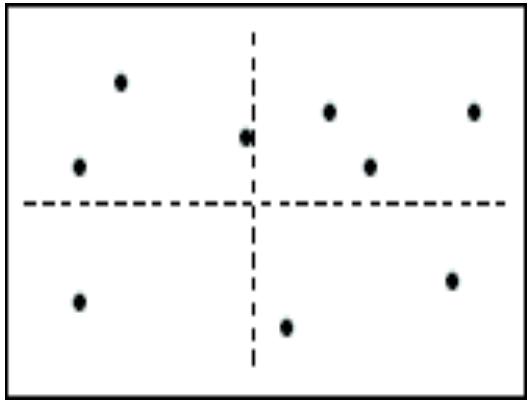
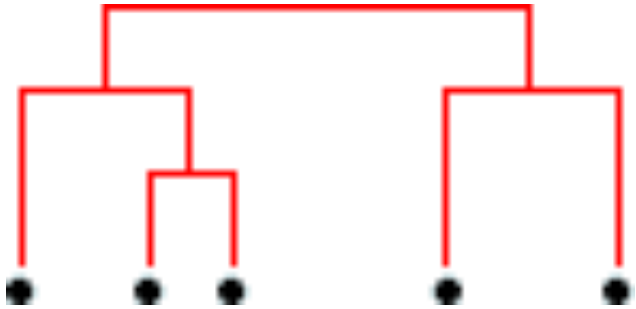
Η Γεωμετρική Ανάλυση Δεδομένων





Άγγελος Μάρκος (Επίκουρος Καθηγητής ΠΤΔΕ ΔΠΘ)

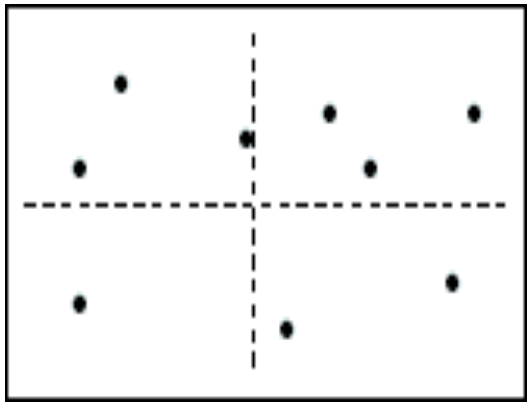
Η Γεωμετρική Ανάλυση Δεδομένων

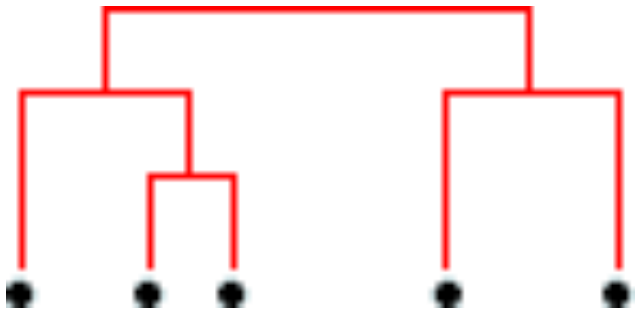
- Η Γεωμετρική Ανάλυση Δεδομένων (Geometric Data Analysis στα αγγλικά και L' Analyse des Données στα γαλλικά) αποτελεί *κλάδο της Πολυμεταβλητής Στατιστικής Ανάλυσης* και εκφράζει μια μεθοδολογική και φιλοσοφική προσέγγιση στη στατιστική συμπερασματολογία, η οποία έρχεται σε αντίθεση (και συχνά σε ρήξη) με την κλασική αγγλοσαξονική παράδοση του *στατιστικού ελέγχου υποθέσεων* (Καραπιστόλης 1999, Μπεχράκης 1999, Παπαδημητρίου 2007).

Η Γεωμετρική Ανάλυση Δεδομένων

Η ΓΑΔ απελευθερώνει τον ερευνητή από δεσμεύσεις, που ενδεχομένως επιβάλλουν εξωγενείς, σε σχέση με την έρευνα, παράγοντες, αφήνοντάς του τη φροντίδα και την ευθύνη να εξάγει ο ίδιος τις ερμηνείες των φαινομένων και τις συνέπειές τους (Benzécri & Collaborateurs, 1973).

Οι δύο βασικές μέθοδοι της ΑΔ είναι η **Παραγοντική Ανάλυση των Αντιστοιχιών** (Correspondence Analysis) και η **Ιεραρχική Ταξινόμηση** (Hierarchical Cluster Analysis), και περιλαμβάνονται σε όλα τα διαδεδομένα στατιστικά πακέτα, όπως SPSS, SAS, Stata, Statistica και R. Αλλά κ σε πιο εξειδικευμένα: CHIC Analysis (Markos et al., 2009).





Η Γεωμετρική Ανάλυση Δεδομένων

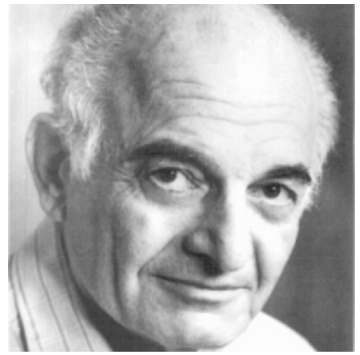
Σκοπός των μεθόδων της ΑΔ είναι να αναδείξουν και να περιγράψουν λανθάνουσες δομές που ενδεχομένως εμπεριέχονται σε πολυδιάστατους πίνακες δεδομένων. Αυτό επιτυγχάνεται μέσα από διαδικασίες αλλαγής και ελάττωσης των διαστάσεων του αρχικού χώρου, στον οποίο μπορεί να περιγραφεί το υπό εξέταση φαινόμενο.

