ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
AGRICULTURAL UNIVERSITY OF ATHENS

Ανάλυση Συστάδων

Κατσιλέρος Αναστάσιος

2020

ΑΝΑΛΥΣΗ ΣΥΣΤΑΔΩΝ

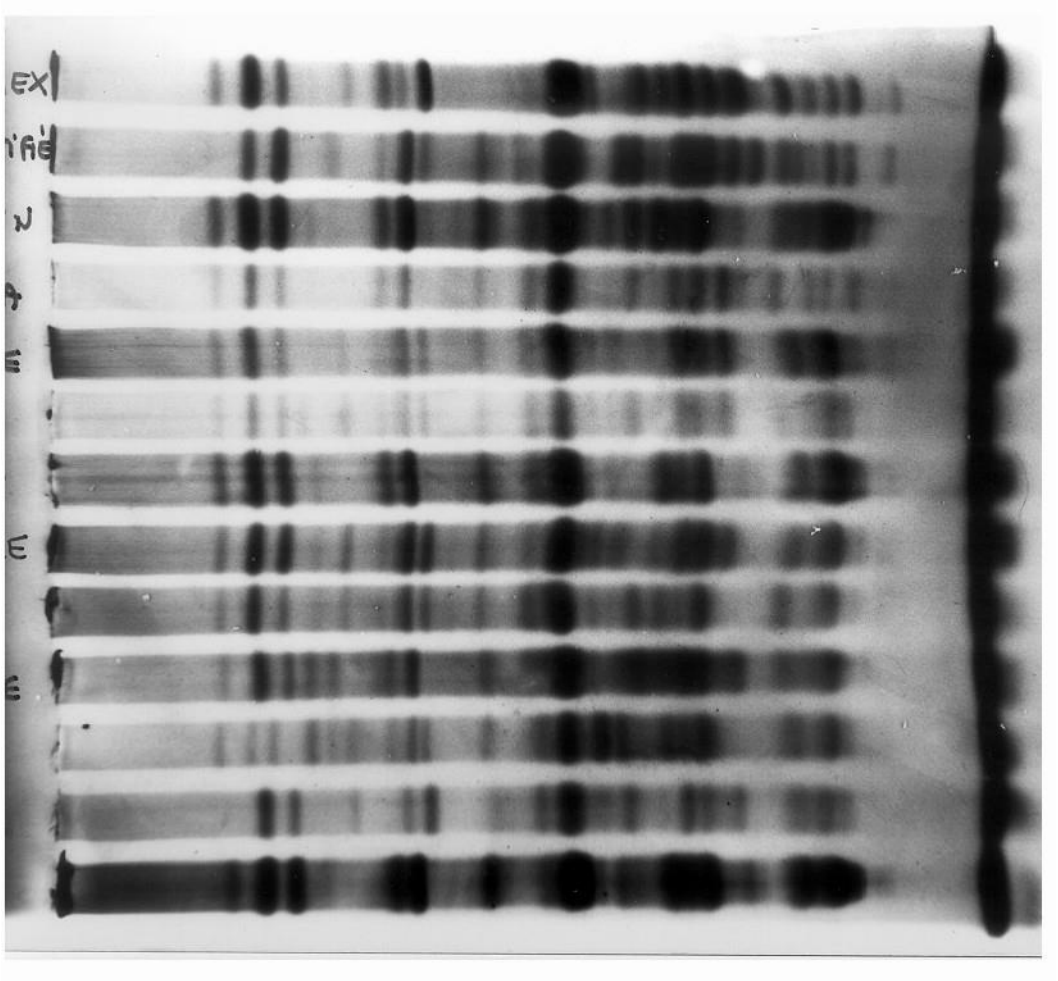
Η ανάλυση συστάδων (cluster analysis) είναι μία μέθοδος η οποία ταξινομεί ένα σύνολο στοιχείων (άτομα, αντικείμενα, ποικιλίες κ.α.) σε ομάδες - συστάδες, χρησιμοποιώντας διάφορα χαρακτηριστικά – μεταβλητές των στοιχείων αυτών. Η ταξινόμηση των στοιχείων σε ομάδες γίνεται κατά τέτοιο τρόπο ώστε τα στοιχεία κάθε ομάδας να παρουσιάζουν μεγάλη ομοιότητα – συγγένεια, ενώ μεταξύ των ομάδων να υπάρχει όσο το δυνατόν μεγαλύτερη διαφοροποίηση.

- ✓ Το πρώτο και σημαντικότερο βήμα στην ανάλυση συστάδων είναι η περιγραφή των στοιχείων με την επιλογή των κατάλληλων χαρακτηριστικών.
- ✓ Στη συνέχεια, πρέπει να οριστεί το **μέτρο ομοιότητας ή απόστασης**, με το οποίο θα γίνονται οι συγκρίσεις μεταξύ των στοιχείων.
- ✓ Τέλος, πρέπει να επιλεγεί η **μέθοδος ομαδοποίησης** που ακολουθείται για την τελική παραγωγή των ομάδων.

Ανάλογα με την επιλογή του μέτρου ομοιότητας - απόστασης και της μεθόδου ομαδοποίησης, οι ομάδες που προκύπτουν είναι διαφορετικές.

