισσότερες εξαρτημένες μεταβλητές που σχετίζονται με αυτές. Τέτοια σύνολα δεδομένων αφορούν συνήθως συνεχή υποδείγματα. Από την άλλη ένα πολυμεταβλητό σύνολο δεδομένων έχει πολλές εξαρτημένες μεταβλητές οι οποίες σχετίζονται μεταξύ τους σε μικρό ή μεγαλύτερο βαθμό. Τα πολυμεταβλητά σύνολα δεδομένων αφορούν συνήθως διακριτά υποδείγματα.

Η οπτικοποίηση μπορεί να εφαρμοσθεί εύκολα σε παρατηρήσεις που εξαρτώνται από δύο ή και τρείς μεταβλητές αφου τότε μπορούμε απλά να αντιστοιχίσουμε κάθε μεταβλητή σε μια διάσταση του χώρου, δισδιάστατου και τρισδιάστατου αντίστοιχα. Σε περισσότερες των τριών μεταβλητών όμως, το πρόβλημα γίνεται πολύ πιο σύνθετο. Στόχος μας είναι να κωδικοποιηθούν οι πολλές διαστάσεις με τέτοιο τρόπο ώστε να μπορούν να αναπαρασταθούν στη συνέχεια στο διδιάστατο επίπεδο ή στον τρισδιάστατο χώρο. Φυσικά στις περιπτώσεις αυτές δε

μπορεί να γίνει αντιστοίχηση των μεταβλητών στις διαστάσεις του χώρου αλλά προηγείται μια τροποποίηση των δεδομένων ώστε αυτά να αποδόσουν στη συνέχεια όσο πιο παραστατικά γίνεται την πληροφορία που περιέχουν.

1.2 Ιστορική Αναδρομή

Ιστορικά η οπτικοποίηση πολυδιάστατων δεδομένων είχε αρχίσει να απασχολεί τους επιστήμονες, και όχι μόνο της Στατιστικής αλλά και άλλων κλάδων, εδώ και αρκετά χρόνια, χωρίς όμως οι προτεινόμενες μέθοδοι να μπορούν να εφαρμοσθούν στην πράξη. Μόνο με τη διάδοση των προσωπικών ηλεκτρονικών υπολογιστών τη δεκαετία του 1980 δημιουργήθηκε το κατάλληλο έδαφος για την ανάπτυξη και εφαρμογή τεχνικών οπτικοποίησης και γραφικής αναπαράστασης των δεδομένων. Σύμφωνα με τους Wong and Bergeron (1997), η εξέλιξη της οπτικοποίησης πολυδιάστατων ή πολυμεταβλητών δεδομένων μπορεί να διαχωριστεί σε 4 περιόδους, με βάση κάποια επιστημονικά γεγονότα - ορόσημα στην ιστορία της επιστήμης. Αρχικά έχουμε την περίοδο μέχρι το 1976 όταν κυρίως οι Στατιστικοί χρησιμοποιούν κάποια γραφήματα δύο διαστάσεων, απλά στην κατασκευή τους, με τη χρήση χρωμάτων και ειδικά διαμορφωμένων φύλλων. Τα γραφήματα αυτά αφορούν μικρά σχετικά σύνολα δεδομένων και έχουν ως σκοπό να αναδείξουν τα κύρια χαρακτηριστικά τους και να προτείνουν την κατάλληλη στατιστική μέθοδο επεξεργασίας τους. Η περίοδος από το 1977 μέχρι το 1985 χαρακτηρίζεται ως η περίοδος «αφύπνισης» για τον τομέα αυτό και επηρεάστηκε τα μέγιστα από το έργο του Tukey για την Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis) το 1977. Πρόκειται για έναν νέο τρόπο σκέψης, έναν τρόπο που θα βοηθήσει τον επιστημονικό (και όχι μόνο) κόσμο να αποκωδικοποιήσει τις πληροφορίες που εμπεριέχουν τα δεδομένα. Με την παράλληλη άφιξη των ηλεκτρονικών υπολογιστών, οι πολύπλοκοι και χρονοβόροι υπολογισμοί γίνονται τώρα άμεσα και άκοπα ενώ και η δυνατότητα βελτιωμένων γραφικών εξυπηρετεί τη διαδικασία της οπτικοποίησης. Τα δεδομένα που κυρίως απασχολούν τους μελετητές και τους ερευνητές είναι τα διδιάστατα και τρισδιάστατα, χωρίς να λείπουν οι αναφορές και σε πολυδιάστατα δεδομένα. Θα ακολουθήσει χρονικά η περίοδος 1986 με 1991 η οποία χαρακτηρίζεται ως περίοδος ανακάλυψης. Το 1986 το Εθνικό Ίδρυμα Επιστημών των Η.Π.Α. (National Science Foundation – N.S.F.) χρηματοδοτεί μια επιστημονική συνάντηση

με σκοπό να αναδειχθούν τρόποι γραφικής αναπαράστασης και σχετικά προγράμματα για υπολογιστές που χρησιμοποιούνται στην επιστημονική έρευνα. Εκεί διατυπώνεται για πρώτη φορά η εφαρμογή γραφικών τεχνικών στην επιστήμη των υπολογιστών και εισάγεται ο όρος ViSC (Visualization in Scientific Computing). Την επόμενη χρονιά οι επιστήμονες δηλώνουν την αναγκαιότητα της χρήσης της Οπτικοποίησης στις επιστήμες, στηριζόμενη ασφαλώς στην τεχνολογία των ηλεκτρονικών υπολογιστών αλλά και τη χρηματοδότηση των σχετικών ερευνών προς αυτήν την κατεύθυνση. Το γεγονός αυτό δίνει την απαιτούμενη ώθηση στους επιστήμονες να ανακαλύψουν πολλές νέες μεθόδους οπτικοποίησης πολυδιάστατων δεδομένων με τη βοήθεια εμπλουτισμένων γραφικών και αμφίδρομων τεχνικών αναπαράστασης. Ακόμα και η Εικονική Πραγματικότητα κάνει την εμφάνιση της την περίοδο αυτή στη σχετική βιβλιογραφία. Τέλος η περίοδος από το 1992 μέχρι και σήμερα χαρακτηρίζεται από την επεξεργασία και την αποτίμηση της δουλειάς που έχει προηγηθεί. Στο διάστημα αυτό έχουν προκύψει λίγες καινούριες τεχνικές και η όποια πρόοδος έχει υπάρξει αφορά σε βελτίωση παλαιοτέρων μεθόδων και προσπάθειες συνδυασμού διαφόρων τεχνικών αναπαράστασης με σκοπό καλύτερα αποτελέσματα. Επίσης έχει η εμφανιστεί η ανάγκη να αξιολογηθούν όλες αυτές οι τεχνικές οπτικοποίησης πολυδιάστατων δεδομένων ως προς την αξιοπιστία, την αποτελεσματικότητα και την πιστότητά τους. Κάποιες από τις καινοτομίες της σημερινής εποχής που θα μπορούσαν να προσθέσουν νέες τεχνικές είναι οι τριδιάστατες ταχείες προβολές (animation) αλλά και η χρήση ηχητικών σημάτων στη διαδικασία της οπτικοποίησης των δεδομένων.

Όπως και να έχει πάντως η διαχρονική εξέλιξη της οπτικοποίησης πολυδιάστατων δεδομένων, το σίγουρο είναι ότι η αναζήτηση νέων μεθόδων όπως και η επανεξέταση των παλαιοτέρων συνεχίζεται και θα συνεχιστεί όσο οι ανάγκες για τη μελέτη των δεδομένων αυτών αυξάνονται και όσο η τεχνολογική πρόοδος επιτρέπει να αξιοποιηθούν όλο και πιο σύγχρονα γραφικά εργαλεία. Μια σημαντική πτυχή λοιπόν της οπτικοποίησης δεδομένων παραμένει η αξιολόγηση και αναβάθμιση των υφιστάμενων μεθόδων γραφικής αναπαράστασης, μια διαδικασία από την οποία μπορεί να προκύψουν βελτιώσεις, νέες ιδέες και προτάσεις καθώς και χρήσιμα συμπεράσματα. Στην παρούσα εργασία θα επιχειρηθεί μια παρουσίαση των σημαντικότερων τεχνικών αναπαράστασης πολυδιάστατων δεδομένων σε διδιάστατο χώρο (επίπεδο). Στην επόμενη παράγραφο θα παρουσιάσουμε μια ταξινόμηση των μεθόδων αυτών, ενώ στα κεφάλαια που ακολουθούν θα δούμε τα κυριότερα

χαρακτηριστικά τους, τα πλεονεκτήματα και τα μειονεκτήματά τους καθώς και εφαρμογές τους σε σύνολα δεδομένων με χρήση εμπορικών προγραμμάτων για ηλεκτρονικούς υπολογιστές.

1.3 Ταξινόμηση Μεθόδων Οπτικοποίησης Δεδομένων

Έχουν προταθεί κατά καιρούς διάφοροι τρόποι ταξινόμησης των μεθόδων οπτικοποίησης πολυδιάστατων δεδομένων, με επικρατέστερη όλων αυτή που πρότειναν οι Keim and Kriegel το (1996). Σύμφωνα με την ταξινόμηση αυτή υπάρχουν οι παρακάτω 5 κατηγορίες τεχνικών:

- Γεωμετρικές Τεχνικές (Geometric Techniques),
- Τεχνικές Εικονογραφημάτων ή Εικονογραφικές Τεχνικές (Icon-Based Techniques),
- Ιεραρχικές Τεχνικές (Hierarchical Techniques),
- Τεχνικές Εικονοστοιχείων (Pixel-Oriented Techniques),
- Τεχνικές Γραφημάτων (Graph-Based Techniques).

Η κατηγοριοποίηση αυτή στηρίζεται στον τρόπο με τον οποίο παρουσιάζονται τα δεδομένα. Πέραν αυτών όμως έχουν προταθεί και οι παρακάτω κατηγορίες τεχνικών:

- Τυπικές Τεχνικές (Standard Techniques),
- Τεχνικές Παραμόρφωσης (Distortion Techniques),
- Τεχνικές Προβολής (Projection Techniques),
- Υβριδικές Τεχνικές (Hybrid Techniques).

Στη συνέχεια, θα βασιστούμε στις παραπάνω κατηγοριοποιήσεις για να παρουσιάσουμε τις κυριότερες τεχνικές αναπαράστασης πολυδιάστατων δεδομένων, με τα πλεονεκτήματα και τα μειονεκτήματά τους. Πρέπει να τονισθεί πάντως ότι η επιλογή κατάλληλης μεθόδου έχει να κάνει με τη φύση των δεδομένων, τα ερωτήματα που καλούμαστε να απαντήσουμε κατά τη μελέτη τους αλλά και με τα διαθέσιμα εργαλεία που έχουμε στα χέρια μας κάθε φορά.

Στις Τυπικές Τεχνικές θα συναντήσουμε μερικές από τις πιο διαδεδομένες και αποτελεσματικές τεχνικές όπως τα:

- Πολλαπλά σημειογραφήματα (multiple line graphs)
- Πίνακες μεταθέσεων (permutation matrix)
- Διαγράμματα survey (survey plot)
- Πίνακας διαγραμμάτων διασποράς (scatter plot matrix)
- Διαγράμματα trellis (trellis plot)
- Διαγράμματα profile (profile plot)

Οι Εικονογραφικές Τεχνικές όπου η κάθε παρατήρηση απεικονίζεται σε ένα εικονογράφημα (glyph), του οποίου κάθε χαρακτηριστικό αντιστοιχεί σε μια μεταβλητή, περιλαμβάνουν τις εξής τεχνικές:

- Πρόσωπα Chernoff (Chernoff faces)
- Stick figure icons
- Κωδικοεικόνες (shape coding)
- Χρωμοεικόνες (color icons)
- Γλυφογραφήματα (metroglyphs)
- Διαγράμματα αστέρων (star plot)
- Δένδρα των Kleiner & Hartigan (Kleiner & Hartigan trees)

Στις Τεχνικές Εικονοστοιχείων η τιμή κάθε παρατήρησης αντιστοιχεί σε ένα εικονοστοιχείο, του οποίου η διάταξη σε ένα συγκεκριμένο πλαίσιο αλλά και ο χρωματισμός του πολλές φορές, μας δίνει τις απαραίτητες πληροφορίες. Χαρακτηριστικές τεχνικές αυτής της κατηγορίας είναι οι εξής:

- Διαγράμματα κυκλικών τομέων (circle segments)
- Τεχνικές σπείρας / τεχνικές αξόνων (spiral / axes techniques)
- Τεχνικές αναδρομικών σχηματισμών (recursive pattern technique)
- Θηκογράμματα εικονοστοιχείων (pixel bar charts)
- Attribute blocks

Στις Γεωμετρικές Τεχνικές οι παρατηρήσεις τοποθετούνται σε γεωμετρικούς σχηματισμούς με τέτοιο τρόπο ώστε να αναδεικνύονται σχέσεις και αλληλεπιδράσεις

μεταξύ των μεταβλητών. Στα επόμενα κεφάλαια θα παρουσιάσουμε τις ακόλουθες τεχνικές αυτής της κατηγορίας:

- Διαγράμματα παράλληλων συντεταγμένων (parallel coordinate plot)
- Καμπύλες του Andrews (Andrews curves)
- Οπτικοποίηση πολικών συντεταγμένων (radial coordinate visualization-RadViz)
- Διαγράμματα hammock (Hammock plot)
- Διαγράμματα μωσαϊκού (mosaic plot)
- Πτερυγογράμματα (pinion plot)
- Υπερθηκόγραμμα (hyperbox)
- Υφαντόγραμμα (textile plot)

Στις Ιεραρχικές Τεχνικές γίνεται εφαρμογή της οπτικοποίησης σε ιεραρχικά δεδομένα. Χαρακτηριστικές ιεραρχικές τεχνικές είναι οι ακόλουθες:

- Δενδροχάρτες (treemaps)
- Στιβογράμματα (dimensional stacking)
- Διαγράμματα Venn (Venn diagrams)
- Δενδρογράμματα (dendrograms)
- Σμηνογράμματα (clustergrams)
- Θερμοχάρτες (heatmaps)
- RINGS

Στις Μεθόδους Μείωσης Διαστάσεων (Dimensionality Reduction Techniques), σκοπός μας είναι η μείωση των μεταβλητών στις πιο σημαντικές, μέσω μιας κατάλληλης διαδικασίας. Αντιπροσωπευτικές τεχνικές της κατηγορίας αυτής είναι οι επόμενες:

- Projection Pursuit / Grand Tour Method
- Ανάλυση κύριων συνιστωσών (principal components analysis)
- Biplots
- Πολυδιάστατη κλιμάκωση (multidimensional scaling)
- Χάρτες του Sammon (Sammon’s mapping)
- Χάρτες αυτοοργάνωσης (self-organizing maps)

ΚΕΦΑΛΑΙΟ 2

Κλασσικές Τεχνικές

2.1 Εισαγωγή

Σε αυτήν την κατηγορία τεχνικών για την οπτικοποίηση πολυμεταβλητών δεδομένων συναντάμε μερικές από τις πρωτοποριακές και βασικές μεθόδους, ικανές να παρέχουν στον ερευνητή μια πρώτη επισκόπηση των δεδομένων. Τα γραφήματα που θα δούμε στη συνέχεια εφαρμόζονται συνήθως στα πρώτα στάδια ανάλυσης των δεδομένων και παρέχουν καλή πληροφόρηση για κάποια πρώτα συμπεράσματα, ιδιαίτερα σε περιπτώσεις που τα δεδομένα μας είναι μεσαίου μεγέθους και περιέχουν μικρό σχετικά αριθμό μεταβλητών. Λόγω της ευκολίας κατασκευής τους αλλά και της απλότητας στην ερμηνεία τους είναι πολύ διαδεδομένα και περιέχονται σε αρκετά στατιστικά πακέτα και προγράμματα ανάλυσης δεδομένων. Από την άλλη μεριά δε συνιστώνται για λεπτομερή ανάλυση των δεδομένων καθώς οι δυνατότητές τους είναι περιορισμένες και ο ερευνητής θα πρέπει να ανατρέξει σε άλλες τεχνικές για την αναζήτηση πιο σύνθετων δομών και σχέσεων.

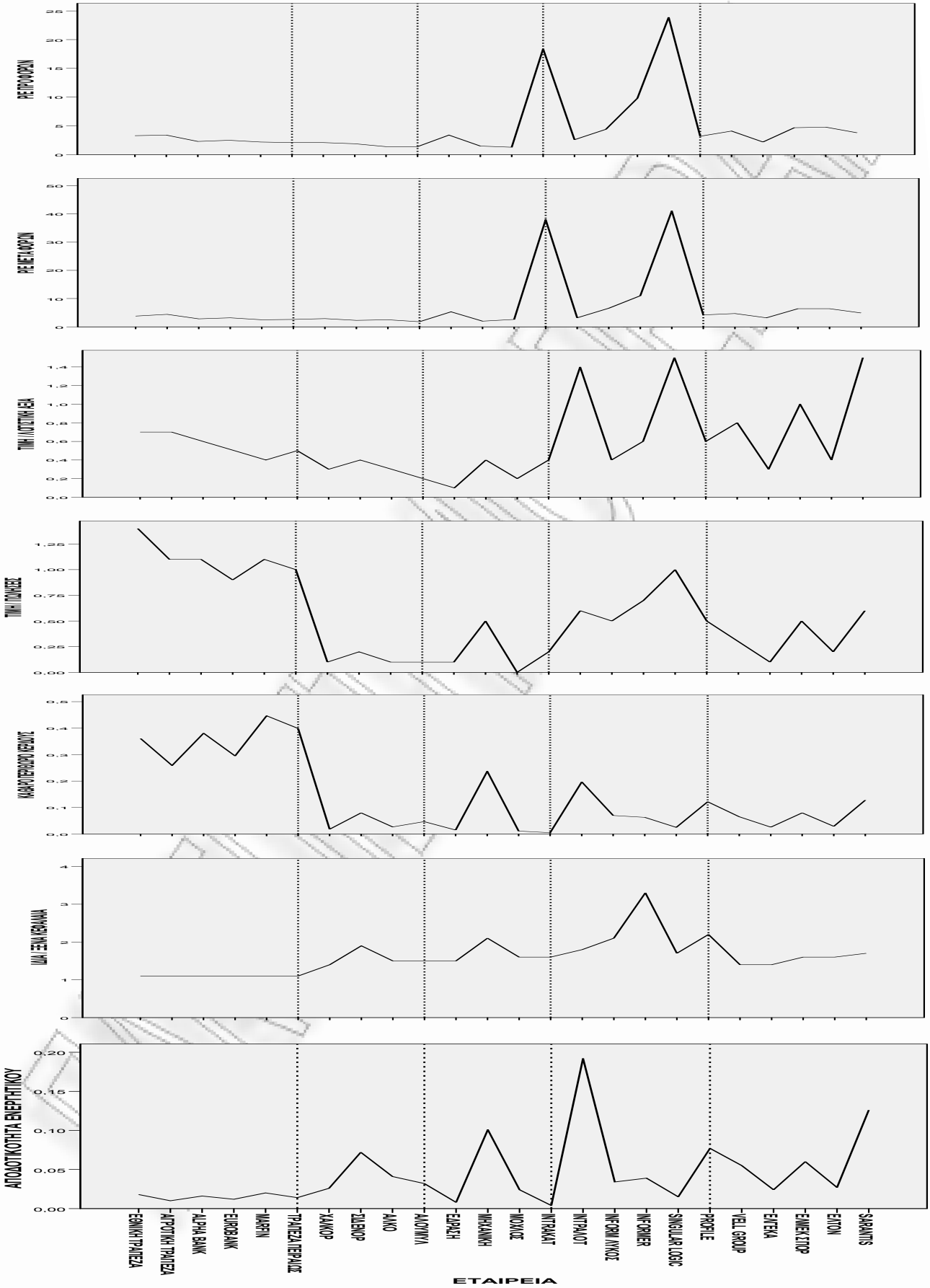
2.2 Πολλαπλά σημειογραφήματα

Είναι η γενίκευση της περίπτωσης του μονού γραφήματος που αναπαριστά τη σχέση μεταξύ 2 μεταβλητών. Στην πολυμεταβλητή ανάλυση χρησιμοποιούνται πολλά γραφήματα για τις επιπλέον μεταβλητές, τοποθετημένα το ένα πάνω από το άλλο. Δημιουργείται ένας οριζόντιος άξονας στον οποίο τοποθετούμε την ανεξάρτητη μεταβλητή ή τη μεταβλητή που μας ενδιαφέρει και ένας κάθετος άξονας στον οποίο τοποθετούμε διαδοχικά τα γραφήματα για τις υπόλοιπες μεταβλητές. Το γεγονός ότι η ανεξάρτητη μεταβλητή ή γενικότερα η μεταβλητή που τοποθετούμε στον οριζόντιο

άξονα είναι μοναδική, δίνει μια ξεχωριστή σημασία στη μεταβλητή αυτή. Ουσιαστικά αυτή η μεταβλητή αντιπροσωπεύει και τη διάταξη των δεδομένων στο γράφημά μας, μια διάταξη που συνδέεται συχνά με κάποια άλλη μεταβλητή των δεδομένων όπως π.χ. το χρόνο. Γενικά αποφεύγουμε να απεικονίσουμε μαζί πολλές μεταβλητές γιατί από ένα πλήθος και μετά δε θα είναι ευδιάκριτα τα γραφήματα και θα δυσκολεύεται πολύ η ανάγνωση και χρήση του διαγράμματος.

Τα δεδομένα που θα χρησιμοποιήσουμε στη συνέχεια αφορούν 24 εταιρείες εισηγμένες στο Χρηματιστήριο Αξιών Αθηνών. Πρόκειται για 6 εταιρείες από τον τραπεζικό κλάδο, 4 από τον κλάδο των μετάλλων, 4 από τον κατασκευαστικό κλάδο, 5 από τον κλάδο της πληροφορικής και 5 από τον κλάδο του χονδρικού εμπορίου. Οι μεταβλητές μας είναι 7 χρηματοοικονομικοί αριθμοδείκτες καθώς και μια κατηγορική μεταβλητή, ο κλάδος στον οποίο ανήκει κάθε εταιρεία. Τα στοιχεία έχουν αντληθεί από την ιστοσελίδα http://www.capital.gr/ και αφορούν τα αποτελέσματα εννεαμήνου 2008. Στην επόμενη σελίδα δίνονται διάγραμμα για κάθε μια από τις 7 μεταβλητές

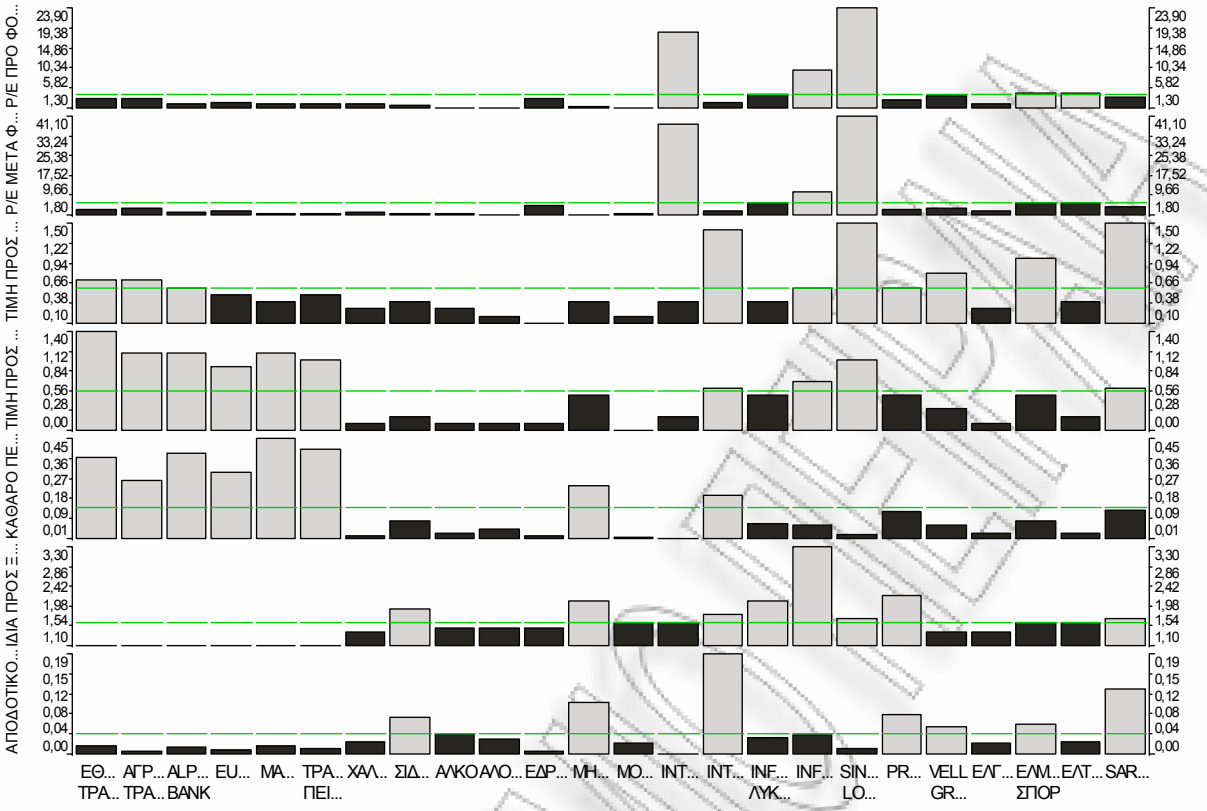
Α/Α	ΕΤΑΙΡΕΙΑ	Ρ/Ε ΠΡΟ ΦΟΡΩΝ	Ρ/Ε ΜΕΤΑ ΦΟΡΩΝ	ΤΙΜΗ ΠΡΟΣ ΛΟΓΙΣΤΙΚΗ ΑΞΙΑ	ΤΙΜΗ ΠΡΟΣ ΠΩΛΗΣΕΙΣ	ΚΑΘΑΡΟ ΠΕΡΙΘΩΡΙΟ ΚΕΡΔΟΥΣ	ΙΔΙΑ ΠΡΟΣ ΞΕΝΑ ΚΕΦΑΛΑΙΑ	ΑΠΟΔΟΤΙΚΟΤΗΤΑ ΕΝΕΡΓΗΤΙΚΟΥ
1	ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ	3,3	3,8	0,7	1,4	0,361	1,1	0,018
2	ΑΓΡΟΤΙΚΗ ΤΡΑΠΕΖΑ	3,4	4,4	0,7	1,1	0,259	1,1	0,010
3	ALPHA BANK	2,3	2,8	0,6	1,1	0,381	1,1	0,016
4	EUROBANK	2,5	3,2	0,5	0,9	0,295	1,1	0,012
5	MARFIN	2,2	2,5	0,4	1,1	0,447	1,1	0,020
6	ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ	2,1	2,6	0,5	1	0,399	1,1	0,014
7	ΧΑΛΚΟΡ	2,1	2,9	0,3	0,1	0,018	1,4	0,026
8	ΣΙΔΕΝΟΡ	1,9	2,3	0,4	0,2	0,080	1,9	0,072
9	ΑΛΚΟ	1,4	2,5	0,3	0,1	0,027	1,5	0,041
10	ΑΛΟΥΜΥΛ	1,4	1,8	0,2	0,1	0,047	1,5	0,032
11	ΕΔΡΑΣΗ	3,4	5,3	0,1	0,1	0,015	1,5	0,008
12	ΜΗΧΑΝΙΚΗ	1,5	2	0,4	0,5	0,237	2,1	0,101
13	ΜΟΧΛΟΣ	1,3	2,6	0,2	0	0,011	1,6	0,024
14	INTRAKAT	18,4	38	0,4	0,2	0,005	1,6	0,004
15	ΙΝΤΡΑΛΟΤ	2,6	3,2	1,4	0,6	0,197	1,8	0,192
16	INFORM ΛΥΚΟΣ	4,4	6,6	0,4	0,5	0,070	2,1	0,034
17	INFORMER	9,8	11	0,6	0,7	0,063	3,3	0,039
18	SINGULAR LOGIC	23,9	41,1	1,5	1	0,025	1,7	0,015
19	PROFILE	3,2	4,2	0,6	0,5	0,122	2,2	0,077
20	VELL GROUP	4,1	4,7	0,8	0,3	0,065	1,4	0,055
21	ΕΛΓΕΚΑ	2,2	3,2	0,3	0,1	0,026	1,4	0,024
22	ΕΛΜΕΚ ΣΠΟΡ	4,7	6,4	1	0,5	0,080	1,6	0,060
23	ΕΛΤΟΝ	4,8	6,4	0,4	0,2	0,029	1,6	0,027
24	SARANTIS	3,8	4,9	1,5	0,6	0,128	1,7	0,126


ΑΠΟΔΟΤΙΚΟΤΗΤΑ ΕΝΕΡΓΗΤΙΚΟΥ
ΜΕΣΗ ΑΞΙΑ ΕΝΕΡΓΗΤΙΚΟΥ
ΜΕΣΗ ΑΞΙΑ ΙΔΙΩΝ ΚΕΦΑΛΑΙΩΝ
ΜΕΣΗ ΑΞΙΑ ΜΕΤΟΧΗΣ
ΜΕΣΗ ΑΞΙΑ ΠΑΘΗΤΙΚΟΥ
ΜΕΣΗ ΑΞΙΑ ΠΡΟΒΛΕΨΕΩΝ
ΜΕΣΗ ΑΞΙΑ ΠΩΛΗΣΕΩΝ
ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ
ΑΓΡΟΤΙΚΗ ΤΡΑΠΕΖΑ
ALPHA BANK
EUROBANK
MARFIN
ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ
ΧΑΛΚΟΡ
ΣΙΔΕΝΟΡ
ΑΛΚΟ
ΑΛΟΥΜΥΛ
ΕΛΒΑΛΧ
ΜΗΧΑΝΙΚΗ
ΜΟΤΟΡΟΣ
ΝΤΡΑΚΑΤ
ΝΤΡΑΛΟΤ
INFORM ΛΥΚΟΣ
INFORMER
SINGULAR LOGIC
PROFILE
ΥΕΓΕΙΑ
ΕΝΤΕΚΑ
ΕΛΜΕΚΣΤΟΡ
ΕΛΤΟΝ
SARANTIS
ΕΤΑΙΡΕΙΑ

(αριθμοδείκτες) έχοντας τοποθετήσει στον οριζόντιο άξονα τις εταιρείες με τη σειρά που εμφανίζονται στον πίνακα με τα δεδομένα. Τα διαγράμματα έγιναν με τη βοήθεια του στατιστικού πακέτου SPSS. Οι κάθετες διακεκομμένες γραμμές που διατρέχουν τα γραφήματα τοποθετήθηκαν για να διαχωρίσουν τις εταιρείες κάθε κλάδου μεταξύ τους.

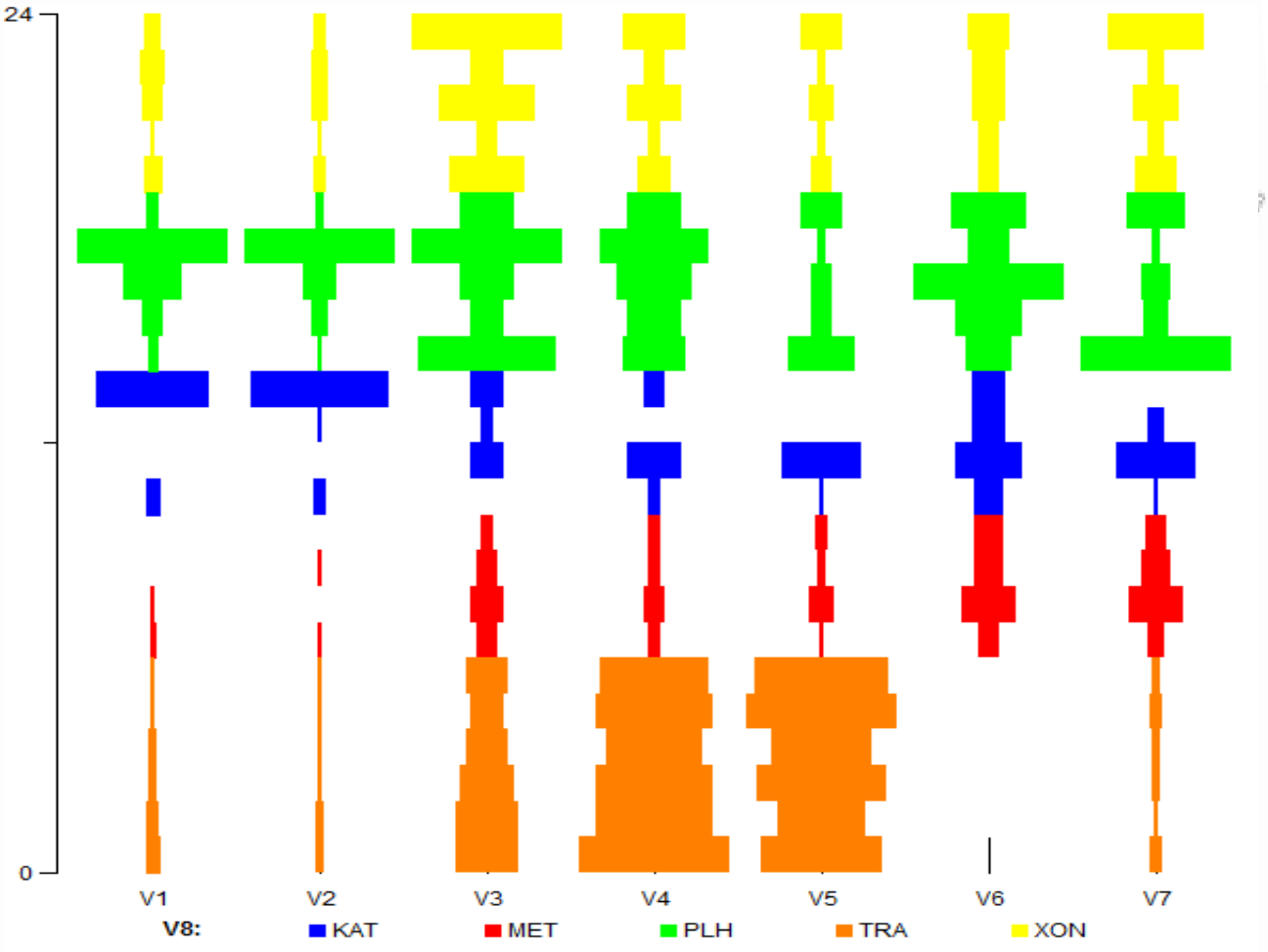
2.3 Πίνακες μεταθέσεων

Εισήχθηκε ως μέθοδος από τον Jacque Bertin (1967) με σκοπό να αναδείξει τις πιθανές σχέσεις μεταξύ πολυμεταβλητών δεδομένων μικρού ή μεσαίου μεγέθους. Πρόκειται ουσιαστικά για πολλαπλά ραβδογράμματα, από τα οποία το καθένα αντιστοιχεί στην τιμή κάθε παρατήρησης σε κάθε μεταβλητή. Το ύψος του ραβδογράμματος υποδηλώνει την τιμή της κάθε παρατήρησης. Ο οριζόντιος άξονας του γραφήματος έχει την ίδια πληροφορία για όλες τις μεταβλητές π.χ. χρόνος ή συνηθέστερα αποτελεί μια διάταξη των δεδομένων με βάση κάποια από τις μεταβλητές. Οι τιμές των δεδομένων οι οποίες βρίσκονται πάνω από το μέσο όρο είναι λευκές ενώ αντίθετα, εκείνες που βρίσκονται κάτω από το μέσο όρο είναι μαύρες. Τοποθετείται επίσης στο διάγραμμα μια πράσινη διακεκομμένη γραμμή που αντιστοιχεί στο μέσο όρο των δεδομένων για κάθε μεταβλητή. Κατά την εφαρμογή της μεθόδου αυτής είναι δυνατή η αναδιάταξη των παρατηρήσεων ως προς οποιαδήποτε μεταβλητή ώστε να αναδειχθούν ενδιαφέροντα στοιχεία των δεδομένων. Στο γράφημα που δίνεται στην επόμενη σελίδα χρησιμοποιήσαμε τα δεδομένα των 24 εταιρειών και κατασκευάστηκε με τη χρήση του προγράμματος Visulab (http://www.inf.ethz.ch/personal/hinterbe/Visulab/).



2.4 Διαγράμματα survey

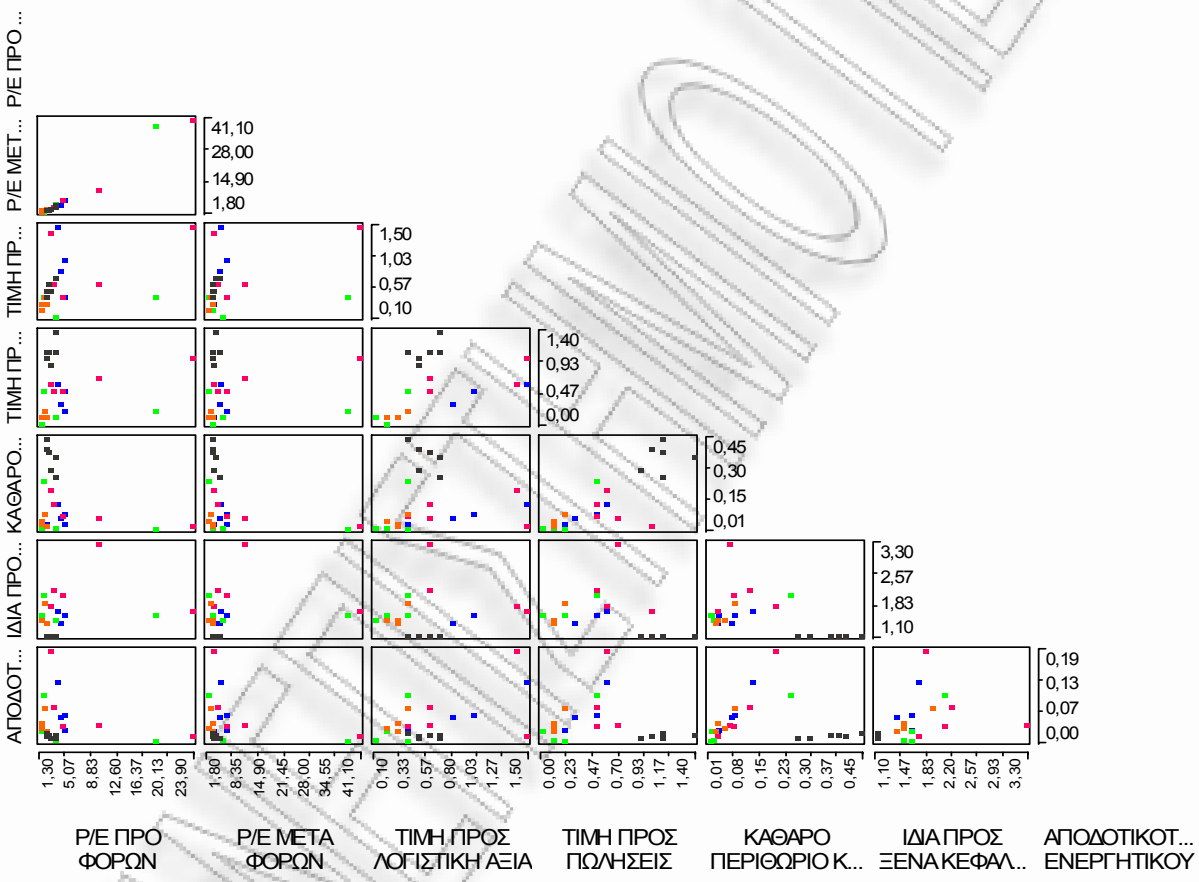
Η μέθοδος αυτή εισήχθηκε με τη συγκεκριμένη ονομασία από τον Lohninger (1994) στο πρόγραμμα επεξεργασίας δεδομένων Inspect και είναι παραλλαγή του πίνακα μεταθέσεων. Αν περιστρέψουμε κατά 90 μοίρες έναν πίνακα μεταθέσεων, "κοντύνουμε" κατά το ήμισυ τις ράβδους και τις προεκτείνουμε προς την αντίθετη κατεύθυνση του άξονα κατά το ίδιο μήκος θα έχουμε ένα διάγραμμα survey. Αυτή η μέθοδος οπτικοποίησης μας επιτρέπει να δούμε συσχετίσεις μεταξύ δυο μεταβλητών, ιδιαίτερα όταν τα δεδομένα μας είναι διατεταγμένα. Σε αρκετές περιπτώσεις χρησιμοποιούνται και χρώματα ώστε να διευκολυνθεί ο εντοπισμός των καλύτερων από τις μεταβλητές σε ταξινομημένα δεδομένα. Έχουμε χρησιμοποιήσει και πάλι τα χρηματοοικονομικά δεδομένα των 24 εταιρειών για να κατασκευάσουμε ένα διάγραμμα survey με τη βοήθεια του Orange (http://www.ailab.si/orange/). Οι διάφοροι χρωματισμοί αντιστοιχούν στους κλάδους που ανήκουν οι εταιρείες.



