Κεφάλαιο 5: Ανάλυση Συστάδων

Σύνοψη

Η ανάλυση συστάδων διευθετεί ένα σύνολο μεταβλητών ή παρατηρήσεων σε συγκεκριμένες ομάδες οι οποίες διαθέτουν κατ΄ ιδίαν κοινά χαρακτηριστικά, ευκρινώς διαφοροποιημένα από εκείνα των άλλων ομάδων. Η απόσταση των στοιχείων στο χώρο μετρείται με τους ειδικούς συντελεστές ομοιότητας και η σύνδεσή τους προς δημιουργία συστάδων με ομοειδές περιεχόμενο τιμών εκάστη πραγματοποιείται με ειδικές μεθόδους διασύνδεσης, ιεραρχικού ή μη χαρακτήρα. Περιγράφονται διεξοδικά οι τόποι εκτίμησης των αποστάσεων των στοιχείων και η συνένωσή τους σε μικρές συστάδες με ολοένα αυξανόμενο αριθμό στοιχείων μέχρι την ολοκλήρωσή τους σε μία τελική σύνθεση συστάδας. Επισημαίνονται επίσης τα κριτήρια της ορθής επιλογής του αριθμούς των συστάδων και η αποτελεσματικότητά της επιλεγμένης μεθόδου. Η ανάλυση συστάδων δρα επικουρικά με τις αναλύσεις κοινών παραγόντων και κύριων συνιστωσών, και η μελέτη περίπτωσης της ανάλυσης συστάδων αποτελεί επέκταση και συγκερασμό των παραπάνω αναλύσεων, σχετικά με τη συσταδοποίηση των ελληνικών τυριών και την ταξιδόμηση των ερωτήσεων αναφορικά με την ίδρυση τριτοβάθμιου ιδρύματος.

Η μελέτη ταξιδόμησης των στοιχείων, όπως ήδη διαφαίνεται, απαιτεί επιτακτικά τη συνδυαστική γνώση των αναλύσεων κοινών παραγόντων και κύριων συνιστωσών και επιπρόσθετα την ανάλυση διακύμανσης μεταξύ των ομάδων για την στατιστική εκτίμηση των διαφορετικών δράσεων των μεταβλητών μεταξύ των ομάδων.

Η ανάλυση συστάδων ή ταξιδόμησης των στοιχείων (Cluster analysis) εφαρμόζεται με τέτοιο τρόπο ώστε να εντάσσονται σε ίδιες συστάδες (ομάδες) στοιχεία (παρατηρήσεις) περισσότερα όμοια μεταξύ τους παρά σε οποιεσδήποτε άλλες (Aldenderfer & Blashfield, 1984, Everitt, 1993). Αυτό επιτυγχάνεται με την επισταμένη επιλογή και διευθέτηση των στοιχείων σε ομάδες παρατηρήσεων με συγγενικά χαρακτηριστικά και με τις ακόλουθες ιδιότητες:

- Κάθε ομάδα διαθέτει ομοειδή σύσταση σε σχέση με κάποια χαρακτηριστικά, δηλαδή οι παρατηρήσεις σε αυτές έχουν τιμές σχεδόν όμοιες μεταξύ τους
- Κάθε ομάδα οφείλει να διαφέρει από τις υπόλοιπες ως προς ίδια χαρακτηριστικά, δηλαδή οι τιμές μιας ομάδας θα πρέπει να διαφέρουν σε μέγεθος κλίμακας από τις τιμές άλλων ομάδων.

Η ανάλυση συστάδων πραγματοποιείται με τη χρήση πολυάριθμων αλγορίθμων με τελείως διαφορετικές ιδιότητες μεταξύ τους ως προς τον τρόπο λειτουργίας και το βαθμό απόδοσής τους και επεξεργάζεται συστάδες οι οποίες εννοιολογικά σημαίνουν αποστάσεις μεταξύ των στοιχείων, πυκνές περιοχές με σημεία στο χώρο, ειδικές κατανομές στοιχείων κτλ. Η διαλογή των στοιχείων στις ομάδες γίνεται με τέτοιο τρόπο ώστε η σύνδεση μεταξύ δύο στοιχείων να μεγιστοποιείται στην περίπτωση που ανήκουν στην ίδια ομάδα, ειδάλλως να ελαχιστοποιείται. Με τον τρόπο αυτόν η ανάλυση συστάδων προάγει την ανεύρεση ειδικών σχέσεων μεταξύ των στοιχείων, χωρίς να παρέχει ανάλογες εξηγήσεις ή ερμηνείες, χωρίς δηλαδή να εξηγεί την ύπαρξη σχέσεων. Ένα άλλο χαρακτηριστικό της μεθόδου είναι ότι δεν απαιτεί καμία *a priori* υπόθεση για να ξεκινήσει τη διερευνητική διαδικασία στα στοιχεία γι' αυτό και δεν απαιτείται η εφαρμογή στατιστικών ελέγχων για τη σημαντικότητα των αποτελεσμάτων που εξάγονται.

Πέραν της προσεκτικής επιλογής ενός αλγόριθμου ακολουθεί και η ρύθμιση ορισμένων παραμέτρων, όπως ο τύπος μέτρησης των αποστάσεων, κάποιο αριθμητικό όριο στοιχείων που πρέπει να έχει μία συστάδα ή ο επιτρεπτός αριθμός των συστάδων τελικής αποδοχής. Επομένως, η ανάλυση ταξιδόμησης δεν αναμένεται να τελεσφορεί ως μία αυτόματη διαδικασία αλλά ως μία επαναληπτική διαδραστική διεργασία βελτιστοποίησης της ταυτοποίησης των στοιχείων με την εφαρμογή δοκιμασίας και αποτυχίας, η οποία ενδέχεται να απαιτήσει και επαναρρύθμιση των αρχικών παραμέτρων.

Η συσταδοποίηση συχνά επιχειρείται και ως ενδιάμεσο στάδιο μεταξύ της παραγοντικής ανάλυσης και της διακριτικής. Αρχικά οργανώνεται παραγοντική ανάλυση για να περιορίσει τις διαστάσεις των δεδομένων και κατ' επέκταση των μεταβλητών, διευκολύνοντας ποιοτικά την εκτέλεση της συσταδοποίησης αφού συμβάλλει επιπρόσθετα και στη μείωση της πολυσυγγραμικότητας των μεταβλητών. Ακολουθεί η ταυτοποίηση των μεταβλητών σε συστάδες και, τελικά, αναλαμβάνει η διακριτική ανάλυση να ελέγξει την προσαρμογή του μοντέλου της συσταδοποίησης και να περιγράψει στατιστικά τις συστάδες. Η χρήση της τελευταίας ενθαρρύνεται πολύ, καθόσον η συσταδοποίηση δεν διαθέτει κριτήρια μέτρησης της καλής προσαρμογής του επιχειρούμενου μοντέλου και βασίζεται ουσιαστικά στη διακριτική ανάλυση να διαπιστώσει αν οι δημιουργούμενες ομάδες είναι στατιστικά σημαντικές και επίσης αν οι μεταβλητές διαφοροποιούνται με σαφήνεια ως προς τη δράση τους μεταξύ των συστάδων. Επισημαίνεται, πάραυτα, ότι ο διαχωρισμός των δεδομένων σε ομάδες μπορεί να μην καταλήγει σε κάποια ουσιαστική ερευνητική ερμηνεία, και επομένως η ορθή επιλογή των συστάδων να έχει απλώς υποθετικό χαρακτήρα. Επιπρόσθετα, η συνεπικουρία της διακριτικής ανάλυσης προσβλέπει στην οικοδόμηση ενός προβλεπτικού μοντέλου συσταδοποίησης επιτρέποντας την εισαγωγή τιμών από νέες παρατηρήσεις και την αξιολόγηση της αξιοπιστίας διαχωρισμού των στοιχείων (μελών) στις συστάδες.

Η ανάλυση συστάδων βρίσκει πρόσφορο έδαφος στη διερεύνηση και άντληση δεδομένων (data mining), στη μάθηση των μηχανών (machine learning), στην ανάπτυξη νευρωνικών δικτύων (neural networks), στην ιατρική (ανάλυση εικόνας, συσταδοποίηση ασθενών, θεραπειών ή συμπτωμάτων), στην ψυχιατρική (σωστή διάγνωση των συμπτωμάτων όπως, παράνοια, σχιζοφρένεια, ουσιαστική για επιτυχημένη θεραπεία), στην αρχαιολογία (πέτρινα εργαλεία και αγγεία διαφορετικών εποχών), στη βιολογία (ταξιδόμηση των ειδών με βάση περιβαλλοντικά γεγονότα), στην επιστήμη τροφίμων (ταξιδόμηση προϊόντων με βάση τη χημική τους σύσταση) και στη βιοπληροφορική.

Με τον όρο της ταξιδόμησης συμπλέουν και κάποιες άλλες ονοματολογίες, όπως είναι η αυτόματη ταξινόμηση (automatic classification), αριθμητική ταξινομία, τυπολογική ανάλυση κτλ., με κύρια διαφορά ως προς την κατεύθυνση χρήσης. Για παράδειγμα, στη διερεύνηση των παρατηρήσεων το ενδιαφέρον στρέφεται στη δημιουργία ποιοτικών συστάδων ενώ στην αυτόματη ταξινόμηση στη διαχωριστική ισχύ των σχηματιζόμενων συστάδων.

Στο παρόν κεφάλαιο δίνεται ιδιαίτερη έμφαση στην ιεραρχική ταξιδόμηση επειδή αποτελεί το συνηθέστερο και περισσότερο ποικίλο στη χρήση τρόπο ομαδοποίησης των στοιχείων, ιδιαίτερα σε κάποιες σύνθετες εφαρμογές της όπως είναι η συσταδοποίηση με διπλή κατεύθυνση (δειγμάτων και μεταβλητών) και κατά δεύτερο λόγο σε κάποιες σπουδαίες μη ιεραρχικές διαδικασίες όπως είναι αυτές των k μέσων και του αλγόριθμου E-M (προσδοκίας-μεγιστοποίησης, expectation-maximization algorithm), με συναφή των μεθόδων αυτών παραδείγματα εφαρμογής.

Η ιεραρχική ανάλυση συστάδων είναι κατάλληλη για μικρό αριθμό παρατηρήσεων και μεταβλητές ίδιου τύπου (μεταξύ ποσοτικών, καταμέτρησης ή ποιοτικών διμερών), ενώ η ανάλυση συστάδων k μέσων συνιστάται στην περίπτωση μεγάλου μεγέθους δείγματος παρατηρήσεων, αλλά για ποσοτικές μόνο μεταβλητές. Η διβηματική ανάλυση συστάδων (two-step cluster analysis) είναι μία νεώτερη τεχνική, κατάλληλη επίσης για μεγάλο αριθμό παρατηρήσεων (ιδίως όταν ως μέτρο ομοιότητας μεταξύ των συστάδων χρησιμοποιείται ο λογάριθμος της μέγιστης πιθανοφάνειας των αποστάσεων), υπερτερεί μοναδικά όμως των άλλων στο χειρισμό, ταυτόχρονα, ονομαστικών και ποσοτικών μεταβλητών.

5.1 Στρατηγικές ταξιδόμησης

Τρεις διαφορετικές τακτικές διασύνδεσης των στοιχείων είναι γνωστές, η καθεμία των οποίων ακολουθεί συγκεκριμένους κανόνες επιλογής:

1. Ομαδοποίηση με άγνωστο αριθμό τελικών συστάδων, εφαρμοζόμενη είτε στον πίνακα ομοιότητας των δειγμάτων SxS είτε στον πίνακα ομοιότητας των μεταβλητών VxV (δενδρική ταξινόμηση-tree clustering).
2. Ομαδοποίηση με συγκεκριμένο αριθμό k συστάδων (ταξιδόμηση k τελικών ομάδων- k-means clustering).
3. Ταυτόχρονη ομαδοποίηση δειγμάτων και μεταβλητών (διασύνδεση διπλής κατεύθυνσης -two-way joining).

Ανεξαρτήτως επιλογής αλγορίθμων, επιζητείται πάντοτε η ίδια στρατηγική δηλαδή η δημιουργία ομάδων στοιχείων, αλλά με τη σύνθεση διαφορετικών μοντέλων ομαδοποίησης, η οποία οδηγεί αναπόφευκτα και στην επιλογή εξειδικευμένων αλγορίθμων. Έτσι, κάθε ταξιδόμηση διατηρεί τα δικά της χαρακτηριστικά, τα οποία αναλόγως μπορούμε να τυποποιήσουμε σε:

- **Μοντέλα διασύνδεσης**, όπως τα ιεραρχικά με σύνδεση διαφόρων τύπων μέτρησης των αποστάσεων.

- **Κεντρικά μοντέλα**, όπως ο αλγόριθμος των k μέσων ο οποίος αντιστοιχεί σε κάθε συστάδα ως μέσο διάνυσμα.
- **Μοντέλα κατανομής** των στοιχείων, στα οποία η ταξιδόμηση ενεργοποιείται με την εκτίμηση στατιστικών παραμέτρων όπως η πολυμεταβλητή κανονική κατανομή με τη χρήση του αλγόριθμου E-M.
- **Μοντέλα πύκνωσης**, τα οποία διασαφηνίζουν στο χώρο τα στοιχεία ως περιοχές με πυκνή ή αραιή δομή.

ή αναλόγως του βαθμού αυτοτέλειας των στοιχείων σε:

- **Ισχυρή (hard) ταξιδόμηση**, όπου κάθε στοιχείο ανήκει σε μία μόνο ομάδα.
- **Ασαφή (fuzzy) ταξιδόμηση** στην οποία κάθε στοιχείο μπορεί να ανήκει σε κάθε ομάδα κατά κάποιο βαθμό.
- ή και της φύσης των στοιχείων σε:
- **Ποσοτική** ταξιδόμηση, σε στοιχεία μετρηθέντα ποσοτικά ή με διαβαθμίσεις.
- **Ποιοτική** ταξιδόμηση, σε στοιχεία που αποτελούν δυαδικές μεταβλητές (απουσία-παρουσία).
- **Διβηματική** ταξιδόμηση σε μικτής προέλευσης στοιχεία, ποσοτικών και ονομαστικών (πολυωνυμικών) μεταβλητών.

5.2 Ιεραρχική ταξιδόμηση

Η ιεραρχική ταξιδόμηση (μοντέλο διασύνδεσης) βασίζεται στην κεντρική ιδέα ότι κάποια στοιχεία σχετίζονται περισσότερο με κάποια γειτονικά τους παρά με άλλα κείμενα μακρύτερα. Έτσι, οργανώνονται διάφοροι αλγόριθμοι που συνδέουν τα στοιχεία προς ομαδοποίηση με βάση το βάθος (μέτρο) της απόστασής τους π.χ. μία συγκεκριμένη συστάδα περιγράφεται με τη μέγιστη απόσταση που απαιτείται να συνδέσει μέρη της ολικής ομαδοποίησης. Σε διαφορετικές αποστάσεις, διαφορετικές συστάδες δημιουργούνται και το σύνολο αυτών παρίσταται τελικά με ένα ιεραρχικό δενδρόγραμμα (βλ. σχήμα 5.1). Δηλαδή, οι αλγόριθμοι δεν αποδίδουν μία απλή απόσταση των στοιχείων αλλά μία εκτεταμένη ιεραρχία συστάδων οι οποίες συνενώνονται μεταξύ τους σε κάποιες αποστάσεις. Στο δενδρόγραμμα, ο άξονας Y (οριζόντιος στο σχήμα 5.1) σημειώνει την απόσταση όπου οι ομάδες συγχωνεύονται ενώ τα στοιχεία τοποθετούνται κατά μήκος του άξονα X με τρόπο ώστε οι ομάδες να διαφοροποιούνται.

Ειδικότερα, υπό τον όρο ιεραρχική ταξιδόμηση ανελίσσεται μία ολόκληρη οικογένεια μεθόδων που διαφέρουν κυρίως στον τρόπο εκτίμησης της απόστασης σύνδεσης των στοιχείων και την επιλογή των κριτηρίων συνένωσης των συστάδων. Η βασική τους ονοματολογία έχει ως εξής:

- **Ιεραρχική** (hierarchical) ταξιδόμηση, με ομαδοποιήσεις υπό μορφή διακλάδωσης πολυσχιδών κλάδων από τον κορμό ενός δέντρου. Η ταξιδόμηση αυτή παρίσταται και υπό μορφή επικαλυπτόμενων δικτυώσεων (πλέγμα).
- **Διαιρετή** (divisive) ταξιδόμηση, εκκινώντας με όλα τα στοιχεία να αποτελούν ενιαίο σύνολο (μία μεγάλη ομάδα) και ακολούθως να αποσπώνται αυτά κατά τμήματα.
- **Συσσωρευτική** (agglomerative) ταξιδόμηση, εκκινώντας από δύο οποιαδήποτε στοιχεία και ομαδοποιώντας αυτά σε ολοένα μεγαλύτερες ομάδες με ομοειδή στοιχεία (συστάδες).
- **Μονοθεσική** (monothetic) ταξιδόμηση, ομαδοποιώντας τα στοιχεία με βάση ένα μόνο χαρακτηριστικό (μία μόνο παρούσα μεταβλητή).
- **Πολυθεσική** (polythetic) ταξιδόμηση, ομαδοποιώντας τα στοιχεία με βάση όλα τα υπάρχοντα χαρακτηριστικά (όλες τις μεταβλητές παρούσες).

Η ανάλυση ταξιδόμησης ακολουθεί κατά κανόνα τα παρακάτω βήματα:

1. Επιλογή του τύπου της ομαδοποίησης των παρατηρήσεων, αν είναι ιεραρχική ή μη.
2. Επιλογή του τύπου μέτρησης της ομοιότητας των παρατηρήσεων.
3. Επιλογή του τρόπου συνένωσης των ομάδων μεταξύ τους.
4. Επιλογή του αναγκαίου αριθμού των ομάδων των στοιχείων.

Η συνηθέστερη περίπτωση ταξιδόμησης, με βάση την προηγούμενη ονοματολογία αποτελεί το συγκερασμό ιεραρχικής, συσωρευτικής, πολυθεσικής, ποσοτικής ταξιδόμησης, η οποία και θα αποτελέσει το επίκαιρο θέμα της ενότητας αυτής.

Τα στοιχεία μίας έρευνας μπορούν να καταγραφούν σε ένα λογιστικό φύλο με δύο εναλλακτικούς τρόπους. Στον πρώτο, τα δείγματα αποτελούν τις στήλες και οι μεταβλητές τις σειρές (Πίν. 5.1) και συνιστά τον τύπο Q, και στο δεύτερο τρόπο, συμβαίνει αλληλομετάθεση μεταξύ των δειγμάτων και των μεταβλητών και συνιστά τον τύπο R (Πίν. 5.2). Ο δεύτερος τρόπος παρατηρείται συνηθέστερα στην επιστήμη τροφίμων όπου οι μεταβλητές συνιστούν τα χημικά ή οργανοληπτικά χαρακτηριστικά κτλ. και στις βιολογικές επιστήμες όπου οι μεταβλητές αντικαθίστανται με την έννοια των ειδών (Gauch & Whittaker, 1981).

	Δείγματα (στήλες)								
Μεταβλητή	S1	S2	S3	S4	S5	S6	S7	S8	S9
V1	1.2	3.4	1.2	1.2	1.1	0.3	6.6	0.3	7.8
V2	2.4	2.0	1.5	8.9	8.9	0.8	5.2	0.5	8.5
V3	3.6	3.0	9.6	5.6	6.5	1.5	6.5	0.9	8.6
V4	7.7	6.7	3.5	2.3	7.7	2.6	7.4	1.2	9.2
V5	6.8	4.5	6.2	4.5	4.5	4.5	7.3	1.4	7.4
V6	5.7	5.6	6.5	1.6	6.2	7.8	4.8	1.3	7.6
V7	0.1	8.9	2.3	0.3	0.4	9.9	5.6	1.8	8.3

Πίνακας 5.1*. Καταγραφή VxS (μεταβλητές x δείγματα). Τύπος Q.*

Η ομοιότητα των παρατηρήσεων προσεγγίζεται με την εκτίμηση του κατά πόσο κοντά κείνται δύο παρατηρήσεις μεταξύ τους. Ως μέτρηση της απόστασης της ομοιότητας τιμών δύο στοιχείων (παρατηρήσεων) μεταξύ τους χρησιμοποιούνται διάφοροι δείκτες γνωστοί ως συντελεστές μέτρησης της απόστασης ομοιότητας ή ανομοιότητας (Similarity-dissimilarity metric coefficients) των στοιχείων και διακρίνονται σε ποσοτικούς και δυαδικούς συντελεστές ανάλογα με τη φύση των αρχικών στοιχείων. Αυτοί εφαρμόζονται είτε μεταξύ των δειγμάτων είτε μεταξύ των μεταβλητών και προκύπτει έτσι κατ` αντιστοιχία, ο πίνακας των δειγμάτων SxS και ο πίνακας των μεταβλητών VxV.