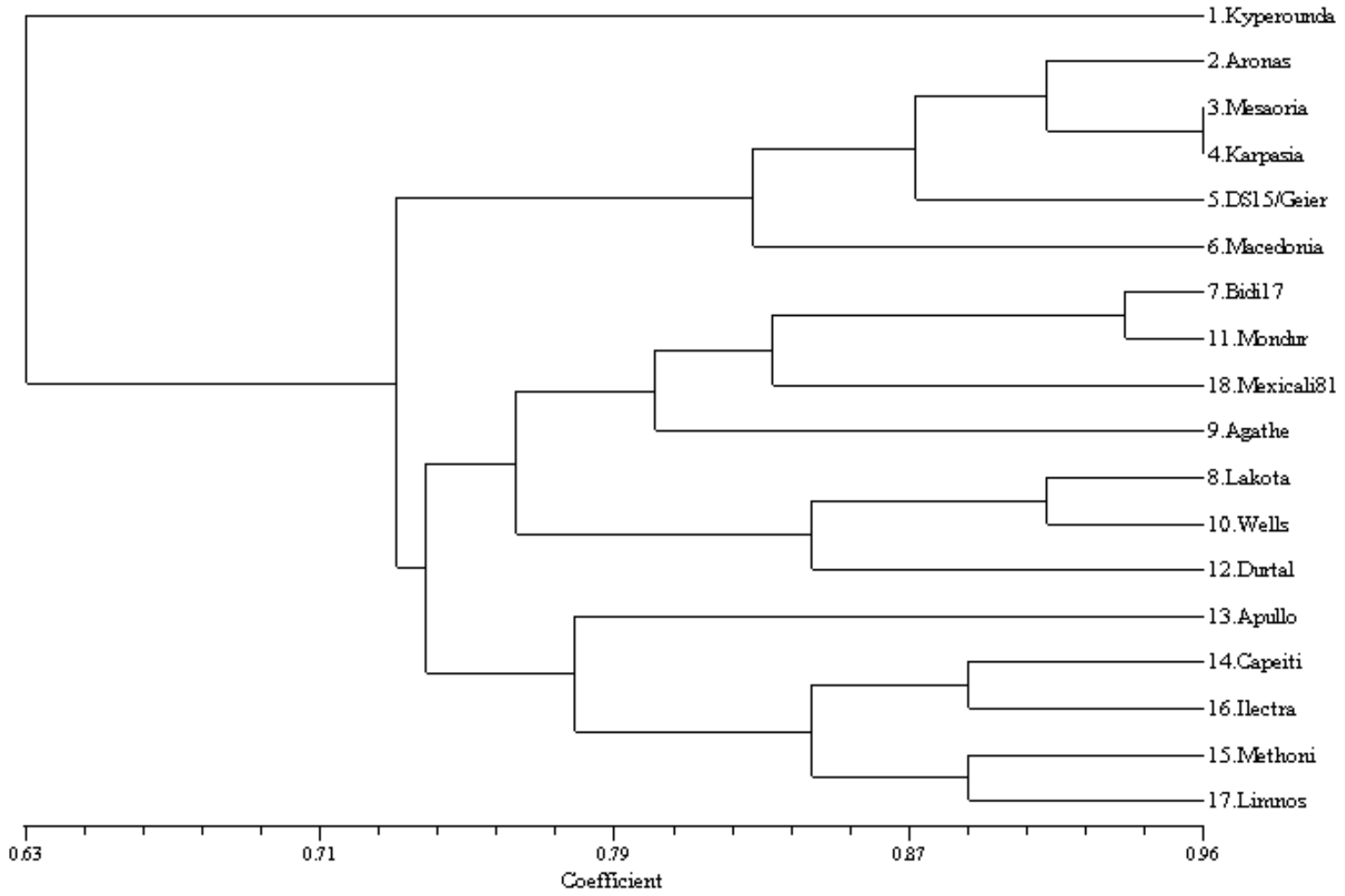
Ο ερευνητής επιλέγει την κατάλληλη ομαδοποίηση ανάλογα με τη φύση των δεδομένων και το πρόβλημα που εξετάζει. Πρέπει να γίνουν πολλές δοκιμές οι οποίες θα περιλαμβάνουν διαφορετικές μεταβλητές και διαφορετικά μέτρα σύγκρισης, ώστε να καταλήξει στην κατάλληλη ομαδοποίηση η οποία να μπορεί να ερμηνευτεί.

Παράδειγμα: Ηλεκτροφόρηση γλιανιδών σιταριού (πηγή: Συμιλλίδης)



Πειραματικοί Σχεδιασμοί

Διαχωρισμός 18 ποικιλιών σίτου με βάση τις γλιαδίνες τους (UPGMA) (πηγή: Συμιλλίδης)



Μέτρα απόστασης

1. Ευκλείδεια απόσταση

$$d_{ij} = \sqrt{\Sigma(X_{ij} - X_{ik})^2}$$

2. Τετραγωνική ευκλείδεια απόσταση

$$d_{ij} = \Sigma(X_{ij} - X_{ik})^2$$

3. Απόσταση Manhattan

$$d_{ij} = \Sigma\left|X_{ij} - X_{ik}\right|$$

4. Μέγιστη απόσταση

$$d_{ij} = \max_i\left|X_{ij} - X_{ik}\right|$$

5. Απόσταση Mahalanobis

$$d_{ij} = \sqrt{(X_{ij} - X_{ik})^T \Sigma^{-1} (X_{ij} - X_{ik})}$$

Μέτρα ομοιότητας

	Παρουσία	**Απουσία**
Παρουσία	a	b
Απουσία	c	d

1. Συντελεστής του Jaccard

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

2. Συντελεστής του Dice - Sorenson

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

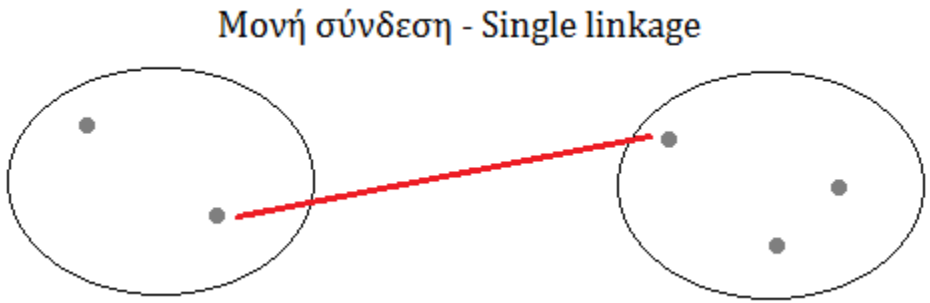
3. Συντελεστής Simple Matching (Sokal & Michener)

