

ΓΙΑΝΝΗ ΠΑΠΑΔΗΜΗΤΡΙΟΥ – ΠΑΝΝΟΥΛΑΣ ΦΛΩΡΟΥ

ΣΥΜΒΟΛΗ ΤΗΣ ΕΥΚΛΕΙΔΕΙΑΣ ΚΑΙ χ^2 ΜΕΤΡΙΚΗΣ
ΣΤΟΝ ΠΡΟΣΔΙΟΡΙΣΜΟ ΤΗΣ ΙΔΑΝΙΚΟΤΕΡΗΣ ΤΑΞΙΝΟΜΗΣΗΣ
ΚΑΤΑ ΑΥΞΟΥΣΑ ΙΕΡΑΡΧΙΑ

ΔΙΑΓΡΑΜΜΑ

Περίληψη

1. Εισαγωγή

2. Τα δεδομένα

3. Αποστάσεις

4. Δημιουργία της ιεραρχίας

5. Δενδρόγραμμα ιεραρχίας

6. Κριτήριο για κάθε διαμελισμό σε r κλάσεις

7. Κριτήριο συνολικής ιεραρχικής ταξινόμησης

8. Επιλογή διαμελισμού για συγκεκριμένο πλήθος κλάσεων

9. Συμπεράσματα

Abstract

Βιβλιογραφία

ΠΕΡΙΛΗΨΗ

Σκοπός της εργασίας είναι να προσδιορισθεί ποια από τις τρεις μεθόδους της *ταξινόμησης κατά αύξουσα ιεραρχία (C.A.H.)*, (Min, Max, Moyenne), είναι ιδανικότερη για κάθε συγκεκριμένο πίνακα δεδομένων, ώστε η αδράνεια των κλάσεων κάθε διαμελισμού να πλησιάζει όσο γίνεται περισσότερο στην αρχική ολική αδράνεια. Επίσης επιχειρείται ο προσδιορισμός του ιδανικότερου διαμελισμού των δεδομένων για κάθε συγκεκριμένο πλήθος κλάσεων. Από τον πίνακα των αρχικών δεδομένων, υπολογίζεται ένας νέος πίνακας με τις αποστάσεις ανά δύο όλων των στοιχείων του. Για τον υπολογισμό των αποστάσεων, εφαρμόζεται η *ευκλείδεια μετρική*, αν οι μεταβλητές είναι ποσοτικές, και η *μετρική του X^2* , όταν αυτές είναι ποιοτικές ή όταν αναλύεται πίνακας συχνοτήτων, ενώ για τη δημιουργία του δένδρογράμματος και τον υπολογισμό των αποστάσεων μεταξύ των κλάσεων χρησιμοποιούνται οι τρεις μέθοδοι: Min, Max και Moyenne.

Αρχικά, προσδιορίζεται ποια από τις τρεις μεθόδους προσαρμόζεται καλύτερα στα δεδομένα. Αυτό επιτυγχάνεται με τον υπολογισμό του ποσοστού της ερμηνεύσιμης αδράνειας διαδοχικά για ένα πλήθος κλάσεων από n (πλήθος δεδομένων) μέχρι 2 (υψηλότερος κόμβος του δένδρογράμματος). Η μεταβολή του ποσοστού αυτού για κάθε μία από τις μεθόδους δίνεται γραφικά σε ένα κοινό διάγραμμα, όπου σχετικά εύκολα παρατηρείται η γενική εικόνα των μεταβολών του.

Στη συνέχεια, για κάθε επιθυμητό τελικό διαμελισμό, εξετάζεται σε ποια από τις τρεις μεθόδους διατηρείται το μεγαλύτερο ποσοστό της αρχικής ολικής αδράνειας, πράγμα που δηλώνει την καλύτερα προσαρμοζόμενη μέθοδο, για το συγκεκριμένο αριθμό κλάσεων.

Σημειώνεται ότι είναι δυνατό μία από τις τρεις μεθόδους να θεωρηθεί, σύμφωνα με τη γενική εικόνα της ιεράρχησης, η καλύτερα προσαρμοζόμενη στα δεδομένα, ενώ για κάποιο συγκεκριμένο αριθμό κλάσεων μια άλλη να ερμηνεύει μεγαλύτερο ποσοστό της αρχικής ολικής αδράνειας.

1. ΕΙΣΑΓΩΓΗ

Με την ταξινόμηση κατά αύξουσα ιεραρχία (C.A.H.) (Benzecri, *La taxinomie*, 1973), όπως και με όλες τις μεθόδους της Ανάλυσης Δεδομένων, επιχειρείται η απεικόνιση σε όσο το δυνατό απλούστερο σχήμα ενός πίνακα δεδομένων, του οποίου οι γραμμές είναι οι παρατηρήσεις (αντικείμενα ή μεταβλητές) που περιγράφονται από ένα σύνολο μεταβλητών που αντιστοιχούν στις στήλες του.

Αντικειμενικός σκοπός της ταξινόμησης είναι να επιτευχθεί διαμελισμός του συνόλου των παρατηρήσεων σε ομογενείς ομάδες, ως προς το σύνολο των μεταβλητών, κάθε μία των οποίων να διαφέρει σημαντικά από τις υπόλοιπες. Μ' αυτό τον τρόπο επιτυγχάνεται μια ιεράρχηση των παρατηρήσεων, δηλαδή μια σειρά διαμελισμών "ο ένας μέσα στον άλλο", που όσο προχωρά, τόσο πιο λεπτομερής γίνεται (Παπαδημητρίου, 1993).

Τα ερωτήματα όμως που τίθενται είναι:

1. Η χρησιμοποιούμενη μέθοδος δίνει ικανοποιητικά αποτελέσματα: Ποιο κριτήριο είναι εκείνο που θα μας βοηθήσει να αποφασίσουμε;
2. Σε ποιο σημείο της ιεράρχησης πρέπει να σταματήσουμε παίρνοντας τον ιδανικό διαμελισμό των δεδομένων, έτσι ώστε αφενός μεν τα αποτελέσματα να ανταποκρίνονται στην πραγματικότητα, και αφετέρου οι ομάδες του διαμελισμού να μην είναι πάρα πολλές;

Με την εφαρμογή της μεθόδου C.A.H., το σύνολο των παρατηρήσεων χωρίζεται αρχικά σε δύο ομάδες, με όσο γίνεται πιο ομογενή στοιχεία η καθεμιά. Κάθε μία απ' αυτές, στη συνέχεια, διασπάται σε δύο νέες ομάδες κατά το δυνατόν ομογενείς, και η διαδικασία συνεχίζεται μέχρι να καταλήξουμε στις αρχικές παρατηρήσεις. Δημιουργείται, δηλαδή, μια "ιεραρχία" από διαδοχικούς διαμελισμούς των παρατηρήσεων, όλο και καλύτερα "προσαρμοζόμενους" στο αρχικό σύνολο παρατηρήσεων (Roux, 1985).

Σχηματικά αυτό παριστάνεται με ένα "δενδρόγραμμα ιεραρχίας", του οποίου κάθε κόμβος παριστάνει μία ομάδα, που χωρίζεται στη συνέχεια σε δύο υποομάδες. Τερματικοί κόμβοι είναι οι αρχικές παρατηρήσεις, οι οποίες εμφανίζονται μόνο μία φορά.

Στην παρούσα εργασία ασχολούμαστε με πίνακες δεδομένων ποσοτικών και ποιοτικών μεταβλητών. Για τον υπολογισμό των αποστάσεων, όταν οι μεταβλητές είναι ποσοτικές, χρησιμοποιούμε την *ευκλείδεια μετρική*, ενώ όταν είναι ποιοτικές, χρησιμοποιούμε τη *μετρική του χ^2* (Benzecri, *L'analyse des correspondances*, 1973). Θεωρούμε ότι οι γραμμές του πίνακα δεδομένων αντιστοιχούν σε αντικείμενα ή παρατηρήσεις και οι στήλες στις μεταβλητές που τα χαρακτηρίζουν.

Για να υπολογίσουμε το πόσο "κοντά" βρίσκονται δύο ομάδες, χρησιμοποιούμε τρεις διαφορετικές μεθόδους, Min, Max, Moyenne (Diday κ.ά., 1982). Κάθε μία από τις μεθόδους αυτές καταλήγει σε διαφορετική ιεράρχηση των παρατηρήσεων. Δεν μπορούμε, γενικά, να συγκρίνουμε μεταξύ τους τις ιεραρ-

χρήσεις των μεθόδων αυτών, αφού μία μέθοδος που είναι ιδανική για ένα σύνολο δεδομένων, δεν είναι τόσο ικανοποιητική για κάποιο άλλο.

Η σύγκριση λοιπόν των μεθόδων μπορεί να πραγματοποιηθεί μόνο εμπειρικά, όταν εφαρμόζονται σ' ένα συγκεκριμένο σύνολο παρατηρήσεων. Για να επιλέξουμε την καταλληλότερη από τις τρεις μεθόδους για το συγκεκριμένο αυτό σύνολο, συγκρίνουμε την αδράνεια (διασπορά) των παρατηρήσεων μετά την ιεράρχηση, με την αρχική ολική αδράνεια (διασπορά) (Escofier κ.ά., 1990). Η μέθοδος κατά την οποία οι δύο αυτές τιμές βρίσκονται πλησιέστερα είναι η καταλληλότερη.

Πολλές φορές δεν ενδιαφερόμαστε για τη συνολική ιεράρχηση των παρατηρήσεων, αλλά επιθυμούμε απλώς να προσδιορίσουμε τον διαμελισμό τους στο ιδανικό πλήθος κλάσεων ή σε συγκεκριμένο πλήθος κλάσεων. Για κάθε μέθοδο, λοιπόν, και για κάθε διαφορετικό πλήθος κλάσεων υπολογίζουμε το ποσοστό της αρχικής ολικής αδράνειας που ερμηνεύει κάθε διαμελισμός, και με τη σύγκριση αυτών μεταξύ τους επιλέγουμε τον ιδανικότερο.

2. ΤΑ ΔΕΔΟΜΕΝΑ

Τα δεδομένα περιέχονται σε ένα πίνακα A διαστάσεων $n \times p$. Όταν μελετάμε ποσοτικές μεταβλητές, στον πίνακα $A(n, p)$ υπάρχουν οι τιμές των n παρατηρήσεων για τις p μεταβλητές. Όταν μελετάμε ποιοτικές μεταβλητές, αναλύουμε τον πίνακα A με στοιχεία 0 και 1, όπου το σύνολο των στηλών p αντιστοιχεί στο σύνολο των κλάσεων όλων των μεταβλητών ή τον πίνακα συχνοτήτων $A(n, p)$, στου οποίου στο στοιχείο $A(i, j)$ εμφανίζεται η συχνότητα συνύπαρξης της κλάσης i μιας μεταβλητής με την j μιας άλλης στον αρχικό πίνακα δεδομένων (παρατηρήσεις $\times$ μεταβλητές).

$$\begin{matrix} & 1 & \dots & j & \dots & p \\ \begin{matrix} 1 \\ \vdots \\ i \\ \vdots \\ n \end{matrix} & \left[\begin{array}{ccccc} \vdots & & & & \\ \dots & A(i, j) & \dots & & \\ \vdots & & & & \end{array} \right]
 \end{matrix}$$

3. ΑΠΟΣΤΑΣΕΙΣ

Από τον πίνακα των δεδομένων A διάστασης $n \times p$ υπολογίζουμε αρχικά έναν άλλο συμμετρικό πίνακα $B(n, n)$, που περιέχει τις αποστάσεις μεταξύ των παρατηρήσεων. Η μετρική που χρησιμοποιείται για τον υπολογισμό της από-

στασης μεταξύ δύο στοιχείων, i και i' , είναι η ευκλείδεια d_e^2 (Pages, 1976) στην περίπτωση ποσοτικών μεταβλητών και η του χ^2 d_x^2 (Escofier–Pages, 1990) όταν αυτές είναι ποιοτικές.

ευκλείδεια απόσταση:

$$d_e^2(i, i') = \sum_{j=1}^p [A(i, j) - A(i', j)]^2$$

χ^2 απόσταση:

$$d_x^2(i, i') = \sum_{j=1}^p \frac{1}{A(., j)} \left[\frac{A(i, j)}{A(i, .)} - \frac{A(i', j)}{A(i', .)} \right]^2$$

4. ΔΗΜΙΟΥΡΓΙΑ ΤΗΣ ΙΕΡΑΡΧΙΑΣ

Αρχικά βρίσκουμε τα δύο "πλησιέστερα" στοιχεία (αυτά που έχουν τη μικρότερη απόσταση μεταξύ τους), και στη θέση τους θέτουμε ένα νέο, με συντεταγμένες (τιμές για τις p μεταβλητές), τις συντεταγμένες του κέντρου βάρους των στοιχείων που αντικαθίστανται. Στη συνέχεια υπολογίζουμε τις αποστάσεις μεταξύ του στοιχείου αυτού και των υπολοίπων με μία από τις ακόλουθες τρεις μεθόδους: Min, Max, Moyenne.

