Ποσοτικές Μέθοδοι Ανάλυσης στις Κοινωνικές Επιστήμες

Ενότητα 4: Ανάλυση κατά Συστάδες.

Θεόδωρος Χατζηπαντελής
Τμήμα Πολιτικών Επιστημών

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Ανάλυση κατά Συστάδες

Cluster Analysis

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Περιεχόμενα ενότητας

1. Ανάλυση κατά Συστάδες.
   i. Η απόσταση.
   ii. Τα βήματα.
   iii. Δενδρόγραμμα.
2. Παραδείγματα.

Ερωτήσεις 1

- **Ποια** είναι η κλίμακα μέτρησης των μεταβλητών;
- **Ποιες** μεταβλητές θα χρησιμοποιήσουμε;

Όπως έχουμε αναφέρει προηγουμένως ο πίνακας δεδομένων που αναλύουμε έχει Ν γραμμές και k στήλες. Οι γραμμές αντιστοιχούν σε υποκείμενα και οι στήλες σε μεταβλητές.

Ερωτήσεις 2

- Σε μια ανάλυση ομαδοποίησης των γραμμών **δεν είναι απαραίτητο** να συμμετέχουν όλες οι στήλες (δηλ. οι μεταβλητές).
- Επίσης είναι δυνατό να κάνουμε ανάλυση ομαδοποίησης **μόνο ενός υποσυνόλου των γραμμών**. (Αν για παράδειγμα έχουμε στοιχεία για φοιτητές ενός τμήματος μπορεί να επιλέξουμε μόνο όσους είναι στο Β' εξάμηνο ή να κάνουμε διαφορετικές αναλύσεις για κάθε εξάμηνο).

Ερωτήσεις 3

Στο παράδειγμα παρακάτω έχουμε ίδια κλίμακα μέτρησης στις μεταβλητές μας με τιμές {1,2,3,4} που ορίζουν διάταξη και την τιμή 9 (ΔΞ/ΔΑ), τότε προκύπτουν μεταβλητές σε κλίμακα μερικής διάταξης.

1. Ομαδοποίηση **αντικειμένων.**
2. Ομαδοποίηση **μεταβλητών.**

Δεδομένα

Πίνακας 1: Κλίμακα μέτρησης.

AA	SEX	E1_1	E1_2	E1_3	E1_4	E1_5	E1_6
1	2	2	3	1	1	3	3
2	2	**9**	1	1	1	1	1
3	2	1	1	2	2	3	1
4	2	1	2	**9**	2	2	2
5	2	1	1	1	1	2	1
6	1	2	1	1	1	3	2
7	1	2	4	3	3	3	**9**
8	1	1	1	1	1	2	3
9	2	2	1	3	1	2	2
10	1	1	2	4	4	2	1

Ποιος είναι ο κατάλληλος συντελεστής;

Block: Άθροισμα απόλυτων τιμών διαφοράς.

Κατασκευή συντελεστή (πχ σε πόσα ταυτίζονται).

Block 1

Πίνακας 2: Block.

	1	2	3	4	5	6	7	8	9	10
1		6	7	5	6	3	5	4	6	11
2	**6**		**4**	**4**	**1**	**3**	**9**	**3**	**4**	**8**
3	7	4		3	3	4	6	5	5	6
4	5	4	3		3	4	5	3	3	3
5	**6**	**1**	**3**	**3**		**3**	**9**	**2**	**4**	**7**
6	3	3	4	4	3		7	3	3	10
7	5	9	6	5	9	7		9	6	6
8	4	3	5	3	2	3	9		4	9
9	6	4	5	3	4	3	6	4		7
10	11	8	6	3	7	10	6	9	7	

Block 2

- Αποφασίζουμε να ενώσουμε το 2 με το 5 γιατί έχουν τη μικρότερη απόσταση.
- Στη συνέχεια προκύπτει ένας πίνακας 9Χ9 όπου οι αποστάσεις των υπολοίπων ομάδων από την ομάδα (2,5) υπολογίζονται είτε θέτοντας τη **μικρότερη** είτε τη **μεγαλύτερη** (ανάλογα με το κριτήριο που χρησιμοποιούμε). Στο παράδειγμα παίρνουμε τη μεγαλύτερη.

Block 3

Πίνακας 3: Block.

	1	(2,5)	3	4	6	7	8	9	10
1		**6**	**7**	**5**	**3**	**5**	**4**	**6**	**11**
(2,5)	6		4	4	3	9	3	4	8
3	7	4		3	4	6	5	5	6
4	5	4	3		4	5	3	3	3
6	3	3	4	4		7	3	3	10
7	5	9	6	5	7		9	6	6
8	**4**	**3**	**5**	**3**	**3**	**9**		**4**	**9**
9	6	4	5	3	3	6	4		7
10	11	8	6	3	10	6	9	7	

Block 4

- Αποφασίζουμε να ενώσουμε το 1 με το 6 γιατί έχουν τη μικρότερη απόσταση.
- Στη συνέχεια προκύπτει ένας πίνακας 8Χ8 όπου οι αποστάσεις των υπολοίπων ομάδων από την ομάδα (1,6) υπολογίζονται θέτοντας τη **μεγαλύτερη.**

Block 5

Πίνακας 4: Block.

	(1,6)	(2,5)	3	4	7	8	9	10
(1,6)		6	7	5	7	4	6	11
(2,5)	**6**		**4**	**4**	**9**	**3**	**4**	**8**
3	7	4		3	6	5	5	6
4	5	4	3		5	3	3	3
7	7	9	6	5		9	6	6
8	**4**	**3**	**5**	**3**	**9**		**4**	**9**
9	6	4	5	3	6	4		7
10	11	8	6	3	6	9	7	

Block 6

- Αποφασίζουμε να ενώσουμε το (2,5) με το 8 γιατί έχουν τη μικρότερη απόσταση.
- Στη συνέχεια προκύπτει ένας πίνακας 7Χ7 όπου οι αποστάσεις των υπολοίπων ομάδων από την ομάδα (2,5,8) υπολογίζονται θέτοντας τη **μεγαλύτερη.**

Συντελεστής ομοιότητας 1

Πίνακας 5: Συντελεστής Ομοιότητας.

	1	2	3	4	5	6	7	8	9	10
1		2	1	0	2	4	2	3	2	0
2	2		2	0	4	3	0	3	2	1
3	1	2		2	3	2	1	2	1	2
4	0	0	2		2	1	0	2	2	3
5	**2**	**4**	**3**	**2**		**3**	**0**	**5**	**3**	**3**
6	4	3	2	1	3		2	3	4	0
7	2	0	1	0	0	2		0	2	0
8	**3**	**3**	**2**	**2**	**5**	**3**	**0**		**3**	**2**
9	2	2	1	2	3	4	2	3		1
10	0	1	2	3	3	0	0	2	1	

Συντελεστής ομοιότητας Α

- Αποφασίζουμε να ενώσουμε το 5 με το 8 γιατί έχουν τη μικρότερη απόσταση.
- Στη συνέχεια προκύπτει ένας πίνακας 9Χ9 όπου οι αποστάσεις των υπολοίπων ομάδων από την ομάδα (5,8) υπολογίζονται είτε θέτοντας τη **μικρότερη** είτε τη **μεγαλύτερη** (ανάλογα με το κριτήριο που χρησιμοποιούμε). Τη μεγαλύτερη στην περίπτωση αυτή.