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

Ιεραρχικές Μέθοδοι Διασύνδεσης

1. Ομαδοποίηση με **απλή διασύνδεση** ή πλησιέστερης γειτνίασης διασύνδεση (single linkage). Η απόσταση μεταξύ δύο συστάδων είναι η ελάχιστη απόσταση (ή μεγαλύτερη ομοιότητα) μεταξύ μιας παρατήρησης σε μια συστάδα και μιας παρατήρησης στην άλλη συστάδα.

$$D_{KL} = \min_{i \in C_K} \min_{j \in C_L} d(x_i - x_j)$$

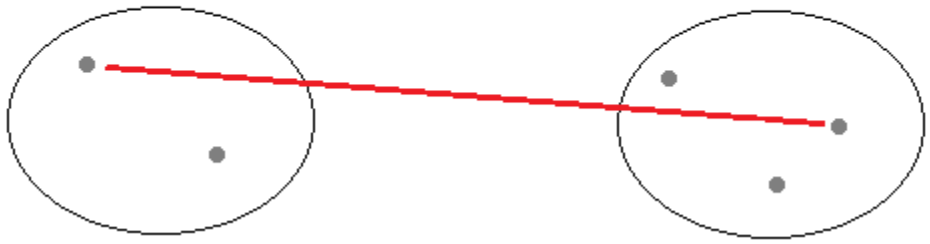


Όπου: n ο αριθμός των παρατηρήσεων, v ο αριθμός των μεταβλητών, x_i η *i* παρατήρηση, C_K η K συστάδα , N_K ο αριθμός των παρατηρήσεων στο C_K , x ο μέσος κάθε παρατήρησης, x_k ο μέσος της συστάδας C_K

2. Ομαδοποίηση με **πλήρη διασύνδεση** ή απομακρυσμένης γειτνίασης διασύνδεση (complete linkage). Η απόσταση μεταξύ δύο συστάδων είναι η μέγιστη απόσταση (ή μικρότερη ομοιότητα) μεταξύ μιας παρατήρησης σε μια συστάδα και μιας παρατήρησης στην άλλη συστάδα. Είναι ευαίσθητη σε ακραίες τιμές.

$$D_{KL} = \max_{i \in C_K} \max_{i \in C_L} d(x_i - x_j)$$

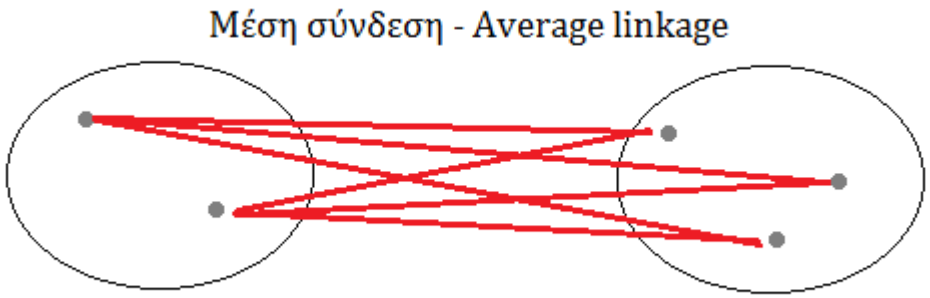
Πλήρη σύνδεση - Complete linkage



Όπου: n ο αριθμός των παρατηρήσεων, v ο αριθμός των μεταβλητών, x_i η *i* παρατήρηση, C_K η K συστάδα , N_K ο αριθμός των παρατηρήσεων στο C_K , x ο μέσος κάθε παρατήρησης, x_k ο μέσος της συστάδας C_K

3. Ομαδοποίηση με **μέση διασύνδεση** (average linkage). Η απόσταση μεταξύ δύο συστάδων είναι η μέση απόσταση μεταξύ ζευγών παρατηρήσεων.

$$D_{KL} = \sum_{i \in c} \sum_{j \in c} \frac{d(x_i - x_j)}{N_K N_L}$$



4. Ομαδοποίηση με **κεντροειδή μέθοδο** (centroid method). Η απόσταση μεταξύ δύο συστάδων ορίζεται ως η τετραγωνική ευκλείδεια απόσταση μεταξύ των μέσων τους.

$$D_{KL} = \left\| \bar{x}_K - \bar{x}_L \right\|^2$$

5. Ομαδοποίηση κατά **Ward**. Η απόσταση μεταξύ δύο συστάδων είναι το άθροισμα των τετραγώνων μεταξύ των δύο συστάδων με προστιθέμενες όλες τις μεταβλητές.

$$D_{KL} = \frac{\left\| \bar{x}_K - \bar{x}_L \right\|^2}{\frac{1}{N_K} + \frac{1}{N_L}}$$

Μη Ιεραρχική Μέθοδος Διασύνδεσης k-means

Η μη ιεραρχική μέθοδος είναι μια επαναληπτική διαδικασία προσαρμογής για τον σχηματισμό ενός αριθμού συγκεκριμένων συστάδων. Η μέθοδος k-mean αρχικά επιλέγει ένα σύνολο n σημείων που ονομάζονται κεντροειδή και είναι μία πρώτη προσέγγιση των μέσων των συστάδων. Κάθε παρατήρηση αποδίδεται στο πλησιέστερο κεντροειδές σημείο και σχηματίζονται ένα σύνολο προσωρινών συστάδων. Τα κεντροειδή σημεία στη συνέχεια αντικαθίστανται από τα μέσα των συστάδων και οι παρατηρήσεις ανακατανέμονται. Η διαδικασία συνεχίζεται μέχρις ότου δεν εμφανιστούν περαιτέρω αλλαγές στις συστάδες.

