ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΥΔΡΟΧΗΜΙΚΗ ΜΕΛΕΤΗ ΥΠΟΓΕΙΩΝ ΝΕΡΩΝ ΜΕ ΧΡΗΣΗ ΓΕΩΣΤΑΤΙΚΩΝ ΜΕΘΟΔΩΝ

ΟΝΟΜΑΤΕΠΩΝΥΜΟ ΣΠΟΥΔΑΣΤΩΝ **ΑΜ**

ΧΑΛΛΑΣ ΑΝΤΩΝΙΟΣ **6348**

ΠΑΠΠΑΣ ΔΗΜΗΤΡΙΟΣ **6077**

ΟΝΟΜΑΤΕΠΩΝΥΜΟ-ΤΙΤΛΟΣ ΕΠΙΒΛΕΠΟΝΤΟΣ

ΠΑΝΑΓΟΠΟΥΛΟΣ ΓΕΩΡΓΙΟΣ-ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ

ΠΑΤΡΑ 2020

ΠΡΟΛΟΓΟΣ

Πανεπηστήμιο Πελοποννήσου Τμήμα Μηχανολόγων Μηχανικών

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η διερεύνηση του υδρογεωλογικού και υδροχημικού καθεστώτος της ευρύτερης περιοχής της Ζαχάρως με σκοπό την μελέτη των υπόγειων νερών με χρήση γεωστατικών μεθόδων.Η ανάθεση του θέματος έγινε την 2-6-2020 από τον υπεύθυνο καθηγητή χημείας και υπο την πλήρη επίβλεψη του κ.Παναγόπουλο Γεώργιο.Ο τίτλος της διπλωματικής εργασίας είναι: <<Υδροχημική μελέτη υπόγειων νερών με χρήση γεωστατικών μεθόδων>>.Ο σκοπός της συγκεκριμένης εργασίας είναι να κατανοήσουμε πως λειτουργεί το κάθε χημικό στοιχείο στο νερό και πως το επηρεάζει και να βοηθήσουμε την περιοχή με αυτόν τον τρόπο στο να βρει λύσεις για τη βελτίωση του υδροφόρου ορίζοντα.Στη συγκεκριμένη εργασία περιλαμβάνοντε κάποιες λεπτομερείς διαδικασίες με τις οποίες αναλύουμε παραστατικά τη σύσταση των υπόγειων νερών.Τέλος θα θέλαμε να ευχαριστήσουμε τον επιβλέποντα καθηγητή μας κ.Παναγόπουλο Γεώργιο για την ανάθεση του θέματος,την εμπιστοσύνη που έδειξε στο προσωπό μας,την καθοδήγησή του κατά τη διάρκεια της διατριβής,την πολύτιμη βοηθειά του και τις καθοριστικές υποδείξεις του στη διαμόρφωση του τελικού κειμένου καθώς και τη συνεργασία του σε όλα τα θέματα που αφορούσαν τη συγκεκριμένη διπλωματική εργασία.

<u>Υπεύθυνη Δήλωση Φοιτητών</u>: Οι κάτωθι υπογεγραμμένοι Φοιτητές έχουμε επίγνωση των συνεπειών του Νόμου περί λογοκλοπής και δηλώνουμε υπεύθυνα ότι είμαστε συγγραφείς αυτής της Διπλωματικής Εργασίας, αναλαμβάνοντας την ευθύνη επί ολοκλήρου του κειμένου εξ ίσου, έχουμε δε αναφέρει στην Βιβλιογραφία μας όλες τις πηγές τις οποίες χρησιμοποιήσαμε και λάβαμε ιδέες ή δεδομένα. Δηλώνουμε επίσης ότι, οποιοδήποτε στοιχείο ή κείμενο το οποίο έχουμε ενσωματώσει στην εργασία μας προερχόμενο από Βιβλία ή άλλες εργασίες ή το διαδίκτυο, γραμμένο ακριβώς ή παραφρασμένο, το έχουμε πλήρως αναγνωρίσει ως πνευματικό έργο άλλου συγγραφέα και έχουμε αναφέρει ανελλιπώς το όνομά του και την πηγή προέλευσης.

Οι Φοιτητές

(Ονοματεπώνυμο) (Ονοματεπώνυμο)

…… ……

(Υπογραφή) (Υπογραφή)

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία μελετά τα υπόγεια νερά στην περιοχή της Ζαχάρως στην Ηλεία με τη χρήση γεωστατικών μεθόδων με σκοπό τη δημιουργία ενός μοντέλου διαχείρησης για τη σωστή λειτουργία των υπόγειω νερων.Στο κύριο μερος της συγκεκριμένης εργασίας ξεκινάμε με την εισαγωγή στην παραγοντική ανάλυση ξεκινόντας απο την ανάλυση της φύση της και προχωρώντας στο ορθογώνιο μοντέλο της παραγοντικής ανάλυσης,στη μέθοδο εκτίμησης παραγόντων που αποτελείται από τη μέθοδο των κυρίων συνιστωσών για την εκτίμηση παραγόντων και τη μέθοδο της μέγιστης πιθανοφάνειας.Έπειτα προχωράμε στα βασικά βήματα της παραγοντικής ανάλυσης τα οποία είναι : η ύπαρξη συσχετίσεων,η περιστροφή και ο υπολογισμός των σκορ των παραγόντων.Τέλος στον τομέα της παραγοντικής ανάλυσης έχουμε την ερμηνεία αποτελεσμάτων και την επαληθευτική παραγοντική ανάλυση.

Στη συνέχεια προχωράμε με την εισαγωγή μας στην ανάλυση κατά συστάδες έχοντας τις εξής παραμέτρους : απόσταση και ομοιότητα,ιεραρχική ομαδοποίηση,τη μέθοδο k-means,τη διακριτή ανάλυση η οποία αποτελείται απο τις ακόλουθες διαδικασίες: αντικείμενο της διακριτής ανάλυσης,ερευνητικός σχεδιασμός για τη διακριτή ανάλυση,προϋποθέσεις για τη διαχωριστική ανάλυση και διορθωτικές ενέργειες, εκτίμηση των διαχωριστικών συναρτήσεων και αποτίμηση της καταλληλότητας του μοντέλου,την ερμηνεία των αποτελεσμάτων της διαχωριστικής ανάλυσης και τέλος την επικύρωση των αποτελεσμάτων της διαχωριστικής ανάλυσης.

Το τρίτο σκέλος είναι η εισαγωγή στα δέντρα ταξινόμησης και παλινδρόμησης που αποτελείται από τους κανόνες διαχωρισμού και την περικοπή του δέντρου.Εδώ τελειώνει το θεωρητικό μέρος της παρούσας διπλωματικής εργασίας.

Τώρα στο πρακτικό μερος ξεκινάμε με την δημιουργια ενός πίνακα τιμών με όλα τα δείγματα υπόγειων νερών από την περιοχή της Ζαχάρως.Στη συνέχεια έγινε επιλογή των παραγόντων με βάση υποκειμενικά κριτήρια και έχοντας γνώση των γεωλογικών συνθηκών που επικρατούν στην περιοχή μελέτης.
Με βάση τα παραπάνω και με τη χρήση του λογισμικού SPSS 23.0, εφαρμόστηκε η παραγοντική ανάλυση R
τύπου σε δείγματα υπόγειου νερού του ελεύθερου υδροφόρου ορίζοντα σε 19 μεταβ λητές συνολικά.

Τέλος για την ανάλυση κατά συστάδες χρησιμοποιήθηκε το πρόγραμμα IBM SPSS 23.Εδώ επεξηγούμε την χρήση της ανάλυσης συστάδων χρησιμοποιώντας τα δεδομένα από τις μετρήσεις που έχουε κανει.Πρώτα ομαδοποιούμε τις παρατηρήσεις χρησιμοποιώντας τις μεθόδους απλής σύνδεσης,τη μέθοδο centroid και του ward.

Σχετικά με τα αποτελέσματα στην παραγοντική ανάλυση βλέπουμε οτι έχουμε τέσσερις τελικούς παράγοντες που ειναι οι εξής: «**Παράγοντας υφαλμύρινσης**», «**Παράγοντας απολιθωμένων νερών**», «**Παράγοντας οργανικής ύλης**» και «**Παράγοντας νιτρορύπανσης**».Οι συγκεκριμένοι παράγοντες μας δείχνουν πως επηρεάζεται η σύσταση των υδάτων στη συγκεκριμένη περιοχή.

Στην ανάλυση κατά συστάδες προκύπτουν τρεις μεγάλες ομάδες, οι οποίες φανερώνουν διακριτές υδροχημικές διαδικασίες.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1

ΧΑΡΑΚΤΗΡΙΣΤΗΚΑ ΔΕΔΟΜΕΝΩΝ ΠΟΙΟΤΗΤΑΣ ΚΑΙ ΠΡΟΕΛΕΥΣΗΣ ΤΟΥ ΝΕΡΟΥ-ΠΡΟΒΛΗΜΑΤΑ ΚΑΙ ΠΡΟΣΕΓΓΙΣΕΙΣ ΣΤΑΤΙΣΤΙΚΗΣ ΑΝΑΛΥΣΗΣ

1.1. ΕΙΣΑΓΩΓΗ

Οι υποθέσεις των κλασσικών στατιστικών παραμετρικών μεθόδων συνήθως δεν ικανοποιούνται από τα δεδομένα ποιότητας νερού των οποίων μερικές φορές μπορεί να εκλείπουν στοιχεία ως ένα ορισμένο βαθμό ή και να είναι περικεκομμένα **(Van Belle και Hughes, 1984)**. Επιπλέον, οι κατανομές των ποιοτικών δεδομένων του επιφανειακού νερού είναι συχνά ασύμμετρες λόγω ακραίων τιμών με αποτέλεσμα να παραβιάζονται συχνά οι υποθέσεις των παραμετρικών στατιστικών ελέγχων. Αυτά τα ζητήματα απαιτούν περιπλοκότερες στατιστικές τεχνικές συμπεριλαμβανομένων της θεώρησης των μη παραμετρικών στατιστικών προσεγγίσεων. Οι μη παραμετρικές μέθοδοι μπορούν να είναι μια καταλληλότερη επιλογή για την ανάλυση των ποιότητας στοιχείων του επιφανειακού νερού. Οι μη παραμετρικές στατιστικές τεχνικές δεν υποθέτουν μια ιδιαίτερη μορφή κατανομής, και μπορούν να χειριστούν τις απομακρυσμένες τιμές και τις περικεκομμένες που είναι συνήθεις στα ποιοτικά δεδομένα του νερού.

1.2. ΑΚΡΑΙΕΣ ΤΙΜΕΣ

Οι ακραίες τιμές είναι τιμές οι οποίες είναι φανερά μεγαλύτερες ή μικρότερες από την πλειονότητα των δεδομένων. Οι ακραίες τιμές μπορεί να οφείλονται **(Helsel και Hirsch, 1992)** σε:

• Λανθασμένη μέτρηση ή λάθος στην καταγραφή της τιμής.

• Μία παρατήρηση από έναν πληθυσμό μη παρόμοιο προς αυτόν της πλειονότητας των δεδομένων.

• Ένα σπάνιο γεγονός από έναν μοναδικό ασύμμετρο πληθυσμό.

Οι ακραίες τιμές οι οποίες είναι αποτέλεσμα κάποιας από τις δυο τελευταίες αιτίες θεωρούνται ως πραγματικές (αληθείς) παρατηρήσεις και δεν πρέπει να εξαιρούνται. Οι λανθασμένες παρατηρήσεις μπορεί να είναι το αποτέλεσμα αρκετών παραγόντων όπως η επιμόλυνση του δείγματος, η ανεπάρκεια του εξοπλισμού του εργαστηρίου, λάθη των αναλυτών, μη προσεκτική καταχώρηση των δεδομένων. Όταν συμπεριλαμβάνονται στην ανάλυση δεδομένων λανθασμένες παρατηρήσεις, οι στατιστικές μέθοδοι μπορεί να δώσουν παραπλανητικές πληροφορίες. Όταν μια ακραία τιμή αποδεικνύεται ότι είναι λανθασμένη παρατήρηση, πρέπει να εξαιρείται από την ανάλυση των δεδομένων. Σε κάθε περίπτωση, ο αριθμός των λανθασμένων παρατηρήσεων μπορεί να μειωθεί σημαντικά όταν ακολουθείται πρόγραμμα διασφάλισης ποιότητας και ελέγχου ποιότητας (Quality Assurance – QA, Quality Control – QC).

1.3. ΕΛΛΕΙΠΟΥΣΕΣ ΤΙΜΕΣ

Οι ελλείπουσες τιμές μπορεί να είναι τυχαίες ή συστηματικές. Οι τυχαίες ελλείπουσες τιμές μπορεί να οφείλονται σε διάφορους παράγοντες όπως αποτυχία του εξοπλισμού, χαμένα δείγματα ή αποτελέσματα της ανάλυσης, έντονα καιρικά φαινόμενα που δεν επιτρέπουν τη διεξαγωγή δειγματοληψίας, έκτακτη έλλειψη προσωπικού κλπ. Οι συστηματικές ελλείπουσες τιμές επιφανειακών δειγμάτων νερού οφείλονται συχνά στο κλίμα μιας περιοχής. Για παράδειγμα, ρέματα και χείμαρροι μπορεί να είναι στεγνά κατά τη διάρκεια του θέρους και απρόσιτα κατά τη χειμερινή περίοδο. Επιπροσθέτως, οι αλλαγές στην συχνότητα δειγματοληψίας επιφέρουν συστηματικές ελλείπουσες τιμές **(Lettenmaier και λοιποί, 1991)**.

Αρκετές τεχνικές στατιστικές αναλύσεις απαιτούν οι δειγματοληψίες να είναι σε ίσα χρονικά διαστήματα με αποτέλεσμα να μην είναι δυνατόν να εφαρμοστούν σε αρχεία δεδομένων τα οποία έχουν ελλείπουσες τιμές. Σε αυτή την περίπτωση, οι μη παραμετρικές μέθοδοι μπορούν συνήθως να διευθετήσουν το πρόβλημα των τυχαίων ελλειπουσών τιμών. Παρόλα αυτά, οι συστηματικές ελλείπουσες τιμές μπορεί να παρουσιάσουν ιδιαίτερα προβλήματα **(Lettenmaier και λοιποί, 1991)**.

Η αντικατάσταση των ελλειπουσών τιμών με αριθμητικές εκτιμήσεις τους δεν συνιστάται λόγω πιθανών σφαλμάτων (μεροληψίας) των στατιστικών αποτελεσμάτων. Η χρησιμοποίηση μη παραμετρικών στατιστικών μεθόδων είναι η καλύτερη εναλλακτική λύση για την πλειονότητα των περιστάσεων. Στην περίπτωση που το ποσοστό των ελλειπουσών τιμών είναι μεγάλο ή οι ελλείπουσες τιμές είναι συστηματικές, η καλύτερη εναλλακτική μπορεί να είναι η μείωση της συχνότητας των δεδομένων πριν την στατιστική ανάλυση.

Ο Harcum (1990) διερεύνησε την επίδραση των ελλειπουσών τιμών στον εντοπισμό τάσεων. Το ερώτημα που έθεσε αφορούσε στο ποσοστό των ελλειπουσών τιμών των δεδομένων που χρειαζόταν ώστε να ήταν αναγκαία η μείωση των δεδομένων από μηνιαία σε τριμηνιαία. Τα συμπεράσματα στα οποία κατέληξε ήταν:

• Κατά την εφαρμογή του ελέγχου Mann-Kendall ή του εποχικού ελέγχου Kendall, απαιτείται η μείωση των δεδομένων από μηνιαία σε τριμηνιαία όταν το ποσοστό των μηνιαίων δεδομένων που λείπουν είναι μεγαλύτερο από 50%.

• Κατά την εφαρμογή του εποχικού ελέγχου Kendall με διόρθωση για την σειριακή συσχέτιση, απαιτείται η μείωση των δεδομένων από μηνιαία σε τριμηνιαία όταν το ποσοστό των μηνιαίων δεδομένων που λείπουν είναι μεγαλύτερο από 40%.

• Στην περίπτωση που τα καταγεγραμμένα δεδομένα είναι για διάστημα μόλις πέντε ετών και επιπλέον το ποσοστό των ελλειπουσών τιμών είναι μεγαλύτερο από 50% δεν υπάρχει καλή εναλλακτική λύση. Οι υποδείξεις του **Harcum** είναι οι πλέον κατάλληλες στην προσπάθεια ανάλυσης ήδη διαθέσιμων δεδομένων. Σε κάθε περίπτωση όμως, ένα ορθά δομημένο και λειτουργικό πρόγραμμα παρακολούθησης οφείλει να προλαμβάνει την εμφάνιση μεγάλου ποσοστού ελλειπουσών τιμών.

1.4. ΠΕΡΙΚΕΚΟΜΜΕΝΑ ΔΕΔΟΜΕΝΑ

Περικοπή ή λογοκρισία τιμών (censoring) είναι η αντικατάσταση αριθμητικών μετρήσεων με ποιοτικές επεξηγήσεις όπως «δεν ανιχνεύθηκε», «μικρότερο από …», «μικρότερο από το όριο ανίχνευσης». Γενικά, στην ανάλυση της ποιότητας του νερού η περικοπή των τιμών εμφανίζεται στα χαμηλά επίπεδα συγκέντρωσης όπου η αξιοπιστία των μετρήσεων τίθεται υπό αμφισβήτηση. Η περικοπή των δεδομένων είναι αποτέλεσμα της έλλειψης εμπιστοσύνης στο αριθμητικό αποτέλεσμα ή και του φόβου ότι τα αβέβαια αριθμητικά αποτελέσματα μπορεί να χρησιμοποιηθούν ή να ερμηνευθούν λανθασμένα **(Ward, 1999)**. Αρκετοί ερευνητές προτείνουν ότι δεν πρέπει να επιτρέπεται η περικοπή των τιμών από τα αναλυτικά εργαστήρια αλλά να παρέχονται οι μετρούμενες τιμές συγκεντρώσεων και τα όρια ανίχνευσης ή τα διαστήματα εμπιστοσύνης. Οι **Porter και Ward (1991)** συγκρίνανε την απώλεια πληροφορίας που αποδίδεται στην περικοπή των δεδομένων με αυτήν που οφείλεται στο θόρυβο της μέτρησης και τη βαθμονόμηση. Συμπέραναν ότι όταν η περικοπή των δεδομένων υπερβαίνει περίπου το 50%, τα μη περικομμένα δεδομένα προσφέρουν περισσότερη πληροφορία για την κεντρική θέση του υπό εξέταση πληθυσμού απ’ ότι τα περικομμένα. Οι **Gilliom και λοιποί (1984)** διερεύνησαν την επίδραση της περικοπής των δεδομένων στην ικανότητα εντοπισμού τάσεων. Κατέληξαν ότι οι τάσεις εντοπίζονται πιο

αποτελεσματικά σε μη περικεκομμένα δεδομένα απ' ότι σε περικεκομμένα ακόμα κι όταν τα δεδομένα που είχαν περικοπεί ήταν πολύ ανακριβή.

Η επίδραση των περικεκομμένων τιμών στην στατιστική ανάλυση δεδομένων ποιότητας του νερού εξαρτάται από το ποσοστό των περικεκομμένων τιμών **(Gilliom και λοιποί, 1984; Porter και λοιποί, 1988; Porter και Ward, 1991; Helsel, 1990)**, τη στατιστική προσέγγιση που επιλέγεται για το χειρισμό τους **(Helsel, 1990)** και την ποιότητα των περικεκομμένων δεδομένων **(Gilliom και λοιποί, 1984; Taylor, 1987)**.

Ο **Helsel (1990)** πραγματοποιεί μια πολύ ενδιαφέρουσα αναφορά σχετική με τις κατάλληλες στατιστικές προσεγγίσεις για τη διαχείριση περικομμένων δεδομένων. Για τους ελέγχους υποθέσεων προτείνει τους μη παραμετρικούς ελέγχους οι οποίοι χρησιμοποιούν τις τάξεις μεγέθους (ranks) των δεδομένων αντί για τα ίδια τα δεδομένα. Όλες οι περικεκομμένες τιμές αντιπροσωπεύονται από τάξεις μεγέθους οι οποίες ισοβαθμούν με τιμές οι οποίες είναι μικρότερες από την χαμηλότερη παρατηρούμενη τιμή πάνω από το όριο αναφοράς. Στην περίπτωση που υπάρχουν περισσότερα του ενός όρια αναφοράς, οι έλεγχοι υποθέσεων όπως οι συγκρίσεις μεταξύ διαφορετικών ομάδων δεδομένων είναι πολύ περίπλοκοι. Ο Helsel προτείνει ότι μια ανθεκτική μέθοδος, που μπορεί πάντα να χρησιμοποιηθεί, είναι να περικοπούν όλα τα δεδομένα στο υψηλότερο όριο αναφοράς, και στη συνέχεια να εφαρμοστεί ο κατάλληλος μη παραμετρικός έλεγχος. Επιπλέον, όταν εμφανίζεται μεγάλο ποσοστό περικεκομμένων τιμών (κοντά στο 50% μεγαλύτερο), ακόμη και οι μη παραμετρικοί έλεγχοι έχουν μικρή ικανότητα να ανιχνεύσουν διαφορές στις κεντρικές τιμές των δεδομένων. Σε αυτήν την περίπτωση, οι πίνακες συνάφειας μπορούν να χρησιμοποιηθούν για τον έλεγχο ύπαρξης διαφορών στα ποσοστά των δεδομένων που υπερβαίνουν το όριο ανίχνευσης σε κάθε ομάδα **(Conover, 1980)**. Ο έλεγχος προσδιορίζει εάν το ποσοστό των δεδομένων που εμπίπτει σε κάθε κατηγορία διαφέρει συναρτήσει των διαφορετικών επεξηγηματικών ομάδων, π.χ. διαφορετικές θέσεις δειγματοληψίας.

1.5. ΕΛΛΕΙΨΗ ΚΑΝΟΝΙΚΟΤΗΤΑΣ

Τα δεδομένα των ποιοτικών χαρακτηριστικών του νερού παρουσιάζουν συχνά θετική ασυμμετρία και συνεπώς παραβιάζουν την υπόθεση της κανονικότητας (Adkins, 1993; Twenter και Loftis, 2003). Οι περισσότερες παραμετρικές στατιστικές τεχνικές οι οποίες χρησιμοποιούνται για την ανάλυση δεδομένων ποιότητας του νερού υποθέτουν ότι τα δεδομένα ακολουθούν κανονική κατανομή. Στην περίπτωση που παραβιάζεται η υπόθεση της κανονικότητας η ισχύς των παραμετρικών μεθόδων μειώνεται δραματικά.

Για την αποφυγή του προβλήματος της μείωσης της ισχύος λόγω της μη κανονικότητας των δεδομένων, προτείνεται η χρήση μη παραμετρικών στατιστικών τεχνικών. Παρόλα αυτά, δεν υπάρχουν διαθέσιμες μη παραμετρικές τεχνικές για όλες τις εφαρμογές. Επιπλέον, σε κάποιες περιπτώσεις υπάρχουν οδηγίες σύμφωνα με τις οποίες απαιτείται η χρήση παραμετρικών μεθόδων. Στην περίπτωση που πρέπει να χρησιμοποιηθούν παραμετρικές τεχνικές, χρειάζεται να διερευνηθεί η κατανομή των δεδομένων.