Συντελεστής ομοιότητας 2

Πίνακας 6: Συντελεστής Ομοιότητας.

	1	2	3	4	(5,8)	6	7	9	10
1		**2**	**1**	**0**	**2**	**4**	**2**	**2**	**0**
2	2		2	0	3	3	0	2	1
3	1	2		2	2	2	1	1	2
4	0	0	2		2	1	0	2	3
(5,8)	2	3	2	2		3	0	3	2
6	**4**	**3**	**2**	**1**	**3**		**2**	**4**	**0**
7	2	0	1	0	0	2		2	0
9	2	2	1	2	3	4	2		1
10	0	1	2	3	2	0	0	1	

Συντελεστής ομοιότητας Β

- Αποφασίζουμε να ενώσουμε το 1 με το 6 γιατί έχουν τη μικρότερη απόσταση.
- Στη συνέχεια προκύπτει ένας πίνακας 8Χ8 όπου οι αποστάσεις των υπολοίπων ομάδων από την ομάδα (1,6) υπολογίζονται θέτοντας τη **μεγαλύτερη**.

Συντελεστής ομοιότητας 3

Πίνακας 7: Συντελεστής Ομοιότητας.

	(1,6)	2	3	4	(5,8)	7	9	10
(1,6)		2	1	0	2	2	2	0
2	**2**		**2**	**0**	**3**	**0**	**2**	**1**
3	1	2		2	3	1	1	2
4	0	0	2		2	0	2	3
(5,8)	**2**	**3**	**3**	**2**		**0**	**3**	**2**
7	2	0	1	0	0		2	0
9	2	2	1	2	3	2		1
10	0	1	2	3	2	0	1	

Συντελεστής ομοιότητας Γ

- Αποφασίζουμε να ενώσουμε το 2 με το (5,8) γιατί έχουν τη μικρότερη απόσταση.
- Στη συνέχεια προκύπτει ένας πίνακας 7Χ7 όπου οι αποστάσεις των υπολοίπων ομάδων από την ομάδα (2,5,8) υπολογίζονται θέτοντας τη **μεγαλύτερη.**

Συντελεστής ομοιότητας 4

Πίνακας 8: Συντελεστής Ομοιότητας.

	(1,6)	(2,5,8)	3	4	7	9	10
(1,6)		2	1	0	2	2	0
(2,5,8)	2		2	0	0	2	1
3	1	2		2	1	1	2
4	**0**	**0**	**2**		**0**	**2**	**3**
7	2	0	1	0		2	0
9	2	2	1	2	2		1
10	**0**	**1**	**2**	**3**	**0**	**1**	

Συντελεστής ομοιότητας 5

Πίνακας 9: Συντελεστής Ομοιότητας.

	(1,6)	(2,5,8)	3	(4,10)	7	9
(1,6)		**2**	**1**	**0**	**2**	**2**
(2,5,8)	**2**		**2**	**0**	**0**	**2**
3	1	2		2	1	1
(4,10)	0	0	2		0	1
7	2	0	1	0		2
9	2	2	1	1	2	

Συντελεστής ομοιότητας 6

Πίνακας 10: Συντελεστής Ομοιότητας.

	(1,6,2,5,8)	3	(4,10)	7	9
(1,6,2,5,8)		**1**	**0**	**0**	**2**
3	1		2	1	1
(4,10)	0	2		0	1
7	0	1	0		2
9	**2**	**1**	**1**	**2**	

Συντελεστής ομοιότητας 7

Πίνακας 11: Συντελεστής Ομοιότητας.

	(1,6,2,5,8,9)	3	(4,10)	7
(1,6,2,5,8,9)		1	0	0
3	**1**		**2**	**1**
(4,10)	**0**	**2**		**0**
7	0	1	0	

Συντελεστής ομοιότητας 8

Πίνακας 12: Συντελεστής Ομοιότητας.

	(1,6,2,5,8,9)	(3, 4,10)	7
(1,6,2,5,8,9)		**0**	**0**
(3, 4,10)	**0**		**0**
7	0	0	

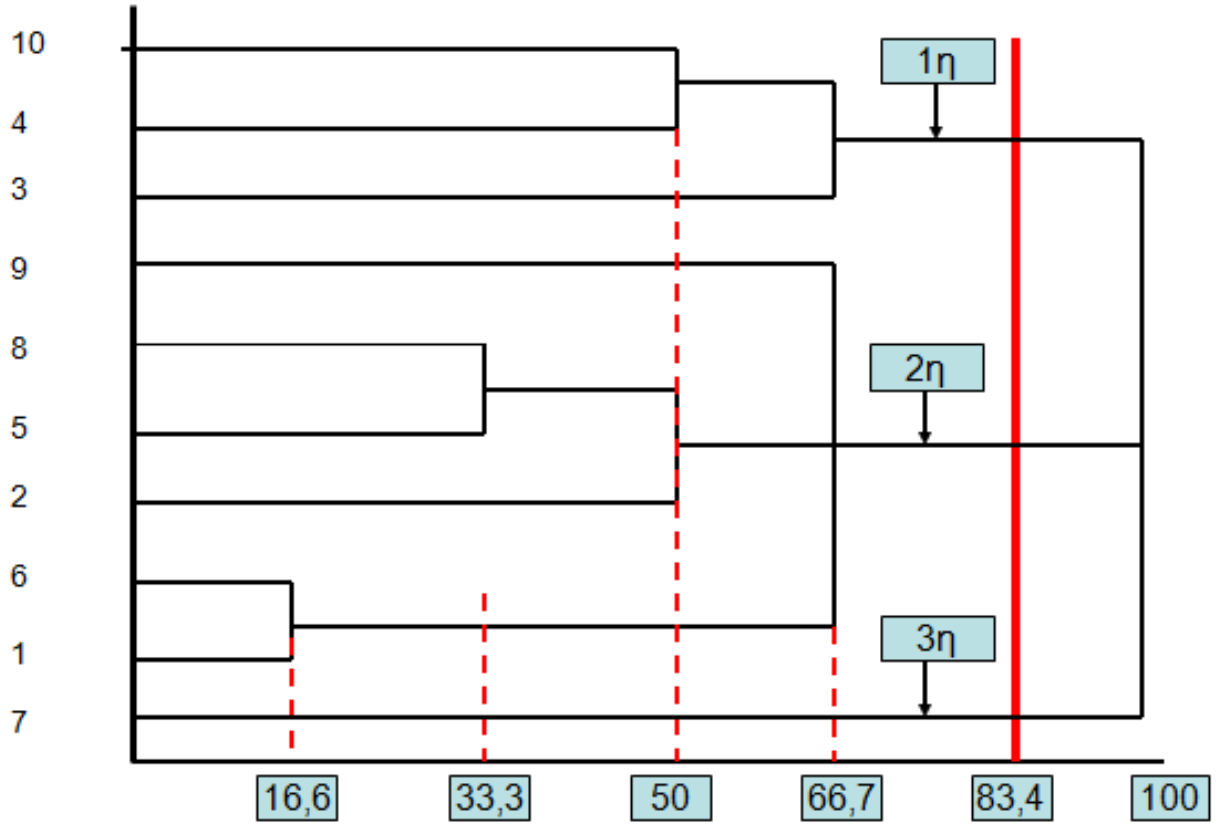
Σύνοψη Βημάτων

Πίνακας 13: Σύνοψη Βημάτων.

