ΠΟΛΥΤΕΧΝΕΙΟ
ΚΡΗΤΗΣ

Σχολή Μηχανικών Παραγωγής και Διοίκησης

Πρόγραμμα Μεταπτυχιακών Σπουδών
"Εφαρμοσμένα μαθηματικά για μηχανικούς"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εφαρμοσμένη Ανάλυση Συστάδων

(Applied Cluster Analysis)

Στρατινάκης Νικόλαος

Α.Μ. 2014019052

Επιβλέπων Καθηγητής: Τρύφων Δάρας

Χανιά, 2018

Π Ε Ρ Ι Λ Η Ψ Η

Η ανάλυση συστάδων (cluster analysis) είναι μια μέθοδος που έχει σαν σκοπό να κατατάξει σε ομάδες υπάρχουσες παρατηρήσεις χρησιμοποιώντας την πληροφορία που υπάρχει σε κάποιες μεταβλητές. Κοιτάζοντας δηλαδή τις παρατηρήσεις μπορεί να πει κανείς πόσο "όμοιες" είναι ως προς κάποιον αριθμό μεταβλητών, δημιουργώντας ομάδες (από παρατηρήσεις) που μοιάζουν μεταξύ τους. Μια επιτυχημένη ανάλυση θα πρέπει να καταλήξει σε ομάδες για τις οποίες οι παρατηρήσεις μέσα σε κάθε ομάδα θα είναι όσο γίνεται πιο ομοιογενείς αλλά παρατηρήσεις διαφορετικών ομάδων θα διαφέρουν όσο γίνεται περισσότερο. Η ομαδοποίηση γίνεται με τη βοήθεια της έννοιας της απόστασης ή της ομοιότητας

Η ανάλυση συστάδων είναι σημαντική όχι μόνο σε επιστήμες όπως η κοινωνιολογία, η βιολογία και η στατιστική, αλλά και σε πολλούς τομείς της πληροφορικής όπως η αναγνώριση προτύπων, η εξόρυξη γνώσης, η ανάκτηση δεδομένων, η τεχνητή νοημοσύνη και η μηχανική μάθηση.

Στο 1ο κεφάλαιο της διπλωματικής εργασίας δίνεται μια εισαγωγή στην Ανάλυση Συστάδων (φιλοσοφία της Α.Σ., μέθοδοι, πλεονεκτήματα/μειονεκτήματα της Α.Σ. και τέλος προβλήματα εφαρμογής της). Στο 2^{ο} κεφάλαιο περιγράφονται τα κυριότερα μέτρα απόστασης και ομοιότητας (ανάλογα με το είδος των μεταβλητών) που χρησιμοποιεί κανείς στην Α.Σ. και δίνονται αναλυτικά παραδείγματα. Στο 3^{ο} κεφάλαιο αναφέρονται οι βασικές μέθοδοι σύνδεσης ομάδων και περιγράφονται οι ιεραρχικές μέθοδοι ταξινόμησης, με αναλυτικά παραδείγματα. Στο 4^{ο} κεφάλαιο περιγράφεται η μέθοδος μη ιεραρχικής ταξινόμησης k-means. Τέλος, στο 5^{ο} κεφάλαιο δίνεται μια εφαρμογή της Α.Σ. στην ταξινόμηση σταθμών μέτρησης ατμοσφαιρικής ρύπανσης στην περιοχή του Λεκανοπεδίου Αττικής.

A B S T R A C T

Cluster analysis is a method designed to classify existing observations using the information that exists in some variables. Looking at the observations, one can say how similar they are to a number of variables, creating groups (from observations) that resemble each other. A successful analysis should result in groups for which the observations within each group will be as homogeneous as possible, but observations of different groups will vary as much as possible. Grouping takes place with the help of the concept of distance or similarity.

Cluster analysis is important not only in many sciences such as sociology, biology and statistics, but also in many areas of information technology such as pattern recognition, knowledge mining, data recovery, artificial intelligence, and mechanical learning.

In the 1st chapter of the thesis there is an introduction to the Cluster Analysis (philosophy of the C.A., methods, advantages / disadvantages of the C.A. and finally problems of its implementation). The 2nd chapter describes the main measures of distance and similarity (depending on the type of variables) used by the C.A. and detailed examples are given. Chapter 3 lists the basic methods of group linkage and describes hierarchical classification methods, with examples. The 4th chapter describes the non-hierarchical k-means method. Finally, in the 5th chapter an application of the C.A., in the classification of atmospheric pollution measuring stations in the Attica region, is given.

Π Ε Ρ Ι Ε Χ Ο Μ Ε Ν Α

Κεφάλαιο 5: Ταξινόμηση σταθμών μέτρησης ατμοσφαιρικής ρύπανσης

Εικόνες

Κεφάλαιο 1: Εισαγωγή στην Ανάλυση Συστάδων

1.1 Εισαγωγή

Όταν μελετάμε πολυμεταβλητά δεδομένα - δηλαδή ένα σύνολο περιπτώσεων/αντικειμένων (cases) ως προς ένα σύνολο μεταβλητών του ενδιαφέροντός μας – είναι αρκετές φορές δύσκολο να διακρίνουμε ομοιότητες ή σχέσεις μεταξύ των αντικειμένων Σκοπός της λεγόμενης **Ανάλυσης Συστάδων** (Α.Σ.) (**cluster analysis**) είναι η δημιουργία ομάδων ομοειδών αντικειμένων με τη βοήθεια των μεταβλητών. Π.χ.:

Οι γεωπόνοι θέλουν να κατατάξουν τα διάφορα είδη φυτών σε ομάδες, ανάλογα με κάποιο(α) συγκεκριμένο(α) χαρακτηριστικό(α) τους, για την αποτελεσματικότερη μελέτη τους.

Τα άτομα που σχεδιάζουν ιστοσελίδες θέλουν να κατατάξουν τους χρήστες σε ομάδες ανάλογα με τις προτιμήσεις που έχουν ως προς την πλοήγησή (το σερφάρισμα) τους, για τη δημιουργία στοχευμένων διαφημίσεων κ.λ.π..

1.2 Καταγραφή δεδομένων

(α) Είδη μεταβλητών

Βασικό ρόλο στην Ανάλυση Συστάδων παίζουν οι μεταβλητές του ενδιαφέροντός μας, δηλαδή τα χαρακτηριστικά ως προς τα οποία εξετάζουμε καθένα από τα αντικείμενά μας (δείγμα). Υπενθυμίζουμε παρακάτω τα διαφορετικά είδη μεταβλητών που συναντά κανείς κατά τη διάρκεια μιας (στατιστικής) έρευνας.

Οι **μεταβλητές** ενός πληθυσμού χωρίζονται σε τρία είδη:

- **Ποσοτικές**: όταν οι τιμές της μεταβλητής είναι **αριθμητικές**. Π.χ.ο αριθμός των παιδιών μιας οικογένειας,

  Οι ποσοτικές μεταβλητές χωρίζονται σε:

  - **Διακριτές**: αν το πλήθος των τιμών της μεταβλητής είναι πεπερασμένο ή το πολύ αριθμήσιμο. Π.χ. Ο αριθμός των παιδιών μιας οικογένειας

  - **Συνεχείς**: αν το πλήθος των τιμών της μεταβλητής είναι μη-αριθμήσιμο. Π.χ. ο χρόνος που απαιτείται για τη συμπλήρωση ενός γραπτού τεστ (η μεταβλητή μπορεί θεωρητικά να πάρει όλες τις τιμές μεταξύ δύο τιμών).

- **Ποιοτικές ή διάταξης (ordinal)**: οι τιμές της μεταβλητής είναι μη-αριθμητικές και επιδέχονται κάποιου είδους διάταξη π.χ. η διάκριση των καθηγητών ενός πανεπιστημίου σε τακτικούς, αναπληρωτές, επίκουρους ή λέκτορες (υπάρχει ιεράρχηση στα κλιμάκια αυτά).

- **Κατηγορικές ή ονομαστικές (nominal)**: οι τιμές της μεταβλητής και σ' αυτή την περίπτωση είναι μη-αριθμητικές αλλά δεν επιδέχονται κάποιου είδους διάταξη. Π.χ. το χρώμα των μαλλιών ενός ατόμου (μαύρο, καστανό, ξανθό κ.ο.κ.).

  Ένα είδος κατηγορικών μεταβλητών είναι και οι **διχοτομικές**. Παίρνουν δύο τιμές (συνήθως 1 και 0, όπου με 1 συμβολίζουμε την ύπαρξη ενός χαρακτηριστικού και 0 την απουσία του). Π.χ. φύλο (άνδρας, γυναίκα), ύπαρξη συγκεκριμένης ασθένειας σε άτομο ή όχι. Έάν οι τιμές της διχοτομικής μεταβλητής έχουν την ίδια αξία (π.χ. άνδρας/γυναίκα) τότε η μεταβλητή καλείται **συμμετρική**, εάν όχι (π΄χ. ύπαρξη συμπτώματος αθένειας) τότε καλείται **μη συμμετρική**.

(β) Τρόποι καταγραφής δεδομένων

(Ι) (Ποσοτικές μεταβλητές)

Έστω ότι όλες οι μεταβλητές, ως προς τις οποίες εξετάζουμε καθένα από τα αντικείμενα στο δείγμα μας, είναι ποσοτικές. Τα δεδομένα μας λοιπόν, μπορούμε να περιγράψουμε με τη βοήθεια ενός πίνακα με οριζόντιες γραμμές τις τιμές κάθε μιάς των **n** περιπτώσεών/ αντικειμένων **(cases)** μας ως προς (κάθε μια από) τις μεταβλητές και στήλες τις ***p*** μεταβλητές του ενδιαφέροντός μας. Ο πίνακας των δεδομένων μας έχει δηλαδή έχει την παρακάτω γενική μορφή.

		Μεταβλητές (variables)			
		M_1	M_2	…	M_p
Περιπτώσεις (cases)	x_1	α_{11}	a_{12}	…	α_{1p}
	x_2	α_{21}	α_{22}	…	α_{2p}
	…	…	…	…	…
	x_n	α_{n1}	α_{n2}	…	α_{np}

Εικόνα 1: πίνακας καταγραφής δεδομένων

Ένα τέτοιος πίνακας μπορεί να περιέχει και τις τιμές ποιοτικών/ κατηγορικών μεταβλητών.

(ΙΙ) Διχοτομικές μεταβλητές

Στην περίπτωση διχοτομικών μεταβλητών, δηλαδή μεταβλητών με δύο δυνατές τιμές, έχουμε ανά δύο χαρακτηριστικά έναν πίνακα της μορφής

		αντικείμενο y		
		1	0	
αντικείμενο x	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	

Εικόνα 2: πίνακας καταγραφής διχοτομικών δεδομένων

Όπου a : το πλήθος των μεταβλητών που το αντικείμενο *x* έχει την τιμή 1 και το *y* την τιμή 1, (τα b, c, d ορίζονται ανάλογα).

1.3 Σκοπός της Ανάλυσης Συστάδων

Η **ανάλυση συστάδων** (cluster analysis) είναι μία οικογένεια μεθόδων ανάλυσης (πολυμεταβλητών) δεδομένων που στοχεύει στην ταξινόμηση/δημιουργία ομάδων ενός δείγματος ατόμων/αντικειμένων με βάση όπως είδαμε μια σειρά από μεταβλητές. Οι ομάδες καλούνται **τάξεις** (clusters), **κλάσεις ή συστάδες**. Η ταξινόμηση/δημιουργία ομάδων γίνεται κατά τέτοιο τρόπο ώστε τα αντικείμενα κάθε ομάδας να έχουν «ομοειδή» χαρακτηριστικά. Μ' άλλα λόγια, σε κάθε ομάδα θέλουμε να υπάρχει εσωτερική ομοιογένεια (όσο το δυνατόν μεγαλύτερη σχέση μεταξύ των στοιχείων της) ενώ μεταξύ των ομάδων να υπάρχει όσο το δυνατόν μεγαλύτερη ανομοιογένεια (σχέση των στοιχείων διαφορετικών ομάδων ελάχιστη δυνατή). Ακόμα, η ανάλυση, προσπαθεί να ανακαλύψει τον αριθμό και τη σύνθεση των ομάδων.

Η αρχική εικόνα που έχει κανείς όταν του δοθούν πολυμεταβλητά δεδομένα (και ιδιαίτερα όταν έχουμε μεγάλο αριθμό περιπτώσεων ή μεταβλητών/χαρακτηριστικών) είναι ασαφής και η εξαγωγή συμπερασμάτων γι αυτά ιδιαίτερα δύσκολη. Με τη βοήθεια λοιπόν της ανάλυσης συστάδων προσπαθούμε να αποκτήσουμε κάποιες (επιπλέον) γνώσεις για τα δεδομένα μας, όπως ομοιότητες μεταξύ τους, παρουσία ή απουσία χαρακτηριστικών κ.λ.π. Θέλουμε να διαπιστώσουμε εάν υπάρχουν σχέσεις/τάσεις που τυχόν τα χαρακτηρίζουν.

1.3.1 Παρατήρηση (Ανάλυση Συστάδων – Διαχωριστική Ανάλυση).

Μια μέθοδος ανάλογη της Ανάλυσης Συστάδων είναι και η λεγόμενη **Πολλαπλή Διακριτική Ανάλυση** (**multiple discriminant analysis**). Οι διαφορές των μεθόδων περιγράφονται παρακάτω.

Η **Ανάλυση Συστάδων** είναι μέθοδος διερευνητική. Παίρνει ένα σύνολο δεδομένων και αναζητά την «καλύτερη» λύση/τρόπο ταξινόμησης τους. Το "καλύτερο" εξαρτάται από τη μέθοδο που χρησιμοποιούμε, αλλά στην ουσία προσπαθούμε να βρούμε εάν υπάρχουν μοτίβα ταξινόμησης. Για παράδειγμα, θέλουμε πολλές φορές να αναλύσουμε ένα σύνολο δεδομένων αγοραστικής συμπεριφοράς και να «βρούμε» αν υπάρχουν ομάδες καταναλωτών που είναι περισσότερο παρόμοιες με αυτές της ομάδας τους από ό, τι σε καταναλωτές σε άλλες ομάδες.

Η **Πολλαπλή Διακριτική Ανάλυση** είναι διαφορετική. Είναι ένα στατιστικό εργαλείο που έχει σα στόχο να εκτιμηθεί η καταλληλότητα μιας ταξινόμησης (οι ομάδες εδώ είναι γνωστές εκ των προτέρων) ή να ταξινομήσει ένα νέο αντικείμενο σε κάποια από τις γνωστές ομάδες. Χρησιμοποιείται συνήθως μετά την Ανάλυση Συστάδων, επειδή η Α.Σ. δεν διαθέτει κριτήρια μέτρησης της καλής προσαρμογής του επιχειρούμενου μοντέλου. Η Α.Σ. βασίζεται ουσιαστικά στη Διακριτική Ανάλυση για 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.

1.3.2 Παρατήρηση (ταξινόμηση μεταβλητών)

Κατά τρόπο ανάλογο της ταξινόμησης των παρατηρήσεων, είναι δυνατή και η **ομαδοποίηση των μεταβλητών** (γίνεται κατά κάποιο τρόπο μια «αντιμετάθεση» περιπτώσεων και μεταβλητών). Εδώ να σημειώσει κανείς ότι, για να έχει ουσιαστικό νόημα η ομαδοποίηση αυτή απαιτούνται τουλάχιστον 3 μεταβλητές.

1.4 Μέθοδοι ταξινόμησης (ανάλογα με τον τρόπο δημιουργίας των ομάδων)

Υπάρχουν διαφορετικοί τρόποι αντιμετώπισης (μέθοδοι ταξινόμησης) του προβλήματος της ομαδοποίησης δεδομένων. Ορισμένοι από αυτούς αναφέρονται στη συνέχεια.

- **Ιεραρχικές** (hierarchical)

  Είναι μέθοδοι στις οποίες υπάρχει ένα είδος ιεράρχισης στις ομάδες που δημιουργούνται σε κάθε επίπεδο εφαρμογής της μεθόδου.

  Π.χ. μια ιεραρχική ταξινόμηση των ζώων φαίνεται στον παρακάτω πίνακα

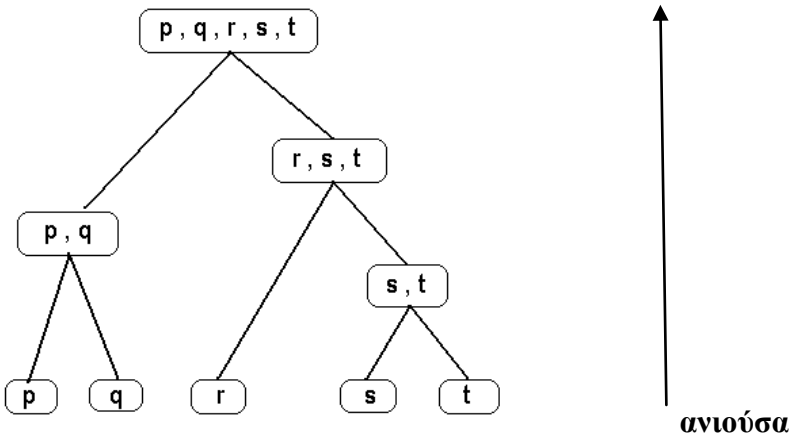
Εικόνα 3: Ιεραρχική ταξινόμηση ζώων

http://daskalosa.eu/physics_st/st_fysika_05_zoa.html

Οι ιεραρχικές ταξινομήσεις είναι κατάλληλες για μικρό συνήθως αριθμό περιπτώσεων και για το ίδιο είδος, κάθε φορά, μεταβλητών (ποσοτικές, διχοτομικές κλπ).

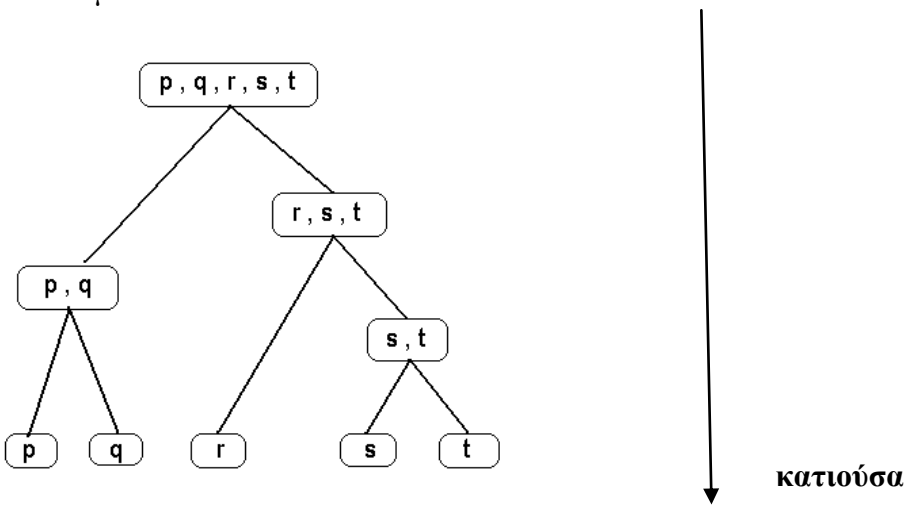
Οι ιεραρχικές μέθοδοι χωρίζονται σε δύο υπομεθόδους:

- την **ιεραρχική συσσωρευτική μέθοδο** ταξινόμησης (hierarchical agglomarative clustering) ή **την ανιούσα ιεραρχική ταξινόμηση:** μέθοδος κατά την οποία δημιουργούνται (νέες) ομάδες με (επαναληπτικές) ενώσεις στοιχείων/ομάδων. Ακόμα έχουμε ιεράρχιση αυτών των ομάδων. Είναι η δημοφιλέστερη από τις μεθόδους ταξινόμησης.



Εικόνα 4: ανιούσα ιεραρχική ταξινόμηση

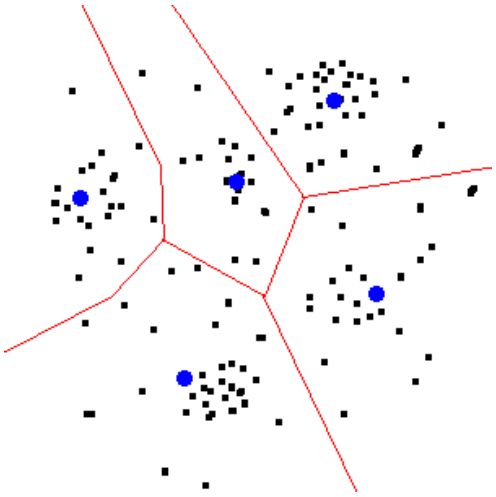
- την **ιεραρχική διαχωριστική μέθοδο ταξινόμησης** (hierarchical divisive clustering) ή **κατιούσα ιεραρχική ταξινόμηση:** μέθοδος κατά την οποία δημιουργούνται (νέες) ομάδες με (επαναληπτικές) διχοτομήσεις στοιχείων/ομάδων. Ακόμα έχουμε και εδώ ιεράρχιση αυτών των ομάδων.



Εικόνα 5: κατιούσα ιεραρχική ταξινόμηση

- Μ**η ιεραρχικές** (partitioning).

  Εδώ δημιουργούμε ομάδες δεδομένων με τη βοήθεια κάποιων στοιχείων που λαμβάνονται σαν **κέντρα.** Σπουδαιότερη των μεθόδων είναι η λεγόμενη **k-means**



Εικόνα 6: μέθοδος k-means

https://www.youtube.com/watch?v=_aWzGGNrcic

Η μη ιεραρχική ταξινόμηση είναι κατάλληλη για μεγάλο αριθμό περιπτώσεων και για ποσοτικές μόνον μεταβλητές.

- **Στατιστικές**: οι παραπάνω δυό κατηγορίες μεθόδων ταξινόμησης είναι αλγοριθμικές (δηλαδή κατατάσσουν τα δεδομένα σε ομάδες με τη βοήθεια επαναληπτικών διαδικασιών). Υπάρχουν μέθοδοι (στατιστικές) όπου κατατάσσουμε τις παρατηρήσεις μας με τη βοήθεια κάθε φορά συγκεκριμένων υποθέσεων. Δεν προσφέρονται, λόγω της πολυπλοκότητας τους, από τα στατιστικά πακέτα

Παρατήρηση 1.4.1

Επιπλέον διαχωρισμοί των μεθόδων είναι και οι εξής:

(α) **Μονοθεσική** όπου η ταξινόμηση των στοιχείων γίνεται με τη βοήθεια ενός μόνον χαρακτηριστικού/μεταβλητής – **πολυθεσική** όπου η ταξινόμηση γίνεται με τη βοήθεια όλως των μεταβλητών που διαθέτουμε.

(β) **Ισχυρή** όταν κάθε στοιχείο ανήκει σε μια μόνον ομάδα – **ασαφής** όταν κάθε στοιχείο ανήκει σε μια ομάδα με ορισμένο βαθμό (συνάφειας).

(γ) **ποσοτική** όταν οι μεταβλητές/χαρακτηριστικά είναι ποσοτικές – **ποιοτική** όταν τα δεδομένα λαμβάνονται από διχοτομικές μεταβλητές (παρουσία - απουσία χαρακτηριστικού)

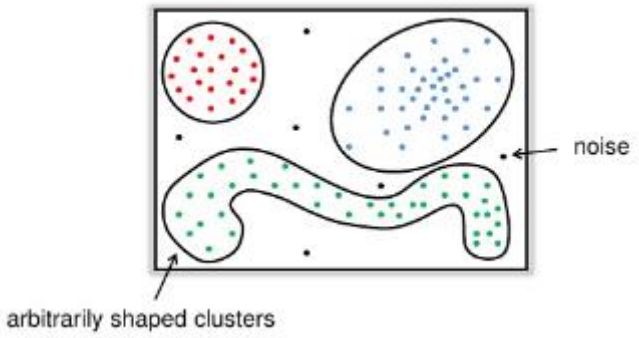
Παρατήρηση 1.4.2

Οι μέθοδοι ταξινόμησης (Ανάλυση Συστάδων) είναι αλγόριθμοι δημιουργίας ομάδων και όχι (Στατιστικά) τεστ σημαντικότητας. Είναι διερευνητικές μέθοδοι και χρησιμοποιούνται στα αρχικά στάδια των ερευνών. Η ταξινόμηση ανακαλύπτει τυχόν δομές που υπάρχουν στα δεδομένα, χωρίς όμως να είναι σε θέση να εξηγήσει την ύπαρξη αυτών των δομών.

Παρατήρηση 1.4.3

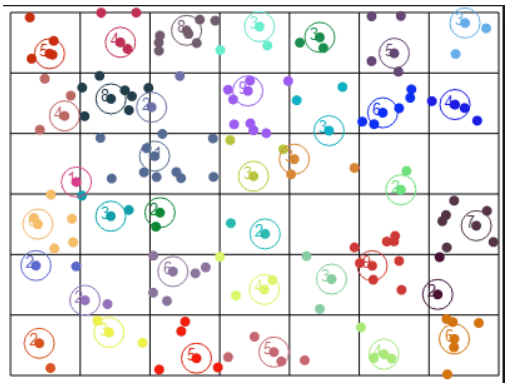
Υπάρχουν ακόμα μέθοδοι ταξινόμησης:

(α) **Μέθοδοι βασισμένες στην πυκνότητα (density based methods)**. Ταξινομούν τα αντικείμενα στο χώρο σε περιοχές υψηλής πυκνότητας, οι οποίες χωρίζονται από περιοχές χαμηλής πυκνότητας. Για κάθε παρατήρηση ορίζεται η γειτονιά της, η οποία είναι καθορισμένης διαμέτρου, και η οποία πρέπει να περιλαμβάνει έναν ελάχιστο αριθμό παρατηρήσεων. Η ομάδα/συστάδα επεκτείνεται όσο η γειτονιά των παρακείμενων σημείων διαθέτει την απαιτούμενη πυκνότητα. Οι μέθοδοι αυτές μπορούν να δημιουργήσουν ομάδες οποιουδήποτε σχήματος.



Εικόνα 7: Ταξινόμηση βασισμένη στην πυκνότητα
www.slideshare.net/ssakpi/density-based-clustering

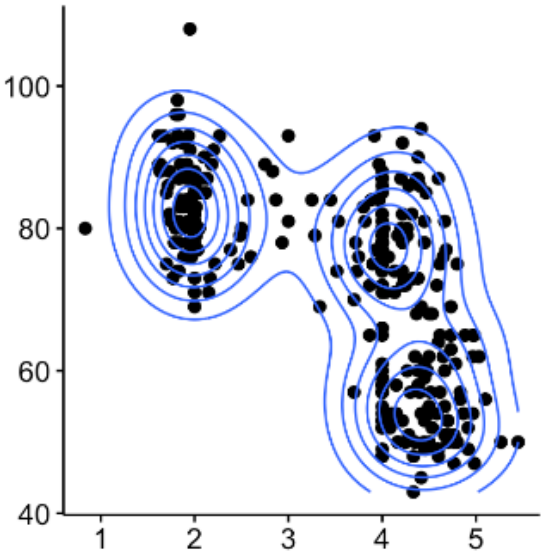
(β) **Μέθοδοι πλέγματος (grid based methods)** Ο χώρος (χωρικός) των δεδομένων χωρίζεται σε διακριτά κελιά (grids), τα οποία συγκροτούν ένα πλέγμα. Τα αντικείμενα πλέον αντιπροσωπεύονται από τα κελιά στα οποία ανήκουν, η δε αναζήτηση των ομάδων γίνεται στα κελιά του πλέγματος. Οι μέθοδοι εξαρτώνται από τον αριθμό και το μέγεθος των κελιών, τον προσανατολισμό τους κ.λ.π.



Εικόνα 8: Ταξινόμηση πλέγματος

https://kunuk.wordpress.com/2011/09/17/clustering-k-means-and-grid-with-c-example-and-html-canvas-part-2/

(γ) **Μέθοδοι βασισμένες σε μοντέλα** (model based methods). Εδώ θεωρούμε ότι τα δεδομένα προέρχονται/περιγράφονται από δύο ή περισσότερες πιθανοθεωρητικές κατανομές (κάθε ομάδα περιγράφεται από μια κατανομή). Αναζητούμε λοιπόν μια κατανομή η οποία είναι **μίξη** των κατανομών αυτών. Προσπαθούμε μ' άλλα λόγια να βελτιστοποιήσουμε την προσαρμογή μεταξύ των δεδομένων και κάποιου μαθηματικού μοντέλου.



Εικόνα 9: ταξινόμηση βασισμένη σε μοντέλα

http://www.sthda.com/english/articles/30-advanced-clustering/104-model-based-clustering-essentials/

(στο σχήμα φαίνεται ότι η κατανομή που περιγράφει τα δεδομένα είναι μίξη $3^{ων}$ ομοειδών κατανομών πιθανότητας).

1.5 Φιλοσοφία των μεθόδων

(α) Ανιούσα ιεραρχική ταξινόμηση

Η φιλοσοφία δημιουργίας ομάδων με τη βοήθεια της **ανιούσας ιεραρχικής ταξινόμησης** είναι η εξής: σε κάθε βήμα γίνεται συνένωση των δυό «πλησιέστερων» στοιχείων (στοιχεία θεωρούνται τα αρχικά δεδομένα όσο και οι ομάδες που έχουν δημιουργηθεί κατά το προηγούμενο βήμα της ταξινόμησης).

Για τη δημιουργία λοιπόν ομάδων και το χαρακτηρισμό στοιχείων σαν **«πλησιέστερα»,** χρειαζόμαστε να δώσουμε έναν ορισμό της **«κοντινότητας»** (δυό στοιχείων). Χρειαζόμαστε δηλαδή ένα είδος **απόστασης** (για ποσοτικά δεδομένα) **ή ομοιότητας**. Χρειαζόμαστε ακόμα και ένα **κριτήριο** συνένωσης των ομάδων (χαρακτηρισμός της κοντινότητας). Τα δύο αυτά χαρακτηριστικά *απόσταση/ομοιότητα* και *κριτήριο* δημιουργίας (συνένωσης) ομάδων δίνουν περαιτέρω υποδιαιρέσεις στις μεθόδους ταξινόμησης.

Παρατήρηση 1.5.1

(i) Εάν χρησιμοποιήσει κανείς δυό διαφορετικές αποστάσεις (κριτήρια) δεν είναι απαραίτητο να δημιουργηθούν οι ίδιες ομάδες.

(ii) «Κοντινότερο» εδώ δεν σημαίνει αυτό που απέχει σε απόσταση λιγότερο (το κοντινότερο είναι συνάρτηση του κριτηρίου/χαρακτηρισμού της κοντινότητας που χρησιμοποιούμε κάθε φορά).

(Β) K-means

Ο αλγόριθμος ξεκινά με την τυχαία επιλογή k αντικειμένων (από το σύνολο δεδομένων ή κάποια άλλη επιλογή) που θα χρησιμεύσουν σαν αρχικά κέντρα για τις ομάδες.. Στη συνέχεια, κάθε ένα από τα υπόλοιπα αντικείμενα αντιστοιχίζονται στο πλησιέστερο (αναφορικά με τη λεγόμενη **Ευκλείδεια απόσταση**) κέντρο βάρους/ομάδα. Κατόπιν, ο αλγόριθμος υπολογίζει το νέο κέντρο βάρους κάθε ομάδας Μετά τον υπολογισμό, κάθε παρατήρηση ελέγχεται και αντιστοιχίζεται και πάλι σ' ένα κέντρο βάρους/ομάδα. Τα βήματα επαναλαμβάνονται, μέχρι να μην γίνονται αλλαγές στις αντιστοιχίες των σημείων με τα κέντρα βάρους, δηλαδή οι ομάδες που σχηματίστηκαν σε ένα βήμα εφαρμογής του αλγόριθμου να είναι ίδιες (ή περίπου ίδιες) με αυτές που θα σχηματίζονταν σε μιά επιπλέον εφαρμογή του.

1.6 Πλεονεκτήματα – μειονεκτήματα Ανάλυσης Συστάδων

Πλεονεκτήματα

(α) Όπως αναφέραμε, με τη βοήθεια της Α.Σ. ανακαλύπτουμε την (τυχόν) ύπαρξη σχέσεων μεταξύ των στοιχείων/περιπτώσεων, με αποτέλεσμα να έχουμε σαφέστερη εικόνα για τα δεδομένα μας.

(β) Δεν απαιτούνται προϋποθέσεις για την εφαρμογή των μεθόδων (όπως π.χ. κανονικότητα δεδομένων/ μεταβλητών), με αποτέλεσμα να μην γίνονται στατιστικοί έλεγχοι παραμέτρων, υποθέσεων πριν την ανάλυση.

(γ) Με την ευρεία χρήση των υπολογιστών και την τήρηση δεδομένων σε πολλές δραστηριότητες της καθημερινότητάς μας, έχει γίνει επιτακτική η ανάγκη μελέτης μεγάλων βάσεων δεδομένων και ακόμα πιο επιτακτική η ανάγκη μείωσης των διαστάσεων αυτών των δεδομένων (επιλέγοντας κάθε φορά τις πιο «σχετικές/ ουσιαστικές» μεταβλητές/ χαρακτηριστικά). Με τη βοήθεια της Α.Σ και ομαδοποίησης των δεδομένων, μπορεί κανείς να δει ποιες από αυτές τις μεταβλητές παίζουν σημαντικό ρόλο στην ομαδοποίηση (διαφοροποιούν σημαντικά τις ομάδες) να επικεντρωθεί σε αυτές αγνοώντας τις υπόλοιπες. Ακόμα, λιγότερες μεταβλητές σημαίνει μικρότερο χρόνο επεξεργασίας των δεδομένων, με συνέπεια μικρότερο (λειτουργικό) κόστος

(δ) Με τη βοήθεια της διαχωριστικής ανάλυσης, μπορούν να καταταγούν στις ήδη υπάρχουσες ομάδες, νέες τιμές/παρατηρήσεις.

Μειονεκτήματα

(α) Η ανάλυση δεν είναι σε θέση να ερμηνεύσει τις τυχόν σχέσεις που μπορεί να υπάρχουν ανάμεσα στα στοιχεία μας. Μπορούμε δηλαδή να καταλήξουμε σε (υποθετικές) ομάδες χωρίς καμιά ουσιαστική ερμηνεία.

(β) Δεν είναι δυνατοί στατιστικοί έλεγχοι για τη σημαντικότητα των αποτελεσμάτων. Δεν είναι δυνατή δηλαδή μια «στατιστική» περιγραφή των ομάδων.

(γ) Δεν διαθέτει ένα κριτήριο μέτρησης/ποσοτικοποίησης τους πόσο καλή είναι η ομαδοποίηση στην οποία καταλήγει κανείς μετά την εφαρμογή μιας (συγκεκριμένης) μεθόδου ταξινόμησης. Γι αυτό εφαρμόζεται συνήθως, όπως αναφέραμε, μαζί με τη διαχωριστική ανάλυση. Με τη βοήθεια της Δ.Α., μπορούμε να διαπιστώσουμε εάν οι ομάδες που δημιουργήσαμε είναι στατιστικά σημαντικές και αν οι μεταβλητές/χαρακτηριστικά διαφοροποιούνται ουσιαστικά ανάμεσα στις (σχηματιζόμενες) ομάδες.

(δ) Κατά την εφαρμογή της διαδικασίας υπάρχουν αρκετά σημεία που θα πρέπει να αποφασίσουμε υποκειμενικά (π.χ. αριθμός ομάδων, επιλογή απόστασης, επιλογή κριτηρίου σύνδεσης κλπ) με αποτέλεσμα για τα ίδια δεδομένα να έχουμε διαφορετικές λύσεις. Η ύπαρξη όμως διαφορετικών λύσεων είναι ίσως ένδειξη ότι δεν υπάρχουν (σημαντικά διαφορετικές) ομοιογενείς ομάδες στα δεδομένα μας.

Παρατήρηση 1.6.1

Αν και όπως αναφέραμε δεν υπάρχουν προϋποθέσεις εφαρμογής των μεθόδων ομαδοποίησης, εν τούτοις θα πρέπει να λαμβάνουμε υπόψη τα παρακάτω:

(α) Το δείγμα/παρατηρήσεις πρέπει να είναι αντιπροσωπευτικό του πληθυσμού που μελετάμε.

(β) Να μην υπάρχουν ακραίες τιμές.

(γ) Η πολυσυγγραμμικότητα μεταξύ των μεταβλητών (συσχέτισή τους) να είναι η μικρότερη δυνατή.

(δ) Ο αριθμός n των παρατηρήσεων να είναι μεγαλύτερος (αρκετά) από τον αριθμό των μεταβλητών p .

1.7. Προβλήματα εφαρμογής Ανάλυσης Συστάδων

Πριν προχωρήσουμε στην περιγραφή των μεθόδων ανάλυσης συστάδων, θα αναφερθούμε σε κάποια από τα προβλήματα που αφορούν όλες τις μεθόδους. Οποιαδήποτε μέθοδο και αν χρησιμοποιήσουμε υπάρχουν σημεία που, όπως αναφέραμε και παραπάνω, θα χρειαστεί ίσως να λειτουργήσουμε υποκειμενικά με αποτέλεσμα από τα ίδια δεδομένα να εξαχθούν διαφορετικά αποτελέσματα. Εάν υπάρχουν στα δεδομένα ομοιογενείς ομάδες περιμένουμε οποιαδήποτε μέθοδος και αν επιλεχθεί να τις εντοπίσει. Επομένως, οι διαφορετικές λύσεις μας προδίδουν ότι ίσως δεν υπάρχει κάποια ομοιογένεια ανάμεσα στις ομάδες.