Γενικά, θεωρείται ότι είναι σύνηθες για τα δεδομένα των ποιοτικών χαρακτηριστικών του νερού να μην κατανέμονται κανονικά. Τα δεδομένα παρουσιάζουν συχνά θετική ασυμμετρία λόγω του κατώτατου ορίου (μηδέν) ή της παρουσίας τιμών κάτω από το όριο ανίχνευσης και της εμφάνισης σπάνιων υψηλών τιμών. Οι ιδιότητες κατανομής των ποιοτικών δεδομένων του νερού έχουν διερευνηθεί από αρκετούς ερευνητές και τα αποτελέσματά τους παρουσιάζονται στη συνέχεια. Οι Montgomery και λοιποί (1987) έλεγξαν 172 αρχεία δεδομένων ποιότητας υπόγειων νερών όσον αφορά στην κανονικότητα των παραμέτρων. Χρησιμοποίησαν ιστογράμματα συχνοτήτων, διαγράμματα κανονικής πιθανότητας (normal probability plots) τον έλεγχο Χ2 καλής προσαρμογής (chi-squared goodness-of-fit test) και τον έλεγχο ασσυμετρίας. Το επίπεδο σημαντικότητας που χρησιμοποίησαν στους δυο ελέγχους υποθέσεων ήταν 5%. Σύμφωνα με τα ευρήματά τους, στο 62% των αρχείων τα δεδομένα δεν ακολουθούσαν την κανονική κατανομή ενώ παρουσίαζαν δεξιά κλίση (θετική ασυμμετρία) με βαθμό κυμαινόμενο σε μεγάλο εύρος.

Οι Gilliom και Helsel (1986) εξέτασαν τα στατιστικά χαρακτηριστικά των συγκεντρώσεων ιχνοστοιχείων σε δείγματα που συλλέχθηκαν σε σταθμούς παρακολούθησης της ποιότητας νερού ποταμών από την γεωλογική επιθεώρηση των ΗΠΑ **(U.S. Geological Survey)**. Σε 482 σετ μη περικομμένων δεδομένων, ο συντελεστής ασυμμετρίας κυμαίνονταν από -0,8 έως 5,2 με διάμεση τιμή 1,8. Το ποσοστό των συντελεστών ασυμμετρίας που είχαν αρνητική τιμή ήταν μόλις 6%. Δεδομένου ότι ο συντελεστής ασυμμετρίας σε μια κανονική κατανομή (η οποία είναι εξ ορισμού συμμετρική) είναι 0, οι Gilliom και Helsel συμπέραναν ότι τα δεδομένα ποιότητας του νερού παρουσιάζουν συχνά θετική ασυμμετρία.

Οι Loftis και λοιποί (1989), εξέτασαν την κανονικότητα των ποιοτικών παραμέτρων νερού λιμνών. Χρησιμοποίησαν αμφίπλευρους ελέγχους υποθέσεων για τους συντελεστές ασυμμετρίας και κύρτωσης σε δυο επίπεδα σημαντικότητας (10% και 2%) σε κάθε έλεγχο. Σύμφωνα με τα ευρήματά τους, στο 20% των αρχείων τα δεδομένα παρέκλιναν σημαντικά από την κανονικότητα. Επανέλαβαν την στατιστική ανάλυση στις δεκαδικά λογαριθμισμένες τιμές των δεδομένων (ο λογαριθμικός μετασχηματισμός μπορεί να φέρει τα δεδομένα πιο κοντά στην κανονικότητα) καθώς επίσης και μετά την απαλοιφή των τριμηνιαίων μέσων (η απαλοιφή των τριμηνιαίων μέσων μπορεί να

μειώσει την επίδραση της εποχικότητας). Καμία από τις δυο αυτές ενέργειες δεν επέφερε σημαντική αλλαγή στο ποσοστό των παραμέτρων οι οποίες δεν ακολουθούσαν κανονική κατανομή.

Η εφαρμογή στατιστικών ελέγχων οι οποίοι υποθέτουν κανονικότητα σε μη κανονικά κατανεμημένα δεδομένα μπορεί να επιφέρει δυο ανεπιθύμητα αποτελέσματα. Το πρώτο είναι η διαστροφή της ονομαστικής τιμής του λάθους τύπου Ι **(Snedecor και Cochran, 1989)**, ενώ το δεύτερο είναι η μείωση της ισχύος **(Conover και Iman, 1976; Helsel, 1987)**. Η μείωση της ισχύος των ελέγχων είναι ενδεχομένως πολύ σοβαρό πρόβλημα γιατί μπορεί να μην αναγνωριστούν σημαντικές αλλαγές ή διαφορές στην ποιότητα του νερού. Γενικά, όταν ένας έλεγχος θεωρείται ότι είναι ανθεκτικός στην έλλειψη κανονικότητας σημαίνει ότι η έλλειψη κανονικότητας δεν προκαλεί μεγάλη διαστροφή στο λάθος τύπου Ι **(Conover και Iman, 1976)**. Για παράδειγμα, ο έλεγχος t συχνά θεωρείται ανθεκτικός στην έλλειψη κανονικότητας γιατί το ονομαστικό επίπεδο σημαντικότητας του διατηρείται για μεγάλο μέγεθος δείγματος. Αυτός όμως ο ισχυρισμός αγνοεί την απώλεια της ισχύος η οποία μπορεί να εμφανιστεί όταν δεν ικανοποιείται η υπόθεση της κανονικότητας **(Helsel, 1987)**.

Η μεταχείριση δεδομένων που ενδεχομένως δεν ακολουθούν κανονική κατανομή μπορεί να γίνει με τους παρακάτω τρόπους:

• Με τη χρησιμοποίηση μη παραμετρικών μεθόδων στατιστικής ανάλυσης. Η αποτελεσματικότητα των μη παραμετρικών ελέγχων είναι ελαφρώς μικρότερη των αντίστοιχων παραμετρικών όταν οι υποκείμενοι πληθυσμοί κατανέμονται κανονικά ενώ είναι πολύ μεγαλύτερη όταν οι υποκείμενοι πληθυσμοί δεν κατανέμονται κανονικά **(Hollander και Wolfe, 1973; Bradley, 1968; Hirsch και λοιποί, 1991)**. Επιπλέον οι μη παραμετρικές μέθοδοι είναι ανθεκτικές στις ακραίες τιμές, μπορούν να χρησιμοποιηθούν για την ανάλυση δεδομένων στα οποία εμπεριέχονται αρκετές περικομμένες τιμές, να χειριστούν τις ελλείπουσες τιμές και να εφαρμοστούν σε διατεταγμένα δεδομένα και ενίοτε σε κατηγορικά δεδομένα **(Daniel, 1991; Gibbons, 1985; Helsel, 1987; Hollander και Wolfe, 1973)**.

• Με τη διεξαγωγή ελέγχων κανονικότητας και, εφόσον τα δεδομένα ακολουθούν κανονική κατανομή, τη χρησιμοποίηση παραμετρικών μεθόδων.

• Με τη διεξαγωγή ελέγχων κανονικότητας και, εφόσον τα δεδομένα δεν ακολουθούν κανονική κατανομή, τη χρησιμοποίηση μετασχηματισμών. Με έλεγχο της κανονικότητας των μετασχηματισμένων δεδομένων και, εφόσον ακολουθούν κανονική κατανομή, τη χρησιμοποίηση παραμετρικών μεθόδων στα μετασχηματισμένα δεδομένα **(Helsel,**

1987). Το πρόβλημα όμως είναι ότι συχνά οι απλοί μετασχηματισμοί (π.χ. λογαρίθμηση, τετραγωνική ρίζα, αντιστροφή) ενώ μπορεί να έχουν φυσική ερμηνεία σε αρκετές περιπτώσεις, δεν οδηγούν απαραίτητα στην κανονικότητα. Από την άλλη, πιο πολύπλοκοι μετασχηματισμοί που προσεγγίζουν την κανονικότητα, παράγουν συχνά νέες μεταβλητές που δεν έχουν καμία φυσική σημασία. Η χρησιμοποίηση μετασχηματισμών είναι γενικά επίπονη και χρονοβόρα διαδικασία στην περίπτωση πολλαπλών σετ δεδομένων **(Hirsch και λοιποί, 1991)**. Κάθε σετ δεδομένων πρέπει να εξεταστεί ξεχωριστά για την επιλογή του καλύτερου δυνατού μετασχηματισμού. Επιπλέον, στην περίπτωση που επιλέγονται αρκετοί διαφορετικοί μετασχηματισμοί, υπάρχει δυσκολία στη σύγκριση των αποτελεσμάτων της στατιστικής ανάλυσης μεταξύ των σετ δεδομένων.

• Με τη διεξαγωγή ανάλυσης ευαισθησίας για την ποσοτικοποίηση των επιδράσεων της έλλειψης κανονικότητας στον παραμετρικό έλεγχο που πρόκειται να χρησιμοποιηθεί **(Hirsch και Slack, 1984)**. Αν οι επιδράσεις είναι αποδεκτές, ο έλεγχος μπορεί να προχωρήσει. Η συγκεκριμένη προσέγγιση θεωρείται δαπανηρή και έχει πρακτική αξία όταν πρόκειται να αναλυθούν πολύ σημαντικά δεδομένα με μεθόδους για τις οποίες δεν υπάρχουν μη παραμετρικές εναλλακτικές μέθοδοι.

1.6. ΕΠΟΧΙΚΟΤΗΤΑ

Η εποχικότητα είναι ένα χαρακτηριστικό των δεδομένων ποιότητας του νερού το οποίο απεικονίζει μια αναγνωρίσιμη κυκλική διάταξη που εμφανίζεται στα δεδομένα **(Ward, 1999)**. Με άλλα λόγια, η εποχικότητα είναι η αλλαγή στις κατανομές των ποιοτικών παραμέτρων του νερού η οποία μπορεί να αποδοθεί στην χρονική περίοδο του έτους.

Η εποχικότητα μεγαλώνει τη διακύμανση των ποιοτικών δεδομένων του νερού με συνέπεια την αύξηση του πλάτους των διαστημάτων εμπιστοσύνης στις διαδικασίες εκτίμησης και τη μείωση της ισχύος στους ελέγχους υποθέσεων.

Η εποχικότητα για μια μεταβλητή μπορεί να ελεγχθεί με κάποιον από τους παρακάτω τρόπους:

• Οπτική επιθεώρηση του διαγράμματος των συγκεντρώσεων έναντι του χρόνου (Montgomery και λοιποί, 1987).

• Εξέταση των ετήσιων κύκλων σε ένα διάγραμμα αυτοσυσχετίσεων (correlogram) (Loftis και λοιποί, 1989).

• Με την κατασκευή θηκογραμμάτων (box plots) για κάθε εποχή (Montgomery και λοιποί, 1987). Στην περίπτωση που τα ορθογώνια πλαίσια δεν επικαλύπτονται, θεωρείται ότι η εποχικότητα πιθανόν να αποτελεί σημαντική πηγή της διακύμανσης των δεδομένων (Ward και Loftis, 1989).

• Με ομαδοποίηση των δεδομένων σύμφωνα με την εποχή και τον υπολογισμό του αριθμητικού μέσου για κάθε εποχή. Στη συνέχεια, με τον υπολογισμό του λόγου του μέγιστου προς τον ελάχιστο αριθμητικό μέσο. Η εποχικότητα αυξάνει ανάλογα με τον λόγο αυτόν. Αντίστοιχη διαδικασία μπορεί να γίνει χρησιμοποιώντας τις τυπικές αποκλίσεις (Loftis και λοιποί, 1989).

• Με ομαδοποίηση των δεδομένων σύμφωνα με την εποχή και διεξαγωγή ανάλυσης διακύμανσης (ANOVA) ή του αντίστοιχου μη παραμετρικού ελέγχου KruskalWallis **(Montgomery και λοιποί, 1987)**.

• Με την εξέταση του διαγράμματος της συνάρτησης αυτοσυσχέτισης **(Reckhow και λοιποί, 1993; Yürekli και Kurunç, 2005)**. Η ύπαρξη εποχικότητας στα δεδομένα διαπιστώνεται με την εμφάνιση στατιστικά σημαντικών αυτοσυσχετίσεων στις εποχικές υστερήσεις.

Στην περίπτωση ύπαρξης εποχικότητας στα δεδομένα, υπάρχουν δυο συνήθεις τρόποι στατιστικής προσέγγισης:

• Η αποεποχικοποίηση, δηλαδή ο μετασχηματισμός των δεδομένων ώστε να εξαλειφθεί ο ποσοτικός εποχικός κύκλος. Αυτό επιτυγχάνεται με την αφαίρεση των εποχικών διαμέσων από όλα τα δεδομένα μέσα σε μια εποχή. Στη συνέχεια ακολουθεί η στατιστική ανάλυση των αποεποχικοποιημένων δεδομένων.

• Η χρησιμοποίηση στατιστικών μεθόδων οι οποίες λαμβάνουν υπόψη τους την εποχικότητα των δεδομένων.

Στην περίπτωση που ο σκοπός της στατιστικής ανάλυσης είναι ο εντοπισμός τάσεων σε δεδομένα τα οποία είναι ύποπτα για την ύπαρξη εποχικότητας, συνιστάται ο εποχικός έλεγχος Kendall (Seasonal Kendall test), ο οποίος μπορεί επιπλέον να χειριστεί ελλείπουσες τιμές, περικεκομμένες τιμές και ισοβαθμούσες τιμές δεδομένων. Ο εποχικός έλεγχος Kendall χρησιμοποιείται σε μεγάλο βαθμό στην ανάλυση δεδομένων ποιότητας του νερού.

1.7. ΣΕΙΡΙΑΚΗ ΣΥΣΧΕΤΙΣΗ

Συνήθως, στα προγράμματα παρακολούθησης της ποιότητας του νερού η δειγματοληψία γίνεται ανά τακτά χρονικά διαστήματα (πχ. ημερησίως, μηνιαίως κλπ). Η σειριακή συσχέτιση ή αυτοσυσχέτιση μπορεί να θεωρηθεί γενικά ως περιττή πληροφορία μεταξύ διαδοχικών παρατηρήσεων μιας χρονολογικής σειράς που προκύπτει από δείγματα τα οποία λαμβάνονται ανά μικρά χρονικά διαστήματα σε σχέση με την ενδιαφέρουσα χρονική περίοδο. Όταν η σειριακή συσχέτιση είναι θετική, παρουσιάζεται η τάση υψηλές τιμές να ακολουθούν υψηλές τιμές και χαμηλές τιμές να ακολουθούν χαμηλές τιμές. Αντίθετα, αρνητική σειριακή συσχέτιση παρουσιάζεται όταν υψηλές τιμές ακολουθούν χαμηλές και χαμηλές ακολουθούν υψηλές. Τα δεδομένα ποιότητας του νερού παρουσιάζουν κυρίως θετική σειριακή συσχέτιση.

Οι Loftis και λοιποί (1991) δηλώνουν ότι το διάστημα εμπιστοσύνης της μέσης τιμής για ένα συγκεκριμένο χρονικό διάστημα (πχ. ένα έτος), όταν εκτιμάται μέσω ενός δεδομένου αριθμού ισαπέχουσων χρονικά παρατηρήσεων που καλύπτουν χρονικά το υπό εξέταση διάστημα, είναι στενώτερο όταν τα δεδομένα είναι σειριακά συσχετισμένα απ΄ ότι όταν είναι ανεξάρτητα. Στην περίπτωση όμως που αποτελεί ενδιαφέρον η εξαγωγή συμπερασμάτων για μια μακρά χρονική περίοδο χρησιμοποιώντας ένα συγκεκριμένο χρονικό διάστημα της περιόδου αυτής, μια σειρά σειριακά συσχετισμένων δεδομένων θα περιέχει μικρότερη ποσότητα πληροφορίας σε σχέση με μια σειρά ανεξάρτητων δεδομένων του ίδιου μεγέθους. Στην περίπτωση αυτή, οι έλεγχοι για εντοπισμό τάσης ή για αλλαγές στη διάρκεια του χρόνου τείνουν να απορρίπτουν την μηδενική υπόθεση, υποδηλώνοντας ότι έχει συμβεί κάποια αλλαγή, συχνότερα απ' ότι θα έπρεπε. Επιπλέον, οι Loftis και λοιποί (1991) επισήμαναν ότι το κατά πόσο το παρατηρούμενο πρότυπο μιας δεδομένης σειράς δεδομένων ισαπέχουσων χρονικά παρατηρήσεων μπορεί να αποδοθεί σε σειριακή συσχέτιση ή τάση εξαρτάται από την χρονική κλίμακα.

ΚΕΦΑΛΑΙΟ 2

ΠΟΛΥΜΕΤΑΒΛΗΤΕΣ ΣΤΑΤΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ

2.1. Εισαγωγή-Στόχοι της πολυμεταβλητής παραγοντικής ανάλυσης

Οι πολυμεταβλητές στατιστικές τεχνικές έχουν πεδίο εφαρμογής την ταυτόχρονη ανάλυση πολυμεταβλητών δεδομένων, τα οποία προκύπτουν όταν αρκετές αλληλοσχετιζόμενες μεταβλητές μετρώνται για κάθε διερευνώμενη μονάδα **(Hair, 1998)**.

Οι στόχοι της Πολυμεταβλητής Ανάλυσης είναι **(Καρλής, 2005- Ντζούφρας, 2001)**:

1. Μελέτη των σχέσεων μεταξύ των μεταβλητών. Εξετάζονται οι σχέσεις μεταξύ των μεταβλητών για την κατανόηση διάφορων φαινόμενων.

2. Μείωση των υπό διερεύνηση δεδομένων/ μεταβλητών. Η πληροφορία που περιέχεται σε πολλές μεταβλητές συμπυκνώνεται σε λίγες περιληπτικές μεταβλητές, με τις οποίες παρέχεται η περαιτέρω δυνατότητα της ερμηνείας αθέατων χαρακτηριστικών της πραγματικότητας.

3. Δημιουργία ομάδων. Με κατάλληλες προσεγγίσεις, τα υπό μελέτη αντικείμενα (παρατηρήσεις ή μεταβλητές) μπορούν να ταξινομηθούν σε ομοειδείς ομάδες ως προς κάποια χαρακτηριστικά, με στόχο την καλύτερη διαχείριση των δεδομένων.

4. Πρόβλεψη. Δημιουργία κατάλληλων μοντέλων που λαμβάνουν υπόψη τους την πολυπλοκότητα της πραγματικότητας και προβλέπουν τη μελλοντική συμπεριφορά ενός συστήματος.

5. Έλεγχος πολυμεταβλητών υποθέσεων και πολύπλοκων επιστημονικών σεναρίων (υποθέσεων). Μέσω σύνθετων μαθηματικών υποδειγμάτων μπορούν να επεκταθούν οι έλεγχοι υποθέσεων σε πολυμεταβλητά δεδομένα με σκοπό τον έλεγχο πιο ρεαλιστικών υποθέσεων.

2.2. Εισαγωγή στην Παραγοντική Ανάλυση

2.2.1 Εισαγωγή

Η παραγοντική ανάλυση (Factor Analysis) ανήκει στην ευρύτερη ομάδα των πολυμεταβλητών στατιστικών μεθόδων. Πρόκειται για μια ευρέως διαδομένη τεχνική με εφαρμογές τόσο στην ψυχομετρία και στις κοινωνικές επιστήμες γενικότερα, όσο στην οικονομία και το μάρκετινγκ ειδικότερα (Steward, 1981). Το 1888 πρώτος ο **Galton** εισήγαγε τον σπόρο της παραγοντικής ανάλυσης υποστηρίζοντας ότι σε ένα σύνολο μεταβλητών υπάρχουν κρυμμένοι παράγοντες που επηρεάζουν την διακύμανση και ερμηνεύουν τη σχέση των αρχικών μεταβλητών. Το 1904 οι **Pearson και Spearman** διατύπωσαν το μοντέλο της παραγοντικής ανάλυσης και έκτοτε αρκετοί ερευνητές συνέβαλαν έτσι ώστε το μοντέλο να πάρει την τελική του μορφή **(Morrison, 1990)**. Αν και οι ρίζες της μεθόδου είναι αρκετά παλιές έγινε ιδιαίτερα δημοφιλής τις τελευταίες δεκαετίες εξαιτίας της γρήγορης τεχνολογικής εξέλιξης των ηλεκτρονικών υπολογιστών. Πραγματικά, η ραγδαία πρόοδος που σημειώθηκε στον τομέα επεξεργασίας της πληροφορίας τις τελευταίες δεκαετίες του 20ου αιώνα προκάλεσε αλυσιδωτές εξελίξεις στην στατιστική επιστήμη. Η εισαγωγή νέων εξειδικευμένων στατιστικών πακέτων λογισμικού βοήθησε στην απλοποίηση, έως ένα βαθμό, των πολυμεταβλητών στατιστικών μεθόδων (Principal Components Method, Regression Analysis, Cluster Analysis κ.τ.λ.), μεταξύ των οποίων συγκαταλέγεται και η παραγοντική ανάλυση. Το γεγονός καθ' αυτό είχε ως αποτέλεσμα να γίνει η μέθοδος πιο προσιτή σε άτομα που δεν έχουν βαθύ υπόβαθρο στη στατιστική και στα μαθηματικά, πράγμα πολύ σημαντικό για την διάδοση της μεθόδου ιδιαίτερα σε στελέχη επιχειρήσεων (Garavaglia, Sarma & Thompson S. 2000). Η παραγοντική ανάλυση αποτελεί στις μέρες μας ένα πολύτιμο εργαλείο του μάρκετινγκ, ιδιαίτερα για την τμηματοποίηση της αγοράς αλλά και τον προσδιορισμό μη μετρήσιμων παραγόντων που επιδρούν στην συμπεριφορά του καταναλωτή και συνεπώς καθορίζουν την εκάστοτε πολιτική, στρατηγική της κάθε επιχείρησης (MacLean & Grey, 1998).

2.2.2 Η φύση της Παραγοντικής Ανάλυσης

Εφαρμόζοντας την παραγοντική ανάλυση οι ερευνητές ανακαλύπτουν ένα πολύτιμο σύμμαχο στο δύσκολο έργο τους. Κύρια λειτουργία της παραγοντικής ανάλυσης είναι μ έσα από ένα σύνολο πολύπλοκα συσχετιζόμενων μεταβλητών, η δημιουργία ή καλύτερα η ανακάλυψη νέων κατά κανόνα ασυσχέτιστων μεταξύ τους μεταβλητών-παραγόντων **(Wells & Shetn, 1971)**. Στόχος αυτής της διαδικασίας είναι η μεγιστοποίηση της κατανόησης των δεδομένων και η ερμηνεία τους κατά τον καλύτερο δυνατό τρόπο **(Joreskog, 2003, SPSS INC., 1999**). Τα παραπάνω αποτελούν την βασική ιδέα για την διερευνητική παραγοντική ανάλυση (Exploratory Factor Analysis).