Βήμα	Κόμβος 1	Κόμβος 2	ομοιότητα	Σύνολο
1	1	6	5	16,6%
2	5	8	4	33,3%
3	2	(5,8)	3	50%
4	4	10	3	50%
5	(1,6)	(2,5,8)	2	66,7%
6	(1,6,2,5,8)	9	2	66,7%
7	3	(4,10)	2	66,7%
8	(1,6,2,5,8,9)	(3, 4,10)	0	100%
9	(1,6,2,5,8,9, 3,4,10)	7	0	100%

Δενδρόγραμμα

Γράφημα 1: η τομή.



Προσοχή

- Προσεκτική μελέτη των μεταβλητών. Πρέπει να γνωρίζουμε αν πρόκειται για **ποσοτικές** (συνεχείς ή διακριτές) ή **ποιοτικές** (διάταξης ή ονομαστικές).
- Πρέπει να αποφασίσουμε **πόσο** συνεισφέρει κάθε μεταβλητή στη διαφοροποίηση μεταξύ υποκειμένων.

Η μέθοδος (παράδειγμα για στήλες)

- Επιλέγουμε ένα υποσύνολο των μεταβλητών.
- Υπολογίζουμε ένα συντελεστή **ομοιότητας** ή **ανομοιότητας** χρησιμοποιώντας τις μεταβλητές για κάθε δυάδα υποκειμένων.
- Προκύπτει από τον (N x k) πίνακα δεδομένων ένας (N x N) πίνακας συντελεστών ομοιότητας ή ανομοιότητας.

Η μέθοδος

- Η μέθοδος αυτή στηρίζεται στη σωστή επιλογή συντελεστή. Οι συντελεστές που χρησιμοποιούμε καθορίζονται από τον τύπο των μεταβλητών.

Ποσοτικές μεταβλητές

Αν χρησιμοποιήσουμε μόνο **ποσοτικές μεταβλητές** τότε:

1. **Ευκλείδεια** απόσταση.
2. **Ευκλείδεια** απόσταση με **συνάρτηση βάρους.**
3. **City block** άθροισμα των απόλυτων τιμών των διαφορών.
4. **Συντελεστής συσχέτισης** (των **τυποποιημένων** τιμών).

Ευκλείδεια απόσταση

- Ας δούμε ένα παράδειγμα: Γνωρίζουμε για κάθε ΟΤΑ του Νομού Θεσσαλονίκης (ανάμεσα σε άλλα στοιχεία) το Μέσο φορολογητέο εισόδημα και το ποσοστό κατοίκων που έχουν ολοκληρώσει τριτοβάθμια εκπαίδευση.

Ένα παράδειγμα

- Οι γραμμές αντιστοιχούν σε Ο.Τ.Α του Νομού Θεσσαλονίκης.
- Οι στήλες σε μεταβλητές που περιγράφουν την κοινωνική σύνθεση των περιοχών.

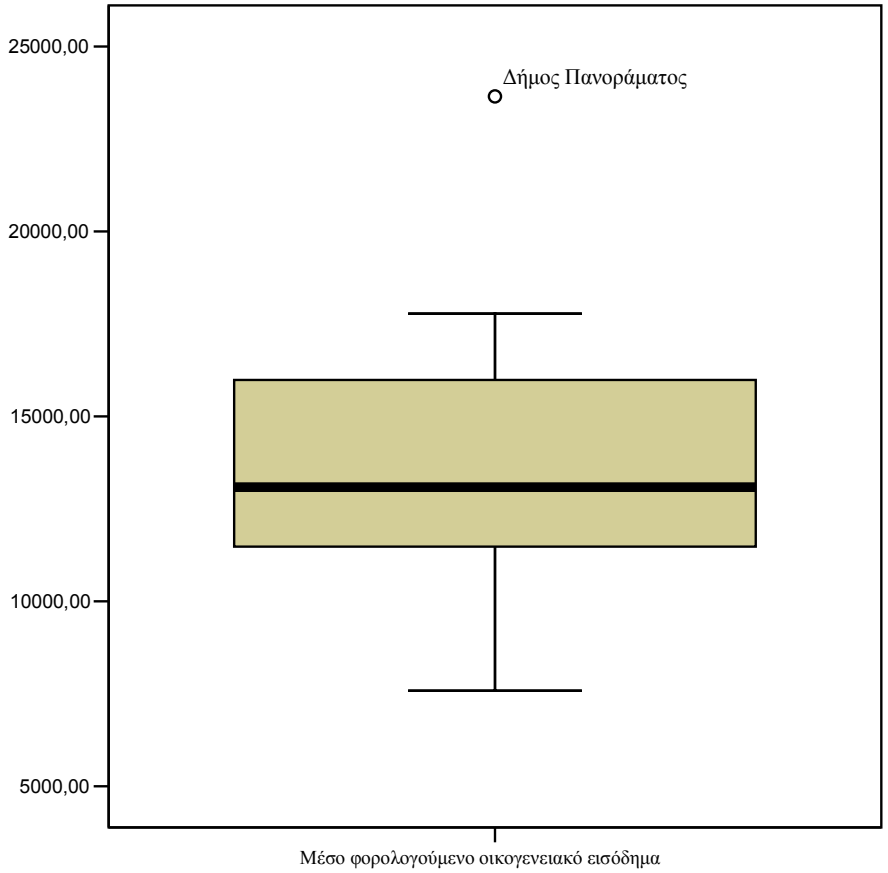
Πάράδειγμα 1

Πίνακας 14: ΟΤΑ.

