ΠΥΘΑΓΟΡΕΙΟΣ ΑΚΑΔΗΜΙΑ

Εμπνευστής: Dr. Δημήτριος Ν. Καραπιστόλης

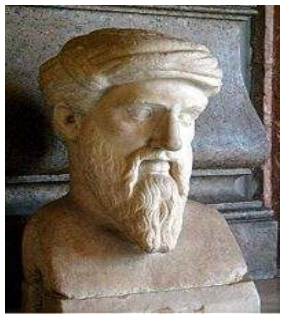
Η Πυθαγόρειος Ακαδημία έχει ως στόχο να εκπαιδεύσει ερευνητές οι οποίοι θα αποκτήσουν τα κατάλληλα εφόδια να πραγματοποιούν έρευνες οι οποίες θα βασίζονται στο Πυθαγόρειο θεώρημα σε συνδυασμό με την μέθοδο της Παραγοντικής Ανάλυσης των Αντιστοιχιών, δημιουργία του Γάλλου καθηγητή Jean Paul Benzecri.

Η κύρια συνεισφορά της Ακαδημίας στον Ελλαδικό χώρο, συνίσταται στο ότι συγκεντρώνει τις δύο βασικές μεθόδους της Παραγοντικής Ανάλυσης των Αντιστοιχιών και της Ανιούσας Ιεραρχικής Ταξινόμησης, όπως αυτές εφαρμόζονται από την Γαλλική Σχολή, σε μία μόνο διαδικασία, εφαρμόζοντας το Πυθαγόρειο Θεώρημα στον Ευκλείδειο ν-διάστατο διανυσματικό χώρο R^{v}.

Α) Βασικές Αρχές στις οποίες στηρίζεται η ερευνητική σκέψη της Ακαδημίας

1. Στο Πυθαγόρειο θεώρημα

Το Πυθαγόρειο θεώρημα αποτελεί τη βάση υπολογισμού της απόστασης μεταξύ δύο σημείων σε οποιονδήποτε διανυσματικό χώρο R^{v} και αν ανήκουν.



«Εν τοις ορθογωνίοις τριγώνοις το από της την ορθήν γωνίαν υποτεινούσης πλευράς τετράγωνον ίσον εστί τοις από των την ορθήν γωνίαν περιεχουσών πλευρών τετραγώνοις».

Πυθαγόρας περίπου 572 π.Χ-490 π.Χ

2. Στη ρήση του Επίκουρου

Για τους διαφωνούντες σχετικά με την χρησιμότητα των εφαρμογών των μεθόδων της Ανάλυσης Δεδομένων, κατάλληλη είναι η ρήση του φιλοσόφου Επίκουρου:

«Ποτέ δεν επιθύμησα να γίνω αρεστός στους πολλούς. Αφ' ενός, δεν κάθισα να μάθω τι αρέσει στους πολλούς κι αφ' ετέρου, τα όσα ήξερα εγώ βρίσκονταν μακριά από τη δική τους αντίληψη».

Επίκουρος 341 π.Χ- 270 π.Χ

3. Στη ρήση του J.P BENZECRI

Ο J.P Benzecri, πρωτοπόρος της Γαλλικής Σχολής της Ανάλυσης Δεδομένων και εμπνευστής της Παραγοντικής Ανάλυσης των Αντιστοιχιών, στην αρχή του μαθήματος στο Πανεπιστήμιο Pierre et Marie Curie (Jussieu Paris VI) σχετικά με την παραγοντική ανάλυση ανέφερε τα εξής:

«Στη καρδιά κάθε ανάλυσης δεδομένων υπάρχει διαγωνοποίηση ενός συμμετρικού τετραγωνικού πίνακα».

Jean Paul Benzecri

Β) Θεωρητικό πλαίσιο στο οποίο στηρίζεται η ερευνητική σκέψη της Σχολής:

1. **Διακύβευμα: Η πληροφορία**.
2. **Μαθηματικός χώρος επεξεργασίας των πληροφοριών: Ο ν-διάστατος Ευκλείδειος διανυσματικός χώρος**
3. **Μέθοδοι επεξεργασίας των πληροφοριών: Οι μέθοδοι της Ανάλυσης Δεδομένων της Γαλλικής Σχολής.**

α) Διακύβευμα: Η πληροφορία

Είναι γνωστό ότι στους περισσότερους επιστημονικούς κλάδους σήμερα κυριαρχεί το τεχνολογικό σύνδρομο. Η κοινωνία καθοδηγείται δίχως να έχει συνήθως άλλη επιλογή, στους τεχνολογικούς ρυθμούς που επιβάλλει η σύγχρονη αντίληψη στις σχέσεις και στη συμπεριφορά του ανθρώπου. Αυτό έχει ως αποτέλεσμα να αναθεωρούνται και να αναπροσαρμόζονται πολλοί τομείς της ανθρώπινης δράσης και σκέψης. Συνέπεια αυτής της νέας τάσης είναι να αναδυθεί ένας νέος οικονομικός πόρος η **πληροφορία**. Η επεξεργασία του πόρου αυτού ως γνωστό είναι αντικείμενο μιας νέας επιστήμης που δεσπόζει στον $20^{ο}$ αιώνα, της **πληροφορικής**.

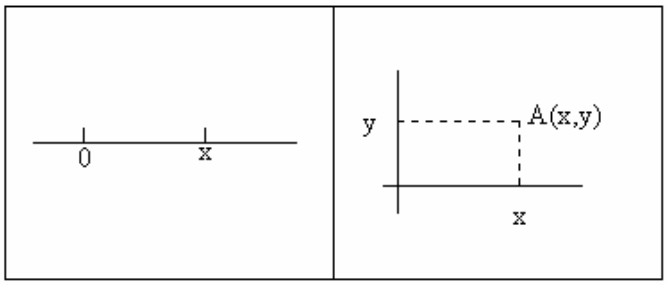
β) Ο ν-διάστατος Ευκλείδειος διανυσματικός χώρος

Ως γνωστό τα ευθύγραμμα τμήματα, οι ορθογώνιοι άξονες, οι κύκλοι, οι σφαίρες δημιούργησαν μία μοναδική Γεωμετρία που ενέπνευσε όχι μόνο την περίφημη φιλοσοφία της πλατωνικής αρμονίας, αλλά και χιλιάδες καλλιτέχνες και επιστήμονες, ανά τους αιώνες.

Η Ευκλείδεια Γεωμετρία όμως άφησε μια βαριά κληρονομιά στον τρόπο διερεύνησης της πραγματικότητας :την έννοια της ακεραίας διάστασης. Μάθαμε δηλαδή να ζούμε, από την εποχή του Ευκλείδη, στο τριδιάστατο χώρο να σχεδιάζουμε στο επίπεδο, να κινούμαστε επί ευθείας ή τεθλασμένης γραμμής.

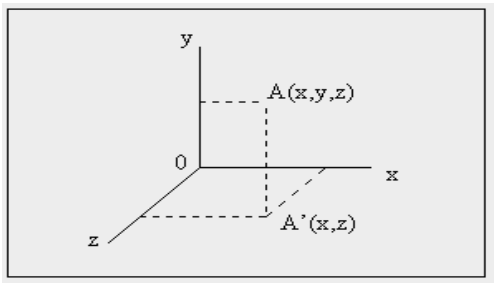
Γνωρίζουμε πως ένας αριθμός x μπορεί να χρησιμοποιηθεί για να παρασταθεί ένα σημείο μιας ευθείας, αφού προηγουμένως έχουμε καθορίσει μια μονάδα μήκους. Δύο

αριθμοί, όπως π.χ. το ζεύγος (x, y) μπορούν να χρησιμοποιηθούν για να παραστήσουν ένα σημείο του επιπέδου. Σχηματικά έχουμε:



σχήμα 1.α σχήμα 1.β

Σημειώνουμε τέλος, ότι μπορούμε να χρησιμοποιήσουμε μια τριάδα αριθμών (x, y, z) για να παρασταθεί ένα σημείο στο χώρο, όπως τον αντιλαμβανόμαστε με τις αισθήσεις μας (σχ. 1.γ).



σχήμα 1.γ

Η ευθεία ονομάζεται **χώρος μιας διάστασης,** το επίπεδο **χώρος δύο διαστάσεων** και ο χώρος που αντιλαμβανόμαστε με τις αισθήσεις μας **χώρος τριών διαστάσεων.**

Στη συνέχεια παρ' ότι δεν μπορούμε να κάνουμε μια γραφική παράσταση δεν μας εμποδίζει να θεωρήσουμε μια τετράδα αριθμών (x_1, x_2, x_3, x_4) και να θεωρήσουμε ότι πρόκειται για ένα σημείο ενός χώρου τεσσάρων διαστάσεων. Γενικά, ορίζουμε ένα σημείο ενός χώρου ν διαστάσεων ως μια διατεταγμένη ν-ιάδα αριθμών (x_1, x_2, ...x_v) όπου ν ακέραιος αριθμός.

Ορισμός: Με τον όρο διατεταγμένη ν-ιάδα εννοούμε ότι η αλλαγή της τάξεως αναγραφής των αριθμών προσδιορίζει εν γένει, διαφορετικό διάνυσμα.

Αντίστροφα, σε κάθε διατεταγμένη ν-ιάδα αριθμών αντιστοιχεί μοναδικό σημείο του χώρου των ν διαστάσεων.

Αν V είναι ο χώρος των σημείων των ν διαστάσεων, ορίζουμε μια αμφιμονοσήμαντη αντιστοιχία των σημείων του χώρου V στις διατεταγμένες ν-ιάδες του

$$R^v f: X \in V \longrightarrow (x_1, x_2, \ldots x_v) \in R^v$$

Τότε το σύνολο V είναι ισοδύναμο προς το σύνολο R^v. Για το λόγο αυτό ταυτίζουμε τον χώρο V με τον R^v όπου

$$R^v = \underbrace{RxRx...xR}_{\text{ν φορές}} = \{(x_1, x_2, ... x_v) / x_1, x_2, ... x_v \in R\}$$

Μια τέτοια ν-ιάδα την συμβολίζουμε μ' ένα κεφαλαίο γράμμα π.χ Α, ενώ με τα αντίστοιχα μικρά γράμματα τους πραγματικούς αριθμούς που ονομάζουμε **συντεταγμένες** του σημείου Α.

γ. Ανάλυση Δεδομένων, μία οικογένεια πολυδιάστατων μη παραμετρικών στατιστικών μεθόδων

Πολλές φορές ένας ερευνητής πρέπει να πάρει μια απόφαση μελετώντας ένα φαινόμενο του οποίου όμως αγνοεί τους συγκεκριμένους μηχανισμούς λειτουργίας του. Ο φυσικός αντιθέτως, όταν θέλει σ' ένα αέριο να διπλασιάσει την πίεσή του, αρκεί να υποδιπλασιάσει τον όγκο του αερίου, υπό σταθερά θερμοκρασία (Νόμος Boyle Mariotte).

Ο φυσικός γνωρίζει πράγματι ένα συγκεκριμένο μοντέλο του φαινομένου, το οποίο περιγράφει ο προαναφερόμενος νόμος. Ο οικονομέτρης χρησιμοποιεί και αυτός μοντέλα, όπου τουλάχιστον μια παράμετρος είναι άγνωστη, την οποία με κατάλληλες μεθόδους προσπαθεί να προσδιορίσει.

Έτσι λ.χ αν επιθυμεί κάποιος να προβλέψει τις πωλήσεις μεγεθών ενός σχεδίου ρούχων, υποθέτει εκ των προτέρων ότι η κατανομή των υψών των ανθρώπων ακολουθεί το νόμο του Gauss, δηλαδή τον κανονικό νόμο. Ακολούθως πραγματοποιεί δειγματοληπτική έρευνα σε διάφορα άτομα, για να εκτιμήσει τις άγνωστες παραμέτρους "μέσο" και "διακύμανση" των υψών των καταναλωτών. Στη συνέχεια με βάση τις παρατηρήσεις που συνέλλεξε και μετά την μαθηματική τους επεξεργασία, καταλήγει σ' ένα μοντέλο πρόβλεψης των πωλήσεων ρούχων, χρήσιμο για τις μελλοντικές του επιχειρηματικές δραστηριότητες.

Βέβαια ο κανονικός νόμος, όπως είναι γνωστό, ασχολείται με τη φύση του τυχαίου, αλλά ως μέσον για να βρεθούν μονοπάτια στο άγριο δάσος της οικονομίας, μάλλον δεν ανταποκρίνεται στην πραγματικότητα. Για το θέμα αυτό αξιοσημείωτη είναι η ρήση του νομπελίστα Βασίλη Λεόντιεφ.

«**Σε κανένα πεδίο εμπειρικής αναζήτησης δεν έχει χρησιμοποιηθεί σε τέτοια έκταση και τόσο επιτηδευμένα μια στατιστική τεχνική με τόσο πενιχρά αποτελέσματα**». [Gleick J.,1990 σ.122]

Την διαπίστωση αυτή την συνειδητοποίησαν αρκετοί ερευνητές με αποτέλεσμα να θεωρούν ότι δεν είναι δυνατόν σε κάθε περίπτωση να χρησιμοποιούνται υποδείγματα. Αυτό μπορεί να οφείλεται είτε στο ότι αγνοείται ο νόμος των πιθανοτήτων που διέπει το φαινόμενο είτε ο θεωρούμενος νόμος δεν ικανοποιεί πλήρως την περιγραφή του φαινομένου.

Ως παράδειγμα μπορούμε να αναφέρουμε τις πωλήσεις ρούχων, όπου δεν είναι απίθανο η κατανομή του ύψους των ανθρώπων να μην ακολουθεί τον κανονικό νόμο, αφού μπορεί να σκεφθεί κανείς ότι δεν υπάρχει άτομο ούτε με ύψος μηδέν, αλλά ούτε και με ύψος 2,60 μέτρα κάτι βέβαια που προβλέπεται θεωρητικά από τον κανονικό νόμο.

Η ανάγκη λοιπόν να μη θεωρείται εκ των προτέρων ότι ένα φαινόμενο ακολουθεί κάποιο συγκεκριμένο νόμο, οδήγησε στην εφαρμογή νέων στατιστικών μη παραμετρικών μεθόδων, κάτω από την ονομασία Ανάλυση Δεδομένων ή όπως αλλιώς μπορεί να την αποκαλέσουμε **Στατιστική δίχως μοντέλα**.

ΠΟΛΥΔΙΑΣΤΑΤΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

- **Ιστορική αναδρομή**

- ✓ Στην αρχή του 20ου αιώνα οι Ευρωπαίοι ψυχολόγοι έψαχναν με τα τέστ που υπέβαλαν στους αρρώστους και τους βαθμούς που συγκέντρωναν από διάφορες μεταβλητές που χρησιμοποιούσαν (μνήμη, ευφυΐα, κ.ά), να βρουν σύνθετες μεταβλητές οι οποίες όχι μόνο δεν παρατηρούνται απ' ευθείας από τα αρχικά δεδομένα, αλλά θα ερμήνευαν κατά τον καλύτερο τρόπο τη συμπεριφορά των αρρώστων όσων βέβαια παρουσίαζαν τα ίδια συμπτώματα. Τις μεταβλητές αυτές τις ονόμασαν «παράγοντες».
- ✓ Οι κυριότεροι επιστήμονες που δημιούργησαν τις προϋποθέσεις ανάδυσης μιας οικογένειας στατιστικών μεθόδων, με την ονομασία Ανάλυση Δεδομένων (Data analysis) είναι οι παρακάτω**:**
- ✓ C. Spearman (1904) Εργασίες πάνω στη μήτρα Διακυμάνσεων-Συνδιακυμάνσεων
- ✓ H. Hottelling (1933) Ανάλυση σε κύριες Συνιστώσες
- ✓ R.A. Fisher (1940) διατυπώνει τους διακριτικούς παράγοντες
- ✓ L.Guttman (1941) Το ομώνυμο φαινόμενο
- ✓ C. Burt (1950) Τον ομώνυμο πίνακα
- ✓ Πάντως η πιο πρόσφατη και αποτελεσματικότερη μέθοδος της οικογένειας δημιουργήθηκε κατά τη δεκαετία του '60 από το Γάλλο καθηγητή J. P Benzecri (1973) του Πανεπιστημίου Paris VI με την ονομασία Παραγοντική Ανάλυση των Αντιστοιχιών (Analyse Factorielle des Correspondances -A.F.C-), η οποία επεξεργάζεται κυρίως ποιοτικά δεδομένα που παρουσιάζονται υπό μορφή πολυδιάστατων πινάκων συμπτώσεων
- ✓ Στα τέλη της δεκαετίας του '70, αρχές της δεκαετίας του '80, μαθητές του J.P. Benzecri, όπως οι Γάλλοι Lebart, Roux, Escoffier, Morineau, Fenelon συντελέσαν όχι μόνο στη διάδοση αλλά και την καθιέρωση των μεθόδων αυτών στη συνείδηση πολλών ερευνητών σ' ολόκληρο τον κόσμο, δημιουργώντας την Γαλλική Σχολή της Ανάλυσης Δεδομένων.
- ✓ Αλλά και στην άλλη μεριά του Ατλαντικού δημιουργήθηκε η λεγόμενη Αμερικανική σχολή με τους J.D Carrol, J. B. Kruskal, R. S Sheppard, G. Yang κ.λ.π κάτω από το όνομα «multidimensional scaling», της οποίας η ευρηματικότητα δεν συγκρίνεται με εκείνη της Γαλλικής Σχολής.
- ✓ Αλλά και στην άλλη μεριά του Ατλαντικού δημιουργήθηκε η λεγόμενη Αμερικανική σχολή με τους J.D Carrol, J. B. Kruskal, R. S Sheppard, G. Yang κ.λ.π κάτω από το όνομα «multidimensional scaling», της οποίας η ευρηματικότητα δεν συγκρίνεται με εκείνη της Γαλλικής Σχολής.
- ✓ Αλλά και στην άλλη μεριά του Ατλαντικού δημιουργήθηκε η λεγόμενη Αμερικανική σχολή με τους J.D Carrol, J. B. Kruskal, R. S Sheppard, G. Yang κ.λ.π κάτω από το

όνομα «multidimensional scaling», της οποίας η ευρηματικότητα δεν συγκρίνεται με εκείνη της Γαλλικής Σχολής.

- ✓ Σύμφωνα με αυτόν η Ανάλυση Δεδομένων περιλαμβάνει:
- ✓ Τρόπους σχεδιασμού και συλλογής των δεδομένων προκειμένου η ανάλυση να καταστεί πιο εύκολη, ορθή και ακριβής
- ✓ Διαδικασίες ανάλυσης
- ✓ Τεχνικές ερμηνείας και αποτελεσμάτων

Οι πιο σημαντικές αρχές της Γαλλικής Σχολής της Ανάλυσης Δεδομένων, όπως τις αναφέρει ο ιδρυτής καθηγητής Jean Paul Benzecri είναι οι εξής:

1η αρχή: Η Στατιστική δεν πρέπει να συγχέεται με την θεωρία των πιθανοτήτων. Πολλοί στατιστικοί έχουν δομήσει τη μαθηματική στατιστική με πολλές υποθέσεις που σπάνια ή ποτέ δεν ικανοποιούνται στην πράξη.

2η αρχή: Το μοντέλο θα πρέπει να διαμορφώνεται ακολουθώντας τα δεδομένα και όχι το αντίστροφο. Αγνοώντας την αρχή αυτή οδηγείται κανείς σ' ένα μεγάλο λάθος της εφαρμογής των μαθηματικών στις επιστήμες της ανθρώπινης συμπεριφοράς. Επιχειρείται δηλαδή η προσαρμογή των δεδομένων στα μοντέλα που έχουν δημιουργηθεί εκ των προτέρων. Ο στόχος αυτός τις περισσότερες φορές δεν υλοποιείται, καθόσον στην ανθρώπινη συμπεριφορά πρωτεύοντα ρόλο διαδραματίζει η μνήμη, οπότε συμπεριφορές του παρελθόντος πολλές φορές δεν είναι παρόμοιες με εκείνες του μέλλοντος, διαδικασία που προκαθορίζεται με την θέσπιση οποιουδήποτε μοντέλου.

3η αρχή: Είναι πρωταρχικής σημασίας να χειρίζεται κανείς πληροφορίες σε όσο το δυνατόν περισσότερες διαστάσεις, γεγονός πολύ δύσκολο με την επίλυση των μαθηματικών μοντέλων.

ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΤΗΣ ΓΑΛΛΙΚΗΣ ΣΧΟΛΗΣ

- ➢ 1ον) Μεγάλο πλεονέκτημα της Γαλλικής σχολής είναι η εξέταση ενός φαινομένου χρησιμοποιώντας περισσότερες από δύο μεταβλητές, αντιμετωπίζοντας έτσι ένα μεγάλο μειονέκτημα πολλών μεθόδων της κλασικής Στατιστικής. Για τον λόγο αυτό χρειάστηκε να χρησιμοποιηθούν τα μαθηματικά του Ευκλείδειου ν-διάστατου χώρου με αποτέλεσμα:
- ➢ 2ον) Κάθε φαινόμενο επειδή από την φύση του είναι σύνθετο στο οποίο υπεισέρχονται πληθώρα παραγόντων, να εξετάζεται με βάση την συνολική αλληλεξάρτηση των παραγόντων αυτών και όχι πλέον δύο-δύο χωριστά.
- ➢ 3ον) Την διανυσματοποίηση ποιοτικών χαρακτηριστικών, δίχως την ανάγκη να χρησιμοποιηθούν οι λεγόμενες ψευδομεταβλητές, που από την ονομασία τους και μόνο καταδεικνύεται η λογική αφαίρεση που πραγματοποιείται από τους θιασώτες της κλασικής Στατιστικής (κυρίως από τους Οικονομέτρες).

ΚΛΑΣΙΚΗ ΣΤΑΤΙΣΤΙΚΗ VS ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Η μέθοδος των ελαχίστων τετραγώνων, βασική μέθοδος κατασκευής υποδειγμάτων αποτίμησης κυρίως οικονομικών δεδομένων, θεωρείται από πάρα πολλούς μελετητές ανεπαρκής, επειδή παρουσιάζει αμερόληπτες και συνεπείς εκτιμήσεις των συντελεστών παλινδρόμησης μόνο κάτω από αυστηρές υποθέσεις οι οποίες συνήθως δεν ανταποκρίνονται στη πραγματικότητα.

Η συμπεριφορά επομένως ενός κοινωνικοοικονομικού συστήματος δε μπορεί να ερμηνευθεί παρά μόνο με αφετηρία τη συνολική θεώρηση του υπό μελέτη φαινομένου, η οποία θα επιτρέψει, όπως προαναφέραμε, την κατανόηση των διαρθρωτικών λειτουργιών του, δίχως να τίθενται προϋποθέσεις στις αλληλεξαρτήσεις των στοιχείων του, ούτε και στο πλήθος των παραγόντων που υπεισέρχονται στη μελέτη.

Πολλές φορές ένας ερευνητής πρέπει να πάρει μια απόφαση μελετώντας ένα φαινόμενο του οποίου όμως αγνοεί τους συγκεκριμένους μηχανισμούς λειτουργίας του.

Έτσι λ.χ. αν επιθυμεί κάποιος να προβλέψει τις πωλήσεις μεγεθών ενός σχεδίου ρούχων, υποθέτει εκ των προτέρων ότι η κατανομή των υψών των ανθρώπων ακολουθεί το νόμο του Gauss, δηλαδή τον κανονικό νόμο. Ακολούθως πραγματοποιεί δειγματοληπτική έρευνα σε διάφορα άτομα, για να εκτιμήσει τις άγνωστες παραμέτρους «μέσο» και «διακύμανση» των υψών των καταναλωτών.

Στη συνέχεια με βάση τις παρατηρήσεις που συνέλλεξε και μετά την μαθηματική τους επεξεργασία, καταλήγει σ' ένα μοντέλο πρόβλεψης των πωλήσεων ρούχων, χρήσιμο για τις μελλοντικές του επιχειρηματικές δραστηριότητες.

Βέβαια ο κανονικός νόμος επειδή ασχολείται με τη φύση του τυχαίου μάλλον δεν ανταποκρίνεται στην πραγματικότητα, αφού μπορεί να σκεφθεί κανείς ότι δεν υπάρχει άτομο ούτε με ύψος μηδέν, αλλά ούτε και με ύψος 2,80 μέτρα κάτι βέβαια που προβλέπεται θεωρητικά από τον κανονικό νόμο.

Η ανάγκη λοιπόν να μη θεωρείται εκ των προτέρων ότι ένα φαινόμενο ακολουθεί κάποιο συγκεκριμένο νόμο, οδήγησε στην εφαρμογή νέων στατιστικών μη παραμετρικών μεθόδων, κάτω από την ονομασία **Ανάλυση Δεδομένων** ή όπως αλλιώς μπορεί να την αποκαλέσουμε **Στατιστική δίχως μοντέλα.**

Το βασικό πλεονέκτημα, λοιπόν, του ερευνητή με τις μεθόδους της ανάλυσης δεδομένων ως μη παραμετρικές είναι ότι δεν απαιτείται να λάβει υπόψη του καμιά υπόθεση ως προς τις παραμέτρους που διαμορφώνουν το υπό μελέτη φαινόμενο**.**

Απλά γίνεται προσπάθεια να ανακαλυφθεί η συμπεριφορά των στοιχείων που συνθέτουν το φαινόμενο και να προσδιοριστεί η δομή του συστήματος που αντιστοιχεί στο φαινόμενο**.**

Ο εντοπισμός, πάντως, των αλληλεξαρτημένων παραγόντων (ποσοτικών και ποιοτικών) που επηρεάζουν τα στοιχεία ενός υποσυστήματος (δείγματος), εξετάζοντας τις ροές των πληροφοριών που δημιουργούν οι αλληλεπιδράσεις αυτές, δίνει την δυνατότητα στον ερευνητή να προσεγγίσει συνολικά την αρχιτεκτονική δομή του συστήματος (πληθυσμού).

Ως επακόλουθο αυτής της δυνατότητας είναι να αποκτά ο ερευνητής μια σφαιρική άποψη του συνόλου I των στατιστικών μονάδων (σύνολο των στοιχείων του συστήματος), το οποίο περιγράφεται από ένα σύνολο J ιδιοτήτων (σύνολο των μεταβλητών του προβλήματος), θέτοντας σε γεωμετρική μορφή το σύστημα των σχέσεων που υφίστανται μεταξύ των στοιχείων των δύο αυτών συνόλων, χρησιμοποιώντας τις ιδιότητες του Ευκλείδειου διανυσματικού χώρου R^N.

Η παραγοντική ανάλυση, πάνω στην οποία βασίζονται οι μέθοδοι της Ανάλυσης δεδομένων, αναπαριστά γραφικά διαφαινόμενες ή μη εκ των προτέρων σχέσεις που υφίστανται μεταξύ των συνόλων I και J τις οποίες προβάλλει πάνω σ' ένα επίπεδο, με στόχο την ευκολότερη ερμηνεία των σχέσεων αυτών.

Με βάση αυτή τη διαδικασία, το κύριο χαρακτηριστικό των μεθόδων της ανάλυσης δεδομένων μπορεί να αποδοθεί με το εξής σλόγκαν:

«Αποδέχομαι μία απώλεια πληροφορίας προκειμένου να επιτύχω ένα όφελος σε σημασία.»

Η προσέγγιση αυτή είναι καθαρά ποιοτικού περιεχομένου και επιτυγχάνεται με αμιγείς αναλυτικές μεθόδους και αποτελεί τη νέα αντίληψη που εισάγεται τις από τις μεθόδους της Ανάλυσης Δεδομένων.

Η τοποθέτηση της Ανάλυσης των Δεδομένων σε σχέση με τις κλασικές επιστήμες μπορεί να αποτυπωθεί με τον παρακάτω πίνακα.

Πίνακας 1

Συγκριτικές τοποθετήσεις μεταξύ κλασικών μεθόδων και μεθόδων ανάλυσης δεδομένων

Φυσικός	Χρησιμοποιεί μοντέλα που επιβεβαιώνονται από πειράματα	Χρησιμοποιεί δείγματα για να επιβεβαιώσει κάποια θεωρία
Κλασικός στατιστικός	Κάνει υποθέσεις πάνω σε νόμους πιθανοτήτων ή μοντέλα	Χρησιμοποιεί δείγματα για να εκτιμήσει παραμέτρους νόμων
Αναλυτής δεδομένων	Δεν κάνει καμιά υπόθεση	Από το δείγμα ψάχνει να βρει τις σχέσεις των στοιχείων βάσει των οποίων θα περιγράψει τη συμπεριφορά τους και τη δομή του φαινομένου

Η Πολυδιάστατη Στατιστική Ανάλυση ή απλά Ανάλυση Δεδομένων συγκεντρώνει πολυάριθμες στατιστικές μεθόδους διαφορετικές μεταξύ τους. Μπορούμε όμως να διακρίνουμε δύο κύριες κατευθύνσεις:

α) Την παραγοντική ανάλυση η οποία επιτρέπει να αποκαλύψουμε τη δομή και τις αλληλεξαρτήσεις ενός συνόλου, εξετάζοντας κάποιο παραγοντικό επίπεδο. Η παραγοντική ανάλυση είναι ένα εργαλείο που χρησιμοποιείται από τον ερευνητή για να παρατηρήσει αυτό που θέλει με τον ίδιο τρόπο που το τηλεσκόπιο ή το μικροσκόπιο χρησιμοποιείται σε άλλους επιστημονικούς χώρους.

β) Την αυτόματη ταξινόμηση, η οποία συνίσταται στο να κατατάξει τις στατιστικές μονάδες σε ομοιογενείς ομάδες, οι οποίες επηρεάζονται ταυτόχρονα από διάφορους παράγοντες, με την βοήθεια κάποιου Αλγορίθμου ταξινόμησης.

ΠΗΓΑΙΝΟΝΤΑΣ ΕΝΑ ΒΗΜΑ ΠΑΡΑΠΕΡΑ

Για να μπορέσει κάποιος να αντιληφθεί τις **Βασικές Αρχές στις οποίες στηρίζεται η ερευνητική σκέψη της Πυθαγορείου Ακαδημίας**, θα πρέπει να μελετήσει τα παρακάτω κείμενα, ώστε να κατανοήσει καλύτερα την επέκταση των σκέψεων που προτείνει η

Ακαδημία, όσον αφορά στις βασικές αρχές του θεμελιωτή της Γαλλικής Σχολής της Ανάλυσης Δεδομένων καθηγητού Jean Paul Benzecri.

ΜΕΘΟΔΟΛΟΓΙΑ ΤΗΣ ΔΙΕΡΕΥΝΗΣΗΣ ΔΕΔΟΜΕΝΩΝ ΜΕ ΒΑΣΗ ΤΗΝ ΠΑΡΑΓΟΝΤΙΚΗ ΑΝΑΛΥΣΗ ΠΟΛΥΔΙΑΣΤΑΤΩΝ ΔΕΔΟΜΕΝΩΝ

Γενικά

Στόχος μιας παραγοντικής ανάλυσης χρησιμοποιώντας μία από τις μεθόδους της, είναι η **οπτικοποίηση** των αλληλεξαρτήσεων και των αλληλεπιδράσεων μεταξύ των «αντικειμένων» (γραμμών) και των μεταβλητών (στηλών) ενός πίνακα δεδομένων.

Πρέπει να τονιστεί ιδιαίτερα ότι η δημιουργία ενός παραγοντικού άξονα **ποσοτικοποιεί** την συμπεριφορά των «αντικειμένων» και των μεταβλητών, με αποτέλεσμα να οδηγεί στην διάταξή τους πάνω στον άξονα, οπότε ανάλογα με την τιμή των συντεταγμένων τους να προσδιορίζουν μία «κατεύθυνση», η οποία να υποδηλώνει τις περισσότερες φορές την **υποβόσκουσα τάση** των δεδομένων του πίνακα.

Ανάλυση ενός πίνακα δεδομένων

Όπως είναι γνωστό σε κάθε έρευνα αγοράς οι περισσότερες ερωτήσεις αντιστοιχούν σε ποιοτικές μεταβλητές και λιγότερο σε ποσοτικές μεταβλητές (ασυνεχείς ή συνεχείς) οι οποίες τελικώς μετασχηματίζονται σε ποιοτικές δημιουργώντας ένα συγκεκριμένο πλήθος διαβαθμίσεων.

Ο ερευνητής για να αναλύσει τις απαντήσεις των ερωτώμενων, έχει στη διάθεσή του δύο διαδικασίες στατιστικής ανάλυσης.

Α) Η πρώτη αναφέρεται σε μεθόδους της κλασικής στατιστικής ανάλυσης.

Χρησιμοποιώντας, λοιπόν, τις μεθόδους της κλασικής στατιστικής ανάλυσης έχει την δυνατότητα δημιουργίας της κατανομής συχνοτήτων των μεταβλητών, τον υπολογισμό των παραμέτρων αυτών των κατανομών, όπως μέσο όρο, διακύμανση και τυπική απόκλιση, εύρεση ποσοστών, τη συσχέτιση μεταξύ δύο μεταβλητών, τη διαπίστωση αν υφίσταται ανεξαρτησία ή εξάρτηση μεταξύ δύο μεταβλητών από το πλήθος των k μεταβλητών, είτε να εφαρμόσει κάποιους ελέγχους παραμετρικούς ή μη παραμετρικούς χρησιμοποιώντας μία ή δύο το πολύ μεταβλητές.

Π.χ Έστω ότι έχουμε ένα ερωτηματολόγιο με 10 ερωτήματα το οποίο απάντησαν 557 ερωτώμενοι και θέλουμε να αναλύσουμε δύο από αυτά τα ερωτήματα. Το ένα ερώτημα αντιστοιχεί στην Ηλικία του ερωτώμενου και το άλλο στο ερώτημα για το «Πώς βλέπει ο ερωτώμενος την καθαριότητα της πόλης της Θεσσαλονίκης;»

Ο ερευνητής καθόρισε για τις ηλικίες τις ακόλουθες τέσσερις διαβαθμίσεις με τους αντίστοιχους κωδικούς **Ηλικία κάτω των 18 ετών** (κωδ:Η1=1),**Ηλικία 19-35 ετών** (κωδ:Η2=2), **Ηλικία 36-45 ετών** (κωδ:Η3 =3) και **Ηλικία πάνω από 46 ετών** (κωδ :Η4=4), ενώ για την στάση των ερωτηθέντων για την καθαριότητα πρότεινε τρεις διαβαθμίσεις με τους αντίστοιχους κωδικούς. Την διαβάθμιση **Αρνητική στάση** (κωδ: Α=1), την διαβάθμιση **Ουδέτερη στάση** (κωδ:Ο=2) και την διαβάθμιση

Θετική στάση (κωδ: Θ=3). Τα δεδομένα όπως προέκυψαν παρουσιάζονται στον παρακάτω πίνακα ως εξής:

Πίνακας 4:Τμήμα του πίνακα δεδομένων

IND	Καθαριότητα	Ηλικία
I1	3	2
I2	2	4
I3	1	1
I4	2	2
I5	2	3
I6	1	2
I7	1	1
I8	2	1
I9	2	2
I10	3	1

Επειδή τα δεδομένα είναι κωδικοποιημένα, δεν έχει κανένα νόημα να εξαχθούν οι γνωστές παράμετροι της κλασικής στατιστικής, όπως λ.χ μέσος όρος, διακύμανση κ.λ.π.

Βέβαια ο ερευνητής μπορεί αρχικά να υπολογίσει, αν υφίσταται συσχέτιση μεταξύ των μεταβλητών **Ηλικία** ερωτώμενων και **Στάση** ως προς την καθαριότητα υπολογίζοντας τον συντελεστή συσχέτισης ρ, η τιμή του οποίου βρέθηκε ίση με ρ= -0,02625.

Με δεδομένο τον σχετικά μικρό αριθμό δείγματος (n=557) ο έλεγχος συσχέτισης H_0:ρ=0 γίνεται δεκτός σε επίπεδο σημαντικότητας α=5%. Επομένως από την συγκεκριμένη τιμή του δείγματος συμπεραίνεται ότι οι δύο μεταβλητές είναι **ασυσχέτιστες**.

Ακολούθως ο ερευνητής δημιουργεί ένα απλό πίνακα διπλής εισόδου, όπου διασταυρώνονται οι τρεις διαβαθμίσεις της Στάσης (Α,Ο,Θ) ως προς την καθαριότητα με τις τέσσερις διαβαθμίσεις της Ηλικίας (Η1,Η2,Η3,Η4), ώστε να διαπιστωθεί αν οι δύο μεταβλητές είναι ανεξάρτητες.