Μερικά προβλήματα που πρέπει να απαντηθούν σχετικά με την ανάλυση είναι:

(1) Ποιες μεταβλητές πρέπει να χρησιμοποιηθούν;

Το πρώτο κρίσιμο και βασικό βήμα στην ανάλυση συστάδων είναι το ποιες μεταβλητές θα χρησιμοποιήσουμε. Είδαμε παραπάνω ότι π.χ. στην περίπτωση των ποσοτικών μεταβλητών, κάθε αντικείμενο θεωρείται σαν ένα διάνυσμα στο χώρο των μεταβλητών. Το ποιές και πόσες μεταβλητές θα χρησιμοποιήσουμε επηρεάζει την αναπαράσταση του αντικειμένου στο χώρο αυτό και επομένως και τη μετέπειτα ομαδοποίηση. Δεν υπάρχει κάποια συγκεκριμένη

(αυστηρή μαθηματικά) θεωρία που να μας οδηγεί στην κατάλληλη επιλογή των μεταβλητών πριν την ανάλυση. Οι μεταβλητές πρέπει να εξετάζονται διεξοδικά, μιάς και οι ομάδες που δημιουργούνται είναι (ίσως) συνάρτηση των μεταβλητών που περιλαμβάνονται στην ανάλυση. Πριν από την Α.Σ. συνήθως χρησιμοποιούμε τη λεγόμενη **Ανάλυση Κύριων Συνιστωσών**, με τη βοήθεια της οποίας απαλλασσόμαστε από τις συσχετισμένες μεταβλητές (περιττή πληροφορία / περιορισμός προβλήματος πολυσυγγραμμικότητας). Επιλέγουμε κατόπιν τις μεταβλητές που δημιουργούν ή πιστεύουμε ότι δημιουργούν σημαντικά διαφοροποιημένες ομάδες. Μετά την ανάλυση θα είμαστε σε θέση να αποφασίσουμε ποιες μεταβλητές μας είναι λιγότερο χρήσιμες και ποιες όχι. Αυτές που οι τιμές τους είναι (σχεδόν) ίδιες για όλες τις ομάδες μπορούν να διαγραφούν. Ύστερα επαναλαμβάνουμε την ανάλυση, και εάν το αποτέλεσμα δεν αλλάξει σημαίνει ότι όντως δεν τις χρειαζόμασταν. Τέλος, η επιλογή των μεταβλητών εξαρτάται από παράγοντες όπως (α) ο σκοπός της ανάλυσής μας (β) τα ίδια τα δεδομένα (πηγή τους, είδος) κ.λ.π..

Παρατήρηση 1.7.1

(α) Η απαλοιφή μεταβλητών συνεπάγεται απώλεια πληροφορίας.

(β) Η χρήση πολλών μεταβλητών είναι μερικές φορές απαγορευτική από το μέγεθος του δείγματος (περιπτώσεων) που χρησιμοποιούμε. Μειώνεται η ταχύτητα επεξεργασίας των δεδομένων, χρειάζεται μεγαλύτερος χώρος αποθήκευσης τους και αυξάνεται το υπολογιστικό κόστος εφαρμογής της μεθόδου

(ΙΙ) Ποια απόσταση/ομοιότητα θα χρησιμοποιήσουμε;

Ο τύπος των δεδομένων, τα ίδια τα δεδομένα αλλά και η μέθοδος που θα χρησιμοποιηθεί είναι καθοριστικής σημασίας στη σωστή επιλογή της απόστασης. Ακόμα χρήσιμο είναι να ξέρουμε το σκοπό της ανάλυσης αλλά και κάποια επιμέρους χαρακτηριστικά. Επομένως η επιλογή της απόστασης είναι ίσως υπόθεση περίπλοκη.

Παρατήρηση 1.7.2

(α) Η ανάλυση συστάδων είναι μία μέθοδος που βασίζεται, για την εφαρμογή της, στους ηλεκτρονικούς υπολογιστές και ιδιαίτερα στα στατιστικά πακέτα που υπάρχουν σε αυτούς. Άρα πρέπει να ξέρουμε εκ των προτέρων αν αυτή η απόσταση που θέλουμε είναι διαθέσιμη από το εκάστοτε στατιστικό πακέτο.

(β) Όταν χρησιμοποιούμε μια απόσταση, η μονάδα μέτρησης των μεταβλητών παίζει σημαντικό ρόλο. Αλλαγή της μονάδας σημαίνει διαφορετική απόσταση, άρα και διαφορετικές ίσως ομάδες. Αν μια από τις μεταβλητές έχει αρκετά μεγάλο εύρος (διαφορά ανάμεσα στη μεγαλύτερη και μικρότερη τιμή), τότε επηρεάζει (κυριαρχεί) σημαντικά τον υπολογισμό της απόστασης (π.χ. ύψος ατόμου σε μέτρα και βάρος σε κιλά), επομένως και την ανάλυση. Γι αυτό συνήθως χρησιμοποιούμε «κανονικοποιημένα» δεδομένα για κάθε μεταβλητή. Η κανονικοποίηση (μετασχηματισμός) των δεδομένων όμως μερικές φορές δημιουργεί και αυτή προβλήματα (μικραίνει τις αποστάσεις ανάμεσα στις ομάδες). Εάν μια μεταβλητή είναι βασική στο διαχωρισμό των ομάδων (διαχωρίζει δηλαδή «καλά» τις ομάδες), τότε θα έχει μεγάλη διασπορά τιμών και θα πρέπει να διατηρηθεί ως έχει χωρίς να τυποποιηθεί. Η διασπορά αυτή θα πρέπει να ληφθεί υπόψη. Τυχόν κανονοικοποίησή της θα έχει σαν αποτέλεσμα τον όχι εμφανή ορισμό ομάδων. Κάτι που θα μπορούσαμε να κάνουμε εδώ είναι, να γίνει η ανάλυση με και χωρίς κανονικοποιήση και να δούμε εάν υπάρχουν σημαντικές διαφορές στο σχηματισμό των ομάδων. Αν ναι, να μην κανονικοποιηθούν οι μεταβλητές με μεγάλες διασπορές.

(III) Πόσες ομάδες θα δημιουργηθούν;

Ο σκοπός της ανάλυσης συστάδων είναι, όπως αναφέραμε, να δημιουργήσει (όσο το δυνατόν πιο) ομοιογενείς ομάδες δεδομένων. Εάν ερώτημα που προκύπτει κατά την ανάλυση είναι, πόσες ομάδες θα πρέπει να δημιουργηθούν; Στις ιεραρχικές μεθόδους, οποιονδήποτε αριθμός μεταξύ του 1 και του αριθμού *n* των αρχικών στοιχείων είναι ένας αριθμός δυνατών ομάδων που θα μπορούσαν να δημιουργηθούν. Δεν είναι φανερό όμως ποιος από αυτούς τους αριθμούς των ομάδων θα βοηθούσε καλύτερα τους σκοπούς της ανάλυσης των δεδομένων μας. Ακόμα, σε ορισμένες μεθόδους πρέπει να γνωρίζουμε τον αριθμό των ομάδων από την αρχή. Θα δούμε αργότερα με ποιους τρόπους/κριτήρια μπορεί να προσδιοριστεί ο αριθμός αυτός ο οποίος εξαρτάται από τη μορφή των δεδομένων. Αυτό που συνήθως ακολουθεί κανείς, είναι να χρησιμοποιήσει μια ιεραρχική μέθοδο για να έχει εικόνα για τον αριθμό των ομάδων που υπάρχουν (κοιτάζοντας τις αποστάσεις μεταξύ των ομάδων στα διάφορα στάδια σχηματισμού τους // μεγάλες αποστάσεις σημαίνουν τη δημιουργία καλά διαχωρισμένων ομάδων) και κατόπιν γνωρίζοντας τον αριθμό των ομάδων να χρησιμοποιήσει τη μέθοδο κ means για την κατασκευή των ομάδων.

(IV) Ποια μέθοδο θα χρησιμοποιήσουμε;

Οι ιεραρχικές μέθοδοι δεν είναι η καλύτερη δυνατή επιλογή σε μεγάλο πλήθος δεδομένων διότι χρειάζονται πολύ χρόνο και υπολογιστική ισχύ. Επίσης έχουν την τάση να δημιουργούν ομάδες με ανομοιογενή μεγέθη. Αυτά τα προβλήματα γίνονται όλο και πιο έντονα σήμερα αφού χρησιμοποιείτε μεγάλο πλήθος δεδομένων στις σύγχρονες εφαρμογές και οι υπολογιστικές ανάγκες είναι μεγάλες.

Η μέθοδος K-means από την άλλη δεν έχει τα παραπάνω προβλήματα αλλά εξαρτάται πολύ από τις αρχικές τιμές/κέντρα που θα χρησιμοποιήσουμε.

Λαμβάνοντας σοβαρά υπόψη μας όλα αυτά θα δούμε παρακάτω κάποιες από τις μεθόδους που χρησιμοποιούνται περισσότερο.

Κεφάλαιο 2: Μέτρα απόστασης και ομοιότητας

2.1 Εισαγωγή

Είδαμε παραπάνω ότι η ανάλυση συστάδων έχει σα σκοπό τη δημιουργία ομάδων με τη μέγιστη δυνατή ομοιογένεια των στοιχείων εντός των ομάδων (ελάχιστη δυνατή απόσταση μεταξύ των στοιχείων κάθε ομάδας) και τη μέγιστη ταυτόχρονα ανομοιογένεια μεταξύ ομάδων (μέγιστη δυνατή απόσταση μεταξύ των ομάδων).

Ένας τρόπος καθορισμού του βαθμού ομοιότητας/ανομοιότητας δύο παρατηρήσεων είναι με τη χρήση της απόστασης τους και καθορίζεται ποσοτικά με δείκτες που ονομάζονται **συντελεστές ομοιότητας** (ή ανομοιότητας) ή **αποστάσεις** (συνήθως η απόσταση είναι αντίστροφη της ομοιότητας: παρατηρήσεις με μεγάλη απόσταση παρουσιάζουν μικρή ομοιότητα και αντίστροφα) . Υπάρχουν αρκετά είδη συντελεστών ομοιότητας/αποστάσεων που μας δίνουν τη δυνατότητα να επιλέξουμε κάθε φορά τον/την κατάλληλο/η για την ανάλυσή μας. Ποιος είναι ο πιο κατάλληλος εξαρτάται από τον τύπο των δεδομένων που εξετάζουμε και από το είδος της ομοιότητας που ενδιαφερόμαστε. Η επιλογή του μέτρου απόστασης/ομοιότητας είναι βήμα κρίσιμο στην ανάλυση συστάδων, αφού θα ορίσει το βαθμό ομοιότητας δύο στοιχείων και επομένως θα επηρεάσει το σχηματισμό των ομάδων.

Για τον υπολογισμό της απόστασης υπάρχει διαφοροποίηση ανάλογα με το εάν τα γνωρίσματα περιέχουν αριθμητικές, δυαδικές, ή ονομαστικές τιμές.

2.2. Η απόσταση

Η ομοιότητα ανάμεσα σε αντικείμενα συχνά εκφράζεται σαν απόσταση (μέτρο απόστασης). Η έννοια της **απόστασης** είναι πολύ βασική στην πολυμεταβλητή στατιστική. Γενικά θα αναφερόμαστε στον όρο απόσταση, όταν υπάρχει μια **μετρική**, δηλαδή μία συνάρτηση που (μετράει πόσο απέχουν δυο παρατηρήσεις μεταξύ τους και) ικανοποιεί τις παρακάτω προϋποθέσεις:

1. $d(x,y) = d(y,x) \geq 0$ (συμμετρική ιδιότητα)
2. $d(x,y) \leq d(x,z) + d(y,z)$ (τριγωνική ιδιότητα)
3. $d(x,y) \neq 0 \Leftrightarrow x \neq y$
4. $d(x,x) = 0$

Η πιο σημαντική από τις παραπάνω ιδιότητες είναι η τριγωνική. Η τιμή $d(x,y)$ καλείται **απόσταση** των x, y.

(Α) Ποσοτικά δεδομένα

Εάν όλες οι μεταβλητές ως προς τις οποίες εξετάζονται οι παρατηρήσεις είναι ποσοτικές, τότε κάθε μια από τις παρατηρήσεις μπορεί να θεωρηθεί σαν ένα σημείο στον *p*-διάστατο χώρο των *p* μεταβλητών ($x=\left(x_1, x_2, ..., x_p\right)$). Δύο σημεία, τα οποία βρίσκονται κοντά στον χώρο αυτό, θεωρούνται όμοια, ενώ δύο σημεία, τα οποία βρίσκονται μακριά στον χώρο, θεωρούνται ανόμοια. Επομένως, η ομοιότητα τους υπολογίζεται από την απόσταση τους σε αυτό το χώρο.

Παράδειγμα 2.2.1

Ας υποθέσουμε ότι έχουμε τέσσερις μαθητές και ξέρουμε γι' αυτούς το βάρος και την ηλικία τους. Πιο αναλυτικά έχουμε:

Μαθητής	Ηλικία	Βάρος
Α	14	56
Β	10	41
Γ	17	100
Δ	8	25

Θέλουμε να μετρήσουμε την απόσταση ανάμεσα στο μαθητή Α και στον μαθητή Γ. Για να το πετύχουμε αυτό θα χρησιμοποιήσουμε τα διάφορα μέτρα απόστασης που περιγράφονται στη συνέχεια.

(Ι) Ευκλείδεια απόσταση

Η **ευκλείδεια απόσταση** αποτελεί ίσως την πιο γνωστή απόσταση ανάμεσα σε ποσοτικά δεδομένα. Εάν τα χαρακτηριστικά: $x=\left(x_1, x_2, ..., x_p\right)$, $y=\left(y_1, y_2, ..., y_p\right)$, τότε η **Ευκλείδεια απόσταση** τους ορίζεται από τη σχέση:

$$d(x,y)=\sqrt{\sum_{i=1}^{p}(x_i-y_i)^2}$$

όπου *d* η απόσταση και *p* ο αριθμός των μεταβλητών

Παρατήρηση 2.2.2

(α) Η ευκλείδεια απόσταση χρησιμοποιείτε πολύ συχνά, ειδικά σε περιπτώσεις λίγων διαστάσεων και τα αποτελέσματα της είναι πολύ καλά όταν υπάρχουν συμπαγείς ή απομεμακρυσμένες ομάδες.

(β) Η απόσταση μεταξύ δύο οποιωνδήποτε παρατηρήσεων δεν επηρεάζεται από την ύπαρξη παρατηρήσεων με μεγάλες αποστάσεις (ακραίες τιμές).

(γ) Η απόσταση μεταξύ δυό στοιχείων δεν επηρεάζεται από την προσθήκη νέων (επιπλέον) δεδομένων.

(δ) Η απόσταση μεταξύ δυό σημείων επηρεάζεται από τη διαφορετική κλίμακα μέτρησης των μεταβλητών (μεταβλητές με μεγάλη κλίμακα μέτρησης «κυριαρχούν» στο μέγεθος της απόστασης). και μπορεί να οδηγήσει σε σημαντικά διαφορετικές ομάδες. Ακόμα μεταβλητές με μεγάλη διασπορά τιμών επηρεάζουν την απόσταση. Γι αυτό, όπως έχουμε αναφέρει και πριν, «κανονικοποιούμε» τα δεδομένα πριν τα ομαδοποιήσουμε.

Παράδειγμα 2.2.1 (συνέχεια)

$$d(\text{Α},\Gamma) = \sqrt{(14-17)^2 + (56-100)^2} = \sqrt{9+1936} = \sqrt{1945} = 44.1$$

Παρατήρηση 2.2.3 («κανονικοποίηση»)

Αξίζει να σημειωθεί ότι η κλίμακα των δύο παρατηρήσεων είναι διαφορετική. Το εύρος των τιμών της ηλικίας είναι 14-17 ενώ το εύρος των τιμών του βάρους είναι 56-100. Κατά συνέπεια την απόσταση θα την καθορίσει σε ένα μεγάλο βαθμό το βάρος, γιατί η διαφορά στην κλίμακα του είναι κατά πολύ μεγαλύτερη από την ηλικία.

Μια λύση στο παραπάνω πρόβλημα είναι να φέρουμε τις παρατηρήσεις σε μία συγκρίσιμη κλίμακα. Αυτό θα γίνει αν διαιρέσουμε κάθε μεταβλητή με την τυπική της απόκλιση. Δηλαδή, αν συμβολίσουμε με s_1^2 την διακύμανση της i μεταβλητής, τότε η ευκλείδεια απόσταση θα γίνει:

$$d(x,y) = \sqrt{\sum_{i=1}^{p}\left(\frac{x_i - y_i}{s_i}\right)^2}$$

όπου s_j η τυπική απόκλιση της j μεταβλητής.

Παράδειγμα 2.2.1 (συνέχεια)

$$d(\text{Α},\Gamma) = \sqrt{\left(\frac{14-17}{1.5}\right)^2 + \left(\frac{56-100}{22}\right)^2} = \sqrt{4+4} = \sqrt{8} = 2.82$$

(II) Τετραγωνική Ευκλείδεια απόσταση:

Η **τετραγωνική Ευκλείδεια απόσταση** (τετράγωνο της ευκλείδειας απόστασης) χρησιμοποιείται όταν επιθυμούμε να προσδώσουμε μεγαλύτερο βάρος σε στοιχεία που σχετικά είναι απομακρυσμένα μεταξύ τους:

$$d(x,y)=\sum_{i=1}^{p}\left(x_i-y_i\right)^2$$

Παράδειγμα 2.2.1 (συνέχεια)

$$d(\Α,\Γ)=\left(14-17\right)^2+\left(56-100\right)^2=9+1936=1945$$

(III) Μετρική City-block (Manhattan)

Η **μετρική city-block** χρησιμοποιεί το άθροισμα των απόλυτων διαφορών των τιμών των μεταβλητών.

$$d\left(x,y\right)=\sum_{i=1}^{n}\ \left|x_i-y_i\ \right|$$

Τα αποτελέσματα που παίρνουμε με αυτή την απόσταση μοιάζουν με αυτά της ευκλείδειας, αλλά επειδή δεν υψώνεται στο τετράγωνο η διαφορά, μειώνεται η επίδραση των ακραίων τιμών.

Παρατήρηση 2.2.4

Ονομάζεται και **Manhattan** καθώς προσομοιάζει την απόσταση μεταξύ δύο σημείων στην περιοχή του Manhattan της Νέας Υόρκης, όπου η απόσταση μεταξύ των σημείων ισοδυναμεί με τον αριθμό των οικοδομικών τετραγώνων που θα πρέπει να διανυθούν στις κατευθύνσεις Βόρεια-Νότια και Ανατολικά-Δυτικά.

Παράδειγμα 2.2.1 (συνέχεια)

$$d(\Α,\Γ)=\left|14-17\right|+\left|56-100\right|=3+44=47$$

(IV) Μετρική του Minkowski

Η **μετρική του Minkowski** για τα χαρακτηριστικά: $x=\left(x_1,x_2,...,x_p\right)$, $y=\left(y_1,y_2,...,y_p\right)$, ορίζεται από τη σχέση:

$$d(x,y)=\left[\sum_{i=1}^{p}\left(\left|x_i-y_i\right|\right)^q\right]^{\frac{1}{q}}$$

Ουσιαστικά η απόσταση Minkowski γενικεύει την ευκλείδεια απόσταση και την απόσταση Manhattan (με $q=2$ και $q=1$ αντίστοιχα). Η τιμή που θα πάρει η παράμετρος q είναι χρήσιμη στο να δώσει βάρος σε κάποιες αποκλίσεις.

(V) Μετρική Chebychev

Η **μετρική Chebychev** ισούται με τη μέγιστη απόλυτη διαφορά των τιμών των χαρακτηριστικών:

$$d(x,y)=\max\left\{\left|x_i-y_i\right|, i=1,...,p\right\}$$

Η παραπάνω απόσταση είναι χρήσιμη αν θέλει κάποιος να εξετάσει αν δυο διαφορετικές παρατηρήσεις διαφέρουν τουλάχιστον ως προς μία μεταβλητή. Ένα μειονέκτημα όμως της συγκεκριμένης απόστασης είναι ότι εξαρτάται από τις διαφορές στη κλίμακα των μεταβλητών. Λόγω του ότι υπάρχει μόνο μία απόκλιση, η μεγαλύτερη, αν οι μεταβλητές έχουν διαφορετική κλίμακα ο τύπος θα αντικατοπτρίζει τη διαφορά στη μεταβλητή με τη μεγαλύτερη κλίμακα.

Παρατήρηση 2.2.5

Οι παραπάνω μετρικές δεν λαμβάνουν υπ' όψιν τυχόν συσχετίσεις ανάμεσα στις μεταβλητές. Αν δυό μεταβλητές είναι πολύ συσχετισμένες η απόσταση τους οφείλεται μόνο στην πρώτη αφού η δεύτερη «ακολουθεί» την πρώτη λόγω της συσχέτισης. Μια απόσταση που προσπαθεί να αντιμετωπίσει το πρόβλημα είναι η **μετρική Mahalanobis** που δίνεται από τον τύπο:

$$d^2(x,y)=(x-y)'S^{-1}(x-y)$$

όπου S ο πίνακας συνδιασπορών των μεταβλητών. Η απόσταση Mahalanobis δεν επηρεάζεται από την κλίμακα μέτρησης τους.

Στην περίπτωση δυό μεταβλητών ο τύπος γίνεται

$$d^2(x,y)=\frac{1}{1-r^2}\left[\frac{(x_1-y_1)^2}{s_1^{\,2}}+\frac{(x_2-y_2)^2}{s_2^{\,2}}-\frac{2r(x_1-y_1)(x_2-y_2)}{s_1 s_2}\right]$$

όπου r ο συντελεστής συσχέτισης των δυό μεταβλητών (Pearson).

Παρατήρηση 2.2.6 (απόσταση μεταβλητών)

Για την ταξινόμηση μεταβλητών έχουμε τα παρακάτω μέτρα απόστασης/ομοιότητας.

(α) Συντελεστής συσχέτισης κατά Pearson

Έστω, ότι έχουμε δύο μεταβλητές X και Y και θέλουμε να εκτιμήσουμε το βαθμό ρ της γραμμικής σχέσης που τυχόν υπάρχει ανάμεσα τους. Μιά εκτίμηση του ρ προκύπτει ως εξής: επιλέγουμε τυχαία ένα δείγμα μεγέθους n από τον πληθυσμό για τον οποίο ενδιαφερόμαστε και για κάθε μέλος του δείγματος παίρνουμε μετρήσεις αναφορικά με τις X και Y. Προκύπτει ένας πίνακας της μορφής:

Αύξων αριθμός στοιχείου	**X**	**Y**
1	x_1	y_1
2	x_2	y_2
...	…	…
n	x_n	y_n

Ένας δείκτης του βαθμού της γραμμικής σχέσης μεταξύ των δύο μεταβλητών X και Y (ή του σ.σ. ρ) είναι ο δειγματικός **συντελεστής συσχέτισης κατά Pearson** (σ.σ) που δίνεται από τον τύπο:

$$r=\frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sqrt{\sum_{i=1}^{x}(x-\bar{x})^2}\sqrt{\sum_{i=1}^{x}(y-\bar{y})^2}}=\frac{\sum_{i=1}^{n}x_iy_i-\frac{\sum_{i=1}^{n}x_i\cdot\sum_{i=1}^{n}y_i}{n}}{\sqrt{\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}\cdot\sqrt{\sum_{i=1}^{n}y_i^2-\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}}}$$

Εάν χρησιμοποιήσουμε την ποσότητα:

$$Cov(X,Y)=E(X-EX)(Y-EY)=\frac{\sum_{i=1}^{n}x_iy_i}{n}-\frac{\sum_{i=1}^{n}x_i\cdot\sum_{i=1}^{n}y_i}{n^2},$$

η οποία ως γνωστόν καλείται **συνδιασπορά** των μεταβλητών X και Y, ο σ.σ. μπορεί ισοδύναμα να γραφεί:

$$\rho=\frac{Cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$$

Πρόταση 2.2.7
Ο συντελεστής συσχέτισης r παίρνει τιμές από –1 έως 1 ($-1\le r\le 1$). Εάν:

- r =1 λέμε ότι έχουμε **τέλεια θετική συσχέτιση,**
- r = -1 έχουμε **τέλεια αρνητική συσχέτιση**

- εάν r = 0 δεν υπάρχει σχέση μεταξύ των δύο μεταβλητών (σχέση **ανύπαρκτη**).

Παρατήρηση 2.2.8
Σαν απόσταση των X και Y θεωρούμε την ποσότητα: $d(X,Y)=1-r$

(β) Συντελεστής συσχέτισης κατά Spearman

Σε μερικά προβλήματα οι τιμές μιάς τουλάχιστον των μεταβλητών X,Y δεν δίνονται σε αριθμητικές τιμές αλλά σε **τάξη μεγέθους (rank)**. Στην περίπτωση αυτή σα δείκτη του βαθμού της σχέσης ανάμεσα στις μεταβλητές, χρησιμοποιούμε το λεγόμενο **συντελεστή συσχέτισης κατά Spearman**, ο οποίος ορίζεται από τον τύπο:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n^3 - n}$$

όπου: n είναι το πλήθος των ζευγαριών των δεδομένων και d_i είναι η διαφορά των τάξεων των δυό μεταβλητών για το i μέλος του δείγματος

Παράδειγμα 2.2.9
Έξι τουριστικά κέντρα (θέρετρα) βαθμολογήθηκαν από άνδρες και γυναίκες. Η σειρά κατάταξής τους ανάλογα με το φύλο φαίνεται στον παρακάτω πίνακα

Θέρετρο	Άνδρες	Γυναίκες
Α	5	2
Β	1	3
Γ	3	1
Δ	2	4
Ε	4	5
Ζ	6	6

Για να υπολογίσουμε τον βαθμό της σχέσης ανάμεσα στην βαθμολογία των ανδρών και των γυναικών, χρησιμοποιούμε τον σ.σ. κατά Spearman.

Θέρετρο	Άνδρες (X)	Γυναίκες (Y)	d_i	d_i^2
Α	5	2	5-2=3	9
Β	1	3	-2	4
Γ	3	1	2	4
Δ	2	4	-2	4
Ε	4	5	1	1
Ζ	6	6	0	0
				22

Άρα, ο σ.σ. κατά Spearman είναι ίσος με

$$r_s = 1 - \frac{6\sum_{i=1}^{6} d_i^{\,2}}{n^3 - n} = 1 - \frac{6 \cdot 22}{6^3 - 6} = 0{,}37$$

(ο σ.σ. κατά Pearson είναι ίσος με r =0,42)

(Β) Διχοτομικές μεταβλητές / Δυαδικά δεδομένα

Ας υποθέσουμε ότι τα δεδομένα που μας ενδιαφέρουν είναι **διχοτομικές μεταβλητές** (δηλαδή για την κάθε παρατήρηση έχουμε μια δυαδική προσέγγιση, ο αριθμός 1 θα υποδηλώνει την παρουσία αυτού του χαρακτηριστικού/μεταβλητής και το 0 την απουσία του). Για παράδειγμα στη ιατρική επιστήμη το 1 ή 0 μπορεί να δηλώνει την παρουσία ή μη κάποιον συμπτωμάτων μιας συγκεκριμένης ασθένειας. Για τέτοιου είδους δεδομένα, ο κάθε ερευνητής έχει στη διάθεση του πληθώρα αποστάσεων που μπορεί να χρησιμοποιήσει.

Για κάθε δυό παρατηρήσεις/αντικείμενα x και y κατασκευάζουμε έναν **πίνακα ομοιότητας** s(x,y) ή ανομοιότητας *d*(x,y) έτσι ώστε να υπολογίσουμε την απόσταση τους

			Παρατήρηση y		
			1	0	
Παρατήρηση	x	1	a	b	a+b
		0	c	d	c+d
			a+c	b+d	

όπου τα a, b, c, d:

a : το πλήθος των μεταβλητών που το αντικείμενο *x* έχει την τιμή 1 και το *y* την τιμή 1,

b : το πλήθος των μεταβλητών που το αντικείμενο *x* έχει την τιμή 1 και το *y* την τιμή 0,

c : το πλήθος των μεταβλητών που το αντικείμενο *x* έχει την τιμή 0 και το *y* την τιμή 1,

d: το πλήθος των μεταβλητών που το αντικείμενο *x* έχει την τιμή 0 και το *y* την τιμή 0.

Οι μεταβλητές μπορούν να χωριστούν σε δύο κατηγορίες: τις **συμμετρικές** και τις **ασύμμετρες**. Στις συμμετρικές η ύπαρξη ή μη ενός χαρακτηριστικού έχει την ίδια αξία/ σημασία. Παραδείγματος χάρη, αν η μεταβλητή που μας ενδιαφέρει είναι το είδος του επαγγέλματος ενός ατόμου και πιο συγκριμένα αν ασχολείται με κάποιο χειρωνακτικό

επάγγελμα (1 αν ΝΑΙ και 0 αν ΟΧΙ). Από την άλλη στις ασύμμετρες μεταβλητές η ύπαρξη ή μη ενός χαρακτηριστικού δεν έχει την ίδια σημασία/αξία. Για παράδειγμα, στην ιατρική η απουσία κάποιον συμπωμάτων δεν είναι το ίδιο σημαντική με την παρουσία αυτών έτσι ώστε να βγει ένα ιατρικό πόρισμα. Επομένως στην κατηγορία των συμμετρικών μεταβλητών όλα τα κελιά έχουν την ίδια βαρύτητα ενώ στην κατηγορία των ασύμμετρων περισσότερη βαρύτητα έχει το κελί (1,1) δηλαδή η κοινή παρουσία των χαρακτηριστικών.

(i) Συμμετρικές μεταβλητές

Στον παρακάτω πίνακα φαίνονται οι αποστάσεις (ομοιότητες) για συμμετρικά δεδομένα.

Ονομασία	**s(x,y)**	**d(x,y)**
απλός συντελεστής αντιστοίχισης (Sokal & Michener, 1958)	$\frac{a+d}{a+b+c+d}$	$\frac{b+c}{a+b+c+d}$
Rogers και Tarimoto(1960)	$\frac{a+d}{(a+d)+2(b+c)}$	$\frac{2(b+c)}{(a+d)+2(b+c)}$
Sokal και Sneath (1963)	$\frac{2(a+d)}{2(a+d)+(b+c)}$	$\frac{(b+c)}{2(a+d)+(b+c)}$

Παρατηρούμε ότι ο πρώτος συντελεστής ομοιότητας εστιάζει στα κελιά (0,0) και (1,1) δηλαδή στην κοινή απουσία ή παρουσία ενός συγκεκριμένου χαρακτηριστικού. Οι δύο επόμενοι συντελεστές εστιάζουν στα ίδια κελιά αλλά δίνουν διαφορετική βαρύτητα στα κελιά (0,0) και (1,1) από ότι στα (1,0) και (0,1).

(ii) Ασύμμετρες μεταβλητές

Ονομασία	**s(x,y)**	**d(x,y)**
Jaccard (1908)	$\frac{a}{a+b+c}$	$\frac{b+c}{a+b+c}$
Dice και Sorensen (1948)	$\frac{2a}{2a+b+c}$	$\frac{b+c}{2a+b+c}$
Sokal και Sneath (1963)	$\frac{a}{a+2(b+c)}$	$\frac{2(b+c)}{a+2(b+c)}$

Μπορούμε να παρατηρήσουμε ότι το κελί (0,0), όπου εκεί βρίσκεται το στοιχείο d, δεν εμφανίζεται καθόλου στους παραπάνω συντελεστές. Ο συντελεστής Jaccard είναι ίδιος με τον απλό συντελεστή με τη διαφορά ότι δεν υπάρχει το στοιχείο d. Οι δύο επόμενοι εστιάζουν στις ίδιες ποσότητες αλλά δίνουν διαφορετικό βάρος σε αυτές.

Παράδειγμα 2.2.7

Έστω ότι διαλέγουμε 4 μαθητές από μία τάξη και θέλουμε να διαπιστώσουμε αν επισκέφθηκαν κάποιες συγκεκριμένες σελίδες στο διαδίκτυο ή όχι. Αυτές οι ιστοσελίδες είναι 10 στον αριθμό. Αν ο μαθητής την επισκέφθηκε θα δώσουμε την τιμή 1 και αν όχι τη τιμή 0. Τα δεδομένα φαίνονται στο παρακάτω πίνακα.

		ιστοσελίδες									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	Α	1	1	1	0	0	0	1	0	0	1
Μαθητές	Β	0	1	1	0	1	1	0	0	0	0
	Γ	0	0	0	0	0	1	1	0	0	0
	Δ	0	1	1	0	0	1	0	0	0	0

Σύμφωνα με τον παραπάνω πίνακα ο μαθητής Α επισκέφθηκε τις ιστοσελίδες {1,2,3,7,10} ενώ ο Β μόνο τις {2,3,5,6}. Για να μετρήσουμε την απόσταση ανάμεσα στους δύο μαθητές θα χρησιμοποιήσουμε τις αποστάσεις για συμμετρικά δεδομένα. Άρα ο πίνακας ομοιότητας θα είναι ο εξής:

		A	
		1	0
B	1	2	2
	0	3	3
		5	5

Σε αυτόν τον πίνακα μπορούμε να διακρίνουμε ότι υπάρχον 2 ίδιες ιστοσελίδες όπου και ο Α και ο Β επισκέφθηκαν και 3 ίδιες ιστοσελίδες όπου δεν τις επισκέφθηκε κανείς. Με βάση αυτόν τον πίνακα μπορούμε να υπολογίσουμε τα μέτρα ομοιότητας.

Συμμετρικές αποστάσεις

Δείκτης ομοίοτητας	**τιμή**
απλός συντελεστής αντιστοίχισης (Sokal & Michener, 1958)	$\frac{3+2}{1+2+3+3}=0.5$
Rogers και Tarimoto(1960)	$\frac{3+2}{5+2\cdot 5}=0.33$
Sokal και Sneath (1963)	$\frac{2\cdot 5}{2\cdot 5+5}=0.66$

Ασύμμετρες αποστάσεις

Δείκτης ομοίοτητας	**τιμή**
Jaccard (1908)	$\frac{2}{2+2+3}=0.28$
Dice και Sorensen (1948)	$\frac{2\cdot 2}{2\cdot 2+2+3}=0.44$
Sokal και Sneath (1963)	$\frac{2}{2+2\cdot 5}=0.16$

Σε αυτό το παράδειγμα μπορούμε να χρησιμοποιήσουμε και τις ασύμμετρες αποστάσεις διότι δεν σημαίνει κάτι ιδιαίτερο αν και οι δύο χρήστες δεν έχουν επισκεφθεί μία σελίδα στο διαδίκτυο (επίσκεψη μιάς ιστοσελίδας ή όχι δεν έχει την ίδια αξία). Είναι τόσες πολλές οι σελίδες αυτές που είναι πολύ πιθανό δύο χρήστες να μην έχουν επισκεφθεί την ίδια ιστοσελίδα.

Αν πάλι επιλέξουμε τους μαθητές Β και Γ, θα πάρουμε τον παρακάτω πίνακα:

		B		
		1	0	
Γ	1	1	1	2
	0	3	5	8
		4	6	

Υπολογίζω και πάλι τα μέτρα ομοιότητας τα οποία δίνονται παρακάτω:

Συμμετρικές αποστάσεις

Δείκτης ομοιότητας	**τιμή**
απλός συντελεστής αντιστοίχισης (Sokal & Michener, 1958)	$\frac{1+5}{1+1+3+5}=0.6$
Rogers και Tarimoto(1960)	$\frac{1+5}{6+2\cdot 4}=0.42$
Sokal και Sneath (1963)	$\frac{2\cdot 6}{2\cdot 6+4}=0.75$

Ασύμμετρες αποστάσεις

Δείκτης ομοιότητας	**τιμή**
Jaccard (1908)	$\frac{1}{1+1+3}=0.2$
Dice και Sorensen (1948)	$\frac{2\cdot 1}{2\cdot 1+1+3}=0.33$
Sokal και Sneath (1963)	$\frac{1}{1+2\cdot 4}=0.11$

Παρατήρηση 2.2.8

Σε αυτό το σημείο μπορούμε να παρατηρήσουμε ότι ο απλός συντελεστής ομοιότητας των Α και Β είναι $s_1(\text{A},\text{B})=0.5$ και των Β και Γ είναι $s_1(\text{B},\Gamma)=0.6$, ενώ ο συντελεστής Jaccard των Α και Β είναι $s_2(\text{A},\text{B})=0.28$ και των Β και Γ είναι $s_2(\text{B},\Gamma)=0.2$. Επομένως διαπιστώνουμε ότι με τη χρήση του απλού συντελεστή, οι μαθητές Β και Γ έχουν περισσότερη ομοιότητα από τους Α και Β, ενώ με τη χρήση του συντελεστή Jaccard το συμπέρασμα είναι ακριβώς το αντίθετο. Κοιτώντας προσεκτικά τα αρχικά δεδομένα βλέπουμε ότι οι μαθητές Α και Β έχουν επισκεφθεί περισσότερες όμοιες σελίδες, σε σύγκριση με τους μαθητές Β και Γ. Έτσι καταλήγουμε στην απόφαση ότι καταλληλότερος συντελεστής ομοιότητας είναι ο Jaccard.