	Μεταβλητές (στήλες)						
Δείγμα	V1	V2	V3	V4	V5	V6	V7
S1	1.2	2.4	3.6	7.7	6.8	5.7	0.1
S2	3.4	2.0	3.0	6.7	4.5	5.6	8.9
S3	1.2	1.5	9.6	3.5	6.2	6.5	2.3
S4	1.2	8.9	5.6	2.3	4.5	1.6	0.3
S5	1.1	8.9	6.5	7.7	4.5	6.2	0.4
S6	0.3	0.8	1.5	2.6	4.5	7.8	9.9
S7	6.6	5.2	6.5	7.4	7.3	4.8	5.6
S8	0.3	0.5	0.9	1.2	1.4	1.3	1.8
S9	7.8	8.5	8.6	9.2	7.4	7.6	8.3

Πίνακας 5.2*. Καταγραφή SxV (δείγματα x μεταβλητές). Τύπος R.*

Η συσταδοποίηση τύπου R στηρίζεται σε μετρήσεις ομοιότητας των μεταβλητών (συσχέτιση κατά Pearson, απόλυτη συσχέτιση) που εμφανίζονται μέσα στις παρατηρήσεις και περιγράφουν την παρουσία των μεταβλητών στη χωροδιάσταση των παρατηρήσεων. Η συσταδοποίηση τύπου Q βασίζεται σε μετρήσεις ομοιότητας ή ανομοιότητας μεταξύ των παρατηρήσεων (Ευκλείδεια απόσταση, Manhattan, Pearson κ.ά.) και προσβλέπει στην παρουσίαση των παρατηρήσεων στη χωροδιάσταση των μεταβλητών.

Σε αντιστοίχιση με την παραγοντική ανάλυση, ο περισσότερο γνωστός τύπος παραγοντικής ανάλυσης αφορά τον τύπο R, όπου οι παρατηρήσεις αντιμετωπίζονται ως σειρές και οι μεταβλητές ως στήλες για τα δεδομένα του προβλήματος. Στην ανάλυση του τύπου αυτού οι παράγοντες συνιστούν τις συστάδες των μεταβλητών μιας ομάδας παρατηρήσεων. Αντιθέτως, η παραγοντική ανάλυση τύπου Q, γνωστή και ως ανάστροφη παραγοντική ανάλυση, δημιουργεί συστάδες παρατηρήσεων, δηλαδή οι μεταβλητές αντιμετωπίζονται ως σειρές και οι παρατηρήσεις ως στήλες. δεν θεωρείται κατάλληλη ως τεχνική συσταδοποίησης.

Α. Συντελεστές ποσοτικών στοιχείων

1. Ευκλείδεια απόσταση:

$$d_{j,k}=\sqrt{\sum (X_{ij}-X_{ik})^2}$$

2. Μέση Ευκλείδεια απόσταση:

$$d_{j,k}=\sqrt{\frac{\sum (X_{ij}-X_{ik})^2}{i}}$$

3. Τετραγωνική Ευκλείδεια απόσταση:

$$d_{j,k}=\sum (X_{ij}-X_{ik})^2$$

4. Απόσταση Manhattan:

$$d_{j,k}=\sum |X_{ij}-X_{ik}|$$

5. Απόσταση των Bray-Curtis:

$$d_{j,k}=\frac{\sum |X_{ij}-X_{ik}|}{\sum (X_{ij}+X_{ik})}$$

6. Συντελεστής Canberra :

$$d_{j,k}=\frac{1}{i}\cdot\frac{\sum |X_{ij}-X_{ik}|}{(X_{ij}+X_{ik})}$$

7. Απόσταση του Chebyshev:

$$d_{j,k}=max_i|X_{ij}-X_{ik}|$$

8. Απόσταση του Minkowski:

$$d_{j,k}=\sum (|X_{ij}-X_{ik}|^p)^{\frac{1}{p}}$$

9. Εκθετική απόσταση:

$$d_{j,k} = \sum (|X_{ij} - X_{ik}|^p)^{\frac{1}{r}}$$

10. Απόσταση του Pearson:

$$d_{j,k} = \sqrt{\frac{\sum (X_{ij} - X_{ik})^2}{v_i}}$$

11. Τετραγωνική απόσταση του Pearson:

$$d_{j,k} = \frac{\sum (X_{ij} - X_{ik})^2}{v_i}$$

12. Ποσοστό ομοιότητας:

$$p_{j,k} = \sum min(p_{ij}, p_{ik})$$

13. Συντελεστές συσχέτισης: r του Pearson, τ του Kendall και rs του Spearman
14. Απόλυτος συντελεστής συσχέτισης του Pearson |r|

Β. Συντελεστές δυαδικών στοιχείων

Αυτοί υπολογίζονται αφού πρώτα τα στοιχεία, που έχουν τις κατηγορίες παρόν-1 και απόν-0, συγκροτηθούν σε συχνότητες σύμφωνα με τον πίνακα κατάταξης 2x2 του πίνακα 5.3.

Α. Πίνακας κατάταξης 2x2 μεταξύ των δειγμάτων

		Δείγμα 1	
		Παρόν	Απόν
Δείγμα 2	Παρόν	a	b
	Απόν	c	d

Β. Πίνακας κατάταξης 2x2 μεταξύ των μεταβλητών

		Μεταβλητή 1	
		Παρόν	Απόν
Μεταβλητή 2	Παρόν	a	b
	Απόν	c	d

N=a+b+c+d

Πίνακας 5.3. *Διευθέτηση των συχνοτήτων δυαδικών στοιχείων σε πίνακες 2x2.*

1. Συντελεστής του Jaccard:

$$S_J = \frac{a}{a+b+c}$$

2. Συντελεστής του Sorensen ή του Dice ή του Czekanowski:

$$S_S = \frac{2a}{2a + b + c}$$

3. Συντελεστής απλής συμφωνίας των στοιχείων:

$$S_m = \frac{a + d}{a + b + c + d}$$

4. Συντελεστής των Baroni-Urbani και Buser:

$$S_m = \frac{\sqrt{ad} + a}{a + b + c + \sqrt{ad}}$$

5. Συντελεστής ανάλογος της Ευκλείδειας απόστασης:

$$d = \sqrt{b + c}$$

6. Συντελεστής ανάλογος της τετραγωνικής Ευκλείδειας απόστασης:

$$d = b + c$$

7. Συντελεστής των Lance και Williams:

$$S_{LW} = \frac{b + c}{2a + b + c}$$

8. Συντελεστής των Russel και Rao:

$$S_{RR} = \frac{a}{a + b + c + d}$$

9. Συντελεστής του Kulczynski:

$$S_K = \frac{a}{b + c}$$

10. Συντελεστής του Ochiai:

$$S_O = \sqrt{\left(\frac{a}{a + b}\right)\left(\frac{a}{a + c}\right)}$$

11. Συντελεστής Y του Yules:

$$S_Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

12. Συντελεστής Q του Yules:

$$S_Q = \frac{ad - bc}{ad + bc}$$

13. Συντελεστής της διασποράς:

$$S_d = \frac{ad + bc}{(a + b + c + d)^2}$$

14. Διάφοροι άλλοι δυαδικοί συντελεστές:

$$S_1 = \frac{(b - c)^2}{N^2}$$

$$S_2 = \frac{bc}{N^2}$$

$$S_3 = \frac{b + c}{4\mathrm{N}}$$

$$S_4 = \frac{N(b + c) - (b - c)^2}{N^2}$$

Η ορολογία για αποστάσεις που εφαρμόζονται μεταξύ των δειγμάτων έχει ως εξής:
Xij = τιμή που αντιστοιχεί στο δείγμα j της μεταβλητής i
Xik = τιμή που αντιστοιχεί στο δείγμα k της μεταβλητής i
i = πλήθος των μεταβλητών
vi= διακύμανση της μεταβλητής i
p= παράμετρος που ορίζεται από τον ερευνητή
r = παράμετρος που ορίζεται από τον ερευνητή
pij = ποσοστό της μεταβλητής i που αντιστοιχεί στο δείγμα j
pik=ποσοστό της μεταβλητής i που αντιστοιχεί στο δείγμα k

Στην παραπάνω ορολογία οι όροι αντιμετατίθενται όταν οι αποστάσεις ομοιότητας εφαρμόζονται μεταξύ των μεταβλητών.

Στις ποσοτικές μεταβλητές, η διασύνδεση των στοιχείων σε ομάδες, εφαρμόζεται συνήθως με συγκεκριμένες αποστάσεις ομοιότητας (ή ανομοιότητας), οι σημαντικότερες των οποίων είναι:

1. Η Ευκλείδεια απόσταση έχει το πλεονέκτημα ότι η απόσταση μεταξύ δύο οποιωνδήποτε στοιχείων δεν επηρεάζεται από την ύπαρξη στοιχείων με μεγάλες αποστάσεις (ακραίες τιμές), σε αντίθεση με το συντελεστή συσχέτισης που επηρεάζεται έντονα.
2. Η τετραγωνική Ευκλείδεια απόσταση χρησιμοποιείται όταν επιθυμούμε να προσδώσουμε μεγαλύτερο βάρος σε στοιχεία που σχετικά είναι απομακρυσμένα μεταξύ τους.
3. Η απόσταση Manhattan εξάγεται ως μέση τιμή όλων των διευθύνσεων και δίνει παρόμοια αποτελέσματα με

την Ευκλείδεια απόσταση. Η τεχνική αυτή δεν ανιχνεύει αν υπάρχουν μεγάλες διαφορές μεταξύ των αποστάσεων των στοιχείων (λείπει ο εκθέτης 2 στην εξίσωση).

4. Η απόσταση Chebychev η οποία μεγιστοποιεί το αποτέλεσμα της απόστασης των στοιχείων.
5. Το ποσοστό ομοιότητας ή ανομοιότητας ή δυσαρμονίας το οποίο εφαρμόζεται μόνο σε στοιχεία αναλογιών (ποσοστών).

Στις κατηγορικές μεταβλητές, οι συντελεστές Jaccard, Sorensen και Ochiai εξάγουν συγκριτικά τα ίδια περίπου αποτελέσματα και από αυτούς ο πρώτος επιλέγεται συχνότερα από τους ερευνητές. Οι πίνακες των μεταβλητών VxV, όταν η μελέτη αφορά ανάλυση των κύριων συνιστωσών (PCA), υπολογίζονται συνήθως με το συντελεστή συσχέτισης r του Pearson και με την Ευκλείδεια απόσταση. Η δεύτερη τεχνική επιλέγεται όταν επιθυμούμε να αυξήσουμε το βαθμό διαφοροποίησης μεταξύ των μεταβλητών.

Χρειάζεται επίσης λεπτομερής περιγραφή κάθε εξίσωσης και τα όρια εντός των οποίων κυμαίνονται οι συντελεστές.

Οι συντελεστές της Ευκλείδειας απόστασης και Manhattan είναι οι συχνότερα απαντώμενοι σε βιβλιογραφικά δεδομένα και των Bray-Curtis αρκετά λιγότερο λόγω κυρίως της μεγάλης υπολογιστικής ευαισθησίας που παρουσιάζει στην ύπαρξη έντονα ακραίων τιμών στα στοιχεία.

Ο συντελεστής συσχέτισης του Pearson χρησιμοποιείται με επιτυχία στην επιστήμη ταξιδόμησης των ειδών.

5.2.1 Πίνακες ομοιότητας ή ανομοιότητας (similarity or dissimilarity matrix)

Η ομαδοποίηση των στοιχείων εφαρμόζεται στους πίνακες ομοιότητας διπλής κατεύθυνσης:

- *Πίνακας ομοιότητας των δειγμάτων (σειρών)* SxS με βάση την επίδραση των μεταβλητών και χρήση κάποιου συντελεστή μέτρησης της απόστασης ομοιότητας (ή ανομοιότητας)
- *Πίνακας ομοιότητας των μεταβλητών (στηλών)* VxV με βάση την επίδραση των δειγμάτων και χρήση κάποιου συντελεστή μέτρησης της απόστασης ομοιότητας.

	S1		S2		S3		S4		S5		S6	
S2	9	8										
	11	7										
S3	12	10	22	2								
	12	1	11	0								
S4	17	10	3	8	2	13						
	3	5	21	3	18	2						
S5	8	4	6	18	12	8	9	9				
	22	1	11	0	11	4	9	8				
S6	18	12	8	12	20	2	17	8	4	12		
	4	0	10	5	7	6	3	7	11	8		
S7	10	10	4	18	12	11	1	16	5	17	15	9
	14	1	13	0	11	1	15	3	12	1	11	0

Πίνακας 5.4*. Πίνακες κατάταξης 2x2 μεταξύ 7 κατηγορικών δειγμάτων (πίνακας SxS).*

	S1	S2	S3	S4	S5	S6
S2	0.321					
S3	0.353	0.629				
S4	0.567	0.094	0.061			
S5	0.235	0.171	0.387	0.333		
S6	0.543	0.267	0.690	0.607	0.147	
S7	0.294	0.114	0.353	0.031	0.147	0.429

Πίνακας 5.5. *Πίνακας ομοιότητας των δειγμάτων SxS του πίνακα 5.4 μετρηθέντων με το συντελεστή του Jaccard.*

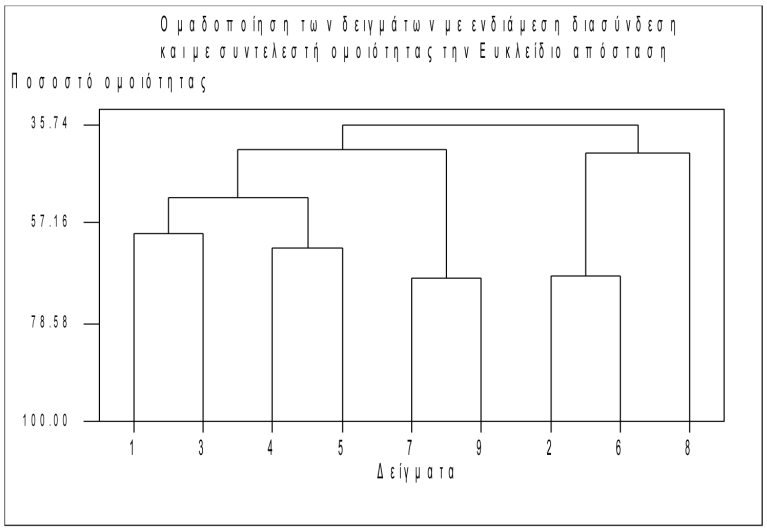
Στον πίνακα 5.4 δείχνεται η καταγραφή των συχνοτήτων σε πίνακα 2x2, 7 δειγμάτων δυαδικής μορφής και στον πίνακα 5.5, ο πίνακας ομοιότητας των δειγμάτων αυτών. Η ομοιότητα των δειγμάτων S1 και S2 με τη χρήση του συντελεστή Jaccard προκύπτει ως: 9/(9+8+11)=0,321.

Στον πίνακα 5.6 παρουσιάζεται η ομοιότητα 9 δειγμάτων ποσοτικής μορφής (Πίν. 5.2) μετρηθέντων με την Ευκλείδεια απόσταση και το δενδρόγραμμα του πίνακα ομοιότητας. Η Ευκλείδεια ομοιότητα των δειγμάτων S1 και S2 υπολογίζεται ως:

$$d_{1,2}=\sqrt{(1,2-3,4)^2+(2,4-2)^2+(3,6-3)^2+(7,7-6,7)^2+(6,8-4,5)^2+(5,7-5,6)^2+(0,1-8,9)^2}=\sqrt{89,1}=9,44$$

Στον πίνακα 5.7 παρουσιάζεται ο πίνακας ομοιότητας 7 μεταβλητών ποσοτικής μορφής (Πίν. 5.2) με τη χρήση του συντελεστή συσχέτισης του Pearson και το δενδρόγραμμα του πίνακα ομοιότητας.

	S1	S2	S3	S4	S5	S6	S7	S8
S2	9.44							
S3	7.76	10.30						
S4	9.88	12.98	10.15					
S5	7.50	11.78	9.41	7.15				
S6	11.81	5.99	11.40	14.62	14.49			
S7	8.76	7.25	9.09	10.76	8.99	11.93		
S8	10.26	11.18	11.54	10.32	13.43	10.95	13.89	
S9	13.37	10.58	12.83	14.45	11.16	14.85	5.93	19.01



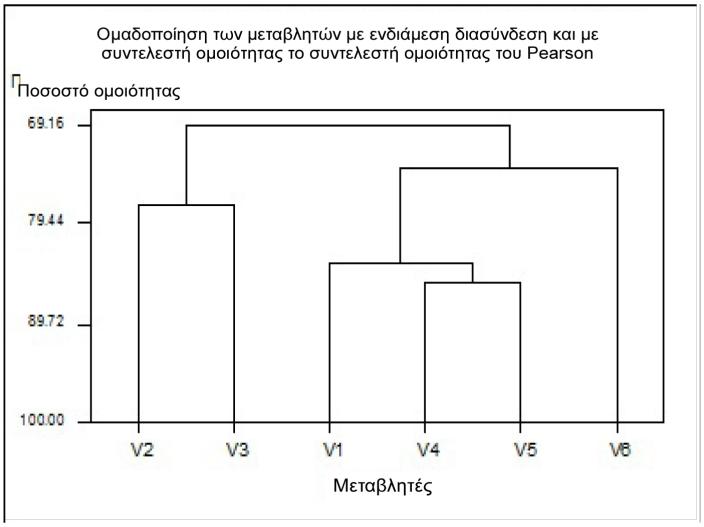
Πίνακας 5.6*. Πίνακας ομοιότητας των δειγμάτων SxS του πίνακα 5.2, μετρηθέντων με την Ευκλείδεια απόσταση και δενδρόγραμμα της ενδιάμεσης (μέσης) διασύνδεσης.*

5.2.2 Τρόποι ταξιδόμησης των στοιχείων

Η σύνδεση των ομάδων μεταξύ τους υπολογίζεται με βάση τον ήδη καταρτισθέντα πίνακα ομοιότητας των στοιχείων με την επιλογή μιας από τις παρακάτω μεθόδους διασύνδεσης:

1. Ομαδοποίηση με ***απλή διασύνδεση ή πλησιέστερης γειτνίασης*** **διασύνδεση** (single linkage or nearest neighbor linkage). Η απόσταση μεταξύ δύο ομάδων προσδιορίζεται από την απόσταση των δύο κοντινότερων στοιχείων που το καθένα ανήκει σε διαφορετική ομάδα. Η μέθοδος αυτή τείνει να δημιουργεί μεγάλο αριθμό διακλαδιζόμενων ομάδων.
2. Ομαδοποίηση με ***πλήρη διασύνδεση ή απομακρυσμένης γειτνίασης*** **διασύνδεση**(complete linkage or furthest neigbor linkage). Η απόσταση μεταξύ δύο ομάδων προσδιορίζεται από την απόσταση των δύο πλέον απομακρυσμένων στοιχείων που το καθένα ανήκει σε διαφορετική ομάδα. Η μέθοδος αυτή είναι κατάλληλη σε στοιχεία που εμφανίζουν φυσικώς ευδιάκριτες δέσμες διαφοροποίησης.
3. Ομαδοποίηση με ***μη σταθμισμένη κατά ζεύγη μέση διασύνδεση ή μέση πλήρη διασύνδεση*** (unweighted pair-group average linkage or average complete linkage). Η απόσταση μεταξύ δύο ομάδων υπολογίζεται ως η μέση απόσταση μεταξύ όλων των ζευγών των στοιχείων στις δύο διαφορετικές ομάδες. Η μέση ή ενδιάμεση διασύνδεση αποτελεί πλεονεκτικό συνδυασμό των δύο προηγουμένων.