Οι μέθοδοι, σε ένα πρώτο επίπεδο, δεν απαιτούν την *a priori* παραδοχή ύπαρξης κάποιας θεωρητικής κατανομής ή κάποια υπόθεση σχετικά με τις παραμέτρους του υπό εξέταση πληθυσμού ή πληθυσμών, δηλαδή την ύπαρξη κάποιου στοχαστικού υποδείγματος (Benzécri, 1991).

Οι μέθοδοι της ΑΔ εμφανίστηκαν και αναπτύχθηκαν ανεξάρτητα και, σε ορισμένες περιπτώσεις, σχεδόν ταυτόχρονα σε αρκετές χώρες, όπως οι Η.Π.Α., η Μεγάλη Βρετανία, ο Καναδάς, η Γαλλία, η Ολλανδία και η Ιαπωνία (Clausen, 1998).

Η Γεωμετρική Ανάλυση Δεδομένων - Ιστορική Ανασκόπηση

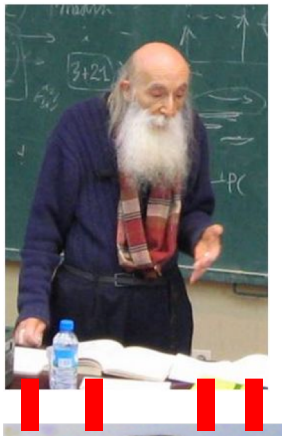
1900 – 1910 Karl Pearson



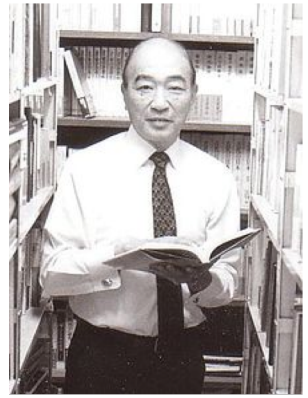
1940 Ronald Fisher (1940), Luis Guttman (1941)

1950 Cyril Burt (Αγγλία)

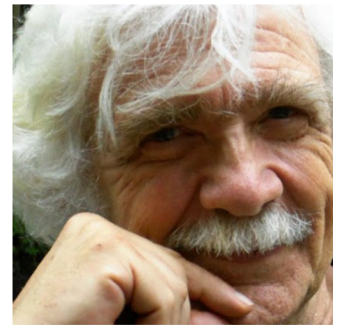
1960 Jean-Paul Benzécri (Γαλλία)



Chikio Hayashi (Ιαπωνία)