2.5 Πίνακας διαγραμμάτων διασποράς

Ο πίνακας διαγραμμάτων διασποράς είναι μια εύχρηστη και αρκετά διαδεδομένη μέθοδος γραφικής αναπαράστασης δεδομένων με πολλές μεταβλητές. Στηρίζεται στο απλό διάγραμμα διασποράς με το οποίο απεικονίζουμε γραφικά τη σχέση μεταξύ δυο μεταβλητών. Κατασκευάζοντας έναν πίνακα με τόσα διαγράμματα διασποράς ώστε να απεικονίζονται όλες οι ανά ζεύγη σχέσεις μεταξύ των μεταβλητών που ερευνούμε, παίρνουμε την τελική μορφή του πίνακα διαγραμμάτων διασποράς. Όπως είναι λογικό, η διαγώνιος του διαγράμματος δε δίνει κάποια πληροφορία καθώς αποτελεί το διάγραμμα διασποράς της κάθε μεταβλητής με την ίδια τη μεταβλητή, για αυτό και αρκετές φορές παραλείπεται, είτε αντικαθίσταται από κάποιες άλλες πληροφορίες. Ακόμα, εφόσον το διάγραμμα διασποράς κάποιας μεταβλητής x ως προς κάποια μεταβλητή y είναι ισοδύναμο με το διάγραμμα διασποράς της y έναντι της x,

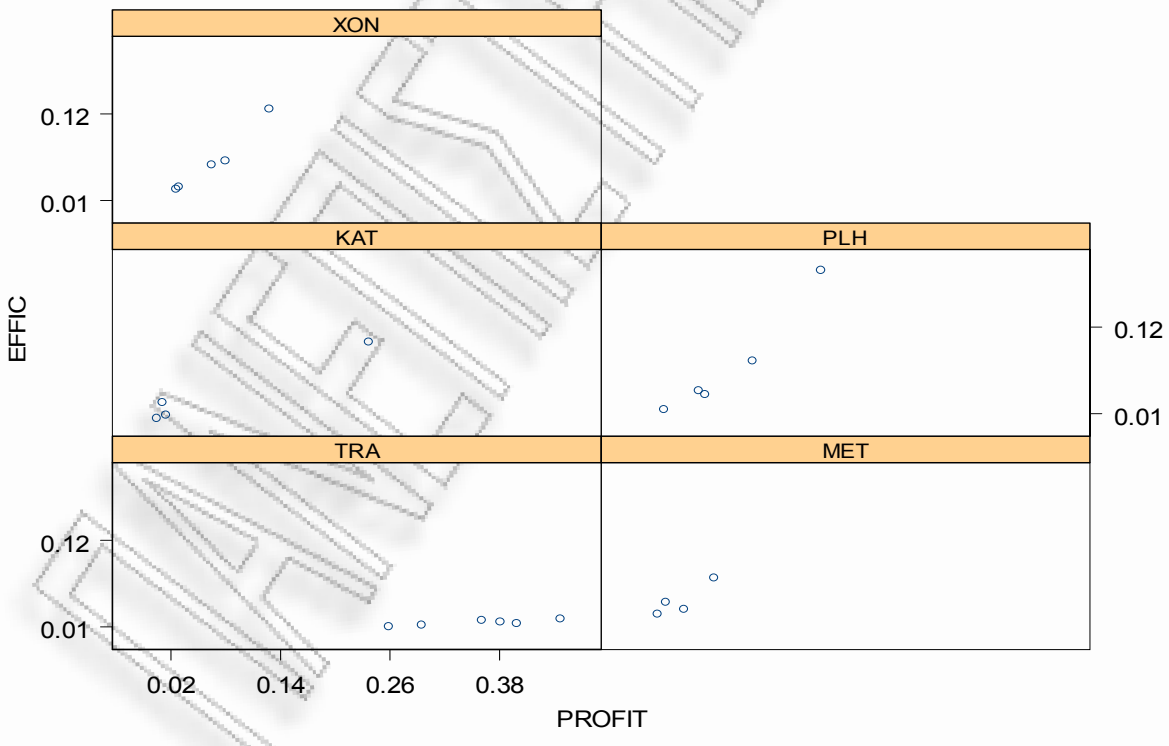
προτιμάται πολλές φορές να περιορίζεται πίνακας διαγραμμάτων διασποράς στα στοιχεία που βρίσκονται κάτω από τη διαγώνιο, ώστε να εξοικονομείται χώρος αλλά και να απλοποιείται η ανάγνωση του γραφήματος. Ο πίνακας διαγραμμάτων διασποράς είναι αποτελεσματικό για σχετικά μικρό αριθμό μεταβλητών αφού σε περίπτωση πολλών μεταβλητών γίνεται αρκετά πολύπλοκο και δυσανάγνωστο. Χρησιμοποιήθηκε και πάλι το λογισμικό Visulab για να κατασκευάσουμε το παρακάτω πίνακας διαγραμμάτων διασποράς για τα δεδομένα των 24 εταιρειών, ενώ ο κάθε χρωματισμός στις παρατηρήσεις αντιστοιχεί στον κλάδο που ανήκει η κάθε εταιρεία.



2.6 Διαγράμματα trellis

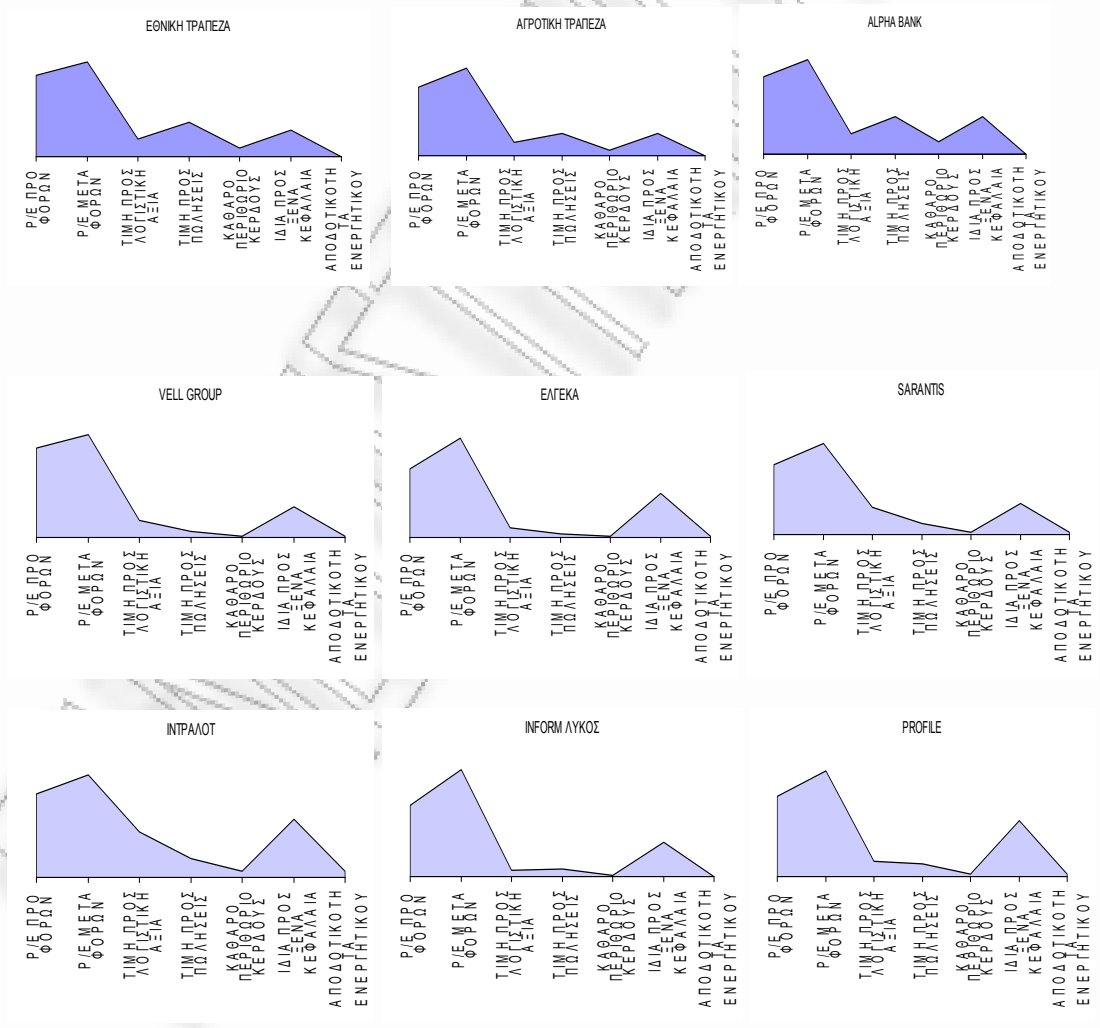
Τα διαγράμματα trellis βασίζονται στα Co-Plots που εισήχθησαν από τον Chambers (1992). Στη συνέχεια ο Cleveland (1993) γενίκευσε τα Co-Plots στα διαγράμματα trellis. Είναι μέθοδος για τη γραφική παρουσίαση πολυδιάστατων

δεδομένων μέσα από ένα πλέγμα γραφημάτων, από όπου προέκυψε και το όνομα διάγραμμα trellis (trellis ονομάζεται ένα ξύλινο πλέγμα που χρησιμοποιείται για τη στήριξη αναρριχόμενων φυτών). Μας επιτρέπει τις συγκρίσεις μεταξύ των τιμών διαφόρων μεταβλητών, υπό κάποια ή κάποιες κατηγορικές μεταβλητές. Οι κατηγορίες βέβαια των μεταβλητών αυτών δε θα πρέπει να είναι πολλές για να μη μεγαλώνει υπερβολικά το διάγραμμα trellis και δυσκολεύει η ανάγνωσή του. Μπορούν να χρησιμοποιηθούν διάφορα είδη γραφημάτων (π.χ. ιστογράμματα, διαγράμματα διασποράς κ.α.) για να απεικονιστούν οι σχέσεις μεταξύ των μεταβλητών. Ένα σημαντικό στοιχείο στο διάγραμμα trellis είναι ότι η κλίμακα μέτρησης των επιμέρους γραφημάτων θα πρέπει να είναι κοινή ώστε να διευκολύνονται οι συγκρίσεις μεταξύ τους. Στο παράδειγμα που έχουμε κατασκευάσει με το στατιστικό πακέτο S-PLUS (http://www.insightful.com/), βλέπουμε τις τιμές των μεταβλητών «Καθαρό Περιθώριο Κέρδους» και «Αποδοτικότητα Ενεργητικού» για τις 24 εταιρείες, ενώ τα δεδομένα μας έχουν κατηγοριοποιηθεί κατά κλάδο δραστηριότητας των εταιρειών.



2.7 Διαγράμματα profile

Τα διαγράμματα profile είναι από τους πιο απλούς τρόπους για να παρουσιαστούν γραφικά πολυδιάστατα δεδομένα. Σε κάθε πολυδιάστατη παρατήρηση αντιστοιχίζουμε ένα γράφημα το οποίο μπορεί να είναι είτε απλό ιστόγραμμα είτε ένα area plot. Προφανώς η μορφή των διαγραμμάτων profile εξαρτάται από τη διάταξη των μεταβλητών στο γράφημα και είναι ευνόητο ότι η διάταξη αυτή θα πρέπει να παραμένει η ίδια για όλες τις παρατηρήσεις ώστε να επιτρέπονται οι συγκρίσεις μεταξύ τους. Με τη βοήθεια του Excel (http://office.microsoft.com/en-us/excel/FX100487621033.aspx) κατασκευάσαμε τα διαγράμματα profile 3 εταιρειών από τον κλάδο των Τραπεζών, 3 από τον κλάδο του Χονδρικού Εμπορίου και 3 από τον κλάδο της Πληροφορικής.



ΚΕΦΑΛΑΙΟ 3

Εικονογραφικές Τεχνικές

3.1 Εισαγωγή

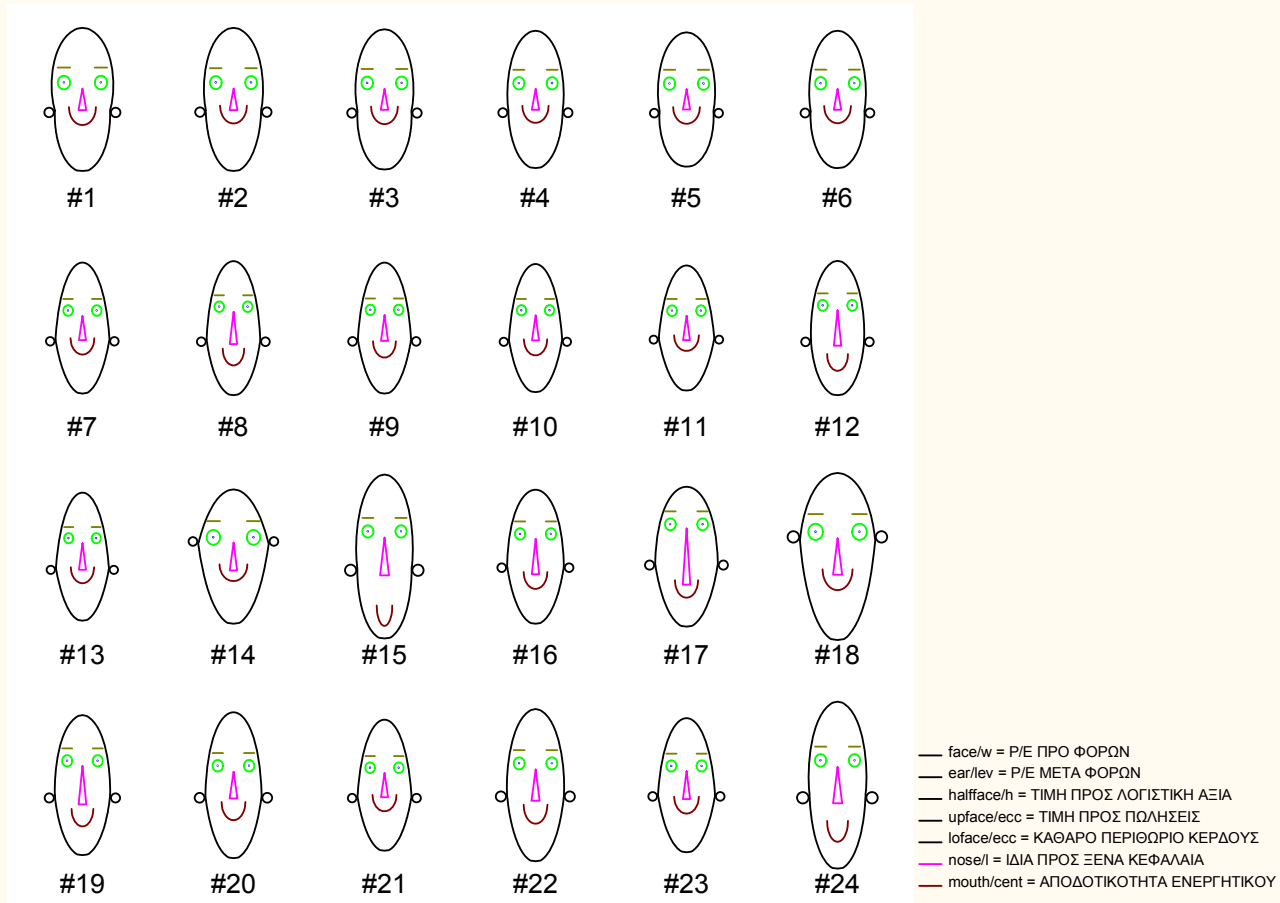
Οι τεχνικές της κατηγορίας αυτής στηρίζονται στο γεγονός ότι κάθε πολυμεταβλητή παρατήρηση αντιστοιχίζεται με ένα εικονογράφημα. Στη συνέχεια τα γεωμετρικά χαρακτηριστικά (σχήμα, μέγεθος κ.α.), οι χρωματισμοί ή η θέση των εικονογραφημάτων σε μια επιφάνεια προβολής χρησιμοποιούνται για να οπτικοποιηθούν οι τιμές των μεταβλητών των δεδομένων. Με τον τρόπο αυτό ο ερευνητής μπορεί να εντοπίσει ομοιότητες ή διαφορές μεταξύ των δεδομένων παρατηρώντας τα αντίστοιχα εικονογραφήματα στο σύνολό τους. Οι τεχνικές αυτές είναι αποτελεσματικές για μεσαίου μεγέθους σύνολα δεδομένων με περιορισμένο αριθμό μεταβλητών.

3.2 Πρόσωπα του Chernoff

Ο Herman Chernoff (1973) εισήγαγε τα διάσημα πρόσωπα του Chernoff με τα οποία μπορούμε να παρουσιάσουμε πολυδιάστατα δεδομένα μέσω εικονογραφημάτων που μοιάζουν με ανθρώπινα πρόσωπα. Η λογική πίσω από την κατασκευή τους είναι ότι σε κάθε χαρακτηριστικό του προσώπου (μέγεθος ματιών, μέγεθος μύτης, ύψος κεφαλιού κ.α.) αντιστοιχίζουμε μια μεταβλητή. Έτσι προκύπτουν τόσα «πρόσωπα» όσες είναι και οι παρατηρήσεις μας, με τόσα χαρακτηριστικά όσες είναι και οι μεταβλητές των δεδομένων. Προφανώς στα πρόσωπα του Chernoff δεν παρουσιάζονται οι πραγματικές τιμές των μεταβλητών καθώς προηγείται μια κατάλληλη μετατροπή των τιμών αυτών ώστε να μπορούν να αντιστοιχηθούν στο κάθε χαρακτηριστικό του «προσώπου». Είναι όμως εξαιρετικά χρήσιμα για να εντοπίσουμε συσχετίσεις μεταξύ των μεταβλητών αλλά και την ύπαρξη ακραίων παρατηρήσεων. Η χρησιμότητά τους έγκειται στο γεγονός πως το

ανθρώπινο μάτι είναι ιδιαίτερα ευαίσθητο στο να εντοπίζει έστω και ελάχιστες διαφορές στα χαρακτηριστικά του προσώπου. Βέβαια τα χαρακτηριστικά ενός προσώπου είναι περιορισμένα και κατά συνέπεια δε μπορούν να απεικονιστούν πολλές μεταβλητές. Στην αντιμετώπιση του προβλήματος αυτού ήρθαν να συμβάλλουν οι Flury & Riedwyl (1981) σημειώνοντας ότι αν χρησιμοποιούμε συμμετρικά πρόσωπα του Chernoff, «χάνουμε» τη δυνατότητα να διπλασιάσουμε τις μεταβλητές που μπορούμε να αναπαραστήσουμε και πρότειναν τη χρήση ασύμμετρων προσώπων. Εκμεταλλευόμενοι τη συμμετρικότητα του προσώπου ως προς ένα νοητό κάθετο άξονα μπορούμε ουσιαστικά να διπλασιάσουμε τις μεταβλητές που οπτικοποιούνται μέσω του γραφήματος. Ένα ακόμα μειονέκτημα της μεθόδου αυτής είναι ότι η τελική μορφή κάθε «προσώπου» εξαρτάται από την αντιστοίχηση των χαρακτηριστικών του στις μεταβλητές η οποία τις περισσότερες φορές γίνεται με υποκειμενικά κριτήρια. Έχει αποδειχθεί επίσης ότι κάποια χαρακτηριστικά του προσώπου απομνημονεύονται πιο εύκολα σε σχέση με κάποια άλλα με συνέπεια ο παρατηρητής των προσώπων του Chernoff να έχει την τάση να τα ομαδοποιεί με συγκεκριμένα κριτήρια και όχι με την αντικειμενικότητα που θα περιμέναμε. Για το λόγο αυτό θα πρέπει να προηγείται μια αξιολόγηση των μεταβλητών και να χρησιμοποιούνται τα χαρακτηριστικά που προκαλούν μεγαλύτερη εντύπωση (όπως τα μάτια ή το σχήμα του προσώπου) για τις σημαντικότερες μεταβλητές. Να σημειωθεί ότι στις μέρες μας γίνεται εφαρμογή και τριδιάστατων προσώπων του Chernoff τα οποία βελτιώνουν την αποτελεσματικότητα της μεθόδου μέσα από την πειστικότερη αναπαράσταση των «προσώπων».

Στη συνέχεια θα δούμε μια εφαρμογή των προσώπων του Chernoff στα χρηματοοικονομικά δεδομένα των 24 εταιρειών, με τη βοήθεια του στατιστικού πακέτου Statistica (http://www.statsoft.com/). Πρόσωπα του Chernoff μπορούν να κατασκευαστούν και με το Mathematica (http://www.wolfram.com/) ή την R (http://www.r-project.org/). Στο κάτω δεξιό μέρος δίνεται η αντιστοίχιση των μεταβλητών στα χαρακτηριστικά του προσώπου που έχουν χρησιμοποιηθεί.

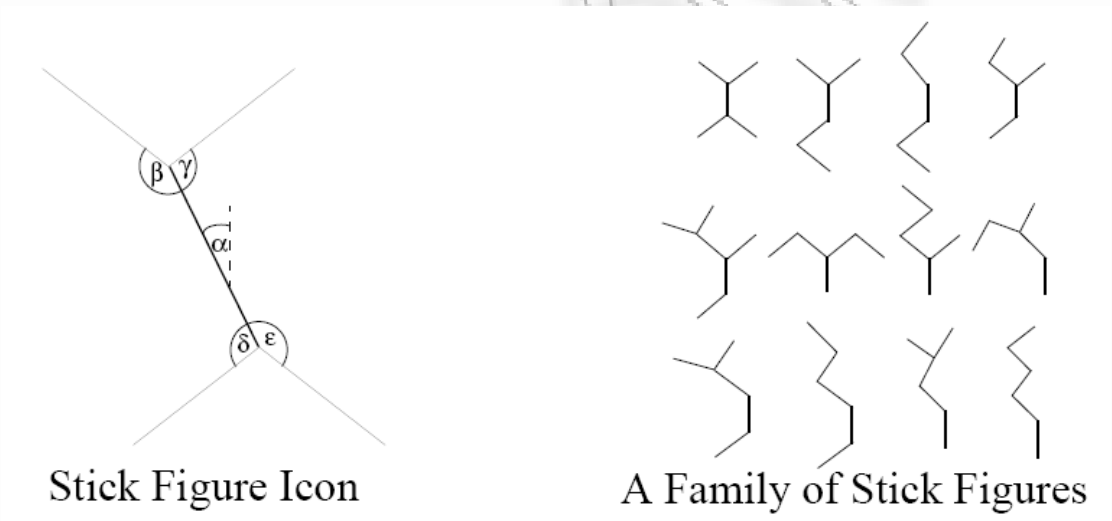


3.3 Stick Figure Icons

Η μέθοδος αυτή εισήχθηκε από τους Pickett & Grinstein (1988) και αρχικά εφαρμόστηκε για την οπτικοποίηση δεδομένων με 5 μεταβλητές καθώς το εικονογράφημα που χρησιμοποιείται αποτελείται από 5 ευθύγραμμα τμήματα, με το κάθε τμήμα να αντιστοιχεί σε μια μεταβλητή. Ένα από αυτά τα τμήματα αποτελεί το κύριο σώμα του εικονογραφήματος και τα υπόλοιπα 4 αποτελούν τα άκρα του. Η τιμή της παρατήρησης για κάθε μια από τις 4 μεταβλητές μετασχηματίζεται με τέτοιο τρόπο ώστε η γωνία κλίσης του κάθε ευθύγραμμου τμήματος αλλά και ο τρόπος που τοποθετούνται τα άκρα πάνω στο σώμα να υποδηλώνει την τιμή αυτή. Η τιμή της $5^{ης}$ μεταβλητής χαρακτηρίζεται με την κλίση του ευθύγραμμου τμήματος το οποίο αποτελεί το σώμα. Αργότερα προτάθηκε να απεικονιστούν και περισσότερες των 5 μεταβλητών, χρησιμοποιώντας το πάχος, το μήκος ή και κάποιους χρωματισμούς των άκρων. Η υποκειμενικότητα όμως με την οποία γίνεται η αντιστοίχιση των

μεταβλητών στα χαρακτηριστικά των stick figure icons αποτελεί μειονέκτημα της μεθόδου καθώς μπορεί να οδηγήσει σε λανθασμένη συμπερασματολογία. Επίσης χρειάζεται ιδιαίτερη εξοικείωση του αναγνώστη με τον τρόπο κατασκευής των εικονογραφημάτων αυτών ώστε να είναι σε θέση να αποκωδικοποιήσει πληροφορίες από αυτά.

Πιο κάτω δίνουμε ένα χαρακτηριστικό για τη λογική κατασκευής Stick Figure Icon (αριστερά) αλλά και μια ομάδα από Stick Figure Icons (δεξιά), τα οποία μπορούν να κατασκευαστούν με τα λογισμικά ExVis (Grinstein & Pickett (1991)), VisDB (http://www.dbs.informatik.uni-muenchen.de/dbs/projekt/visdb/visdb.html) και XmdvTool (http://davis.wpi.edu/~xmdv/).



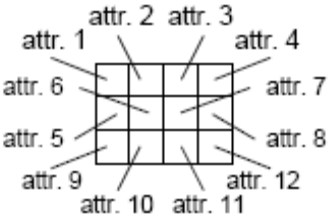
Stick Figure Icon

A Family of Stick Figures

3.4 Κωδικοεικόνες

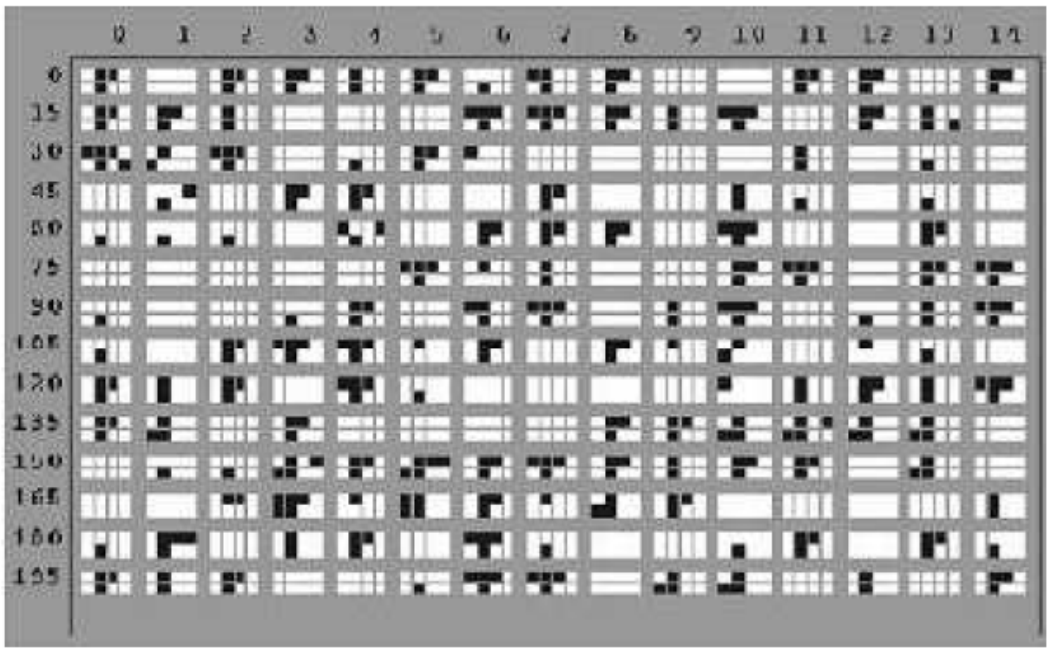
Η μέθοδος αυτή εισήχθηκε από τον Beddow (1990) με σκοπό να υποστηρίξει την οπτικοποίηση μεγάλων συνόλων επιστημονικών δεδομένων και να αναδείξει σχέσεις μεταξύ των μεταβλητών τις οποίες δε μπορούσαν να φανερώσουν στους ερευνητές οι αριθμητικές μέθοδοι. Στηρίζεται στην ιδέα ότι κάθε πολυμεταβλητή παρατήρηση αναπαρίσταται με ένα ορθογώνιο παραλληλόγραμμο το οποίο με τη σειρά του επιμερίζεται σε τόσα ισομεγέθη τμήματα, συνήθως ορθογώνια παραλληλόγραμμα, όσες είναι και οι μεταβλητές των δεδομένων. Στη συνέχεια οι τιμές της κάθε

μεταβλητής κατηγοριοποιούνται σε επίπεδα (π.χ. μικρές, μεσαίες και μεγάλες) με βάση κάποιο κριτήριο που θέτουμε (πολύ συχνά γίνεται σύγκριση με τη διακύμανση των τιμών της μεταβλητής). Ακολούθως δίνεται ένας χρωματισμός σε κάθε κατηγορία (μικρές τιμές – άσπρο, μεσαίες τιμές – γκρί, μεγάλες τιμές – μαύρο) και έτσι προκύπτει ένα χρωματισμένο ορθογώνιο για την τιμή κάθε παρατήρησης στην κάθε μεταβλητή. Δίνεται η δυνατότητα στον ερευνητή να καθορίσει τον αριθμό των επιπέδων που κατηγοριοποιούνται οι μεταβλητές, τα κριτήρια με τα οποία γίνεται η κατηγοριοποίηση αυτή αλλά και τους χρωματισμούς που θα χρησιμοποιηθούν. Είναι προφανές ότι δε μπορούν να αναπαρασταθούν πάρα πολλές μεταβλητές καθώς θα γίνει περίπλοκο και δυσανάγνωστο το γράφημα, για την ανάγνωση του οποίου απαιτείται εξοικείωση από τον χρήστη. Παρόλα αυτά τα εικονογραφήματα που προκύπτουν με τη μέθοδο των κωδικοεικόνων είναι αποτελεσματικά για πολύ μεγάλα σύνολα δεδομένων, ειδικά στην περίπτωση που οι μεταβλητές έχουν μικρό εύρος τιμών ή είναι δίτιμες.

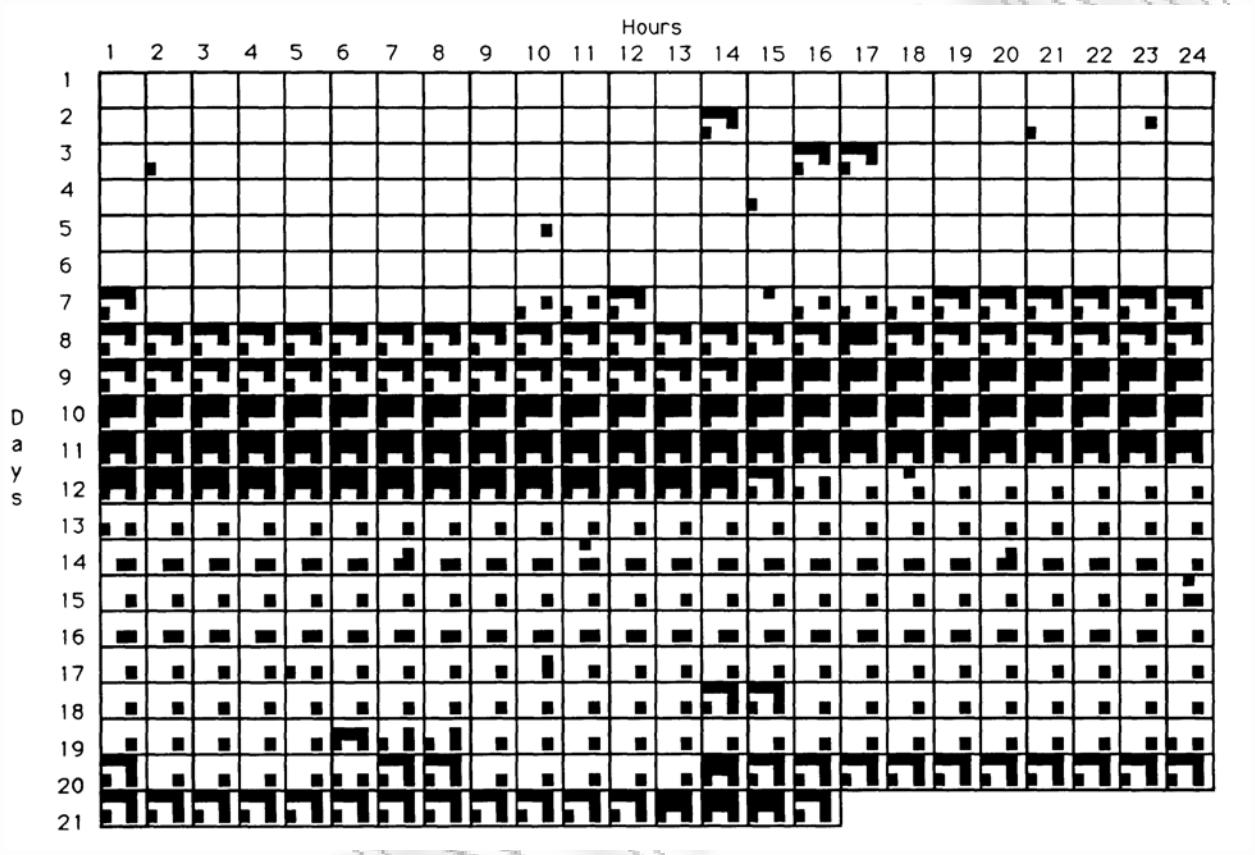


Στο διπλανό σχήμα δίνεται μια κλασσική δομή ενός εικονογραφήματος με 12 μεταβλητές αλλά και δύο παραδείγματα οπτικοποίησης δεδομένων με κωδικοεικόνες, κωδικοποιώντας τις παρατηρήσεις με τη χρήση 2 χρωμάτων (λευκό – μαύρο).

Το γράφημα που ακολουθεί αφορά μικροβιολογικά δεδομένα 8 μεταβλητών για κάθε



ένα από 210 άτομα, ενώ το επόμενο γράφημα αφορά ωριαίες μετεωρολογικές μετρήσεις για 8 χαρακτηριστικά. Τέτοια γραφήματα μπορούν να κατασκευαστούν με το λογισμικό XmdvTool (http://davis.wpi.edu/~xmdv/).

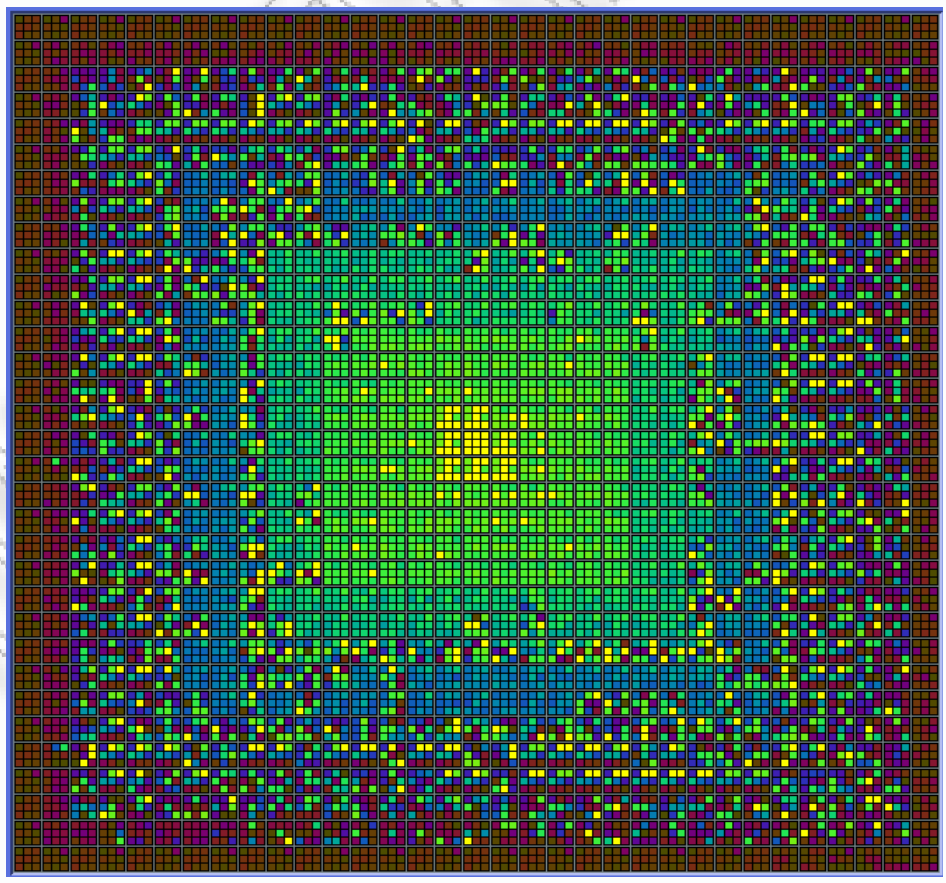


3.5 Χρωμοεικόνες

Η μέθοδος αυτή η οποία αναπτύχθηκε από τον Levkowitz (1991), χρησιμοποιεί μια περιοχή, συνήθως ένα τετράγωνο, για να αναπαραστήσει κάθε πολυδιάστατη παρατήρηση. Το τετράγωνο χωρίζεται σε υποπεριοχές στις οποίες δίνονται διαφορετικοί χρωματισμοί ενώ η παράμετρος που αντιστοιχεί σε κάθε υποπεριοχή λαμβάνει απόχρωση που να αντιστοιχεί στην τιμή της. Πολλές φορές χρησιμοποιείται σήμανση για συγκεκριμένες πλευρές ώστε να τονιστούν κάποιες παράμετροι με ιδιαίτερη σημασία.

Στο διπλανό σχήμα δίνεται ένα παράδειγμα κατασκευής χρωμοεικόνας για 8 μεταβλητές.

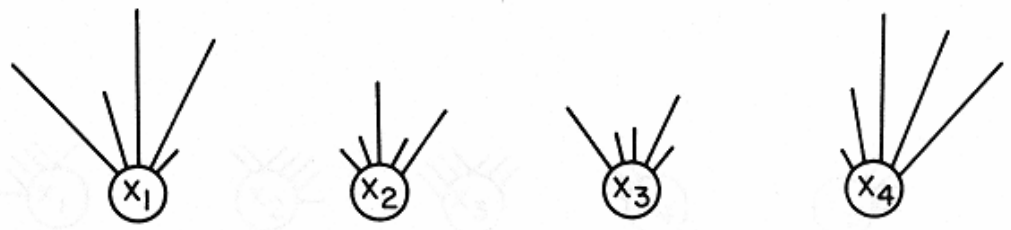
Ένας άλλος τρόπος για να κατασκευαστεί μια χρωμοεικόνα είναι να δοθούν διάφοροι χρωματισμοί στις πλευρές και στη συνέχεια να συμπληρωθούν τα τμήματα με τις κατάλληλες μίξεις των χρωμάτων. Η τακτική αυτή δίνει καλύτερο συνδυασμό των παραμέτρων και βελτιωμένο οπτικό αποτέλεσμα καθώς οι τιμές τους συνενώνονται με ομαλό τρόπο. Εκτός του τετραγώνου μπορούν να χρησιμοποιηθούν και πολύγωνα άλλης μορφής, ανάλογα με τον αριθμό των παραμέτρων αλλά και τη φύση των δεδομένων. Ο αριθμός των μεταβλητών που αναπαριστούνται συνήθως μέσω μιας τετραγωνικής χρωμοεικόνας περιορίζεται στις 8 καθώς ο επιμερισμός σε περισσότερα τμήματα κάνει το γράφημα πολύ δύσκολο στην ερμηνεία. Το πρόβλημα αυτό μπορεί να αντιμετωπιστεί εν μέρει χρησιμοποιώντας και μια τρίτη διάσταση στα εικονογραφήματα, π.χ. το ύψος, που είναι δυνατόν να αυξήσει το πλήθος των παραμέτρων που αναπαρίστανται. Παρακάτω δίνεται ένα παράδειγμα γραφήματος με χρωμοεικόνες, με δεδομένα που έχουν προκύψει από προσομοίωση, το οποίο έχει κατασκευαστεί με τη βοήθεια του λογισμικού VisDB (http://www.dbs.informatik.uni-muenchen.de/dbs/projekt/visdb/visdb.html).



3.6 Γλυφογραφήματα

Τα γλυφογραφήματα εφαρμόστηκαν από τον Anderson (1957) σε βοτανολογικά δεδομένα με σκοπό την μελέτη διαφόρων ποικιλιών καλαμποκιού. Στη συνέχεια χρησιμοποιήθηκαν και σε άλλες μορφές δεδομένων με πολλές μεταβλητές. Πρόκειται για εικονογραφήματα τα οποία έχουν ως κεντρικό σώμα ένα κύκλο από τον οποίο προεκτείνονται ακτίνες, που αντιστοιχούν στις μεταβλητές των δεδομένων. Η αρχική εφαρμογή τους χρησιμοποιούσε μια κατηγοριοποίηση των τιμών των μεταβλητών σε μικρές, μεσαίες και μεγάλες. Φυσικά η κατηγοριοποίηση μπορεί να γίνει και με διαφορετικά κριτήρια, αρκεί τα επίπεδα να μην είναι πολλά έτσι ώστε να διευκολύνεται η άντληση πληροφοριών από το κάθε εικονογράφημα. Το μήκος της κάθε ακτίνας αντιστοιχεί στο μέγεθος της τιμής της κάθε μεταβλητής. Αφού καθοριστεί και η αντιστοίχιση των ακτίνων στις μεταβλητές, θα έχουμε κατασκευάσει τόσα γλυφογραφήματα όσες είναι και οι παρατηρήσεις μας, με τόσες ακτίνες το καθένα όσες είναι οι μεταβλητές. Το ανθρώπινο μάτι μπορεί να εντοπίσει σχέσεις και ομαδοποιήσεις μεταξύ των δεδομένων, εφόσον όμως οι μεταβλητές δεν υπερβαίνουν τις 7 τον αριθμό. Για το λόγο αυτό, σε περίπτωση ύπαρξης πολλών μεταβλητών, είναι προτιμότερο να προηγείται μια μέθοδος επιλογής των σημαντικότερων ή ακόμα και να μετατρέπονται οι συσχετισμένες μεταβλητές σε ενιαίους δείκτες. Θεωρητικά οι ακτίνες μπορούν να τοποθετηθούν περιμετρικά του κυκλικού σώματος του εικονογραφήματος αλλά είναι προτιμότερο να τις τοποθετούμε προς μια συγκεκριμένη κατεύθυνση ώστε να διευκολύνουμε την ανάγνωσή τους. Ακόμα υπάρχει η δυνατότητα χρήσης επιπλέον σημάνσεων (π.χ. κουκίδες) για να τονίσουμε κάποια χαρακτηριστικά που παρουσιάζουν ιδιαίτερο ενδιαφέρον. Μια γενική διαπίστωση είναι ότι ο ερευνητής θα πρέπει να είναι εξοικειωμένος με τη μορφή των metroglyphs σε κάθε περίπτωση, καθώς και με την αντιστοίχιση των μεταβλητών στις ακτίνες ώστε να μπορεί γρήγορα και εύκολα να αποκωδικοποιήσει και να εντοπίσει τις συσχετίσεις και τις ομαδοποιήσεις των δεδομένων.

