ΑΝΑΛΥΣΗ ΣΥΣΤΑΔΩΝ

Η ανάλυση συστάδων κατανέμει ένα σύνολο μεταβλητών ή παρατηρήσεων σε συγκεκριμένες ομάδες οι οποίες διαθέτουν κοινά χαρακτηριστικά, ευκρινώς διαφοροποιημένα από εκείνα των άλλων ομάδων.

Η απόσταση των στοιχείων στο χώρο μετρείται με τους ειδικούς συντελεστές ομοιότητας και η σύνδεσή τους προς δημιουργία συστάδων πραγματοποιείται με ειδικές μεθόδους διασύνδεσης, ιεραρχικού ή μη χαρακτήρα.

- Το πρώτο και σημαντικότερο βήμα κατά τη διαδικασία της ανάλυσης συστάδων είναι η περιγραφή των δεδομένων και η επιλογή των κατάλληλων χαρακτηριστικών.
- Στη συνέχεια, πρέπει να οριστεί το μέτρο ομοιότητας με το οποίο θα γίνονται οι συγκρίσεις μεταξύ των παρατηρήσεων.
- Τέλος, πρέπει να επιλεγεί η μέθοδος ομαδοποίησης που ακολουθείται για την τελική παραγωγή των ομάδων. Οι ιεραρχικοί και k-means αλγόριθμοι είναι οι συνηθέστερες επιλογές.

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

Ο ερευνητής πρέπει να κάνει τις επιλογές αυτές ανάλογα με τη φύση των δεδομένων και το πρόβλημα που εξετάζει. Συχνά απαιτούνται πολλές δοκιμές της ανάλυσης συστάδων, περιλαμβάνοντας διαφορετικές μεταβλητές ή αφαιρώντας κάποιες παρατηρήσεις και χρησιμοποιώντας διαφορετικά μέτρα σύγκρισης, ώστε να εξακριβωθεί η σταθερότητα της ομαδοποίησης.

Το τελικό αποτέλεσμα πρέπει να μπορεί να ερμηνευτεί. Για τον σκοπό αυτό μελετώνται οι τιμές των μεταβλητών σε κάθε ομάδα και με βάση την εμπειρία του ερευνητή εξετάζεται αν οι ομάδες υπάρχουν στην πραγματικότητα ή αποτελούν απλά το αποτέλεσμα ενός αλγορίθμου.

Συντελεστές ποσοτικών στοιχείων

1. Ευκλείδεια απόσταση

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

Η Ευκλείδεια απόσταση έχει το πλεονέκτημα ότι η απόσταση μεταξύ δύο οποιωνδήποτε στοιχείων δεν επηρεάζεται από την ύπαρξη στοιχείων με μεγάλες αποστάσεις (ακραίες τιμές).

2. Τετραγωνική Ευκλείδεια απόσταση

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

Η τετραγωνική Ευκλείδεια απόσταση χρησιμοποιείται όταν επιθυμούμε να προσδώσουμε μεγαλύτερο βάρος σε στοιχεία που σχετικά είναι απομακρυσμένα μεταξύ τους.

3. Απόσταση Manhattan

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

Η τεχνική αυτή δεν ανιχνεύει αν υπάρχουν μεγάλες διαφορές μεταξύ των αποστάσεων των στοιχείων.

4. Απόσταση Chebychev

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

Η απόσταση Chebychev η οποία μεγιστοποιεί το αποτέλεσμα της απόστασης των στοιχείων.

5. Ποσοστό ομοιότητας

$$p_{ij} = \Sigma \min(p_{ij}, p_{ik})$$

Το ποσοστό ομοιότητας ή ανομοιότητας ή δυσαρμονίας το οποίο εφαρμόζεται μόνο σε στοιχεία αναλογιών (ποσοστών).