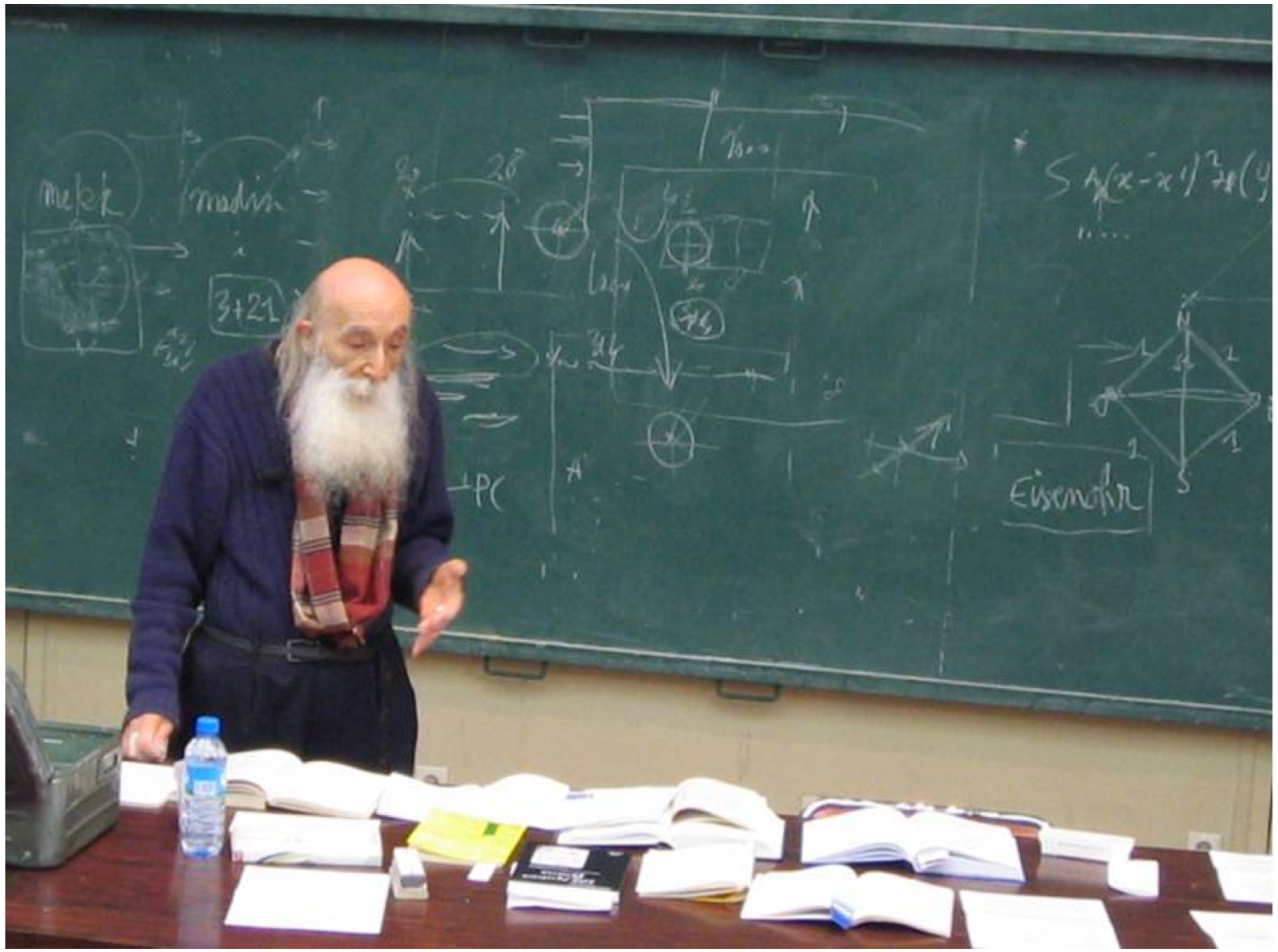


Jan De Leeuw (Ολλανδία)



1984 – Γιάννης Παπαδημητρίου (Ελλάδα)

Jean-Paul Benzécri (1932 –)
Ο Δάσκαλος του Δασκάλου μου

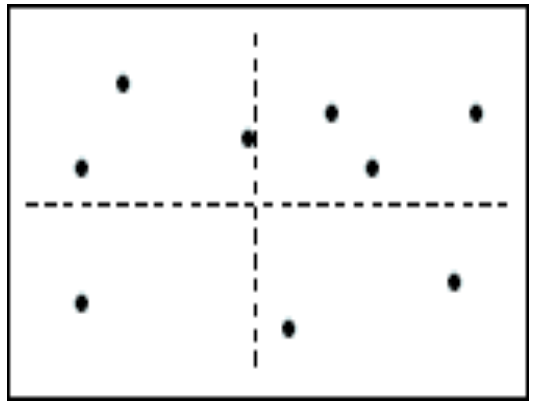


Σε αυτό το μάθημα επικεντρωνόμαστε σε δύο μεθόδους της Ανάλυσης Δεδομένων

Μέθοδος που αποκαλύπτει συνεχείς δομές (παράγοντες, άξονες, διαστάσεις…)

Παραγοντική Ανάλυση των Αντιστοιχιών

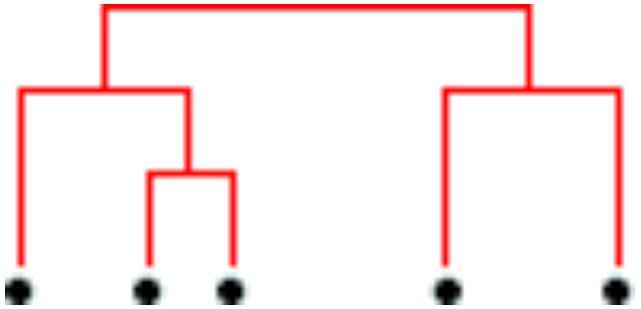
(Correspondence Analysis)



Μέθοδος που αποκαλύπτει διακριτές δομές (συστάδες, ομάδες, διαμερίσεις…)

<u>Ιεραρχική Ταξινόμηση</u>

(Hierarchical Cluster Analysis)



Κεντρική Έννοια: **Απόσταση**

Ένα Παράδειγμα

- Ένας ερευνητής μελέτησε το φαινόμενο των διακοπών ενός δείγματος 138 φοιτητών με τις 3 παρακάτω ερωτήσεις-μεταβλητές:

Ερώτηση/ Μεταβλητή Α (είδος διακοπών)	**Ερώτηση/ Μεταβλητή Β (επάγγελμα πατέρα)**	**Ερώτηση/ Μεταβλητή Γ (φύλο)**
Α1: Ξενοδοχείο (1) Α2: Οργανωμένη εκδρομή (2) Α3: Ενοικιαζόμενο δωμάτιο (3) Α4: Κάμπινγκ οργανωμένο (4) Α5: Ελεύθερο κάμπινγκ (5) Α6: Δεν πήγα διακοπές (6) Α7: Στο εξοχικό της οικογένειας (7)	Β1: Μισθωτός (1) Β2: Ελ.επ. επιστήμονας (2) Β3: Συνταξιούχος (3) Β4: Εργάτης, τεχνίτης, αγρότης (4) Β5: Εισοδηματίας (5) Β6: Ελεύθερος επαγγελματίας (6)	Γ1: Άντρας (1) Γ2: Γυναίκα (2)

Αρχικός Πίνακας Δεδομένων

- Ο αρχικός πίνακας δεδομένων που περιγράφει το φαινόμενο των διακοπών 138 φοιτητών με τις 3 μεταβλητές.