Πιο κάτω δίνονται ενδεικτικά 4 γλυφογραφήματα που αντιστοιχούν σε 4 παρατηρήσεις 5 μεταβλητών.

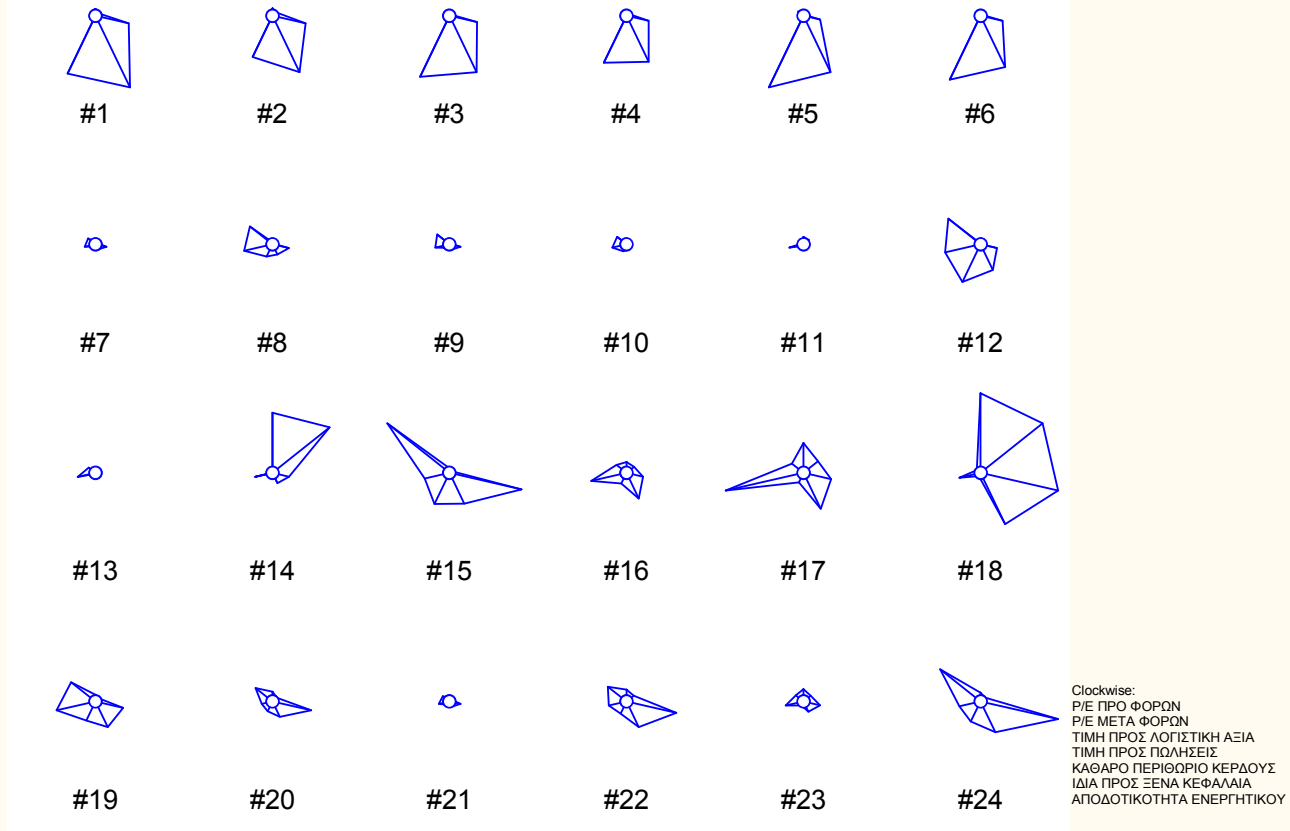


3.7 Διαγράμματα Αστέρων

Τα διαγράμματα αστέρων παρουσιάστηκαν από τον Chambers (1983) και αποτελούν μια από τις πιο διαδεδομένες τεχνικές οπτικοποίησης πολυδιάστατων δεδομένων. Αναφέρονται συχνά και ως διαγράμματα ραντάρ (radar plots) ή διαγράμματα αράχνης (spider plots). Η κάθε πολυμεταβλητή παρατήρηση απεικονίζεται με τη μορφή ενός αστεριού, το οποίο έχει τόσες ακτίνες όσες είναι και οι μεταβλητές. Οι ακτίνες ισαπέχουν μεταξύ τους και το μήκος της κάθε μιας υποδηλώνει την τιμή της αντίστοιχης μεταβλητής. Βέβαια θα πρέπει να προηγείται μια τυποποίηση των τιμών των μεταβλητών για να είναι πιο ομοιόμορφη η απεικόνιση των διαγραμμάτων αστέρων. Για να βελτιωθεί η αποτελεσματικότητα των διαγραμμάτων αστέρων, τα άκρα των ακτινών ενώνονται μεταξύ τους σχηματίζοντας περιμετρικά ένα πολύγωνο. Μεγάλη σημασία έχει και σε αυτή τη μέθοδο η σειρά με την οποία αντιστοιχίζονται οι μεταβλητές στις ακτίνες του γραφήματος αλλά και το πλήθος των μεταβλητών. Όσο αυξάνεται ο αριθμός τους τόσο μειώνεται η αποτελεσματικότητα των διαγραμμάτων αστέρων και δυσκολεύει η ανάγνωσή τους. Με ένα μέτριο αριθμό μεταβλητών πάντως, είναι πολύ εύκολη η σύγκριση μεταξύ παρατηρήσεων και η ύπαρξη ή όχι ομοιοτήτων μεταξύ τους, ο εντοπισμός των επικρατέστερων μεταβλητών αλλά και η ύπαρξη ακραίων παρατηρήσεων. Η σύγκριση μπορεί να διευκολυνθεί με τον τονισμό των παρατηρήσεων με τα πιο έντονα χαρακτηριστικά (σκίαση, διακεκομμένες γραμμές κ.α.). Επίσης είναι προτιμότερο η αντιστοίχιση των μεταβλητών στις ακτίνες να γίνεται με συγκεκριμένο και όχι με τυχαίο τρόπο (π.χ. οι θετικά συσχετισμένες μεταβλητές να τοποθετούνται σε γειτονικές ακτίνες).

Σε πολλές περιπτώσεις είναι προτιμότερο να γίνονται αναδιατάξεις των μεταβλητών και να συγκρίνονται τα αποτελέσματα ώστε να υπάρχει η δυνατότητα να εξεταστούν όσο περισσότερες σχέσεις μεταξύ των δεδομένων. Ακόμα είναι καλύτερα από άποψη ερμηνείας, τα διαγράμματα αστέρων να τοποθετούνται στο διάγραμμα με κάποια λογική διαδικασία (π.χ. τα εικονογραφήματα ομαδοποιημένων παρατηρήσεων να βρίσκονται σε κοντινές θέσεις). Σε γενικές γραμμές, τα διαγράμματα αστέρων είναι χρήσιμα για μικρά ή μεσαία σύνολα δεδομένων καθώς για μεγάλο πλήθος παρατηρήσεων δεν παρέχουν ουσιαστική πληροφόρηση. Αυτό προκαλείται σε αρκετές περιπτώσεις από τον περιορισμένο χώρο απεικόνισης του διαγράμματος που έχουμε στη διάθεση μας.

Στη συνέχεια θα δούμε μια εφαρμογή των διαγραμμάτων αστέρων στα χρηματοοικονομικά δεδομένα των 24 εταιρειών, με τη βοήθεια του στατιστικού πακέτου Statistica. Στο κάτω δεξιά μέρος αναγράφεται ο τρόπος που έχει γίνει η αντιστοίχιση των μεταβλητών στις ακτίνες των εικονογραφημάτων, σύμφωνα με τη φορά των δεικτών του ρολογιού. Ανάλογα εικονογραφήματα μπορούν να κατασκευαστούν και με το πρόγραμμα SAS/GRAPH (http://www.sas.com/technologies/bi/query_reporting/graph/index.html), το Dataplot (http://www.itl.nist.gov/div898/software/dataplot.html/), το Ggobi (http://www.ggobi.org/), την R, το Mathematica, το Matlab (http://www.mathworks.com/), το XmdvTool και άλλα στατιστικά προγράμματα.

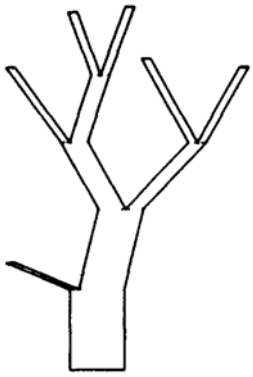


3.8 Δένδρα των Kleiner & Hartigan

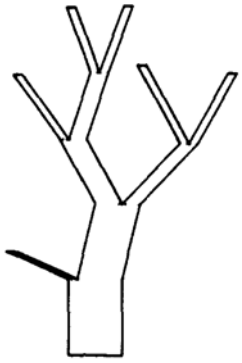
Η μέθοδος αυτή εισήχθηκε από τους Kleiner & Hartigan (1981) και στηρίζεται στην αναπαράσταση πολυδιάστατων δεδομένων με τη μορφή εικονογραφημάτων σε σχήμα δένδρου. Κάθε παρατήρηση οπτικοποιείται μέσω ενός δένδρου το οποίο παρουσιάζει με ιεραρχική δομή τις τιμές των διαφόρων μεταβλητών. Συνήθως απαιτείται μια τυποποίηση των τιμών πριν προχωρήσει η κατασκευή των δένδρων. Το πλάτος κάθε κλαδιού στο δένδρο είναι ανάλογο του πλήθους των μεταβλητών που βρίσκονται πάνω από αυτό. Επομένως κάθε φύλλο, δηλαδή κάθε ακραίο κλαδί, έχει πλάτος ίσο με την ελάχιστη μονάδα μέτρησης εφόσον αντιστοιχεί σε μια μεταβλητή. Η γωνία μεταξύ των δύο κλαδιών σε μια διακλάδωση είναι γραμμική συνάρτηση του μέγιστου λογαρίθμου της απόστασης μεταξύ των μεταβλητών που βρίσκονται πάνω από τα δύο αυτά κλαδιά. Επίσης είναι σημαντικό να προαποφασίζεται η σειρά με την οποία θα τοποθετηθούν οι μεταβλητές, και κατά συνέπεια τα κλαδιά, πάνω στο δένδρο. Ακόμα, το μήκος κάθε κλαδιού είναι ανάλογο της μέσης τιμής των μεταβλητών που βρίσκονται από πάνω του.

Το κύριο μειονέκτημα της μεθόδου είναι η δυσκολία να γίνουν συγκρίσεις μεταξύ μεταβλητών στο ίδιο δένδρο, ακόμα κι αν αυτές βρίσκονται σε κοντινά κλαδιά. Πρόκειται όμως για αποτελεσματική μέθοδο στο να εντοπίζει τις ομαδοποιήσεις των δεδομένων, τις κύριες τάσεις αλλά και τις ακραίες τιμές.

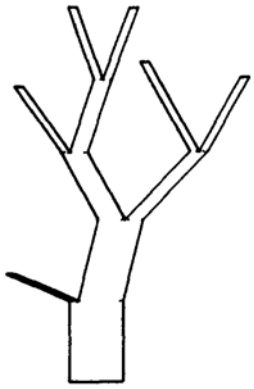
Στην επόμενη σελίδα δίνεται το παράδειγμα δένδρων των Kleiner & Hartigan για δεδομένα που αφορούν εκλογικά αποτελέσματα σε έξι αμερικάνικες πολιτείες. Η κατασκευή του διαγράμματος έγινε με χρήση του προγράμματος SAS/GRAPH (http://www.sas.com/technologies/bi/query_reporting/graph/index.html).



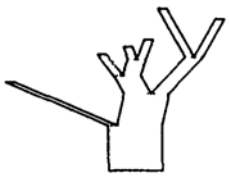
MISSOURI



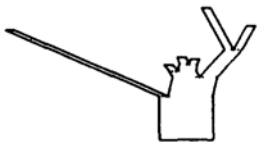
MARYLAND



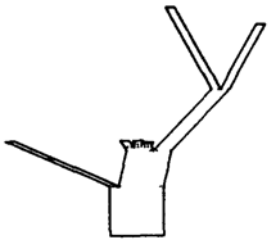
KENTUCKY



LOUISIANA



MISSISSIPPI



S.C.

ΚΕΦΑΛΑΙΟ 4

Τεχνικές Εικονοστοιχείων

4.1 Εισαγωγή

Η κατασκευή γραφημάτων για την οπτικοποίηση πολυμεταβλητών δεδομένων με τη χρήση εικονοστοιχείων βασίζεται στο γεγονός ότι μπορούμε να αναπαραστήσουμε κάθε τιμή των πολυμεταβλητών παρατηρήσεων με ένα χρωματισμένο εικονοστοιχείο. Αντιστοιχίζοντας κάθε τιμή με ένα εικονοστοιχείο, καταφέρνουμε να εκμεταλλευτούμε πλήρως την επιφάνεια προβολής και να οπτικοποιήσουμε πολύ μεγάλα σύνολα δεδομένων. Οι τεχνικές αυτές, αν και είναι ιδανικές για μεγάλο όγκο δεδομένων, παρουσιάζουν αδυναμίες στο να εντοπίσουν σύνθετες δομές των δεδομένων.

4.2 Διαγράμματα Κυκλικών Τομέων

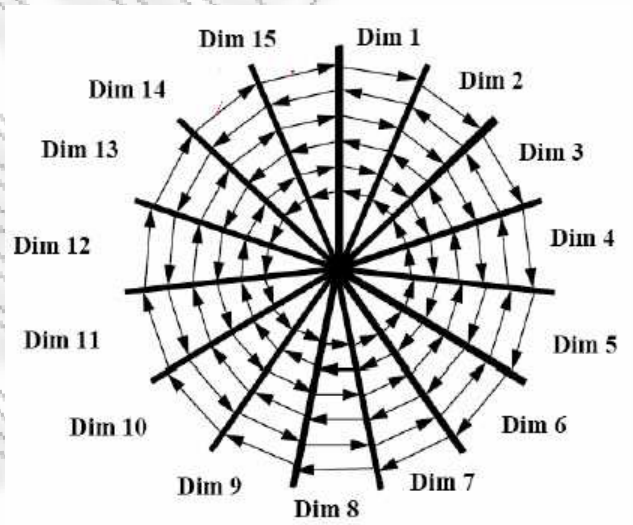
Η τεχνική των διαγραμμάτων κυκλικών τομέων παρουσιάστηκε από τους Ankerst, Keim & Kriegel (1996) με σκοπό να συμβάλλει στην οπτικοποίηση πολύ μεγάλων σε όγκο πολυμεταβλητών δεδομένων. Η βασική ιδέα κατασκευής ενός τέτοιου γραφήματος είναι ο επιμερισμός του κύκλου σε τόσα τμήματα, όσες είναι και οι μεταβλητές των δεδομένων. Σε κάθε ένα από τα κυκλικά τμήματα αντιστοιχίζεται και μια μεταβλητή ενώ η τιμή κάθε παρατήρησης παρουσιάζεται με τη βοήθεια ενός κατάλληλα χρωματισμένου εικονοστοιχείου. Συνήθως χρησιμοποιείται μια χρωματική κλίμακα όπου τα πιο ανοιχτά χρώματα αντιστοιχούν σε μεγαλύτερες τιμές και τα πιο σκούρα χρώματα αντιστοιχούν σε μικρότερες τιμές.

Η διάταξη των τιμών των παρατηρήσεων σε κάθε τμήμα του κύκλου γίνεται με συγκεκριμένο τρόπο, ξεκινώντας από το κέντρο του κύκλου και κινούμενοι προς το εξωτερικό, αλλάζοντας κατεύθυνση κάθε φορά που συναντάμε τα όρια του

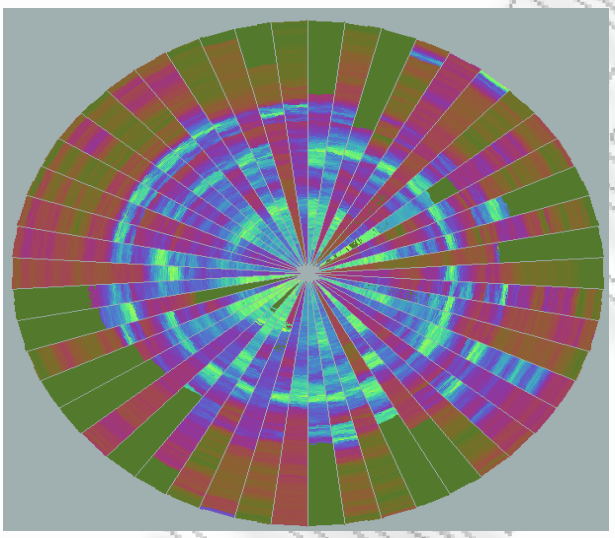
συγκεκριμένου τμήματος, ακολουθώντας έτσι μια φυσική διάταξη που έχουν τα δεδομένα εκ των προτέρων. Αφού ολοκληρωθεί η διαδικασία για μια μεταβλητή, επαναλαμβάνεται με τον ίδιο ακριβώς τρόπο για τις υπόλοιπες μέχρι να συμπληρωθούν όλα τα τμήματα του κύκλου και να απεικονιστούν έτσι όλες οι μεταβλητές. Με τον τρόπο αυτό, εάν τα δεδομένα μας είναι ιστορικά, οι τιμές των παλαιότερων παρατηρήσεων θα βρίσκονται κοντά στο κέντρο του κύκλου ενώ οι πιο πρόσφατες θα βρίσκονται στο εξωτερικό μέρος του.

Η χρήση της μεθόδου αυτής δίνει τη δυνατότητα να αναδιατάξουμε τα τμήματα (και κατά συνέπεια τις μεταβλητές των δεδομένων) ώστε να διευκολυνθεί η ανακάλυψη σχέσεων μεταξύ τους αλλά και η αποτελεσματική διερεύνηση ομαδοποιήσεων των πολυπληθών δεδομένων που άλλες τεχνικές οπτικοποίησης δεν επιτρέπουν να πραγματοποιηθεί. Σε γενικές γραμμές η μέθοδος των διαγραμμάτων κυκλικών τομέων θεωρείται πιο αποτελεσματική από τις υπόλοιπες μεθόδους της ίδιας κατηγορίας, καθώς προσφέρει ένα σημείο αναφοράς στον αναγνώστη του γραφήματος, το κέντρο του κύκλου δηλαδή, αναβαθμίζοντας έτσι την οπτική σύγκριση των δεδομένων.

Πιο κάτω δίνεται σχηματικά ο τρόπος διάταξης των τιμών των αποστάσεων (εικονοστοιχείων) των παρατηρήσεων με τη μέθοδο των διαγραμμάτων κυκλικών τομέων, σε δεδομένα με 15 μεταβλητές.



Ακολούθως δίνεται ένα παράδειγμα εφαρμογής των διαγραμμάτων κυκλικών τομέων σε δεδομένα 50 χρηματιστηριακών μετοχών, με τη χρήση του λογισμικού VisDB (http://www.dbs.informatik.uni-muenchen.de/dbs/projekt/visdb/visdb.html), όπου τα πιο ανοιχτά χρώματα αντιστοιχούν σε υψηλές τιμές ενώ τα πιο σκούρα χρώματα σε χαμηλές τιμές.



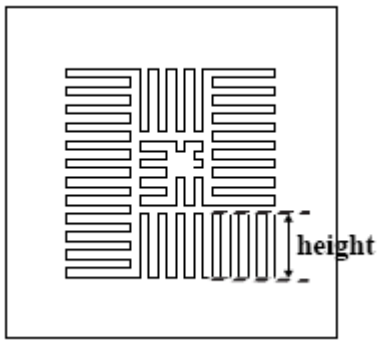
4.3 Τεχνικές Σπείρας / Τεχνικές Αξόνων

Οι τεχνικές αυτές αναπτύχθηκαν από τους Keim & Kriegel (1994) με σκοπό να αναπαραστήσουν μεγάλα σύνολα πολυδιάστατων δεδομένων, με χρήση εικονοστοιχείων. Ο αρχικός χώρος προβολής χωρίζεται σε τόσα τμήματα, όσα είναι και τα χαρακτηριστικά των δεδομένων. Η βασική ιδέα είναι να οπτικοποιηθούν τα δεδομένα σε ένα πλαίσιο συγκεκριμένου ερωτήματος από τον ερευνητή, ώστε στη συνέχεια το γράφημα να του παρέχει πληροφόρηση σύμφωνα με την κατεύθυνση που αυτός έχει δώσει.

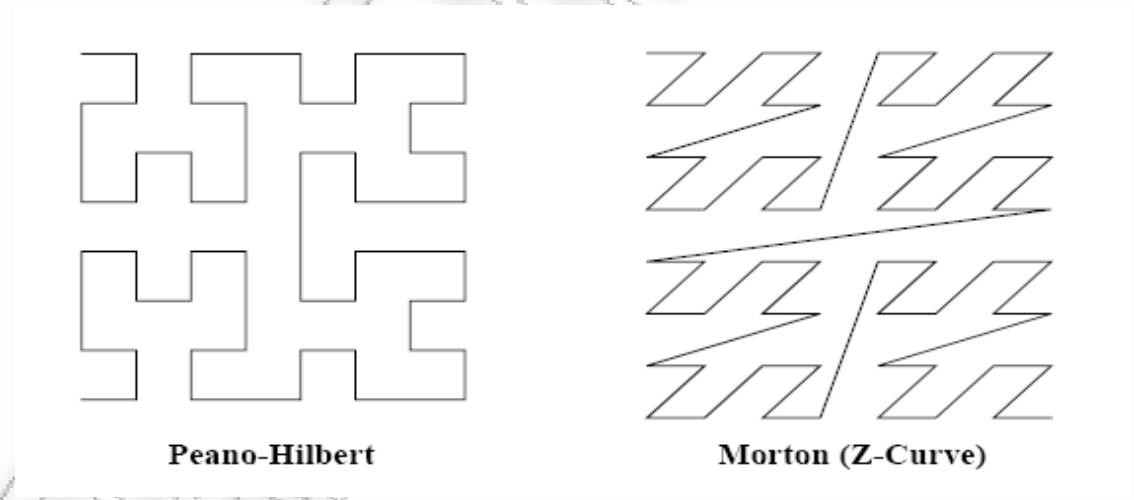
Τα εικονοστοιχεία δεν αντιστοιχούν στις τιμές των παρατηρήσεων για κάθε χαρακτηριστικό των δεδομένων αλλά εκφράζουν τις αποστάσεις ή αλλιώς τη σχετικότητα των παρατηρήσεων με βάση το ερώτημα που έχει τεθεί από τον ερευνητή. Κατά συνέπεια ο χρωματισμός των εικονοστοιχείων έχει να κάνει αποστάσεις αυτές και όχι με τις τιμές των δεδομένων. Επίσης η διάταξη των

εικονοστοιχείων στο χώρο γίνεται με βάση τις αποστάσεις που έχουν οριστεί, δηλαδή τη σχέση τους με το ερώτημα που έχει τεθεί. Μια λογική διάταξη είναι να τοποθετούνται κοντά στο κέντρο οι παρατηρήσεις με τη μεγαλύτερη συνάφεια και όσο αυτή μειώνεται να απομακρυνόμαστε από το κέντρο.

Πιο κάτω δίνεται ο απλός σπειροειδής τρόπος διάταξης των εικονοστοιχείων για την κατασκευή του γραφήματος.

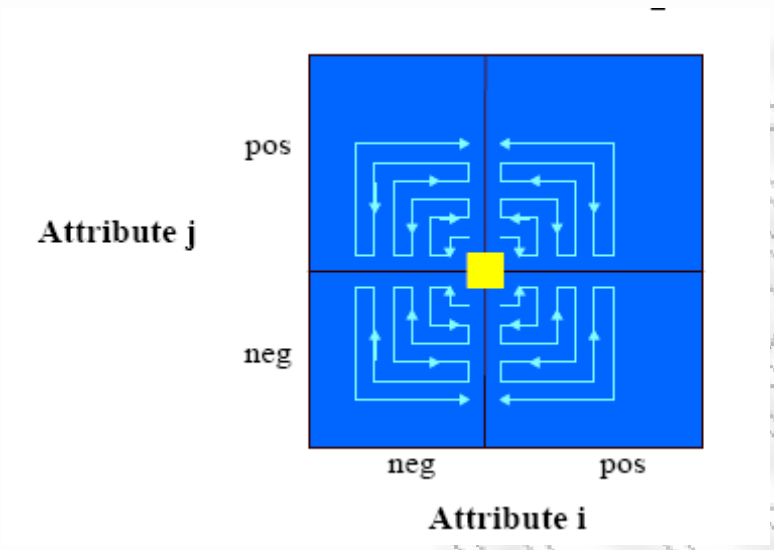


Υπάρχουν και άλλες μέθοδοι διάταξης των εικονοστοιχείων με κυριότερες την διάταξη Peano – Hilbert και τη διάταξη Morton, των οποίων δίνουμε παρακάτω μια σχηματική αναπαράσταση.

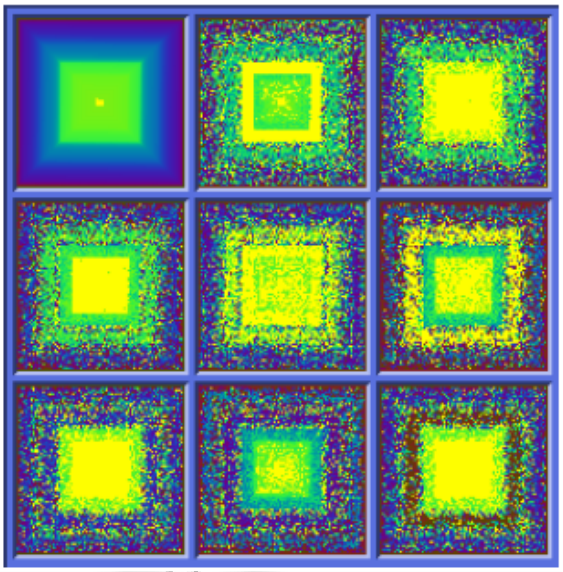


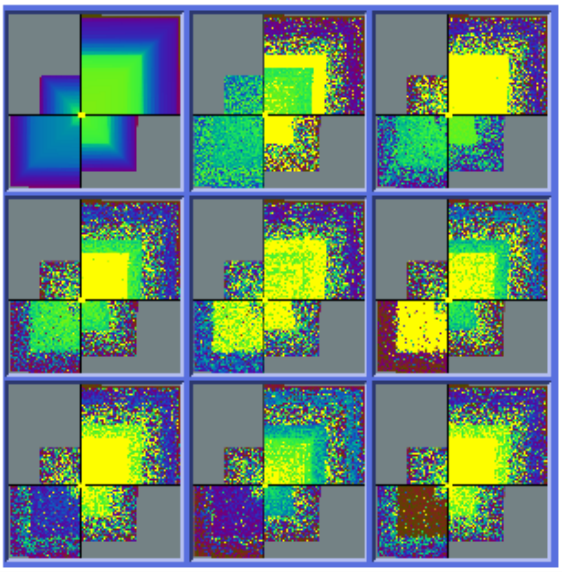
Μια βελτιωμένη εκδοχή της τεχνικής σπείρας είναι η τεχνική αξόνων, κατά την οποία το διάγραμμα χωρίζεται με δύο κάθετους άξονες σε τέσσερα τμήματα, σύμφωνα με το πρόσημο των αποστάσεων μεταξύ δύο μεταβλητών. Έτσι για τη μια μεταβλητή οι αρνητικές αποστάσεις τοποθετούνται αριστερά και οι θετικές δεξιά ενώ για τη δεύτερη μεταβλητή οι θετικές τοποθετούνται στο πάνω μέρος και οι αρνητικές

στο κάτω. Ακολουθεί και μια σχηματική αναπαράσταση του τρόπου διάταξης των εικονοστοιχείων με την τεχνική των αξόνων.



Τέλος δίνονται δύο παραδείγματα διαγραμμάτων των τεχνικών σπείρας και αξόνων τα οποία κατασκευάστηκαν με τη βοήθεια του λογισμικού VisDB (http://www.dbs.informatik.uni-muenchen.de/dbs/projekt/visdb/visdb.html). Τα δεδομένα που χρησιμοποιήθηκαν είναι 7.000 παρατηρήσεις 8 μεταβλητών οι οποίες κατασκευάστηκαν μέσω προσομοίωσης από την ομοιόμορφη κατανομή.





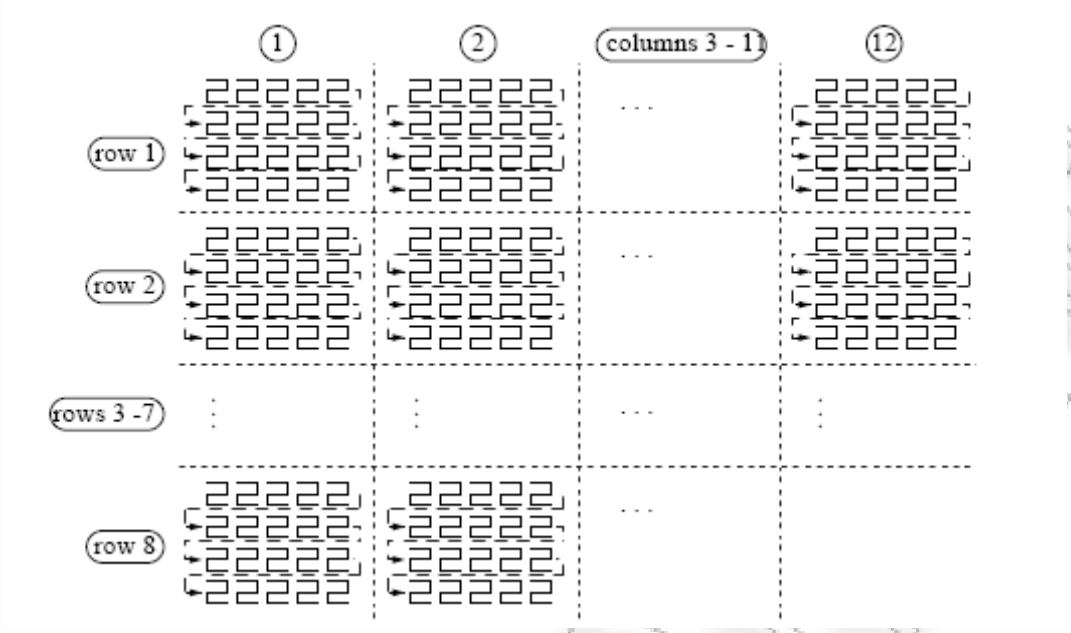
4.4 Τεχνικές Αναδρομικών Σχηματισμών

Πρόκειται για τεχνική η οποία βασίζεται στην αντιστοίχιση της τιμής κάθε μεταβλητής σε ένα χρωματισμένο εικονοστοιχείο και την οργάνωση των εικονοστοιχείων αυτών σε ομάδες. Η οργάνωση των ομάδων αυτών μπορεί να γίνει με πολλούς διαφορετικούς τρόπους και δίνεται η δυνατότητα στον ερευνητή να επιλέξει την κατάλληλη για κάθε περίπτωση.

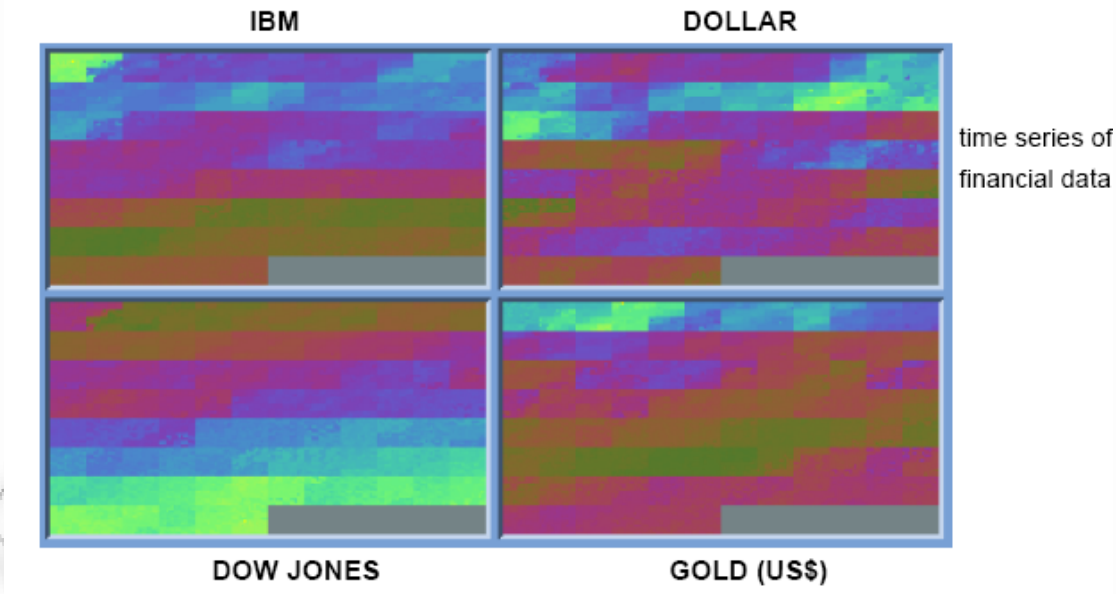
Ο χρωματισμός των εικονοστοιχείων γίνεται έτσι ώστε η απόχρωση να αντιστοιχεί στην τιμή κάθε παρατήρησης. Στη συνέχεια τα ομαδοποιημένα εικονοστοιχεία τοποθετούνται σε ένα πλαίσιο το οποίο αποτελεί και το τελικό γράφημα. Κάθε χαρακτηριστικό των δεδομένων παρουσιάζεται σε ένα ξεχωριστό υποπλαίσιο του τελικού πλαισίου και η διάταξη των εικονοστοιχείων σε αυτό γίνεται με συγκεκριμένο τρόπο, ο οποίος θα πρέπει να είναι ίδιος για όλα τα υποπλαίσια. Μια χαρακτηριστική διάταξη των εικονοστοιχείων προκύπτει ξεκινώντας από το πάνω αριστερά μέρος του υποπλαισίου, τοποθετώντας προς τα δεξιά τα επόμενα, και αλλάζοντας κατεύθυνση όταν συμπληρωθεί η πάνω σειρά, συνεχίζοντας με την ακριβώς επόμενη. Το μέγεθος του τελικού πλαισίου, των υποπλαισίων και κατά συνέπεια των εικονοστοιχείων είναι στην ευχέρεια του ερευνητή αλλά περιορίζεται από τον χώρο παρουσίασης του διαγράμματος (π.χ. οθόνη υπολογιστή).

Η μέθοδος αυτή χρησιμοποιείται κυρίως για δεδομένα τα οποία είναι διατεταγμένα από τη φύση τους π.χ. χρονοσειρές αλλά μπορεί να εφαρμοστεί και σε δεδομένα τα οποία δεν έχουν κάποια φυσική διάταξη. Σε τέτοιες περιπτώσεις ο ερευνητής μπορεί να διατάξει τα δεδομένα με βάση κάποιο χαρακτηριστικό που τον ενδιαφέρει και να κάνει τις συγκρίσεις που επιθυμεί με βάση αυτό.

Ακολουθεί μια σχηματική αναπαράσταση ενός χαρακτηριστικού τρόπου διάταξης των εικονοστοιχείων στα υποπλαίσια αλλά και των υποπλαισίων στο τελικό πλαίσιο.



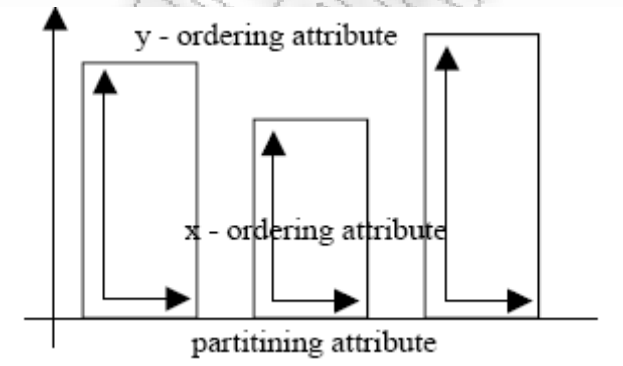
Δίνεται παρακάτω ένα παράδειγμα της τεχνικής αναδρομικών σχηματισμών με χρήση ιστορικών χρηματιστηριακών δεδομένων, το οποίο κατασκευάστηκε με τη βοήθεια του λογισμικού VisDB (http://www.dbs.informatik.uni-muenchen.de/dbs/projekt/visdb/visdb.html).



4.5 Θηκογράμματα Εικονοστοιχείων

Η μέθοδος αυτή προτάθηκε από τους Keim, Hao, Ladisch, Hsu & Dayal (2001) και πρόκειται για μια παραλλαγή των κλασσικών θηκογραμμάτων ώστε να οπτικοποιούνται δεδομένα με περισσότερες των 2 μεταβλητές αλλά και να παρουσιάζονται περισσότερες πληροφορίες που αφορούν τα δεδομένα. Σε αυτήν την περίπτωση η κάθε παρατήρηση αντιστοιχεί σε ένα εικονοστοιχείο. Σκοπός είναι να παρουσιάζονται οι τιμές όλων των παρατηρήσεων για κάποιο χαρακτηριστικό των δεδομένων και όχι οι τιμές τους αθροιστικά όπως συμβαίνει με τα κλασσικά θηκογράμματα.

Η βασική ιδέα είναι να χρησιμοποιούνται ένα ή δύο χαρακτηριστικά για να διαχωρίζονται τα δεδομένα σε ράβδους και στη συνέχεια να χρησιμοποιούνται δύο επιπλέον χαρακτηριστικά των δεδομένων για να επιβάλλεται μια διάταξη μέσα στις ράβδους. Ακολουθεί μια σχηματική αναπαράσταση του τρόπου με τον οποίο υλοποιείται η κατασκευή των θηκογραμμάτων εικονοστοιχείων.

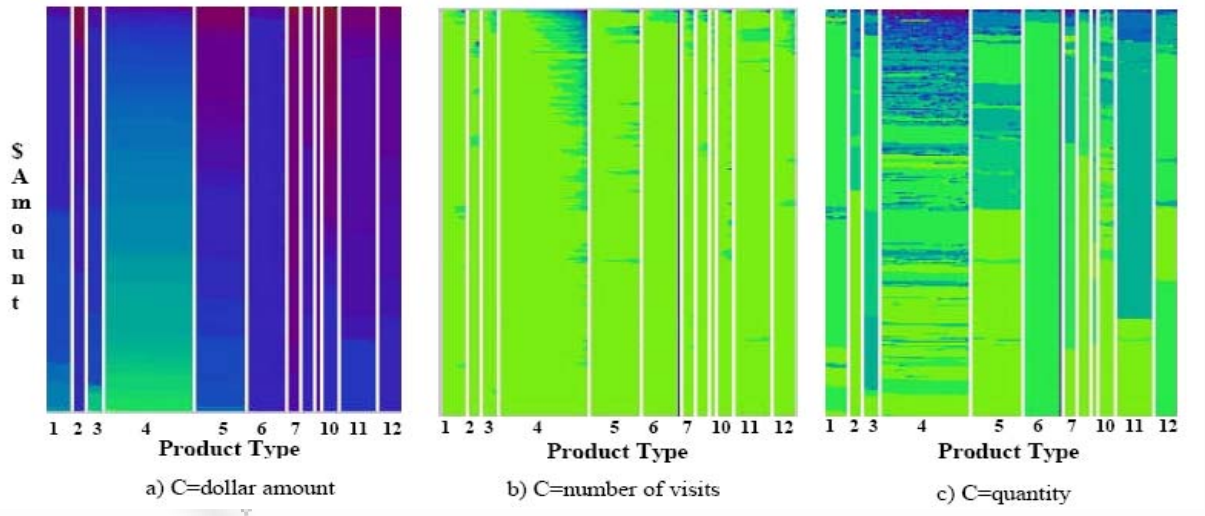


Η ένα προς ένα αντιστοιχία μεταξύ παρατηρήσεων και εικονοστοιχείων μας επιτρέπει να χρησιμοποιήσουμε το χρώμα των εικονοστοιχείων για να απεικονίσουμε ένα επιπλέον χαρακτηριστικό των δεδομένων. Επίσης τα θηκογράμματα εικονοστοιχείων επιτυγχάνουν τη μέγιστη αξιοποίηση του διαθέσιμου χώρου προβολής, καθώς δε χρησιμοποιούν ράβδους ίδιου πλάτους με διαφορετικά ύψη όπως συμβαίνει στα κλασσικά θηκογράμματα αλλά ισοϋψείς ράβδους με κυμαινόμενο πλάτος.