6. Μέση Ευκλείδεια απόσταση

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

7. Απόσταση των Bray-Curtis

$$d_{ij} = \frac{\Sigma\left|X_{ij} - X_{ik}\right|}{\Sigma(X_{ij} + X_{ik})}$$

8. Συντελεστής Canberra

$$d_{ij} = \frac{1}{i}\frac{\Sigma\left|X_{ij} - X_{ik}\right|}{\Sigma(X_{ij} + X_{ik})}$$

9. Απόσταση του Minkowski

$$d_{ij} = \Sigma\left|(X_{ij} - X_{ik})^p\right|^{\frac{1}{p}}$$

10. Απόσταση Mahalanobis

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

11. Απόσταση του Pearson

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

12. Τετραγωνική απόσταση του Pearson

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

Συντελεστές δυαδικών στοιχείων

	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)

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

4. Συντελεστής Watson, Williams & Lance

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

5. Συντελεστής Levandowsky

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

6. Συντελεστής Q του Yules

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

7. Συντελεστής Russel & Rao

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

Ιεραρχική Μέθοδος Διασύνδεσης

1. Ομαδοποίηση με **απλή διασύνδεση** ή πλησιέστερης γειτνίασης διασύνδεση (single linkage). Η απόσταση μεταξύ δύο ομάδων προσδιορίζεται από την απόσταση των δύο κοντινότερων στοιχείων που το καθένα ανήκει σε διαφορετική ομάδα. Η μέθοδος αυτή τείνει να δημιουργεί μεγάλο αριθμό διακλαδιζόμενων ομάδων.

2. Ομαδοποίηση με **πλήρη διασύνδεση** ή απομακρυσμένης γειτνίασης διασύνδεση (complete linkage). Η απόσταση μεταξύ δύο ομάδων προσδιορίζεται από την απόσταση των δύο πλέον απομακρυσμένων στοιχείων που το καθένα ανήκει σε διαφορετική ομάδα. Η μέθοδος αυτή είναι κατάλληλη σε στοιχεία που εμφανίζουν φυσικώς ευδιάκριτες δέσμες διαφοροποίησης.

3. Ομαδοποίηση με **μη σταθμισμένη κατά ζεύγη μέση διασύνδεση ή μέση πλήρη διασύνδεση** (unweighted pairgroup average linkage or average complete linkage). Η απόσταση μεταξύ δύο ομάδων υπολογίζεται ως η μέση απόσταση μεταξύ όλων των ζευγών των στοιχείων στις δύο διαφορετικές ομάδες. Η μέση ή ενδιάμεση διασύνδεση αποτελεί πλεονεκτικό συνδυασμό των δύο προηγουμένων ταξινόμησης των ομάδων. Θεωρείται ως η πλέον αποτελεσματική μέθοδος, έχει όμως το μειονέκτημα να σχηματίζει ομάδες πολύ μικρού μεγέθους.

4. Ομαδοποίηση με **σταθμισμένη κατά ζεύγη μέση διασύνδεση** (weighted pair-group average linkage or weighted average linkage), γνωστή και ως ομαδοποίηση του McQuitty. Η απόσταση μεταξύ δύο ομάδων υπολογίζεται, όπως και προηγουμένως, με την προσθήκη του μεγέθους κάθε ομάδας (αριθμός στοιχείων ανά ομάδα) ως συντελεστή στάθμισης. Η μέθοδος αυτή αντικαθιστά την προηγούμενη όταν τα μεγέθη των ομάδων εμφανίζονται ιδιαίτερα άνισα.

5. Ομαδοποίηση με **μη σταθμισμένη κεντροειδή διασύνδεση** (average centroid linkage or unweighted pair-goup centroid). Η απόσταση μεταξύ δύο ομάδων υπολογίζεται από τη διαφορά της απόστασης μεταξύ των δύο κεντρικών σημείων. Το κεντρικό σημείο μιας ομάδας είναι το ενδιάμεσο σημείο που ορίζεται από το σύνολο των διαστάσεων (μεταβλητών) που συμμετέχουν στην ομαδοποίηση και αντιστοιχεί στο κέντρο βάρους της ομάδας.

6. Ομαδοποίηση με **σταθμισμένη κεντροειδή διασύνδεση** (weighted average centroid linkage or weighted pair-goup centroid). Αν οι σχηματιζόμενες ομάδες συντίθενται από άνισο αριθμό στοιχείων, τότε εισάγεται στην προηγούμενη μέθοδο και ένας συντελεστής στάθμισης που λαμβάνει υπόψη το διαφορετικό μέγεθος των ομάδων.