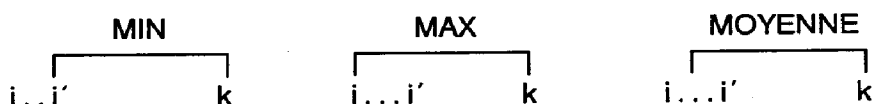
Αν i και i' είναι τα δύο συνενοούμενα στοιχεία σε μία κλάση, που συμβολίζεται iUi' , και k ένα άλλο στοιχείο, η απόσταση του k από την κλάση αυτή είναι:

$$\text{Min} \quad d_m(iUi', k) = \min \{d(i, k), d(i', k)\}$$

$$\text{Max} \quad d_M(iUi', k) = \max \{d(i, k), d(i', k)\}$$

$$\text{Moyenne} \quad d_\mu(iUi', k) = \frac{p(i) d(i, k) + p(i') d(i', k)}{p(i) + p(i')}$$

(όπου $p(x)$ είναι το πλήθος στοιχείων της κλάσης x)
εποπτικά οι τρεις μέθοδοι παριστάνονται στο σχήμα 1:



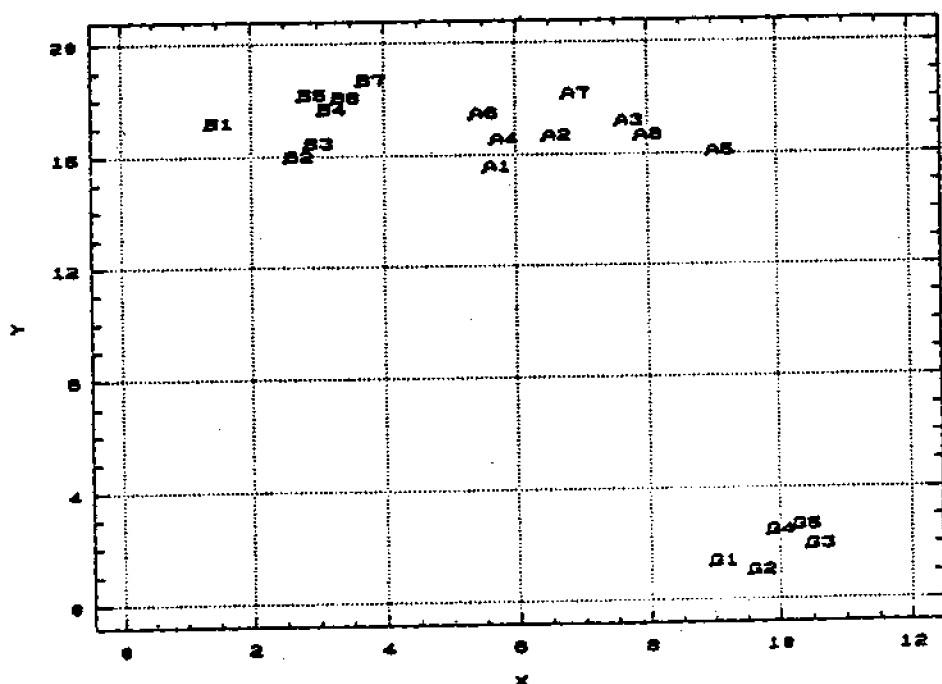
Σχήμα 1. Σχηματική παράσταση των τριών μεθόδων.

Συνεχίζουμε έτσι, ενώνοντας διαδοχικά τις δύο πλησιέστερες κλάσεις, μέχρι να καταλήξουμε σε δύο μόνο, τις οποίες ενώνουμε σε μία τελική κλάση, που περιέχει όλα τα δεδομένα.

Για την εποπτικότερη παρουσίαση των τριών μεθόδων, θεωρούμε ένα πίνακα δεδομένων 20 παρατηρήσεων και 2 ποσοτικών μεταβλητών (πίνακας 1).

ΠΙΝΑΚΑΣ 1
20 παρατηρήσεις ως προς 2 ποσοτικές μεταβλητές

Αντικείμενο	X	Y
A1	5.5	15.4
A2	6.4	16.5
A3	7.5	17.0
A4	5.6	16.4
A5	8.9	15.9
A6	5.3	17.3
A7	6.7	18.0
A8	7.8	16.5
B1	1.3	17.0
B2	2.5	15.8
B3	2.8	16.3
B4	3.0	17.5
B5	2.7	18.0
B6	3.2	17.9
B7	3.6	18.5
Γ1	8.9	1.2
Γ2	9.5	0.9
Γ3	10.4	1.8
Γ4	9.8	2.3
Γ5	10.2	2.5



Σχήμα 2. Τα σημεία των 20 παρατηρήσεων στους άξονες των 2 μεταβλητών.

Τα δεδομένα επιλέχτηκαν έτσι ώστε να σχηματίζουν τρεις ομάδες στο επίπεδο των δύο διαστάσεων (σχήμα 2). Σε κάθε βήμα, και για τις τρεις μεθόδους, με την προτεινόμενη διαδικασία, τυπώνεται ο πίνακας αποστάσεων, ώστε να φαίνεται ποια είναι τα "κοντινότερα" στοιχεία, τα οποία ενώνονται για να σχηματίσουν τον νέο κόμβο.

Αποτελέσματα του πρώτου παραδείγματος ταξινόμησης (απόσταση με την ευκλείδεια μετρική)

Ο πρώτος πίνακας αποστάσεων (πίνακας 2) είναι (ίδιος και για τις τρεις μεθόδους, και η ελάχιστη απόσταση είναι 0.44721 για τα στοιχεία B6 και B4. Στο επόμενο βήμα τα στοιχεία αυτά ενώνονται σε ένα που συμβολίζεται με τον κόμβο νούμερο 21.

Ο δεύτερος πίνακας αποστάσεων (πίνακας 3) διαφέρει στις τρεις μεθόδους ως προς τη γραμμή και τη στήλη που βρίσκεται ο κόμβος 21. Έτσι π.χ. η απόσταση μεταξύ 21 και B3, με τη μέθοδο Max είναι 1.64924 (δηλαδή η απόσταση B6-B3 του πίνακα 2), με τη μέθοδο Min είναι 1.21655 (δηλαδή η από-

σταση B4-B3), ενώ με τη Moyenne είναι 1.43290 (δηλαδή ο μέσος όρος των αποστάσεων B6-B3 και B4-B3).

Η ελάχιστη απόσταση στον δεύτερο πίνακα αποστάσεων (πίνακας 3) είναι και για τις τρεις μεθόδους 0.44721, μεταξύ των στοιχείων Γ4-Γ5. Στη συνέχεια ενώνονται τα στοιχεία αυτά και σχηματίζουν τον κόμβο 22. Ο επόμενος πίνακας αποστάσεων, ο τρίτος για κάθε μέθοδο, διαφέρει από τον αντίστοιχο δεύτερο στη γραμμή και στηλη του κόμβου 22.

Στον τρίτο πίνακα αποστάσεων διαπιστώνουμε ότι: Με τη μέθοδο Max η ελάχιστη απόσταση 0.58310 είναι των τριών ζευγών B5-21, B3-B2, και των A3-A8. Ενώνονται λοιπόν τα A3 και A8 σχηματίζοντας τον κόμβο 23, στο επόμενο βήμα τα B3-B2, σχηματίζοντας τον κόμβο 24 και στο μεθεπόμενο τα B5 και 21 που σχηματίζουν τον κόμβο 25.

Με τη μέθοδο Min η ελάχιστη απόσταση είναι 0.50990 μεταξύ B5 και 21, τα οποία ενώνονται στον κόμβο 23. Στο επόμενο βήμα ενώνονται και εδώ τα A3-A8 με απόσταση 0.58310 που σχηματίζουν τον κόμβο 24, ενώ στο μεθεπόμενο τα B3-B2 με την ίδια απόσταση σχηματίζουν τον κόμβο 25.

Με τη μέθοδο Moyenne η ελάχιστη απόσταση είναι 0.54650 για τα B5 και 21, τα οποία σχηματίζουν τον κόμβο 23. Στη συνέχεια ενώνονται τα A3-A8 με απόσταση 0.58310, δημιουργώντας τον κόμβο 24. Τέλος στο επόμενο βήμα ενώνονται τα B3-B2 με απόσταση πάλι 0.58310 και δίνουν τον κόμβο 25.

Στο 6ο βήμα, και με τις τρεις μεθόδους, ενώνονται τα Γ1-Γ2 με απόσταση 0.67082 και δίνουν τον κόμβο 26.

Στο 7ο βήμα, για τη μέθοδο Max, ενώνεται το Γ3 με τον κόμβο (Γ4-Γ5) 22, ($d(\Gamma3, \Gamma4)=0.78102$) και δίνουν τον κόμβο 27. Με τη μέθοδο Min, ενώνεται το B7 με τον κόμβο (B5-21) 23 ($d(B7, 21)=0.72111=d(B7, B6)$), ενώ στο 8ο βήμα ενώνεται το Γ3 με (Γ4-Γ5) τον 22 ($d(\Gamma3, \Gamma5)=0.72801$). Με τη μέθοδο Moyenne, το Γ3 με τον (Γ4-Γ5) 22 ενώνονται στο 7ο βήμα ($0.75452=\text{μέσος όρος των } d(\Gamma3, \Gamma4) \text{ και } d(\Gamma3, \Gamma5)$).

Στο 8ο βήμα, με τη μέθοδο Max και Moyenne, ενώνονται τα A2-A4 ($d(A2, A4)=0.80623$) δημιουργώντας τον κόμβο 28. Το ίδιο συμβαίνει με τη μέθοδο Min, αλλά στο βήμα 9.

Στο 9ο βήμα, με τη μέθοδο Max, συνδέεται ο κόμβος (B5-21) 25 με το B7 ($d(B7, 21)=1.16619=d(B7, B4)$), ενώ με τη μέθοδο Moyenne ο κόμβος (B5-21) 23 με το B7 (απόσταση=0.97229).

Στο 10ο βήμα, με τη μέθοδο Max, ενώνεται το A6 με τον κόμβο (A2-A4) 28 ($d(A6, A2)=1.36015$). Με τη μέθοδο Min ενώνεται το A6 με τον κόμβο (A2-A4) 29 ($d(A6, A4)=0.94868$). Τέλος, με τη μέθοδο Moyenne, ενώνεται το A6 με τον κόμβο (A2-A4) 28 (μέσος όρος (των $d(A6, A2)$, $d(A6, A4)=1.15442$)).

Στο 11ο βήμα, με τη μέθοδο Max, ενώνονται τα B1 και (B2-B3) 24 ($d(B1, B2)=1.69706$). Με τις άλλες δύο μεθόδους, στο 11ο βήμα, ενώνεται το A1 με τον κόμβο 30 (με αποστάσεις 1.00499, 1.44558 αντίστοιχα).

ΠΙΝΑΚΑΣ 2
Ο πρώτος πίνακας αποστάσεων μεταξύ των 20 στοιχείων

	A1	A2	A3	A4	A5	A6	A7	A8	B1	B2	B3	B4	B5	B6	B7	Γ_1	Γ_2	Γ_3	Γ_4
A2	1.42127																		
A3	2.56125	1.20830																	
A4	1.00499	0.80623	1.99249																
A5	3.43657	2.57099	1.78045	3.33766															
A6	1.91050	1.36015	2.22036	0.94868	3.86264														
A7	2.86356	1.52971	1.28062	1.94165	3.04138	1.56525													
A8	2.54951	1.40000	0.58310	2.20227	1.25300	2.62488	1.86011												
B1	4.49444	5.12445	6.20000	4.34166	7.67919	4.01123	5.49181	6.51920											
B2	3.02655	3.96232	5.14198	3.15753	6.40078	3.17648	4.74131	5.34603	1.69706										
B3	2.84605	3.60555	4.75184	2.80179	6.11310	2.69258	4.25441	5.00400	1.65529	0.58310									
B4	3.26497	3.54401	4.52769	2.82312	6.11310	2.30868	3.73363	4.90306	1.77200	1.77200	1.21655								
B5	3.82099	3.98249	4.90306	3.31210	6.54599	2.69258	4.00000	5.31601	1.72047	2.20907	1.70294	0.58310							
B6	3.39706	3.49285	4.33318	2.83019	6.04070	2.18403	3.50143	4.80833	2.10238	2.21359	1.64924	0.44727	0.50990						
B7	3.63593	3.44093	4.17852	2.90000	5.90339	2.08087	3.14006	4.65188	2.74591	2.91548	2.34094	1.16619	1.02956	0.72111					
Γ_1	14.60137	15.50290	15.86190	15.55410	14.70000	16.49758	16.94344	15.33949	17.53283	15.94114	16.28558	17.33494	17.90754	17.64596	18.08365				
Γ_2	15.04161	15.90503	16.22375	15.98312	15.01200	16.92926	17.32772	15.89235	18.06793	16.46238	16.79434	17.82723	18.40245	18.12981	18.56260	0.67082			
Γ_3	14.45579	15.23450	15.47417	15.36880	14.17956	16.31748	16.61716	14.92816	17.71581	16.07514	-16.37101	17.35655	17.93683	17.63661	18.03136	1.61555	1.27279		
Γ_4	13.78768	14.60137	14.87884	14.71224	13.62975	15.66046	16.00312	14.34015	16.98058	15.34731	15.65248	16.65173	17.23079	16.93871	17.34589	1.42127	1.43178	0.78102	
Γ_5	13.72953	14.50655	14.74924	14.64138	13.46291	15.59006	15.89025	14.20422	17.01352	15.36815	15.65866	16.63851	17.21918	16.91626	17.30780	1.83848	1.74642	0.72801	0.44721