α/α	Α	Β	Γ
1	1	2	1
2	2	2	1
3	3	1	2
4	3	3	2
5	4	3	2
6	4	1	2
.	.	.	.
136	6	6	2
137	4	1	1
138	4	1	2

Κατασκευή Λογικού Πίνακα 0-1

1. Δημιουργούμε μία στήλη για κάθε κατηγορία της κάθε μεταβλητής

Μετ. Α	Μετ. Β	Μετ. Γ
⇨7 κατηγορίες	⇨6 κατηγορίες	⇨2 κατηγορίες

2. Μεταφορά κάθε γραμμής του αρχικού πίνακα δεδομένων σε ένα νέο πίνακα (0-1). Κάθε αριθμός γίνεται σχετική θέση στον 0-1, ως εξής:

Α/Α	**Α**	**Β**	**Γ**
1	1	2	1

	1								**2**					**1**	
	Α1	**Α2**	**Α3**	**Α4**	**Α5**	**Α6**	**Α7**	**Β1**	**Β2**	**Β3**	**Β4**	**Β5**	**Β6**	**Γ1**	**Γ2**
1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0

Ο Λογικός Πίνακας 0-1

	Α1	Α2	Α3	Α4	Α5	Α6	Α7	Β1	Β2	Β3	Β4	Β5	Β6	Γ1	Γ2
1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0
2	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0
3	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1
4	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
136	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1
137	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0
138	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1

Κατασκευή του Πίνακα Burt

	A1	A2	A3	A4	A5	A6	A7	B1	B2	B3	B4	B5	B6	Γ1	Γ2
1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0

	A1	A2	A3	A4	A5	A6	A7	B1	B2	B3	B4	B5	B6	Γ1	Γ2
A1	**12**	0	0	0	0	0	0	0	8	0	0	4	0	9	3
A2	0	**6**	0	0	0	0	0	0	4	0	0	2	0	2	4
A3	0	0	**38**	0	0	0	0	16	0	10	6	0	6	21	17
A4	0	0	0	**34**	0	0	0	14	0	10	4	0	6	18	16
A5	0	0	0	0	**10**	0	0	4	4	0	0	0	2	4	6
A6	0	0	0	0	0	**24**	0	8	0	4	0	0	12	14	10
A7	0	0	0	0	0	0	**14**	6	0	2	6	0	0	8	6
B1	0	0	16	14	4	8	6	**48**	0	0	0	0	0	24	24
B2	8	4	0	0	4	0	0	0	**16**	0	0	0	0	10	6
B3	0	0	10	10	0	4	2	0	0	**26**	0	0	0	20	6
B4	0	0	6	4	0	0	6	0	0	0	**16**	0	0	8	8
B5	4	2	0	0	0	0	0	0	0	0	0	**6**	0	3	3
B6	0	0	6	6	2	12	0	0	0	0	0	0	**26**	11	15
Γ1	9	2	21	18	4	14	8	24	10	20	8	3	11	**76**	0
Γ2	3	4	17	16	6	10	6	24	6	6	8	3	15	0	**62**

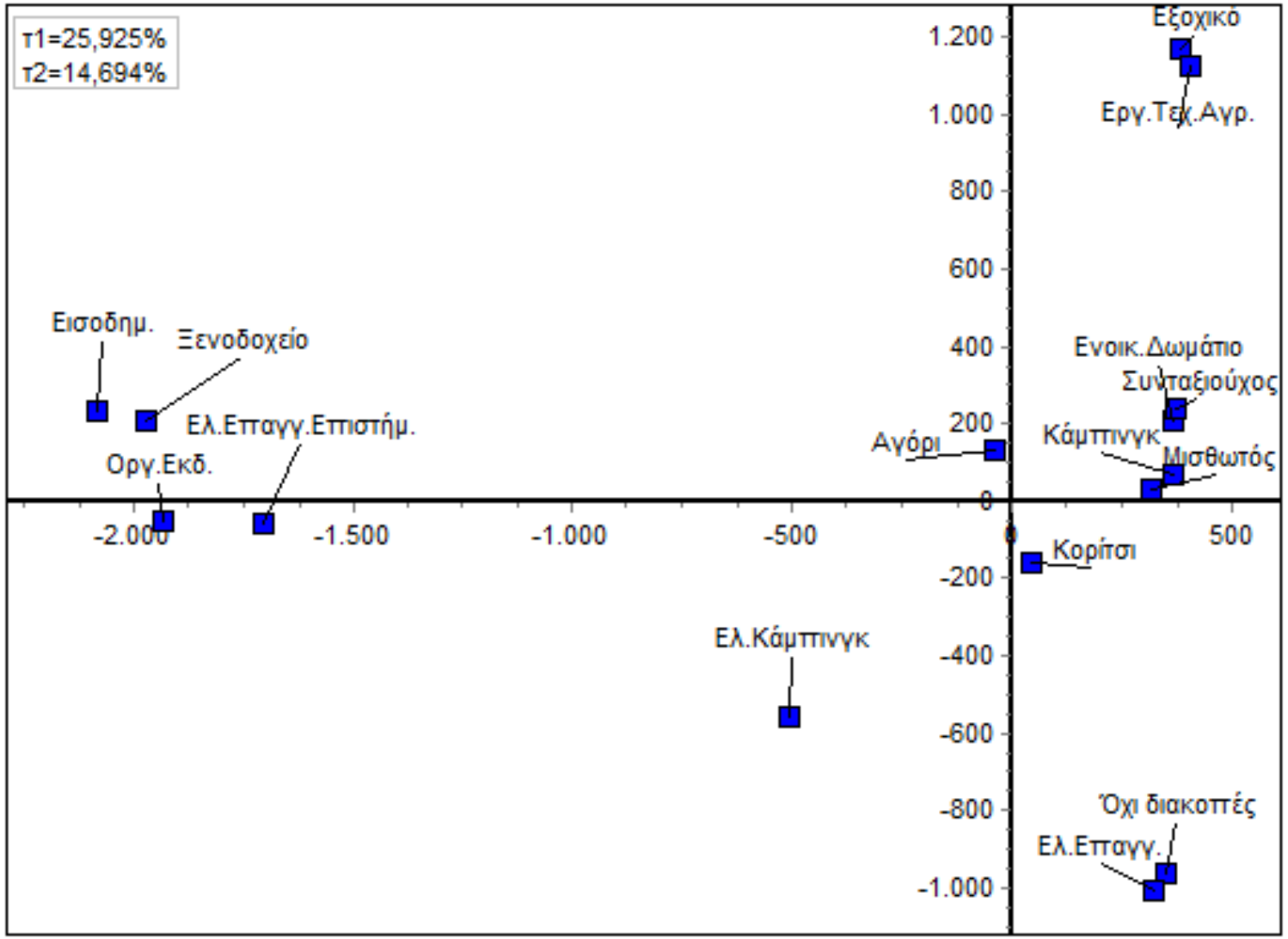
Το Παραγοντικό Επίπεδο 1x2 (σημεία στηλών)