7. Ομαδοποίηση κατά **Ward**. Βασίζεται στην εφαρμογή της ανάλυσης της διακύμανσης στις παρατηρήσεις των ομάδων με σκοπό την εκτίμηση των αποστάσεων μεταξύ των ομάδων. Ουσιαστικά, η μέθοδος αυτή αποσκοπεί στην ελαχιστοποίηση της μεταβλητότητας μεταξύ δύο εξεταζόμενων ομάδων που σχηματίζονται σε κάθε διαδοχικό στάδιο της ιεραρχικής

Μη Ιεραρχική Μέθοδος Διασύνδεσης k-means

Η μη ιεραρχική μέθοδος προϋποθέτει ότι θέλουμε να δημιουργήσουμε έναν συγκεκριμένο αριθμό k ομάδων. Η μέθοδος ξεκινά με έναν αριθμό σημείων k ή με έναν αριθμό ομάδων παρατηρήσεων k. Αν η μέθοδος ξεκινήσει με k σημεία κάθε παρατήρηση τοποθετείται σε μία ομάδα με το πλησιέστερο προς αυτή σημείο.

Αν ξεκινήσουμε με k αριθμό ομάδων τότε αρχικά υπολογίζονται τα κεντροειδή των ομάδων. Στη συνέχεια ακολουθεί μια διαδικασία διαδοχικών προσεγγίσεων με κάποια κριτήρια βέλτιστου διαχωρισμού των ομάδων, υπολογίζοντας είτε νέα σημεία είτε νέες ομάδες, μέχρις ότου δεν υπάρχει θέμα μετακίνησης των παρατηρήσεων από μία ομάδα σε άλλη.

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	25000	**1**	-0,16	0,82	0,14	1,36	1,31
2	1000	10,9	600	10	10000	**2**	-1,59	-0,29	-1,46	-1,01	-1,15
3	3400	8,8	1000	10	9000	**3**	-0,86	-1,53	-1,12	-1,01	-1,31
4	3800	13,6	1700	140	25000	**4**	-0,74	1,29	-0,53	0,17	1,31
5	4000	12,8	1600	140	25000	**5**	-0,68	0,82	-0,62	0,17	1,31
6	8200	8,3	2600	60	12000	**6**	0,59	-1,82	0,22	-0,55	-0,82
7	1200	11,4	400	10	16000	**7**	-1,53	0,00	-1,63	-1,01	-0,16
8	9100	11,5	3300	60	14000	**8**	0,87	0,06	0,81	-0,55	-0,49
9	9900	12,5	3400	180	18000	**9**	1,11	0,65	0,90	0,54	0,16
10	9600	13,7	3600	390	25000	**10**	1,02	1,35	1,07	2,45	1,31
11	9600	9,6	3300	80	12000	**11**	1,02	-1,06	0,81	-0,37	-0,82
12	9400	11,4	4000	100	13000	**12**	0,96	0,00	1,40	-0,19	-0,66
Mean	6241,7	11,4	2333,3	120,8	17000,0						
SD	3293,5	1,7	1188,4	110,0	6096,4						

Τύπος τυποποίησης 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	-0,16	0,82	0,14	1,36	1,31		**1**												
2	-1,59	-0,29	-1,46	-1,01	-1,15		**2**	4,00											
3	-0,86	-1,53	-1,12	-1,01	-1,31		**3**	4,28	1,41										
4	-0,74	1,29	-0,53	0,17	1,31		**4**	1,48	3,24	3,89									
5	-0,68	0,82	-0,62	0,17	1,31		**5**	1,43	3,06	3,58	0,46								
6	0,59	-1,82	0,22	-0,55	-0,82		**6**	3,79	3,06	2,02	3,95	3,62							
7	-1,53	0,00	-1,63	-1,01	-0,16		**7**	3,51	0,99	1,99	2,54	2,34	3,29						
8	0,87	0,06	0,81	-0,55	-0,49		**8**	2,87	3,31	3,04	2,98	2,84	1,92	3,32					
9	1,11	0,65	0,90	0,54	0,16		**9**	1,96	4,04	3,97	2,59	2,53	2,85	3,86	1,36				
10	1,02	1,35	1,07	2,45	1,31		**10**	1,84	5,57	5,69	3,14	3,19	4,72	5,21	3,58	2,24			
11	1,02	-1,06	0,81	-0,37	-0,82		**11**	3,43	3,46	2,73	3,73	3,48	1,02	3,63	1,13	2,07	4,09		
12	0,96	0,00	1,40	-0,19	-0,66		**12**	2,99	3,79	3,46	3,36	3,23	2,13	3,86	0,68	1,31	3,42	1,17	