ΠΙΝΑΚΑΣ 3
Πίνακες αποστάσεων με τις τρεις μεθόδους
Δεύτερος πίνακας αποστάσεων με τη μέθοδο Moyenne

A1	A2	A3	A4	A5	A6	A7	A8	B1	B2	B3	21	B5	F5	B7	F1	F2	F3
A2	1.42127																
A3	2.56125	1.20830															
A4	1.00499	0.80623	1.99249														
A5	3.43657	2.57099	1.78045	3.33766													
A6	1.91050	1.36015	2.22036	0.94868	3.86264												
A7	2.86356	1.52971	1.28062	1.94165	3.04138	1.55525											
A8	2.54951	1.40000	0.58310	2.20227	1.25300	2.62488	1.96011										
B1	4.49444	5.12445	6.20000	4.34166	7.67919	4.01123	5.49181	6.51920									
B2	3.02655	3.96232	5.14198	3.15753	6.40078	3.17648	4.74131	5.34603	1.69706								
B3	2.84605	3.60555	4.75184	2.80179	6.11310	2.69258	4.25441	5.00400	1.65529	0.58310							
21	3.33101	3.51843	4.46043	2.82556	6.07690	2.24636	3.61753	4.85569	1.93719	1.99280	1.43290						
B5	3.82099	3.99249	4.90306	3.31210	6.54599	2.69258	4.00000	5.31601	1.72047	2.20907	1.70294	0.54650					
F5	13.72953	14.50655	14.74924	14.64138	13.46291	15.59006	15.89025	14.20422	17.01352	15.36815	16.77739	17.21918					
B7	3.63593	3.44093	4.17852	2.90000	5.90339	2.08087	3.14006	4.65188	2.74591	2.91548	0.94365	1.02956					
F1	14.60137	15.50290	15.86190	15.55410	14.70000	16.49758	16.94344	15.33949	17.53283	15.94114	16.28558	17.90754	17.30780	18.09365			
F2	15.04161	15.90503	16.22375	15.98312	15.01200	16.92926	17.32772	15.69235	18.06793	16.79434	17.97852	18.40245	1.74642	18.56280	0.67082		
F3	14.45579	15.23450	15.47417	15.36880	14.17956	16.31748	16.61716	14.92816	17.71581	16.07514	17.49658	17.93683	0.72801	18.03136	1.61555	1.27279	
F4	13.78768	14.60137	14.87884	14.71224	13.62975	15.66046	16.00312	14.34015	16.98058	15.34731	16.79522	17.23079	0.44721	17.34589	1.42127	1.43178	0.78102

ΠΙΝΑΚΑΣ 3 (συνέχεια)
Πίνακες αποστάσεων με τις τρεις μεθόδους
Δεύτερος πίνακας αποστάσεων με τη μέθοδο Min

	A1	A2	A3	A4	A5	A6	A7	A8	B1	B2	B3	21	B5	T5	B7	T1	T2	T3
A2	1.42127																	
A3	2.56125	1.20830																
A4	1.00489	0.80623	1.98249															
A5	3.43857	2.57099	1.78045	3.33766														
A6	1.91050	1.36015	2.22036	0.94868	3.86264													
A7	2.86356	1.52971	1.28062	1.94165	3.04138	1.56525												
A8	2.54951	1.40000	0.58310	2.20227	1.25300	2.62488	1.86011											
B1	4.739444	5.12445	6.20000	4.34166	7.67919	4.01123	5.49181	6.51920										
B2	3.02655	3.96232	5.14198	3.15753	6.40078	3.17648	4.74131	5.34603	1.69706									
B3	2.94805	3.60555	4.75184	2.80179	6.11310	2.69258	4.25441	5.00400	1.65529	0.58310								
21	3.26497	3.49285	4.39318	2.82312	6.04070	2.18403	3.50143	4.80833	1.77200	1.77200	1.21655							
B5	3.82089	3.99249	4.90306	3.31210	6.54599	2.69258	4.00000	5.31601	1.72047	2.20907	1.70294	0.50990						
T5	13.72953	14.50655	14.74924	14.64138	13.46291	15.59006	15.89025	14.20422	17.01352	15.36815	15.65886	16.63851	17.21918					
B7	3.63593	3.44093	4.17852	2.90000	5.90339	2.08087	3.14006	4.65188	2.74591	2.91548	1.02956	0.72111	1.02956	17.30780				
T1	14.80137	15.50290	15.86190	15.55410	14.70000	16.49758	16.94344	15.33949	17.53283	15.94114	16.28558	17.33494	17.90754	1.83848	18.09365			
T2	15.04161	15.90503	16.22375	15.98312	15.01200	16.92926	17.32772	15.69235	18.06793	16.46238	-16.79434	17.82723	18.40245	1.74642	18.56260	0.67082		
T3	14.45579	15.23450	15.47417	15.36880	14.17956	16.31748	16.61716	14.92816	17.71581	16.07514	16.37101	17.35655	17.93683	0.72801	18.03136	1.61555	1.27279	
T4	13.78768	14.60137	14.87984	14.71224	13.62975	15.66046	16.00312	14.34015	16.98058	15.34731	15.65248	16.65173	17.23079	0.44727	17.34689	1.42127	1.43178	0.78102

ΠΙΝΑΚΑΣ 3 (συνέχεια)
Πίνακες αποστάσεων με τις τρεις μεθόδους
Δεύτερος πίνακας αποστάσεων με τη μέθοδο Max

	A1	A2	A3	A4	A5	A6	A7	A8	B1	B2	B3	21	B5	Γ5	B7	Γ1	Γ2	Γ3
A2	1.42127																	
A3	2.56125	1.20830																
A4	1.00499	0.80623	1.99249															
A5	3.43857	2.57089	1.78045	3.33766														
A6	1.91050	1.36015	2.22036	0.94868	3.86264													
A7	2.96356	1.52971	1.28062	1.94165	3.04138	1.56525												
A8	2.54951	1.40000	0.59310	2.20227	1.25300	2.62468	1.86011											
B1	4.49444	5.12445	6.20000	4.34166	7.67919	3.17648	5.49181	6.51920										
B2	3.02655	3.96232	5.14198	3.15753	6.40078	3.17648	4.74131	5.34603	1.69706									
B3	2.84605	3.60555	4.75184	2.80179	6.11310	2.69258	4.25441	5.00400	1.65529	0.58310								
21	3.39706	3.54401	4.52769	2.83019	6.11310	2.30868	3.73363	4.90306	2.10236	2.21359	1.64924							
B5	3.82099	3.99249	4.90306	3.31210	6.54599	2.69258	4.00000	5.31601	1.72047	2.20907	1.70294	0.58310						
Γ5	13.72963	14.50655	14.74824	14.64138	13.46291	15.59006	15.89025	14.20422	17.01352	15.36815	15.65686	16.91626	17.21918					
B7	3.63593	3.44093	4.17852	2.90000	5.90339	2.08087	3.14006	4.65188	2.74591	2.91548	2.34094	1.16619	1.02956	17.30780				
Γ1	14.60137	15.50290	15.86190	15.55410	14.70000	16.49758	16.94344	15.33949	17.53283	15.94114	16.28558	17.64596	17.90754	1.83848	18.09365			
Γ2	15.04161	15.90503	16.22375	15.98312	15.01200	16.92926	17.32772	15.69235	18.06793	16.46238	-16.79434	18.12981	18.40245	1.74642	18.56260	0.67082		
Γ3	14.45579	15.23450	15.47417	15.96880	14.17966	16.31748	16.61716	14.92816	17.71581	16.07514	16.37101	17.63661	17.93683	0.72801	18.03136	1.61555	1.27279	
Γ4	13.78768	14.60137	14.67884	14.71224	13.62375	15.66046	16.00312	14.34015	16.98058	15.34731	15.65248	16.93871	17.23079	0.44721	17.34589	1.42127	1.43178	0.78102

Στο 12ο βήμα, με τη μέθοδο Max, ενώνονται τα στοιχεία A5 και (A8-A3) 23 ($d(A5, A3) = 1.78045$). Με τη μέθοδο Min οι κόμβοι (A8-A3) 24 και (A1-30) 31 (απόσταση 1.20830), ενώ με τη μέθοδο Moyenne ενώνονται τα A5 και (A8-A3) 24 (απόσταση 1.51672).

Στο 13ο βήμα, με τη μέθοδο Max, ενώνονται οι κόμβοι (Γ1-Γ2) 26 και (Γ3-22) 27 ($d(\Gamma1, 22) = 1.83848$), με τη μέθοδο Min οι κόμβοι (B2-B3) 25 και (B7-23) 27 ($d(B3, 23) = 1.21655$), ενώ με τη μέθοδο Moyenne οι κόμβοι (Γ1-Γ2) 26 και (Γ3-22) 27 (απόσταση 1.55438).

Στο 14ο βήμα, με τη μέθοδο Max, το στοιχείο A1 ενώνεται με τον κόμβο (A6-28) 30 ($d(A1, A6) = 1.91050$). Με τη μέθοδο Min ενώνεται το A5 με τον κόμβο 32 (απόσταση 1.253). Τέλος, με τη μέθοδο Moyenne, ενώνεται το B1 με το (B3-B2) 25 (απόσταση 1.67618).

Στο 15ο βήμα, και με τη μέθοδο Max και με τη Moyenne, ενώνεται το A7 με τον κόμβο (30-A1) (αντίστοιχοι κόμβοι 34-31), αλλά με διαφορετική τιμή απόστασης. Για τη μέθοδο Max $d(A7, A1) = 2.86356$, ενώ για τη Moyenne (μέσος των $d(A7, A1)$ και $d(A7, 30)$) $= 1.97504$. Στο ίδιο βήμα με τη μέθοδο Min ενώνονται οι κόμβοι 26-28, ενώ με τη μέθοδο αυτή η ένωση του A7 με τον κόμβο (A5-32) 34 γίνεται στο επόμενο βήμα. Επίσης στο 16ο βήμα, με τη μέθοδο Max, ενώνονται οι κόμβοι 29 και 31, με τη Moyenne οι 29 και 34.

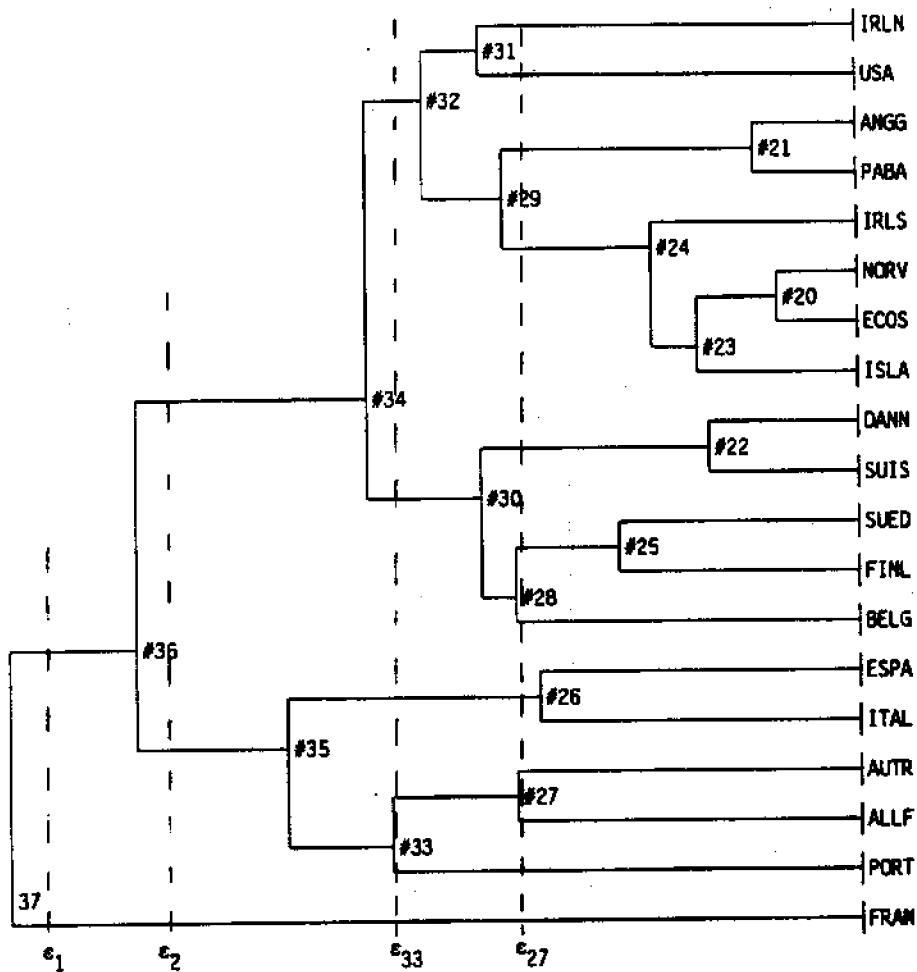
Στο βήμα 18, ενώ με τις άλλες μεθόδους ενώνονται ανώτεροι κόμβοι, με τη Min ενώνεται το αρχικό στοιχείο B1 με τον κόμβο 33.