Επαναλαμβάνοντας την ίδια διαδικασία και για τους μαθητές Γ, Δ και για τους Α, Δ και για τους Α, Γ και για τους Β, Δ, προκύπτει ο παρακάτω (συμμετρικός) **πίνακας ομοιότητας**:

$$\begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{Γ} \\ \text{Δ} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{B} & \text{Γ} & \text{Δ} \end{array} \\ \begin{pmatrix} 1 & & & \\ 0.28 & 1 & & \\ 0.16 & 0.2 & 1 & \\ 0.33 & 0.75 & 0.25 & 1 \end{pmatrix} \end{array}$$

Παρατήρηση 2.2.9

(α) Ανάλογα με τον πίνακα ομοιότητας, υπολογίζεται και ο πίνακας αποστάσεων των δεδομένων, ο οποίος θα χρησιμοποιηθεί στην ομαδοποίησή τους.

(β) Το πρόβλημα υπολογισμού του πίνακα ομοιότητας/απόστασης στη γενική περίπτωση, μπορεί είναι πιο πολύπλοκο. Δηλαδή να χρειαστεί να συνδυάσουμε διαφορετικούς συντελεστές (συμμετρικούς και ασύμμετρους) για να πάρουμε τον σωστό πίνακα ομοιότητας.

(Γ) Δεδομένα σε ονομαστική κλίμακα

Έστω τώρα ότι κάθε μια από τις μεταβλητές μας παίρνει **ονομαστικές** τιμές (λέξεις) δηλαδή οι μεταβλητές μας είναι **ονομαστικές** (εθνικότητα, θρήσκευμα, χώρα κ.α.), οι οποίες δεν επιδέχονται καμιά μορφή ιεράρχησης. Εάν x, y τα γνωρίσματα, τα οποία εξετάζουμε ως προς τις p-μεταβλητές και έστω ότι οι τιμές τους συμπίπτουν σε m από αυτά τα γνωρίσματα, τότε οι συντελεστές ομοιότητας και απόστασης τους ορίζονται από τις σχέσεις.

$$s(x,y)=\frac{m}{p}, \quad d(x,y)=\frac{p-m}{p}$$

Παρατήρηση 2.2.10

Ένας δεύτερος τρόπος υπολογισμού της απόστασης μεταξύ των x, y είναι με τη βοήθεια διχοτομικών ψευδομεταβλητών. Πιο συγκεκριμένα, για κάθε ονομαστική μεταβλητή με k δυνατές τιμές, εισάγουμε k ψευδομεταβλητές (μια για κάθε τιμή της μεταβλητής). Εάν μια παρατήρηση έχει μια συγκεκριμένη τιμή στη μεταβλητή, τότε η (αντίστοιχη) ψευδομεταβλητή παίρνει την τιμή 1, ενώ οι υπόλοιπες ψευδομεταβλητές παίρνουν την τιμή 0. Π.χ. ένα ενδιαφερόμαστε για το χρώμα των ματιών ενός ατόμου και {καστανό, μαύρο, γαλανό, πράσινο} είναι οι δυνατές τιμές της μεταβλητής, τότε χρησιμοποιούμε 4

ψευδομεταβλητέες μια για κάθε χρώμα ματιών (μια για το καστανό με τιμή 1 εάν το άτομο έχει χρώμα ματιών καστανό και 0 αν όχι κ.ο.κ).

Αφού με αυτόν τον τρόπο οι ονομαστικές μεταβλητές μπορούν να μετατραπούν σε (πολλαπλές) διχοτομικές, η απόστασή τους μπορεί να ορισθεί όπως σε προηγούμενη παράγραφο.

(Δ) Μεταβλητές σε κλίμακα διάταξης/κατάταξης

Έστω ότι μια μεταβλητή του ενδιαφέροντός μας είναι μια **μεταβλητή διάταξης**, παίρνει δηλαδή k δυνατές τιμές οι οποίες μπορούν να διαταχθούν κατά κάποιο τρόπο. Τέτοιου είδους μεταβλητές χρησιμοποιούμε όταν θέλουμε να δημιουργήσουμε ερωτήσεις κλειστού τύπου (π.χ. δε συμφωνώ 0, συμφωνώ λίγο 1, συμφωνώ αρκετά 2, συμφωνώ πολύ 3, συμφωνώ πάρα πολύ 4).

Επειδή υπάρχει μιά σειρά διάταξης, μπορούμε ν' αντιστοιχήσουμε σε κάθε μια από τις τιμές της μεταβλητής μια αριθμητική τιμή, ως εξής: στην τιμή που δηλώνει τη χαμηλότερη θέση στη σειρά διάταξης (σε μια αύξουσα σειρά διάταξης) αντιστοιχούμε την τιμή 1, στην επόμενη θέση αντιστοιχούμε την τιμή 2, κ.ο.κ. στην τελευταία αντιστοιχούμε την τιμή k. Έστω μια μεταβλητή διάταξης μετατρέπεται σε μια μεταβλητή ποσοτική/αριθμητική.

Παρατήρηση 2.2.11

Αν μια μεταβλητή διάταξης παίρνει αρκετές τιμές, τότε οι τιμές αυτές μπορούν να επηρεάσουν τις αποστάσεις μεταξύ των παρατηρήσεων με αποτέλεσμα τη διαφοροποίηση στο σχηματισμό των ομάδων. Για την αντιμετώπιση του προβλήματος καινονικοποιούμε τις τιμές της μεταβλητής (τις ανάγουμε σε τιμές στο διάστημα [0,1]), με τη βοήθεια του τύπου:

$$m_{new} = \frac{m-1}{n-1}$$

όπου m_{new} η νέα τιμή της μεταβλητής, m η τιμή της μεταβλητής πριν την κανονικοποίηση και n το πλήθος των τιμών της μεταβλητής.

Παράδειγμα 2.2.12

Έστω ότι ενδιαφερόμαστε για τη βαθμίδα ενός μέλους Δ.Ε.Π. Τότε η μεταβλητή παίρνει τις τιμές: {λέκτορας, επίκουρος καθηγητής, αναπληρωτής καθηγητής, καθηγητής} στις οποίες μπορούμε να αντιστοιχήσουμε τις (αριθμητικές) τιμές 1,2,3,4. Οι κανονικοποιημένες τιμές φαίνονται στη δεξιά στήλη του παρακάτω πίνακα.

Διατακτικές τιμές	Αριθμητικές τιμές	Κανονικοποιημένες τιμές
Λέκτορας	**1**	$\frac{1-1}{4-1}=0$
Επίκουρος καθηγητής	**2**	$\frac{2-1}{4-1}=0,33$
Αναπληρωτής καθηγητής	**3**	$\frac{3-1}{4-1}=0,66$
Καθηγητής	**4**	$\frac{4-1}{4-1}=1$

Αφού κάθε μεταβλητή διάταξης μπορεί να μετασχηματιστεί σε μια (ποσοτική) αριθμητική με τιμές στο διάστημα [0,1], ο υπολογισμός της απόστασης δύο χαρακτηριστικών μπορεί να γίνει με τη βοήθεια κάποιας απόστασης, ποσοτικών μεταβλητών, που περιγράφηκαν παραπάνω.

(Ε) Μεταβλητές διάφορων τύπων

Η μέχρι τώρα ανάλυση αφορούσε περιπτώσεις όπου όλες οι μεταβλητές ήταν του ίδιου τύπου δηλαδή της ίδιας κατηγορίας. Όμως κάτι τέτοιο δεν ανταποκρίνεται στην πραγματικότητα αφού οι μεταβλητές που μπορεί να έχουμε σε ένα πρόβλημα μπορεί να είναι συνεχείς όπως η ηλικία, δυαδικής φύσης όπως το φύλο ή κατηγορικής μορφής όπως είναι η πόλη. Επομένως αναφερόμαστε σε δεδομένα μεικτού τύπου.

Για να διαχειριστούμε μεικτού τύπου δεδομένα μπορούμε να χρησιμοποιήσουμε τριών ειδών τεχνικές.

Η πρώτη μέθοδος κατατάσσει αρχικά τις μεταβλητές σε ομοειδής ομάδες. Για κάθε ομάδα μεταβλητών εφαρμόζουμε μια ανάλυση συστάδων, ελπίζοντας ότι οι ομάδες που θα προκύψουν (σε κάθε ομαδοποίηση) να μοιάζουν μεταξύ τους. Συνήθως δεν προκύπτουν ανάλογες ομάδες.

Η δεύτερη τεχνική δημιουργεί ψευδομεταβλητές για κάθε τύπου πληροφορία που υπάρχει, έτσι ώστε να προκύψουν διχοτομικές/δυαδικές μεταβλητές. Όμως στην κατηγορία των ποσοτικών/συνεχών μεταβλητών για παράδειγμα θα πρέπει οι μεταβλητές να γίνουν ποιοτικές, δηλαδή να ορίσουμε μικρά διαστήματα στα οποία παίρνουν τιμές (με αντίστοιχη

συχνότητα) και μετά να ορίσουμε και εδώ ψευδομεταβλητές. Κατά τη διάρκεια του μετασχηματισμού όμως ενδέχεται να θα χαθεί σημαντική πληροφορία.

Η τρίτη τεχνική αναπτύχθηκε από τον Gower (1971) και είναι αυτή που χρησιμοποιείτε περισσότερο. Δυστυχώς όμως αυτή δεν υπάρχει στα περισσότερα στατιστικά πακέτα με αποτέλεσμα ο ερευνητής να πρέπει να καταφεύγει σε πολύπλοκους υπολογισμούς έτσι ώστε να την χρησιμοποιήσει.

Η υλοποίηση αυτή στηρίζεται στον υπολογισμό μίας απόστασης για μεικτού τύπου δεδομένα. Ο τύπος Gower της ομοιότητας δυό παρατηρήσεων x και y είναι ο παρακάτω:

$$s(x,y) = \frac{\sum_{i=1}^{p} w_i(x,y)s_i(x,y)}{\sum_{i=1}^{p} w_i(x,y)}$$

$s_i(x,y)$: είναι ο συντελεστής ομοιότητας μεταξύ των παρατηρήσεων x και y για τη μεταβλητή *i*.

$w_i(x,y)$: είναι τα βάρη που παίρνουν την τιμή 1 ή 0.

Για την ποσότητα $s_i(x,y)$:

- Αν η μεταβλητή είναι κατηγορικής κλίμακας, συμμετρική ή και όχι, τότε η τιμή που θα πάρει το $s_i(x,y)$ θα είναι 1 αν οι παρατηρήσεις ταυτίζονται και 0 αν όχι.
- Αν υπάρχει συνεχής μεταβλητή η τιμή που θα πάρει το $s_i(x,y)$ θα είναι $1-\frac{|x_i-y_i|}{R_i}$, όπου R_i συμβολίζει το εύρος της τιμής της μεταβλητής.

Τα βάρη $w_i(x,y)$ παίρνουν την τιμή 1 ή 0:

- 0 αν το x_i ή το y_i απουσιάζει
- 0 αν η μεταβλητή είναι μη-συμμετρική και $x_i = y_i = 0$
- 1 σε οποιαδήποτε άλλη περίπτωση.

Παράδειγμα 2.2.13

Θα εμπλουτίσουμε το παραπάνω παράδειγμα βάζοντας κάποια δημογραφικά στοιχεία.

		Ιστοσελίδα											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Ηλικία**	**Πόλη**
	Α	1	1	1	0	0	0	1	0	0	1	18	Ηράκλειο
Μαθητές	Β	0	1	1	0	1	1	0	0	0	0	21	Ηράκλειο
	Γ	0	0	0	0	0	1	1	0	0	0	23	Χανιά
	Δ	0	1	1	0	0	1	0	0	0	0	41	Ρέθυμνο

Από ότι βλέπουμε έχουν ορισθεί επιπλέον δύο στήλες στον πίνακα, μία που αναφέρεται στην ηλικία του κάθε μαθητή και ακόμα μία που μας πληροφορεί για την πόλη από την οποία κατάγεται. Η μεταβλητή ηλικία είναι συνεχής, η πόλη ονομαστική ενώ κάθε μια από τις (10 συνολικά) μεταβλητές ιστοσελίδα είναι διχοτομική. Χρησιμοποιώντας το δείκτη του Gower για τους μαθητές Α και Β θα πάρουμε τα εξής:

Για τη συνεχή μεταβλητή ηλικία:

$$s_{11}(x,y)=1-\frac{|x_{11}-y_{11}|}{R_{11}}=1-\frac{|21-18|}{41-18}=1-\frac{3}{23}=0.87 \quad \text{και} \quad w_{11}(x,y)=1$$

Για την ονομαστική μεταβλητή πόλη: $s_{12}(x,y)=1$ (διότι οι παρατηρήσεις ταυτίζονται) και

$$w_{12}(x,y)=1$$

Για την $1^{η}$ μεταβλητή ιστοσελίδα: $s_1(x,y)=0$ γιατί $x_1 \neq y_1$ και $w_1(x,y)=1$. Ακόμα, για τη $2^{η}$ μεταβλητή ιστοσελίδα: $s_2(x,y)=1$ γιατί $x_2=y_2$ και $w_2(x,y)=1$. Ανάλογα, ορίζονται οι ποσότητες $s_j(x,y)$ και $w_j(x,y),\ j=3,4,...,10$.

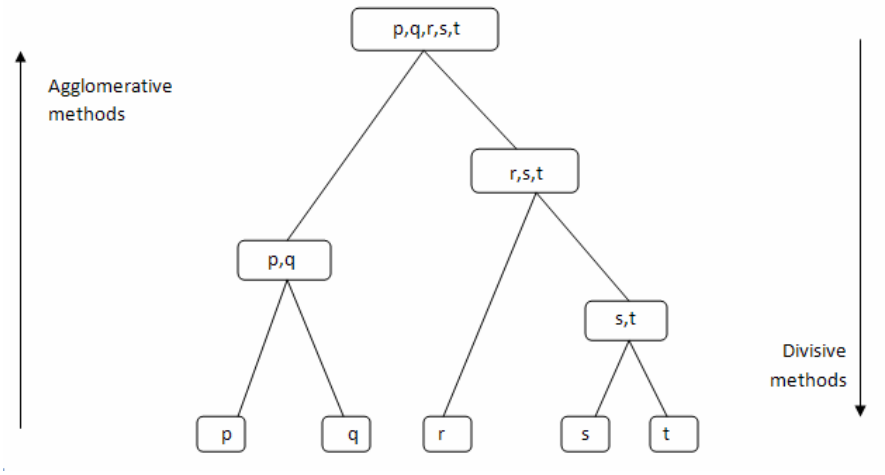
Τελικά, ο συντελεστής ομοιότητας δίνεται από τη σχέση:

$$s(x,y)=\frac{\sum_{i=1}^{12} w_i(x,y)s_i(x,y)}{\sum_{i=1}^{12} w_i(x,y)}=\frac{1\cdot 0+1\cdot 1+1\cdot 1+0\cdot 1+1\cdot 0+1\cdot 0+1\cdot 0+0\cdot 1+0\cdot 1+1\cdot 0+1\cdot 0.87+1\cdot 1}{1+1+1+0+1+1+1+0+0+1+1+1}=0.43$$

Κεφάλαιο 3 Ιεραρχικές μέθοδοι ομαδοποίησης

3.1 Εισαγωγή

Η πρώτη και ίσως πιο δημοφιλής κατηγορία μεθόδων ανάλυσης συστάδων είναι οι λεγόμενες **ιεραρχικές μέθοδοι** (hierarchical methods). Είναι αλγόριθμοι οι οποίοι σε κάθε στάδιο εφαρμογής τους δημιουργούν έναν αριθμό ομάδων (συγκεκριμένο τρόπος διαμοιρασμού των αντικειμένων σε ομάδες/συστάδες). Οι ομάδες που δημιουργούνται στα διάφορα στάδια μπορούν να ιεραρχηθούν (ταξινομηθούν) κατά κάποιο τρόπο. Στο κατώτατο επίπεδο της ιεραρχίας βρίσκονται τα μεμονωμένα αντικείμενα, ενώ στο ανώτατο επίπεδο βρίσκεται μια «υπερσυστάδα» (υπερομάδα) η οποία περιλαμβάνει όλα τα αντικείμενα. Οι ομάδες (ιεραρχία) προκύπτουν είτε από συγχώνευση ομάδων του προηγούμενου επιπέδου εφαρμογής του αλγορίθμου ή από διάσπαση ομάδων του προηγούμενου επιπέδου. Στην $1^{η}$ περίπτωση οι μέθοδοι ονομάζονται **συσσωρευτικές** (agglomerative) ενώ στην $2^{η}$ **διαιρετικές** (divisive).



Εικόνα 10: ιεραρχικές μέθοδοι ταξινόμησης

Στις **συσσωρευτικές** μεθόδους λοιπόν αρχικά θεωρούμε κάθε αντικείμενο σαν μια ομάδα. Κατόπιν, τα πιο όμοια αντικείμενα επιλέγονται και συγχωνεύονται, δημιουργώντας μια νέα ομάδα. Συνεχίζοντας, από τις ομάδες που προκύπτουν, επιλέγονται οι πιο όμοιες και συγχωνεύονται. Η διαδικασία επαναλαμβάνεται μέχρι να ενταχθούν όλα τα αντικείμενα σε μια ενιαία ομάδα.

Στις **διαιρετικές** μεθόδους αρχικά θεωρούμε όλα τα αντικείμενα σαν μέλη μιας ενιαίας ομάδας. Κατόπιν, η αρχική αυτή ομάδα διαιρείται σε δύο υποομάδες κατά τέτοιο τρόπο ώστε οι υποομάδες οι οποίες θα προκύψουν να έχουν τη μεγαλύτερη ανομοιότητα. Η διαδικασία των διαδοχικών διασπάσεων επαναλαμβάνεται μέχρι κάθε αντικείμενο να αποτελεί μια

ξεχωριστή υποομάδα. Εάν στα δεδομένα μας υπάρχουν *n* αντικείμενα/σημεία, τότε και στις δύο κατηγορίες μεθόδων υπάρχουν *n*-1 επίπεδα.

Το ποιο είναι το πιο κατάλληλο σύνολο των ομάδων από αυτά που δημιουργούνται, το οποίο περιγράφει έναν φυσικό τρόπο διαμοιρασμού των αντικειμένων, εναπόκειται στον ερευνητή.

Τα βασικά **πλεονεκτήματα** των ιεραρχικών μεθόδων είναι τα εξής:

- Είναι σχετικά εύκολες στην εφαρμογή τους
- Μπορούν να εντοπίσουν καλά διαχωρισμένες, επιμήκεις και ομόκεντρες ομάδες.
- Δημιουργούνται πολλά (ιεραρχικά δομημένα) σύνολα ομάδων που επιτρέπουν στον ερευνητή να επιλέξει το (πιο κατάλληλο) επίπεδο που αυτός επιθυμεί.

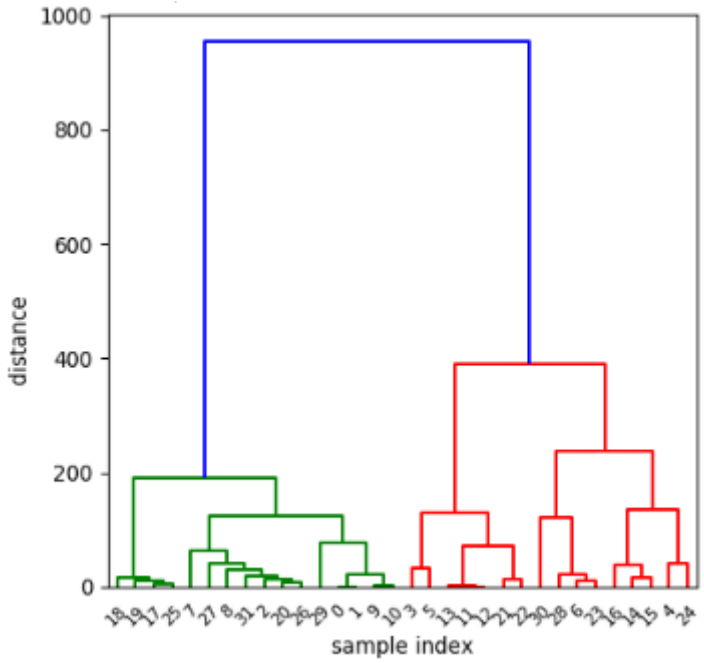
Μειονεκτήματα των ιεραρχικών μεθόδων είναι τα εξής:

- Κάθε εφαρμογή της μεθόδου (επίπεδο) είναι διαδικασία μη αντιστρεπτή. Από τη στιγμή που δύο αντικείμενα ενταχθούν στην ίδια ομάδα, παραμένουν στην (ίδια) ομάδα, και δεν υπάρχει δυνατότητα να διαχωριστούν αργότερα και να ενταχθούν σε διαφορετικές ομάδες.
- Για την εφαρμογή των μεθόδων, χρειάζεται σε κάθε επίπεδο υπολογισμός της απόστασης/ ομοιότητας μεταξύ των ομάδων του προηγούμενου επιπέδου. Η εργασία αυτή απαιτεί χρόνο και κόστος επεξεργασίας από τους Η.Υ. Γι αυτό οι ιεραρχικές μέθοδοι δεν θεωρούνται κατάλληλες μέθοδοι ταξινόμησης για μεγάλα σύνολά δεδομένων. Οι ιεραρχικές μέθοδοι χρειάζεται να ελέγξουν πολλές αποστάσεις, και για το λόγο αυτό καθυστερούν όταν χρειάζεται να επεξεργαστούν μεγάλο αριθμό αντικειμένων. Το υπολογιστικό κόστος/πολυπλοκότητα αλγορίθμου είναι τουλάχιστον της τάξης $O(N^2)$ *σε μνήμη και* $O(N^3)$ *σε χρόνο,* όπου *N* το πλήθος των αντικειμένων.
- Τα αρχικά δεδομένα έχουν «ισχυρή» επίδραση στο τελικό αποτέλεσμα.
- Οι μέθοδοι είναι ευαίσθητες στην ύπαρξη ακραίων τιμών.

3.2. Δενδρόγραμμα

Ένας διαγραμματικός τρόπος αναπαράστασης μιας ιεραρχικής μεθόδου ταξινόμησης είναι το λεγόμενο **δενδρόγραμμα (dendrogramm)**. Το εν λόγω διάγραμμα έχει τη μορφή ενός ανεστραμμένου δένδρου. Τα φύλλα του δένδρου (το κατώτερο επίπεδο), είναι τα μεμονωμένα

αντικείμενα (συνήθως άξονας των x). Από κάθε αντικείμενο/ομάδα ξεκινά ένας κατακόρυφος κλάδος. Κάθε δύο τέτοιοι κλάδοι ενώνονται σε έναν κόμβο (ένωση αντικειμένων σε ομάδα).



Εικόνα 11: δενδρόγραμμα

https://python-graph-gallery.com/400-basic-dendrogram/

Στη συσσωρευτική ομαδοποίηση, ένας κόμβος με τους κλάδους συμβολίζει τη συγχώνευση δύο ομάδων του προηγούμενο επιπέδου. Στη διαιρετική ομαδοποίηση, ένας κόμβος με τους κλάδους συμβολίζει τη διάσπαση μιάς ομάδας σε δύο. Στις συσωρευτικές μεθόδους ο βαθμός ανομοιότητας/απόστασης αυξάνεται μονότονα με το επίπεδο εφαρμογής της μεθόδου. Ο σχεδιασμός του δένδρου γίνεται με τέτοιον τρόπο, ώστε η διαφορά ύψους των επιπέδων να αποτυπώνει την αύξηση της ανομοιότητας/απόστασης. Στον άξονα λοιπόν των y συνήθως υπάρχουν οι αποστάσεις (συνένωσης) των ομάδων.

Χρησιμοποιούμε το δενδρόγραμμα για να επιλέξουμε ένα επίπεδο εφαρμογής της μεθόδου.

3.3. Ιεραρχική συσσωρευτική μέθοδος συστάδων

Έστω λοιπόν ότι θέλουμε να ταξινομήσουμε/ομαδοποιήσουμε μια συλλογή από αντικείμενα με τη βοήθεια ορισμένων χαρακτηριστικών (τα οποία έχουμε επιλέξει αρχικά). Έστω ακόμα ότι έχουμε αποφασίσει::

(α) τη μέθοδο ταξινόμησης που θα ακολουθήσουμε και έστω ότι αυτή είναι μιά ιεραρχική συσσωρευτική μέθοδος.

(β) τη μέθοδο απόστασης που θα χρησιμοποιήσουμε (κάποια από αυτές που περιγράψαμε στο προηγούμενο κεφάλαιο, για να υπολογίσουμε την ομοιότητα μεταξύ των αντικειμένων). Με τη βοήθεια της απόστασης κατασκευάζουμε τον ακόλουθο πίνακα αποστάσεων (proximity matrix) μεταξύ των αντικειμένων:

$$\begin{array}{c} \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{array}\begin{array}{c} \begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array} \\ \begin{pmatrix} d(x_1,x_1) & & & \\ d(x_2,x_1) & d(x_2,x_2) & & \\ \vdots & \vdots & \vdots & \\ d(x_n,x_1) & d(x_n,x_2) & \cdots & d(x_n,x_n) \end{pmatrix} \end{array} = \begin{pmatrix} 0 & & & \\ d(x_2,x_1) & 0 & & \\ \vdots & \vdots & \vdots & \\ d(x_n,x_1) & d(x_n,x_2) & \cdots & 0 \end{pmatrix}$$

Τα δύο πρώτα βήματα που ακολουθεί κανείς, σε οποιαδήποτε μέθοδο ιεραρχικής συσσωρευτικής ταξινόμησης, είναι τα εξής:

Βήμα 1ο: θεωρούμε καθένα από τα' αντικείμενα σαν μια ξεχωριστή ομάδα, και υπολογίζουμε τον πίνακα αποστάσεων μεταξύ τους, όπως περιγράφηκε παραπάνω.

Βήμα 2ο: Βρίσκουμε τη μικρότερη δυνατή απόσταση (μεταξύ των αποστάσεων των αντικειμένων) και συνενώνουμε σε μια (νέα) αμάδα τα αντικείμενα με τη μικρότερη (αυτή) δυνατή απόσταση.

Στη συνέχεια, θα πρέπει να ορίσουμε τον τρόπο μέτρησης της απόστασης μεταξύ δύο ομάδων (ομάδες θα θεωρούνται από το σημείο αυτό και τα μεμονωμένα αντικείμενα). Θα πρέπει να ορίσουμε ένα **κριτήριο σύνδεσης** (linkage criteria) δυό ομάδων. Το κριτήριο αυτό καθορίζει την (αν)ομοιότητα δυό ομάδων, υπολογίζοντας την απόσταση ανάμεσα στις ομάδες αντικειμένων σαν συνάρτηση κάθε φορά των αποστάσεων μεταξύ των αντικειμένων των ομάδων. Υπάρχουν περισσότερα από ένα τέτοια κριτήρια, που περιγράφονται παρακάτω, με αποτέλεσμα τον παραπέρα διαχωρισμό των μεθόδων ιεραρχικής ταξινόμησης

Αφού έχει ορισθεί και το κριτήριο σύνδεσης δύο ομάδων, τα επόμενα βήματα που ακολουθούμε στη συσσωρευτική ταξινόμηση είναι

Βήμα 3ο: Υπολογίζουμε ξανά τον πίνακα απόστασης, ανάμεσα στις ομάδες που έχουν προκύψει (από την συνένωση αντικειμένων) με τη βοήθεια του κριτηρίου σύνδεσης που χρησιμοποιούμε (για να καθορίσουμε την απόσταση ανάμεσα σε

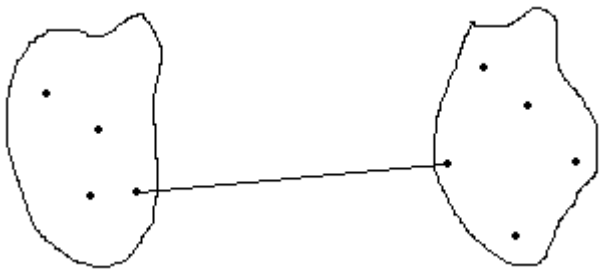
δύο ομάδες). Βρίσκουμε και πάλι τη μικρότερη απόσταση του νέου πίνακα (απόστασης) και ενώνουμε τις δύο ομάδες που αντιστοιχούν σε αυτή την απόσταση σε μία.

Βήμα 4^{ο}: εάν όλα τα αντικείμενα ανήκουν σε μια ομάδα, η διαδικασία σταματά εδώ. Αν όχι επαναλαμβάνουμε τα βήματα 3 και 4.

3.4 Κριτήρια σύνδεσης

(A) Απλή σύνδεση ή πλησιέστερου γείτονα (single linkage / nearest neighbor)

Η ομαδοποίηση με τη χρήση της **απλής σύνδεσης** γνωστή και ως τεχνική **πλησιέστερου ή κοντινότερου γείτονα (**Florek (1951), Sneath (1957), (1967)) έχει σα χαρακτηριστικό της: *η απόσταση μεταξύ των ομάδων ορίζεται να είναι αυτή του πλησιέστερου ζευγαριού αντικειμένων (ένα μέρος του ζευγαριού από κάθε ομάδα)*. Δηλαδή, η απόσταση μεταξύ δύο ομάδων είναι η απόσταση των δύο κοντινότερων αντικειμένων τους.



Εικόνα 12: μέθοδος απλής σύνδεσης

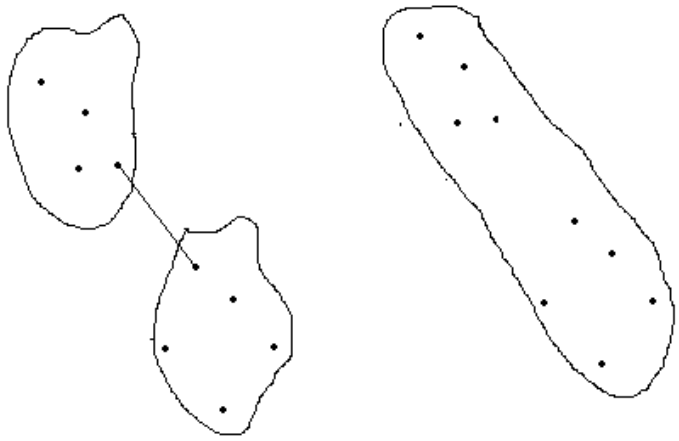
Ο τύπος που δίνει την απόσταση, μεταξύ δύο ομάδων, είναι:

$$d\left(O_1, O_2\right) = \min_{x_\alpha \in O_1, x_\beta \in O_2} d\left(x_\alpha, x_\beta\right)$$

όπου O_1, O_2 οι ομάδες και x_α, x_β αντίστοιχα στοιχεία τους.

Πλεονέκτημα της μεθόδου είναι ότι μπορεί να δημιουργήσει και μη ελλειπτικές ομάδες. Μειονέκτημα είναι ότι, λόγω του ότι πρέπει να υπολογίσουμε όλες τις αποστάσεις μεταξύ των στοιχείων, έχει μεγάλο υπολογιστικό κόστος. Ένα ακόμα μειονέκτημα είναι (μερικές φορές), το λεγόμενο **φαινόμενο της αλυσίδας** (chaining phenomenon): η μέθοδος συνενώνει ομάδες με δύο κοντινά σημεία και (αρκετά) σημεία που βρίσκονται σε μεγάλες αποστάσεις.

Συνεχίζοντας τη διαδικασία μπορεί να συμβεί να προστίθενται διαρκώς νέα σημεία στην «ουρά» της ομάδας, με αποτέλεσμα της δημιουργία μιάς επιμήκους ομάδας (αλυσίδα).



Εικόνα 13: φαινόμενο αλυσίδας

Για να ξεκινήσουμε την εφαρμογή της μεθόδου, κατασκευάζουμε τον πίνακα των αποστάσεων $D=\left(d_{ij}\right)_{i,j=1}^{N}$ των αντικειμένων, και στη συνέχεια βρίσκουμε τη μικρότερη απόσταση. Στη συνέχεια, θα συγχωνεύσουμε τις δύο παρατηρήσεις/αντικείμενα (στα οποία αντιστοιχεί η μικρότερη απόσταση) σε μία. Έστω λοιπόν ότι μιλάμε για τα αντικείμενα Α και Β, τότε συγχωνεύονται και θα γίνουν η ομάδα (ΑΒ). Με τη βοήθεια του κριτηρίου του *κοντινότερου γείτονα*, και χρησιμοποιώντας τον τύπο $d_{(AB)C}=\min(d_{AB},d_{AC})$ θα υπολογίσουμε την απόσταση της συστάδας (ΑΒ) από τα υπόλοιπα αντικείμενα κ.ο.κ., και θα σχηματίσουμε ένα νέο πίνακα αποστάσεων μεταξύ των ομάδων. Ενώνουμε τις δύο ομάδες με τη μικρότερη απόσταση. Συνεχίζουμε ανάλογα τη διαδικασία.

Παράδειγμα 3.4.1

Το παράδειγμα αφορά πέντε χώρες και δείκτες τους σε γεννήσεις, θανάτους και βρεφική θνησιμότητα.

Χώρα	**Γεννήσεις/1000 κατοίκους**	**Θάνατοι/1000 κατοίκους**	**Θάνατοι βρεφών/ 1000 γεννήσεις**
Γαλλία	24.7	5.7	30.8
Αγγλία	12.5	11.9	14.4
Γερμανία	11.6	13.4	14.8
Ισπανία	14.3	10.2	16
Πολωνία	13.6	10.7	26.9

Θα χρησιμοποιήσουμε τις τρεις μεταβλητές και την <u>ευκλείδεια απόσταση</u> για να πάρουμε τον πίνακα τις απόστασης. Για λόγους ευκολίας θα ορίσουμε: Γαλλία=1, Αγγλία=2, Γερμανία=3, Ισπανία=4, Πολωνία=5. Ο πίνακας αποστάσεων των αντικειμένων (χωρών) είναι ίσος με:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 21.35 & 0 & & & \\ 22.06 & 1.79 & 0 & & \\ 18.64 & 2.94 & 4.35 & 0 & \\ 12.78 & 12.60 & 12.55 & 10.93 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ζευγαριών είναι της Αγγλίας με τη Γερμανία δηλαδή:

$$\min_{i,j}(d_{i,j}) = d_{32} = 1,79$$

Άρα τα αντικείμενα 3 και 2 θα σχηματίσουν μια ομάδα την οποία συμβολίζουμε με (23).

Στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της ομάδας (23) από τις υπόλοιπες χώρες δηλαδή τις 1, 4, 5. Η απόσταση αυτή υπολογίζεται με το κριτήριο του κοντινότερου γείτονα:

$$d_{(23)1} = \min\{d_{31,}d_{21}\} = \min\{22.06, 21.35\} = 21.35$$

$$d_{(23)4} = \min\{d_{34,}d_{24}\} = \min\{4.35, 2.94\} = 2.94$$

$$d_{(23)5} = \min\{d_{35,}d_{25}\} = \min\{12.55, 12.60\} = 12.55$$

Κατασκευάζουμε τώρα έναν νέο πίνακα αποστάσεων (θα διαγράψουμε από τον προηγούμενο πίνακα τις γραμμές 3, 2 και τις στήλες 3, 2 και θα προσθέσουμε μια γραμμή και μια στήλη για την ομάδα (23)). Ο νέος πίνακας αποστάσεων είναι ίσος με:

$$\begin{array}{c} \\ 1 \\ (23) \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & (23) & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & \\ 21.35 & 0 & & \\ 18.64 & 2.94 & 0 & \\ 12.78 & 12.55 & 10.93 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ομάδων/ζευγαριών είναι αυτή ανάμεσα στην (Αγγλία, Γερμανία) με την Ισπανία δηλαδή $\min_{i,j}(d_{ij}) = d_{(23)4} = 2.94$. Άρα οι ομάδες (23) και 4 θα σχηματίσουν τη νέα ομάδα (234).

Συνεχίζοντας, στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της ομάδας (234) από τις υπόλοιπες δηλαδή τις 1, 5. Υπολογίζονται ανάλογα, από τους τύπους:

$$d_{(234)1} = \min\{d_{21}, d_{31}, d_{41}\} = \min\{21.35, 22.06, 18.64\} = 18.64$$

$$d_{(234)5} = \min\{d_{25}, d_{35}, d_{45}\} = \min\{12.60, 12.55, 10.93\} = 10.93$$

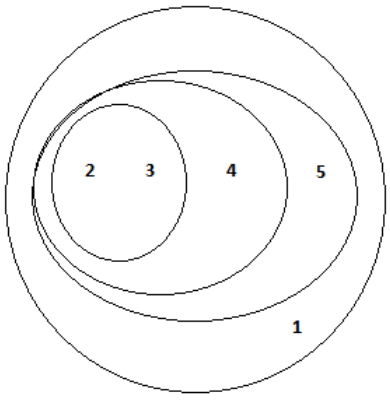
Τώρα, θα διαγράψουμε στον πίνακα αποστάσεων τις γραμμές (23), 4 και τις στήλες (23), 4 και θα προσθέσουμε μια γραμμή και μια στήλη για την ομάδα (234).

$$\begin{array}{c} \\ 1 \\ (234) \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & (234) & 5 \end{array} \\ \begin{pmatrix} 0 & & \\ 18.64 & 0 & \\ 12.78 & 10.93 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή της ομάδας (Αγγλία, Γερμανία, Ισπανία) με την Πολωνία δηλαδή $\min_{i,j}(d_{ij}) = d_{(234)5} = 10.93$ Άρα τα αντικείμενα (234) και 5 θα σχηματίσουν την ομάδα (2345).