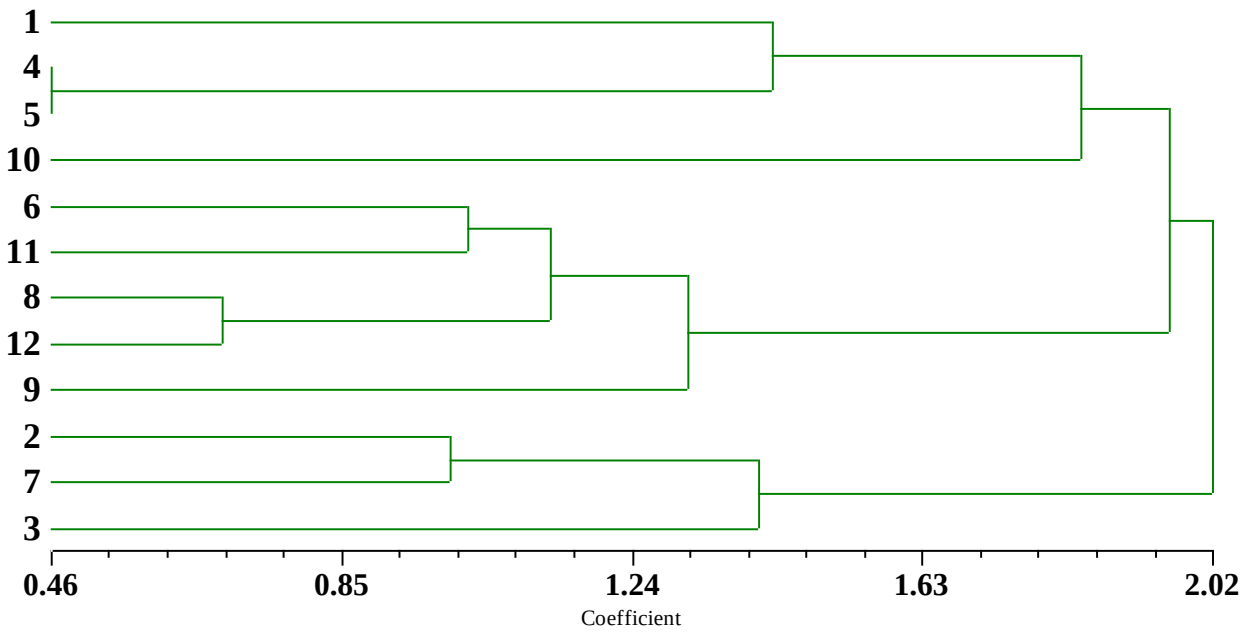
Ευκλείδεια απόσταση $d_{ij} = \sqrt{\Sigma(X_{ij} - X_{ik})^2}$

$$d_{1,2} = \sqrt{(-0{,}16-(-1{,}59))^2+(0{,}82-(-0{,}29))^2+(0{,}14-(-1{,}46))^2+(1{,}36-(-1{,}01))^2+(1{,}31-(-1{,}15))^2} = 4{,}0$$