Αριθμητικό Παράδειγμα: 12 γονότυποι και 5 ποσοτικές μεταβλητές

Γονότυποι	V1	V2	V3	V4	V5	Γονότυποι	V1	V2	V3	V4	V5
1	5700	12,8	2500	270	25	**1**	-0,17	0,79	0,13	1,30	1,26
2	1000	10,9	600	10	10	**2**	-1,62	-0,28	-1,40	-0,96	-1,10
3	3400	8,8	1000	10	9	**3**	-0,88	-1,46	-1,07	-0,96	-1,26
4	3800	13,6	1700	140	25	**4**	-0,75	1,24	-0,51	0,17	1,26
5	4000	12,8	1600	140	25	**5**	-0,69	0,79	-0,59	0,17	1,26
6	8200	8,3	2600	60	12	**6**	0,60	-1,74	0,21	-0,53	-0,79
7	1200	11,4	400	10	16	**7**	-1,56	0,00	-1,56	-0,96	-0,16
8	9100	11,5	3300	60	14	**8**	0,88	0,06	0,78	-0,53	-0,47
9	9900	12,5	3400	180	18	**9**	1,13	0,62	0,86	0,52	0,16
10	9600	13,7	3600	390	25	**10**	1,04	1,29	1,02	2,34	1,26
11	9600	9,6	3300	80	12	**11**	1,04	-1,01	0,78	-0,36	-0,79
12	9400	11,4	4000	100	13	**12**	0,97	0,00	1,34	-0,18	-0,63
Mean	6241,7	11,4	2333,3	120,8	17,0						
SD	3239,9	1,78	1241,2	114,9	6,36						

Τυποποίηση τιμών:

$$Z = \frac{X - \mu}{\sigma}$$

Υπολογισμός Ευκλείδειας απόστασης

Γονότυποι	V1	V2	V3	V4	V5			1	2	3	4	5	6	7	8	9	10	11	12
1	-0,17	0,79	0,13	1,30	1,26		**1**												
2	-1,62	-0,28	-1,40	-0,96	-1,10		**2**	4,00											
3	-0,88	-1,46	-1,07	-0,96	-1,26		**3**	4,29	1,42										
4	-0,75	1,24	-0,51	0,17	1,26		**4**	1,48	3,25	3,91									
5	-0,69	0,79	-0,59	0,17	1,26		**5**	1,43	3,06	3,60	0,46								
6	0,60	-1,74	0,21	-0,53	-0,79		**6**	3,81	3,05	2,02	3,98	3,64							
7	-1,56	0,00	-1,56	-0,96	-0,16		**7**	3,51	0,99	1,99	2,55	2,34	3,30						
8	0,88	0,06	0,78	-0,53	-0,47		**8**	2,87	3,31	3,05	2,98	2,84	1,92	3,32					
9	1,13	0,62	0,86	0,52	0,16		**9**	1,96	4,05	3,95	2,59	2,53	2,85	3,86	1,36				
10	1,04	1,29	1,02	2,34	1,26		**10**	1,82	5,55	5,65	3,14	3,18	4,68	5,21	3,55	2,21			
11	1,04	-1,01	0,78	-0,36	-0,79		**11**	3,44	3,46	2,73	3,73	3,50	1,03	3,63	1,14	2,08	4,05		
12	0,97	0,00	1,34	-0,18	-0,63		**12**	3,00	3,79	3,47	3,36	3,24	2,15	3,86	0,68	1,32	3,82	1,19	

Ευκλείδεια απόσταση $d_{ij} = \sqrt{\Sigma(X_{ij} - X_{ik})^2}$

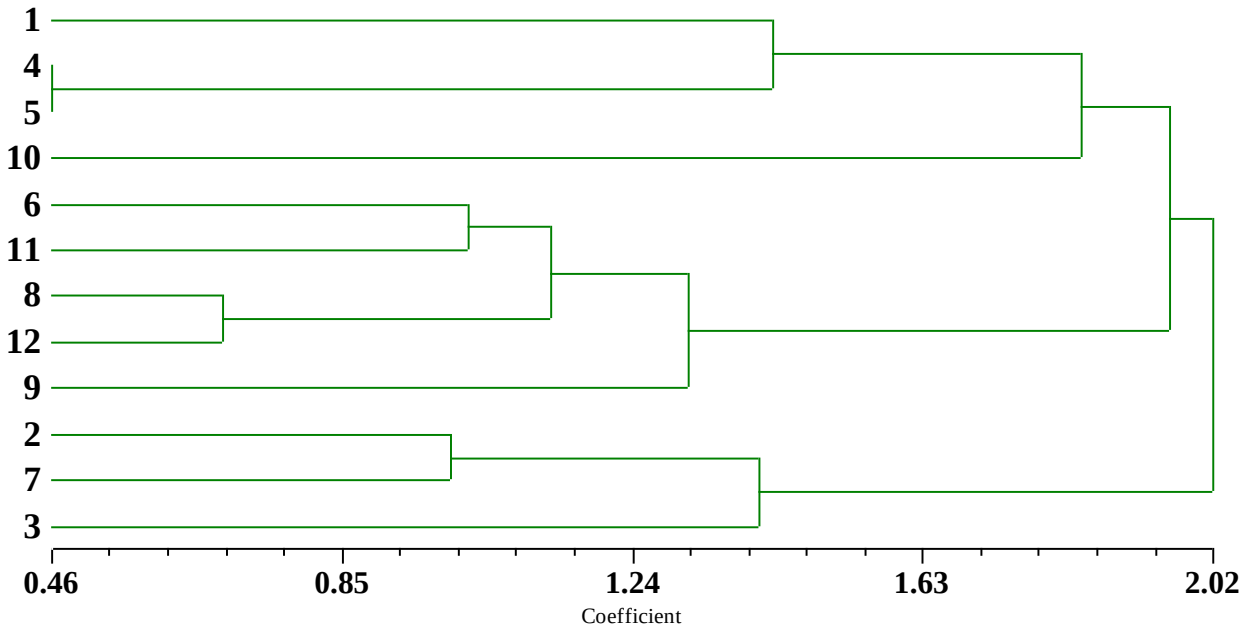
$$d_{1,2} = \sqrt{(-0,17-(-1,62))^2 + (0,79-(-0,28))^2 + (0,13-(-1,40))^2 + (1,30-(-0,96))^2 + (1,26-(-1,10))^2} = 4,0$$