- Σκοπός της Παραγοντικής Ανάλυσης των Αντιστοιχιών είναι να απεικονίσει τα σημεία γραμμών (άτομα) και στηλών (κατηγορίες) του πίνακα εισόδου ως σημεία σε ένα χώρο – συνήθως δύο διαστάσεων - με τη μικρότερη δυνατή απώλεια πληροφορίας.

- Σημεία που βρίσκονται «κοντά» μεταξύ τους στο επίπεδο, συνιστούν ομάδες ατόμων με παρόμοιο προφίλ απαντήσεων (και το αντίστροφο).

- Σκοπός του ερευνητή είναι να ερμηνεύσει το περιεχόμενο της κάθε ομάδας, καθώς και να εντοπίσει αντιπαραθέσεις μεταξύ των ομάδων.

- Η ερμηνεία ξεκινάει από τους άξονες και επεκτείνεται στο επίπεδο.

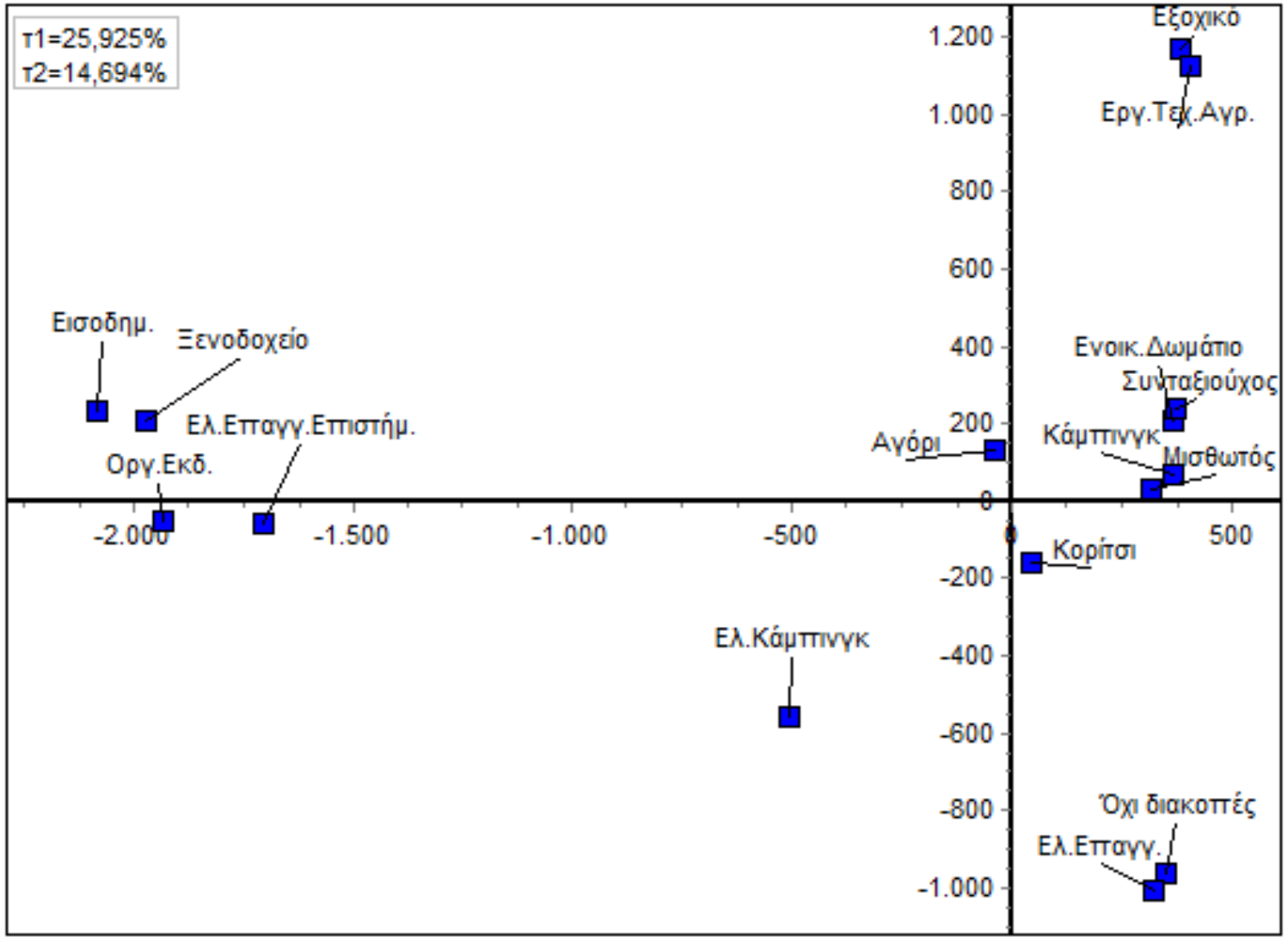


Παραγοντικό επίπεδο 1x2

Το Παραγοντικό Επίπεδο 1x2 (σημεία στηλών)

- Παρατηρήστε τα δύο ποσοστά (%) που εμφανίζονται επάνω και αριστερά στο γράφημα.