ΟΤΑ	Μέσο μέγεθος νοικοκυριού (αριθμός ατόμων)	Μέση ηλικία (χρόνια)	Απασχολούμενοι εκτός Θεσ/νίκης (%)	Έγγαμοι (%)	Μέσο φορολογούμενο οικογενειακό εισόδημα	Μόρφωση Δημοτικού Γυμνασίου (%)	Ανώτατη Μόρφωση (%)	Οικονομικά Ενεργοί (%)	Άνεργοι (%)	Πρωτογενής τομέας (%)	Δευτερογενής τομέας (%)	Τριτογενής τομέας (%)	Αποχή	Άκυρα / Λευκά	Ν.Δ	ΠΑ.ΣΟ.Κ	Κ.Κ.Ε	ΛΑ.Ο.Σ	ΣΥΝ
Δήμος Θεσσαλονίκης	2,46	39,19	18,40	44,40	16.052	43,50	19,50	42,20	11,00	0,80	22,30	72,00	20,20	2,20	37,30	26,30	4,30	3,10	3,60
Δήμος Αγίου Παύλου	2,63	37,60	87,30	47,60	12.217	54,30	13,50	44,60	12,50	0,70	27,70	66,30	17,00	2,60	28,60	34,40	7,20	3,80	3,00
Δήμος Αμπελοκήπων	2,83	37,31	90,20	48,60	11.435	59,60	10,00	44,90	14,30	1,20	33,40	60,50	12,80	3,50	33,60	33,50	6,50	4,10	2,50
Δήμος Ελευθερίου Κορ	3,11	34,58	56,30	50,20	7.581	63,60	8,00	45,80	14,00	1,90	38,90	53,50	10,80	3,30	32,60	37,60	5,40	5,30	1,30
Δήμος Ευόσμου	3,01	33,02	90,70	51,20	15.720	60,40	9,80	45,80	12,50	1,20	33,80	59,00	10,80	3,30	37,90	31,40	6,00	4,80	2,10
Δήμος Καλαμαριάς	2,77	37,98	17,70	49,70	15.919	45,70	20,70	44,60	9,70	1,50	19,50	73,30	12,30	2,40	33,30	34,40	6,50	3,50	3,80
Δήμος Μενεμένης	3,12	34,77	28,30	47,60	12.498	68,90	6,00	42,40	14,80	1,20	34,00	58,10	10,70	3,20	34,80	37,80	4,30	4,20	1,70
Δήμος Νεαπόλεως	2,73	38,27	20,60	49,30	10.768	58,70	10,30	44,20	12,60	0,90	32,50	60,00	20,00	2,80	30,60	32,10	5,00	3,20	2,30
Δήμος Πολίχνης	3,02	34,86	46,60	50,70	10.098	62,30	9,40	44,30	12,90	1,10	34,90	58,40	10,00	2,90	33,10	35,20	8,50	4,00	2,50
Δήμος Σταυρουπόλεως	2,96	35,36	88,70	49,90	12.369	63,10	9,00	44,60	12,80	1,20	37,30	56,30	13,40	3,10	31,30	35,30	6,40	4,20	2,60
Δήμος Συκεών	2,88	36,60	24,40	50,20	8.324	58,50	11,70	44,90	12,70	0,70	29,90	62,80	17,90	2,50	28,10	34,20	7,20	3,50	2,80
Δήμος Τριανδρίας	2,64	37,22	18,70	45,80	11.520	50,50	14,70	42,30	11,70	0,60	25,00	71,00	11,20	2,30	36,70	32,00	7,30	3,50	3,50
Κοινότητα Ευκαρπίας	3,41	34,52	81,30	50,20	12.848	68,90	6,50	44,10	15,80	1,60	48,50	44,80	0,70	2,20	36,50	39,30	3,80	4,50	1,20
Κοινότητα Πεύκων	3,33	32,75	74,60	52,30	14.962	46,40	22,10	47,40	9,00	1,30	21,80	70,00	5,00	2,50	43,00	29,30	7,50	4,70	4,70
Δήμος Θέρμης	3,17	34,82	80,30	51,40	17.775	52,40	16,70	45,20	8,60	5,90	23,70	64,40	9,10	1,90	41,50	32,50	5,40	4,10	2,50
Δήμος Θερμαϊκού	2,90	35,12	92,50	51,10	16.957	49,70	16,30	46,20	10,80	3,40	24,20	66,70	0,80	3,00	38,10	35,00	4,20	4,40	2,60
Δήμος Εχεδώρου	3,24	34,59	44,00	50,10	13.327	69,90	5,80	42,80	10,60	12,00	37,00	44,30	10,40	2,40	39,30	34,30	5,10	3,70	1,70
Δήμος Πανοράματος	3,15	36,85	48,80	49,90	23.651	34,70	32,70	44,90	7,00	0,90	14,30	74,10	11,80	1,70	51,00	23,90	3,60	2,80	3,40
Δήμος Πυλαίας	3,00	36,54	85,70	50,60	17.343	47,50	20,60	44,70	9,80	1,20	21,70	71,60	9,70	3,00	40,50	28,40	7,60	3,90	3,60
Δήμος Μίκρας	3,30	37,90	43,70	52,20	13.749	55,00	15,90	41,50	8,00	10,60	21,20	61,40	7,60	2,40	41,80	31,80	5,90	4,10	3,60

Συνέχεια

Γράφημα 2: ΟΤΑ.



Παράδειγμα 2

Πίνακας 15: Εισόδημα.

Κλιμάκια εισοδήματος		Αριθμός φορολογικών δηλώσεων	Αριθμός ατόμων που περιλαμβάνονται στις φορ. Δηλώσεις	φορολογούμενο οικογενειακό εισόδημα	μέσο φορολογούμενο οικογ. εισόδημα
(σε εκατομμύρια δρχ.)	σε ευρώ (€)			σε ευρώ (€)	σε ευρώ (€)
< 1	< 2.934,70	681.087	848.758	831.761.217	1.221
1 έως 2	2.934,70 έως 5.869,41	783.399	1.042.514	3.457.914.446	4.414
2 έως 3	5.869,41 έως 8.804,11	1.070.130	1.478.506	7.749.924.014	7.242
3 έως 4	8.804,11 έως 11.738,81	665.396	1.008.589	6.751.516.179	10.147
4 έως 5	11.738,81 έως 14.673,51	483.944	766.359	6.353.885.987	13.129
5 έως 6	14.673,51 έως 17.608,22	354.119	585.416	5.687.697.955	16.062
6 έως 8	17.608,22 έως 23.477,62	440.710	767.739	8.909.169.443	20.215
8 έως 10	23.477,62 έως 29.347,03	243.573	443.799	6.374.701.247	26.172
10 έως 15	29.347,03 έως 44.020,54	306.324	577.941	10.813.540.151	35.301
> 15	> 44.020,54	149.366	282.085	10.096.285.429	67.594
ΣΥΝΟΛΟ		**5.178.048**	**7.801.706**	**67.026.396.068**	**12.944**

Μία μεταβλητή

- Αν θέλαμε να χωρίσουμε τους 20 ΟΤΑ μόνο σε σχέση με το μέσο εισόδημα θα χρησιμοποιούσαμε μόνο το εισόδημα και κατά συνέπεια θα είχαμε μαζί τους ΟΤΑ με «πλησιέστερο» εισόδημα στην ίδια ομάδα.

Δύο μεταβλητές;

- Αν όμως θέλουμε να ταξινομήσουμε τους ΟΤΑ με βάση δύο μεταβλητές (π.χ. εισόδημα και ανώτατη μόρφωση) πρέπει να σχηματίσουμε ένα σύνθετο δείκτη.

Δύο μεταβλητές

- Οι δύο μεταβλητές μας (ΜΦΕ και ΤΕ) για Θεσσαλονίκη και Καλαμαριά είναι:
- 16052 και 19,5.
- 15919 και 20,7.
- Η ευκλείδεια απόσταση είναι ίση με:
- $\sqrt{(16052-15919)^2+(19,5-20,7)^2}=133$.