Σε περίπτωση που θέλουμε να συγκρίνουμε περισσότερα χαρακτηριστικά των δεδομένων μπορούμε να χρησιμοποιήσουμε πολλαπλά θηκογράμματα εικονοστοιχείων με τον ίδιο χρωματισμό στα εικονοστοιχεία και τον ίδιο διαχωρισμό στις ράβδους. Είναι ιδιαίτερα σημαντικό να διατηρούνται οι θέσεις των εικονοστοιχείων στα διάφορα γραφήματα σταθερές, ώστε τοποθετώντας τα στη σειρά να μπορέσουμε να ανακαλύψουμε συσχετίσεις μεταξύ των μεταβλητών.

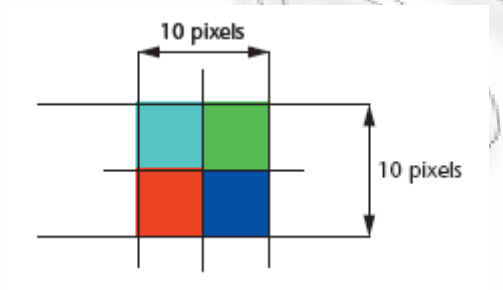
Κάποιες σημαντικοί παράμετροι κατά την κατασκευή των θηκογραμμάτων εικονοστοιχείων είναι να συμπληρωθούν πλήρως με εικονοστοιχεία οι ράβδοι, εκτός ίσως από την τελευταία και να μην υπάρχουν πολλαπλές καταχωρήσεις, δηλαδή κάθε εικονοστοιχείο να καταλαμβάνει μια και μοναδική θέση στη ράβδο. Ακόμα πρέπει να εξασφαλίζεται ότι παρόμοιες παρατηρήσεις θα βρίσκονται σε κοντινές θέσεις αλλά και ότι η διάταξη των ράβδων γίνεται με τον κατάλληλο τρόπο ώστε να αντιστοιχεί στα χαρακτηριστικά που διαχωρίζουν τα δεδομένα. Ένα τελευταίο σημείο που χρίζει προσοχής είναι η διαδικασία χρωματισμού των εικονοστοιχείων και η αντιστοιχία των αποχρώσεων στις τιμές των παρατηρήσεων.

Δίνεται παρακάτω ένα χαρακτηριστικό παράδειγμα πολλαπλών θηκογραμμάτων εικονοστοιχείων μέσω των οποίων ερμηνεύεται η συμπεριφορά 3 διαφορετικών χαρακτηριστικών των δεδομένων. Τέτοια γραφήματα μπορούν να κατασκευαστούν με το λογισμικό VisMine που αναπτύχθηκε στα εργαστήρια της Hewlett Packard (http://www.hpl.hp.com/).



4.6 Attribute blocks

Η μέθοδος αυτή εισήχθηκε από τον Miller (2007), κυρίως με σκοπό την αναπαράσταση πολυδιάστατων δεδομένων σε γεωγραφικό χώρο, όπως τα κλιματολογικά δεδομένα. Κεντρική ιδέα της κατασκευής του ομώνυμου γραφήματος είναι τα Attribute Blocks τα οποία αντιστοιχούν στις πολυδιάστατες παρατηρήσεις. Κάθε block αποτελείται από τόσα κελιά όσες είναι και οι μεταβλητές που θέλουμε να αναπαραστήσουμε. Με τον τρόπο αυτό κάθε κελί του block αντιστοιχεί στην τιμή της μεταβλητής για τη συγκεκριμένη παρατήρηση. Δίνεται παρακάτω ένα παράδειγμα κατασκευής attribute block για 4 μεταβλητές.

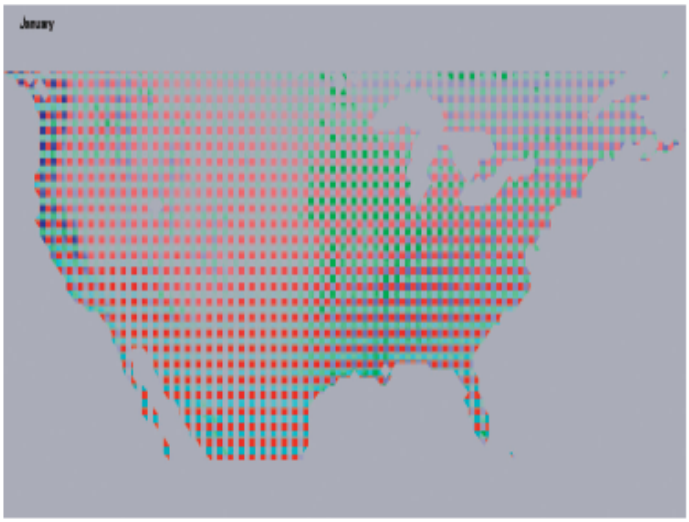


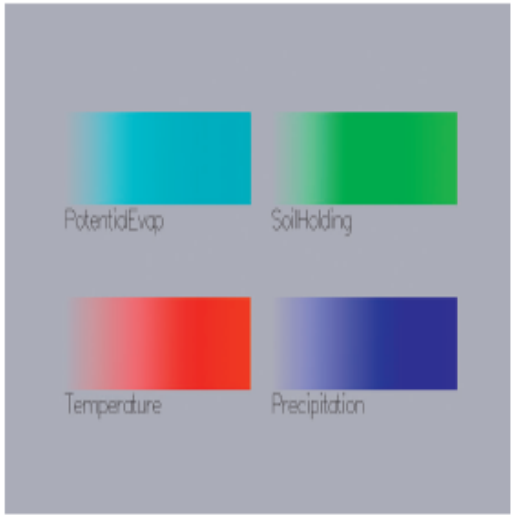
Αφού επιλέξουμε ένα χρώμα για κάθε μεταβλητή, η τιμή της μεταβλητής στο κελί λαμβάνει την κατάλληλη απόχρωση με την εξής λογική: οι μηδενικές τιμές χρωματίζονται με λευκό, χρώμα το οποίο είναι συνήθως στο φόντο της επιφάνειας που θα τοποθετηθούν τα blocks, και όσο η τιμή αυξάνεται, αυξάνεται και ο τόνος του χρώματος μέχρι να φθάσουμε στις μέγιστες τιμές όπου και έχουμε την πιο έντονη απόχρωση. Στη συνέχεια τα attribute blocks τοποθετούνται διατεταγμένα στο γεωγραφικό χώρο ή στην επιφάνεια η οποία μας ενδιαφέρει και σχηματίζεται το τελικό γράφημα.

Με τη διαδικασία χρωματισμού που έχουμε επιλέξει οι υψηλότερες τιμές θα έχουν εντονότερο χρώμα στο γράφημα ενώ οι χαμηλές τιμές θα τείνουν στο χρώμα του φόντου, οπότε θα περνούν απαρατήρητες από τον αναγνώστη του γραφήματος. Είναι προφανές ότι το μέγεθος των κελιών μέσα στα blocks και κατά συνέπεια των ίδιων των blocks, ο χρωματισμός που θα αντιστοιχεί σε κάθε μεταβλητή αλλά και το πλήθος των μεταβλητών που θα συμπεριληφθούν σε κάθε block είναι στην ευχέρεια του ερευνητή να τα καθορίσει.

Υπάρχουν κάποια σημεία τα οποία απαιτούν προσοχή κατά τον καθορισμό των παραμέτρων αυτών, καθώς η αύξηση του πλήθους των μεταβλητών αλλά και του μεγέθους των blocks θα μειώσει την αποτελεσματικότητα του γραφήματος στο να εντοπίζει μεταβολές των τιμών σε σχετικά μικρές περιοχές της επιφάνειας προβολής αλλά και θα δυσκολέψει το διαχωρισμό των μεταβλητών από τον παρατηρητή. Πρακτικά ένας αριθμός 8 μεταβλητών είναι αποδεκτός για την ικανοποιητική απόδοση του γραφήματος. Επίσης η επιλογή των χρωμάτων θα πρέπει να γίνεται έτσι ώστε να υποβοηθείται η αναγνώριση των μεταβλητών και να αποφεύγονται συγχύσεις με τη χρήση παρόμοιων χρωματισμών.

Ακολουθεί το παράδειγμα ενός γραφήματος με attribute blocks για 4 μεταβλητές σε κλιματικά δεδομένα, με επιφάνεια προβολής μια γεωγραφική περιοχή, το οποίο έχει κατασκευαστεί με το λογισμικό SimVis (http://www.simvis.at/). Δίνεται επίσης και η κλίμακα χρωμάτων που χρησιμοποιείται για κάθε μεταβλητή.





ΚΕΦΑΛΑΙΟ 5

Γεωμετρικές Τεχνικές

5.1 Εισαγωγή

Στις γεωμετρικές τεχνικές εκμεταλλευόμαστε γεωμετρικούς μετασχηματισμούς και γεωμετρικές προβολές των δεδομένων για να εντοπίσουμε σχέσεις που πιθανόν να υπάρχουν ανάμεσά τους. Στην κατηγορία αυτή συμπεριλαμβάνονται μερικές από τις πιο διαδεδομένες τεχνικές οπτικοποίησης πολυμεταβλητών δεδομένων, ικανές να χειριστούν διάφορους τύπους δεδομένων και να απαντήσουν σε ποικίλα ερωτήματα και ζητήματα που θέτει ο ερευνητής.

5.2 Διαγράμματα παράλληλων Συντεταγμένων

Το διάγραμμα παράλληλων συντεταγμένων παρουσιάστηκε για πρώτη φορά από τον Inselberg (1985) αλλά ως εργαλείο οπτικοποίησης πολυδιάστατων δεδομένων χρησιμοποιήθηκε από τον Wegman (1990). Πρόκειται για μια πολύ διαδεδομένη μέθοδο, λόγω της ευκολίας υλοποίησής της αλλά και λόγω της αποτελεσματικότητάς της. Αν και σε πρώτη ματιά δε φαίνεται ικανή για ικανοποιητική οπτικοποίηση δεδομένων, με μια προσεκτικότερη ανάγνωση του σχετικού διαγράμματος αναδεικνύονται οι σημαντικές πληροφορίες που παρέχει στον ερευνητή.

Η βασική ιδέα κατασκευής του διαγράμματος παράλληλων συντεταγμένων στηρίζεται στη μετατροπή του καρτεσιανού συστήματος συντεταγμένων σε σύστημα παράλληλων συντεταγμένων. Οι μεταβλητές αναπαρίστανται ως παράλληλοι άξονες και οι τιμές των παρατηρήσεων τοποθετούνται επάνω στους άξονες αυτούς, ενώ μια τεθλασμένη γραμμή ενώνει τα σημεία μεταξύ των αξόνων. Με τον τρόπο αυτό κάθε πολυδιάστατη παρατήρηση παρουσιάζεται ως τεθλασμένη γραμμή στο επίπεδο χωρίς

να έχουμε απώλεια πληροφορίας. Γειτονικά σημεία επί των αξόνων υποδηλώνουν ομαδοποίηση των παρατηρήσεων ενώ και οι ακραίες τιμές είναι πολύ εύκολο να εντοπιστούν.

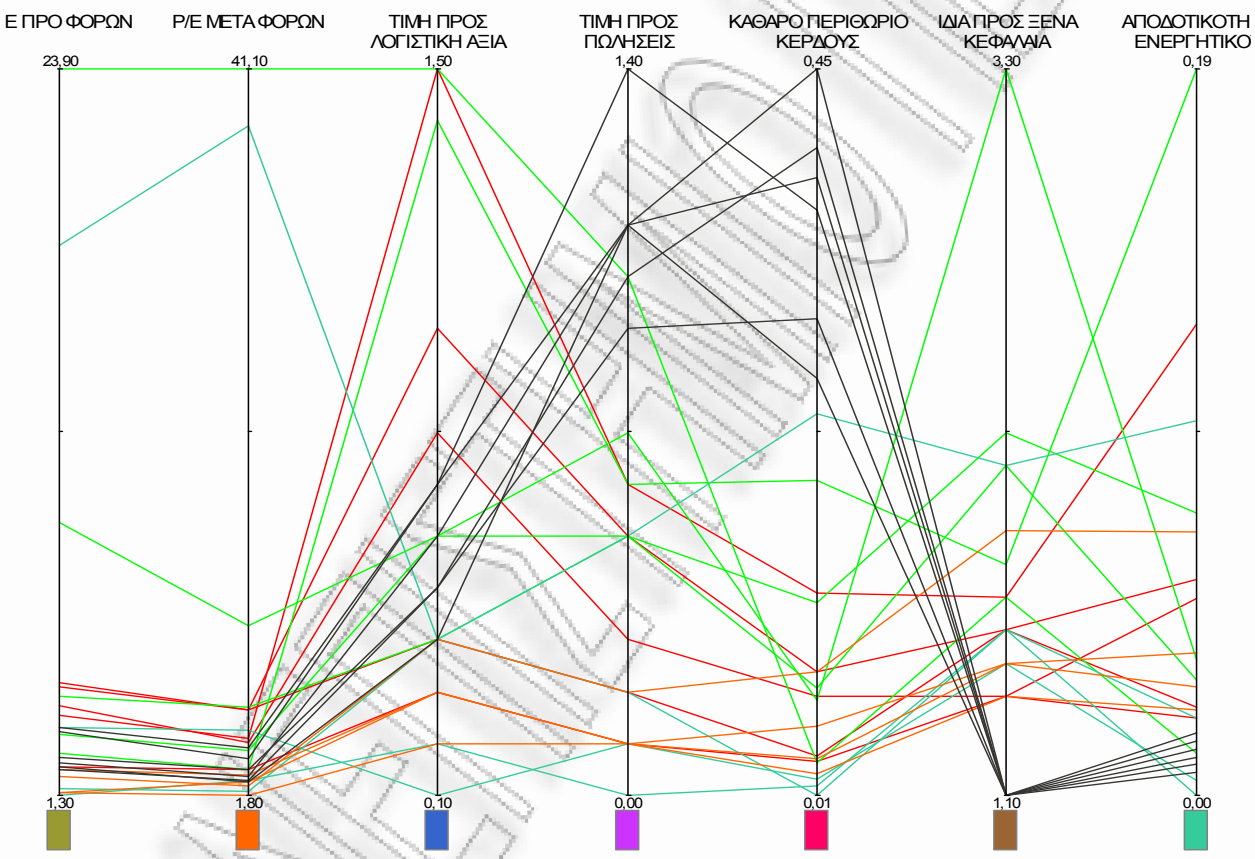
Θεωρητικά μπορούμε να απεικονίσουμε πολλές μεταβλητές σχεδιάζοντας τους αντίστοιχους άξονες αλλά πρακτικά κάτι τέτοιο δυσκολεύει την ανάγνωση του διαγράμματος, για αυτό και συνιστάται να απεικονίζονται μέχρι 10 μεταβλητές. Ακόμα, η σύγκριση των θέσεων των σημείων σε γειτονικούς άξονες μπορεί να αναδείξει πιθανές γραμμικές συσχετίσεις μεταξύ των αξόνων αλλά και να προσδιορίζει το βαθμό ανεξαρτησίας των αντίστοιχων μεταβλητών. Αν τοποθετήσουμε τους άξονες με τυχαίο τρόπο σημαίνει ότι δε δίνεται ιδιαίτερη βαρύτητα σε κάποιες μεταβλητές και επομένως θα πρέπει να αναδιατάξουμε τους άξονες με όλους τους πιθανούς τρόπους που προκύπτουν από τους συνδυασμούς των μεταβλητών, ώστε να έχουμε πλήρη εικόνα των δεδομένων. Αντί για όλη αυτή τη διαδικασία, η οποία μπορεί για μεγάλο πλήθος μεταβλητών να είναι κοπιαστική, μπορεί να προηγηθεί μια αξιολόγηση των μεταβλητών με βάση κάποιο κριτήριο σημαντικότητας και να τοποθετηθούν οι άξονες με φθίνουσα σειρά σημαντικότητας στο διάγραμμα.

Ένα σοβαρό πρόβλημα που προκύπτει κατά τη χρήση των διαγραμμάτων παράλληλων συντεταγμένων είναι όταν ο αριθμός των παρατηρήσεων είναι πολύ μεγάλος και η πυκνότητα του διαγράμματος αυξάνεται τόσο ώστε να μην ξεχωρίζουν οι γραμμές μεταξύ τους. Το πρόβλημα αυτό μπορεί να αντιμετωπισθεί με την ομαδοποίηση των παρατηρήσεων και το χρωματισμό των ομάδων που προκύπτουν με διαφορετικό χρώμα. Ακόμα μπορούν οι διάφορες ομάδες παρατηρήσεων να απεικονιστούν σε διαφορετικά διαγράμματα ή ανά ζεύγη ώστε να μειωθεί η πυκνότητα των διαγραμμάτων και να διευκολυνθεί η ανάγνωσή τους.

Αξίζει να σημειωθεί ότι το εύρος τιμών που υποδεικνύει ο κάθε άξονας μπορεί να είναι το φυσικό εύρος που έχουν τα δεδομένα ή να προηγηθεί κάποιος μετασχηματισμός των τιμών ώστε όλοι οι άξονες να έχουν κοινή διαβάθμιση τιμών. Επίσης υπάρχει η δυνατότητα οι άξονες να τοποθετηθούν σε κάθετη και όχι οριζόντια διάταξη, σε περίπτωση που κάτι τέτοιο εξυπηρετεί το σκοπό μιας συγκεκριμένης έρευνας ή υπάρχει περιορισμός στο διαθέσιμο χώρο προβολής του διαγράμματος.

Έχουμε χρησιμοποιήσει τα χρηματοοικονομικά δεδομένα των 24 εταιρειών για να κατασκευάσουμε το παρακάτω διάγραμμα παράλληλων συντεταγμένων. Κάθε κλάδος εταιρειών έχει χρωματιστεί διαφορετικά, ενώ οι άξονες έχουν τοποθετηθεί με

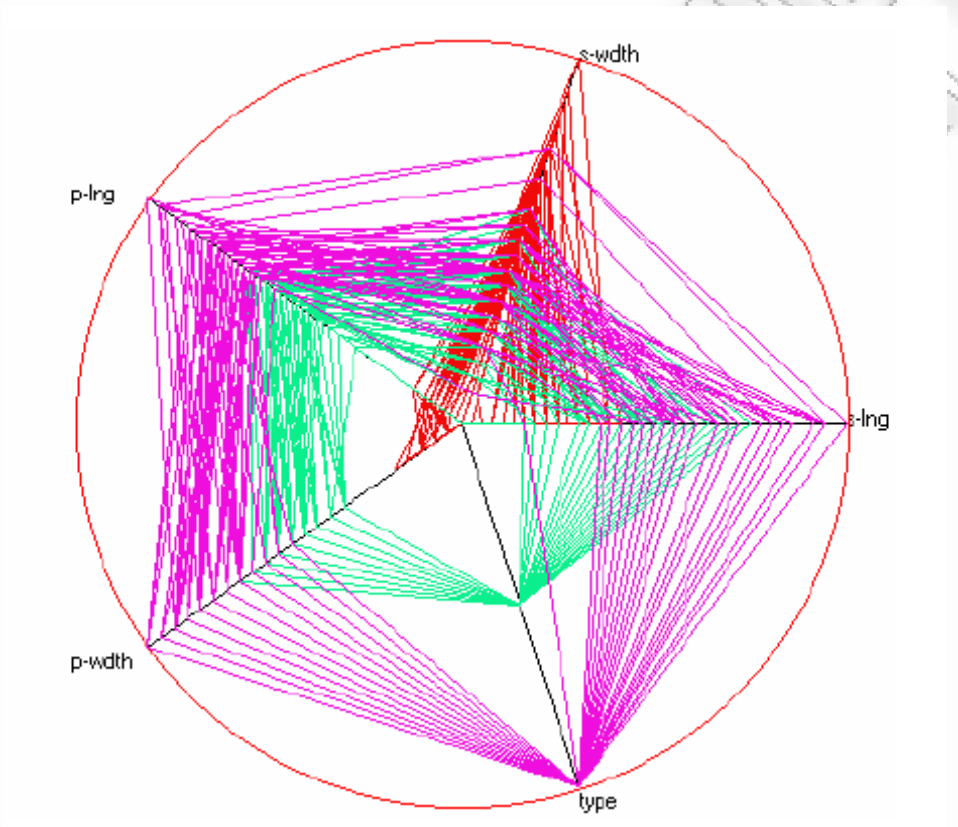
τη σειρά εμφάνισης των μεταβλητών και αντικατοπτρίζουν το φυσικό εύρος τιμών των δεδομένων. Για την κατασκευή του διαγράμματος χρησιμοποιήθηκε το λογισμικό Visulab αν και πολλά στατιστικά προγράμματα έχουν ενσωματωμένη την επιλογή κατασκευής διαγράμματος παράλληλων συντεταγμένων. Μερικά από αυτά είναι το Parallax (http://www.kdnuggets.com/software/parallax.html), το VisDB (http://www.dbs.informatik.uni-muenchen.de/dbs/projekt/visdb/visdb.html), το XmdvTool (http://davis.wpi.edu/~xmdv/), το Ggobi (http://www.ggobi.org/) και το Orange (http://www.ailab.si/orange/).



Μια παραλλαγή του διαγράμματος παράλληλων συντεταγμένων είναι το διάγραμμα κυκλικών συντεταγμένων (circular coordinate plot). Στηρίζεται στις ίδιες αρχές μόνο που τώρα οι άξονες που αντιστοιχούν στις μεταβλητές τοποθετούνται σαν ακτίνες που ξεκινούν από το κέντρο ενός κύκλου και εκτείνονται μέχρι την περίμετρο. Με τον τρόπο αυτό μπορούν να απεικονιστούν περισσότερες μεταβλητές όμως υπάρχει πιθανότητα να συγκεντρωθούν πολλά ευθύγραμμα τμήματα κοντά στο κέντρο του κύκλου, κάνοντας δύσκολη την ανάγνωση του γραφήματος. Και στην

περίπτωση αυτή μπορούν να χρησιμοποιηθούν διαφορετικοί χρωματισμοί για τις διάφορες ομάδες δεδομένων.

Πιο κάτω δίνουμε ένα παράδειγμα διαγράμματος κυκλικών συντεταγμένων για τα γνωστά Iris data, τα οποία συγκεντρώθηκαν από τον Anderson (1935) και αφορούν σε μετρήσεις 4 χαρακτηριστικών (μήκος και πλάτος πετάλου, μήκος και πλάτος σέπαλου) 3 διαφορετικών ποικιλιών ίριδας (setosa, versicolor, virginica).



5.3 Καμπύλες του Andrews

Μια πολύ εύχρηστη μέθοδος αναπαράστασης πολυδιάστατων δεδομένων προτάθηκε από τον Andrews (1972) και είναι από τις πιο διαδεδομένες ακόμα και σήμερα. Δεν απαιτεί την εξοικείωση του αναγνώστη του διαγράμματος με στατιστικές ή μαθηματικές μεθόδους αφού είναι πολύ απλή η κατασκευή των καμπυλών. Σύμφωνα με τη μέθοδο αυτή, κάθε πολυδιάστατο διάνυσμα $\mathrm{y} = \left(y_1, y_2, ..., y_q \right)^{\mathrm{T}}$ με q μεταβλητές μπορεί να αναπαρασταθεί στο επίπεδο με τη μορφή μια καμπύλης η οποία ορίζεται από την ακόλουθη συνάρτηση:

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

με $-\pi \le t \le \pi$.

Έτσι σε κάθε πολυδιάστατη παρατήρηση αντιστοιχεί και μια ξεχωριστή καμπύλη. Οι αποστάσεις μεταξύ των καμπυλών αλλά και η μορφή τους μπορούν να αναδείξουν ομαδοποιήσεις των παρατηρήσεων. Οι καμπύλες του Andrews έχουν μερικές πολύ χρήσιμες ιδιότητες, τις οποίες και αναφέρουμε παρακάτω:

α. Οι καμπύλες του Andrews διατηρούν τις αποστάσεις μεταξύ των διανυσμάτων των παρατηρήσεων. Πιο συγκεκριμένα η απόσταση μεταξύ δύο καμπυλών που ορίζονται από τις συναρτήσεις $f_x(t)$ και $f_y(t)$ που αντιστοιχούν στα διανύσματα *x* και *y* ορίζεται από τη σχέση

$$\left\| f_x(t) - f_y(t) \right\| = \sqrt{\int_{-\pi}^{\pi} \left[f_x(t) - f_y(t) \right]^2 dt}$$

και αποδεικνύεται ότι είναι ανάλογη της Ευκλείδειας απόστασης μεταξύ των διανυσμάτων *x* και *y*.

β. Η μέση τιμή μιας παρατήρησης διατηρείται και στο διάγραμμα των καμπυλών Andrews, δηλαδή, αν το διάνυσμα $\bar{y}$ αντιστοιχεί στη μέση παρατήρηση των $y_1, y_2, \ldots, y_n$, θα ισχύει η σχέση

$$f_{\bar{y}}(t) = \bar{f}_y(t) = n^{-1} \sum_{i=1}^{n} f_{y_i}(t).$$

γ. Η Ευκλείδεια απόσταση μεταξύ δύο σημείων διατηρείται και μεταξύ των καμπυλών που αντιστοιχούν σε αυτά τα σημεία. Αυτό σημαίνει ότι γειτονικά σημεία θα έχουν καμπύλες που πλησιάζουν μεταξύ τους.

δ. Η συνάρτηση που περιγράφει τις καμπύλες Andrews διατηρεί τις γραμμικές σχέσεις μεταξύ των παρατηρήσεων. Δηλαδή αν τρία σημεία *x*, *y* και *z* βρίσκονται πάνω σε μια ευθεία με το *y* ανάμεσα στα *x* και *z* τότε η καμπύλη $f_y(t)$ θα βρίσκεται ανάμεσα στις $f_x(t)$ και $f_z(t)$.

ε. Αν οι μεταβλητές είναι ασυσχέτιστες με κοινή διακύμανση σ^2, τότε οι καμπύλες Andrews διατηρούν τη διακύμανση σταθερή. Πιο συγκεκριμένα ισχύει ότι

$$Var\left(f_y(t)\right) = \sigma^2 \left(\frac{1}{2} + \sin^2 t + \cos^2 t + \sin^2 2t + \cos^2 2t + \ldots. \right).$$

Όταν ο αριθμός των διαστάσεων του διανύσματος *y* είναι περιττός, τότε η διακύμανση της αντίστοιχης καμπύλης $f_y(t)$ είναι ίση με $\sigma^2(q/2)$. Όταν ο αριθμός των διαστάσεων είναι άρτιος με $q=2r$, τότε ισχύει ότι

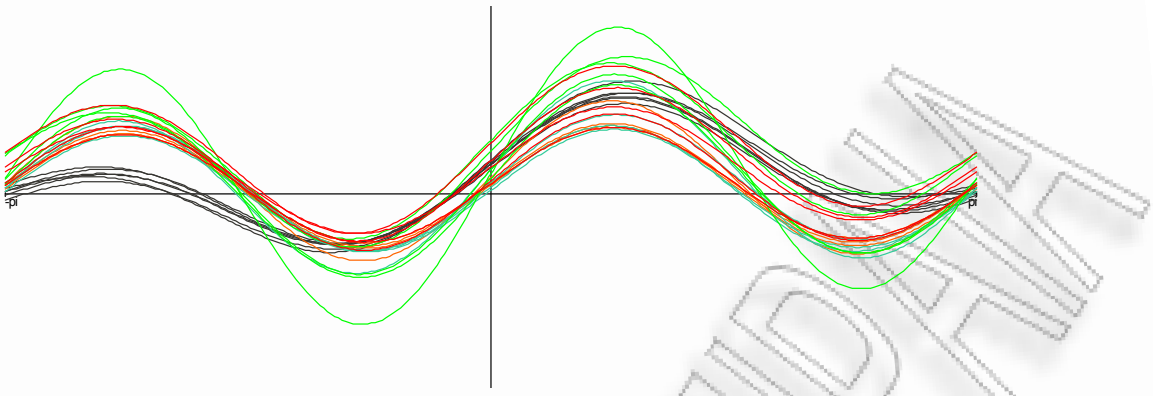
$$\sigma^2\left(r-\frac{1}{2}\right) \le Var\left(f_y(t)\right) \le \sigma^2\left(r+\frac{1}{2}\right) \text{ , με } -\pi \le t \le \pi .$$

Πρέπει να τονίσουμε πως οι παραπάνω ιδιότητες κάνουν τις καμπύλες του Andrews ιδιαίτερα χρήσιμο εργαλείο στην εξαγωγή συμπερασμάτων για τις σχέσεις των δεδομένων. Τα δεδομένα μπορούν να ομαδοποιηθούν εύκολα αφού οι καμπύλες των συσχετισμένων παρατηρήσεων θα έχουν παρόμοια μορφή. Επίσης οι ακραίες τιμές ξεχωρίζονται πολύ εύκολα εφόσον οι αντίστοιχες καμπύλες θα διαφοροποιούνται σημαντικά από τις υπόλοιπες. Επίσης, λόγω του ότι δεν υπάρχει απώλεια πληροφορίας, αυξάνεται η αντικειμενικότητα του διαγράμματος.

Ένα σημαντικό μειονέκτημα των καμπυλών Andrews είναι ότι, όταν ο αριθμός των παρατηρήσεων αυξάνεται, δυσχεραίνει η ανάγνωση του διαγράμματος. Πρακτικά η μελέτη μέχρι 20 παρατηρήσεων επιτρέπει την εξαγωγή συμπερασμάτων για τα δεδομένα. Επίσης ο βαθμός ομοιότητας μεταξύ των καμπυλών και τα κριτήρια ομαδοποίησής τους αποτελεί μια υποκειμενική διαδικασία και ποικίλει μεταξύ των ερευνητών. Τέλος, λόγω της τριγωνομετρικής φύσης της συνάρτησης που περιγράφει τις καμπύλες του Andrews, έχει μεγάλη σημασία ποιές μεταβλητές εισέρχονται πρώτες στη συνάρτηση και ποιές ακολουθούν. Καλό θα είναι επομένως να προηγείται μια αξιολόγηση της σημαντικότητας των μεταβλητών ώστε να τοποθετούνται πρώτα αυτές με τη μεγαλύτερη βαρύτητα και ακολούθως οι λιγότερο σημαντικές, εκτός της περίπτωσης βεβαίως που υπάρχει μια φυσική διάταξη των μεταβλητών.

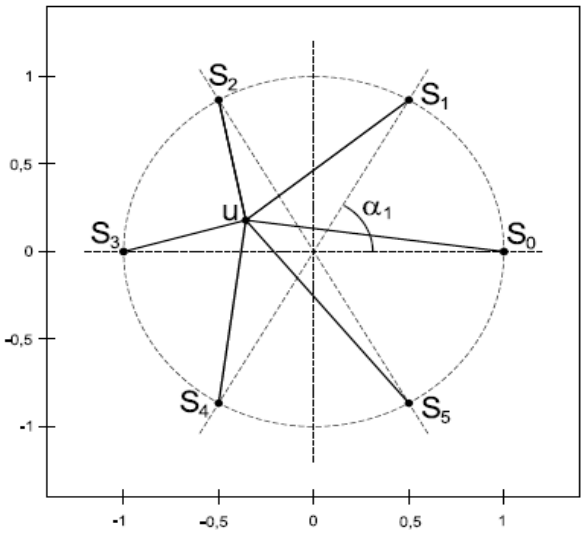
Ένα στοιχείο που μπορεί να βοηθήσει στην ανάγνωση του διαγράμματος είναι ο χρωματισμός των καμπυλών με βάση κάποιο κριτήριο όπως μια εκ των προτέρων κατηγοριοποίηση των δεδομένων. Τέλος θα πρέπει να σημειωθεί ότι έχουν προταθεί πολλές παραλλαγές και προεκτάσεις των καμπυλών του Andrews με σκοπό να αντιμετωπιστούν συγκεκριμένα προβλήματα οπτικοποίησης από τους ερευνητές.

Έχουμε κατασκευάσει στην επόμενη σελίδα ένα διάγραμμα καμπυλών του Andrews για τα δεδομένα των 24 εταιρειών χρησιμοποιώντας 5 μεταβλητές με τη βοήθεια του λογισμικού Visulab. Οι καμπύλες έχουν διαφορετικό χρώμα για κάθε κλάδο δραστηριοποίησης των εταιρειών.



5.4 Οπτικοποίηση Πολικών Συντεταγμένων

Η μέθοδος της οπτικοποίησης πολικών συντεταγμένων παρουσιάστηκε από τους Hoffman, Grinstein, Marx, Grosse & Stanley (1997) και χρησιμοποιεί το νόμο του Hook περί επιμήκυνσης και συσπείρωσης ελατηρίου από τη Φυσική για να οπτικοποιήσει πολυδιάστατες παρατηρήσεις στο επίπεδο. Η κατασκευή του γραφήματος γίνεται ως εξής: οι μεταβλητές τοποθετούνται περιμετρικά επάνω σε έναν κύκλο και σε ίσες αποστάσεις μεταξύ τους. Στη συνέχεια κάθε πολυμεταβλητή παρατήρηση τοποθετείται στο εσωτερικό του κύκλου, υποθέτοντας ότι από τα περιμετρικά σημεία που αντιστοιχούν στις μεταβλητές την «τραβούν ελατήρια» με δύναμη ανάλογη της τιμής της παρατήρησης στην κάθε μεταβλητή. Η τελική θέση της παρατήρησης θα είναι το σημείο εξισορρόπησης όλων των δυνάμεων που ασκούνται σε αυτήν από τα «ελατήρια» των μεταβλητών. Προφανώς η παρατήρηση θα τείνει να τοποθετηθεί πιο κοντά σε κάποια μεταβλητή για την οποία έχει υψηλότερη τιμή. Στο διπλανό σχήμα δίνεται σχηματικά ο τρόπος με τον οποίο λειτουργεί το γράφημα της οπτικοποίησης πολικών συντεταγμένων.



Η μέθοδος της οπτικοποίησης πολικών συντεταγμένων είναι ιδιαίτερα χρήσιμο εργαλείο

στην ανεύρεση ομαδοποιήσεων μεταξύ των πολυμεταβλητών δεδομένων αν και παρουσιάζει ένα σοβαρό μειονέκτημα: επειδή για την τελική θέση της κάθε παρατήρησης χρησιμοποιείται η εξισορρόπηση των διαφόρων δυνάμεων από τις μεταβλητές, είναι πολύ πιθανό παρατηρήσεις με διαφορετικές τιμές να βρεθούν στο ίδιο ακριβώς σημείο πάνω στο γράφημα και κατά συνέπεια να μη φανούν κάποιες ομαδοποιήσεις που υπάρχουν στο σύνολο δεδομένων. Ακόμα είναι πιθανό μια παρατήρηση με υψηλές τιμές για δύο μεταβλητές οι οποίες είναι τοποθετημένες απέναντι η μια από την άλλη στο γράφημα, να ισορροπήσει στο κέντρο του γραφήματος, αποκρύπτοντας έτσι το μέγεθος των τιμών των μεταβλητών και τελικά να εμφανίζεται ως μια παρατήρηση με ίσες τιμές για όλες τις μεταβλητές.

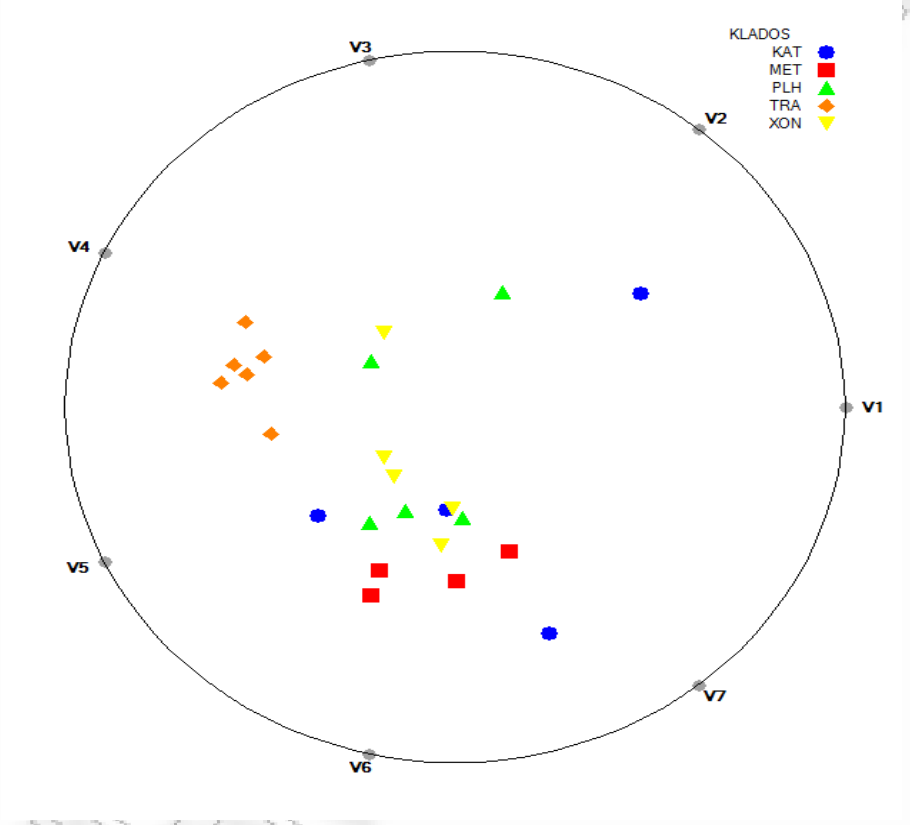
Πολύ σημαντικός παράγοντας στην κατασκευή ενός διαγράμματος οπτικοποίησης πολικών συντεταγμένων είναι και η διάταξη των μεταβλητών στην περίμετρο του κύκλου. Μια αναδιάταξη των σημείων που αντιστοιχούν στις μεταβλητές μπορεί να αλλάξει κατά πολύ τη μορφή του γραφήματος. Αυτό που συνιστάται πάντως ώστε να έχουμε πλήρη εικόνα των δεδομένων είναι να κατασκευαστούν και να μελετηθούν όλες οι πιθανές διατάξεις των μεταβλητών στην περίμετρο του κύκλου, κάτι ιδιαίτερα χρονοβόρο στην περίπτωση ύπαρξης πολλών μεταβλητών. Ακόμα, συνηθίζεται οι τιμές των μεταβλητών να μετασχηματίζονται σε κλίμακα από 0 μέχρι 1 πριν τοποθετηθούν στο γράφημα. Συμβαίνει όμως κάποιες φορές, ανάλογα με τα δεδομένα, να χρησιμοποιείται διαφορετική κλίμακα ώστε να γίνουν πιο ευκρινείς οι αποστάσεις μεταξύ των σημείων.

Μια ακόμα παράμετρος που πρέπει να ληφθεί υπόψη είναι και ο αριθμός των παρατηρήσεων που θέλουμε να οπτικοποιήσουμε. Υπάρχει περίπτωση όταν είναι μεγάλο το πλήθος, να συγκεντρωθούν τόσα πολλά σημεία στο εσωτερικό του κύκλου ώστε να μην είναι δυνατή πλέον η ανάγνωση του γραφήματος. Μερικές λύσεις σε αυτό το πρόβλημα είναι ο χρωματισμός ή η σήμανση των παρατηρήσεων με βάση κάποια κατηγορική μεταβλητή ή κάποια άλλη ομαδοποίηση που μπορεί να υπάρχει στα δεδομένα. Αν σε κάθε ομάδα δοθεί διαφορετικό χρώμα ή σχήμα, θα διευκολυνθεί αρκετά ο αναγνώστης του γραφήματος ώστε να διακρίνει τις παρατηρήσεις.

Τέλος, ενώ θεωρητικά ο αριθμός των μεταβλητών που μπορούμε να αναπαραστήσουμε είναι άπειρος, πρακτικά δε μπορεί να είναι πολύ μεγάλος για να μη μειωθεί η αποτελεσματικότητα του γραφήματος. Έτσι, σε περιπτώσεις πολύ μεγάλου αριθμού μεταβλητών συνιστάται η επιλογή των σημαντικότερων μέσα από μια

διαδικασία επιλογής η οποία θα καθορίζεται από την φύση των δεδομένων αλλά και το περιεχόμενο της έρευνας.