Πριν το1970 ο όρος παραγοντική ανάλυση δεν συνοδευόταν από τον όρο διερευνητική. Αυτός ο όρος εισήχθη το 1969 για να γίνει η διάκριση από την επαληθευτική παραγοντική ανάλυση **(Confirmatory Factor Analysis, Joreskog, 1969)**. Λεπτομέρειες για την συγκεκριμένη μέθοδο παρατίθενται στην προτελευταία ενότητα του τρέχοντος κεφαλαίου. Κύριο αντικείμενο της παρούσας εργασίας αποτελεί η Δ.Π.Α., στις επόμενες σελίδες γίνεται εκτενής αναφορά για τις λειτουργίες και τον τρόπο εφαρμογής της μεθόδου. Συνοπτικά θα μπορούσαμε να υποστηρίξουμε ότι η μέθοδος της παραγοντικής ανάλυσης (Π.Α) **(Johnson, 1998, σελ.147, SPSS INC., 1999, σελ.317)**:

- ερμηνεύει και εξετάζει τις σχέσεις μεταξύ ενός μεγάλου αριθμού μεταβλητών μέσω της ανάδειξης παραγόντων που επηρεάζουν και καθορίζουν την φύση αλλά και την εξέλιξη των αρχικών μεταβλητών.

- καθορίζει τον αριθμό τον νέων παραγόντων και προσπαθεί να δώσει την ερμηνεία τους.

- μειώνει τον όγκο των δεδομένων χωρίς την απώλεια σημαντικής πληροφορίας, διευκολύνοντας ταυτόχρονα την επεξεργασία τους και συνεπώς την εξόρυξη συμπερασμάτων.

- Αναδεικνύει και μετράει μεταβλητές-παράγοντες που διέπουν την συμπεριφορά των ατόμων άλλα δεν είναι έμμεσα μετρήσιμα μ ε άπλα ερωτήματα σε ερωτηματολόγια.

` Όλα τα παραπάνω διευκολύνουν κατά πολύ το έργο των ερευνητών αφού απλοποιούν τις σχέσεις μεταξύ των μεταβλητών και μειώνουν σημαντικά τις διαστάσεις ενός προβλήματος (Ντζούφρας, 2001). Παρόλα αυτά υπάρχουν επικριτές της μεθόδου που υποστηρίζουν ότι μπορεί να οδηγήσει σε λανθασμένες υποθέσεις προκαλώντας σύγχυση και παραπληροφόρηση (Kline, 1998, σελ.11). Είναι γεγονός ότι η μέθοδος έχει δεχτεί δριμύ κριτική για την ερμηνευτική ικανότητα της και για την υποκειμενικότητα των αποτελεσμάτων. Οι ανησυχίες των επιστημών πηγάζουν κυρίως από το γεγονός ότι η παραγοντική ανάλυση πολλές φορές στηρίζεται στην υποκειμενικότητα του εκάστοτε ερευνητή. Η άποψη αυτή είναι εύλογη καθώς κατά κανόνα ο αριθμός των παραγόντων καθορίζεται αυθαίρετα και χωρίς αντικειμενικά κριτήρια. Ακόμη ενδέχεται να δοθούν διαφορετικές ερμηνείες στους παράγοντες από διαφορετικούς επιστήμονες ανάλογα με τις αντιλήψεις και τα πιστεύω του καθενός από αυτούς (Johnson, 1998, σελ.148). Για παράδειγμα ένας φοιτητής διοίκησης επιχειρήσεων δεν είναι δυνατόν να δώσει πανομοιότυπη ερμηνεία σε ένα πρόβλημα παραγοντικής ανάλυσης μ ε ένα φτασμένο ανώτατο στέλεχος επιχειρήσεως. Αυτό συμβαίνει αφενός γιατί ένας φοιτητής δεν έχει την εργασιακή εμπειρία ενός στελέχους και αφετέρου γιατί ένα στέλεχος ενδεχόμενος να

μην έχει το χρόνο να παρακολουθήσει τις επιστημονικές εξελίξεις. Δεν έχουν άδικο λοιπόν αυτοί που υποστηρίζουν ότι τα αποτελέσματα αυτής της μεθόδου αποτελούν τον καθρέφτη των απόψεων αλλά και των πεποιθήσεων του κάθε ερευνητή **(Kline, 1994)**. Επιπρόσθετα, τροφή για δυσμενή σχόλια δίνει το γεγονός ότι η παραγοντική ανάλυση επιδέχεται πολλές μεθόδους επίλυσης (μέθοδος κυρίων συνιστωσών, ελάχιστων τετράγωνων, μεγίστης πιθανοφάνειας κ.α.). Συνεπώς στα ίδια δεδομένα είναι πιθανόν να δοθούν πολλές ερμηνείες ανάλογα με την μέθοδο που θα ακολουθηθεί. Δεν υπάρχει, δηλαδή, μοναδική λύση για κάθε πρόβλημα (Steward, 1981). Παρόλα τα μειονεκτήματα της μεθόδου η συνεισφορά της είναι σημαντική και μερικές φορές πολύτιμη. Τα αρνητικά στοιχεία που προκύπτουν κατά την εφαρμογή της παραγοντικής ανάλυσης θα μπορούσαν να ελαχιστοποιηθούν ή έστω να περιορισθούν σε κάθε περίπτωση υπό τις εξής προϋποθέσεις **(Steward, 1981)**:

- οι ερευνητές πρέπει να ελέγχουν κατά την εφαρμογή της μεθόδου αν το δείγμα είναι κατάλληλο, η παραγοντική ανάλυση δεν αποτελεί πανάκεια για κάθε πρόβλημα.
- Δεν πρέπει να συγχέεται η μέθοδος μ ε την ανάλυση κατά συστάδες (Cluster analysis) ή άλλες πολυμεταβλητές μεθόδους. Ο ρόλος της Π.Α. δεν είναι απλά να χωρίζει τις μεταβλητές σε ομάδες-συστάδες αλλά να ερμηνεύει μ ε τον καλύτερο δυνατό τρόπο την διακύμανση και τη δομή τους.

Τέλος, συμπερασματικά, η Δ.Π.Α. αν και παρουσιάζει μειονεκτήματα μπορεί να αναδειχτεί ένα πανίσχυρο εργαλείο στα χέρια ενός έμπειρου επιστήμονα **(Sheth & Tigert, 2003)**

2.2.3 Το ορθογώνιο μοντέλο της Παραγοντικής Ανάλυσης

Όπως έχει ήδη προαναφερθεί στόχος της Π.Α. αποτελεί η ανάδειξη παραγόντων και η ταυτόχρονη αναπαράσταση της διακύμανσης των αρχικών μεταβλητών μέσα από αυτούς (Afifi,1990, σελ. 398). Έτσι λοιπόν αν θεωρήσουμε ένα σύνολο μεταβλητών x j από ένα πληθυσμό μ ε μέσο μ και διακύμανση-συνδιακυμανση Σ, σύμφωνα μ ε το γενικό μοντέλο της παραγοντικής ανάλυσης υπάρχουν κρυφοί παράγοντες m οι οποίοι απεικονίζονται ως f1,f2,……., fm, όποτε από τα παραπάνω:

$$X_1 = \mu + \lambda_{11}F_1 + \lambda_{12}F_2 + \ldots\ldots\ldots + \lambda_{1m}F_m + u_1$$

$$X_2 = \mu + \lambda_{21}F_1 + \lambda_{22}F_2 + \ldots\ldots\ldots + \lambda_{2m}F_m + u_2$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$X_j = \mu + \lambda_{j1}F_1 + \lambda_{j2}F_2 + \ldots\ldots\ldots + \lambda_{jm}F_m + u_j$$

για j=1,2,…..,p

Όπου m είναι ο αριθμός των παραγόντων, λjm η επιβάρυνση (loading) του fm παράγοντα στην χj μεταβλητή και τέλος uj ο μοναδικός παράγοντας ή σφάλμα που αντιστοιχεί ένας για κάθε αρχική μεταβλητή. Οι μοναδικοί παράγοντες εκφράζουν το σύνολο της συνδιακύμανσης των αρχικών μεταβλητών που δεν μπορεί να ερμηνευτεί από τους Fm παράγοντες **(Johnson E.D. 1998, σελ:150)**. Η (1.1) εκφράζει τις μεταβλητές που παρατηρούμε (x1….xp) ως γραμμικούς συνδυασμούς των παραγόντων F1….Fm. Για να έχει νόημα η παραγοντική ανάλυση πρέπει m<p, διαφορετικά δεν πετυχαίνουμε περικοπή του όγκου των δεδομένων αλλά απλά ένα μετασχηματισμό τους (Ντζούφρας, 2001). Επίσης μπορούμε να υποθέσουμε ότι όλες οι μεταβλητές έχουν μέσο ίσο με το μηδέν όποτε το μ παραλείπεται από τη σχέση (1.1) και έχουμε: Xj = λj1F1+ λj2F2+……+ λjmFm + uj για j=1,2,…..,p (1.2). Για να λειτουργήσει το μοντέλο θα πρέπει να ισχύουν οι παρακάτω υποθέσεις. Θεωρούμε ότι ισχύει (Johnson & Wicher, 1998, σελ:516): 1. Οι Fk παράγοντες για k=1,2,…,m είναι ανεξάρτητοι μεταξύ τους, με μέσο μ=0 και διακύμανση ίση με τη μονάδα. 2. Οι μοναδικοί παράγοντες uj είναι επίσης ασυσχέτιστοι μεταξύ τους με μέσο μ=0 και διακύμανση ίση με Ψi για i=1,2,….,p. 3. Οι παράγοντες Fk και οι μοναδικοί παράγοντές uj ακολουθούν διαφορετικές κατανομές για όλους τους συνδυασμούς των j και k, δηλαδή δεν σχετίζονται μεταξύ τους σε καμία περίπτωση. 4. Επίσης μόνο για την περίπτωση εκτίμησης μ ε τη μέθοδο της μέγιστης πιθανοφάνειας θα πρέπει να υποθέσουμε ότι οι δεδομένα μας προέρχονται από πολυμεταβλητή κανονική κατανομή.

Η Διανυσματική μορφή της σχέσης (1.2) έχει ως εξής: Το μοντέλο της Π.Α. μπορεί και να γραφτεί και με τη μορφή πινάκων ως έξης (Johnson, 1998, σελ:150):

$$\underset{(\rho\times 1)}{X} = \underset{(\rho\times m)}{\Lambda}\underset{(m\times 1)}{F} + \underset{(\rho\times 1)}{n} \quad (1.3)$$

$\mathbf{X}= [X_1, X_2, \ldots\ldots, X_\rho]$

$\mathbf{F}= [F_1, F_2, \ldots\ldots, F_m]$

$\mathbf{U}= [u_1, u_2, \ldots\ldots, u_\rho]$

$$\mathbf{\Lambda}= \begin{Bmatrix} \lambda_{11} & \lambda_{12} \ldots\ldots \lambda_{1m} \\ \lambda_{21} & \lambda_{22} \ldots\ldots \lambda_{2m} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \lambda_{\rho 1} & \lambda_{\rho 2} \ldots\ldots \lambda_{\rho m} \end{Bmatrix}$$

Αντίστοιχα οι παραπάνω υποθέσεις σε μορφή πινάκων είναι η εξής:

1. E(F) = 0 **Cov(F) = E [FF΄] = I**

(m×1) (m×m)

2. E(n) =0, **Cov(n) = E [nn΄] = Ψ, όπου Ψ ένας**

(ρ×1) διαγώνιος πίνακας (ρ×ρ)

3. Cov(n,F) = E [nFT] = 0 (ρ×ρ)

Από τις παραπάνω σχέσεις μπορούμε να συμπεράνουμε ότι: **X =ΛF+u** ή

Σ= Cov(X)= Cov(ΛF+n) =Λ× Cov(F+n)×Λ^T+Ψ = ΛIΛ^T+Ψ= ΛΛ^T+Ψ άρα Σ= ΛΛ^T+Ψ (1.3)

ή **Var (χj) = λj12 +……. λjm2 + ψj**

Communality specific variance

Παρατηρούμε ότι η συνολική διακύμανση συντίθεται από δυο μέρη, από την εταιρικότητα (communality) το ποσοστό δηλαδή της διακύμανσης που ερμηνεύεται από τους παράγοντες και από την επονομαζόμενη ειδικότητα (specificity), το ποσοστό που

δεν μπορεί να ερμηνευτεί από το μοντέλο και εκφράζεται από τον μοναδικό παράγοντα u **(Afifi, 1990, σελ.399)**. Αντικειμενικός σκοπός της παραγοντικής ανάλυσης είναι ο υπολογισμός των επιβαρύνσεων καθώς και της εταιρικότητας μέσα από την σχέση (1.4) (Johnson, 1998, σελ.154). Τα διάφορα στατιστικά πακέτα, όπως για παράδειγμα το SPSS (πάνω στο οποίο βασίστηκε και η ανάλυση περιπτώσεων στα κεφαλαία 3-4) προσφέρουν λύσεις στα παραπάνω προβλήματα. Σ' αυτό το σημείο θα πρέπει να τονίσουμε ότι τα αριθμητικά αυτά στοιχεία παίζουν καθοριστικό ρόλο τόσο στον αριθμό των παραγόντων που θα προκύψουν όσο και στην επιλογή της μεθόδου εκτίμησης παραγόντων **(Tabachnick & Fidel 1989, σελ.647)**. Το κομμάτι της Π.Α αναφορικά μ ε την εκτίμηση των παραγόντων, για το οποίο γίνεται εκτενής αναφορά στην επόμενη ενότητα του κεφαλαίου, είναι ίσως το πιο κρίσιμο στην όλη διαδικασία.

2.2.4 Μέθοδοι εκτίμησης παραγόντων

Όπως ήδη έχει προαναφερθεί, η ερμηνευτική ικανότητα καθώς και η λειτουργικότητα ενός παραγοντικού μοντέλου εξαρτάται απόλυτα από το πλήθος των παραγόντων. Επομένως, μείζονος σημασίας ρόλο παίζει η επιλογή της κατάλληλης μεθόδου εκτίμησης παραγόντων **(Steward, 1981)**. Οι δύο μέθοδοι που χρησιμοποιούνται κατά κόρον για την εκτίμηση των παραγόντων είναι η μέθοδος των κυρίων συνιστωσών (Principal Factor Method) και η μέθοδος της μέγιστης πιθανοφάνειας **(Maximum Likelihood, Johnson, 1998, σελ.156)**. Βέβαια κατά καιρούς και σε περιπτώσεις που η εφαρμογή των δυο αυτών μεθόδων δεν ενδείκνυται, προτιμούνται και άλλες μέθοδοι όπως αυτή των ελαχίστων τετραγώνων (Least-Squares), των κύριων αξόνων (principal Axes), της γενικευμένης μεθόδου των ελαχίστων τετράγωνων (Tabachnick & Fidel, 1989, σελ.623). Αξίζει να αναφέρουμε ότι για να έχουμε καλύτερη εικόνα των δεδομένων μας είναι απαραίτητο να δοκιμάσουμε περισσότερες από μία μεθόδους πριν καταλήξουμε οριστικά **(Johnson & Wicher, 1998, σελ.552)**. Στα πλαίσια αυτής της εργασίας θα ακολουθήσει παρουσίαση των δυο σημαντικότερων μεθόδων που θα βοηθήσει στην καλύτερη κατανόηση τους.

2.2.4.1 Η μέθοδος των κυρίων συνιστωσών για την εκτίμηση παραγόντων

Πρόκειται για μια τροποποίηση της πολυμεταβλητής στατιστικής μεθόδου των κυρίων συνιστωσών έτσι ώστε μέσα από αυτή να εξυπηρετηθούν οι ανάγκες της παραγοντικής ανάλυσης. Με άλλα λόγια η μέθοδος των κυρίων συνιστωσών επιστρατεύεται στην διάθεση της παραγοντικής ανάλυσης. Για αυτό το λόγο δεν θα πρέπει να δημιουργείται σύγχυση στον αναγνώστη όταν αναφέρεται η μέθοδος των κυρίων συνιστωσών στην Π.Α.. Δεν πρόκειται για μια αυτοτελής και ανεξάρτητη μέθοδο άλλα απλά για ένα εργαλείο που τίθεται στην υπηρεσία της παραγοντικής ανάλυσης **(SPSS, σελ: 319)**. Η μέθοδος των κυρίων συνιστωσών βασίζεται στην φασματική ανάλυση του πίνακα διακύμανσης ή συσχέτισης **(Ντζούφρας, 2001)**. Το γεγονός, δηλαδή ότι ένας αρχικός τετραγωνικός πίνακας διακύμανσης δύναται να μετατραπεί σε ένα διαγώνιο πίνακα πολλαπλασιάζοντας τον μ ε ένα τρίτο κατάλληλο πίνακα, ο οποίος περιέχει τις ιδιότιμες του αρχικού πίνακα. Καταλήγουμε δηλαδή σε ένα διάνυσμα μεταβλητών Χ ασυσχέτιστων μεταξύ τους και αυτό είναι εύλογο αφού το διάνυσμα Χ έχει διαγώνιο πίνακα διακύμανσης. Με άλλα λόγια οι επιβαρύνσεις όπως και οι μοναδικοί παράγοντες (σφάλματα) περιέχουν τις ιδιοτιμές του αρχικού πίνακα διακύμανσης **(Johnson, 1998, σελ.159)**. Στη μέθοδο αυτή οι εταιρικότητες υπολογίζονται από το άθροισμα των τετραγώνων των επιβαρύνσεων της κάθε μεταβλητής. Επίσης οι μοναδικοί παράγοντες προκύπτουν αν αφαιρέσουμε από την μονάδα τις εταιρικότητες **(Afifi, 1990, σελ.403)**. Πρόκειται ουσιαστικά για ένα μετασχηματισμό των δεδομένων που διευκολύνει την στατιστική ερευνά και την εξαγωγή συμπερασμάτων. Αρχικός στόχος της μεθόδου αποτελεί ο υπολογισμός του συνόλου της εταιρικότητας και των επιβαρύνσεων αφενός και αφετέρου ο υπολογισμός των μοναδικών παραγόντων ψ1,ψ2,…….ψp. Ανάλογα με την φύση των δεδομένων άλλα και αυτό που διευκολύνει την εργασία μας στην εκάστοτε περίπτωση μας δίνεται η ευχέρεια να προχωρήσουμε σε παραγοντική ανάλυση χρησιμοποιώντας τον πίνακα διακύμανσης είτε τον πίνακα συσχετίσεων. Ο πίνακας διακύμανσης επιλέγεται όταν οι μεταβλητές μας δεν μετριούνται με διαφορετικές μονάδες και το εύρος των τιμών τους δεν έχει μεγάλη έκταση. Στην αντίθετη περίπτωση, όταν δηλαδή παρατηρούμε μεγάλες διαφοροποιήσεις στις τιμές του πίνακα διακύμανσης τότε θα ήταν προτιμότερο να χρησιμοποιήσουμε τον πίνακα συσχετίσεων ή να τυποποιήσουμε τις τιμές μας ώστε να εξουδετερωθεί η επίδραση από την διάφορα στις μονάδες μέτρησης **(Johnson & Wicher, 1998)**. Ένα από τα πλεονεκτήματα της μεθόδου των κυρίων συνιστωσών είναι ότι οι τιμές των επιβαρύνσεων δεν μεταβάλλονται σε κάθε αυξομείωση του αριθμού των παραγόντων. Η σταθερότητα αυτή μας επιτρέπει να προσθέτουμε παράγοντες στο μοντέλο χωρίς να αλλάζει η ερμηνεία των υπόλοιπων παραγόντων. Επιπρόσθετα αυτή η μέθοδος καταλήγει πάντα σε κάποια λύση γιατί ουσιαστικά όπως έχει προαναφερθεί αποτελεί ένα μαθηματικό μετασχηματισμό των αρχικών δεδομένων πράγμα που συνεπάγεται ότι υπάρχουν αυξημένες πιθανότητες σφάλματος στην ερμηνεία του μοντέλου. Ένα από τα μειονεκτήματα αυτής της μεθόδου είναι ότι δεν παρέχει στατιστικούς έλεγχους καλής προσαρμογής του μοντέλου με αποτέλεσμα να

χρησιμοποιούνται υποκειμενικά κριτήρια στον προσδιορισμό το αριθμού των παραγόντων άλλα και στην ερμηνεία του μοντέλου.

2.2.4.2 Η μέθοδος της μεγίστης πιθανοφάνειας

Η μέθοδος της μεγίστης πιθανοφάνειας στηρίζεται σε πιο σύνθετες υπολογιστικές διαδικασίες εκτίμησης των παραμέτρων. Βέβαια, η δυσκολία αυτή, στις μέρες μας παρακάμπτεται αφού αυτή η εργασία διεκπεραιώνεται από Η/Υ **(Kline, 1994, σελ.48)**. Η συγκεκριμένη μέθοδος βασίζεται στην υπόθεση ότι τα δεδομένα μας ακολουθούν την κανονική κατανομή **(Johnson & Wicher, 1998, σελ.530)**. Συνεπώς θα πρέπει επίσης οι κοινοί παράγοντες και οι μοναδικοί παράγοντες να ακολουθούν την πολυμεταβλητή κανονική κατανομή, μόνο σε αυτή την περίπτωση είναι εφικτός ο υπολογισμός των επιβαρύνσεων και των σφαλμάτων (specific variances). Όπως συμβαίνει στη μέθοδο των κυρίων συνιστωσών έτσι και στη μέθοδο της μεγίστης πιθανοφάνειας, για να εφαρμόσουμε την παραγοντική ανάλυση, μπορούμε να χρησιμοποιήσουμε είτε τον πίνακα διακύμανσης S είτε τον πίνακα συσχετίσεων R. Θα πρέπει να τονίσουμε ότι εδώ δεν έχει μεγάλη σημασία ποιος από τους δύο πίνακες θα χρησιμοποιηθεί καθώς η μέθοδος της μεγίστης πιθανοφάνειας δεν είναι ευάλωτη στις διάφορες των μονάδων μέτρησης όμως σχεδόν πάντα επιλέγεται ο πίνακας διακύμανσης (Johnson & Wicher, 1998, σελ.531-32). Η μέθοδος αυτή μας δίνει ελαφρώς διαφοροποιημένα αποτελέσματα από την μέθοδο των κυρίων συνιστωσών. Επίσης πρέπει να τονίσουμε ότι οι επιβαρύνσεις (Factor Loadings) σε αυτή την περίπτωση δεν μένουν αμετάβλητες καθώς αυξομειώνεται ο αριθμός των παραγόντων, πράγμα που ισχύει για τη μέθοδο των κυρίων συνιστωσών (Ντζούφρας, 2001). Σημαντικό πλεονέκτημα της προσέγγισης με τη μέθοδο μεγίστης πιθανοφάνειας αποτελεί το γεγονός ότι είναι δυνατός ο στατιστικός έλεγχος καλής προσαρμογής του μοντέλου μέσω του ελέγχου του λόγου πιθανοφανειών (Likelihood Ratio Test, LRT). Επιπλέον είναι δυνατή η επιλογή του κατάλληλου αριθμού παραγόντων βασισμένη σε στατιστικά και όχι σε υποκειμενικά κριτήρια όπως συμβαίνει μ ε τη μέθοδο των κυρίων συνιστωσών (Johnson, 1998, σελ.157). Τα περισσότερα πακέτα στατιστικού λογισμικού, συνήθως παρουσιάζουν δυο τέτοια πληροφοριακά κριτήρια το Akaike's information criterion (AIC, Akaike, 1974) και το Swarrz's Bayesian criterion (BIC). Έτσι λοιπόν θα πρέπει να επιλέγεται ο αριθμός των παραγόντων που αντιστοιχεί στη μικρότερη τιμή των δυο παραπάνω κριτηρίων. Η επιλογή της μικρότερης τιμής για κάθε κριτήριο στηρίζεται στο γεγονός ότι στην ουσία τα δυο κριτήρια επιβάλουν κάποια ποινή σε κάθε μοντέλο, έτσι είναι λογικό να υιοθετείται στην εκάστοτε περίπτωση το μοντέλο μ ε την μικρότερη ποινή (Ντζούφρας, 2001). Από τα παραπάνω συμπεραίνουμε ότι, η μέθοδος της μεγίστης πιθανοφάνειας είναι κατάλληλη όταν έχουμε την υποψία ότι τα δεδομένα μας προέρχονται από κανονική κατανομή. Το υπέρτατο της πλεονέκτημα είναι ότι παρέχει στατιστικούς έλεγχους για την

επιλογή του μοντέλου. Παρόλα αυτά ο έλεγχος πιθανοφάνειας τείνει να παρουσιάζει σχετικά ασήμαντους παράγοντες για την ανάλυση (Kline, 1994, σελ.50). Για αυτό το λόγο και εναλλακτικά έχουν καθιερωθεί και τα κριτήρια AIC και BIC που αναφέρθηκαν πιο πάνω.