Στο τελευταίο επίπεδο ομαδοποίησης, η ομάδα (2345) ενώνεται με το αντικείμενο 1 και σχηματίζουν μια ομάδα. Η απόσταση μεταξύ τους είναι:

$$d_{(2345)1} = \min\{d_{21}, d_{31}, d_{41}, d_{51}\} = \min\{21.35, 22.06, 18.64, 12.78\} = 12.78$$

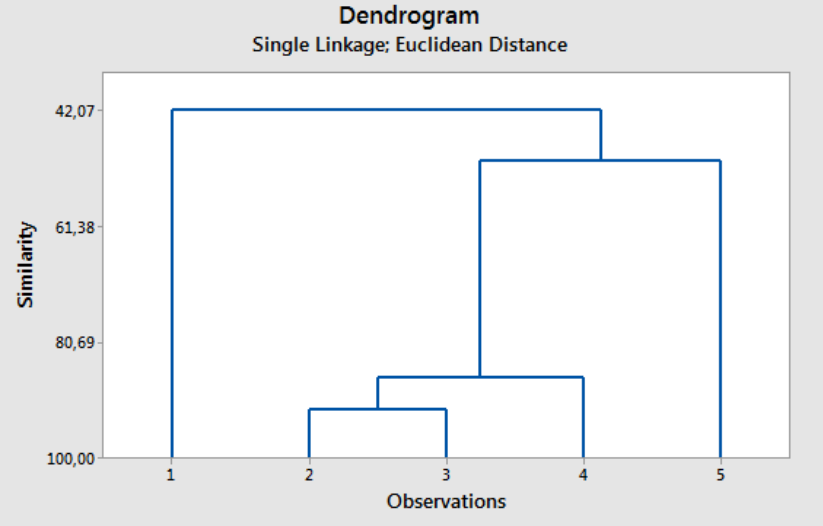


Εικόνα 14: ταξινόμηση με το κριτήριο κοντινότερου γείτονα

Οι διαμερίσεις/ ομάδες που παράγονται σε κάθε επίπεδο εφαρμογής της μεθόδου είναι οι εξής:

Επίπεδο	Ομάδες
E5	[1], [2], [3], [4], [5]
E4	[2 3], [1], [4], [5]
E3	[2 3 4], [1], [5]
E2	[1], [2 3 4 5]
E1	[1 2 3 4 5]

Το δενδρόγραμμα που απεικονίζει τη διαδικασία σε κάθε επίπεδο φαίνεται στο παρακάτω σχήμα. Το ύψος στο διάγραμμα αντιπροσωπεύει την απόσταση στην οποία γίνεται κάθε συνένωση.



Εικόνα 15: δενδρόγραμμα (μέθοδος απλής σύνδεσης)

Παρατήρηση 3.4.2

Το στατιστικό πρόγραμμα SPSS, δίνει τα παρακάτω αποτελέσματα/πίνακες:

(α) τον πίνακα των αποστάσεων:

Proximity Matrix

Case	Euclidean Distance				
	1:Γαλλία	2:Αγγλία	3:Γερμανία	4:Ισπανία	5:Πολωνία
1:Γαλλία	,000	21,360	22,066	18,640	12,784
2:Αγγλία	21,360	,000	1,794	2,948	12,606
3:Γερμανία	22,066	1,794	,000	4,355	12,558
4:Ισπανία	18,640	2,948	4,355	,000	10,934
5:Πολωνία	12,784	12,606	12,558	10,934	,000

This is a dissimilarity matrix

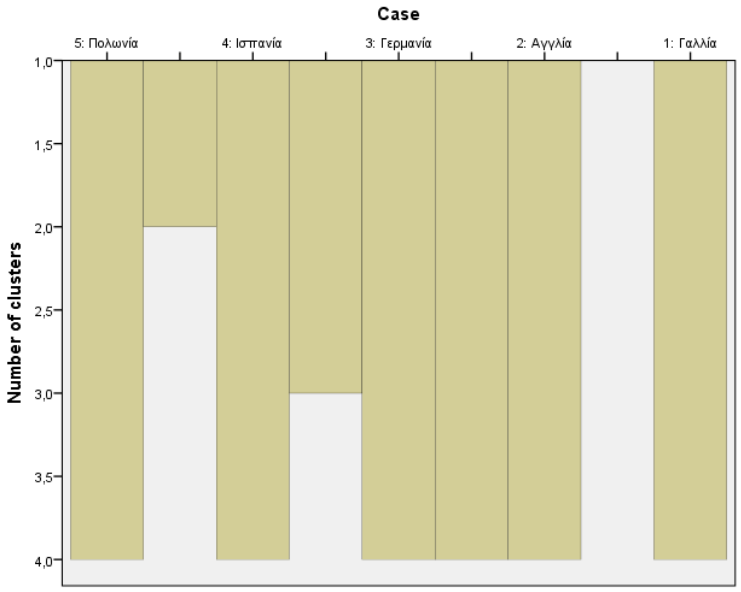
(β) πίνακα που περιγράφει τα στάδια της διαδικασίας

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	3	1,794	0	0	2
2	2	4	2,948	1	0	3
3	2	5	10,934	2	0	4
4	1	2	12,784	0	3	0

Στη $2^{η}$ και $3^{η}$ στήλη φαίνεται ποια αντικείμενα ενώνονται και στην $4^{η}$ η απόσταση.

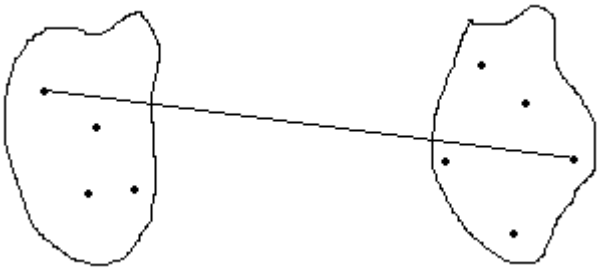
(γ) Αναπαράσταση της διαδικασίας με τη βοήθεια ενός γραφήματος που λέγεται **icicle**



Εικόνα 16: διάγραμμα icicle (μέθοδος απλής σύνδεσης)

(Β) Ομαδοποίηση με τη χρήση της πλήρους σύνδεσης ή απώτερου γείτονα (complete linkage / furthest neighbor**)**

Η ομαδοποίηση με τη χρήση της **πλήρης σύνδεσης ή του απώτερου ή μακρινότερου γείτονα** είναι το αντίθετο της απλής σύνδεσης, με την έννοια ότι η απόσταση μεταξύ των ομάδων ορίζεται τώρα ως *η απόσταση του πιο απομακρυσμένου ζευγαριού αντικειμένων, ένα από κάθε ομάδα*. Δηλαδή, η απόσταση μεταξύ δύο ομάδων είναι η απόσταση των δυό μακρινότερων αντικειμένων τους



Εικόνα 17: σύνδεση απώτερου γείτονα

Ο τύπος που δίνει την απόσταση, ανάμεσα σε δύο ομάδς, είναι :

$$d\left(O_1, O_2\right) = \max_{x_\alpha \in O_1, x_\beta \in O_2} d\left(x_\alpha, x_\beta\right)$$

όπου O_1, O_2 οι ομάδες και x_α, x_β αντίστοιχα στοιχεία τους.

Η μέθοδος εφαρμόζεται όταν γνωρίζουμε ότι αντικείμενα της ίδιας ομάδας πιθανόν να βρίσκονται σε μεγάλες αποστάσεις μεταξύ τους. Δημιουργεί συμπαγείς και σφαιρικές συστάδες με συγκρίσιμη διάμετρο. Το κριτήριο της μεθόδου δεν είναι τοπικό. Ολόκληρη η δομή της εκάστοτε ομάδας θα επηρεάσει την απόφαση για τη συνένωση.

Πλεονέκτημα της μεθόδου είναι ότι αποφεύγονται προβλήματα, όπως εκείνο του φαινομένου της αλυσίδας, που παρουσιάζονται στην απλή σύνδεση, Μειονέκτημα της μεθόδου είναι η ευαισθησία της στην ύπαρξη αντικειμένων με ακραίες τιμές. Εάν σε μια ομάδα υπάρχει ένα αντικείμενο με ακραίες τιμές, τότε δύσκολα θα συγχωνευθεί με κάποιαν άλλη.

Για να ξεκινήσουμε την εφαρμογή της μεθόδου κατασκευάζουμε τον πίνακα των αποστάσεων $D = \{d_{ij}\}$ (με απόσταση την οποία έχουμε προκαθορίσει) και από τον πίνακα βρίσκουμε τη μικρότερη απόσταση, και έστω ότι αντιστοιχεί στα αντικείμενα Α και Β. Συγχωνεύουμε τα αντικείμενα αυτά σε μιά ομάδα, έστω την (ΑΒ). Με τη βοήθεια του κριτηρίου του *μακρύτερου γείτονα*, και χρησιμοποιώντας τον τύπο $d_{(AB)c} = \max\{d_{AB}, d_{AC}\}$ θα υπολογίσουμε τις αποστάσεις (ΑΒ) από όλα τα υπόλοιπα αντικείμενα και θα σχηματίσουμε ένα νέο πίνακα αποστάσεων μεταξύ των ομάδων. Ενώνουμε τις δύο ομάδες με τη μικρότερη απόσταση. Συνεχίζουμε με τον ίδιο τρόπο.

Παράδειγμα 3.4.1 (συνέχεια)

Χώρα	Γεννήσεις/1000 κατοίκους	Θάνατοι/1000 κατοίκους	Θάνατοι βρεφών/ 1000 γεννήσεις
Γαλλία	24.7	5.7	30.8
Αγγλία	12.5	11.9	14.4
Γερμανία	11.6	13.4	14.8
Ισπανία	14.3	10.2	16
Πολωνία	13.6	10.7	26.9

Θα χρησιμοποιήσουμε και πάλι τις τρεις μεταβλητές και την ευκλείδεια απόσταση για να πάρουμε τον πίνακα τις απόστασης. Για λόγους ευκολίας θα ισχύει και πάλι : Γαλλία=1, Αγγλία=2, Γερμανία=3, Ισπανία=4, Πολωνία=5. Ο πίνακας αποστάσεων είναι ίσος με:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 21.35 & 0 & & & \\ 22.06 & 1.79 & 0 & & \\ 18.64 & 2.94 & 4.35 & 0 & \\ 12.78 & 12.60 & 12.55 & 10.93 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ζευγαριών είναι πάλι αυτή μεταξύ Αγγλίας και Γερμανίας, δηλαδή

$$\min_{i,j}(d_{ij}) = d_{32} = 1.79 .$$

Άρα τα αντικείμενα 3 και 2 ενώνονται και θα σχηματίσουν την ομάδα (23).

Στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της ομάδας (23) με τα υπόλοιπα αντικείμενα 1, 4, 5:

$$d_{(23)1} = \max\{d_{21}, d_{31}\} = \max\{21.35, 22.06\} = 22.06$$

$$d_{(23)4} = \max\{d_{24}, d_{34}\} = \max\{2.94, 4.35\} = 4.35$$

$$d_{(23)5} = \max\{d_{25}, d_{35}\} = \max\{12.60, 12.55\} = 12.60$$

Διαγράφουμε, στον προηγούμενο πίνακα αποστάσεων, τις γραμμές 2, 3 και τις στήλες 2, 3 και προσθέτουμε μια γραμμή και μια στήλη για την ομάδα (23), δημιουργώντας έναν νέο πίνακα αποστάσεων:

$$\begin{array}{c} \\ 1 \\ (23) \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & (23) & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & \\ 22.06 & 0 & & \\ 18.64 & 4.35 & 0 & \\ 12.78 & 12.60 & 10.93 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή ανάμεσα στην (Αγγλία, Γερμανία) με την Ισπανία δηλαδή $\min_{i,j}\left(d_{ij}\right) = d_{(23)4} = 4.35$

Άρα τα αντικείμενα (23) και 4 συγχωνεύονται και θα σχηματίσουν την ομάδα (234).

Συνεχίζοντας, στο επόμενο επίπεδο θα χρειαστούμε τις αποστάσεις της ομάδας (234) από τα αντικείμενα 1, 5:

$$d_{(234)1} = \max\{d_{21}, d_{31}, d_{41}\} = \max\{21.35, 22.06, 18.64\} = 22.06$$

$$d_{(234)5} = \max\{d_{25}, d_{35}, d_{45}\} = \max\{12.60, 12.55, 10.93\} = 12.60$$

Διαγράφουμε, στον προηγούμενο πίνακα αποστάσεων, τις γραμμές (23), 4 και τις στήλες (23), 4 και προσθέτουμε μια γραμμή και μια στήλη για τη συστάδα (234)

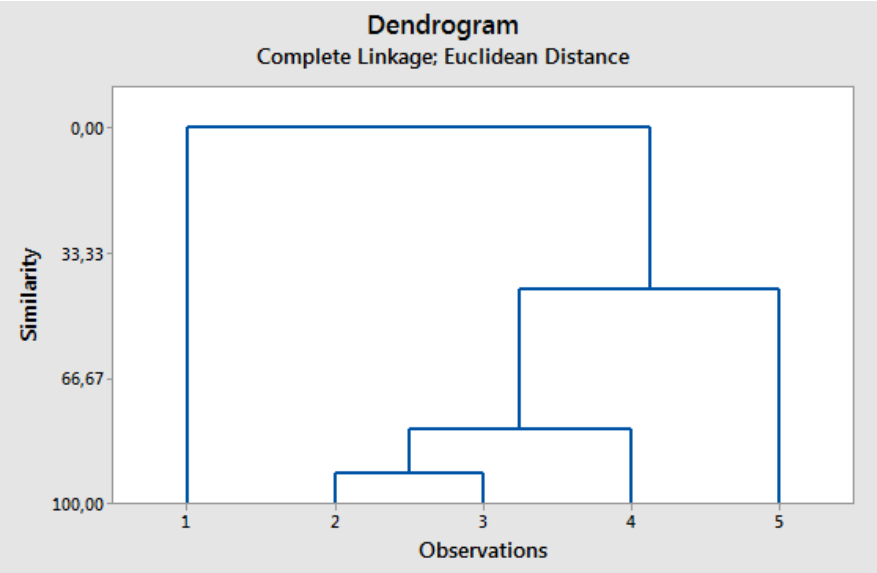
$$\begin{array}{c} \\ 1 \\ (234) \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & (234) & 5 \end{array} \\ \begin{pmatrix} 0 & & \\ 22.06 & 0 & \\ 12.78 & 12.60 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή ανάμεσα στη (Αγγλία, Γερμανία, Ισπανία) με την Πολωνία δηλαδή $\min_{i,j}\left(d_{ij}\right) = d_{(234)5} = 12.60$.

Άρα οι ομάδες (234) και 5 ενώνονται και θα σχηματίσουν την συστάδα (2345).

Στο τελευταίο επίπεδο ομαδοποίησης, η ομάδα (2345) ενώνεται με το αντικείμενο 1 και σχηματίζουν μια ομάδα την (12345). Η απόσταση μεταξύ τους είναι:

$$d_{(2345)1} = \max\{d_{21}, d_{31}, d_{41}, d_{51}\} = \max\{21.35, 22.06, 18.64, 12.78\} = 22.06$$



Εικόνα 18: δενδρόγραμμα (μέθοδος πλήρους σύνδεσης)

Παρατήρηση 3.4.3

Το στατιστικό πρόγραμμα SPSS, δίνει τα παρακάτω αποτελέσματα:

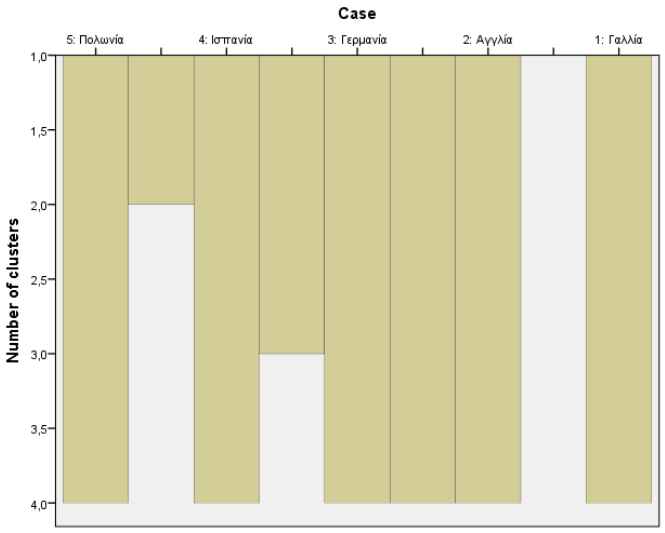
(α) πίνακα που περιγράφει τα στάδια της διαδικασίας

Complete Linkage

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	3	1,794	0	0	2
2	2	4	4,355	1	0	3
3	2	5	12,606	2	0	4
4	1	2	22,066	0	3	0

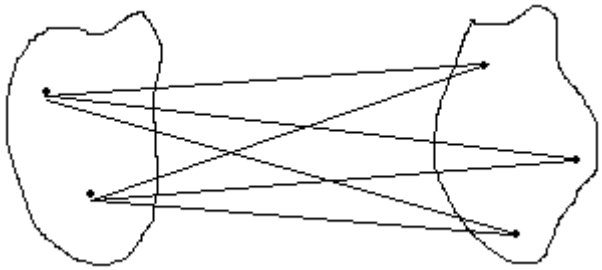
(β) Αναπαράσταση της διαδικασίας με τη βοήθεια ενός γραφήματος που λέγεται **icicle**



Εικόνα 19: διάγραμμα icicle (μέθοδος πλήρους σύνδεσης)

(Γ) Μέση πλήρης σύνδεση ή μη σταθμισμένη κατά ζεύγη σύνδεση (average linkage between groups / unweighted pair group average linkage)

Στην ομαδοποίηση με τη χρήση της **μέσης πλήρης σύνδεσης** (μέθοδος μέσης απόστασης), απόσταση μεταξύ δύο ομάδων υπολογίζεται ως *η μέση απόσταση μεταξύ όλων των ζευγαριών των αντικειμένων των ομάδων, με ένα μέλος του ζευγαριού ανά ομάδα.*



Εικόνα 20: μέση πλήρης σύνδεση

Ο τύπος που δίνει την απόσταση, ανάμεσα σε δύο ομάδες, είναι :

$$d\left(O_1, O_2\right) = \frac{\sum_{x_\alpha \in O_1} \sum_{x_\beta \in O_2} d\left(x_\alpha, x_\beta\right)}{N_{O_1} \cdot N_{O_2}}$$

όπου O_1, O_2 οι ομάδες, x_α, x_β αντίστοιχα στοιχεία τους και N_{O_1}, N_{O_2} ο αριθμός των στοιχείων των ομάδων.

Πλεονέκτημα της μεθόδου είναι ότι δεν δημιουργεί μεγάλες σε μήκος αλυσίδες/ομάδες και ούτε έχουμε τόσο έντονο πρόβλημα με ακραίες τιμές. Μειονέκτημα, είναι ότι λόγω του ότι

έχουμε να υπολογίσουμε μέσες αποστάσεις μεταξύ των ομάδων, έχει μεγαλύτερο υπολογιστικό κόστος.

Για να ξεκινήσουμε την εφαρμογή της μεθόδου της μέσης απόστασης κατασκευάζουμε τον πίνακα των αποστάσεων $D=\{d_{ij}\}$ (με απόσταση την οποία έχουμε προκαθορίσει), από τον πίνακα βρίσκουμε τη μικρότερη απόσταση και έστω ότι αντιστοιχεί στα αντικείμενα Α και Β. Συγχωνεύουμε τα αντικείμενα αυτά σε μιά ομάδα, έστω την (ΑΒ). Με τη βοήθεια του κριτηρίου της *μέσης απόστασης*, και χρησιμοποιώντας τον τύπο $d_{(AB)C}=\frac{\sum_i \sum_j d_{ij}}{N_{(AB)}N_C}$ θα υπολογίσουμε τις αποστάσεις της ομάδας (AB) με την ομάδα/αντικείμενο C (d_{ij}: η απόσταση μεταξύ του αντικειμένου i της (AB) από το j της C, $N_{(AB)}$ και N_C οι αριθμοί των στοιχείων στις ομάδες) και σχηματίζουμε ένα νέο πίνακα αποστάσεων μεταξύ των ομάδων. Ενώνουμε τις δύο ομάδες με τη μικρότερη απόσταση. Συνεχίζουμε με τον ίδιο τρόπο.

Παράδειγμα 3.4.1 (συνέχεια)

Χώρα	**Γεννήσεις/1000 κατοίκους**	**Θάνατοι/1000 κατοίκους**	**Θάνατοι βρεφών/ 1000 γεννήσεις**
Γαλλία	24.7	5.7	30.8
Αγγλία	12.5	11.9	14.4
Γερμανία	11.6	13.4	14.8
Ισπανία	14.3	10.2	16
Πολωνία	13.6	10.7	26.9

Θα χρησιμοποιήσουμε τις τρεις μεταβλητές και την ευκλείδεια απόσταση για να πάρουμε τον πίνακα τις απόστασης ο οποίος είναι ίσος με:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 21.35 & 0 & & & \\ 22.06 & 1.79 & 0 & & \\ 18.64 & 2.94 & 4.35 & 0 & \\ 12.78 & 12.60 & 12.55 & 10.93 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ζευγαριών είναι πάλι αυτή μεταξύ Αγγλίας και Γερμανίας, δηλαδή $\min_{i,j}(d_{ij}) = d_{32} = 1.79$. Άρα τα αντικείμενα 2 και 3 θα σχηματίσουν την ομάδα (23).

Στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της συστάδας (23) με τα υπόλοιπα αντικείμενα 1, 4, 5:

$$d_{(23)1} = \left(d_{21} + d_{31}\right)/2 = 21,71$$

$$d_{(23)4} = \left(d_{24} + d_{34}\right)/2 = 3,65$$

$$d_{(23)5} = (d_{25} + d_{35})/2 = 12.58$$

Διαγράφουμε, στον προηγούμενο πίνακα αποστάσεων, τις γραμμές 2, 3 και τις στήλες 2, 3 και προσθέτουμε μια γραμμή και μια στήλη για τη συστάδα (23), δημιουργώντας έναν νέο πίνακα αποστάσεων:

$$\begin{array}{c} \\ 1 \\ (23) \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{cccc} 1 & (23) & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & \\ 21.71 & 0 & & \\ 18.64 & 3.65 & 0 & \\ 12.78 & 12.58 & 10.93 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή ανάμεσα στην (Αγγλία, Γερμανία) με την Ισπανία δηλαδή $\min_{i,j}\left(d_{ij}\right) = d_{4(23)} = 3.65$. Άρα τα αντικείμενα (23) και 4 θα σχηματίσουν την ομάδα (234).

Συνεχίζοντας, στο επόμενο επίπεδο θα χρειαστούμε τις αποστάσεις της ομάδας (234) από τα αντικείμενα 1, 5:

$$d_{(234)1} = (d_{21} + d_{31} + d_{41})/3 = 20.68$$

$$d_{(234)5} = (d_{25} + d_{35} + d_{45})/3 = 12.03$$

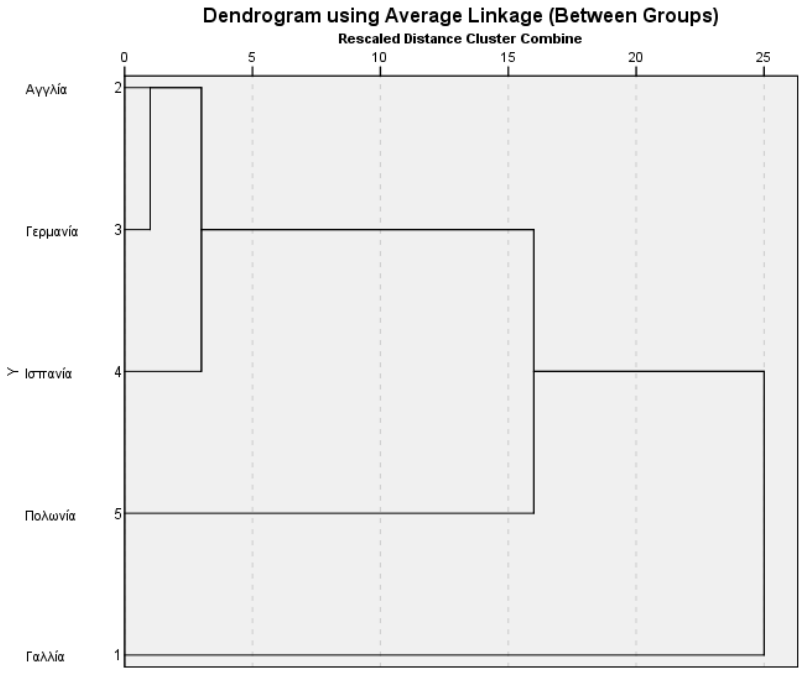
Διαγράφουμε, στον προηγούμενο πίνακα αποστάσεων, τις γραμμές (23), 4 και τις στήλες (23), 4 και προσθέτουμε μια γραμμή και μια στήλη για την ομάδα (234)

$$\begin{array}{c} \\ 1 \\ (234) \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & (234) & 5 \end{array} \\ \begin{pmatrix} 0 & & \\ 20.68 & 0 & \\ 12.78 & 12.03 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή ανάμεσα στη (Αγγλία, Γερμανία, Ισπανία) με την Πολωνία δηλαδή $\min_{i,j}\left(d_{ij}\right) = d_{(234)5} = 12.03$. Άρα οι ομάδες (234) και 5 ενώνονται και θα σχηματίσουν την συστάδα (2345).

Στο τελευταίο επίπεδο ομαδοποίησης, η ομάδα (2345) ενώνεται με το αντικείμενο 1 και σχηματίζουν μια ομάδα την (12345). Η απόσταση μεταξύ τους είναι:

$$d_{(2345)1} = (d_{21} + d_{31} + d_{41} + d_{51}) / 4 = 18.71$$



Εικόνα 21: δενδρόγραμμα (μέση πλήρης σύνδεση)

Παρατήρηση 3.4.4

Το στατιστικό πρόγραμμα SPSS, δίνει τα παρακάτω αποτελέσματα:

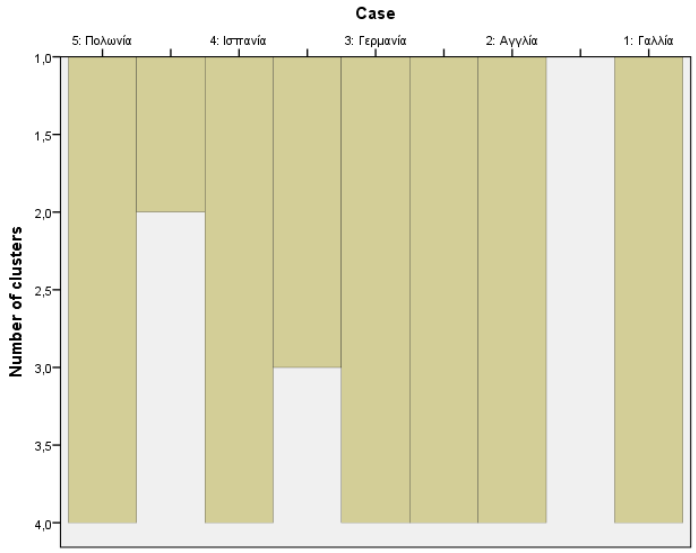
(α) πίνακα που περιγράφει τα στάδια της διαδικασίας

Average Linkage (Between Groups)

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	3	1,794	0	0	2
2	2	4	3,652	1	0	3
3	2	5	12,032	2	0	4
4	1	2	18,712	0	3	0

(β) Αναπαράσταση της διαδικασίας με τη βοήθεια ενός γραφήματος που λέγεται **icicle**



Εικόνα 22: διάγραμμα icicle (μέση πλήρης σύνδεση)

Παρατήρηση 3.4.5

Μια παραλλαγή της παραπάνω μεθόδου είναι και η λεγόμενη **σταθμισμένη κατά ζεύγη μέση σύνδεση** (weighted pair group average linkage). Η απόσταση μεταξύ δύο ομάδων, στη μέθοδο αυτή, υπολογίζεται όπως και προηγουμένως με τη διαφορά ότι χρησιμοποιούμε κάθε φορά σα συντελεστή στάθμισης (της μέσης απόστασης) το μέγεθος της αντίστοιχης ομάδας (αριθμός στοιχείων της ομάδας). Η μέθοδος χρησιμοποιείται συνήθως όταν υπάρχει σημαντική διαφορά στα μεγέθη των ομάδων (ιδιαίτερα άνισα).

(Δ) Μέθοδος σύνδεσης μέσης απόστασης μέσα στις ομάδες (average linkage within groups)

Στη συγκεκριμένη μέθοδο, ομάδες ενώνονται κατά τέτοι τρόπο ώστε η μέση απόσταση, ανάμεσα στα στοιχεία της ομάδας που θα προκύψει, να είναι η ελάχιστη δυνατή.
Η απόσταση μεταξύ δύο ομάδων ορίζεται σαν η μέση τιμή όλων των δυνατών ζευγαριών αντικειμένων στην ομάδα που θα προκύψει, εάν οι δύο αυτές ομάδες ενωθούν.

Εικόνα 23: μέθοδος μέσης απόστασης μέσα στις ομάδες

Παράδειγμα 3.4.1 (συνέχεια)

Θα χρησιμοποιήσουμε τις τρεις μεταβλητές και την ευκλείδεια απόσταση για να πάρουμε τον πίνακα τις απόστασης ο οποίος είναι ίσος με:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 21.35 & 0 & & & \\ 22.06 & 1.79 & 0 & & \\ 18.64 & 2.94 & 4.35 & 0 & \\ 12.78 & 12.60 & 12.55 & 10.93 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ζευγαριών είναι αυτή ανάμεσα στην Αγγλία και τη Γερμανία δηλαδή $\min_{i,j}(d_{ij}) = d_{32} = 1.79$. Άρα τα αντικείμενα 2 και 3 θα σχηματίσουν την συστάδα (23).

Στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της συστάδας (23) με τα υπόλοιπα αντικείμενα 1, 4, 5:

$$d_{(23)4} = \left(d_{23} + d_{24} + d_{34}\right)/3 = 3,03$$
$$d_{(23)5} = \left(d_{23} + d_{25} + d_{35}\right)/3 = 8,98$$
$$d_{(23)1} = \left(d_{23} + d_{21} + d_{31}\right)/3 = 15,06$$

Διαγράφουμε, στον προηγούμενο πίνακα αποστάσεων, τις γραμμές 2, 3 και τις στήλες 2, 3 και προσθέτουμε μια γραμμή και μια στήλη για την ομάδα (23), δημιουργώντας έναν νέο πίνακα αποστάσεων:

$$\begin{array}{c} \\ 1 \\ (23) \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & (23) & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & \\ 15.06 & 0 & & \\ 18,64 & 3,03 & 0 & \\ 12,78 & 8,98 & 10,93 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή ανάμεσα στην (Αγγλία, Γερμανία) με την Ισπανία δηλαδή $\min_{i,j}(d_{ij}) = d_{(23)4} = 3.03$. Άρα τα αντικείμενα (23) και 4 θα σχηματίσουν την ομάδα (234).

Συνεχίζοντας, στο επόμενο επίπεδο θα χρειαστούμε τις αποστάσεις της συστάδας (234) από τια αντικείμενα 1, 5:

$$d_{(234)5} = \left(d_{23} + d_{24} + d_{25} + d_{34} + d_{35} + d_{45}\right)/6 = 7,53$$
$$d_{(234)1} = \left(d_{23} + d_{24} + d_{21} + d_{34} + d_{31} + d_{41}\right)/6 = 11,85$$

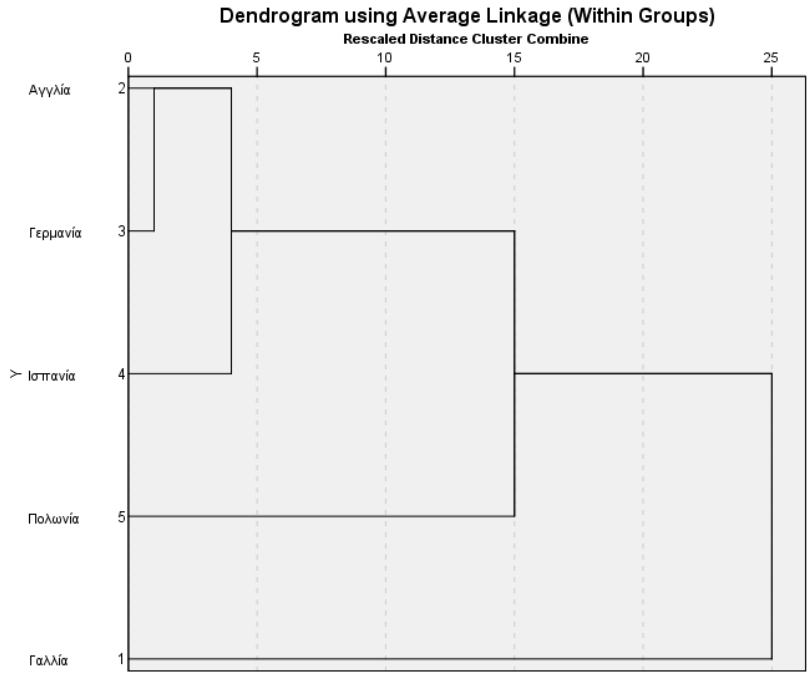
Διαγράφουμε, στον προηγούμενο πίνακα αποστάσεων, τις γραμμές (23), 4 και τις στήλες (23), 4 και προσθέτουμε μια γραμμή και μια στήλη για τη συστάδα (234).

$$\begin{array}{c} \\ 1 \\ (234) \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & (234) & 5 \end{array} \\ \begin{pmatrix} 0 & & \\ 11,85 & 0 & \\ 12,78 & 7,53 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή ανάμεσα στη (Αγγλία, Γερμανία, Ισπανία) με την Πολωνία δηλαδή $\min_{i,j}(d_{ij}) = d_{(234)5} = 7.53$. Άρα οι ομάδες (234) και 5 ενώνονται και θα σχηματίσουν την ομάδα (2345).

Στο τελευταίο επίπεδο ομαδοποίησης, η ομάδα (2345) ενώνεται με το αντικείμενο 1 και σχηματίζουν μια ομάδα την (12345). Η απόσταση μεταξύ τους είναι:

$$d_{(12345)5} = \left(d_{12} + d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25} + d_{34} + d_{35} + d_{45}\right)/10 = 12$$



Εικόνα 24: δενδρόγραμμα (μέθοδος μέσης απόστασης μέσα στις ομάδες)

Παρατήρηση 3.4.6

Το στατιστικό πρόγραμμα SPSS δίνει τ' αποτελέσματα:

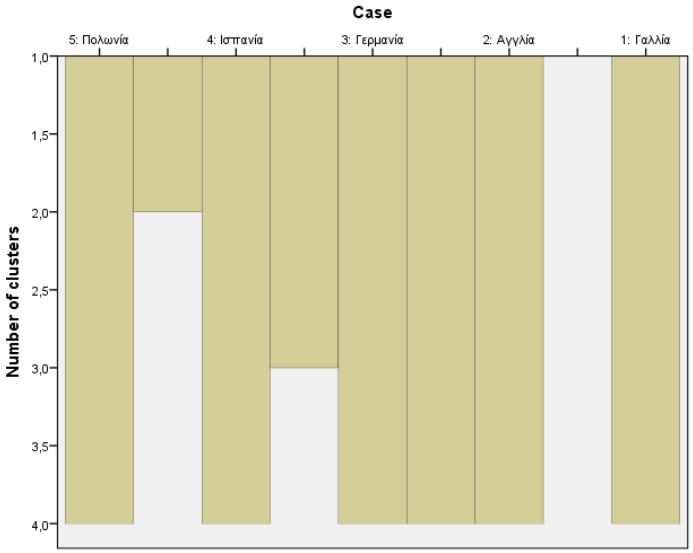
(α) πίνακα που περιγράφει τα στάδια της διαδικασίας

Average Linkage (Within Groups)

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	3	1,794	0	0	2
2	2	4	3,033	1	0	3
3	2	5	7,533	2	0	4
4	1	2	12,004	0	3	0

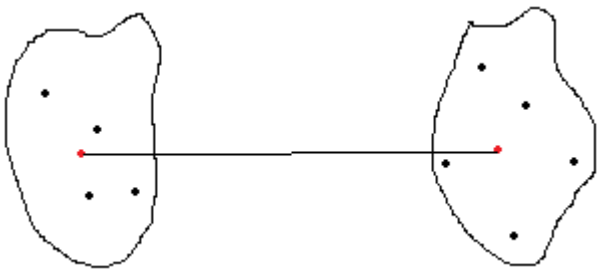
(β) Αναπαράσταση της διαδικασίας με τη βοήθεια ενός γραφήματος που λέγεται **icicle**



Εικόνα 25: διάγραμμα icicle (μέση απόσταση μέσα στις ομάδες)

(E) Μη σταθμισμένη σύνδεση κέντρων βάρους (average centroid linkage or unweighted pair-group centroid).