Αυτό γίνεται επειδή δύο μεταβλητές εφόσον είναι ασυσχέτιστες, δεν σημαίνει ότι είναι και ανεξάρτητες (Καραπιστόλης Δ, 2011 Στατιστική Επιχειρήσεων).Δημιουργεί λοιπόν τον παρακάτω πίνακα στον οποίο εφαρμόζεται ο έλεγχος ανεξαρτησίας.

Πίνακας 5:Πίνακας διπλής εισόδου

Διαβαθμίσεις	Η1	Η2	Η3	Η4	Άθροισμα γραμμής
Α	9	55	27	17	108
Ο	3	83	29	58	173
Θ	11	136	69	60	276
Άθροισμα στήλης	23	274	125	135	577

Ο έλεγχος ανεξαρτησίας σε επίπεδο σημαντικότητας 5% δίνει τα ακόλουθα αποτελέσματα.

Για ν=6 βαθμούς ελευθερίας => X^2=20,85

Σε επίπεδο σημαντικότητας 5% έχουμε $X^2_{6,0.05} = 12,59$

Επειδή $X^2_{6,0.05} < X^2$ συμπεραίνεται ότι η **Στάση ως προς την καθαριότητα** με την **Ηλικία** είναι **εξαρτημένες**. Αυτό σημαίνει ότι οι διαφορετικές ηλικίες έχουν συγκεκριμένη στάση ως προς την καθαριότητα. Βέβαια αυτό **το συμπέρασμα δεν μας πληροφορεί ποια ηλικία συνδέεται με ποια στάση, βασικό ερώτημα προς διερεύνηση.**

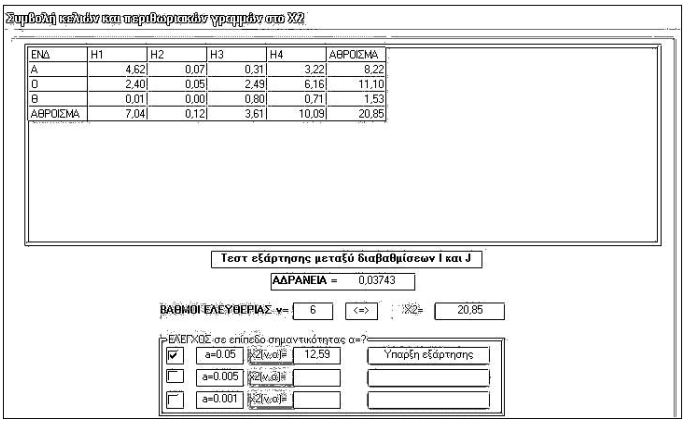
Β) Η δεύτερη διαδικασία στατιστικής ανάλυσης είναι αυτή που αναφέρεται σε μεθόδους της πολυπαραγοντικής στατιστικής ανάλυσης και ιδιαίτερα στις μεθόδους της Παραγοντικής Ανάλυσης των Αντιστοιχιών (Analyse Factorielle des Corespondances -AFC-) και της Ανιούσας Ιεραρχικής Ταξινόμησης (Classification Ascendante Hierarchique-CAH-), οι οποίες όχι μόνο μας προσδιορίσουν αν συσχετίζονται ή όχι οι δύο ερωτήσεις, αλλά και ποια ηλικία έχει συγκεκριμένη στάση, Οι μέθοδοι αυτές οπωσδήποτε προσφέρουν πληρέστερη πληροφόρηση από τις κλασικές μεθόδους και είναι ιδιαίτερα χρήσιμες όταν μελετά κάποιος ποιοτικές μεταβλητές.

Εφαρμογή της Παραγοντικής Ανάλυσης των Αντιστοιχιών

Εφόσον αναλύσουμε ένα πίνακα διπλής εισόδου με την μέθοδο της Παραγοντικής Ανάλυσης των Αντιστοιχιών, έχουμε στη διάθεσή μας, μετά την δημιουργία του παραγοντικού επιπέδου, τις ακόλουθες πληροφορίες.

1. το ποσοστό της αντλούμενης πληροφορίας που παρέχει το παραγοντικό επίπεδο
2. τις αλληλεπιδράσεις μεταξύ μεταβλητών που παρίστανται από τις στήλες
3. τις αλληλεπιδράσεις μεταξύ ιδιοτήτων που παρίστανται από τις γραμμές
4. ποια μεταβλητή στήλη συνδέεται με ποια ιδιότητα γραμμή

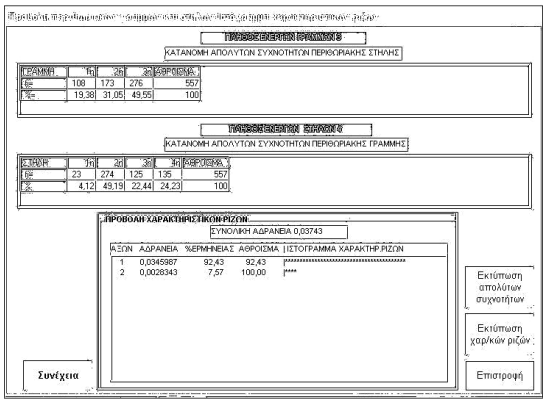
Η ανάλυση του πίνακα 5 βάσει του λογισμικού MAD δίνει τις παρακάτω πληροφορίες:



εικόνα 1

Από την εικόνα 1 προκύπτει ότι οι δύο μεταβλητές είναι **εξαρτημένες** σε επίπεδο σημαντικότητας 0,05.

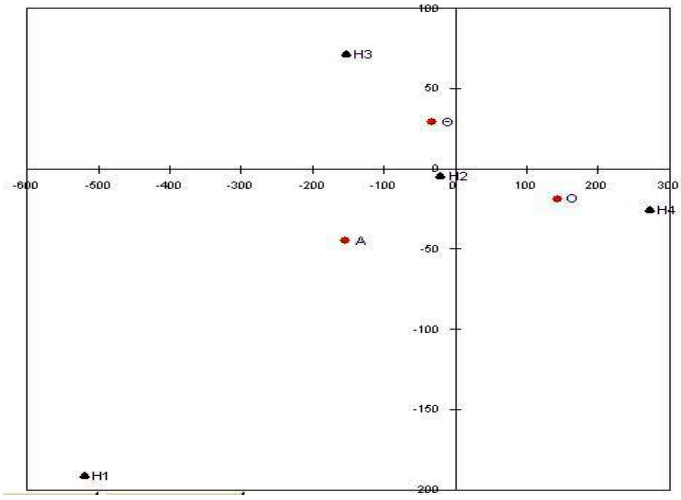
Στη συνέχεια ακολουθεί η εικόνα 2. Με δεδομένο ότι η τάξη του πίνακα 5 είναι ίση με 3, η ανάλυση με την AFC παρέχει δύο παραγοντικούς άξονες, καθόσον η τάξη ενός πίνακα διπλής εισόδου όταν είναι k, η παραγοντική ανάλυση προσδιορίζει k-1 παραγοντικούς άξονες.



εικόνα 2

Η εικόνα 2 πληροφορεί ότι ο 1^{ος} παραγοντικός άξονας ερμηνεύει το 92,43% της συνολικής πληροφορίας που αντλείται από τα δεδομένα του πίνακα 5, ενώ ο 2^{ος} παραγοντικός άξονας το υπόλοιπο 7,57%.

Ακολουθεί η εικόνα 3



εικόνα 3

Η εικόνα 3 παρουσιάζει το παραγοντικό επίπεδο 1x2, από το οποίο μπορεί να διαπιστωθεί η αλληλεπίδραση μεταξύ μεταβλητών και ιδιοτήτων.

Για την ερμηνεία των πληροφοριών που μπορεί να εξαχθούν από το παραγοντικό επίπεδο, γίνεται με τη βοήθεια των ενδείξεων του πίνακα 6, ο οποίος παρουσιάζει τις τιμές d του **δείκτη έλξης- άπωσης** μεταξύ «γραμμών» και στηλών του πίνακα.

Πίνακας 6 : Τιμές του δείκτη έλξης- άπωσης

ΕΝΔ	**Η1**	**Η2**	**Η3**	**Η4**
Α	2,0181	1,0352	1,1140	0,6494
Ο	0,4199	0,9752	0,7469	1,3832
Θ	0,9651	1,0016	1,1140	0,8969

Όταν η τιμή του δείκτη d έλξης-άπωσης είναι μεγαλύτερη του 1, σημαίνει ότι υφίσταται έλξη μεταξύ των διαβαθμίσεων των γραμμών και των στηλών, ενώ όταν η τιμή είναι μικρότερη του 1, υποδεικνύει άπωση. Στην περίπτωση που η τιμή είναι ίση με 1 δεν υφίσταται καμία εξάρτηση.

Γενικές αρχές εφαρμογής της Π.Α.Α

Οι γενικές αρχές που διέπουν την Παραγοντική Ανάλυση των Αντιστοιχιών (Π.Α.Α) είναι οι εξής:

- Οι τιμές των φατνίων ενός πίνακα πρέπει να είναι μη αρνητικοί αριθμοί

- Οι περιθωριακές κατανομές ενός πίνακα δεδομένων πρέπει να έχουν φυσική σημασία. Με άλλα λόγια τα φατνία του πίνακα δεδομένων πρέπει να είναι συχνότητες. Αν τα δεδομένα προέρχονται από ετερογενείς μεταβλητές λ.χ τιμές ποσοστών, τιμές μιας ποσοτικής μεταβλητής και διαβαθμίσεις ενός ποιοτικού χαρακτηριστικού, τότε για να αναλυθεί ένας παρόμοιος πίνακας δεδομένων με την Π.Α.Α πρέπει προηγουμένως τα δεδομένα να ομογενοποιηθούν ώστε να δημιουργηθεί ο αντίστοιχος λογικός πίνακας 0-1.

- Ο 1ος παράγοντας είναι μία κατ' αρχή προσέγγιση των δεδομένων, η οποία όμως χρειάζεται διόρθωση. Η διόρθωση επιτυγχάνεται με τον 2ο παράγοντα. Εάν δεν ικανοποιούμαστε από την προσέγγιση που προσφέρουν οι δύο πρώτοι παράγοντες συνεχίζουμε με τους υπόλοιπους που διαθέτουμε.

- Στον χώρο R^p των σημείων-αντικειμένων $\{I_k=(k=1,2,\ldots n)\}$ όταν δύο **σημεία-αντικείμενα** είναι γειτονικά, σημαίνει ότι τα δύο αυτά αντικείμενα χαρακτηρίζονται από τις ίδιες περίπου τιμές για κάθε μία από τις p μεταβλητές, οι δε αποστάσεις μεταξύ των σημείων υπολογίζονται με βάση την Ευκλείδεια απόσταση.

- Στον χώρο R^n των σημείων-μεταβλητών $\{J_m=(m=1,2,\ldots p)\}$ όταν δύο **σημεία-μεταβλητές** είναι κοντά το ένα από το άλλο, αυτό σημαίνει πως το σύνολο των αντικειμένων έδωσε στις μεταβλητές αυτές περίπου τις ίδιες τιμές.

- Όταν ένα σημείο-αντικείμενο και ένα σημείο-μεταβλητή είναι γειτονικά σημαίνει ότι το αντικείμενο χαρακτηρίζεται από την ιδιότητα της μεταβλητής.

- Όταν δύο μεταβλητές βρίσκονται σε αντίθετα τεταρτημόρια του παραγοντικού επιπέδου, σημαίνει ότι έχουν αντίθετες ιδιότητες.

Συνεπώς από την απεικόνιση των σημείων πάνω στο παραγοντικό επίπεδο και την επικύρωση με βάση τις τιμές του δείκτη d προκύπτουν οι εξής παρατηρήσεις: οι ηλικίες **κάτω των 18** ετών (Η1) και οι ηλικίες **19 έως 35** ετών (Η2) συνδέονται με την **Αρνητική στάση** (Α), οι ηλικίες **35 έως 45** ετών **(**Η3) συνδέονται με την **Θετική στάση** (Θ)**,** ενώ τέλος **οι ηλικίες άνω των 45** ετών συνδέονται με την **Ουδέτερη στάση** (Ο).

Η ανάγλυφη αυτή διαπίστωση προκύπτει από την εξής λογική που διέπει την Παραγοντική Ανάλυση των Αντιστοιχιών.

Ως γνωστόν όταν από μια σειρά αριθμών υπολογιστεί ο μέσος όρος, οι αποκλίσεις κάθε τιμής από αυτή την παράμετρο θέσης, προσδιορίζουν κάτι συγκεκριμένο. Έτσι όσο η θετική απόκλιση είναι μεγαλύτερη τόσο πιο χαρακτηριστική είναι η ιδιότητα που μετρά η μεταβλητή που αντιστοιχεί στον μέσο όρο. Ανάλογη ερμηνεία δίνεται για την αρνητική απόκλιση. Δηλαδή όσο μεγαλύτερη είναι η απόκλιση τόσο λιγότερο χαρακτηρίζει η μεταβλητή την ιδιότητα που μετρά.

Συνεπώς αν από ένα δείγμα 100 ατόμων προέκυψε ότι το μέσο ετήσιο εισόδημα είναι 12.000 ευρώ, όταν κάποιος από αυτούς έχει εισόδημα 35.000 ευρώ θεωρείται εύπορος, ενώ αν παρουσιάζει εισόδημα 55.000 ευρώ θεωρείται πλούσιος. Στην αντίθετη περίπτωση ένας άλλος που δηλώνει 5.000 ευρώ θεωρείται φτωχός.

Ανάλογη είναι η διαδικασία ερμηνείας στην Παραγοντική Ανάλυση των Αντιστοιχιών.

Αντί του μέσου όρου, επειδή διασταυρώνονται δύο τουλάχιστον μεταβλητές, υπολογίζεται ο πίνακας ανεξαρτησίας, δηλαδή το πως θα παρουσιαζόταν ο πίνακας διπλής εισόδου, αν η αλληλεπίδραση των δύο μεταβλητών ήταν ιδανική. Η ιδανική αυτή σχέση υπολογίζεται βάσει της σχέσης (1)

$$k_{ij} = \frac{n(A_i) \cdot n(B_j)}{k} \quad (1)$$

Όπου

k_{ij} = το κελί του ιδανικού πίνακα Θ(ixj) που αντιστοιχεί στην i γραμμή και j στήλη
k = το γενικό σύνολο
$n(A_i)$ = το άθροισμα της γραμμής i
$n(B_j)$ = το άθροισμα της στήλης j

Στη συνέχεια αφαιρούμε από την παρατηρούμενη απόλυτη συχνότητα που υπάρχει σε κάθε κελί $A_i \cap B_j$ του πίνακα δεδομένων την ιδανική συχνότητα k_{ij}, οπότε προκύπτει η ζητούμενη απόκλιση από την ιδανική κατάσταση, η οποία θα χαρακτηρίζει την αλληλεπίδραση των μεταβλητών A_i και B_j.

Επομένως με βάση το παραπάνω σκεπτικό από τα στοιχεία του πίνακα 2, για κάθε ηλικία και για κάθε στάση Α= Αρνητική ,Ο= Ουδέτερη, Θ= Θετική έχουμε

1. για την Ηλικία κάτω των 18 ετών (Η1)

$$A \longrightarrow 9 - \frac{23 \cdot 108}{557} = 9 - 4,46 = +4,54$$
$$O \longrightarrow 3 - \frac{23 \cdot 173}{557} = 3 - 7,14 = -4,14$$
$$\Theta \longrightarrow 11 - \frac{23 \cdot 276}{557} = 11 - 11,39 = -0,39$$

Από τις τρεις αυτές σχέσεις προκύπτει ότι θετική απόκλιση παρατηρείται μόνο στην Αρνητική στάση, ενώ στις υπόλοιπες έχουμε αρνητικές αποκλίσεις. Επομένως οι Ηλικίες κάτω των 18 ετών συνδέονται αρνητικά με την συγκεκριμένη διαβάθμιση, γεγονός που υπέδειξε το παραγοντικό επίπεδο 1x2.

Συνεχίζοντας τους ίδιους υπολογισμούς για τις υπόλοιπες τρεις κατηγορίες ηλικιών έχουμε:

2. για την Ηλικία από 19 έως 35 ετών (Η2)

$$A \longrightarrow 55 - \frac{274 \cdot 108}{557} = 55 - 53,13 = +1,87$$
$$O \longrightarrow 83 - \frac{274 \cdot 173}{557} = 83 - 85,1 = -2,1$$
$$\Theta \longrightarrow 136 - \frac{274 \cdot 276}{557} = 136 - 135,7 = +0,3$$

Από τους υπολογισμούς αυτούς προκύπτουν δύο θετικές αποκλίσεις, με μεγαλύτερη εκείνη που αφορά στην Αρνητική στάση. Επομένως οι Ηλικίες από 19-35 ετών συνδέονται με την συγκεκριμένη διαβάθμιση, διαπίστωση που υπέδειξε και το παραγοντικό επίπεδο 1x2.

3. για την Ηλικία από 36 έως 45 ετών (Η3)

$$A \longrightarrow 29 - \frac{125 \cdot 108}{557} = 29 - 24,24 = +4,76$$
$$O \longrightarrow 29 - \frac{125 \cdot 173}{557} = 29 - 38,82 = -9,82$$
$$\Theta \longrightarrow 69 - \frac{125 \cdot 276}{557} = 69 - 61,94 = +7,06$$

Από αυτές τις τρεις σχέσεις προκύπτουν δύο θετικές αποκλίσεις, με μεγαλύτερη εκείνη που αφορά στην Θετική στάση. Επομένως οι Ηλικίες από 36 έως 45 ετών συνδέονται περισσότερο με την συγκεκριμένη διαβάθμιση, γεγονός που υπέδειξε το παραγοντικό επίπεδο 1x2.

4. για τις Ηλικίες Πάνω από 45 ετών (Η4)

$$A \longrightarrow 17 - \frac{135 \cdot 108}{557} = 17 - 26,18 = -9,18$$

$$O \longrightarrow 58 - \frac{135 \cdot 173}{557} = 58 - 41,93 = +16,07$$

$$\Theta \longrightarrow 60 - \frac{135 \cdot 276}{557} = 60 - 66,89 = -6,89$$

Στην περίπτωση των ηλικιών Άνω των 45 ετών παρατηρείται θετική απόκλιση μόνο στην Ουδέτερη στάση, ενώ στις υπόλοιπες έχουμε αρνητικές αποκλίσεις. Επομένως οι Ηλικίες πάνω από 45 ετών συνδέονται με την συγκεκριμένη διαβάθμιση, στοιχείο που υπέδειξε το παραγοντικό επίπεδο 1x2.

Τελικά γίνεται φανερό το πόσο γρήγορα ο αναλυτής παρατηρώντας ένα παραγοντικό επίπεδο, μπορεί να κάνει τις διαπιστώσεις του, χωρίς να υποχρεωθεί να εκτελέσει όλες τις παραπάνω επίπονες πράξεις καταλήγοντας στα ίδια συμπεράσματα.

Μέχρι τώρα έχουμε εντοπίσει την εξάρτηση κάθε μεταβλητής-στήλη με την ιδιότητα που παριστάνει κάθε γραμμή. Οφείλουμε όμως να υπολογίσουμε την **ένταση της εξάρτησης** αυτής, η οποία μπορεί να εντοπιστεί με την διαδικασία ενός απλού ελέγχου της διαφοράς αναλογιών σε επίπεδο σημαντικότητας 5%, όπως θα παρουσιαστεί στη συνέχεια.

Πριν αναφερθούμε στον υπολογισμό της έντασης της εξάρτησης, σκόπιμο είναι να αναφέρουμε εισαγωγικά το τι θεωρούμε «δημιουργία» μιας Ανιούσας Ιεραρχικής Ταξινόμησης ενός πλήθους n στατιστικών μονάδων.

Μία ανιούσα ιεραρχική ταξινόμηση των στοιχείων ενός συνόλου I με πληθάριθμο card(I)=n, είναι μία διαδικασία που παράγει μια ακολουθία διαμελισμών του αρχικού συνόλου σε υποσύνολα μη κενά και ξένα ανά δύο μεταξύ τους, τις λεγόμενες **κλάσεις**, τη μία μέσα στην άλλη, συνενώνοντας κάθε φορά δύο μόνο κλάσεις οι οποίες βάσει κάποιας μετρικής παρουσιάζουν σε κάθε βήμα ομαδοποίησης την μικρότερη απόσταση. Όσο απομακρύνεται κανείς από τον αρχικό διαμελισμό (ο οποίος περιλαμβάνει τόσες κλάσεις όσα είναι τα αντικείμενα που ταξινομούνται), τόσο αυτός γίνεται λιγότερο λεπτομερής.

Η τελευταία κλάση περιλαμβάνει το σύνολο των κλάσεων που δημιουργήθηκαν από τις συνενώσεις στοιχείων και κλάσεων, το πλήθος των οποίων είναι 2n-1.

Η βασική, λοιπόν, θεώρηση της ανιούσας ιεραρχικής ταξινόμησης ξεκινά από το ότι κάθε μια από τις n στατιστικές μονάδες προς ταξινόμηση, αποτελεί μια διακεκριμένη κλάση με τα δικά της χαρακτηριστικά και καταλήγει σε μια μόνο η οποία συμπεριλαμβάνει το σύνολο των στατιστικών μονάδων.

Απ’ ότι γίνεται αντιληπτό στόχος της ανιούσας ιεραρχικής ταξινόμησης είναι να ομαδοποιήσει τις στατιστικές μονάδες ενός πληθυσμού σ' ένα περιορισμένο πλήθος ομοιογενών κλάσεων, ως προς την συμπεριφορά ορισμένων μεταβλητών, λαμβάνοντας υπόψη το σύνολο των μεταβλητών, ώστε κάθε μία να διαφέρει από τις άλλες, όσο το δυνατόν περισσότερο.

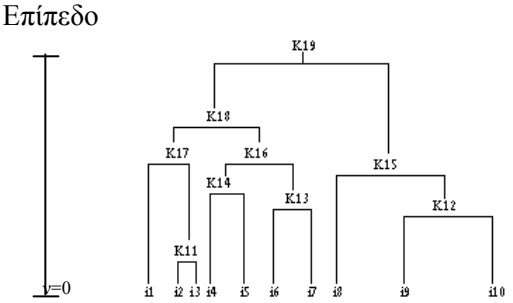
Βέβαια όπως όλες οι μέθοδοι της ανάλυσης δεδομένων έτσι και η CAH απεικονίζει μ' ένα απλό σχήμα το αποτέλεσμα της ταξινόμησης του πίνακα δεδομένων T(nxp), το καλούμενο **δενδρόγραμμα**, του οποίου οι γραμμές αποτελούν τις n παρατηρήσεις που περιγράφονται από το σύνολο των p μεταβλητών που αντιστοιχούν στις στήλες του.

Οι κλάσεις δημιουργούνται βάσει ενός αντικειμενικού αλγορίθμου, πέρα από τις υποκειμενικές μεθόδους που μπορεί να αναπτύξει κάθε ερευνητής. Λέμε αντικειμενικό αλγόριθμο γιατί η ομαδοποίηση των στατιστικών μονάδων γίνεται χωρίς καμιά a priori υπόθεση στον αρχικό πίνακα δεδομένων και βάσει μιας συγκεκριμένης μετρικής.

Ο τρόπος με τον οποίο πραγματοποιείται ο διαμελισμός των γραμμών του πίνακα δεδομένων είναι που κάνει την ουσιαστική διαφορά μεταξύ της **ταξινόμησης** και της **κατάταξης**.

Με την κατάταξη τοποθετούνται οι γραμμές του πίνακα (διαθέσιμες παρατηρήσεις) σε **προκαθορισμένες ομάδες** (διακριτική ανάλυση),ενώ με την ταξινόμηση αναζητούνται και προσδιορίζονται με την βοήθεια ενός αλγόριθμου οι ομάδες διαμελισμού των παρατηρήσεων, λαμβάνοντας υπόψη το **σύνολο των μεταβλητών που τις χαρακτηρίζουν**. Χρησιμοποιώντας την **διαδικασία VACOR** βρίσκουμε όχι μόνο ποιες μεταβλητές συγκεκριμένα χαρακτηρίζουν τις κλάσεις, αλλά και την ένταση της εξάρτησης, καθώς επίσης ποιες από τις στατιστικές μονάδες ταξινομούνται σε κάθε κλάση.

Η περιγραφή της ταξινόμησης γίνεται, όπως προαναφέραμε, με το δενδρόγραμμα (σχήμα 1) του οποίου οι κόμβοι συμβολίζουν τις υποδιαιρέσεις του πληθυσμού και το επίπεδο ν του κάθε κόμβου δείχνει τον βαθμό ομοιότητας των παρατηρήσεων [Benzecri J.P et al.1980].



σχήμα 2: Δενδρόγραμμα ταξινόμησης 10 στατιστικών μονάδων κατ' αύξουσα ιεραρχία

Στο σχήμα 2 τα i1,i2,…,i10 αντιπροσωπεύουν 10 στατιστικές μονάδες οι οποίες ομαδοποιούνται σε εννέα κλάσεις με πληθάριθμο μεγαλύτερο ή ίσο του 2.

Συγκεκριμένα έχουμε:

την κλάση K11 ={i2,i3} την κλάση K12={i9,i10} την κλάση K13={i6,i7}

την κλάση K14={i4,i5} την κλάση K15={K12,i8} την κλάση K16={K13,K14}

την κλάση K17={i1,K11} την κλάση K18={K16,K17} και την κλάση K19={K15,K18}

Το σύνολο των εννέα κλάσεων (Κ11 έως Κ19} ονομάζεται **τυπολογία** της ιεραρχίας

Επειδή δεν είναι σκόπιμο να αναφερθεί η ταξινόμηση των 557 ερωτηθέντων στα 10 ερωτήματα που τους ετέθησαν, εφαρμόζουμε την διαδικασία VACOR αποκλειστικά στα δεδομένα του πίνακα 2, για να βρούμε την **ένταση** που χαρακτηρίζει (θετικά ή αρνητικά) τις διαβαθμίσεις των ηλικιών με τις στάσεις των ερωτώμενων (βλέπε πίνακα 4).

Η διαδικασία αυτή επιβάλλεται να γίνεται μετά από κάθε ανάλυση με την AFC, ώστε να διευκολύνει την ερμηνεία των αλληλεξαρτήσεων μεταξύ μεταβλητών (στηλών) και στατιστικών μονάδων (γραμμών), ώστε να αποφευχθεί οποιαδήποτε σύγχυση που μπορεί να προκαλέσουν οι προβολές των σημείων που αντιπροσωπεύουν γραμμές και στήλες, πάνω στο παραγοντικό επίπεδο όταν αυτό δεν συγκεντρώνει το σύνολο της πληροφορίας που παρέχει ο πίνακας δεδομένων.

Πίνακας 4:

IND	H1	H2	H3	H4
A	4,9828	0,8157	1,4472	-4,6798
O	-2,846	-0,577	-3,2142	5,1126
Θ	-0,1771	0,0367	1,4472	-1,3808

Στον πίνακα 4 οι τιμές υπολογίζονται με βάση τον στατιστικό έλεγχο σε επίπεδο σημαντικότητας 5% της διαφοράς των αναλογιών μεταξύ της αναλογίας P_δ που προκύπτει από το «δείγμα» που αντιπροσωπεύει κάθε κελί του πίνακα δεδομένων και της αναλογίας P_M κάθε μεταβλητής ως προς το σύνολο k των απαντήσεων.

Έτσι η τιμή του κελιού (A,H1)=4,9822 υπολογίζεται ως εξής:

Θεωρούμε τον ακόλουθο στατιστικό έλεγχο σε επίπεδο σημαντικότητας 5%

$$H_0 : P_\delta - P_M = 0$$
$$H_1 : P_\delta - P_M > 0$$

Στη συνέχεια υπολογίζεται η τιμή του z

$$z = \frac{P_\delta - P_M}{s_p} \qquad \textbf{(2)}$$

Όπου $s_p = \sqrt{\frac{P \cdot Q}{n}}$ $\quad \mu\varepsilon \quad$ $Q = 1 - P$

Όταν από την σχέση (2) προκύπτει z>1,645 τότε ισχύει η εναλλακτική υπόθεση H_1, δηλαδή η μεταβλητή M με αναλογία P_M χαρακτηρίζει έντονα την διαβάθμιση με αναλογία P_δ.

Σε διαφορετικές τιμές του z όπου -1,645<z<1,645 απλά η μεταβλητή παρουσιάζει μέτρια ένταση (θετική ή αρνητική) ανάλογα με την τιμή που παρουσιάζει), ενώ όταν

προκύπτει τιμή z<-1,645 τότε η απουσία εξάρτησης της μεταβλητής Μ με την διαβάθμιση αυτή θεωρείται έντονη.

Με τα δεδομένα του πίνακα 4 για το κελί (Α,Η1) έχουμε την τιμή 1,0125. Το πώς προέκυψε η τιμή αυτή ακολουθούμε τους παρακάτω υπολογισμούς.

Η αναλογία του κελιού (Α,Η1) είναι $\frac{9}{108} = 0,0833$

Η αναλογία της μεταβλητής Η1 ως προς το σύνολο των απαντήσεων είναι $\frac{23}{557} = 0,0413$

Στη συνέχεια υπολογίζουμε την τιμή $z = \frac{0,0833 - 0,0413}{0,00843} = 4,98221$

όπου $s_p = \sqrt{\frac{0,0413 \cdot (1 - 0,0413)}{557}} = \quad 0,00843$

Εφόσον η τιμή του z είναι θετική αλλά μεγαλύτερη της τιμής 1,645 η επίδραση της διαβάθμισης Ηλικία κάτω των 18 ετών (Η1) δείχνει ότι συνδέεται θετικά με την αρνητική στάση για την καθαριότητα (Α) με σημαντική ένταση.

Από τον πίνακα 4 όσον αφορά στην αρνητική στάση των ηλικιών ως προς την καθαριότητα προκύπτει η **κλιμακούμενη** Αρνητική στάση των ηλικιών κατά φθίνουσα τιμή του z. Ήτοι οι ηλικίες κάτω των 18 ετών (Η1), οι ηλικίες 36 έως 45 ετών (Η3), οι ηλικίες από 19 έως 35 ετών (Η2) οι οποίες παρουσιάζουν μέτρια αρνητική ένταση, ενώ οι ηλικίες άνω των 45 ετών (Η4) εμφανίζουν αντίθετη στάση ως προς τις υπόλοιπες ηλικίες, δηλαδή δεν συμφωνούν (και μάλιστα έντονα), ότι η καθαριότητα της πόλης της Θεσσαλονίκης είναι βρώμικη.

Εν κατακλείδι πρέπει να τονιστεί ιδιαίτερα πως σ' ένα τόσο απλό πίνακα δεδομένων διπλής εισόδου, οι εξαρτήσεις μεταξύ μεταβλητών (στηλών) και ιδιοτήτων (γραμμών) μπορεί να είναι εύκολες να διαπιστωθούν με απλές πράξεις. Αλλά όμως σ' ένα πίνακα λ.χ διαστάσεων 50x20 οι προηγούμενες διαπιστώσεις με απλές πράξεις όπως παρουσιάστηκαν στο παράδειγμα, είναι πολύ δύσκολες και πολύπλοκες.

Αντιθέτως η ευκολία ανάγνωσης της συμπεριφοράς των μεταβλητών, όπου το πλήθος τους μπορεί να είναι οσοδήποτε μεγάλο, καθιστά τις συγκεκριμένες μεθόδους πολυδιάστατης στατιστικής ανάλυσης δημοφιλείς στους ερευνητές οποιουσδήποτε επιστημονικού πεδίου, αναγνωρίζοντας την μεγάλη συμβολή τους στην εξαγωγή πολυσύνθετων συμπερασμάτων που δεν προκύπτουν με την εφαρμογή συγκεκριμένων μεθόδων της κλασικής στατιστικής ανάλυσης.

Σε κάθε ταξινόμηση ανεξάρτητα με ποια μετρική πραγματοποιείται, δημιουργούνται κλάσεις που κάθε μια περιέχει ένα συγκεκριμένο πλήθος «αντικειμένων», το οποίο παριστάνουν οι γραμμές του πίνακα δεδομένων. Έτσι σ' ένα ερωτηματολόγιο έρευνας αγοράς σε κάθε γραμμή αντιστοιχούν οι απαντήσεις του ερωτηθέντα στο σύνολο των ερωτημάτων που του ετέθησαν.

Πρέπει να τονιστεί ιδιαίτερα, ότι σ' ένα ερωτηματολόγιο έρευνας αγοράς, οι περισσότερες ερωτήσεις είναι ποιοτικού χαρακτήρα, επομένως η ταξινόμηση πρέπει να γίνεται με τον αλγόριθμο του Ward, καθ' όσον είναι η μόνη μέθοδος που χρησιμοποιεί για την συνένωση δύο κλάσεων, την απώλεια αδράνειας, αφού προηγουμένως υπολογίζονται οι αποστάσεις μεταξύ δύο «γραμμών-αντικειμένων», βάσει της μετρικής του x^2 (μετρική J.P. Benzecri)

Μελετώντας, λοιπόν, κάθε κλάση χωριστά, βρίσκουμε ποιοι ερωτηθέντες την απαρτίζουν και συγχρόνως ποιες μεταβλητές επηρέασαν αλλά και κατά πόσο, ώστε να δημιουργηθεί η συγκεκριμένη κλάση. Εάν βέβαια επιθυμούμε να εντοπίσουμε ποιοι ερωτηθέντες επηρεάζονται από μια συγκεκριμένη μεταβλητή, η ανιούσα ιεραρχική ταξινόμηση, εστιάζει στον εντοπισμό της μικρότερης απόστασης των κέντρων των κλάσεων των ερωτηθέντων από τις μεταβλητές πάνω στον κάθε παραγοντικό άξονα χωριστά, με συγκεκριμένο ποσοστό ερμηνείας, συνεπώς η απάντηση δεν είναι απόλυτα ακριβής.

Το ερώτημα όμως, του εντοπισμού κάθε ερωτώμενου με ποια μεταβλητή συνδέεται, απαντάται στην επόμενη εργασία με τίτλο «Μέθοδος KARAP ως εργαλείο εξόρυξης δεδομένων»

Η ΜΕΘΟΔΟΣ KARAP ΩΣ ΕΡΓΑΛΕΙΟ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Γενικά

Μία ανιούσα ιεραρχική ταξινόμηση των n «υποκειμένων» (γραμμές) ενός πίνακα δεδομένων T(n,p) είναι μία διαδικασία που παράγει μια ακολουθία διαμελισμών του αρχικού συνόλου σε υποσύνολα μη κενά και ξένα ανά δύο μεταξύ τους, τις λεγόμενες **κλάσεις**, τη μία μέσα στην άλλη, συνενώνοντας κάθε φορά δύο μόνο κλάσεις οι οποίες βάσει κάποιας μετρικής παρουσιάζουν σε κάθε βήμα ομαδοποίησης την μικρότερη απόσταση.

Απ' ότι γίνεται αντιληπτό στόχος της ανιούσας ιεραρχικής ταξινόμησης είναι να ομαδοποιήσει το σύνολο των στατιστικών μονάδων ενός πληθυσμού σ' ένα περιορισμένο πλήθος ομοιογενών κλάσεων, τις επονομαζόμενες **«συστάδες»** ως προς την συμπεριφορά ορισμένων μεταβλητών, λαμβάνοντας υπόψη το σύνολο των μεταβλητών, ώστε κάθε μία να διαφέρει από τις άλλες, όσο το δυνατόν περισσότερο.

Συγκεκριμένα στη ταξινόμηση με την διαδικασία VACOR οι κλάσεις δημιουργούνται βάσει ενός αντικειμενικού αλγορίθμου (αλγόριθμος του Ward), πέρα από τις υποκειμενικές μεθόδους που μπορεί να αναπτύξει κάθε ερευνητής. Λέμε αντικειμενικό αλγόριθμο γιατί η ομαδοποίηση των στατιστικών μονάδων γίνεται χωρίς καμιά a priori υπόθεση στον αρχικό πίνακα δεδομένων βάσει της μετρικής x^2.

Στις κλάσεις που δημιουργεί η μέθοδος VACOR εντοπίζεται το πλήθος των μεταβλητών που τις χαρακτηρίζουν. Αυτό γίνεται με την βοήθεια που παρέχει το πρόγραμμα MAD με τον πίνακα «Συμβολή των μεταβλητών στο χαρακτηρισμό των κλάσεων» σε συνδυασμό με τον πίνακα «Συμβολή των μεταβλητών στη διάσπαση των k υψηλότερων κόμβων». Κατά συνέπεια τα «υποκείμενα» που συμμετέχουν στη διαμόρφωση των κλάσεων συνδέονται επίσης με τις μεταβλητές που χαρακτηρίζουν κάθε κλάση.