Πειραματικοί Σχεδιασμοί

Μέθοδος της απλής διασύνδεσης ή πλησιέστερης γειτνίασης

	1	2	3	4	5	6	7	8	9	10	11
2	4,00										
3	4,29	1,42									
4	1,48	3,25	3,91								
5	1,43	3,06	3,60	0,46							
6	3,81	3,05	2,02	3,98	3,64						
7	3,51	0,99	1,99	2,55	2,34	3,30					
8	2,87	3,31	3,05	2,98	2,84	1,92	3,32				
9	1,96	4,05	3,95	2,59	2,53	2,85	3,86	1,36			
10	1,82	5,55	5,65	3,14	3,18	4,68	5,21	3,55	2,21		
11	3,44	3,46	2,73	3,73	3,50	1,03	3,63	1,14	2,08	4,05	
12	3,00	3,79	3,47	3,36	3,24	2,15	3,86	0,68	1,32	3,82	1,19

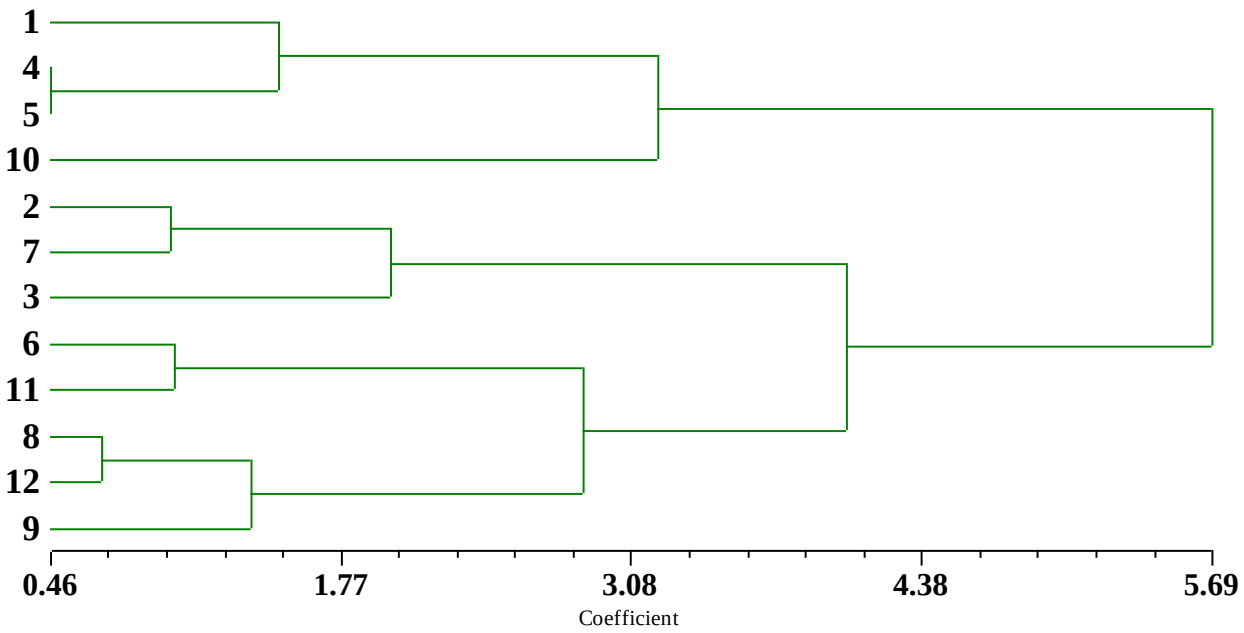
Απόσταση												
0,46	4	5										
0,68	8	12										
0,99	2	7										
1,03	6	11										
1,14	8	12	6	11								
1,32	8	12	6	11	9							
1,42	2	7	3									
1,43	4	5	1									
1,82	4	5	1	10								
1,96	8	12	6	11	9	4	5	1	10			
2,02	8	12	6	11	9	4	5	1	10	2	7	3



Μέθοδος της πλήρης διασύνδεσης ή απομακρυσμένης γειτνίασης

	1	2	3	4	5	6	7	8	9	10	11
2	4,00										
3	4,29	1,42									
4	1,48	3,25	3,91								
5	1,43	3,06	3,60	0,46							
6	3,81	3,05	2,02	3,98	3,64						
7	3,51	0,99	1,99	2,55	2,34	3,30					
8	2,87	3,31	3,05	2,98	2,84	1,92	3,32				
9	1,96	4,05	3,95	2,59	2,53	2,85	3,86	1,36			
10	1,82	5,55	5,65	3,14	3,18	4,68	5,21	3,55	2,21		
11	3,44	3,46	2,73	3,73	3,50	1,03	3,63	1,14	2,08	4,05	
12	3,00	3,79	3,47	3,36	3,24	2,15	3,86	0,68	1,32	3,82	1,19