	V1	V2	V3	V4	V5	V6
V2	0.417					
V3	0.487	0.551				
V4	0.687	0.445	0.429			
V5	0.657	0.315	0.675	0.709		
V6	0.291	-0.030	0.329	0.545	0.588	
V7	0.475	-0.220	-0.140	0.170	0.175	0.545



Πίνακας 5.7*. Πίνακας ομοιότητας των μεταβλητών VxV του πίνακα 5.2, μετρηθέντων με το συντελετή συσχέτισης r του Pearson και δενδρόγραμμα της ενδιάμεσης διασύνδεσης.*

4. Ομαδοποίηση με ***σταθμισμένη κατά ζεύγη μέση διασύνδεση*** (weighted pair-group average linkage or weighted average linkage), γνωστή και ως ***ομαδοποίηση του McQuitty*** Η απόσταση μεταξύ δύο ομάδων υπολογίζεται, όπως και προηγουμένως, με την προσθήκη του μεγέθους κάθε ομάδας (αριθμός στοιχείων ανά ομάδα) ως συντελεστή στάθμισης. Η μέθοδος αυτή αντικαθιστά την προηγούμενη όταν τα μεγέθη των ομάδων εμφανίζονται ιδιαίτερα άνισα.
5. Ομαδοποίηση με ***μη σταθμισμένη κεντροειδή διασύνδεση*** (average centroid linkage or unweighted pair-goup centroid). Η απόσταση μεταξύ δύο ομάδων υπολογίζεται από τη διαφορά της απόστασης μεταξύ των δύο κεντρικών σημείων. Το κεντρικό σημείο μιας ομάδας είναι το ενδιάμεσο σημείο που ορίζεται από το σύνολο των διαστάσεων (μεταβλητών) που συμμετέχουν στην ομαδοποίηση και αντιστοιχεί στο κέντρο βάρους της ομάδας.
6. Ομαδοποίηση με ***σταθμισμένη κεντροειδή διασύνδεση*** (weighted average centroid linkage or weighted pair-goup centroid). Αν οι σχηματιζόμενες ομάδες συντίθενται από άνισο αριθμό στοιχείων, τότε εισάγεται στην προηγούμενη μέθοδο και ένας συντελεστής στάθμισης που λαμβάνει υπόψη το διαφορετικό μέγεθος των ομάδων.
7. Ομαδοποίηση κατά **Ward**. Βασίζεται στην εφαρμογή της ανάλυσης της διακύμανσης στις παρατηρήσεις των ομάδων με σκοπό την εκτίμηση των αποστάσεων μεταξύ των ομάδων. Ουσιαστικά, η μέθοδος αυτή αποσκοπεί στην ελαχιστοποίηση της μεταβλητότητας μεταξύ δύο εξεταζόμενων ομάδων που σχηματίζονται σε κάθε διαδοχικό στάδιο της ιεραρχικής ταξιδόμησης των ομάδων. Θεωρείται ως η πλέον αποτελεσματική μέθοδος, έχει όμως το μειονέκτημα να σχηματίζει ομάδες πολύ μικρού μεγέθους.

5.2.3 Διαδικασία υπολογισμού διασύνδεσης των στοιχείων

Οποιαδήποτε μέθοδος ταξιδόμησης των στοιχείων μπορεί να επιλεγεί και να δημιουργηθεί με οποιαδήποτε απόσταση ομοιότητας. Εννοείται ότι, διαφορετικές αποστάσεις ομοιότητας δίνουν διαφορετικά αποτελέσματα διασύνδεσης ακόμα και αν η ίδια μέθοδος διασύνδεσης επιλεγεί. Στον πίνακα 5.8 παρίστανται 6 δειγματοληπτικές μονάδες και 5 μεταβλητές, από τον οποίο υπολογίζοντας την Ευκλείδεια απόσταση, προκύπτει ο πίνακας ομοιότητας SxS.

S	V1	V2	V3	V4	V5
S1	4,8	6,1	5,8	6,4	4,0
S2	4,7	6,2	6,3	6,2	3,9
S3	4,5	6,3	7,5	6,0	4,8
S4	4,9	4,4	6,0	7,2	5,9
S5	5,1	6,6	6,6	6,6	5,0
S6	5,6	6,8	3,9	5,9	5,4

Πίνακας ομοιότητας των Ευκλείδιων αποστάσεων

	S1	S2	S3	S4	S5
S2	0,57				
S3	1,95	1,53			
S4	2,68	2,89	2,94		
S5	1,42	1,33	1,29	2,53	
S6	2,64	3,04	3,85	3,55	2.87

Πίνακας 5.8*. Φύλλο εργασίας σειρών–στηλών (ειδών-μεταβλητών) και πίνακας ομοιότητας SxS της Ευκλείδειας απόστασης.*

Κρίνεται σκόπιμο, πριν την περιγραφή των τεχνικών διασύνδεσης, η παράθεση ορισμένων στοιχείων ορολογίας για την πληρέστερη κατανόηση των εννοιών:

- Πλησιέστερη απόσταση D_{ij} είναι η ελάχιστη από τις $n_i n_j$ αποστάσεις (βλ. πίνακα 5.8 των αποστάσεων) μεταξύ κάθε στοιχείου της ομάδας i και j.
- Μεγαλύτερη απόσταση D_{ij} είναι η μέγιστη από τις $n_i n_i$ αποστάσεις μεταξύ κάθε στοιχείου της ομάδας i και j.
- Μέση απόσταση D_{ij} είναι η απόσταση μεταξύ των αποστάσεων των ζευγών των μονάδων στις ομάδες i και j.
- Διάμεση απόσταση D_{ij} είναι η απόσταση της διαμέσου των αποστάσεων των ζευγών των μονάδων στις ομάδες i και j, και μειώνει τη δράση των ύποπτων τιμών (μακρινές τιμές συγκριτικά με τις άλλες)
- Κέντρο βάρους D_{ij} είναι η απόσταση μεταξύ των κέντρων βάρους των ομάδων i και j ή των μέσων τιμών αυτών χωριστά ανά ομάδα.
- Ελάχιστη διακύμανση του Ward είναι το μεταξύ των ομάδων άθροισμα των τετραγώνων των τιμών αποκλίσεων από το μέσο όρο των μεταβλητών που ανήκουν στις ομάδες i και j.

Μέθοδος της πλησιέστερης γειτνίασης

Η διαδικασία για τη συγκρότηση των ομάδων με τη μέθοδο της πλησιέστερης γειτνίασης περιγράφεται ως εξής:

1. Εκκινούμε με n ομάδες όσες είναι και οι δειγματοληπτικές μονάδες (παρατηρήσεις), δηλαδή κάθε ατομική παρατήρηση είναι και μία ομάδα.
2. Επιλέγουμε από τον πίνακα των αποστάσεων ομοιότητας τη μέγιστη τιμή ομοιότητας (δηλαδή στην ουσία τη μικρότερη απόσταση) και συνδέουμε το ζεύγος των επικείμενων δειγματοληπτικών μονάδων, ας υποθέσουμε i και j, σε μία πρώτη ομάδα αφήνοντας έτσι n-1 ομάδες ελεύθερες.
3. Επανερχόμαστε στον πίνακα των αποστάσεων ομοιότητας και ελέγχουμε τις αποστάσεις της νέας ομάδας με κάθε άλλη μονάδα k, την μέγιστη τιμή ομοιότητας των οποίων επιλέγουμε ως μέτρο σύνδεσης της απόστασης i και k και επίσης j και k.
4. Συνδέουμε τις δύο πλέον ομοιόμορφες ομάδες καταγράφοντας εκείνη την απόσταση από τις δύο που δίνει τη μικρότερη τιμή.
5. Συνεχίζουμε τη διαδικασία ελέγχου των τιμών ομοιότητας και συνδέουμε τις ομάδες με τη μικρότερη απόσταση κοκ. και μέχρις ότου όλες οι ομάδες να ενσωματωθούν σε μία.

Με βάση τις Ευκλείδιες αποστάσεις του πίνακα 5.8, θα έχουμε τα εξής:

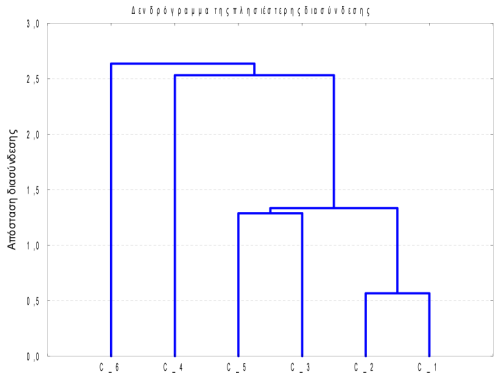
1. Από τον πίνακα των αποστάσεων ομοιότητας η μικρότερη απόσταση βρίσκεται στο ζεύγος 1 και 2 (0,57) και

οι δύο αυτές μονάδες συγχωνεύονται, σχηματίζοντας την πρώτη ομάδα και χρησιμοποιώντας ως κριτήριο ομαδοποίησης την τιμή 1.

2. Από τον πίνακα προκύπτει ότι η επόμενη μικρότερη απόσταση 1,29 αντιστοιχεί στο ζευγάρι 3 και 5. Έτσι συγχωνεύουμε τις μονάδες 3 και 5 σε μία δεύτερη ομάδα χρησιμοποιώντας ως κριτήριο ομαδοποίησης την τιμή 2 (Πίν. 5.9).
3. Η επόμενη μικρότερη απόσταση του πίνακα εντοπίζεται στο ζευγάρι των μονάδων 2 και 5 (1,33), οι οποίες ήδη αποτελούν μέλη των ομάδων (1+2) και (3+5) και η τιμή 3 ορίζεται ως το κριτήριο ομαδοποίησης. Έτσι, οι δύο ομάδες συγχωνεύονται δημιουργώντας τη νέα ομάδα (1+2+3+5).
4. Στη συνέχεια ελέγχουμε την ελάχιστη απόσταση μεταξύ της προηγούμενης ομάδας και των μονάδων 4 και 6, όπου προκύπτει ότι η μονάδα 4 έχει την μικρότερη απόσταση (2,53) και έτσι συγχωνεύεται ώστε να προκύψει η νέα ομάδα (1+2+3+4+5) με κριτήριο ομαδοποίησης το 4.
5. Τέλος συνδέουμε τη νέα ομάδα με την μονάδα 6 (απόσταση 2,64) σε μία τελική, χρησιμοποιώντας ως κριτήριο σύνδεσης την τιμή 5.

Όλη η παραπάνω διαδικασία παρίσταται στο δενδρόγραμμα του πίνακα 5.9.

ΑΠΟΣΤΑΣΗ	S1	S2	S3	S4	S5	S6
0,57	C_1	C_2				
1,29	C_3	C_5				
1,33	C_1	C_2	C_3	C_5		
2,53	C_1	C_2	C_3	C_5	C_4	
2,64	C_1	C_2	C_3	C_5	C_4	C_6



Πίνακας 5.9*. Σχέδιο διαδοχικής συγχώνευσης των παρατηρήσεων σε ομάδες με τη μέθοδο της πλησιέστερης γειτνίασης και δενδρόγραμμα διασύνδεσης της μεθόδου.*

Μέθοδος της απομακρυσμένης γειτνίασης

Η διαδικασία της δημιουργίας ομάδων με τη μέθοδο της μακρύτερης γειτνίασης έχει ως εξής:

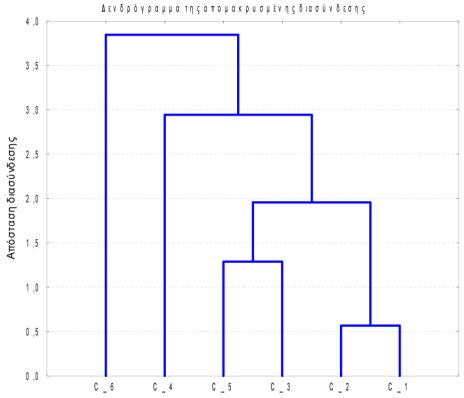
1. Εκκινούμε με n ομάδες όσες και οι δειγματοληπτικές μονάδες.
2. Επιλέγουμε από τον πίνακα των αποστάσεων ομοιότητας την ελάχιστη απόσταση (η πρώτη ομάδα πάντα υπολογίζεται με την ελάχιστη τιμή απόστασης) που αντιστοιχεί στο ζεύγος μονάδων i και j, υπολείποντας έτσι n-1 μονάδες.
3. Επανελέγχουμε τις αποστάσεις μεταξύ της ομάδας i και j με κάθε άλλη ομάδα k χρησιμοποιώντας την ελάχιστη τιμή ομοιότητας (μέγιστη απόσταση) ως μέτρο σύνδεσης.
4. Συνδέουμε τις δύο πλέον όμοιες ομάδες επιλέγοντας τη μέγιστη απόσταση μεταξύ τους και χρησιμοποιώντας ως μέτρο σύνδεσης τις μέγιστες αποστάσεις μεταξύ i και k και j και k.
5. Συνεχίζουμε τη διαδικασία ελέγχου των τιμών ομοιότητας και συνδέουμε τις ομάδες με τη μέγιστη απόσταση κοκ. και μέχρις ότου όλες οι ομάδες να ενσωματωθούν σε μία.

Η τακτική που ακολουθούμε είναι ίδια όπως με την προηγούμενη μέθοδο, με τη διαφορά ότι όταν υπολογίζουμε εκ νέου κάθε φορά την απόσταση μεταξύ κάποιου σχηματισμού ομάδας με μία νέα, χρησιμοποιούμε ως μέτρο διασύνδεσης τη μέγιστη απόσταση αντί της ελάχιστης (Πίν. 5.10):

1. Επιλέγουμε τις μονάδες με την ελάχιστη τιμή απόστασης και δημιουργούμε την πρώτη ομάδα με κριτήριο σύνδεσης την τιμή 1 και η οποία είναι η ομάδα (1+2) με απόσταση 0,57 (πίνακες 5.8 και 5.10).
2. Συγχωνεύουμε τις μονάδες 3 και 5 με μέτρο σύνδεσης την τιμή 1,29 και με κριτήριο ομαδοποίησης την τιμή

2. Επισημαίνεται ότι και εδώ χρησιμοποιείται αναγκαστικά, η μικρότερη απόσταση επειδή διαφορετικά το ζεύγος αυτό απομακρύνεται από την περαιτέρω διαδικασία.
3. Συγχωνεύουμε τις ομάδες (1+2) και (3+5) με κριτήριο ομαδοποίησης την τιμή 3 και με μέτρο σύνδεσης τη μέγιστη τιμή που προκύπτει από τα ζεύγη του πίνακα 5.8, 1-3:1,95, 1-5:1,42, 2-3:1,53, 2-5:1,33. Έτσι η νέα ομάδα (1+2+3+5) έχει απόσταση 1,95.

ΑΠΟΣΤΑΣΗ	S1	S2	S3	S4	S5	S6
0,57	C_1	C_2				
1,29	C_3	C_5				
1,95	C_1	C_2	C_3	C_5		
2,94	C_1	C_2	C_3	C_5	C_4	
3,85	C_1	C_2	C_3	C_5	C_4	C_6



Πίνακας 5.10*. Σχέδιο διαδοχικής συγχώνευσης των παρατηρήσεων σε ομάδες με τη μέθοδο της απομακρυσμένης γειτνίασης και δενδρόγραμμα διασύνδεσης της μεθόδου.*

4. Στη συνέχεια ελέγχουμε τη μέγιστη απόσταση μεταξύ της προηγούμενης ομάδας και των μονάδων 4 και 6, όπου προκύπτει ότι η μονάδα 6 έχει τη μέγιστη απόσταση (3,84), όμως για να προχωρήσει η διαδικασία της μεθόδου μέχρι το τελικό στάδιο, αναγκαστικά, επιλέγεται η μονάδα 4 (2,94) με κριτήριο ομαδοποίησης το 4, ειδάλλως περατώνεται στο σημείο αυτό.
5. Συνδέουμε όλες τις μονάδες για να σχηματιστεί μία τελική ομάδα με κριτήριο ομαδοποίησης το 5.
6. Η παραπάνω διαδικασία απλουστεύεται αν σχηματίσουμε το δενδρόγραμμα στον πίνακα 5.10.

Τα δενδρογράμματα των δύο μεθόδων διαφέρουν ουσιαστικά στο μέγεθος της κλίμακας διασύνδεσης που είναι μεγαλύτερο στην απομακρυσμένη γειτνίαση.

Τονίζεται ιδιαίτερα ότι η μέθοδος εκκινεί πάντα, συνδέοντας αρχικά όλες τις μονάδες που κατέχουν μικρή Ευκλείδεια απόσταση και εφαρμόζεται ακολούθως. Δεν μπορούμε να συνδέουμε εξαρχής μονάδες με μεγάλη απόσταση επειδή έτσι αφήνουμε αναγκαστικά εκτός όλες τις λοιπές με μικρές τιμές.

Μέθοδος της μέσης τιμής ή ενδιάμεσης διασύνδεσης (μη σταθμισμένης)

Η απόσταση μεταξύ δύο ομάδων υπολογίζεται ως η μέση απόσταση μεταξύ όλων των ζευγών των μονάδων μέσα στις δύο ομάδες, μία ανά ομάδα (Πίν. 5.11). Για παράδειγμα, από τον πίνακα 5.8 των αποστάσεων ομοιότητας και εκκινώντας πάντα με το ζεύγος των μονάδων που έχει τη μικρότερη Ευκλείδεια απόσταση, η πρώτη ομάδα σχηματίζεται από τη συγχώνευση των μονάδων 1 και 2 (0,57). Το ίδιο ισχύει και για το ζεύγος 3 και 5 (1,29). Στο επόμενο στάδιο οι δύο ομάδες, οι οποίες συνδυάζουν τη μικρότερη απόσταση, συγκροτούνται σε μία (1+2+3+5) υπολογίζοντας τη μέση τιμή διασύνδεσης του αθροίσματος των αποστάσεων D_{13}, D_{15}, D_{23} και D_{25} ως εξής: (1,95+1,42+1,53+1,33)/4=6,23/4=1,56. Στη συνέχεια ακολουθεί η ενσωμάτωση της μονάδας 4 διαιρώντας το άθροισμα των αποστάσεων D_{14}, D_{24}, D_{34} και D_{45} δια του 4: (2,68+2,89+2,94,2,53)/4=11,04/4=2,76.

Η τελική ομάδα συγκροτείται ενσωματώνοντας την μονάδα 6 ακολουθώντας την ίδια διαδικασία (3,19).

Γενικεύοντας, η απόσταση μεταξύ της ομάδας k και t δίνεται από τη μέση τιμή $n_k x n_i$ αποστάσεων, όπου n_k και n_i είναι ο αριθμός των μονάδων στις ομάδες k και t αντίστοιχα.

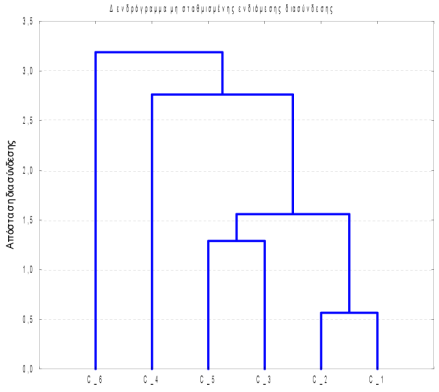
Οι τέσσερις πρώτες μέθοδοι υπολογίζονται άμεσα αριθμητικά, ενώ οι υπόλοιπες χρειάζονται τη βοήθεια της γεωμετρίας ή άλγεβρας.

Μέθοδος κεντρικής διασύνδεσης ή του κέντρου βάρους (μη σταθμισμένης)

Με τη μέθοδο αυτή κάθε ομάδα αντικαθίσταται από το κέντρο βάρους (μέση τιμή) αυτής της ομάδας. Στον πίνακα 5.12 δείχνεται το φύλλο εργασίας SxV (5x4) καθώς και οι τετραγωνικές Ευκλείδιες αποστάσεις αυτών. Η απόσταση 11 του πίνακα δίνει τη μικρότερη τιμή που αντιστοιχεί στις μονάδες S2 και S3 και έτσι δημιουργείται η πρώτη ομάδα. Η ομάδα αυτή αντιπροσωπεύεται από το κέντρο βάρους των αρχικών τιμών των αντίστοιχων μεταβλητών, δηλαδή (5+8)/2 για τη μεταβλητή V1, (7+6)/2 για τη V2, (4+4)/2 για τη V3 και (5+4)/2 για τη V4. Από το τροποποιημένο φύλο εργασίας προκύπτει ο νέος πίνακας ομοιότητας στον οποίο οι μονάδες S1 και S5 έχουν τη μικρότερη Ευκλείδεια απόσταση (13) και άρα συγκροτούν τη δεύτερη ομάδα τροποποιώντας ανάλογα και το φύλο εργασίας με τις νέες τιμές που προκύπτουν. Ο νέος πίνακας ομοιότητας δίνει τη μικρότερη απόσταση (21,5) στο κελί συντεταγμένων των δύο προηγούμενων ομάδων οι οποίες έτσι συγχωνεύονται σε μία τρίτη ομάδα (S1+S3+S2+S5). Το νέο φύλλο εργασίας το οποίο περιέχει τις μέσες τιμές των δύο προηγούμενων ομάδων διαμορφώνει και τον τελικό πίνακα στον οποίο η τρίτη ομάδα ενώνεται με τη μονάδα 4 με τιμή απόστασης 25,63 για να αποτελέσει και την τελική ομάδα.