Είναι γνωστό ότι χρησιμοποιώντας την μέθοδο VACOR με τα γνωστά

προγράμματα που υλοποιούν την Ανιούσα Ιεραρχική Ταξινόμηση, δεν είναι εφικτή η σύνδεση κάθε «υποκειμένου» με μια συγκεκριμένη μεταβλητή, εκτός και αν χρησιμοποιηθεί ο στατιστικός έλεγχος της διαφοράς των αναλογιών σε επίπεδο σημαντικότητας 5%, μεταξύ της αναλογίας P_δ κάθε μιας από τις p τιμές που αντιστοιχεί στο «δείγμα» που αναφέρεται σε κάθε «υποκείμενο» και της αναλογίας P_M κάθε μεταβλητής ως προς το σύνολο των p μεταβλητών (Morineau A. 1984).

$$\begin{aligned} H_0 &: P_\delta - P_M = 0 \\ H_1 &: P_\delta - P_M > 0 \end{aligned} \qquad (1)$$

Στη συνέχεια υπολογίζεται η τιμή του z

$$z = \frac{P_\delta - P_M}{s_p} \qquad (2)$$

Όπου $s_p = \sqrt{\frac{P_M \cdot Q_M}{n}}$ $\mu\varepsilon$ $Q_M = 1 - P_M$

Όπως αναφέρει ο καθηγητής Θ. Μπεχράκης «Με βάση την τιμή του ελέγχου κατατάσσουμε τα προβλήματα για κάθε ομάδα. Η τιμή του ελέγχου αποτελεί κριτήριο που χρησιμοποιείται για την επιλογή των προβλημάτων τα οποία χαρακτηρίζουν την κάθε ομάδα. Όσο μεγαλύτερη είναι η τιμή του ελέγχου για μια συγκεκριμένη ομάδα και ένα συγκεκριμένο πρόβλημα , τόσο περισσότερο χαρακτηριστικό είναι το πρόβλημα αυτό για τη συγκεκριμένη ομάδα» (Μπεχράκης Θ. σελ. 74)

Όταν, λοιπόν, από την σχέση (2) προκύπτει η τιμή $z>1,645$ τότε ισχύει η εναλλακτική υπόθεση H_1, δηλαδή η μεταβλητή M με αναλογία P_M χαρακτηρίζει έντονα την κλάση ή το υποκείμενο με αναλογία P_δ. Σε διαφορετικές τιμές του z όπου $-1,645<z<1,645$ απλά η μεταβλητή παρουσιάζει μέτρια σύνδεση (θετική ή αρνητική) ανάλογα με την τιμή που παρουσιάζει), ενώ όταν προκύπτει τιμή $z<-1,645$ τότε η απουσία εξάρτησης της μεταβλητής με την κλάση (ή το αντικείμενο) θεωρείται έντονη.

Από την άλλη η σύνδεση των υποκειμένων με κάθε μεταβλητή μπορεί να πραγματοποιηθεί εφ' όσον αναλυθεί ο πίνακας δεδομένων T(n,p) με την Παραγοντική Ανάλυση των Αντιστοιχιών (-Π.Α.Α-), εξαχθούν οι συντεταγμένες των p μεταβλητών και των n αντικειμένων πάνω στους p-1 παραγοντικούς άξονες και στη συνέχεια να υπολογιστούν οι αποστάσεις κάθε αντικειμένου A από κάθε μεταβλητής B χρησιμοποιώντας την παρακάτω σχέση η οποία υπολογίζει την απόσταση μεταξύ δύο διανυσμάτων του διανυσματικού χώρου R^n (Serge Lang σελ. 16)

$$\left\| A - B \right\| = \sqrt{(A - B) \cdot (A - B)} \qquad (3)$$

Αρχικά θα συγκριθούν τα αποτελέσματα της σύνδεσης των «υποκειμένων» με τις μεταβλητές που προέκυψαν αφενός μετά την ταξινόμηση με την μέθοδο VACOR, χρησιμοποιώντας τον προαναφερόμενο έλεγχο με την z κατανομή, αφετέρου με την τοποθέτηση όλων των σημείων (γραμμών και στηλών) του πίνακα δεδομένων T(n,p) στον Ευκλείδειο διανυσματικό χώρο $R^{(p-1)}$, βάσει των **παραγόντων** που προκύπτουν μετά την εφαρμογή της Παραγοντικής Ανάλυσης των Αντιστοιχιών. Η εργασία αυτή θα αποτελέσει επίσης μια νέα διαδικασία ταξινόμησης των «υποκειμένων».

Η επιλογή του Ευκλείδειου χώρου R^{p-1} έγινε αφού ως γνωστό οι παραγοντικοί άξονες

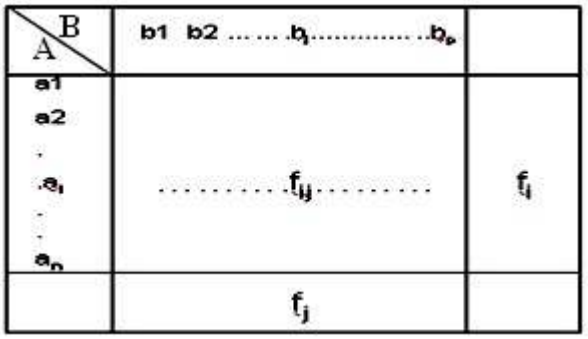
δημιουργούν μία ορθοκανονική βάση στον χώρο $R^{(p-1)}$, όπου βάσει των συντεταγμένων τοποθετούνται οι p μεταβλητές και οι n γραμμές του πίνακα δεδομένων στις πραγματικές τους θέσεις, απ' όπου αντλείται το σύνολο της πληροφόρησης που παρέχει ο πίνακας δεδομένων που αναλύεται.

Στο σημείο αυτό θα παραθέσουμε ένα παράδειγμα προς επιβεβαίωση των προαναφερόμενων προτάσεων.

Σχέσεις μεταξύ παραγόντων

Με βάση ένα πίνακα σύμπτωσης δύο ποιοτικών μεταβλητών Α και Β δημιουργείται ο αντίστοιχος πίνακας σχετικών συχνοτήτων.

Πίνακας 1:Πίνακας σχετικών συχνοτήτων



Η παραγοντική ανάλυση των αντιστοιχιών επιτρέπει όχι μόνο τη γεωμετρική και αλγεβρική διαπίστωση της απόκλισης από την κατάσταση της ανεξαρτησίας των δύο ποιοτικών μεταβλητών Χ και Υ, αλλά και τη διερεύνηση των ομοιοτήτων που παρουσιάζουν μεταξύ τους οι σχετικές κατανομές (προφίλ) των γραμμών ή των στηλών του πίνακα, που αντιστοιχούν στο σύνολο των διαβαθμίσεων των δύο μεταβλητών Χ και Υ.

Στη συνέχεια δημιουργείται ο πίνακας των προφίλ των γραμμών

Πίνακας 2: Πίνακας f_J^I των προφίλ των γραμμών

Διαβαθμίσεις	**b_1……..b_j……b_p**	
A_1	.	
.	.	
a_i	.....$f^i_j = f_{ij}/f_i$	1
.	.	
a_n	.	

Για να γίνει κατανοητό για πιο λόγο είναι προτιμότερο να χρησιμοποιείται το προφίλ μιας γραμμής ενός πίνακα, ως διανυσματική έκφραση της αντίστοιχης στατιστικής μονάδας i, παρά η γραμμή με τα αρχικά δεδομένα, η απάντηση έχει ως εξής: Εφόσον δύο γραμμές είναι ανάλογες μεταξύ τους, τα προφίλ τους θα είναι ταυτόσημα και όταν παρασταθούν σ' ένα γράφημα οι γραφικές παραστάσεις των διανυσμάτων στα οποία αντιστοιχούν θα συμπέσουν, ενώ αντιθέτως οι γραμμές με τα αρχικά δεδομένα θα παριστάνουν δύο συγγραμμικά διανύσματα.

Αυτή η διαπίστωση είναι πολύ σημαντική καθόσον στην Παραγοντική Ανάλυση των Αντιστοιχιών (-Π.Α.Α-) το ενδιαφέρον εστιάζεται στις αναλογίες των «υποκειμένων» μέσα στις διαβαθμίσεις των «μεταβλητών».

Οι προβολές των σημείων f_J^i του νέφους $N(I)_J$ των γραμμών του πίνακα δεδομένων πάνω στους παραγοντικούς άξονες Δ_a (a=1,...p-1),οι οποίες συμβολίζονται με $F_a(i)$ (όπου $i=f_j^i$ τυχόν προφίλ γραμμή), αποτελούν τις συντεταγμένες των σημείων αυτών πάνω στους άξονες Δ_a. Κάθε συντεταγμένη $F_a(i)$ σχετική με τον παραγοντικό άξονα a ονομάζεται **παράγοντας a του προφίλ i** (J-P & F. Benzecri 1980 σελ. 65).

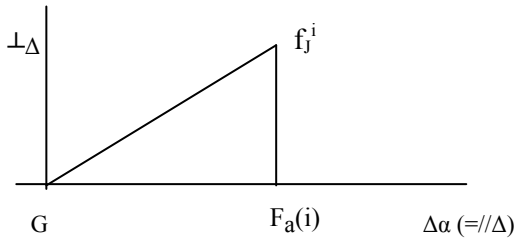
Για να προσδιορίσουμε τους παραγοντικούς άξονες Δ_a σ' ένα επίπεδο (a=1,2) χρησιμοποιούμε το θεώρημα του Huyghens, το οποίο αναφέρει ότι η ολική αδράνεια $I_{O\lambda}$ του νέφους N(I) μπορεί να αναλυθεί σε δύο μέρη. Το ένα μέρος αφορά στην αδράνεια $I_{//\Delta}$ κατά μήκος μιας ευθείας Δ_a η οποία περνά από το βαρύκεντρο $G\{=f_J\}$ του νέφους και το άλλο στην αδράνεια $I_{\perp\Delta}$ κάθετα της ευθείας Δ_α.

Από τις άπειρες ευθείες που διέρχονται από το σημείο G, λαμβάνεται εκείνη η οποία καθιστά την αδράνεια $I_{//\Delta}$ μέγιστη και συνεπώς ελάχιστη την αδράνεια $I_{\perp\Delta}$.

Ήτοι

$$I_{O\lambda} = I_{//\Delta} + I_{\perp\Delta} \qquad (4)$$

Σχηματικά έχουμε:



σχήμα 1: *Διάσπαση της ολικής αδράνειας*

Ως γνωστόν σε κάθε χαρακτηριστική ρίζα λ_a ενός τετραγωνικού πίνακα $S_J^J = f_J^I \, o \, f_I^J$ αντιστοιχεί ένα χαρακτηριστικό διάνυσμα $u_a=\varphi_a{}^J$ που συνδέεται με το παραγοντικό άξονα Δ_a.

Για κάθε διάνυσμα $\varphi_a{}^J$ (j=1,...p) οι συντεταγμένες του ικανοποιούν τις παρακάτω σχέσεις:

$$\sum_{j=1}^{p} f_j \cdot F_a^J = 0 \qquad (5)$$

$$\sum_{j=1}^{p} f_j \cdot (F_a^J)^2 = 1 \qquad (6)$$

Κάθε παράγοντας $F_a(i)$, ο οποίος είναι διάνυσμα, υπολογίζεται με την παρακάτω σχέση

$$F_a(i) = \sum_{j=1}^{p} f_j^i \cdot \varphi(j) \qquad (7)$$

Σε κάθε παραγοντικό άξονα Δ_a ισχύουν οι εξής σχέσεις:

$$\sum_{i=1}^{n} f_i \cdot F_a(i) = 0 \qquad (8)$$

$$\sum_{i=1}^{n} f_i \cdot (F_a(i))^2 = \lambda_a \qquad (9)$$

ενώ για δύο διαφορετικούς παραγοντικούς άξονες Δ_r και Δ_s ισχύει:

$$\sum_{i=1}^{n} f_i \cdot F_r(i) F_s(i) = 0 \qquad (10)$$

Ο αριθμός $F_a(i)$ σε απόλυτη τιμή μετρά την απόσταση που χωρίζει το κέντρο βάρους $G=\{f_J\}$ του νέφους $N(I)_J$ από τη προβολή του προφίλ f_{ij} (που παριστάνει η γραμμή i του πίνακα δεδομένων) πάνω στον άξονα Δ_a.

Γενικώς ισχύει

$$d^2\left(f_J^i,\ G\right) = \sum_{a=1}^{p-1} F_a^2(i) \qquad (11)$$

Επομένως η απόσταση που χωρίζει το κέντρο βάρους $G=\{f_J\}$ από την προβολή του προφίλ γραμμή f_j^i λ.χ στο παραγοντικό επίπεδο $\Delta_1 x \Delta_2$ είναι η υποτείνουσα του ορθογωνίου τριγώνου με πλευρές τα $F_1(i)$ και $F_2(i)$. Δηλαδή για το παραγοντικό επίπεδο 1x2 ισχύει η σχέση

$$d^2\left(f_J^i,\ G\right) = F_1^2(i) + F_2^2(i) \qquad (12)$$

Η απόσταση $d^2(f_J^i, f_J)$ υπολογίζεται επίσης με τη χρήση της παρακάτω σχέσης

$$d^2\left(f_J^i,\ G\right) = \sum_{j=1}^{p} \frac{1}{f_i} (f_j^i - f_J)^2 \qquad (13)$$

Η σχέση 12 υποδεικνύει ότι οι παραγοντικοί άξονες πάνω στους οποίους υπολογίζονται οι παράγοντες $F_a(i)$ είναι ορθογώνιοι, οπότε το σύστημα των p-1 παραγοντικών αξόνων αποτελεί εκ κατασκευής ένα ορθοκανονικό σύστημα συντεταγμένων στον διανυσματικό χώρο R^{p-1}. Στη συνέχεια θα παραθέσουμε ένα αριθμητικό παράδειγμα προς επιβεβαίωση της σχέσης 13.

Αριθμητικό παράδειγμα

Χρησιμοποιώντας ένα απλό αριθμητικό παράδειγμα διαπιστώνει κανείς εύκολα την ισχύει των σχέσεων 12 και 13.
Δίνεται ο παρακάτω πίνακας σύμπτωσης

Πίνακας 3:Πίνακας δεδομένων

	J_1	**J_2**	**J_3**	
I_1	0	1	0	1
I_2	1	0	1	2
I_3	1	1	0	2
I_4	0	0	1	1
	2	**2**	**2**	**6**

Βρίσκουμε καταρχήν τα προφίλ των γραμμών f_J^I και των στηλών f_I^J .Ήτοι

$$f_J^I = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{και} \quad f_I^J = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \end{pmatrix}$$

Τα προφίλ f_I^J προέκυψαν αφού προηγουμένως πήραμε τον ανάστροφο πίνακα του T(4,3)

Για την βασική εφαρμογή της Π.Α.Α απαιτείται η εύρεση του συμμετρικού τετραγωνικού

πίνακας S_J^J ο οποίος υπολογίζεται με το παρακάτω γινόμενο

$$S_J^J = f_J^I \circ f_I^J = \frac{1}{2}\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \cdot \frac{1}{2}\begin{pmatrix} 0 & 2 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \frac{1}{4} \cdot \begin{pmatrix} 2 & 1 & 1 \\ 1 & 3 & 0 \\ 1 & 0 & 3 \end{pmatrix}$$

Στη συνέχει ακολουθεί η εύρεση των τριών χαρακτηριστικών ριζών του τετραγωνικού πίνακα S_J^J οι οποίες είναι οι εξής : $\lambda_0 = \frac{4}{4} = 1$ $\lambda_1 = \frac{3}{4}$ $\lambda_2 = \frac{1}{4}$

Ως γνωστόν σε κάθε χαρακτηριστική ρίζα λ_i (εκτός της τερτιμένης ρίζας λ_0) αντιστοιχεί ένα χαρακτηριστικό διάνυσμα $u_a=\varphi_a^J$ που συνδέεται με το παραγοντικό άξονα Δ_a.

Για κάθε χαρακτηριστικό διάνυσμα φ_a^J (j=1,...p) οι συντεταγμένες του ικανοποιούν όπως προαναφέραμε τις σχέσεις 5 και 6

Οι τιμές των χαρακτηριστικών διανυσμάτων παρουσιάζονται στις σχέσεις 14 και 15

$$\varphi_{11}=0 \text{ , } \varphi_{12}=\frac{\sqrt{6}}{2} \text{ και } \varphi_{13}= -\frac{\sqrt{6}}{2} \qquad (14)$$

$$\varphi_{21}=-\sqrt{2} \quad \varphi_{22}=\frac{\sqrt{2}}{2} \text{ και } \varphi_{23}=\frac{\sqrt{2}}{2} \qquad (15)$$

Χρησιμοποιώντας τη σχέση 7 οι παράγοντες F_a^I έχουν τις παρακάτω τιμές.

$$F_a^I = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & -\sqrt{2} \\ \frac{\sqrt{6}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{6}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{6}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{6}}{4} & -\frac{\sqrt{2}}{4} \\ \frac{\sqrt{6}}{4} & -\frac{\sqrt{2}}{4} \\ -\frac{\sqrt{6}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}$$

Διαπιστώνουμε ότι επαληθεύεται η σχέση 9

$$\lambda_1 = \frac{1}{6}\cdot\left(\frac{\sqrt{2}}{2}\right)^2 + \frac{2}{6}\cdot\left(-\frac{\sqrt{2}}{4}\right)^2 + \frac{2}{6}\cdot\left(-\frac{\sqrt{2}}{4}\right)^2 + \frac{1}{6}\cdot\left(\frac{\sqrt{2}}{2}\right)^2 = \frac{1}{4}$$

$$\lambda_2 = \frac{1}{6}\cdot\left(\frac{\sqrt{2}}{2}\right)^2 + \frac{2}{6}\cdot\left(-\frac{\sqrt{2}}{4}\right)^2 + \frac{2}{6}\cdot\left(-\frac{\sqrt{2}}{4}\right)^2 + \frac{1}{6}\cdot\left(\frac{\sqrt{2}}{2}\right)^2 = \frac{1}{4}$$

Οι παράγοντες των μεταβλητών G_a^J βρίσκονται χρησιμοποιώντας την σχέση

$$G_a^J = \frac{1}{\sqrt{\lambda_\alpha}} F_a^J \text{ o } f_I^J \qquad (16)$$

Απ'όπου προκύπτει

$$G_a^J = \begin{pmatrix} 0 & -\frac{\sqrt{2}}{2} \\ \frac{3\sqrt{2}}{4} & \frac{\sqrt{2}}{4} \\ -\frac{3\sqrt{2}}{4} & \frac{\sqrt{2}}{4} \end{pmatrix}$$

Για τις τέσσερις γραμμές i_1,i_2,i_3 και i_4 οι τιμές των συντεταγμένων στους δύο παραγοντικούς άξονες Δ_1 και Δ_2 έχουν ως εξής:

$$F_1(i_1)=\frac{\sqrt{6}}{2}, F_1(i_2)=-\frac{\sqrt{6}}{4}, F_1(i_3)=\frac{\sqrt{6}}{4}, F_1(i_4)=-\frac{\sqrt{6}}{2}$$

$$F_2(i_1)=\frac{\sqrt{2}}{2}, F_2(i_2)=-\frac{\sqrt{2}}{4}, F_2(i_3)=-\frac{\sqrt{2}}{4}, F_2(i_4)=\frac{\sqrt{2}}{2}$$

ενώ οι αντίστοιχες μάζες των γραμμών ισούνται με

$$f_1=\frac{1}{6}, f_2=\frac{2}{6}, f_3=\frac{2}{6}, f_4=\frac{1}{6}$$

Το ότι οι δύο παραγοντικοί άξονες Δ_1 και Δ_2 είναι κάθετοι, θα προκύψει, με την επαλήθευση της σχέσης 12, η οποία αποτελεί τη διατύπωση του Πυθαγόρειου Θεωρήματος στο επίπεδο.

Π.χ Για την γραμμή i_2 έχουμε

$$F_1(i_2)=-\frac{\sqrt{6}}{4}=-0,612, F_2(i_2)=-\frac{\sqrt{2}}{4}=-0,354$$

Χρησιμοποιώντας τη σχέση 13 έχουμε ότι

$$d^2(i_2,f_J)=\frac{1}{\frac{2}{6}}\left(\frac{1}{2}-\frac{2}{6}\right)^2+\frac{1}{\frac{2}{6}}\left(0-\frac{2}{6}\right)^2+\frac{1}{\frac{2}{6}}\left(\frac{1}{2}-\frac{2}{6}\right)^2=3\cdot\frac{1}{36}+3\cdot\frac{4}{36}+3\cdot\frac{1}{36}=\frac{18}{36}=\frac{1}{2}$$

Οπότε $$d^2(i_2,f_J)=\frac{1}{2}\rightarrow \quad d(i_2,f_J)=0,707$$

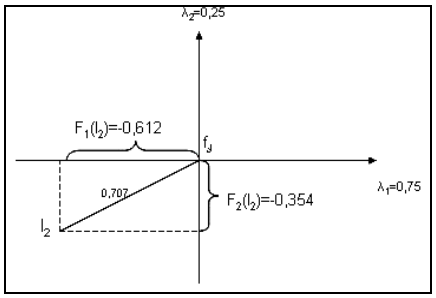
Χρησιμοποιώντας τις τιμές των $F_1(i_2)$ και $F_2(i_2)$ της γραμμής i_2 στους δύο παραγοντικούς άξονες Δ_1 και Δ_2 έχουμε

$$F_1^2(i_2)+F_2^2(i_2)=\left(-\frac{\sqrt{6}}{4}\right)^2+\left(-\frac{\sqrt{2}}{4}\right)^2=\frac{6}{16}+\frac{2}{16}=\frac{1}{2}$$

Επομένως επαληθεύεται η σχέση 12

$$d^2(i_2,f_J)=F_1^2(i_2)+F_2^2(i_2) \qquad (17)$$

Σχηματικά το παραγοντικό επίπεδο 1x2 είναι το ακόλουθο.



σχήμα 2: Επαλήθευση της σχέσης $d^2(i_2,f_J)=F_1^2(i_2)+F_2^2(i_2)$

Προφανώς η σχέση 17 ισχύει και στο πολυδιάστατο χώρο R^{p-1}

$d^2(i_k,f_J)=F_1^2(i_k)+F_2^2(i_k)+.....+F_s^2(i_k)$ όπου k=1,…n και s=1,…,p-1 (18)

Η σχέση 18 είναι η έκφραση του Πυθαγόρειου θεωρήματος στο πολυδιάστατο χώρο R^{p-1}

Όσον αφορά στην απόσταση μεταξύ του προφίλ μιας γραμμής f_J^i και του προφίλ μιας μεταβλητής f_I^j ισχύει η παρακάτω σχέση.

$$d^2\left(f_J^{\,i},\ f_I^{\,j}\right)=\sum_{p=1}^{p-1}\left[F_p(i)-G_p(j)\right]^2 \qquad (19)$$

Με τα δεδομένα του παραδείγματος για την γραμμή i_1 και την στήλη j_1 έχουμε

$$d^2(i_1,j_1)=\left[F_1(i_1)-G_1(j_1)\right]^2+\left[F_2(i_1)-G_2(j_1)\right]^2=\left[\frac{\sqrt{6}}{2}-0\right]^2+\left[\frac{\sqrt{2}}{2}-(-\frac{\sqrt{2}}{2})\right]^2=3,5$$

Οπότε

$$d(i_1,j_1)=\sqrt{3,5}=1,871$$

Σύνδεση μεταξύ των «γραμμών» και των «στηλών» ενός διδιάστατου πίνακα δεδομένων

Έστω σ' ένα πίνακα δεδομένων T(n,p) οι n γραμμές αντιστοιχούν σε n ερωτώμενους, ενώ στις p γραμμές του πίνακα οι τιμές των p ερωτήσεων που αντιστοιχούν σε p μεταβλητές. Στο ερώτημα του εντοπισμού ενός ερωτώμενου με ποια μεταβλητή συνδέεται περισσότερο, θα γίνει λεπτομερής αναφορά, χρησιμοποιώντας ένα συγκεκριμένο παράδειγμα έξι ποιοτικών μεταβλητών (για την μέτρηση των οποίων χρησιμοποιήθηκε η 5βάθμια κλίμα Likert, όπου το 5 αφορούσε την άριστη εντύπωση), στο οποίο απάντησαν 99 άτομα, τα οποία αποτελούν μία από τις πέντε κλάσεις, που δημιουργήθηκαν με την εφαρμογή της Ανιούσας Ιεραρχικής ταξινόμησης με την διαδικασία VACOR, σε 1721 ξένους επισκέπτες της Θεσσαλονίκης. Το ερωτηματολόγιο αφορούσε έξι ερωτήματα σχετικά με το πώς βαθμολόγησαν οι ξένοι επισκέπτες α) τα αξιοθέατα της πόλης της Θεσσαλονίκης β) την Ελληνική κουζίνα γ) την νυχτερινή ζωή της πόλης δ) το αρχιτεκτονικό της στυλ ε) την ασφάλειά της και στ) την φιλικότητα των ντόπιων.

Οι έξι μεταβλητές παρίστανται αντιστοίχως ως εξής : Δ4,Δ5,Δ6,Δ7,Δ8,Δ9. Με δεδομένη την ταξινόμηση των 99 ατόμων σε μία από τις ομάδες που δημιούργησε η ταξινόμηση των 1721 ατόμων με την μέθοδο VACOR ο πίνακας 4 παρουσιάζει τμήμα των απαντήσεών τους.

Πίνακας 4: Τμήμα του πίνακα δεδομένων

A/A	ind	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
1	11	5	3	3	5	5	5
2	20	2	2	1	2	3	3
3	60	4	2	3	5	3	5
.	.	.	.	.	.	.	.
21	315	4	3	4	5	3	5
.	.	.	.	.	.	.	.
67	1085	3	4	1	3	4	5
.	.	.	.	.	.	.	.
94	1623	4	0	3	5	4	3
.	.	.	.	.	.	.	.
99	1712	4	2	1	5	3	2

Με βάση τον στατιστικό έλεγχο της διαφοράς των αναλογιών (σχέσεις 1 και 2) σε επίπεδο σημαντικότητας 5%, προσδιορίζεται η σύνδεση κάθε κλάσης ή ερωτώμενου με μία ή περισσότερες μεταβλητές.

Έχοντας εφαρμόσει την Ανιούσα Ιεραρχική Ταξινόμηση με την μέθοδο VACOR στα δεδομένα του πίνακα 4, δημιουργήθηκε η συγκεκριμένη τυπολογία με τις εξής πέντε

ομοιογενείς συστάδες :180,186,191,192 και 193 στις οποίες χρησιμοποιώντας τον προαναφερόμενο έλεγχο υποθέσεων προκύπτει ο πίνακας 5, όπου η συστάδα 186 φαίνεται να συνδέεται με δύο μεταβλητές τις Δ5 και Δ6, ενώ η συστάδα 192 συνδέεται περισσότερο με τις μεταβλητές Δ5, Δ7 και Δ9, με μεγαλύτερη ένταση όμως με την Δ5 ($Z_{\Delta 5}$=6,3422).

Πίνακας 5: Παρουσιάζει την σύνδεση μεταξύ των 5 συστάδων και των έξι μεταβλητών

ind	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
180	-0,4215	-14,4224	13,6994	0,9611	2,2438	-2,414
186	0,9345	1,8842	5,5692	-1,3116	-3,1378	-2,3919
191	12,4939	-15,3903	-15,8726	11,0566	-2,7345	3,528
192	-4,0938	6,3422	-6,8836	1,9318	0,1465	2,51
193	-0,2294	0,1392	-1,2727	-1,6656	2,2553	0,5183

Από την εφαρμογή του ελέγχου στις τιμές των έξι μεταβλητών για κάθε ερωτώμενο προκύπτει ο πίνακας 6,από τον οποίο διαπιστώνεται ότι ο ερωτώμενος 20 συνδέεται περισσότερο με τις μεταβλητές Δ5, Δ8 και Δ9 με μεγαλύτερη ένταση όμως με την Δ5 ($Z_{\Delta 5}$=7,1634), ενώ ο ερωτώμενος 315 φαίνεται να συνδέεται με τρεις μεταβλητές τις Δ5, Δ6 και Δ7 με μεγαλύτερη ένταση όμως με την Δ6 ($Z_{\Delta 6}$=7,9739).

Πίνακας 6: Παρουσιάζει την σύνδεση μεταξύ ερωτώμενων και μεταβλητών

ind	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
11	-0,5345	1,5176	0,631	0,0817	0,5959	-1,7833
20	-4,8848	7,1634	-4,8654	-4,3151	5,021	2,4658
60	-1,7209	-2,0604	3,6369	4,0674	-5,8574	2,0785
315	-3,4385	2,9401	7,9739	1,909	-7,1596	-0,0129
1185	21,1945	-15,3903	-15,8726	13,2493	-21,5643	10,9749
1623	1,522	-15,3903	6,7143	8,1673	2,6932	-5,6008

Βέβαια με την διαδικασία αυτή υπάρχουν και πολλοί ερωτώμενοι που δεν συνδέονται ιδιαίτερα με καμία μεταβλητή, όπως λ.χ ο ερωτώμενος 11, επειδή η τιμή της z κατανομής περιέχεται μεταξύ των τιμών -1,645<z<1,645, δηλαδή οφείλεται σε τυχαίους παράγοντες, αφού ισχύει η H_0.

Από τον πίνακα 6, λοιπόν, όπως προαναφέρθηκε θεωρούμε ότι η μεγαλύτερη τιμή του z που αντιστοιχεί σε κάποια μεταβλητή από την σειρά των έξι τιμών, προσδιορίζει την σύνδεσή του ερωτώμενου με την συγκεκριμένη μεταβλητή.

Χρησιμοποιώντας ως εναλλακτική προσέγγιση του προβλήματος της σύνδεσης των ερωτωμένων με τις μεταβλητές, σκόπιμο είναι να εφαρμοστεί στα δεδομένα του πίνακα 4 η Παραγοντική Ανάλυση των Αντιστοιχιών (-Π.Α.Α-) και στη συνέχεια βάσει των συντεταγμένων των ερωτωμένων και των μεταβλητών στους p-1 παραγοντικούς άξονες, να προσδιοριστεί με την χρήση της Ευκλείδειας μετρικής, ποιος ερωτώμενος συνδέεται κυρίως με ποια μεταβλητή.

Η εξαγωγή των αποτελεσμάτων με την χρήση της Π.Α.Α προϋποθέτει σε κάθε περίπτωση να ληφθεί υπόψη η ερμηνευτική ικανότητα των παραγοντικών αξόνων.

Προς τούτο υπάρχουν τρεις περιπτώσεις ερμηνείας, ανάλογα με την ερμηνευτική ικανότητα των παραγοντικών αξόνων. Δηλαδή να χρησιμοποιηθούν δύο παραγοντικοί άξονες ή τρεις ή να χρησιμοποιηθούν όλοι οι παραγοντικοί άξονες που προσδιορίζονται από το σύνολο των μεταβλητών

1. Με δύο παραγοντικούς άξονες.

Αναλύοντας τα δεδομένα του πίνακα 4 με την Παραγοντική Ανάλυση των Αντιστοιχιών, προκύπτουν τα εξής:

Α) Προβολή χαρακτηριστικών ριζών

Πίνακας 7: Ιστόγραμμα χαρακτηριστικών ριζών

ΑΞΩΝ	ΑΔΡΑΝΕΙΑ	%ΕΡΜΗΝΕΙΑΣ	ΑΘΡΟΙΣΜΑ	ΙΣΤΟΓΡΑΜΜΑ ΧΑΡΑΚΤΗΡ.ΡΙΖΩΝ
1	0,0305713	33,07	33,07	|**********
2	0,0285268	30,86	63,93	|**********
3	0,0149623	16,19	80,11	|**********
4	0,0097252	10,52	90,63	|**********
5	0,0086587	9,37	100,00	|**********

Από τον πίνακα 7 προκύπτει ότι με τους δύο πρώτους παραγοντικούς άξονες αντλείται το 63,93% της συνολικής πληροφόρησης που προέρχεται από τα δεδομένα του πίνακα 4.

Πίνακας 8: Συντεταγμένες των ερωτώμενων και των μεταβλητών για το σύνολο των παραγοντικών αξόνων

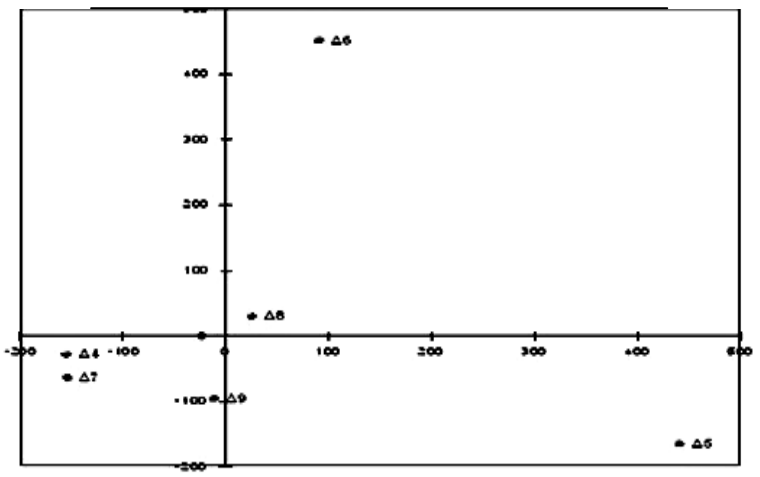
Συντεταγμένες των ερωτώμενων

ind	FA1	FA2	FA3	FA4	FA5
11	34	12	-3	-15	-34
20	182	-122	-111	46	12
60	-49	51	80	-107	90
73	-3	-111	-135	-103	-238
82	162	73	-101	-60	-33
88	-81	-35	-182	-103	-108
99	-7	104	92	-48	24

Συντεταγμένες των μεταβλητών

ind	GA1	GA2	GA3	GA4	GA5
Δ4	-153	-29	136	114	-78
Δ5	441	-165	97	-34	-63
Δ6	91	451	70	-34	37
Δ7	-152	-65	-1	-177	-29
Δ8	27	29	-229	57	-64
Δ9	-10	-97	-8	38	169

Στη συνέχεια παρατηρούμε το παραγοντικό επίπεδο 1x2

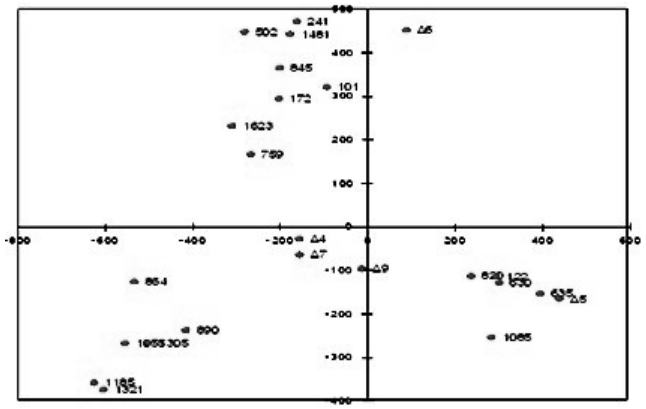


σχήμα 3: Παραγοντικό επίπεδο 1x2 των μεταβλητών

Με βάση τους δύο πρώτους παραγοντικούς άξονες το επίπεδο 1x2 χωρίζεται σε τέσσερις υποχώρους. Στον 1^{ο} υποχώρο τοποθετούνται τα σημεία τα οποία έχουν και τις δύο συντεταγμένες θετικές. Με την ίδια διαδικασία ανάλογα με τα πρόσημα των συντεταγμένων των σημείων τοποθετούνται στον 2^{ο}, στον 3^{ο}, είτε στον 4^{ο} υποχώρο.

Έτσι στον 1ο υποχώρο βρίσκονται οι μεταβλητές Δ6 και Δ8, στον 2ο δεν βρίσκεται καμία μεταβλητή, στον 3ο υποχώρο βρίσκονται οι μεταβλητές Δ4,Δ7 και Δ9, ενώ στον 4ο υποχώρο βρίσκεται η μεταβλητή Δ5.