Τέλος, στο 19ο βήμα και με τις τρεις μεθόδους ενώνονται οι κόμβοι 36 και 37 δίνοντας τον κόμβο 38, ο οποίος στο επόμενο βήμα ενώνεται με τον κόμβο 33, δίνοντας ένα μόνο κόμβο που περιέχει και τις 20 αρχικές παρατηρήσεις.

5. ΔΕΝΔΡΟΓΡΑΜΜΑ ΙΕΡΑΡΧΙΑΣ

Η ταξινόμηση κατά αύξουσα ιεραρχία (C.A.H.) σχηματικά παριστάνεται με ένα δενδρόγραμμα, όπου κάθε κόμβος αντιστοιχεί σε μία κλάση (ομάδα παρατηρήσεων) και στο επόμενο επίπεδο χωρίζεται σε δύο υποκλάσεις (υποομάδες) (στις κλάσεις που ενώθηκαν για να σχηματίσουν τον κόμβο). Η κάθε μία από τις υποκλάσεις χωρίζεται με τη σειρά της σε δύο υποκλάσεις, μέχρι να καταλήξουμε στους τερματικούς κόμβους, οι οποίοι αντιστοιχούν στις αρχικές παρατηρήσεις και εμφανίζονται μόνο μία φορά.

Ένα δενδρόγραμμα ιεραρχίας φαίνεται στο σχήμα 3 και παρουσιάζει την ιεραρχική ταξινόμηση κατ' αύξουσα ιεραρχία, με τη μέθοδο Moyenne για 19 χώρες.



Σχήμα 3. Δενδρόγραμμα ιεραρχίας με τη μέθοδο Mogenne για 19 χώρες.

Κάθε φορά που "κόβουμε" το δενδρόγραμμα με μία ευθεία γραμμή σε κάποιο επίπεδο, οι κόμβοι που απομένουν σχηματίζουν ένα διαμελισμό του συνόλου των αρχικών παρατηρήσεων.

"Κόβουμε" λοιπόν διαδοχικά το δενδρόγραμμα με την ευθεία ϵ_1 , σχηματίζοντας στην αρχή διαμελισμό σε δύο κλάσεις, κατόπιν με την ευθεία ϵ_2 σε τρεις κλάσεις κτλ., μέχρι να φθάσουμε στο διαμελισμό σε n κλάσεις, που είναι τα αρχικά μας δεδομένα.

Οι αποστάσεις των επιπέδων, που ορίζουν οι κόμβοι, είναι ανάλογες με τις πραγματικές αποστάσεις τους. Π.χ. η απόσταση μεταξύ των κόμβων 33 και 27 είναι ανάλογη με την απόσταση των επιπέδων ϵ_{33} και ϵ_{27} .

Οι σχηματιζόμενοι διαμελισμοί

Οι διαδοχικοί διαμελισμοί που σχηματίζονται και με τις τρεις μεθόδους, για το προηγούμενο παράδειγμα, φαίνονται στα αντίστοιχα δένδρογράμματα (σχήματα 4, 5, 6). Όσο προχωράμε από τις αρχικές παρατηρήσεις προς την κορυφή, παρατηρούμε ότι η απόσταση των στοιχείων που ενώνονται αυξάνει και επομένως μειώνεται η ομοιογένεια των ομάδων (πίνακες 4, 5, 6).

Εξετάζουμε τώρα τα τρία δένδρογράμματα, αρχίζοντας την ανάλυσή τους από την κορυφή προς τα κάτω.

Για τον διαμελισμό σε τρεις κλάσεις, και με τις τρεις μεθόδους, οδηγούμαστε σε τρεις ομάδες. Η πρώτη περιέχει τα στοιχεία A1, A2, A3, A4, A5, A6, A7, η δεύτερη τα B1, B2, B3, B4, B5, B6 και η τρίτη τα Γ1, Γ2, Γ3, Γ4, Γ5.

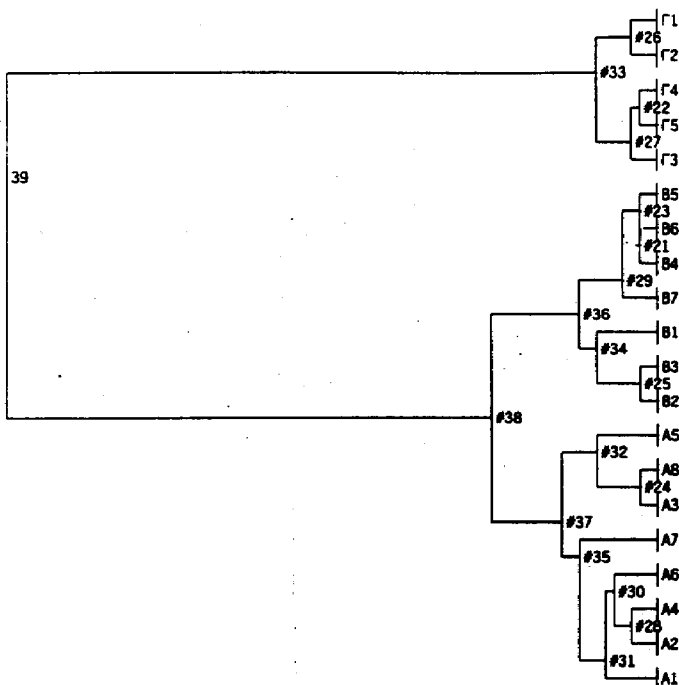
Ο χωρισμός σε τέσσερις κλάσεις είναι διαφορετικός για κάθε μέθοδο. Ενώ με τη Moyenne και τη Max η ομάδα A χωρίζεται σε δύο υποομάδες (A1-A3-A8) και (A2-A4-A5-A6-A7), με τη μέθοδο Min αποχωρίζεται το B1 από την υπόλοιπη ομάδα B.

Στις πέντε κλάσεις, με τις μεθόδους Max και Moyenne διασπάται η ομάδα B σε δύο υποομάδες (B1-B2-B3) και (B4-B5-B6-B7), ενώ με τη μέθοδο Min αποχωρίζεται το A7 από την ομάδα A.

Στις έξι κλάσεις, με τις μεθόδους Max και Moyenne αποχωρίζεται το A7 από την υποομάδα της A, ενώ με τη μέθοδο Min διασπάται η ομάδα Γ σε δύο υποομάδες: τη (Γ1-Γ2) και (Γ3-Γ4-Γ5).

Σε επτά κλάσεις η εικόνα είναι διαφορετική και με τις τρεις μεθόδους. Με τη μέθοδο Moyenne χωρίζεται το B1 από τα B2-B3, με τη μέθοδο Max το A1 από τα A2-A4-A6, ενώ με τη μέθοδο Min το A5 από την υπόλοιπη ομάδα A.

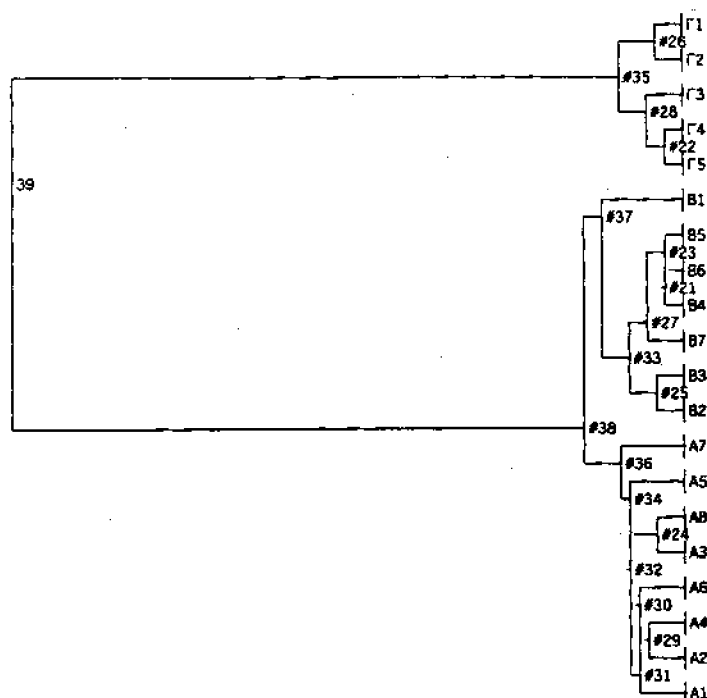
Στις οκτώ κλάσεις με τις μεθόδους Max και Moyenne διασπάται η ομάδα Γ σε δύο υποομάδες, τη (Γ1-Γ2) και (Γ3-Γ4-Γ5), ενώ με τη μέθοδο Min αποχωρίζεται το A5 από την υπόλοιπη ομάδα A.



Σχήμα 4. Δενδρόγραμμα ιεραρχίας με τη μέθοδο Μογεννη.

ΠΙΝΑΚΑΣ 4
Κόμβοι της ταξινόμησης με τη μέθοδο Μογεννη

Κόμβος	Αριστερό	Δεξιό	Βάρος	Απόσταση
21	B6	B4	2	0.447214
22	Γ4	Γ5	2	0.447214
23	B5	21	3	0.546499
24	A8	A3	2	0.583095
25	B3	B2	2	0.583095
26	Γ1	Γ2	2	0.670820
27	22	Γ3	3	0.754518
28	A4	A2	2	0.806226
29	23	B7	4	0.972288
30	A6	28	3	1.154415
31	30	A1	4	1.445584
32	A5	24	3	1.516723
33	26	27	5	1.554382
34	B1	25	3	1.676175
35	A7	31	5	1.975042
36	29	34	7	2.030048
37	32	35	8	2.409936
38	36	37	15	4.236877
39	33	38	20	16.112806

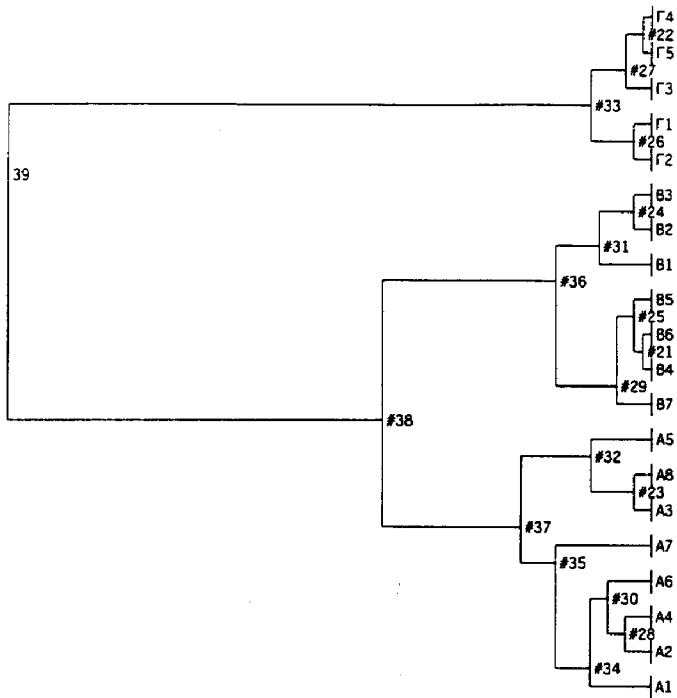


Σχήμα 5. Δενδρογράμμο ιεραρχίας με τη μέθοδο Min.

ΠΙΝΑΚΑΣ 5

Κόμβοι της ταξινόμησης με τη μέθοδο Min

Κόμβος	Αριστερό	Δεξιό	Βάρος	Απόσταση
21	B6	B4	2	0.447214
22	Γ4	Γ5	2	0.447214
23	B5	21	3	0.509902
24	A8	A3	2	0.583095
25	B3	B2	2	0.583095
26	Γ1	Γ2	2	0.670820
27	23	B7	4	0.721110
28	Γ3	22	3	0.728011
29	A4	A2	2	0.806226
30	A6	29	3	0.948683
31	30	A1	4	1.004988
32	24	31	6	1.208305
33	27	25	6	1.216553
34	A5	32	7	1.252996
35	26	28	5	1.272792
36	A7	34	8	1.280625
37	B1	33	7	1.655295
38	37	36	15	2.080865
39	35	38	20	13.462912



Σχήμα 6. Δενδρογράμμο ιεραρχίας με τη μέθοδο Max.

ΠΙΝΑΚΑΣ 6
Κόμβοι της ταξινόμησης με τη μέθοδο Max

Κόμβος	Αριστερό	Δεξιό	Βάρος	Απόσταση
21	B6	B4	2	0.447214
22	Γ4	Γ5	2	0.447214
23	A8	A3	2	0.583095
24	B3	B2	2	0.583095
25	B5	21	3	0.583095
26	Γ1	Γ2	2	0.670820
27	22	Γ3	3	0.781025
28	A4	A2	2	0.806226
29	25	B7	4	1.166190
30	A6	28	3	1.360147
31	24	B1	3	1.697056
32	A5	23	3	1.780449
33	27	26	5	1.838478
34	30	A1	4	1.910497
35	A7	34	5	2.863564
36	31	29	7	2.915476
37	32	35	8	3.862642
38	36	37	15	7.679193
39	33	38	20	18.562597

Ο χωρισμός σε εννέα κλάσεις, με τις μεθόδους Moyenne και Max, γίνεται διαχωρίζοντας το A5 από την υποομάδα των A3-A8, ενώ με τη μέθοδο Min, διασπώντας την υποομάδα της A, σε δύο νέες υποομάδες, τις A3-A8 και A1-A2-A4-A6.