- Το άθροισμά τους είναι περίπου 40%, κάτι που δείχνει ότι το 40% της ολικής πληροφορίας (ή αδράνειας) του πίνακα Burt αναπαρίσταται σε αυτό το γράφημα.



Παραγοντικό επίπεδο 1x2

Πίνακας Ιδιοτιμών της Ανάλυσης του Πίνακα Burt

**** Εφαρμογή της Πολλαπλής Ανάλυσης Αντιστοιχιών στον Πίνακα Burt***

Άξονας	Αδράνεια	% Ερμηνείας	Αθρ. % Ερμηνείας	Ιστόγραμμα χαρακτ. ριζών
1	0.417	25.9	**25.9**	**********
2	0.237	14.6	**40.6**	**********
3	0.194	12.0	**52.6**	**********
4	0.158	9.7	**62.4**	**********
5	0.139	8.6	**71.0**	**********
6	0.112	6.9	**78.0**	**********
7	0.110	6.8	**84.9**	**********
8	0.091	5.1	**90.5**	**********
9	0.067	4.5	**94.7**	**********
10	0.051	2.2	**97.9**	**********
11	0.033	0.2	**99.9**	**********
12	0.000	0.2	**100**	**********

Συντεταγμένες και δείκτες ερμηνείας γραμμών/στηλών

** Εφαρμογή της Πολλαπλής Ανάλυσης Αντιστοιχιών στον Πίνακα Burt*

	#F1	COR	CTR	#F2	COR	CTR	#F3	COR	CTR	#F4	COR	CTR
ΞΕΝΟΔ	-1970	724	260	208	8	5	536	53	41	-115	2	2
ΟΡΓ.ΕΚ	-1932	405	125	-54	0	0	-586	37	24	-663	47	38
ΔΩΜΑΤ	368	143	29	207	45	16	110	12	5	72	5	3
ΚΑΜΠ	369	124	26	69	4	1	211	40	18	124	14	8
ΕΛ.ΚΑΜ	-510	55	14	-559	67	31	-982	207	119	1352	393	278
ΟΧΙ	347	65	16	-964	502	223	183	18	9	-420	95	63
ΕΞΟΧΙΚΟ	384	44	11	1164	403	189	-637	120	68	-364	39	27
ΜΙΣΘ	318	148	27	28	1	0	-237	82	33	235	81	40
ΕΛ.ΕΠ.ΕΠ	-1707	738	264	-59	0	0	-185	8	6	703	125	119
ΣΥΝΤ	371	85	20	241	36	15	908	512	263	198	24	15
ΕΡ.ΤΕΧ.ΑΓ	404	55	14	1125	432	203	-586	116	67	-477	77	54
ΕΙΣΟΔ	-2081	452	145	232	5	3	368	14	9	-1623	275	233
ΕΛ.ΕΠ	322	61	15	-1004	593	263	-84	4	2	-400	94	62
ΑΓΟΡΙ	-39	4	0	129	57	13	388	511	142	112	42	14
ΚΟΡΙΤΣΙ	47	4	0	-160	57	15	-477	511	174	-139	42	18

Δείκτες Ερμηνείας των Γραμμών / Στηλών

COR, το μέρος της μεταβλητότητας (αδράνειας) της γραμμής ή στήλης που εξηγεί κάθε παραγοντικός άξονας.

CTR (%), το ποσοστό της μεταβλητότητας (αδράνειας) που εξηγεί κάθε κατηγορία γραμμής ή στήλης σε κάθε άξονα.

Συνήθως θέλουμε (εμπειρικά):

