ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΧΗΜΕΙΑΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΕΦΑΡΜΟΓΕΣ ΠΟΛΥΠΑΡΑΜΕΤΡΙΚΩΝ ΣΤΑΤΙΣΤΙΚΩΝ ΤΕΧΝΙΚΩΝ ΣΤΗ ΧΗΜΙΚΗ ΑΝΑΛΥΣΗ

IMPLEMENTATION OF MULTIVARIATE TECHNIQUES IN CHEMICAL ANALYSIS

ΕΛΕΝΗ ΦΑΡΜΑΚΗ

ΧΗΜΙΚΟΣ

ΑΘΗΝΑ

ΙΟΥΝΙΟΣ 2012

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Εφαρμογές Πολυπαραμετρικών Στατιστικών Τεχνικών στη Χημική Ανάλυση

ΕΛΕΝΗ ΦΑΡΜΑΚΗ

Α.Μ.: 102707

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:

Κ. Ευσταθίου, Καθηγητής ΕΚΠΑ

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ ΠΑΡΑΚΟΛΟΥΘΗΣΗΣ:

Κ. Ευσταθίου, Καθηγητής ΕΚΠΑ

Ν. Θωμαΐδης, Επίκουρος Καθηγητής ΕΚΠΑ

Μ. Κουππάρης, Καθηγητής ΕΚΠΑ

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Κ. Ευσταθίου, Καθηγητής ΕΚΠΑ

Ν. Θωμαΐδης, Επίκουρος Καθηγητής ΕΚΠΑ

Μ. Κουππάρης, Καθηγητής ΕΚΠΑ

Α. Καλοκαιρινός, Καθηγητής ΕΚΠΑ

Α. Τσαντίλη - Κακουλίδου, Καθηγήτρια ΕΚΠΑ

Ε. Μπακέας, Επίκουρος Καθηγητής ΕΚΠΑ

Α. Οικονόμου, Επίκουρος Καθηγητής ΕΚΠΑ

ΗΜΕΡΟΜΗΝΙΑ ΕΞΕΤΑΣΗΣ 2/7/2012

ΠΕΡΙΛΗΨΗ

Σε αυτή τη διατριβή μελετήθηκε η εφαρμογή πολυπαραμετρικών τεχνικών σε μεγάλες βάσεις δεδομένων ταξινόμησης, με σκοπό τη θεωρητική τους παρουσίαση, τη σύγκριση αυτών και την εξαγωγή συμπερασμάτων, σχετικά με το πεδίο εφαρμογής τους και το χειρισμό τους, τις δυνατότητες και τους περιορισμούς τους.

Χρησιμοποιήθηκαν μη επιβλεπόμενες τεχνικές όπως Principal Components Analysis/Factor Analysis (PCA/FA) και Cluster Analysis (CA) αλλά και επιβλεπόμενες όπως Discriminant Analysis (DA), Classification Trees (CT) και Artificial Neural Networks (ANN). Ιδιαίτερη έμφαση δόθηκε στις τεχνικές CT και ANN (μελετήθηκαν τρεις μέθοδοι και αρχιτεκτονικές αντίστοιχα για καθεμιά από αυτές). Ερευνήθηκαν τα πλεονεκτήματα, μειονεκτήματα και ιδιαιτερότητες τους και βελτιστοποιήθηκαν τα μοντέλα ταξινόμησης των τεχνικών. Όλες οι τεχνικές συγκρίθηκαν μεταξύ τους, με κριτήριο τα αποτελέσματα τους (της ορθής ταξινόμησης των δειγμάτων) σε τρεις βάσεις δεδομένων οι οποίες αφορούσαν τους προσδιορισμούς α) μετάλλων-μεταλλοειδών στους τρεις ταμιευτήρες που χρησιμοποιούνται για την ύδρευση της πρωτεύουσας (Υλίκη, Μόρνο και Μαραθώνα), β) μετάλλων-μεταλλοειδών και ανόργανων στοιχείων σε θαλάσσια δείγματα ιζημάτων από μεγάλες ιχθυοκαλλιέργειες της χώρας, γ) σπανίων γαιών σε δείγματα ελαιολάδων από διάφορες περιοχές.