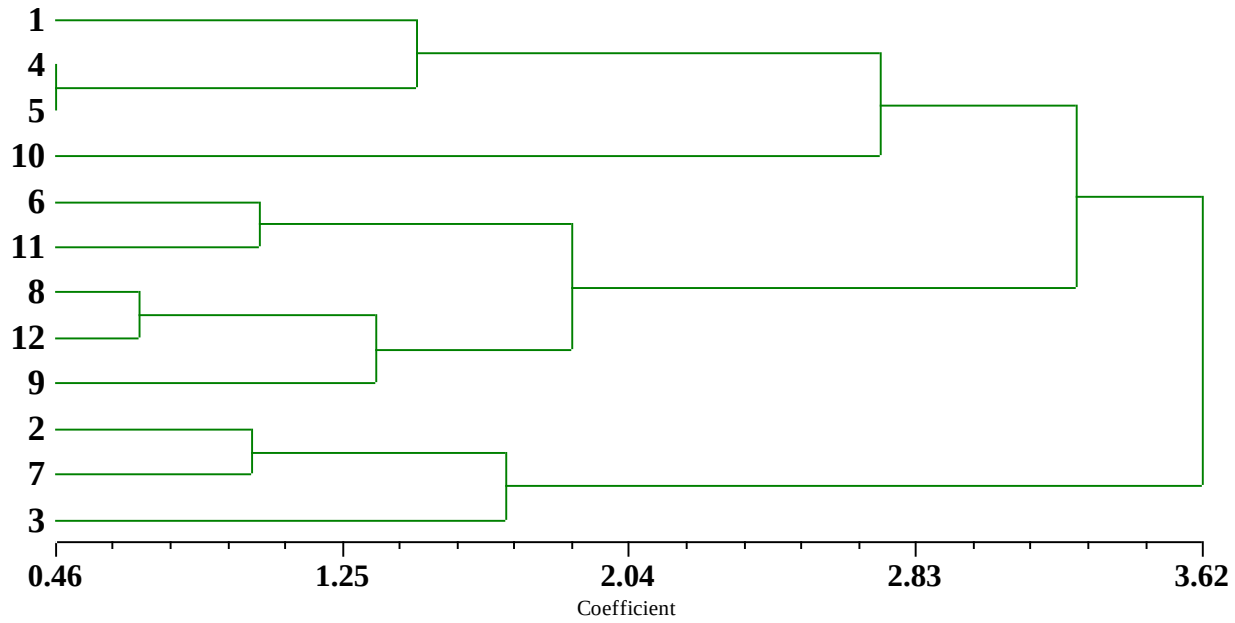
Απόσταση												
0,46	4	5										
0,68	8	12										
0,99	2	7										
1,03	6	11										
1,36	8	12	9									
1,48	4	5	1									
1,99	2	7	3									
2,85	8	12	9	6	11							
3,18	4	5	1	10								
4,05	8	12	6	11	9	4	5	1	10			
5,65	8	12	6	11	9	4	5	1	10	2	7	3



Μέθοδος της μέσης διασύνδεσης (UPGMA)

	1	2	3	4	5	6	7	8	9	10	11
2	4,00										
3	4,29	1,42									
4	1,48	3,25	3,91								
5	1,43	3,06	3,60	0,46							
6	3,81	3,05	2,02	3,98	3,64						
7	3,51	0,99	1,99	2,55	2,34	3,30					
8	2,87	3,31	3,05	2,98	2,84	1,92	3,32				
9	1,96	4,05	3,95	2,59	2,53	2,85	3,86	1,36			
10	1,82	5,55	5,65	3,14	3,18	4,68	5,21	3,55	2,21		
11	3,44	3,46	2,73	3,73	3,50	1,03	3,63	1,14	2,08	4,05	
12	3,00	3,79	3,47	3,36	3,24	2,15	3,86	0,68	1,32	3,82	1,19

Απόσταση												
0,46	4	5										
0,68	8	12										
0,99	2	7										
1,03	6	11										
1,36	8	12	9									
1,48	4	5	1									
1,71	2	7	3									
1,97	8	12	9	6	11							
2,73	4	5	1	10								
3,33	8	12	6	11	9	4	5	1	10			
3,72	8	12	6	11	9	4	5	1	10	2	7	3



Αριθμητικό Παράδειγμα: 5 γονότυποι και 5 δυαδικές μεταβλητές (παρουσία – απουσία)

Γονότυποι	V1	V2	V3	V4	V5
1	1	1	0	1	0
2	1	0	0	0	1
3	1	1	1	1	0
4	0	1	1	0	1
5	0	1	0	1	0

Γονότυποι 1 & 2

	Παρουσία (1) στο γον. 1	Απουσία (0) στο γον. 1
Παρουσία (1) στο γον. 2	a = 1	b = 1
Απουσία (0) στο γον. 2	c = 2	d = 1