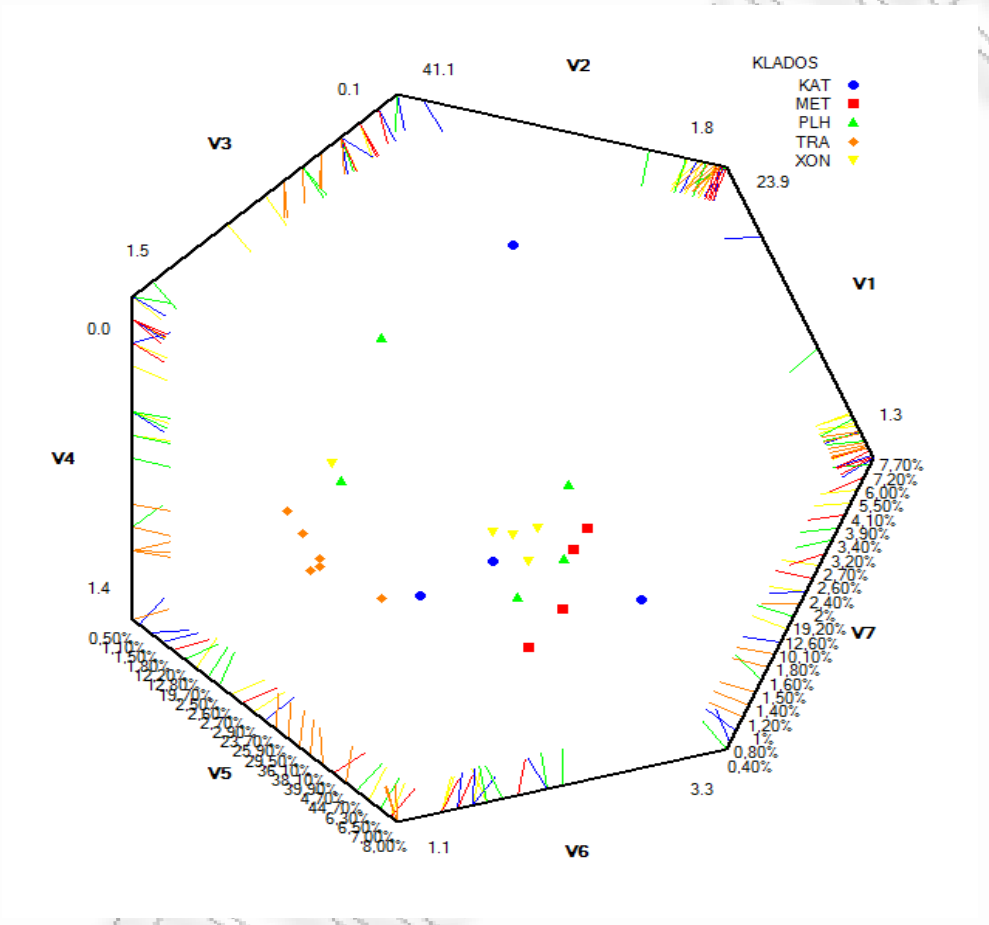
Δίνεται στη συνέχεια ένα παράδειγμα εφαρμογής της μεθόδου οπτικοποίησης πολικών συντεταγμένων στα χρηματοοικονομικά δεδομένα των 24 εταιρειών, το οποίο κατασκευάστηκε με το λογισμικό Orange. Για κάθε κλάδο δραστηριότητας έχει χρησιμοποιηθεί διαφορετικό χρώμα και σχήμα στις παρατηρήσεις. Μπορούμε εύκολα να παρατηρήσουμε την ομαδοποίηση που έχει επιτευχθεί κυρίως στις τράπεζες αλλά και στις εταιρείες μετάλλων.



Μια παραλλαγή της μεθόδου οπτικοποίησης πολικών συντεταγμένων είναι η μέθοδος της οπτικοποίησης πολυωνυμικών συντεταγμένων (polynomial coordinate visualization – PolyViz). Στη μέθοδο αυτή οι μεταβλητές δεν τοποθετούνται στην περίμετρο ενός κύκλου, αλλά στις πλευρές ενός πολυγώνου. Προφανώς ο αριθμός των μεταβλητών θα ορίζει και το είδος του πολυγώνου που χρησιμοποιούμε.

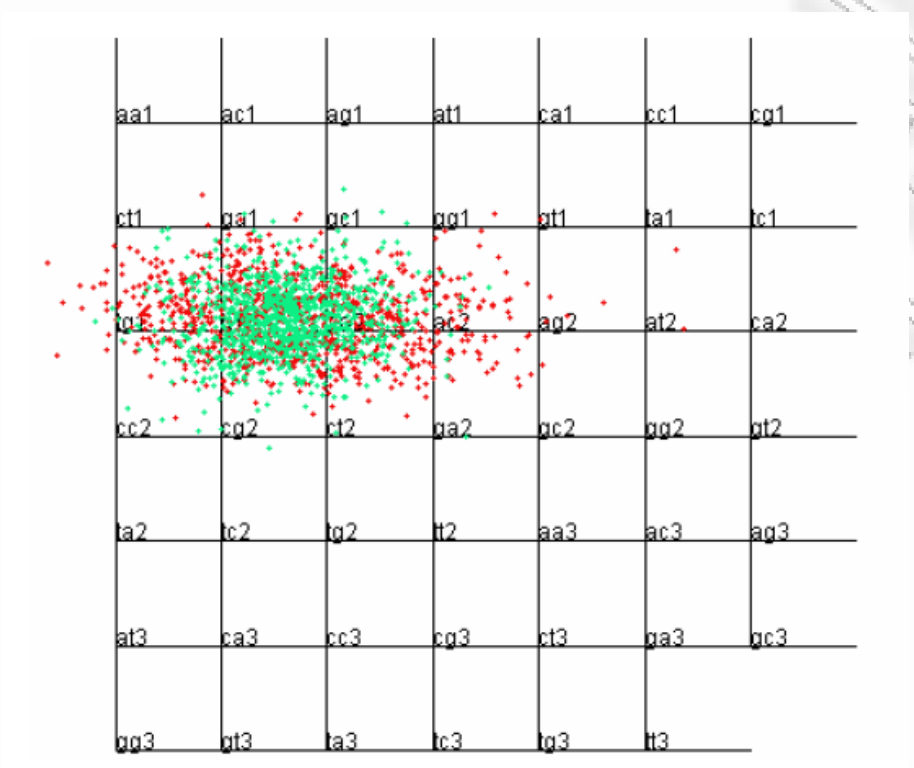
Το πλεονέκτημα αυτής της μεθόδου έναντι της οπτικοποίησης πολικών συντεταγμένων είναι ότι οι πλευρές του πολυγώνου μπορούν να χρησιμοποιηθούν ως άξονες που περιέχουν όλο το εύρος τιμών της κάθε μεταβλητής. Με αυτόν τον τρόπο ελαχιστοποιείται η πιθανότητα να συμπίπτουν στο ίδιο σημείο παρατηρήσεις με

αρκετά διαφορετικές τιμές, όπως συμβαίνει στην οπτικοποίηση πολικών συντεταγμένων. Δίνεται στη συνέχεια ένα παράδειγμα γραφήματος πολυωνυμικών συντεταγμένων στα χρηματοοικονομικά δεδομένα των 24 εταιρειών, το οποίο κατασκευάστηκε με το λογισμικό Orange. Για κάθε κλάδο δραστηριότητας έχει χρησιμοποιηθεί διαφορετικό χρώμα και σχήμα στις παρατηρήσεις.



Μια ακόμα παραλλαγή της οπτικοποίησης πολικών συντεταγμένων είναι η οπτικοποίηση πλέγματος (grid visualization – GridViz), στην οποία οι μεταβλητές απεικονίζονται ως σημεία ενός ορθογώνιου πλέγματος και όχι ως περιμετρικά σημεία ενός κύκλου. Με τον τρόπο αυτό μπορούμε να αυξήσουμε σημαντικά τον αριθμό των μεταβλητών που εισέρχονται στο γράφημα, σε σχέση με τη χρήση της οπτικοποίησης πολικών συντεταγμένων. Και σε αυτή την περίπτωση πάντως η τοποθέτηση των παρατηρήσεων γίνεται με τον ίδιο τρόπο με την οπτικοποίηση πολικών συντεταγμένων, δηλαδή η τελική θέση κάθε παρατήρησης προκύπτει από την εξισορρόπηση των δυνάμεων που ασκούνται από τα σημεία – μεταβλητές. Δίνουμε παρακάτω ένα παράδειγμα οπτικοποίησης πλέγματος για δεδομένα γενετικής 48

μεταβλητών το οποίο μπορεί να κατασκευαστεί με το RapidMiner (http://rapid-i.com/).



5.5 Διαγράμματα hammock

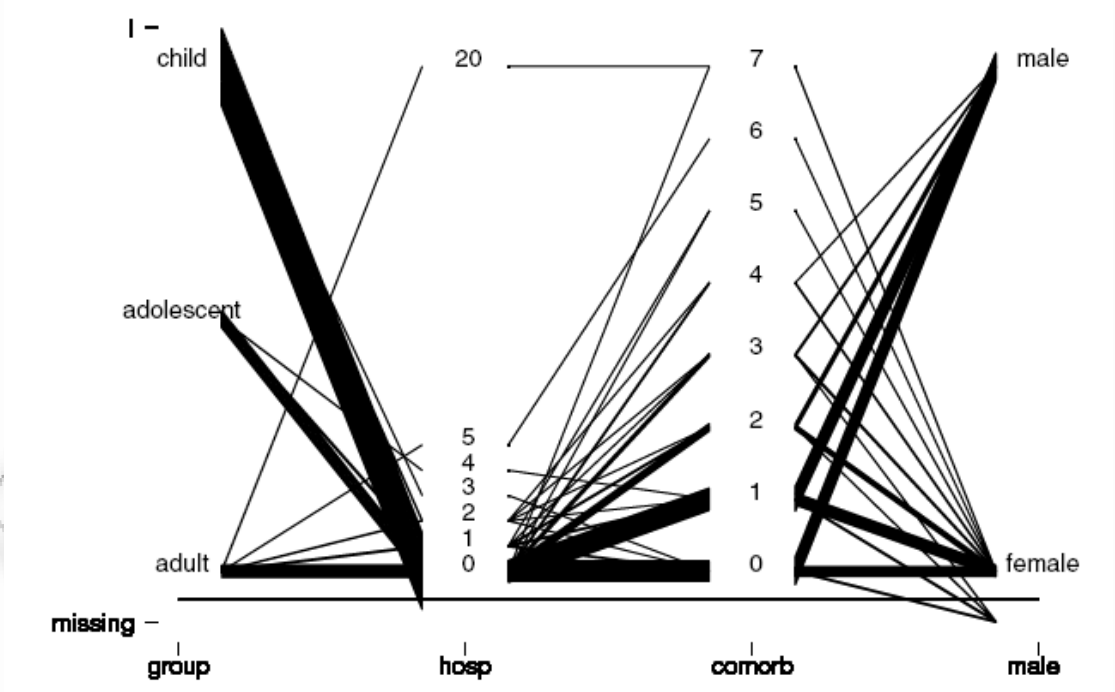
Η μέθοδος αυτή η οποία είναι παραλλαγή του διαγράμματος παράλληλων συντεταγμένων εισήχθηκε από τον Schonlau (2003) με στόχο την ταυτόχρονη παρουσίαση συνεχών και κατηγορικών πολυδιάστατων δεδομένων. Όπως και στο διάγραμμα παράλληλων συντεταγμένων, οι μεταβλητές αναπαρίστανται ως παράλληλοι άξονες και οι τιμές των παρατηρήσεων τοποθετούνται πάνω στους άξονες αυτούς. Σε κάθε άξονα χρησιμοποιείται το φυσικό εύρος τιμών της αντίστοιχης μεταβλητής, ενώ κάτω από το διάγραμμα μπορεί να τοποθετηθεί μια ξεχωριστή κατηγορία για τις ελλείπουσες τιμές, η οποία διαχωρίζεται από τις υπόλοιπες με μια οριζόντια γραμμή.

Σε αντίθεση με το διάγραμμα παράλληλων συντεταγμένων, στο διάγραμμα hammock τα σημεία στους άξονες ενώνονται μεταξύ τους όχι με μια τεθλασμένη γραμμή αλλά με παραλληλόγραμμα, των οποίων το πλάτος αντιστοιχεί στο πλήθος

των παρατηρήσεων με τη συγκεκριμένη τιμή για τη μεταβλητή. Έτσι αν σε κάποια μεταβλητή υπάρχει μια μόνο παρατήρηση με την σχετική τιμή, το παραλληλόγραμμο θα έχει το ελάχιστο πλάτος και θα εκφυλισθεί σε ευθεία. Το γεγονός αυτό βοηθάει και στον εντοπισμό ακραίων τιμών, οι οποίες αναμένουμε να συνδέονται με παραλληλόγραμμα μικρού πλάτους αλλά και να διαφέρουν σχηματικά από τις υπόλοιπες παρατηρήσεις.

Όπως και στα διαγράμματα παράλληλων συντεταγμένων, υπάρχει περιορισμός στον αριθμό των μεταβλητών που μπορούμε να αναπαραστήσουμε ώστε το διάγραμμα να μη γίνεται πολύπλοκο και δυσανάγνωστο. Επίσης θα πρέπει να αναδιατάσσονται οι άξονες ώστε να έχουμε πλήρη εικόνα των σχέσεων μεταξύ των μεταβλητών. Τέλος υπάρχει η δυνατότητα να χρησιμοποιηθούν χρωματισμοί με σκοπό να διευκολυνθεί ο αναγνώστης του διαγράμματος, ειδικά όταν ο όγκος των δεδομένων είναι πολύ μεγάλος.

Παρουσιάζουμε στη συνέχεια ένα χαρακτηριστικό δείγμα διαγράμματος hammock για δεδομένα με συνεχείς και κατηγορικές μεταβλητές. Διαγράμματα hammock μπορούν να κατασκευαστούν με το Stata (http://www.stata.com/) και με την R.



5.6 Διαγράμματα Μωσαϊκού

Το διάγραμμα μωσαϊκού παρουσιάστηκε από τους Hartigan & Kleiner (1981), εξελίχθηκε από τον Friendly (1994) και πρόκειται για την γραφική απεικόνιση ενός πίνακα συχνοτήτων με δύο ή περισσότερες κατηγορικές μεταβλητές, χωρίς μάλιστα να έχουμε απώλεια πληροφορίας των δεδομένων. Η κατασκευή του στηρίζεται στο γεγονός ότι η συχνότητα κάθε κελιού ενός πίνακα συχνοτήτων μπορεί να αναπαρασταθεί με ένα ορθογώνιο παραλληλόγραμμο. Κατά συνέπεια, όσο μεγαλύτερη είναι η παρατηρούμενη συχνότητα του κελιού τόσο μεγαλύτερο θα είναι το εμβαδόν του αντίστοιχου παραλληλόγραμμου.

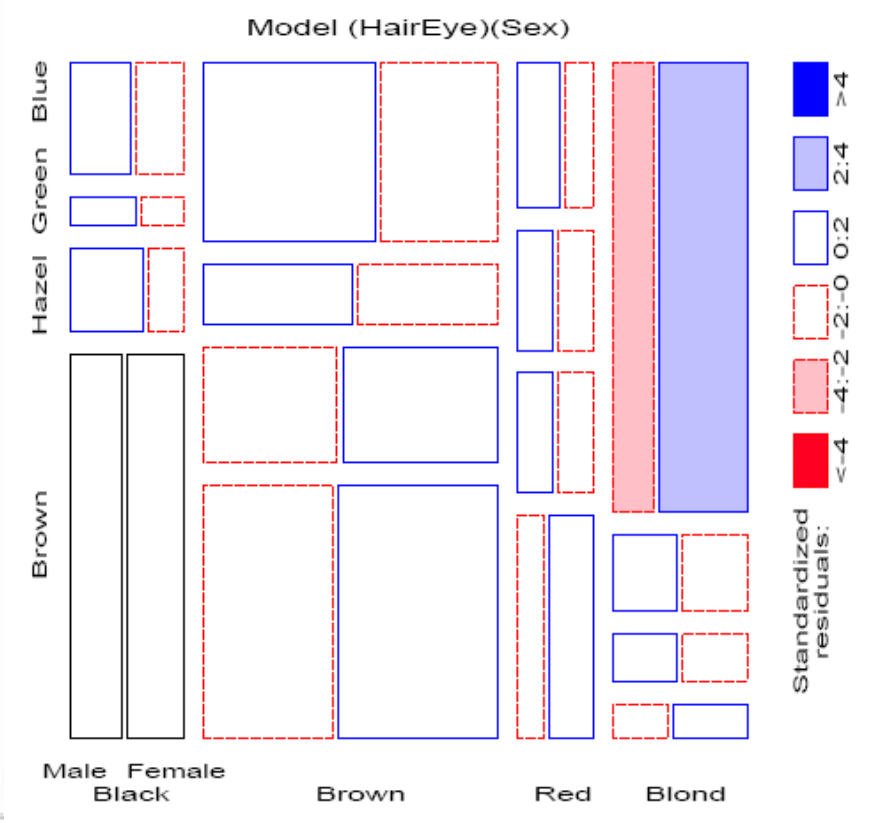
Το διάγραμμα μωσαϊκού διευκολύνει τον εντοπισμό υπερβολικά μικρών ή μεγάλων συχνοτήτων αλλά και σχέσεων μεταξύ των μεταβλητών. Η χρήση του διαγράμματος γενικεύεται από την περίπτωση των δύο μεταβλητών και στην πολυμεταβλητή περίπτωση, διαιρώντας τα παραλληλόγραμμα τόσες φορές, όσες είναι και οι επιπλέον μεταβλητές που εισέρχονται σε αυτό. Ο επιμερισμός αυτός θεωρητικά μπορεί να συνεχίζεται επ' άπειρον αλλά πρακτικά δεν συνιστάται η αναπαράσταση υπερβολικά μεγάλου αριθμού μεταβλητών. Στην περίπτωση χρήσης περισσότερων των δύο μεταβλητών, τα κενά ανάμεσα στα παραλληλόγραμμα κάθε επιπέδου είναι μεγαλύτερα ώστε να υπογραμμίζεται η κατηγοριοποίηση των δεδομένων και να διευκολύνεται ο αναγνώστης του γραφήματος.

Η αποτελεσματικότητα του διαγράμματος μωσαϊκού μπορεί να βελτιωθεί σημαντικά αν χρησιμοποιηθεί σκίαση και χρωματισμός των παραλληλογράμμων για να τονιστούν κάποια επιπλέον χαρακτηριστικά των δεδομένων. Επίσης συνιστάται η αναδιάταξη των παραλληλογράμμων για διάφορους συνδυασμούς των μεταβλητών με σκοπό να γίνεται ευκολότερος ο εντοπισμός συσχετίσεων μεταξύ των μεταβλητών.

Είναι προφανές ότι τα παραλληλόγραμμα των συχνοτήτων θα ευθυγραμμίζονται στο διάγραμμα όταν οι μεταβλητές είναι στατιστικά ανεξάρτητες. Τα διαγράμματα μωσαϊκού, εκτός από τις πιθανές σχετίσεις ή ανεξαρτησίες μεταξύ των μεταβλητών που μπορούν να αναδείξουν, συμβάλλουν ως τεχνική και στην κατασκευή λογαριθμογραμμικών μοντέλων, αναπαριστώντας τα κατάλοιπα του υποδείγματος. Πολλές φορές μάλιστα αναπαρίστανται με χρήση διαγραμμάτων μωσαϊκού και τα τυποποιημένα κατάλοιπα του Pearson από την υπόθεση της ανεξαρτησίας, ώστε να εξυπηρετούνται δύο σκοποί ταυτόχρονα: τόσο η οπτική

επισκόπηση των δεδομένων μέσω των συχνοτήτων όσο και η καταλληλότητα ενός δεδομένου υποδείγματος μέσω των καταλοίπων.

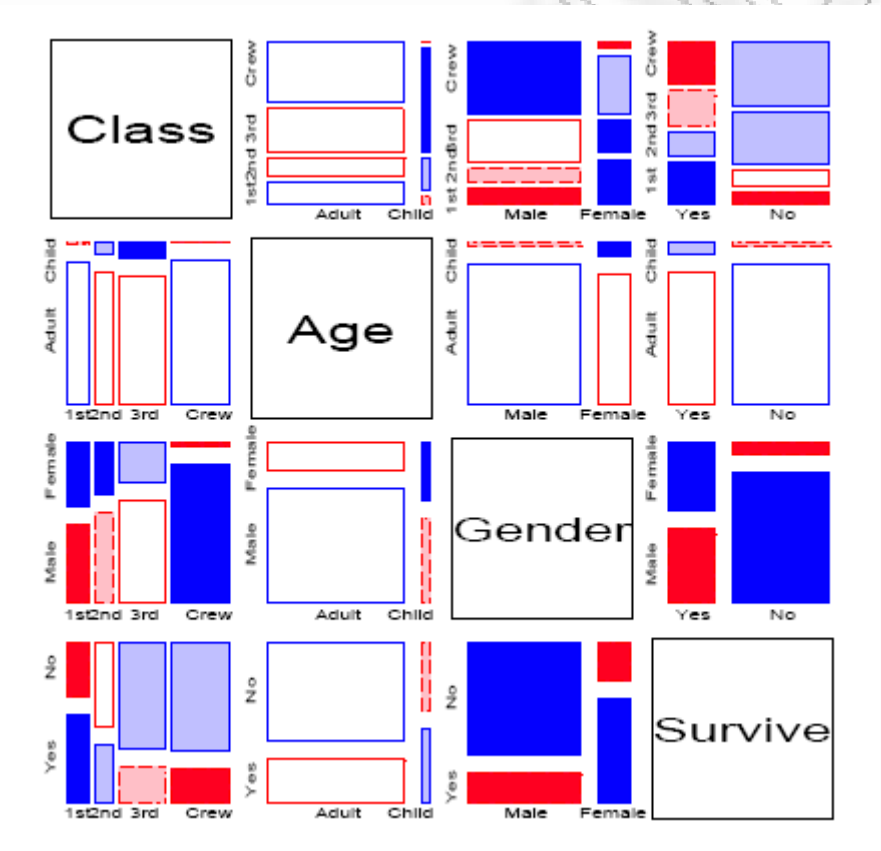
Ακολουθεί ένα χαρακτηριστικό παράδειγμα διαγράμματος μωσαϊκού όπου υπάρχουν και τα τυποποιημένα κατάλοιπα του Pearson. Τα δεδομένα αφορούν τους επιβαίνοντες στον Τιτανικό (http://www.encyclopedia-titanica.org/) και έχουν χρησιμοποιηθεί τρείς κατηγορικές μεταβλητές ενώ τα κατάλοιπα παίρνουν μια απόχρωση σύμφωνα με τη διαβάθμιση τιμών που υπάρχει στο πλαϊνό μέρος του διαγράμματος. Διαγράμματα μωσαϊκού μπορούν να κατασκευαστούν με τα προγράμματα Orange και Statgraphics (http://www.statgraphics.com/).



Πολλές φορές μας ενδιαφέρει να δούμε όλους τους ανά δύο συνδυασμούς των κατηγορικών μεταβλητών με τις αντίστοιχες περιθώριες συχνότητες και να εντοπίσουμε στους συνδυασμούς αυτούς τις αλληλεπιδράσεις ή τις συνθήκες ανεξαρτησίας που ισχύουν. Για το σκοπό αυτό υπάρχει ο πίνακας διαγραμμάτων μωσαϊκού που είναι ο αντίστοιχος του πίνακα διαγραμμάτων διασποράς μόνο που στη θέση των διαγραμμάτων διασποράς τοποθετούνται τα διαγράμματα μωσαϊκού για διάφορα ζεύγη κατηγορικών μεταβλητών.

Όπως και στον πίνακα διαγραμμάτων διασποράς, έτσι και στον πίνακα διαγραμμάτων μωσαϊκού θεωρητικά μπορούμε να αναπαραστήσουμε όσες μεταβλητές επιθυμούμε αλλά πρακτικά χρησιμοποιούμε λόγω περιορισμένου χώρου προβολής, 3 ή το πολύ 4 μεταβλητές.

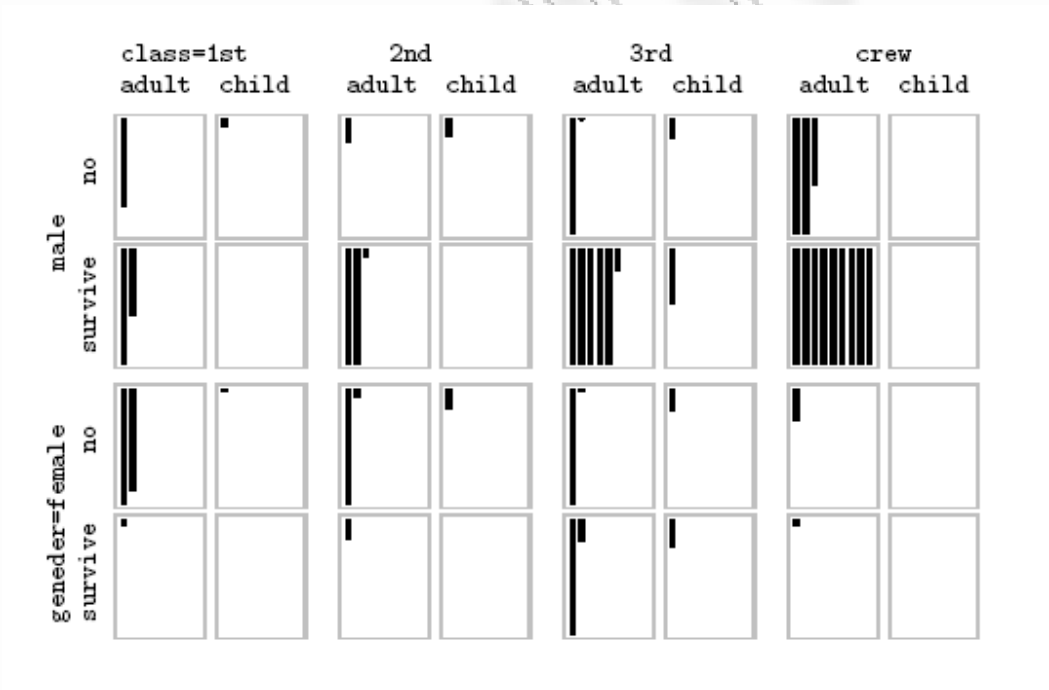
Ακολουθεί ένα χαρακτηριστικό παράδειγμα πίνακα διαγραμμάτων μωσαϊκού με εφαρμογή των δεδομένων των επιβαινόντων στον Τιτανικό και χρήση 4 κατηγορικών μεταβλητών. Ο χρωματισμός των παραλληλογράμμων γίνεται με την ίδια λογική που ίσχυε και στο προηγούμενο διάγραμμα μωσαϊκού και υποδηλώνει τις τιμές των τυποποιημένων καταλοίπων.



Μια επιπλέον παραλλαγή του διαγράμματος μωσαϊκού που έχει προταθεί από τον Moon (2004) είναι το λεγόμενο διάγραμμα ευθειών μωσαϊκού (line mosaic plot). Εδώ οι συχνότητες των κελιών δεν οπτικοποιούνται μέσω παραλληλογράμμων αλλά μέσω γραμμών, το ύψος των οποίων αντικατοπτρίζει και το μέγεθος των αντίστοιχων συχνοτήτων. Όλα τα πεδία τα οποία περιέχουν τις γραμμές είναι ίδιου μεγέθους και διαχωρίζονται από οριζόντιους και κάθετους άξονες ώστε να διευκολύνεται η αντίληψη για το μέγεθος των γραμμών.

Υπάρχουν δύο τρόποι για να κατασκευαστεί το διάγραμμα ευθειών μωσαϊκού: ο πρώτος είναι να πάρουμε ως σημείο αναφοράς τις συχνότητες μιας μεταβλητής και στη συνέχεια να ακολουθήσουμε ανάλογη διαδικασία για τις υπόλοιπες μεταβλητές. Ο δεύτερος είναι μέσω μιας περιοδικά επαναλαμβανόμενης διαδικασίας. Και σε αυτή την παραλλαγή μπορούμε να χρησιμοποιήσουμε χρώματα για να τονίσουμε κάποια επιπλέον χαρακτηριστικά των δεδομένων.

Με τη μέθοδο του διαγράμματος ευθειών μωσαϊκού προκύπτει ένα πλεονέκτημα έναντι του διαγράμματος μωσαϊκού, ότι αυξάνεται ο αριθμός των μεταβλητών που μπορούμε να αναπαραστήσουμε μέσα από το διάγραμμα ενώ παράλληλα απλοποιείται και η ανάγνωση των δεδομένων λόγω της δομής του διαγράμματος, χωρίς να απαιτείται ιδιαίτερη εξοικείωση του αναγνώστη.
Πιο κάτω δίνουμε ένα χαρακτηριστικό παράδειγμα διαγράμματος ευθειών μωσαϊκού με βάση τα δεδομένα των επιβαινόντων στον Τιτανικό.



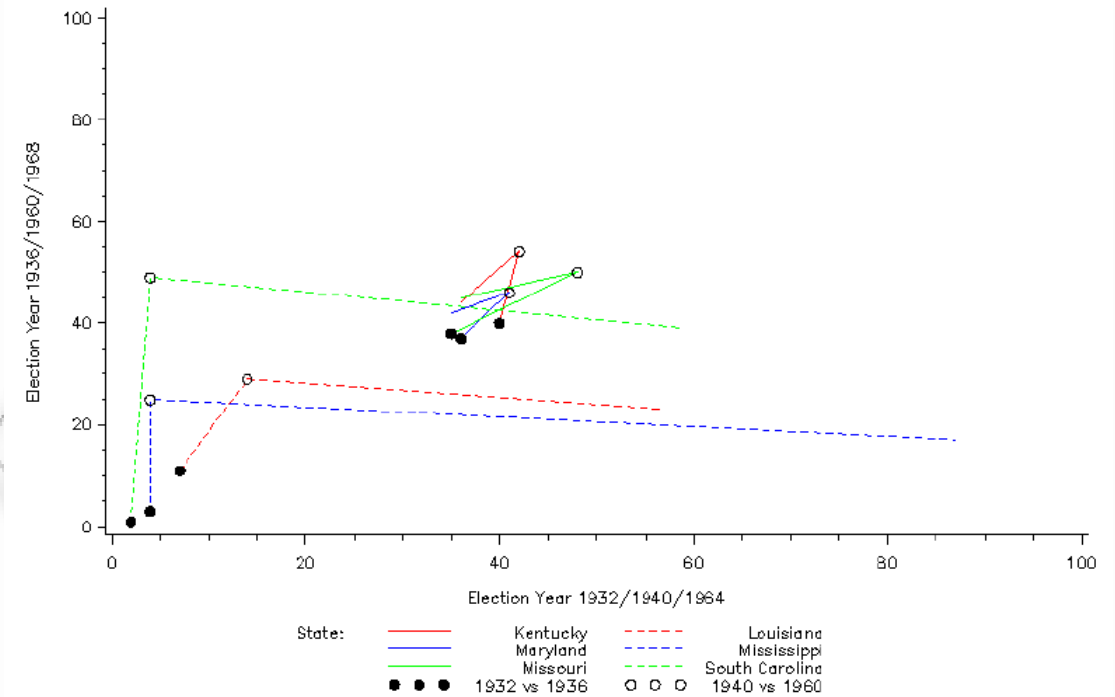
5.7 Πτερυγογράμματα

Η μέθοδος αυτή εισήχθηκε από του Schwenke & Fergen (2002) και στηρίζεται στο απλό διδιάστατο διάγραμμα διασποράς, με τη διαφορά ότι επιτρέπει την επαναχρησιμοποίηση των αξόνων για να συγκριθούν και άλλα ζεύγη μεταβλητών εκτός των δυο αρχικών. Λόγω του ότι χρησιμοποιούνται οι ίδιοι άξονες για

περισσότερες μεταβλητές, θα πρέπει οι μεταβλητές να έχουν ίδια κλίμακα μέτρησης είτε να τυποποιούνται οι τιμές τους σε μια κοινή κλίμακα. Όταν τελειώσουν οι συγκρίσεις μεταξύ των ζευγών μεταβλητών, τα σημεία ενώνονται με γραμμές και προκύπτει το τελικό διάγραμμα. Ο τρόπος με τον οποίο συνενώνονται οι γραμμές έχει δώσει και το όνομα πτερυγόγραμμα (pinion plot) καθώς το σχήμα θυμίζει πτερύγια πτηνών.

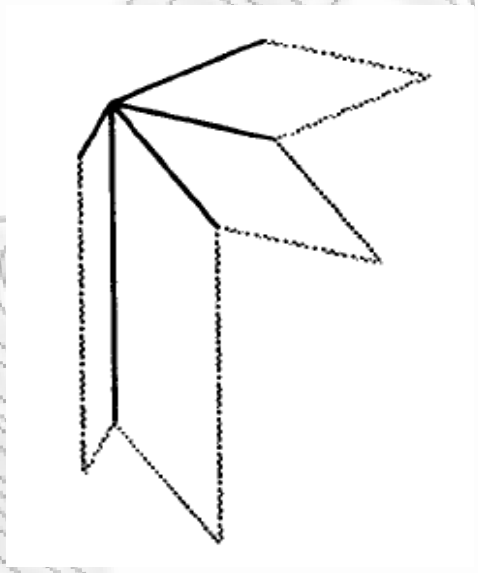
Η μορφή και η ομαδοποίηση των γραμμών που σχηματίζονται μας δίνει μια εικόνα για τη συμπεριφορά των δεδομένων. Για παράδειγμα μια συγκέντρωση γραμμών με όμοια χαρακτηριστικά θα υποδηλώνει ομοιόμορφη συμπεριφορά από τις αντίστοιχες μεταβλητές. Είναι συνήθης τακτική να χρησιμοποιούνται διάφορα σύμβολα για να σημειώνονται τα σημεία ώστε να μπορούμε να ξεχωρίζουμε ποιά συνδέονται με πιο ζευγάρι μεταβλητών. Είναι προφανές ότι ο τρόπος που γίνεται το «ζευγάρωμα» των μεταβλητών επηρεάζει την τελική εικόνα του διαγράμματος. Αν αλλάξουμε τα ζεύγη των μεταβλητών μπορεί το αποτέλεσμα να είναι τελείως διαφορετικό, για αυτό και πρέπει ο ερευνητής να δοκιμάζει, αν είναι δυνατόν σε ένα διαδραστικό περιβάλλον, διάφορους συνδυασμούς των μεταβλητών ώστε να καταλήξει στο πιο χρήσιμο οπτικό αποτέλεσμα.

Δίνουμε παρακάτω ένα χαρακτηριστικό παράδειγμα πτερυγογράμματος που έχει κατασκευαστεί με το λογισμικό SAS και αφορά σε δεδομένα εκλογικών αναμετρήσεων σε 6 αμερικάνικες πολιτείες .



5.8 Υπερθηκόγραμμα

Η μέθοδος του υπερθηκογράμματος εισήχθηκε από τους Alpern & Carter (1991). Το Hyperbox είναι η διδιάστατη απεικόνιση ενός πολυδιάστατου ορθογώνιου παραλληλεπίπεδου. Το κύριο χαρακτηριστικό της μεθόδου είναι ότι περιέχει ένα παραλληλόγραμμο για κάθε ζεύγος μεταβλητών, επιτρέποντας με αυτό τον τρόπο την ταυτόχρονη απεικόνιση όλων των ανά δύο σχέσεων ενός πολυμεταβλητού συνόλου δεδομένων.

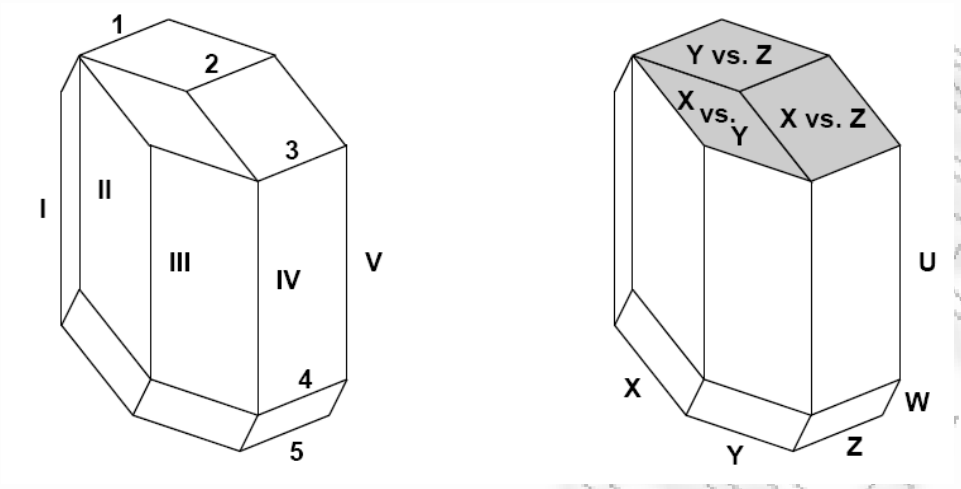


Η κατασκευή του υπερθηκογράμματος είναι πολύ απλή: για *n* μεταβλητές, σχεδιάζουμε *n* ευθύγραμμα τμήματα από ένα κοινό σημείο και μέσα σε μια γωνία 180 μοιρών το μέγιστο, όπως φαίνεται και στο διπλανό σχήμα.

Το μήκος των τμημάτων αυτών είτε επιλέγεται αυθαίρετα είτε επιλέγεται με τέτοιο τρόπο ώστε να περιέχει κάποια πληροφόρηση, ενώ το ίδιο συμβαίνει και με τις γωνίες που σχηματίζονται μεταξύ των ευθύγραμμων τμημάτων. Στη συνέχεια σχεδιάζουμε παραλληλόγραμμα για κάθε ζευγάρι παρακείμενων τμημάτων όπως φαίνεται και στο σχήμα που ακολουθεί, και προχωράμε με αυτό τον τρόπο μέχρι να μην είναι δυνατόν να κατασκευαστούν άλλα παραλληλόγραμμα.

Το υπερθηκόγραμμα που θα σχηματιστεί, για τις *n* μεταβλητές θα έχει n^2 ευθύγραμμα τμήματα ενώ θα περιέχει συνολικά $n(n-1)/2$ παραλληλόγραμμα. Για κάθε ευθύγραμμο τμήμα στο υπερθηκόγραμμα υπάρχουν *n-1* ακόμα ευθύγραμμα τμήματα που έχουν το ίδιο μήκος και την ίδια διεύθυνση. Τα ομοειδή αυτά ευθύγραμμα τμήματα σχηματίζουν ένα σύνολο διεύθυνσης και συνήθως η κάθε μεταβλητή αντιστοιχίζεται σε μια διεύθυνση. Υπάρχει η δυνατότητα να «κοπεί» το υπερθηκόγραμμα κατά μήκος κάποιας από τις διευθύνσεις του ώστε να επιτρέψει την αναλυτικότερη μελέτη συγκεκριμένων μεταβλητών, γεγονός ιδιαίτερα χρήσιμο σε περιπτώσεις δεδομένων από χρονοσειρές.

Ακολουθεί ένα παράδειγμα κατασκευής υπερθηκογράμματος με 5 μεταβλητές όπως και ένας πιθανός τρόπος αντιστοίχισης των μεταβλητών στις διευθύνσεις.



5.9 Υφαντόγραμμα

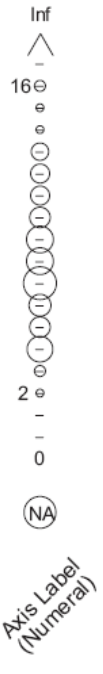
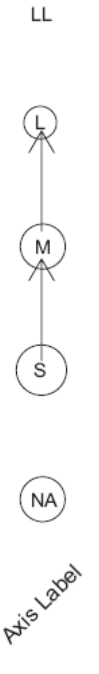
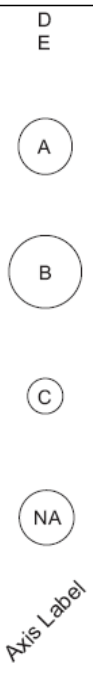
Το υφαντόγραμμα παρουσιάστηκε από τους Kumasaka & Shibata (2006) και πρόκειται για μέθοδο οπτικοποίησης πολυδιάστατων δεδομένων η οποία στηρίζεται σε μεγάλο βαθμό στο διάγραμμα παράλληλων συντεταγμένων. Και σε αυτή την περίπτωση οι μεταβλητές αναπαριστώνται ως κατακόρυφοι άξονες Οι τιμές των παρατηρήσεων τοποθετούνται στους άξονες αυτούς ενώ μια τεθλασμένη γραμμή ενώνει τα σημεία για να σχηματιστεί το προφίλ της κάθε πολυμεταβλητής παρατήρησης.

Πολλές φορές, στα διαγράμματα παράλληλων συντεταγμένων ο εντοπισμός των ιδιαίτερων χαρακτηριστικών των δεδομένων γίνεται δύσκολος, ιδιαίτερα σε περιπτώσεις που οι γραμμές που αντιστοιχούν στις παρατηρήσεις τέμνονται αρκετές φορές μεταξύ τους. Το πρόβλημα αυτό αντιμετωπίζεται με την επιλογή της θέσης αλλά και της κλίμακας μέτρησης των αξόνων ώστε οι γραμμές να ευθυγραμμίζονται όσο πιο οριζόντια γίνεται. Έτσι βελτιώνεται αρκετά η αποτελεσματικότητα του γραφήματος κάνοντας πιο εύκολη και ξεκάθαρη την άντληση των πληροφοριών.