2.2.5 Τα βασικά βήματα της Factor Analysis

Σε αυτό το σημείο και αφού έχουμε παρουσιάσει τον τρόπο λειτουργίας των δύο πιο διαδεδομένων μεθόδων εκτίμησης παραγόντων, είμαστε έτοιμοι να κάνουμε λόγο για την ουσία των πραγμάτων, δηλαδή τον τρόπο με τον οποίο εφαρμόζουμε παραγοντική ανάλυση. Υπάρχουν τέσσερα κομβικά σημεία στα οποία θα πρέπει να σταθούμε έτσι ώστε να πετύχουμε το στόχο μας. Τα τέσσερα αυτά διαδοχικά βήματα τα οποία θα πρέπει να ακολουθήσουμε κατά γράμμα είναι τα εξής (SPSS INC., 1999, σελ.319, Johnson,1998, σελ.147):

- Υπολογισμός του πίνακα συσχετίσεων έτσι ώστε να διαπιστωθεί αν οι μεταβλητές σχετίζονται μεταξύ τους, καθώς μόνο σ' αυτή την περίπτωση είναι δυνατή η εξαγωγή κοινών παραγόντων

- Υπολογισμός των επιβαρύνσεων των παραγόντων στις αρχικές μεταβλητές. Σ' αυτό το σημείο θα πρέπει να επίσης να επιλεγεί και η μέθοδος εκτίμησης των παραγόντων.

- Περιστροφή των επιβαρύνσεων προκειμένου να αυξηθεί η ερμηνευτική ικανότητα τους.

- Υπολογισμός των σκορ των παραγόντων (factor scores), καθώς και αποθήκευση τους για τη δυνατότητα περαιτέρω επεξεργασίας. Πριν γίνει εκτενής αναφορά και ανάλυση σε μερικά από τα σημεία κλειδιά για τα οποία μόλις έγινε λόγος, κρίνεται απαραίτητο να γίνουν μερικές διευκρινήσεις για τον προσδιορισμό του αριθμού των παραγόντων. Αρχικά ίσως να χρειαστεί να μαντέψουμε τον αριθμό των παραγόντων και είναι βέβαιο ότι θα πρέπει να πειραματιστούμε και να κάνουμε αρκετές δοκιμές έως ότου καταλήξουμε στο τελικό μοντέλο (Johnson & Wicher, 1998, σελ.557). Πολλές φορές τα στατιστικά προγράμματα μας υποχρεώνουν να δηλώσουμε από την αρχή των αριθμό των παραγόντων (Tryfos, 1997). Στη συνεχεία θα πρέπει να επαναλάβουμε αρκετές φορές τη διαδικασία φυσικά με

διαφορετικό αριθμό παραγόντων κάθε φορά (Johnson, 1998, σελ.155). Ο αριθμός των παραγόντων είναι αλληλένδετος άλλα και εξαρτώμενος από την μέθοδο εκτίμησης που θα επιλεγεί. Όσον αφορά την μέθοδο των κύριων συνιστωσών δεν υπάρχει περιορισμός στον αριθμό παραγόντων, στην περίπτωση όμως που υπάρχουν πολλοί παράγοντες, ναι μεν απεικονίζεται το σύνολο της διακύμανσης άλλα στην ουσία δεν επιτυγχάνεται ο στόχος της απλοποίησης του προβλήματος μέσα από την μείωση των μεταβλητών (Tabachnick & Fidel 1989, σελ.599). Επιπλέον δεν υπάρχει σταθερό και ταυτόχρονα αντικειμενικό κριτήριο για τον καθορισμό του αριθμού παραγόντων. Συνήθως διαμορφώνεται σύμφωνα με τις αρχικές ενδείξεις που παίρνουμε για το ποσοστό της συνολικής διακύμανσης που εκφράζεται από τους παράγοντες, το ποσοστό της διακύμανσης των αρχικών μεταβλητών που ερμηνεύεται και το scree plot που ίσως είναι και το πιο αξιόπιστο κριτήριο **(Afifi, 1990, σελ.400**). Αντίθετα στην μέθοδο της μεγίστης πιθανοφάνειας περιορίζεται αριθμός των παραγόντων χρησιμοποιώντας είτε τον έλεγχο πιθανοφάνειας είτε τα κριτήρια AIC/BIC. Αντιλαμβανόμαστε λοιπόν ότι η διαδικασία επιλογής αριθμού παραγόντων δεν είναι αυτοτελής και φυσικά δεν μας δίνει μοναδική λύση καθώς κάθε μέθοδος εκτίμησης δίνει διαφορετικά αποτελέσματα. Επίσης χρήσιμη είναι η σύγκριση των αποτελεσμάτων μεταξύ τους έτσι ώστε να καταλήξουμε στο μοντέλο που εκφράζει με τον καλύτερο δυνατόν τρόπο τα αρχικά δεδομένα μας **(Johnson & Wicher 1998, σελ.557)**. Στην περίπτωση που οι λύσεις διαφέρουν μεταξύ τους κατά πολύ, ίσως είναι ένα σημάδι ότι τα δεδομένα μας δεν είναι κατάλληλα για την εφαρμογή της μεθόδου (Ντζούφρας, 2001).

2.2.5.1 ΄Υπαρξη συσχετίσεων

Για να προχωρήσουμε σε παραγοντική ανάλυση θα πρέπει να είμαστε σίγουροι ότι τα δεδομένα μας συσχετίζονται σε βαθμό τέτοιο έτσι ώστε να είναι δυνατή η όρυξη κοινών παραγόντων. Σε αντίθετη περίπτωση δεν είναι καθόλου φρόνιμο να προχωρήσουμε σε παραγοντική ανάλυση καθώς δεν θα είναι εφικτή η εξαγωγή συμπερασματών. Η αναγκαιότητα λοιπόν για τον έλεγχο περιγραφικών δεδομένων είναι ύψιστης σημασίας (Steward, 1981). Η διαδικασία αυτή ξεκινάει με την μελέτη του πίνακα συσχετίσεων, σε μεγάλα δείγματα, ικανοποιητικές θεωρούνται συσχετίσεις μεγαλύτερες του 0,4. Όταν μια μεταβλητή δε σχετίζεται πολύ με τις υπόλοιπες είναι καλό να παραλείπεται από την ανάλυση γιατί στο τέλος προκύπτει σαν ξεχωριστός παράγοντας από μόνης της (Kline, 1994). Δεν έχει νόημα λοιπόν να την συμπεριλάβουμε στην ανάλυση περιπλέκοντας την εργασία μας αφού ήδη λειτουργεί σαν αυτόνομος παράγοντας (Ντζούφρας, 2001). Ο απλός συντελεστής συσχέτισης εκφράζει τη σχέση μεταξύ δύο μεταβλητών. Πολλές φορές όμως αυτή οι πληροφορία από μόνη της υπάρχει περίπτωση να είναι παραπλανητική. Αυτό συμβαίνει διότι δύο μεταβλητές μπορεί να φαίνεται ότι σχετίζονται

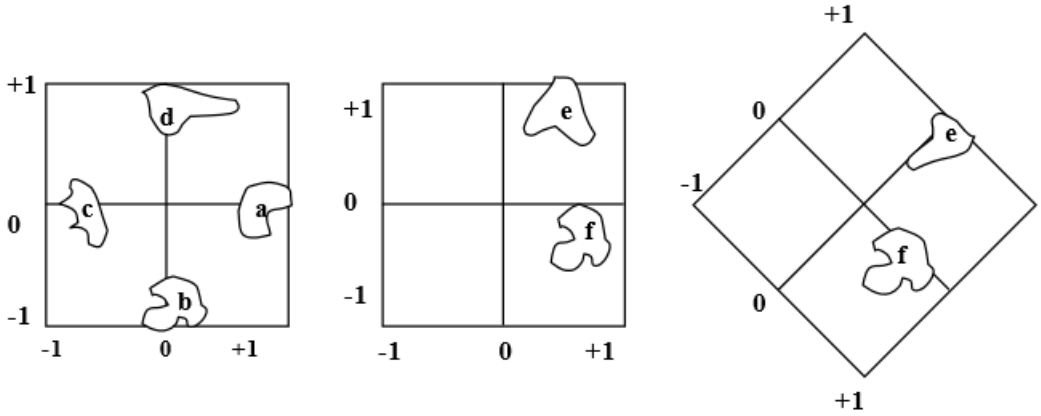
ίσως γιατί μία τρίτη έχει πολύ μεγάλη συσχέτιση με μια από τις δύο αρχικές. Είναι απαραίτητη δηλαδή η ύπαρξη ενός συντελεστή που να απεικονίζει τη σχέση μεταξύ των μεταβλητών χωρίς την επιπρόσθετη επιρροή των υπολειπόμενων μεταβλητών. Αυτή την υπηρεσία διατελεί ο μερικός συντελεστής συσχέτισης (Ντζούφρας, 2001).

Ένα μέτρο που συγκρίνει τους συντελεστές συσχέτισης με τους μερικούς συντελεστές συσχέτισης είναι το Kaiser-Meyer-Olkin (KMO). Μια υψηλή τιμή αυτού του δείκτη υποδηλώνει την καταλληλότητα των δεδομένων μας, ικανοποιητικές τιμές θεωρούνται αυτές που είναι μεγαλύτερες του 0,8. Τέλος ο δείκτης δειγματικής καταλληλότητας (measure of sampling adequacy, MSA) ελέγχει κατά πόσο μια από της αρχικές μεταβλητές είναι κατάλληλη, ιδεατές τιμές θεωρούνται αυτές που κυμαίνονται γύρο στη μονάδα. Σε περίπτωση που κάποια μεταβλητή δεν πληρεί αυτό το κριτήριο θα ήταν καλύτερα να παραβλεφθεί από την μελέτη. Σε τελική ανάλυση θα μπορούσαμε να πούμε ότι αν τα δεδομένα μας δεν καλύπτουν όλες τις παραπάνω προϋπόθεσεις δεν θα ήταν σε καμία περίπτωση καθόλου καλή ιδέα να προχωρήσουμε στην ανάλυση (Steward, 1981).

2.2.5.2 Περιστροφή

Ο κύριος σκοπός της Π.Α. είναι η ανάδειξη μέσα από τα αρχικά δεδομένα ξεκάθαρών και εύκολα ερμηνεύσιμων παραγόντων. Συνήθως όμως αυτή η διαδικασία δεν είναι καθόλου εύκολη καθώς οι παράγοντές που προκύπτουν δεν είναι ευδιάκριτοι μ ε την πρώτη ματιά **(Johnson, 1998, σελ.170)**. Στην περίπτωση όμως που οι μεγάλες επιβαρύνσεις (factor loadings) μπορούσαν να γίνουν μεγαλύτερες και οι μικρές μικρότερες οι παράγοντες θα σχετιζόταν με λιγότερες μεταβλητές και το έργο μας θα ήταν πιο εύκολο. Αυτό ακριβώς πετυχαίνουμε μ ε την περιστροφή των παραγόντων (SPSS INC., 1999, σελ.321). Στην ουσία με την περιστροφή δημιουργούμε καινούργιους παράγοντες που είναι εύκολα ερμηνεύσιμοι. Μ' αυτό τον τρόπο κάθε παράγοντας θα έχει πολύ υψηλές επιβαρύνσεις κοντά στη μονάδα και υπολειπόμενες πολύ μικρές επιβαρύνσεις κοντά στο μηδέν (Afifi, 1990, σελ.409). Αυτή αποτελεί η ιδεατή περίπτωση που οι παράγοντες είναι ευδιάκριτοι και εύκολα αναγνωρίσιμοι. Το σχήμα 1.1 απεικονίζει γραφικά την διαδικασία της περιστροφής, στην πρώτη περίπτωση φαίνεται καθαρά ποιες συστάδες μεταβλητών αποτελούν των πρώτο και δεύτερο παράγοντα. Είναι ευδιάκριτο ότι οι συστάδες a, c διαμορφώνουν τον πρώτο παράγοντα ενώ οι συστάδες b, d το δεύτερο παράγοντα. Αντίθετα στην δεύτερη περίπτωση τίποτα δεν είναι ξεκάθαρο και δεν μπορούμε να εξάγουμε εύκολα συμπέρασμα. Το πρόβλημα μπορεί να λυθεί αν περιστρέψουμε τους άξονες των παραγόντων κατά 45 μοίρες. Τέλος στο τρίτο σχήμα, μετά την περιστροφή, είναι σαφές ότι οι ο πρώτος παράγοντας

συγκροτείται από τη συστάδα των αρχικών μεταβλητών e ενώ ο δεύτερος παράγοντας από τη συστάδα των μεταβλητών f. Αυτή αποτελεί γραφική απεικόνιση της περιστροφής των παραγόντων, αυτό όμως δεν συνεπάγεται ότι οι περιστροφή διεξάγεται μόνο διαγραμματικά. Αυτό που γίνεται στην πράξη είναι ο αριθμητικός υπολογισμός των νέων επιβαρύνσεων με τη βοήθεια υπολογιστών (Johnson, & Wicher, 1998, σελ.543).



Σχήμα 3.2, περιστροφή παραγόντων (πηγή:SPSS INC., 1999)

Θεωρητικά υπάρχουν άπειρες τεχνικές άλλα και συνδυασμοί μ ε τους οποίους μπορεί να πραγματοποιηθεί η περιστροφή των παραγόντων. Η πιο διαδεδομένη και ευρέως γνωστή μέθοδος περιστροφής ή οποία συμπεριλαμβάνεται στην πλειοψηφία των στατιστικών πακέτων ονομάζεται Varimax. Η συγκεκριμένη μέθοδος είναι ορθογώνια δηλαδή η γωνία που σχηματίζουν οι παράγοντες κατά την περιστροφή είναι ορθή **(Afifi, 1990, σελ.419)**. Το γεγονός αυτό εξασφαλίζει την προϋπόθεση για τη μη συσχέτιση των νέων παραγόντων. Η Varimax περιστροφή επιτυγχάνεται μ ε την μεγιστοποίηση του συνόλου της μεταβλητότητας των επιβαρύνσεων ανάμεσα στους παράγοντες **(Afifi, 1990, σελ.411)**. Άλλοι τύποι ορθογώνιων περιστροφών αποτελούν οι Quartimax, Equimax Promax **(Johnson, 1998, σελ.172)**. Στην περίπτωση όμως που δεν επιθυμούμε οι παράγοντες να είναι ασυσχέτιστοι μετ60 αξύ τους οι παραπάνω τεχνικές δεν έχουν καμία χρησιμότητα. Σ΄αυτές τις περιπτώσεις ενδείκνυται η Oblique περιστροφή ή Direct Oblimin περιστροφή όπου οι άξονες των παραγόντων δεν είναι απαραίτητο να σχηματίζουν γωνία 90 μοιρών.

2.2.5.3 Υπολογισμός των Σκορ των Παραγόντων

Μετά την εκτίμηση των παραγόντων, των επιβαρύνσεων και την διενέργεια της τυχόν περιστροφής ακολουθεί ο υπολογισμός των σκορ των παραγόντων **(Afifi, 1990, σελ.415)**. Αυτή η διαδικασία είναι αρκετά δύσκολη καθώς στηρίζεται στην σχέση (1.3) (Johnson, 1998, σελ.180). Η δυσκολία του υπολογισμού έγκειται στο γεγονός ότι οι άγνωστες παρατηρήσεις των F και u ξεπερνούν σε αριθμό τις δεδομένες γνώστες παρατηρήσεις των X **(Afifi, 1990, σελ.550)**. Ουσιαστικά στην προσπάθεια μας να εκτιμήσουμε τα σκορ των παραγόντων παρουσιάζουμε τους παράγοντες ως γραμμικό συνδυασμό των αρχικών μεταβλητών **(Ντζούφρας, 2001)**. Διαφορές στον υπολογισμό των σκορ των παραγόντων εντοπίζονται ανάλογα με το ποια μέθοδο εκτίμησης παραγόντων θα χρησιμοποιηθεί. Όσον αφορά την μέθοδο των κυρίων συνιστωσών τα σκορ είναι δυνατόν να προκύψουν μ ε ακρίβεια. Στη μέθοδο τις μεγίστης πιθανοφάνειας η όλη διαδικασία είναι περισσότερο περίπλοκη και οι εκτιμήσεις γίνονται κατά προσέγγιση αφού ακριβής υπολογισμός των παραγοντικών σκορ είναι αδύνατος (Ντζούφρας, 2001) . Κατόπιν του προσδιορισμού του παραγοντικού μοντέλου και ανεξάρτητα από ποια μέθοδος εκτίμησης έχει χρησιμοποιηθεί υπάρχουν αρκετές τεχνικές για τον προσδιορισμό των σκορ των παραγόντων. Οι πιο γνωστές μέθοδοι που χρησιμοποιούνται από την πλειοψηφία των στατιστικών πακέτων είναι η μέθοδος του Bartlett ή αλλιώς γνωστή ως μέθοδος σταθμισμένων ελαχίστων τετραγώνων (weighted least-squares) και μέθοδός Thomson ή παλινδρόμησης (Regression method) **(Johnson, 1998, σελ.181-82)**. Η σπουδαιότητα της εξαγωγής των σκορ των παραγώντων πηγάζει από την ανάγκη για περαιτέρω στατιστική ανάλυση. Μια από τις λειτουργίες της παραγοντικής ανάλυσης είναι και η προετοιμασία του εδάφους για περαιτέρω στατιστική μελέτη των δεδομένων. Τα σκορ των παραγόντων στην ουσία διαμορφώνουν και περιγράφουν τις καινούργιες μεταβλητές οι οποίες αποτελούν την βάση για περαιτέρω μελέτη **(Tabachnick & Fidel, 1989, σελ.640)**. Τώρα τα αρχικά δεδομένα μας τροποποιημένα είναι έτοιμα να χρησιμοποιηθούν σε περαιτέρω στατιστική ανάλυση και έρευνα. Συνήθως μετά από κάθε παραγοντική ανάλυση ακολουθεί περαιτέρω επεξεργασία των δεδομένων με άλλές πολυμεταβλητές μεθόδους ή έστω γίνεται παρουσίαση περιγραφικών στοιχείων. Φυσικά δεν μπορούσε να ισχύει κάτι διαφορετικό για την παρούσα διπλωματική εργασία, έτσι λοιπόν στα κεφάλαια 3-4 υπάρχει εκτενής μελέτη των δεδομένων μετά την παραγοντική ανάλυση.

2.2.6 Ερμηνεία αποτελεσμάτων

Ίσως πιο δύσκολο και επίπονο κομμάτι τις Π.Α. είναι η ερμηνεία των αποτελεσμάτων. Η διαδικασία αυτή είναι κατά κάποιο τρόπο υποκειμενική αφού έγκειται αποκλειστικά στην κρίση του κάθε επιστήμονα και ουσιαστικά αποτυπώνονται σε αυτή η νοοτροπία και οι απόψεις του ερευνητή αναφορικά με το πρόβλημα. Δεν είναι σπάνιο αλλά ούτε και παράδοξο δύο ή και περισσότερα άτομα να δίνουν διαφορετική ερμηνεία σε ακριβώς ίδια αριθμητικά αποτελέσματα **(Steward, 1981)**. Οι επιβαρύνσεις (Factor Loadings) κατά κανόνα εκφράζουν το βαθμό συσχέτισης των παραγόντων μ ε τις αρχικές μεταβλητές. Οι τιμές των επιβαρύνσεων που ξεπερνούν το 0,30 (ή άλλες φορές του 0,40) θεωρείται ότι αντικατοπτρίζουν ένα ικανοποιητικό βαθμό συσχέτισης των παραγόντων με τις αρχικές μεταβλητές. Σκοπός αυτής της κατηγοριοποίησης των επιβαρύνσεων είναι η ομαδοποίηση των μεταβλητών κατά παράγοντα. Η ομαδοποίηση αυτή μας διευκολύνει να ανακαλύψουμε τις βαθύτερες σχέσεις μεταξύ μεταβλητών και παραγόντων άλλα και ταυτόχρονα να δώσουμε έννοια και οντότητα σε καθένα παράγοντα ξεχωριστά. Στην ουσία βαφτίζουμε τους παράγοντες μ ε το όνομα της καθοριστικής, για την εξέλιξη της διακύμανσης της ομάδας των μεταβλητών, κρυμμένης συνιστώσας. Φυσικά οι ομάδες μεταβλητών που συγκροτούν τους παράγοντες δημιουργούνται με βάση το κριτήριο συσχέτισης 0.30 ή 0,40 όπως έχει ήδη προαναφερθεί **(Tabachnick, & Fidel, 1989, σελ.634)**. Οι μεγάλες επιβαρύνσεις αποτελούν καθοριστικοί δείκτες για την διαμόρφωση των παραγόντων. Επιβαρύνσεις μεγαλύτερες του 0,70 θεωρούνται εξαιρετικές ερμηνεύοντας κατά προσέγγιση το 50% της διακύμανσης της εκάστοτε μεταβλητής. Επιβαρύνσεις μεγαλύτερες του 0,63 θεωρούνται πολύ καλές αφού εξηγούν το 40% της διακύμανσης περίπου, μεγαλύτερες του 0,55 θεωρούνται απλά καλές ερμηνεύοντας το 30% της διακύμανσης, μεγαλύτερες του 0,45 είναι μέτριες εκφράζοντας το 20% της διακύμανσης, και τέλος επιβαρύνσεις μεγαλύτερες του 0,32 φτωχές και ανεπαρκείς αφού ερμηνεύουν, πάντα προσεγγιστικά, μόλις το 10% της διακύμανσης της εκάστοτε μεταβλητής. Η επιλογή του ορίου συμφώνα με το οποίο θεωρείται μια επιβάρυνση αμελητέα για ένα παράγοντα και κατά κανόνα παραλείπεται εξαρτάται και από το ποσοστό της διακύμανσης που ερμηνεύεται. Η τελική απόφαση όμως για την απόρριψη ή όχι της μεταβλητής εναπόκειται στην κρίση του κάθε ερευνητή. Για αυτό το λόγο υπάρχουν περιπτώσεις που επιλέγονται επιβαρύνσεις μ ε χαμηλές τιμές ποτέ όμως μικρότερες από το όριο του 0,30 (Comrey, 1962). Τέλος, ανακεφαλαιώνοντας υπενθυμίσουμε ότι η απόπειρα για ερμηνεία παραγοντικών μοντέλων δεν είναι τίποτα άλλο παρά η προσπάθεια αναγνώρισης της δομής των παραγόντων η και ταυτοποίηση τους μ ε ένα χαρακτηρισμό-ονομασία. Είναι μια δυναμική διαδικασία που θα μπορούσε να θεωρηθεί περισσότερο τέχνη από επιστήμη καθώς ο ανθρωπινός παράγοντας παίζει καθοριστικό ρόλο στη σύνθεση των τελικών αποτελεσμάτων **(Rummel, 1970)**.