Στις δέκα κλάσεις με τη μέθοδο Moyenne και Min διαχωρίζεται το A1, ενώ με τη μέθοδο Max το B1. Έτσι προκύπτει ίδια ομαδοποίηση σε 10 ομάδες, και με τις τρεις μεθόδους.

Στις έντεκα κλάσεις και με τις τρεις μεθόδους διαχωρίζεται το A6 από την ομάδα που ανήκε.

Στις δώδεκα κλάσεις με τις μεθόδους Max και Moyenne διαχωρίζεται το B7, ενώ με τη μέθοδο Min διαχωρίζονται τα A2 και A4.

Ο διαχωρισμός των A2 και A4 γίνεται με τις μεθόδους Max και Moyenne στις δεκατρείς κλάσεις, ενώ με τη μέθοδο Min αποχωρίζεται το Γ3 από την υποομάδα Γ4-Γ5.

Ο χωρισμός του Γ3 γίνεται με τις μεθόδους Max και Moyenne στις δεκατέσσερις κλάσεις, ενώ με τη μέθοδο Min αποχωρίζεται το B7.

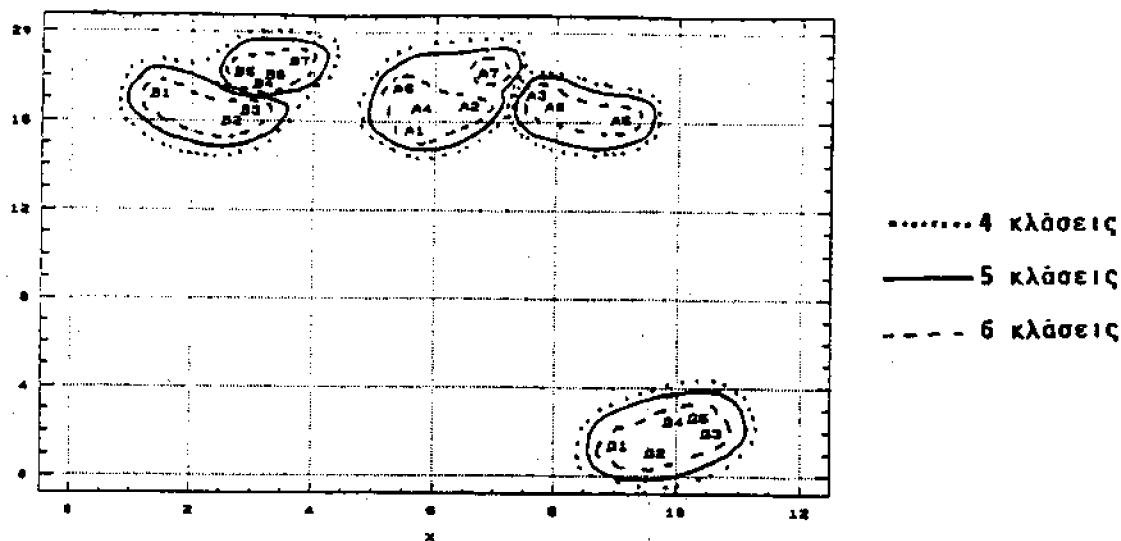
Ο διαμελισμός σε κλάσεις είναι ίδιος τώρα και στις τρεις μεθόδους και ο χωρισμός σε δεκαπέντε κλάσεις γίνεται σε όλες, διαχωρίζοντας τα Γ1 και Γ2.

Αν θελήσουμε να παρατηρήσουμε τα αποτελέσματα της ιεράρχησης και με τις τρεις μεθόδους, στο επίπεδο με άξονες τις δύο μεταβλητές (σχήματα 7, 8, 9), διαπιστώνουμε ότι για το συγκεκριμένο παράδειγμα οι μέθοδοι Max και Moyenne δίνουν περίπου τις ίδιες ομάδες, με κάποιες μικρές διαφοροποιήσεις, ενώ η μέθοδος Min διαχωρίζει από την αρχή κάποιο στοιχείο, που είναι απομακρυσμένο από τα υπόλοιπα (π.χ. το B1 από την υπόλοιπη ομάδα B).

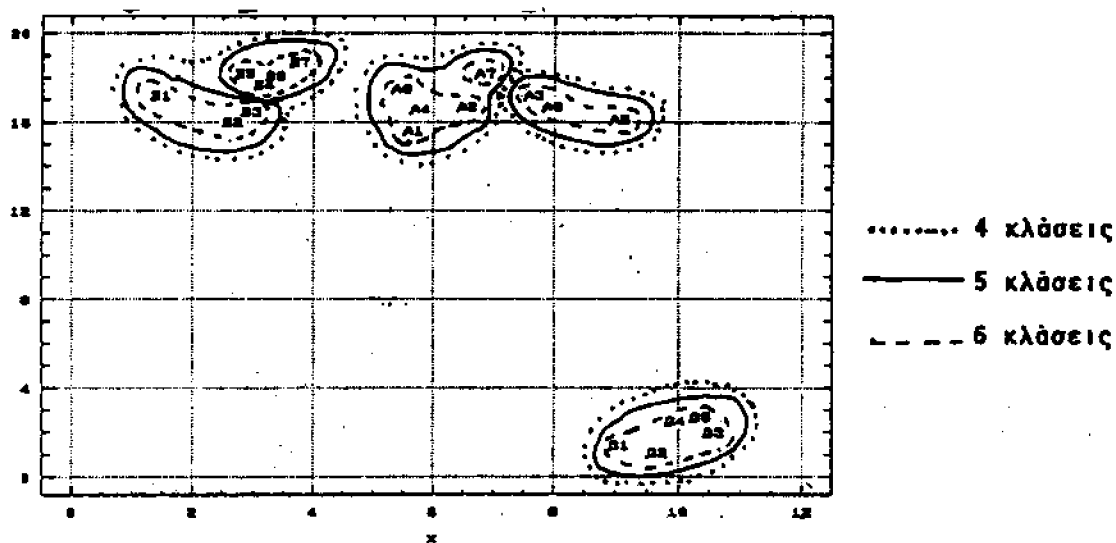
Πάντως ο διαμελισμός σε τρεις κλάσεις είναι ίδιος και στις τρεις μεθόδους και συμπίπτει με τις τρεις ομάδες των αρχικών δεδομένων (σχήμα 2).

Οι μέθοδοι Max και Moyenne δίνουν τα ίδια αποτελέσματα για 4, 5 ή 6 κλάσεις και σαφώς καλύτερα από τη μέθοδο Min (σχήμα 8). Καλύτερος διαμελισμός σε 7 κλάσεις είναι αυτός που παίρνουμε με τη μέθοδο Moyenne, ακολουθεί της μεθόδου Max με παρόμοια αποτελέσματα και τέλος ο διαμελισμός της μεθόδου Min (σχήμα 4).

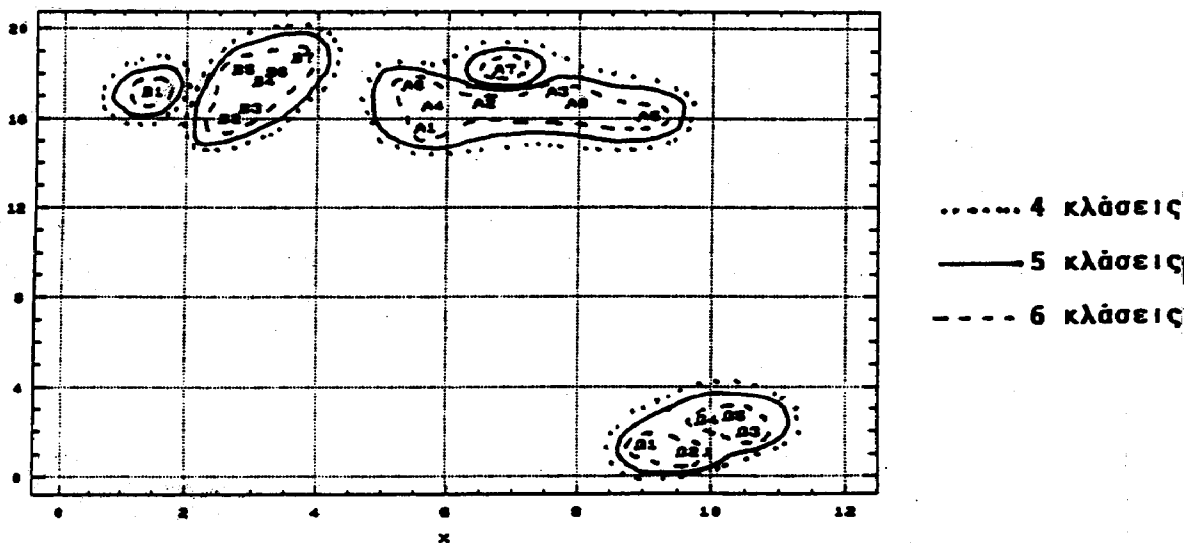
Διαμελισμοί με τη Μογερνε



Διαμελισμοί με τη Max



Διαμελισμοί με τη Min



Σχήματα 7,8,9. Η παρουσίαση των ομάδων στο επίπεδο για τις μεθόδους *Moyenne, Max, Min*.

Ένας διαμελισμός θεωρείται τόσο καλύτερος, όσο η αδράνεια του πλησιάζει στην αρχική ολική αδράνεια, όπως εξηγείται στην επόμενη παράγραφο.

6. ΚΡΙΤΗΡΙΟ ΓΙΑ ΚΑΘΕ ΔΙΑΜΕΛΙΣΜΟ ΣΕ r ΚΛΑΣΕΙΣ

Για το σύνολο των n παρατηρήσεων, στο χώρο των p διαστάσεων, υπολογίζουμε την αδράνεια ως προς το κέντρο βάρους τους g , με έναν από τους παρακάτω τύπους, ανάλογα αν τα δεδομένα είναι ποσοτικά ή ποιοτικά. Ευκλείδεια μετρική:

$$E_r = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p [A(i,j) - g(j)]^2$$

Μετρική του χ^2 :

$$E_r = \sum_{i=1}^n \sum_{j=1}^p \frac{(A(i,j) - A(i,.) \cdot A(.,j))}{A(i,.) \cdot A(.,j)}^2$$

όπου $A(i, \cdot)$ και $A(\cdot, j)$ παριστάνουν τα αθροίσματα της i γραμμής και j στήλης αντίστοιχα.

Κάθε διαμελισμός ενός συνόλου παρατηρήσεων δίνει διαφορετικό πλήθος κλάσεων, και τα στοιχεία των κλάσεων διαφέρουν από μέθοδο σε μέθοδο. Προσπαθούμε, λοιπόν, να βρούμε το κριτήριο εκείνο, που θα μας βοηθήσει να ελέγξουμε αν ένας διαμελισμός ομαδοποιεί ικανοποιητικά τα αρχικά δεδομένα και αν είναι καλύτερος ή όχι από κάποιον άλλον.

Για κάθε διαμελισμό, που χωρίζει τα δεδομένα σε r πλήθος κλάσεις, υπολογίζουμε την αδράνεια (διασπορά) των r κλάσεων (E^g ή E^x), ως προς το κέντρο βάρους των αρχικών παρατηρήσεων (Ρουχ, 1985), με τις παρακάτω σχέσεις όταν χρησιμοποιούμε την ευκλείδεια μετρική:

$$E_r^g = \frac{1}{n} \sum_{k=1}^r p(k) \sum_{j=1}^p [g_k(j) - g(j)]^2$$

ή

$$E_r^g = \frac{1}{n} \left[\sum_{k=1}^r \sum_{j=1}^p p(k) g_k^2(j) - n \sum_{j=1}^p g^2(j) \right]$$

και όταν χρησιμοποιούμε τη μετρική του X^2 τη σχέση:

$$E_r^x = \frac{1}{n} \sum_{k=1}^r p(k) \sum_{j=1}^p \frac{g_k(j) - \frac{A(\cdot, j)}{n} \sum_{k=1}^r p(k)}{\frac{A(\cdot, j)}{n} \sum_{k=1}^r p(k)}$$

όπου g_k είναι το διάνυσμα του κέντρου βάρους της k κλάσης, $p(k)$ το πλήθος στοιχείων της και g το διάνυσμα του κέντρου βάρους όλων των δεδομένων.

Θεωρούμε ότι κάθε κλάση αποτελείται από ένα μόνο σημείο (κέντρο βάρους) και το σημείο αυτό σταθμίζεται με έναν αριθμό $p(k)$, που αντιστοιχεί στο πλήθος των στοιχείων (βάρους) της κλάσης (θεωρούμε, δηλαδή, ότι όλα τα στοιχεία της κλάσης βρίσκονται στο κέντρο βάρους της). Την αδράνεια, που προαναφέραμε, τη συγκρίνουμε με την αρχική αδράνεια των παρατηρήσεων (πριν γίνει οποιαδήποτε ομαδοποίηση). Όσο περισσότερο πλησιάζουν οι δύο αυτές τιμές (ο λόγος των αδρανειών πλησιάζει στο 1), τόσο περισσότερο η δημιουργούμενη ομαδοποίηση προσαρμόζεται καλύτερα στα αρχικά δεδομένα.