Η ευκλείδεια απόσταση

- Η ευκλείδεια απόσταση επηρεάζεται από τους αριθμούς.
- Έτσι για το ΜΦΕ έχουμε

  (16052-15919)=133 ,
- και για το ΤΕ

  (20,7-19,5)=1,2,
- Άρα το ΜΦΕ εισόδημα μετράει κατά 133/(134,2) δηλαδή κατά 99% στον υπολογισμό της απόστασης.

Ίδια κλίμακα

- Για να διορθώσουμε το παραπάνω (να μετρούν το ίδιο ο δύο μεταβλητές) πρέπει να μετατρέψουμε τους αριθμούς στην ίδια κλίμακα. Για να το πετύχουμε αυτό πρέπει να μετασχηματίσουμε τις τιμές των μεταβλητών χρησιμοποιώντας για κάθε μία **το μέσο όρο της και την τυπική απόκλιση της.**

Μετασχηματισμοί

- **Z scores**

  Αφαιρώ από κάθε τιμή **τη μέση τιμή** (μέσο όρο) και διαιρώ με την τυπική απόκλιση.

Έτσι όλες οι τιμές (κατά στήλη) έχουν διασπορά (άρα και τυπική απόκλιση) ίση με 1 και μέση τιμή 0.

Οι τυποποιημένες τιμές

- Οι τυποποιημένες τιμές γίνονται:

  0,62 και 0,82,

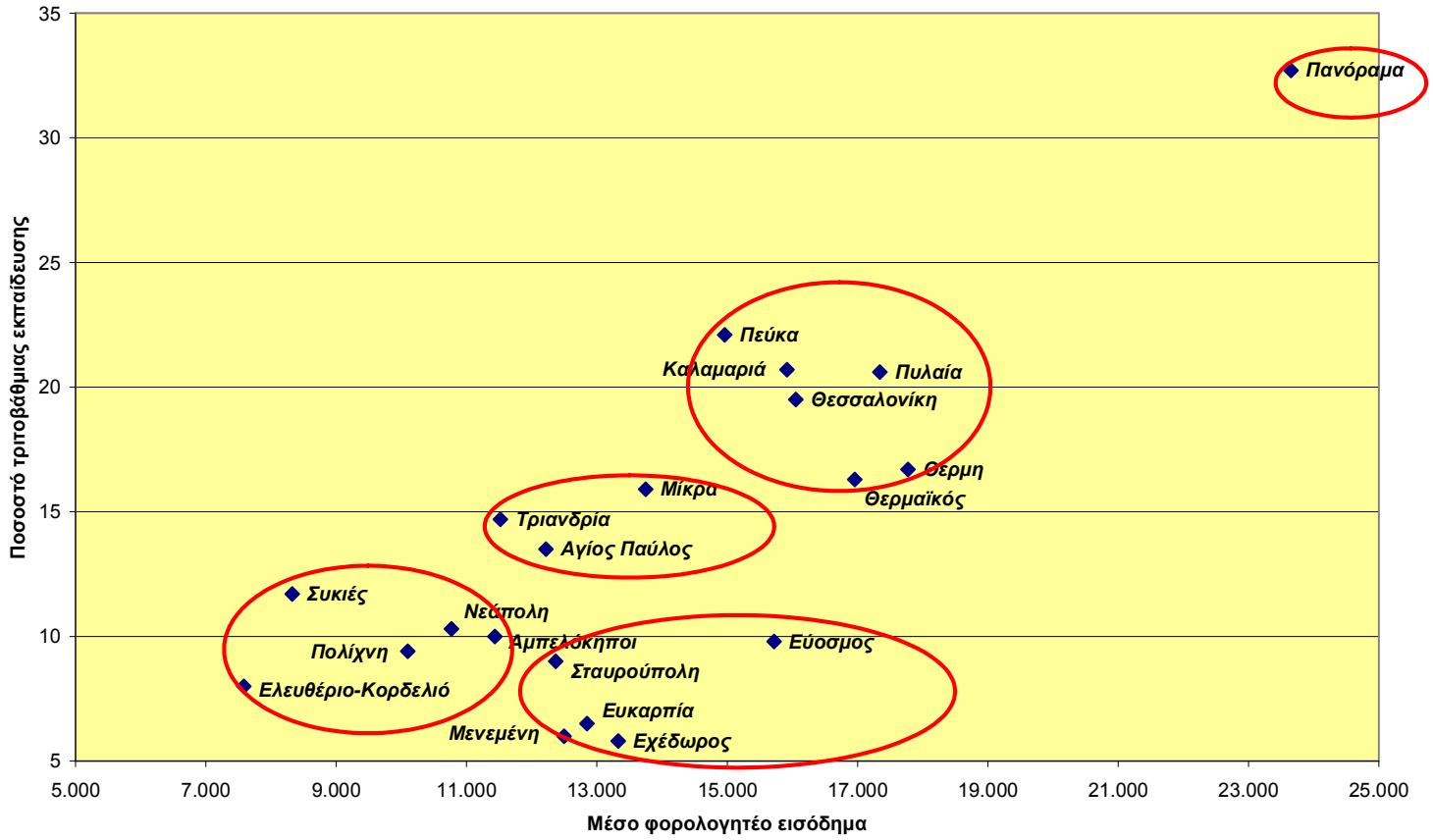
  0,58 και 0,99.

- Η απόσταση είναι τότε:
- $\sqrt{(0,62-0,58)^2+(0,82-0,99)^2}=0,18$.

  Δηλ. είναι 0.0016 για το εισόδημα και 0.0289 για το μορφωτικό επίπεδο και το εισόδημα υπολογίζεται στην απόσταση κατά 23%.

Ποσοστό Τριτοβάθμιας εκπαίδευσης/ Μέσο φορολογητέο εισόδημα

Γράφημα 3: Εκπαίδευση/εισόδημα.



Υπολογίζουμε

Maximum magnitude of 1

- Διαιρώ κάθε τιμή με την μέγιστη τιμή. **Έτσι έχουμε μέγιστη τιμή 1.**

Mean of 1

- Διαιρώ κάθε τιμή με το μέσο όρο. **Έτσι ο μέσος όρος γίνεται 1.**

Standard deviation of 1

- Διαιρώ κάθε τιμή με την τυπική απόκλιση.
  Έτσι η τυπική απόκλιση γίνεται 1.

Range

Range -1 to 1

- Διαιρώ κάθε τιμή με το εύρος. **Έτσι οι τιμές είναι από -1 μέχρι 1.**

Range 0 to 1

- Αφαιρώ την μικρότερη από όλες και διαιρώ με το εύρος. **Έτσι οι τιμές γίνονται από 0 μέχρι 1.**

Συντελεστές

- Ποιοτικές μεταβλητές: $\mathbf{X^{2.}}$

Προσοχή. Για να χρησιμοποιήσουμε στο SPSS τον συντελεστή Χ τετράγωνο πρέπει να δεδομένα μας να είναι στη μορφή πίνακα συχνοτήτων. Δηλαδή για κάθε περίπτωση να έχουμε μετρήσει τη συχνότητα εμφάνισης κάποιου χαρακτηριστικού. Παράδειγμα: τον αριθμό ψήφων που πήραν κόμματα σε εκλογές. Αν οι γραμμές αντιστοιχούν σε εκλογικές περιφέρειες και οι στήλες σε κόμματα τότε τα στοιχεία στον πίνακα είναι ο αριθμός ψήφων (δηλαδή απόλυτες συχνότητες).