2.2.7 Επαληθευτική παραγοντική ανάλυση

Όπως ήδη έχει αναφερθεί στην δεύτερη ενότητα του τρέχοντος κεφαλαίου, από της αρχές της δεκαετίας του 1970 υπήρξε διαχωρισμός της έννοιας παραγοντική ανάλυση σε διερευνητική (exploratory) και επαληθευτική (confirmatory). Αν και αντικείμενο της παρούσας εργασίας αποτελεί Δ.Π.Α. σε αυτό το σημείο θα περιγράψουμε την Ε.Π.Α. έτσι ώστε να είναι πλήρης η αναφορά μας στις τεχνικές Π.Α..

Ουσιαστικά οι δύο αυτές τεχνικές είναι συμπληρωματικές, πολλοί ερευνητές όμως υποστηρίζουν ότι κατά κάποιο τρόπο η Ε.Π.Α. παρουσιάζει κάποια ανωτερότητα συγκριτικά μ ε την Δ.Π.Α. **(Kline, 1994, σελ,81)**. Η άποψη αυτή στηρίζεται στο γεγονός ότι η Ε.Π.Α. ελέγχει την ορθότητα των υποθέσεων της Δ.Π.Α. (Joreskog, 1969). Όταν γίνεται αναφορά στον έλεγχο ή στην επαλήθευση μιας υποθετικής δομής παραγόντων το μυαλό μ ας συνειρμικά θα επρεπε να πηγαίνει κατευθείαν στη μέθοδο της μεγίστης πιθανοφάνειας (Maximum Likelihood Analysis) και αυτό γιατί αυτή είναι το κύριο εργαλείο της Ε.Π.Α.. Και σε αυτή την περίπτωση επίσης για να επιλυθούν οι πολύπλοκες αλγεβρικές σχέσεις απαραίτητη είναι η βοήθεια στατιστικών πακέτων. Τα προγράμματα που χρησιμοποιούνται κατά κόρον για την εφαρμογή της μεθόδου είναι το LISREL και το AMOS **(Ντζούφρας, 2001)**. Για να προχωρήσουμε σε Ε.Π.Α. θα πρέπει να έχουν ολοκληρωθεί όλες οι διεργασίες της Δ.Π.Α., με άλλα λόγια θα πρέπει το μοντέλο της παραγοντικής ανάλυσης να είναι πλήρης διαμορφωμένο **(Macleans & Gray, 1998)**. Στην περίπτωση που κάποιο παραγοντικό μοντέλο απορριφθεί από την μέθοδο σημαίνει ότι δεν απεικονίζει σωστά τα αρχικά δεδομένα. Στην αντίθετη περίπτωση όμως δεν συνεπάγεται, κατά πάσα πιθανότητα, ότι και οι υποθέσεις που έχουν γίνει είναι απαραίτητα ορθές άλλα απλά και μόνο ότι τα δεδομένα μας απεικονίζονται επαρκώς **(Kline, 1994, σελ.98)**. Τέλος, δεν θα πρέπει σε καμία περίπτωση να υποβαθμίζεται η αναγκαιότητα μεγάλου δείγματος όπως επίσης και να παραγκωνίζεται η σπουδαιότητα των διαστημάτων εμπιστοσύνης. Παρόλα αυτά η Ε.Π.Α. μπορεί να αποδειχθεί πολύτιμη σε περιπτώσεις που η επαλήθευση των αρχικών υποθέσεων είναι αναγκαία **(Kline 1994, σελ.99**).

2.3. ΑΝΑΛΥΣΗ ΚΑΤΑ ΣΥΣΤΑΔΕΣ

2.3.1. Εισαγωγή

Η ανάλυση κατά συστάδες (Cluster analysis, CA) είναι μια πολυμεταβλητή μέθοδος για τον εντοπισμό φυσικών ομαδοποιήσεων των δεδομένων. Οι παρατηρήσεις ομαδοποιούνται σε ομάδες – συστάδες κατά τέτοιο τρόπο ώστε κάθε ομάδα να είναι όσο το δυνατόν πιο ομοιογενής (παρόμοια αντικείμενα να ανήκουν στην ίδια ομάδα) και οι ομάδες να διαφέρουν μεταξύ τους όσο το δυνατόν περισσότερο. Η ανάλυση κατά συστάδες δεν έχει προϋποθέσεις για τις κατανομές των μεταβλητών. Στην περίπτωση διαφορετικών μονάδων μέτρησης στις μεταβλητές, συνιστάται η τυποποίηση των δεδομένων πριν την ανάλυση κατά συστάδες **(Rogerson, 2001)**. Η τυποποίηση των δεδομένων εξασφαλίζει ότι κάθε μεταβλητή έχει την ίδια επίδραση στην ανάλυση.

Η ανάλυση κατά συστάδες αποσκοπεί στην κατάταξη υπαρχουσών παρατηρήσεων σε ομάδες χρησιμοποιώντας την πληροφορία που υπάρχει σε κάποιες μεταβλητές. Η μέθοδος, εξετάζοντας πόσο όμοιες είναι κάποιες παρατηρήσεις ως προς κάποιο αριθμό μεταβλητών, τείνει να δημιουργεί ομάδες από παρατηρήσεις που μοιάζουν μεταξύ τους **(Massart και Kaufman, 1983)**. Μια επιτυχημένη ανάλυση καταλήγει σε ομάδες για τις οποίες οι παρατηρήσεις μέσα σε κάθε ομάδα να είναι όσο γίνεται πιο ομοιογενείς αλλά οι παρατηρήσεις διαφορετικών ομάδων διαφέρουν όσο γίνεται περισσότερο. Οι βασικότερες προσεγγίσεις για την ομαδοποίηση των δεδομένων είναι οι ιεραρχικές μέθοδοι, η μέθοδος K-Means και παραλλαγές τους.

Η σημαντική διαφορά της μεθόδου από τη διαχωριστική ανάλυση, η οποία εξετάζεται σε επόμενη ενότητα, είναι πως στη διαχωριστική ανάλυση είναι γνωστή η ομαδοποίηση ως προς κάποιο χαρακτηριστικό των παρατηρήσεων και επιζητείται η δημιουργία συγκεκριμένου κανόνα που θα βοηθήσει στην κατάταξη καινούριων παρατηρήσεων. Καθώς λοιπόν οι δύο μέθοδοι έχουν κάποια κοινά χαρακτηριστικά ως προς τον τρόπο που λειτουργούν, μπορούν να λειτουργήσουν συμπληρωματικά.

2.3.2. Απόσταση και ομοιότητα

Δύο πολύ βασικές έννοιες για την ανάλυση κατά συστάδες είναι οι έννοιες της απόστασης και της ομοιότητας. Οι δύο αυτές έννοιες είναι αντίθετες. Παρατηρήσεις που είναι όμοιες θα έχουν μεγάλη ομοιότητα και μικρή απόσταση. Οι έννοιες αυτές είναι πολύ χρήσιμες καθώς επιτρέπουν τη μέτρηση της ομοιότητας των παρατηρήσεων μεταξύ τους και επομένως την τοποθέτηση τους στην ίδια ομάδα. Η επιλογή της απόστασης έχει να κάνει με τη μέθοδο που θα χρησιμοποιηθεί, τον τύπο των δεδομένων καθώς και τα ίδια τα δεδομένα. Για δεδομένα συνεχή η ευκλείδια απόσταση ή η τετραγωνική ευκλείδια απόσταση είναι οι συχνότερα χρησιμοποιούμενες. Αν μία από τις μεταβλητές έχει όμως τεράστια διακύμανση σε σχέση με τις υπόλοιπες, αυτή θα παίζει σπουδαιότερο ρόλο και άρα θα κατευθύνει και τα αποτελέσματα. Σε αυτή την περίπτωση καλό είναι τα δεδομένα να τυποποιηθούν ώστε να έχουν ίδια μέση τιμή και διακύμανση (άρα και ειδικό βάρος).

2.3.3. Ιεραρχική ομαδοποίηση

Οι πιο διαδεδομένοι ιεραχικοί αλγόριθμοι είναι οι συσσωρευτικοί (agglomerative). Στις συσσωρευτικές ιεραρχικές μεθόδους, αρχικά κάθε παρατήρηση είναι από μόνη της μια ομάδα. Σε κάθε βήμα ενώνονται δύο παρατηρήσεις ή μία παρατήρηση με μία προϋπάρχουσα ομάδα ή δύο προϋπάρχουσες ομάδες, που έχουν την πιο μικρή απόσταση. Η μορφή των αποτελεσμάτων προσδιορίζει τον αριθμό των ομάδων που θα επιλεγούν. Στην πραγματικότητα τα κριτήρια επιλογής του αριθμού των ομάδων είναι πολλά, αλλά μερικές φορές η ερμηνεία που μπορεί να δοθεί είναι ο καλύτερος οδηγός για την επιλογή του αριθμού των ομάδων. Το κρίσιμο σημείο για τον ιεραρχικό αλγόριθμο είναι πως θα υπολογιστεί η απόσταση της ομάδας που κατασκευάστηκε (είτε από συγχώνευση άλλων ομάδων είτε από συγχώνευση παρατηρήσεων).

Υπάρχουν πολλές μέθοδοι, όπως:

• Η μέθοδος του κοντινότερου γείτονα (nearest neighbour or single linkage)

• Η μέθοδος του μακρινότερου γείτονα (furthest neighbour or complete linkage)

• Η μέθοδος του μέσου ανάμεσα στις ομάδες (Average between groups)

• Η μέθοδος του μέσου μέσα στις ομάδες (Average within groups)

• Η μέθοδος του Ward

Από αυτές, η πιο απλή είναι η μέθοδος του κοντινότερου γείτονα η οποία όμως έχει το μειονέκτημα πως δίνει ομάδες με μεγάλες διαφορές ως προς το μέγεθός τους. Η μέθοδος του Ward έχει το πλεονέκτημα ότι δίνει περίπου ισοπληθείς ομάδες και για αυτό χρησιμοποιείται συχνά στην πράξη. Από πειράματα προσομοίωσης έχει βρεθεί ότι η μέθοδοι με την καλύτερη επίδοση είναι η Ward's και η Average Linkage ενώ η μέθοδος του κοντινότερου γείτονα παρουσιάζει τη χειρότερη επίδοση (Καρλής, 2005).

2.3.4. Η μέθοδος k-means

Στην μέθοδο K-Means, ο αριθμός των ομάδων πρέπει να είναι γνωστός εκ των προτέρων. Η μέθοδος δουλεύει επαναληπτικά. Χρησιμοποιεί την έννοια του κέντρου της ομάδας (centroid) και στη συνέχεια κατατάσσει τις παρατηρήσεις ανάλογα με την απόστασή τους από τα κέντρα όλων των ομάδων. Το κέντρο της ομάδας δεν είναι τίποτα άλλο από τη μέση τιμή, ή άλλο μέτρο θέσης, για κάθε μεταβλητή όλων των παρατηρήσεων της ομάδας. Συνήθως χρησιμοποιείται η ευκλείδεια απόσταση. Ο αλγόριθμος K-Means δουλεύει ικανοποιητικά για μεγάλα σετ δεδομένων επειδή σε αυτή την περίπτωση είναι πολύ πιο γρήγορος από την ιεραρχική ομαδοποίηση.

2.4. Διακριτή ανάλυση

2.4.1. Αντικείμενο της διακριτής ανάλυσης

Η διακριτή ή διαχωριστική ανάλυση (Discriminant Analysis, DA) είναι μια πολυμεταβλητή στατιστική μέθοδος, η οποία σχετίζεται με την πολλαπλή παλινδρόμηση, μέσω της οποίας προσδιορίζονται γραμμικές εξισώσεις οι οποίες διαχωρίζουν με τον καλύτερο δυνατό τρόπο τις παρατηρήσεις σε δύο ή περισσότερες φυσικά προκαθορισμένες ομάδες. Η διαχωριστική ανάλυση χρησιμοποιείται για τους παρακάτω ερευνητικούς σκοπούς:

• Τον προσδιορισμό τυχόν στατιστικών σημαντικών διαφορών που υπάρχουν μεταξύ των μέσων επιδόσεων δύο ή περισσοτέρων προκαθορισμένων ομάδων δεδομένων.

• Τον προσδιορισμό των ανεξάρτητων μεταβλητών που ευθύνονται περισσότερο για τις διαφορές των μέσων επιδόσεων δύο ή περισσοτέρων προκαθορισμένων ομάδων δεδομένων.

• Για τον καθορισμό διαδικασιών για την ταξινόμηση αντικειμένων σε ομάδες βάσει των σκορ που επιδεικνύουν σε ένα σύνολο ανεξαρτήτων μεταβλητών.

• Για τον καθορισμό του αριθμού και τη σύνθεση των διαστάσεων της διάκρισης μεταξύ ομάδων, βάσει ενός συνόλου ανεξάρτητων μεταβλητών.

Συνεπώς, η διαχωριστική ανάλυση είναι χρήσιμη τόσο στην κατανόηση των διαφορών μεταξύ ομάδων όσο και στην ορθή ταξινόμηση αντικειμένων σε ομάδες ή κλάσεις.

2.4.2. Ερευνητικός σχεδιασμός για τη διακριτή ανάλυση

Η διακριτή ανάλυση είναι αρκετά ευαίσθητη όσον αφορά στο λόγο του μεγέθους του δείγματος προς τον αριθμό των ανεξάρτητων μεταβλητών. Τα αποτελέσματα τείνουν να είναι μη σταθερά όταν το μέγεθος του δείγματος μικραίνει σε σχέση με τον αριθμό των ανεξάρτητων μεταβλητών. Το ελάχιστο μέγεθος του δείγματος συνιστάται να είναι τριπλάσιο του αριθμού των ανεξάρτητων μεταβλητών **(Massart και λοιποί, 1988)**. Επιπλέον, το μέγεθος της μικρότερης ομάδας θα πρέπει να υπερβαίνει τον αριθμό των ανεξάρτητων μεταβλητών. Σαν ένας πρακτικός οδηγός, κάθε ομάδα θα πρέπει να έχει τουλάχιστον 20 παρατηρήσεις. Επιπροσθέτως, τα σχετικά μεγέθη των επιμέρους ομάδων δεν θα πρέπει να διαφέρουν κατά πολύ καθώς οι μεγάλες σχετικές διαφορές μπορεί να επηρεάσουν την εκτίμηση των διαχωριστικών συναρτήσεων και την ταξινόμηση των παρατηρήσεων (οι μεγαλύτερες ομάδες παρουσιάζουν δυσανάλογα μεγαλύτερη τυχαία ταξινόμηση των παρατηρήσεων σε αυτές).

2.4.3. Προϋποθέσεις για τη διαχωριστική ανάλυση και διορθωτικές ενέργειες

Η σωστή εφαρμογή της διαχωριστικής ανάλυσης προϋποθέτει οι ανεξάρτητες μεταβλητές να έχουν πολυμεταβλητή κανονική κατανομή και άγνωστες πλην όμως ίσες διακυμάνσεις και συνδιακυμάνσεις σε κάθε ομάδα που προκαθορίζεται από την εξαρτημένη κατηγορική μεταβλητή **(Green, 1978; Harris, 1975; Hair και λοιποί, 1998)**. Η έλλειψη πολυμεταβλητής κανονικότητας μπορεί να προκαλέσει προβλήματα στην εκτίμηση των διαχωριστικών συναρτήσεων. Παρόλα αυτά, η προϋπόθεση της πολυμεταβλητής κανονικότητας δεν είναι κρίσιμα σημαντική. Σύμφωνα με τον **Ainsworth (2004)**, εάν στη μικρότερη ομάδα υπάρχουν τουλάχιστον 20 παρατηρήσεις,

η διαχωριστική ανάλυση είναι ανθεκτική στις παραβιάσεις της πολυμεταβλητής κανονικότητας των δεδομένων ακόμα και στην περίπτωση που οι ομάδες είναι άνισου μεγέθους. Όταν σκοπός της διαχωριστικής ανάλυσης είναι η ταξινόμηση, τότε η ανάλυση επηρεάζεται από παραβιάσεις της υπόθεσης της ομοιογένειας επειδή τα αντικείμενα τείνουν να ταξινομούνται στις ομάδες με τη μεγαλύτερη διασπορά. Αυτό μπορεί να αποτιμηθεί σχεδιάζοντας το διάγραμμα των διαχωριστικών συναρτήσεων για τις δύο πρώτες τουλάχιστον συναρτήσεις και συγκρίνοντας τις ομάδες προκειμένου να φανεί αν έχουν περίπου το ίδιο μέγεθος και διάχυση. Ο έλεγχος της υπόθεσης ομοιογένειας μπορεί επίσης να ελεγχθεί με τον στατιστικό έλεγχο Box's M, ο οποίος ελέγχει την μηδενική υπόθεση ότι οι πίνακες διακύμανσης – συνδιακύμανσης των ομάδων είναι ίσοι. Θα πρέπει να επισημανθεί ότι ο έλεγχος Box's M είναι ευαίσθητος στην έλλειψη κανονικότητας των μεταβλητών και το μεγάλο μέγεθος του δείγματος με αποτέλεσμα να δίνει συχνά στατιστικά σημαντικό αποτέλεσμα. Για αυτόν το λόγο, συνήθως χρησιμοποιείται σαν επίπεδο σημαντικότητας του ελέγχου το α=0,001 αντί του συνηθισμένου α=0,05. Στην περίπτωση παραβίασης της υπόθεσης της ομοιογένειας, συνιστάται ο μετασχηματισμός των δεδομένων, η χρησιμοποίηση ξεχωριστών πινάκων συνδιακύμανσης, η τετραγωνική διαχωριστική ανάλυση (quadratic discriminant analysis – QDA) ή μη παραμετρικές προσεγγίσεις για την ταξινόμηση των παρατηρήσεων **(Ainsworth, 2004; Hair και λοιποί, 1998)**.

Η επίδραση των άνισων πινάκων συνδιακύμανσης μπορεί να ελαχιστοποιηθεί με την αύξηση του μεγέθους του δείγματος καθώς επίσης με τη χρησιμοποίηση ξεχωριστών πινάκων συνδιακύμανσης της κάθε ομάδας κατά το στάδιο της ταξινόμησης (Hair και λοιποί, 1998). Εάν η ταξινόμηση χρησιμοποιώντας ξεχωριστούς πίνακες συνδιακύμανσης είναι κατά 2% ή περισσότερο ακριβής σε σχέση με την ταξινόμηση που βασίζεται στον κοινό πίνακα συνδιακύμανσης μέσα στις ομάδες, θα πρέπει να αναφέρονται τα αποτελέσματα της ταξινόμησης που βασίζεται στο μοντέλο των ξεχωριστών πινάκων συνδιακύμανσης **(Schwab, 2007)**. Η τετραγωνική διαχωριστική ανάλυση (quadratic discriminant analysis – QDA), κατά την οποία δεν χρησιμοποιείται κοινός πίνακας συνδιακύμανσης, παρέχεται από αρκετά στατιστικά προγράμματα **(Gessner και λοιποί, 1988; Huberty, 1984; Johnson και Wichern, 1998)**.

Οι δυσκολίες λόγω των μη ευνοϊκών στατιστικών χαρακτηριστικών επηρεάζουν μόνο τα όρια και επομένως την ταξινόμηση με τη διαχωριστική ανάλυση και όχι τον προσδιορισμό των διαχωριστικών συναρτήσεων. Επομένως οι διαχωριστικές συναρτήσεις μπορούν να χρησιμοποιηθούν τελείως ικανοποιητικά για την παρουσίαση των δεδομένων **(Massart και λοιποί, 1988)**.