Ο τρόπος κατασκευής του υφαντογράμματος επιτρέπει την οπτικοποίηση, εκτός των αριθμητικών και κατηγορικών δεδομένων αλλά και την αντιμετώπιση ελλειπουσών τιμών. Για να κατασκευάσουμε το γράφημα πρέπει αρχικά να ταξινομήσουμε τα δεδομένα σε αριθμητικά και μη αριθμητικά. Στη συνέχεια τα αριθμητικά ταξινομούνται σε συνεχή και διακριτά ενώ τα μη αριθμητικά ταξινομούνται σε διατάξιμα, μη διατάξιμα και λογικά. Στην περίπτωση των συνεχών

αριθμητικών δεδομένων, οι τιμές των παρατηρήσεων αναγράφονται πάνω σε μια συνεχή γραμμή ενώ στα διακριτά δεδομένα οι τιμές αναγράφονται σε μια σειρά από σημάδια. Η μύτη ενός βέλους τοποθετείται στην κατάλληλη άκρη ώστε να υποδηλώσει την κατεύθυνση που έχουν οι συντεταγμένες ενώ οι τιμές αντιστοιχίζονται με κύκλους και όχι με σημεία. Τέλος σημειώνονται η ελάχιστη και η μέγιστη τιμή κάθε μεταβλητής ώστε να έχουμε μια εικόνα του εύρους τιμών.

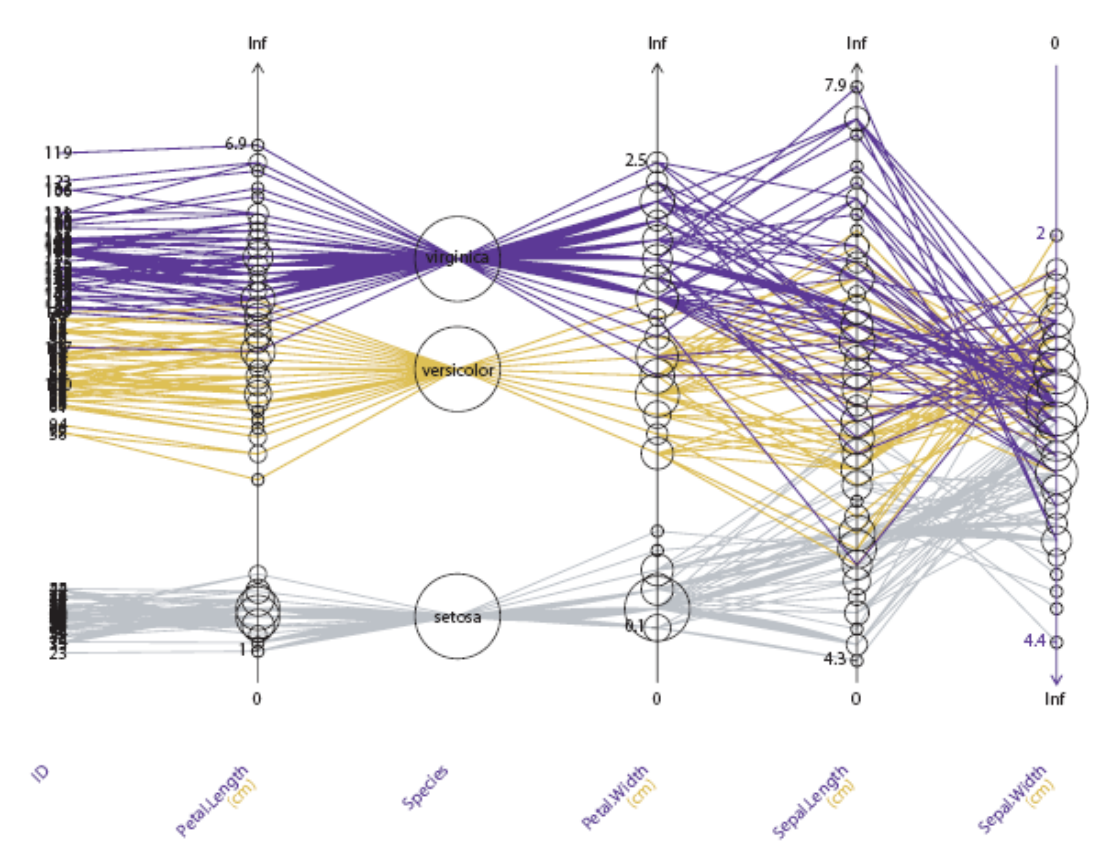
Στην περίπτωση που έχουμε μη αριθμητικά δεδομένα, δεν έχουμε κάποιον άξονα για τις τιμές αλλά χρησιμοποιούνται κύκλοι για κάθε επίπεδο της μεταβλητής, με το όνομα του επιπέδου να αναγράφεται πάνω στον κύκλο και το εμβαδόν αυτού να υποδηλώνει την αντίστοιχη συχνότητα. Αν πρόκειται για διατάξιμα δεδομένα, τοποθετείται μια σειρά από βέλη για να τονιστεί η φυσική διάταξη των επιπέδων της μεταβλητής ενώ αν έχουμε λογικά δεδομένα, ο κύκλος που περιέχει την επιλογή «false» χρωματίζεται με μαύρο χρώμα για να ξεχωρίζουν από τα υπόλοιπα δεδομένα. Σε κάθε περίπτωση η ύπαρξη ελλειπουσών τιμών υποδηλώνεται με έναν κύκλο ο οποίος τοποθετείται στο κάτω μέρος του γραφήματος ενώ το εμβαδόν του αντιστοιχεί στο πλήθος των τιμών που λείπουν. Πιο κάτω δίνεται ένας πίνακας με την αντιστοίχιση των τιμών των δεδομένων και τη σήμανση αυτών σε περίπτωση που είναι είτε αριθμητικά είτε μη αριθμητικά.

Numerical data		Non-numerical data		
Continuous	Discrete	Ordered	Unordered	Logical
				

Συνήθως στο αριστερό μέρος του γραφήματος τοποθετείται και μια στήλη με την ονομασία ID, όπου υποδηλώνεται η ταυτότητα κάθε πολυμεταβλητής παρατήρησης μέσα από έναν κωδικό ή κάποιον αριθμό και χρησιμοποιείται ως βοηθητικό εργαλείο στην ανάγνωση του γραφήματος.

Στο υφαντόγραμμα δύο ξεχωριστά χαρακτηριστικά συναντώνται, οι κόμβοι οι οποίοι υποδηλώνουν ότι ένα διάνυσμα των δεδομένων είναι απομονωμένο από τα υπόλοιπα διανύσματα και οι παράλληλες «υφάνσεις» οι οποίες υποδηλώνουν ύπαρξη γραμμικής σχέσης μεταξύ των μετρήσεων ή ισοδυναμία μεταξύ κατηγορικών δεδομένων.

Στη συνέχεια δίνεται ένα υφαντόγραμμα που έχει κατασκευαστεί με το DandD (http://www.stat.math.keio.ac.jp/DandDIV/) για τα γνωστά Iris data με 5 μεταβλητές, 1 κατηγορική και 4 συνεχείς.



ΚΕΦΑΛΑΙΟ 6

Ιεραρχικές Τεχνικές

6.1 Εισαγωγή

Οι τεχνικές της κατηγορίας αυτής αναπτύχθηκαν για την οπτικοποίηση πολυμεταβλητών δεδομένων με ιεραρχική δομή. Ένα κύριο χαρακτηριστικό στοιχείο των τεχνικών αυτών είναι ότι προσπαθούν να αξιοποιήσουν όσο το δυνατόν πιο αποτελεσματικά τη διαθέσιμη επιφάνεια προβολής ώστε να είναι δυνατή η αναπαράσταση μεγάλου όγκου δεδομένων. Είναι ιδανικές τεχνικές για οπτικοποίηση δεδομένων αρχείου και βοηθούν ιδιαίτερα τον εντοπισμό ομαδοποιήσεων.

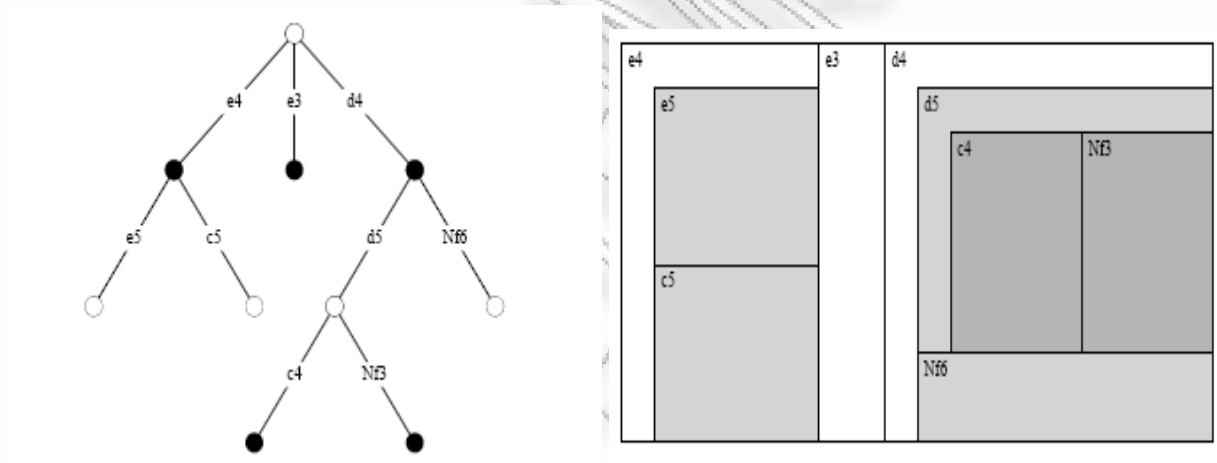
6.2 Δενδροχάρτες

Η μέθοδος των δενδροχαρτών εισήχθηκε από τους Johnson & Shneiderman (1991) με σκοπό να βοηθήσει στην οπτικοποίηση της δομής και του μεγέθους των αρχείων στο σκληρό δίσκο ενός ηλεκτρονικού υπολογιστή. Αναπτύχθηκε με χρήση μιας λογικής παράλληλης με αυτήν των δενδροδιαγραμμάτων (node & link diagrams) τα οποία και χρησιμοποιούνταν σαν κύρια εργαλεία οπτικοποίησης δεδομένων με ιεραρχική δομή μέχρι τότε αλλά παρουσίαζαν πολλούς περιορισμούς. Μεγάλες και πολύπλοκες δομές δεδομένων ήταν αδύνατον να αναπαρασταθούν με τη χρήση των δενδροδιαγράμματων αφού γίνονταν εξαιρετικά δύσχρηστα. Με τους δενδροχάρτες δίνεται η δυνατότητα να αξιοποιηθεί στο ακέραιο ο διαθέσιμος χώρος προβολής των ιεραρχικών δεδομένων αλλά και να ελεγχθούν διαδραστικά τα παρουσιαζόμενα αποτελέσματα. Επίσης βελτιώνεται η κατανόηση της δομής των δεδομένων, χωρίς να

απαιτείται μεγάλη εξοικείωση του παρατηρητή, καθώς και η αισθητική του γραφήματος.

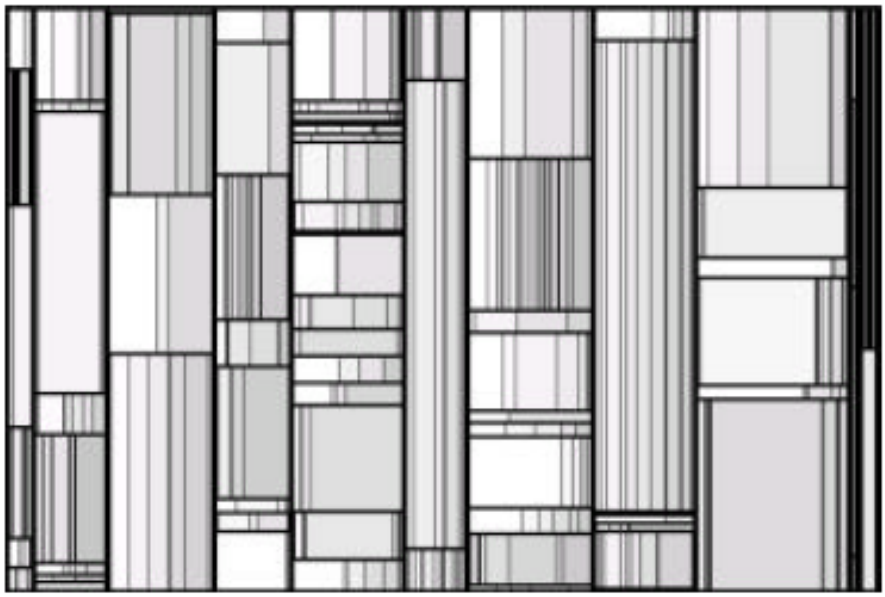
Η τεχνική κατασκευής των δενδροχαρτών στηρίζεται στο ότι αναπαριστούμε το διάγραμμα δένδρου με ένα ορθογώνιο παραλληλόγραμμο. Στη συνέχεια το παραλληλόγραμμο αυτό επιμερίζεται σε τόσα τμήματα, όσοι είναι και οι κόμβοι του δένδρου. Στο πρώτο ιεραρχικό επίπεδο, ο επιμερισμός γίνεται κάθετα, στο δεύτερο γίνεται οριζόντια και συνεχίζεται με αυτή τη λογική μέχρι να φθάσουμε στο τελευταίο επίπεδο.

Πρόκειται λοιπόν για μια επαναλαμβανόμενη διαδικασία με την οποία καταλήγουμε στο ότι η επιφάνεια κάθε επιμερισμένου παραλληλογράμμου θα είναι ανάλογη των κόμβων που περιέχει το αντίστοιχο επίπεδο. Ακολουθεί ένα παράδειγμα αναπαράστασης των ιεραρχικών δεδομένων ενός δενδροδιαγράμματος (node and link diagram) μέσω ενός αντίστοιχου δενδροχάρτη.



Υπάρχει η δυνατότητα να χρησιμοποιηθούν βοηθήματα στην οπτικοποίηση των δεδομένων μέσω ενός δενδροχάρτη όπως χρωματισμοί, σκίαση ή έντονα πλαίσια ώστε να διευκολυνθεί η ανάγνωσή τους και η ανάδειξη συγκεκριμένων χαρακτηριστικών. Έχουν προταθεί αρκετοί αλγόριθμοι για τον τρόπο επιμερισμού των παραλληλογράμμων με σκοπό το καλύτερο κάθε φορά οπτικό αποτέλεσμα. Ο αρχικός αλγόριθμος που εφαρμόστηκε για την κατασκευή των δενδροχαρτών λέγεται slice and dice algorithm και στηρίζεται στην εναλλαγή κατεύθυνσης (οριζόντια ή κάθετα) του επιμερισμού των παραλληλογράμμων κάθε φορά που αλλάζουμε επίπεδο

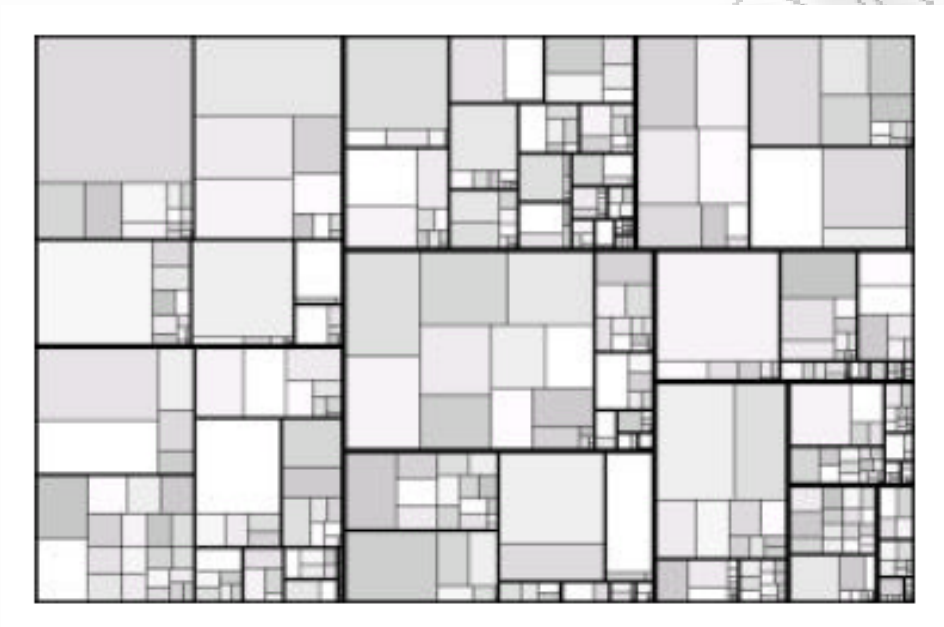
ιεραρχίας. Είναι πολύ απλός στην υλοποίησή του αλλά μειονεκτεί ως προς την οπτική αποτελεσματικότητα καθώς μπορεί να δημιουργήσει παραλληλόγραμμα με υψηλό aspect ratio, δηλαδή παραλληλόγραμμα με άνιση κατανομή του εμβαδού στις δύο διαστάσεις. Δίνουμε ένα παράδειγμα Treemap που έχει κατασκευαστεί με τον αλγόριθμο slice and dice. Δενδροχάρτες μπορούν να κατασκευαστούν με τα προγράμματα Ilog (http://www.ilog.com/products/ilogelixir/demos/) και Openviz (http://www.openviz.com/).



Οι Bruls, Huizing & vanWijk (2000) πρότειναν έναν εναλλακτικό αλγόριθμο μέσω του οποίου προκύπτουν οι τετραγωνισμένοι δενδροχάρτες (squarified treemaps). Ο αλγόριθμος αυτός επιμερίζει τα παραλληλόγραμμα με τέτοιο τρόπο ώστε να επιτυγχάνει το ελάχιστο aspect ratio κάθε φορά, δηλαδή να προσεγγίζουν τα τετράγωνα. έτσι λύνεται το πρόβλημα των μακρόστενων παραλληλογράμμων που σχηματίζονταν με τον αλγόριθμο slice and dice και βελτιώνεται η αποτελεσματικότητα του γραφήματος κάνοντας τις συγκρίσεις πιο εύκολες.

Επειδή ο υπολογιστικός χρόνος που απαιτείται για να γίνουν όλοι οι επιμερισμοί ταυτόχρονα αυξάνεται δραματικά, προτιμάται να γίνεται ο επιμερισμός τμηματικά για ένα επίπεδο κάθε φορά. Πολλές φορές χρησιμοποιούνται πλαίσια ώστε να τονιστεί η δομή των δεδομένων. Αν και ο αλγόριθμος αυτός βελτιώνει το οπτικό

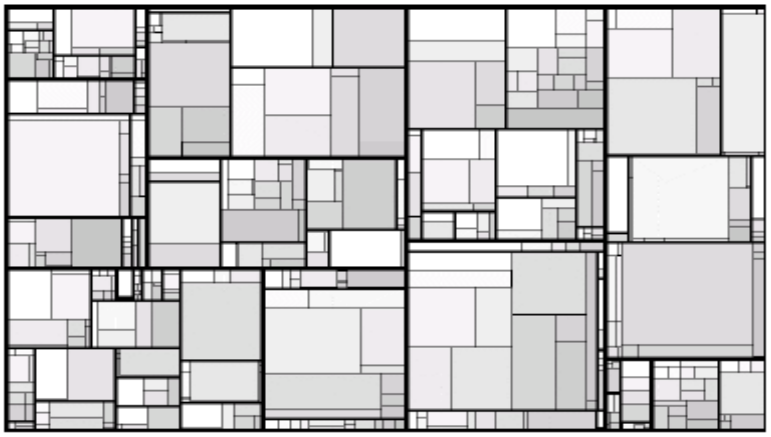
αποτέλεσμα, ιδιαίτερα όσον αφορά τα μεγέθη των παραλληλογράμμων, εμπεριέχει ένα σοβαρό μειονέκτημα: δε διατηρεί την αρχική διάταξη των δεδομένων, κάτι που συμβαίνει με τον αλγόριθμο slice and dice. Δίνουμε παρακάτω ένα παράδειγμα τετραγωνισμένου δενδροχάρτη.



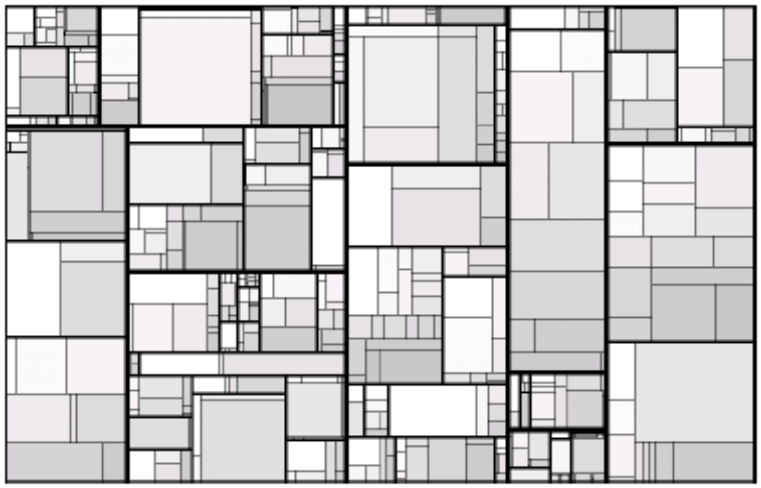
Για να αντιμετωπιστούν τα προβλήματα που προκύπτουν από τους δύο προηγούμενους αλγόριθμους, οι Berderson, Shneiderman & Wattenberg (2002) πρότειναν τους διατεταγμένους δενδροχάρτες (ordered treemaps). Ο σχετικός αλγόριθμος εξασφαλίζει ότι η αρχική διάταξη των ιεραρχικών δεδομένων θα διατηρηθεί και στο γράφημα, επιτυγχάνοντας παράλληλα και σχετικά χαμηλά aspect ratios για τα παραλληλόγραμμα.

Η τεχνική κατασκευής των διατεαγμένων δενδροχαρτών στηρίζεται στην επιλογή ενός σταθερού παραλληλογράμμου που καλείται pivot, γύρω από το οποίο γίνεται ο επιμερισμός με συγκεκριμένη επαναλαμβανόμενη διαδικασία. Η επιλογή του σταθερού παραλληλογράμμου μπορεί να γίνει με 3 κριτήρια: με βάση το μεγαλύτερο εμβαδόν (pivot-by-size), με βάση την πιο κεντρική θέση (pivot-by-middle) ή με βάση τον πιο ισομεγέθη διαχωρισμό (pivot-by-split-size). Δίνονται παρακάτω παραδείγματα διατεταγμένων δενδροχαρτών για τους 3 τρόπους επιλογής του pivot.

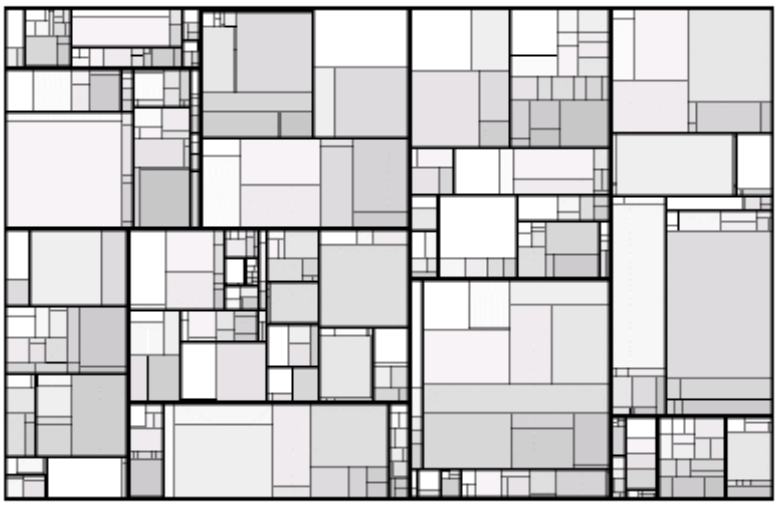
α. Με το κριτήριο pivot-by-size:



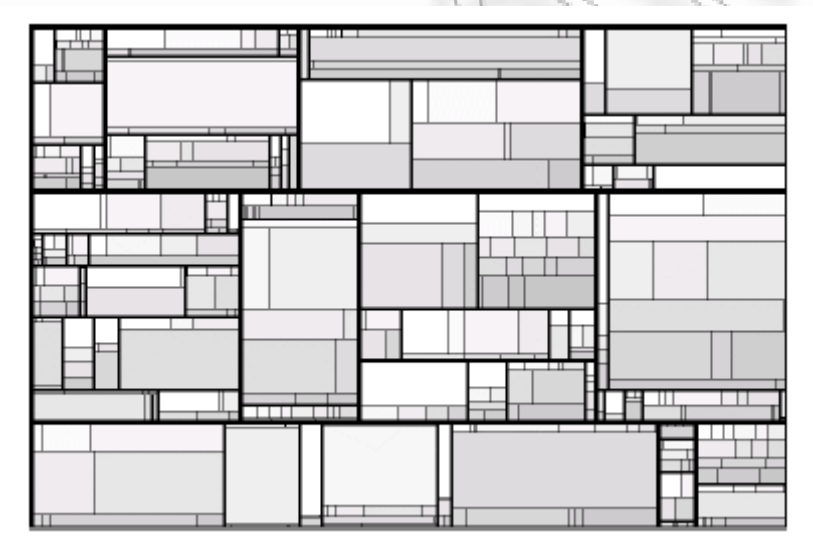
β. Με το κριτήριο pivot-by-middle:



γ. Με το κριτήριο pivot-by-split-size:

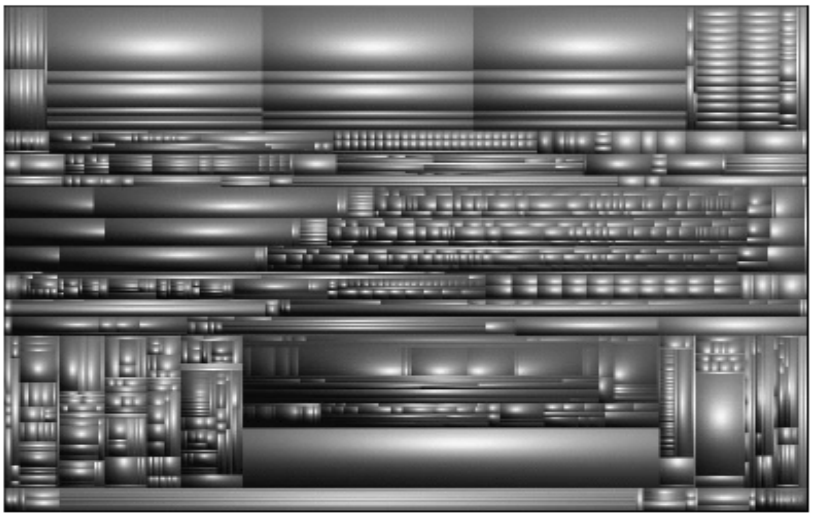


Μια ακόμα εναλλακτική πρόταση έγινε και πάλι από τους Berderson, Shneiderman & Wattenberg (2002) και ονομάζεται αλγόριθμος δενδροχαρτών «λωρίδων» (strip treemap algorithm). Η κατασκευή του αντίστοιχου γραφήματος γίνεται με τη διαδοχική προσθήκη των παραλληλογράμμων σε οριζόντιες ή κάθετες γραμμές, των οποίων το πλάτος ποικίλει. Η προσθήκη γίνεται με βάση την αρχική διάταξη των δεδομένων ενώ μεταφερόμαστε στην επόμενη γραμμή όταν διαπιστωθεί ότι η προσθήκη του επόμενου παραλληλογράμμου αυξάνει το aspect ratio των υπολοίπων πάνω από το όριο που έχουμε θέσει. Είναι ουσιαστικά ένας συνδυασμός των διατεταγμένων δενδροχαρτών και των τετραγωνισμένων δενδροχαρτών, επιτυγχάνοντας τη διατήρηση της διάταξης των δεδομένων αλλά και αρκετά χαμηλά aspect ratios. Ένα παράδειγμα δενδροχάρτη «λωρίδων» παρουσιάζεται παρακάτω.

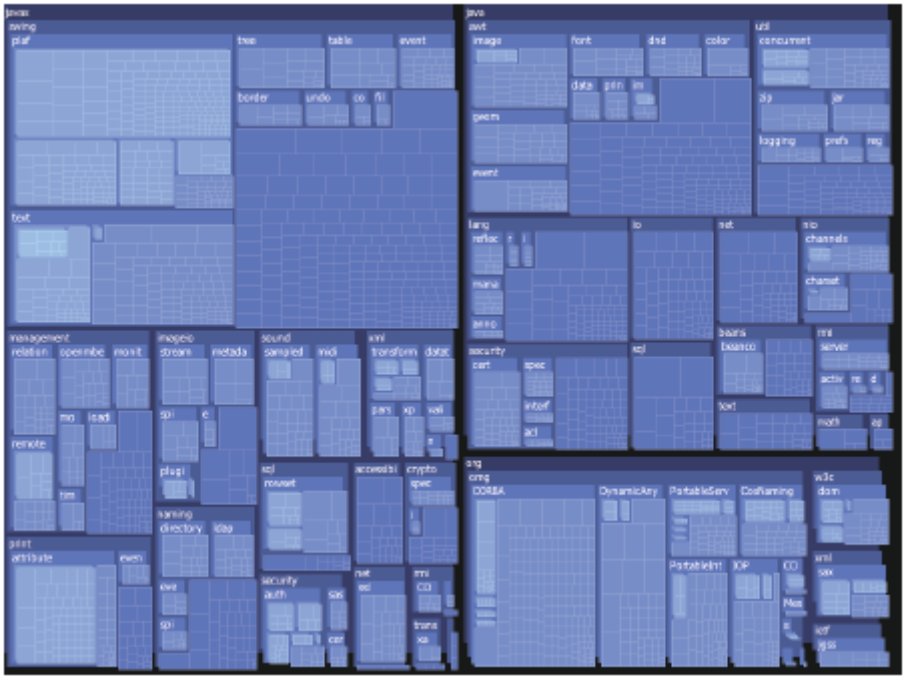


Οι van Wijk & van de Wetering (1999) έχουν προτείνει τους λεγόμενους δενδροχάρτες «μαξιλαριών» (cushion treemaps) που προκύπτουν τοποθετώντας στα παραλληλόγραμμα κορυφές, οι οποίες στη συνέχεια φωτοσκιάζονται με κατάλληλο τρόπο ώστε να αποκτά το γράφημα μια τρισδιάστατη οπτική. Η τεχνική εκμεταλλεύεται το γεγονός πως ο άνθρωπος αντιλαμβάνεται πολύ ευκολότερα τις διαφορές μέσα από μια σκιαγραφημένη επιφάνεια από ότι στο επίπεδο.

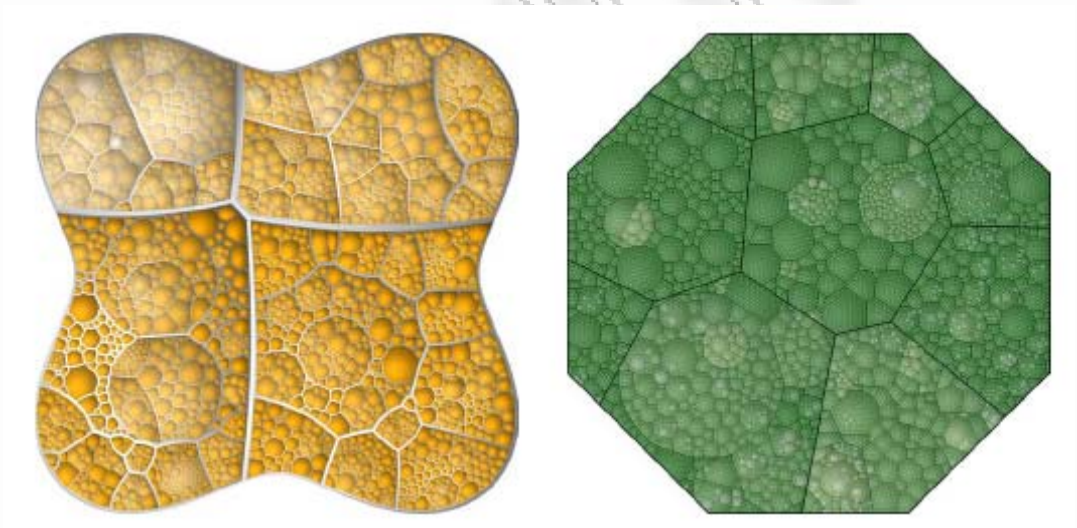
Η παραπάνω τεχνική μπορεί να εφαρμοστεί σε όλα τα είδη δενδροχαρτών εύκολα και γρήγορα με μια κατάλληλη τροποποίηση του αλγορίθμου αυξάνοντας κατά πολύ την αποτελεσματικότητα του αντίστοιχου γραφήματος. Παρουσιάζουμε στη συνέχεια ένα παράδειγμα δενδροχάρτη «μαξιλαριών».



Μια ακόμα συμπληρωματική τεχνική με την ονομασία «επικαλυπτόμενοι» δενδροχάρτες (cascaded treemaps) παρουσιάστηκε από τους Lü & Fogarty (2008). Τα παραλληλόγραμμα εμφανίζονται ως πλακίδια τοποθετημένα το ένα πάνω από το άλλο, δίνοντας έτσι και μια αίσθηση βάθους στο γράφημα. Η τεχνική δίνει έμφαση στην ανάδειξη της δομής των δεδομένων και επιτυγχάνει εξοικονόμηση χώρου προβολής. Συνήθως χρησιμοποιούνται και αποχρώσεις ώστε να τονιστούν τα διαφορετικά επίπεδα ιεραρχίας των δεδομένων. Ένα χαρακτηριστικό δείγμα «επικαλυπτόμενου δενδροχάρτη» δίνεται παρακάτω.



Τέλος μια παραλλαγή των δενδροχαρτών είναι και οι δενδροχάρτες Voronoi που προτάθηκαν από τους Balzer, Deussen & Lewerentz (2005). Η διαφορά της τεχνικής αυτής με τους υπόλοιπους δενδροχάρτες έγκειται στο γεγονός ότι δε χρησιμοποιούνται παραλληλόγραμμα για την οπτικοποίηση των ιεραρχικών δεδομένων, αλλά μορφές ψηφιδωτών οι οποίες μπορεί να έχουν κυκλικό, τριγωνικό, πολυγωνικό ή οποιοδήποτε άλλο σχήμα. Έτσι δεν υφίσταται πλέον το πρόβλημα βελτιστοποίησης του aspect ratio των παραλληλογράμμων και βελτιώνεται η αναπαράσταση της δομής των δεδομένων. Και σε αυτή την περίπτωση χρησιμοποιούνται αποχρώσεις για να τονιστούν τα επίπεδα της ιεραρχίας αλλά και φωτοσκίαση για να δοθεί τρισδιάστατη οπτική στο γράφημα. Δίνουμε πιο κάτω δύο παραδείγματα δενδροχαρτών Voronoi.



6.3 Στοιβογράμματα

Πρόκειται για μια μέθοδο που προτάθηκε από τους LeBlanc, Ward & Wittels (1990) με σκοπό την οπτικοποίηση πολυδιάστατων δεδομένων με ιεραρχική δομή. Η ιδέα στην οποία βασίζεται η μέθοδος των στοιβογραμμάτων είναι η ενσωμάτωση ενός συστήματος συντεταγμένων στο εσωτερικό ενός άλλου συστήματος συντεταγμένων. Κατά την απεικόνιση των μεταβλητών στο γράφημα, θεωρούμε ότι οι μεταβλητές

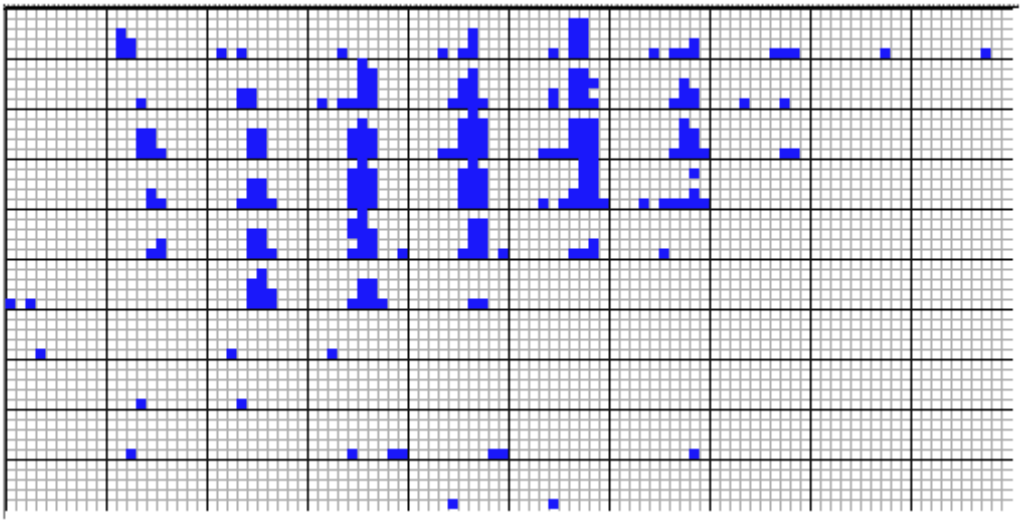
«κινούνται» με διαφορετική ταχύτητα, με τις εσωτερικές μεταβλητές να κινούνται ταχύτερα, δηλαδή να επαναλαμβάνονται περισσότερες φορές από τις εξωτερικές μεταβλητές. Αυτό σημαίνει ότι οι δύο εξωτερικές μεταβλητές ή αλλιώς οι πιο «αργές» μεταβλητές, χωρίζουν την επιφάνεια σε τμήματα ανάλογα με το πλήθος στοιχείων της κάθε μεταβλητής. Στη συνέχεια οι αμέσως πιο «γρήγορες» μεταβλητές χωρίζουν κάθε ένα από τα τμήματα αυτά σε υποτμήματα και η διαδικασία επαναλαμβάνεται μέχρι να ολοκληρωθεί η ενσωμάτωση όλων των μεταβλητών.

Θα πρέπει επομένως να έχει δοθεί εξ' αρχής σε κάθε μεταβλητή ένα επίπεδο ή «ταχύτητα» και μια διεύθυνση, κάθετη ή οριζόντια. Επίσης θα πρέπει να ορίσουμε στις μεταβλητές πεπερασμένο πλήθος τιμών, σε περίπτωση που δεν είναι κατηγορικές. Εφόσον ο διαχωρισμός του επιπέδου γίνεται με βάση ζεύγη μεταβλητών, πρέπει στην περίπτωση ύπαρξης μονού αριθμού μεταβλητών να προσθέσουμε μια «κενή» μεταβλητή για να εξισορροπήσουμε το γράφημα.

Όταν ολοκληρωθεί η διαδικασία ενσωμάτωσης όλων των μεταβλητών, υπάρχει η δυνατότητα χρήσης χρωμάτων για τον τονισμό κάποιων μεταβλητών. Επίσης είναι δυνατή η αναδιάταξη των μεταβλητών από τον ερευνητή ώστε να πάρουμε διάφορα γραφήματα αλλά ο καθορισμός των δυνατών τιμών των μεταβλητών.

Η αποτελεσματικότητα του γραφήματος εξαρτάται σε μεγάλο βαθμό από τις μεταβλητές που χρησιμοποιούνται στο εξωτερικό σύστημα συντεταγμένων και για το λόγο αυτό πρέπει αυτές να επιλέγονται πολύ προσεκτικά. Συνίσταται να επιλέγονται για τη θέση αυτή οι σημαντικότερες μεταβλητές των δεδομένων.

Πιο κάτω δίνουμε ένα χαρακτηριστικό παράδειγμα της μεθόδου Dimensional Stacking που αφορά δεδομένα εξόρυξης πετρελαίου, με 4 μεταβλητές. Στον εξωτερικό οριζόντιο άξονα αναπαριστάται το γεωγραφικό μήκος της περιοχής εξόρυξης ενώ στον εσωτερικό οριζόντιο άξονα το μέγεθος του κοιτάσματος πετρελαίου. Στον εξωτερικό κάθετο άξονα είναι το γεωγραφικό πλάτος της περιοχής εξόρυξης και στον εσωτερικό κάθετο άξονα είναι το βάθος εξόρυξης. Παρόμοια γραφήματα μπορούν να κατασκευαστούν με το XmdvTool.