Εύκολα γίνεται αντιληπτό, ότι η μέθοδος αυτή συνδυάζει αλλεπάλληλους υπολογισμούς και του αρχικού φύλλου εργασίας και των επικείμενων διορθωμένων αποστάσεων ομοιότητας. Αντίθετα, οι τρεις προαναφερθείσες μέθοδοι βασίζονται αποκλειστικά στον πίνακα ομοιότητας των στοιχείων.

ΑΠΟΣΤΑΣΗ	S1	S2	S3	S4	S5	S6
0,57	C_1	C_2				
1,29	C_3	C_5				
1,56	C_1	C_2	C_3	C_5		
2,76	C_1	C_2	C_3	C_5	C_4	
3,19	C_1	C_2	C_3	C_5	C_4	C_6



Πίνακας 5.11*. Σχέδιο διαδοχικής συγχώνευσης των παρατηρήσεων σε ομάδες με τη μέθοδο της ενδιάμεσης διασύνδεσης και δενδρόγραμμα διασύνδεσης της μεθόδου.*

Μέθοδος της ελάχιστης διακύμανσης του Ward

Η μέθοδος αυτή δεν υπολογίζει αποστάσεις μεταξύ των ομάδων αλλά σχηματίζει ομάδες μεγιστοποιώντας την ομοιογένεια μέσα στις ομάδες. Αυτό σημαίνει ότι ελαχιστοποιεί τη μεταβλητότητα μεταξύ των τιμών σε κάθε ομάδα με μία υπολογιστική διαδικασία γνωστή και ως σφάλμα μέσα στις ομάδες

$$ESS = \sum (X_{ij} - \bar{X}_i)^2$$

Ας θεωρήσουμε για παράδειγμα τα στοιχεία στο φύλλο εργασίας SxV (6x2) στον πίνακα 5.13. Αρχικά κάθε στοιχείο θεωρείται ως μία ομάδα και επομένως ισχύει ESS=0. Στο επόμενο βήμα θεωρούμε τις εξής μονάδες ως πέντε ομάδες: η πρώτη είναι ζευγάρι με διαφορετικούς συνδυασμούς και οι υπόλοιπες τέσσερις είναι μοναδι-

αίες (ατομικές) ομάδες. Αναφορικά με την ομάδα (S1,S2) θα έχουμε

$$ESS=(4-4{,}5)^2+(5-4{,}5)^2+(8-7{,}5)^2+(7-7{,}5)^2=1{,}0$$

όπου τα ψηφία 4,5 και 7,5 είναι οι μέσες τιμές των μονάδων S1 και S3 για τις μεταβλητές V1 και V2 αντίστοιχα. Η διαδικασία αυτή επαναλαμβάνεται για όλους τους πιθανούς συνδυασμούς των 5 ομάδων.

Από τον πίνακα 5.13 προκύπτει ότι αμφότερες οι ομάδες (S1,S2) και (S3,S5) έχουν την ελάχιστη διακύμανση (ESS=1,0) και επιλέγουμε τυχαία ως πρώτη ομάδα εκείνη με τις μονάδες S1 και S2.

Μεταβλητές

	V1	V2	V3	V4
S1	4	8	7	6
S2	5	7	4	5
S3	8	6	4	4
S4	9	5	8	8
S5	7	8	9	6

Τετραγωνισμένη Ευκλείδια απόσταση

	S1	S2	S3	S4	S5
S1	0	12	33	39	13
S2	12	0	11	45	31
S3	33	11	0	34	34
S4	39	45	34	0	18
S5	13	31	34	18	0

	V1	V2	V3	V4
S1	4	8	7	6
S2+S3	6,5	6,5	4	4,5
S4	9	5	8	8
S5	7	8	9	6

	S1	S2+S3	S4	S5
S1	0	19,75	39	13
S2+S3	19,75	0	36,75	29,75
S4	39	36,75	0	18
S5	13	29,75	18	0

	V1	V2	V3	V4
S1+S5	5,5	8	8	6
S2+S3	6,5	6,5	4	4,5
S4	9	5	8	8

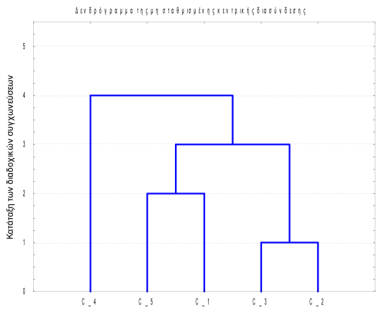
	S1+S5	S2+S3	S4
S1+S5	0	21,5	25,25
S2+S3	21,5	0	36,75
S4	25,25	36,75	0

	V1	V2	V3	V4
S1+S5 S2+S3	6	7,25	6	5,25
S4	9	5	8	8

	S1+S5 S2+S3	S4
S1+S5 S2+S3	0	25,63
S4	25,63	0

Σχέδιο διαδοχικής συγχώνευσης των παρατηρήσεων σε ομάδες με τη μέθοδο της κεντρικής διασύνδεσης και δενδρόγραμμα διασύνδεσης της μεθόδου.

ΑΠΟΣΤΑΣΗ	S1	S2	S3	S4	S5
11	C_2	C_3			
13	C_1	C_5			
21,5	C_1	C_5	C_2	C_3	
25,63	C_1	C_5	C_2	C_3	C_4



Πίνακας 5.12*. Φύλλα εργασίας ειδών-στηλών με τις τετραγωνικές Ευκλείδιες αποστάσεις τους και τρόπος ενσωμάτωσης των δειγματοληπτικών μονάδων (παρατηρήσεων).*

Ακολούθως, διαμορφώνουμε τους νέους συνδυασμούς 4 ομάδων πλέον συνδέοντας την ομάδα (S1,S2) με τις υπόλοιπες μονάδες εκλαμβάνοντας τες είτε ως μοναδιαίες ομάδες (ESS=0) είτε ως ανά δύο (βλ. πίνακα 5.13, π.χ. (S1,S2,S3), (S1,S2,S4), (S1,S2,S5), (S1,S2,S6). Έτσι, για την ομάδα (S1,S2,S3) προκύπτει,

$$ESS=(4-5,7)^2+(5-5,7)^2+(8-5,7)^2+(8-6,33)^2+(7-6,33)^2+(4-6,33)^2=17,34$$

εξίσωση, όπου οι τιμές 5,7 και 6,33 είναι οι μέσες τιμές των μονάδων S1, S2, S3 στις μεταβλητές V1 και V2 αντίστοιχα. Με τον τρόπο αυτό προκύπτει συνολικά ο πίνακας με όλους τους πιθανούς συνδυασμούς των τεσσάρων ομάδων ο οποίος προτείνει τη λύση 6 όπου έχουμε την μικρότερη τιμή ESS=2,0.

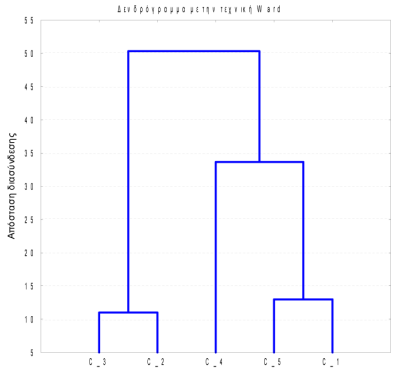
Η παραπάνω διαδικασία επαναλαμβάνεται και με όλους τους συνδυασμούς των τριών ομάδων κοκ. μέχρις ότου όλες οι ομάδες συνενωθούν σε μία τελική.

	1	2	3	4	5	ESS
Α/Α	ΟΛΟΙ ΟΙ ΣΥΝΔΥΑΣΜΟΙ 5 ΟΜΑΔΩΝ					
1	S1,S2	S3	S4	S5	S6	1,0
2	S1,S3	S2	S4	S5	S6	16,0
3	S1,S4	S2	S3	S5	S6	14,5
4	S1,S5	S2	S3	S4	S6	9,0
5	S1,S6	S1	S3	S4	S5	8,5
6	S2,S3	S1	S4	S5	S6	9,0
7	S2,S4	S1	S3	S5	S6	8,5
8	S2,S5	S1	S3	S4	S6	4,0
9	S2,S6	S1	S3	S4	S5	6,5
10	S3,S4	S1	S2	S3	S6	2,5
11	S3,S5	S1	S2	S4	S6	1,0
12	S3,S6	S1	S2	S4	S5	12,5
13	S4,S5	S1	S2	S3	S6	2,5
14	S4,S6	S1	S2	S3	S5	5,0
15	S5,S6	S1	S2	S3	S4	8,5
Α/Α	ΟΛΟΙ ΟΙ ΣΥΝΔΥΑΣΜΟΙ 4 ΟΜΑΔΩΝ					
1	S1,S2,S3	S4	S5	S6		17,3
2	S1,S2,S4	S3	S5	S6		16,0
3	S1,S2,S5	S3	S4	S6		9,3
4	S1,S2,S6	S3	S4	S5		10,7
5	S1,S2	S3,S4	S5	S6		3,5
6	S1,S2	S3,S5	S4	S6		2,0
7	S1,S2	S3,S6	S4	S5		13,5
8	S1,S2	S4,S5	S3	S6		3,5
9	S1,S2	S4,S6	S3	S5		6,0
10	S1,S2	S5,S6	S3	S4		9,5

	V1	V2
S1	4	8
S2	5	7
S3	8	4
S4	9	6
S5	7	5
S6	8	9

Σχέδιο διαδοχικής συγχώνευσης των παρατηρήσεων σε ομάδες με τη μέθοδο Ward

ΑΠΟΣΤΑΣΗ	S1	S2	S3	S4	S5
11.00	C_2	C_3			
13.00	C_1	C_5			
33.67	C_1	C_5	C_4		
50.33	C_1	C_5	C_4	C_2	C_3

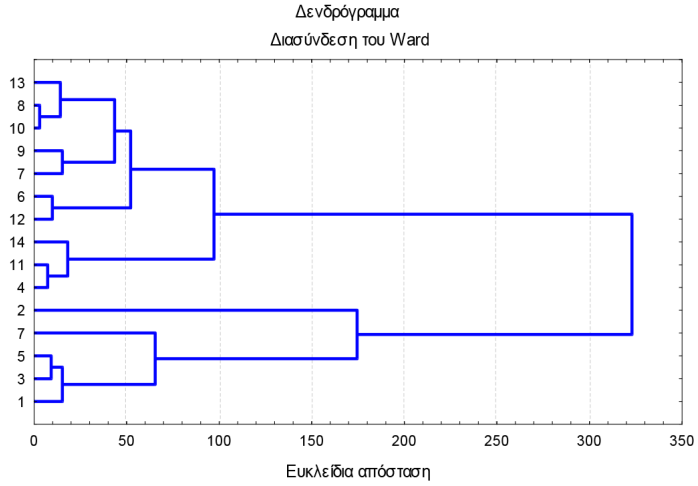


Πίνακας 5.13*. Βήματα υπολογισμού της ταξιδόμησης του Ward. Αποτελέσματα της ανάλυσης διακύμανσης στα στοιχεία του φύλλου εργασίας 6x2. Σχέδιο συγχώνευσης των παρατηρήσεων σε ομάδες με τις τετραγωνικές Ευκλείδιες αποστάσεις ομοιότητας και δενδρόγραμμα της τεχνικής του Ward. Λόγω εξοικονόμησης χώρου παραλήφθηκε η περαιτέρω διαδικασία της ανάλυσης των συνδυασμών 3 και 2 ομάδων.*

Τεχνική σύνθεσης δενδρογράμματος

Η λειτουργία των αλγόριθμων ταξιδόμησης στηρίζεται στη συνένωση στοιχείων (παρατηρήσεων) σε ολοένα μεγαλύτερες ομάδες χρησιμοποιώντας ως μέτρο σύνδεσης την ομοιότητα ή αλλιώς την απόσταση μεταξύ των στοιχείων. Ένα τυπικό αποτέλεσμα της ταξιδόμησης είναι η δημιουργία του ιεραρχικού δέντρου (Σχ. 5.1). Στο γράφημα και στα αριστερά αυτού, κάθε στοιχείο αποτελεί και μία ιεραρχική κλάση (ομάδα). Στη συνέχεια, τα στοιχεία συνδέονται όλο και περισσότερο μεταξύ τους, έτσι συγχωνεύονται σταδιακά σε μεγαλύτερες ομάδες με ολοένα και περισσότερα ανόμοια πλέον στοιχεία και τελικά όλα τα στοιχεία συνενώνονται μαζί. Ο οριζόντιος άξονας περιγράφει την κλίμακα σύνδεσης και σε κάθε κόμβο (σημείο συνένωσης δύο ομάδων) αναγνωρίζουμε την τιμή της σύνδεσης στην οποία τα συγκεκριμένα στοιχεία συνδέθηκαν μαζί για να δημιουργήσουν μία νέα ομάδα. Όταν τα στοιχεία περιέχουν μία φανερή δομή σχετικά με ομάδες ομοειδών στοιχείων τότε αυτή η δομή καταφαίνεται στο

ιεραρχικό δέντρο με τη μορφή διακριτών κλάδων. Με τον τρόπο αυτό εύκολα διακρίνουμε τις ομάδες (κλάδους) και μπορούμε να εξάγουμε συμπεράσματα με βάση την ιδιομορφία αυτή.



Σχήμα 5.1. *Τυπικό δενδρόγραμμα ταξιδόμησης 15 στοιχείων με την τεχνική του Ward και την Ευκλείδια απόσταση ομοιότητας.*

Σύνδεση προς δύο κατευθύνσεις (two-way joining)

Η ανάλυση της ταξιδόμησης οφείλει τη δημοφιλία της στη θεαματική απεικόνιση των στοιχείων υπό μορφή ομάδων με κοινά χαρακτηριστικά, το γνωστό δενδρόγραμμα. Το διάγραμμα αυτό δείχνει τη δομή των στοιχείων ταξιδόμησης μεμονωμένα είτε των στηλών είτε των σειρών και προφανώς η σύνδεση αυτή δεν βοηθάει να αναπτυχθεί μία ταυτόχρονη σχέση μεταξύ σειρών και στηλών δηλαδή μία σχέση αντιστοίχισης.

Πρόσφατα αναπτύχθηκαν μέθοδοι διασύνδεσης οι οποίες βοηθούν την ταξιδόμηση να εκτελεστεί ταυτόχρονα και στις δύο περιπτώσεις και αυτό μπορεί να αποδειχθεί εξαιρετικά χρήσιμο όταν ο ερευνητής ενδιαφέρεται να συνδυάσει τη δράση ομάδων στοιχείων με ομάδες μεταβλητών π.χ. όταν ο γιατρός θέλει να ομαδοποιήσει ασθενείς (παρατηρήσεις) και παράλληλα να διακρίνει τυχόν συναθροίσεις αυτών που εμφανίζουν διαφορετικές ομάδες συμπτωμάτων (μεταβλητές). Δύο τεχνικές είναι ικανές να πραγματοποιήσουν την διπλή αυτή σύνδεση:

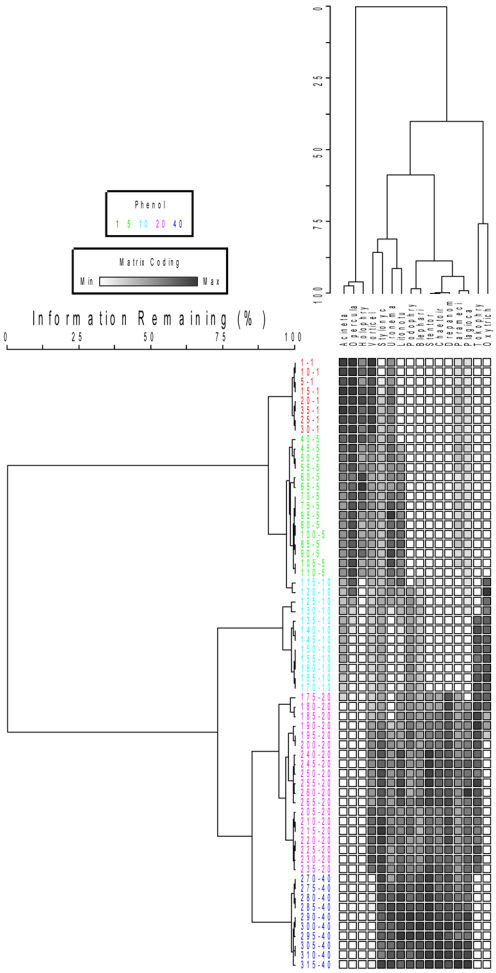
α) Ταξιδόμηση δύο κατευθύνσεων (Two-way clustering)

Σκοπός της ταξιδόμησης αυτής είναι να εκθέσει πιθανές σχέσεις μεταξύ ομάδων ατομικών στοιχείων, οι οποίες απορρέουν από διαφορές ή ομοιότητες α) μεταξύ των σειρών στην ίδια ομάδα ή και μεταξύ σειρών σε διαφορετικές ομάδες και β) μεταξύ στηλών στην ίδια ομάδα ή και σε διαφορετικές ομάδες (Peck, 2010). Στο επόμενο στάδιο, οι συνδυασμοί αυτοί σχετίζονται γραφικά μεταξύ τους.

Η ταξιδόμηση δύο κατευθύνσεων εκτελεί ανάλυση ομαδοποίησης στις σειρές και τις στήλες απεικονίζοντας δύο δενδρογράμματα ταυτόχρονα. Οι στήλες και σειρές του αρχικού πίνακα αναδιοργανώνονται έτσι ώστε να ταιριάζουν στην κατάταξη των παρατηρήσεων στο διπλής κατεύθυνσης δενδρόγραμμα. Σημαντικό στοιχείο της επιτυχημένης εκτέλεσης της μεθόδου είναι η σχετικοποίηση (ενότητα 1.8.2, κεφάλαιο 1) των ομάδων ταξιδόμησης των στηλών αμέσως μετά την ομαδοποίηση των σειρών. Η διαδικασία αυτή είναι εξαιρετικά σημαντική στην οικολογία των κοινωνιών. Χωρίς τη σχετικοποίηση των στηλών, τα κοινά είδη ενδέχεται να διαχωρίζονται σε άλλες ομάδες από τα σπάνια είδη ακόμα και αν κορυφώνουν στο ίδιο σημείο απόκρισης κατά μήκος της περιβαλλοντικής διαβάθμισης. Η σχετικοποίηση των ειδών, που καταλαμβάνουν τη θέση των στηλών, δεν εξαφανίζει το διαχωρισμό μεταξύ των κοινών και σπάνιων ειδών αλλά μετριάζει το αποτέλεσμά τους.

Το δενδρόγραμμα διπλής κατεύθυνσης συντίθεται από το δενδρόγραμμα των σειρών, το δενδρόγραμμα

των στηλών και την παρουσίαση των στοιχείων του αρχικού πίνακα διαμορφωμένο κατάλληλα για να ανταποκριθεί στην διαρθρωτική κατάταξη του δενδρογράμμοτος (Σχ. 5.2).



Σχήμα 5.2. *Διάγραμμα ταξιδόμησης δύο κατευθύνσεων στο οποίο συνδυάζονται τα διαφορετικά επίπεδα προσθήκης τοξικής φαινόλης στο οικοσύστημα (δείγματα-σειρές) με την παρουσία των λιγότερο ή περισσότερο ανθεκτικών ειδών (στήλες). Οι σκιάσεις δείχνουν μεγαλύτερη αφθονία των ειδών.*

β) Ανάλυση του δείκτη διπλής κατεύθυνσης των ειδών (Two-Way-INdicator SPecies Analysis-TWINSPAN)

Είναι μία τεχνική διπλής ταξιδόμησης πολύ δημοφιλής στην οικολογία των κοινωνιών η οποία ταυτόχρονα ταξινομεί είδη και δείγματα (Hill, 1979b) και βασίζεται στο διαχωρισμό των διαστάσεων της αμοιβαίας μεσοστάθμισης (βλ. κεφάλαιο 9).