2.4.4. Εκτίμηση των διαχωριστικών συναρτήσεων και αποτίμηση της καταλληλότητας του μοντέλου

Η διαχωριστική ανάλυση εμπεριέχει την εξαγωγή των διαχωριστικών συναρτήσεων οι οποίες είναι γραμμικοί συνδυασμοί δύο ή περισσότερων ανεξάρτητων μεταβλητών που διαχωρίζουν καλύτερα τις προκαθορισμένες ομάδες. Μία διαχωριστική συνάρτηση λαμβάνεται σύμφωνα με τον παρακάτω τύπο:

$$Z_{jk} = \alpha + W_1 X_{1k} + W_2 X_{2k} + \ldots + W_n X_{nk}$$

Όπου, Z_{jk} = η διαχωριστική τυποποιημένη τιμή της διαχωριστικής συνάρτησης j για το αντικείμενο k

α = σταθερά

W_i = ο διαχωριστικός συντελεστής της ανεξάρτητης μεταβλητής i

X_{ik} = η τιμή της ανεξάρτητης μεταβλητής i για το αντικείμενο k

Ο αριθμός των διαχωριστικών συναρτήσεων οι οποίες εκτιμώνται με τη διαχωριστική ανάλυση είναι κατά μία μονάδα μικρότερος από τον αριθμό των ομάδων. Μία διαχωριστική συνάρτηση εκτιμάται χρησιμοποιώντας τις εξής δύο προσεγγίσεις: Την ταυτόχρονη εισαγωγή όλων των μεταβλητών και τη βηματική προσέγγιση (stepwise). Με την προσέγγιση της ταυτόχρονης εκτίμησης, η διαχωριστική συνάρτηση εκτιμάται κατά τέτοιο τρόπο ώστε όλες οι ανεξάρτητες μεταβλητές να θεωρείται ότι δρουν ταυτόχρονα. Συνεπώς ο υπολογισμός της διαχωριστικής συνάρτησης βασίζεται σε ολόκληρο το σετ των ανεξάρτητων μεταβλητών ανεξάρτητα από τη διαχωριστική ισχύ κάθε ανεξάρτητης μεταβλητής (Hair και λοιποί, 1998). Η προσέγγιση αυτή είναι κατάλληλη, όταν ο ερευνητής, για θεωρητικούς λόγους, θέλει να συμπεριλάβει όλες τις ανεξάρτητες μεταβλητές στην ανάλυση και δεν ενδιαφέρεται να δει ενδιάμεσα αποτελέσματα που βασίζονται μόνο στις μεταβλητές με τη μεγαλύτερη διαχωριστική ικανότητα. Εναλλακτικά, κατά τη βηματική προσέγγιση, οι ανεξάρτητες μεταβλητές εισέρχονται στη διαχωριστική συνάρτηση μία κάθε φορά βάσει της διαχωριστικής τους ισχύος. Κατά τη βηματική προσέγγιση, επιλέγεται αρχικά η μία μοναδική καλύτερη διαχωριστική μεταβλητή. Στη συνέχεια, η αρχική αυτή μεταβλητή συνδυάζεται ξεχωριστά με κάθε μία από τις υπόλοιπες ανεξάρτητες μεταβλητές και επιλέγεται από αυτές η μεταβλητή η οποία είναι η πιο ικανή, σε συνδυασμό με την πρώτη μεταβλητή, να βελτιώσει τη διαχωριστική ικανότητα της συνάρτησης. Η τρίτη καθώς και κάθε επόμενη μεταβλητή επιλέγονται με παρόμοιο τρόπο. Καθώς επιπρόσθετες μεταβλητές συμπεριλαμβάνονται, κάποιες από τις μεταβλητές που είχαν επιλεγεί σε προηγούμενο

στάδιο μπορεί να απομακρυνθούν εάν οι πληροφορίες που περιέχουν σχετικά με τις διαφορές μεταξύ των ομάδων είναι διαθέσιμες σε κάποιο συνδυασμό των άλλων μεταβλητών που έχουν συμπεριληφθεί σε επόμενα στάδια. Τελικά, ή όλες οι ανεξάρτητες μεταβλητές θα έχουν εισαχθεί στην συνάρτηση ή οι μεταβλητές που έχουν εξαιρεθεί θα έχουν κριθεί ως μη σημαντικά συμβάλλουσες σε επιπλέον διαχωρισμό ή διάκριση. Η βηματική προσέγγιση είναι χρήσιμη όταν ο ερευνητής θέλει να θεωρήσει έναν σχετικά μεγάλο αριθμό ανεξάρτητων μεταβλητών προς εισαγωγή στην συνάρτηση. Η τεχνική αυτή προσδιορίζει τις μεταβλητές που επιτρέπουν τον καλύτερο διαχωρισμό δύο ή περισσότερων ομάδων **(Dixon και λοιποί, 1981; Dillon και λοιποί, 1984; Hair και λοιποί, 1998; Johnson και Wichern, 1998; Sharma, 1996)**. Σύμφωνα με αυτήν την προσέγγιση, σε κάθε βήμα επιλέγεται η αμέσως επόμενη μεταβλητή που συμβάλλει σε καλύτερο διαχωρισμό έτσι ώστε οι μεταβλητές που δεν είναι χρήσιμες στον διαχωρισμό ανάμεσα στις ομάδες να αποκλείονται και τελικά να προσδιορίζεται ένα μειωμένο σετ μεταβλητών. Το μειωμένο αυτό σετ μεταβλητών είναι τυπικά σχεδόν τόσο καλό ή και μερικές φορές καλύτερο από το συνολικό σετ των μεταβλητών. Παρόλα αυτά αξίζει να σημειωθεί ότι η βηματική προσέγγιση μπορεί να δώσει λιγότερο σταθερά και γενικεύσιμα αποτελέσματα όταν το μέγεθος του δείγματος είναι μικρό σε σχέση με τον αριθμό των ανεξάρτητων μεταβλητών. Συνεπώς, θεωρείται σημαντική η επικύρωση των αποτελεσμάτων.

Τα στατιστικά κριτήρια τα οποία χρησιμοποιούνται για την εισαγωγή ή την απομάκρυνση μεταβλητών είναι τα: Wilks' lambda, unexplained variance, Mahalanobis D2, μικρότερος F λόγος και Rao's V. Σύμφωνα με τη μέθοδο Wilks' lambda, οι μεταβλητές επιλέγονται προς εισαγωγή στη διαχωριστική συνάρτηση βάσει του πόσο πολύ μειώνουν το Wilks' lambda (Petalas και Anagnostopoulos, 2006; Αποστολάκης και λοιποί, 2003; Σιάρδος, 2000). Σε κάθε βήμα εισάγεται η μεταβλητή που ελαχιστοποιεί το συνολικό Wilks' lambda. Με τη μέθοδο της ανεξήγητης διακύμανσης, σε κάθε βήμα εισέρχεται η μεταβλητή που ελαχιστοποιεί το άθροισμα της ανεξήγητης διακύμανσης ανάμεσα στις ομάδες. Το κριτήριο επιλογής μεταβλητών με τη μέθοδο Mahalanobis D2 είναι η μεγιστοποίηση του Mahalanobis D2 ανάμεσα στις ομάδες. Με τη μέθοδο του μικρότερου F λόγου επιλέγονται οι μεταβλητές που μεγιστοποιούν τον F λόγο που υπολογίζεται από την απόσταση Mahalanobis ανάμεσα στις ομάδες. Το Rao's V είναι ένα μέτρο των διαφορών μεταξύ των ομάδων. Συχνά αποκαλείται και Lawley-Hotelling ίχνος. Σύμφωνα με τη μέθοδο αυτή, σε κάθε βήμα, εισέρχεται η μεταβλητή που μεγιστοποιεί την αύξηση του Rao's V. Η περισσότερο συχνά χρησιμοποιούμενη μέθοδος είναι η Wilks' lambda.

Η στατιστική σημαντικότητα της διαχωριστικής ισχύος κάθε διαχωριστικής συνάρτησης αποτιμάται μέσω των τιμών κατάλληλων κριτηρίων, όπως τα Wilks’ lambda, Hottelling’s trace και Pillai’s, με περισσότερο συχνά χρησιμοποιούμενο το Wilks’ lambda το οποίο ισούται με το ποσοστό της συνολικής διακύμανσης στις τυποποιημένες τιμές της διαχωριστικής συνάρτησης που δεν εξηγείται από τις διαφορές ανάμεσα στις ομάδες.

Μικρότερες τιμές του Wilks’ lambda υποδεικνύουν μεγαλύτερη διαχωριστική ικανότητα της συνάρτησης. Ο έλεγχος Χ2 που συνοδεύει το Wilks’ lambda ελέγχει τη μηδενική υπόθεση ότι οι μέσοι της διαχωριστικής συνάρτησης είναι ίσοι ανάμεσα στις ομάδες. Η απόρριψη της μηδενικής υπόθεσης υποδεικνύει ότι η διαχωριστική συνάρτηση συμβάλλει στατιστικά σημαντικά στο διαχωρισμό των ομάδων.

2.4.5. Ερμηνεία των αποτελεσμάτων της διαχωριστικής ανάλυσης

Η ερμηνεία των αποτελεσμάτων της διαχωριστικής ανάλυσης μπορεί να βασιστεί:

α) Στο απόλυτο μέγεθος των τυποποιημένων διαχωριστικών συντελεστών. Οι ανεξάρτητες μεταβλητές με σχετικά μεγαλύτερους συντελεστές κατά απόλυτη τιμή συνεισφέρουν περισσότερο στη διαχωριστική ισχύ της συνάρτησης απ΄ ότι οι μεταβλητές με μικρότερους συντελεστές. Η ερμηνεία των διαχωριστικών συντελεστών είναι παρόμοια με την ερμηνεία των συντελεστών των ανεξάρτητων μεταβλητών στη γραμμική παλινδρόμηση και συνεπώς υποβάλλεται στις ίδιες κριτικές. Ένας μικρός συντελεστής μπορεί να υποδηλώνει ή ότι η σχετική ανεξάρτητη μεταβλητή είναι άσχετη με τον προσδιορισμό μιας σχέσης ή ότι έχει μερικώς εξαιρεθεί από την σχέση λόγω πολυσυγγραμμικότητας.

β) Στις φορτίσεις διάκρισης ή διαχωρισμού (discriminant loadings) οι οποίες ονομάζονται επίσης και συντελεστές δομής (structure coefficients). Οι φορτίσεις είναι μέτρο της γραμμικής συσχέτισης μεταξύ κάθε ανεξάρτητης μεταβλητής και της διαχωριστικής συνάρτησης και μπορούν να ερμηνευθούν όπως οι παραγοντικές φορτίσεις στην αποτίμηση της σχετικής συνεισφοράς κάθε ανεξάρτητης μεταβλητής στη διαχωριστική συνάρτηση.

Παρόλο που τόσο οι φορτίσεις όσο και οι διαχωριστικοί συντελεστές μπορεί να μην είναι σταθεροί, οι φορτίσεις θεωρούνται περισσότερο έγκυρες για την ερμηνεία της διαχωριστικής ισχύος των ανεξάρτητων μεταβλητών λόγω της φύσης τους να συνδέονται με τη συσχέτιση.

2.4.6. Επικύρωση των αποτελεσμάτων της διαχωριστικής ανάλυσης

Η αποτίμηση της ακρίβειας εκτίμησης της ορθής ταξινόμησης, μέσω της διαχωριστικής ανάλυσης, γίνεται βάσει του πίνακα ταξινόμησης. Ένα υψηλό ποσοστό ορθής ταξινόμησης υποδεικνύει καλή ακρίβεια του μοντέλου. Παρόλα αυτά, η χρησιμοποίηση των ίδιων δεδομένων τόσο στην κατασκευή του μοντέλου όσο και στον προσδιορισμό των αποτελεσμάτων της ταξινόμησης έχει ως αποτέλεσμα τις επονομαζόμενες «τιμές σφάλματος επαναληπτικής αντικατάστασης» (reconstitution error rates). Αυτές τείνουν να υποεκτιμούν τις πραγματικές τάξεις σφάλματος του μοντέλου. Περισσότερο ακριβείς εκτιμήσεις μπορούν να ληφθούν με την εκτίμηση της ταξινόμησης δεδομένων τα οποία δεν είχαν χρησιμοποιηθεί στην κατασκευή του μοντέλου. Η μέθοδος αυτή απαιτεί μεγάλο μέγεθος δείγματος και χρησιμοποιεί ένα μέρος των δεδομένων για την κατασκευή του μοντέλου (σετ εκμάθησης) ενώ ένα άλλο μέρος (σετ επικύρωσης) για την αποτίμηση της ικανότητας διαχωρισμού του μοντέλου. Μία άλλη μέθοδος επικύρωσης των αποτελεσμάτων είναι αυτή της διασταυρούμενης επικύρωσης (cross-validation) **(Crask και λοιποί, 1977, Dillon και λοιποί, 1984)**. Σύμφωνα με τη μέθοδο αυτή, το μοντέλο κατασκευάζεται παρακρατώντας ένα από τα δεδομένα και στη συνέχεια εκτιμάται η ομάδα στην οποία ανήκει αυτό. Η ίδια διαδικασία επαναλαμβάνεται για καθένα από τα δεδομένα.

2.5. Δέντρα ταξινόμησης και παλινδρόμησης

2.5.1. Εισαγωγή

Τα δέντρα ταξινόμησης και παλινδρόμησης (Classification and regression trees - CART) χρησιμοποιούνται ως εναλλακτική μη-παραμετρική μέθοδος προς τη διαχωριστική ανάλυση. Ο αλγόριθμος των δέντρων ταξινόμησης και παλινδρόμησης χρησιμοποιείται για τον προσδιορισμό ενός σετ από «εάν – τότε» (if-then) λογικών συνθηκών (διαχωριστικό σημείο - split) οι οποίες επιτρέπουν την ακριβή πρόβλεψη ή ταξινόμηση των περιπτώσεων (**Breiman και λοιποί, 1984**). Σκοπός των δέντρων ταξινόμησης και παλινδρόμησης είναι η πρόβλεψη ή η επεξήγηση της απόκρισης μιας κατηγορικής μεταβλητής (δέντρα ταξινόμησης) ή μιας συνεχούς μεταβλητής (δέντρα παλινδρόμησης) από ένα σετ επεξηγηματικών μεταβλητών με τη χρησιμοποίηση δυαδικών περιοδικά επαναλαμβανόμενων διαχωριστικών κανόνων, οι οποίοι βασίζονται

σε όρια – κατώφλια (thresholds) στις επεξηγηματικές μεταβλητές **(Breiman και λοιποί, 1984; Steinberg και Cola, 1997)**.

Η μέθοδος διάκρισης που χρησιμοποιείται από τα δέντρα ταξινόμησης μπορεί να θεωρηθεί ότι προσεγγίζει καλύτερα την ανθρώπινη λογική καθώς χρησιμοποιείται ξεχωριστά η κάθε μεταβλητή κάθε φορά.

Υπάρχουν δύο υποκείμενες βασικές ιδέες για την μέθοδο CART των δέντρων ταξινόμησης. Η πρώτη αφορά στον επαναλαμβανόμενο διαχωρισμό του διαστήματος των ανεξάρτητων μεταβλητών. Η δεύτερη αφορά στην περικοπή του δέντρου με τη χρησιμοποίηση των δεδομένων επικύρωσης ή ελέγχου.

2.5.2. Κανόνες διαχωρισμού για τα δέντρα ταξινόμησης

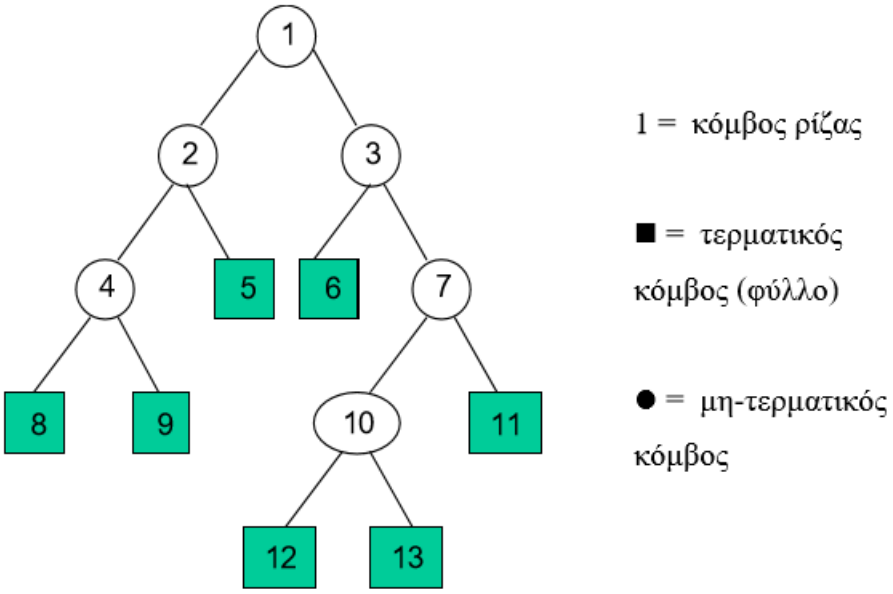
Η πρώτη βασική ιδέα της μεθόδου CART για την ανάπτυξη ενός δέντρου ταξινόμησης πραγματοποιείται μέσω της παρακάτω διαδικασίας. Καθεμία μεταβλητή και όλες οι πιθανές διαχωριστικές τιμές για κάθε μεταβλητή εξετάζονται ώστε να βρεθεί ο καλύτερος διαχωρισμός. Οι πιθανές διαχωριστικές τιμές για μία μεταβλητή είναι απλώς τα ενδιάμεσα σημεία μεταξύ ζευγών διαδοχικών τιμών της μεταβλητής. Αυτά τα διαχωριστικά σημεία ταξινομούνται σύμφωνα με το πόσο πολύ μειώνουν τη μη-αγνότητα (impurity) δηλαδή την ετερογένεια της σύστασης του κόμβου. Η μείωση της μη-αγνότητας ορίζεται ως η μη - αγνότητα του κόμβου πριν τη διάσπαση μείον το ποσόν των μη -αγνοτήτων των δύο κόμβων που προκύπτουν από μια διάσπαση. Το κριτήριο διαχωρισμού (διαχωριστικός κανόνας) θα είναι αυτό που επιφέρει τη μεγαλύτερη μείωση στην μη - αγνότητα ενός κόμβου. Υπάρχουν διάφοροι τρόποι με τους οποίους μπορεί να μετρηθεί η μη-αγνότητα ενός κόμβου. Το δημοφιλέστερο μέτρο μη-αγνότητας είναι ο δείκτης μη - αγνότητας Gini (Gini impurity index**) (Moisen, 2008)**. Εάν οι κλάσεις δηλωθούν ως k με k=1, 2, … , J, όπου J είναι ο συνολικός αριθμός των κλάσεων για την εξαρτημένη μεταβλητή, ο δείκτης μη- αγνότητας Gini για έναν κόμβο προσδιορίζεται από τον τύπο:

$$I(A) = 1 - \sum_{k=1}^{J} p_k^2$$

Όπου, pk είναι το κλάσμα των παρατηρήσεων του κόμβου που ανήκουν στην κλάση k. Στην περίπτωση κατά την οποία όλες οι παρατηρήσεις ανήκουν σε μία μοναδική κλάση

θα είναι I(A) = 0. Η τιμή του I(A) μεγιστοποιείται όταν όλες οι κλάσεις εμφανίζονται στον κόμβο με ίσες αναλογίες. Η μέγιστη τιμή του δείκτη είναι (J-1)/J. Ο δείκτης Gini είναι ένα μέτρο του συνολικού σφάλματος σε κάθε κόμβο (Loh και Shih, 1997). Η μεταβλητή που επιλέγεται θα είναι αυτή που μειώνει περισσότερο το σφάλμα μεταξύ του γονικού κόμβου και του ποσού των σφαλμάτων των δύο νέων κόμβων τέκνων.

Ο λόγος για τον οποίο η μέθοδος καλείται αλγόριθμος δέντρων ταξινόμησης είναι ότι κάθε διάσπαση μπορεί να απεικονιστεί ως διάσπαση ενός κόμβου σε δύο κόμβους διαδόχων (κόμβοι-τέκνα). Η πρώτη διάσπαση παρουσιάζεται ως διακλάδωση του κόμβου - ρίζας ενός δέντρου. Στο **σχήμα 3.1** παρουσιάζεται ένα παράδειγμα δέντρου ταξινόμησης.



Σχήμα 3.1. Παράδειγμα δομής δέντρου ταξινόμησης.

2.5.3 Περικοπή του δέντρου

Η ακρίβεια είναι το σπουδαιότερο χαρακτηριστικό ενός δέντρου ταξινόμησης. Ο καλύτερος τρόπος για να ελεγχθεί η προβλεπτική ακρίβεια ενός δέντρου ταξινόμησης βασίζεται στον προσδιορισμό του ποσοστού λαθεμένης ταξινόμησης των περιπτώσεων ενός ανεξάρτητου σετ δεδομένων ελέγχου, με γνωστές κατανομές κλάσεων, το οποίο διατρέχει το δέντρο ταξινόμησης **(Yohannes και Webb, 1999)**. Τα μεγάλα δέντρα

ταξινόμησης παρουσιάζουν δύο προβλήματα. Πρώτον, παρόλο που είναι υψηλής ακρίβειας, με χαμηλά ή μηδενικά ποσοστά λαθεμένης ταξινόμησης, τα αποτελέσματα που παρουσιάζουν όταν εφαρμόζονται σε νέα δεδομένα είναι ανεπαρκή **(Steinberg και Colla, 1997)**. Δεύτερον, η κατανόηση και επεξήγηση των δέντρων με μεγάλο αριθμό τερματικών κόμβων αποτελεί μια πολύπλοκη διαδικασία. Επομένως, τα μεγάλα δέντρα αναφέρονται ως πολύπλοκα δέντρα. Η πολυπλοκότητα ενός δέντρου μετρείται με τον αριθμό των τερματικών κόμβων.

Η δεύτερη βασική ιδέα της μεθόδου CART, ήταν η χρησιμοποίηση των δεδομένων επικύρωσης για την εκ των υστέρων «περικοπή» του δέντρου το οποίο αναπτύχθηκε από τα δεδομένα εκμάθησης. Η μέθοδος CART χρησιμοποιεί τα δεδομένα επικύρωσης για την εκ των υστέρων «περικοπή» του δέντρου το οποίο έχει υπεραναπτυχθεί σκοπίμως. Η ιδέα πίσω από τη διαδικασία «περικοπής» είναι να αναγνωριστεί ότι ένα πολύ μεγάλο δέντρο είναι πιθανόν να κάνει υπερ-προσαρμογή των δεδομένων εκμάθησης. Για παράδειγμα, οι τελευταίες διασπάσεις μπορεί να οδηγούν σε ορθογώνια (κόμβους) με πολύ λίγα σημεία (ακόμη και μόνο ένα). Τότε, αυτοί οι τελευταίοι διαχωρισμοί είναι πιθανόν να συλλαμβάνουν απλώς τον «θόρυβο» στο σετ εκμάθησης παρά να απεικονίζουν τα πρότυπα που θα απεικονίζονταν σε μελλοντικά δεδομένα όπως είναι τα δεδομένα επικύρωσης.

Η «περικοπή» αναφέρεται στη διαδοχική επιλογή ενός κόμβου απόφασης και τον επαναπροσδιορισμό του σαν έναν κόμβο φύλλου. Με αυτόν τον τρόπο «περικόπτονται» οι κλάδοι που εκτείνονται πέρα από αυτόν τον κόμβο απόφασης (δηλαδή το υπο-δέντρο του) με αποτέλεσμα τη μείωση του μεγέθους του δέντρου. Με τη διαδικασία της «περικοπής» γίνεται ουσιαστικά μια αντιστάθμιση μεταξύ του σφάλματος της λαθεμένης ταξινόμησης των δεδομένων του σετ επικύρωσης έναντι του αριθμού των κόμβων απόφασης στο «περικεκομμένο» δέντρο μέχρι να προσδιοριστεί ένα δέντρο που να συλλαμβάνει τα πρότυπα αλλά όχι τον «θόρυβο» στα δεδομένα εκμάθησης.