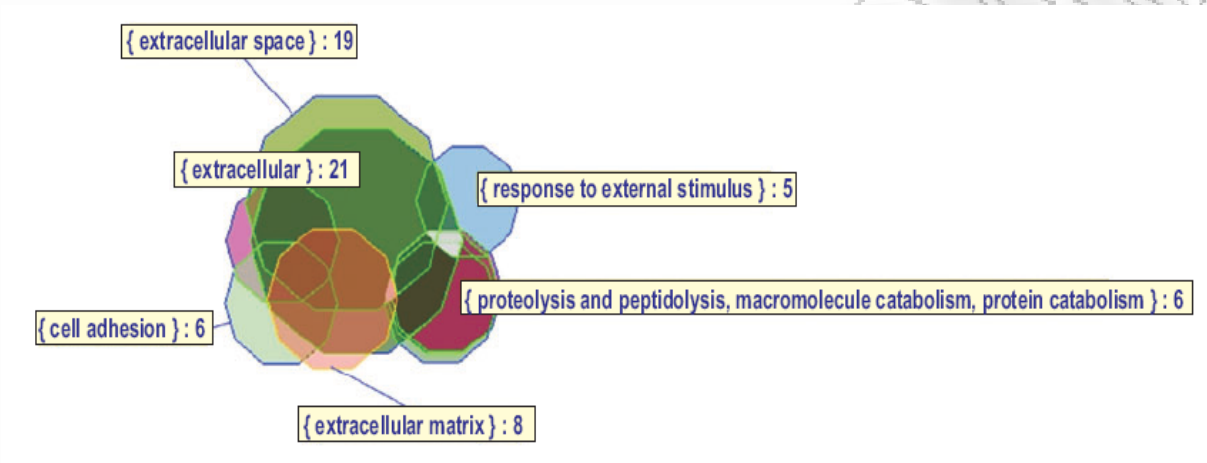
6.4 Διαγράμματα Venn

Πρόκειται για τεχνική η οποία κάνει χρήση των γνωστών διαγραμμάτων Venn για να αναδείξει τις διάφορες σχέσεις και αλληλεπιδράσεις που υπάρχουν σε σύνολα δεδομένων. Βέβαια τα διαγράμματα Venn που χρησιμοποιούνται είναι ελαφρώς τροποποιημένα ως προς το εξής: η επιφάνειά τους είναι ανάλογη της ποσότητας του συνόλου που αντιπροσωπεύουν. Έτσι, εκτός των αλληλεπιδράσεων και των σχέσεων μεταξύ των δεδομένων που γίνονται εύκολα ορατές από το γράφημα, μπορούμε να λάβουμε πληροφόρηση και για το μέγεθος των ίδιων των δεδομένων. Επιπλέον, υπάρχει η δυνατότητα χρήσης χρωματισμών αλλά και ετικετών στα αντίστοιχα σύνολα ώστε να βελτιώνεται το οπτικό αποτέλεσμα.

Υπάρχουν όμως κάποιοι περιορισμοί στα διαγράμματα Venn: σε περίπτωση που έχουμε να αναπαραστήσουμε πολλά σύνολα δεδομένων, το τελικό γράφημα μπορεί να γίνει εξαιρετικά πολύπλοκο και ουσιαστικά να μη μπορεί να αναγνωστεί, επομένως θα πρέπει να περιοριζόμαστε σε μικρό αριθμό συνόλων, ειδικά στην περίπτωση που υπάρχουν πολλές αλληλεπιδράσεις. Επίσης υπάρχει το πρόβλημα της ασυνέπειας όταν περιοχές στο διάγραμμα υποδηλώνουν σχέσεις μεταξύ των συνόλων, οι οποίες στην πραγματικότητα δεν υπάρχουν. Τις περιπτώσεις αυτές τις αντιμετωπίζουμε χρωματίζοντας με μαύρο χρώμα τις περιοχές αυτές.

Τα διαγράμματα Venn δεν έχουν αναγκαστικά κυκλικό σχήμα αλλά μπορεί να είναι πολύγωνα, όπως συμβαίνει και στο παράδειγμα που παραθέτουμε παρακάτω, το

οποίο έχει υλοποιηθεί με το λογισμικό VennMaster και αφορά γενετικά δεδομένα. Παρόμοια διαγράμματα μπορούν να κατασκευαστούν και με το Visokio Omniscope (http://www.visokio.com/omniscope) .



6.5 Δενδρογράμματα

Τα δενδρογράμματα είναι γραφήματα που εξυπηρετούν την οπτικοποίηση συστάδων σε πολυδιάστατα δεδομένα με ιεραρχική δομή. Στο κάτω μέρος του διαγράμματος βρίσκονται όλες οι παρατηρήσεις τοποθετημένες σε έναν οριζόντιο άξονα, σε ίσες αποστάσεις μεταξύ τους. Στον κάθετο άξονα τοποθετείται μια κλίμακα μέτρησης των αποστάσεων ή του μέτρου ανομοιότητας σύμφωνα με το οποίο ενώνονται δύο οποιεσδήποτε συστάδες.

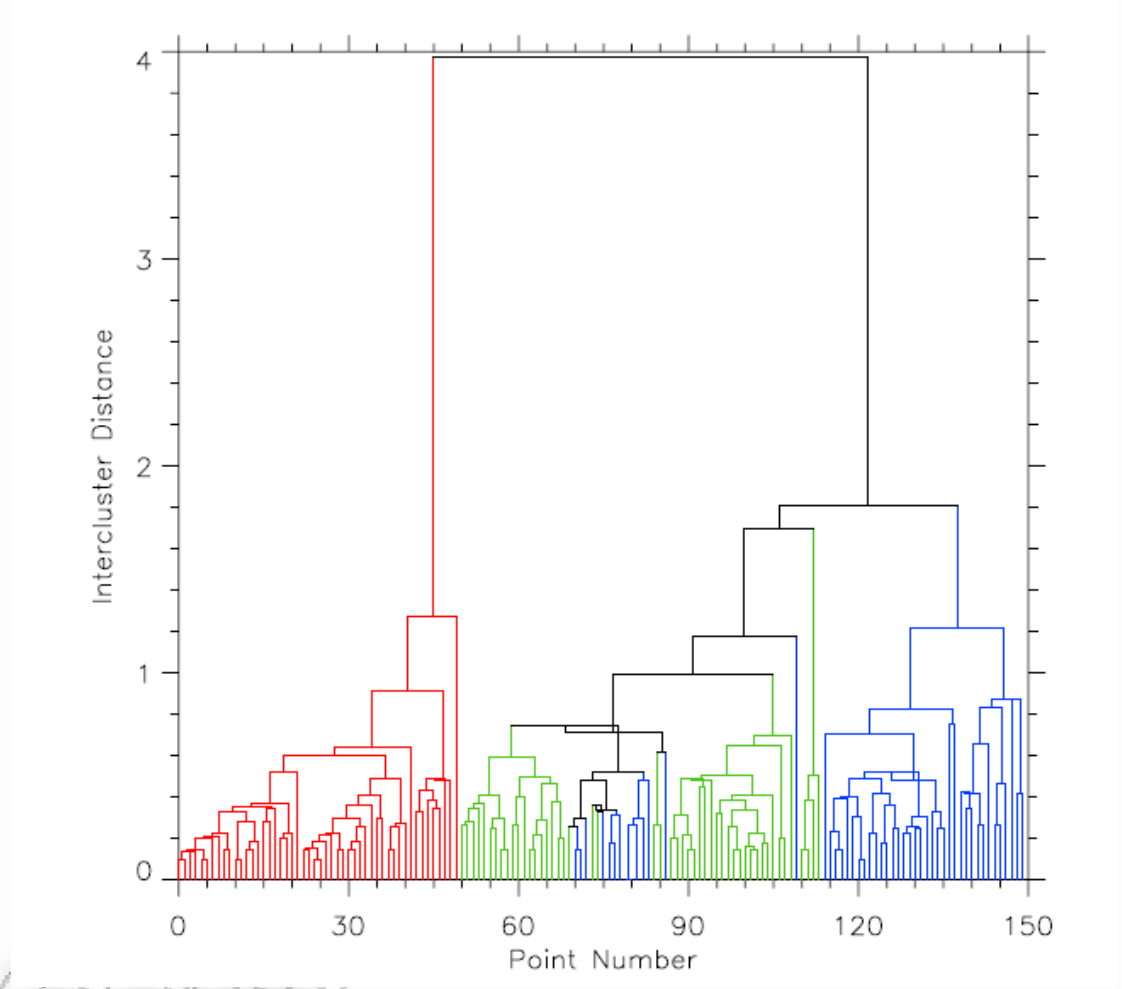
Για τον υπολογισμό των αποστάσεων συνένωσης υπάρχουν πολλοί τρόποι, κάθε ένας από τους οποίους δίνει και διαφορετική ομαδοποίηση. Ο αριθμός των συστάδων που θα σχηματιστούν μπορεί να μειωθεί ή να αυξηθεί αν θέσουμε ως όριο μια συγκεκριμένη απόσταση συνένωσης. Όταν δύο παρατηρήσεις ομαδοποιούνται στο διάγραμμα, χρησιμοποιείται μια οριζόντια και δύο κάθετες γραμμές για να δηλώσουν το σχηματισμό της αντίστοιχης συστάδας.

Ο τρόπος με τον οποίο τοποθετούνται οι παρατηρήσεις στον οριζόντιο άξονα εξαρτάται από το σχηματισμό των συστάδων, καθώς θα πρέπει να αποφεύγονται οι διασταυρώσεις των συστάδων μεταξύ τους ώστε να παραμένει ευανάγνωστο το

διάγραμμα. Συνήθως δίνεται όμως η δυνατότητα στον χρήστη να χρωματίσει τις συστάδες ώστε να διευκολύνει την ανάγνωση του διαγράμματος.

Ένα μειονέκτημα των δενδρογραμμάτων είναι ότι δε μπορούν να χρησιμοποιηθούν αποτελεσματικά σε μεγάλα σύνολα δεδομένων, καθώς απαιτούν πολύ μεγάλη επιφάνεια οπτικοποίησης.

Παρακάτω δίνεται ένα παράδειγμα εφαρμογής δενδρογράμματος στο σύνολο δεδομένων iris data. Παρόμοια διαγράμματα μπορούν να κατασκευαστούν με το Minitab (http://www.minitab.com/), το Statgraphics και το Stata.



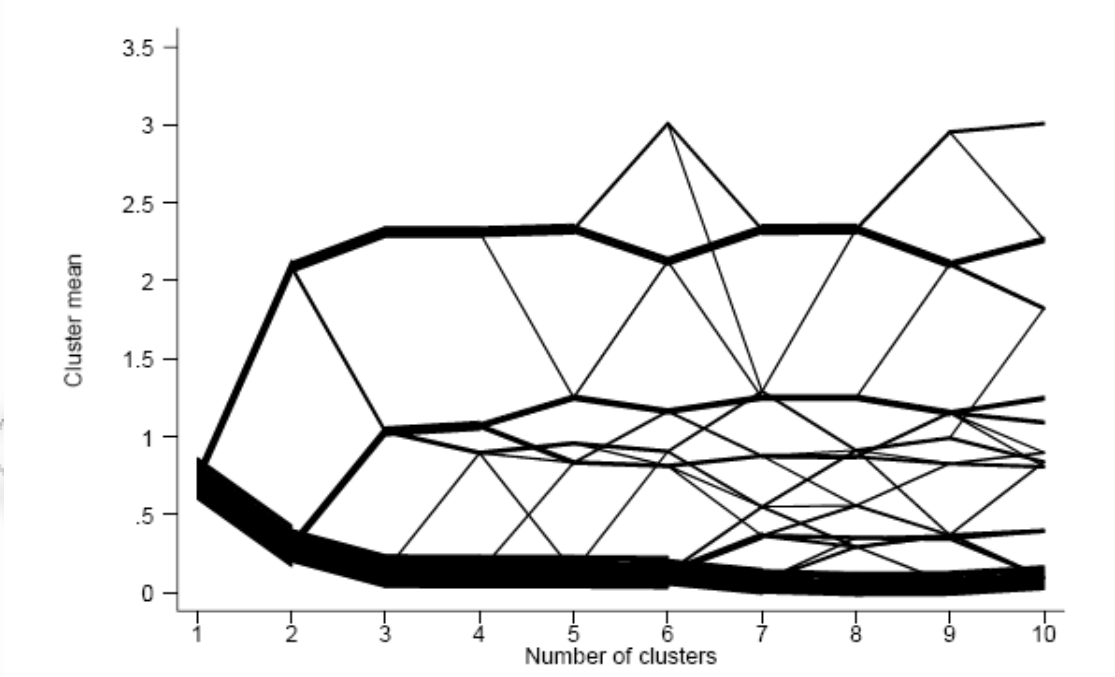
6.6 Σμηνογράμματα

Η μέθοδος των σμηνογραμμάτων εισήχθηκε από τον Schonlau (2002) και αποτελεί μια επέκταση των δενδρογραμμάτων για την οπτικοποίηση της δομής των

ομάδων σε πολυδιάστατα δεδομένα. Το κύριο πλεονέκτημα που έχουν τα σμηνογράμματα έναντι των δενδρογραμμάτων είναι ότι μπορούν να χρησιμοποιηθούν και για την αναπαράσταση μη ιεραρχικών δεδομένων.

Η ουσιαστική διαφορά μεταξύ σμηνογράμματος και δενδρογράμματος είναι ότι αντί για τις αποστάσεις των συστάδων που χρησιμοποιούμε στα δενδρογράμματα, στα σμηνογράμματα απεικονίζουμε στον έναν άξονα τον αριθμό των συστάδων. Αν και οι αποστάσεις αυτές δίνουν καλή πληροφορία για τις συστάδες, δε μπορούν να εφαρμοσθούν σε μη ιεραρχικά δεδομένα. Πάντως σε περίπτωση σχετικά μικρού αριθμού ιεραρχικών δεδομένων είναι προτιμότερη η χρήση των δενδρογραμμάτων λόγω αυτής της επιπλέον πληροφόρησης που παρέχουν στον ερευνητή. Στον άλλο άξονα του σμηνογράμματος τοποθετούμε τον μέσο των παρατηρήσεων για όλες τις μεταβλητές στην κάθε συστάδα. Οι μέσοι αυτοί των συστάδων ενώνονται με παραλληλόγραμμα των οποίων το πλάτος υποδεικνύει το πλήθος των παρατηρήσεων που περιέχονται στην αντίστοιχη συστάδα.

Και στην περίπτωση των σμηνογραμμάτων μπορούν να χρησιμοποιηθούν διαφορετικοί χρωματισμοί ώστε να διευκολυνθεί ο εντοπισμός των συστάδων, όπως συμβαίνει και στα δενδρογράμματα. Ένα χαρακτηριστικό παράδειγμα σμηνογράμματος που έχει κατασκευαστεί με το Stata παρουσιάζεται στη συνέχεια.



6.7 Θερμοχάρτες

Οι θερμοχάρτες ως όρος αναφέρονται σε οποιοδήποτε γράφημα χρησιμοποιεί χρώματα για την αναπαράσταση ποσοτικών δεδομένων. Οι συνηθέστερες μορφές τέτοιων γραφημάτων είναι οι μετεωρολογικοί χάρτες, οι οποίοι χρησιμοποιούν χρωματισμούς για να οπτικοποιήσουν ποσότητες όπως θερμοκρασία, ποσοστά βροχοπτώσεων κ.α.

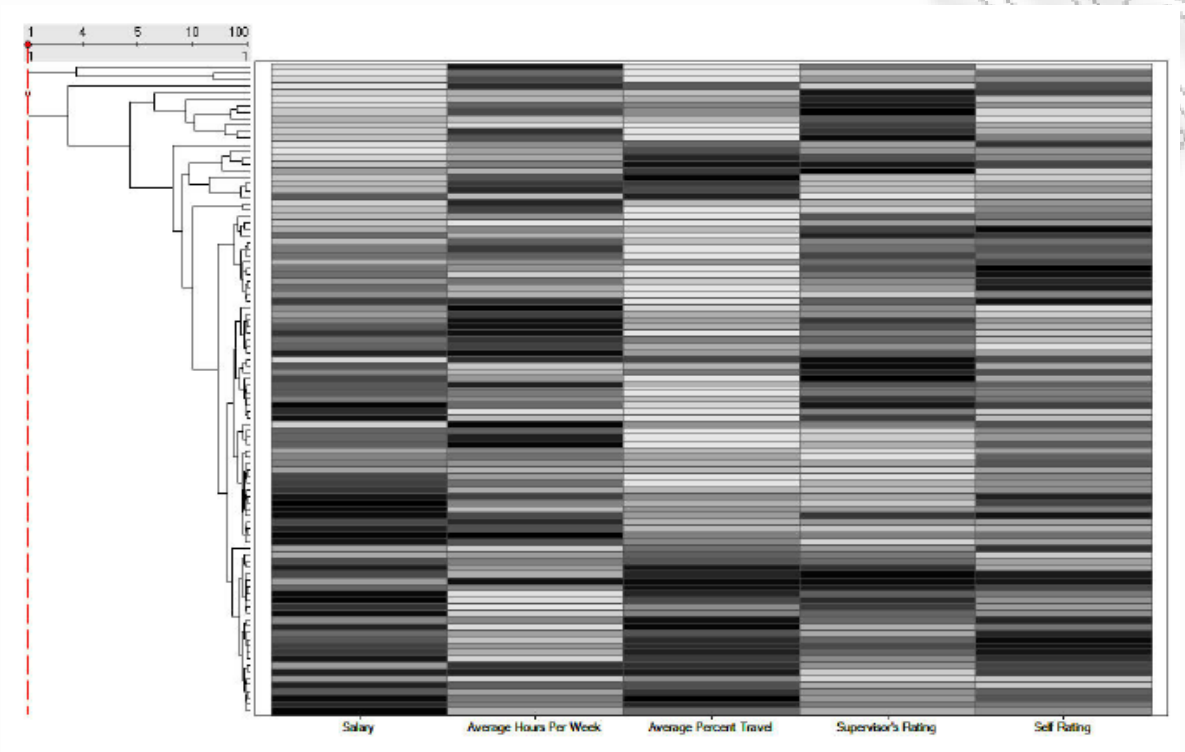
Στην περίπτωση που οι θερμοχάρτες χρησιμοποιούνται για την αναπαράσταση πολυδιάστατων δεδομένων, οι παρατηρήσεις παρουσιάζονται σε γραμμές και στήλες με τέτοιο τρόπο ώστε οι στήλες να αντιστοιχούν στις μεταβλητές και οι γραμμές να αντιστοιχούν στις πολυμεταβλητές παρατηρήσεις. Με αυτόν τον τρόπο, επιλέγοντας μια στήλη μπορούμε να δούμε τις τιμές των παρατηρήσεων για τη συγκεκριμένη μεταβλητή ενώ, επιλέγοντας μια συγκεκριμένη γραμμή, μπορούμε να δούμε τη συμπεριφορά της αντίστοιχης παρατήρησης για όλες τις μεταβλητές. Η τιμή κάθε παρατήρησης σε κάθε μεταβλητή, είτε αυτή είναι συνεχής είτε είναι κατηγορική, χρωματίζεται σύμφωνα με μια συνεχή κλίμακα από το άσπρο μέχρι το μαύρο (grayscale), με το άσπρο να υποδηλώνει τη χαμηλότερη μέτρηση και το μαύρο να υποδηλώνει την υψηλότερη.

Δίνεται η δυνατότητα στον χρήστη του γραφήματος να διατάξει κατά αύξουσα ή φθίνουσα σειρά τις τιμές των παρατηρήσεων για κάποια συγκεκριμένη μεταβλητή ώστε να αναδειχθούν πιθανές συσχετίσεις μεταξύ των μεταβλητών. Ακόμα, είναι δυνατόν να απομονωθούν συγκεκριμένες παρατηρήσεις ή μεταβλητές ώστε να μελετηθεί πιο προσεκτικά η συμπεριφορά τους.

Οι θερμοχάρτες χρησιμοποιούνται αρκετές φορές σε συνδυασμό με τα δενδρογράμματα για να οργανώσουν σε ιεραρχική δομή τις παρατηρήσεις, με βάση τις ομοιότητες που έχουν αυτές μεταξύ τους και τις ομάδες που σχηματίζονται. Ο τρόπος και τα κριτήρια με τα οποία γίνεται η ομαδοποίηση των δεδομένων μπορούν να επιλεγούν από τον ερευνητή σύμφωνα με τη φύση των δεδομένων αλλά και τους στόχους της ανάλυσης.

Πιο κάτω δίνουμε ένα παράδειγμα θερμοχάρτη με το αντίστοιχο δενδρόγραμμα για την ιεραρχική οργάνωση δεδομένων αξιολόγησης των συνθηκών εργασίας κάποιων εργαζομένων. Η κάθε γραμμή αντιστοιχεί σε έναν εργαζόμενο ενώ η κάθε

στήλη αντιστοιχεί σε μια από τις πέντε μεταβλητές των δεδομένων. Ανάλογα γραφήματα μπορούν να κατασκευαστούν με το Visokio Omniscope.



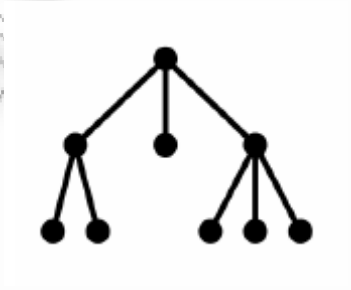
Εκτός από την χρήση της χρωματικής κλίμακας ανάμεσα στο άσπρο και το μαύρο (grayscale) υπάρχει η δυνατότητα να χρησιμοποιηθεί και μια εναλλασσόμενη κλίμακα μεταξύ δύο χρωμάτων, με ένα ουδέτερο χρώμα στο ενδιάμεσο όπως γκρι ή λευκό. Ένα παράδειγμα μιας τέτοιας κλίμακας δίνεται παρακάτω:

Οι χρωματικές κλίμακες χρησιμοποιούνται σε περιπτώσεις που οι τιμές εμπεριέχουν έναν λογικό διαχωρισμό όπως θετικές και αρνητικές, με τις μηδενικές τιμές να βρίσκονται στη μέση. Επίσης μπορεί να χρησιμοποιηθεί ως σημείο διαχωρισμού η μέση τιμή των παρατηρήσεων ή κάποιο άλλο κριτήριο που θα τεθεί από τον ερευνητή.

Στη συνέχεια δίνουμε ένα παράδειγμα θερμοχάρτη με τη χρήση της χρωματικής κλίμακας που παρουσιάστηκε πιο πάνω, σε δεδομένα μεταβολών των τιμών μετοχών, όπου τα δεδομένα αυτά έχουν ταξινομηθεί κατά αύξουσα σειρά.

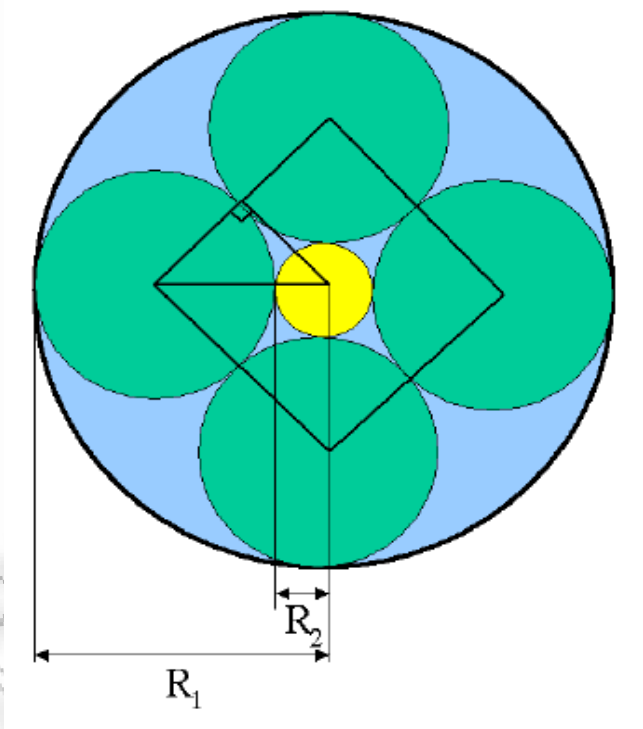
6.8 RINGS

Η μέθοδος RINGS (Ringed Interactive-Navigation Graph System) αναπτύχθηκε από τους Teoh & Ma (2002) με σκοπό την οπτικοποίηση μεγάλων συνόλων δεδομένων με ιεραρχική δομή που εμφανίζονται με τη μορφή δενδροδιαγραμμάτων (node & link diagrams) όπως αυτό που φαίνεται στο ακόλουθο σχήμα.



Στην τεχνική αυτή κάθε κόμβος και οι αντίστοιχες διακλαδώσεις του αναπαρίστανται σε έναν κύκλο, ο οποίος τοποθετείται με συγκεκριμένη διαδικασία στο εσωτερικό ενός αρχικού κύκλου. Συνήθως οι διακλαδώσεις χρωματίζονται ανάλογα με την απόσταση του κόμβου από τον αρχικό κόμβο.

Ενώ σε προηγούμενες μεθόδους οι κύκλοι τοποθετούνταν μόνο στην περίμετρο του αρχικού κύκλου, στη μέθοδο RINGS οι κύκλοι τοποθετούνται με τέτοιο τρόπο ώστε να γίνεται εκμετάλλευση του χώρου στο εσωτερικό του αρχικού κύκλου που πριν έμενε κενός και αναξιοποίητος. Οι κόμβοι με τις περισσότερες διακλαδώσεις τοποθετούνται περιμετρικά του κέντρου του αρχικού κύκλου σε ισομεγέθεις κύκλους, ενώ ο χώρος που μένει ελεύθερος στο εσωτερικό χρησιμοποιείται για να τοποθετηθούν οι κόμβοι με τις λιγότερες διακλαδώσεις, όπως φαίνεται και στο σχήμα που ακολουθεί.

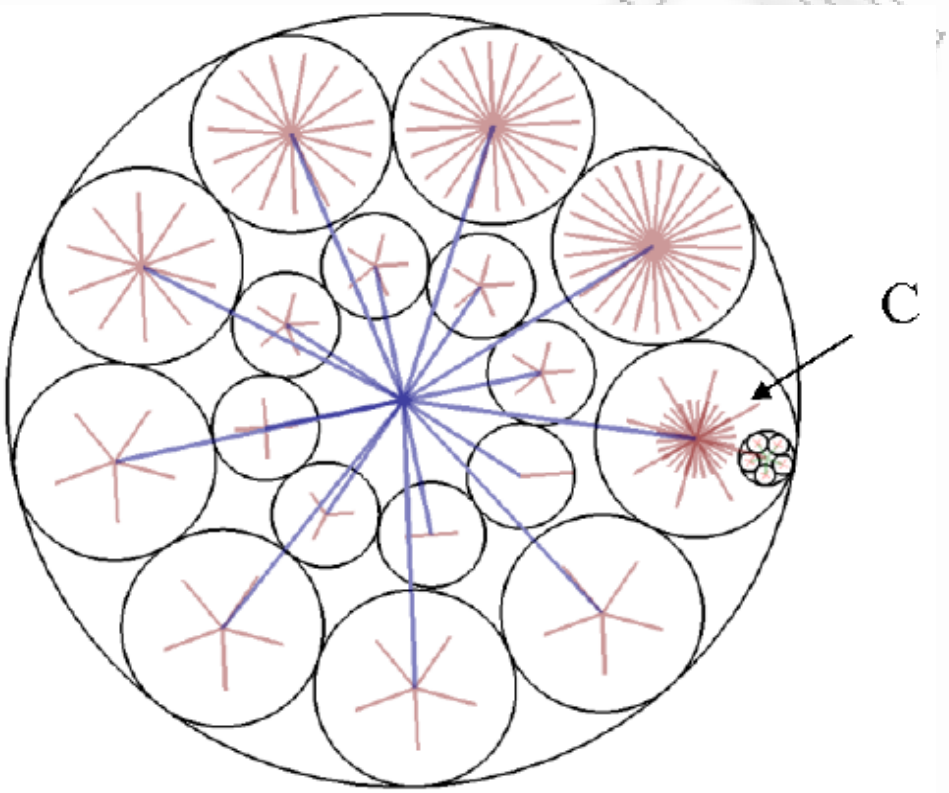


Οι 4 πράσινοι κύκλοι αντιστοιχούν στους 4 μεγαλύτερους κόμβους του αρχικού κόμβου που αναπαριστάται με το μεγάλο μπλε κύκλο. Ο μικρός κίτρινος κύκλος στο εσωτερικό είναι η επιφάνεια που απομένει για να τοποθετηθούν ο υπόλοιποι μικρότεροι κόμβοι.

Δίνεται επίσης η διαδραστική δυνατότητα στον χρήστη του γραφήματος να προβάλει κάποιο συγκεκριμένο κόμβο με τις αντίστοιχες διακλαδώσεις του, χωρίς να επηρεαστεί η ιεραρχική δομή των δεδομένων. Τέλος μπορεί να γίνει αναδιάταξη των κύκλων στο εσωτερικό του αρχικού κύκλου με σκοπό να βελτιωθεί η αποτελεσματικότητα του γραφήματος, αντιμετωπίζοντας έτσι το πρόβλημα της υπερκάλυψης των κόμβων.

Με χρήση της μεθόδου RINGS επιτυγχάνεται η μέγιστη δυνατή αξιοποίηση του διαθέσιμου χώρου ενώ έτσι γίνεται δυνατή και η οπτικοποίηση πολύ μεγάλων συνόλων δεδομένων με πολλούς κόμβους, γεγονός που θα ήταν πολύ δύσκολο με άλλες παρόμοιες τεχνικές.

Πιο κάτω δίνεται ένα παράδειγμα της μεθόδου RINGS σε ιεραρχικά δεδομένα, όπου με C σημειώνεται ο μεγαλύτερος κόμβος.



ΚΕΦΑΛΑΙΟ 7

Μέθοδοι Μείωσης Διαστάσεων

7.1 Εισαγωγή

Οι τεχνικές μείωσης διαστάσεων χρησιμοποιούνται με σκοπό να ελαττωθεί ο αρχικός αριθμός μεταβλητών των δεδομένων ώστε να μπορέσουμε στη συνέχεια να προβάλουμε τα τις παρατηρήσεις σε χώρο χαμηλότερης διάστασης, συνήθως διδιάστατο ή τρισδιάστατο. Η εφαρμογή των τεχνικών αυτών προέκυψε από την ανάγκη να αντιμετωπιστεί ένα πρόβλημα των πολυδιάστατων δεδομένων, η λεγόμενη «κατάρα» της διάστασης (dimensionality curse), το γεγονός δηλαδή ότι το μεγαλύτερο μέρος του πολυδιάστατου χώρου παραμένει κενό και δυσκολεύει την επισκόπηση των δεδομένων. Με κατάλληλους μετασχηματισμούς και τροποποιήσεις προσπαθούμε να αναπαραστήσουμε όσο πιο πιστά γίνεται τα δεδομένα στον χώρο χαμηλότερης διάστασης ώστε να έχουμε τη μικρότερη δυνατή απώλεια πληροφορίας των αρχικών δεδομένων.

7.2 Projection Pursuit / Grand Tour

Η μέθοδος Projection Pursuit εισήχθηκε από τους Friedman & Tukey (1974) ενώ μελετήθηκε αρκετά και από τον Huber (1985). Πρόκειται για αλγόριθμο που χρησιμοποιεί τις αποστάσεις μεταξύ των δεδομένων αλλά και τη διακύμανση του πλήθους των δεδομένων ώστε να επιτύχει τις βέλτιστες γραμμικές προβολές τους από τον πολυδιάστατο στο διδιάστατο χώρο.

Ο αλγόριθμος αυτός σχετίζει την κάθε κατεύθυνση στον πολυδιάστατο χώρο με έναν συνεχή δείκτη, που καλείται δείκτης προβολής (projection index) και μετρά τη χρησιμότητα της συγκεκριμένης κατεύθυνσης ως άξονα προβολής και στη συνέχεια μεταβάλει την κατεύθυνση προβολής ώστε να μεγιστοποιηθεί ο δείκτης. Η μέθοδος

μπορεί να φανεί ιδιαίτερα χρήσιμη στον εντοπισμό ομαδοποιήσεων και το διαχωρισμό των δεδομένων. Σε περίπτωση που βρεθούν προβολές που διαχωρίζουν εμφανώς τα δεδομένα σε δύο ή περισσότερες ομάδες, μπορούν τα δεδομένα αυτά να απομονωθούν και η μέθοδος να εφαρμοστεί πλέον στις σχηματισμένες ομάδες, οδηγώντας σε νέες προβολές που ίσως αποκαλύψουν περαιτέρω ομαδοποιήσεις.

Η επιλογή του σκοπού που θέλουμε να εξυπηρετήσει ο δείκτης προβολής είναι πολύ κρίσιμη για την επιτυχία του αλγορίθμου. Ο δείκτης αυτός συνήθως είναι ένα μέτρο μη-κανονικότητας, με την εντροπία να εμφανίζεται ως μια πρώτη επιλογή. Υπενθυμίζουμε ότι η εντροπία H ενός τυχαίου διανύσματος $\boldsymbol{y}$ με συνάρτηση πυκνότητας πιθανότητας f ορίζεται ως εξής:

$$H(\boldsymbol{y}) = -\int f(\boldsymbol{y}) \log f(\boldsymbol{y}) d\boldsymbol{y}.$$

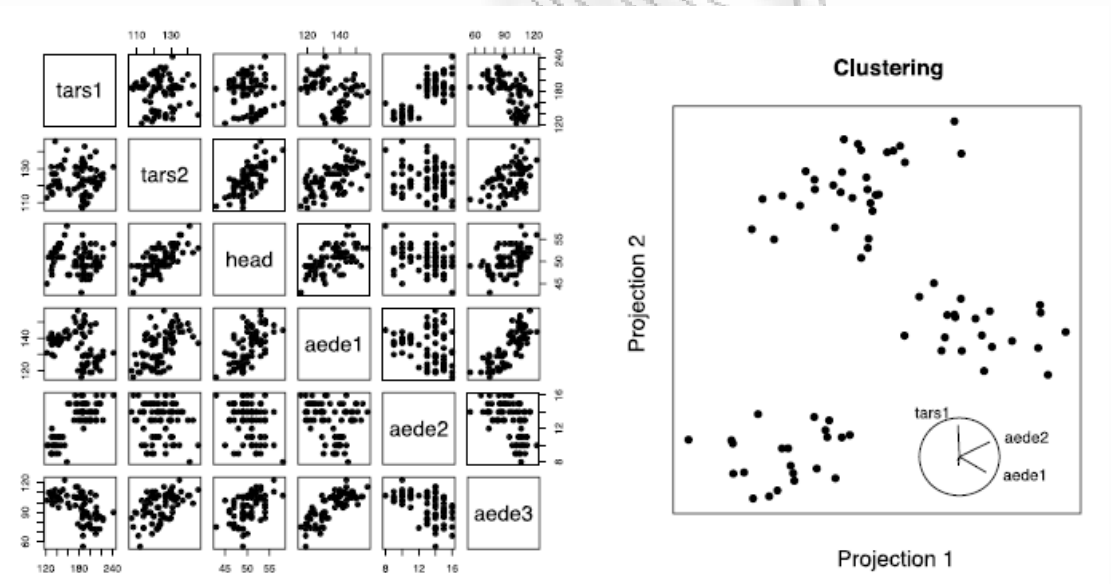
Η εντροπία μεγιστοποιείται ως προς την f όταν η συνάρτηση πυκνότητας πιθανότητας είναι η κανονική. Εφόσον σκοπός μας είναι να βρούμε κατευθύνσεις $\boldsymbol{w}$ τέτοιες ώστε η προβολή των δεδομένων $\boldsymbol{x}$ σε αυτήν την κατεύθυνση $\boldsymbol{w}^T\boldsymbol{x}$ να παρουσιάζει κάποια ενδιαφέρουσα συμπεριφορά, μπορούμε να ελαχιστοποιήσουμε την $H(\boldsymbol{w}^T\boldsymbol{x})$ ως προς $\boldsymbol{w}$, θεωρώντας τη διακύμανση της $\boldsymbol{w}^T\boldsymbol{x}$ σταθερή, για να εντοπίσουμε τις κατευθύνσεις αυτές. Το πρόβλημα με τη διαφορική εντροπία είναι ότι για την εκτίμησή της απαιτείται ο υπολογισμός της πυκνότητας της $\boldsymbol{w}^T\boldsymbol{x}$, πράγμα πολύ δύσκολο τόσο θεωρητικά όσο και πρακτικά. Για το λόγο αυτό έχουν προταθεί και άλλα μέτρα μη-κανονικότητας όπως σταθμισμένες αποστάσεις μεταξύ της πυκνότητας του διανύσματος των δεδομένων $\boldsymbol{x}$ και της πολυμεταβλητής κανονικής κατανομής ή ακόμα και προσεγγίσεις της διαφορικής εντροπίας.

Καθώς όμως η μέθοδος Projection Pursuit περιορίζεται στην επιλογή μιας προβολής από το σύνολο των δυνατών πολυδιάστατων προβολών των δεδομένων, δεν υπάρχει η δυνατότητα για ταυτόχρονη μελέτη περισσότερων προβολών. Στο πρόβλημα αυτό της στατικότητας ήρθαν να βοηθήσουν οι Asimov (1985) και Asimov & Buja (1986) προτείνοντας τη μέθοδο Grand Tour η οποία αποτελεί ένα δυναμικό εργαλείο στην εξέταση των πολυδιάστατων δεδομένων μέσα από τις διάφορες προβολές τους.

Ο σκοπός της μεθόδου είναι να έχουμε εικόνα για τα δεδομένα από διάφορες οπτικές γωνίες. Αυτό επιτυγχάνεται περιστρέφοντας διδιάστατες προβολές των δεδομένων με βάση έναν θεωρητικό άξονα ο οποίος επιλέγουμε να είναι μια από τις μεταβλητές. Αν δηλαδή έχουμε p+1 μεταβλητές, επιλέγουμε τη μια ως μεταβλητή

χρόνου και με τις υπόλοιπες κατασκευάζουμε διδιάστατες προβολές. Έτσι με τις μεταβολές της μεταβλητής-χρόνου παρατηρούμε τις διάφορες διαδοχικές προβολές προσπαθώντας να ανιχνεύσουμε ενδιαφέρουσες δομές στα δεδομένα. Ένα σημαντικό ζήτημα που προκύπτει στη μέθοδο Grand Tour είναι ο τρόπος που γίνεται η εναλλαγή των προβολών καθώς μεταβάλλεται η μεταβλητή-χρόνος, αφού η κίνηση θα πρέπει να γίνεται ομαλά και με διακριτά βήματα ώστε ο ερευνητής να έχει ξεκάθαρη εικόνα των διαφόρων προβολών και η αναπαράσταση να έχει αποτελεσματικότητα.

Εφαρμογές της μεθόδου Grand Tour μπορούν να γίνουν με το στατιστικό πακέτο ggobi. Στη συνέχεια δίνεται ένα παράδειγμα της μεθόδου σε δεδομένα 6 μεταβλητών όπου αριστερά φαίνεται ο πίνακας όλων των δυνατών διδιάστατων διαγραμμάτων διασποράς και αριστερά μια κατάλληλα επιλεγμένη προβολή από στην οποία είναι εμφανείς οι ομάδες των δεδομένων που σχηματίζονται.



7.3 Ανάλυση Κύριων Συνιστωσών

Η μέθοδος της Ανάλυσης Κύριων Συνιστωσών προτάθηκε από τον Pearson (1901) και έχει ως σκοπό τη μείωση της διάστασης ενός συνόλου δεδομένων το οποίο αποτελείται από ένα μεγάλο πλήθος συσχετισμένων μεταβλητών. Αυτό που επιχειρείται είναι μέσω ενός γραμμικού μετασχηματισμού των αρχικών μεταβλητών,

να προκύψουν κάποιες νέες μεταβλητές, οι κύριες συνιστώσες, από τις οποίες θα επιλεγούν αυτές που ερμηνεύουν το μεγαλύτερο μέρος της διακύμανσης των δεδομένων.

Τα πλεονεκτήματα που αποκομίζουμε από τη μέθοδο αυτή είναι πολλά, με κυριότερο το ότι από τις αρχικές συσχετισμένες μεταβλητές μεταβαίνουμε σε νέες ασυσχέτιστες, αντιμετωπίζοντας έτσι το πρόβλημα της πολυσυγγραμικότητας στην ανάλυση παλινδρόμησης. Επισημαίνεται ότι, μειώνοντας τη διάσταση των δεδομένων, μπορούμε να αποθηκεύσουμε πιο εύκολα μεγάλο όγκο δεδομένων.