7. ΚΡΙΤΗΡΙΟ ΣΥΝΟΛΙΚΗΣ ΙΕΡΑΡΧΙΚΗΣ ΤΑΞΙΝΟΜΗΣΗΣ

Εφαρμόζοντας τη μέθοδο C.A.H. σ' ένα πίνακα δεδομένων, προκύπτει η ολική ιεραρχική ταξινόμηση των παρατηρήσεων, που παρουσιάζεται εποπτικά με το δένδρογραμμα της ιεράρχησης, του οποίου η αναλυτική ερμηνεία παρέχεται από το πρόγραμμα VACOR (Benzecri, 1985). Για κάθε διαμελισμό, με αριθμό κλάσεων r , εφαρμόζουμε το κριτήριο της παραγράφου 6, ξεκινώντας από τις δύο τελικές κλάσεις και φθάνοντας στις n αρχικές κλάσεις.

Στη συνέχεια, για κάθε διαμελισμό, υπολογίζουμε τον λόγο

$$\lambda_r = \frac{\text{αδράνεια διαμελισμού}}{\text{αρχική αδράνεια}} \quad (\text{το } r \text{ από } 2 \text{ μέχρι } n)$$

για την ευκλείδεια μετρική:

$$\lambda_r = \frac{\frac{1}{n} \sum_{k=1}^r p(k) \sum_{j=1}^p [g_k(j) - g(j)]^2}{\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p [A(i,j) - g(j)]^2}$$

και για τη μετρική του χ^2 :

$$\lambda_r = \frac{\frac{1}{n} \sum_{k=1}^r p(k) \sum_{j=1}^p \frac{g_k(j) - \frac{A(.,j)}{n} \sum_{j=1}^p \frac{g_k(j)}{p(k)}}{\frac{A(.,j)}{n} \sum_{j=1}^p g_k(j)}}{\sum_{i=1}^n \sum_{j=1}^p \frac{(A(i,j) - A(i,.) \frac{A(.,j)}{n})^2}{A(i,.) A(.,j)}}$$

όπου, g_k είναι το διάνυσμα του κέντρου βάρους της k κλάσης με πλήθος στοιχείων $p(k)$ και g είναι το διάνυσμα του κέντρου βάρους όλων των δεδομένων. Ο λόγος αυτός εκφράζει το ποσοστό της ολικής αδράνειας που ερμηνεύεται από τον συγκεκριμένο διαμελισμό (Jambu κ.ά., 1980).

Με τις τιμές αυτού του λόγου σχηματίζουμε ένα διάγραμμα, όπου εποπτικά παρουσιάζεται η μεταβολή του λόγου αδράνειας του διαμελισμού ως προς την αρχική ολική αδράνεια, σε σχέση με το πλήθος κλάσεων κάθε διαμελισμού. Παρατηρώντας το διάγραμμα που εκφράζει το ποσοστό της ερμηνευόμενης αδράνειας, ως προς το πλήθος κλάσεων του διαμελισμού, από δεξιά

προς τα αριστερά, δηλαδή από τον διαμελισμό σε n κλάσεις προς τη μία κλάση (που περιλαμβάνει όλες τις παρατηρήσεις), διαπιστώνεται ότι ο λόγος λ_r μειώνεται όσο ελαττώνεται ο αριθμός των κλάσεων κάθε διαμελισμού.

Η μεταβολή του λόγου αυτού μεταξύ δύο διαδοχικών διαμελισμών προκύπτει από τον παρακάτω τύπο:

d_r = λόγος διαμελ. $r+1$ κλάσεις – λόγος διαμελ. r κλάσεων

Ειδικά για την ευκλείδεια μετρική είναι

$$d_r = \frac{1}{\sum_{i=1}^n \sum_{j=1}^p [A(i,j) - g(j)]^2} \left\{ \left[\frac{1}{n} \left[\sum_{k=1}^{r+1} \sum_{j=1}^p p(k) g_k^2(j) - n \sum_{j=1}^p g^2(j) \right] \right] - \right. \\ \left. - \left[\frac{1}{n} \left[\sum_{k=1}^r \sum_{j=1}^p p'(k) g_k'^2(j) - n \sum_{j=1}^p g^2(j) \right] \right] \right\}$$

Επειδή οι ομάδες του διαμελισμού σε r κλάσεις είναι ίδιες μ' αυτές σε $r+1$ κλάσεις, εκτός από μία ομάδα (έστω τη x), η οποία διασπάται σε δύο υποομάδες, x' και x'' , στο διαμελισμό σε $r+1$ κλάσεις, ο τύπος της μεταβολής του λόγου d_r λαμβάνει τη μορφή:

$$d_r = \frac{1}{\sum_{i=1}^n \sum_{j=1}^p [A(i,j) - g(j)]^2} \left[p(x') \sum_{j=1}^p g_{x'}^2(j) + p(x'') \sum_{j=1}^p g_{x''}^2(j) - p'(x) \sum_{j=1}^p g^2(j) \right]$$

και επειδή $p'(x) = p(x') + p(x'')$, ο τύπος λαμβάνει την τελική μορφή:

$$d_r = \frac{1}{\sum_{i=1}^n \sum_{j=1}^p [A(i,j) - g(j)]^2} \left\{ p(x') \left[\sum_{j=1}^p g_{x'}^2(j) - \sum_{j=1}^p g^2(j) \right] + p(x'') \left[\sum_{j=1}^p g_{x''}^2(j) - \sum_{j=1}^p g^2(j) \right] \right\}$$

Παρατηρούμε, λοιπόν, ότι η μεταβολή του ερμηνευόμενου ποσοστού αδράνειας μεταξύ δύο διαδοχικών διαμελισμών εξαρτάται μόνο από τη διασπώμενη κλάση x καθώς και από τις δύο νέες υποκλάσεις στις οποίες διασπάται.

Όταν, μεταβαίνοντας από το διαμελισμό σε $r+1$ κλάσεις στο διαμελισμό σε r κλάσεις, ο λόγος λ_r μεταβάλλεται απότομα, σημαίνει ότι επιλέγοντας το διαμελισμό με μία κλάση λιγότερη, χάνουμε σημαντικό ποσοστό πληροφορί-

ας. Στο σημείο αυτό, δηλαδή στις $r+1$ κλάσεις, έχουμε, χωρίς καμία a priori υπόθεση, τον πιο ικανοποιητικό διαμελισμό.

Επειδή δεν είναι εμφανής ο εντοπισμός των απότομων μεταβολών του ποσοστού της ερμηνευόμενης αδράνειας, από τη γραφική παράσταση του λόγου λ , σε σχέση με τον αριθμό κλάσεων, προτείνουμε τη μελέτη της γραφικής παράστασης των διαφορών d , μεταξύ των ποσοστών ερμηνευόμενης αδράνειας όλων των διαδοχικών διαμελισμών, όπου εμφανίζονται πολύ πιο έντονα οι διακυμάνσεις του.

Από το παιας μεθόδου, μεταξύ των Min, Max, Moyenne, η μεταβολή του λόγου λ , είναι πιο ομαλή, συμπεραίνουμε ποια απ' αυτές δίνει την καλύτερα προσαρμοσμένη στα δεδομένα συνολική ιεράρχηση. Καλύτερη είναι η μέθοδος εκείνη στην οποία ο λόγος έχει τιμή κοντά στο 1 για μεγαλύτερο αριθμό διαμελισμών (από δεξιά προς τα αριστερά στο διάγραμμα), και οι μεταβολές του λόγου από διαμελισμό σε διαμελισμό είναι μικρές.

*Παράδειγμα: Ταξινόμηση 19 χωρών ως προς τις αιτίες θανάτου
(ευκλείδεια μετρική)*

Εφαρμόσαμε τις τρεις μεθόδους, σε ένα πίνακα διαστάσεων 19×6 (πίνακας 10), χρησιμοποιώντας την ευκλείδεια μετρική. Κάθε γραμμή αντιστοιχεί σε μία χώρα, και κάθε στήλη σε μία αιτία θανάτου. Στη διασταύρωση της γραμμής i και της στήλης j αντιστοιχεί το πλήθος ατόμων της χώρας i , που πέθαναν λόγω της j αιτίας (βιβλίο Todd, *Le fou et le proletaire*, 1979, annexe 2, σελ. 283) (βιβλίο Roux, 1985).

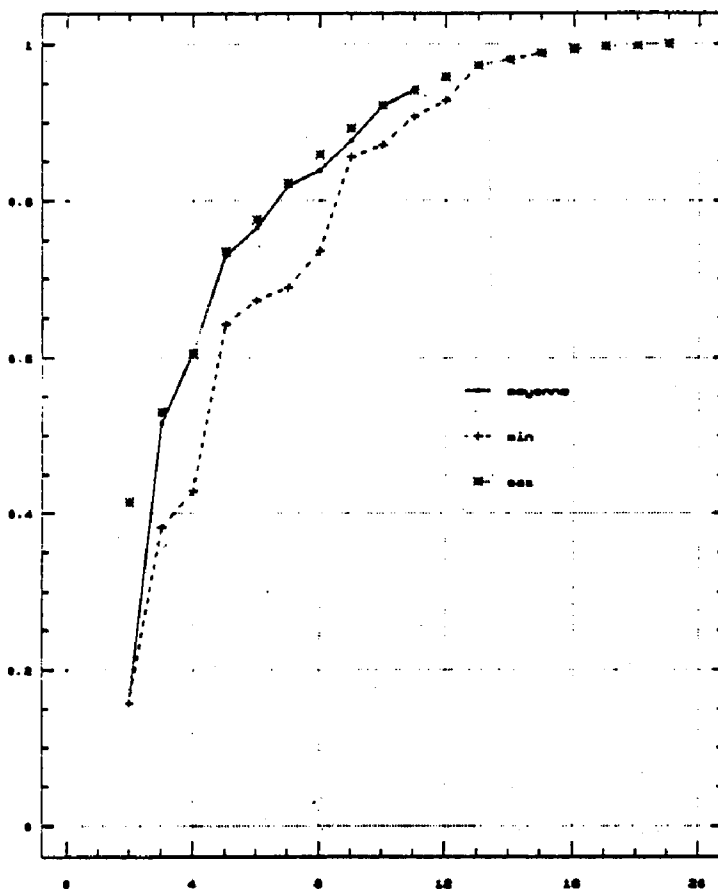
ΠΙΝΑΚΑΣ 10
Οι 6 μεταβλητές για τις 19 χώρες

AUTRICHE	241	16	330	43	363	325
FRANCE	156	9	225	10	535	328
PORTUGAL	85	19	349	7	281	345
ALLFEDER	210	12	230	21	298	269
BELGIQUE	156	10	260	13	367	144
FINLANDE	251	26	180	29	387	55
SUEDE	194	11	151	13	384	122
SUISSE	225	9	195	26	276	128
ITALIE	54	11	219	19	224	319
IRLNORD	40	136	215	18	320	43
DANNEMAR	241	6	168	11	230	107
ISLANDE	101	5	179	23	380	9
ECOSSE	82	15	155	18	342	59
ESPAGNE	40	4	136	17	237	225
NORVEGE	104	6	138	22	346	41
IRLSUD	38	7	182	32	314	37
PABAS	89	7	169	10	218	47
ANGGALLE	79	10	130	14	203	36
USA	121	102	220	26	273	158

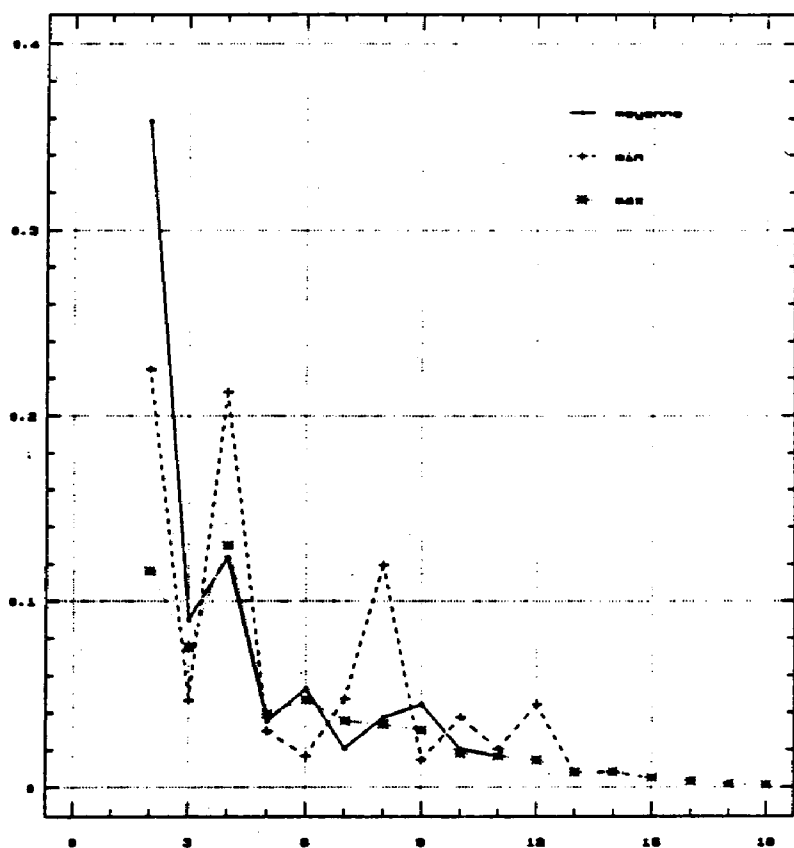
Το δένδρο ιεράρχησης (Moyenne) φαίνεται στο σχήμα 3.