Συντελεστής Jaccard

$$S_{J(1,2)} = \frac{a}{a+b+c} = \frac{1}{1+1+2} = 0,25$$

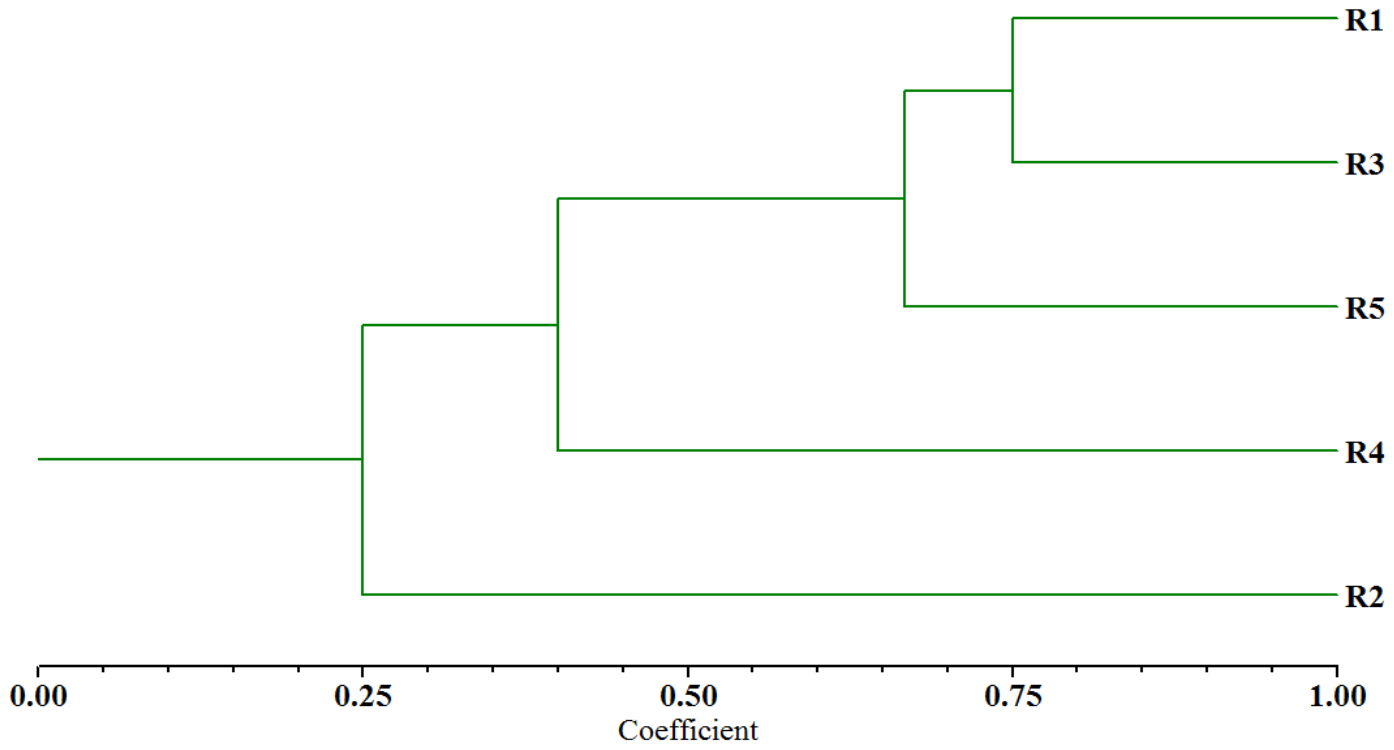
Συντελεστής Dice

$$S_{D(1,2)} = \frac{2a}{2a+b+c} = \frac{2*1}{2*1+1+2} = 0,4$$

Συντελεστής Jaccard

	1	2	3	4	5
1	1				
2	0,25	1			
3	0,75	0,2	1		
4	0,2	0,25	0,4	1	
5	0,66	0	0,5	0,25	1