Το βασικό χαρακτηριστικό της μεθόδου είναι η δημιουργία ενός πίνακα διπλής κατάταξης (Πίν. 5.14). Τα είδη διατάσσονται κατά μήκος της αριστερής πλευράς του πίνακα και τα δείγματα στην οριζόντια πάνω θέση. Στα δεξιά και κάτω του πίνακα, διατάσσονται οι δυαδικοί αριθμοί 0 και 1 οι οποίοι καθορίζουν ριζικά την ταυτόχρονη ταξιδόμηση των ειδών και δειγμάτων. Το εσωτερικό περιεχόμενο του πίνακα καταγράφει τα επίπεδα αφθονίας ανά είδος και δείγμα. Οι αφθονίες έχουν μετρηθεί ως ποσοτικές τιμές πληθυσμιακής εμφάνισης αλλά εδώ υπολογίζονται ως κλάσεις (επίπεδα) των **εικονικών ειδών** (pseudospecies). Τα είδη αυτά στην πραγματικότητα αποτελούν μία μεταβλητή με 5 συνήθως κλάσεις αφθονίας εκφρασμένες σε ποσοστά: 0, 2, 5, 10 και 20 ή αλλιώς εικονικά είδη

1, 2, 3, 4, 5 κατ΄αντιστοιχία. Οι κλάσεις αυτές οι οποίες εμφανίζουν το αποτέλεσμά τους στον πίνακα, λειτουργούν ως εξής: η απουσία ενός είδους τιμάται με την κλάση 1 (0%), σε ένα είδος με αφθονία 9% παράγονται τρεις κλάσεις τιμών 1, 2 και 3 (δηλαδή 0%, 2% και 5%), ενώ αφθονία 1,9% παράγει την κλάση 1 μόνο. Εύκολα συνάγεται ότι η τεχνική απαιτεί την εισαγωγή των στοιχείων στον αρχικό πίνακα υπό μορφή ποσοστιαίας αφθονίας και όχι αναλογιών.

```
                5555566666344444444455555333333332232222111111111112222
                5678901234901234567890123467812354890456701234567890123123456789

 6  Stentor     55555555555555555555555555555----------------------------------  000
 7  Chaetoir    55555555555555555555555555555----------------------------------  000
11  Drepanom    55555555555555555555555555555----------------------------------  000
 4  Podophry    55555555555555555555544455555555445-455----------------------  001
 5  Tokophry    ----------555555555555555555555555555----------------------------  001
16  Blephari    55555555555555555555555555554444444444444----------------------  001
10  Parameci    44555555554444444444455444444------------4432333443333-333343333  01
14  Litonotu    55555555555555555555555555553333243355544555555555555555---------  01
15  Plagioca    555555555555555555555555-55--------333344433333333443433333333  01
 2  Vorticel    55555555555555555555555555555555555555555555555555555555555555555  10
 8  Stylonyc    55555555555555555555555555555555555555555555555555555555555555555  10
 9  Uronema     45555555553444444444444444-----------444254444455555555444454444  10
12  Oxytrich    ----------33--------------443444444444444----------------------  10
 1  Acineta     --------------------------5--55555555555555555555555555555555555  11
 3  Opercula    ----------------------------------5545555555555555555555555555  11
13  Holophry    ----------------------------------------333444323333333333444334  11

                00000000000000000000000000000000011111111111111111111111111111
                00000000000000000000000000000111110000000111111111111111111111
                00000000000000000000000000111000010001111000000000000000111111111
                000000000011111111111111111
```

Πίνακας 5.14. *Πίνακας κατάταξης διπλής κατεύθυνσης σειρών (ειδών) και στηλών (δειγμάτων). Ο πρώτος διαχωρισμός (κατάτμηση του πίνακα) εκκινεί από το σημείο απόκλισης των δυαδικών ψηφίων 0 και 1 στα δεξιά του πίνακα και ταυτόχρονα στο αντίστοιχο σημείο απόκλισης στο κάτω μέρος του πίνακα.*

Σημαντικά μειονεκτήματα της μεθόδου είναι:

α) επειδή η τεχνική βασίζεται υπολογιστικά στην αμοιβαία μεσοστάθμιση (RA), παράγει αξιόλογη πληροφόρηση μόνο όταν εδραιώνεται μία σημαντική περιβαλλοντική διαβάθμιση (χρήση του άξονα 1 των συνιστωσών της RA)

β) σε ένα απλό πίνακα διπλής κατεύθυνσης είναι αδύνατη η παρουσία και περιγραφή στοιχείων μεγάλου όγκου και περίπλοκης υπολογιστικά διαχείρισης.

Η τεχνική έχει πρόσφατα βελτιώσει κάποιες αδυναμίες στην απόδοση των αποτελεσμάτων της με την οργανώνοντας πιο αυστηρά τα κριτήρια απόφασης διαχωρισμού των ειδών και του αριθμού εισόδου των δοκιμαστικών επαναληπτικών τιμών.

5.2.4 Διαγνωστικά κριτήρια της ορθής επιλογής του αριθμού των συστάδων

Αμέσως μετά την επιλογή του τελικού αριθμού των ομάδων προς επικείμενη ερμηνεία των αποτελεσμάτων, επιβάλλεται να ελέγξουμε την εγκυρότητα ή μη της επιλογής μας. Τέσσερα κριτήρια, τα οποία ανταποκρίνονται σε κάθε βήμα επιλογής των τελικών ομάδων, προτάθηκαν για τον σκοπό αυτό:

1. Η ρίζα του μέσου τετραγώνου της τυπικής απόκλισης RMSSTD (root mean squared standard deviation),

$$RMSSTD = \sqrt{\frac{\sum_{j=1}^{p} S_i^2}{p}}$$

s_i^2 είναι η κοινή τυπική απόκλιση όλων των μεταβλητών που σχηματίζουν την ομάδα, p είναι ο αριθμός των μεταβλητών. Όσο μικρότερη είναι η τιμή του κριτηρίου τόσο πιο ομοιογενείς είναι οι παρατηρήσεις στις αντίστοιχες μεταβλητές της ομάδας. Το κριτήριο είναι έγκυρο μόνο όταν εφαρμόζεται η τυποποίηση των μεταβλητών πριν την ταξιδόμηση.

2. Το κριτήριο R^2 το οποίο προκύπτει ως

$$R^2 = \frac{SS_b}{SS_b + SS_w}$$

όπου SS_b είναι η διακύμανση (άθροισμα των τετραγώνων) μεταξύ των ομάδων και SS_w η διακύμανση μέσα στις ομάδες. Το κριτήριο μετρά το μέγεθος της διαφοράς που αναπτύσσεται μεταξύ των ομάδων και παίρνει τιμές από 0 (δεν υπάρχουν διαφορές μεταξύ των ομάδων) μέχρι 1 (αναπτύσσονται μέγιστες διαφορές μεταξύ των ομάδων).

3. Το ημιμερικό R^2 (semipartial R squared) το οποίο προκύπτει ως η διαφορά μεταξύ της διακύμανσης SS_w της νέας ομάδας και του αθροίσματος των διακυμάνσεων όλων των ομάδων που συνδέονται για τη δημιουργία της νέας ομάδας. Το κριτήριο αυτό εκφράζει το βαθμό της απώλειας της ομοιογένειας (loss of homogeneity) και όταν η απώλεια είναι μηδέν τότε η νέα ομάδα προέρχεται από δύο απόλυτα ομοιογενείς ομάδες. Αν η απώλεια είναι μεγάλη τότε η νέα ομάδα δημιουργείται από τη συγχώνευση ετερογενών ομάδων.
4. Η Ευκλείδεια απόσταση μεταξύ δύο ομάδων υποψήφιων προς συγχώνευση, η οποία υπολογίζεται διαφοροποιημένα, ανάλογα με τον τρόπο διασύνδεσης που επιλέχθηκε. Μικρές τιμές του κριτηρίου δηλώνουν υψηλή ομοιογένεια μεταξύ των δύο ομάδων και μεγάλες τιμές δηλώνουν υψηλό βαθμό ανομοιομορφίας.

Τα τέσσερα παραπάνω κριτήρια, από τα οποία τα τρία μετρούν την ομοιογένεια των ομάδων και το τέταρτο (R^2) την ετερογένεια, μπορούν να παρασταθούν γραφικά, σε συνάρτηση με τον αριθμό των ομάδων σε κάθε διαδοχικό βήμα επιλογής (Σχ. 5.3). Στο γράφημα αυτό αναζητούνται απότομες μεταβολές ανοδικές ή καθοδικές, ικανές να σχηματίσουν το άνοιγμα του αγκώνα. Εκεί, παρατηρείται σημαντική μεταβολή κατά τη μετάβαση από το βήμα συνένωσης 14 μέχρι το βήμα 12 και άρα 3 ομάδες είναι αρκετές για την ερμηνεία των αποτελεσμάτων των στοιχείων.

5.2.5 Αξιολόγηση της αποτελεσματικότητας της επιλεγμένης ταξιδόμησης

Η μέτρηση της ομοιότητας μεταξύ δϋο ταξιδομήσεων χρησιμοποιείται ως μέτρο σύγκρισης διαφορετικών αλγόριθμων ταξιδόμησης των ίδιων στοιχείων. Η αξιολόγηση αυτή αναφέρεται και ως εγκυρότητα της ταξιδόμησης και διακρίνεται σε εσωτερική και εξωτερική.

α) Εσωτερική αξιολόγηση

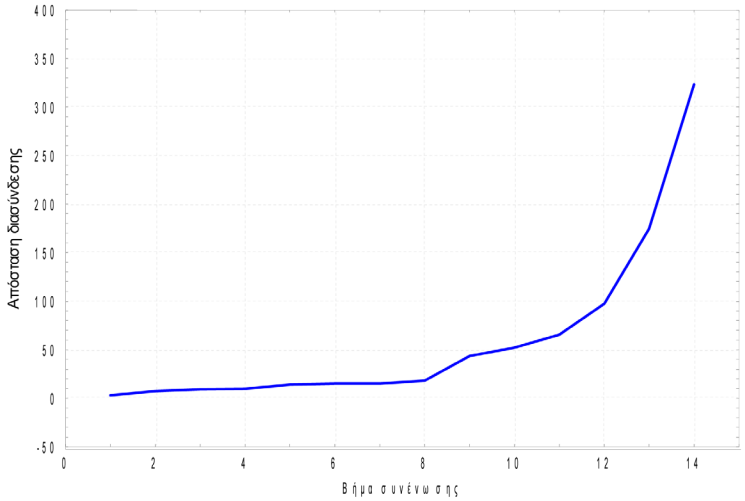
Εφαρμόζεται αποκλειστικά στα στοιχεία της ταξιδόμησης που μόλις σχηματίστηκε, εξ ου και το όνομα εσωτερική, και αποδίδει τον καλύτερο βαθμό στον αλγόριθμο εκείνο που παράγει συστάδες με υψηλή ομοιότητα στοιχείων μέσα σε αυτές και χαμηλή ομοιότητα μεταξύ των συστάδων. Μειονέκτημα της αξιολόγησης αυτής είναι ότι ενίοτε αποδίδονται υψηλοί βαθμοί επίδοσης χωρίς να προάγεται η άντληση σοβαρής πληροφόρησης από την ταξιδόμηση ή να αξιολογεί εσφαλμένα αλγόριθμους που μετρούν το ίδιο μοντέλο ταξιδόμησης.

Στην πράξη, η εσωτερική αξιολόγηση ενδείκνυται στις περιπτώσεις που ένας αλγόριθμος αποδίδει καλύτερα από κάποιον άλλο, χωρίς όμως κατ' ανάγκη να παράγει και πιο αξιόπιστα συγκριτικά αποτελέσματα.

Δύο τεχνικές αξιολόγησης προτάσσονται για την εκτίμηση της ποιότητας των αλγόριθμων ταξιδόμησης των στοιχείων:

- Ο δείκτης **Davies-Bouldin (DB)**,

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{i \neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$



Σχήμα 5.3. *Γράφημα της απόστασης διασύνδεσης με τα βήματα συνένωσης σε ομάδες. Το γράφημα είναι αποτέλεσμα της δημιουργίας του δενδρογράμματος του σχήματος 5.1.*

όπου n είναι ο αριθμός των συστάδων c_x το κέντρο της συστάδας x, σ_x η μέση απόσταση των στοιχείων στη συστάδα x και $d(c_i, c_j)$ η απόσταση μεταξύ των κέντρων c_i και c_j. Χαμηλές τιμές του δείκτη DB στις συστάδες δηλώνουν ότι εξεταζόμενος αλγόριθμος παράγει ένα σύνολο (άθροισμα) συστάδων με μικρές αποστάσεις των στοιχείων μέσα στις συστάδες (υψηλή ομοιότητα) και μεγάλες αποστάσεις (χαμηλή ομοιότητα) μεταξύ των συστάδων.

- Ο δείκτης του **Dunn**, ο οποίος αναγνωρίζει πυκνές και επαρκώς διαχωριζόμενες συστάδες μετρώντας το λόγο μεταξύ της ελάχιστης εσωτερικής απόστασης των στοιχείων εντός της συστάδας προς τη μέγιστη μεταξύ των συστάδων απόσταση. Έτσι, για κάθε νέα συστάδα ο δείκτης υπολογίζεται ως

$$D = \min_{1 \le i \le n}\left[\min_{1 \le j \le n,\, i \neq j}\left(\frac{d(i,j)}{\max_{1 \le k \le n} d'(k)}\right)\right]$$

όπου ο όρος d(i,j) μετρά την απόσταση μεταξύ των συστάδων i και j και ο όρος d'(k) την εσωτερική απόσταση μιας συστάδας k. Η δεύτερη απόσταση εκτιμάται με διάφορους τύπους μέτρησης απόστασης (Ευκλείδεια, Pearson) και η πρώτη με μία από τις γνωστές μεθόδους συνένωσης των συστάδων (πλησιέστερη ή απομακρυσμένη γειτνίαση κτλ.). Το κριτήριο αξιολόγησης D ερευνά συστάδες με υψηλή ομοιότητα των στοιχείων τους και χαμηλή ομοιότητα μεταξύ των συστάδων, επομένως προσδοκώνται υψηλές τιμές του δείκτη για να θεωρείται μία αξιολόγηση ικανοποιητική.

β) Εξωτερική αξιολόγηση

Η μέθοδος αυτή αξιολογεί ταξιδομήσεις με την συγκρότηση στοιχείων που δεν συμμετέχουν στην ταξιδόμηση αλλά σταχυολογούνται από επιστήμονες ολκής για να εξυπηρετήσουν υπό μορφή ειδικών κλάσεων ορισμένα μέτρα αναφοράς της εγκυρότητας κάθε ταξιδόμησης (benchmarks). Ουσιαστικά, οι εξωτερικές αξιολογήσεις μετρούν πόσο ικανοποιητικά μία υποβληθείσα διαδικασία ταξιδόμησης προσεγγίζει τις προκαθορισμένες κλάσεις

αξιολόγησης και έτσι τα στοιχεία υποβάλλονται στη διαδικασία των επαναληπτικών δοκιμαστικών τιμών. Το ζητούμενο του χρυσού κανόνα της αξιολόγησης (gold standard) εστιάζεται στον αριθμό που εκφράζει πόσες επαναληπτικές φορές μία κλάση αποδόθηκε σωστά σε ένα στοιχείο της ταξιδόμησης (αληθώς θετικό) ή αλλιώς στο ζεύγος εκείνο των στοιχείων (της διπλής αξιολόγησης) που ανήκουν πραγματικά στην ίδια συστάδα.

Τρεις μέθοδοι εξωτερικής αξιολόγησης προτείνονται:

- Ο δείκτης **RI** του **Rand** ο οποίος υπολογίζει πόσο όμοια οι ταξιδομημένες συστάδες αντικαθιστούν τις πρότυπες ταξιδομήσεις ή καλύτερα το ποσοστό των ορθών επιλογών που εξάγονται από την εφαρμογή του αλγόριθμου:

$$RI = \frac{TP+TN}{TP+TN+FN+TN}$$

όπου TP είναι ο αριθμός των αληθώς θετικών στοιχείων, TN ο αριθμός των ψευδώς θετικά ανταποκριθέντων στοιχείων, FN ο αριθμός των ψευδώς αρνητικών και FP των ψευδώς θετικών στοιχείων. Υψηλές τιμές εκφράζουν υψηλό ποσοστό ομοιότητας και κατ' επέκταση εγκυρότητας της μεθόδου ταξιδόμησης.

Ο δείκτης αυτός διορθώνεται σταθμικά ως προς τις συχνότητες FP και FN και εκφράζεται ακριβέστερα από την μέτρηση $\mathbf{F}_{\beta}$,

$$F_{\beta} = \frac{(\beta^2+1)\cdot P\cdot R}{\beta^2\cdot P+R}$$

όπου P=TP/(TP+FP) και R=TP/TP+FN) και β είναι μία σταθμισμένη παράμετρος (β≥0).

- Ο δείκτης **J** του **Jaccard**,

$$J = \frac{TP}{TP+FP+FN}$$

ο οποίος συγκρίνει δύο ομάδες στοιχείων και παίρνει τιμές 0-1. Τιμές κοντά στο 1 δηλώνουν ότι οι δύο συστάδες ταυτίζονται και κοντά στο 0 ότι δεν έχουν κοινά σημεία.

- Ο δείκτης **FM** των **Folkes-Mallows**,

$$FM = \frac{TP}{\sqrt{(TP+FP)\cdot(TP+FN)}}$$

ο οποίος συγκρίνει την ομοιότητα των συστάδων που παράγονται από τον αλγόριθμο της ταξιδόμησης με τον αλγόριθμο (κλάσεις αξιολόγησης) της πρότυπης ταξιδόμησης. Υψηλές τιμές ειδοποιούν για υψηλότερη ομοιότητα στην εκτίμηση των δύο μεθόδων ταξιδόμησης.

5.3 Μη ιεραρχική ανάλυση των συστάδων

Όταν με τις προηγούμενες ιεραρχικές μεθόδους δεν παράγεται σημαντικό πλεονέκτημα στην ερμηνεία των αποτελεσμάτων, τότε καταφεύγουμε στις μεθόδους μη ιεραρχικής ταξιδόμησης. Όμως, οι μέθοδοι αυτές βασίζονται στην προβλέψιμη επιλογή συγκεκριμένου αριθμού ομάδων αιτιολογημένη από πρότερη επιστημονική γνώση και εμπειρία στο αντικείμενο της έρευνας. Τέτοιες μέθοδοι είναι η ταξιδόμηση των k μέσων και των μικτών κατανομών E-M. Όταν μία τέτοια επιλογή δεν είναι εφικτή ως προς την ερμηνεία των αποτελεσμάτων, τότε καταφεύγουμε σε τυχαίο αριθμό προεπιλεγμένων ομάδων. Η υπολογιστική διαδικασία περιλαμβάνει δύο επαναλαμβανόμενες ενέργειες, τη μεταφορά των μονάδων από μία ομάδα σε άλλη και την αμοιβαία ανταλλαγή μονάδων από μία ομάδα σε άλλη. Η άριστη απόσπαση των μονάδων συντελείται συνήθως με τη μέθοδο της ελάχιστης διακύμανσης του Ward μέσα στις ομάδες. Η εγκυρότητα της επιλογής ελέγχεται με την εκτίμηση των τεσσάρων στατιστικών κριτηρίων (κριτήρια RMSSTD και R^2, ημιμερικό R^2, Ευκλείδεια απόσταση).

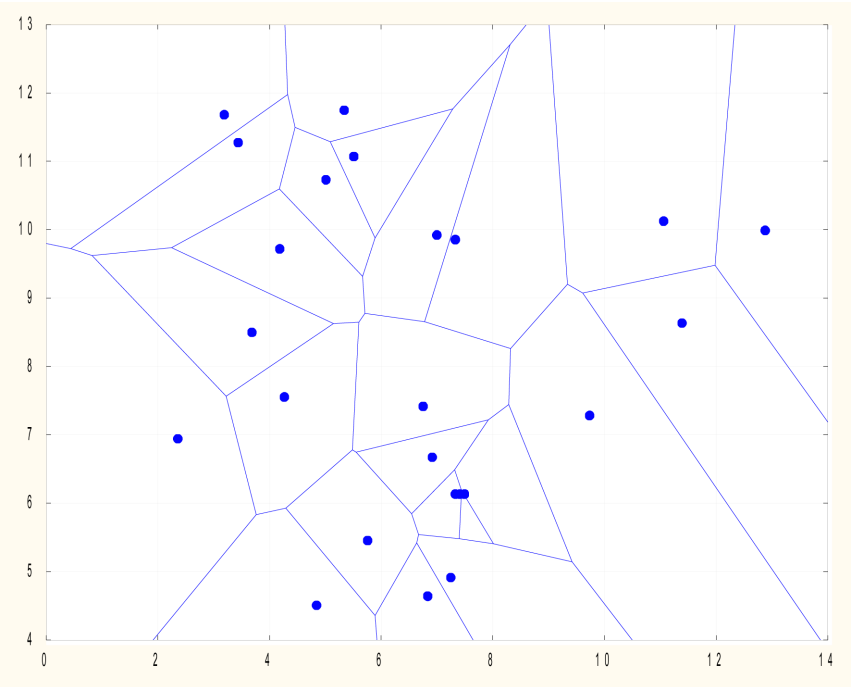
Στη μη ιεραρχική κατηγορία εντάσσεται και η βηματική συσταδοποίηση η οποία χρησιμοποιείται αποκλειστικά στις περιπτώσεις που οι μελετώμενες μεταβλητές είναι ταυτόχρονα κατηγορικές και ποσοτικές.