Στο διάγραμμα 1 παριστάνεται το ερμηνευόμενο ποσοστό αδράνειας, σε σχέση με το πλήθος κλάσεων, κάθε διαμελισμού και για τις τρεις μεθόδους. Παρατηρούμε ότι η μέθοδος Max παρουσιάζει μεγαλύτερες τιμές του ερμηνευόμενου ποσοστού αδράνειας, για κάθε πλήθος κλάσεων. Συμπεραίνουμε, λοιπόν, ότι για τα συγκεκριμένα δεδομένα συνολικά καλύτερη θεωρείται η ιεράρχηση με τη μέθοδο Max.

Στη συνέχεια, στο διάγραμμα 2 παριστάνονται γραφικά οι μεταβολές (διαφορές) του ποσοστού ερμηνευόμενης αδράνειας μεταξύ δύο διαδοχικών διαμελισμών, ως προς το πλήθος κλάσεών τους.



Διάγραμμα 1. Γραφική παράσταση του λόγου λ ως προς το πλήθος κλάσεων διαμελισμού για τις χώρες.



Διάγραμμα 2. Γραφική παράσταση των διαφορών του λόγου λ ως προς το πλήθος κλάσεων διαμελισμού για τις χώρες.

Στο πρώτο διάγραμμα (διάγραμμα 1) οι μεταβολές των ποσοστών είναι πολύ μικρές, σχεδόν ανεπαίσθητες, ενώ στο δεύτερο (διάγραμμα 2) παρατηρούμε εύκολα ότι:

Με τη μέθοδο Moyenne, και για διαμελισμό σε πέντε ή περισσότερες κλάσεις, η καμπύλη των διαφορών είναι αρκετά ομαλή και οι μεταβολές παρουσιάζουν ελάχιστες διαφοροποιήσεις. Όσο όμως ελαττώνεται το πλήθος των κλάσεων, οι διαφοροποιήσεις γίνονται πιο σημαντικές. Αυτό δείχνει πιο φανερά ότι, χρησιμοποιώντας τη μέθοδο Moyenne, ο ιδανικότερος διαμελισμός των παρατηρήσεων είναι σε πέντε κλάσεις, όπου ερμηνεύεται ποσοστό 72.92% (πίνακας 7) της αρχικής ολικής αδράνειας των παρατηρήσεων.

Με τη μέθοδο Max η καμπύλη των διαφορών είναι πιο ομαλή και οι διαφοροποιήσεις είναι μικρότερες. Ο ιδανικότερος διαμελισμός, ως προς αυτή τη μέθοδο, είναι σε πέντε κλάσεις, όπου ερμηνεύεται ποσοστό 73.55% (πίνακας 8) της αρχικής ολικής αδράνειας των παρατηρήσεων.

Τέλος, για τη μέθοδο Min, παρατηρώντας το διάγραμμα 2, βλέπουμε τη μη ομαλότητα της καμπύλης. Αυτό κάνει ακόμη πιο έντονη τη μη καταλληλότητα της μεθόδου για τα συγκεκριμένα δεδομένα.

Διαμελισμοί των 19 χωρών με τις τρεις μεθόδους

ΠΙΝΑΚΑΣ 7 Διαμελισμοί με τη Moyenne			ΠΙΝΑΚΑΣ 8 Διαμελισμοί με τη Max			ΠΙΝΑΚΑΣ 9 Διαμελισμοί με τη Min		
Κλάσεις	Αδράνεια διαμελισμού	Λόγος ολ. αδρ.	Κλάσεις	Αδράνεια διαμελισμού	Λόγος ολ. αδρ.	Κλάσεις	Αδράνεια διαμελισμού	Λόγος ολ. αδρ.
2	4590.9207	0.1569	2	12115.9651	0.4140	2	4590.9207	0.1569
3	15077.3545	0.5152	3	15513.3993	0.5301	3	11173.9845	0.3818
4	17719.7808	0.6055	4	17719.7808	0.6055	4	12542.0108	0.4286
5	21339.3542	0.7292	5	21522.8251	0.7355	5	18769.0966	0.6414
6	22388.6700	0.7650	6	22683.6277	0.7751	6	19656.8827	0.6717
7	23935.8652	0.8179	7	24063.4503	0.8223	7	20152.1985	0.6886
8	24547.4441	0.8388	8	25112.7661	0.8581	8	21540.7400	0.7361
9	25649.1985	0.8765	9	26107.0161	0.8921	9	25042.7599	0.8557
10	26953.8783	0.9210	10	27008.0450	0.9229	10	25467.5757	0.8703
11	27547.2555	0.9413	11	27547.2555	0.9413	11	26569.3301	0.9079
12	28042.5713	0.9582	12	28042.5713	0.9582	12	27162.7073	0.9282
13	28467.3871	0.9728	13	28467.3871	0.9728	13	28467.3871	0.9728
14	28706.0450	0.9809	14	28697.2073	0.9806	14	28706.0450	0.9809
15	28935.8652	0.9888	15	28935.8652	0.9888	15	28935.8652	0.9888
16	29081.0055	0.9937	16	29081.0055	0.9937	16	29081.0055	0.9937
17	29180.3740	0.9971	17	29180.3740	0.9971	17	29180.3740	0.9971
18	29232.7950	0.9989	18	29232.7950	0.9989	18	29232.7950	0.9989
19	29264.6371	1.0000	19	29264.6371	1.0000	19	29264.6371	1.0000

Παράδειγμα: Ταξινόμηση των 52 νομών της Ελλάδας ως προς την απασχόληση (χ^2 μετρική)

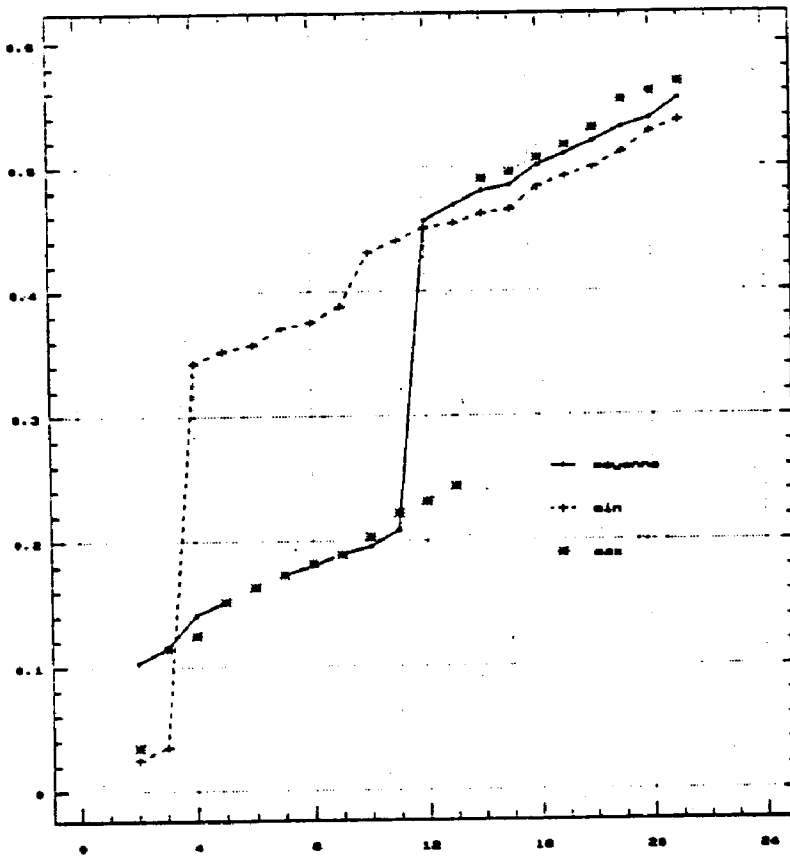
Εφαρμόζουμε πάλι τις τρεις μεθόδους σ' ένα πίνακα δεδομένων διαστάσεων 52×6 (πίνακας 11), χρησιμοποιώντας τώρα τη μετρική χ^2 , αφού τα δεδομένα που έχουμε είναι συχνότητες. Κάθε γραμμή αντιστοιχεί σε ένα νομό της Ελλάδας και κάθε στήλη στα είδη απασχόλησης (εργαζόμενοι, άνεργοι, εργοδότες, υπάλληλοι κτλ.). Στη διασταύρωση της γραμμής i και στήλης j βρίσκεται το πλήθος ατόμων του i νομού για τη j μεταβλητή.

ΠΙΝΑΚΑΣ 11

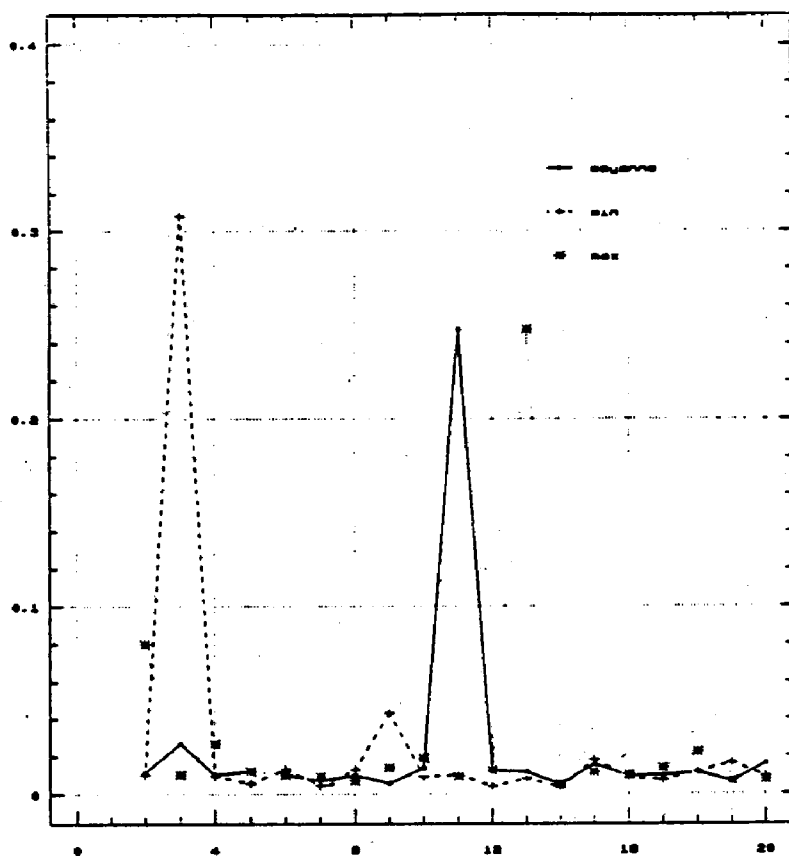
Οι συχνότητες των 6 μεταβλητών για τους 52 νομούς