Πάνω στο παραγοντικό επίπεδο 1x2 τοποθετούνται και τα προφίλ των 99 ερωτώμενων, ανάλογα με τα πρόσημα των συντεταγμένων τους, όπου μία ομάδα «υποκειμένων» στο 2ο υποχώρο είναι «ορφανή» από μεταβλητή. (Η επιλογή των σημείων έγινε βάσει των κριτηρίων COR και CTR)



σχήμα 4: Παραγοντικό επίπεδο 1x2 των «αντικειμένων» και των μεταβλητών

Επειδή οι δύο πρώτοι παραγοντικοί άξονες δημιουργούν ένα ορθογώνιο σύστημα συντεταγμένων, για να μετρηθεί μία απόσταση μεταξύ μιας μεταβλητής και ενός «υποκειμένου» χρησιμοποιείται η Ευκλείδεια μετρική μεταξύ δύο σημείων κάνοντας χρήση των συντεταγμένων τους πάνω στους δύο άξονες. Με την χρήση του λογισμικού MAD υπολογίζεται η μικρότερη απόσταση κάθε «υποκειμένου» μεταξύ των έξι μεταβλητών, οπότε προκύπτει το παρακάτω αποτέλεσμα.

Πίνακας 9: Σύνδεση μεταξύ «υποκειμένων» και μεταβλητών βάσει δύο παραγόντων

ΜΕΤ	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
ΠΛΗΘΟΣ	8	7	5	14	45	20
	88	122	101	358	11	20
	172	238	241	389	60	73
	554	420	502	629	82	116
	644	630	845	645	99	209
	693	635	1481	864	176	213
	759	820		890	246	378
	1548	1085		1006	298	596
	1623			1055	301	641
				1185	312	643
				1279	315	678
				1305	355	703
				1321	368	704
				1433	399	705
				1712	452	879
					536	882
					553	1086
					555	1105
					567	1328
					702	1503
					745	1696
					914	
					922	
					990	
					1000	
					1019	
					1020	
					1114	
					1127	
					1140	
					1156	
					1172	
					1200	
					1269	
					1307	
					1383	
					1482	
					1530	
					1540	
					1546	
					1570	
					1580	
					1619	
					1626	
					1633	
					1694	

Από τον πίνακα 9 διαπιστώνεται ότι 7 ερωτώμενοι {122,238,420,630,635,820 και 1085} συνδέονται με την μεταβλητή Δ5, καθόσον βρίσκονται στη μικρότερη απόσταση, από τις υπόλοιπες πέντε μεταβλητές, ενώ με την μεταβλητή Δ4 συνδέονται 8 ερωτώμενοι.

2) Με τρεις παραγοντικούς άξονες

Εφόσον χρησιμοποιηθούν οι τρεις πρώτοι παραγοντικοί άξονες και με δεδομένο ότι αποτελούν στον χώρο των τριών διαστάσεων ένα τρισορθογώνιο σύστημα συντεταγμένων, τότε δημιουργούνται οκτώ υποχώροι, στους οποίους ανάλογα με τα πρόσημα των συντεταγμένων των σημείων τοποθετούνται σ' έναν απ' αυτούς.

Συνεπώς αν συμβολίσουμε τα σημεία που βρίσκονται στο $1^{ο}$,$2^{ο}$,$3^{ο}$ και $4^{ο}$ υποχώρο με το σύμβολο ↑ δίπλα από την ταυτότητα του σημείου, ενώ για τα σημεία στον $5^{ο}$, $6^{ο}$, $7^{ο}$ και $8^{ο}$ υποχώρο θέσουμε το σύμβολο ↓ ,έχουμε για πρώτη φορά στη παγκόσμια βιβλιογραφία, μια απεικόνιση του τριδιάστατου χώρου στο επίπεδο, (χωρίς να χρησιμοποιείται η μέθοδος της προοπτικής), οπότε τα διαγράμματα αυτά στο εξής θα αναφέρονται ως **διαγράμματα Καραπιστόλη**.

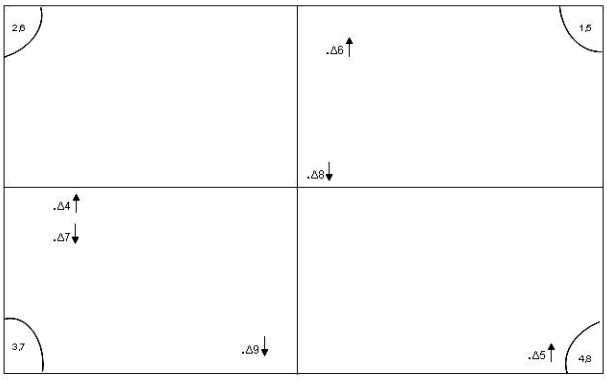
Εφαρμόζοντας την συγκεκριμένη διαδικασία. προκύπτουν αφ' ενός ο πίνακας 10 με την τοποθέτηση των μεταβλητών στους 8 υποχώρους, αφετέρου ο πίνακας 11 με την τοποθέτηση των 99 ερωτώμενων στους αντίστοιχους υποχώρους, καθώς και το αντίστοιχο διάγραμμα Καραπιστόλη.

Πίνακας 10: Οι οκτώ υποχώροι με τις μεταβλητές που εντοπίζονται σ' αυτούς

A/A	1ος	2ος	3ος	4ος	5ος	6ος	7ος	8ος
1	Δ6		Δ4	Δ5	Δ8		Δ7	
2							Δ9	

Πίνακας 11: Οι οκτώ υποχώροι με τους ερωτώμενους που εντοπίζονται σ' αυτούς

ΜΕΤ	1ος	2ος	3ος	4ος	5ος	6ος	7ος	8ος
ΠΛΗΘΟΣ	23	6	10	9	11	10	15	15
	298	60	358	116	11	101	73	20
	312	99	645	209	82	172	88	122
	315	301	879	630	246	241	213	176
	399	502	1055	635	355	554	389	238
	567	536	1105	643	452	759	596	368
	702	1546	1185	703	553	845	629	378
	745		1321	704	555	1383	644	420
	914		1328	705	1140	1481	693	641
	922		1433	882	1172	1548	864	678
	1019		1712		1307	1623	890	820
	1020				1694		1006	990
	1114						1086	1000
	1127						1279	1085
	1156						1305	1482
	1200						1503	1696
	1269							
	1530							
	1540							
	1570							
	1580							
	1619							
	1626							
	1633							



σχήμα 5: Παραγοντικός χώρος 1x2x3 των μεταβλητών

Από το **διάγραμμα Καραπιστόλη** διαπιστώνεται ότι οι μεταβλητές Δ4 και Δ7, καθώς και οι μεταβλητές Δ6 και Δ8 στον τριδιάστατο χώρο βρίσκονται σε διαφορετικούς υποχώρους, με ότι αυτό μπορεί να σημαίνει για την ερμηνεία τους σε ποσοστό 80,11%, έναντι της ερμηνείας τους σε ποσοστό 63,93% που παρέχει το παραγοντικό επίπεδο 1x2.

Από τον πίνακα 11 προκύπτει με βάση τα πρόσημα των τριών συντεταγμένων των σημείων, ότι με την τοποθέτησή τους στους οκτώ υποχώρους, η μεταβλητή Δ5 και οι 9 ερωτώμενοι που ανήκουν στον 4^{ο} υποχώρο όφειλαν με την χρήση των τριών παραγόντων να συνδέονται μεταξύ τους. Υπολογίζοντας όμως τις ελάχιστες αποστάσεις μεταξύ των ερωτώμενων και των μεταβλητών προκύπτει ο πίνακας 12, ο οποίος πληροφορεί ότι γενικώς μόνο 4 ερωτώμενοι (οι 122,630,635 και 1085) συνδέονται με την μεταβλητή Δ5.

Το ενδιαφέρον είναι ότι από τους εννέα ερωτώμενους που ανήκουν στο 4^{ο} υποχώρο μαζί με την μεταβλητή Δ5 (πίνακας 11), μόνο δύο εξ' αυτών ο 630 και ο 635 συνδέονται πράγματι με την μεταβλητή Δ5, επειδή βρίσκονται σε μικρότερη απόσταση, ενώ οι άλλοι 7 συνδέονται με άλλες μεταβλητές όπως λ.χ ο ερωτώμενος 116 συνδέεται με την μεταβλητή Δ9 που βρίσκεται στον 7^{ο} υποχώρο, λόγω μικρότερης απόστασης από εκείνη που έχει με την μεταβλητή Δ5.

Πίνακας 12: Σύνδεση των ερωτώμενων με τις μεταβλητές βάσει τριών παραγόντων

ΜΕΤ	**Δ4**	**Δ5**	**Δ6**	**Δ7**	**Δ8**	**Δ9**	
ΠΛΗΘΟΣ	**15**	**4**	**6**	**16**	**11**	**47**	
	60	122	101	389	82	11	705
	99	630	241	554	88	20	745
	301	635	502	629	172	73	820
	358	1085	845	644	452	116	879
	536		1269	693	553	176	882
	645		1481	759	555	209	914
	702			864	990	213	922
	1019			890	1000	238	1086
	1020			1006	1140	246	1114
	1105			1055	1503	298	1127
	1156			1279	1694	312	1172
	1185			1305		315	1200
	1321			1383		355	1307
	1433			1548		368	1328
	1546			1623		378	1482
				1712		399	1530
						420	1540
						567	1570
						596	1580
						641	1619
						643	1626
						678	1633
						703	1696
						704	

Αυτή η διαφορετική σύνδεση των ερωτώμενων με την μεταβλητή Δ5, οφείλεται σε ποσοστό 16,19% της πληροφορίας που παρέχει ο $3^{ος}$ παραγοντικός άξονας.

3.Με το σύνολο των παραγοντικών αξόνων

Χρησιμοποιώντας το σύνολο των πέντε παραγοντικών αξόνων δημιουργείται μία ορθοκανονική βάση στον R^5, όπου τοποθετούνται όλες οι μεταβλητές και όλοι οι ερωτώμενοι στις πραγματικές τους θέσεις, απ' όπου αντλείται το σύνολο της πληροφόρησης που παρέχει ο πίνακας δεδομένων.

Η σύνδεση των ερωτώμενων με βάση τη ελάχιστη απόσταση μεταξύ μεταβλητών και «υποκειμένων» χρησιμοποιώντας όλες τις συντεταγμένες παρουσιάζεται στον πίνακα 13, σε αντιδιαστολή με τις συνδέσεις τους χρησιμοποιώντας την z κατανομή.

Από τον πίνακα 13 προκύπτει ότι 7 ερωτώμενοι {122,238,420,630,635,820,1085} συνδέονται με την μεταβλητή Δ5, (όπως και στη περίπτωση των δύο παραγόντων), ενώ όσον αφορά τους ερωτώμενους που συνδέονται με την μεταβλητή Δ4 διαπιστώνουμε ότι αυξάνοντας το ποσοστό πληροφόρησης από 63,11% στο 100%, διαφοροποιείται το πλήθος και οι ερωτώμενοι που συνδέονται με την μεταβλητή, παρατήρηση η οποία ισχύει και για κάποιες από τις υπόλοιπες μεταβλητές.

Έτσι με την μεταβλητή Δ4 με τους δύο πρώτους παράγοντες συνδέονται 8 ερωτώμενοι, με τους τρεις παράγοντες συνδέονται 15 και με όλους τους παράγοντες συνδέονται 16. Αξιοσημείωτο είναι ότι κανείς από τους 8 ερωτώμενους που φαίνεται να συνδέονται με την μεταβλητή Δ4,βάσει των δύο παραγόντων δεν φαίνεται να συνδέονται με τους 15 ερωτώμενους, όταν γίνεται χρήση των τριών παραγόντων, ενώ όταν γίνεται χρήση του συνόλου των παραγόντων μόνο 10 από τους 15 συνεχίζουν να συνδέονται με την μεταβλητή Δ4.

Αυτή η διαφοροποίηση στο πλήθος και στο ποιοι ερωτώμενοι συνδέονται με κάθε μεταβλητή, ανάλογα με το πλήθος των παραγόντων που λαμβάνονται υπόψη, σημαίνει ότι η μόνη ορθή διαδικασία εντοπισμού της σύνδεσης των ερωτώμενων με τις μεταβλητές ενός πίνακα δεδομένων, είναι να χρησιμοποιούνται όλοι οι παράγοντες που εξάγονται από την Παραγοντική Ανάλυση των Αντιστοιχιών, αφού σε αυτή την περίπτωση δημιουργείται ένας ν-διάστατος ορθοκανονικός χώρος, ο οποίος απεικονίζει την πραγματική εικόνα των σχέσεων μεταξύ των σημείων που απεικονίζουν τους ερωτώμενους και των σημείων που απεικονίζουν τις μεταβλητές, προσφέροντας το 100% της πληροφορίας ανεξάρτητα από την συμμετοχή κάθε «υποκειμένου» στη διαμόρφωση των παραγοντικών αξόνων.

Η σύνδεση των υποκειμένων με τις μεταβλητές βάσει της Ευκλείδειας μετρικής και της μέγιστης τιμής της z κατανομής παρουσιάζονται στον παρακάτω πίνακα.

Στόχος της εργασίας τελικά είναι να αποδειχθεί η πιο αποτελεσματική λύση εντοπισμού της σύνδεσης μεταξύ μεταβλητών και υποκειμένων, έχοντας χρησιμοποιήσει την z κατανομή και την Ευκλείδεια μετρική. Ο στόχος αυτός μπορεί να υλοποιηθεί εφόσον εκπαιδευτούν τα δεδομένα που απορρέουν από τον πίνακα 13 χρησιμοποιώντας ταξινομητές μηχανικής μάθησης και ειδικότερα την μηχανή μάθησης SVM

Πίνακας 13: Σύνδεση υποκειμένων και μεταβλητών με τις δύο διαφορετικές διαδικασίες

	Συνδεση με Ευκλείδεια μετρική							Συνδεση με z μετρική					
IND	**Δ4**	**Δ5**	**Δ6**	**Δ7**	**Δ8**	**Δ9**		**Δ4**	**Δ5**	**Δ6**	**Δ7**	**Δ8**	**Δ9**
Πλήθος	**16**	**7**	**9**	**20**	**26**	**21**		**6**	**36**	**21**	**11**	**22**	**3**
	99	122	101	60	11	116		536	11	99	60	73	213
	301	238	172	358	20	209		645	20	101	358	82	1433
	536	420	241	389	73	213		1055	116	172	629	88	1696
	645	630	502	567	82	298		1105	122	241	879	176	
	702	635	845	629	88	312		1185	209	301	1086	368	
	759	820	1269	644	176	315		1548	238	315	1279	378	
	1019	1085	1481	693	246	378			246	399	1305	389	
	1020		1570	864	355	596			298	502	1321	452	
	1055		1633	879	368	641			312	845	1328	554	
	1105			890	399	643			355	1114	1623	644	
	1127			914	452	678			420	1156	1712	693	
	1185			1006	553	703			553	1172		759	
	1200			1086	554	704			555	1269		864	
	1546			1279	555	705			567	1481		890	
	1548			1305	990	745			596	1530		990	
	1626			1321	1000	882			630	1540		1000	
				1328	1140	922			635	1546		1006	
				1383	1172	1114			641	1570		1140	
				1623	1307	1156			643	1580		1383	
				1712	1482	1433			678	1619		1482	
					1503	1696			702	1633		1503	
					1530				703			1694	
					1540				704				
					1580				705				
					1619				745				
					1694				820				
									882				
									914				
									922				
									1019				
									1020				
									1085				
									1127				
									1200				
									1307				
									1626				

Γενικά περί εκπαίδευσης δεδομένων με ταξινομητές μηχανικής μάθησης

Η **μηχανική μάθηση** (machine learning) είναι μια περιοχή της τεχνικής νοημοσύνης η οποία αφορά αλγορίθμους και μεθόδους που επιτρέπουν στους υπολογιστές να «μαθαίνουν». Στόχος της Μηχανικής μάθησης είναι με την χρήση ενός υπολογιστικού συστήματος, η δημιουργία μοντέλων χρησιμοποιώντας ένα σύνολο δεδομένων.

Έχουν αναπτυχθεί πολλές τεχνικές μηχανικής μάθησης που χρησιμοποιούνται ανάλογα με τη φύση του προβλήματος και εμπίπτουν σε ένα από τα παρακάτω δυο είδη:

1.Μάθηση με επίβλεψη (supervised learning)

2.Μάθηση χωρίς επίβλεψη (unsupervised learning

Στη μάθηση με επίβλεψη το σύστημα καλείται να «μάθει» μια έννοια ή συνάρτηση από ένα σύνολο δεδομένων, η οποία αποτελεί περιγραφή ενός μοντέλου.

Στη μάθηση χωρίς επίβλεψη το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας πρότυπα, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι.

Στη παρούσα εργασία θα χρησιμοποιηθεί η μάθηση με επίβλεψη, όπου το σύστημα πρέπει να «μάθει» επαγωγικά μια συνάρτηση που ονομάζεται **συνάρτηση στόχος** (target function) και αποτελεί έκφραση του μοντέλου που περιγράφει τα δεδομένα.

Η συνάρτηση στόχος χρησιμοποιείται για την πρόβλεψη της τιμής μίας μεταβλητής, που ονομάζεται **μεταβλητή εξόδου**, βάσει των τιμών ενός συνόλου μεταβλητών, που ονομάζονται **μεταβλητές εισόδου** ή **χαρακτηριστικά**

Στην μάθηση με επίβλεψη διακρίνονται δυο είδη προβλημάτων (learning tasks), τα προβλήματα ταξινόμησης και τα προβλήματα παρεμβολής.

Η ταξινόμηση (classification) αφορά στη δημιουργία μμοντέλων πρόβλεψης διακριτών τάξεων (κλάσεων/κατηγοριών).

Η παρεμβολή (regression) αφορά στη δημιουργία μμοντέλων πρόβλεψης αριθμητικών τιμών.

Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ). Support Vector Machines (SVM)

Οι **Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ)** χαρακτηρίζονται ως μμηχανές μμάθησης και στηρίζονται στη Θεωρία Στατιστικής Μάθησης (Statistical Learning Theory) και στα νευρωνικά δίκτυα τύπου Perceptron. Προτάθηκαν από τον Vladimir Vapnik.

Στην περίπτωση της ταξινόμησης, οι ΜΔΥ προσπαθούν να βρουν μια υπέρ-επιφάνεια (hypersurface) που να διαχωρίζει στο χώρο των παραδειγμάτων τα αρνητικά από τα θετικά παραδείγματα.

Τα ΜΔΥ χαρακτηρίζονται από τα ακόλουθα στάδια:

1 Εκπαίδευση: Στη φάση αυτή γίνονται οι υπολογισμοί των παραμέτρων του μμοντέλου μμάθησης με χρήση κατάλληλου συνόλου δεδομένων μμάθησης.

2 Δοκιμή : Το μμοντέλο παραμέτρων (support vectors) που υπολογίστηκε δοκιμάζεται για τη δυνατότητα επιτυχημένης εκτίμησης αποτελέσματος σε ένα σετ δεδομένων που δεν έχει εκπαιδευθεί.

3 Εκτίμηση επίδοσης: Υπολογίζονται κατάλληλοι δείκτες επίδοσης του μοντέλου, κυρίως του ποσοστού σφάλματος, με στόχο την διερεύνηση της δυνατότητας γενίκευσης του μμοντέλου.

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVM, Support Vector Machines) ανήκουν στους αλγόριθμους Επιβλεπόμενης Μηχανικής Μάθησης με αξιοσημείωτη επιτυχία σε προβλήματα κατάταξης. Όπως και οι περισσότεροι αλγόριθμοι μηχανικής μάθησης, παριστάνουν τα προς κατάταξη αντικείμενα ως διανύσματα ιδιοτήτων.

Στην περίπτωσή μας, τα προς κατάταξη αντικείμενα είναι ερωτώμενοι και οι ιδιότητες παρέχουν πληροφορίες όπως αν ο προς κατάταξη ερωτώμενος συνδέεται με την Α ή Β μεταβλητή.

Προκειμένου να χρησιμοποιηθούν οι ΜΔΥ σε προβλήματα ταξινόμησης με περισσότερες από δύο τάξεις έχουν προταθεί δύο κατηγορίες προσεγγίσεων

- Άμεσες: Εύρεση των διαχωριζόντων υπέρ-επιπέδων σε ένα βήμα (Vapnik (1998), Crammer and Singer (2000))
- Έμμεσες: Συνδυασμός των αποτελεσμάτων ενός συνόλου δυαδικών ΜΔΥ: ένας-εναντίον-ενός, ένας-εναντίον-όλων (Vapnik (1998))

Οι έμμεσες προσεγγίσεις είναι απλούστερες και υλοποιούνται ευκολότερα, πάντως καμία από τις προσεγγίσεις δεν επιστρέφει πιθανότητες.

Εφαρμογή της μηχανής μάθησης SVM

Για την εφαρμογή της προτεινόμενης σύγκρισης χρησιμοποιούνται εκ νέου οι έξι ποιοτικές μεταβλητές Δ4 έως Δ9 οι τιμές των οποίων αφορούν στις απαντήσεις που δόθηκαν από τους 99 ξένους επισκέπτες της Θεσσαλονίκης.

Ο πίνακας 14 παρουσιάζει την σύνδεση των αντικειμένων με τις αντίστοιχες μεταβλητές, αφενός βάσει της ελάχιστης απόστασης χρησιμοποιώντας την Ευκλείδεια μετρική, αφετέρου με την μέγιστη τιμή της z κατανομής.

Πίνακας 14: Σύνδεση των 99 αντικειμένων και των 6 μεταβλητών βάσει της Ευκλείδειας μετρικής και της μέγιστης τιμής της z κατανομής

ΕΝΔ	Δ4	.....	Δ9	MET1
11	5		5	5
20	2		3	5
60	4		5	4
73	3		2	5
82	3		4	5
88	2		2	5
99	4		4	1
101	5		5	3
116	3		5	6
122	3		5	2
172	4		5	3

ΕΝΔ	Δ4	.....	Δ9	MET2
11	5		5	2
20	2		3	2
60	4		5	4
73	3		2	5
82	3		4	5
88	2		2	5
99	4		4	3
101	5		5	3
116	3		5	2
122	3		5	2
172	4		5	3

Σημείωση 1:Οι τιμές των μεταβλητών MET1 και MET2 από 1 έως 6 αντιστοιχούν στις μεταβλητές Δ4 έως Δ9. Η σύμπτωση τιμών μεταξύ MET1 και MET2 ανέρχεται στο 49,5%

Εκπαίδευση των δεδομένων με την Μηχανή Διανυσμάτων Υποστήριξης ΜΔΥ

Εκπαιδεύοντας τα δεδομένα του πίνακα 13 με την χρήση της Μηχανής Διανυσμάτων Υποστήριξης (ΜΔΥ, Support Vector Machines) διαπιστώνεται ότι η διαδικασία σύνδεσης αντικειμένων και μεταβλητών με την Ευκλείδεια μετρική υπερέχει από την αντίστοιχη με την z κατανομή. Αυτή η διαπίστωση προκύπτει επειδή το ποσοστό επίδοσης στην εκμάθηση των δεδομένων του πίνακα 13, που αφορά στη σύνδεση των αντικειμένων με την Ευκλείδεια μετρική είναι υψηλότερο (78,89% πίνακας 16) από εκείνο που προκύπτει με την κατανομή z (71,11% Πίνακας 15).

Επί πλέον με την Ευκλείδειο μετρική στις 20 επαναλήψεις εκμάθησης των δεδομένων τα ποσοστά πάνω από την μέση τιμή είναι κατά πολύ υψηλότερα (7 στις 20 επαναλήψεις πάνω από 80% με μέγιστη τιμή 100% και κατώτερη 61,11%) από τα αντίστοιχα ποσοστά με την z κατανομή που παρουσιάζουν μέγιστη τιμή μόλις στο 88,89% και κατώτερη 44,44%.

Πίνακας 15: Εκπαίδευση με την ΜΔΥ βάσει της z κατανομής

MET_Z (20%)	
0,7778	
0,6667	
0,7778	
0,6667	
0,7778	
0,6667	
0,4444	
0,5556	71,11%
0,6111	
0,8333	
0,7778	
0,6667	
0,6667	
0,7778	
0,7222	
0,7778	
0,7778	
0,6667	
0,8889	
0,7222	

Πίνακας 16: Εκπαίδευση με την ΜΔΥ βάσει της Ευκλείδειας μετρικής

MET_EU (20%)	
0,8333	
0,7778	
0,7778	
0,7778	
0,8333	
0,8889	
0,7222	
0,7778	78,89%
0,9444	
0,7778	
1,0000	
0,7778	
0,6111	
0,7222	
0,7222	
0,8889	
0,7222	
0,8333	
0,6667	
0,7222	

Η νέα διαδικασία ταξινόμησης των γραμμών ενός πίνακα δεδομένων. Μέθοδος KARAP

Η προτεινόμενη διαδικασία ταξινόμησης απαντά στον προβληματισμό που υφίσταται σε κάθε Ανιούσα Ιεραρχική Ταξινόμηση με τη μέθοδο VACOR, ότι δηλαδή δεν εντοπίζονται με ακρίβεια τα αντικείμενα που συνδέονται με τις μεταβλητές των κλάσεων.

Για τον λόγο αυτό η προτεινόμενη νέα διαδικασία ταξινόμησης των n αντικειμένων ενός πίνακα δεδομένων T(n,p), επιτυγχάνει στο βαθμό που επιθυμεί ο ερευνητής, την

ομοιογένεια των αντικειμένων ως προς την αποκλειστική σύνδεσή τους με κάθε μία μεταβλητή ή αν επιθυμεί ως προς ένα συνδυασμό μεταβλητών. Ο αλγόριθμος της προτεινόμενης ταξινόμησης με την ονομασία **KARAP**, ο οποίος υλοποιείται με το πρόγραμμα MAD είναι ο εξής:

1. Δημιουργείται ο λογικός πίνακας 0-1 ο προερχόμενος από τον πίνακα δεδομένων T(n,p), είτε χρησιμοποιώντας διαβαθμίσεις των μεταβλητών, είτε κλίμακες Likert.
2. Η αρίθμηση κάθε αντικειμένου αντιστοιχεί από το 1 έως το n.
3. Ο λογικός πίνακας 0-1 αναλύεται με την -Π.Α.Α- για την εξαγωγή των συντεταγμένων των μεταβλητών και των αντικειμένων στους παραγοντικούς άξονες
4. Χρησιμοποιώντας τις συντεταγμένες Fa και Ga εντοπίζεται με βάση την Ευκλείδεια μετρική η σύνδεση κάθε αντικειμένου με κάθε μεταβλητή.

Εάν επιθυμεί ο ερευνητής ταξινόμηση των αντικειμένων ως προς ένα συνδυασμό μεταβλητών, λόγω του μεγάλου πλήθους των, συνεχίζει με τα παρακάτω βήματα.

5. Δημιουργείται ο πίνακας Burt που αντιστοιχεί στον λογικό πίνακα 0-1
6. Εφαρμόζεται η Ανιούσα Ιεραρχική Ταξινόμηση με την μέθοδο VACOR στα δεδομένα του πίνακα Burt
7. Βάσει της τυπολογίας της ιεραρχίας, η οποία προκύπτει από την τομή του δενδρογράμματος σε k συστάδες, κατατάσσονται τα αντικείμενα ανάλογα με τις μεταβλητές με τις οποίες συνδέονται, σύμφωνα με το βήμα 3.

Εφαρμόζοντας την προτεινόμενη ταξινόμηση με βάση τα βήματα 1 έως 3 στα 84 δεδομένα του πίνακα 4 (δεν ελήφθησαν σκόπιμα υπόψη όσοι επισκέπτες δεν απάντησαν σε κάποιο κριτήριο), προέκυψαν τα ακόλουθα αποτελέσματα.

Πίνακας 17: Ταξινόμηση των 84 ερωτώμενων βάσει της μεθόδου karap

	ΑΞΙΟΘΕΑΤΑ				
IND	Δ41	Δ42	Δ43	Δ44	Δ45
ΠΛΗΘΟΣ	0	2	4	0	5
1		20	420		554
2		1140	630		643
3			820		705
4			1085		1114
5					1628
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					

	ΚΟΥΖΙΝΑ			
Δ51	Δ52	Δ53	Δ54	Δ55
3	2	8	0	0
101	879	11		
629	1696	246		
1503		298		
		315		
		355		
		399		
		922		
		1462		

	ΝΥΚΤΕΡΙΝΗ ΖΩΗ			
Δ61	Δ62	Δ63	Δ64	Δ65
6	4	5	0	0
73	122	967		
88	635	702		
358	990	914		
1279	1000	1020		
1433		1546		
1712				

	ΑΡΧΙΤΕΚΤΟΝΙΚΟ ΣΤΥΛ			
Δ71	Δ72	Δ73	Δ74	Δ75
2	0	0	10	6
452			82	60
1105			301	312
			388	678
			1269	745
			1520	1207
			1540	1363
			1570	
			1580	
			1619	
			1633	

	ΑΣΦΑΛΕΙΑ			
Δ81	Δ82	Δ83	Δ84	Δ85
0	1	1	1	3
	645	1158	238	553
				555
				693

	ΦΙΛΙΚΟΤΗΤΑ			
Δ91	Δ92	Δ93	Δ94	Δ95
0	0	0	6	15
			99	116
			536	176
			1019	209
			1127	213
			1172	378
			1200	598
				641
				644
				703
				704
				882
				1006
				1086
				1548
				1694

Από τον πίνακα 17 προκύπτει με σαφήνεια η αντίληψη που έχουν οι 84 ερωτηθέντες της συγκεκριμένης κλάσης σχετικά με την εικόνα της Θεσσαλονίκης, ως προς τα έξι κριτήρια που χρησιμοποιήθηκαν στην έρευνα. Π.χ δεν τους άρεσε ή ήταν αδιάφοροι για την Ελληνική κουζίνα (Δ_{51},Δ_{52} και Δ_{53} ποσοστό 13/84=15,48%), ενώ υπερθεμάτιζαν για την φιλικότητα των ντόπιων (Δ_{94} και Δ_{95} 21/84=25%), όπως επίσης ήταν τελείως αρνητική για την νυκτερινή ζωή της πόλης σε ποσοστό 17,86%

Συμπέρασμα

1. Με τα διαγράμματα Καραπιστόλη απεικονίζεται παραστατικά ο τριδιάστατος υποχώρος στο επίπεδο, ώστε ο ερευνητής να μην χρειάζεται να πραγματοποιήσει ταξινόμηση , ώστε να μην συγχέει γειτονικά σημεία τα οποία ανήκουν σε διαφορετικούς υποχώρους του R^3

2. Με την μηχανή μάθησης SVM αξιολογείται αντικειμενικά κάθε μορφή ταξινόμησης με οποιαδήποτε μετρική και αν δημιουργήθηκε, βοηθώντας τον ερευνητή στη σύγκριση των αποτελεσμάτων μεταξύ δύο διαφορετικών μεθόδων ταξινόμησης, αλλά και στην διαπίστωση της ομοιογένειας των κλάσεων κάθε ταξινόμησης**.**

3. Κρίνεται σκόπιμο, όταν πρόκειται να ερευνηθεί η σύνδεση των αντικειμένων με συγκεκριμένες μεταβλητές, αφενός να χρησιμοποιηθεί το σύνολο των παραγόντων που προκύπτουν μετά την Παραγοντική Ανάλυση των Αντιστοιχιών, αφετέρου να χρησιμοποιηθεί η προτεινόμενη διαδικασία ταξινόμησης με την ονομασία KARAP, αφού ο ερευνητής εντοπίζει την μοναδικότητα της σύνδεσης των αντικειμένων εντός των κλάσεων με κάθε διαβάθμιση των μεταβλητών, ερμηνεύοντας με ευκολότερο τρόπο την συμπεριφορά του συνόλου των αντικειμένων.

4.Η μέθοδος KARAP εξασφαλίζει στον ερευνητή συμπαγείς κλάσεις, ως προς την μοναδικότητα της συμπεριφοράς των αντικειμένων κάθε συστάδας μεταβλητών, αφού περιλαμβάνει μόνο αντικείμενα των οποίων το προφίλ συνδέεται με συγκεκριμένες μεταβλητές της κάθε συστάδας.

Στη συνέχεια θα παρουσιαστεί μία ΝΕΑ μέθοδος ταξινόμησης με την ονομασία BENKAR, η οποία αξιολογεί τα αποτελέσματα κυρίως της Ανιούσας Ιεραρχικής Ταξινόμησης με την μέθοδο FACOR, βασική μέθοδο ταξινόμησης της Γαλλικής Σχολής.

ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΑΝΙΟΥΣΑΣ ΙΕΡΑΡΧΙΚΗΣ ΤΑΞΙΝΟΜΗΣΗΣ ΜΕΘΟΔΟΣ BENKAR

Περίληψη

Με αυτή την εργασία προτείνεται μια πρωτότυπη μέθοδος αξιολόγησης της Ανιούσας Ιεραρχικής Ταξινόμησης η οποία δημιουργείται μετά από ανάλυση ενός πίνακα δεδομένων Τ(n,p) με την Παραγοντική Ανάλυση των Αντιστοιχιών, χρησιμοποιώντας την διαδικασία FACOR και τον αλγόριθμο Ward.

Η δημιουργία μιας ανιούσας ιεραρχικής ταξινόμησης εστιάζει στον εντοπισμό, πάνω στον κάθε παραγοντικό άξονα χωριστά, της μικρότερης απόστασης κάθε κέντρου νέας κλάσης που πρόκειται να συνενωθεί, από τα κέντρα των κλάσεων που ήδη έχουν δημιουργηθεί, αφού προηγουμένως κάθε κέντρο κλάσης έχει σταθμιστεί ανάλογα με το βάρος των «αντικειμένων» που περιλαμβάνει η κάθε κλάση.

Όταν όμως σε μία κλάση περιλαμβάνονται «αντικείμενα» με ακραίες τιμές, τότε η ομοιογένεια της κλάσης το οποίο είναι ζητούμενο σε κάθε ταξινόμηση, λόγω της στάθμισης παραμορφώνεται σχετικά με τις τιμές των «αντικειμένων» που περιλαμβάνει, γι' αυτό κρίνεται σκόπιμο να αξιολογηθεί η συνολική ομοιογένεια των κλάσεων ως προς ένα καθορισμένο πλήθος κλάσεων της ιεραρχίας.

Γενικά

Μία ανιούσα ιεραρχική ταξινόμηση των «αντικειμένων» ενός συνόλου Ι με πληθάριθμο

card(I)=n, είναι μία διαδικασία που παράγει μια ακολουθία διαμελισμών του αρχικού συνόλου σε υποσύνολα μη κενά και ξένα ανά δύο μεταξύ τους, τις λεγόμενες **κλάσεις**, τη μία μέσα στην άλλη, συνενώνοντας κάθε φορά δύο μόνο κλάσεις οι οποίες βάσει κάποιας μετρικής παρουσιάζουν σε κάθε βήμα ομαδοποίησης την μικρότερη απόσταση.

Απ' ότι γίνεται αντιληπτό στόχος της ανιούσας ιεραρχικής ταξινόμησης είναι να ομαδοποιήσει το σύνολο των στατιστικών μονάδων ενός πληθυσμού σ' ένα περιορισμένο πλήθος ομοιογενών κλάσεων, ως προς την συμπεριφορά ορισμένων μεταβλητών, λαμβάνοντας υπόψη το σύνολο των μεταβλητών, ώστε κάθε μία να διαφέρει από τις άλλες, όσο το δυνατόν περισσότερο.

Οι κλάσεις δημιουργούνται βάσει ενός αντικειμενικού αλγορίθμου, πέρα από τις υποκειμενικές μεθόδους που μπορεί να αναπτύξει κάθε ερευνητής. Λέμε αντικειμενικό αλγόριθμο γιατί η ομαδοποίηση των στατιστικών μονάδων γίνεται χωρίς καμιά a priori υπόθεση στον αρχικό πίνακα δεδομένων και βάσει μιας συγκεκριμένης μετρικής.

Βέβαια ένας πίνακας λ.χ που περιέχει βαθμολογίες διαφόρων κριτηρίων μπορεί να περιλαμβάνει και ακραίες τιμές, οι οποίες υποχρεωτικά θα ληφθούν υπόψη στη διαδικασία ταξινόμησης των «αντικειμένων».

Αποτέλεσμα αυτού το γεγονότος είναι οι τιμές των «αντικειμένων» στα διάφορα κριτήρια, που περιλαμβάνονται σε μια συγκεκριμένη κλάση, να είναι αρκετά διαφορετικές μεταξύ τους, κάτι που αναιρεί σε κάποιο βαθμό την ομοιογένεια των απαντήσεων της κλάσης. Ακριβώς αυτή την ομοιογένεια των κλάσεων μιας Ανιούσας Ιεραρχικής Ταξινόμησης επιδιώκει να αξιολογήσει η προτεινόμενη μέθοδος.