Στη λεγόμενη **μη σταθμισμένη σύνδεση κέντρων βάρους** *η απόσταση μεταξύ δύο ομάδων είναι ίση με την απόσταση των κέντρων βάρους (centroids) των ομάδων* (οι συντεταγμένες του κέντρου βάρους μιας ομάδας είναι οι μέσοι όροι των συντεταγμένων των αντικειμένων της ομάδας). Οι ομάδες που δημιουργούνται με τη μέθοδο είναι συμπαγείς και ελλειπτικές.



Εικόνα 26: μη σταθμισμένη σύνδεση κέντρων βάρους

Ο τύπος που δίνει την απόσταση, ανάμεσα σε δύο ομάδες, είναι :

$$d\left(O_1, O_2\right) = d\left(K_1, K_2\right)$$

όπου O_1, O_2 οι ομάδες και K_1, K_2 οι αντίστοιχες συντεταγμένες των κέντρων βάρους των ομάδων.

Παράδειγμα 3.4.1 (συνέχεια)

Χώρα	Γεννήσεις/1000 κατοίκους	Θάνατοι/1000 κατοίκους	Θάνατοι βρεφών/ 1000 γεννήσεις
Γαλλία	24.7	5.7	30.8
Αγγλία	12.5	11.9	14.4
Γερμανία	11.6	13.4	14.8
Ισπανία	14.3	10.2	16
Πολωνία	13.6	10.7	26.9

Θα χρησιμοποιήσουμε τις τρεις μεταβλητές και την ευκλείδεια απόσταση για να πάρουμε τον πίνακα τις απόστασης ο οποίος είναι ίσος με:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 21.35 & 0 & & & \\ 22.06 & 1.79 & 0 & & \\ 18.64 & 2.94 & 4.35 & 0 & \\ 12.78 & 12.60 & 12.55 & 10.93 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ζευγαριών είναι αυτή ανάμεσα στην Αγγλία και τη Γερμανία δηλαδή $\min_{i,j}(d_{ij}) = d_{32} = 1.79$. Άρα τα αντικείμενα 3 και 2 θα σχηματίσουν την συστάδα (23)

Στο επόμενο επίπεδο ομαδοποίησης υπολογίσουμε το μέσο/κέντρο βάρους των αντικειμένων 2 και 3 δηλαδή της Αγγλίας και της Γερμανίας για κάθε μεταβλητή δηλαδή, ξεχωριστά για τις Γεννήσεις/1000 κατοίκους, Θάνατοι/1000 κατοίκους, Θάνατοι βρεφών/ 1000 γεννήσεις. Επομένως ο μέσος όρος των Γεννήσεων/1000 κατοίκους της Αγγλίας και Γερμανίας είναι $(12.5+11.6)/2 = 12.05$, των Θανάτων/1000 κατοίκους είναι $(11.9+13.4)/2 = 12.65$ και των Θανάτων βρεφών/1000 γεννήσεις είναι $(14.4+14.8)/2 = 14.6$. Άρα ο πίνακας των δεδομένων μου θα μετασχηματιστεί ως εξής:

Χώρα	Γεννήσεις/1000 κατοίκους	Θάνατοι/1000 κατοίκους	Θάνατοι βρεφών/ 1000 γεννήσεις
Γαλλία	24.7	5.7	30.8
{Αγγλία, Γερμανία}	12.05	12.65	14.6
Ισπανία	14.3	10.2	16
Πολωνία	13.6	10.7	26.9

Υπολογίζουμε εκ νέου τον πίνακα των αποστάσεων με τη χρήση της ευκλείδειας απόστασης

$$\begin{array}{c} \\ 1 \\ (23) \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & (23) & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & \\ 21.69 & 0 & & \\ 18.64 & 3.6 & 0 & \\ 12.78 & 12.54 & 10.93 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ζευγαριών τώρα είναι η (Αγγλία, Γερμανία) με την Ισπανία δηλαδή $\min_{i,j}(d_{ij}) = d_{(23)4} = 3.6$. Άρα τα αντικείμενα (23) και 4 ενώνονται και θα σχηματίσουν την ομάδα (234).

Στο επόμενο επίπεδο ομαδοποίησης υπολογίζουμε το κέντρο βάρους των αντικειμένων (23) και 4 δηλαδή της (Αγγλίας, Γερμανίας) και της Ισπανίας. Επομένως ο μέσος όρος των Γεννήσεων/1000 κάτοικους της (Αγγλίας, Γερμανίας) και της Ισπανίας είναι $(12.5+11.6+14.3)/3 = 12.8$, των Θανάτων/1000 κάτοικους είναι $(11.9+13.4+10.2)/3 = 11.83$ και των Θανάτων βρεφών/1000 γεννήσεις είναι $(14.4+14.8+16)/3 = 15.06$. Άρα ο πίνακας των δεδομένων μου θα μετασχηματιστεί ως εξής:

Χώρα	Γεννήσεις/1000 κάτοικους	Θάνατοι/1000 κάτοικους	Θάνατοι βρεφών/ 1000 γεννήσεις
Γαλλία	24.7	5.7	30.8
{Αγγλία, Γερμανία, Ισπανία}	12.8	11.83	15.06
Πολωνία	13.6	10.7	26.9

Ο πίνακας των αποστάσεων είναι:

$$\begin{array}{c} \\ 1 \\ (234) \\ 5 \end{array}\begin{array}{c} \begin{array}{ccc} 1 & (234) & 5 \end{array} \\ \begin{pmatrix} 0 & & \\ 20.66 & 0 & \\ 12.78 & 11.92 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ζευγαριών είναι η (Αγγλία, Γερμανία, Ισπανία) με την Πολωνία δηλαδή

$$\min_{i,j}(d_{ij}) = d_{(234)5} = 11.92$$

Άρα τα αντικείμενα (234) και 5 θα σχηματίσουν την συστάδα (2345)

Στο επόμενο επίπεδο ομαδοποίησης υπολογίζουμε το κέντρο βάρους των αντικειμένων (234) και 5 δηλαδή της (Αγγλίας, Γερμανίας, Ισπανίας) και της Πολωνία. Επομένως ο μέσος όρος των Γεννήσεων/1000 κάτοικους της (Αγγλίας, Γερμανίας, Ισπανίας) και της Πολωνίας είναι $(12.5+11.6+14.3+13.6)/4=13$, των Θανάτων/1000 κάτοικους είναι $(11.9+13.4+10.2+10.7)/4=11.55$ και των Θανάτων βρεφών/1000 γεννήσεις είναι $(14.4+14.8+16+26.9)/4=18.02$. Άρα ο πίνακας των δεδομένων μου θα μετασχηματιστεί ως εξής:

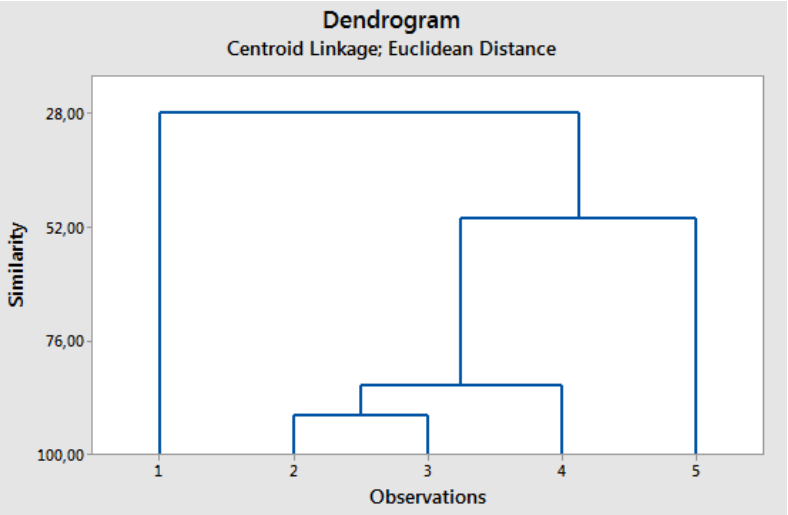
Χώρα	**Γεννήσεις/1000 κατοικους**	**Θάνατοι/1000 κατοικους**	**Θανατοι βρεφών/ 1000 γεννησεων**
Γαλλία	24.7	5.7	30.8
{Αγγλία, Γερμανία,Ισπανία, Πολωνία}	13	11.55	18.02

Ο πίνακας των αποστάσεων είναι:

$$\begin{array}{c} \\ 1 \\ (2345) \end{array}\begin{array}{c} \begin{array}{cc} 1 & (2345) \end{array} \\ \begin{pmatrix} 0 & \\ 18.28 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ζευγαριών είναι η (Αγγλία, Γερμανία, Ισπανία, Πολωνία) με την Γαλλία δηλαδή $\min_{i,j}(d_{ij}) = d_{(2345)1} = 18.28$

Άρα τα αντικείμενα (2345) και 1 θα σχηματίσουν την συστάδα την (12345)



Εικόνα 27: δενδρόγραμμα (μη σταθμισμένη σύνδεση κέντρων βάρους)

(ΣΤ) Μέθοδος σύνδεσης του Ward

Η μέθοδος **σύνδεσης του Ward** διαφέρει από τις μεθόδους που περιγράψαμε μέχρι τώρα, γιατί δεν υπολογίζει αποστάσεις ανάμεσα στις ομάδες. Θεωρεί την ανάλυση συστάδων σαν ένα πρόβλημα διασποράς αντί για χρήση μετρικών και μέτρων ομοιότητας. Κριτήριο για τη δημιουργία ομάδων είναι η μεγιστοποίηση της ομοιογένειας στο εσωτερικό τους. Ένα μέτρο ομοιογένειας που χρησιμοποιείται είναι το άθροισμα των *τετραγώνων των σφαλμάτων*:

$$ESS = \sum_i \sum_j \sum_k \left(x_{ijk} - K_{i.k} \right)^2$$

όπου x_{ijk} είναι η τιμή της j παρατήρησης, ως προς την k μεταβλητή που ανήκει στην i ομάδα και $K_{i.k}$ είναι το κέντρο βάρους της των παρατηρήσεων της i ομάδας ως προς την k μεταβλητή K_i. Ακόμα, μπορούμε να ορίσουμε το *ολικό άθροισμα τετραγώνων*:

$$TSS = \sum_i \sum_j \sum_k \left(x_{ijk} - K_{..k} \right)^2$$

όπου $K_{..k}$ είναι το κέντρο βάρους ως προς την k μεταβλητή. Το:

$$r^2 = \frac{TSS - ESS}{TSS}$$

Επιδίωξη της μεθόδου είναι η ελαχιστοποίηση του ESS ή μεγιστοποίηση του r^2.

Το ίδιο κριτήριο όπως θα δούμε σε επόμενο κεφάλαιο, χρησιμοποιείται και από τον αλγόριθμο k-Means, οπότε η μέθοδος Ward μπορεί να θεωρηθεί ανάλογο της . Η μέθοδος είναι κατάλληλη για ποσοτικές μεταβλητές.

Η μέθοδος, για να συνενώσει δύο ομάδες από ένα πλήθος *k ομάδων*, ελέγχει όλα τα δυνατά

$$\binom{k}{2} = \frac{k!}{2!(k-2)!} = \frac{k(k-1)}{2}$$

ζευγάρια ομάδων τα οποία μπορούν να δημιουργηθούν (από συνενώσεις δυό ομάδων), και επιλέγει εκείνο το ζευγάρι, το οποίο όταν ενωθεί θα μας δώσει την ομάδα με το μικρότερο τετραγωνικό σφάλμα.

Η μέθοδος του Ward έχει την τάση να παράγει ισοπληθείς ομάδες.

Το συγκεκριμένο μέτρο είναι πολύ γνωστό κριτήριο στην ανάλυση συστάδων και χρησιμοποιείτε κατά κόρον στις μεθόδους βελτιστοποίησης.

Παράδειγμα 3.4.1 (συνέχεια)

Χώρα	**Γεννήσεις/1000 κάτοικους**	**Θάνατοι/1000 κάτοικους**	**Θάνατοι βρεφών/ 1000 γεννήσεις**
Γαλλία	24.7	5.7	30.8
Αγγλία	12.5	11.9	14.4
Γερμανία	11.6	13.4	14.8
Ισπανία	14.3	10.2	16
Πολωνία	13.6	10.7	26.9

Για λόγους ευκολίας γράφουμε:

	M1	**M2**	**M3**
A1	24.7	5.7	30.8
A2	12.5	11.9	14.4
A3	11.6	13.4	14.8
A4	14.3	10.2	16
A5	13.6	10.7	26.9

Αρχικά θεωρούμε κάθε παρατήρηση ως μία ομάδα, οπότε έχουμε 5 διαφορετικές ομάδες Στο επόμενο στάδιο, δύο αντικείμενα θα ενωθούν για το σχηματισμό μιας ομάδας. Υπάρχουν:

$$\binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{5(5-1)}{2} = 10$$

τέτοιες συνενώσεις από τις οποίες θα επιλέξουμε εκείνη που δίνει το μικρότερο τετραγωνικό σφάλμα.

Το τετραγωνικό σφάλμα για την ομάδα που αποτελείται από τα αντικείμενα A1 και A2 υπολογίζεται ως εξής:

$$ESS(A1, A2) = (24,7-18,6)^2 + (12,5-18,6)^2 + (5,7-8,8)^2 + (11,9-8,8)^2 + (30,8-22,6)^2 + \\ + (14,4-22,6)^2 = 228,12$$

Οι υπόλοιπες ομάδες/αντικείμενα έχουν σφάλμα μηδέν γιατί αποτελούνται από μία μόνο παρατήρηση. Έτσι το συνολικό σφάλμα στη συγκεκριμένη περίπτωση είναι 228,12.

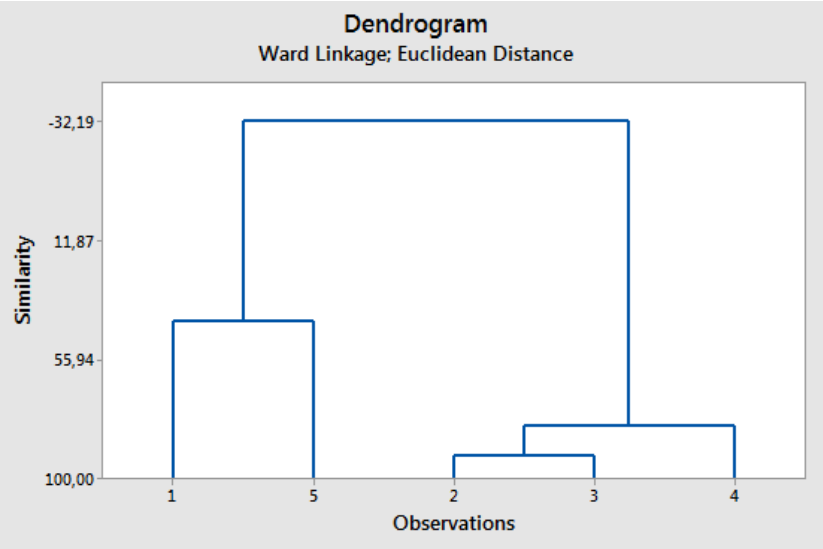
Συνεχίζοντας, εφαρμόζουμε την ίδια διαδικασία και για τις υπόλοιπες 9 επιλογές. Τα αποτελέσματα (αναφορικά με το τετραγωνικό σφάλμα) φαίνονται στον παρακάτω πίνακα.

Λύση	**1**	**2**	**3**	**4**	**ESS**
1	A1A2	A3	A4	A5	228.12
2	A2A3	A1	A4	A5	1.79
3	A3A4	A1	A2	A5	9.5
4	A4A5	A1	A2	A3	59.82
5	A1A3	A2	A4	A5	243.48
6	A1A4	A2	A3	A5	173.74
7	A1A5	A2	A3	A4	81.74
8	A2A4	A1	A3	A5	4.36
9	A2A5	A1	A3	A4	79.48
10	A3A5	A1	A2	A4	78.88

Βλέπουμε ότι το μικρότερο τετραγωνικό σφάλμα εντοπίζεται στην ένωση των αντικειμένων A2,A3. Έτσι τα αντικείμενα A2 και A3 ενώνονται σε μια ομάδα ενώ όλα τ΄ άλλα αντικείμενα αποτελούν από μόνα τους μια ομάδα το καθένα.

Συνεχίζουμε τη διαδικασία στο επόμενο βήμα, δημιουργούμε ή 2 ομάδες των δύο αντικειμένων (η μια είναι αυτή των A2,A3) και τα υπόλοιπα αντικείμενα αποτελούν από μόνα τους μια ομάδα το καθένα ή μια ομάδα των τριών αντικειμένων, δυο από τα οποία είναι τα A2, A3 ενώ τα υπόλοιπα αντικείμενα και εδώ αποτελούν το καθένα μια ομάδα. Και εδώ

ψάχνουμε για την ομαδοποίηση που δίνει το μικρότερο τετραγωνικό σφάλμα. Συνεχίζουμε τη διαδικασία με ανάλογο τρόπο.



Εικόνα 28: δενδρόγραμμα (μέθοδος σύνδεσης του Ward)

Παρατήρηση 3.4.6

Το στατιστικό πρόγραμμα SPSS δίνει τ' αποτελέσματα:

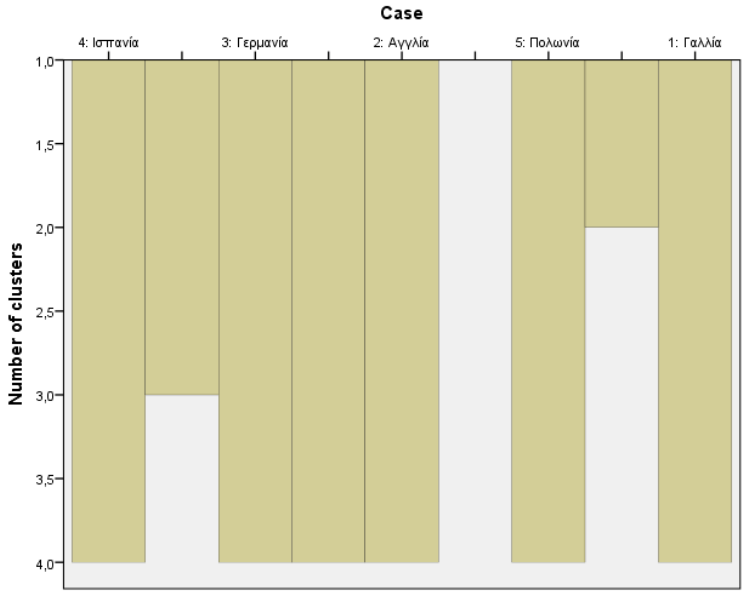
(α) πίνακα που περιγράφει τα στάδια της διαδικασίας

Ward Linkage

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	3	,897	0	0	2
2	2	4	3,033	1	0	4
3	1	5	9,424	0	0	4
4	1	2	24,009	3	2	0

(β) Αναπαράσταση της διαδικασίας με τη βοήθεια ενός γραφήματος που λέγεται **icicle**



Εικόνα 29: διάγραμμα icicle (Μέθοδος σύνδεσης του Ward)

Κεφάλαιο 4: Μη ιεραρχική ομαδοποίηση / Διαχωριστικές μέθοδοι

4.1 Εισαγωγή

Μιά δεύτερη κατηγορία μεθόδων ομαδοποίησης είναι οι λεγόμενες **μη ιεραρχικές** μέθοδοι ή **διαχωριστικές (partitioning methods)**. Στις μεθόδους αυτές, έχουμε ένα πλήθος από *N* σημεία/αντικείμενα τα οποία, με τη βοήθειά των μεθόδων, διαμερίζονται/κατανέμονται σε k ομάδες. Ο αριθμός αυτός των k ομάδων προκαθορίζεται από τον ερευνητή. Οι μέθοδοι ξεκινούν από έναν αρχικό διαχωρισμό σε ομάδες (με τη βοήθεια ορισμένων αρχικών συνθηκών/σημείων), και με τη χρήση ενός αλγορίθμου, τα αντικείμενα επανακατανέμονται σε ομάδες. Η δημιουργία των ομάδων γίνεται κατά τρόπο τέτοιο, ώστε να βελτιστοποιείται μιά συνάρτηση κριτηρίου διαχωρισμού (όπως το άθροισμα τετραγώνων των σφαλμάτων) είτε τοπικά είτε καθολικά. Οι διαχωριστικές μέθοδοι δημιουργούν ένα σύνολο από k ομάδες, σε αντίθεση με τις ιεραρχικές μεθόδους, οι οποίες δημιουργούν μια (ιεραρχική) σειρά διαδοχικών επιπέδων ομαδοποίησης, με κάθε επίπεδο να ορίζει ένα σύνολο ομάδων (μιά δυνατή ομαδοποίηση).

Μειονέκτημα των μεθόδων αποτελεί η ευαισθησία τους στις αρχικές τους συνθήκες/σημεία όπως επίσης ο καθορισμός, πριν την εφαρμογή τους, του πλήθους των ομάδων. Ακόμα, για την εύρεση της καθολικά βέλτιστης λύσης θα πρέπει να δοκιμαστούν όλοι οι δυνατοί διαχωρισμοί, πράγμα αδύνατο (τουλάχιστον από πλευράς κόστους). Στην πράξη, εκτελούμε τον αλγόριθμο ένα ικανό αριθμό φορών, με διαφορετικά αρχικά σημεία κάθε φορά, και το καλύτερο αποτέλεσμα που λαμβάνουμε είναι η επιθυμητή ομαδοποίηση.

Η πιο γνωστή μέθοδος διαχωριστικής ανάλυσης συστάδων είναι ο αλγόριθμός k-means και η μέθοδος των μικτών κατανομών E-M.

4.2 Μέθοδος k-means

Η μέθοδος k-means είναι ίσως από τους πιό γνωστούς και περισσότερο χρησιμοποιούμενους αλγόριθμους ομαδοποίησης. Κατατάσσει αντικείμενα σε προκαθορισμένο αριθμό ομάδων, έτσι ώστε τα αντικείμενα που ανήκουν στην ίδια ομάδα να είναι όσο το δυνατόν πιο ομοιόμορφα (δηλαδή, υψηλή ενδοκλασική ομοιότητα), ενώ τα αντικείμενα σε διαφορετικές ομάδες να είναι όσο το δυνατόν πιο διαφορετικά. Στην ομαδοποίηση k-means, κάθε ομάδα αντιπροσωπεύεται από το κέντρο της (δηλαδή το κέντρο βάρους της) που αντιστοιχεί στο μέσο όρο των συνεταγμένων των σημείων που αντιστοιχούν στην ομάδα.

Η βασική ιδέα πίσω από την ομαδοποίηση k-means βρίσκεται στο ότι οι ομάδες ορίζονται κατά τέτοιο τρόπο ώστε να ελαχιστοποιείται η συνολική διακύμανση μέσα στις ομάδες. (κριτήριο διαχωρισμού). Υπάρχουν αρκετοί αλγόριθμοι k-means. Ο πρότυπος αλγόριθμος είναι αυτός των Hartigan-Wong (Hartigan and Wong 1979), ο οποίος ορίζει τη συνολική μεταβολή/διακύμανση μέσα σε μιά ομάδα σαν το άθροισμα των τετραγωνικών αποστάσεων (Euclidean) μεταξύ των αντικειμένων της ομάδας και του αντίστοιχου κέντρου βάρους (centroid) της:

$$W\left(C_j\right) = ESS_{C_j} = \sum_{x_i \in C_j} \left(x_i - \mu_j\right)^2$$

όπου C_j η j ομάδα, x_i αντικείμενο της ομάδας και μ_j το κέντρο βάρους της.

Κάθε παρατήρηση x_i αντιστοιχίζεται σε μια δεδομένη ομάδα έτσι ώστε η απόσταση της από το κένρο βάρους της ομάδας να είναι η μικρότερη. Το **συνολικό άθροισμα τετραγώνων των αποκλίσεων** (συνολική διακύμανση της ομαδοποίησης) είναι ίσο με:

$$W\left(C\right) = ESS_{total} = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left(x_i - \mu_j\right)^2$$

ποσότητα που θέλουμε να είναι η ελάχιστη δυνατή.

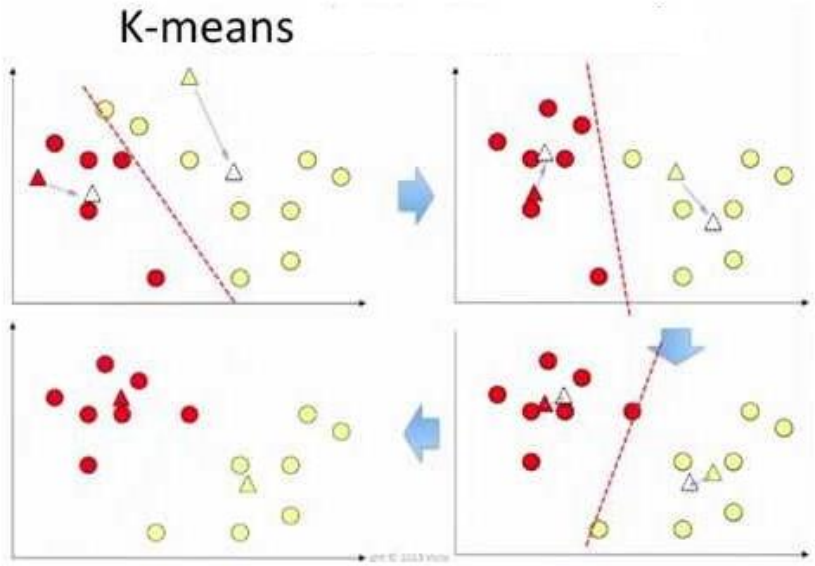
Αλγόριθμος k-means

Το πρώτο βήμα λοιπόν κατά την ομαδοποίηση k-means είναι ο καθορισμός του αριθμού των ομάδων (k) που θα δημιουργηθούν (σαν τελική λύση).

Ο αλγόριθμος ξεκινά με την (τυχαία) επιλογή k αντικειμένων από το σύνολο δεδομένων (ή κάποια άλλη επιλογή σημείων) που θα χρησιμεύσουν σαν αρχικά κέντρα για τις ομάδες (γνωστά ως μέσα ή κεντροειδή).

Στη συνέχεια, κάθε ένα από τα υπόλοιπα αντικείμενα αντιστοιχίζεται στο πλησιέστερο (αναφορικά με την Ευκελίδια απόσταση) κέντρο βάρους/ομάδα. Κατόπιν, ο αλγόριθμος υπολογίζει το νέο κέντρο βάρους κάθε συστάδας ("update centroid"). Μετά τον υπολογισμό αυτό, κάθε παρατήρηση ελέγχεται και αντιστοιχίζεται και πάλι σ' ένα κένρο βάρους/ομάδα. Τα βήματα επαναλαμβάνονται, μέχρι να μην γίνονται αλλάγές στις αντιστοιχίες των σημείων με τα κέντρα βάσους (σύγλκιση του αλγόριθμου), δηλαδή οι ομάδες που σχηματίστηκαν σε ένα βήμα εφαρμογής του αλγόριθμου να είναι ίδιες (ή περίπου ίδιες) με αυτές που θα

σχηματίζονταν σε μιά επιπλέον εφαρμογή του. Πολλές φορές ο αλγόριθμος σταματά μετά από έναν ορισμένο αριθμό επαναλήψεων/εφαρμογών του (επιλεγμένο από τον ερευνητή).



Εικόνα 30: η μέθοδος k-means

Τα βήματα λοιπόν του αλγόριθμου k-means μπορεί να συνοψιστούν ως εξής:

- **Βήμα 1**: Καθορίζουμε τον αριθμό των ομάδων (k) που θα δημιουργηθούν (καθορισμός από τον ερευνητή)
- **Βήμα 2**: Επιλέγουμε (τυχαία) k αντικείμενα ως αρχικά κέντρα (βάρους) των ομάδων.
- **Βήμα 3**: Αντιστοιχούμε κάθε παρατήρηση στο πλησιέστερο κέντρο βάρους (με βάση συνήθως την ευκλείδεια απόσταση μεταξύ του αντικειμένου και του κέντρου βάρους)
- **Βήμα 4**: Για κάθε ομάδα, επαναϋπολογίζουμε το κέντρο της χρησιμοποιώντας τις νέες μέσες τιμές όλων των σημείων/δεδομένων στην ομάδα.
- **Βήμα 5**: Προσπαθούμε να ελαχιστοποιήσουμε το συνολικό άθροισμα τετραγώνων των σφαλάματων (διακύμανση) μέσα στις ομάδες. Δηλαδή, επαναλαμ-βάνουμε τα βήματα 3 και 4 μέχρι να σταματήσουν οι ανακατανομές των σημείων στις ομάδες που δημιουργούνται ή να επιτευχθεί ένα μέγιστος (προεπιλεγμένος) αριθμός επαναλήψεων.

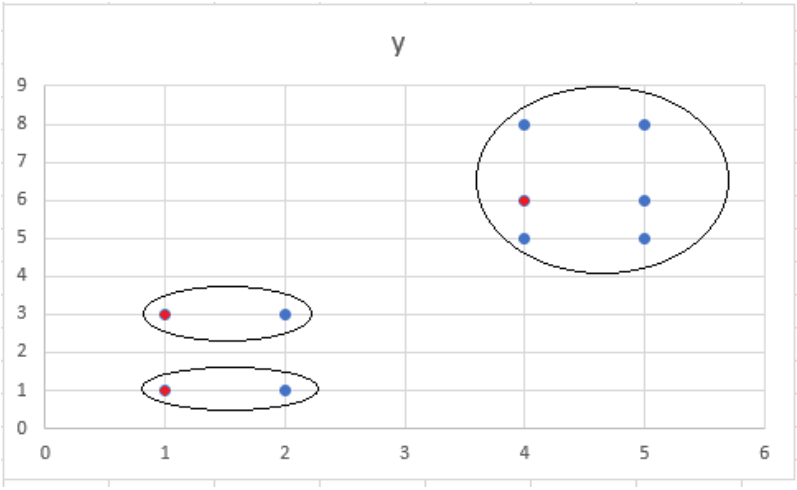
Τα κυριότερα **πλεονεκτήματα** του αλγόριθμου είναι τα εξής:

- Είναι απλός και εύκολος στην εφαρμογή του (τις περισσότερες φορές δεν απαιτεί ούτε καν πλήρη γνώση του, αφού περιλαμβάνεται στα περισσότερα στατιστικά προγράμματα).
- Η πολυπλοκότητα του αλγόριθμου είναι γραμμική της τάξης $O(knl)$ όπου k το πλήθος των ομάδων, n το πλήθος των αντικειμένων και l το πλήθος των επαναλήψεων (δεν χρειάζεται να αποθηκευθούν οι αποστάσεις μεταξύ των αντικειμένων σε έναν πίνακα και με κάθε συγχώνευση ή διαίρεση να υπολογίζεται εκ νέου ο πίνακας όπως συμβαίνει στις ιεραρχικές μεθόδους). Αυτό σημαίνει ότι είναι κατάλληλος (μικρότερος χρόνος και κόστος επεξεργασίας) για μεγάλα σύνολα δεδομένων σε αντίθεση με τις περισσότερες από τις ιεραρχικές μεθόδους.
- Δημιουργεί συμπαγείς ομάδες, σφαιρικές με ίσο αριθμό αντικειμένων.

Τα βασικότερα **μειονεκτήματα** του αλγόριθμου είναι τα εξής

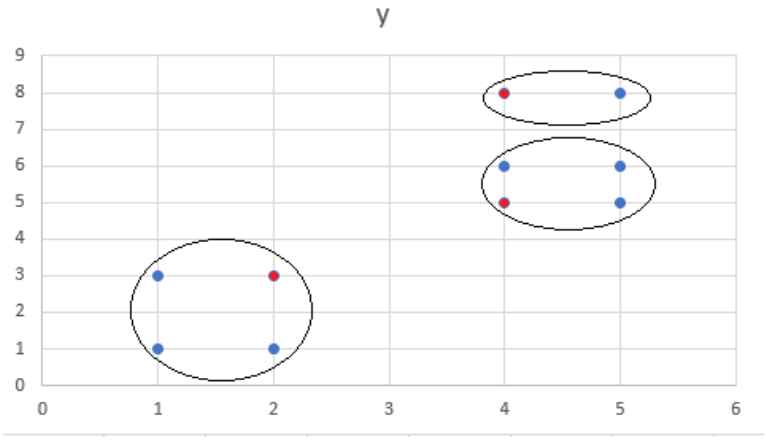
- Ο αλγόριθμος μπορεί να εφαρμοστεί μόνον σε ποσοτικές μεταβλητές
- Εφαρμόζεται σε δεδομένα που σχηματίζουν σφαιρικές, συμπαγείς ομάδες με ίσο αριθμό αντικειμένων (και ίσες διασπορές). Η παραβίαση αυτής της υπόθεσης δεν κάνει αδύνατη τη χρήση της μεθόδου, αλλά πρέπει να είμαστε προσεκτικοί στην ερμηνεία των αποτελεσμάτων.
- Απαιτεί την επιλογή του αριθμού των ομάδων (k) πριν την εφαρμογή του.
- Εξαρτάται από την επιλογή των αρχικών σημείων (κέντρων). Διαφορετικά αρχικά κέντρα δίνουν, συνήθως, διαφορετική ομαδοποίηση.

  Π.χ έστω ότι έχουμε τα σημεία (1,1), (1,3), (3,1), (3,3), (4,5), (4,6), (4,8), (5,5), (5,6), (5,8). Εάν επιλέξουμε σαν κέντρα τα (κόκκινα) σημεία (1,1), (1,3) και (4,6) τότε δημιουργούνται οι τρείς ομάδες που φαίνονται στο παρακάτω σχήμα.



Εικόνα 31: παράδειγμα 1 εφαρμογής αλγορίθμου k-means

Εάν τώρα επιλέξουμε σαν κέντρα τα (κόκκινα) σημεία (2,3), (4,5) και (4,8) τότε δημιουργούνται οι τρείς ομάδες που φαίνονται στο παρακάτω σχήμα.



Εικόνα 32: παράδειγμα 2 εφαρμογής αλγορίθμου k-means

Είναι προφανές ότι οι ομάδες των δύο παραπάνω σχημάτων (για τα ίδια δεδομένα) είναι διαφορετικές.

- Είναι ευαίσθητος στην ύπαρξη ακραίων τιμών. Αντικείμενα με μεγάλες τιμές/ συντεταγμένες επηρεάζουν συνήθως τον καθορισμό των νέων κέντρων γεγονός που έχει επίδραση στο σχηματισμό και τη διαμόρφωση των τελικών ομάδων.
- Είναι ευαίσθητος στην κλίμακα μέτρησης των δεδομένων (π.χ. κανονικοποίηση των δεδομένων μπορεί να αλλάξει ριζικά την ομαδοποίηση).

Παράδειγμα 4.2.1

	V1 (γεννήσεις)	V2 (θάνατοι)	V3 (παιδική θνη/τα)
S1 (Γαλλία)	24.7	5.7	30.8
S2 (Αγγλία)	12.5	11.9	14.4
S3 (Γερμανία)	11.6	13.4	14.8
S4 (Ισπανία)	14.3	10.2	16
S5 (Πολωνία)	13.6	10.7	26.9

Θα εφαρμόσουμε τη μέθοδο k- means για k=2. Με τη βοήθεια του προγράμματος SPSS έχουμε τα παρακάτω αποτελέσματα:

Cluster Membership

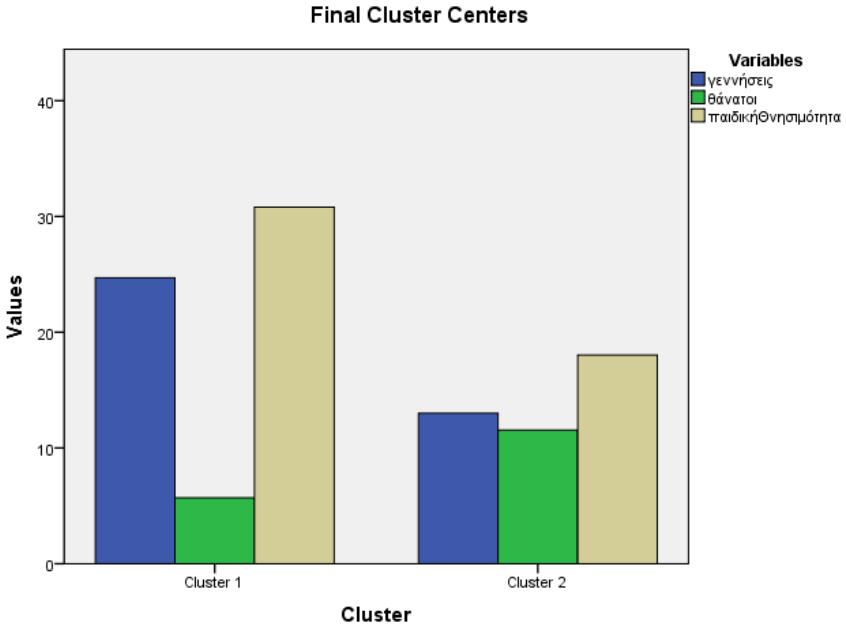
Case Number	ID	Cluster	Distance
S1	Γαλλία	1	,000
S2	Αγγλία	2	3,676
S3	Γερμανία	2	3,973
S4	Ισπανία	2	2,759
S5	Πολωνία	2	8,936

Από την τρίτη στήλη του πίνακα βλέπουμε ότι σχηματίζονται δύο ομάδες, οι {S1}, {S2,S3,S4,S5}.