Μέθοδος της απλής διασύνδεσης ή πλησιέστερης γειτνίασης

	1	2	3	4	5	6	7	8	9	10	11
2	4,00										
3	4,28	**1,41**									
4	1,48	3,24	3,89								
5	**1,43**	3,06	3,58	**0,46**							
6	3,79	3,06	**2,02**	3,95	3,62						
7	3,51	**0,99**	1,99	2,54	2,34	3,29					
8	2,87	3,31	3,04	2,98	2,84	1,92	3,32				
9	**1,96**	4,04	3,97	2,59	2,53	2,85	3,86	1,36			
10	**1,84**	5,57	5,69	3,14	3,19	4,72	5,21	3,58	2,24		
11	3,43	3,46	2,73	3,73	3,48	**1,02**	3,63	**1,13**	2,07	4,09	
12	2,99	3,79	3,46	3,36	3,23	2,13	3,86	**0,68**	**1,31**	3,42	1,17

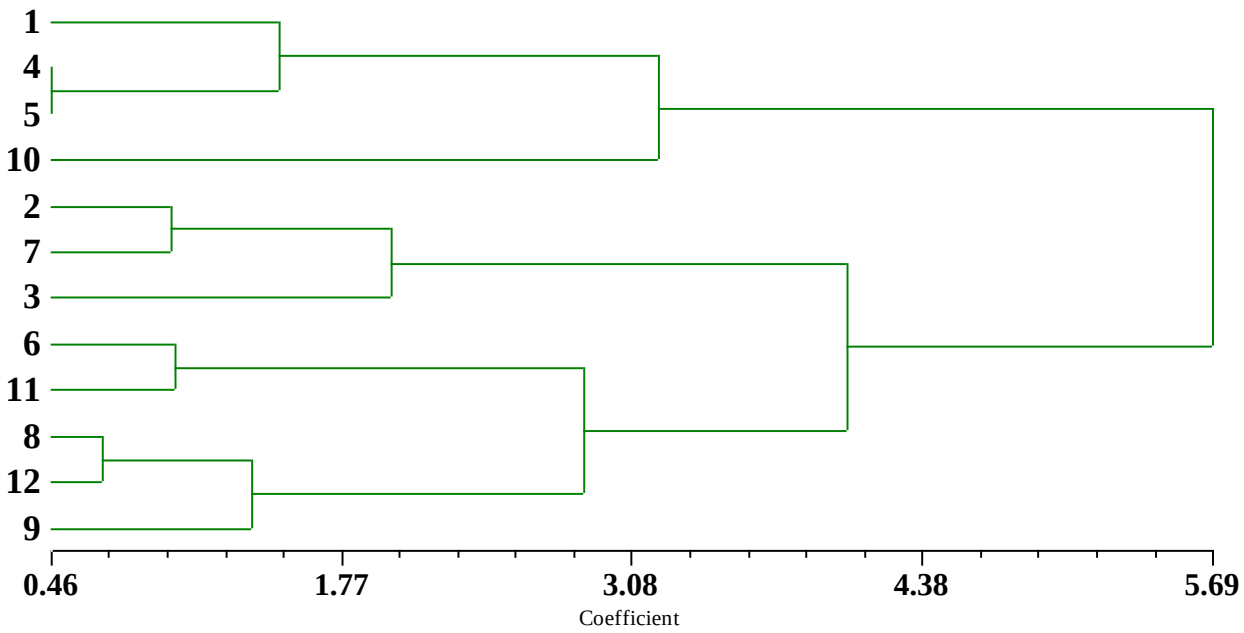
Απόσταση												
0,46	4	5										
0,68	8	12										
0,99	2	7										
1,02	6	11										
1,13	8	12	6	11								
1,31	8	12	6	11	9							
1,41	2	7	3									
1,43	4	5	1									
1,84	4	5	1	10								
1,96	8	12	6	11	9	4	5	1	10			
2,02	8	12	6	11	9	4	5	1	10	2	7	3



Μέθοδος της πλήρης διασύνδεσης ή απομακρυσμένης γειτνίασης

	1	2	3	4	5	6	7	8	9	10	11
2	4,00										
3	4,28	1,41									
4	**1,48**	3,24	3,89								
5	1,43	3,06	3,58	**0,46**							
6	3,79	3,06	2,02	3,95	3,62						
7	3,51	**0,99**	**1,99**	2,54	2,34	3,29					
8	2,87	3,31	3,04	2,98	2,84	1,92	3,32				
9	1,96	4,04	3,97	2,59	2,53	**2,85**	3,86	**1,36**			
10	1,84	5,57	**5,69**	3,14	**3,19**	4,72	5,21	3,58	2,24		
11	3,43	3,46	2,73	3,73	3,48	**1,02**	3,63	1,13	2,07	**4,09**	
12	2,99	3,79	3,46	3,36	3,23	2,13	3,86	**0,68**	1,31	3,42	1,17