Δημιουργία της Ανιούσας Ιεραρχικής Ταξινόμησης

Έστω ο πίνακας Τ(nxp) με n γραμμές και p στήλες. Σε κάθε ταξινόμηση ανεξάρτητα με ποια μετρική πραγματοποιείται, δημιουργούνται κλάσεις που κάθε μια περιέχει ένα συγκεκριμένο πλήθος «αντικειμένων», το οποίο παριστάνουν οι n γραμμές του πίνακα δεδομένων. Έτσι λ.χ σ' ένα ερωτηματολόγιο έρευνας αγοράς σε κάθε γραμμή του πίνακα αντιστοιχούν οι απαντήσεις του κάθε ερωτηθέντα στο σύνολο των p ερωτημάτων που του ετέθησαν.

Ως γνωστόν σε μία ταξινόμηση με την μέθοδο FACOR, χρησιμοποιείται ο αλγόριθμος του Ward, όταν λοιπόν περνάμε από ένα διαμελισμό με λ+1 κλάσεις σ' ένα άλλο διαμελισμό που έχει λ κλάσεις, συγχωνεύοντας δύο κλάσεις σε μία με κριτήριο την μείωση της διαταξικής αδράνειας σύμφωνα με το θεώρημα του Huggens. Επί πλέον θεωρούμε ότι όλα τα στοιχεία συγκεντρώνονται στο **κέντρο βάρους** της κλάσης, το οποίο **σταθμίζεται** με το βάρος των στοιχείων της k κλάσης, το οποίο αποτελεί το βαρύκεντρο της κλάσης

Η συνένωση δύο παρατηρήσεων ή δύο κλάσεων σε μία κλάση δημιουργεί αυτό που ονομάζουμε **κόμβο** της ιεραρχίας. Ο κάθε κόμβος της ιεραρχίας συμβολίζει το κέντρο βάρους των «αντικειμένων» που συμμετέχουν σ' αυτόν, η δε περιγραφή της ταξινόμησης γίνεται με το δενδρόγραμμα, του οποίου οι κόμβοι συμβολίζουν τις υποδιαιρέσεις του πληθυσμού.

Αν δεν ενδιαφερόμαστε για την συνολική ιεραρχία των n «αντικειμένων», αλλά μόνο για ένα περιορισμένο αριθμό k κλάσεων, δεν έχουμε παρά να πάρουμε μία **«τομή»** του δενδρογράμματος στο επίπεδο ε_1, δηλαδή να **"κόψουμε"** το δενδρόγραμμα με μία ευθεία γραμμή, στο σημείο όπου οι κλάδοι που απομένουν να ικανοποιούν τον αριθμό των k κλάσεων που επιθυμούμε να διατηρήσουμε.

Η μέθοδος BENKAR

Ως γνωστό μέχρι σήμερα μέθοδοι αξιολόγησης μιας ταξινόμησης μπορούν να πραγματοποιηθούν με διαδικασίες που προβλέπουν είτε την χρήση των νευρωνικών δικτύων, είτε χρησιμοποιώντας ταξινομητές μηχανικής μάθησης, οι οποίοι δεν αποδίδουν πιθανότητες αλλά μόνο εκτίμηση της επίδοσης μάθησης για το αποτέλεσμα που προκύπτει.

Με την προτεινόμενη μέθοδο, χρησιμοποιούνται εκτός από τις βασικές αρχές της Παραγοντικής Ανάλυσης των Αντιστοιχιών και των ιδιοτήτων του Ευκλείδειου διανυσματικού χώρου R^n, αποδίδεται η κατανομή πιθανοτήτων των αντικειμένων να ανήκουν σε συγκεκριμένες κλάσεις της ιεραρχίας..

Ειδικότερα οι k κόμβοι (δηλαδή τα k κέντρα των κλάσεων) μιας συγκεκριμένης τυπολογίας της ιεραρχίας, δημιουργούνται αφού πρώτα αθροίσουμε για κάθε στήλη τις τιμές των γραμμών του πίνακα T(n,p) που ανήκουν σε κάθε κλάση και στη συνέχεια τις k κλάσεις τις θεωρήσουμε ως νέες γραμμές, δημιουργώντας ένα επαυξημένο πίνακα T(n+k,p) τον οποίο αναλύουμε με την Παραγοντική Ανάλυση των Αντιστοιχιών, οπότε υπολογίζουμε τις συντεταγμένες τους πάνω στους p-1 παραγοντικούς άξονες.

Χρησιμοποιώντας τις συντεταγμένες του συνόλου των σημείων του νέφος N(I) των γραμμών, του νέφους N(J) των στηλών και των k κόμβων της ταξινόμησης πάνω στους p-1 παραγοντικούς άξονες, που προκύπτουν από την ανάλυση του πίνακα T(n+k,p), δημιουργείται μία ορθοκανονική βάση στον χώρο $R^{(p-1)}$, όπου τοποθετούνται οι p μεταβλητές, οι n στατιστικές μονάδες και τα κέντρα των k κλάσεων, στις πραγματικές τους θέσεις, απ' όπου αντλείται το σύνολο της πληροφόρησης που παρέχει ο πίνακας δεδομένων.

Ακολούθως με την χρήση της Ευκλείδειας μετρικής μπορούμε να υπολογίσουμε τις αποστάσεις της κάθε στατιστικής μονάδας από τα k κέντρα των κλάσεων, όπως προτείνει και ο αλγόριθμος του Ward. Τις k αποστάσεις της κάθε στατιστικής μονάδας τις μετατρέπουμε σε k πιθανότητες, όπου η πιο μικρή από τις k αποστάσεις, αντιστοιχεί στη μεγαλύτερη πιθανότητα που έχει η στατιστική μονάδα να είναι πλησίον του κέντρου της κλάσης με την μικρότερη απόσταση, όπου a priori στη μικρότερη αυτή απόσταση δεν αντιστοιχεί πάντοτε η κλάση που προσδιόρισε η διαδικασία της ταξινόμησης βάσει του αλγορίθμου του Ward.

Τούτου διότι, όταν σε μία κλάση περιλαμβάνονται στατιστικές μονάδες με ακραίες τιμές η **ομοιογένεια** της κλάσης αυτής, σχετικά με τις τιμές των άλλων στατιστικών μονάδων που περιλαμβάνει αλλοιώνεται, καθόσον το κέντρο της κάθε κλάσης, όπως προαναφέρθηκε, σταθμίζεται με το βάρος των στοιχείων της k κλάσης σε κάθε βήμα της δημιουργίας ενός κόμβου της ιεραρχίας.

Γι' αυτό κρίνεται σκόπιμο να αξιολογηθεί η ορθή τοποθέτηση των στατιστικών μονάδων ως προς ένα καθορισμένο πλήθος κλάσεων της ταξινόμησης χρησιμοποιώντας τις μέγιστες πιθανότητες που προκύπτουν από την μετατροπή των ελάχιστων αποστάσεων κάθε στατιστικής μονάδος από τα κέντρα των κλάσεων, βάσει της Ευκλείδειας μετρικής.

Η μετατροπή, των p-1 αποστάσεων των n σημείων-γραμμών από τα κέντρα των k κλάσεων σε αντίστοιχες πιθανότητες, προκύπτει από την σχέση:

$$P(i,j) = tdis(i,j) / \sum_{j=1}^{k} tdis(i,j) \text{ για } \quad \text{κάθε} \quad i \in n \quad j \in k$$

με

$$tdis(i,j) = 1/\left[dis(i,j)\right]^2 \text{ για } \quad \text{κάθε} \quad i \in n \quad j \in k$$

όπου $dis(i,j)$ η απόσταση κάθε $i \in n$ από κάθε κέντρο κλάσης $j \in k$

Ακολούθως σχηματίζεται η κατανομή των n μέγιστων πιθανοτήτων σε m ίσες τάξεις, απ' όπου προκύπτει η ζητούμενη αξιολόγηση της αρχικής ταξινόμησης με την μέθοδο FACOR.

Μελετώντας την κατανομή των μέγιστων πιθανοτήτων, αν η αθροιστική συχνότητα (η οποία μετατρέπεται σε ποσοστό) των δύο τελευταίων τάξεων, που προσδιορίζει το πλήθος των στατιστικών μονάδων τα οποία κατατάχθηκαν με τις δύο μεθόδους στις ίδιες κλάσεις είναι σχετικά μικρό και αν ακόμη το ποσοστό που προσδιορίζει το εύρος των δύο τελευταίων τάξεων είναι αρκούντως ικανοποιητικό, η ταξινόμηση σε k ομοιογενείς κλάσεις θεωρείται ότι δεν είναι ικανοποιητική, επειδή η ασυμφωνία στην κατάταξη των αντικειμένων στις k κλάσεις υποδηλώνει ότι οι k κλάσεις περιέχουν ανομοιογενείς στατιστικές μονάδες ως προς τις τιμές των p κριτηρίων.

Αν βέβαια το ποσοστό των αντικειμένων των δύο τελευταίων τάξεων είναι ικανοποιητικό, σε συνάρτηση με το ποσοστό που προσδιορίζει το εύρος των δύο τελευταίων τάξεων της κατανομής πιθανοτήτων, καθορίζουν την αξιολόγηση της Ανιούσας Ιεραρχικής Ταξινόμησης με την μέθοδο FACOR.

Εφαρμογή της μεθόδου BENKAR

Για την εφαρμογή της προτεινόμενης μεθόδου θα χρησιμοποιηθεί ένα συγκεκριμένο ερωτηματολόγιο ποιοτικών μεταβλητών (για την μέτρηση των οποίων χρησιμοποιήθηκε η 5βάθμια κλίμα Likert, όπου το 5 αφορούσε την άριστη εντύπωση), στο οποίο απάντησαν 1721 άτομα. Ένα τμήμα του ερωτηματολογίου αφορούσε έξι ερωτήσεις σχετικά με το πώς βαθμολογούν οι ξένοι επισκέπτες α) τα αξιοθέατα της πόλης της Θεσσαλονίκης β) την Ελληνική κουζίνα γ) την νυχτερινή ζωή της πόλης δ) το αρχιτεκτονικό της στυλ ε) την ασφάλειά της και στ) την φιλικότητα των ντόπιων.

Στα δεδομένα που προέκυψαν εφαρμόστηκε η Ανιούσα Ιεραρχική Ταξινόμηση (-CAH-) με την διαδικασία FACOR. Για την αξιολόγηση της ταξινόμησης και με κριτήριο διαμελισμού το λ_r προκρίθηκε η τομή του δενδρογράμματος σε πέντε κλάσεις. Τα στοιχεία αφορούν ξένους επισκέπτες της Θεσσαλονίκης και τα δεδομένα περιέχονται στην έρευνα που διεξήχθη στα πλαίσια του προγράμματος ΑΡΧΙΜΗΔΗΣ ΙΙΙ με τίτλο «Τεχνολογίες Ανάλυσης Δεδομένων και Διαχείρισης Γνώσης στο σχεδιασμό τουριστικών προϊόντων»

Οι έξι μεταβλητές παρίστανται αντιστοίχως ως εξής : Δ4,Δ5,Δ6,Δ7,Δ8,Δ9. Με δεδομένη την ταξινόμηση των 1721 ατόμων με την διαδικασία FACOR ο πίνακας 1 παρουσιάζει τις απαντήσεις τους και τις πέντε κλάσεις στις οποίες ανήκουν οι ερωτώμενοι.

Πίνακας 1: Τιμές των έξι μεταβλητών και οι πέντε κλάσεις στις οποίες ανήκουν οι ερωτώμενοι μετά την ταξινόμηση με την διαδικασία FACOR

IND	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9	Class FACOR
I1	4	4	0	4	5	5	3
I2	5	4	5	5	4	5	4
I3	3	4	3	1	2	3	2
.							.
1690	4	3	2	2	2	2	5
‘							.
1719	5	5	5	2	2	5	2
1720	5	4	4	2	5	3	5
1721	5	5	4	3	4	5	5

Ο πίνακας 2 παρουσιάζει τον επαυξημένο πίνακα που δημιουργείται στο βήμα 3.

Πίνακας 2: Τμήμα του επαυξημένου πίνακα T(n+k,p)

IND	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9
I1	4	4	0	4	5	5
I2	5	4	5	5	4	5
.						
I1690	4	3	2	2	2	2
‘						
I1721	5	5	4	3	4	5
K1	398	212	224	387	378	421
K2	727	750	810	592	391	620
K3	548	545	35	482	430	521
K4	3679	3612	3192	3600	3403	3437
K5	1580	1883	1907	1548	1788	1963

Ακολουθώντας τα βήματα 4 και 5 δημιουργείται, ο πίνακας T3(n,2k) ο οποίος παρουσιάζει για κάθε σημείο-γραμμή τις αποστάσεις K_i (i=1,…5) των σημείων από τα κέντρα των πέντε κλάσεων και τις αντίστοιχες πιθανότητες P_i (i=1,…5)

Πίνακας 3:Τμήμα του πίνακα αποστάσεων K_i των σημείων-γραμμών με τις αντίστοιχες πιθανότητες

IND	K1	P1	K2	P2	K3	P3	K4	P4	K5	P5
I1	12,769	0,077	41,496	0,007	3,816	0,865	20,147	0,031	25,590	0,019
I2	5,770	0,032	3,719	0,078	21,824	0,002	1,264	0,672	2,229	0,216
I3	26,919	0,024	5,980	0,485	29,984	0,019	10,966	0,144	7,280	0,327
.	.	.	.	.	.	.	.	.	.	.
I1690	11,800	0,051	4,899	0,298	12,839	0,043	3,795	0,497	8,065	0,110
‘	.	.	.	.	.	.	.	.	.	.
I1719	25,445	0,019	4,003	0,755	35,039	0,010	12,399	0,079	9,378	0,138
I1720	9,281	0,003	5,912	0,006	21,472	0,001	1,164	0,162	0,514	0,829
I1721	9,253	0,015	4,365	0,069	15,999	0,005	2,087	0,302	1,470	0,609

Στη συνέχεια στο βήμα 6 προκύπτει ο πίνακας T4(n,3) ο οποίος παρουσιάζει για κάθε σημείο-γραμμή στη $1^η$ στήλη την ταξινόμησή του με την διαδικασία FACOR, στη $2^η$ στήλη την νέα ταξινόμησή του με την διαδικασία της ελάχιστης απόστασης και στη $3^η$ στήλη την μέγιστη πιθανότητα να ανήκει στην κλάση που αντιστοιχεί στην ελάχιστη απόσταση.

Πίνακας 4: Ταξινόμηση με διαδικασία FACOR και την Ευκλείδεια μετρική στο χώρο R^5

IND	FACOR	DIS	maxProb
I1	3	3	0,8651
I2	4	4	0,6717
I3	2	2	0,4851
.	.	.	.
I1690	5	4	0,497
‘	.	.	.
I1720	5	5	0,829
I1721	5	5	0,6086

Από το λογισμικό MAD δίδονται τα ακόλουθα αποτελέσματα:

Τακτοποιήθηκαν στις ΙΔΙΕΣ κλάσεις :1416 άτομα

Τακτοποιήθηκαν σε ΔΙΑΦΟΡΕΤΙΚΕΣ κλάσεις : 305 άτομα

Ποσοστό καλής προσαρμογής : 82,28%

Συνεχίζοντας με το βήμα 7 έχουμε τις τρεις κατανομές σε πέντε κλάσεις που προήλθαν μετά από ταξινόμηση των 1721 ατόμων

α) με την Ανιούσα Ιεραρχική Ταξινόμηση (-CAH-)

β) με την Ευκλείδειο μετρική (DIS) στο χώρο R^5 που δημιούργησαν οι πέντε παραγοντικοί άξονες μετά την εφαρμογή της Παραγοντικής Ανάλυσης των Αντιστοιχιών (-AFC-) στον πίνακα δεδομένων Τ(1726,6)

γ) την κατανομή των μέγιστων πιθανοτήτων να ανήκουν τα 1721 άτομα στις πέντε διαφορετικές κλάσεις της Ανιούσας Ιεραρχικής Ταξινόμησης (-CAH-).

Πίνακας 5: Οι τρεις κατανομές του 7ου βήματος

CAH	n_i	f_i	DIS	n_i	f_i	Probability distribution	ni
K1	99	0,0575	**K1**	136	0,079	**T1: 0,2305 - 0,3855**	48
K2	181	0,1051	**K2**	225	0,1307	**T2: 0,3855 - 0,5408**	230
K3	129	0,0749	**K3**	142	0,0825	**T3: 0,5408 -0, 6962**	315
K4	848	0,4927	**K4**	764	0,4439	**T4: 0,6962 - 0,8515**	335
K5	464	0,2696	**K5**	454	0,2638	**T5: 0,8515 - 1,0000**	488
	1721	1		1721	1		1416

Παρατήρηση: Η ερμηνεία της κλάσης Τ5 είναι η εξής: 488 άτομα από τα 1416 που τακτοποιήθηκαν στις ίδιες κλάσεις, δηλαδή ποσοστό 34,5% έχει πιθανότητα από 0,8515 και άνω να ανήκει σε ΜΙΑ από τις πέντε κλάσεις της ταξινόμησης CAH.

Προχωρώντας στο βήμα 8 η αξιολόγηση της συγκεκριμένης ταξινόμησης των 1721 επισκεπτών με την μέθοδο BENKAR, χρησιμοποιώντας τις δύο τελευταίες τάξεις της κατανομής πιθανοτήτων είναι η εξής: 823 (=488+335) άτομα από τα 1721, δηλαδή το 47,82% του συνόλου των ερωτηθέντων έχει πιθανότητα από 0,6962 και άνω να ανήκει σε ΜΙΑ από τις πέντε κλάσεις της ταξινόμησης.

Επειδή το ποσοστό 47,82% του συνόλου των ερωτηθέντων δεν είναι ικανοποιητικό, παρά το ότι το ποσοστό 82,28% της προσαρμογής των «αντικειμένων» στις πέντε κλάσεις των δύο ταξινομήσεων είναι αρκετά υψηλό, τα δεδομένα του πίνακα Τ(1721,6) εντός των κλάσεων φαίνεται ότι δεν έχουν την απαραίτητη επιθυμητή ομοιογένεια, σε σχέση με τις τιμές των έξι μεταβλητών.

Εφαρμόζοντας στη συνέχεια την μέθοδο BENKAR για το πώς αξιολογούν οι 1721 επισκέπτες της Θεσσαλονίκης τα τρία πρώτα ερωτήματα της έρευνας Δ1=την καθαριότητα της πόλης, Δ2=τις φυσικές ομορφιές και Δ3=τις τιμές των προϊόντων και υπηρεσιών, προέκυψαν τα εξής αποτελέσματα:

Πίνακας 6: Τιμές των τριών μεταβλητών και οι κλάσεις στις οποίες ανήκουν οι ερωτώμενοι μετά την ταξινόμηση με την διαδικασία FACOR

IND	Δ1	Δ2	Δ3	Class FACOR
I1	3	5	4	5
I2	2	5	3	5
I3	1	3	4	5
.				
1690	2	3	5	5
‘				
1719	1	2	3	5
1720	1	5	2	3
1721	2	5	3	5

Από το λογισμικό MAD δίδονται τα ακόλουθα αποτελέσματα:

Τακτοποιήθηκαν στις ΙΔΙΕΣ κλάσεις :1638 άτομα

Τακτοποιήθηκαν σε ΔΙΑΦΟΡΕΤΙΚΕΣ κλάσεις : 83 άτομα

Ποσοστό καλής προσαρμογής : 95,18%

Πίνακας 7: Οι τρεις κατανομές του 7ου βήματος

CAH	n_i	f_i	DIS	n_i	f_i	Probability distribution	ni
K1	119	0,0691	**K1**	135	0,0784	**T1: 0,3700 -0, 4970**	68
K2	294	0,1708	**K2**	348	0,2022	**T2: 0,4970 - 0,6241**	201
K3	221	0,1284	**K3**	219	0,1272	**T3: 0,6241 - 0,7512**	161
K4	691	0,4015	**K4**	681	0,3957	**T4: 0,7512 - 0,8784**	254
K5	396	0,23	**K5**	338	0,1963	**T5: 0,8784 - 1,0000**	954
	1721	1		1721	1		1638

Από τον πίνακα 7 προκύπτει η εξής αξιολόγηση: 1208 άτομα από τα 1721, δηλαδή το 70,18% του συνόλου των ερωτηθέντων έχει πιθανότητα από 0,7512 και άνω να ανήκει σε ΜΙΑ από τις πέντε κλάσεις της ταξινόμησης..

Επειδή το ποσοστό 70,18% του συνόλου των ερωτηθέντων είναι ικανοποιητικό, ενώ και το ποσοστό 95,18% της προσαρμογής των «αντικειμένων» στις πέντε κλάσεις των δύο ταξινομήσεων είναι υψηλό, τα δεδομένα του πίνακα Τ(1721,3) εντός των κλάσεων αυτών μπορεί να θεωρηθούν ότι έχουν την απαραίτητη επιθυμητή ομοιογένεια, σε σχέση με τις τιμές των τριών μεταβλητών.

Συγκρίνοντας τις δύο αξιολογήσεις των ταξινομήσεων αφενός με τα κριτήρια Δ1-Δ3, αφετέρου με τα κριτήρια Δ4-Δ9, αναδεικνύεται ότι οι 1721 επισκέπτες της Θεσσαλονίκης είχαν ομοιόμορφη εικόνα της πόλης ως προς τα τρία πρώτα κριτήρια που κλήθηκαν να αξιολογήσουν, ενώ για τα υπόλοιπα έξι κριτήρια δεν παρουσίαζαν ανάλογη ομοιογένεια απαντήσεων. Αυτό μπορεί να οφείλεται επειδή οι επισκέπτες προερχόταν από 51 διαφορετικές χώρες του πλανήτη, οπότε κρίνεται φυσιολογικό στα κριτήρια Δ1 έως Δ3 να έχουν περίπου την ίδια αντίληψη, ενώ σε κάποια από τα κριτήρια Δ4 έως Δ9 να παρουσιάζουν τελείως διαφορετικές απόψεις λόγω διαφορετικών παραδόσεων που ισχύουν στον τόπο του κάθε επισκέπτη.

Εκπαίδευση των δεδομένων με την Μηχανή Διανυσμάτων Υποστήριξης –SVM-

Εκπαιδεύοντας με 20 επαναλήψεις τα δεδομένα του πίνακα 1, με την χρήση της Μηχανής SVM, διατηρώντας σε κάθε επανάληψη ένα τυχαίο δείγμα 20% των 1721 τιμών, μια φορά με τις κατατάξεις των «αντικειμένων» με την μέθοδο FACOR και την άλλη με την μέθοδο BENKAR έχουμε τα ακόλουθα αποτελέσματα:

Πίνακας 8: Ποσοστά εκμάθησης των 20 επαναλήψεων μετά την εκπαίδευση των ταξινομήσεων των δεδομένων βάσει της μεθόδου FACOR και της μεθόδου BENKAR

FACOR (20%)		BENKAR (20%)	
0,6950		0,8363	
0,6862		0,9620	
0,7595		0,7105	
0,7830		0,9415	
0,7009		0,9444	
0,7185		0,9094	
0,7859		0,7515	
0,8211		0,9532	
0,7889	76,54%	0,6462	87,88%
0,8710		0,9035	
0,8328		0,7076	
0,8182		0,9064	
0,6862		0,9474	
0,7830		0,9474	
0,7067		0,8918	
0,8065		0,9737	
0,8035		0,9386	
0,7155		0,7982	
0,7595		0,9561	
0,7859		0,9503	

Από τον πίνακα 8 διαπιστώνεται ότι η μέθοδος BENKAR υπερέχει στην σωστή ταξινόμηση των αντικειμένων. Αυτό προκύπτει επειδή το ποσοστό εκτίμησης της εκμάθησης των δεδομένων του πίνακα 1, που αφορά την ταξινόμηση με την μέθοδο BENKAR (87,88%) είναι υψηλότερο από εκείνο που προκύπτει με την Ανιούσα Ιεραρχική Ταξινόμηση με την μέθοδο FACOR (76,54%).Επί πλέον με την μέθοδο BENKAR στις 20 επαναλήψεις εκμάθησης των δεδομένων τα ποσοστά εκτίμησης πάνω από την μέση τιμή είναι κατά πολύ υψηλότερα (13 στις 20 επαναλήψεις πάνω από 90% με μέγιστη τιμή 97,37%) από τα αντίστοιχα ποσοστά εκτίμησης με την μέθοδο FACOR που η μέγιστη τιμή ανέρχεται μόλις στο 87,10%.

Με δεδομένο ότι όπως προαναφέρθηκε η ΜΔΥ δεν επιστρέφει πιθανότητες, ενώ η μέθοδος BENKAR υπολογίζει την πιθανότητα κάθε στατιστικής μονάδας να ανήκει σε μία ορισμένη κλάση, ως εκ τούτου η κατανομή των μέγιστων πιθανοτήτων που προκύπτει από την προτεινόμενη μέθοδο, μπορεί να θεωρηθεί ότι αξιολογεί αντικειμενικά την Ανιούσα Ιεραρχική Ταξινόμηση που απορρέει με την μέθοδο FACOR.

Συμπέρασμα

Η μέθοδος BENKAR αξιοποιώντας αντικειμενικά κριτήρια, όπως είναι οι συντεταγμένες των σημείων πάνω στους παραγοντικούς άξονες μετά την εφαρμογή στον πίνακα δεδομένων της Παραγοντικής Ανάλυσης των Αντιστοιχιών, αλλά και της τοποθέτησής τους βάσει αυτών των συντεταγμένων στον Ευκλείδειο διανυσματικό χώρο R^p, με την χρήση της Ευκλείδειας μετρικής, παρέχει την δυνατότητα, αντικειμενικής αξιολόγησης της ομοιογένειας των «αντικειμένων» που συμμετέχουν στη διαμόρφωση των κλάσεων της Ανιούσας Ιεραρχικής Ταξινόμησης.

Η δυνατότητα αυτή της μεθόδου BENKAR μπορεί να εφαρμοστεί σε οποιοδήποτε πίνακα Τ(n,p) του οποίου τα δεδομένα έχουν ταξινομηθεί με οποιοδήποτε κριτήριο συνένωσης, δεδομένου ότι σε κάθε περίπτωση η μέθοδος BENKAR ταξινομεί τα «αντικείμενα», αφού προηγουμένως τα τοποθετήσει σ' ένα ορθοκανονικό σύστημα συντεταγμένων R^p που δημιουργούν οι παραγοντικοί άξονες μετά την εφαρμογή της Παραγοντικής Ανάλυσης των Αντιστοιχιών στον πίνακα δεδομένων.

Η υπεροχή της μεθόδου BENKAR στην κατάταξη των «αντικειμένων» σε k κλάσεις που υπέδειξε η Ανιούσα Ιεραρχική Ταξινόμηση, διαπιστώνεται και με την χρήση της Μηχανής Διανυσμάτων Υποστήριξης, η οποία αξιολογεί την εκμάθηση των δεδομένων σε ποσοστό υψηλότερο από εκείνο που παρέχει η ίδια μηχανή για την ταξινόμηση των ίδιων δεδομένων με την Ανιούσα Ιεραρχική Ταξινόμηση, χρησιμοποιώντας την μέθοδο FACOR.

Επιπροσθέτως η μέθοδος BENKAR σε αντίθεση με την ΜΔΥ, αλλά και με κάθε άλλη μέθοδο ταξινόμησης, **υπολογίζει την πιθανότητα των «αντικειμένων» να ανήκουν στις διαμορφούμενες κλάσεις**, στοιχείο που αποτελεί βασικό πλεονέκτημα της συγκεκριμένης μεθόδου, με τελική κατάληξη την αντικειμενική αξιολόγηση της ομοιογένειας των κλάσεων, που είναι ένα από τα ζητούμενα σε κάθε ταξινόμηση.

Το γιατί βασιζόμαστε στα αποτελέσματα της **Μηχανής Διανυσμάτων Υποστήριξης ΜΔΥ** (SVM) και όχι σε μία κλασσική μέθοδο ταξινόμησης απαντά η παρακάτω εργασία της Δρος Ειρήνης Καραπιστόλη Ηλεκτρολόγο μηχανικό και μηχανικό Ηλεκτρονικών Υπολογιστών.

ΣΥΓΚΡΙΣΗ ΤΕΧΝΙΚΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΒΑΣΕΙ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΤΗΣ ΑΝΙΟΥΣΑΣ ΙΕΡΑΡΧΙΚΗΣ ΤΑΞΙΝΟΜΗΣΗΣ

Δρ. Ειρήνη Καραπιστόλη

Περίληψη

Η παρούσα εργασία διερευνά τη χρήση διαφορετικών τεχνικών επιβλεπόμενης μηχανικής μάθησης για να αξιολογήσει τα αποτελέσματα που προκύπτουν από έναν πίνακα δεδομένων στον οποίο εφαρμόστηκε η Ανιούσα Ιεραρχική Ταξινόμηση. Τα αποτελέσματα της σύγκρισης τονίζουν την καλή απόδοση των Μηχανών Υποστήριξης Διανυσμάτων (Support Vector Machines - SVM), και υποδεικνύουν ότι είναι μια πολλά υποσχόμενη τεχνική για την αποτελεσματική αξιολόγηση της ταξινόμησης των αντικειμένων με τη μέθοδο της Ανιούσας Ιεραρχικής Ταξινόμησης.

1. Εισαγωγή - Ανάλυση Δεδομένων και Μηχανική Μάθηση

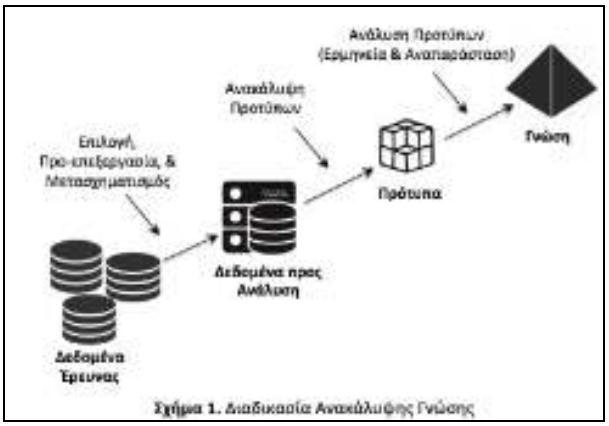
Κατά τη διάρκεια των τελευταίων δεκαετιών, υπήρξε μια απίστευτη αύξηση των δυνατοτήτων μας για παραγωγή και αποθήκευση μεγάλου όγκου δεδομένων (Big Data). Γενικά, υπάρχει ένα ανταγωνιστικό πλεονέκτημα στο να είμαστε σε ικανοί να χρησιμοποιήσουμε σωστά την αφθονία των δεδομένων που συγκεντρώνονται σήμερα. Η αποτελεσματική ανάλυση των δεδομένων που συλλέγονται μπορεί να προσφέρει σημαντικά πλεονεκτήματα όπως βελτίωση στην κατανόηση πολλών παραγωγικών διαδικασιών, σχεδιασμός καλύτερων συστημάτων, κ.ο.κ., και έχει πολύ χρήσιμες εφαρμογές, όπως καλύτερες προβλέψεις, διαγνώσεις, ταξινομήσεις, κ.ο.κ. (Gillblad et al., 2003).

Αυτός είναι και ο λόγος που η τεχνητή νοημοσύνη και ειδικότερα η μηχανική μάθηση (machine learning) έχουν αποκτήσει μεγάλο εύρος εφαρμογής τα τελευταία χρόνια. Κύριος στόχος είναι η αντιμετώπιση του προβλήματος της υπερ-πληροφόρησης (information overload) μέσω της ανάπτυξης συστημάτων τα οποία θα μπορούν να φιλτράρουν και να αναλύουν σε βάθος τον ολοένα και αυξανόμενο όγκο δεδομένων, αναζητώντας σχετική πληροφορία για τον τελικό χρήστη. Παρόλο που η διαδικασία εκμάθησης στους υπολογιστές απέχει αρκετά από τη διαδικασία εκμάθησης στους ανθρώπους, πληθώρα εφαρμογών έχουν επιτυχώς αναπτυχθεί τα τελευταία χρόνια οι οποίες χρησιμοποιούν τη μηχανική μάθηση σε διάφορους τομείς όπως για παράδειγμα την εξόρυξη δεδομένων (data mining) ή αλλιώς ανακάλυψη γνώσης (knowledge discovery) σε μεγάλες βάσεις δεδομένων.

Στο Σχήμα 1 απεικονίζονται τα βήματα μιας τυπικής διαδικασίας ανάλυσης δεδομένων και ανακάλυψης γνώσης. Η πρώτη φάση περιλαμβάνει την προετοιμασία των δεδομένων, ώστε να απομακρυνθούν τα δεδομένα που θεωρούνται θόρυβος και τα οποία μπορεί να οδηγήσουν σε λανθασμένα συμπεράσματα. Στη συνέχεια, χρησιμοποιώντας τα δεδομένα που έχουν ουσιαστική πληροφορία, εφαρμόζονται στατιστικές μέθοδοι καθώς και τεχνικές εξόρυξης δεδομένων οι οποίες αποσκοπούν στον εντοπισμό προτύπων, και συνήθως αφορούν στην εφαρμογή κανόνων συσχέτισης (association rules), αλγορίθμων ομαδοποίησης (clustering) και ταξινόμησης (classification). Στο τελικό στάδιο, αυτό της ανάλυσης προτύπων, πραγματοποιείται η μετάφραση των προτύπων που ανακτήθηκαν στο προηγούμενο βήμα σε ουσιαστική γνώση και γίνεται η αναπαράσταση της εξαγόμενης γνώσης.

Επιστρέφοντας στο αρχικό ερώτημα, και με δεδομένη την πληθώρα τεχνικών που έχουν αναπτυχθεί για να αντιμετωπίσουν επιτυχώς τις παραπάνω προκλήσεις, τίθεται το ερώτημα αν η ανάλυση δεδομένων, που είναι το πρώτο στάδιο της διαδικασίας εξόρυξης γνώσης, έχει κάνει καλά τη δουλειά της.

Στα πλαίσια της παρούσας εργασίας, προτείνεται μια εναλλακτική μέθοδος για την αξιολόγηση των αποτελεσμάτων που προκύπτουν από έναν πίνακα δεδομένων στον οποίο εφαρμόστηκε η Ανιούσα Ιεραρχική Ταξινόμηση (Hierarchical Cluster Analysis, HCA) μέσω της σύγκρισης των αποτελεσμάτων εκπαίδευσης πέντε ευρέως διαδεδομένων τεχνικών μηχανικής μάθησης (Lim *et al.*, 2000). Οι τεχνικές αυτές είναι: τα δένδρα απόφασης (Decision Trees, DT), τα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks, ANN), οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines, SVM), ο αφελής πιθανοτικός ταξινομητής Bayes (Naive Bayes, NB), και ο ταξινομητής των K-πλησιέστερων γειτόνων (K2 Nearest Neighbor, KNN). Εξ΄ όσων γνωρίζω, η μεθοδολογία αυτή είναι η πρώτη που εφαρμόζει πολλαπλές τεχνικές μηχανικής μάθησης για να αξιολογήσει τα αποτελέσματα κατάταξης που προκύπτουν από έναν πίνακα δεδομένων στον οποίο εφαρμόστηκε ταξινομήσαμε τη διαδικασία VACOR.



Σχήμα 1. Διαδικασία Ανακάλυψης Γνώσης

2. Μεθοδολογία – Ανάλυση των Δεδομένων Έρευνας

Τα στοιχεία της παρούσας εργασίας αφορούν ξένους επισκέπτες της Θεσσαλονίκης και τα δεδομένα περιέχονται στην έρευνα που διεξήχθη στα πλαίσια υλοποίησης του προγράμματος ΑΡΧΙΜΗΔΗΣ ΙΙΙ με τίτλο «Τεχνολογίες Ανάλυσης Δεδομένων και Διαχείρισης Γνώσης στο σχεδιασμό τουριστικών προϊόντων» (http://www.mkt.teithe.gr/dankman/). Η έρευνα πραγματοποιήθηκε από τις 15-4-2013 έως τις 15-10-2013 σε 1721 ξένους επισκέπτες της Θεσσαλονίκης, και είχε ως στόχο να συλλέξει πληροφορίες σχετικά με την αντιληπτή εικόνα της πόλης της Θεσσαλονίκης, όπως αυτή προκύπτει κατά τη διάρκεια της επίσκεψής τους, να μελετήσει τις προσδοκίες τους και τους παράγοντες απόφασης να επισκεφτούν την πόλη, και τέλος, να αναλύσει τις ανάγκες τους και τους παράγοντες που συμβάλλουν στην ικανοποίηση τους.

Η επεξεργασία των στοιχείων έγινε με την εφαρμογή της Ανιούσας Ιεραρχικής Ταξινόμησης βάσει της διαδικασίας VACOR (Benzecri, 1992). Για την αξιολόγηση της ταξινόμησης προκρίθηκε η τομή του δενδρογράμματος σε πέντε (5) κλάσεις.