Final Cluster Centers

	Cluster	
	1	2
γεννήσεις	24,70	13,00
θάνατοι	5,70	11,55
παιδικήΘνησιμότητα	30,80	18,03

Οι συντεταγμένες των τελικών κέντρων των ομάδων είναι (24,7, 5,7, 30,8) και (13, 11,55, 18,03) αντίστοιχα. Από τον πίνακα βλέπουμε ότι η $1^{η}$ ομάδα χαρακτηρίζεται από μεγάλο αριθμό γεννήσεων (24,70) και μεγάλη παιδική θνησιμότητα (30,80), ενώ η $2^{η}$ από μεγάλο αριθμό θανάτων (οι διαφορές μεταξύ των ομάδων μπορεί να μην είναι στατιστικά σημαντικές, γεγονός που πρέπει να ελεγχθεί). Τα συμπεράσματα αυτά μπορούν να παρατηρηθούν και στο παρακάτω γράφημα.



Εικόνα 33: τελικά κέντρα βάρους ομάδων

Distances between Final Cluster Centers

Cluster	1	2
1		18,284
2	18,284	

Η απόσταση ανάμεσα στα τελικά κέντρα των ομάδων είναι 18,284

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
γεννήσεις	109,512	1	1,420	3	77,121	,003
θάνατοι	27,378	1	2,030	3	13,487	,035
παιδικήΘνησιμότητα	130,560	1	35,469	3	3,681	,151

.

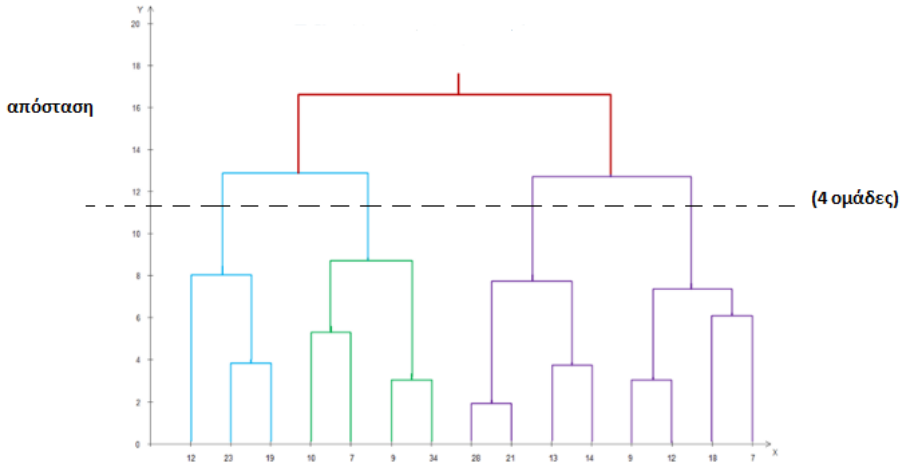
Φαίνεται ότι η διαφορά στις γεννήσεις και στους θανάτους να είναι στατιστικά σημαντική, ενώ στην παιδική θνησιμότητα όχι. Επομένως η $1^{η}$ ομάδα χαρακτηρίζεται από μεγάλο αριθμό γεννήσεων ενώ η $2^{η}$ από μεγάλο αριθμό θανάτων.

4.3 Καθορισμός του βέλτιστου αριθμού ομάδων

Ένα από τα προβλήματα που αντιμετωπίζει κανείς στην ανάλυση συστάδων (και ιδιαίτερα στη μέθοδο k-means) είναι ο καθορισμός του *βέλτιστου αριθμού των ομάδων*. Δεν υπάρχει

οριστική απάντηση στο πρόβλημα, ο βέλτιστος αυτός αριθμός είναι υποκειμενικός και εξαρτάται μεταξύ άλλων, από την απόσταση που χρησιμοποιούμε, το κριτήριο ένωσης, τις μεταβλητές που χρησιμοποιούμε για την ομαδοποίηση κ.λ.π.

Μια πρώτη απλή και αρκετά δημοφιλής προσέγγιση στο ερώτημα είναι: εφαρμόζουμε ιεραρχική ομαδοποίηση στα δεδομένα, δημιουργούμε και παρατηρούμε προσεκτικά το δενδρόγραμμα που προσπαθούμε να δούμε εάν υποδεικνύει έναν συγκεκριμένο αριθμό ομάδων (το σημείο που ίσως υπάρχει ένα άλμα στην απόσταση συνένωσης δυό ομάδων είναι ίσως ενδεικτικό του αριθμού των ομάδων που θα πρέπει να σχηματιστούν) .



Εικόνα 34: καθορισμός ομάδων με ιεραρχική ταξινόμηση

Περιγράφουμε παρακάτω μερικές από τις μεθόδους καθορισμού του (βέλτιστου) αριθμού των ομάδων k. Οι μέθοδοι αυτές χωρίζονται σε δύο ομάδες:

(I) **Άμεσες**: στην περίπτωση τους προσπαθούμε να βελτιστοποιήσουμε ένα κριτήριο διαχωρισμού όπως το ολικό άθροισμα τετραγώνων των σφαλμάτων.

(II) **Στατιστικές**: έλεγχος κατάλληλων υποθέσεων με χρήση μηδενικής και εναλλακτικής υπόθεσης.

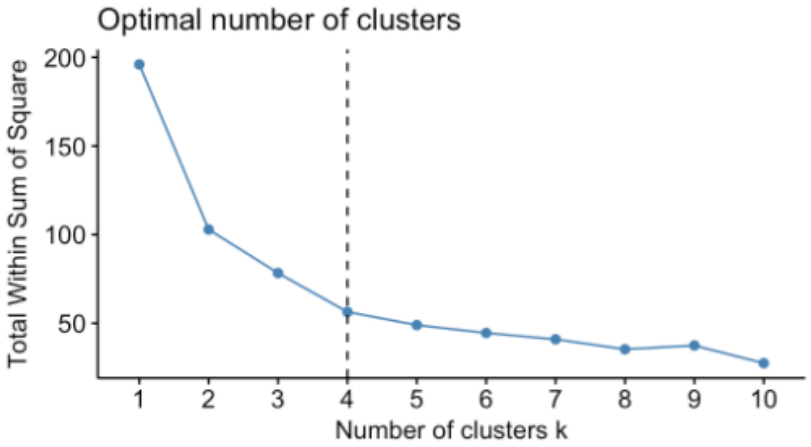
4.4 Άμεσες μέθοδοι καθορισμού του βέλτιστου αριθμού ομάδων

(Α) Μέθοδος elbow («μέθοδος του αγκώνα»)

Μια από τις απλούστερες μεθόδους καθορισμού του βέλτιστου αριθμού των ομάδων είναι και η λεγόμενη **μέθοδος elbow** («**μέθοδος του αγκώνα**»). Εφαρμόζουμε τον αλγόριθμο k-means για ένα εύρος τιμών του k (π.χ. k=1,2,.. μέχρι ένα συγκεκριμένο k), και για κάθε k υπολογίζουμε το άθροισμα τετραγώνων των σφαλμάτων ESS. Δημιουργούμε ένα διάγραμμα

όπου στον άξονα των x υπάρχουν οι τιμές του k και στον άξονα των y οι αντίστοιχες τιμές του ESS.

Θέλουμε να επιλέξουμε το k έτσι ώστε να έχουμε ένα μικρό ESS. Αλλά καθώς αυξάνουμε το k, το ESS τείνει να μειωθεί προς το 0 (εάν το k ισούται με τον αριθμό των παρατηρήσεων/ αντικειμένων, κάθε σημείο/ παρατήρηση θεωρείται μια ομάδα και έτσι το ESS είναι 0). Επομένως, επιλέγουμε το k έτσι ώστε το ESS να είναι αρκετά μικρό, αλλά ο ρυθμός αλλαγής του ESS είναι σχετικά υψηλός. Αυτό το σημείο είναι συνήθως ο «αγκώνας» (elbow) της καμπύλης.



Εικόνα 35: μέθοδος elbow

Παρατήρηση 4.4.1

Στην παραπάνω εικόνα μπορούμε να δούμε ότι είναι δύσκολο να προσδιορίσουμε που είναι στην πραγματικότητα ο «αγκώνας» της καμπύλης. Η μέθοδος του αγκώνα δίνει μιά ένδειξη που μπορεί να είναι ο βέλτιστος αριθμός k, αλλά είναι μια πολύ υποκειμενική μέθοδος και για ορισμένα σύνολα δεδομένων δεν δίνει σαφές αποτέλεσμα.

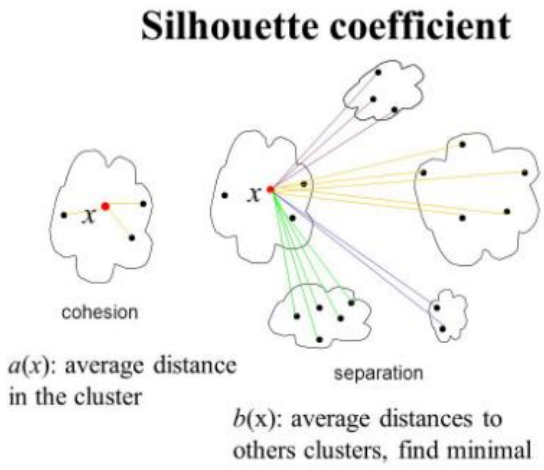
(Β) Μέθοδος average silhouette.

Μια δεύτερη μέθοδος, που μετρά την ποιότητα της ομαδοποίησης, είναι η λεγόμενη **μέθοδος average silhouette**. Καθορίζει πόσο καλά κάθε αντικείμενο βρίσκεται μέσα στην ομάδα του. Η μέθοδος υπολογίζει την ποσότητα **average silhouette** των παρατηρήσεων για διαφορετικές τιμές του k. Μεγάλη τιμής της ποσότητας **average silhouette** υποδηλώνει καλή ομαδο-ποίηση.

Αρχικά, υπολογίζεται ο δείκτης **silhouette coefficient** $s(i)$ για κάθε ένα σημείο i του δείγματός μας από τη σχέση:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

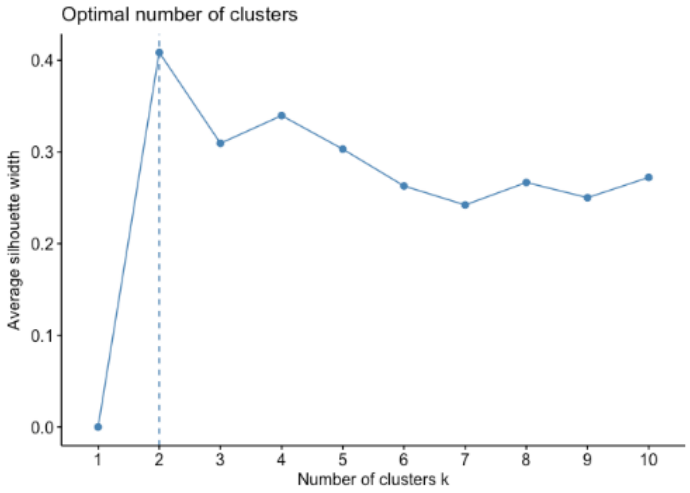
όπου $a(i)$ η μέση απόσταση του αντικειμένου i από τα στοιχεία της ομάδας που ανήκει, και για το $b(i)$ υπολογίζουμε τη μέση απόσταση του αντικειμένου i από τα στοιχεία κάθε μιάς από τις ομάδες στις οποίες δεν ανήκει, και από τις αποστάσεις αυτές $b(i)$ είναι η μικρότερη.



Εικόνα 36: υπολογισμός δείκτη silhouette coefficient
http://www.mtechprojects.org/silhouette-coefficient-projects.html

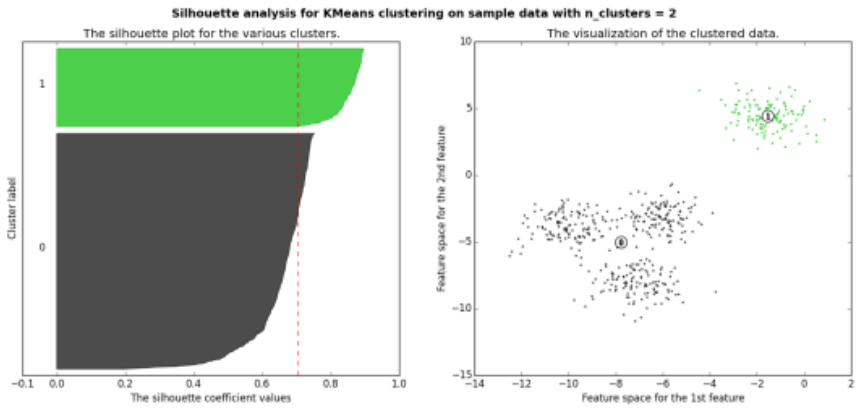
Οι τιμές του δείκτη silhouette coefficient βρίσκονται στο διάστημα [-1, 1]. Τιμή +1 υποδεικνύει ότι το αντικείμενο βρίσκεται πολύ μακριά από γειτονική ομάδα και πολύ κοντά στην ομάδα που έχει αντιστοιχιστεί. Ανάλογα, η τιμή του -1 υποδεικνύει ότι το σημείο είναι κοντά σε γειτονική ομάδα από την ομάδα που έχει αντιστοιχιστεί. Μια τιμή 0 σημαίνει ότι βρίσκεται στο όριο της απόστασης μεταξύ των δύο ομάδων. Η τιμή μας δίνει έναν τρόπο να δούμε του πόσο καλά κάθε σημείο ανήκει σε ομάδα που έχει αντιστοιχιστεί, αλλά δεν μας δίνει μια εικόνα για το πόσο καλή είναι η ομαδοποίηση.

Ορίζουμε λοιπόν το λεγόμενο δείκτη **average silhouette** που είναι ο μέσος όρος των δεικτών $s(i)$, όταν το i διατρέχει όλα τα αντικείμενα του δείγματος μας. Ο βέλτιστος αριθμός k είναι αυτός που μεγιστοποιεί την τιμή του δείκτη average silhouette σε μια σειρά πιθανών τιμών για το k.



Εικόνα 37: δείκτης average silhouette
https://uc-r.github.io/kmeans_clustering

Παράδειγμα 4.4.2



Εικόνα 38: ομαδοποίηση k-means (δεξιά) και δείκτης average silhouette (αριστερά)
http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Στο παραπάνω δεξιά διάγραμμα έχουμε την ομαδοποίηση των δεδομένων μας με τη μέθοδο k-means, σε δύο ομάδες (τη μαύρη που την ονομάζουμε 0 και την πράσινη που την ονομάζουμε 1). Αριστερά με μαύρο είναι οι τιμές του δείκτη silhouette coefficient για καθένα από τα στοιχεία της ομάδας 0 και με πράσινο για καθένα από τα στοιχεία της ομάδας 1. Η τιμή του δείκτη average silhouette είναι 0,7 (κόκκινη γραμμή). Για να είναι αυτή μια καλή τιμή (για το συγκεκριμένο αριθμό των ομάδων), θα πρέπει να λάβουμε υπόψη τα ακόλουθα σημεία

• η μέση αυτή τιμή πρέπει να είναι όσο το δυνατόν πιο κοντά στο 1

- το μεγαλύτερο μέρος της γραφικής παράστασης των δεικτών της κάθε ομάδας πρέπει να είναι πάνω από τη μέση τιμή. Οποιαδήποτε περιοχή διαγράμματος κάτω από τη μέση τιμή δεν είναι επιθυμητή.
- Τέλος, το πλάτος των διαγραμμάτων των ομάδωνπρέπει να είναι όσο το δυνατόν πιο ομοιόμορφο.

Στην περίπτωσή μας η τιμή 0,7 δεν είναι καλή καθώς, το γράφημα της μαύρης ομάδας έχει μεγάλο τμήμα του κάτω από τη μέση τιμή, και επίσης τα γραφήματα δεν είναι ομοιόμορφα.

Κεφάλαιο 5: Ταξινόμηση σταθμών μέτρησης ατμοσφαιρικής ρύπανσης

5.1 Ατμόσφαιρα – ατμοσφαιρική ρύπανση

«**Ατμόσφαιρα**» καλείται το αεριώδες τμήμα που μπορεί να περιβάλλει ένα ουράνιο σώμα και το οποίο ακολουθεί το σύνολο των κινήσεών του. Στην περίπτωση της γης φτάνει σε ύψος περίπου 3.500 χλμ.. Η ατμόσφαιρα αποτελεί: (α) την πηγή του οξυγόνου για την αναπνοή (προστατευτικό σώμα το οποίο συντηρεί τη ζωή στη γη) (β) την πηγή του διοξειδίου του άνθρακα για την φωτοσύνθεση (γ) προμηθεύει με άζωτο τους οργανισμούς για την σύνθεση των δομικών τους μορίων (δ) είναι βασικό τμήμα του υδρολογικού κύκλου μεταφέροντας νερό από τους ωκεανούς στα εδάφη και στους χερσαίους ταμιευτήρες νερού.

Ακόμα, η ατμόσφαιρα:

- κρατάει σταθερή τη θερμοκρασία της γης.
- απορροφά μεγάλο μέρος της κοσμικής ακτινοβολίας και υπεριώδους ακτινοβολίας, προστατεύοντας τους οργανισμούς από τις επιπτώσεις της.
- απορροφά μεγάλο μέρος της ηλεκτρομαγνητικής ακτινοβολίας (εκπέμπεται από τον ήλιο), επιτρέποντας να φτάσουν σημαντικές ποσότητες στη γη.
- επαναπορροφά μεγάλο μέρος της υπέρυθρης ακτινοβολίας που εκπέμπεται από την επιφάνεια της γης.

Οι λειτουργίες αυτές συνδέονται με τις αρνητικές επιπτώσεις της ρύπανσής της

Ατμοσφαιρική ρύπανση είναι η παρουσία στην ατμόσφαιρα κάθε είδους ουσιών που μπορούν να προκαλέσουν αλλοίωση των χαρακτηριστικών της (δομή, σύστασης κ.λ.π.). Οι αλλαγές αυτές μπορούν να έχουν αρνητικές επιπτώσεις στην υγεία, στους ζωντανούς οργανισμούς και στα οικοσυστήματα. Κάτω από ορισμένες συνθήκες είναι πιθανό να φτάσει σε επίπεδα που μπορεί να δημιουργήσουν ανεπιθύμητες συνθήκες διαβίωσης («νέφος»).

Ατμοσφαιρικός ρύπος θεωρείται κάθε ουσία (σε στερεή, υγρή ή αέρια μορφή) η οποία εισέρχεται, από τον άνθρωπο, στον αέρα του περιβάλλοντος άμεσα ή έμμεσα. Οι ουσίες μπορεί να είναι απόρροια της ανθρώπινης δραστηριότητας ή αποτέλεσμα της αλληλεπίδρασής του με το οικοσύστημα και μπορούν να προκαλέσουν επιπτώσεις στην άνεση, στην ευεξία και στην υγεία του ανθρώπου, καθώς και όλων των έμβιων οργανισμών. Η παρουσία τέτοιου είδους ουσιών στην ατμόσφαιρα μπορεί να προκαλέσει την ατμοσφαιρική ρύπανση.

Οι κυριότεροι (**συμβατικοί**) ρύποι, συνηθέστερα μετρούμενοι, που καθορίζουν τη ρύπανση της ατμόσφαιρας (ποιότητα αέρα) είναι οι εξής:

- Μονοξείδιο του Άνθρακα (CO)
- Διοξείδιο του Θείου (SO_2)
- Οξείδια του Αζώτου (NO_x)
- Όζον (O_3)
- Αιωρούμενα Σωματίδια (PM)
- Υδρογονάνθρακες (HC) και τα παράγωγά τους
- Βαρέα μέταλλα

Οι ρύποι αυτοί χωρίζονται σε δύο κατηγορίες: (α) στους πρωτογενείς (SO_2, NO, CO, HC, Pb, Cl_2, F_2, αιωρούμενα σωματίδια) που εκπέμπονται απευθείας από μία αναγνωρισμένη πηγή (βιομηχανική δραστηριότητα, εργοστάσια παραγωγής ηλεκτρικής ενέργειας, θέρμανση κατοικιών) (β) δευτερογενείς (NO_2, O_3, PAN) προέρχονται από διάφορες χημικές μεταβολές στα μόρια των ρύπων μέσω αντιδράσεων.

Το μεγαλύτερο ποσοστό των παραγόμενων αέριων ρύπων προέρχεται από καθαρά φυσικές πηγές (πηγές που δεν οφείλονται στην ανθρώπινη δραστηριότητα όπως: ηφαίστεια, πυρκαγιές δασών, ωκεανοί, αποσάθρωση του εδάφους κ.α.)

Οι ανθρωπογενείς εκπομπές είναι κυρίως υπεύθυνες για τα μεγαλύτερα περιβαλλοντικά προβλήματα γιατί οι εκπομπές αυτές συγκεντρώνονται σε μικρές γεωγραφικές περιοχές (κυρίως αστικές περιοχές και βιομηχανικές ζώνες), προκαλώντας την ένταση του φαινομένου της απόθεσης ρύπων στην ατμόσφαιρα με τις γνωστές συνέπειες.

Οι κυριότερες ανθρωπογενείς πηγές ρύπανσης είναι:

- Βιομηχανικές καύσεις (αιωρούμενα σωματίδια, διοξείδιο του θείου κ.ά.)
- Παραγωγή και μεταφορά ενέργειας (διοξείδιο του θείου, οξείδια του αζώτου, βενζόλιο κ.ά.)
- Μεταφορές (μονοξείδιο του άνθρακα, υδρογονάνθρακες, οξείδια αζώτου, βενζόλιο κ.ά.)
- Κεντρική θέρμανση (διοξείδιο του θείου, οξείδια του αζώτου κ.ά.)
- Εναπόθεση και αποτέφρωση στερεών αποβλήτων (μονοξείδιο του άνθρακα).

Το μεγαλύτερο μέρος της ατμοσφαιρικής ρύπανσης (τις τελευταίες δεκαετίες) δημιουργείται από διαδικασίες καύσης υλικών που ονομάζονται **καύσιμα** (λιθάνθρακας, λιγνίτης, τύρφη, κοκ, ξυλάνθρακες, ξύλα, απορρίμματα - βενζίνη, πετρέλαιο, κηροζίνη -- φυσικό αέριο, υδρογόνο κ.ά.). Σχεδόν όλα τα συμβατικά καύσιμα αποτελούνται κυρίως από χημικές ενώσεις δύο στοιχείων, του άνθρακα και του υδρογόνου (υδρογονάνθρακες). Για την καύση του, χρησιμοποιείται αέρας και κατά τη διάρκειά της εκτός από τη θερμότητα δημιουργείται και μία σειρά ρύπων που καταλήγουν στην ατμόσφαιρα.

5.2 Επιπτώσεις ατμοσφαιρικής ρύπανσης

Οι επιπτώσεις παίζουν καθοριστικό ρόλο στην υποβάθμιση των οικοσυστημάτων και της ποιότητας ζωής των ανθρώπων. Ανάλογα με το είδους του ρύπου που εισέρχεται στην ατμόσφαιρα προκαλούνται και οι αντίστοιχες επιδράσεις. Οι κυριότερες συνέπειες της ατμοσφαιρικής ρύπανσης είναι συνοπτικά οι εξής:

- Μείωση της ορατότητας
- Αύξηση της συχνότητας σχηματισμού ομίχλης
- Μείωση της άμεσης ηλιακής ακτινοβολίας
- Μεταβολή του μικροκλίματος
- Συνέπειες παγκοσμίου κλίμακας
- φθορές και στα υλικά (διάβρωση και καταστροφή μνημείων και κτιρίων)
- σημαντικές επιπτώσεις στη χλωρίδα και την πανίδα.

Σε μεγάλες συγκεντρώσεις:

- Το **μονοξείδιο του άνθρακα** μπορεί να προκαλέσει μείωση των αντανακλαστικών, πονοκεφάλους, απώλεια αισθήσεων, ακόμη και θάνατο (μειώνει την ικανότητα του αίματος να μεταφέρει οξυγόνο σε βασικούς ιστούς του οργανισμού, επιδρώντας κυρίως στο καρδιοαγγειακό και νευρικό σύστημα).
- το **διοξείδιο του θείου** επιφέρουν αυξημένη συχνότητα ασθενειών του αναπνευστικού συστήματος και υψηλότερη θνησιμότητα.
- **το διοξείδιο του αζώτου** μπορεί να προκαλέσει αναπνευστικές ασθένειες στα παιδιά και δυσκολία στην αναπνοή στους ασθματικούς.

- **το όζον** (εξαιρετικά τοξικό) μπορεί να προκαλέσει ζάλη, εμετούς, ερεθισμό στην αναπνευστική οδό, διαταραχή της αναπνευστικής λειτουργίας, άσθμα, φλεγμονή στους πνεύμονες, πιθανή επιδεκτικότητα σε μολύνσεις του αναπνευστικού, ερεθισμό των ματιών κ.λπ.
- τα **αιωρούμενα σωματίδια** (ανάλογα με τη χημική σύσταση και το μέγεθός τους), μπορούν να εισχωρήσουν στο αναπνευστικό σύστημα και να προκαλέσουν διάφορες ασθένειες καθώς και προβλήματα στην αναπνοή.
- Κάποιες ενώσεις υδρογονανθράκων έχουν καρκινογόνο δράση.

5.3. Κυριότεροι ρύποι

(α) Μονοξείδιο του άνθρακα

Το μονοξείδιο του άνθρακα (CO) είναι αέριο άοσμο, άγευστο και άχρωμο παράγεται από την ατελή καύση του άνθρακα και είναι πολύ τοξικό. Το CO οξειδώνεται από την ελεύθερη ρίζα του OH σε CO_2 έχοντας χρόνος ζωής 2-4 μήνες και παίζει ρόλο στη χημεία του όζοντος.

Κυριότερες πηγές προέλευσης του CO είναι οι εξατμίσεις των μηχανών των βενζινοκίνητων αυτοκινήτων, καθώς και οι εξατμίσεις πάσης φύσεως μηχανών όταν συντελείται ατελής καύση της καύσιμης ύλης. Άλλες ανθρωπογενείς πηγές είναι η απόθεση στερεών αποβλήτων, η παραγωγή σιδήρου κ.λπ. Η σημαντικότερη φυσική πηγή του CO στην ατμόσφαιρα είναι η οξείδωση του ατμοσφαιρικού μεθανίου.

(β) Διοξείδιο του θείου (SO_2)

Το διοξείδιο του θείου (SO_2) είναι ένα αέριο άχρωμο, άοσμο σε χαμηλές συγκεντρώσεις, με έντονη ερεθιστική μυρωδιά σε πολύ υψηλές συγκεντρώσεις. Στην ατμόσφαιρα αντιδρά για να σχηματίσει τριοξείδιο του θείου (SO_3). Πρόκειται για έναν αρκετά τοξικό ρύπο για το φυσικό περιβάλλον.

Φυσικές πηγές των θειούχων ενώσεων αποτελούν οι κοιλότητες συγκέντρωσης βιολογικής ύλης, η αναερόβια σήψη, η διάχυση σταγονιδίων από τη θάλασσα, οι ηφαιστειακές εκρήξεις, και οι θερμές πηγές. Ανθρωπογενείς πηγές προέλευσης είναι τα εργοστάσια παραγωγής ενέργειας, οι βιομηχανίες, τα διυλιστήρια πετρελαίου, οι κεντρικές θερμάνσεις, οι χημικές βιομηχανίες, καθώς και τα πετρελαιοκίνητα αυτοκίνητα που χρησιμοποιούν καύσιμο με ψηλή περιεκτικότητα σε θείο.

Το SO_2 σε αστικές περιοχές προκαλεί βλάβες στις επιφάνειες των κτιρίων, ενώ η μετατροπή του σε θειικό οξύ και η εναπόθεσή του στο έδαφος μέσω της όξινης βροχής καταστρέφει δασικές εκτάσεις, προκαλώντας την αύξηση της οξύτητας των λιμνών και των ποταμών. Επίσης, είναι ο κύριος υπεύθυνος για τη διάβρωση μετάλλων, την υποβάθμιση προστατευτικών επιστρωμάτων, τη φθορά οικοδομικών υλικών, καθώς επίσης και για την υποβάθμιση τους ποιότητας του χαρτιού, των δερμάτινων ειδών και των έργων και μνημείων ιστορικού ενδιαφέροντος.

(γ) Οξείδια του αζώτου (NO_x)

Τα πιο σημαντικά οξείδια του αζώτου που εμπλέκονται στη ρύπανση του αέρα είναι το μονοξείδιο του αζώτου (NO) (πρωτογενής) και το διοξείδιο του αζώτου (NO_2) (δευτερογενής αντίδραση με όζον). Το NO είναι αέριο άχρωμο, άοσμο, άγευστο και μη τοξικό, ενώ το NO_2 είναι αέριο με κιτρινωπό-καφέ χρώμα, διαλυτό στο νερό, ισχυρό οξειδωτικό, με οξεία ερεθιστική οσμή. Το NO_2 σε υψηλές συγκεντρώσεις δίνει το χαρακτηριστικό χρώμα του στην όψη του ουρανού στις αστικές περιοχές. Τα οξείδια του αζώτου NO και NO_2 παίζουν καθοριστικό ρόλο στον έλεγχο του όζοντος.

Οι κυριότερες πηγές NO_x είναι οι καύσεις ορυκτών καυσίμων σε εγκαταστάσεις παραγωγής ηλεκτρικής ενέργειας και εργοστάσια, καθώς και τα μεταφορικά μέσα. Η κύρια πηγή NO_2 είναι η οξείδωση του NO, ενώ πηγές NO_2 εσωτερικών χώρων αποτελούν οι συσκευές που λειτουργούν με αέριο, οι θερμάστρες κηροζίνης, οι ξυλόσομπες και το τσιγάρο. Οι ποσότητες NO_2 που εκλύονται πρωτογενώς στην ατμόσφαιρα είναι περιορισμένες συγκριτικά με αυτές του NO. Το NO_2 θεωρείται πιο σημαντικός ρύπος αναφορικά με την επίδραση στον άνθρωπο.

Το NO_2 σε συνδυασμό με άλλους παράγοντες συμβάλλει στη δημιουργία του φωτοχημικού νέφους, Επίσης τα NO_x θεωρούνται από τους πιο σημαντικούς ρύπους καθώς καταστρέφουν το στρώμα του όζοντος και συμμετέχουν στο σχηματισμό όξινης βροχής, ενώ συμβάλλουν και στην έξαρση του φαινομένου του θερμοκηπίου.

(δ) Όζον

Το όζον (O_3) είναι αέριο άχρωμο, βαρύτερο του αέρα με δριμεία οσμή. Πρόκειται για ένα αέριο στοιχείο που παράγεται στην στρατόσφαιρα (15-50 χλμ.) όπου και βρίσκεται περίπου το 90% του ολικού όζοντος της ατμόσφαιρας της γης. Το στρατοσφαιρικό όζον είναι το λεγόμενο «καλό» όζον, καθώς προστατεύει (σα φίλτρο) από την επιβλαβή υπεριώδη ηλιακή

ακτινοβολία. Η μείωση του όζοντος στην στρατόσφαιρα από τη χρήση ανθρωπογενών χημικών στοιχείων όπως οι χλωροφθοράνθρακες, αποτελεί τα τελευταία χρόνια ένα παγκόσμιο πρόβλημα.

Το υπόλοιπο 10% του όζοντος βρίσκεται στο χαμηλότερο στρώμα της ατμόσφαιρας, την τροπόσφαιρα (0-15 χλμ.). Χαμηλά στο έδαφος είναι ένας ρύπος που συνδέεται με το φωτοχημικό νέφος σε αστικά κέντρα («κακό» όζον). Στην τροπόσφαιρα το όζον είναι δευτερογενής ρύπος που σχηματίζεται ως αποτέλεσμα αλυσίδας χημικών αντιδράσεων μεταξύ του οξυγόνου, πτητικών οργανικών ενώσεων (VOCs) και οξειδίων του αζώτου (NO_x) σε συνθήκες έντονης ηλιακής ακτινοβολίας και υψηλών θερμοκρασιών. Πηγές εκπομπής των ρύπων που συντελούν στη δημιουργία του όζοντος είναι τα οχήματα, τα εργοστάσια, οι χωματερές, χημικά διαλυτικά και πολλές άλλες μικρές πηγές όπως βενζινάδικα, κ.λπ..

Το τροποσφαιρικό όζον έχει πολλαπλή σημασία για την ατμόσφαιρα της γης: (α) αποτελεί τη βασική πηγή του πιο σημαντικού οξειδωτικού μέσου, της ρίζας του υδροξυλίου (OH). Το OH αποτελεί το ισχυρότερο «απορρυπαντικό» της ατμόσφαιρας που την καθαρίζει από μια σειρά οργανικών/ανόργανων ενώσεων που εκπέμπονται από φυσικές ή ανθρωπογενείς πηγές. Εάν έλειπε η ρίζα του OH όλες αυτές οι ενώσεις θα είχαν μεγάλο χρόνο ζωής και θα συσσωρεύονταν στα ανώτερα στρώματα της τροπόσφαιρας, δρώντας επικουρικά στο γνωστό φαινόμενο του θερμοκηπίου (β) το όζον που βρίσκεται στα υψηλότερα στρώματα της τροπόσφαιρας είναι από μόνο του ένα θερμοκηπικό αέριο που σημαίνει ότι δρα και αυτό επικουρικά στο φαινόμενο του θερμοκηπίου, απορροφώντας τη γήινη υπέρυθρη ακτινοβολία. Το O_3 είναι επίσης ισχυρότατο οξειδωτικό μέσο π.χ. για το σχηματισμό της όξινης βροχής (οξειδώνει SO_2 προς H_2SO_4).

Οι ανεβασμένες συγκεντρώσεις του όζοντος σε περιοχές τεχνολογικά ανεπτυγμένες οφείλονται στις ανθρωπογενείς εκπομπές των NO_x και των υδρογονανθράκων (HCs). Οι συγκεντρώσεις του O_3 αντανακλούν μία αλληλεπίδραση των εκπομπών των NO_x, HCs, της μετεωρολογίας μεταφοράς και της ατμοσφαιρικής χημείας. Μετά την αύξηση των NO_x ακολουθεί αύξηση και του O_3 σαν αποτέλεσμα των φωτοχημικών αντιδράσεων που μείνουν τα επίπεδα των πρωτογενών ρύπων.

Ένας άλλος παράγοντας που επηρεάζει τη συγκέντρωση του τροποσφαιρικού όζοντος είναι η φωτοχημική δραστηριότητα. Σε επίπεδο ημερήσιας διακύμανσης η μέγιστη συγκέντρωση O_3 σημειώνεται νωρίς το απόγευμα, καθώς η ηλιακή ακτινοβολία στην επιφάνεια της γης είναι

μέγιστη κατά τις μεσημβρινές ώρες. Σε επίπεδο εποχιακής διακύμανσης, μεγάλες συγκεντρώσεις παρατηρούνται την άνοιξη και το καλοκαίρι και μικρότερες το χειμώνα.

Το όζον έχει αρνητικές επιπτώσεις στις αγροτικές καλλιέργειες, δασική και άλλη βλάστηση καθώς είναι φυτο-τοξικό στοιχείο (όταν βρίσκεται σε μεγάλες συγκεντρώσεις γίνεται επικίνδυνο για φυτά και δάση, επηρεάζοντας την ικανότητα τους να παράγουν και να αποθηκεύουν τροφή). Ως οξειδωτικό μέσο επιδρά και σε διάφορα οργανικά υλικά όπως οργανικά χρώματα που χρησιμοποιούνται στην ζωγραφική, στις εξωτερικές ζωγραφισμένες διακοσμήσεις κτιρίων, ή για βαφή υφασμάτων, όπως το χαρτί, διάφορα εκθέματα των μουσείων φυσικής ιστορίας όπως φτερά, δέρμα ζώων, πάπυρο κ.α.

(ε) Αιωρούμενα σωματίδια

Τα αιωρούμενα σωματίδια (PM) είναι μικρά τεμάχια ύλης σε στερεή ή υγρή μορφή, που βρίσκονται στην ατμόσφαιρα. Τα βασικά χαρακτηριστικά τους είναι το μέγεθος, η χημική σύσταση και η φάση στην οποία βρίσκονται (υγρή/αέρια). Οι κυριότερες πηγές εκπομπής τους είναι οι διάφορες βιομηχανικές δραστηριότητες, τα αυτοκίνητα, οι πυρκαγιές, γεωργικές δραστηριότητες, οι κατασκευές, η επαναιώρηση σκόνης λόγω ισχυρών ανέμων κ.λπ.