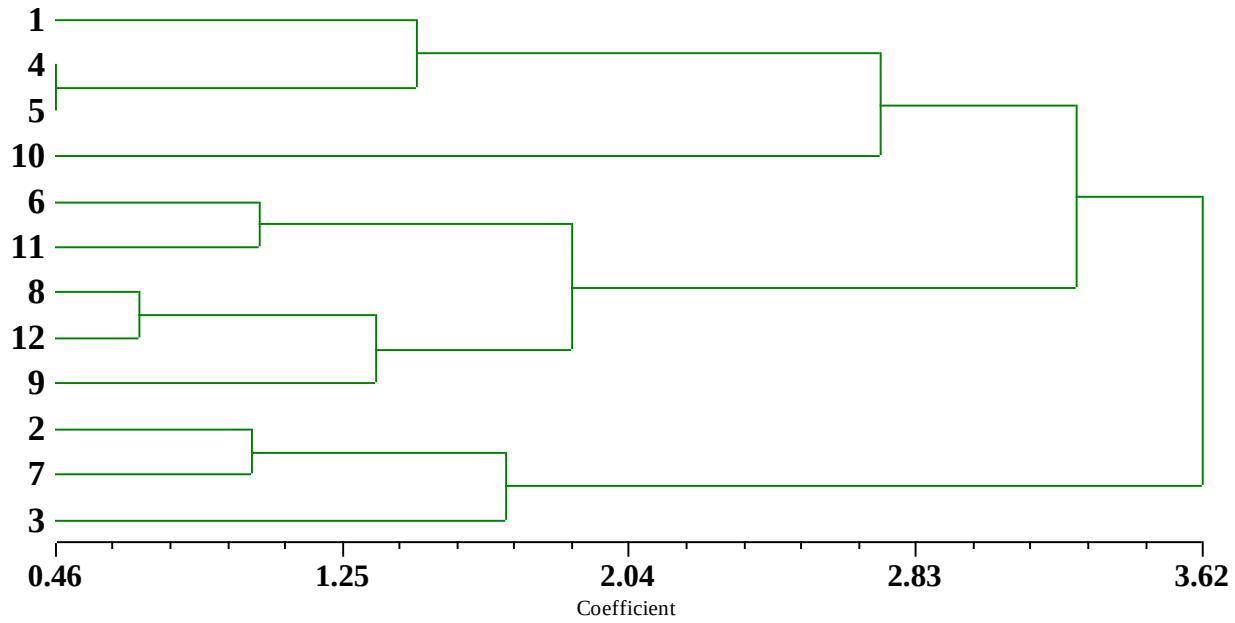
Απόσταση												
0,46	4	5										
0,68	8	12										
0,99	2	7										
1,02	6	11										
1,36	8	12	9									
1,48	4	5	1									
1,99	2	7	3									
2,85	8	12	9	6	11							
3,19	4	5	1	10								
4,09	8	12	6	11	9	4	5	1	10			
5,69	8	12	6	11	9	4	5	1	10	2	7	3



Μέθοδος της μη σταθμισμένης κατά ζεύγη μέσης διασύνδεσης (UPGMA)

	1	2	3	4	5	6	7	8	9	10	11
2	4,00										
3	4,28	1,41									
4	1,48	3,24	3,89								
5	1,43	3,06	3,58	0,46							
6	3,79	3,06	2,02	3,95	3,62						
7	3,51	0,99	1,99	2,54	2,34	3,29					
8	2,87	3,31	3,04	2,98	2,84	1,92	3,32				
9	1,96	4,04	3,97	2,59	2,53	2,85	3,86	1,36			
10	1,84	5,57	5,69	3,14	3,19	4,72	5,21	3,58	2,24		
11	3,43	3,46	2,73	3,73	3,48	1,02	3,63	1,13	2,07	4,09	
12	2,99	3,79	3,46	3,36	3,23	2,13	3,86	0,68	1,31	3,42	1,17