Η V-πλή διασταυρούμενη επικύρωση (V-fold cross validation) είναι μία εντατικά υπολογιστική μέθοδος για την επικύρωση μίας διαδικασίας για την κατασκευή μοντέλου, με την οποία αποφεύγεται η απαίτηση για ένα νέο ή ανεξάρτητο σετ δεδομένων επικύρωσης. Σύμφωνα με τη μέθοδο αυτή, το σετ δεδομένων εκμάθησης διαχωρίζεται τυχαία σε V τμήματα ώστε να διασφαλιστεί ότι η κατανομή των αποτελεσμάτων είναι παρόμοια σε καθένα από τα V υπο-τμήματα των δεδομένων. Ένα από αυτά τα υπο-τμήματα των δεδομένων παρακρατείται για να χρησιμοποιηθεί ως ανεξάρτητο σετ δεδομένων ελέγχου, ενώ τα υπόλοιπα V-1 υπο-τμήματα ενώνονται για να χρησιμοποιηθούν ως σετ δεδομένων εκμάθησης κατά τη διαδικασία κατασκευής του μοντέλου. Όλη η διαδικασία κατασκευής του μοντέλου επαναλαμβάνεται V φορές κρατώντας κάθε φορά ένα διαφορετικό υπο-τμήμα ως σετ δεδομένων ελέγχου. Με

αυτόν τον τρόπο, παράγονται V διαφορετικά μοντέλα, καθένα από τα οποία μπορεί να ελεγχθεί έναντι ενός ανεξάρτητου υπο-τμήματος των δεδομένων. Η μέση επίδοση αυτών των V μοντέλων είναι μία εξαιρετική εκτίμηση της επίδοσης του αρχικού μοντέλου (που παράγεται με την χρησιμοποίηση ολόκληρου του σετ των δεδομένων εκμάθησης) σε ένα μελλοντικό ανεξάρτητο σετ δεδομένων, ως μία συνάρτηση του αριθμού των τερματικών κόμβων ή της πολυπλοκότητας. Αυτό επιτρέπει μία εκτίμηση της πολυπλοκότητας του δέντρου, βασισμένη στα δεδομένα, η οποία καταλήγει στην βέλτιστη επίδοση όσον αφορά σε ένα ανεξάρτητο σετ δεδομένων. Με τη χρησιμοποίηση αυτής της μεθόδου, εμφανίζεται ένα ελάχιστο ποσοστό λαθεμένης ταξινόμησης όταν το δέντρο είναι αρκετά σύνθετο ώστε να προσαρμόσει την πληροφορία του σετ δεδομένων εκμάθησης, αλλά όχι τόσο σύνθετο ώστε να προσαρμόσει και το «θόρυβο» που υπάρχει στα δεδομένα.

Προκειμένου να δημιουργηθεί μια ακολουθία από ολοένα και απλούστερα δέντρα, καθένα από τα οποία είναι το καλύτερο μεταξύ όσων έχουν το ίδιο μέγεθος, χρησιμοποιείται το κριτήριο «κόστους πολυπλοκότητας» (cost complexity). Το κριτήριο «κόστους πολυπλοκότητας» εκφράζεται ως Err(T) + α |L(T)|, όπου Err(T) είναι το σφάλμα λαθεμένης ταξινόμησης ενός δέντρου T (βάσει των δεδομένων εκμάθησης), |L(T)| είναι ο αριθμός φύλλων (τερματικών κόμβων) στο δέντρο T και α

είναι το κόστος ποινής ανά κόμβο (ένας αριθμός που θα μεταβάλλεται προς τα πάνω από το μηδέν). Όταν α = 0, δεν υπάρχει καμία ποινή για την ύπαρξη πάρα πολλών κόμβων σε ένα δέντρο και το καλύτερο δέντρο με τη χρησιμοποίηση του κριτηρίου «πολυπλοκότητας κόστους» είναι το πλήρως αναπτυγμένο δέντρο (χωρίς περικοπή). Όταν η τιμή του α αυξάνεται πάρα πολύ, η συνιστώσα κόστους ποινής αντισταθμίζει τη συνιστώσα σφάλματος λαθεμένης ταξινόμησης της συνάρτησης του κριτηρίου «πολυπλοκότητας κόστους» και το καλύτερο δέντρο είναι απλά το δέντρο με τα λιγότερα φύλλα, δηλαδή το δέντρο με απλά έναν κόμβο. Οι **Breiman και λοιποί (1984)**, έδειξαν ότι για κάθε α≥0 υπάρχει ένα μοναδικό δέντρο ελάχιστου μεγέθους το οποίο ελαχιστοποιεί την ποσότητα Err(T) + α |L(T)|. Καθώς η τιμή του α αυξάνεται από το 0 σε κάποια τιμή, αρχικά για κάποιο δέντρο T που διαμορφώνεται με την περικοπή του υποδέντρου σε έναν κόμβο απόφασης, εξισορροπείται ακριβώς το πρόσθετο κόστος του αυξανόμενου σφάλματος λαθεμένης ταξινόμησης (λόγω λιγότερων φύλλων) ενάντι του κόστους ποινής που κερδίζεται από την ύπαρξη λιγότερων φύλλων. Το πλήρες δέντρο «περικόπτεται» στον συγκεκριμένο κόμβο απόφασης με την περικοπή του υποδέντρου του και τον αναπροσδιορισμό αυτού του κόμβου απόφασης ως κόμβο φύλλου. Έστω T το νέο δέντρο που προκύπτει. Η διαδικασία επαναλαμβάνεται στο νέο δέντρο T με την περαιτέρω αύξηση της τιμής του α. Συνεχίζοντας με τον ίδιο τρόπο, δημιουργείται μία ακολουθία δέντρων με διαρκώς μειωμένο αριθμό κόμβων μέχρις του δέντρου που αποτελείται από μόνο έναν κόμβο.

Το κριτήριο «κόστους πολυπλοκότητας» είναι ιδιαίτερα χρήσιμο στην περίπτωση κατά την οποία χρησιμοποιείται η διαδικασία της διασταυρούμενης επικύρωσης **(Moisen, 2008 De'Ath and Fabricius, 2000)**. Με αυτόν τον τρόπο, η V-πλή διασταυρούμενη επικύρωση εφαρμόζεται σε ένα σχετικά μικρότερο σετ δέντρων. Από την ακολουθία των δέντρων φαίνεται φυσικό να επιλεχτεί αυτό που έδωσε το ελάχιστο σφάλμα λαθεμένης ταξινόμησης στο σύνολο των δεδομένων επικύρωσης. Το δέντρο αυτό καλείται δέντρο ελάχιστου σφάλματος. Εναλλακτικά, οι **Breiman και λοιποί (1984)** πρότειναν τον κανόνα του ενός τυπικού σφάλματος (1 – SE Rule) για την επιλογή του βέλτιστου μεγέθους του δέντρου. Σύμφωνα με τον κανόνα αυτόν επιλέγεται το μικρότερο από τα δέντρα των οποίων το ποσοστό λαθεμένης ταξινόμησης απέχει μέχρι ένα τυπικό σφάλμα από το δέντρο ελάχιστου σφάλματος. Με αυτόν τον τρόπο επιλέγονται ακόμα πιο φειδωλά δέντρα (parsimonious) με μικρή «θυσία» όσον αφορά στο σφάλμα ταξινόμησης.

ΚΕΦΑΛΑΙΟ 3

ΥΔΡΟΧΗΜΕΙΑ

3.1 Γενικά

Η χημική σύσταση του νερού μιας περιοχής συνδέεται άμεσα με την ορυκτοχημική σύσταση των πετρωμάτων που κυριαρχούν στην περιοχή και διαμορφώνεται κατά την επαφή του με αυτά, τόσο στην επιφάνεια του εδάφους, όσο και στον υδροφόρο ορίζοντα. Επίσης σημαντικό ρόλο στη διαμόρφωση τηςποιότητας του νερού παίζουν και άλλοι παράγοντες, όπως ο χρόνος παραμονή του νερού στον υδροφόρο, η τροφοδοσία,αλλά και οι ανθρωπογενείς δραστηριότητες. Η υδροχημική έρευνα είναι ένα χρήσιμο εργαλείο όχι μόνο για την ανάδειξη της ποιότητας του νερού, αλλά και των χημικών διεργασιών που την διαμορφώνουν κατά τη διάρκεια της κίνησης του στους υδροφόρους. Συμβάλλει επίσης καθοριστικά στην ορθολογική διαχείριση των υδατικών πόρων.

3.2 Δειγματοληψία υπόγειου νερού

Στα πλαίσια της παρούσας πτυχιακής πραγματοποιήθηκαν δειγματοληψίες και χημικές αναλύσεις υπόγειων νερών από γεωτρήσεις της δυτικής Αχαίας και συγκεκριμένα στην περιοχή της Ζαχάρως.Τα σημειά της υδροληψείας περιλαμβάνονται στον ακόλουθο πίνακα με τις συντεταγμένες τοποθεσίας.

Geology	X	Y
Alluvial	293908	4151298
Alluvial	292126	4147153
Alluvial	293882	4142093
Alluvial	297926	4140779
Alluvial	294395	4140956
Alluvial	294553	4140419
Alluvial	295032	4143602
Alluvial	292236	4153400
Alluvial	290131	4149895
Alluvial	292415	4147915
Alluvial	292487	4145638
Alluvial	295074	4139741

Alluvial	295883	4140444
Alluvial	289365	4151854
Alluvial	292428	4145756
Alluvial	292983	4144339
Alluvial	293694	4143195
Alluvial	294883	4139214
Alluvial	291842	4150409
Alluvial	291360	4148580
Karstic	296903	4152388
Karstic	300778	4149941
Karstic	299672	4147193
Karstic	298216	4145563
Karstic	301309	4154412
Neogene	297057	4141181
Neogene	295266	4148186
Neogene	292993	4147586
Neogene	293286	4145255
Neogene	291402	4152595
Neogene	293915	4143717
Neogene	295177	4145756
Neogene	294193	4152411
Neogene	292260	4153776
Neogene	296769	4141465
Neogene	296086	4142807
Neogene	293553	4146573
Neogene	292890	4148348
Neogene	292904	4154606
Neogene	293434	4145312
Neogene	298005	4141842
Neogene	293519	4144338
Neogene	292155	4150133
Neogene	295055	4141094
Neogene	295082	4149862
Neogene	293635	4153060

Η δειγματοληψία έγινε σύμφωνα με τα πρότυπα της Υπηρεσίας Περιβαλλον τικής Προστασίας των Η.Π.Α. (U.S.E.P.A.,
1976). Για να εξασφαλιστεί ότι τα δείγματα νερού αντιπροσωπεύουν το νερό του υδροφόρου ορίζοντα, τα δείγματα ελήφθησαν μετά από άντληση τουλάχιστον μισής ώρας στις γεωτρήσεις.
Χρησιμοποιήθηκαν 2 φιάλες πολυαιθυλενίου (1000ml και 100ml) για κάθε

δείγμα
νερού, οι οποίες προηγουμένως είχαν πλυθεί με υδροχλωρικό οξύ 5% κ.β. και απιονισμένο νερό, ενώ και πριν τη δειγματοληψία ξεπλένονταν με το νερό της γεώτρησης τουλάχιστον 5 φορές.
Στην φιάλη όγκου 1000ml αποθηκευόταν το νερό, που προοριζόταν για ανάλυση ανιόντων,
ενώ στη φιάλη 100ml αποθηκευόταν διηθημένο και οξυνισμένο νερό για τη ν ανάλυση κατιόντων και ιχνοστοιχείων. Η διήθηση γινόταν επιτόπου με τη χρήση αντλίας κενού και φίλτρων 0,45μm
(Millipore®). Η οξύνιση των δειγμάτων έγινε με προσθήκη 0.5ml υπερκάθαρου νιτρικού οξέος (Merck®), ώστε το pH να ελαττωθεί στο 2 και να αποφευχθούν φαινόμενα καθίζησης κατιόντων.
Κατά τη διάρκεια της δειγματοληψίας μετρήθηκαν επί τόπου οι ασταθείς παράμετροι του υπόγειου νερού. Η θερμοκρασία (T), η ηλεκτρική αγωγιμότητα (E.C.), το pH και το δυναμικό οξειδοαναγωγής (Eh) μετρήθηκα ν με φορητό εξοπλισμό της Consort®, ενώ η αλκαλικότητα και το διοξείδιο του άνθρακα προσδιορίστηκαν με τη μέθοδο της τιτλοδότησης (Hach).
Όλες οι υπόλοιπες χημικές αναλύσεις πραγματοποιήθηκαν αμέσως μετά τη συλλογή στο Εργαστήριο Χημείας του Πανεπιστημίου Πελοποννήσου.

3.3 Πολυμεταβλητή παραγοντική ανάλυση.

Στα πλαίσια της παρούσας έρευνας κρίθηκε χρήσιμη η εφαρμογή της πολυμεταβλητής παραγοντικής ανάλυσης R-
τύπου με στόχο τη μείωση των μεταβλητών και την καλύτερη αξιοποίηση του όγκου των δεδομένων. Πιο συγκεκριμένα, με την εφαρμογή της R-τύπου παραγοντικής ανάλυσης οι μεταβλητές (φυσικοχημικές παράμετροι, κύρια ιόντα και ιχνοστοιχεία) μετατρέπονται σε παράγοντες που ομαδοποιούν τον ίδιο όγκο δεδομένων.
Στον τομέα της υδρογεωλογίας έχει εφαρμοστεί με επιτυχία σε πολλές περιπτώσεις η παραγοντική ανάλυση R-τύπου (Voudouris et al. 1997, Lambrakis et al. 2004, Μανδηλαράς,
2005) αναδεικνύοντας περισσότερο τις υδροχημικές διαδικασίες. Ως προς την επιλογή του αριθμού των παραγόντων, έχουν προταθεί διάφορα κριτήρια (Παπα θεοδώρου,
1996), με τα κυριότερα να είναι τα ακόλουθα, τα οποία και χρησιμοποιήθηκαν στην παρούσα διατριβή: ¾
Επιλογή των παραγόντων με βάση το ποσοστό της ολικής διακύμανσης που εκφράζουν συνολικά. ¾

Επιλογή των παραγόντων, οι οποίοι σχετίζονται με ιδιοτιμές μεγαλύτερες από τη μονάδα. ¾
Επιλογή του αριθμού των παραγόντων, σύμφωνα με το διάγραμμα αλλαγή κλίση ς του Cattel. ¾
Επιλογή των παραγόντων με βάση υποκειμενικά κριτήρια και έχοντας γνώση των γεωλογικών συνθηκών που επικρατούν στην περιοχή μελέτης.
Με βάση τα παραπάνω και με τη χρήση του λογισμικού SPSS 23.0, εφαρμόστηκε η παραγοντική ανάλυση R-τύπου σε δείγματα υπόγειου νερού του ελεύθερου υδροφόρου ορίζοντα σε 19 μεταβλητές συνολικά.

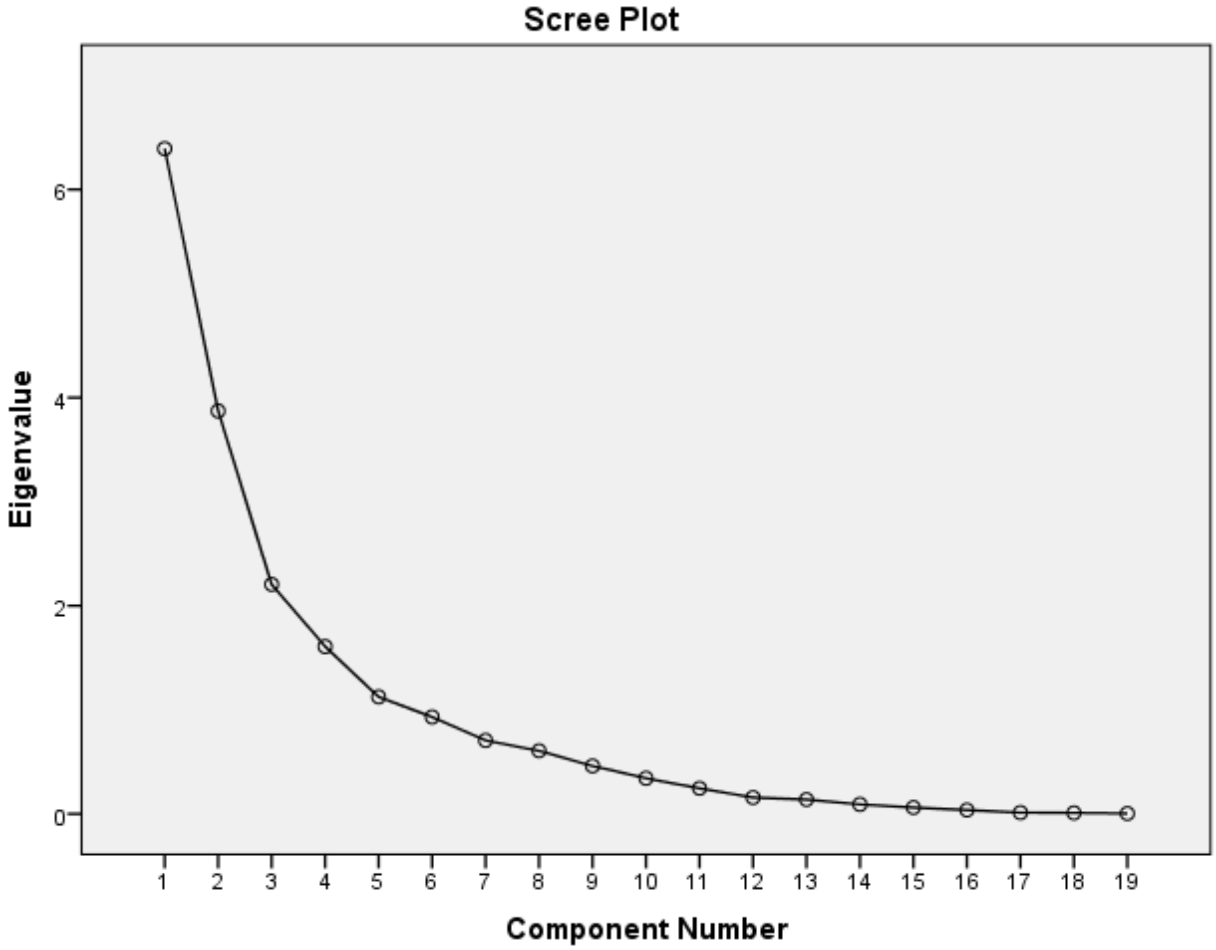
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.392	33.644	33.644	6.392	33.644	33.644	5.307	27.931	27.931

2	3.872	20.378	54.022	3.872	20.378	54.022	3.274	17.234	45.165
3	2.205	11.605	65.627	2.205	11.605	65.627	2.605	13.708	58.873
4	1.608	8.465	74.092	1.608	8.465	74.092	2.141	11.268	70.142
5	1.124	5.917	80.009	1.124	5.917	80.009	1.875	9.867	80.009
6	.931	4.899	84.908						
7	.706	3.715	88.623						
8	.606	3.189	91.812						
9	.460	2.421	94.233						
10	.342	1.802	96.035						
11	.246	1.293	97.329						
12	.157	.826	98.155						
13	.138	.724	98.879						
14	.091	.479	99.359						
15	.060	.318	99.676						
16	.036	.192	99.868						
17	.012	.063	99.931						
18	.010	.050	99.981						
19	.004	.019	100.000						

Extraction Method: Principal Component Analysis.

,


Scree Plot
Eigenvalue
6
4
2
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Component Number

Ύστερα από πολλαπλές δοκιμές και με βάση τα κριτήρια που αναφέρονται παραπάνω, καταλήξαμε σε 4 παράγοντες, οι οποίοι εκφράζουν συνολικά το 80,009% της ολικ ής διακύμανσης

Component Matrix[a]

	Component				
	1	2	3	4	5
SiO2	.343	.169	.819	-.022	.106
pH_field	-.129	.817	-.291	.291	.033
Cond	.928	.015	-.271	-.145	-.031
TOC	.638	.341	.408	.115	-.354
TN	.418	-.403	.024	.739	-.304
Ca	.503	-.656	.049	-.185	-.238
Mg	.675	-.442	.099	.003	.382
Na	.698	.487	-.390	-.088	-.009
K	.499	-.193	-.123	.321	.454
NH4-N	.594	.311	.409	.105	-.442
HCO3	.804	.277	-.077	.027	-.111
Cl	.739	-.052	-.209	-.385	-.072
SO4	.681	-.526	-.225	-.075	.146
F	.090	.784	.038	.220	.176
NO3	.050	-.593	-.221	.718	-.017
Ba_tot	.151	.269	.628	.179	.427
Li_tot	.696	.364	-.013	.053	.127
Sr_tot	.875	-.160	.122	-.130	.141
B_tot	.279	.589	-.550	.130	.023

Extraction Method: Principal Component Analysis.

a. 5 components extracted.

Rotated Component Matrix[a]

	Component				
	1	2	3	4	5
SiO2	.072	-.159	.487	-.143	**.735**
pH_field	-.202	**.896**	.061	-.071	-.041
Cond	**.915**	.105	.301	.052	-.123
TOC	.269	.129	**.827**	.025	.234
TN	.143	-.130	.359	**.899**	-.042
Ca	.496	-.649	.201	.197	-.174
Mg	**.721**	-.314	-.063	.269	.335
Na	**.677**	**.560**	.277	-.118	-.147
K	**.537**	.095	-.189	.443	.286
NH4-N	.213	.083	**.862**	.025	.168
HCO3	**.651**	.286	.484	.053	.024
Cl	**.791**	-.072	.233	-.168	-.179
SO4	**.789**	-.321	-.068	.287	-.076
F	-.046	**.748**	.184	-.131	.303
NO3	.000	-.165	-.175	**.919**	-.123
Ba_tot	-.018	.115	.136	-.028	**.819**
Li_tot	**.583**	.378	.323	.004	.222
Sr_tot	**.820**	-.165	.279	.073	.245
B_tot	.297	**.770**	.049	-.018	-.247

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.[a]
a. Rotation converged in 10 iterations.

Component Transformation Matrix

Component	1	2	3	4	5
1	.873	.042	.440	.163	.130
2	-.134	.835	.298	-.419	.145
3	-.291	-.397	.444	-.128	.738
4	-.254	.339	.091	.883	.180
5	.268	.170	-.716	-.041	.621

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

3.3.1 Αποτελέσματα παραγοντικής ανάλυσης

Ο πρώτος παράγοντας εκφράζει το 33,644% της συνολικής διακύμανσης και συνίσταται από τις θετικές φορτίσεις στις ακόλουθες χημικές παραμέτρους του υπόγειου νερού: Cond (0,915), Sr (0,820), Cl (0,791), SO_4 (0,789), Mg (0,721), Na (0,677), HCO_3 (0,651), Li (0,583) και K (0,537). Η συμμετοχή των προαναφερόμενων χημικών παραμέτρων του υπόγειου νερού φανερώνει ότι ο πρώτος αυτός παράγοντας αποτελεί τον «**Παράγοντας υφαλμύρινσης**» του υπόγειου νερού κι επιβεβαιώνει ότι η αυξημένη αλατότητα στους υδροφόρους ορίζοντες της περιοχής συνδέεται με τη διείσδυση του θαλασσινού νερού προς την ενδοχώρα.