Η χημική σύσταση των σωματιδίων είναι συνάρτηση της περιοχής (αστική, υπαίθρια, θαλάσσια) από την οποία έχουν προέλθει. Ο χρόνος ζωής των σωματιδίων είναι αντιστρόφως ανάλογος με το μέγεθός τους. Τα μεγάλα σωματίδια έχουν χρόνο ζωής μερικές ώρες, ενώ τα μικρά σωματίδια μερικές ημέρες. Τα αιωρούμενα σωματίδια διακρίνονται σε διάφορες κατηγορίες: (α) σωματίδια με αεροδυναμική διάμετρο >10μm (χονδρόκκοκα), τα οποία δεν μπορούν να εισέλθουν στον ανθρώπινο οργανισμό, συνήθως μένουν στην ρινική κοιλότητα (β) σωματίδια με διάμετρο <10μm (PM_{10}). Στην κατηγορία αυτή ανήκουν και τα $PM_{2,5}$, σωματίδια με διάμετρο <2.5μm, τα οποία διακρίνονται σε αυτά με διάμετρο <0.1μm (ultrafine particles) και αυτά με διάμετρο >0.1μm.

5.4 Δίκτυο σταθμών μέτρησης ατμοσφαιρικής ρύπανσης

Σύμφωνα με την εθνική και κοινοτική νομοθεσία, η λειτουργία δικτύου σταθμών μέτρησης της ατμοσφαιρικής ρύπανσης αποτελεί υποχρέωση της χώρας. Το Εθνικό Δίκτυο Παρακολούθησης Ατμοσφαιρικής Ρύπανσης (ΕΔΠΑΡ) ξεκίνησε να λειτουργεί στα τέλη του 2000. Σήμερα (2017), αποτελείται από 34 σταθμούς, 15 από τους οποίους είναι

εγκατεστημένοι και λειτουργούν στην ευρύτερη περιοχή της Αθήνας, 7 στην περιοχή της Θεσσαλονίκης και 12 στην υπόλοιπη χώρα.

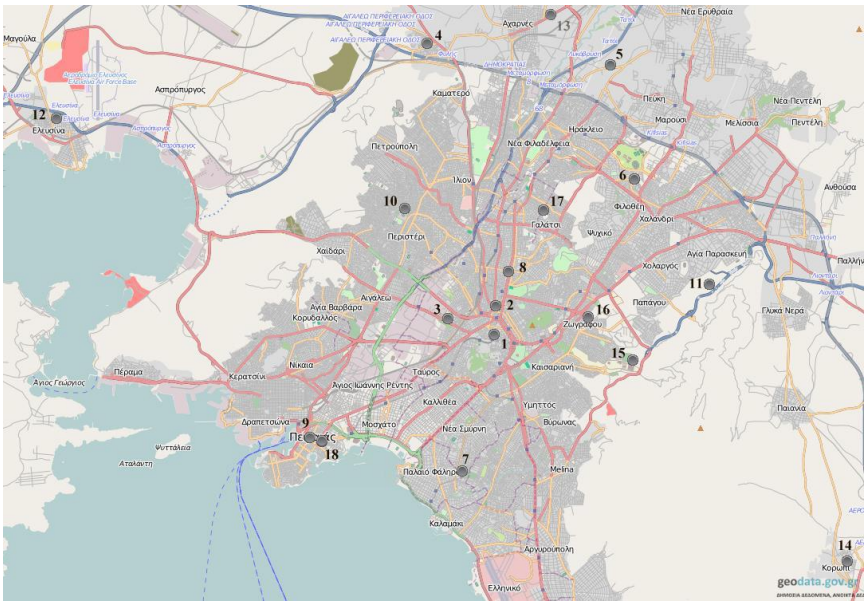
Η περιοχή της Αθήνας

Το πολεοδομικό συγκρότημα της Αθήνας περιλαμβάνει την περιοχή της Αθήνας, του Πειραιά και των προαστίων τους, με συνολικό πληθυσμό 3.181.872 κατοίκων (απογραφή 2011). Έχει έκταση 412 χλμ^2 και περιλαμβάνει συνολικά 40 Δήμους και πέντε Περιφερειακές Ενότητες (ΠΕ): Πειραιώς, Κεντρικού Τομέα Αθηνών, Βορείου Τομέα Αθηνών, Δυτικού Τομέα Αθηνών και Νοτίου Τομέα Αθηνών, που αποτελεί το κύριο μέρος του πολεοδομικού συγκροτήματος. Η Αθήνα και ο Πειραιάς αποτελούν τα δύο μεγάλα αστικά κέντρα της Αττικής.

Το συγκρότημα καλύπτει το Λεκανοπέδιο Αττικής, περιβάλλεται από τα τέσσερα ορεινούς όγκους: το όρος Αιγάλεω στα βορειοδυτικά (468m), την Πάρνηθα στα βόρεια (1453m), την Πεντέλη στα βορειοανατολικά (1109m) και τον Υμηττό στα ανατολικά (1026m). Τα ανοίγματα που υπάρχουν μεταξύ της Πεντέλης και της Πάρνηθας καθώς και του Υμηττού και της Πεντέλης βοηθούν στην απομάκρυνση των ρύπων από το λεκανοπέδιο Αττικής. Αντιθέτως, από το άνοιγμα ανάμεσα στο όρος Αιγάλεω και την Πάρνηθα επιτρέπεται η μεταφορά αέριων μαζών από το Θριάσιο Πεδίο, οι οποίες λόγω της βιομηχανικής δραστηριότητας που συντελείται στην περιοχή έχουν συχνά υψηλές συγκεντρώσεις στους διάφορους ρύπους.

Πηγές ατμοσφαιρικής ρύπανσης στο λεκανοπέδιο Αττικής αποτελούν η βιομηχανική ζώνη στα νοτιοδυτικά του Θριάσιου πεδίου, ο αερολιμένας στα ανατολικά, το λιμάνι του Πειραιά, καθώς και η ρύπανση που οφείλεται στην κυκλοφορία των οχημάτων και στην κεντρική θέρμανση των κτιρίων κατά τη διάρκεια των χειμερινών μηνών. Παρά το γεγονός ότι η χρήση του φυσικού αερίου έχει αυξηθεί τα τελευταία χρόνια, στις περισσότερες περιπτώσεις χρησιμοποιείται πετρέλαιο για την κεντρική θέρμανση.

Το 2016, η Δ/νση ΚΑΠΑ (Τμήμα Ποιότητας Ατμόσφαιρας), λειτούργησε 14 σταθμούς μέτρησης ατμοσφαιρικής ρύπανσης στην ευρύτερη περιοχή της Αθήνας, καθώς και 1 σταθμό στην Αλίαρτο Βοιωτίας για τις ανάγκες του Προγράμματος Διασυνοριακής Μεταφοράς της Ρύπανσης (ΕΜΕΡ).



Εικόνα 39: Χάρτης σταθμών μέτρησης ατμοσφαιρικής ρύπανσης του ΕΔΠΑΡ στην ευρύτερη περιοχή της Αθήνα Πηγή: ΥΠΕΝ

Οι σταθμοί κατηγοριοποιήθηκαν σε σταθμούς Κέντρου και Περιφέρειας, με βάση τη γεωγραφική τους θέση αλλά και την εκτίμηση των πηγών ρύπανσης που τους επηρεάζουν.

Πίνακας: Ταξινόμηση σταθμών μέτρησης ατμοσφαιρικής ρύπανσης

Σταθμός	Χαρακτηρισμός
Αθηνάς (ΑΘΗ)	Κέντρο
Αριστοτέλους (ΑΡΙ)	Κέντρο
Γεωπονική (ΓΕΩ)	Κέντρο
Λιόσια (ΛΙΟ)	Περιφέρεια
Λυκόβρυση (ΛΥΚ)	Περιφέρεια
Μαρούσι (ΜΑΡ)	Περιφέρεια
Νέα Σμύρνη (ΣΜΥ)	Κέντρο
Πατησίων (ΠΑΤ)	Κέντρο
Πειραιάς (ΠΕΙ-1)	Κέντρο
Περιστέρι (ΠΕΡ)	Κέντρο
Αγ. Παρασκευή (ΑΓ.ΠΑΡ)	Περιφέρεια
Ελευσίνα (ΕΛΕ)	Περιφέρεια
Θρακομακεδόνες (ΘΡΑ)	Περιφέρεια
Κορωπί (ΚΟΡ)	Περιφέρεια

5.5 Ταξινόμηση σταθμών μέτρησης ατμοσφαιρικής ρύπανσης (περιοχή Αθηνών)

Θα προσπαθήσουμε παρακάτω να ομαδοποιήσουμε τους σταθμούς μέτρησης των ρύπων με βάση τις μέσες τιμές/συγκεντρώσεις, κατά τη διάρκεια του 2016, μερικών από τους αέριους ρύπους.

Παράδειγμα 5.5.1

Οι μέσες συγκεντρώσεις $4^{ων}$ αερίων ρύπων για τους αντίστοιχους σταθμούς μέτρησης φαίνονται στον παρακάτω πίνακα.

Θα προσπαθήσουμε νε ομαδοποιήσουμε τους σταθμούς (αναφορικά) με τους 4 ρύπους

(α) με τη μέθοδο του κοντινότερου γείτονα και την ευκλείδεια απόσταση

(β) με τη με τη μέθοδο του Ward γείτονα και την ευκλείδεια απόσταση

(γ) με τη μέθοδο k-means

	NO_2	PM_{10}	O_3	NO
Λιόσια	20	34	67	9
Λυκόβρυση	20	29	55	9
Μαρούσι	27	32	62	15
Νέα Σμύρνη	31	30	65	12
Πειραιάς Ι.	64	43	24	54
Περιστέρι	29	35	65	10
Αγ. Παρασκευή	14	22	86	2
Ελευσίνα	29	31	61	21
Θρακομακεδόνες	8	21	91	3
Κορωπί	28	31	53	18

Λύση

Θα προσπαθήσουμε να ομαδοποιήσουμε τους σταθμούς (αναφορικά) με τους 4 ρύπους

(α) με τη μέθοδο του **κοντινότερου γείτονα και την ευκλείδεια απόσταση**

Θα χρησιμοποιήσουμε τις τέσσερις μεταβλητές και την ευκλείδεια απόσταση για να πάρουμε τον πίνακα τις απόστασης. Για λόγους ευκολίας θα ορίσουμε: Λιόσια=1, Λυκόβρυση=2, Μαρούσι=3, Νέα Σμύρνη=4, Πειραιάς Ι.=5, Περιστέρι=6, Αγ. Παρασκευή=7, Ελευσίνα=8, Θρακομακεδόνες=9, Κορωπί=10. Ο πίνακας αποστάσεων των αντικειμένων (περιοχών) είναι ίσος

με:

	1	2	3	4	5	6	7	8	9	10
1	0									
2	13.00	0								
3	10.67	11.95	0							
4	12.24	15.19	6.16	0						
5	76.75	71.54	66.74	68.57	0					
6	9.32	14.76	6.85	5.74	70.04	0				
7	24.29	33.09	31.84	29.90	97.41	29.98	0			
8	16.43	16.27	6.48	10.10	61.86	12.36	35.94	0		
9	30.41	39.24	38.30	36.97	103.48	36.90	7.93	42.01	0	
10	18.70	12.36	9.59	13.78	59.80	15.00	40.27	8.60	46.57	0

Η μικρότερη απόσταση μεταξύ των ζευγαριών είναι της Νέας Σμύρνης με το Περιστέρι δηλαδή:

$$\min_{i,j}(d_{i,j}) = d_{64} = 5.74$$

Άρα τα αντικείμενα 4 και 6 θα ενωθούν και θα σχηματίσουν μια ομάδα την οποία συμβολίζουμε με (46).

Στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της ομάδας (46) από τις υπόλοιπες χώρες δηλαδή τις 1, 2, 3, 5, 7, 8, 9, 10. Η απόσταση αυτή υπολογίζεται με το κριτήριο του κοντινότερου γείτονα:

$$d_{(64)1} = \min\{d_{41,}d_{61}\} = \min\{12.24, 9.32\} = 9.32$$
$$d_{(64)2} = \min\{d_{42,}d_{62}\} = \min\{15.19, 14.76\} = 14.76$$
$$d_{(64)3} = \min\{d_{43,}d_{63}\} = \min\{6.16, 6.85\} = 6.16$$
$$d_{(64)5} = \min\{d_{54,}d_{65}\} = \min\{68.57, 70.04\} = 68.57$$
$$d_{(64)7} = \min\{d_{74,}d_{76}\} = \min\{29.90, 29.98\} = 29.90$$
$$d_{(64)8} = \min\{d_{84,}d_{86}\} = \min\{10.10, 12.36\} = 10.10$$
$$d_{(64)9} = \min\{d_{94,}d_{96}\} = \min\{36.97, 36.90\} = 36.90$$
$$d_{(64)10} = \min\{d_{104,}d_{106}\} = \min\{13.78, 15.00\} = 13.78$$

Κατασκευάζουμε τώρα έναν νέο πίνακα αποστάσεων (θα διαγράψουμε από τον προηγούμενο πίνακα τις γραμμές 4, 6 και τις στήλες 4, 6 και θα προσθέσουμε μια γραμμή και μια στήλη για την ομάδα (46)). Ο νέος πίνακας αποστάσεων είναι ίσος με:

	1	2	3	(46)	5	7	8	9	10
1	0								
2	13	0							
3	10.67	11.95	0						
(46)	9.32	14.76	6.16	0					
5	76.75	71.54	66.74	68.57	0				
7	24.29	33.09	31.84	29.90	97.41	0			
8	16.43	16.27	6.48	10.10	61.86	35.94	0		
9	30.41	39.24	38.30	36.90	103.48	7.93	42.01	0	
10	18.70	12.36	9.59	13.78	59.80	40.27	8.60	46.57	0

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή ανάμεσα στην (Νέα Σμύρνη, Περιστέρι) με το Μαρούσι δηλαδή $\min_{i,j}(d_{ij}) = d_{(46)3} = 6.16$. Άρα οι ομάδες (46) και 3 θα θα ενωθούν και θα σχηματίσουν τη νέα ομάδα (346).

Συνεχίζοντας, στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της ομάδας (346) από τις υπόλοιπες δηλαδή τις 1, 2, 5, 7, 8, 9, 10. Υπολογίζονται ανάλογα, από τους τύπους:

$$d_{(346)1} = \min\{d_{31}, d_{41}, d_{61}\} = \min\{10.67, 12.24, 9.32\} = 9.32$$
$$d_{(346)2} = \min\{d_{32}, d_{42}, d_{62}\} = \min\{11.95, 15.19, 14.76\} = 11.95$$
$$d_{(346)5} = \min\{d_{53}, d_{54}, d_{65}\} = \min\{66.74, 68.57, 70.04\} = 66.74$$
$$d_{(346)7} = \min\{d_{73}, d_{74}, d_{76}\} = \min\{31.84, 29.90, 29.98\} = 29.90$$
$$d_{(346)8} = \min\{d_{83}, d_{84}, d_{86}\} = \min\{6.48, 10.10, 12.36\} = 6.48$$
$$d_{(346)9} = \min\{d_{93}, d_{94}, d_{96}\} = \min\{38.30, 36.97, 36.90\} = 36.90$$
$$d_{(346)10} = \min\{d_{103}, d_{104}, d_{106}\} = \min\{9.59, 13.78, 15.00\} = 9.59$$

Τώρα, θα διαγράψουμε στον πίνακα αποστάσεων τις γραμμές (46), 3 και τις στήλες (46), 3 και θα προσθέσουμε μια γραμμή και μια στήλη για την ομάδα (346).

	1	2	(346)	5	7	8	9	10
1	0							
2	13.00	0						
(346)	9.32	11.95	0					
5	76.75	71.54	66.74	0				
7	24.29	33.09	29.90	97.41	0			
8	16.43	16.27	6.48	61.86	35.94	0		
9	30.41	39.24	36.90	103.48	7.93	42.01	0	
10	18.70	12.36	9.59	59.80	40.27	8.60	46.57	0

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή της ομάδας (Νέας Σμύρνη, Περιστέρι, Μαρούσι) με την Ελευσίνα δηλαδή $\min_{i,j}(d_{ij}) = d_{(346)8} = 6.48$ Άρα τα αντικείμενα (346) και 8 θα σχηματίσουν την ομάδα (3468).

Συνεχίζοντας, στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της ομάδας (3468) από τις υπόλοιπες δηλαδή τις 1, 2, 5, 7, 9, 10. Υπολογίζονται ανάλογα:

$$d_{(3468)1} = \min\{d_{31}, d_{41}, d_{61}, d_{81}\} = \min\{10.67, 12.24, 9.32, 16.43\} = 9.32$$
$$d_{(3468)2} = \min\{d_{32}, d_{42}, d_{62}, d_{82}\} = \min\{11.95, 15.19, 14.76, 16.27\} = 11.95$$
$$d_{(3468)5} = \min\{d_{53}, d_{54}, d_{65}, d_{85}\} = \min\{66.74, 68.57, 70.04, 61.86\} = 61.86$$
$$d_{(3468)7} = \min\{d_{73}, d_{74}, d_{76}, d_{87}\} = \min\{31.84, 29.90, 29.98, 35.94\} = 29.90$$
$$d_{(3468)9} = \min\{d_{93}, d_{94}, d_{96}, d_{98}\} = \min\{38.30, 36.97, 36.90, 42.01\} = 36.90$$
$$d_{(3468)10} = \min\{d_{103}, d_{104}, d_{106}, d_{108}\} = \min\{9.59, 13.78, 15.00, 8.60\} = 8.60$$

Θα διαγράψουμε στον πίνακα αποστάσεων τις γραμμές (346), 8 και τις στήλες (346), 8 και θα προσθέσουμε μια γραμμή και μια στήλη για την ομάδα (3468).

	1	2	(3468)	5	7	9	10
1	0						
2	13.00	0					
(3468)	9.32	11.95	0				
5	76.75	71.54	61.86	0			
7	24.29	33.09	29.90	97.41	0		
9	30.41	39.24	36.90	103.48	7.93	0	
10	18.70	12.36	8.60	59.80	40.27	46.57	0

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή της Αγ. Παρασκευής με τους Θρακομακεδόνες δηλαδή $\min_{i,j}(d_{ij}) = d_{79} = 7.93$ Άρα τα αντικείμενα 7 και 9 θα σχηματίσουν την ομάδα (79).

Συνεχίζοντας, στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της ομάδας (79) από τις υπόλοιπες δηλαδή τις 1, 2, (3468), 5, 10. Υπολογίζονται ανάλογα:

$$d_{(79)1} = \min\{d_{71}, d_{91}\} = \min\{24.29, 30.41\} = 24.29$$
$$d_{(79)2} = \min\{d_{72}, d_{92}\} = \min\{33.09, 39.24\} = 33.09$$
$$d_{(79)(3468)} = \min\{d_{73}, d_{74}, d_{76}, d_{87}, d_{93}, d_{94}, d_{96}, d_{98}\} = \min\{31.84, 29.90, 29.98, 35.94,$$
$$38.30, 36.97, 36.90, 42.01\} = 29.90$$
$$d_{(79)5} = \min\{d_{75}, d_{95}\} = \min\{97.41, 103.48\} = 97.41$$
$$d_{(79)10} = \min\{d_{107}, d_{109}\} = \min\{40.27, 46.57\} = 40.27$$

Τώρα, θα διαγράψουμε στον πίνακα αποστάσεων τις γραμμές 7, 9 και τις στήλες 7, 9 και θα προσθέσουμε μια γραμμή και μια στήλη για την ομάδα (79).

	1	2	(3468)	5	(79)	10
1	0					
2	13.00	0				
(3468)	9.32	11.95	0			
5	76.75	71.54	61.86	0		
(79)	24.29	33.09	29.90	97.41	0	
10	18.70	12.36	8.60	59.80	40.270	0

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή της (Νέας Σμύρνη, Περιστέρι, Μαρούσι, Ελευσίνα) με το Κορωπί δηλαδή $\min_{i,j}(d_{ij}) = d_{10(3468)} = 8.60$ Άρα τα αντικείμενα 10 και (3468) θα σχηματίσουν την ομάδα (346810).

Στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της ομάδας (346810) από τις υπόλοιπες δηλαδή τις 1, 2, 5, (79). Υπολογίζονται ανάλογα, από τους τύπους:

$$d_{(346810)1} = \min\{d_{31}, d_{41}, d_{61}, d_{81}, d_{101}\} = \min\{10.67, 12.24, 9.32, 16.43, 18.70\} = 9.32$$
$$d_{(346810)2} = \min\{d_{32}, d_{42}, d_{62}, d_{82}, d_{102}\} = \min\{11.95, 15.19, 14.76, 16.27, 12.36\} = 11.95$$
$$d_{(346810)5} = \min\{d_{53}, d_{54}, d_{65}, d_{85}, d_{105}\} = \min\{66.74, 68.57, 70.04, 61.86, 59.80\} = 59.80$$
$$d_{(346810)(79)} = \min\{d_{73}, d_{74}, d_{76}, d_{87}, d_{107}, d_{93}, d_{94}, d_{96}, d_{98}, d_{109}\} = \min\{31.84, 29.90, 29.98,$$
$$35.94, 40.27, 38.30, 36.97, 36.90, 42.01, 46.57\} = 29.90$$

Τώρα, θα διαγράψουμε στον πίνακα αποστάσεων τις γραμμές 10, (3468) και τις στήλες 10, (3468) και θα προσθέσουμε μια γραμμή και μια στήλη για την ομάδα (346810).

	1	2	(346810)	5	(79)
1	0				
2	13.00	0			
(346810)	9.32	11.95	0		
5	76.75	71.54	59.80	0	
(79)	24.29	33.09	29.90	97.41	0

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή της (Νέα Σμύρνη, Περιστέρι, Μαρούσι, Ελευσίνα, Κορωπί) με τα Λιόσια δηλαδή $\min_{i,j}(d_{ij}) = d_{1(346810)} = 9.32$ Άρα τα αντικείμενα 1 και (346810) θα σχηματίσουν την ομάδα (1346810).

Συνεχίζοντας, στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της ομάδας (1346810) από τις υπόλοιπες δηλαδή τις 2, 5, (79). Υπολογίζονται ανάλογα:

$d_{(1346810)2} = \min\{d_{21}, d_{32}, d_{42}, d_{62}, d_{82}, d_{102}\} = \min\{13.00, 11.95, 15.19, 14.76, 16.27, 12.36\} = 11.95$

$d_{(1346810)5} = \min\{d_{51}, d_{53}, d_{54}, d_{65}, d_{85}, d_{105}\} = \min\{76.75, 66.74, 68.57, 70.04, 61.86, 59.80\} = 59.80$

$d_{(1346810)(79)} = \min\{d_{71}, d_{73}, d_{74}, d_{76}, d_{87}, d_{107}, d_{91}, d_{93}, d_{94}, d_{96}, d_{98}, d_{109}\} = \min\{24.29, 31.84, 29.90, 29.98, 35.94, 40.27, 30.41, 38.30, 36.97, 36.90, 42.01, 46.57\} = 24.29$

Τώρα, θα διαγράψουμε στον πίνακα αποστάσεων τις γραμμές 1, (346810) και τις στήλες 1, (346810) και θα προσθέσουμε μια γραμμή και μια στήλη για την ομάδα (1346810).

	(1346810)	2	5	(79)
(1346810)	0			
2	11.95	0		
5	59.80	71.54	0	
(79)	24.29	33.09	97.41	0

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή της (Νέα Σμύρνη, Περιστέρι, Μαρούσι, Ελευσίνα, Κορωπί, Λιόσια) με τη Λυκόβρυση $\min_{i,j}(d_{ij}) = d_{2(1346810)} = 11.95$ Άρα τα αντικείμενα 2 και (1346810) θα σχηματίσουν την ομάδα (12346810).

Συνεχίζοντας, στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της ομάδας (12346810) από τις υπόλοιπες δηλαδή τις 5, (79). Υπολογίζονται ανάλογα, από τους τύπους:

$d_{(12346810)5} = \min\{d_{51}, d_{52}, d_{53}, d_{54}, d_{65}, d_{85}, d_{105}\} = \min\{76.75, 71.54, 66.74, 68.57, 70.04, 61.86, 59.80\} = 59.80$

$d_{(12346810)(79)} = \min\{d_{71}, d_{72}, d_{73}, d_{74}, d_{76}, d_{87}, d_{107}, d_{91}, d_{92}, d_{93}, d_{94}, d_{96}, d_{98}, d_{109}\} = \min\{24.29, 33.09, 31.84, 29.90, 29.98, 35.94, 40.27, 30.41, 39.24, 38.30, 36.97, 36.90, 42.01, 46.57\} = 24.29$

Τώρα, θα διαγράψουμε στον πίνακα αποστάσεων τις γραμμές 2, (1346810) και τις στήλες 2, (1346810) και θα προσθέσουμε μια γραμμή και μια στήλη για την ομάδα (12346810).

	(12346810)	5	(79)
(12346810)	0		
5	59.80	0	
(79)	24.29	97.41	0

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή της (Νέα Σμύρνη, Περιστέρι, Μαρούσι, Ελευσίνα, Κορωπί, Λιόσια, Λυκόβρυση) με την (Αγ. Παρασκευή, Θρακομακεδόνες) δηλαδή $\min_{i,j}(d_{ij}) = d_{(79)(12346810)} = 24.29$ Άρα τα αντικείμενα (12346810) και (79) θα σχηματίσουν την ομάδα (1234678910).

Τέλος, η ομάδα (1234678910) θα ενωθεί με την 5 για το σχηματισμό μιάς ενιαίας ομάδας, σε απόσταση:

$$d_{(1234678910)5} = \min\{d_{51}, d_{52}, d_{53}, d_{54}, d_{65}, d_{75}, d_{85}, d_{95}, d_{105}\} = \min\{76.75, 71.54, 66.74, 68.57, 70.04, 97.41, 61.86, 103.48, 59.80\} = 59.80$$

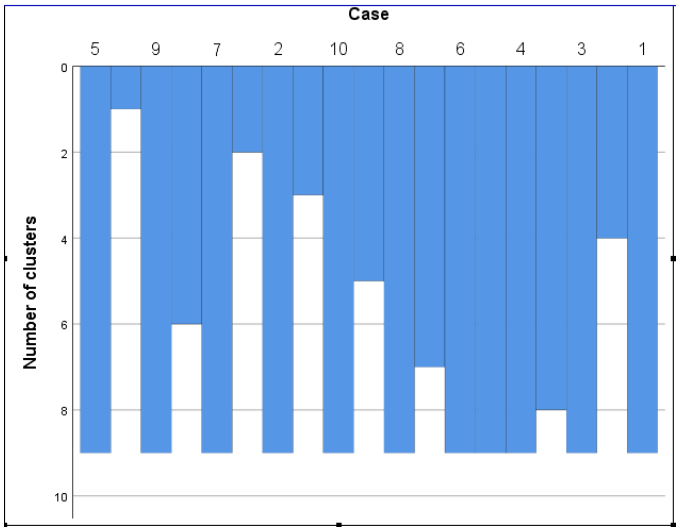
Το στατιστικό πρόγραμμα SPSS δίνει τ' αποτελέσματα:

(α) πίνακα που περιγράφει τα στάδια της διαδικασίας

Agglomeration Schedule

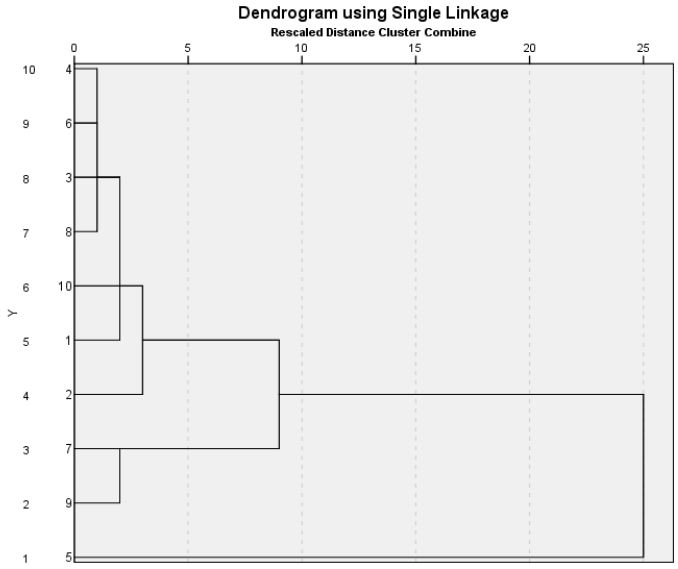
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	6	5,745	0	0	2
2	3	4	6,164	0	1	3
3	3	8	6,481	2	0	5
4	7	9	7,937	0	0	8
5	3	10	8,602	3	0	6
6	1	3	9,327	0	5	7
7	1	2	11,958	6	0	8
8	1	7	24,290	7	4	9
9	1	5	59,808	8	0	0

(β) Αναπαράσταση της διαδικασίας με τη βοήθεια ενός γραφήματος που λέγεται **icicle**



Εικόνα 40: διάγραμμα icicle Π 5.5.1 (Μέθοδος σύνδεσης του Ward)

(γ) Αναπαράσταση της διαδικασίας με τη βοήθεια του γραφήματος **δενδρόγραμμα**



Εικόνα 41: δενδρόγραμμα Π 5.5.1 (μέθοδος σύνδεσης του κοντινότερου γείτονα)

Παρατήρηση 5.5.2

Σύμφωνα λοιπόν με τη μέθοδο του κοντινότερου γείτονα είναι προτιμότερο να χωρίσουμε τις παρατηρήσεις σε τρεις ομάδες. Η μία ομάδα θα έχει τα στοιχεία (Αγ. Παρασκευή, Θρακομακεδόνες), η άλλη (Νέα Σμύρνη, Περιστέρι, Μαρούσι, Ελευσίνα, Κορωπί, Λιόσια, Λυκόβρυση) και η άλλη Πειραιάς Ι.

(β) Θα προσπαθήσουμε να ομαδοποιήσουμε τους σταθμούς (αναφορικά) με τους 4 ρύπους με τη μέθοδο του Ward και την ευκλείδεια απόσταση

Το στατιστικό πρόγραμμα SPSS δίνει τ' αποτελέσματα:

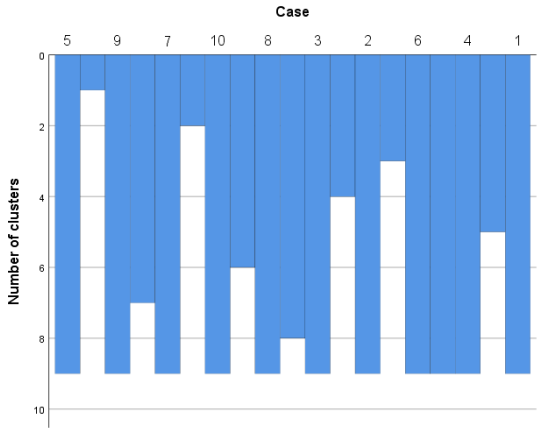
(α) πίνακα που περιγράφει τα στάδια της διαδικασίας

Ward Linkage

Agglomeration Schedule

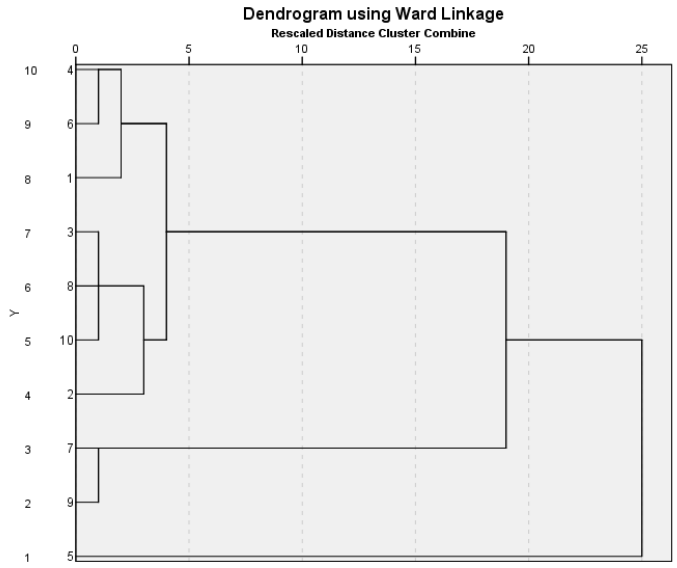
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	6	2,872	0	0	5
2	3	8	6,113	0	0	4
3	7	9	10,081	0	0	8
4	3	10	15,066	2	0	6
5	1	4	21,300	0	1	7
6	2	3	29,395	0	4	7
7	1	2	39,062	5	6	8
8	1	7	83,260	7	3	9
9	1	5	142,557	8	0	0

(β) Αναπαράσταση της διαδικασίας με τη βοήθεια του γραφήματος **icicle**



Εικόνα 42: διάγραμμα icicle Π 5.5.1 (Μέθοδος σύνδεσης του Ward)

(γ) Αναπαράσταση της διαδικασίας με τη βοήθεια του γραφήματος **δενδρόγραμμα**



Εικόνα 43: δενδρόγραμμα Π 5.5.1 (μέθοδος σύνδεσης του Ward)

Σύμφωνα λοιπόν και με τη μέθοδο ward είναι προτιμότερο να χωρίσουμε τις παρατηρήσεις σε τρεις ομάδες. Η μία ομάδα θα έχει τα στοιχεία (Αγ. Παρασκευή, Θρακομακεδόνες) , η άλλη τα στοιχεία (Νέας Σμύρνη, Περιστέρι, Μαρούσι, Ελευσίνα, Κορωπί, Λιόσια, Λυκόβρυση) και η άλλη Πειραιάς Ι.

(γ) Θα προσπαθήσουμε να ομαδοποιήσουμε με τη μέθοδο **k-means**

Για να εφαρμόσουμε τη μέθοδο **k-means** πρέπει να αποφασίσουμε πόσες ομάδες θα σχηματίσουμε. Από τη μέθοδο του κοντινότερου γείτονα, που εφαρμόσαμε παραπάνω,

βλέπουμε ότι είναι προτιμότερο να εφαρμόσουμε τη μέθοδο k- means για k=3. Με τη βοήθεια του προγράμματος SPSS έχουμε τα παρακάτω αποτελέσματα:

Initial Cluster Centers

	Cluster		
	1	2	3
NO2	8,00	28,00	64,00
PM10	21,00	31,00	43,00
O3	91,00	53,00	24,00
NO	3,00	18,00	54,00

Αυτά είναι τα αρχικά κέντρα των ομάδων (8.00, 21.00, 91.00, 3.00), (28.00, 31.00, 53.00, 18.00) και (64.00, 43.00, 24.00, 54.00)

Cluster Membership

Case Number	Cluster	Distance
1	2	9,932
2	2	10,209
3	2	1,948
4	2	6,487
5	3	,000
6	2	6,693
7	1	3,969
8	2	8,076
9	1	3,969
10	2	9,521

Βλέπουμε ότι σχηματίζονται τρεις ομάδες οι {7,9}, {1,2,3,4,6,8,10}, {5}

Final Cluster Centers

	Cluster		
	1	2	3
NO2	11,00	26,29	64,00
PM10	21,50	31,71	43,00
O3	88,50	61,14	24,00
NO	2,50	13,43	54,00

Οι συντεταγμένες των τελικών κέντρων της $1^{ης}$ της $2^{ης}$ και της της $3^{ης}$ ομάδας είναι (11.00, 21.50, 88.50, 2.50), (26.29, 31.71, 61.14, 13.43) και (64.00, 43.00, 24.00, 54.00) αντίστοιχα.

Από τον πίνακα βλέπουμε ότι η 1η ομάδα χαρακτηρίζεται από μεγάλο αριθμό συγκεντρώσεων O_3 (88,50), ενώ η 3η από μεγάλο αριθμό συγκεντρώσεων NO_2 (64.00), PM_{10} (43.00), NO (54.00) (οι διαφορές μεταξύ των ομάδων μπορεί να μην είναι στατιστικά σημαντικές, γεγονός που πρέπει να ελεγχθεί).

Distances between Final Cluster Centers

Cluster	1	2	3
1		34,725	100,418
2	34,725		67,641
3	100,418	67,641	

Η απόσταση ανάμεσα στα τελικά κέντρα της 1ης ομάδας από τη 2η είναι 34.72, της 1ης από τη 3η είναι 100.41 και της 2ης από τη 3η είναι 67.64.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
NO2	942,286	2	19,633	7	47,996	,000
PM10	163,836	2	3,990	7	41,064	,000
O3	1422,771	2	25,908	7	54,916	,000
NO	924,943	2	19,173	7	48,241	,000

Για να εξετάσουμε αν οι διαφορές συγκεντρώσεων, ανάμεσα στις ομάδες, καθενός από τους ρύπους είναι στατιστικά σημαντικές θα κάνουμε ανάλυση διασποράς για καθέναν από τους ρύπους. Τα αποτελέσματα φαίνονται παρακάτω.

NO2

Descriptives

NO2

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	2	11,00	4,243	3,000	-27,12	49,12	8	14
2	7	26,29	4,461	1,686	22,16	30,41	20	31
3	1	64,00	.	.	.	.	64	64
Total	10	27,00	14,989	4,740	16,28	37,72	8	64

Test of Homogeneity of Variances

NO2

Levene Statistic	df1	df2	Sig.
,131[a]	1	7	,728

a. Groups with only one case are ignored in computing the test of homogeneity of variance for NO2.