Βήματα στην Cluster

- Απόσταση μεταξύ γραμμών.
- Κριτήριο σχηματισμού ομάδων.
- Κριτήριο εύρεσης λύσης.

Συντελεστής ομοιότητας ή ανομοιότητας;

- Αν ο πίνακας συντελεστών είναι συντελεστές **ομοιότητας** παίρνουμε το μεγαλύτερο ενώ αν είναι ανομοιότητας το **μικρότερο.**
- Στην περίπτωση της ευκλείδειας απόστασης (δηλαδή της συνήθους απόστασης) παίρνουμε το **μικρότερο**. Αν έχουμε X^2 παίρνουμε το **μεγαλύτερο**. Θυμηθείτε ότι ο συντελεστής X^2 δίνει τη σχέση δύο μεταβλητών. Όσο μεγαλύτερη είναι η τιμή του τόσο πιο συσχετισμένες είναι.

Ο σχηματισμός των ομάδων

- Ξεκινάμε από τη μικρότερη (μεγαλύτερη) απόσταση στον πίνακα αποστάσεων.
- Έτσι σχηματίζονται Ν-1 ομάδες (αν έχω Ν υποκείμενα).
- Στη συνέχεια υπολογίζουμε τις καινούριες αποστάσεις .

Νέες αποστάσεις

- Αν δύο υποκείμενα x και y έχουν συγκροτήσει μια ομάδα τότε το z απέχει από αυτά:
- 1. Min{d(x,z), d(y,z)} ή
- 2. Max{d(x,z), d(y,z)} ή
- 3. (d(x,z)+d(y,z))/2.

Μέθοδοι 1&2

- Βασίζονται μόνο στην διάταξη των αποστάσεων (όχι δηλαδή στην τιμή τους) Αυτό είναι μεγάλο πλεονέκτημα όταν οι αποστάσεις έχουν υπολογιστεί υποκειμενικά (όταν οι μεταβλητές είναι **ποιοτικές**).

Μέθοδος 3 & άλλες

- Αν οι αποστάσεις είναι μετρικές (δηλαδή οι μεταβλητές συνεχείς) είναι εύλογη η χρήση της 3.
- **Μέθοδος του Ward.** Σε κάθε βήμα σχηματίζονται όλες οι πιθανές δυάδες και επιλέγεται εκείνη που προκαλεί τη λιγότερη απώλεια πληροφορίας (επίσης για συνεχείς μεταβλητές).

Ένα ερωτηματολόγιο

Πίνακας 16: Ερώτηση Ε1.

***Ε1* Θα ήθελα να μου πείτε πόσο ευχαριστημένος είσαστε σχετικά με**

	Πάρα Πολύ	Σχετικά	Λίγο	Καθόλου	Δεν Ξέρω
Τις δυνατότητες διασκέδασης στην περιοχή	1	2	3	4	9
Την ζωή στην Θεσσαλονίκη	1	2	3	4	9
Τις σχέσεις που έχετε με τους γείτονες σας	1	2	3	4	9
Την οικογένεια σας	1	2	3	4	9
Τους καθηγητές σας	1	2	3	4	9
Την πολιτιστική κίνηση στην Θεσσαλονίκη	1	2	3	4	9

Αντικείμενο της ανάλυσης

- Ο σκοπός είναι να βρεθούν «ομάδες» μεταβλητών (δηλ. **κριτηρίων**) με κριτήριο την τοποθέτηση των φοιτητών στις ερωτήσεις.
- **Συσχέτιση** των μεταβλητών.
- Έχουμε 803 γραμμές (υποκείμενα) και 6 στήλες (μεταβλητές).

Απόσταση

- Σχηματίζουμε για κάθε δυάδα μεταβλητών τον πίνακα διπλής εισόδου. Υπάρχουν 6 μεταβλητές, άρα σχηματίζονται

$$\binom{6}{2} = \frac{6!}{2!*4!} = \frac{5*6}{1*2} = 15$$

πίνακες.

Πίνακας 17: E1.1/E1.2.

	1	2	3	4	9
1	45	31	57	33	4
2	26	40	85	59	7
3	23	39	132	105	2
4	10	8	33	39	3
9	3	1	2	4	12

Ο παραπάνω πίνακας είναι ο αντίστοιχος για την διασταύρωση της μεταβλητής ερ1.1 με την ερ1.2 από μια άλλη έρευνα.

Υπολογισμοί (πειραματικά μεγέθη)

Πίνακας 18: Υπολογισμοί.

	1	2	3	4	9	
1	45	31	57	33	4	170
2	26	40	85	59	7	217
3	23	39	132	105	2	301
4	10	8	33	39	3	93
9	3	1	2	4	12	22
	107	119	309	240	28	803

Υπολογισμοί (θεωρητικά μεγέθη)

Πίνακας 19: Υπολογισμοί.

	1	2	3	4	9	
1	22.7	25.2	65.4	50.8	5.9	170
2	28.9	32.2	83.5	64.9	7.6	217
3	40.1	44.6	115.8	90.0	10.5	301
4	12.4	13.8	35.8	27.8	3.2	93
9	2.9	3.3	8.5	6.6	0.8	22
	107	119	309	240	28	803

Υπολογισμός της X^2

Πίνακας 20: Υπολογισμοί.

	1	2	3	4	9	
1	22.0	1.3	1.1	6.2	0.6	
2	0.3	1.9	0.0	0.5	0.0	
3	7.3	0.7	2.3	2.5	6.9	
4	0.5	2.4	0.2	4.5	0.0	
9	0.0	1.6	4.9	1.0	164.5	
						233.4

Οι Αποστάσεις

- Παρατηρούμε ότι το μεγαλύτερο μέρος της απόστασης προέρχεται από το συνδυασμό (9,9). Έτσι σκεφτόμαστε να περιορίσουμε το δείγμα μας μόνο σε εκείνες της περιπτώσεις που δεν υπάρχει 9 σε κάποια από τις δύο μεταβλητές (δηλαδή εξαιρούμε τις 38 περιπτώσεις και κρατάμε για την ανάλυση τις υπόλοιπες 765).

Νέοι υπολογισμοί

Πίνακας 21: Υπολογισμοί.

	1	2	3	4	9	
1	22.3	1.1	1.4	6.5		
2	0.2	1.8	0.0	0.5		
3	7.7	1.1	1.2	1.8		
4	0.4	2.5	0.3	4.5		
9						
						53.3

Υπολογισμός των συντελεστών στο σύνολο του δείγματος

Υπολογισμός των συντελεστών στο σύνολο του δείγματος για κάθε δυάδα μεταβλητών.

Στο αρχικό παράδειγμα των 2000 περιπτώσεων μετά τους σχετικούς υπολογισμούς για το σύνολο του δείγματος μας προκύπτει ο παρακάτω πίνακας αποστάσεων.