Για την εφαρμογή της προτεινόμενης σύγκρισης των ταξινομητών επιβλεπόμενης μηχανικής μάθησης χρησιμοποιείται ένα συγκεκριμένο ερωτηματολόγιο με δύο ενότητες που αφορούν: (Α) **στις αιτίες** που προκάλεσαν τον επισκέπτη να επιλέξει τον συγκεκριμένο

προορισμό με 15 ερωτήματα(κριτήρια), και (Β) **στην εικόνα της Θεσσαλονίκης** που δημιουργήθηκε κατά την επίσκεψή τους όπου τέθηκαν εννέα ποιοτικά κριτήρια. Να σημειωθεί ότι για τη μέτρηση των κριτηρίων χρησιμοποιήθηκε η 5βάθμια κλίμακα Likert, όπου το 5 αφορούσε στην άριστη εντύπωση.

Ο Πίνακας 1 παρουσιάζει τις απαντήσεις των ερωτώμενων στις ερωτήσεις της ενότητας «Αιτίες που επιλέχθηκε αυτός ο προορισμός» καθώς και τις πέντε κλάσεις (Κ1, Κ2, Κ3, Κ4 και Κ5) στις οποίες ανήκουν οι ερωτώμενοι. Να σημειωθεί ότι η τιμή 1 δηλώνει θετική απάντηση στο αντίστοιχο κριτήριο.

Όσον αφορά στην ενότητα Β του ερωτηματολογίου «Εικόνα της Θεσσαλονίκης», ένα τμήμα της ενότητας αυτής, αφορά σε τρεις ερωτήσεις σχετικά με το πώς αξιολογούν οι 1721 επισκέπτες της Θεσσαλονίκης α) την καθαριότητα της πόλης (μεταβλητή Δ1), β) τις φυσικές ομορφιές (μεταβλητή Δ2), και γ) τις τιμές των προϊόντων και υπηρεσιών (μεταβλητή Δ3). Με δεδομένη την ταξινόμηση των 1721 ατόμων με τη διαδικασία VACOR, ο Πίνακας 1 παρουσιάζει τις απαντήσεις τους, καθώς και τις πέντε κλάσεις στις οποίες ανήκουν οι ερωτώμενοι.

Πίνακας 1.Οι τιμές των 15 μεταβλητών (Γ1-Γ15) και οι κλάσεις στις οποίες ανήκουν οι ερωτώμενοι μετά την ταξινόμηση με τη διαδικασία VACOR

IND	Γ1 –Φήμη	Γ2 – Φυσική Ομορφιά	Γ3 – Γνώμη Φίλων	Γ4 – Το κλίμα	Γ5 – Ιστορία της περιοχής	Γ6 – Μπάνιο & Παραλία	Γ7 – Επισκέψεις Μουσείων	Γ8 – Νυχτερινή Διασκέδαση	Γ9 – Εκδρομές στη Φύση	Γ10 – Τοπική Κουζίνα	Γ11 – Οργάνωση-Υποδομές	Γ12 – Τοπικές Μεταφορές	Γ13 – Ασφάλεια	Γ14 – Φυσικό Περιβάλλον	Γ15 – Life Style	Κλάση VACOR
I1	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	5
I2	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	3
I3	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	5
.																.
1008	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	4
.																.
1719	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1
1720	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	5
1721	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	5

Το άλλο τμήμα της ενότητας Β αφορά σε έξι ερωτήσεις σχετικά με το πώς βαθμολογούν οι ξένοι επισκέπτες α) τα αξιοθέατα της πόλης της Θεσσαλονίκης, β) την Ελληνική κουζίνα, γ) τη νυχτερινή ζωή της πόλης, δ) το αρχιτεκτονικό της στυλ, ε) την ασφάλειά της, και στ) τη φιλικότητα των ντόπιων. Οι έξι μεταβλητές παρίστανται αντιστοίχως ως εξής: Δ4, Δ5, Δ6, Δ7, Δ8 και Δ9. Με δεδομένη την ταξινόμηση των 1721 ατόμων με τη διαδικασία VACOR, ο Πίνακας 2 και 3 παρουσιάζει τις απαντήσεις τους, καθώς και τις πέντε κλάσεις στις οποίες ανήκουν οι ερωτώμενοι.

Πίνακας 2.Οι τιμές των τριών μεταβλητών Δ1-Δ3 και οι κλάσεις στις οποίες ανήκουν οι ερωτώμενοι μετά την ταξινόμηση με τη διαδικασία VACOR

IND	Δ1	Δ2	Δ3	Κλάση VACOR
I1	3	5	4	5
I2	4	3	3	4
I3	1	3	4	5
.				.
1008	3	3	5	2
.				.
1719	1	2	3	5
1720	1	5	2	3
1721	2	5	3	5

Πίνακας 3.Οι τιμές των έξι μεταβλητών Δ4-Δ9 και οι πέντε κλάσεις στις οποίες ανήκουν οι ερωτώμενοι μετά την ταξινόμηση με τη διαδικασία VACOR

IND	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9	Κλάση VACOR
I1	4	4	0	4	5	5	3
I2	5	4	5	5	4	5	4
I3	3	4	3	1	2	3	2
.							.
1008	4	5	4	4	4	5	5
.							.
1719	5	5	5	2	2	5	2
1720	5	4	4	2	5	3	5
1721	5	5	4	3	4	5	5

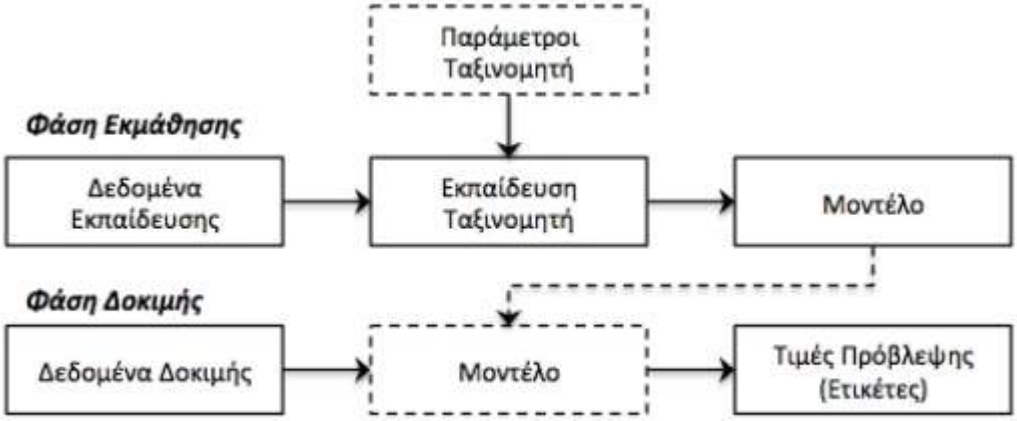
3.Τεχνικές Μηχανικής Μάθησης

Οι τεχνικές μηχανικής μάθησης διακρίνονται σε αυτές που πραγματοποιούνται χωρίς επίβλεψη (unsupervised learning), και σε αυτές που πραγματοποιούνται με επίβλεψη (supervised learning). Στην πρώτη κατηγορία, ο αλγόριθμος μηχανικής μάθησης κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων χωρίς να γνωρίζει επιθυμητές εξόδους για το σύνολο εκπαίδευσης. Χαρακτηριστικό παράδειγμα ανεπίβλεπτης μάθησης αποτελεί η εύρεση κανόνων συσχέτισης (association rules) μεταξύ των τιμών των χαρακτηριστικών στα διανύσματα εκμάθησης.

Ωστόσο, το μεγαλύτερο τμήμα της ερευνητικής δραστηριότητας στο χώρο της μηχανικής μάθησης αφορά στη μάθηση με επίβλεψη ή αλλιώς, *επιβλεπόμενη μάθηση*, τυπικό παράδειγμα της οποίας είναι τα προβλήματα κατηγοριοποίησης ή ταξινόμησης (*classification*). Σε ένα πρόβλημα ταξινόμησης, όπως αυτά με τα οποία ασχολείται η παρούσα εργασία, δίνεται ένα σύνολο στοιχείων $(x_1,x_2,\ldots\ldots,x_n,y)$ όπου $x_1,x_2,\ldots\ldots,x_n$, είναι το σύνολο εκπαίδευσης (διάνυσμα εισόδου), και y είναι μια τιμή κλάσης η οποία περιγράφει την κατηγορία στην οποία ανήκει κάθε στοιχείο του συνόλου δεδομένων.

Το πρόβλημα στην περίπτωση της ταξινόμησης είναι να αναγνωρισθεί η κατηγορία ενός νεοεισερχόμενου, στο σύνολο δεδομένων, στοιχείου (βλ. Σχήμα 2).

Στη συνέχεια, δίνεται μια σύντομη περιγραφή των πέντε (5) ταξινομητών (*classifiers*) που χρησιμοποιήσαμε για τους σκοπούς της ανάλυσης μας.



Σχήμα 2. Διαδικασία Ταξινόμησης

3.1 Ταξινομητές Επιβλεπόμενης Μηχανικής Μάθησης

3.1.1 Δένδρα Αποφάσεων Ταξινόμησης

Μια ευρέως χρησιμοποιούμενη μέθοδος μηχανικής μάθησης είναι και αυτή που βασίζεται σε δέντρα απόφασης (*decision trees, DT*). Σύμφωνα με τη μέθοδο αυτή επιχειρείται η προσέγγιση μιας κατηγορικής συνάρτησης στόχου, ακολουθώντας την τεχνική του 'διαίρει και βασίλευε'. Ο χώρος του προβλήματος χωρίζεται σε περιοχές από στιγμιότυπα που φέρουν την ίδια τιμή ως προς κάποιο χαρακτηριστικό, και η διαδικασία επαναλαμβάνεται αναδρομικά, αναπαριστώντας με τον τρόπο αυτό το παραγόμενο μοντέλο ως δέντρο απόφασης. Ο πιο γνωστός αλγόριθμος για την κατασκευή δέντρων απόφασης είναι ο C4.5 (Quinlan, 1993). Ο C4.5 είναι ο πρώτος αλγόριθμος που χρησιμοποίησε για κριτήριο διαχωρισμού το κέρδος πληροφορίας. Πρόσφατες έρευνες που συγκρίνουν τα δέντρα απόφασης με άλλους αλγορίθμους μηχανικής μάθησης, δείχνουν ότι ο C4.5 έχει έναν πολύ καλό συνδυασμό ακρίβειας και ταχύτητας εκμάθησης.

3.1.2 Τεχνητά Νευρωνικά Δίκτυα

Τα τεχνητά νευρωνικά δίκτυα (*artificial neural networks, ANN*) αποτελούν μια σημαντική μέθοδο

μοντελοποίησης σύνθετων προβλημάτων πρόβλεψης με μεγάλο αριθμό εξαρτημένων μεταβλητών. Αρχικά προτάθηκαν ως ένα μαθηματικό μοντέλο προσομοίωσης της πολύπλοκης λειτουργίας του ανθρώπινου εγκεφάλου (Haykin, 1998). Σε αναλογία με το βιολογικό νευρώνα του εγκεφάλου, ο τεχνητός νευρώνας (artificial neuron) είναι η δομική μονάδα του νευρωνικού δικτύου. Υπάρχουν δύο είδη νευρώνων, οι νευρώνες εισόδου και οι υπολογιστικοί νευρώνες. Οι νευρώνες εισόδου δεν υπολογίζουν τίποτα, μεσολαβούν ανάμεσα στις εισόδους του δικτύου και τους υπολογιστικούς νευρώνες. Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν τις εισόδους τους με τα συναπτικά βάρη και υπολογίζουν το άθροισμα του γινομένου. Το άθροισμα που προκύπτει είναι το όρισμα της συνάρτησης μεταφοράς, η οποία μπορεί να είναι: βηματική (step), γραμμική (linear), μη γραμμική (non-linear), ή στοχαστική (stochastic).

3.1.3 Μηχανές Διανυσμάτων Υποστήριξης

Οι μηχανές διανυσμάτων υποστήριξης (*support vector machines,* SVM) (Cortes & Vapnik, 1995), είναι ένα είδος συγκερασμού γραμμικών μοντέλων και εκπαίδευσης βασισμένη σε στιγμιότυπα. Στόχος αυτής της κατηγορίας αλγορίθμων είναι η επιλογή ενός μικρού αριθμού στιγμιότυπων εκπαίδευσης από κάθε κλάση, των διανυσμάτων υποστήριξης (support vectors), που συνορεύουν στο χώρο του προβλήματος με στιγμιότυπα άλλων κλάσεων. Τα επιλεγμένα στιγμιότυπα χρησιμοποιούνται για την κατασκευή μιας γραμμικής συνάρτησης διάκτισης (discriminant function), ικανής να τα διαχωρίσει όσο το δυνατόν περισσότερο (Cristianini & Shawe-Taylor, 2000).

3.1.4 Πιθανοτικοί Ταξινομητές

Η συλλογιστική κατά Bayes (Bayesian reasoning) βασίζεται στη πιθανοτική θεωρία κατηγοριοποίησης, όπου στόχος είναι να κατηγοριοποιηθεί ένα δείγμα X σε μια από τις δεδομένες κατηγορίες c_1 ,c_2, ,,,,c_n χρησιμοποιώντας ένα μοντέλο πιθανότητας που ορίζεται σύμφωνα με τη θεωρία του Bayes. Πρόκειται επομένως για κατηγοριοποιητές που κάνουν αποτίμηση πιθανοτήτων και όχι πρόβλεψη. Μια ευρέως χρησιμοποιούμενη μέθοδος κατά Bayes είναι ο αφελής ταξινομητής Bayes (*naive Bayes classifier*, NB) (Domingos & Pazzani, 1997). Ο ταξινομητής αυτός εφαρμόζεται σε προβλήματα εκμάθησης όπου τα στιγμιότυπα αναπαρίστανται μέσω του μοντέλου του διανυσματικού χώρου, τα χαρακτηριστικά παίρνουν διακριτές τιμές (αν κάποια είναι συνεχή πρέπει να κβαντιστούν), και η συνάρτηση-στόχος παίρνει τιμές (ετικέτες) σε ένα πεπερασμένο σύνολο V.

3.1.5 Ταξινομητές Κ-Πλησιέστερων Γειτόνων

Ο ταξινομητής Κ-πλησιέστερων γειτόνων (*K-Nearest Neighbor, KNN*) είναι μια καλύτερη προσέγγιση του αλγορίθμου πλησιέστερου γείτονα. Στην πιο απλή του μορφή, χρησιμοποιείται μόνο η απόσταση από τον k-ιοστό κοντινότερο γείτονα. Το μέτρο που χρησιμοποιεί ο ταξινομητής KNN για να υπολογίσει την απόσταση μεταξύ ζευγών στοιχείων δεδομένων είναι η Ευκλείδεια απόσταση. Η προσέγγιση αυτή έχει αρκετά καλά αποτελέσματα σε πολυδιάστατους χώρους. Όταν τα αντικείμενα είναι πολύ κοντά στα δεδομένα-στόχο, χρησιμοποιείται η απόσταση του k-ιοστού πλησιέστερου γείτονα αντί για την απόσταση του πρώτου πλησιέστερου γείτονα. Αυτό το μέτρο ευρωστίας, καθιστά δύσκολο τον εντοπισμό ακραίων αντικειμένων ή αντικειμένων που βρίσκονται μέσα σε μια σφιχτή συστάδα. Επίσης, ο αλγόριθμος απαιτεί να ορίσει ο χρήστης τον αριθμό k των γειτόνων, ο οποίος πρέπει να βρίσκεται μεταξύ των ορίων 2 και 50.Ένα πλεονέκτημα του ταξινομητή KNN είναι η απλότητά του.

3.2 Μέτρα Αξιολόγησης Αλγορίθμων Ταξινόμησης

Για την αξιολόγηση των αποτελεσμάτων των διαφόρων μεθόδων ταξινόμησης σε πραγματικά σύνολα δεδομένων είναι απαραίτητη η ανάπτυξη μετρικών που ποσοτικοποιούν την ικανότητα ενός ταξινομητή να επιλέγει τις σωστές κατηγορίες για τα δεδομένα εισόδου. Στην πράξη, η αξιολόγηση βασίζεται στη μέτρηση εκείνων των εγγραφών δοκιμής που ταξινομήθηκαν σωστά και λανθασμένα, και συνοψίζονται σε έναν πίνακα σύγχυσης

(*confusion matrix*). Κάθε στιγμιότυπο ij στον πίνακα δείχνει τον αριθμό των εγγραφών από την κλάση i που προβλέπεται να ανήκει στην κλάση j. Για παράδειγμα, για το στιγμιότυπο f_{01} είναι ο αριθμός των εγγραφών από την κλάση 0 που εσφαλμένα με την πρόβλεψη τοποθετήθηκε στην κλάση 1. Βασισμένο στα στιγμιότυπα, ο αριθμός τελικά που προβλέπονται σωστά είναι το άθροισμα των f_{00} και f_{11}, ενώ αυτά που προβλέπονται λάθος είναι τα f_{01} και f_{10}.

Πίνακας 4. Παράδειγμα πίνακα σύγχυσης για πρόβλημα 2 κλάσεων και οι τύποι αποφάσεων ενός ταξινομητή

		Προβλεπόμενη Κλάση	
		Κλάση 1	Κλάση 0
Πραγματική Κλάση	Κλάση 1	TP (f_{11})	FN (f_{10})
	Κλάση 0	FP (f_{01})	TN (f_{00})

Κατά τον έλεγχο του πίνακα σύγχυσης υπολογίζονται όλες οι πιθανές περιπτώσεις κατάταξης, οι οποίες εκφράζονται σε σχέση με τον αριθμό των αληθώς θετικών (True Positives, TN), αληθώς αρνητικών (True Negatives, TN), ψευδώς θετικών (False Positives, FP), και ψευδώς αρνητικών (False Negatives, FN) ταξινομήσεων που αφορούν την κάθε κλάση. Στόχος, λοιπόν, είναι η ελαχιστοποίηση των FP, FN και η μεγιστοποίηση των άλλων μεγεθών.

Ένα άλλο κυρίαρχο μέτρο αποτίμησης της αποδοτικότητας ενός ταξινομητή είναι η ορθότητα (*accuracy*) που μετρά την ικανότητα του ταξινομητή να επιλέγει τη σωστή κατηγορία. Χρησιμοποιώντας τα μεγέθη του Πίνακα 4, η ορθότητα υπολογίζεται ως εξής:

$$\text{Accuracy} = (TP + TN) / (P + N).$$

Σε περιπτώσεις συνόλων δεδομένων με κυρτές κατανομές κατηγοριών, η ορθότητα δεν μπορεί να αξιολογήσει πλήρως έναν ταξινομητή. Έτσι, για την αξιολόγηση της ταξινόμησης ανά κατηγορία χρησιμοποιούνται και άλλες μετρικές, η οποίες αναλύονται στη συνέχεια (Tan, Steinbach, Kumar, 2005), (Mitchell, 1997).

Η ανάκληση (*recall*) ή ευαισθησία (*sensitivity*) είναι η ικανότητα του ταξινομητή να ανακαλεί (εντοπίζει) τα δείγματα που ανήκουν σε μια συγκεκριμένη κατηγορία. Είναι δηλαδή το ποσοστό των δειγμάτων που ο ταξινομητής καταφέρνει να κατηγοριοποιήσει ορθά, από το σύνολο των δειγμάτων που ανήκουν στην κατηγορία (true positive rate).

$$\text{Recall} = TP/(TP+FN).$$

Η ακρίβεια (*precision*) ορίζεται ως το ποσοστό των ορθών προβλέψεων (κατηγοριοποιήσεων ενός ταξινομητή) στο σύνολο των αποφάσεων αποδοχής που πραγματοποιεί.

$$\text{Precision} = TP/(TP+FP).$$

Μια μετρική που συνδυάζει την ανάκληση και την ακρίβεια είναι ο αρμονικός μέσος όρος *F* της ανάκλησης και της ακρίβειας ή F-measure.

$$\text{F-measure} = (2 \times \text{Accuracy} \times \text{Recall}) / (\text{Accuracy} + \text{Recall}).$$

Τέλος, η ειδικότητα (*specificity*) δείχνει την ικανότητα ενός ταξινομητή να απορρίπτει σωστά τα δείγματα, δηλαδή το ποσοστό των σωστών απορρίψεων στο σύνολο των δειγμάτων που θα έπρεπε να απορριφθούν (true negative rate).

$$\text{Specificity} = TN/(TN+FP)$$

3.3 Επικύρωση Σφάλματος Ταξινομητών

Σχεδόν πάντα, όλοι οι αλγόριθμοι ταξινόμησης προτύπων έχουν μια ή περισσότερες ελεύθερες παραμέτρους (Arlot & Celisse, 2010). Για παράδειγμα, ο αριθμός νευρώνων σε ένα νευρωνικό δίκτυο, ο αριθμός των πλησιέστερων γειτόνων σε έναν ταξινομητή KNN, κ.ο.κ.. Η ποιότητα των μοντέλων εξετάζεται με την εκτίμηση του *σφάλματος γενίκευσης*, δηλαδή την ικανότητα του μοντέλου να προβλέπει την κατηγορία μιας νέας περίπτωσης. Με βάση την απόδοση του κάθε ταξινομητή στα διαθέσιμα δεδομένα προκύπτει εάν ένα μοντέλο καλύπτει καλά το σύνολο των δεδομένων που διαχειρίζεται. Η πλέον διαδεδομένη επιλογή επικύρωσης είναι η διασταυρωμένη επικύρωση k-τεμαχίων (k-*fold cross validation*)[1], σύμφωνα με την οποία τα διαθέσιμα δεδομένα διαχωρίζονται με τυχαίο τρόπο σε k μη επικαλυπτόμενα τεμάχια (*folds*). Κάθε τεμάχιο έχει τον ίδιο αριθμό προτύπων, ενώ διατηρούνται προσεγγιστικά και οι σχετικές αναλογίες των προτύπων κάθε κλάσης. Η κατηγοριοποίηση επαναλαμβάνεται k φορές, κάθε φορά θεωρώντας τα δεδομένα ενός τεμαχίου ως δεδομένα δοκιμής και τα υπόλοιπα $k - 1$ ως δεδομένα εκπαίδευσης. Το συνολικό σφάλμα κατηγοριοποίησης στα k σύνολα δοκιμής που προκύπτει, ονομάζεται σφάλμα διασταυρωμένης επικύρωσης (*cross-validation error*). Υπάρχει ακόμα ένα είδος σφάλματος που μπορεί να διαπράξει ο ταξινομητής, και ονομάζεται σφάλμα κατάρτισης (*resubstitution*). Το σφάλμα αυτό είναι το ποσοστό των παρατηρήσεων που ταξινομούνται λανθασμένα στο σύνολο εκπαίδευσης. Ένα καλό μοντέλο πρέπει να έχει χαμηλό σφάλμα κατάρτισης, καθώς και χαμηλό σφάλμα διασταυρωμένης επικύρωσης (Mitchell, 1997).

4. Αποτελέσματα Αξιολόγησης

Οι τεχνικές μηχανικής μάθησης που παρουσιάστηκαν στην Ενότητα 3.1 συγκρίθηκαν σε ένα ελεγχόμενο σετ από πειράματα που στόχο είχαν να εξετάσουν την επίδοση τους στην εκμάθηση των δεδομένων της έρευνας.

Η επιτυχία της εκμάθησης αξιολογείται συγκρίνοντας την έξοδο του ταξινομητή με τη μεταβλητή των μελών της συστάδας που παράγεται από την πολυδιάστατη παραγοντική ανάλυση (διαδικασία VACOR) για μια σειρά από τυχαία επιλεγμένες περιπτώσεις που σκόπιμα διατηρούνται έξω από την εκπαιδευτική διαδικασία (δηλ. πρόκειται για περιπτώσεις τις οποίες οι ταξινομητές δεν έχουν ξαναδεί ποτέ).

Ο πίνακας εισόδου περιλάμβανε 1721 περιπτώσεις και διαμοιράστηκε τυχαία στα δεδομένα που θεωρούνται «γνωστά» (training data), καθώς και στα «άγνωστα» δεδομένα (testing data) κάνοντας χρήση της τεχνικής διασταυρωμένης επικύρωσης k=10 τεμαχίων.

Στη συνέχεια της ενότητας, παρουσιάζουμε τα αποτελέσματα που αφορούν στην επικύρωση της απόδοσης των ταξινομητών.

Τα αποτελέσματα της κατάταξης αξιολογήθηκαν βάσει των μετρικών που αναλύθηκαν στην Ενότητα 3.2, και παρουσιάζονται για κάθε περίπτωση στους Πίνακες 5, 6 και 7. Τα αριθμητικά αποτελέσματα επιβεβαιώνουν ότι οι μηχανές διανυσμάτων απόφασης (SVM) μπορούν να μάθουν με μεγαλύτερο ποσοστό επιτυχίας πώς να γενικεύουν τα αποτελέσματα της ανάλυσης των δεδομένων σε ένα πολυδιάστατο πρόβλημα ταξινόμησης.

Επίσης, καταδεικνύουν ότι η συγκεκριμένη τεχνική μηχανικής μάθησης εκτελεί με μεγαλύτερη επιτυχία την ταξινόμηση των αγνώστων επισκεπτών σε όλες τις κατηγορίες

ερωτήσεων (ορθότητα 99,41% για τις ερωτήσεις Δ1-Δ3, 95,88% για τις ερωτήσεις Δ4-Δ9, και 100% για τις ερωτήσεις Γ1-Γ15) .

Πίνακας 5.Απόδοση των 5 ταξινομητών για τα δεδομένα των ερωτήσεων Δ1-Δ3.

Ταξινομητής	DT	ANN	SVM	NB	KNN
Σφάλμα αντικατάστασης	0.006	0.102	0.004	0.203	0.006
Σφάλμα επικύρωσης	0.093	0.112	0.039	0.229	0.065
Ορθότητα	0.924	0.919	0.994	0.843	0.936
Ακρίβεια	0.885	0.903	0.997	0.697	0.877
Ευαισθησία	0.979	0.976	0.999	0.952	0.980
Ειδικότητα	0.905	0.941	0.993	0.887	0.928
F-score	0.927	0.937	0.998	0.772	0.921
Gmean	0.894	0.917	0.995	0.713	0.894

Πίνακας 6.Απόδοση των 5 ταξινομητών για τα δεδομένα των ερωτήσεων Δ4-Δ9.

Ταξινομητής	DT	ANN	SVM	NB	KNN
Σφάλμα αντικατάστασης	0.016	0.099	0.008	0.217	0.058
Σφάλμα επικύρωσης	0.121	0.174	0.061	0.262	0.109
Ορθότητα	0.919	0.866	0.959	0.802	0.901
Ακρίβεια	0.890	0.820	0.941	0.790	0.842
Ευαισθησία	0.974	0.967	0.990	0.938	0.971
Ειδικότητα	0.925	0.804	0.930	0.824	0.877
F-score	0.929	0.890	0.965	0.856	0.900
Gmean	0.905	0.805	0.934	0.799	0.855

Πίνακας 7.Απόδοση των 5 ταξινομητών για τα δεδομένα των ερωτήσεων Γ1-Γ15.

Ταξινομητής	DT	ANN	SVM	NB	KNN
Σφάλμα αντικατάστασης	0.000	0.002	0.000	0.000	0.002
Σφάλμα επικύρωσης	0.000	0.037	0.000	0.000	0.028
Ορθότητα	1.000	0.983	1.000	1.000	0.994
Ακρίβεια	1.000	0.967	1.000	1.000	0.987
Ευαισθησία	1.000	0.996	1.000	1.000	0.997
Ειδικότητα	1.000	0.967	1.000	1.000	0.998
F-score	1.000	0.981	1.000	1.000	0.992
Gmean	1.000	0.964	1.000	1.000	0.992

Τα αποτελέσματα της κατάταξης αξιολογήθηκαν και μέσω του πίνακα σύγχυσης, και παρουσιάζονται λόγω περιορισμένου χώρου μόνο για τον ταξινομητή SVM στον Πίνακα 8. Για την πρώτη περίπτωση ταξινόμησης (ερωτήσεις Δ1-Δ3), το 97.44% των περιπτώσεων της κλάσης K1 και το 100.0% των περιπτώσεων των κλάσεων K2-K5 αντίστοιχα έχουν προβλεφθεί σωστά, και η συνολική ακρίβεια ταξινόμησης που επιτεύχθηκε έφτασε σε ποσοστό το 99.41%. Όσον αφορά στη δεύτερη περίπτωση ταξινόμησης (ερωτήσεις Δ4-Δ9), μόνο στην κατηγορία K3 είχαμε μηδενικές ψευδείς ταξινομήσεις από το συνολικό σετ των δεδομένων ελέγχου, και η συνολική συμφωνία που επιτεύχθηκε μεταξύ των προβλεπόμενων και των πραγματικών κλάσεων έφτασε σε ποσοστό το 95.88%. Τέλος, για την τρίτη περίπτωση ταξινόμησης (ερωτήσεις Γ1-Γ15), οι ορθώς ταξινομημένες περιπτώσεις είναι 100% για όλες τις κλάσεις K1-K5.

Πίνακας 8. Πίνακες σύγχυσης του ταξινομητή SVM. Ταξινόμηση των επισκεπτών σε 5 κλάσεις ανάλογα με τις ερωτήσεις Δ1-Δ3, Δ4-Δ9 και Γ1-Γ15 αντίστοιχα.

Δ1-Δ3		Προβλεπόμενη Κλάση				
		K1	K2	K3	K4	K5
Πραγματική Κλάση	K1	38	0	0	0	1
	K2	0	11	0	0	0
	K3	0	0	69	0	0
	K4	0	0	0	22	0
	K5	0	0	0	0	29

Δ4-Δ9		Προβλεπόμενη Κλάση				
		K1	K2	K3	K4	K5
Πραγματική Κλάση	K1	13	0	1	0	0
	K2	1	74	0	1	0
	K3	0	0	22	0	0
	K4	0	1	0	43	1
	K5	2	0	0	0	11

Γ1-Γ15		Προβλεπόμενη Κλάση				
		K1	K2	K3	K4	K5
Πραγματική Κλάση	K1	107	0	0	0	0
	K2	0	15	0	0	0
	K3	0	0	17	0	0
	K4	0	0	0	15	0
	K5	0	0	0	0	15

5. Συμπεράσματα

Στην παρούσα εργασία αξιολογήσαμε και συγκρίναμε την ικανότητα εκμάθησης και γενίκευσης διαφόρων τεχνικών επιβλεπόμενης μηχανικής μάθησης σε τρία προβλήματα ταξινόμησης επισκεπτών.

΄Οπως φάνηκε από τα αποτελέσματα της σύγκρισης, οι μηχανές SVM αξιοποίησαν καλύτερα τα δεδομένα έρευνας και παρουσίασαν τη μεγαλύτερη πιθανότητα επιτυχημένης εκμάθησης σε όλα τα προβλήματα ταξινόμησης τα οποία τους παρουσιάστηκαν. Όσον αφορά στην απόδοση των δένδρων απόφασης, τα αριθμητικά αποτελέσματα δείχνουν ότι ο ταξινομητής αυτός επιτυγχάνει τη δεύτερη καλύτερη επίδοση απαιτώντας μάλιστα σημαντικά μικρότερο χρόνο εκπαίδευσης. Ο ταξινομητής KNN παρουσιάζει την τρίτη καλύτερη επίδοση, η οποία μάλιστα είναι άμεσα συνυφασμένη με την επιλογή του αριθμού k των πλησιέστερων γειτόνων.

Αναφορικά με τα τεχνητά νευρωνικά δίκτυα, από την πειραματική διαδικασία, παρατηρήσαμε ότι η διαδικασία προσαρμογής τους στα δεδομένα της έρευνας ήταν η πιο δύσκολη συγκρινόμενη με όλους τους υπόλοιπους ταξινομητές. Η κατάλληλη δε επιλογή του αριθμού των νευρώνων αποδεικνύεται πολύ σημαντική για την απόδοση των δικτύων αυτών.

Τέλος, διαπιστώσαμε ότι ο πιθανοτικός ταξινομητής NB χρειάζεται περαιτέρω βελτιστοποίηση των παραμέτρων του προκειμένου να επιτύχει καλύτερα ποσοστά εκμάθησης. Μια πρώτη παρατήρηση που αφορά στον ταξινομητή αυτό, είναι ότι η επιλογή «mvmn» αντί του «Kernel» σαν μοντέλο κατανομής βελτιώνει σημαντικά την απόδοση του.

Συνολικά, και με βάση τα αριθμητικά αποτελέσματα, μπορούμε να ισχυριστούμε ότι οι μηχανές διανυσμάτων υποστήριξης (SVM) είναι μια πολλά υποσχόμενη τεχνική για την αποτελεσματική αξιολόγηση της ταξινόμησης των αντικειμένων με τη μέθοδο της Ανιούσας Ιεραρχικής Ταξινόμησης.

Η παρακάτω εφαρμογή με πραγματικά δεδομένα δείχνει τον τρόπο εφαρμογής της μεθόδου KARAP

Η ΕΦΑΡΜΟΓΗ ΤΗΣ ΜΕΘΟΔΟΥ KARAP ΩΣ ΕΡΓΑΛΕΙΟ ΣΧΕΔΙΑΣΜΟΥ ΔΙΑΦΗΜΙΣΤΙΚΗΣ ΕΚΣΤΡΑΤΕΙΑΣ ΕΝΟΣ ΤΟΥΡΙΣΤΙΚΟΥ ΠΡΟΟΡΙΣΜΟΥ

Περίληψη

Η παρούσα εργασία αφορά μια νέα μέθοδο σχεδιασμού μιας διαφημιστικής εκστρατείας ενός τουριστικού προορισμού. Τα στοιχεία της έρευνας συγκεντρώθηκαν από τους επισκέπτες της περιοχής με τη χρήση ενός ερωτηματολογίου και τα αποτελέσματα προέκυψαν μετά από επεξεργασία των απαντήσεων με πολυπαραγοντικές στατιστικές αναλύσεις και ιδιαίτερα με μια νέα εφαρμογή την αποκαλούμενη μέθοδο karap, η οποία συνδυάζει τα αποτελέσματα της Παραγοντικής Ανάλυσης των Αντιστοιχιών, με τον Ευκλείδειο διανυσματικό χώρο R^n.

Εισαγωγή

Έστω σ' ένα πίνακα δεδομένων Τ(n,p) οι n γραμμές αντιστοιχούν σε n ερωτώμενους, ενώ στις p γραμμές του πίνακα οι τιμές των p ερωτήσεων που αντιστοιχούν σε p κριτήρια. Στο ερώτημα του εντοπισμού ενός ερωτώμενου με ποιο κριτήριο συνδέεται κυρίως, θα γίνει λεπτομερής αναφορά, χρησιμοποιώντας ένα συγκεκριμένο ερωτηματολόγιο στο οποίο απάντησαν 231 Ρώσοι επισκέπτες από τους 1721 ξένους επισκέπτες της Θεσσαλονίκης από 51 χώρες της υφηλίου, που ερωτήθηκαν την περίοδο 15-5-13 έως 15-9-13. Η επιλογή των συγκεκριμένων τουριστών έγινε επειδή το μέγεθός του δείγματος εκτός του ότι ήταν ικανοποιητικό, δηλαδή το 13,42% του συνόλου αλλά και λόγω ιδιαίτερων δεσμών που έχουν οι συγκεκριμένοι επισκέπτες με την πόλη. Τα δεδομένα περιέχονται στην έρευνα που διεξήχθη στα πλαίσια του προγράμματος ΑΡΧΙΜΗΔΗΣ ΙΙΙ με τίτλο «Τεχνολογίες Ανάλυσης Δεδομένων και Διαχείρισης Γνώσης στο σχεδιασμό τουριστικών προϊόντων» με επιστημονικό υπεύθυνο τον καθηγητή Δρ. Δημήτριο Καραπιστόλη.

Στα στοιχεία που θα επικεντρωθούμε είναι σχετικά με τις απαντήσεις των Ρώσων επισκεπτών βάσει τριών ενοτήτων. Η πρώτη ενότητα αφορούσε στα ΑΙΤΙΑ που τους προκάλεσαν να επισκεφθούν την πόλη της Θεσσαλονίκης, ενώ η δεύτερη ενότητα αφορούσε στην κριτική τους για την ΕΙΚΟΝΑ που παρουσιάζει η πόλη. Η τρίτη ενότητα αφορούσε στο ποια θα είναι η ΣΤΑΣΗ που θα κρατήσουν στο μέλλον για μια εκ νέου επίσκεψη στη πόλη με τρεις διαβαθμίσεις (Απίθανο-Πιθανό-Βέβαιο).