$COR > 200$

και

$CTR > 1000/(\text{αριθμός κατηγοριών} + 1)$

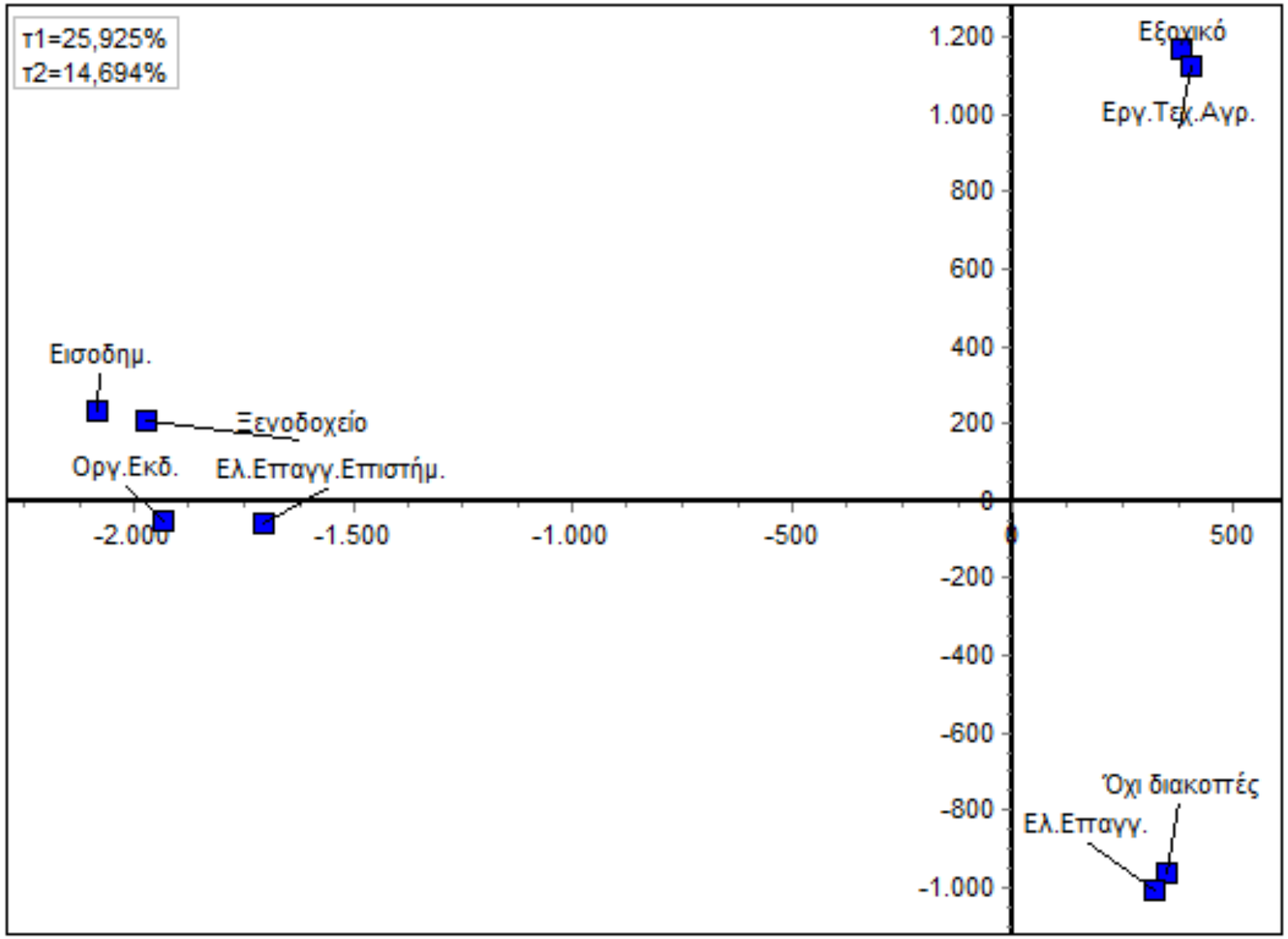
Συντεταγμένες και δείκτες ερμηνείας γραμμών/στηλών

Επιλέγουμε
COR > 200
και
CTR > 1000/(15+1) = 62,5

** Εφαρμογή της Πολλαπλής Ανάλυσης Αντιστοιχιών στον Πίνακα Burt*

	#F1	COR	CTR	#F2	COR	CTR	#F3	COR	CTR	#F4	COR	CTR
ΞΕΝΟΔ	-1970	724	260	208	8	5	536	53	41	-115	2	2
ΟΡΓ.ΕΚ	-1932	405	125	-54	0	0	-586	37	24	-663	47	38
ΔΩΜΑΤ	368	143	29	207	45	16	110	12	5	72	5	3
ΚΑΜΠ	369	124	26	69	4	1	211	40	18	124	14	8
ΕΛ.ΚΑΜ	-510	55	14	-559	67	31	-982	207	119	1352	393	278
ΟΧΙ	347	65	16	-964	502	223	183	18	9	-420	95	63
ΕΞΟΧΙΚΟ	384	44	11	1164	403	189	-637	120	68	-364	39	27
ΜΙΣΘ	318	148	27	28	1	0	-237	82	33	235	81	40
ΕΛ.ΕΠ.ΕΠ	-1707	738	264	-59	0	0	-185	8	6	703	125	119
ΣΥΝΤ	371	85	20	241	36	15	908	512	263	198	24	15
ΕΡ.ΤΕΧ.ΑΓ	404	55	14	1125	432	203	-586	116	67	-477	77	54
ΕΙΣΟΔ	-2081	452	145	232	5	3	368	14	9	-1623	275	233
ΕΛ.ΕΠ	322	61	15	-1004	593	263	-84	4	2	-400	94	62
ΑΓΟΡΙ	-39	4	0	129	57	13	388	511	142	112	42	14
ΚΟΡΙΤΣΙ	47	4	0	-160	57	15	-477	511	174	-139	42	18

Το Παραγοντικό Επίπεδο 1x2 (σημεία στηλών)
μετά την επιλογή των σημαντικών σημείων ως προς COR>200 και CTR>62,5



Παραγοντικό επίπεδο 1x2

Ένα άλλο παράδειγμα

Αντιλήψεις για την Επιστήμη

Πίνακας 2. Αντιλήψεις για την Επιστήμη

ID	A	B	C	D
1	2	3	4	3
2	3	4	2	3
3	2	3	2	4
4	2	2	2	2
5	3	3	3	3
...	...	...	...	...
199	3	4	2	3
200	1	2	2	2

Πηγή: International Social Science Survey, 1994 (Γερμανία)

Κλίμακα
1. Συμφωνώ απόλυτα
2. Συμφωνώ
3. Ούτε συμφωνώ / ούτε διαφωνώ
4. Διαφωνώ
5. Διαφωνώ απόλυτα

Σε ποιο βαθμό συμφωνείτε ή διαφωνείτε με τις παρακάτω προτάσεις;

Α. Εμπιστευόμαστε υπερβολικά τη θρησκευτική πίστη κι όχι αρκετά (όσο θα έπρεπε) την επιστήμη.