Ανάλυση σε στήλες

Πίνακας 22: Υπολογίζουμε τους συντελεστές X^2.

	e1_1	e1_2	e1_3	e1_4	e1_5	e1_6
e1_1		520,48	75,42	29,35	37,01	236,01
e1_2			79,48	114,39	25,56	135,3
e1_3				72,47	182,44	63,65
e1_4					44,1	49,93
e1_5						102,05
e1_6						

Πρώτο βήμα

Πίνακας 23: Πρώτο βήμα.

	e1_1	e1_2	e1_3	e1_4	e1_5	e1_6
e1_1		**520,48**	75,42	29,35	37,01	236,01
e1_2	520,48		79,48	114,39	**25,56**	135,30
e1_3	75,42	79,48		72,47	182,44	63,65
e1_4	29,35	114,39	72,47		44,10	49,93
e1_5	37,01	**25,56**	182,44	44,10		102,05
e1_6	236,01	135,30	63,65	49,93	102,05	

Πρώτο βήμα (προετοιμασία για το δεύτερο)

- Επειδή η τιμή 520,48 είναι η μεγαλύτερη (αντιστοιχεί δηλαδή στη μικρότερο απόσταση) θα ενωθεί η μεταβλητή 1 με τη μεταβλητή 2 (ονομάζεται e1_12). Ο συντελεστής συνάφειας της e_12 με τις υπόλοιπες είναι ο μεγαλύτερος από τους προηγούμενους. Δηλαδή για παράδειγμα ο συντελεστής για (e1_12; e1_3) είναι ίσος με το max{(e1_1;e1,3),(e1_2;e1_3)}= max{75,42;79,48}=79,48.

Δεύτερο βήμα

Πίνακας 24: Δεύτερο βήμα.

		e1_12	e1_3	e1_4	e1_5	e1_6
e1_12			79,48	114,39	37,01	**236,01**
e1_3		79,48		72,47	182,44	63,65
e1_4		114,39	72,47		44,10	49,93
e1_5		37,01	182,44	44,10		102,05
e1_6		236,01	63,65	49,93	102,05	

Τρίτο βήμα

Πίνακας 25: Τρίτο βήμα.

		e1_126	e1_3	e1_4	e1_5	
e1_126			79,48	114,39	102,05	
e1_3		79,48		72,47	**182,44**	
e1_4		114,39	72,47		44,10	
e1_5		102,05	182,44	44,10		

Τέταρτο & Πέμπτο βήμα

Πίνακας 26: Τέταρτο βήμα.

	e1_126	e1_35	e1_4	
e1_126		102,05	114,39	
e1_35	102,05		72,47	
e1_4	114,39	72,47		
				114,39

Πίνακας 27: Πέμπτο βήμα.

	e1_1264	e1_35
e1_1264		102,05
e1_35	102,05	

Σύνοψη βημάτων

- Για κάθε βήμα γράφουμε ποιοι δύο κόμβοι συνδέονται και την αντίστοιχη συνάφεια (δηλαδή το μεγαλύτερο συντελεστή στον πίνακα με βάση τον οποίο αποφασίσαμε να ενώσουμε τους κόμβους). Όταν ολοκληρώσουμε τον πίνακα υπολογίζουμε τις αποστάσεις εκφράζοντας την απόσταση σαν ποσοστό της τελικής συνάφειας επί της συνάφειας κάθε βήματος. Πχ για το βήμα 3 διαιρούμε το 102,05 με το 182,44 και έχουμε 55,9%.

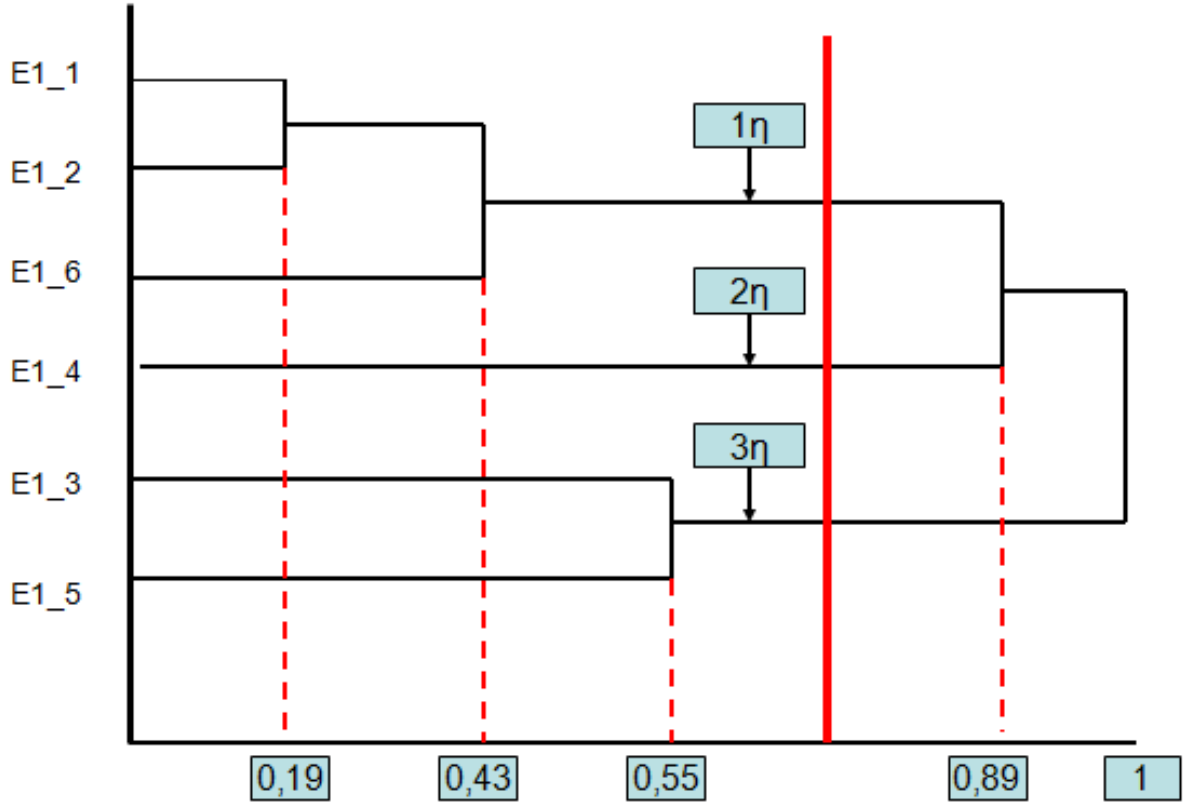
Σύνοψη βημάτων: Πίνακας

Πίνακας 28: Σύνοψη βημάτων.

Βήμα	Κόμβος 1	Κόμβος 2	Συνάφεια	Απόσταση	Αύξηση
1	e1_1	e1_2	520,48	0,196	0,196-0,000=0,196
2	e1_12	e1_6	236,01	0,432	0,432-0,196=0,236
3	e1_3	e1_5	182,44	0,559	0,559-0,432=0,127
4	e1_126	e1_4	114,39	0,892	0,892-0,559=0,333
5	e1_1264	e1_35	102,05	1,000	1,000-0,892=0,108

Το Δενδρόγραμμα

Γράφημα 4: η τομή.