Απόσταση												
0,46	4	5										
0,68	8	12										
0,99	2	7										
1,02	6	11										
1,34	8	12	9									
1,46	4	5	1									
1,71	2	7	3									
1,88	8	12	9	6	11							
2,73	4	5	1	10								
3,27	8	12	6	11	9	4	5	1	10			
3,62	8	12	6	11	9	4	5	1	10	2	7	3



Αριθμητικό Παράδειγμα: 5 γονότυποι και 5 δυαδικές μεταβλητές

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	1	1	0

	Παρουσία	Απουσία
Παρουσία	a	b
Απουσία	c	d

G1/G2	Παρουσία	Απουσία
Παρουσία	1	1
Απουσία	2	1

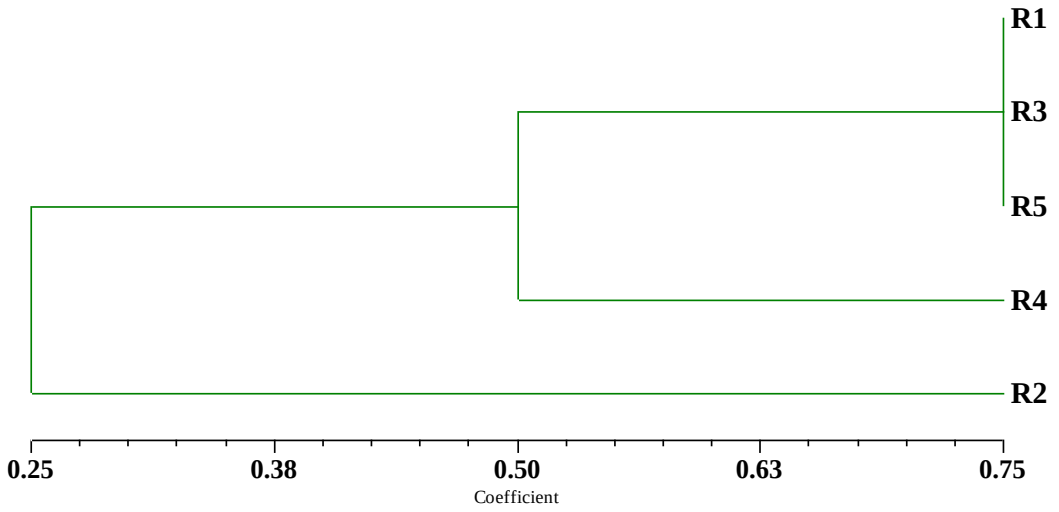
Συντελεστής Jaccard $S_{J(1,2)} = \frac{a}{a+b+c} = \frac{1}{1+1+2} = 0,25$