Β. Γενικά, η επιστήμη σήμερα κάνει περισσότερο κακό παρά καλό.

C. Η οποιαδήποτε αλλαγή προκαλεί ο άνθρωπος στη φύση – ασχέτως με το πόσο επιστημονική είναι ή όχι – θα χειροτερέψει τα πράγματα.

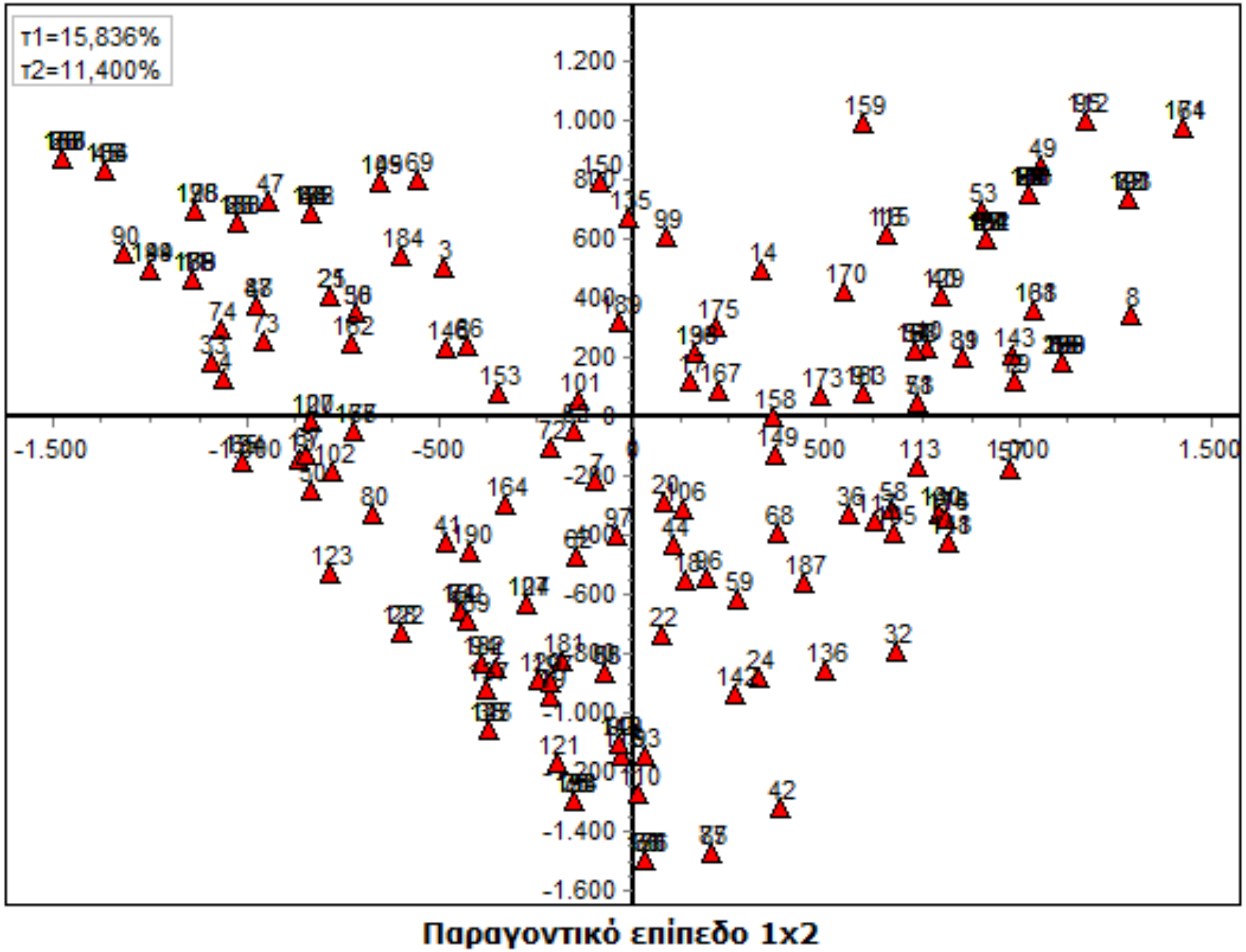
D. Η σύγχρονη επιστήμη θα λύσει τα περιβαλλοντικά μας προβλήματα με μικρές αλλαγές στον τρόπο ζωής μας.

Αποτελέσματα της Πολλαπλής Ανάλυσης Αντιστοιχιών στον Πίνακα 0-1

Πίνακας Ιδιοτιμών

Άξονας	Αδράνεια	%Ερμηνείας	Αθρ. %Ερμην.	Ραβδόγραμμα Χαρ. Ριζών
				**********
1	0,401	31,148	31,148	******
2	0,208	16,141	47,289	********
3	0,102	7,911	55,2	****
4	0,093	7,201	62,4	****
5	0,08	6,213	68,614	***
6	0,074	5,74	74,354	***
7	0,064	4,992	79,346	**
8	0,055	4,266	83,612	**
9	0,051	3,932	87,544	**
10	0,043	3,343	90,886	**
11	0,038	2,945	93,832	*
12	0,023	1,808	95,639	*
13	0,021	1,645	97,284	*
14	0,017	1,322	98,606	*
15	0,011	0,874	99,48	
16	0,007	0,52	100	

Παραγοντικό επίπεδο 1x2
(σημεία γραμμών)



Παραγοντικό επίπεδο 1x2
(σημεία στηλών)