Πόσες ομάδες;

- Επειδή από το 3ο στο 4ο βήμα παρατηρούμε την μεγαλύτερη αύξηση της απόστασης μπορούμε να υποθέσουμε ότι η καλύτερη διαίρεση είναι ανάμεσα σε αυτά τα βήματα. Άρα σχηματίζονται τρεις ομάδες {1,2,6},{4},{3,5}. Δηλαδή:
- {διασκέδαση, Θεσσαλονίκη, πολιτισμός}.
- {οικογένεια}.
- {καθηγητές, γείτονες}.

Διαδοχικά ερωτήματα

- Υπολογισμός απόστασης.
- Θα επιλέξουμε max, min ή μέσο όρο.
- Δενδρόγραμμα.

Γραμμές (παράδειγμα με κόμματα)

- Ομαδοποίηση εκλογικών ενοτήτων.
- Απόλυτες συχνότητες (άρα X^2).
- Μέθοδος max.

Παράδειγμα με κόμματα

Πίνακας 29: Σύνοψη βημάτων.

ΟΤΑ	egg	vot	ak_l	egk	pasok	nd	kke	syn	dhkki	laos	loipa
ΑΓ ΠΑΥΛΟΥ	5817	4836	154	4682	2002	1667	423	176	139	226	49
ΑΜΠΕΛΟΚΗΠΩΝ	23339	20367	811	19556	7839	7863	1520	585	524	971	254
ΕΛΕΥΘΕΡΙΟΥ-ΚΟΡΔΕΛΙΟΥ	15250	13616	513	13103	5747	4981	838	207	361	811	158
ΕΥΟΣΜΟΥ	26445	23596	878	22718	8325	10040	1600	572	625	1271	285
ΘΕΣΣΑΛΟΝΙΚΗΣ 1	4817	3500	125	3375	1251	1513	221	145	72	144	29
ΘΕΣΣΑΛΟΝΙΚΗΣ 11	8856	7222	194	7028	2355	3449	361	339	175	277	72
ΘΕΣΣΑΛΟΝΙΚΗΣ 12	7176	4660	109	4551	1289	2548	201	197	92	164	60
ΘΕΣΣΑΛΟΝΙΚΗΣ 13	7500	6205	192	6013	2297	2613	363	163	126	367	84
ΘΕΣΣΑΛΟΝΙΚΗΣ 14	7338	6065	181	5884	2192	2686	322	235	147	238	64
ΘΕΣΣΑΛΟΝΙΚΗΣ 15	7909	6100	210	5890	2328	2431	418	255	162	228	68
ΘΕΣΣΑΛΟΝΙΚΗΣ 16	5289	3996	142	3854	1389	1829	217	126	93	145	55
ΘΕΣΣΑΛΟΝΙΚΗΣ 17	7064	4914	120	4794	1342	2660	199	282	74	195	42
ΘΕΣΣΑΛΟΝΙΚΗΣ 18	10122	7778	301	7477	2800	3356	404	214	177	418	108
ΘΕΣΣΑΛΟΝΙΚΗΣ 19	33667	27926	830	27096	9081	13574	1307	1331	555	963	285
ΘΕΣΣΑΛΟΝΙΚΗΣ 2	6825	5568	160	5408	2039	2229	440	209	166	254	71
ΘΕΣΣΑΛΟΝΙΚΗΣ 22	6481	5046	130	4916	1509	2543	248	294	92	179	51
ΘΕΣΣΑΛΟΝΙΚΗΣ 23	4972	3116	45	3071	744	1919	82	181	47	76	22
ΘΕΣΣΑΛΟΝΙΚΗΣ 24	5800	4589	119	4470	1544	2116	227	300	80	150	53
ΘΕΣΣΑΛΟΝΙΚΗΣ 25	4753	3578	79	3499	1406	1502	190	130	82	138	51
ΘΕΣΣΑΛΟΝΙΚΗΣ 26	20927	18095	506	17589	5925	8606	923	855	414	681	185
ΘΕΣΣΑΛΟΝΙΚΗΣ 28	8719	7078	243	6835	2491	2993	437	275	210	340	89
ΘΕΣΣΑΛΟΝΙΚΗΣ 29	5147	3210	84	3126	994	1583	164	183	62	108	32
ΘΕΣΣΑΛΟΝΙΚΗΣ 3	10512	8895	238	8657	3253	3555	758	355	209	415	112
ΘΕΣΣΑΛΟΝΙΚΗΣ 30	6672	4980	140	4840	1829	2145	282	194	125	209	56
ΘΕΣΣΑΛΟΝΙΚΗΣ 4	6153	4271	92	4179	1132	2397	138	286	57	124	45
ΘΕΣΣΑΛΟΝΙΚΗΣ 5	16611	13101	353	12748	4103	6483	575	646	260	523	158
ΘΕΣΣΑΛΟΝΙΚΗΣ 7	29323	24950	684	24266	7372	12646	1258	1276	532	899	283
ΘΕΣΣΑΛΟΝΙΚΗΣ 9	16875	14482	379	14103	5119	5927	1204	627	426	611	189
ΚΑΛΑΜΑΡΙΑΣ	46606	40880	1117	39763	16070	15566	3072	1808	1138	1649	460
ΕΥΚΑΡΠΙΑΣ	4594	4197	101	4096	1809	1678	177	58	92	210	72
ΠΕΥΚΩΝ	3126	2971	78	2893	919	1346	236	148	71	150	23
ΜΕΝΕΜΕΝΗΣ	10523	9403	334	9069	3980	3663	456	179	232	445	114
ΝΕΑΠΟΛΕΩΣ	25605	20502	726	19776	8237	7857	1305	593	668	834	282
ΠΟΛΙΧΝΗΣ	20908	18825	604	18221	7378	6931	1781	534	530	850	217
ΣΤΑΥΡΟΥΠΟΛΕΩΣ	22513	19512	704	18808	7950	7067	1463	586	540	951	251
ΣΥΚΕΩΝ	26513	21791	667	21124	9075	7463	1923	750	716	954	243
ΤΡΙΑΝΔΡΙΑΣ	8424	7484	197	7287	2704	3092	623	297	195	297	79

Σημείωμα Χρήσης Έργων Τρίτων (1/2)

- Το Έργο αυτό κάνει χρήση των ακόλουθων έργων:
- Εικόνες/Σχήματα/Διαγράμματα/Φωτογραφίες
- Σχήμα 1-2: Παράδειγμα.
- Γράφημα 1-4: τα διαγράμματα που προέκυψαν από την ομαδοποίηση.

Σημείωμα Χρήσης Έργων Τρίτων (2/2)

- Το Έργο αυτό κάνει χρήση των ακόλουθων έργων:
- Πίνακες
- Πίνακας 1-29: Βήματα Ομαδοποίησης μεταβλητών και οι υπολογισμοί.

Σημείωμα Αναφοράς

Σημείωμα Αδειοδότησης

Τέλος ενότητας

Επεξεργασία: Σωτήρογλου Μαρίνα

Θεσσαλονίκη, Εαρινό Εξάμηνο 2014-2015

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημειώματα

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.