α) Ταξιδόμηση k μέσων

Στις περιπτώσεις που επιζητείται η συσσώρευση των στοιχείων σε συγκεκριμένο αριθμό ομάδων k, τότε εφαρμόζεται η ταξιδόμηση των k μέσων ως τεχνική βέλτιστης αναζήτησης αυτών: ανεύρεση των κέντρων k ομάδων και κατανομή των στοιχείων στο πλησιέστερο κέντρο ομάδας με τρόπο ώστε το τετράγωνο των αποστάσεων των στοιχείων από την ομάδα να ελαχιστοποιείται. Αρχικά, ο αλγόριθμος προσδιορίζει k κέντρα συστάδων αντιπροσωπευτικά N σημείων (k<N), και ακολούθως κάθε σημείο, με τη χρήση των επαναληπτικών δοκιμών, διευθετείται σε μία από τις k συστάδες και κάθε κέντρο αποτελεί το μέσο όρο των ενταγμένων σημείων (Bishop, 1995). Η εκτέλεση πραγματοποιείται με τον αλγόριθμο του Lloyd (Hartigan & Wong, 1979) με μείζον μειονέκτημα της μεθόδου τον ορισμό εξαρχής συγκεκριμένων ομάδων στα στοιχεία, καθότι ο αλγόριθμος προτιμά να δημιουργεί ισομεγέθεις συστάδες (με ίδιο περίπου αριθμό περίπου στοιχείων) οδηγώντας έτσι στη λανθασμένη οριοθέτηση μεταξύ των ομάδων αφού βελτιστοποιεί τα κέντρα των ομάδων και όχι τα όρια αυτών. Διαθέτει όμως σοβαρά πλεονεκτήματα:

α) Όταν οι μεταβλητές είναι πολυπληθείς, η συσταδοποίηση k μέσων αποδεικνύεται υπολογιστικά πολύ ταχύτερη των ιεραρχικών με την προϋπόθεση ότι απαιτείται μικρός αριθμός ομάδων.

β) Παράγει συστάδες πιο ομοιογενείς (συμπαγείς) συγκριτικά με αυτές των ιεραρχικών μεθόδων και ειδικότερα όταν έχουν σφαιρική μορφή, διότι αποσπά τα στοιχεία στο χώρο συγκροτώντας ειδικές δομές οι οποίες συνθέτουν το διάγραμμα ή μωσαϊκό του Voronoi (βλ. ένθετο σχήμα). Αυτό συνίσταται από πεδία σε ένα χώρο στοιχείων μετρούμενων με κάποιο μέτρο απόστασης π.χ. Ευκλείδεια, Manhattan κτλ., όπου κάθε πεδίο συντίθεται από ένα στοιχείο p_k και το κύτταρο ή έκταση του Voronoi R_k καθώς επίσης και κάθε σημείο του οποίου η απόσταση από το p_k είναι μικρότερη ή ίση της απόστασής του σε οποιαδήποτε άλλη περιοχή. Στο σχήμα παρακάτω η ταξιδόμηση k μέσων διαχωρίζει τα στοιχεία σε 24 κύτταρα του Voronoi.



Σύμφωνα με την συσταδοποίηση των k μέσων όλα τα στοιχεία (παρατηρήσεις) τοποθετούνται σε μία αρχική συστάδα και ο αλγόριθμος θα μετακινήσει τα στοιχεία σε διάφορες συστάδες με στόχο την ελαχιστοποίηση των αποστάσεων μέσα σε κάθε συστάδα και τη μεγιστοποίηση αυτών μεταξύ των συστάδων. Υπάρχουν τρεις τρόποι υλοποίησης της παραπάνω διαδικασίας:

1. Επιλογή των N πρώτων παρατηρήσεων ως κέντρα k αρχικών συστάδων. Επομένως η θέση και τιμή των πρώτων παρατηρήσεων είναι ζωτικής σημασίας και θα πρέπει αυτές να τοποθετούνται με περίσκεψη από τον ερευνητή.
2. Επιλογή του αλγόριθμου κάποιων αρχικών παρατηρήσεων ως πρώτα κέντρα συστάδων και με επαναληπτική διαδικασία δοκιμών, τελική διευθέτηση όλων των στοιχείων στις συστάδες εξασφαλίζοντας ελάχιστη απόσταση αυτών από τα κέντρα τους και μέγιστη μεταξύ των συστάδων (συνηθέστερη επιλογή).
3. Οι τιμές των αποστάσεων μεταξύ όλων των στοιχείων διατάσσονται αυξητικά και ακολούθως επιλέγονται μερικές αποστάσεις ως αρχικά κέντρα συστάδων κατά σταθερά διαστήματα της κατάταξης.

β) Ταξιδόμηση της κατανομής των στοιχείων

Η μέθοδος αυτή βασίζεται στην εκτίμηση στατιστικών παραμέτρων που περιγράφουν διάφορες κατανομές, έτσι όμοιες συστάδες μπορούν να σχηματιστούν από στοιχεία που ανήκουν σε συγκεκριμένη κατανομή. Η μέθοδος αυτή δημιουργεί ενίοτε προβλήματα υπερπροσαρμογής των στοιχείων στην κατανομή γι' αυτό και απαιτείται η χρήση περιοριστών όρων.

Μία πολλά υποσχόμενη επιλογή της ταξιδόμησης αυτής είναι τα μοντέλα μικτών κατανομών του Gauss (κανονική κατανομή) με την εφαρμογή του αλγόριθμου E-M (Dempster et al, 1977). Τα στοιχεία αναδιπλώνονται σε ένα συγκεκριμένο αριθμό κατανομών του Gauss (προς αποφυγή υπερπροσαρμογής) ο οποίος εκκινεί τυχαία και οι στατιστικές παράμετροι βελτιώνονται επαναληπτικά για να ταιριάξουν άριστα στα στοιχεία, αποδίδοντας, πάραυτα, ενίοτε διαφορετικά τελικά αποτελέσματα.

Τα μοντέλα ταξιδόμησης κατανομών παράγουν περίπλοκα πρότυπα συστάδων τα οποία μπορούν να αξιοποιήσουν τη συσχέτιση και εξάρτηση μεταξύ των μεταβλητών των στοιχείων, έχουν όμως το μειονέκτημα να ακολουθούν συγκεκριμένη κατανομή χωρίς παρεκκλίσεις.

Σε περίπτωση υιοθέτησης της ισχυρής ταξιδόμησης, τα στοιχεία τοποθετούνται υποχρεωτικά στην κατανομή Gauss που πιθανότερα ανήκουν, πράγμα που δεν συμβαίνει στη χαλαρή διασύνδεση.

Ο αλγόριθμος E-M (Expectation-Maximization), σε στοιχεία με κανονικές κατανομές, εκτιμά μέσους όρους και τυπικές αποκλίσεις για κάθε συστάδα μεγιστοποιώντας την πιθανότητα (αναλογία) κατανομής των παρατηρούμενων στοιχείων. Δηλαδή, προσεγγίζει τις παρατηρούμενες τιμές κατανομής των στοιχείων με βάση τη μίξη διαφορετικών κατανομών στις διάφορες συστάδες. Για παράδειγμα, αν μία μεταβλητή διαχωρίζεται σε δύο συστάδες, και άρα δύο κανονικές κατανομές υπάρχουν ως προς τα στοιχεία, ο αλγόριθμος E-M υπολογίζει τη νέα κατανομή που προκύπτει από το άθροισμα των δύο αρχικών κατανομών και με βάση αυτή υπολογίζει τις στατιστικές παραμέτρους που αναλογούν σε κάθε αρχική κατανομή. Ο αλγόριθμος E-M λειτουργεί επίσης αποτελεσματικά σε λογαριθμο-κανονικές κατανομές και Poisson.

Σε αντίθεση με την ταξιδόμηση των k μέσων, στην οποία τα στοιχεία κατανέμονται μόνο σε συστάδες με μέγιστες αποστάσεις μεταξύ τους, στην ταξιδόμηση E-M, κάθε παρατήρηση ανήκει σε μία συστάδα με συγκεκριμένη τιμή πιθανότητας, διότι η τεχνική βασίζεται αποκλειστικά στην ταξινόμηση των πιθανοτήτων.

γ) Διβηματική ταξιδόμηση των στοιχείων

Η διβηματική ανάλυση συστάδων αποτελεί περισσότερο ένα επεξεργαστικό εργαλείο παρά μία απλή τεχνική ταξιδόμησης, η οποία προσδιορίζει συστάδες παρατηρήσεων με την επινόηση προσυστάδων και εφαρμογή σε αυτές ιεραρχικών τεχνικών (Σιάρδος 2015). Με τη χρήση ενός ευέλικτου αλγορίθμου καταφέρνει να διαχειριστεί μεγάλο όγκο δεδομένων σε χρόνο πολύ μικρότερο από τον απαιτούμενο με την εφαρμογή ιεραρχικών μεθόδων συσταδοποίησης.

Η διβηματική ανάλυση διαχειρίζεται εξίσου ικανοποιητικά ποσοτικές (σε τυποποιημένη μορφή) και ονομαστικές μεταβλητές με την προϋπόθεση ότι οι ποσοτικές κατανέμονται κανονικά και οι κατηγορικές πολυωνυμικά. Ως μέτρο ομοιότητας χρησιμοποιεί τον λογάριθμο της μέγιστης πιθανοφάνειας των αποστάσεων και την Ευκλείδεια απόσταση όταν αποκλειστικά όλες οι μελετώμενες μεταβλητές είναι ποσοτικές. Εκτελείται σε δύο χρόνους:

Αρχικά δημιουργεί προσυστάδες οι οποίες περιορίζουν το εύρος της μήτρας των αποστάσεων μεταξύ των παρατηρήσεων και χρησιμοποιούνται στη θέση των αρχικών παρατηρήσεων εφαρμόζοντας τους αλγόριθμους της ιεραρχικής ανάλυσης προς σχηματισμό δενδρογράμματος συστάδων.

Στη συνέχεια, οι προσυστάδες με βάση τον αλγόριθμο της συσσωρευτικής ιεραρχικής διασύνδεσης διερευνώνται ως προς την καταλληλότητα του αριθμού της τελικής σύστασης της συσταδοποίησης χρησιμοποιώντας ως κριτήρια επιλογής του άριστου αριθμού αυτών τα κριτήρια πληροφόρησης του Bayes (BIC) και του Akaike (AIC) τα οποία στην απλούστερή τους μορφή προσδιορίζονται ως

$$BIC = -2 \cdot \ln(L) + k \cdot \ln(n)$$

και

$$AIC = 2k - 2 \cdot \ln(L)$$

όπου k είναι ο αριθμός των παραμέτρων που πρόκειται να εκτιμηθούν (π.χ. ο αριθμός των βαθμών ελευθερίας), L η μέγιστη πιθανοφάνεια και n ο αριθμός των παρατηρήσεων.

Μειονέκτημα της μεθόδου, απαντώμενο και στην ταξιδόμηση των k μέσων, συνιστά η εξάρτηση των αποτελεσμάτων από τη σειρά διαδοχής των τιμών των παρατηρήσεων στο μοντέλο, η οποία όταν μεταβάλλεται, η βηματική ανάλυση δίνει διαφορετικές λύσεις στο τελικό μοντέλο. Το πρόβλημα αυτό αντιμετωπίζεται με μεθόδους τυχαιοποίησης της σειράς των παρατηρήσεων και επαναληπτικές εφαρμογές του προγράμματος μέχρις ότου η διαφορετική διάταξη των παρατηρήσεων να εξάγει συγκλίνοντα αποτελέσματα.

5.4 Ανακεφαλαίωση - Επισημάνσεις

Η επιλογή της ταξιδόμησης των στοιχείων, ανεξαρτήτως το ποια μέθοδος θα επιλεγεί τελικά, δεν συνιστάται όταν γνωρίζουμε από προγενέστερη εμπειρία το συγκεκριμένο αριθμό των ομάδων. Αντίθετα, προτείνεται ενθέρμως στις περιπτώσεις 'τυφλής' ανεύρεσης του αναγκαίου αριθμού των ομάδων και δρα εξαιρετικά ικανοποιητικά, όταν συνυπολογίζεται επικουρικά με την ανάλυση των κύριων συνιστωσών στα ίδια στοιχεία.

Συνοψίζοντας τη μέχρι τούδε πληροφόρηση, ορισμένες επισημάνσεις είναι αναγκαίο να αναφερθούν, σχετικά με την ανάλυση της ταξιδόμησης:

1. Είναι μία τεχνική η οποία κατανέμει τα στοιχεία σε ομάδες βασιζόμενη στην ομοιότητα των τιμών μεταξύ τους. Ομοιόμορφες μονάδες θα σχηματίσουν ομάδες διακριτές από άλλες ομάδες.
2. Η τεχνική χρησιμοποιεί διαφορετικούς αλγόριθμους ανάλογα με τη στρατηγική επιλογής της σύνδεσης των ομάδων οι οποίοι όμως μπορούν να συνοψιστούν σε μία γενικευμένη εξίσωση γραμμικών συνδυασμών των Lance και Williams (1967).

Δείγμα	DM	FAT/DM	TN/DM	WSN/TN	NPN/TN	BC	pH	FF	FR	REC	FB	DEF	Hard	Elas	Frac
GRD	64,22	50,60	5,51	29,50	11,58	7,31	5,61	24,60	18,20	73,86	79,20	35,20	9,35	10,11	4,90
TEM	49,08	51,95	5,25	14,50	9,50	5,58	4,52	18,00	11,29	62,70	22,70	19,70	5,81	4,72	11,02
KEE	61,67	49,45	5,01	29,50	9,90	11,40	5,38	37,80	27,20	72,00	88,94	33,80	12,31	7,25	8,59
FEK	45,39	52,87	5,71	19,26	25,62	5,42	4,91	8,34	5,11	61,00	9,96	29,50	3,69	3,70	11,75
FEI	45,23	53,06	5,59	18,22	25,05	5,08	4,05	17,46	9,94	57,00	24,50	21,40	3,65	3,30	10,73
GRN	62,70	54,62	6,10	18,95	22,09	4,85	5,48	26,60	19,50	73,30	74,80	48,60	7,87	10,10	4,45
KED	65,09	49,16	5,73	26,59	23,02	11,14	5,19	86,80	61,50	70,80	91,60	20,30	11,49	4,06	10,13
KAM	56,16	49,41	6,87	33,21	26,39	4,27	5,49	14,00	9,60	68,60	31,90	33,30	4,38	7,97	7,51
VIT	62,28	43,35	7,10	52,69	14,53	6,91	5,63	23,65	16,70	70,50	47,00	41,00	7,47	9,24	4,61
FEF	45,32	49,09	6,02	17,83	24,21	3,73	4,52	12,10	7,60	63,30	13,10	21,20	3,55	4,48	8,48
KEDO	67,11	49,17	5,97	22,74	18,60	9,78	5,23	48,40	34,80	72,00	55,90	27,00	10,88	8,57	7,24
LAM	67,88	51,56	6,68	39,57	17,00	7,46	5,55	43,80	31,70	72,40	64,80	33,40	10,35	9,77	6,07
MET	58,19	46,82	7,11	27,99	23,64	7,54	5,16	10,70	7,55	70,50	17,00	23,80	5,13	7,60	6,13
KAP	63,18	46,69	6,71	29,36	23,77	5,78	5,58	19,20	13,90	72,50	27,30	25,10	9,36	8,34	5,40
GRM	62,53	47,97	6,31	21,74	35,84	6,92	4,99	19,20	12,80	66,50	33,40	38,00	7,69	8,66	6,65
KEDE	58,94	39,87	6,47	28,86	22,96	12,93	5,04	22,00	13,90	63,20	46,40	23,50	10,58	5,22	9,94
FED	49,80	48,69	5,60	30,93	32,50	7,07	4,47	5,50	2,84	51,80	6,50	36,80	4,28	5,43	9,92
KAL	50,42	53,55	6,78	26,97	19,50	4,19	4,37	5,12	2,90	56,70	13,05	37,40	5,91	5,97	9,69
ROM	61,60	46,67	6,00	24,97	15,62	7,40	5,10	54,40	37,10	68,20	66,00	23,50	9,16	5,55	9,80
KAG	57,78	51,48	5,95	35,66	25,14	7,40	5,10	9,40	6,45	71,50	20,90	42,90	4,66	9,16	6,08
GRC	63,75	48,62	5,89	18,59	25,14	2,40	5,38	15,45	10,91	70,60	49,00	55,60	7,85	6,32	6,93
KEM	61,28	48,95	5,54	33,58	27,61	6,78	5,23	24,80	17,70	71,00	58,20	24,00	8,60	5,45	7,36
FEA	45,64	52,58	6,11	15,44	22,25	3,26	4,24	16,70	10,85	64,60	23,00	23,30	4,08	3,92	11,65
TES	46,08	54,79	4,62	23,09	28,49	3,51	4,37	13,10	8,22	62,60	18,60	32,70	4,19	3,02	11,22

Πίνακας 5.15*. Φύλο εργασίας 24 ποικιλιών τυρού και των χαρακτηριστικών τους: χημική σύσταση, μηχανικές και οργανοληπτικές ιδιότητες. Η μεταβλητή FF απορρίφθηκε από περαιτέρω στατιστική επεξεργασία λόγω της εξαιρετικά υψηλής συσχέτισης που εμφάνισε με την FR (r=0,998).*

3. Η ομαδοποίηση των μονάδων παρίσταται γραφικά με τη χρήση δενδρογράμματος. Η επιλογή συγκεκριμένου αριθμού ομάδων υποβοηθείται με τη χρήση τεσσάρων στατιστικών κριτηρίων, στην πράξη όμως είναι τελείως υποκειμενικό θέμα και έχει κυρίως να κάνει με τη βαθιά γνώση στο αντικείμενο και την εμπειρική λογική. Ένας χρυσός κανόνας είναι να μην αποδεχόμαστε απλές διαιρέσεις του δενδρογράμματος οι οποίες οδηγούν στην επιλογή ομάδων με πολύ λίγες μονάδες σε αυτές. Είναι προτιμότερο η τελική ομαδοποίηση να συνδυάζεται με τα γραφικά αποτελέσματα της ανάλυσης των κύριων συνιστωσών.