	Εργοδότης	Δική του	Συμβοηθεί	Μισθωτοί	Απολυθ.	Νέοι άνεργ.
ΑΘΗΝΑΣ	44912	203909	18729	720097	31438	27709
ΑΙΤΩΛΙΑΣ	845	37609	24999	20414	1174	2187
ΥΠΟΛ. ΑΤΤΙΚΗΣ	3475	31666	4995	64693	2650	2419
ΒΟΙΩΤΙΑΣ	426	15298	4096	20453	460	884
ΕΥΒΟΙΑΣ	1256	21620	5957	33103	738	1336
ΕΥΡΥΤΑΝΙΑΣ	89	4899	1555	3717	72	473
ΦΘΙΩΤΙΔΟΣ	676	25181	10515	22126	492	1379
ΦΩΚΙΔΟΣ	295	6714	2314	6523	165	379
ΑΡΓΟΛΙΔΟΣ	761	15681	7931	11147	372	435
ΑΡΚΑΔΙΑΣ	654	19737	10049	12997	162	1244
ΑΧΑΪΑΣ	2964	32372	10380	49161	1234	2915
ΗΛΕΙΑΣ	663	27326	12123	18382	446	1605
ΚΟΡΙΝΘΙΑΣ	1007	20271	7931	16241	550	961
ΛΑΚΩΝΙΑΣ	433	17973	8716	8264	231	693
ΜΕΣΣΗΝΙΑΣ	977	29016	15388	17324	302	1591
ΖΑΚΥΝΘΟΥ	253	5689	2345	2980	62	406
ΚΕΡΚΥΡΑΣ	1251	13808	9026	15464	557	623
ΚΕΦΑΛΛΗΝΙΑΣ	109	4801	780	3470	153	262
ΛΕΥΚΑΔΟΣ	194	3867	2420	2170	140	170
ΑΡΤΗΣ	251	15274	4216	8968	442	1361
ΘΕΣΠΡΩΤΙΑΣ	202	6503	1941	4395	190	661
ΙΩΑΝΝΙΝΩΝ	1408	20732	6024	21666	978	1768
ΠΡΕΒΕΖΗΣ	192	8722	3815	5998	360	732
ΚΑΡΔΙΤΣΗΣ	448	24569	18863	10737	401	1235
ΛΑΡΙΣΗΣ	1728	36598	13498	35065	1361	2450
ΜΑΓΝΗΣΙΑΣ	1997	20611	4421	31685	925	1601
ΤΡΙΚΑΛΩΝ	941	22542	11837	13060	664	1453
ΓΡΕΒΕΝΩΝ	201	6429	1272	4089	235	622
ΔΡΑΜΑΣ	651	14087	5753	13115	521	1407
ΗΜΑΘΙΑΣ	999	20228	5032	17720	802	1371
ΘΕΣ/ΝΙΚΗΣ	13442	80132	12654	176307	8194	8796
ΚΑΒΑΛΑΣ	861	17210	6157	20721	666	1205
ΚΑΣΤΟΡΙΑΣ	1513	7037	2473	10198	53	184
ΚΙΛΚΙΣ	334	14063	4705	9726	482	994
ΚΟΖΑΝΗΣ	927	18603	5666	22410	936	1365
ΠΕΛΛΗΣ	797	24054	11470	12740	461	1562
ΠΙΕΡΙΑΣ	677	16724	8282	11208	754	963
ΞΕΡΡΩΝ	1194	35794	21324	19608	1121	2304
ΦΛΩΡΙΝΑΣ	194	8527	3722	5750	81	620
ΧΑΛΚΙΔΙΚΗΣ	388	12071	1848	10407	351	482
ΕΒΡΟΥ	505	24389	8789	15980	653	1457
ΞΑΝΘΗΣ	643	12377	8884	11874	661	624
ΡΟΔΟΠΗΣ	694	19393	13333	10560	571	720
ΔΩΔΕΚΑΝΗΣΟΥ	1797	12943	1999	27748	593	739
ΚΥΚΛΑΔΩΝ	767	12082	2476	13902	478	394
ΛΕΣΒΟΥ	543	16177	4725	12777	183	532
ΣΑΜΟΥ	259	6129	1481	4754	171	253
ΧΙΟΥ	437	5567	1083	6317	203	245
ΗΡΑΚΛΕΙΟΥ	2910	39140	14705	30450	715	1286
ΛΑΣΙΘΙΟΥ	461	13849	5646	7750	142	602
ΡΕΘΥΜΝΟΥ	424	12115	8193	6376	101	481
ΧΑΝΙΩΝ	1121	17803	6805	17805	393	894

Στο διάγραμμα 3 παριστάνεται το ερμηνευόμενο ποσοστό αδράνειας, σε σχέση με το πλήθος κλάσεων κάθε διαμελισμού, και για τις τρεις μεθόδους και για πλήθος κλάσεων από 2 μέχρι 21. Παρατηρούμε ότι η μέθοδος Min παρουσιάζει σημαντικά μεγαλύτερες τιμές του ποσοστού της ερμηνευόμενης αδράνειας, για πλήθος κλάσεων από 4 μέχρι 12, ενώ για πλήθος κλάσεων από 12 και πάνω οι τρεις μέθοδοι συγκλίνουν ως προς αυτό. Παρόλο που η Min παρουσιάζει μικρότερο ποσοστό ερμηνευόμενης αδράνειας σε σχέση με τη μέθοδο Moyenne για δύο ή τρεις κλάσεις, μπορούμε να θεωρήσουμε ότι η μέθοδος Min είναι συνολικά καλύτερη για τα συγκεκριμένα δεδομένα. Ο χωρισμός σε δύο ή τρεις κλάσεις είναι ανεπαρκής για τα δεδομένα, γιατί το ποσοστό ερμηνευόμενης αδράνειας είναι πολύ μικρό.



Διάγραμμα 3. Γραφική παράσταση του λόγου λ ως προς το πλήθος κλάσεων διαμελισμού για τους νομούς.



Διάγραμμα 4. Γραφική παράσταση των διαφορών του λόγου λ ως προς το πλήθος κλάσεων διαμελισμού για τους νομούς.

Στο διάγραμμα 4 απεικονίζουμε τις μεταβολές του ερμηνευόμενου ποσοστού αδράνειας μεταξύ δύο διαδοχικών διαμελισμών, ως προς το πλήθος κλάσεών τους, και με τις τρεις μεθόδους. Σ' αυτό διακρίνουμε εύκολα τις απότομες μεταβολές του ποσοστού ερμηνευόμενης αδράνειας, που παριστάνονται με τις μεγάλες μεταβολές της γραφικής παράστασης.

Με τη μέθοδο Mougenne ο ιδανικότερος διαμελισμός των συγκεκριμένων δεδομένων επιτυγχάνεται χωρίζοντάς τα σε 12 κλάσεις, με τις οποίες ερμηνεύεται ποσοστό 45.64% της αρχικής ολικής αδράνειας των δεδομένων.

Με τη μέθοδο Max ο ιδανικότερος διαμελισμός επιτυγχάνεται με τον χωρισμό σε 14 κλάσεις, όπου ερμηνεύεται ποσοστό 49.05% της αρχικής ολικής αδράνειας.

Με τη μέθοδο Min ο ιδανικότερος διαμελισμός είναι εκείνος που παίρνουμε χωρίζοντας τα δεδομένα σε 4 κλάσεις με ποσοστό ερμηνευόμενης αδράνειας 34.29%.

8. ΕΠΙΛΟΓΗ ΔΙΑΜΕΛΙΣΜΟΥ ΓΙΑ ΣΥΓΓΕΚΡΙΜΕΝΟ ΠΛΗΘΟΣ ΚΛΑΣΕΩΝ

Πολλές φορές, για διάφορους λόγους (είδος δεδομένων, βιβλιογραφία κλπ.), είναι σκόπιμο να διαμελισθούν τα δεδομένα σε συγκεκριμένο αριθμό, έστω x κλάσεων. Σ' αυτή την περίπτωση, για να βρούμε ποια από τις τρεις μεθόδους δίνει τον καλύτερο διαμελισμό, ελέγχουμε το λόγο αδράνειας του διαμελισμού σε x κλάσεις προς την αρχική ολική αδράνεια, και με τις τρεις μεθόδους. Ο λόγος αυτός υπολογίζεται από τον τύπο:

$$\lambda_x = \frac{\frac{1}{n} \sum_{k=1}^x p(k) \sum_{j=1}^p [g_k(j) - g(j)]^2}{\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p [A(i,j) - g(j)]^2}$$

όταν για την απόσταση χρησιμοποιούμε την ευκλείδεια μετρική, και:

$$\lambda_x = \frac{\frac{1}{n} \sum_{k=1}^x p(k) \sum_{j=1}^p \frac{g_k(j) - \frac{A(.,j)}{n} \sum_{j=1}^p \frac{g_k(j)}{p(k)}}{\frac{A(.,j)}{n} \sum_{j=1}^p g_k(j)}}{\sum_{i=1}^n \sum_{j=1}^p \frac{(A(i,j) - A(i,.) \frac{A(.,j)}{n})^2}{A(i,.) A(.,j)}}$$

για την απόσταση με τη χ^2 μετρική.

Η μέθοδος που παρουσιάζει τη μεγαλύτερη τιμή του λόγου αυτού δίνει τον ιδανικότερο διαμελισμό σε x κλάσεις.

Για να διαπιστωθεί αυτό πιο παραστατικά, απεικονίζεται γραφικά ο λόγος λ_x ως προς τον αριθμό κλάσεων, στο ίδιο διάγραμμα και για τις τρεις μεθόδους. Η μέθοδος της οποίας η γραφική παράσταση για το συγκεκριμένο αριθμό κλάσεων βρίσκεται υψηλότερα από τις υπόλοιπες δίνει τον ιδανικότερο διαμελισμό.

Είναι δυνατό μια μέθοδος να θεωρηθεί συνολικά καλύτερη για τα δεδομένα μας, ενώ μια άλλη να ερμηνεύει μεγαλύτερο ποσοστό αρχικής ολικής αδράνειας για συγκεκριμένο πλήθος κλάσεων. Έτσι, ενώ στο δεύτερο παράδειγμα θεωρήσαμε γενικά καλύτερη τη μέθοδο Min, αν θέλουμε να χωρίσουμε τα δεδομένα σε 13 κλάσεις (διάγραμμα 3), πρέπει να διαλέξουμε τον διαμελισμό που προκύπτει από τη μέθοδο Moyenne και ερμηνεύει ποσοστό 46.89% της ολικής αρχικής αδράνειας (ο διαμελισμός σε 13 κλάσεις της μεθόδου Min ερμηνεύει 45.43%, ενώ της μεθόδου Max 24.34%).

9. ΣΥΜΠΕΡΑΣΜΑΤΑ

Με την εργασία αυτή, για κάθε συγκεκριμένο σύνολο δεδομένων, ποσοτικών ή ποιοτικών μεταβλητών, προτείνουμε ένα τρόπο προσδιορισμού για το ποια από τις τρεις μεθόδους ταξινόμησης (Min, Max, Moyenne) οδηγεί στην ιεράρχηση, που είναι καλύτερα προσαρμοζόμενη σ' αυτά τα δεδομένα. Αυτό επιτυγχάνεται με τον υπολογισμό του λόγου της αδράνειας του κάθε διαμελισμού προς την αρχική ολική αδράνεια, σε συνδυασμό με τη μελέτη δύο γραφικών παραστάσεων. Στην πρώτη απεικονίζεται γραφικά ο λόγος λ_i , που εκφράζει το ποσοστό της ολικής αδράνειας, που ερμηνεύεται από τον κάθε διαμελισμό, ως προς τον αριθμό κλάσεων και για τις τρεις μεθόδους στο ίδιο διάγραμμα, ενώ στη δεύτερη οι μεταβολές d_i του λόγου αυτού μεταξύ διαδοχικών διαμελισμών, πάλι σε κοινό διάγραμμα και για τις τρεις μεθόδους.

Επίσης προτείνουμε για κάθε μία από τις τρεις μεθόδους τον προσδιορισμό του ιδανικότερου διαμελισμού των δεδομένων, χωρίς να είναι γνωστός εκ των προτέρων ο αριθμός των κλάσεων. Ο ιδανικότερος διαμελισμός είναι εκείνος στον οποίο παρατηρείται απότομη μεταβολή του λόγου αδράνειας του διαμελισμού προς την αρχική ολική αδράνεια. Και αυτό γιατί αν δεχθούμε τον διαμελισμό με μία κλάση λιγότερη, χάνουμε σημαντικό ποσοστό ερμηνευόμενης αδράνειας, ενώ προχωρώντας στο διαμελισμό με μία κλάση περισσότερη διαχωρίζουμε τα δεδομένα σε επιπλέον κλάσεις, χωρίς να κερδίζουμε μεγάλο ποσοστό ερμηνευόμενης αδράνειας.

Τέλος, επιθυμώντας τον διαχωρισμό των δεδομένων σε συγκεκριμένο πλήθος κλάσεων, προσδιορίζουμε ποια μέθοδος δίνει διαμελισμό με τον οποίο ερμηνεύεται το μεγαλύτερο ποσοστό της αρχικής ολικής αδράνειας. Αυτό επιτυγχάνεται συγκρίνοντας τους λόγους αδράνειας του διαμελισμού προς την αρχική ολική αδράνεια μεταξύ των τριών μεθόδων, για το συγκεκριμένο πλήθος κλάσεων. Η μέθοδος που παρουσιάζει μεγαλύτερη τιμή του λόγου αυτού, δίνει τον καλύτερο διαμελισμό σ' αυτό το πλήθος των κλάσεων.

ABSTRACT

In this paper, we try to find the best method of classification, between three methods (min, max, moyenne), for certain data.

We calculate the ratio λ , which is: (variance of partition)/(total variance), for every partition with number of classes from n (number of data) to 2.

The method which presents the greater value for λ , is the best fitting method.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Benzecri J. P., Benzecri F., Cheung Y. L., Maiza S. (1985), "Aides a l' interpretation et etiquetage des arbres en classification ascendante hierarchique; listages FACOR, VACOR et INSUP", *Les Cahiers de l' Analyse des Donnees*, n° 3.
- Benzecri J. P. et Collaborateurs (1973), *L' analyse des donnees*, Dunod, Paris.
- Bertier P., Bouroche J.-M. (1981), *Analyse des donnees multidimensionnelles*, Presses Universitaires de France, Paris.
- Diday E., Lemaire J., Pouget J., Testu F. (1982), *Elements d' analyse de donnees*, Dunod, Paris.
- Escofier B., Pages J. (1990), *Analyses factorielles simples et multiples*, Dunod, Paris.
- Jambu M., Benzecri J. P., Lebeaux M. O. (1980), "Aide a l' interpretation en classification automatique", *Les Cahiers de l' Analyse des Donnees*, n° 1.
- Pages J., Cailliez F. (1976), *Introduction a l' Analyse des Donnees*, Smash, Paris.
- Παπαδημητρίου Γ., Κaráκος Α., Κουτρομανίδης Θ. (1993), "Εφαρμογή μεθόδων της ανάλυσης δεδομένων στη χωροταξία των νομών της Ελλάδας ως προς την απασχόληση", *Τιμητικός τόμος στη μνήμη του Καθηγητή Σ. Ζευγαρίδη*, Πανεπιστήμιο Μακεδονίας.
- Roux M. (1985), *Algorithmes de classification*, Masson, Paris.