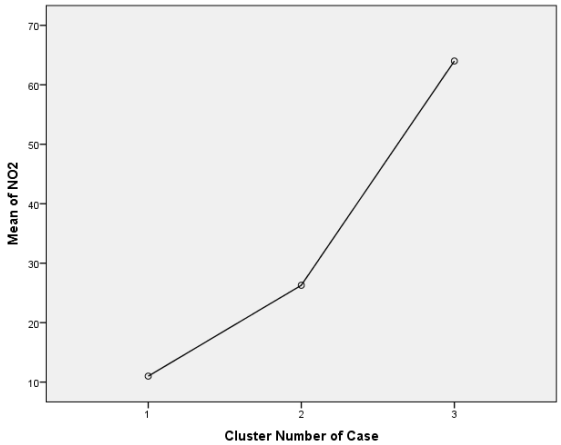
Ικανοποιείται η υπόθεση της ισότητας των διασπορών των συγκεντρώσεων του **NO2** ανάμεσα στις ομάδες.

ANOVA

NO2

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1884,571	2	942,286	47,996	,000
Within Groups	137,429	7	19,633		
Total	2022,000	9			

Η διαφορά των συγκεντρώσεων του NO2 είναι στατιστικά σημαντική sig 0,000 (οι ομάδες διαφέρουν ως προς τις συγκεντρώσεις του NO2, γεγονός που φαίνεται και στο παρακάτω διάγραμμα). Η 3^{η} ομάδα χαρακτηρίζεται από υψηλές συγκεντρώσεις NO2 ενώ η 1^{η} από χαμηλές



Εικόνα 44: μέσες συγκεντρώσεις NO2 (αναφορικά με την ομάδα)

PM10

Descriptives

PM10

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	2	21,50	,707	,500	15,15	27,85	21	22
2	7	31,71	2,138	,808	29,74	33,69	29	35
3	1	43,00	.	.	.	.	43	43
Total	10	30,80	6,286	1,988	26,30	35,30	21	43

Test of Homogeneity of Variances

PM10

Levene Statistic	df1	df2	Sig.
1,916[a]	1	7	,209

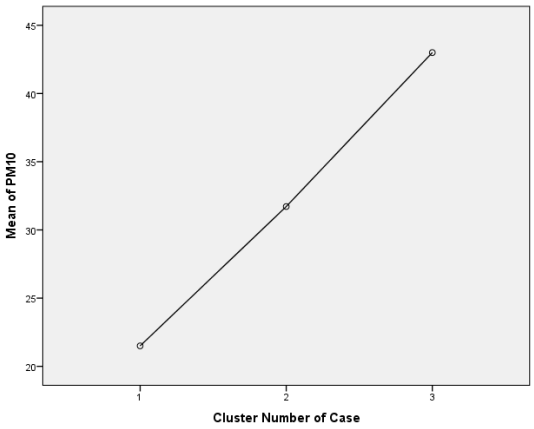
a. Groups with only one case are ignored in computing the test of homogeneity of variance for PM10.

ANOVA

PM10

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	327,671	2	163,836	41,064	,000
Within Groups	27,929	7	3,990		
Total	355,600	9			

Η διαφορά των συγκεντρώσεων του **PM10** είναι στατιστικά σημαντική sig 0,000 (οι ομάδες διαφέρουν ως προς τις συγκεντρώσεις του **PM10**, γεγονός που φαίνεται και στο παρακάτω διάγραμμα). Η 3η ομάδα χαρακτηρίζεται από υψηλές συγκεντρώσεις **PM10** ενώ η 1η από χαμηλές



Εικόνα 45: μέσες συγκεντρώσεις PM10 (αναφορικά με την ομάδα)

O3

Descriptives

O3

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	2	88,50	3,536	2,500	56,73	120,27	86	91
2	7	61,14	5,305	2,005	56,24	66,05	53	67
3	1	24,00	.	.	.	.	24	24
Total	10	62,90	18,339	5,799	49,78	76,02	24	91

Test of Homogeneity of Variances

O3

Levene Statistic	df1	df2	Sig.
,574[a]	1	7	,473

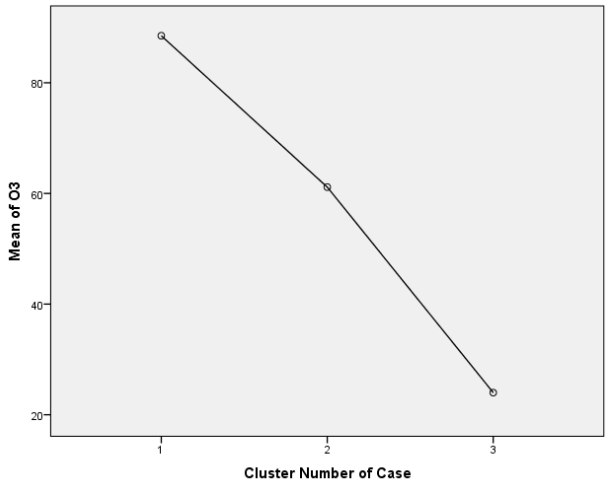
a. Groups with only one case are ignored in computing the test of homogeneity of variance for O3.

ANOVA

O3

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2845,543	2	1422,771	54,916	,000
Within Groups	181,357	7	25,908		
Total	3026,900	9			

Η διαφορά των συγκεντρώσεων του **O3** είναι στατιστικά σημαντική sig 0,000 (οι ομάδες διαφέρουν ως προς τις συγκεντρώσεις του **O3**, γεγονός που φαίνεται και στο). Η 1η ομάδα χαρακτηρίζεται από υψηλές συγκεντρώσεις **O3** ενώ η 3η από χαμηλές



Εικόνα 46: μέσες συγκεντρώσεις O3 (αναφορικά με την ομάδα)

NO

Descriptives

NO

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	2	2,50	,707	,500	-3,85	8,85	2	3
2	7	13,43	4,721	1,784	9,06	17,79	9	21
3	1	54,00	.	.	.	.	54	54
Total	10	15,30	14,848	4,695	4,68	25,92	2	54

Test of Homogeneity of Variances

NO

Levene Statistic	df1	df2	Sig.
4,849[a]	1	7	,064

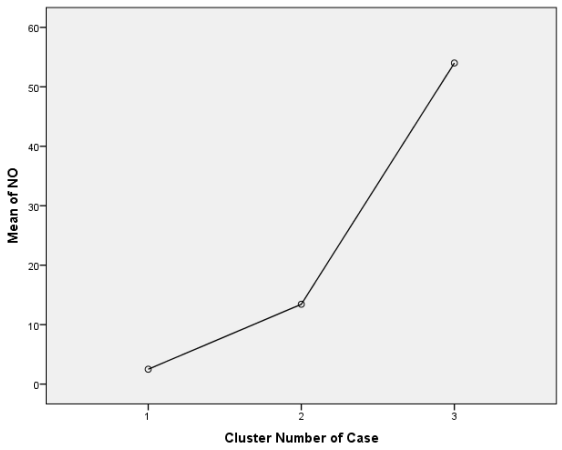
a. Groups with only one case are ignored in computing the test of homogeneity of variance for NO.

ANOVA

NO

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1849,886	2	924,943	48,241	,000
Within Groups	134,214	7	19,173		
Total	1984,100	9			

Η διαφορά των συγκεντρώσεων του **NO** είναι στατιστικά σημαντική sig 0,000 (οι ομάδες διαφέρουν ως προς τις συγκεντρώσεις του N**O**, γεγονός που φαίνεται και στο). Η 3^{η} ομάδα χαρακτηρίζεται από υψηλές συγκεντρώσεις N**O** ενώ η 1^{η} από χαμηλές



Εικόνα 47: μέσες συγκεντρώσεις NO (αναφορικά με την ομάδα)

Παρατήρηση 5.5.4

Από την παραπάνω ανάλυση βλέπουμε ότι, η $1^{η}$ ομάδα χαρακτηρίζεται από χαμηλές συγκεντρώσεις NO2, NO και σωματιδίων PM10 και υψηλές συγκεντρώσεις O3. Η $3^{η}$, αντίστροφα, χαρακτηρίζεται από υψηλές συγκεντρώσεις NO2, NO και σωματιδίων PM10 και χαμηλές συγκεντρώσεις O3. Η $2^{η}$ είναι κάτι ενδιάμεσα της $1^{ης}$ και $3^{ης}$.

Παράδειγμα 5.5.5

Οι μέσες συγκεντρώσεις $2^{ων}$ αερίων ρύπων για τους αντίστοιχους σταθμούς μέτρησης, κατά τη διάρκεια του 2016, φαίνονται στον παρακάτω πίνακα.

	$PM_{2,5}$	PM_{10}
Αριστοτέλους	20	41
Λυκόβρυση	17	29
Πειραιάς Ι.	20	43
Αγ. Παρασκευή	12	22
Ελευσίνα	21	31
Θρακομακεδόνες	13	21
Αλίαρτος	16	34

Θα προσπαθήσουμε νε ομαδοποιήσουμε τους σταθμούς (αναφορικά) με τους 2 ρύπους

(α) με τη μέθοδο του κοντινότερου γείτονα και την ευκλείδεια απόσταση

(β) με τη με τη μέθοδο του Ward γείτονα και την ευκλείδεια απόσταση

(γ) με τη μέθοδο k-means

Λύση

(α) Θα προσπαθήσουμε να ομαδοποιήσουμε τους σταθμούς (αναφορικά) με τους 2 ρύπους (α) με τη **μέθοδο του κοντινότερου γείτονα** και την **ευκλείδεια απόσταση**

Θα χρησιμοποιήσουμε τις δύο μεταβλητές και την ευκλείδεια απόσταση για να πάρουμε τον πίνακα τις απόστασης. Για λόγους ευκολίας θα ορίσουμε: Αριστοτέλους=1, Λυκόβρυση=2, Πειραιάς Ι.=3, Αγ. Παρασκευή=4, Ελευσίνα=5, Θρακομακεδόνες=6, Αλίαρτος=7. Ο πίνακας αποστάσεων των αντικειμένων (περιοχών) είναι ίσος με:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array}\begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 0 & & & & & & \\ 12.36 & 0 & & & & & \\ 2.00 & 14.31 & 0 & & & & \\ 20.61 & 8.60 & 22.47 & 0 & & & \\ 10.05 & 4.47 & 12.04 & 12.78 & 0 & & \\ 21.19 & 8.94 & 23.08 & 1.41 & 12.80 & 0 & \\ 8.06 & 5.09 & 9.84 & 12.64 & 5.83 & 13.34 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ζευγαριών είναι της Αγ. Παρασκευής με τους Θρακομακεδόνες δηλαδή:

$$\min_{i,j}(d_{i,j}) = d_{64} = 1.41$$

Άρα τα αντικείμενα 4 και 6 θα ενωθούν και θα σχηματίσουν μια ομάδα την οποία συμβολίζουμε με (46).

Στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της ομάδας (46) από τις υπόλοιπες χώρες δηλαδή τις 1, 2, 3, 5, 7. Η απόσταση αυτή υπολογίζεται με το κριτήριο του κοντινότερου γείτονα:

$$d_{(46)1} = \min\{d_{41,}d_{61}\} = \min\{20.61, 21.19\} = 20.61$$
$$d_{(46)2} = \min\{d_{42,}d_{62}\} = \min\{8.60, 8.94\} = 8.60$$
$$d_{(46)3} = \min\{d_{43,}d_{63}\} = \min\{22.47, 23.08\} = 22.47$$
$$d_{(46)5} = \min\{d_{54,}d_{65}\} = \min\{12.72, 12.80\} = 12.72$$
$$d_{(46)7} = \min\{d_{74,}d_{76}\} = \min\{12.64, 13.34\} = 12.64$$

Κατασκευάζουμε τώρα έναν νέο πίνακα αποστάσεων (θα διαγράψουμε από τον προηγούμενο πίνακα τις γραμμές 4, 6 και τις στήλες 4, 6 και θα προσθέσουμε μια γραμμή και μια στήλη για την ομάδα (46)). Ο νέος πίνακας αποστάσεων είναι ίσος με:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ (46) \\ 5 \\ 7 \end{array}\begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & (46) & 5 & 7 \end{array} \\ \begin{pmatrix} 0 & & & & & \\ 12.36 & 0 & & & & \\ 2.00 & 14.31 & 0 & & & \\ 20.61 & 8.60 & 22.47 & 0 & & \\ 10.05 & 4.47 & 12.04 & 12.72 & 0 & \\ 8.06 & 5.09 & 9.84 & 12.64 & 5.83 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή ανάμεσα στην Αριστοτέλους με τον Πειραιάς Ι, δηλαδή $\min_{i,j}(d_{ij}) = d_{31} = 2.00$. Άρα οι ομάδες 1 και 3 θα σχηματίσουν τη νέα ομάδα (13).

Συνεχίζοντας, στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της ομάδας (13) από τις υπόλοιπες δηλαδή τις 2, (46), 5, 7. Υπολογίζονται ανάλογα, από τους τύπους:

$$d_{(13)2} = \min\{d_{21}, d_{32}\} = \min\{12.36, 14.31\} = 12.36$$
$$d_{(13)(46)} = \min\{d_{41}, d_{61}, d_{43}, d_{63}\} = \min\{20.61, 21.19, 22.47, 23.08\} = 20.61$$
$$d_{(13)5} = \min\{d_{51}, d_{53}\} = \min\{10.05, 12.04\} = 10.05$$
$$d_{(13)7} = \min\{d_{71}, d_{73}\} = \min\{8.06, 9.84\} = 8.06$$

Θα διαγράψουμε στον πίνακα αποστάσεων τις γραμμές 1, 3 και τις στήλες 1, 3 και θα προσθέσουμε μια γραμμή και μια στήλη για την ομάδα (13).

	(13)	2	(46)	5	7
(13)	0				
2	12.36	0			
(46)	20.61	8.60	0		
5	10.05	4.47	12.72	0	
7	8.06	5.09	12.64	5.83	0

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή της Λυκόβρυση με την Ελευσίνα δηλαδή $\min_{i,j}(d_{ij}) = d_{52} = 4.47$ Άρα τα αντικείμενα 5 και 2 θα σχηματίσουν την ομάδα (52).

Στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της ομάδας (52) από τις υπόλοιπες δηλαδή τις (13), (46), 7. Υπολογίζονται ανάλογα, από τους τύπους:

$$d_{(52)(13)} = \min\{d_{51}, d_{53}, d_{21}, d_{32}\} = \min\{10.05, 12.04, 12.36, 14.31\} = 10.05$$
$$d_{(52)(46)} = \min\{d_{54}, d_{65}, d_{42}, d_{62}\} = \min\{12.72, 12.80, 8.60, 8.94\} = 8.60$$
$$d_{(52)7} = \min\{d_{75}, d_{72}\} = \min\{5.83, 5.09\} = 5.09$$

Τώρα, θα διαγράψουμε στον πίνακα αποστάσεων τις γραμμές 5, 2 και τις στήλες 5, 2 και θα προσθέσουμε μια γραμμή και μια στήλη για την ομάδα (25).

$$\begin{array}{c} \\ (13) \\ (25) \\ (46) \\ 7 \end{array} \begin{array}{c} \begin{array}{cccc} (13) & (25) & (46) & 7 \end{array} \\ \begin{pmatrix} 0 & & & \\ 10.05 & 0 & & \\ 20.61 & 8.60 & 0 & \\ 8.06 & 5.09 & 12.64 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή της ομάδας (Λυκόβρυση, Ελευσίνα) με τον Αλίαρτο δηλαδή $\min_{i,j}(d_{ij}) = d_{7(25)} = 5.09$ Άρα τα αντικείμενα/ομάδες 7 και (25) ενώνονται και θα σχηματίσουν την ομάδα (257).

Συνεχίζοντας, στο επόμενο επίπεδο ομαδοποίησης θα χρειαστούμε τις αποστάσεις της ομάδας (257) από τις υπόλοιπες δηλαδή τις (13), (46). Υπολογίζονται ανάλογα, από τους τύπους:

$$d_{(257)(13)} = \min\{d_{21}, d_{32}, d_{51}, d_{53}, d_{71}, d_{73}\} = \min\{12.36, 14.31, 10.05, 12.04, 8.06, 9.84\} = 8.06$$
$$d_{(257)(46)} = \min\{d_{42}, d_{62}, d_{54}, d_{65}, d_{74}, d_{76}\} = \min\{8.60, 8.94, 12.72, 12.80, 12.64, 13.34\} = 8.60$$

Τώρα, θα διαγράψουμε στον πίνακα αποστάσεων τις γραμμές (25), 7 και τις στήλες (25), 7 και θα προσθέσουμε μια γραμμή και μια στήλη για την ομάδα (257).

$$\begin{array}{c} \\ (13) \\ (257) \\ (46) \end{array} \begin{array}{c} \begin{array}{ccc} (13) & (257) & (46) \end{array} \\ \begin{pmatrix} 0 & & \\ 8.06 & 0 & \\ 20.61 & 8.60 & 0 \end{pmatrix} \end{array}$$

Η μικρότερη απόσταση μεταξύ των ομάδων είναι αυτή της (Λυκόβρυση, Ελευσίνα Αλίαρτο) με την (Αριστοτέλους, Πειραιάς Ι.) δηλαδή $\min_{i,j}(d_{ij}) = d_{(13)(257)} = 8.06$ Άρα τα αντικείμενα (13) και (257) θα σχηματίσουν την ομάδα (12357).

Τέλος, οι ομάδες (12357) και (46) θα ενωθούν για να σχηματίσουν μια ενιαία ομάδα, σε απόσταση:

$$d_{(12357)(46)} = \min\{d_{41}, d_{61}, d_{42}, d_{62}, d_{43}, d_{63}, d_{54}, d_{65}, d_{74}, d_{76}\} = \min\{20.61, 21.19,$$
$$8.60, 8.94, 22.47, 23.08, 12.72, 18.80, 12.64, 13.34\} = 8.60$$

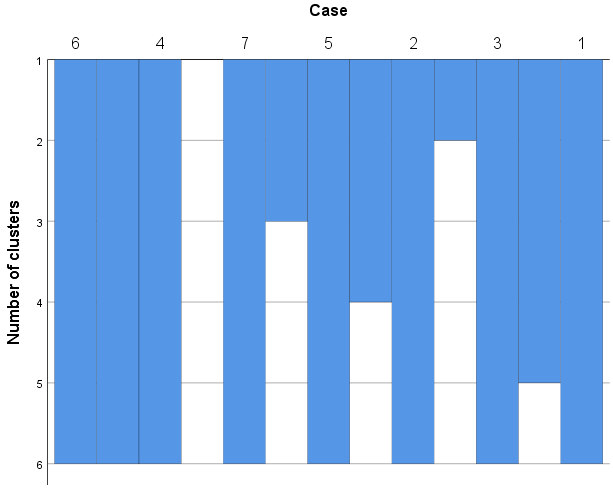
Το στατιστικό πρόγραμμα SPSS δίνει τ' αποτελέσματα:

(α) πίνακα που περιγράφει τα στάδια της διαδικασίας

Agglomeration Schedule

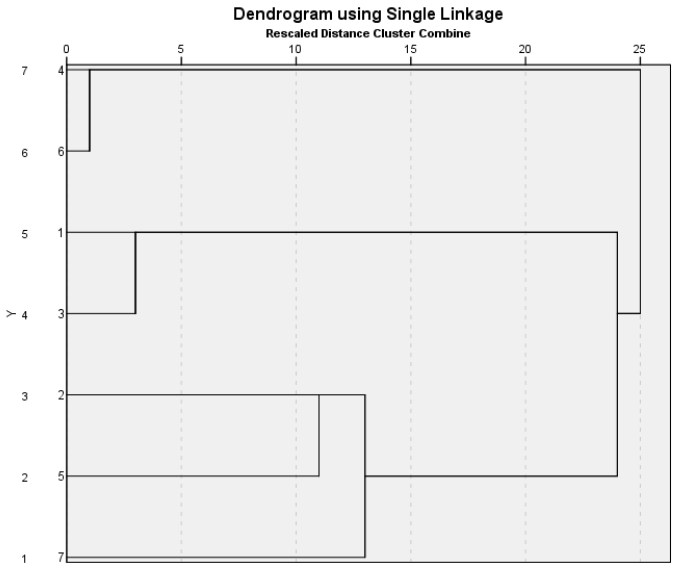
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	6	1,414	0	0	6
2	1	3	2,000	0	0	5
3	2	5	4,472	0	0	4
4	2	7	5,099	3	0	5
5	1	2	8,062	2	4	6
6	1	4	8,602	5	1	0

(β) Αναπαράσταση της διαδικασίας με τη βοήθεια του γραφήματος **icicle**



Εικόνα 48: διάγραμμα icicle Π 5.5.5 (Μέθοδος σύνδεσης του Ward)

(γ) Αναπαράσταση της διαδικασίας με τη βοήθεια του γραφήματος **δενδρόγραμμα**



Εικόνα 49: δενδρόγραμμα Π 5.5.5 (μέθοδος σύνδεσης του κοντινότερου γείτονα)

Παρατήρηση 5.5.6

Σύμφωνα λοιπόν με τη μέθοδο του κοντινότερου γείτονα είναι προτιμότερο να χωρίσουμε τις παρατηρήσεις σε τρεις ομάδες. Η μία ομάδα θα έχει τα στοιχεία (Αγ. Παρασκευή, Θρακομακεδόνες), η άλλη (Λυκόβρυση, Ελευσίνα, Αλίαρτος) και η τελευταία (Αριστοτέλους, Πειραιάς Ι).

(β) Θα προσπαθήσουμε να ομαδοποιήσουμε τους σταθμούς (αναφορικά) με τους 2 ρύπους με τη μέθοδο του **Ward** και την **ευκλείδεια απόσταση**

Το στατιστικό πρόγραμμα SPSS δίνει τ' αποτελέσματα:

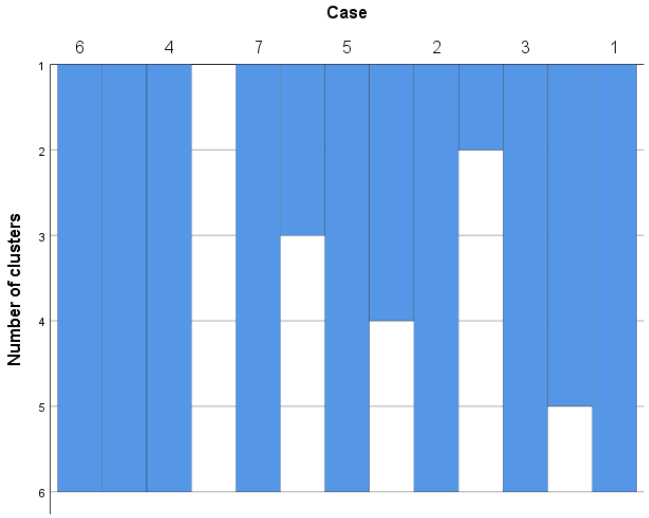
(Ι) πίνακα που περιγράφει τα στάδια της διαδικασίας

Ward Linkage

Agglomeration Schedule

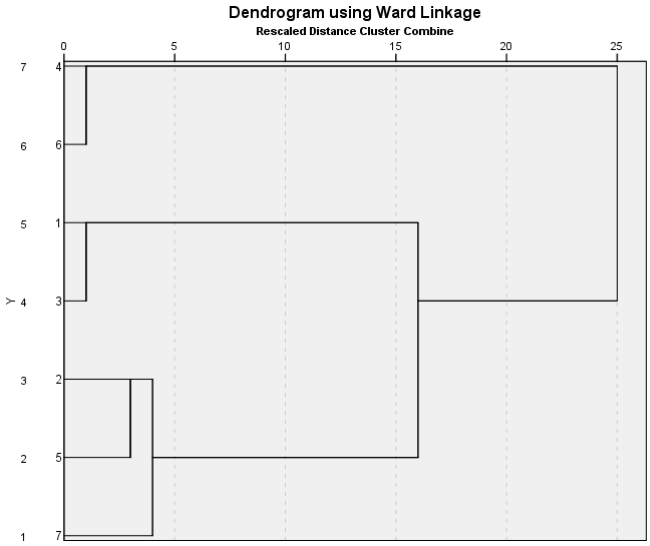
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	6	,707	0	0	6
2	1	3	1,707	0	0	5
3	2	5	3,943	0	0	4
4	2	7	6,841	3	0	5
5	1	2	17,525	2	4	6
6	1	4	34,563	5	1	0

(ΙΙ) Αναπαράσταση της διαδικασίας με τη βοήθεια του γραφήματος **icicle**



Εικόνα 50: διάγραμμα icicle Π. 5.5.5 (Μέθοδος σύνδεσης του Ward)

(γ) Αναπαράσταση της διαδικασίας με τη βοήθεια του γραφήματος **δενδρόγραμμα**



Εικόνα 51: δενδρόγραμμα 5.5.5 (μέθοδος σύνδεσης του Ward)

Παρατήρηση 5.5.7

Σύμφωνα με τη μέθοδο ward είναι προτιμότερο να χωρίσουμε τις παρατηρήσεις σε τρεις ομάδες. Η μία ομάδα θα έχει τα στοιχεία (Αγ. Παρασκευή, Θρακομακεδόνες) , η άλλη τα στοιχεία (Αριστοτέλους, Πειραιάς Ι.) και η άλλη (Λυκόβρυση, Ελευσίνα, Αλίαρτος). Καταλήγουμε δηλαδή στο ίδιο συμπέρασμα όπως αυτό προέκυψε στη μέθοδο του κοντινότερου γείτονα.

(γ) Θα προσπαθήσουμε να ομαδοποιήσουμε τους σταθμούς (αναφορικά) με τους 2 ρύπους με τη μέθοδο **k-means**

Για να εφαρμόσουμε τη μέθοδο k-means πρέπει να αποφασίσουμε πόσες ομάδες θα σχηματίσουμε. Από τη μέθοδο του κοντινότερου γείτονα, που εφαρμόσαμε παραπάνω, βλέπουμε ότι είναι προτιμότερο να εφαρμόσουμε τη μέθοδο k- means για k=3. Με τη βοήθεια του προγράμματος SPSS έχουμε τα παρακάτω αποτελέσματα:

Initial Cluster Centers

	Cluster		
	1	2	3
PM2.5	13,00	21,00	20,00
PM10	21,00	31,00	43,00

Αυτά είναι τα αρχικά κέντρα των ομάδων (13.00, 21.00), (21.00, 31.00) και (20.00, 43.00)

Cluster Membership

Case Number	Cluster	Distance
1	3	1,000
2	2	2,539
3	3	1,000
4	1	,707
5	2	3,018
6	1	,707
7	2	3,333

Βλέπουμε ότι σχηματίζονται τρεις ομάδες οι {1,3}, {2,5,7, {4,6}

Final Cluster Centers

	Cluster		
	1	2	3
PM2.5	12,50	18,00	20,00
PM10	21,50	31,33	42,00

Οι συντεταγμένες των τελικών κέντρων των ομάδων είναι (12.50, 21.50), (18.00, 31.33) και (20.00, 42.00) αντίστοιχα. Από τον πίνακα βλέπουμε ότι η 1η ομάδα χαρακτηρίζεται από χαμηλό αριθμό συγκεντρώσεων $PM_{2,5}$ (12.50) και PM_{10} (21.50) ενώ η 3η από μεγάλο αριθμό συγκεντρώσεων $PM_{2,5}$ (20.00) και PM_{10} (42.00) (οι διαφορές μεταξύ των ομάδων μπορεί να μην είναι στατιστικά σημαντικές, γεγονός που πρέπει να ελεγχθεί).

Distances between Final Cluster Centers

Cluster	1	2	3
1		11,267	21,829
2	11,267		10,853
3	21,829	10,853	

Η απόσταση ανάμεσα στα τελικά κέντρα της 1ης ομάδας από τη 2η είναι 11.26, της 1ης από τη 3η είναι 21.82 και της 2ης από τη 3η είναι 10.85

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
PM2.5	30,750	2	3,625	4	8,483	,036
PM10	210,274	2	3,792	4	55,457	,001

Για να εξετάσουμε αν οι διαφορές συγκεντρώσεων, ανάμεσα στις ομάδες, καθενός από τους δύο τύπους σωματιδίων είναι στατιστικά σημαντικές θα κάνουμε ανάλυση διασποράς για καθέναν από τους ρύπους. Τα αποτελέσματα φαίνονται παρακάτω.

PM2.5

Descriptives

PM2.5

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	2	12,50	,707	,500	6,15	18,85	12	13
2	3	18,00	2,646	1,528	11,43	24,57	16	21
3	2	20,00	,000	,000	20,00	20,00	20	20
Total	7	17,00	3,559	1,345	13,71	20,29	12	21

Test of Homogeneity of Variances

PM2.5

Levene Statistic	df1	df2	Sig.
5,500	2	4	,071

ANOVA

PM2.5

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	61,500	2	30,750	8,483	,036
Within Groups	14,500	4	3,625		
Total	76,000	6			

Η διαφορά των συγκεντρώσεων των PM2.5 είναι στατιστικά σημαντική sig 0,036 (οι ομάδες διαφέρουν ως προς τις συγκεντρώσεις των PM2.5, γεγονός που φαίνεται και στο παρακάτω διάγραμμα). Η $3^{η}$ ομάδα χαρακτηρίζεται από υψηλές συγκεντρώσεις PM2.5 ενώ η $1^{η}$ από χαμηλές (η $2^{η}$ ομάδα με την $3^{η}$ δεν διαφέρουν στατιστικά σημαντικά – δείτε τεστ πολλαπλών συγκρίσεων παρακάτω).

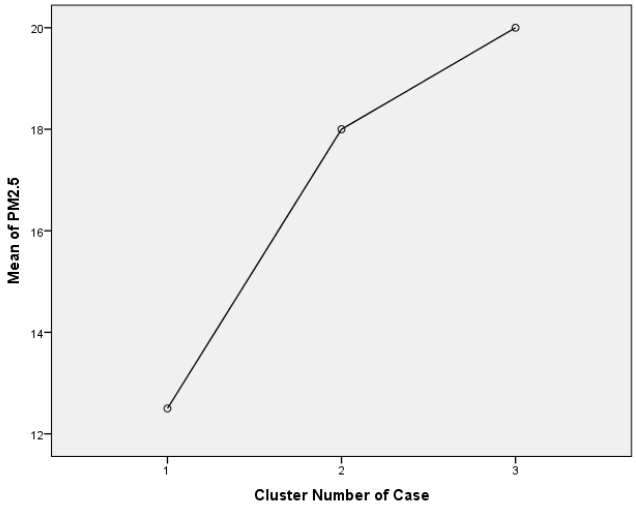
Multiple Comparisons

Dependent Variable: PM2.5

Tukey HSD

(I) Cluster Number of Case	(J) Cluster Number of Case	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-5,500	1,738	,072	-11,69	,69
	3	-7,500*	1,904	,036	-14,29	-,71
2	1	5,500	1,738	,072	-,69	11,69
	3	-2,000	1,738	,538	-8,19	4,19
3	1	7,500*	1,904	,036	,71	14,29
	2	2,000	1,738	,538	-4,19	8,19

*. The mean difference is significant at the 0.05 level.



Εικόνα 52: μέσες συγκεντρώσεις PM2.5 (αναφορικά με την ομάδα)

PM10

Descriptives

PM10

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	2	21,50	,707	,500	15,15	27,85	21	22
2	3	31,33	2,517	1,453	25,08	37,58	29	34
3	2	42,00	1,414	1,000	29,29	54,71	41	43
Total	7	31,57	8,522	3,221	23,69	39,45	21	43

PM10

Levene Statistic	df1	df2	Sig.
1,294	2	4	,369

ANOVA

PM10

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	420,548	2	210,274	55,457	,001
Within Groups	15,167	4	3,792		
Total	435,714	6			

Η διαφορά των συγκεντρώσεων των PM10 είναι στατιστικά σημαντική sig 0,001 (οι ομάδες διαφέρουν ως προς τις συγκεντρώσεις των PM10, γεγονός που φαίνεται και στο παρακάτω διάγραμμα). Η 3η ομάδα χαρακτηρίζεται από υψηλές συγκεντρώσεις PM10 ενώ η 1η από χαμηλές.

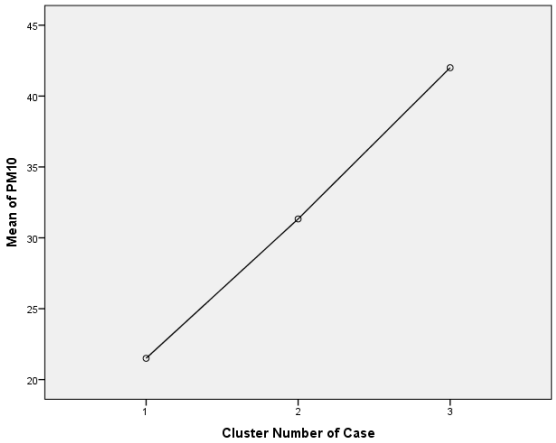
Multiple Comparisons

Dependent Variable: PM10

Tukey HSD

(I) Cluster Number of Case	(J) Cluster Number of Case	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-9,833*	1,778	,011	-16,17	-3,50
	3	-20,500*	1,947	,001	-27,44	-13,56
2	1	9,833*	1,778	,011	3,50	16,17
	3	-10,667*	1,778	,009	-17,00	-4,33
3	1	20,500*	1,947	,001	13,56	27,44
	2	10,667*	1,778	,009	4,33	17,00

*. The mean difference is significant at the 0.05 level.



Εικόνα 53: μέσες συγκεντρώσεις PM10 (αναφορικά με την ομάδα)

Φαίνεται ότι η διαφορά και στους δύο ρύπους, ανάμεσα στις ομάδες, να είναι στατιστικά σημαντική. Επομένως οι μέσες τιμές μεταξύ των ομάδων διαφέρουν σημαντικά και η $1^{η}$ ομάδα χαρακτηρίζεται από χαμηλό αριθμό συγκεντρώσεων $PM_{2,5}$ και PM_{10} ενώ η τρίτη από μεγάλο αριθμό συγκεντρώσεων $PM_{2,5}$ και PM_{10} .

Βιβλιογραφία

Ελληνική

- Αποστολάκης Ι., Καστανιά Α., Πιερράκου Χ., *Στατιστική επεξεργασία δεδομένων στην υγεία*, Παπαζήση 2003.
- Βαλαβανίδης, Α. (2007). *Οικοτοξικολογία και περιβαλλοντική τοξικολογία*. Εκδόσεις Τμήματος Χημείας Πανεπιστημίου Αθηνών, Αθήνα.
- Γεντετάκης, Ι. (2010). *Ατμοσφαιρική Ρύπανση: Επιπτώσεις, έλεγχος και εναλλακτικές τεχνολογίες*. 2η έκδοση, Αθήνα: Κλειδάριθμος.
- Καρλής Δ., Πολυμεταβλητή Στατιστική Ανάλυση, 2005, Σταμούλης.
- Λαζαρίδης, Μ. (2010). *Ατμοσφαιρική ρύπανση με στοιχεία μετεωρολογίας*. Εκδόσεις Τζιόλα, 2η έκδοση, Θεσσαλονίκη.
- Loren Larashi, *Στατιστική ανάλυση δεδομένων ατμοσφαιρικής ρύπανσης στην ευρύτερη περιοχή της Αθήνας*, Διπλωματική εργασία, Σχολή ΜΗ.ΠΕΡ., Πολυτεχνείο Κρήτης, Χανιά 2018.
- Μαυρομάτης Γ., *Στατιστικά μοντέλα και μέθοδοι ανάλυσης δεδομένων*, 1999, University studio press, Θεσσαλονίκη.
- Μπεχράκης Θ., *Πολυδιάστατη Ανάλυση Δεδομένων*, 1999, Λιβάνη.
- Ντζούφρας Ι, *Στοιχεία Πολυμεταβλητής Ανάλυσης Δεδομένων*, 2002, Πανεπιστημιακές παραδόσεις Πανεπιστήμιο Αιγαίου.
- Παπαδημητρίου Γ., *Εισαγωγή στην Ανάλυση δεδομένων*, 1998, Εκδ .Πανεπιστήμιο Μακεδονίας.
- Σιάρδος Γ., *Μέθοδοι πολυμεταβλητής Στατιστικής Ανάλυσης*, εκδόσεις ΖΗΤΗ, 2004, 3η έκδοση
- ΥΠΕΝ (2017). *Ετήσια Έκθεση Ποιότητας της Ατμόσφαιρας 2016*. Υπουργείο Περιβάλλοντος και Ενέργειας. Γεν. Δ/νση Περιβαλλοντικής Πολιτικής, Δ/νση Κλιματικής Αλλαγής & Ποιότητας Ατμόσφαιρας, Τμήμα Ποιότητας της Ατμόσφαιρας,, Ιούνιος 2017.

Ξενόγλωσση

- Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). London: Arnold

Ιστοσελίδες

http://daskalosa.eu/physics_st/st_fysika_05_zoa.html

www.slideshare.net/ssakpi/density-based-clustering

http://www.sthda.com/english/articles/30-advanced-clustering/104-model-based-clustering-essentials/

https://python-graph-gallery.com/400-basic-dendrogram/

https://www.youtube.com/watch?v=_aWzGGNrcic

https://kunuk.wordpress.com/2011/09/17/clustering-k-means-and-grid-with-c-example-and-html-canvas-part-2/

https://uc-r.github.io/kmeans_clustering

http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html