Ο δεύτερος παράγοντας εκφράζει το 20,378% της ολικής διακύμανσης των δεδομένων και σχηματίζεται από τις θετικές φορτίσεις του pH (0,896) και των ιόντων B (0,770), F (0,748) και Na (0,560). Η συμμετοχή των παραμέτρων αυτών δείχνει ότι ο δεύτερος παράγοντας είναι ο «**Παράγοντας απολιθωμένων νερών**». Δείχνει δηλαδή ότι στην περιοχή μελέτης εμφανίζεται υδροφόρος ορίζοντας που περιέχει νερά που δεν ανανεώνονται και παραμένουν για πολλά χρόνια στον υδροφορέα.

Ο τρίτος παράγοντας εκφράζει το 11,605% της συνολικής διακύμανσης και σχηματίζεται από τις θετικές φορτίσεις του TOC (0,827) και του NH_4-N (0,862). Ο παράγοντας αυτός ονομάζεται «**Παράγοντας οργανικής ύλης**» και φανερώνει ότι οι υψηλές συγκεντρώσεις του οργανικού άνθρακα και του αμμωνίου συνδέονται μεταξύ τους. Η αυξημένη περιεκτικότητα σε αμμώνιο είναι οργανικής προέλευσης και προέρχεται από την οργανική ύλη (λιγνίτης) που περιέχεται στα νεογενή ιζήματα της περιοχής.

Ο τέταρτος παράγοντας εκφράζει το 8,465% της ολικής διακύμανσης και σχηματίζεται από τις υψηλές θετικές φορτίσεις του συνολικού Αζώτου, TN (0,899) και των νιτρικών ιόντων, NO_3 (0,919). Ο παράγοντας αυτός ονομάζεται «**Παράγοντας νιτρορύπανσης**» και δείχνει ότι η αυξημένη συγκέντρωση του συνολικού αζώτου στην περιοχή οφείλεται στα νιτρικά ιόντα, τα οποία με τη σειρά τους προέρχονται από τις εντατικές λιπάνσεις των γεωργικών δραστηριοτήτων της περιοχής.

Ο πέμπτος παράγοντας εκφράζει μόλις το 5,917% της ολικής διακύμανσης των δεδομένων και σχηματίζεται από τις αυξημένες φορτίσεις του SiO_2 (0,735) και του Ba (0,819). Ο παράγοντας αυτός δεν εκφράζει κάποια γνωστή υδροχημική διαδικασία

3.4 Ανάλυση κατά συστάδες

Βήματα στη διαδικασία της συσταδιοποίησης

Η διαδικασία της συσταδοποίησης μπορεί να οδηγήσει σε διαφορετικές τμηματοποιήσεις ενός συνόλου δεδομένων, ανάλογα με το κριτήριο που χρησιμοποιείται. Κατά συνέπεια, υπάρχει ανάγκη προεπεξεργασίας των δεδομένων προτού εφαρμοστεί η συσταδοποίηση σε ένα σύνολο δεδομένων. Τα βασικά βήματα για την ανάπτυξη της διαδικασίας της συσταδοποίησης είναι τα παρακάτω : Επιλογή χαρακτηριστικών γνωρισμάτων. Σε αυτό το βήμα βασικός στόχος είναι να επιλεχτούν κατάλληλα τα γνωρίσματα (attributes) στα οποία πρόκειται να εφαρμοστεί η συσταδοποίηση ώστε να κωδικοποιηθεί όσο το δυνατόν περισσότερη πληροφορία σχετικά με την εργασία που μας ενδιαφέρει. Κατά συνέπεια, η προεπεξεργασία των δεδομένων μπορεί να είναι απαραίτητη πριν από την χρησιμοποίηση τους στη διαδικασία της συσταδοποίησης. Επιλογή αλγορίθμου συσταδοποίησης. Αυτό το βήμα αναφέρεται στην επιλογή ενός αλγορίθμου που οδηγεί στον καθορισμό ενός καλού σχήματος συσταδοποίησης (clustering scheme). Το μέτρο γειτνίασης και το κριτήριο συσταδοποίησης που θα χρησιμοποιηθούν χαρακτηρίζουν ουσιαστικά τον αλγόριθμο συσταδοποίησης καθώς και τη δυνατότητα του να καθορίσει ένα σχήμα συσταδοποίησης που να ταιριάζει στο σύνολο δεδομένων. Συνεπώς το βήμα αυτό βασίζεται στα εξής: Στο μέτρο γειτνίασης (proxitity measure) το οποίο προσδιορίζει πόσο «όμοια» είναι δύο αντικείμενα (δηλαδή διανύσματα γνωρισμάτων). Στις περισσότερες από τις περιπτώσεις πρέπει να εξασφαλίσουμε ότι όλα τα επιλεγμένα γνωρίσματα συμβάλλουν εξίσου στον υπολογισμό του μέτρου εγγύτητας και δεν υπάρχει κανένα γνώρισμα που να υπερισχύει των άλλων. Κριτήριο συσταδοποίησης. Σε αυτό το βήμα, πρέπει να καθορίσουμε το κριτήριο συσταδοποίησης το οποίο μπορεί να εκφραστεί μέσω μιας συνάρτησης κόστους ή κάποιου άλλου τύπου κανόνων. Πρέπει να τονίσουμε ότι πρέπει να λάβουμε υπόψη τον τύπο των συστάδων που αναμένονται να εμφανιστούν στο σύνολο δεδομένων. Κατά συνέπεια, πρέπει να καθορίσουμε «ένα καλό» κριτήριο συσταδοποίησης που να οδηγεί σε μια τμηματοποίηση που να ταιριάζει καλά στο σύνολο δεδομένων. Εγκυρότητα αποτελεσμάτων συσταδοποίησης. Η ακρίβεια των αποτελεσμάτων του αλγορίθμου συσταδοποίησης εξακριβώνεται χρησιμοποιώντας τα κατάλληλα κριτήρια και τεχνικές. Εφόσον οι αλγόριθμοι συσταδοποίησης καθορίζουν τις

συστάδες που δεν είναι γνωστές εκ των προτέρων, ανεξάρτητα από τις μεθόδους συσταδοποίησης, η τελική τμηματοποίηση των δεδομένων απαιτεί κάποιου είδους αξιολόγηση στις περισσότερες εφαρμογές. Ερμηνεία αποτελεσμάτων. Σε πολλές περιπτώσεις, οι εμπειρογνώμονες στην περιοχή της εφαρμογής που αναφέρεται η συσταδοποίηση πρέπει να ενσωματώσουν τα αποτελέσματα τη; συσταδοποίησης με αλλά πειραματικά στοιχεία

Διαδικασία clustering

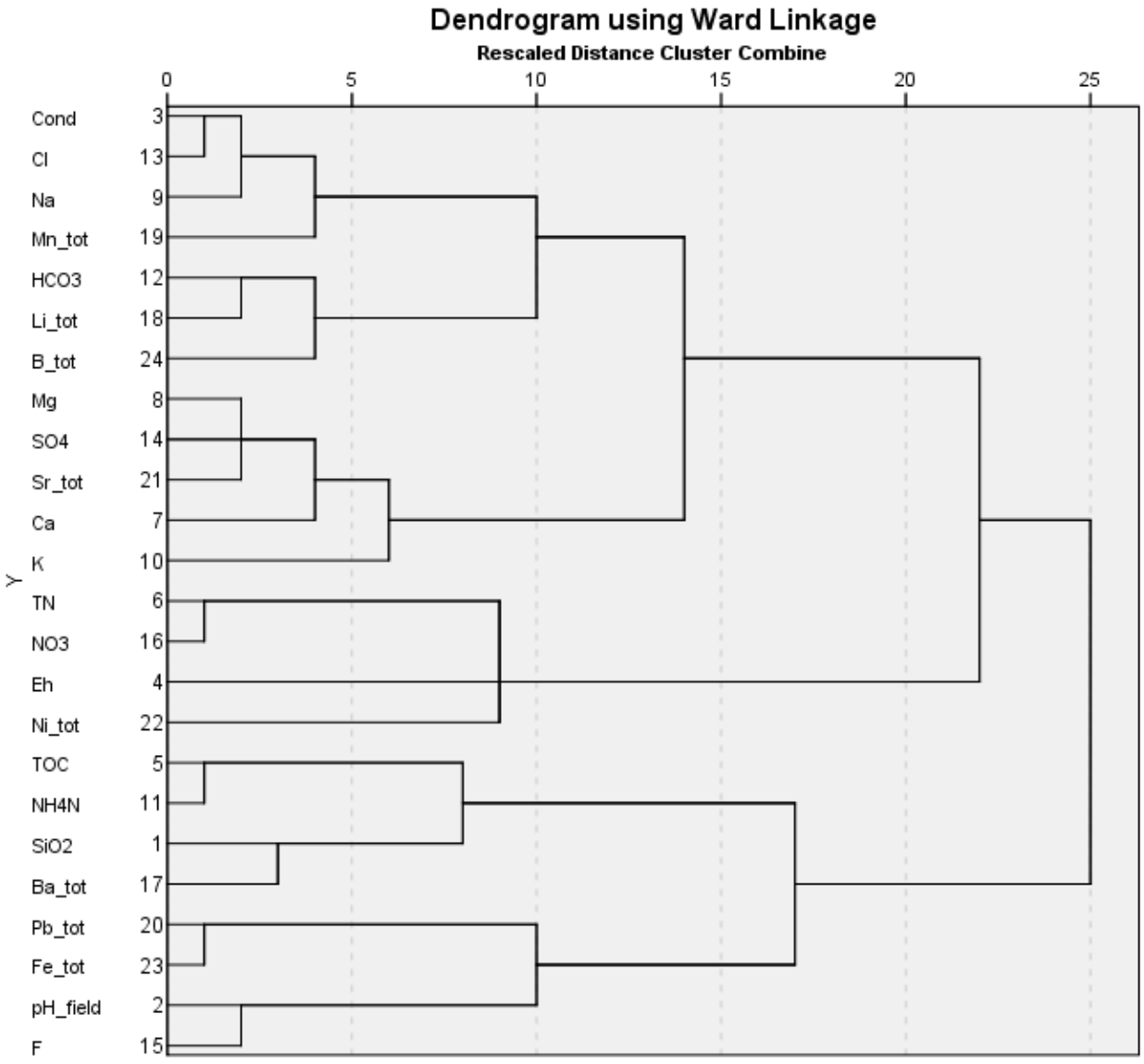
Συνήθως τα βήματα που ακολουθούνται κατά την διαδικασία του Clustering είναι: (a) Αναπαράσταση των στοιχείων η οποία μπορεί να περιλαμβάνει παραγωγή νέων χαρακτηριστικών η επιλογή μέρους των χαρακτηριστικών των στοιχείων. (Pattern representation) (b) Ορισμός του μέτρου ομοιότητας μεταξύ των στοιχείων. (Similarity measure definition) (c) Η καθαυτή διαδικασία της ομαδοποίησης. (Clustering) (d) Αφαίρεση δεδομένων όταν χρειάζεται. (Data abstraction) (e) Προσδιορισμός και εκτίμηση του αποτελέσματος. (Assessment of output)
Η αναπαράσταση των στοιχείων αφορά στον αριθμό των κλάσεων, τον αριθμό των διαθέσιμων στοιχείων, στον αριθμό και τύπο των χαρακτηριστικών τα οποία ενδιαφέρουν τον αλγόριθμο του Clustering. Μερικά από τα προηγούμενα δεν είναι άμεσα διαθέσιμα. Ενδιαφέρον παρουσιάζει η διαδικασία της επιλογής χαρακτηριστικών κατα την οποία βρίσκονται και επιλέγονται τα καταλληλότερα χαρακτηριστικά των στοιχείων τα οποία θα χρησιμοποιηθούν για το Clustering. Εξάλλου, η διαδικασία της εξαγωγής χαρακτηριστικών χρησιμοποιεί μετασχηματισμούς υπαρχόντων χαρακτηριστικών για την παραγωγή άλλων τα οποία πιθανόν να είναι πιο ενδιαφέροντα. Οποιαδήποτε από τις τεχνικές αυτές μπορεί να χρησιμοποιηθεί για την επιλογή των κατάλληλων χαρακτηριστικών γνωρισμάτων για την αναπαράσταση των στοιχείων προς ομαδοποίηση. Το μέτρο ομοιότητας μεταξύ των στοιχείων καθορίζεται από μια συνάρτηση απόστασης. Ένα απλό μέτρο απόστασης όπως η Ευκλείδεια απόσταση μπορεί να χρησιμοποιηθεί για να αντικατοπτρίσει την διαφορά-ανομοιότητα μεταξύ δύο στοιχείων, ενώ άλλα μέτρα απόστασης ποσοτικοποιούν την ομοιότητα των στοιχείων. Αναφορά στα μέτρα απόστασης γίνεται παρακάτω. Η διαδικασία του Clustering μπορεί να πραγματοποιηθεί με πολλούς τρόπους. Το αποτέλεσμα μπορεί να είναι απόλυτα καθορισμένο (ομαδοποίηση των δεδομένων σε ξένες μεταξύ τους κλάσεις), ή fuzzy (όπου κάθε στοιχείο δεν ανήκει μόνο σε μία κλάση αλλά είναι μέλος όλων των κλάσεων με κάποιο βαθμό σε κάθε μια). Οι αλγόριθμοι ιεραρχικού Clustering παράγουν μια σειρά από εμφωλευμένες κλάσεις μετά από διαδικασίες διαχωρισμού ή συγχώνευσης με βάση το μέτρο ομοιότητας μεταξύ των στοιχείων διαφορετικών ομάδων. Οι αλγόριθμοι διαμέρισης από την μεριά τους στοχεύουν στο να διαχωρίσουν τα δεδομένα με τέτοιο τρόπο ώστε να βελτιστοποιείται το κριτήριο με το οποίο γίνεται το Clustering, πιθανόν κάποιο μέτρο ομοιότητας η διαφοροποίησης. Άλλες τεχνικές Clustering βασίζονται στην θεωρία των πιθανοτήτων και άλλες σε θεωρία

γράφων. Η αφαίρεση δεδομένων είναι διαδικασία κατά την οποία το σύνολο των δεδομένων αποκτά μια απλή και συμπαγή αναπαράσταση. Ο όρος απλή αναπαράσταση μπορεί να εξηγηθεί είτε από την μεριά της αυτοποιημένης ανάλυσης είτε από την μεριά του ανθρώπου. Στην πρώτη περίπτωση θα θέλαμε τα δεδομένα μας να αναπαρίστανται με τέτοιο σαφή και απλό τρόπο ώστε μια περαιτέρω υπολογιστική επεξεργασία να είναι εξίσου εφικτή. Στην δεύτερη περίπτωση η απλή αναπαράσταση των δεδομένων τα κάνει πιο κατανοητά σε αυτούς που πρόκειται να τα επεξεργαστούν και να βγάλουν συμπεράσματα. Συνήθως η αφαίρεση δεδομένων στο Clustering είναι μια συνοπτική αναπαράσταση κάθε κλάσης με την βοήθεια κάπου αντιπροσώπου στοιχείου το οποίο καλείται κεντροειδές (centroid). Τέλος η εκτίμηση (validation) της διαδικασίας του Clustering προσπαθεί να εκτιμήσει το αποτέλεσμα ενός αλγορίθμου, να βρει τι χαρακτηρίζει μια καλή διαδικασία από μια όχι και τόσο επιτυχή. Κυρίως αυτό που εκτιμάται είναι το τελικό αποτέλεσμα, δηλαδή κατά πόσο οι κλάσσεις που δημιουργήθηκαν έχουν νόημα και κατά ποσό αυτές δεν δημιουργήθηκαν κατά τυχαίο τρόπο εξαιτίας του συγκεκριμένου αλγορίθμου που χρησιμοποιήθηκε.

Για την ανάλυση κατά συστάδες χρησιμοποιήθηκε το πρόγραμμα IBM SPSS 23.Σε αυτό το κεφάλαιο επεξηγούμε την χρήση της ανάλυσης συστάδων χρησιμοποιώντας τα δεδομένα από τις μετρήσεις που έχουε κανει.Πρώτα ομαδοποιούμε τις παρατηρήσεις χρησιμοποιώντας τις μεθόδους απλής σύνδεσης,τη μέθοδο centroid και του ward.Πολλαπλές μέθοδοι χρησιμοποιούνται για να καθορίσουμε αν οι διαφορετικές αυτές μέθοδοι παράγουν παρόμοιες λύσεις των cluster.Ακολουθείται στη συνέχεια μια τεχνική μη ιεραρχικής ομαδοποίησης.Οι καλύτερες λύσεις που λαμβάνονται από την ιεραρχική διαδικασία θα χρησιμοποιηθούν σαν λύσεις αφετηρίας ή αρχικές λύσεις.

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	13	5.882	0	0	6
2	6	16	14.214	0	0	16
3	5	11	22.680	0	0	15
4	20	23	32.490	0	0	18
5	8	14	43.245	0	0	8
6	3	9	54.614	1	0	12
7	2	15	66.711	0	0	18
8	8	21	81.340	5	0	13
9	12	18	96.666	0	0	11
10	1	17	115.919	0	0	15
11	12	24	137.594	9	0	19
12	3	19	160.113	6	0	19
13	7	8	184.353	0	8	14
14	7	10	216.396	13	0	20
15	1	5	257.924	10	3	21
16	4	6	302.839	0	2	17
17	4	22	349.470	16	0	22
18	2	20	400.920	7	4	21
19	3	12	454.287	12	11	20
20	3	7	525.501	19	14	22
21	1	2	611.594	15	18	23
22	3	4	721.342	20	17	23
23	1	3	846.856	21	22	0


Dendrogram using Ward Linkage
Rescaled Distance Cluster Combine
0
5
10
15
20
25
Y
Cond 3
Cl 13
Na 9
Mn_tot 19
HCO3 12
Li_tot 18
B_tot 24
Mg 8
SO4 14
Sr_tot 21
Ca 7
K 10
TN 6
NO3 16
Eh 4
Ni_tot 22
TOC 5
NH4N 11
SiO2 1
Ba_tot 17
Pb_tot 20
Fe_tot 23
pH_field 2
F 15

3.4.1 Αποτελέσματα ανάλυσης κατά συστάδες

Από το δενδρόγραμμα της ανάλυσης κατά συστάδες προκύπτουν τρεις μεγάλες ομάδες, οι οποίες φανερώνουν διακριτές υδροχημικές διαδικασίες.

Η πρώτη ομάδα του δενδρογράμματος περιλαμβάνει τις παραμέτρους από την Cond έως και το K. Από τη συμμετοχή των παραμέτρων Cond, Cl, Na, Mn, HCO_3, Li, B, Mg, SO_4, Sr, Ca και K προκύπτει ότι η σημαντικότερη υδροχημική διαδικασία στην περιοχή μελέτης είναι η υφαλμύρινση των υπογείων υδάτων από τη διείσδυση του θαλασσινού νερού προς την ενδοχώρα.

Η δεύτερη ομάδα παραμέτρων του δενδρογράμματος περιλαμβάνει τις παραμέτρους TN, NO_3, Eh και Ni και δείχνει ότι μία επίσης σημαντική υδροχημική διαδικασία είναι η νιτρορύπανση των υπόγειων νερών από την εντατική εφαρμογή αζωτούχων λιπασμάτων σε γεωργικές δραστηριότητες.

Η τρίτη ομάδα παραμέτρων του δενδρογράμματος περιλαμβάνει τις παραμέτρους TOC, NH_4-N, SiO_2, Ba, Pb, Fe, pH και F. Ο σχηματισμός αυτής της ομάδας δείχνει την επίδραση της οργανικής ύλης στη χημική σύσταση των υπόγειων νερών, η οποία εξαρτάται κυρίως από το pH.

ΚΕΦΑΛΑΙΟ 4

ΣΥΜΠΕΡΑΣΜΑΤΑ

Βασικός στόχος της συγκεκριμένης διπλωματικής εργασίας είναι η διερεύνηση του υδρογεωλογικού και υδροχημικού καθεστώτος της ευρύτερης περιοχής της Ζαχάρως με σκοπό την μελέτη των υπόγειων νερών με χρήση γεωστατικών μεθόδων.Πρώτον τα σημαντικότερα αποτελέσματα της έρευνας που μπορούν να συνοψιστούν στα παρακάτω συμπεράσματα μέσω της παραγοντικής ανάλυσης τα οποία ειναι τα εξής:

- Η αυξημένη αλατότητα στους υδροφόρους ορίζοντες της περιοχής συνδέεται με τη διείσδυση του θαλασσινού νερού προς την ενδοχώρα.
- Στην περιοχή μελέτης εμφανίζεται υδροφόρος ορίζοντας που περιέχει νερά που δεν ανανεώνονται και παραμένουν για πολλά χρόνια στον υδροφορέα.
- Η αυξημένη περιεκτικότητα σε αμμώνιο είναι οργανικής προέλευσης και προέρχεται από την οργανική ύλη (λιγνίτης) που περιέχεται στα νεογενή ιζήματα της περιοχής.
- Η αυξημένη συγκέντρωση του συνολικού αζώτου στην περιοχή οφείλεται στα νιτρικά ιόντα, τα οποία με τη σειρά τους προέρχονται από τις εντατικές λιπάνσεις των γεωργικών δραστηριοτήτων της περιοχής.

Δεύτερον τα σημαντικότερα αποτελέματα της έρευνας που μπορούν να συνοψιστούν στα παρακάτω συμπεράσματα μέσω της ανάλυσης κατά συστάδες τα οποία είναι τα εξής:

- Είναι η υφαλμύρινση των υπογείων υδάτων από τη διείσδυση του θαλασσινού νερού προς την ενδοχώρα.
- Η νιτρορύπανση των υπόγειων νερών από την εντατική εφαρμογή αζωτούχων λιπασμάτων σε γεωργικές δραστηριότητες.
- Η επίδραση της οργανικής ύλης στη χημική σύσταση των υπόγειων νερών, η οποία εξαρτάται κυρίως από το pH.

Σύμφωνα με τα εξής συμπεράσματα θα πρέπει να ληφθούν κάποια απαραίτητα μέτρα έτσι ώστε να λυθούν τα παραπάνω προβλήμτα στον υδροφόρο ορίζοντα της περιοχής.Οι λύσεις κατά τη γνώμη μου είναι να βρεθέι ένας τρόπος έτσι ώστε να ανανεώνονται τα νερά στον υδροφορέα μέσω ανακύκλωσης των υδάτων και επιπλέον να σταματήσει η εντατική χρήση των λιπασμάτων στις γεωργικές καλλιέργιες και να χρησιμοποιούνται άλλοι οικολογικοί τρόποι ετσι ώστε να αποφευχθεί η μόλυνση του υδροφόρου ορίζοντα από το άζωτο.Επιπλέον να βρεθεί ένας τρόπος έτσι ώστε να μπορεί να ρυθμιστεί το pH με αποτέλεσμα την βελτίωση της καθαρότητας των υπόγειων νερών.

ΒΙΒΛΙΟΓΡΑΦΙΑ

(Johnson, 1998, SPSS INC., 1999)

(Kline, 1998, σελ.11)

(Afifi,1990, σελ. 398)

(Tabachnick & Fidel 1989)

(Maximum Likelihood, Johnson, 1998)

(Johnson & Wicher, 1998)

(Ντζούφρας, 2001)