4. Οι περισσότερες από τις τεχνικές ταξιδόμησης δίνουν σχεδόν ίδια αποτελέσματα (ίδιο περίπου δενδρόγραμμα) μόνο όταν τα αρχικά στοιχεία χαρακτηρίζονται από έντονα εμφανείς διαφοροποιημένες τάσεις. Όταν όμως συσσωρεύεται μεγάλος όγκος στοιχείων και η φύση των δειγματοληπτικών μονάδων εμφανίζει κάποιο βαθμό ασάφειας στο τρόπο επιλογής αυτών, τότε είναι δυνατόν να μην εμφανίζονται στο δενδρόγραμμα αναμενόμενες τάσεις που προβλέπονται από προηγούμενη εμπειρία στο αντικείμενο μελέτης. Στην περίπτωση αυτή συνιστάται η δοκιμή διαφορετικών τεχνικών ταξιδόμησης για συγκριτικούς λόγους ή στροφή σε άλλες μεθόδους πολυμεταβλητής ανάλυσης.
5. Οι μέθοδοι διασύνδεσης της ελάχιστης και μέγιστης απόστασης των στοιχείων (πλησιέστερη και απομακρυσμένη γειτνίαση) συνήθως καταστρέφουν τη δομή των αρχικών στοιχείων γιατί μετά τη συγχώνευση των ομάδων οι υπολογιζόμενες νέες αποστάσεις διαφέρουν πάρα πολύ από τον πίνακα των αρχικών αποστάσεων ομοιότητας. Γενικά, οι δύο αυτές μέθοδοι θα πρέπει να αποφεύγονται όταν τα συλλεγόμενα στοιχεία έχουν βιολογικό ή οικολογικό χαρακτήρα.
6. Οι μέθοδοι της απομακρυσμένης γειτνίασης, μέσης τιμής των ομάδων και των κεντρικών τιμών, παράγουν ομάδες με ισχυρή συγγένεια των τιμών μέσα στις ομάδες. Αντίθετα, η μέθοδος της πλησιέστερης γειτνίασης δημιουργεί επιμήκεις ομάδες στα δενδρογράμματα στις οποίες συγχωνεύονται μονάδες με ελάχιστη συγγένεια μεταξύ τους. Έτσι, η χρήση της μεθόδου αυτής θεωρείται αποδεκτή στις περιπτώσεις που αναζητούνται ύποπτες τιμές στα στοιχεία. Από την άλλη πλευρά, η μέθοδος της ελάχιστης διακύμανσης του Ward τείνει να σχηματίσει ομάδες με μονάδες ιδίου μεγέθους περίπου, η μέθοδος της μέσης τιμής διασύνδεσης των ομάδων δημιουργεί ομάδες ίσων περίπου αποστάσεων, ενώ η μέθοδος της διαμέσου (κεντρική διασύνδεση) αποδίδει μεγαλύτερη βαρύτητα στις μονάδες που προστίθενται τελευταίες σε κάθε ομάδα.

Γενικά, οι ιεραρχικές μέθοδοι υποφέρουν πολύ από την παρουσία ακραίων τιμών των στοιχείων με αποτέλεσμα αυτά να εμφανίζονται ατομικά ως επιπρόσθετες και άχρηστες συστάδες ή ακόμα και να αναγκάζουν άλλες συστάδες να συνενώνονται επισύροντας το φαινόμενο της αλυσιδωτής συνένωσης στην απλή διασύνδεση.

Η τελική επιλογή μιας μεθόδου ταξιδόμησης βασίζεται σε συγκεκριμένο τύπο διασύνδεσης:

- Οι ιεραρχικές μέθοδοι βρίσκουν ομάδες που συντίθενται από υποομάδες και έτσι δημιουργείται το δενδρόγραμμα.
- Η συσσωρευτική μέθοδος οικοδομεί ομάδες με ιεραρχικό τρόπο από κάτω προς τα πάνω σύμφωνα με κάποιο στοιχείο ομοιότητας ή ανομοιότητας των παρατηρήσεων. Στην αρχή όλες οι παρατηρήσεις αποτελούν ατομικές ομάδες οι οποίες συνενώνονται διαδοχικά σε μεγαλύτερες ομάδες.

Η ιεραρχική συσσωρευτική ανάλυση περιλαμβάνει τα εξής στάδια συγκρότησης σε ομάδες:

1. Υπολογισμός του πίνακα απόστασης μεταξύ κάθε ζεύγους παρατηρήσεων. Από τον πίνακα των στοιχείων n παρατηρήσεων και p μεταβλητών υπολογίζουμε τον πίνακα απόστασης nxn.
2. Επιλογή και ένωση δύο ομάδων με βάση κάποιο κριτήριο ελάχιστης απόστασης και συγχώνευση των παρατηρήσεων.
3. Συνένωση των επόμενων δύο ομάδων κοκ.
4. Δημιουργία δενδρογράμματος των ομάδων και υποομάδων με βάση την κλίμακα των αποστάσεων.

Οι πολυθεσικές μέθοδοι χρησιμοποιούν πολυάριθμες μεταβλητές (ή είδη ή άλλα χαρακτηριστικά) ως βάση λήψης απόφασης η οποία καθορίζει πότε θα συμβεί συγχώνευση ή διαίρεση των παρατηρήσεων.

Οι μονοθεσικές μέθοδοι βασίζουν τη συγχώνευση ή διαίρεση των παρατηρήσεων σε ένα μόνο είδος ή μεταβλητή.

Η διαιρετή μέθοδος εκκινεί με όλες τις παρατηρήσεις σε μία ομάδα η οποία διαιρείται σε δύο υποομάδες, αυτή σε τέσσερις κοκ. Πολυθεσική διαιρετή είναι η τεχνική TWINSPAN.

Οι μη ιεραρχικές μέθοδοι αναζητούν μία άριστη σύνθεση για ένα συγκεκριμένο αριθμό ομάδων χωρίς αυτή να σχετίζεται απαραίτητα με άλλα επίπεδα ομάδων. Η άριστη σύνθεση ορίζεται στατιστικά με την ελαχιστοποίηση της εντός της ομάδας διακύμανσης. Μη ιεραρχική μέθοδος είναι η συσταδοποίηση κατανομής των στοιχείων και των k μέσων (k-means). Επειδή οι υπολογιστικές απαιτήσεις της δεύτερης μεθόδου δεσμεύουν μικρό μόνο μέρος της μνήμης RAM του υπολογιστή και άρα είναι ελάχιστες, χρησιμοποιείται ευρέως στις περιπτώσεις όπου παρατηρείται ροή τεράστιου όγκου δεδομένων.

Η πλησιέστερη γειτνίαση είναι η απόσταση μεταξύ δύο ομάδων που ορίζεται ως η ελάχιστη απόσταση δύο στοιχείων, ένα από κάθε ομάδα. Η τεχνική αυτή 'πνίγει' τις ομάδες και αποτυγχάνει όταν χρησιμοποιείται σε μεγάλο όγκο δεδομένων.

Η απομακρυσμένη γειτνίαση είναι η απόσταση μεταξύ δύο ομάδων που ορίζεται ως η μέγιστη μεταξύ δύο στοιχείων, ένα από κάθε ομάδα. Μειονεκτεί στο γεγονός ότι μερικές φορές δημιουργεί διακριτές ομάδες εκεί που δεν υπάρχουν.

Η μέση απόσταση διασύνδεσης μεταξύ των ομάδων υπολογίζεται ως η μέση τιμή όλων των αποστάσεων

όλων των ζευγών παρατηρήσεων, μιας από κάθε ομάδα.

Το κέντρο βάρους διασύνδεσης (centroid) μεταξύ των ομάδων υπολογίζεται ως η απόσταση μεταξύ των κεντρικών βαρών (μέσα ανύσματα συντεταγμένων) των δύο ομάδων.

Το διάμεσο κέντρο βάρους διασύνδεσης (median) αποτελεί εναλλακτική επιλογή της προηγούμενης τεχνικής. Ορίζει την απόσταση μεταξύ μιας ομάδας Α και της συνενωτικής ομάδας Β+Γ ως την απόσταση από το κέντρο βάρους της Α μέχρι το μεσοδιάστημα της ευθείας που συνδέει τα κεντρικά βάρη των ομάδων Β και Γ.

Η διασύνδεση του Ward στηρίζεται στην ελαχιστοποίηση του αθροίσματος των τετραγώνων του σφάλματος, το οποίο ορίζεται ως το άθροισμα των τετραγώνων των αποστάσεων από κάθε παρατήρηση προς το κέντρο βάρους της ομάδας του.

5.5 Μελέτη περίπτωσης ανάλυσης συστάδων

Η στατιστική επεξεργασία διεκπεραιώθηκε με τη χρήση των προγραμμάτων STATISTICA 12.0 και MINITAB 16.0.

Τα παραδοσιακά ελληνικά τυριά μελετήθηκαν ως προς τη χημική τους σύσταση, τις μηχανικές ιδιότητες και τα οργανοληπτικά χαρακτηριστικά (Πίν. 5.15). Σκοπός της μελέτης είναι η διερεύνηση ιδιαζουσών συμπεριφορών των τυριών, η παρασκευή των οποίων εξασφαλίζει ένα μεγάλα φάσμα διαφορών σε αρκετά από τα επιλεγέντα χαρακτηριστικά. Αυτό προδικάζεται, επίσης, και από μία πρότερη γενικότερη γνώση διαχωρισμού αυτών σε μαλακά (φέτες), σκληρά (κεφαλοτύρια) και ημίσκληρα (κασέρια και κεφαλογραβιέρες). Επομένως, το ερώτημα που τίθεται είναι αν τα μελετώμενα χαρακτηριστικά μπορούν να τεκμηριώσουν ένα τέτοιο διαχωρισμό ή αλλιώς μία ομαδοποίηση των τυριών με κοινές ιδιότητες ανά ομάδα. Για το σκοπό αυτόν εφαρμόστηκαν τρεις αναλύσεις ταξιδόμησης, μία ιεραρχική και δύο μη ιεραρχικές.

Ιεραρχική ταξιδόμηση

Ως συντελεστής ομοιότητας των στοιχείων επιλέχθηκε η Ευκλείδεια απόσταση (Πίν. 5.16) και τα στοιχεία ακολούθως υποβλήθηκαν στην ανάλυση ταξιδόμησης με την τεχνική του Ward (Πίν. 5.17). Τα στοιχεία (δείγματα) διευθετήθηκαν σε 3 ομάδες τυριών με βάση το κρημνογράφημα (Σχ. 5.4, σχηματισμός αγκώνα) αλλά και τη μορφολογία του δενδρογράμματος της ταξιδόμησης (Σχ. 5.5):

Ομάδα 1- **Ημίσκληρα**
Ομάδα 2- **Μαλακά**
Ομάδα 3- **Σκληρά**

Τα μαλακά τυριά διαχωρίζονται πλήρως από τα υπόλοιπα αν και θα μπορούσε να απαιτηθεί και δεύτερος διαχωρισμός σε δύο υποομάδες, η δεύτερη με δύο μόνο τυριά, αν εγειρόταν κάποιος σοβαρός λόγος. Οι αποστάσεις μεταξύ των ομάδων διευκρινίζονται στον πίνακα 5.17, όπως και κάποια άλλα χαρακτηριστικά της ταξιδόμησης.

Με βάση την παραπάνω ομαδοποίηση εξάγονται και οι μέσες τιμές των χαρακτηριστικών στους τρεις τύπους τυριών (Πίν. 5.18 και 5.19) οι οποίες διαφοροποιούνται στατιστικά σημαντικά μεταξύ τους συγκρίνοντας τα 95% όρια εμπιστοσύνης αυτών (Σχ. 5.6).

Όσον αφορά τη χημική τους σύσταση και επικεντρώνοντας στο σχήμα 5.6 και πίνακα 5.18 παρατηρούμε τα ακόλουθα:

- Τα μαλακά τυριά περιέχουν λιγότερη ξηρή ουσία, οξύτητα (pH), υδατοδιαλυτό άζωτο και περισσότερο λίπος (όλα επί ξηρής ουσίας) από τα υπόλοιπα.
- Τα ημίσκληρα εμφανίζουν υψηλότερο pH και ολικό άζωτο (επί ξηρού βάρους).
- Τα σκληρά τυριά διακρίνονται από τα υπόλοιπα επειδή περιέχουν μεγαλύτερη συγκέντρωση άλμης.

Σχετικά με τα μηχανικά χαρακτηριστικά, η εικόνα των τυριών έχει ως εξής:Τα μαλακά εμφανίζουν χαμηλές τιμές σκληρότητας FB και ανάνηψης (REC).

- Τα ημίσκληρα υψηλές τιμές παραμόρφωσης (DEF).
- Τα σκληρά μεγαλύτερη σκληρότητα FR και FB.

Τέλος, τα οργανοληπτικά χαρακτηριστικά διαφοροποιούνται σαφέστερα μεταξύ των ομάδων:

- Τα μαλακά εμφανίζουν μικρή σκληρότητα και ελαστικότητα και υψηλή ευθρυπτότητα.
- Τα ημίσκληρα μεγάλη ελαστικότητα, μικρή ευθρυπτότητα και μέτρια σκληρότητα.
- Τα σκληρά τυριά εμφανίζουν μέτρια ευθρυπτότητα και ελαστικότητα και έντονη σκληρότητα.

Συμπερασματικά, η ανάλυση ταξιδόμησης του Ward είναι ικανή να διαχωρίσει τις ποικιλίες των τυριών σε τρεις μεγάλες ομάδες με ισχυρά διαφοροποιημένα χαρακτηριστικά μεταξύ τους. Η επιλογή των συγκεκριμένων χαρακτηριστικών κρίνεται ως σημαντικό υπόβαθρο κατατομής για μελλοντικές διερευνητικές μελέτες.

	GRD	TEM	KEE	FEK	FEI	GRN	KED	KAM	VIT	FEF	KEDO	LAM	MET	KAP	GRM	KEDE	FED	KAL	ROM	KAG	GRC	KEM	FEA	TES
GRD		6,12	3,04	6,92	7,55	3,09	5,64	4,68	4,56	6,62	2,98	2,73	4,89	3,79	4,93	6,05	7,33	6,48	4,07	4,10	4,51	3,83	7,08	7,35
TEM	6,12		5,84	3,19	3,05	6,61	6,69	5,55	7,89	3,16	5,56	6,85	5,40	6,00	6,00	6,09	5,02	4,03	4,44	5,70	6,08	5,08	2,77	3,82
KEE	3,04	5,84		6,96	7,43	5,12	4,08	6,12	5,90	7,09	2,98	4,05	6,05	5,22	5,87	5,10	7,40	6,99	3,40	5,51	5,71	4,10	7,20	7,27
FEK	6,92	3,19	6,96		2,25	6,50	7,23	4,44	7,79	2,35	6,28	7,12	4,92	5,85	4,96	6,14	3,18	3,15	5,32	4,89	5,43	4,84	2,16	2,42
FEI	7,55	3,05	7,43	2,25		7,25	7,34	5,52	8,56	2,34	6,72	7,75	5,64	6,66	5,62	6,52	3,44	3,41	5,60	5,84	6,45	5,36	1,86	2,52
GRN	3,09	6,61	5,12	6,50	7,25		6,56	4,25	5,69	6,35	4,12	3,75	5,08	4,29	4,05	7,20	7,03	5,89	5,20	3,78	3,18	4,55	6,64	6,87
KED	5,64	6,69	4,08	7,23	7,34	6,56		6,69	7,27	7,23	3,58	5,05	6,50	5,88	6,14	5,22	7,75	7,68	3,14	6,93	6,85	4,39	7,16	7,58
KAM	4,68	5,55	6,12	4,44	5,52	4,25	6,69		4,19	4,06	4,67	4,11	2,27	2,73	3,08	5,44	4,74	3,97	4,51	2,59	3,86	3,44	4,83	5,62
VIT	4,56	7,89	5,90	7,79	8,56	5,69	7,27	4,19		7,22	5,24	3,60	4,36	3,87	5,48	5,68	7,23	6,60	5,43	4,44	5,68	5,06	8,07	8,67
FEF	6,62	3,16	7,09	2,35	2,34	6,35	7,23	4,06	7,22		5,96	6,86	3,98	5,08	4,59	5,82	3,66	3,29	4,96	4,81	5,39	4,54	1,94	3,42
KEDO	2,98	5,56	2,98	6,28	6,72	4,12	3,58	4,67	5,24	5,96		2,81	4,20	3,26	3,99	4,43	6,65	6,08	2,48	4,49	4,86	3,15	6,29	6,94
LAM	2,73	6,85	4,05	7,12	7,75	3,75	5,05	4,11	3,60	6,86	2,81		4,39	3,21	4,63	5,72	7,26	6,17	3,91	4,13	5,01	3,79	7,21	7,79
MET	4,89	5,40	6,05	4,92	5,64	5,08	6,50	2,27	4,36	3,98	4,20	4,39		2,23	3,22	4,42	5,13	4,47	4,30	3,24	4,86	3,62	5,03	6,31
KAP	3,79	6,00	5,22	5,85	6,66	4,29	5,88	2,73	3,87	5,08	3,26	3,21	2,23		3,14	4,63	6,06	5,40	3,88	3,55	4,24	2,99	6,01	6,88
GRM	4,93	6,00	5,87	4,96	5,62	4,05	6,14	3,08	5,48	4,59	3,99	4,63	3,22	3,14		4,84	4,40	4,68	4,59	3,11	3,48	3,33	5,35	5,67
KEDE	6,05	6,09	5,10	6,14	6,52	7,20	5,22	5,44	5,68	5,82	4,43	5,72	4,42	4,63	4,84		5,62	6,19	3,84	5,85	6,40	4,35	6,48	7,32
FED	7,33	5,02	7,40	3,18	3,44	7,03	7,75	4,74	7,23	3,66	6,65	7,26	5,13	6,06	4,40	5,62		3,43	5,98	4,70	5,77	5,09	4,28	3,72
KAL	6,48	4,03	6,99	3,15	3,41	5,89	7,68	3,97	6,60	3,29	6,08	6,17	4,47	5,40	4,68	6,19	3,43		5,41	4,40	5,14	5,23	3,16	4,21
ROM	4,07	4,44	3,40	5,32	5,60	5,20	3,14	4,51	5,43	4,96	2,48	3,91	4,30	3,88	4,59	3,84	5,98	5,41		5,09	4,95	2,98	5,11	6,05
KAG	4,10	5,70	5,51	4,89	5,84	3,78	6,93	2,59	4,44	4,81	4,49	4,13	3,24	3,55	3,11	5,85	4,70	4,40	5,09		3,87	3,60	5,50	5,43
GRC	4,51	6,08	5,71	5,43	6,45	3,18	6,85	3,86	5,68	5,39	4,86	5,01	4,86	4,24	3,48	6,40	5,77	5,14	4,95	3,87		4,25	5,77	5,77
KEM	3,83	5,08	4,10	4,84	5,36	4,55	4,39	3,44	5,06	4,54	3,15	3,79	3,62	2,99	3,33	4,35	5,09	5,23	2,98	3,60	4,25		5,14	5,16
FEA	7,08	2,77	7,20	2,16	1,86	6,64	7,16	4,83	8,07	1,94	6,29	7,21	5,03	6,01	5,35	6,48	4,28	3,16	5,11	5,50	5,77	5,14		2,98
TES	7,35	3,82	7,27	2,42	2,52	6,87	7,58	5,62	8,67	3,42	6,94	7,79	6,31	6,88	5,67	7,32	3,72	4,21	6,05	5,43	5,77	5,16	2,98	

Πίνακας 5.16. *Πίνακας ομοιότητας των Ευκλείδιων αποστάσεων.*

Βήμα	Ομάδες	Ομοιότητα %	Απόσταση	Σύνδεση ομάδων		Νέα ομάδα	Παρατηρήσεις
1	23	78,53	1,86	5	23	5	2
2	22	74,32	2,23	13	14	13	2
3	21	74,29	2,23	5	10	5	3
4	20	72,78	2,36	4	5	4	4
5	19	71,43	2,48	11	19	11	2
6	18	70,16	2,59	8	13	8	3
7	17	68,52	2,73	1	12	1	2
8	16	64,13	3,11	15	20	15	2
9	15	63,34	3,18	6	21	6	2
10	14	62,59	3,24	4	24	4	5
11	13	62,41	3,26	11	22	11	3
12	12	60,45	3,43	17	18	17	2
13	11	57,18	3,71	2	4	2	6
14	10	56,88	3,74	8	15	8	5
15	9	56,12	3,81	3	11	3	4
16	8	51,97	4,17	3	7	3	5
17	7	47,8	4,53	1	9	1	3
18	6	39,91	5,21	2	17	2	8
19	5	38,17	5,36	3	16	3	6
20	4	27,59	6,28	1	6	1	5
21	3	16,99	7,2	1	8	1	10
22	2	-37,23	11,9	1	3	1	16
23	1	-226,42	28,31	1	2	1	24

Πίνακας 5.17. *Βήματα συγχώνευσης των παρατηρήσεων σε ομάδες με τις Ευκλείδιες αποστάσεις και το ποσοστό ομοιότητας των ομάδων καθώς και λοιπά χαρακτηριστικά της.*

ΜΕΤΑΒΛΗΤΕΣ	ΣΚΛΗΡΑ	ΗΜΙΣΚΛΗΡΑ	ΜΑΛΑΚΑ
ΧΗΜΙΚΕΣ			
DM	62,62	61,87	47,12
FAT/DM	47,21	49,11	52,07
TN/DM	5,79	6,42	5,71
WSN/TN	27,71	30,73	20,78
NPN/TN	19,62	22,51	23,39
BC	9,91	6,08	4,73
pH	5,20	5,40	4,43
ΜΗΧΑΝΙΚΕΣ			
FR	32,03	14,73	7,34
REC	69,53	71,03	59,96
FB	67,84	44,53	16,43
DEF	25,35	37,69	27,75
ΟΡΓΑΝΟΛΗΠΤΙΚΕΣ			
Hard	10,50	7,41	4,40
Elas	6,02	8,73	4,32
Frac	8,84	5,87	10,56