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

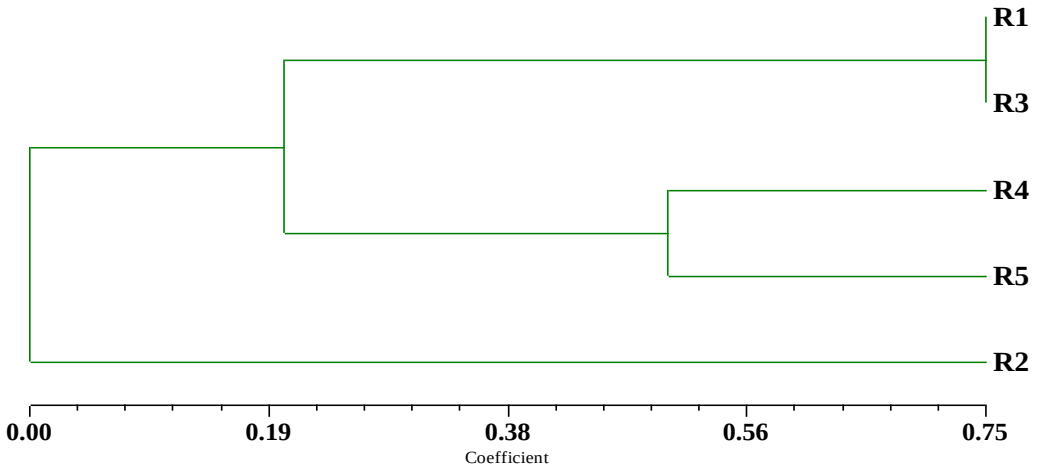
Συντελεστής Jaccard

	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	0,5	0	0,75	0,5	1

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



Μέθοδος απλής διασύνδεσης

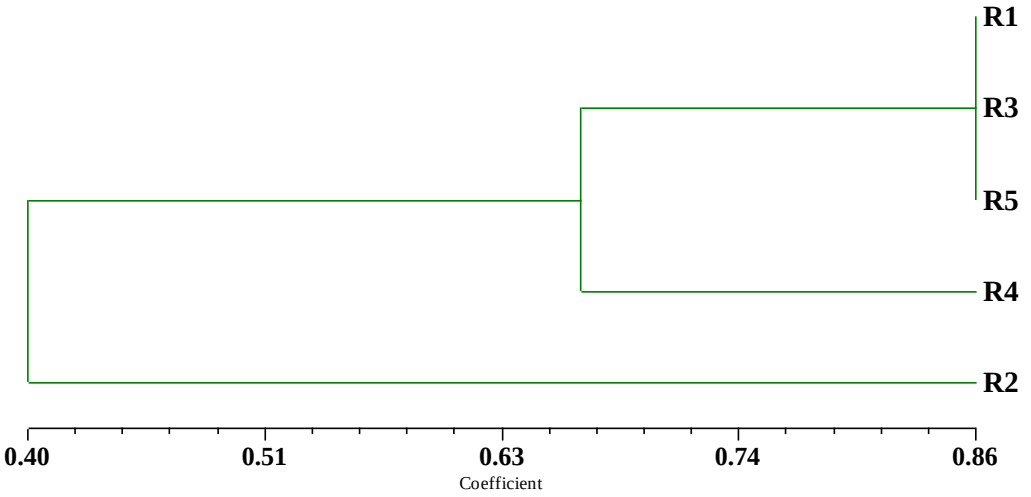


Μέθοδος Πλήρης διασύνδεσης

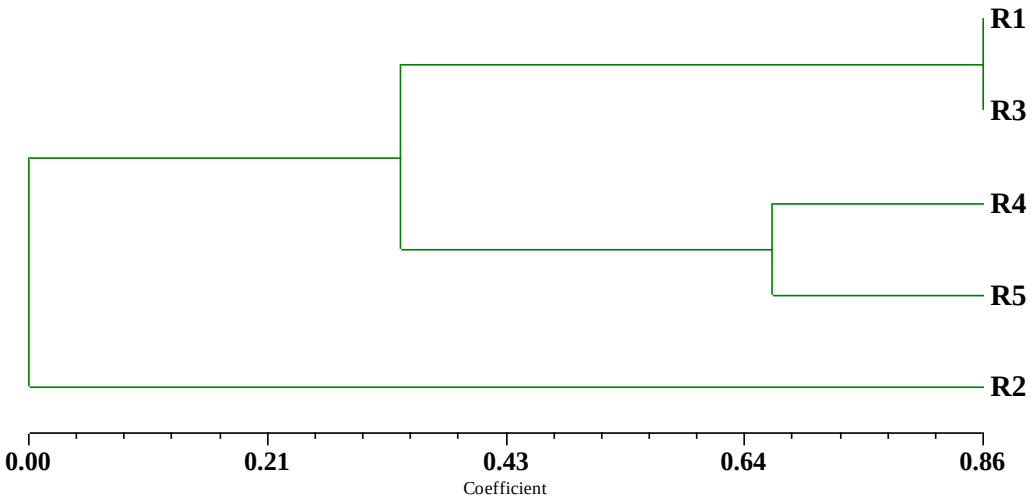
Συντελεστής Dice

	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	0,66	0	0,85	0,66	1

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



Μέθοδος απλής διασύνδεσης



Μέθοδος Πλήρης διασύνδεσης