Σχετικά με τα αίτια που προκάλεσαν την επίσκεψη χρησιμοποιήθηκαν τρεις ενότητες των πέντε κριτηρίων, όπου η ενότητα Γ1 η οποία αφορούσε τους λόγους επιλογής του προορισμού περιλάμβανε τα εξής πέντε κριτήρια Γ11={Η φήμη του ως τουριστικού προορισμού}, Γ12={Οι φυσικές ομορφιές της περιοχής}, Γ13={Οι γνώμες φίλων και γνωστών}, Γ14={Το κλίμα} και Γ15={Ιστορία της περιοχής}.

Η Γ2 ενότητα αφορούσε στα στοιχεία που τους προσέλκυσαν περισσότερο κατά την επίσκεψή τους και περιλάμβανε τα εξής πέντε κριτήρια. Γ21={Μπάνια-Θαλάσσια σπορ}, Γ22={Επισκέψεις μουσείων}, Γ23={Νυχτερινή διασκέδαση}, Γ24={ Εκδρομές-Επαφές με τη φύση} και Γ25={ Η τοπική κουζίνα}.

Η Γ3 ενότητα αφορούσε στους πιο σημαντικούς παράγοντες ενός προορισμού και περιλάμβανε τα εξής πέντε κριτήρια. Γ31={Οργάνωση-Υποδομές}, Γ32={Τοπικές μεταφορές}, Γ33={ Ασφάλεια}, Γ34={Φυσικό περιβάλλον} και Γ35={Life Style}.

Σχετικά με την εικόνα της πόλης χρησιμοποιήθηκαν τα εξής εννέα κριτήρια. α) Δ1={Την καθαριότητα}, Δ2={ Τις φυσικές ομορφιές},Δ3={ Τις τιμές προϊόντων και υπηρεσιών},Δ4={ τα αξιοθέατα της πόλης της Θεσσαλονίκης}, Δ5={την Ελληνική κουζίνα},Δ6={ την νυχτερινή ζωή της πόλης},Δ7={ το αρχιτεκτονικό της στυλ},Δ8={ την ασφάλειά της} και Δ9={ την φιλικότητα των ντόπιων

Εφαρμογή της μεθόδου KARAP

Στον πίνακα 1 παρουσιάζεται τμήμα των κωδικοποιημένων απαντήσεων των 231 Ρώσων, όπου η μεταβλητή ΣΤΑΣΗ θα χρησιμοποιηθεί ως εξαρτημένη μεταβλητή του μοντέλου πρόβλεψης.

Πίνακας 1:Τμήμα του πίνακα των κωδικοποιημένων απαντήσεων

ind	Γ1	Γ2	Γ3	Δ1	Δ2	Δ3	Δ4	Δ5	Δ6	Δ7	Δ8	Δ9	ΣΤ
35	3	1	5	3	4	4	4	4	5	0	4	3	1
112	1	1	4	3	5	3	4	5	3	4	3	5	2
131	5	3	5	3	5	5	5	5	5	5	3	5	3
133	5	3	5	3	4	5	5	5	5	4	4	5	3
134	1	3	3	3	4	4	4	5	5	4	5	5	3
135	5	3	5	3	5	4	4	5	5	5	3	5	3

Σημείωση: Οι απαντήσεις λ.χ του επισκέπτη Νο 35 παρουσιάζονται κωδικοποιημένες ως εξής:

Για την τιμή λ.χ Γ1=3 η επιλογή του ήταν «η φήμη του ως τουριστικού προορισμού», ενώ για την τιμή Δ8=4 η επιλογή του για την ασφάλειά ήταν «Ευχάριστη», ενώ όσο αφορά την μελλοντική του στάση για μια νέα επίσκεψη στη πόλη ήταν ΣΤ=1, δηλαδή «Απίθανο».

Αρχικά η ανάλυση θα περιοριστεί στις απαντήσεις της πρώτης ενότητας ώστε να εντοπιστούν για κάθε προφίλ επισκέπτη με ποιο προφίλ κριτηρίου συνδέεται περισσότερο.

Πίνακας 1α: Τμήμα του πίνακα που αφορούν την 1^{η} ενότητα κριτηρίων

ind	Γ1	Γ2	Γ3
35	3	1	5
112	1	1	4
131	5	3	5
133	5	3	5
134	1	3	3
135	5	3	5
169	2	1	2
188	5	1	1
207	3	4	4

Με την χρήση του λογισμικού MAD δημιουργείται ο λογικός πίνακας 0-1.

Πίνακας 2:Το προφίλ 10 Ρώσων με βάση τις αιτίες προσέλκυσης

Ind	Γ11	Γ12	Γ13	Γ14	Γ15	Γ21	Γ22	Γ23	Γ24	Γ25	Γ31	Γ32	Γ33	Γ34	Γ35
35	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1
112	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0
131	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1
133	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1
134	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0
135	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1
169	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0
188	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0
207	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0

Εφαρμόζουμε την μέθοδο KARAP. Αρχικά παίρνουμε τον πίνακα ταξινόμησης των ερωτώμενων και στη συνέχεια βγάζουμε τον αντίστοιχο λογικό πίνακα 0-1 της ταξινόμησης.

Πίνακας 3:Ο πίνακας ταξινόμησης των ερωτώμενων

IND	Γ11	Γ12	Γ13	Γ14	Γ15	Γ21	Γ22	Γ23	Γ24	Γ25	Γ31	Γ32	Γ33	Γ34	Γ35
	23	**5**	**17**	**22**	**8**	**10**	**25**	**11**	**33**	**13**	**14**	**10**	**19**	**4**	**17**
1	112	417	320	327	220	425	209	131	207	313	188	169	477	307	35
2	365	420	371	359	339	618	214	133	228	743	340	419	592	478	389
3	366	459	372	370	426	983	215	134	229	845	424	456	674	847	543
4	376	612	373	451	852	993	256	135	238	859	672	640	750	1057	546
5	488	1020	1056	466	887	1003	353	460	293	888	848	1186	791		849
6	544		1444	510	964	1007	358	495	294	1059	963	1195	846		961
7	545		1508	585	1044	1046	374	513	295	1187	1010	1370	858		1006
8	635		1544	596	1247	1101	378	514	296	1389	1189	1406	1019		1023
9	673		1567	857		1118	418	1400	297	1390	1265	1474	1050		1115
10	676		1574	1008		1363	427	1509	298	1393	1330	1481	1081		1180
11	751		1618	1011			476	1668	299	1465	1625		1194		1181
12	790		1624	1047			749		300	1480	1660		1371		1409
13	850		1627	1048			752		301	1510	1667		1466		1476
14	908		1647	1062			851		302		1669		1467		1478
15	968		1654	1254			860		303				1485		1479
16	1083		1656	1256			962		304				1486		1511
17	1085		1659	1261			965		305				1614		1626
18	1193			1264			1009		306				1615		
19	1213			1564			1021		308				1655		
20	1280			1613			1069		309						
21	1374			1657			1188		310						
22	1383			1671			1249		311						
23	1408						1279		312						
24							1619		314						
25							1653		315						
26									319						
27									357						
28									909						
29									1196						
30									1278						
31									1443						
32									1475						
33									1484						

Πίνακας 4: Ο λογικός πίνακας 0-1 της ταξινόμησης του πίνακα 3

IND	Γ11	Γ12	Γ13	Γ14	Γ15	Γ21	Γ22	Γ23	Γ24	Γ25	Γ31	Γ32	Γ33	Γ34	Γ35
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
112	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
131	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
133	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
134	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
135	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
169	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
188	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
207	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

Στον πίνακα 4 προσθέτουμε μια νέα στήλη με τις τιμές της στήλης ΣΤΑΣΗ του πίνακα 1. Στη συνέχεια ταξινομούνται οι 231 ερωτώμενοι με βάση τις τιμές 1-3, απ' όπου προκύπτει ο

πίνακας 6, ο οποίος παρουσιάζει για κάθε στάση και κάθε κριτήριο το πλήθος των προφίλ των ερωτώμενων που συνδέονται με κάθε στάση. Ο πίνακας 5 παρουσιάζει την περίπτωση για την ΣΤΑΣΗ 1.

Πίνακας 5: Το σύνολο ΣΤ1 των ερωτώμενων για την ΣΤΑΣΗ 1

IND	Γ11	Γ12	Γ13	Γ14	Γ15	Γ21	Γ22	Γ23	Γ24	Γ25	Γ31	Γ32	Γ33	Γ34	Γ35
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
359	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
366	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
420	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
510	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
…	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1008	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ΣΤ1	**2**	**1**	**0**	**4**	**1**	**0**	**3**	**1**	**0**	**1**	**2**	**0**	**0**	**0**	**2**

Ακολουθεί ο πίνακας 6 στον οποίο εφαρμόζεται η μέθοδος KARAP.

Πίνακας 6: Κατανομή ερωτώμενων στις τρεις ΣΤΑΣΕΙΣ

IND	ΣΤ1	ΣΤ2	ΣΤ3
Γ11	2	**14**	7
Γ12	**1**	2	2
Γ13	0	6	**11**
Γ14	4	**11**	7
Γ15	1	**4**	3
Γ21	0	**8**	2
Γ22	3	**14**	8
Γ23	1	**6**	4
Γ24	0	18	**15**
Γ25	1	**7**	5
Γ31	2	**11**	1
Γ32	0	3	**7**
Γ33	0	11	**8**
Γ34	0	**3**	1
Γ35	2	**10**	5
	17	128	86

Μετά την ανάλυση προκύπτει ο πίνακας 7

IND	ΣΤΑΣΗ_1	ΣΤΑΣΗ_2	ΣΤΑΣΗ_3
ΠΛΗΘΟΣ	1	10	4
1	Γ12	Γ11	Γ13
2		Γ14	Γ24
3		Γ15	Γ32
4		Γ21	Γ33
5		Γ22	
6		Γ23	
7		Γ25	
8		Γ31	
9		Γ34	
10		Γ35	

Από τον συνδυασμό των πινάκων 3 και 7 βρίσκουμε αφενός από τον πίνακα 7 την σύνδεση κάθε στάσης με τα 15 κριτήρια, αφετέρου από τον πίνακα 3 τους ερωτώμενους που συνδέονται με κάθε κριτήριο.

Την ίδια διαδικασία εφαρμόζουμε με τα κριτήρια Δ1-Δ9 και το κριτήριο της ΣΤΑΣΗΣ για τους ίδιους επισκέπτες. Παρακάτω παρουσιάζεται ο πίνακας ταξινόμησης των επισκεπτών,

αφού προηγουμένως δημιουργήσαμε τον λογικό πίνακα 0-1 των τιμών Δ1-Δ9 και τον οποίο αναλύουμε με τη μέθοδο karap.Αρχικά προκύπτει ο πίνακας 8

Πίνακας 8: Σύνδεση ερωτώμενων με διαβαθμίσεις των εννέα κριτηρίων

Δ14	Δ24	Δ25	Δ31	Δ33	Δ34	Δ44	Δ45	Δ54	Δ55	Δ61	Δ63	Δ64	Δ72	Δ74	Δ75	Δ83	Δ84	Δ85	Δ94	Δ95
5	27	24	1	18	8	6	39	3	26	2	1	3	1	15	7	4	6	1	1	33
220	35	295	1264	238	306	460	214	961	131	1189	419	303	1009	207	353	188	169	596	299	112
544	209	309		312	365	909	215	965	133	1279		1567		339	417	888	293			314
1254	228	315		313	513	962	294	1613	134			1574		359	451	1046	635			319
1481	229	418		327	514	1194	297		135					488	672	1374	1008			320
1656	340	456		370	751	1443	298		256					545	845		1056			366
	358	477		510	752	1484	301		296					846	968		1195			371
	389	676		847	1101		302		300					848	1050					373
	420	851		849	1393		307		304					908						374
	427	1019		850			308		305					1011						378
	495	1021		964			310		311					1047						426
	749	1069		1007			372		357					1062						459
	790	1081		1044			376		543					1188						466
	859	1371		1057			424		546					1213						476
	1010	1475		1187			425		592					1624						478
	1023	1479		1330			852		674					1655						585
	1083	1485		1626			858		743											612
	1196	1508		1653			1020		750											618
	1265	1511		1669			1059		791											640
	1363	1544					1115		857											673
	1370	1564					1180		860											983
	1389	1614					1181		887											993
	1390	1625					1186		963											1003
	1400	1667					1256		1278											1006
	1406	1671					1280		1465											1048
	1409						1383		1486											1085
	1627						1408		1509											1118
	1668						1466													1193
							1474													1247
							1476													1249
							1478													1261
							1480													1444
							1615													1467
							1618													1510
							1619													
							1647													
							1654													
							1657													
							1659													
							1660													

Σημειωτέων ότι για κάποιες διαβαθμίσεις των κριτηρίων δεν υπήρξε καμιά σύνδεση με ερωτώμενο λ.χ για τις διαβαθμίσεις Δ11,Δ12,Δ13 και Δ15

Ακολουθώντας την ίδια διαδικασία όπως με τα κριτήρια Γ11-Γ35, προκύπτει αρχικά ο πίνακας 9, τον οποίο αναλύουμε με τη μέθοδο KARAP, με αποτέλεσμα να εντοπιστεί η σύνδεση των τριών στάσεων με τις διαβαθμίσεις των εννέα κριτηρίων.(πίνακας 10)

Πίνακας 9:Πλήθος προφίλ των ερωτώμενων που συνδέονται με τις ΣΤΆΣΕΙΣ 1,2, και 3

IND	ΣΤ1	ΣΤ2	ΣΤ3
Δ14	0	1	**4**
Δ24	2	**21**	3
Δ25	1	7	**16**
Δ31	0	**1**	0
Δ33	**5**	9	4
Δ34	0	**6**	2
Δ44	0	**6**	0
Δ45	0	17	**22**
Δ54	1	**3**	0
Δ55	1	**15**	10
Δ61	**2**	0	0
Δ63	0	**1**	0
Δ64	0	1	**2**
Δ72	0	**1**	0
Δ74	3	**8**	4
Δ75	0	2	**5**
Δ83	0	**4**	0
Δ84	1	**4**	1
Δ85	0	**1**	0
Δ94	0	0	**1**
Δ95	1	**20**	12
	17	**128**	**86**

Πίνακας 10: Σύνδεση των διαβαθμίσεων Δ11-Δ95 με τις ΣΤΆΣΕΙΣ 1,2, και 3

IND	**ΣΤ1**	**ΣΤ2**	**ΣΤ3**
ΠΛΗΘΟΣ	**2**	**13**	**6**
1	Δ33	Δ24	Δ14
2	Δ61	Δ31	Δ25
3		Δ34	Δ45
4		Δ44	Δ64
5		Δ54	Δ75
6		Δ55	Δ94
7		Δ63	
8		Δ72	
9		Δ74	
10		Δ83	
11		Δ84	
12		Δ85	
13		Δ95	

Με τη βοήθεια των πινάκων ταξινόμησης των κριτηρίων Γ11-Γ35 και Δ11-Δ95 για κάθε στάση 1,2, και 3, εντοπίζονται οι επισκέπτες, λ.χ από τον πίνακα 5,που συνδυάζουν τις διαβαθμίσεις των δύο κριτηρίων. Έτσι π.χ από τον πίνακα 6 ένας επισκέπτης από την ΣΤΑΣΗ 1 συνδυάζεται με την διαβάθμιση Γ12 των αιτίων επιλογής. Ο επισκέπτης αυτός εντοπίζεται με τον κωδικό 420. Συνδυάζοντας, λοιπόν, τα αποτελέσματα των πινάκων 6 και 7 καθώς και των πινάκων 9 και 10, βρίσκουμε για κάθε μία στάση (Απίθανο, Πιθανό, Βέβαιο) ποιες αιτίες (Γ11-Γ35) και ποια κριτήρια (Δ11-Δ95) επηρέασαν τους Ρώσους επισκέπτες να αντιδράσουν

με την συγκεκριμένη στάση απέναντι στη πόλη της Θεσσαλονίκης. Συνοπτικά έχουμε τον παρακάτω πίνακα 11.

Πίνακας 11:Σύνδεση των προφίλ κριτηρίων και των επισκεπτών με βάση την στάση τους

ΣΤΑΣΗ 1				**ΣΤΑΣΗ 2**				**ΣΤΑΣΗ 3**			
Γ12	1	Δ33	5	Γ11	14	Δ24	21	Γ13	11	Δ14	4
		Δ61	2	Γ14	11	Δ31	1	Γ24	15	Δ25	16
				Γ15	4	Δ34	6	Γ32	7	Δ45	22
				Γ21	8	Δ44	6	Γ33	8	Δ64	2
				Γ22	14	Δ54	3			Δ75	5
				Γ23	6	Δ55	15			Δ94	1
				Γ25	7	Δ63	1				
				Γ31	11	Δ72	1				
				Γ34	3	Δ74	8				
				Γ35	10	Δ83	4				
						Δ84	4				
						Δ85	1				
						Δ95	20				
	1		**7**		**88**		**91**		**41**		**50**

Συνοπτική αποτύπωση των δεδομένων του πίνακα 11 με τις αντίστοιχες απαντήσεις των 231 Ρώσων, οι οποίες είναι κατάλληλες για τον εντοπισμό της διαφημιστικής εκστρατείας, παρέχει ο πίνακας 12.

Πίνακας 12 : Κατανομή των απαντήσεων των Ρώσων επισκεπτών ανάλογα με την στάση τους σε συνδυασμό με τις αιτίες που τους έφερε στη πόλη και την βαθμολογία τους σε εννέα κριτήρια για την εικόνα που παρουσιάζει η Θεσσαλονίκη

	ΚΡΙΤΗΡΙΑ	**ΣΤΑΣΗ 1 (Απίθανο)**	**ΣΤΑΣΗ 2 (Πιθανό)**	**ΣΤΑΣΗ 3 (Βέβαιο)**	**ΣΥΝΟΛΟ**
1.	**ΑΙΤΙΕΣ ΠΡΟΣΕΛΚΥΣΗΣ**	1	88	41	130
2.	**ΕΙΚΟΝΑ ΠΟΛΗΣ**	7	91	50	148
	ΣΥΝΟΛΟ	**8**	**179**	**91**	**278**

Για να δημιουργηθεί ο διαφημιστικός σχεδιασμός της πόλης της Θεσσαλονίκης που θα αφορά μελλοντικούς Ρώσους τουρίστες, χρησιμοποιώντας τα αίτια και την εικόνα της πόλης όπως τα αποτύπωσαν οι Ρώσοι επισκέπτες σε αυτή την έρευνα, οφείλουμε να επεξεργαστούμε με συγκεκριμένη διαδικασία τα στοιχεία που εμφανίζει ο πίνακας 12.

Επειδή μεταξύ του πλήθους των 278 απαντήσεων ορισμένοι επισκέπτες εμφανίζονται να συνδέονται και με τα κριτήρια των αιτίων και με τα κριτήρια της εικόνας της πόλης, όπως διαφαίνεται από τον πίνακα 11,οφείλουμε να διαπιστώσουμε πόσοι και ποιοι είναι αυτοί, επίσης και πόσοι συνδέονται με άλλα κριτήρια κάθε στάσης, ώστε να προκύψουν οι αντίστοιχες πιθανότητες κάθε περίπτωσης. Αποτέλεσμα αυτής της επεξεργασίας είναι ο πίνακας 13.

Πίνακας 13: Κατανομή ερωτώμενων (και απαντήσεων) για κάθε περίπτωση

Περιπτώσεις	**ΣΤΑΣΗ 1**	**ΣΤΑΣΗ 2**	**ΣΤΑΣΗ 3**	**ΣΥΝΟΛΟ**
Μόνο με κριτήρια των αιτίων	1	25	14	40
Μόνο με κριτήρια της εικόνα	7	28	23	58
Και με τα δύο κριτήρια	0	63 (2x63=126)	27 (2x27=54)	90 (2x90=180)
Με τα υπόλοιπα κριτήρια της στάσης	9	12	22	43
ΣΥΝΟΛΟ (Επισκεπτών) **ΣΥΝΟΛΟ (Απαντήσεων)**	**17** **(8)**	**128** **(179)**	**86** **(91)**	**231** **(278)**

Αρχικά από τον πίνακα 13 εντοπίζεται η πιθανότητα (άρα και το ποσοστό) κάθε στάσης. Έτσι για την Στάση 1, όσοι θεωρούν Απίθανο να επισκεφτούν την πόλη, το ποσοστό είναι 7,36%, για όσους είναι Πιθανό (Στάση 2) το ποσοστό ανέρχεται στο 55,41%, ενώ όσοι είναι βέβαιοι ότι θα επαναλάβουν μια νέα επίσκεψη (Στάση 3) ανέρχονται στο 37,23%. Ακολουθεί ο πίνακας 14, ο οποίος παρουσιάζει την πιθανότητα κάθε στάσης με βάση τις απαντήσεις σύμφωνα με τα δεδομένα του πίνακα 13, λαμβάνοντας υπόψη τα άτομα που εντοπίζονται για κάθε περίπτωση

Πίνακας 14: Πίνακας πιθανοτήτων για κάθε περίπτωση και κάθε στάση

Περιπτώσεις	**ΣΤΑΣΗ 1**	**ΣΤΑΣΗ 2**	**ΣΤΑΣΗ 3**
Μόνο από τις αιτίες	0,0588	0,1953	0,1628
Μόνο από την εικόνα	0,4118	0,2188	0,2675
Και από τα δύο κριτήρια	0	0,4921	0,3139
Με τα υπόλοιπα κριτήρια της στάσης	0,5294	0,0938	0,2558
Σύνολο	1	1	1

Εφαρμόζοντας στα δεδομένα του πίνακα 13 το τεστ ανεξαρτησίας σε επίπεδο σημαντικότητας α=5%, έδειξε το εξής αποτέλεσμα.

βαθμοί ελευθερίας ν=6, $x^2_{(v,0.05)}$=12,591 και x^2=34,289. Επειδή $x^2 > x^2_{(v,0.05)}$ συνεπάγεται ότι η στάση ενός επισκέπτη εξαρτάται από τα κριτήρια που διαμορφώνουν την επιλογή του να επισκεφθεί τη Θεσσαλονίκη.

Συνεπώς

Από τον πίνακα 14, χωρίς να υπολογίζονται οι πιθανότητες που αφορούν στα υπόλοιπα κριτήρια κάθε στάσης, επειδή το προφίλ τους χαρακτηρίζει διαφορετική στάση, όπως προκύπτει από τους πίνακες 6 και 9, διατυπώνονται τα εξής συμπεράσματα.

1) Από το γενικό ποσοστό 7,36% των Ρώσων που δήλωσαν αρνητική στάση, ξεχωρίζει ένα πολύ μικρό ποσοστό (5,88%) το οποίο δεν θα επαναλάβει την επίσκεψη, επειδή δεν άρεσε τις φυσικές ομορφιές της περιοχής, ενώ ένα ποσοστό 41,18% δεν έχει καθόλου καλή εικόνα της Θεσσαλονίκης για τις τιμές των προϊόντων και τη νυκτερινή ζωή της πόλης.
2) Από το γενικό ποσοστό 55,41% των Ρώσων που δήλωσαν ότι είναι πιθανή μια επανάληψη της επίσκεψης, ξεχωρίζει ένα σημαντικό ποσοστό (49,21%) το οποίο ως αίτιο προσέλκυσης υπήρξε η φήμη του προορισμού, το κλίμα, οι επισκέψεις μουσείων, η οργάνωση-υποδομές της περιοχής και το Life Style, ενώ ενθουσιάστηκαν κυρίως με τις φυσικές ομορφιές της πόλης, με την Ελληνική κουζίνα και την φιλικότητα των ντόπιων.
3) Από το γενικό ποσοστό 37,23% των Ρώσων που δήλωσαν ότι είναι βέβαιη μια επανάληψη της επίσκεψης ξεχωρίζει ένα ποσοστό (31,39%) το οποίο ως αίτιο προσέλκυσης υπήρξαν οι γνώμες φίλων και γνωστών, καθώς και η δυνατότητα που παρέχει η περιοχή για εκδρομές, ενώ ενθουσιάστηκαν με τις φυσικές ομορφιές και με τα αξιοθέατα της πόλης. Επίσης ένα ικανοποιητικό ποσοστό 26,75% στηρίζει την συγκεκριμένη απόφαση επειδή έμειναν ικανοποιημένοι από την εικόνα της πόλης.

Λόγω του υψηλού ποσοστού (92,64%) των Ρώσων που έχουν θετική και μάλλον θετική στάση για μια μελλοντική επίσκεψη στη Θεσσαλονίκη, μια σωστή διαφημιστική εκστρατεία για τους ομοεθνείς τους, πρέπει να εστιαστεί κυρίως στα παρακάτω θετικά στοιχεία που προσδιόρισαν οι Ρώσοι επισκέπτες της Θεσσαλονίκης.

Έτσι

Α) Όσον αφορά στις αιτίες που θα προκαλούσαν μια επίσκεψη στη πόλη, πρέπει να δοθεί έμφαση στη φήμη που έχει η πόλη ως τουριστικός προορισμός, στο κλίμα της περιοχής, στις επισκέψεις μουσείων, στο Life Style της πόλης και στη δυνατότητα για εκδρομές στη φύση.

Β) Σε ότι έχει σχέση με την εικόνα της πόλης πρέπει να τονιστούν τα αξιοθέατα της πόλης, η Ελληνική κουζίνα και η φιλικότητα των ντόπιων.

Συμπέρασμα

Η εφαρμογή της μεθόδου KARAP στη διαμόρφωση του σχεδιασμού της διαφημιστικής εκστρατείας ενός τουριστικού προορισμού μετά από μία έρευνα αγοράς με κριτήρια που προσδιορίζουν τον λόγο της επίσκεψης των ερωτώμενων, αλλά και με κριτήρια που καθορίζουν την εικόνα της περιοχής, προσδιορίζει με ακρίβεια

Α) την μελλοντική στάση της επίσκεψης από τους ίδιους ερωτηθέντες

Β) Ποια κριτήρια χαρακτηρίζουν την περιοχή

Έτσι, με βάση τα πραγματικά στοιχεία της ανάλυσης καθορίζεται η διαφημιστική εκστρατεία που πρέπει να ακολουθήσει η τοπική κοινωνία για να αυξήσει την επόμενη περίοδο τους ξένους επισκέπτες.

Η επανάληψη των αναλύσεων με νέα δεδομένα κάθε έτος θα βελτιώσει σημαντικά την επισκεψιμότητα της περιοχής.

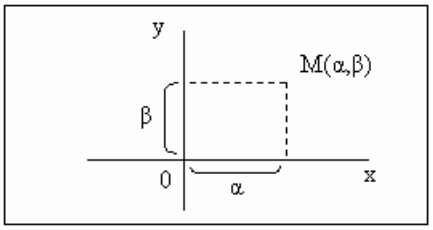
ΣΤΟΙΧΕΙΑ ΑΝΑΛΥΤΙΚΗΣ ΓΕΩΜΕΤΡΙΑΣ ΣΤΟ ΕΠΙΠΕΔΟ

Γενικά

Στόχος της Αναλυτικής Γεωμετρίας είναι να επιλύει γεωμετρικά προβλήματα χρησιμοποιώντας τις συντεταγμένες των σημείων στο καρτεσιανό επίπεδο. Ιστορικά η δημιουργία της μεθόδου των συντεταγμένων οφείλεται στον Έλληνα Απολλώνιο. Η μέθοδός του αναπτύχθηκε στη συνέχεια κατά τον $18^{ο}$ αιώνα από τους Ferma και Descartes (Καρτέσιος), ενώ τεράστια συμβολή στην ανάπτυξη της Αναλυτικής Γεωμετρίας είχε και ο Euler.

Προσδιορισμός ενός σημείου στο καρτεσιανό επίπεδο

Η θέση κάθε σημείου στο επίπεδο καθορίζεται από δύο συντεταγμένες. Την **τετμημένη** και την **τεταγμένη**. Η πρώτη αναφέρεται στη θέση του σημείου Μ ως προς τον οριζόντιο άξονα Χ΄ΟΧ, ενώ η δεύτερη ως προς τον κάθετο άξονα ΥΟΥ΄



σχήμα 1

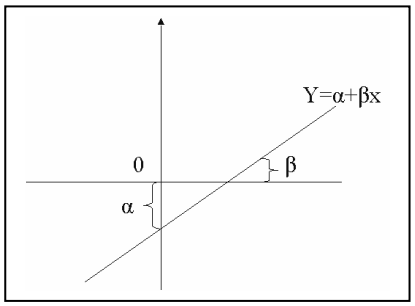
Οι δύο κάθετοι άξονες Χ΄ΟΧ και ΥΟΥ΄ χωρίζουν το επίπεδο σε τέσσερα τεταρτημόρια τα οποία συμβολίζονται με τα λατινικά γράμματα I, II, III, IV.
Τα πρόσημα των συντεταγμένων των σημείων του επιπέδου ανάλογα του τεταρτημορίου που ανήκουν δίδονται στον παρακάτω πίνακα.

Συντεταμένες	Τεταρτημόρια			
	I	II	III	IV
Τετμημένη	+	–	–	+
Τεταγμένη	+	+	–	–

2 Η εξίσωση της ευθείας

Η γενική μορφή της εξίσωσης μιας ευθείας στο επίπεδο δίνεται από τη σχέση

$$y = \alpha x + \beta \qquad (1)$$



σχήμα 2

Η εξίσωση της ευθείας η οποία διέρχεται από τα δύο σημεία $M_1\,(x_1, y_1)$ $M_2\,(x_2, y_2)$ δίνεται από τη σχέση

$$\frac{x - x_1}{x_2 - x_1} = \frac{y - y_1}{y_2 - y_1} \qquad (2)$$

Παράδειγμα
Να βρεθεί η εξίσωση της ευθείας που διέρχεται από τα σημεία $M_1(2, 3)$ και $M_2(1, 4)$ και να παρασταθεί γραφικά.

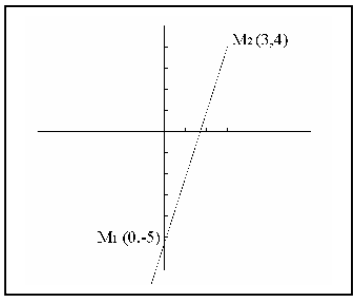
Έχουμε $\dfrac{x-2}{3-2} = \dfrac{y-1}{4-1}$ ή $x - 2 = \dfrac{1}{3}(y-1)$

οπότε $y = 3x - 5.$

Για να κατασκευάσουμε μία ευθεία αρκεί να γνωρίζουμε δύο σημεία της. Στην εξίσωση $y = 3x - 5$ θέτοντας δύο αυθαίρετες τιμές στο x, παίρνουμε αντίστοιχα δύο τιμές του y. Αφού τα δύο ζεύγη (x, y) που προκύπτουν επαληθεύουν την εξίσωση, σημαίνει ότι είναι σημεία της ευθείας.

Επομένως

Το σημείο M_1 με συντεταγμένες $x = 0\,, y = -5$ και το σημείο M_2 με συντεταγμένες $x = 3\,, y = 4$ τα σημειώνουμε ως εξής: $M_1(0, -5)$ και $M_2(3, 4)$. Τα σημεία αυτά τα σχεδιάζουμε στο επίπεδο και στη συνέχεια φέρνουμε την ευθεία που διέρχεται από αυτά.(σχ.3)



σχήμα 3

3 Σχετική θέση δύο ευθειών

Έστω οι εξισώσεις δύο ευθειών $y = \alpha_1 + \beta_1 x$ και $y = \alpha_2 + \beta_2 x$. Οι ίδιες εξισώσεις μπορούν να γραφτούν ως εξής: $A_1 y + B_1 x + C_1 = 0$ και $A_2 y + B_2 x + C_2 = 0$. Αν ισχύει

$$\frac{A_1}{A_2} = \frac{B_1}{B_2} \qquad (3)$$

τότε οι δύο ευθείες είναι παράλληλες. Αν ισχύει

$$\frac{A_1}{A_2} = \frac{B_1}{B_2} = \frac{C_1}{C_2} \qquad (4)$$

τότε οι ευθείες συμπίπτουν.

Στην περίπτωση που οι δύο ευθείες τέμνονται, το σημείο τομής βρίσκεται λύνοντας το σύστημα

$$A_1 y + B_1 x + C_1 = 0$$
$$A_2 y + B_2 x + C_2 = 0$$

Παράδειγμα

Να βρεθεί η σχετική θέση των ευθειών $y = 3x - 5$ και $y = x + 1$.

Οι εξισώσεις των ευθειών γράφονται ως εξής:

$$y - 3x + 5 = 0$$
$$y - x - 1 = 0$$

Επομένως

$A_1=1$, $B_1=-3$, $C_1=5$, $A_2=1$, $B_2=-1$ και $C_2=-1$.

Παίρνουμε τους λόγους

$$\frac{A_1}{A_2} = \frac{1}{1} = 1 \qquad \text{και} \qquad \frac{B_1}{B_2} = \frac{-3}{-1} = 3$$

Επειδή $\frac{A_1}{A_2} \neq \frac{B_1}{B_2}$ οι δύο ευθείες τέμνονται.

Το σημείο τομής Μ είναι:

$$\left.\begin{array}{l} y = 3x - 5 \\ y = x + 1 \end{array}\right\} \Rightarrow \left.\begin{array}{l} x + 1 = 3x - 5 \\ y = x + 1 \end{array}\right\} \Rightarrow \left.\begin{array}{l} x = 3 \\ y = x + 1 \end{array}\right\} \Rightarrow \left.\begin{array}{l} x = 3 \\ y = 4 \end{array}\right\}$$

Δηλαδή Μ(3, 4).

4. Απόσταση δύο σημείων

Η απόσταση δύο σημείων M_1 (x_1,y_1) και $M_2(x_2, y_2)$ δίνεται από τη σχέση

$$d = \sqrt{(x_2 - x_1)^2 + \left(y_2 - y_1^{\,2}\right)} \qquad (5)$$

Παράδειγμα

Να βρεθεί η απόσταση μεταξύ των σημείων $M_1(1, 3)$ και M_2 $(-4, 2)$.
Έχουμε

$$d = \sqrt{(-4-1)^2 + (2-3)^2} = \sqrt{25+1} = \sqrt{26}$$

5 Συνθήκη ώστε δύο ευθείες να είναι κάθετες

Η συνθήκη καθετότητας των δύο ευθειών $y = \alpha_1 x + \beta_1$ και $y = \alpha_2 x + \beta_2$ είναι

$$\alpha_1 \cdot \alpha_2 = 1 \qquad (6)$$

6 Γωνία δύο ευθειών

Η γωνία που σχηματίζουν δύο τεμνόμενες ευθείες $y = \alpha_1 x + \beta_1$ και $y = \alpha_2 x + \beta_2$ βρίσκεται από τη σχέση

$$\varepsilon\phi\theta = \frac{\alpha_2 - \alpha_1}{1 + \alpha_1 \cdot \alpha_2} \qquad (7)$$

7. Συνθήκη ώστε τρία σημεία να κείνται πάνω σε ευθεία

Τρία σημεία $M_1(x_1, y_1)$, $M_2(x_2, y_2)$ και $M_3(x_3, y_3)$ κείνται πάνω σε μία ευθεία όταν ισχύει

$$(x_2 - x_1)\cdot(y_3 - y_1) - (x_3 - x_1)\cdot(y_2 - y_1) = 0 \qquad (8)$$

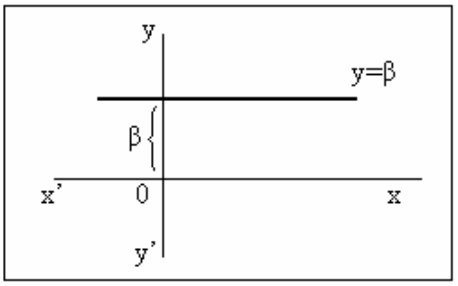
8. Εξίσωση ευθείας διερχόμενης από δοθέν σημείο και παράλληλη προς δοθείσα ευθεία

Η ευθεία που διέρχεται από το σημείο $M(x_1, y_1)$ και είναι παράλληλη προς την ευθεία $y = \alpha x + \beta$ παρίσταται από την εξίσωση

$$y - y_1 = \alpha(x - x_1) \qquad (9)$$

Μια ευθεία παράλληλη προς τον άξονα Χ΄ΟΧ έχει ως εξίσωση

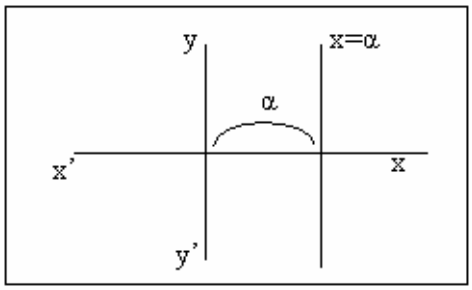
$$y = \beta \qquad (10)$$



σχήμα 4

Μια ευθεία παράλληλη προς τον άξονα Υ΄ΟΥ έχει ως εξίσωση

$$x = \alpha \qquad (11)$$



σχήμα 5

9 Εξίσωση ευθείας διερχόμενης από δοθέν σημείο και κάθετη προς δοθείσα ευθεία

Η ευθεία που διέρχεται από το σημείο $M(x_1, y_1)$ και είναι κάθετη προς την ευθεία $y = \alpha x + \beta$ παρίσταται από την εξίσωση

$$y - y_1 = -\frac{1}{\alpha}(x - x_1) \qquad (12)$$

10 Απόσταση ενός σημείου από μια ευθεία

Η απόσταση d του σημείου $M(x_1, y_1)$ από την ευθεία Ax+By+C=0 είναι ίση με την απόλυτη τιμή της ποσότητας

$$\delta = \frac{Ax_1 + By_1 + C}{\sqrt{A^2 + B^2}} \qquad (13)$$

Δηλαδή

$$d = |\delta|$$

11 Σχετική θέση μιας ευθείας και δύο σημείων

Η σχετική θέση των σημείων $M_1(x_1, y_1)$, $M_2(x_2, y_2)$ και της ευθείας Ax+By+C=0 ορίζεται με βάση τα παρακάτω κριτήρια:

i) Τα σημεία κείνται από την μία πλευρά της ευθείας όταν οι αριθμοί:

$\kappa_1 = Ax_1 + By_1 + C$ και $\kappa_2 = Ax_2 + By_2 + C$ έχουν το ίδιο πρόσημο.

ii) Τα σημεία κείνται εκατέρωθεν της ευθείας αν οι αριθμοί κ_1 και κ_2 έχουν διαφορετικά πρόσημα.

iii) Ένα ή και τα δύο σημεία κείνται επί της ευθείας όταν ένας ή και οι δύο αριθμοί (κ_1 και κ_2) είναι ίσοι με μηδέν.