Πίνακας 5.18. *Μέσες τιμές των χαρακτηριστικών των τυριών ανά τύπο τυριού.*

Τελική κατάτμηση των στοιχείων σε 3 ομάδες

	Παρατηρήσεις	ESS	Μέση απόσταση από το κέντρο βάρους	Μέγιστη απόσταση από το κέντρο βάρους
ΣΚΛΗΡΑ	6	37,89	2,43	3,38
ΗΜΙΣΚΛΗΡΑ	10	73,58	2,67	3,55
ΜΑΛΑΚΑ	8	36,66	2,06	2,91

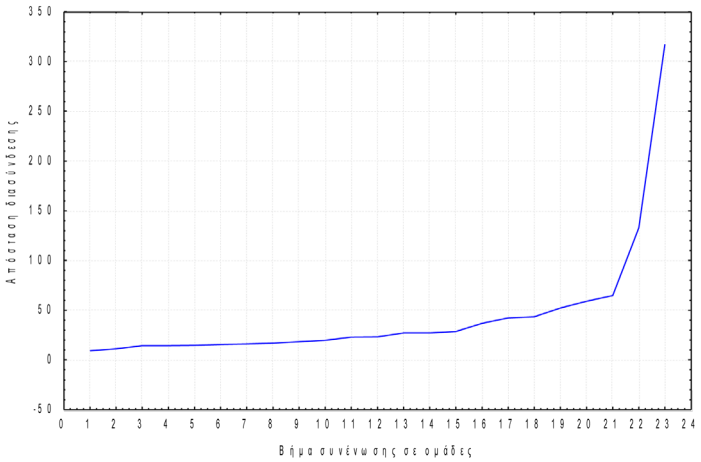
Κεντρικά βάρη (κεντρικές τιμές) των ομάδων

Μεταβλητή	ΣΚΛΗΡΑ	ΗΜΙΣΚΛΗΡΑ	ΜΑΛΑΚΑ
DM	-1,29	0,61	0,7
FAT/DM	0,71	-0,15	-0,69
TN/DM	-0,49	0,62	-0,37
WSN/TN	-0,68	0,47	0,12
NPN/TN	0,20	0,07	-0,37
BC	-0,70	-0,19	1,24
pH	-1,23	0,77	0,35
FR	-0,69	-0,14	1,15
REC	-1,16	0,67	0,43
FB	-0,94	0,14	1,03
DEF	-0,37	0,67	-0,63
Hard	-0,99	0,08	1,18
Elas	-0,98	0,93	-0,24
Frac	1,01	-0,98	0,28

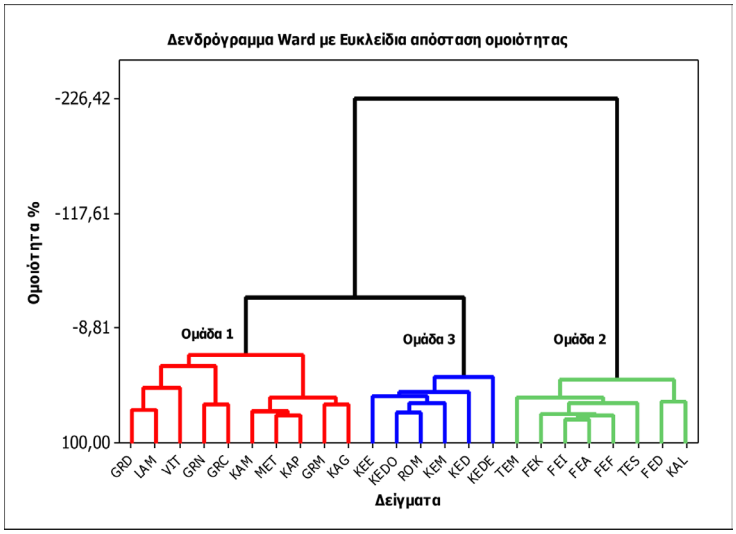
Αποστάσεις μεταξύ των κεντρικών τιμών των ομάδων

	ΣΚΛΗΡΑ	ΗΜΙΣΚΛΗΡΑ	ΜΑΛΑΚΑ
ΣΚΛΗΡΑ	0	3,49	5,36
ΗΜΙΣΚΛΗΡΑ	3,49	0	5,08
ΜΑΛΑΚΑ	5,36	5,08	0

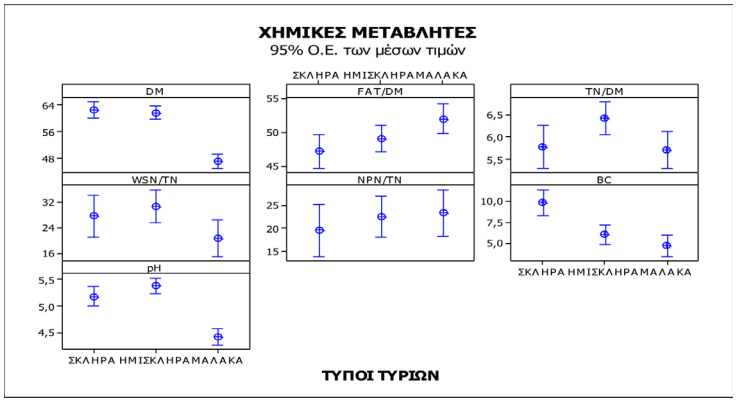
Πίνακας 5.19. *Αποτελέσματα της εκτίμησης των ομάδων ταξιδόμησης κατά Ward.*

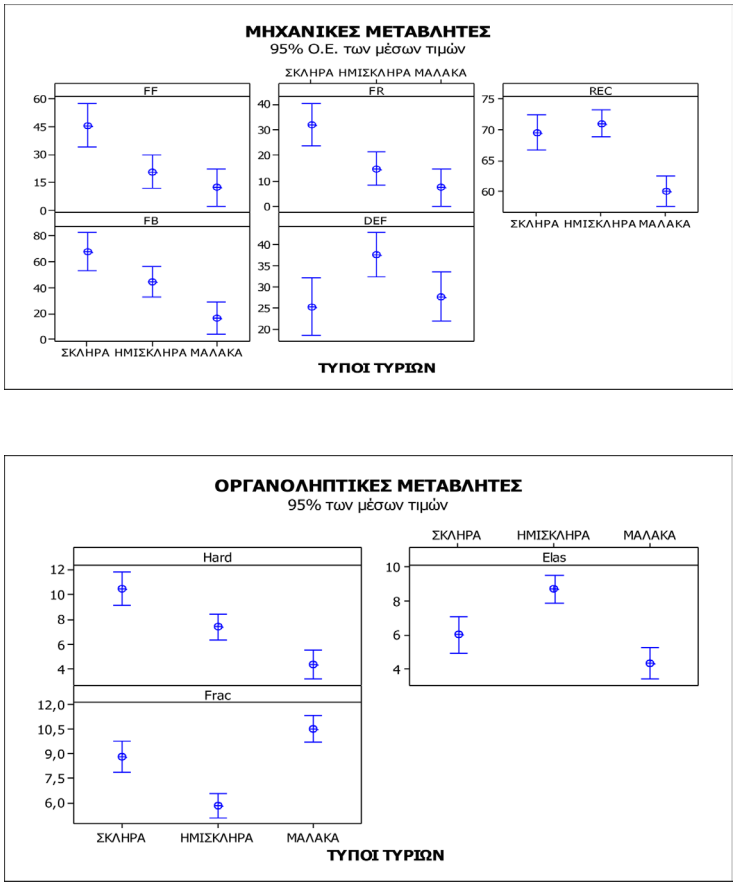


Σχήμα 5.4. *Γράφημα της απόστασης διασύνδεσης με τα βήματα συνένωσης σε ομάδες.Ο σχηματισμός αγκώνα περιλαμβάνει τα βήματα συνένωσης 21-23 και άρα τρεις ομάδες επιλέγονται για περαιτέρω επεξεργασία.*



Σχήμα 5.5. *Δενδρόγραμμα των 15 ποικιλιών τυρού με βάση την κατατομή των χαρακτηριστικών τους.*





Σχήμα 5.6. *Μέσες τιμές των χημικών, μηχανικών και οργανοληπτικών χαρακτηριστικών με βάση τις 3 ομάδες τυριών. Οι κάθετες γραμμές εκφράζουν τα 95% όρια εμπιστοσύνης (Ο.Ε.) των μέσων τιμών υπολογισμένα από την ανάλυση διακύμανσης. Μέσες τιμές των οποίων τα όρια εμπιστοσύνης δεν επικαλύπτονται δηλώνουν στατιστική σημαντικότητα μεταξύ των ομάδων.*

Μη ιεραρχική ταξιδόμηση

Τα στοιχεία υποβλήθηκαν στην ομαδοποίηση των k μέσων (δεύτερη επιλογή στην ενότητα 5.3) με τρεις ομάδες τυριών και στην ταξιδόμηση της κανονικής κατανομής τους με τον αλγόριθμο E-M (Πίν. 5.20).

Η πρώτη τεχνική ανέδειξε την ίδια ταυτότητα συμμετοχής των τυριών στις τρεις συστάδες με την ιεραρχική ανάλυση του Ward (Πίν. 5.20α) εμφανίζοντας μόνο μία αστοχία (η ποικιλία ΚΕΜ μετατοπίστηκε στην ομάδα 1 από την ομάδα 3). Οι αποστάσεις μεταξύ των συστάδων εμφάνισαν παρόμοια συμπεριφορά με την ιεραρχική του Ward (Πίν. 5.20β): μέγιστη διαφορά μεταξύ μαλακών-σκληρών τυριών, λίγο μικρότερη μεταξύ μαλακών-ημίσκληρων και αρκετά χαμηλότερη μεταξύ σκληρών-ημίσκληρων. Το γράφημα της συμπεριφοράς των μέσων τιμών των μεταβλητών στις 3 συστάδες (5.20γ) κατέδειξε χαμηλότερες τιμές στις περισσότερες εξ αυτών στα μαλακά τυριά και ιδιαίτερα στην ξηρή ουσία (DM). Λόγω του σχεδόν ομοιόμορφου διαχωρισμού των μελών στις συστάδες των k μέσων με αυτόν στην ιεραρχική του Ward, συμπεραίνουμε ότι οι τάσεις μεταβολής είναι παρόμοιες.

Ελάχιστα διαφοροποιημένη διευθέτηση των μελών στις 3 συστάδες παρέχεται και από την ανάλυση των κατανομών (Πίν. 20α) στην οποία παρατηρείται μεταπήδηση της ποικιλίας KEDE από τα σκληρά τυριά στα ημίσκληρα (ομάδες 3 και 1). Αξιοσημείωτο είναι ο πλήρης και σαφής διαχωρισμός των μαλακών τυριών και στις τρεις μεθόδους ταξιδόμησης.

Στα σχήματα 5.7α και 5.7β απεικονίζονται οι κανονικές κατανομές των σπουδαιότερων χαρακτηριστικών της μελέτης από τα οποία συμπεραίνουμε τα ακόλουθα:

Τα μαλακά τυριά διαγράφουν ευκρινώς διαχωρισμένες καμπύλες στις μεταβλητές DM, pH και REC με χαμηλές συγκριτικά τιμές. Δεν παρατηρείται σαφής διαχωρισμός μεταξύ των σκληρών και ημίσκληρων τυριών, είναι εμφανής όμως η τάση των σκληρών τυριών να μετατοπίζονται περισσότερα στα υψηλότερα της κλίμακας μέτρησης (προς τα δεξιά) απ' ό,τι τα ημίσκληρα. Η τάση στα σκληρά γίνεται ισχυρότερη στην παρουσία άλμης

(BC), μηχανικής (FB) και οργανοληπτικής σκληρότητας (Hard), ενώ στα ημίσκληρα παρατηρείται μόνο στην οργανοληπτική ελαστικότητα (Elas). Τα μαλακά είναι επίσης περισσότερο εύθρυπτα (Frac).

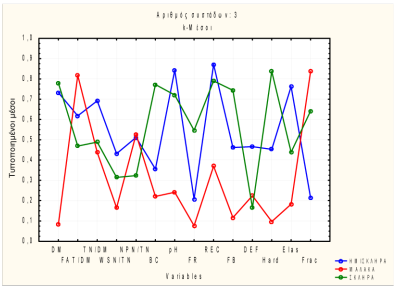
Συνοψίζοντας την αξιολόγηση των τριών ταξιδομήσεων, διαπιστώνουμε ότι όλες οι τεχνικές αποδεικνύονται εξαιρετικά αποτελεσματικές στο διαχωρισμό των στοιχείων σε τρεις διακριτικές ομάδες, με ιδιάζοντα χαρακτηριστικά εκάστη. Οι διαφορετικές επεξεργασίες που παραδοσιακά ακολουθούνται για την παρασκευή των τυριών αυτών βοήθησαν επίσης σημαντικά στη διαδικασία της ορθής ταξιδόμησης των στοιχείων.

Τυριά	Ward	K-μέσοι	Απόσταση	E-M	Πιθανότητα
GRD	1	1	0.81	1	0.999998
GRN	1	1	0.79	1	1.000000
KAM	1	1	0.63	1	1.000000
VIT	1	1	0.87	1	1.000000
LAM	1	1	0.73	1	1.000000
MET	1	1	0.73	1	1.000000
KAP	1	1	0.52	1	1.000000
GRM	1	1	0.67	1	1.000000
KAG	1	1	0.64	1	1.000000
GRC	1	1	0.82	1	1.000000
TEM	2	2	0.68	2	1.000000
FEK	2	2	0.41	2	1.000000
FEI	2	2	0.41	2	1.000000
FEF	2	2	0.48	2	1.000000
FED	2	2	0.75	2	1.000000
KAL	2	2	0.68	2	1.000000
FEA	2	2	0.43	2	1.000000
TES	2	2	0.60	2	1.000000
KEE	3	3	0.68	3	0.999994
KED	3	3	0.69	3	1.000000
KEDO	3	3	0.55	3	0.999926
KEDE	3	3	0.85	1	1.000000
ROM	3	3	0.41	3	0.999991
KEM	3	1	0.72	3	0.981819

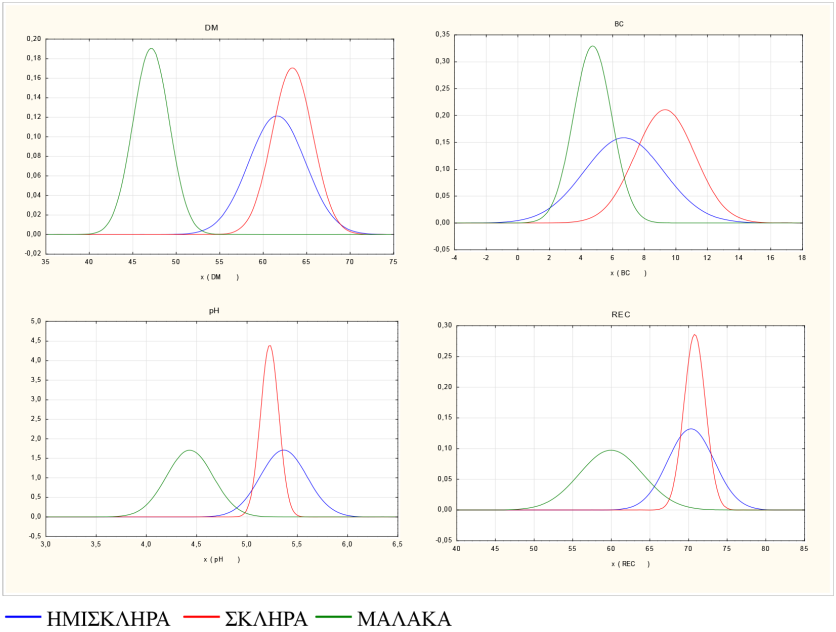
(β)

	ΣΚΛΗΡΑ	ΗΜΙΣΚΛΗΡΑ	ΜΑΛΑΚΑ
ΣΚΛΗΡΑ	0	1.02	1.63
ΗΜΙΣΚΛΗΡΑ	1.02	0	1.51
ΜΑΛΑΚΑ	1.63	1.51	0

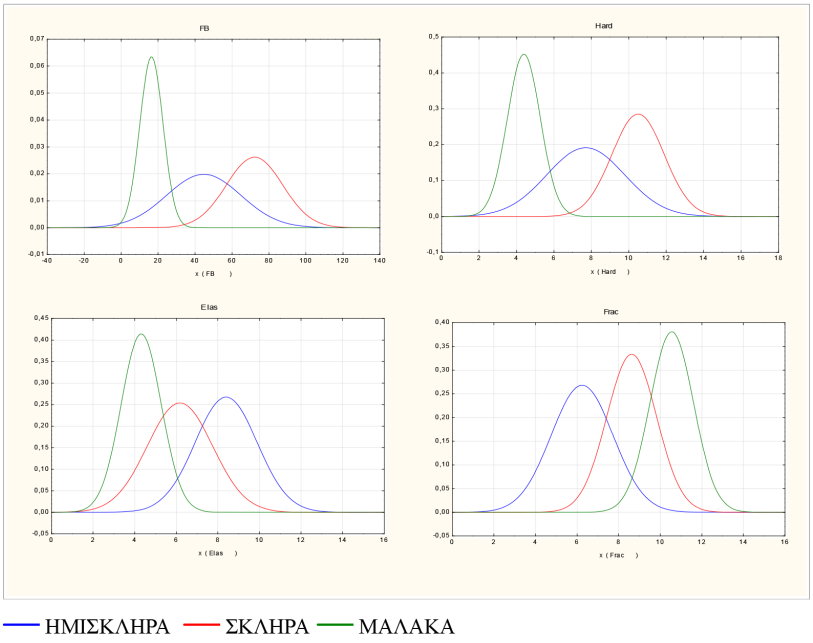
(γ)



Πίνακας 5.20. *Αποτελέσματα των στατιστικών παραμέτρων των μη ιεραρχικών μεθόδων ταξιδόμησης. α) Διευθέτηση των συστάδων των k μέσων, του αλγόριθμου E-M και της ιεραρχικής του Ward στα 24 τυριά της μελέτης. Τα σκιασμένα κελιά δείχνουν τις διαφορές ταξινόμησης συγκριτικά με την ιεραρχική του Ward. β) Αποστάσεις μεταξύ των συστάδων με τη μέθοδο των k μέσων. γ) Κατανομή των μέσων τιμών των μεταβλητών της μελέτης στις τρεις συστάδες των k μέσων.*



Σχήμα 5.7α. *Κανονική κατανομή των στατιστικά σημαντικότερων φυσικοχημικών μεταβλητών και της ανάκαμψης REC στις τρεις συστάδες με βάση υπολογισμού των μικτών κατανομών του αλγόριθμου E-M.*



Σχήμα 5.7β. *Κανονική κατανομή των στατιστικά σημαντικών οργανοληπτικών μεταβλητών και της σκληρότητας FB στις τρεις συστάδες με βάση υπολογισμού των μικτών κατανομών του αλγόριθμου Ε-Μ.*

5.6 Βιβλιογραφία

Aldenderfer M.S. & Blashfield R.K. (1984). *Cluster Analysis.* Sage Publications, Newbury Park, 88 p.

Bishop C.M. (1995). *Neural networks for pattern recognition.* Oxford: Clarendon Press, 482 p.

Dempster A.P, Laird M. and Rubin D.B. (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society, Series B, 39(1), 1–38.

Everitt, B. S. (1993). *Cluster analysis.* 3rd ed. N. York: Halsted Press,

Gauch H.G. & Whittaker. R.H. (1981). Hierarchical classification of community data. *Journal of Ecology,* 69, 135-152.

Hartigan J. A. & Wong M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society,* Series C, 28(1), 100–108.

Hill M.O. (1979b). *TWINSPAN—A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Ecology and Systematics*, Cornell University, Ithaca, NY, 90 p.

Lance G.N. & Williams W.T. (1967). A general theory of classification sorting strategies. I. hierarchical systems. *Computer Journal,* 9, 373-380.

Peck G.E. (2010). *Multivariate analysis for community ecologists.* MjM Software Design, Gleneden Beach, Oregon, 162 p.

Σιάρδος Κ.Γ. (2015). *Μέθοδοι πολυμεταβλητής στατιστικής ανάλυσης, με την επίλυση ασκήσεων μέσω του προγράμματος SPSS.* Εκδόσεις Σταμούλη Α.Ε., Αθήνα, (υπό έκδοση).