Μέθοδος απλής διασύνδεσης

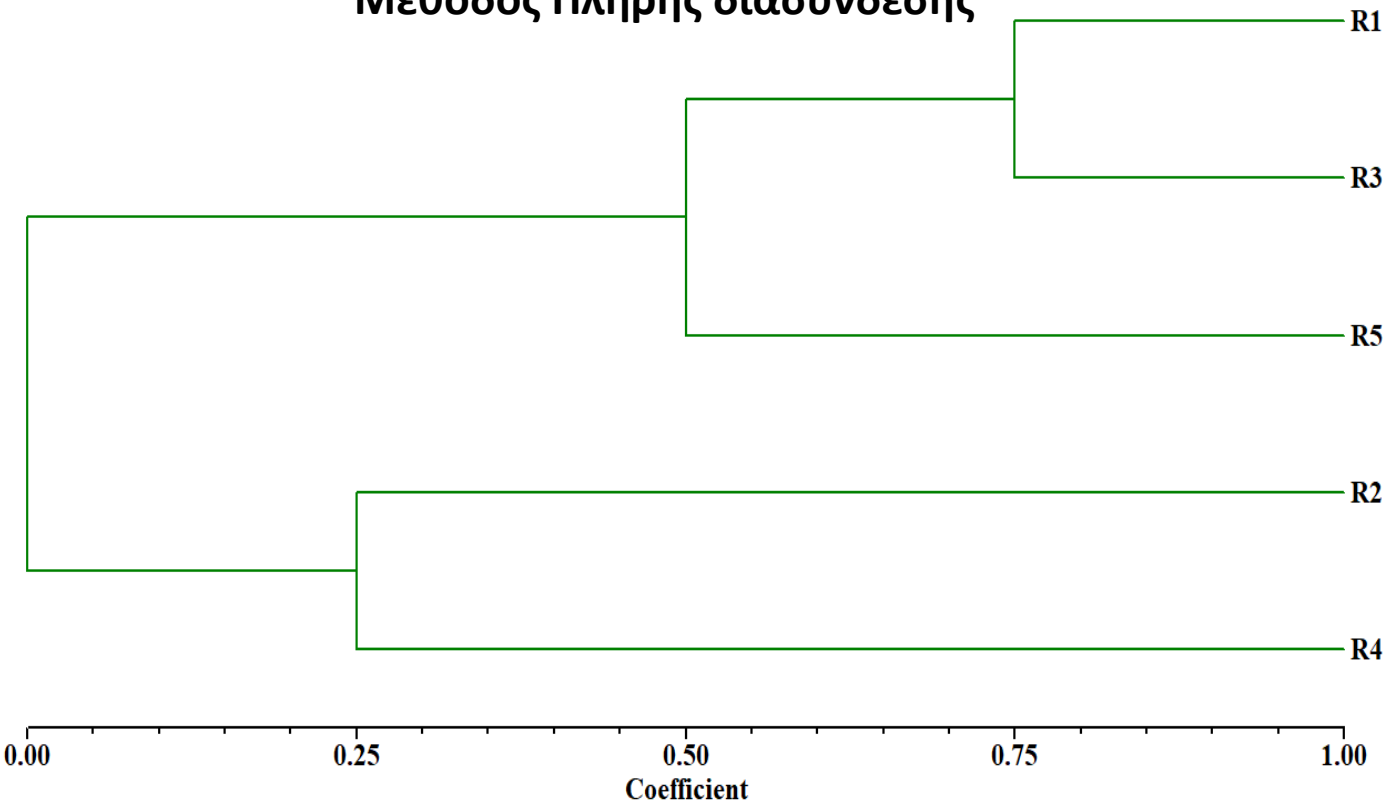


Πειραματικοί Σχεδιασμοί

Συντελεστής Jaccard

	1	2	3	4	5
1	1				
2	0,25	1			
3	0,75	0,2	1		
4	0,2	0,25	0,4	1	
5	0,66	0	0,5	0,25	1

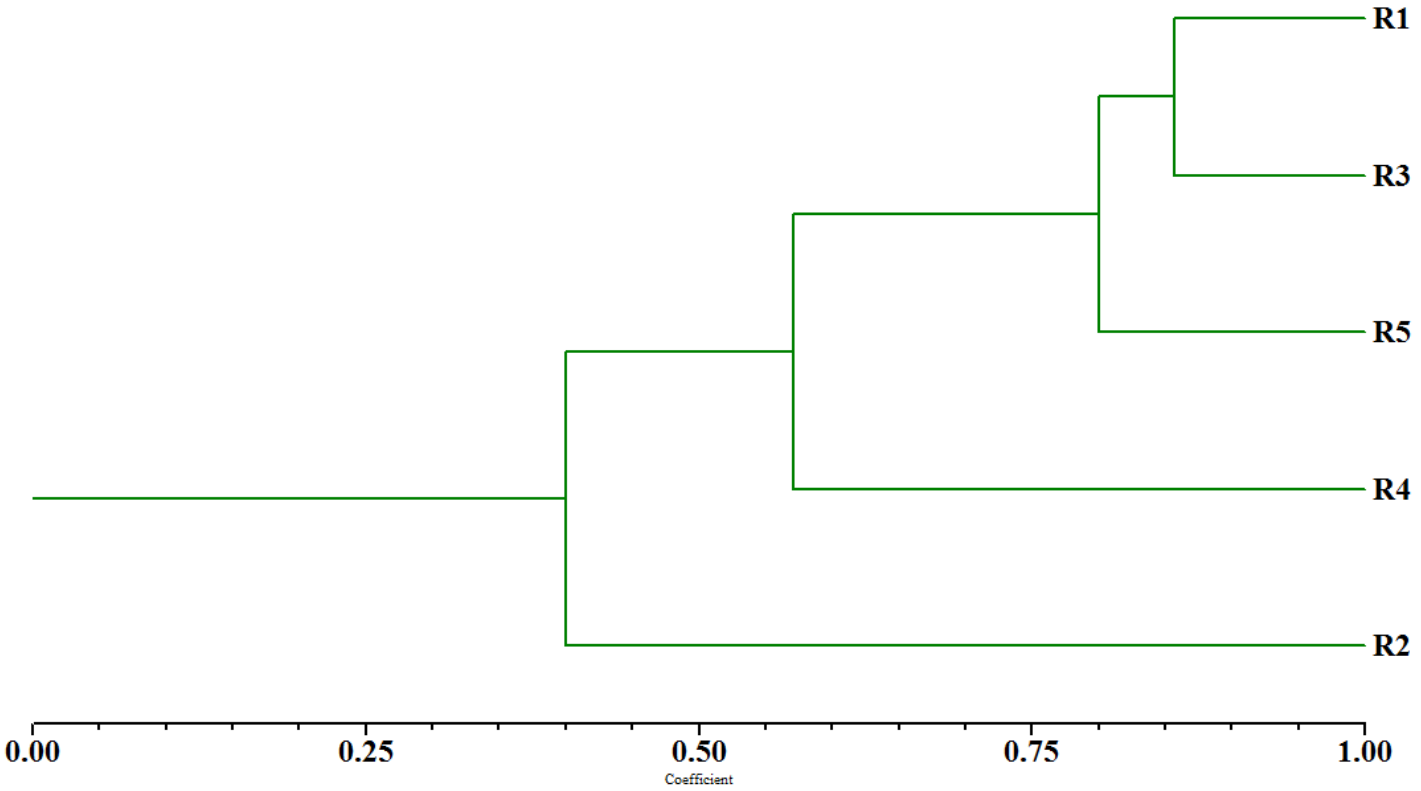
Μέθοδος Πλήρης διασύνδεσης



Συντελεστής Dice

	1	2	3	4	5
1	1				
2	0,4	1			
3	0,86	0,33	1		
4	0,33	0,4	0,57	1	
5	0,8	0	0,66	0,4	1

Μέθοδος απλής διασύνδεσης



Συντελεστής Dice

	1	2	3	4	5
1	1				
2	0,4	1			
3	0,86	0,33	1		
4	0,33	0,4	0,57	1	
5	0,8	0	0,66	0,4	1

Μέθοδος Πλήρης διασύνδεσης