Παράδειγμα

Τα σημεία $M_1(2, 6)$, $M_2(-4, -2)$ βρίσκονται εκατέρωθεν της ευθείας $3x + 5y - 1 = 0$ επειδή οι αριθμοί

$$\kappa_1 = 3 \cdot 2 + 5 \cdot 6 - 1 = 35 \text{ και } \kappa_2 = 3 \cdot (-4) + 5(-2) - 1 = -23$$

έχουν αντίθετο πρόσημο.

12 Διαίρεση ευθυγράμμου τμήματος σε μέρη ανάλογα ως προς δοθέντα λόγο

Δίνονται τα σημεία $M_1(x_1, y_1)$ και $M_2(x_2, y_2)$. Ζητούνται οι συντεταγμένες (x, y) του σημείου Κ που διαιρεί το ευθύγευθύγραμμο τμήμα M_1M_2 έτσι ώστε

$$\frac{M_1K}{KM_2} = \frac{\lambda_1}{\lambda_2} = \rho$$

όπου λ_1, λ_2 δοθέντες αριθμοί.

Οι συντεταγμένες του σημείου Κ(x, y) δίνονται από τους τύπους.

$$x = \frac{\lambda_1 x_1 + \lambda_1 x_2}{\lambda_1 + \lambda_2} \qquad (14)$$

και

$$y = \frac{\lambda_2 y_1 + \lambda_2 y_2}{\lambda_1 + \lambda_2} \qquad (15)$$

ή

$$x = \frac{x_1 + \rho x_2}{1 + \rho} \qquad (16)$$

και

$$y = \frac{y_1 + \rho y_2}{1 + \rho} \qquad (17)$$

Παράδειγμα

Δίνεται το σημείο Β (9, –4) και η αρχή των αξόνων Ο (0,0). Να βρεθεί το σημείο Κ που διαιρεί το ευθύγραμμο τμήμα ΒΟ με λόγο 2:3.
Δίνονται $\lambda_1=2$, $\lambda_2=3$, $x_1=9$, $y_1=-4$, $x_2=0$, $y_2=0$.

Χρησιμοποιώντας του τύπους, 14 και 15 βρίσκουμε ότι

$$x = \frac{18}{5} \qquad \text{και} \qquad y = -\frac{12}{5} \qquad \text{άρα} \quad K\left(\frac{18}{5}, -\frac{12}{5}\right)$$

13 Μετασχηματισμός των ορθογωνίων συντεταγμένων

Η ίδια καμπύλη ανάλογα με το σύστημα συντεταγμένων που χρησιμοποιείται παρουσιάζει διαφορετική εξίσωση.

Μερικές φορές έχουμε την ανάγκη ενώ γνωρίζουμε την εξίσωση μιας καμπύλης σ' ένα συγκεκριμένο σύστημα συντεταγμένων, να βρούμε την εξίσωση της καμπύλης σ' ένα νέο σύστημα το οποίο προέκυψε είτε με παράλληλη μετατόπιση της αρχής των αξόνων είτε με στροφή περί την αρχή Ο κατά γωνία α.

I. Όταν έχουμε παράλληλη μετατόπιση

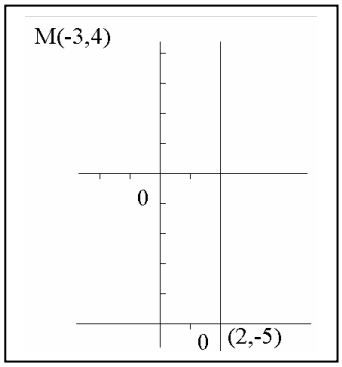
Έστω το σημείο Μ(x, y). Θεωρούμε ότι η αρχή Ο μετατοπίζεται στο σημείο Ο΄(x_o, y_o) και ότι οι νέες συντεταγμένες του σημείου Μ είναι Μ(x΄, y΄). Οι τύποι που συνδέουν τις παλαιές με τις νέες συντεταγμένες είναι

$$x' = x - x_0 \qquad \text{και} \qquad y' = y - y_0 \tag{18}$$

Παράδειγμα

Η αρχή των αξόνων του συστήματος Χ΄ΟΧ΄ μεταφέρθηκε στο σημείο Ο΄(2, –5). Να βρεθούν οι νέες συντεταγμένες του σημείου Μ (–3, 4).
Δίδονται $x_0 = 2$, $y_0 = -5$, $x = -3$, $y = 4$. Άρα

$$x' = x - x_0 = -3 - 2 = -5$$
$$y' = y - y_0 = 4 - (-5) = 9$$



σχήμα 4.6

II. Όταν έχουμε στροφή περί την αρχή Ο κατά γωνία α

Οι τύποι που δίνουν τις νέες συντεταγμένες ως προς τις παλαιές είναι

$$\begin{aligned} x' &= x\sigma\upsilon\nu\alpha + y\eta\mu\alpha \\ y' &= -x\eta\mu\alpha + y\sigma\upsilon\nu\alpha \end{aligned} \tag{19}$$

Ισχύουν επίσης οι σχέσεις:

$$\begin{aligned} x &= x'\sigma\upsilon\nu\alpha - y'\eta\mu\alpha \\ y &= x\eta\mu\alpha + y'\sigma\upsilon\nu\alpha \end{aligned} \tag{20}$$

Παράδειγμα 1

Έστω το σημείο Μ (6, 0) και ότι το σύστημα αξόνων περιστρέφεται κατά γωνία -20^0. Να βρεθούν οι νέες συντεταγμένες.
Έχουμε

$$x' = 6\sigma\upsilon\nu(-20) + 0 \cdot \eta\mu(-20) \cong 5{,}64$$
$$y' = -6\eta\mu(-20) + 0 \cdot \sigma\upsilon\nu(-20) \cong 2{,}05$$

Παράδειγμα 2

Έστω η καμπύλη 2xy=49 το δε σύστημα συντεταγμένων στρέφεται 45^0. Να βρεθεί η εξίσωση της καμπύλης στο νέο σύστημα.
Χρησιμοποιώντας τις εξισώσεις 20 έχουμε:

$$x = x' \cdot \frac{\sqrt{2}}{2} - y' \cdot \frac{\sqrt{2}}{2}$$

$$y = x' \cdot \frac{\sqrt{2}}{2} + y' \cdot \frac{\sqrt{2}}{2}$$

Αντικαθιστούμε τις τιμές αυτές στην εξίσωση της καμπύλης

$$2 \cdot \left(x' \cdot \frac{\sqrt{2}}{2} - y' \cdot \frac{\sqrt{2}}{2} \right) \left(x' \cdot \frac{\sqrt{2}}{2} + y' \cdot \frac{\sqrt{2}}{2} \right) = 49 \qquad \text{ή}$$

$$2 \frac{\sqrt{2}}{2} \cdot \frac{\sqrt{2}}{2} (x' - y')(x' + y') = 49 \qquad \text{ή}$$

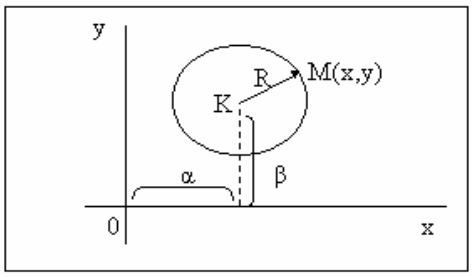
$$x'^2 - y'^2 = 49$$

η οποία αποτελεί την ζητούμενη εξίσωση.

14 Εξίσωση κύκλου

Η εξίσωση ενός κύκλου με ακτίνα R και κέντρο το σημείο K(α, β) δίνεται από την παρακάτω σχέση:

$$(x - \alpha)^2 + (y - \beta)^2 = R^2 \qquad (21)$$



σχήμα 7

Αντιστρόφως κάθε εξίσωση της μορφής

$$Ax^2 + Bx + Ay^2 + Cy + D = 0 \qquad (22)$$

παριστά ένα κύκλο όταν οι συντελεστές A, B, C, D ικανοποιούν την σχέση

$$B^2 + C^2 - 4AD > 0 \qquad (22\alpha)$$

Στην περίπτωση αυτή το κέντρο του κύκλου έχει ως συντεταγμένες τις

$$\alpha = -\frac{B}{2A} \qquad \text{και} \qquad \beta = -\frac{C}{2A} \qquad (23)$$

ενώ η ακτίνα R του κύκλου είναι ίση με

$$R = \sqrt{\frac{B^2 + C^2 - 4AD}{4A^2}} \tag{24}$$

Παράδειγμα

Να βρεθεί η ακτίνα και το κέντρο του κύκλου που δίνεται από την εξίσωση

$$5x^2 - 10x + 5y^2 + 20y - 20 = 0.$$

Η δοθείσα εξίσωση παριστάνει πράγματι κύκλο επειδή ισχύει

$$B^2 + C^2 = 4AD = (-10)^2 + (+20)^2 - 4 \cdot 5(-20) = 100 > 0$$

Επομένως το κέντρο του κύκλου έχει ως συντεταγμένες

$$\alpha = -\frac{(-10)}{2 \cdot 5} = 1 \qquad \beta = -\frac{20}{2 \cdot 5} = -2 \qquad \text{ήτοι} \quad K(1,\ -2)$$

και ακτίνα

$$R = \sqrt{\frac{(-10)^2 + 20^2 - 4 \cdot 5(-20)}{4 \cdot 5^2}} = 1$$

15. Ευθύ άθροισμα

Ορίζουμε το διανυσματικό χώρο V ως ευθύ άθροισμα των υποχώρων του U και W όπου $U \cap W = \{\square\}$ αν για κάθε στοιχείο $\upsilon \in V$ υπάρχουν στοιχεία $u \in U$ και $w \in W$ μοναδικά, τέτοια ώστε

$$\upsilon = u + w \tag{25}$$

οπότε γράφουμε

$$V = U \oplus W \tag{26}$$

Θεώρημα: Αν V είναι ένας διανυσματικός χώρος με πεπερασμένη διάσταση και επιπλέον είναι ευθύ άθροισμα των διανυσματικών υποχώρων U και W, τότε

$$\dim V = \dim U + \dim W \tag{27}$$

Απόδειξη:

Έστω $\{u_1,\ u_2,\ ...,\ u_r\}$ μια βάση του U και $\{w_1,\ w_2,\ ..., w_s\}$ μια βάση του W. Κάθε στοιχείο $u \in W$ εκφράζεται κατά τρόπο μοναδικό ως γραμμικός συνδυασμός των στοιχείων της βάσης του, δηλαδή

$$\alpha_1 u_1 + x_2 u_2 + ... + x_r u_r .$$

Επίσης κάθε στοιχείο $w \in W$ εκφράζεται κατά τρόπο μοναδικό ως γραμμικός συνδυασμός των στοιχείων της βάσης του, δηλαδή

$$w = y_1 w_1 + y_2 w_2 + ... + y_s w_s$$

Από την υπόθεση έχουμε ότι $V = U \oplus W$ δηλαδή κάθε στοιχείο $\upsilon \in V$ εκφράζεται κατά τρόπο μοναδικό ως άθροισμα των u και w.
Δηλαδή $\upsilon = u + w$ ή ακόμη

$$\upsilon = x\ u_1 + x_2 u_2 + ... + x_r u_r + y_1 w_1 + y_2 w_2 + ... + y_s w_s$$

που φανερώνει ότι τα στοιχεία $u_1, u_2, ..., u_r, w_1, w_2, ..., w_r$ είναι μια βάση του V οπότε αποδεικνύεται και το θεώρημα.

Σκόπιμο είναι σε αυτό το σημείο να παρουσιαστεί η μέθοδος GRAM – SCHMIDT–HILBERT η οποία μετασχηματίζει ένα Πλαγιογώνιο σύστημα συντεταγμένων σε ΟΡΘΟΚΑΝΟΝΙΚΟ μια διαδικασία που είναι επιβεβλημένη ώστε να εφαρμοστεί το Πυθαγόρειο θεώρημα στο Ευκλείδειο ν-διάστατο διανυσματικό χώρο R^{v}.

Από την ελεύθερη Εγκυκλοπαίδεια ΒΙΚΙΠΑΙΔΕΙΑ πληροφορούμαστε για τα αξιώματα του Χιλμπερτ, πάνω στα οποία στηρίζεται αποκλειστικά η επίλυση των προβλημάτων χρησιμοποιώντας την μέθοδο KARAP.

ΑΞΙΩΜΑΤΑ ΧΙΛΜΠΕΡΤ ΤΗΣ ΕΥΚΛΕΙΔΕΙΑΣ ΓΕΩΜΕΤΡΙΑΣ

Τα **αξιώματα Χίλμπερτ** της **Ευκλείδειας Γεωμετρίας** ορίζονται ως εξής:

1. Έστω Χ ένα μη κενό **σύνολο** που τα στοιχεία του ονομάζουμε **σημεία** {Α,Β,Γ,...}. Το σύνολο Χ θα το λέμε **Γεωμετρικό χώρο**. Κάθε υποσύνολο του Γεωμετρικού χώρου θα το λέμε **Σχήμα.**
2. Μέσα στο Γεωμετρικό χώρο δεχόμαστε δυο βασικές κατηγορίες από υποσύνολα, τις **ευθείες** {α,β,γ,...} και τα **επίπεδα** {P,Q,R,S,..., }

Τα είδη των μαθηματικών αντικειμένων

Στο σύστημα του **Χίλμπερτ** τα αρχικά μαθηματικά αντικείμενα είναι τριών ειδών: τα «σημεία», οι «ευθείες» και τα «επίπεδα», που συνδέονται μεταξύ τους με τις σχέσεις του «ανήκειν», του «μεταξύ» και της «ισοδυναμίας».

Το σύστημα του Χίλμπερτ εξετάζει τις αρχικές αυτές έννοιες και τις σχέσεις τους και οι πέντε ομάδες αξιωμάτων που εισάγει συνιστούν έμμεσο ορισμό των αρχικών αντικειμένων και των σχέσεων τους.

1. (Ι) Τα **αξιώματα** σύνδεσης («ανήκειν») ορίζουν τις ιδιότητες της αμοιβαίας θέσης μεταξύ σημείων, ευθειών και επιπέδων .
2. (ΙΙ) Τα αξιώματα διάταξης ορίζουν τις ιδιότητες της αμοιβαίας θέσης σημείων πάνω σε μια ευθεία ή ένα επίπεδο.
3. (ΙΙΙ)Τα αξιώματα σύνδεσης ισοδυναμίας ορίζουν την έννοια της «ισότητας» δύο τμημάτων ή γωνιών.
4. (IV) Τα αξιώματα συνέχειας .
5. (V) Το αξίωμα παραλληλίας

Τα αξιώματα σύνδεσης είναι οκτώ

- (I_1) Από οποιαδήποτε δύο σημεία διέρχεται μία μόνο ευθεία.
- (I_2) Σε κάθε ευθεία υπάρχουν τουλάχιστον δύο σημεία.
- (I_3) Υπάρχουν τουλάχιστον τρία σημεία που δεν κείνται στην ίδια ευθεία.
- (I_4) Από οποιαδήποτε τρία σημεία που δεν κείνται στην ίδια ευθεία,διέρχεται ένα μόνο επίπεδο.
- (I_5) Σε οποιοδήποτε επίπεδο υπάρχει πάντοτε ένα σημείο που ανήκει σε αυτό.

- (I_6) Αν δύο σημεία βρίσκονται σε ένα επίπεδο, τότε και η ευθεία που διέρχεται από τα σημεία αυτά βρίσκεται σ' αυτό το επίπεδο.
- (I_7) Αν δύο επίπεδα έχουν κοινό σημείο, τότε έχουν τουλάχιστον ένα ακόμα κοινό σημείο.
- (I_8) Υπάρχουν τουλάχιστον τέσσερα σημεία που δεν βρίσκονται στο ίδιο επίπεδο.

Τα αξιώματα διάταξης είναι τέσσερα:

- (II_1) Από τρία διαφορετικά σημεία μιας ευθείας ένα και μόνον ένα βρίσκεται μεταξύ των δύο άλλων.
- (II_2) Για οποιαδήποτε δύο σημεία Α και Γ υπάρχει τουλάχιστον ένα σημείο Β στην ευθεία ΑΓ τέτοιο, ώστε το σημείο Γ να βρίσκεται μεταξύ του Α και του Β.
- (II_3) Για οποιαδήποτε τρία σημεία μιας ευθείας υπάρχει όχι περισσότερο από ένα σημείο που βρίσκεται μεταξύ των δύο άλλων. Η σχέση του «μεταξύ» για σημεία σε μια ευθεία μας επιτρέπει να ορίσουμε την έννοια του ευθύγραμμου τμήματος.
- (II_4) Έστω Α, Β, Γ τρία σημεία που δε βρίσκονται στην ίδια ευθεία και έστω ε ευθεία στο επίπεδο των Α, Β, Γ που δε διέρχεται από κανένα από τα σημεία Α, Β, Γ. Αν η ευθεία ε διέρχεται από ένα σημείο του ευθύγραμμου τμήματος ΑΒ, τότε πρέπει να διέρχεται κι από ένα σημείο του τμήματος ΑΓ ή από ένα σημείο του τμήματος ΒΓ Αξίωμα (Moritz Pasch.

Τα αξιώματα σύνδεσης ισοδυναμίας είναι πέντε:

- (III_1) Αν Α και Β είναι δύο διαφορετικά σημεία στην ευθεία ε και Α' είναι ένα σημείο της ίδιας ευθείας ή άλλης ευθείας ε', τότε μπορεί πάντοτε να βρεθεί σημείο Β' που βρίσκεται στο δεδομένο από το σημείο Α ' μέρος της ευθείας ε' τέτοιο, ώστε το τμήμα ΑΒ να είναι ισοδύναμο (ίσο) με το τμήμα Α' Β'.
- (III_2) Αν δύο τμήματα είναι ισοδύναμα προς τρίτο, τότε είναι και μεταξύ τους ισοδύναμα.
- (III_3) Έστω ΑΒ και ΒΓ δύο τμήματα της ευθείας ε που δεν έχουν κοινό σημείο και έστω επίσης Α' Β' και Β'Γ' δύο τμήματα της ίδιας ευθείας ή άλλης ευθείας ε' που επίσης δεν έχουν κοινό σημείο. Αν τώρα ΑΒ=Α'Β', ΒΓ=Β'Γ', τότε και ΑΓ=Α'Γ'.

Η γωνία ορίζεται ως το σχήμα που αποτελείται από δύο διαφορετικές ημιευθείες με κοινό αρχικό σημείο.

- (III_4) Από δεδομένη ημιευθεία σε δεδομένο ημιεπίπεδο που ορίζεται από αυτή την **ημιευθεία** και την προέκταση της , μπορεί να σχηματιστεί μια μοναδική γωνία ισοδύναμη με τη δεδομένη γωνία.
- (III_5) Αν δύο τρίγωνα ΑΒΓ και $A_1B_1Γ_1$ έχουν ΑΒ=A_1B_1 ΑΓ=$A_1Γ_1$ και γωνίαA=*γωνία*A_1, τότε και γωνίαB=*γωνία*B_1,γωνία$Γ$=*γωνία*$Γ_1$.

Τα αξιώματα συνέχειας είναι δύο

- (IV_1) Έστω ΑΒ και ΓΔ δύο οποιαδήποτε τμήματα. Τότε στην ευθεία ΑΒ υπάρχει πεπερασμένος αριθμός σημείων A_1, A_2, ..., A_v, τέτοιων ώστε τα τμήματα

$AA_1, A_1A_2, ..., A_{v-1}A_v$ να είναι ισοδύναμα με το τμήμα ΓΑ και το σημείο Β να βρίσκεται μεταξύ Α και Αν **(Αρχιμήδεια ιδιότητα).**

- (IV_2) Τα σημεία μιας ευθείας σχηματίζουν σύστημα, το οποίο, τηρούμενης της γραμμικής διάταξης, του πρώτου αξιώματος ισοδυναμίας και του αξιώματος Ευδόξου-Αρχιμήδη δεν είναι επεκτάσιμο, δηλ. σ' αυτό το σύστημα σημείων δεν είναι δυνατόν να προστεθεί ένα ακόμα σημείο, έτσι ώστε στο επεκτεταμένο σύστημα που αποτελείται από το αρχικό σύστημα και το συμπληρωματικό σημείο να ικανοποιούνται τα παραπάνω αξιώματα (αξίωμα γραμμικής πληρότητας).

Το αξίωμα παραλληλίας

- Έστω ε τυχούσα ευθεία και σημείο Α εκτός αυτής. Στο επίπεδο που ορίζεται από την ευθεία ε και το σημείο Α υπάρχει όχι περισσότερο από μία ευθεία που διέρχεται από το σημείο Α και δεν τέμνει την ευθεία ε.

Τα θεωρήματα (ύπαρξης) του Γεωμετρικού Χώρου

1. Θεώρημα: Αν τρία σημεία είναι πάνω σε ευθεία τότε κάθε ένα από αυτά είναι πάνω στην ευθεία που ορίζουν τα δυο άλλα.

2.Θεώρημα: Υπάρχει σημείο που είναι έξω από δοσμένη ευθεία.

3. Θεώρημα: Αν δυο επίπεδα έχουν ένα κοινό σημείο τότε η τομή τους είναι ευθεία.

4. Θεώρημα: Αν τέσσερα σημεία δεν είναι πάνω στο αυτό επίπεδο, τότε δεν υπάρχει τριάδα από αυτά που να είναι πάνω σε ευθεία.

5. Θεώρημα: Υπάρχει σημείο έξω από δοσμένο επίπεδο.

6. Θεώρημα: Υπάρχουν δυο ευθείες που δεν είναι πάνω στο ίδιο επίπεδο.

7. Θεώρημα: Υπάρχουν δυο επίπεδα.

8. Θεώρημα: Ευθεία και σημείο έξω από αυτή ορίζουν ακριβώς ένα επίπεδο.

9. Θεώρημα: Από δυο τεμνόμενες ή παράλληλες ευθείες ορίζεται ακριβώς ένα επίπεδο.

10. Θεώρημα: (Αστέρος) Εάν ευθείες τέμνονται ανά δυο και δεν ανήκουν στο ίδιο επίπεδο, τότε περνούν από το ίδιο σημείο.

11. Θεώρημα: (Ημιεπίπεδο) Ας θεωρήσουμε ένα επίπεδο S και μία ευθεία του α. Τα σημεία τα διαφορετικά της ευθείας α να ανήκουν σε δυο υποσύνολα, στο S_1 και στο S_2 . Τα υποσύνολα S_1, S_2 τα καθορίζουμε έτσι,

> E_1: Αν το σημείο Α ανήκει στο S_1 και το Β ανήκει στο S_1 το ευθύγραμμο τμήμα ΑΒ δεν έχει κοινά σημεία με την α;

E_2 : Αν Α ανήκει στο S_1 και Γ ανήκει στο S_2 τότε το ΑΓ έχει κοινό σημείο με την α. Τα σύνολα S_1,α,S_2 αποτελούν ένα διαμερισμό του S. Τα σημειοσύνολα S_1, και S_2 ονομάζονται ανοικτά Ημιεπίπεδα ενώ S_1 με την α ή S_2 με την α κλειστά Ημιεπίπεδα με αρχική ευθεία την α και φορέα το S. Τα Ημιεπίπεδα με την αρχική ημιευθεία και τον ίδιο φορέα ονομάζονται αντίθετα.

12. Θεώρημα: (Θέση σημείου και επιπέδου) Ας θεωρήσουμε ένα επίπεδο S και ένα σημείο Α τότε

E_1: το Α ανήκει στο S
E_2: το Α δεν ανήκει στο S

13. Θεώρημα: (Ημίχωρος) Θεωρούμε επίπεδο S τότε και τα σημεία του χώρου διαμερίζονται σε τρία σύνολα X_1 X_2S:

E_1: Αν τα σημεία {Α,Β}ανήκει στο X_1, ή X_2 το ευθύγραμμο τμήμα ΑΒ δεν τέμνει το S και
E_2: Αν Α ανήκει στο X_1Γ ανήκει στο X_2 τότε ΑΓ και S έχουν ένα κοινό σημείο με το S. (ίχνος ευθείας και επιπέδου).
E_3: Τα σημειοσύνολα X_1 X_2 τα ονομάζουμε ανοιχτούς ημίχωρους, ενώ τα X_1 με το S ή X_2με το S κλειστούς ημίχωρους. Τα σημεία Α,Β λέμε ότι είναι προς το ίδιο μέρος του S ενώ τα Α,Γ από τη μια και την άλλη μεριά του S.

14. Θεώρημα: (θέση ευθείας και επίπεδου) Θεωρούμε ευθεία α και επίπεδο S, Τότε,

E_1: Αν ή ευθεία α και το επίπεδο S έχουν δυο κοινά σημεία τότε, από το αξίωμα ΙΙΙ, η α είναι υποσύνολο του S ή βρίσκεται στο S (α⊂S)
E_2: Αν η ευθεία α και το S έχουν ένα κοινό σημείο α τομή S ={Α} (α∩S={Α}) τότε το Α λέγεται κοινό σημείο τής ευθείας με το επίπεδο ή ίχνος, και θα λέμε ότι ή ευθεία και το επίπεδο τέμνονται.
E_3: Εάν α τομή S=κενό τότε λέμε ότι η α είναι παράλληλη στο S (α// S).

15. Θεώρημα: (Θέση δυο ευθειών) Θεωρούμε δυο ευθείες α και β.

E_1: Αν οι α, β είναι υποσύνολα του ίδιου επιπέδου, τότε λέγονται συνεπίπεδες ή συμβατές.

Άρα ή θα είναι παράλληλές ή θα τέμνονται ή, θα ταυτίζονται»..

E_2: Οι α, β δεν ανήκουν στο ίδιο επίπεδο θα τις λέμε ασύμβατες ή στρεβλές.

16. Θεώρημα θέσης δύο επιπέδων Έστω δυο επίπεδα P,S.

E_1: Αν τα επίπεδα περιέχουν τρία σημεία Α,Β,Γ διάφορα που δεν βρίσκονται στην ίδια ευθεία, λέμε ότι ταυτίζονται P τομή S = P=S ή P=S (P∩S=P=S).
Δυο επίπεδα που δεν ταυτίζονται λέγονται διάφορα ή διακεκριμένα (P διάφορο S) .

E_2: Αν σημείο A□P και A□S και P διάφορο του S, τότε έχον κοινή και μία ευθεία και θα λέμε ότι τέμνονται κατά μία ευθεία, έστω α δηλαδή P τομή S=α (P∩S=α).

E_3: Αν P τομή S=κενό τότε τα επίπεδα λέγονται παράλληλα. P// S.

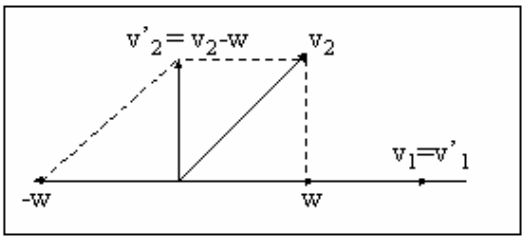
Μέθοδος του GRAM – SCHMIDT–HILBERT

Έστω μια βάση $B=\{\upsilon_1,\upsilon_2...,\upsilon_v\}$ ενός διανυσματικού χώρου V. Με τη μέθοδο του Gram - Schmidt θα κατασκευάσουμε μια κανονική βάση $B'=\{\upsilon_1',\upsilon'_2,...,\upsilon'_v\}$ του V χρησιμοποιώντας τα στοιχεία της B.

Η διαδικασία που ακολουθείται είναι η παρακάτω, αφού προηγουμένως συμβολίσουμε το εσωτερικό γινόμενο δύο διανυσμάτων με <A, B>.

Ως πρώτο διάνυσμα της κανονικής βάσης θεωρούμε ότι είναι το υ_1. Δηλαδή $\upsilon_1'=\upsilon_1$.

Το δεύτερο διάνυσμα της B΄ το βρίσκουμε από το υ_2, αφού του αφαιρέσουμε την προβολή του πάνω στο υ_1' (σχ.8).



σχήμα 8

Το τρίτο διάνυσμα της B΄ το βρίσκουμε από το υ_3, αφού του αφαιρέσουμε την προβολή του πάνω στο υ_2' και την προβολή του στο υ_1', κ.ο.κ.

Δηλαδή η διαδικασία προβλέπει

$$\upsilon_1'=\upsilon_1$$

$$\upsilon_2'=\upsilon_2-\frac{<\upsilon_2,\upsilon_1'>}{<\upsilon_1',\upsilon_1'>}\cdot\upsilon_1'$$

$$\upsilon_3'=\upsilon_3-\frac{<\upsilon_3,\upsilon_2'>}{<\upsilon_2',\upsilon_2'>}\cdot\upsilon_2'-\frac{<\upsilon_3,\upsilon_1'>}{<\upsilon_1',\upsilon_1'>}\cdot\upsilon_1'$$

.

.

$$\upsilon_v'=\upsilon_v-\frac{<\upsilon_v,\upsilon'_{v-1}>}{<\upsilon'_{v-1},\upsilon'_{v-1}>}\cdot\upsilon'_{v-1}-...-\frac{<\upsilon_v,\upsilon_1'>}{<\upsilon_1',\upsilon_1'>}\cdot\upsilon_1'$$

Παράδειγμα:

Δίνονται τα παρακάτω διανύσματα A = (1, 1, 0), B = (1, –2, 0), Γ=(1,0,–1) τα οποία αποτελούν βάση του R^3. Να βρεθεί μια κανονική βάση του R^3, χρησιμοποιώντας τα διανύσματα A, B, Γ.

Θέτουμε

$$A' = A'$$

$$B' = B - \frac{< B, A' >}{< A', A' >} \cdot A'$$

$$\Gamma' = \Gamma - \frac{< \Gamma, B' >}{< B', B' >} \cdot B - \frac{< \Gamma, A' >}{< A', A' >} \cdot A'$$

Συνεπώς, $A' = A = (1, 1, 0)$

Για να βρούμε το Β΄ αφαιρούμε από το Β την προβολή του Β πάνω στο Α.

$$B' = (1,-2,0) - \frac{1-2+0}{1+1+0} \cdot (1,1,0) = (1,-2,0) + \frac{1}{2} \cdot (1,1,0) = \left(\frac{3}{2},-\frac{3}{2},0\right)$$

$$\Gamma' = (1,0,-1) - \frac{\frac{3}{2}+0+0}{\frac{9}{4}+\frac{9}{4}+0} \cdot \left(\frac{3}{2},-\frac{3}{2},0\right) - \frac{1+0+0}{1+1+0} \cdot (1,1,0) = (1,0,-1) -$$

$$-\frac{1}{3} \cdot \left(\frac{3}{2},-\frac{3}{2},0\right) - \frac{1}{2} \cdot (1,1,0) = \left(1-\frac{1}{2}-\frac{1}{2},0+\frac{1}{2}-\frac{1}{2},-1,0,0\right) = (0,0,-1)$$

Παρατηρούμε ότι τα διανύσματα $A' = (1, 1, 0)$, $B' = \left(\frac{3}{2},-\frac{3}{2},0\right)$ και

$\Gamma' = (0, 0, -1)$ είναι ανά δύο ορθογώνια αφού ισχύει:

$$A' \cdot B' = 0,\ A' \cdot \Gamma' = 0,\ B' \cdot \Gamma' = 0$$

Βρίσκουμε ότι είναι και γραμμικώς ανεξάρτητα αφού για να ισχύει

$\lambda_1(1,1,0) + \lambda_2\left(\frac{3}{2},-\frac{3}{2},0\right) + \lambda_3(0,0,-1) = 0$ πρέπει να είναι $\lambda_1 = \lambda_2 = \lambda_3 = 0$.

Η κανονική βάση Α΄, Β΄, Γ΄ για να γίνει ορθοκανονική αρκεί να διαιρεθεί κάθε διάνυσμα με το μέτρο του.
Δηλαδή:

$$A'' = \frac{A'}{\|A'\|} = \frac{1}{\sqrt{2}}(1,1,0) = \left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}},0\right)$$

$$B'' = \frac{B'}{\|B'\|} = \frac{1}{\sqrt{\frac{18}{4}}}\left(\frac{3}{2},-\frac{3}{2},0\right) = \left(\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}},0\right)$$

$$\Gamma'' = \frac{\Gamma'}{\|\Gamma'\|} = \frac{1}{1}(0,0,-1) = (0,0,-1)$$

Η κατασκευή μιας ορθοκανονικής βάσης ξεκινώντας από μια κανονική (ορθογώνια) βάση, αποτελεί την μέθοδο του Hilbert.

Τελειώνοντας την παράγραφο που αναφέρεται στο διανυσματικό λογισμό, αναφέρουμε ένα χρήσιμο θεώρημα χωρίς απόδειξη.

Θεώρημα: Έστω V ένας διανυσματικός χώρος ν διαστάσεων εφοδιασμένος με ένα μη αρνητικό εσωτερικό γινόμενο. Έστω W ένας διανυσματικός υποχώρος p διαστάσεων του V, όπου $0<p<\nu$. Έστω $W^{\perp}$ ο διανυσματικός υποχώρος του V αποτελούμενος απ' όλα τα διανύσματα τα κάθετα στο W. Τότε ο διανυσματικός χώρος V είναι ευθύ άθροισμα των υποχώρων W και $W^{\perp}$ καθώς επίσης η διάσταση του $W^{\perp}$ είναι $\dim W^{\perp} = \nu - p$.
Επομένως

$$\dim W + \dim W^{\perp} = \dim V \qquad (28)$$

Ο υποχώρος $W^{\perp}$ λέγεται ορθογώνιος υποχώρος του W.

Ο αλγόριθμος κατασκευής ορθοκανονικής βάσης με την μέθοδο GRAM – SCHMIDT–HILBERT

Ο ακόλουθος αλγόριθμος MATLAB υλοποιεί την μέθοδο Gram-Schmidt-Hilbert για Ευκλείδεια διανύσματα. Τα διανύσματα $v_1, ..., v_k$ (στήλες της μήτρας V, έτσι ώστε το V (:, j) είναι το j^{th} διάνυσμα) αντικαθίστανται από ορθοκανονικά διανύσματα (στήλες του U) που καλύπτουν τον ίδιο υποχώρο.

```
n = size(V,1);
k = size(V,2);
U = zeros(n,k);
U(:,1) = V(:,1)/sqrt(V(:,1)'*V(:,1));
for i = 2:k
   U(:,i) = V(:,i);
  for j = 1:i-1
   U(:,i) = U(:,i) - ( U(:,i)'*U(:,j) )/( U(:,j)'*U(:,j) )*U(:,j);
  end
U(:,i) = U(:,i)/sqrt(U(:,i)'*U(:,i));
end
```