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 $\mathbf{X}=\left(X_1, X_2, ..., X_k\right)$ από τις οποίες θέλουμε να δημιουργήσουμε τις κύριες συνιστώσες $\mathbf{Y}=\left(Y_1, Y_2, ..., Y_k\right)$. Οι τελευταίες είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών, δηλαδή:

$$
\begin{aligned}
Y_1 &= \alpha_{11}X_1 + \alpha_{12}X_2 + ... + \alpha_{1k}X_k \\
Y_2 &= \alpha_{21}X_1 + \alpha_{22}X_2 + ... + \alpha_{2k}X_k \\
&..... \\
Y_k &= \alpha_{k1}X_1 + \alpha_{k2}X_2 + ... + \alpha_{kk}X_k
\end{aligned}
$$

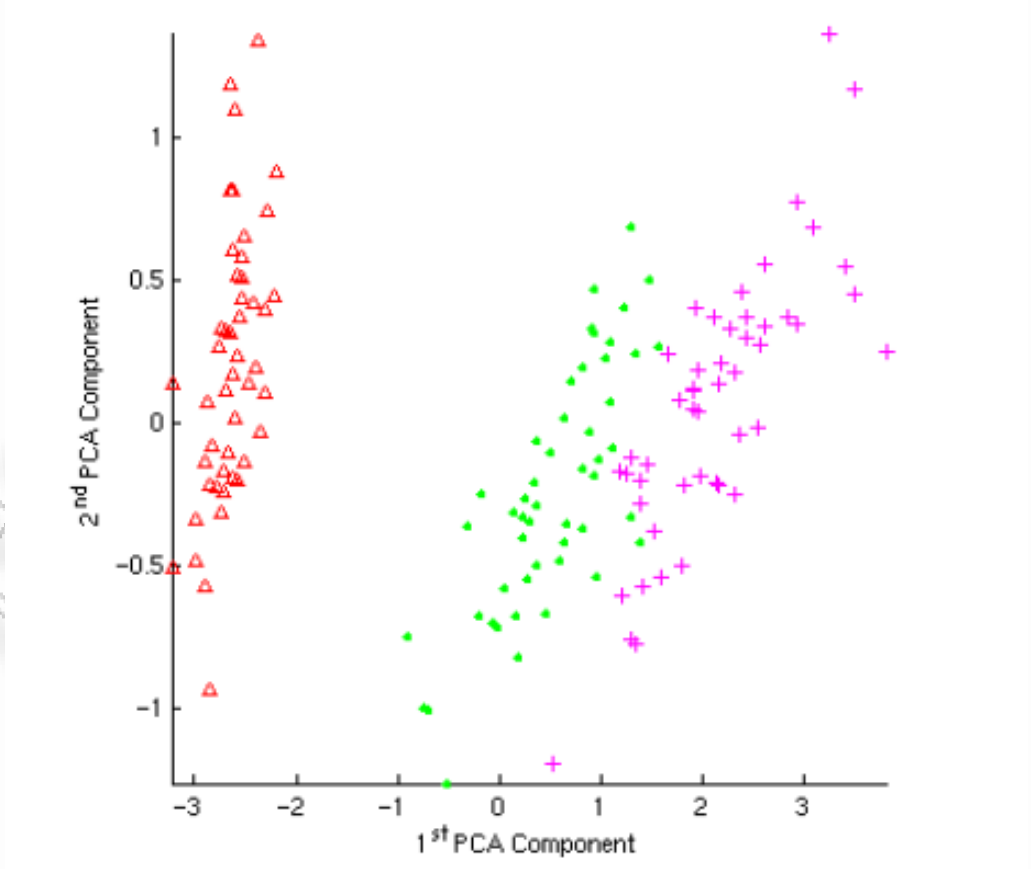
ή με μορφή πινάκων $\mathrm{Y} = \mathrm{AX}$, όπου Α είναι ο τετραγωνικός πίνακας $\mathrm{A} = \left(\alpha_{ij}\right)$ με διάσταση $k \times k$. Η προταθείσα από τον Pearson μεθοδολογία αφορά στον εντοπισμό των στοιχείων του τετραγωνικού πίνακα Α έτσι ώστε οι κύριες συνιστώσες να είναι σε φθίνουσα σειρά ως προς τη διακύμανση, δηλαδή η πρώτη συνιστώσα να έχει τη μεγαλύτερη διακύμανση, η δεύτερη να έχει τη δεύτερη μεγαλύτερη διακύμανση κοκ. Αποδεικνύεται ότι αν Σ είναι ο θετικά ορισμένος πίνακας συνδιακύμανσης του διανύσματος $\boldsymbol{X}$, και $\left(\lambda_1, \boldsymbol{\varepsilon}_1\right), ..., \left(\lambda_k, \boldsymbol{\varepsilon}_k\right)$ τα ζεύγη ιδιοτιμών και ιδιοδιανυσμάτων του Σ με $\lambda_1 \geq ..... \geq \lambda_k$, τότε οι γραμμικοί συνδυασμοί $Y_1 = \boldsymbol{\varepsilon}_1^{'}\mathbf{X}, ..., Y_k = \boldsymbol{\varepsilon}_k^{'}\mathbf{X}$ είναι οι κύριες συνιστώσες, ασυσχέτιστες μεταξύ τους και με συνολική διασπορά ίση με τη συνολική διασπορά των αρχικών μεταβλητών. Μπορούμε στη συνέχεια να υπολογίσουμε πολύ εύκολα το ποσοστό συμμετοχής της κάθε συνιστώσας στην ερμηνεία της συνολικής διασποράς των δεδομένων με στόχο να βρούμε έναν σχετικά μικρό αριθμό αυτών που να εξηγούν μεγάλο μέρος της διακύμανσης.

Αξίζει να σημειωθεί ότι η εφαρμογή της μεθόδου μπορεί να γίνει είτε στον πίνακα συνδιακύμανσης είτε στον πίνακα συσχέτισης, με διαφορετικά πάντως αποτελέσματα. Συνήθως για την οπτικοποίηση των δεδομένων επιλέγεται ο πίνακας συσχέτισης. Η επιλογή πάντως του αριθμού των κύριων συνιστωσών δε γίνεται με μοναδικό τρόπο, αλλά έχουν προταθεί διάφορα κριτήρια όπως το scree plot, το κριτήριο του Kaiser, η μέθοδος broken stick, η μέθοδος του Velicer, η μέθοδος cross validation, έλεγχοι υποθέσεων ή η κανονική προσέγγιση.

Αυτό που μας ενδιαφέρει στη χρήση της Ανάλυσης Κύριων Συνιστωσών για την οπτικοποίηση είναι να κρατήσουμε τις δύο ή τις τρείς πρώτες συνιστώσες ώστε να είμαστε σε θέση να αναπαραστήσουμε τα δεδομένα σε χώρο δύο ή τριών διαστάσεων αντίστοιχα, ελπίζοντας πάντα ότι αυτές οι συνιστώσες θα ερμηνεύουν το μεγαλύτερο μέρος της συνολικής διακύμανσης χωρίς να έχουμε μεγάλη απώλεια πληροφορίας των δεδομένων.

Εφαρμογές της μεθόδου Ανάλυσης Κύριων Συνιστωσών μπορούν να γίνουν με αρκετά προγράμματα, μερικά από τα οποία είναι το Matlab, το S-plus, η R, το ViSta (http://www.geokon.com/products/dataloggers.php) και το SciLab (http://www.scilab.org/).

Πιο κάτω δίνουμε ένα παράδειγμα Ανάλυσης Κύριων Συνιστωσών στο γνωστό σύνολο Iris data.



7.4 Biplots

Το Biplot εισήχθηκε από τον Gabriel (1971) ως ένα γραφικό εργαλείο για την ταυτόχρονη αναπαράσταση των γραμμών και των στηλών ενός πίνακα δεδομένων της μορφής αντικείμενα × μεταβλητές, σε κοινό διδιάστατο χώρο. Το πρόθεμα «Bi» δεν αναφέρεται στον αριθμό των διαστάσεων αλλά στη δυνατότητα ταυτόχρονης απεικόνισης αντικειμένων και μεταβλητών του πίνακα στο ίδιο διάγραμμα. Οι μεταβλητές απεικονίζονται ως διανύσματα και τα αντικείμενα ως σημεία, ώστε η τιμή του αντικειμένου *i* στη μεταβλητή *j* να προσεγγίζεται από το εσωτερικό γινόμενο των συντεταγμένων του σημείου που αναπαριστά το αντικείμενο *i* και του διανύσματος που αντιστοιχεί στη μεταβλητή *j*. Αν *X* είναι ένας $n \times q$ πίνακας δεδομένων της μορφής αντικείμενα × μεταβλητές με $n>q$, τότε μπορεί αυτός ο πίνακας να αναλυθεί ως εξής:

$$X=AB^{\mathrm{T}}$$

όπου *A* είναι ένας $n \times q$ πίνακας που τα στοιχεία των στηλών του περιέχουν τις συντεταγμένες των *n* αντικειμένων σε ένα *p*-διάστατο ορθογώνιο σύστημα συντεταγμένων με $p \le q$, ενώ αντίστοιχα *B* είναι ένας $n \times q$ πίνακας οι γραμμές του οποίου περιέχουν τις συντεταγμένες των *q* μεταβλητών στους ίδιους *p* άξονες. Οι στήλες των *A* και *B* ονομάζονται παραγοντικοί άξονες του Biplot και τα στοιχεία του πίνακα *X* μπορούν να ληφθούν υπολογίζοντας το εσωτερικό γινόμενο των κατάλληλων γραμμών του *A* και των αντίστοιχων στηλών του B^{T} μέσω της ακόλουθης σχέσης:

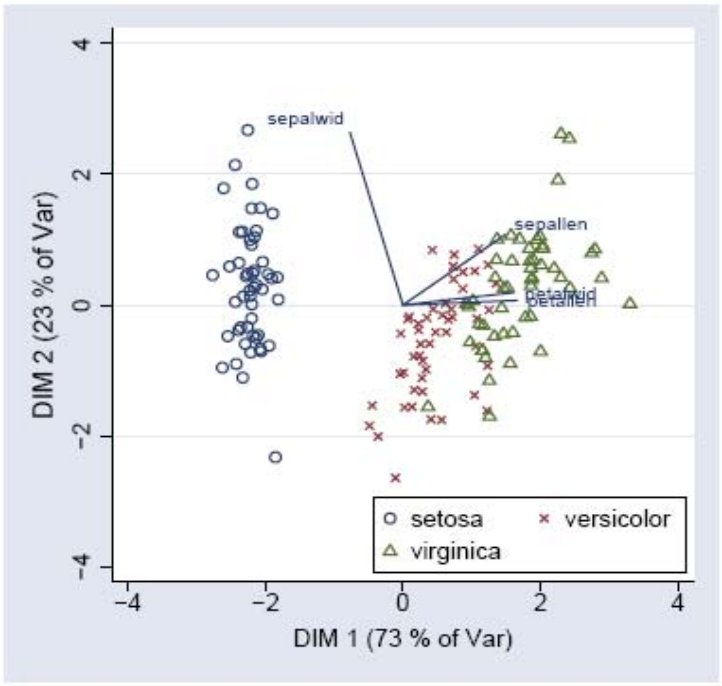
$$x_{ij} = \sum_{r=1}^{p} a_{ir} b_{rj}^{T} \text{ , με } i=1,....,n \text{ και } j=1,....,q.$$

Στη συνέχεια τα στοιχεία των πινάκων *A* και *B* υπολογίζονται εφαρμόζοντας τη μέθοδο Singular Value Decomposition στον πίνακα *X*, από όπου προκύπτουν οι χαρακτηριστικές τιμές και τα χαρακτηριστικά διανύσματα που θα χρησιμοποιηθούν για τις συντεταγμένες του Biplot. Ανάλογα με το πλήθος των χαρακτηριστικών διανυσμάτων που θα χρησιμοποιηθούν για την εκτίμηση των στοιχείων των πινάκων *A* και *B*, επιτυγχάνεται και η αντίστοιχη αναπαράσταση των στοιχείων του πίνακα *X*. Έτσι η αναπαράσταση των σημείων και των διανυσμάτων στο Biplot θα είναι η βέλτιστη δυνατή διδιάστατη προσέγγιση των αρχικών δεδομένων, δηλαδή θα

περιέχεται σε αυτό το μεγαλύτερο τμήμα της αρχικής διακύμανσης των δεδομένων στον q-διάστατο χώρο.

Η αποτελεσματικότητα του Biplot ως προς την ερμηνεία της αρχικής διακύμανσης μπορεί να μετρηθεί μέσω δεικτών καλής προσαρμογής. Το μεγαλύτερο πλεονέκτημα του Biplot είναι η ευκολία ερμηνείας του, καθώς στηρίζεται σε κάποιες σημαντικές ιδιότητες. Ο βαθμός γραμμικής συσχέτισης των μεταβλητών προκύπτει από την εξέταση του συνημιτόνου της γωνίας που σχηματίζουν οι αντίστοιχοι άξονες. Άρα αν η γωνία μεταξύ των διανυσμάτων είναι οξεία οι μεταβλητές συσχετίζονται θετικά, αν η γωνία είναι αμβλεία συσχετίζονται αρνητικά ενώ αν η γωνία είναι ορθή τότε οι μεταβλητές είναι γραμμικά ανεξάρτητες. Επίσης οι αποστάσεις των αντικειμένων όπως αυτά απεικονίζονται στους παραγοντικούς άξονες καθορίζουν το βαθμό ομοιότητας των αντίστοιχων παρατηρήσεων. Τέλος τα διανύσματα είναι προσανατολισμένα προς τα αντικείμενα για τα οποία έχουν τις μεγαλύτερες τιμές.

Γενικά η αποτελεσματικότητα του Biplot εξαρτάται σε μεγάλο βαθμό από την δυνατότητα που έχουμε να αναπαραστήσουμε επαρκώς τον πίνακα δεδομένων σε διδιάστατο χώρο χωρίς να έχουμε μεγάλη απώλεια πληροφορίας, δηλαδή να ερμηνεύεται από το γράφημα ένα αρκετά μεγάλο κομμάτι της αρχικής διασποράς των δεδομένων. Σημειώνεται ότι εφόσον τα Biplots βασίζονται στη μείωση των αρχικών διαστάσεων των δεδομένων, μπορούν να συνδυαστούν με αρκετές από τις σχετικές μεθόδους όπως την Παραγοντική Ανάλυση Αντιστοιχιών (Correspondence Analysis), την Ανάλυση Κύριων Συνιστωσών (Principal Components Analysis) ή την Πολυδιάστατη Κλιμάκωση (Multidimensional Scaling).



Τα Biplots μπορούν να κατασκευαστούν με τα προγράμματα Matlab και Stata. Στο διπλανό σχήμα δίνουμε ένα παράδειγμα Biplot για τα Iris data.

7.5 Πολυδιάστατη Κλιμάκωση

Η μέθοδος της πολυδιάστατης κλιμάκωσης (multidimensional scaling) εισήχθηκε από τον Torgerson (1952) και αναπτύχθηκε περαιτέρω από τους Shepard (1961) και Kruskal (1963). Ουσιαστικά πρόκειται για ένα σύνολο τεχνικών που έχουν ως σκοπό την αναπαράσταση δεδομένων από τον πολυδιάστατο χώρο σε χώρο χαμηλότερης διάστασης, συνήθως διδιάστατο, χρησιμοποιώντας τις εγγύτητες μεταξύ των παρατηρήσεων, δηλαδή τα μέτρα εκείνα που δηλώνουν το βαθμό ομοιότητας ή ανομοιότητας δύο αντικειμένων. Τα μέτρα αυτά χρησιμοποιούνται ώστε να τοποθετηθούν οι παρατηρήσεις ως σημεία σε μια επιφάνεια προβολής με τέτοιο τρόπο ώστε να αναδεικνύονται οι σχέσεις και οι δομές των δεδομένων οι οποίες δεν είναι ανιχνεύσιμες στον πολυδιάστατο χώρο. Για να επιτευχθεί αυτό θα πρέπει το μέγεθος της ομοιότητας ή της ανομοιότητας δύο παρατηρήσεων να είναι ανάλογο της απόστασης των αντίστοιχων σημείων στο χάρτη προβολής, δηλαδή όσο μεγαλύτερη η ανομοιότητα μεταξύ των παρατηρήσεων, τόσο μεγαλύτερη θα πρέπει να είναι και η απόσταση μεταξύ των σημείων.

Οι αλγόριθμοι που έχουν σχεδιαστεί για την ανάλυση των ανομοιοτήτων και κατά συνέπεια στη μείωση των διαστάσεων των δεδομένων, διακρίνονται σε δύο βασικές κατηγορίες: τη μετρική πολυδιάστατη κλιμάκωση (metric multidimensional scaling) και τη μη μετρική πολυδιάστατη κλιμάκωση (non-metric multidimensional scaling). Στη μετρική πολυδιάστατη κλιμάκωση που αποτελεί και την πρωτότυπη ιδέα, χρησιμοποιούνται οι ακριβείς τιμές των αποστάσεων μεταξύ των παρατηρήσεων για να αναπαραστήσουμε τα σημεία στην επιφάνεια προβολής. Αν κάθε πολυδιάστατο διάνυσμα $\boldsymbol{x}_k$ αντιπροσωπεύεται στο διδιάστατο χώρο από ένα διάνυσμα $\boldsymbol{x}'_k$, στόχος είναι οι αποστάσεις στον διδιάστατο χώρο να διατηρηθούν όσο το δυνατόν πιο κοντά στις αρχικές αποστάσεις των δεδομένων. Αν η απόσταση μεταξύ των $\boldsymbol{x}_k$ και $\boldsymbol{x}_l$ είναι $d(k,l)$ και η απόσταση μεταξύ των $\boldsymbol{x}'_k$ και $\boldsymbol{x}'_l$ είναι $d'(k,l)$, τότε η μέθοδος της μετρικής πολυδιάστατης κλιμάκωσης επιχειρεί να προσεγγίσει την απόσταση $d(k,l)$ με την $d'(k,l)$. Σε αυτήν την περίπτωση η συνάρτηση που καλούμαστε να ελαχιστοποιήσουμε είναι η εξής:

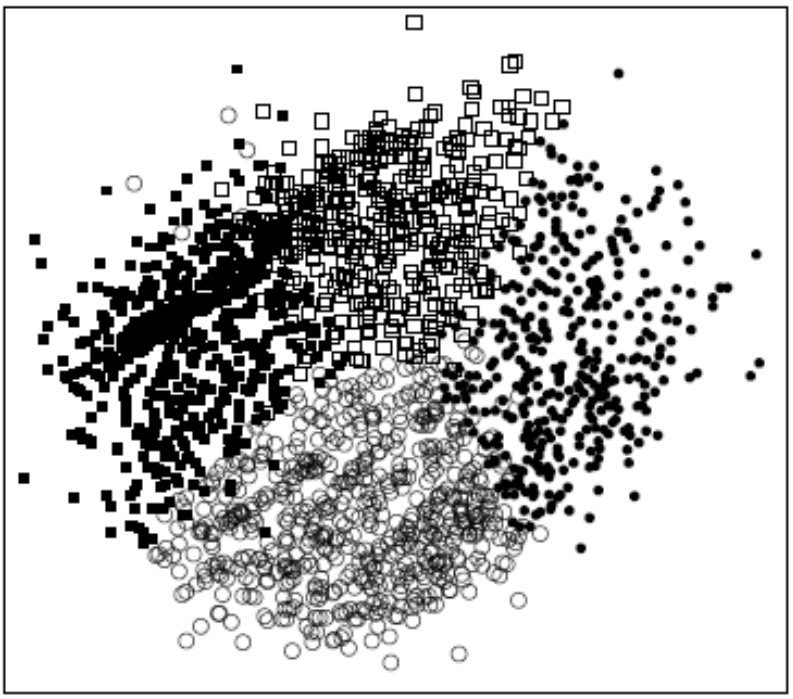
$$E_M = \sum_{k \neq 1} \left[d(k,l) - d'(k,l) \right]^2$$

Σε περιπτώσεις όπου τα δεδομένα είναι διατάξιμα δεν έχει νόημα να λαμβάνονται οι ακριβείς τιμές των αποστάσεων των διανυσμάτων αλλά χρησιμοποιούνται οι βαθμολογικές σειρές (rank orders) των αποστάσεων των διανυσμάτων. Για να επιτευχθεί αυτό εισάγεται μια μονότονα αύξουσα συνάρτηση f τέτοια ώστε να αποτυπώσει τις αποστάσεις με τιμές οι οποίες θα διατηρήσουν τη βαθμολογική σειρά. Στη μη μετρική πολυδιάστατη κλιμάκωση χρησιμοποιούνται τέτοιες συναρτήσεις και η ποσότητα που καλούμαστε να ελαχιστοποιήσουμε γίνεται:

$$E_N = \frac{1}{\sum_{k\neq 1}\left[d'(k,l)\right]^2}\sum_{k\neq 1}\left[f\left(d(k,l)\right)-d'(k,l)\right]^2$$

Παρόλο που η μέθοδος της μη μετρικής πολυδιάστατης κλιμάκωσης αναπτύχθηκε για να διαχειριστεί διατάξιμα δεδομένα, μπορεί να χρησιμοποιηθεί και σε συνεχή δεδομένα με τη διαφορά ότι προσπαθεί να διατηρήσει τη διάταξη των αποστάσεων μεταξύ των διανυσμάτων και όχι τις ακριβείς τιμές τους.

Εφαρμογές της μεθόδου μπορούν να γίνουν με το SPSS, το ViSta, την R, το XGvis (http://www.research.att.com/areas/stat/xgobi/) , το Novospark Visualizer (http://www.novospark.com/) και άλλα προγράμματα. Πιο κάτω δίνουμε ένα παράδειγμα πολυδιάστατης κλιμάκωσης όπου τα δημογραφικά δεδομένα 8 μεταβλητών 1926 νοικοκυριών, διαχωρίζονται εμφανώς σε 4 ομάδες ενώ στις παρατηρήσεις κάθε ομάδας έχει δοθεί και διαφορετικό σύμβολο.



7.6 Χάρτες του Sammon

Η μέθοδος παρουσιάστηκε από τον Sammon (1969) με σκοπό να βοηθήσει στην αναπαράσταση *n*-διάστατων διανυσμάτων από τον *n*-διάστατο χώρο σε χώρο χαμηλότερης διάστασης, έτσι ώστε να διατηρηθεί η δομή των δεδομένων και να εντοπιστούν γεωμετρικές σχέσεις σε υποσύνολα των διανυσμάτων, όπως ομαδοποιήσεις ή γραμμικές σχέσεις. Η διατήρηση της δομής των δεδομένων επιτυγχάνεται μέσω ενός αλγόριθμου ο οποίος σχεδιάζει σημεία στον χώρο χαμηλότερης διάστασης έτσι ώστε οι αποστάσεις μεταξύ τους να προσεγγίζουν τις αποστάσεις που υπάρχουν στον *n*-διάστατο χώρο.

Η διαδικασία έχει ως εξής: έστω ότι έχουμε *k* διανύσματα $\boldsymbol{X}_i$ με $i=1,2,...,k$ στον *n*-διάστατο χώρο και σε αντιστοιχία προς αυτά ορίζουμε *k* διανύσματα $\boldsymbol{Y}_i$ με $i=1,2,...,k$ στον *d*-διάστατο χώρο, όπου $d<n$ και συνήθως $d=2$, καθώς μας ενδιαφέρει περισσότερο η αναπαράσταση στο επίπεδο. Ορίζουμε την απόσταση μεταξύ των διανυσμάτων $\boldsymbol{X}_i$ και $\boldsymbol{X}_j$ να είναι d_{ij}^* και την απόσταση μεταξύ των διανυσμάτων $\boldsymbol{Y}_i$ και $\boldsymbol{Y}_j$ να είναι d_{ij}. Ως μέτρο απόστασης χρησιμοποιείται συνήθως η Ευκλείδεια απόσταση, αν και μπορεί να χρησιμοποιηθεί και κάποιο άλλο από τα γνωστά μέτρα, ανάλογα με τα δεδομένα. Στη συνέχεια θέτουμε τα διανύσματα Y_i στον *d*-διάστατο χώρο έτσι ώστε:

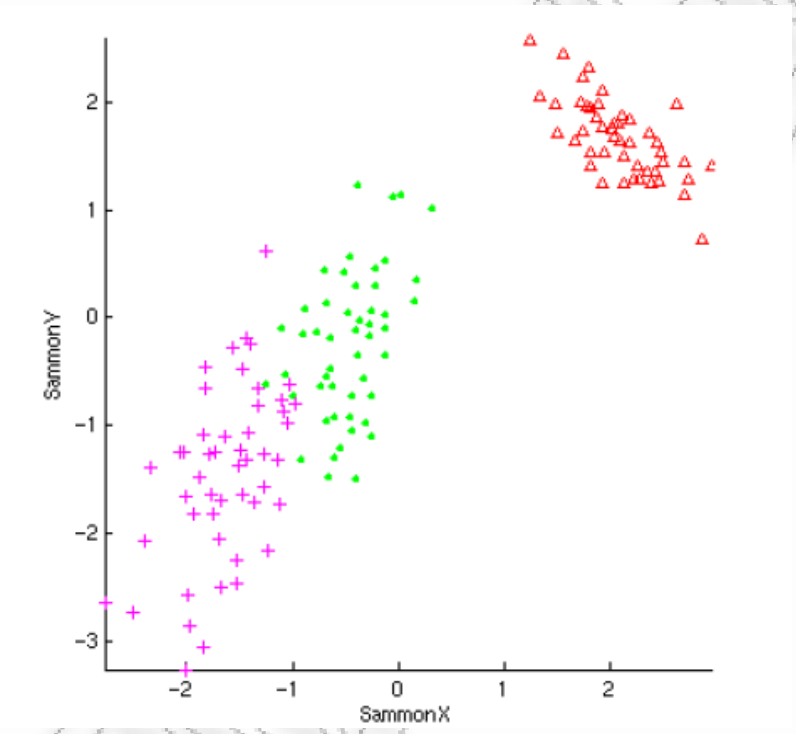
$$Y_1 = \begin{bmatrix} y_{11} \\ . \\ . \\ y_{1d} \end{bmatrix}, \; Y_2 = \begin{bmatrix} y_{21} \\ . \\ . \\ y_{2d} \end{bmatrix}, \; \ldots\ldots, \; Y_k = \begin{bmatrix} y_{k1} \\ . \\ . \\ y_{kd} \end{bmatrix}$$

Έπειτα υπολογίζουμε όλες τις αποστάσεις d_{ij} τις οποίες χρησιμοποιούμε για να ορίσουμε ένα σφάλμα *E* το οποίο δίνει μια εικόνα του πόσο καλά ταιριάζει ο σχηματισμός των k σημείων στον *d*-διάστατο χώρο με τα *k* σημεία στον *n*-διάστατο χώρο. Το σφάλμα αυτό είναι συνάρτηση των *d*x*k* μεταβλητών y_{pq} όπου $p=1,...,k$ και $q=1,...,d$ και υπολογίζεται ως εξής:

$$E = \frac{1}{\sum_{i<j} \left[d_{ij}^* \right]} \sum_{i<j}^{N} \frac{\left[d_{ij}^* - d_{ij} \right]^2}{d_{ij}^*}$$

Το επόμενο βήμα του αλγόριθμου είναι να προσαρμόζουμε τις μεταβλητές y_{pq} ή ισοδύναμα να αλλάζουμε το σχηματισμό των διανυσμάτων Y_i στον d-διάστατο χώρο, ώστε να μειώνουμε το σφάλμα E μέχρι να πετύχουμε την ελαχιστοποίησή του. Το διάστημα τιμών του σφάλματος είναι [0,1], με την τιμή 0 να υποδηλώνει την τέλεια, χωρίς την παραμικρή απώλεια πληροφορίας, αναπαράσταση των δεδομένων σε χώρο χαμηλότερης διάστασης.

Οι χάρτες του Sammon μπορούν να κατασκευαστούν με την R και με το Matlab. Παρακάτω δίνουμε ένα παράδειγμα της μεθόδου για τα δεδομένα Iris data.



7.7 Χάρτες Αυτο-Οργάνωσης

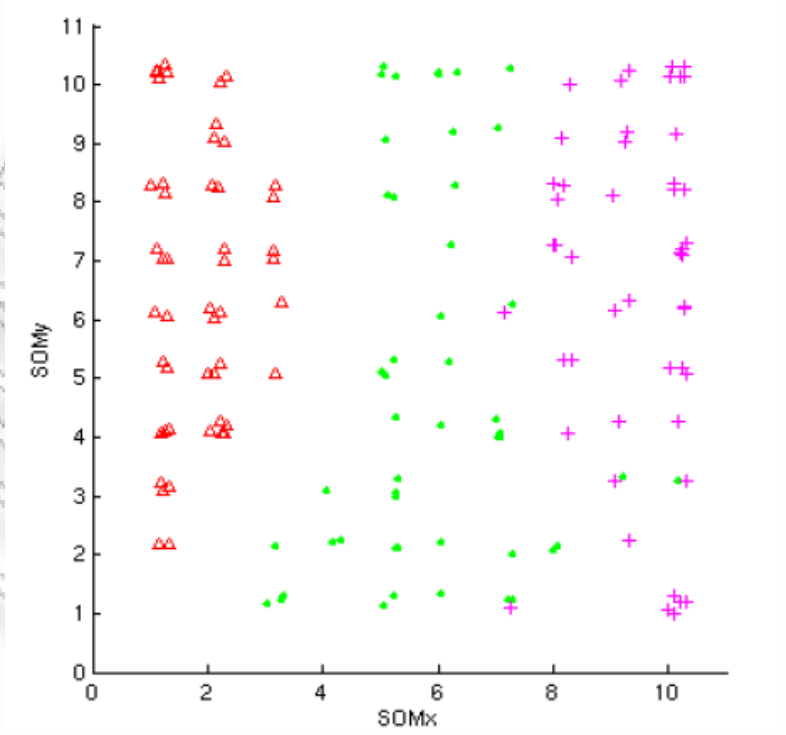
Οι χάρτες αυτο-οργάνωσης (self-organizing maps) παρουσιάστηκαν από τον Kohonen (1990) και πρόκειται για αλγόριθμο νευρωνικών δικτύων ο οποίος χρησιμοποιείται για να ομαδοποιήσει τα δεδομένα και να αναπαραστήσει στη συνέχεια τις ομάδες στο διδάστατο χώρο. Ένα n-διάστατο σύνολο δεδομένων χρησιμοποιείται για να εκπαιδεύσει ένα νευρωνικό δίκτυο όπου οι n διαστάσεις χρησιμοποιούνται ως n κόμβοι εισόδου στο δίκτυο. Οι κόμβοι εξόδου συνήθως είναι τοποθετημένοι σε ένα ευθύγραμμο πλέγμα και συνδέονται με όλους τους κόμβους εισόδου. Σε κάθε νευρώνα του δικτύου αποδίδεται ένα διανυσματικό βάρος διάστασης n ενώ η διαδικασία εκπαίδευσης καθώς και τα βάρη είναι καθορισμένα

κατά τέτοιο τρόπο ώστε μόνο ένας κόμβος εξόδου να λειτουργεί για κάθε στοιχείο των δεδομένων που εισάγεται. Τα διανύσματα εισόδου που βρίσκονται πιο κοντά σε ένας κόμβο εξόδου θα έχουν ενισχυμένα βάρη ενώ αυτά που βρίσκονται πιο μακριά θα λειτουργούν σε άλλους κόμβους εξόδου.

Η εκπαίδευση του δικτύου σταματά είτε όταν παρέλθει συγκεκριμένο χρονικό διάστημα είτε όταν οι κόμβοι εξόδου σταθεροποιηθούν για όλα τα στοιχεία εισόδου. Οι κόμβοι εξόδου του δικτύου παράγουν συντεταγμένες *X* και *Y* για κάθε διάνυσμα εισαγωγής, οι οποίες μπορούν να χρησιμοποιηθούν για την αναπαράσταση σε διδιάστατο χώρο. Ο σχεδιασμός του δικτύου είναι τέτοιος ώστε παρόμοια στοιχεία των δεδομένων να αναπαρίστανται σε παρόμοιες συντεταγμένες *X* και *Y*. Έτσι τα δεδομένα μπορούν να αναπαρασταθούν στη συνέχεια σε ένα διάγραμμα διασποράς με τις συντεταγμένες *X* και *Y* που παρήχθησαν προηγουμένως από το δίκτυο.

Το μειονέκτημα της μεθόδου, όπως και των περισσοτέρων νευρωνικών δικτύων είναι ότι οι ομάδες που σχηματίζονται δε μπορούν εύκολα να περιγραφούν με τους όρους των αρχικών χαρακτηριστικών ή διαστάσεων των δεδομένων.

Χάρτες αυτο-οργάνωσης, οι οποίοι καλούνται και χάρτες Kohonen, μπορούν να κατασκευαστούν με το Visipoint (http://www.visipoint.fi/), το Viscovery (http://www.eudaptics.com/), το Rapanalyst (http://www.raptorinternational.com/) και τα Databionic ESOM Tools (http://databionic-esom.sourceforge.net/). Στη συνέχεια δίνουμε μια εφαρμογή της μεθόδου στα Iris data.



Βιβλιογραφία

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. Anderson, E. (1935). The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society,* **59**, 2–5.
3. Anderson, E. (1957). A Semigraphical Method for the Analysis of Complex Problems, *Mathematics,* **43,** 923-927.
4. Andrews, D. F. (1972). Plots of high dimensional data, *Biometrics*, **28**, 125-136.
5. Ankerst, M., Keim, D. A. and Kriegel, H. P. (1996). Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets, *Proc. Visualization '96, Hot Topic Session, San Francisco, CA.*
6. Asimov, D. (1985). The grand tour: a tool for viewing multidimensional data, *SIAM Journal on Scientific and Statistical Computing,* **6**, 128-143.
7. Balzer, M., Deussen, O. and Lewerentz, C. (2005). Voronoi treemaps for the visualization of software structures, *In Proceedings of the ACM Symposium on Software Visualization. ACM,* 165-172.
8. Beddow, J. (1990). Shape Coding of Multidimensional Data on a Microcomputer Display, *IEEE Visualization '90, San Francisco, CA*, 238-246.
9. Bederson, B. B., Shneiderman, B. and Wattenberg, M. (2002). Ordered and quantum treemaps: Making effective use of 2D space to display hierarchies, *In ACM Transactions on Computer Graphics,* **21**, 833-854.
10. Bertin, J. (1983). *Semiology of Graphics*, University of Wisconsin Press.
11. Bruls, M., Huizing, K. and vanWijk, J. J. (2000). Squarified Treemaps*, In Proceedings of the Joint Eurographics and IEEE TCVG Symposium on Visualization*, 33-42.
12. Buja, A. and Asimov, D. (1986). Grand Tour Methods: An Outline, *Computing Science and Statistics,* **17**, 63-67.
13. Buja, A., Swayne, D. F., Littman, M. L., Dean, N. and Hofmann, H. (2001). Xgvis: Interactive Data Visualization with Multidimensional Scaling, (*www.research.att.com/areas/stat/xgobi/papers/xgvis.ps.gz)*.

14. Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. (1983). *Graphical Methods for Data Analysis*, The Wadsworth statistics/Probability series.
15. Chambers, J. M., Hastie, Trevor J. eds. (1992). *Statistical Models in S,* Wadsworth & Brooks/Cole, Pacific Grove C.
16. Chernoff, H. (1973). The Use of Faces to Represent Points in *k*-Dimensional Space Graphically, *Journal of the American Statistical Association,* **68**, 361-368.
17. Cleveland, W. S. (1993). *Visualizing Data*, Hobart Press.
18. Cook, D., Buja, A., Cabrera, J. and Hurley, C. (1995). Grand Tour and Projection Pursuit, *Journal of Computational and Graphical Statistics,* **4**, 155-172.
19. du Toit, S. H. C., Steyn, A. G. W. and Stumpf, R. H. (1986). *Graphical Exploratory Data Analysis*, Springer Verlag.
20. Everitt, B. S. and Dunn, G. (2001). *Applied Multivariate Data Analysis*, Arnold Publications 2001.
21. Flury, B. and Riedwyl, H. (1981). Graphical Representation of Multivariate Data by Means of Asymmetrical Faces, *Journal of the American Statistical Association,* **76**, 757-765.
22. Friedman, J. H. and Tukey, J. W. (1974). A Projection Pursuit Algorithm for Exploratory Data Analysis, *IEEE Transactions on Computers,* **23**, 881-890.
23. Friendly, M. (1994). Mosaic Displays for Multi-Way Contingency Tables, *Journal of the American Statistical Association*, **89**, 190-200.
24. Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis, *Biometrika,* **58**, 453-467.
25. Grinstein, G. and Pickett, R. (1991). EXVIS: An Exploratory Visualization Environment, *Proceedings of Graphics Interface '91*, 254-261.
26. Hartigan, J. A. and Kleiner, B. (1981). Mosaics for contingency tables, *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface,* 268–273.
27. Hoffman, P., and Grinstein, G. (2008). *Visualizations for High Dimensional Data Mining - Table Visualizations*, Institute for Visualization and Perception Research Computer Science Department, University of Massachusetts Lowell.

28. Hoffman, P. E., Grinstein, G. G., Marx, K., Grosse, I. and Stanley, E. (1997). DNA Visual and Analytic Data Mining, *IEEE Visualization '97, Phoenix, AZ,* 437-441.
29. Huber, P. J. (1985). Projection Pursuit, *The Annals of Statistics,* **13**, 435-475.
30. Huh, M. Y. (2004). Line Mosaic Plot: Algorithm and Implementation, *COMPSTAT '2004 Symposium,* 277-285.
31. Inselberg, A. (1985). The plane with parallel coordinates, *The Visual Computer*, **5**, 69-91.
32. 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.
33. Keim, D Hao, M. C., Ladisch, J., Hsu, M. and Dayal, U. (2001). Pixel Bar Charts: A New Technique for Visualizing Large Multi-Attribute Data Sets without Aggregation, *Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01),* 113.
34. Keim, D. A. and Kriegel, H.-P. (1994). VisDB: Database Exploration using Multidimensional Visualization, *Computer Graphics & Applications,* **14**, 40-49.
35. Keim, D. A. and Kriegel, H.-P. (1996). Visualization Techniques for Mining Large Databases: A Comparison, *IEEE Transactions on Knowledge and Data Engineering,* **8**, 923-928.
36. Kleiner, B. and Hartigan, J. A. (1981). Representing Points in Many Dimensions by Trees and Castles, *Journal of the American Statistical Association,* **76**, 260-269.
37. Kohonen, T. (1990). The Self-Organizing Map, *Procedding of the IEEE,* **78**, 1464-1480.
38. Kromesch, S. and Juhász, S. (2008). *High Dimensional Data Visualization*, Department of Automation and Applied Informatics, Budapest University of Technology and Economics, Budapest, Hungary.
39. Kruskal, J. B. (1963). Nonmetric multidimensional scaling: A numerical method, *Psycometrika,* **29**, 115-129.
40. Krzanowski, W. J. (2000*). Principles of Multivariate Analysis: A User's Perspective*, Oxford University Press.

41. Kumasaka, N. and Shibata, R. (2008). High Dimensional Data Visualization, *Computational Statistics & Data Analysis,* **52**, 3616-3644.
42. LeBlanc, J., Ward, M. O. and Wittels, N. (1990). Exploring n-dimensional databases, *in Proc. Visualization '90, San Francisco, CA,* 230–239.
43. Levkowitz, H. (1991). Color Icons: Merging Color and Texture Perception for Integrated Visualization of Multiple Parameters, *Proceedings of the Visualization '91 Conference, IEEE Computer Society Press, San Diego, CA,* 22-25.
44. Lohninger, H. (1994). INSPECT, a program system to visualize and interpret chemical data, *Chemomet. Intell. Lab. Syst.*, **22***,* 147-153.
45. Lü, H. and Fogarty, J. (2008). Cascaded treemaps: examining the visibility and stability of structure in treemaps, *Proceedings of graphics interface 2008, Ontario Canada,* 259-266.
46. Marghescu, D. (2007). Multidimensional Data Visualization Techniques for Financial Performance Data: A Review, *Turku Centre for Computer Science,* TUCS Technical Report No 810.
47. Miller, J. R. (2007). Attribute Blocks: A Tool for Visualizing Multiple Continuously-Defined Attributes, *IEEE Computer Graphics & Applications,* **27**, 57-69.
48. Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space, *Philosophical Magazine,* **2***,* 559-572.
49. Pickett, R. M. and Grinstein, G. G. (1988). Iconographic Displays For Visualizing Multidimensional Data, *Systems, Man, and Cybernetics, Proceedings of the 1988 IEEE International Conference, China,* **1**, 514-519.
50. Sachinopoulou, A. (2001). Multidimensional Visualization, *Technical Research Centre of Finland.*
51. Sammon, J. W. (1969). A nonlinear mapping for data structure analysis, *IEEE Transactions on Computers,* **18**, 401–409.
52. Schonlau, M. (2003). Visualizing Categorical Data Arising in the Health Sciences Using Hammock Plots, *In Proceedings of the Section on Statistical Graphics, American Statistical Association,* CD-ROM.
53. Schonlau, M. (2002). Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams, *The Stata Journal,* **3**, 316-327.

54. Schwenke, J. R. and Fergen, B. J. (2002). Graphical Techniques for Displaying Multivariate Data Using SAS/GRAPH Software, *Observations,* ftp.**sas**.com/techsup/download/observations/obswww22/obswww22.pdf.
55. Shepard, R. N. (1961). The analysis of proximities: Multidimensional scaling with an unknown distance function. II, *Psycometrika,* **27**, 219-246.
56. Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations, *Department of Computer Science, Human-Computer Interaction Laboratory and Institute for Systems Research University of Maryland.*
57. Shneiderman, B. (1992). Tree Visualization with Tree-Maps: 2-d Space-Filling Approach, *ACM Transactions on Graphics,* **11**, 92-99.
58. 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.
59. Torgerson, S. W. (1952). Multidimensional Scaling: I. Theory and Method, *Psychometrika,* **17**, 401-419.
60. Tory, M. and Möller, T. (2004). Human Factors in Visualization Research, *IEEE Transactions on Visualization and Computer Graphics,* **10**, 72-84.
61. van Wijk, J. J. and van de Wetering, H. (1999). Cushion Treemaps: Visualization of Hierarchical Information, *IEEE Symposium on Information Visualization (INFOVIS'99), San Francisco*, 73-78.
62. van Wijk, J. J. and van Liere, R. (1993). HyperSlice: visualization of scalar functions of many variables. *IEEE Visualization Proceedings of the 4th conference on Visualization '93*, 119-125.
63. Hyper-dimensional data analysis using parallel coordinates, *Journal of the American Statistical Association,* **85**, 664-675.
64. Wegman, E. J. (2003). *Visual Data Mining*, Center for Computational Statistics, George Mason University.
65. Wegman, E. J. and Carr, D. B. (1993). *Statistical Graphics and Visualization*, Center for Computational Statistics, George Mason University.
66. Wong, P. C. and Bergeron, R. D. (1994). 30 Years of Multidimensional Multivariate Visualization*, IEEE Computer Society, Washington, DC, USA,* 3-33.