Η DA αν και είναι παραμετρική τεχνική με πολλούς περιορισμούς στην εφαρμογή της, ανταποκρίθηκε στις ανάγκες των προβλημάτων και παρείχε πάντα μια πρώτη άποψη για το πρόβλημα (δυνατότητα ή όχι γραμμικού διαχωρισμού των ομάδων με βάση το Canonical plot της ανάλυσης και αρχική αξιολόγηση των μεταβλητών). Τα ποσοστά ορθής ταξινόμησης που παρείχε ήταν αρκετές φορές συγκρίσιμα με των πιο προηγμένων τεχνικών. Τα CT με 3 διαφορετικές μεθόδους και αρκετή ευελίξια (παρείχαν πολλές παραμέτρους προς δοκιμή και βελτιστοποίηση), επέτυχαν υψηλά ποσοστά ταξινόμησης με λίγες ή πολλές μεταβλητές (περισσότερες συνήθως των ANN), κατασκευάζοντας επαναλήψιμα μοντέλα με δυνατότητες γενίκευσης. Τα ANN αποδείχθηκαν ιδιαίτερα ευέλικτη τεχνική, με δυνατότητες αποτελεσματικής αξιολόγησης των μεταβλητών και εφαρμογής τους σε απλές αλλά και πολυπλοκότερες βάσεις προσεγγίζοντας γραμμικές και μη γραμμικές συναρτήσεις. Κατασκευάστηκαν ανθεκτικά και ευέλικτα μοντέλα. Μειονέκτημά τους αποτέλεσαν ωστόσο, τα φαινόμενα υπερ-προσαρμογής που παρουσιάζουν και χρειάστηκαν προσεκτικοί χειρισμοί για την αποφυγή τους.

Έτσι, τα διαθέσιμα δείγματα διαχωρίστηκαν σε τρεις ομάδες: χρησιμοποιήθηκαν εκτός της συνήθους ομάδας εκπαίδευσης, επιπλέον ομάδες επικύρωσης και ελέγχου. Με τον τρόπο αυτό, έγινε άμεση ταυτοποίηση των φαινομένων υπερ-προσαρμογής (ώστε να διακόπτεται αυτόματα η εκπαίδευση του μοντέλου), αλλά και δοκιμή των μοντέλων σε νέα, "'αγνωστα" δείγμα-

τα, ώστε να ελέγχεται η δυνατότητα γενίκευσης αυτών. Ο διαχωρισμός σε ομάδες έγινε είτε τυχαία (όπως επιτάσσει η σύγχρονη βιβλιογραφία), είτε με βάση της προκατεργασίας με DA (μέθοδος που δεν έχει χρησιμοποιηθεί ποτέ στο παρελθόν). Επιπλέον, έγινε προσπάθεια εφαρμογής όσο το δυνατόν απλούστερων δομών με λίγες παραμέτρους (μεταβλητές, βάρη) αλλά και λειτουργικές μονάδες επεξεργασίας (νευρώνες).

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Χημειομετρία

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Πολυπαραμετρικές τεχνικές, Δέντρα Ταξινόμησης, Τεχνητά Νευρωνικά Δίκτυα

ABSTRACT

This thesis investigated the implementation of multivariate techniques in large classification data bases, targeting their theoretical presentation, comparison and inference, regarding their application field, handling, potentialities and restrictions.

Unsupervised techniques like Principal Components Analysis/Factor Analysis (PCA/FA) and Cluster Analysis (CA) and supervised ones like Discriminant Analysis (DA), Classification Trees (CTs) and Artificial Neural Networks (ANNs) were used. Emphasis was placed on the techniques of CTs and ANNs (three methods and architectures are studied respectively for each one of them). The advantages, disadvantages and their particularities were exploited and the classification models were optimized. All the techniques were compared to each other in terms of their results (the percentages of samples correctly classified) in three data bases, that concerned the determinations of a) metals-metalloids in the three reservoirs that are used for the water supply of Athens (Iliki, Mornos and Marathon), b) metals-metalloids and nutrients in marine sediments from big aquacultures of the country, c) rare earth elements (REE) in olive oil samples from different regions.

Although DA is a parametric multivariate technique, with many restrictions in its implementation, responded to the needs of all the problems and always provided an initial evaluation for that (capability of linear or not linear discrimination on the basis of the Canonical plot of the analysis and initial evaluation of the variables). The percentages of the correct classification it provided, were frequently compared to that of the most sophisticated techniques. CTs with 3 different methods and enough flexibility (they provided many parameters for trials and optimization), resulted in high percentages with the use of few or more variables (usually more than ANNs), constructing reproducible models with generalization. ANNs were proved to be a particularly flexible technique, with potentialities of efficient variables' evaluation and implementation in simple but also complicated data bases, approximating linear and non-linear functions. Robust and flexible models were constructed. However, over-training phenomena seemed to plague ANN and careful handling was needed for their avoidance.

The available samples were split in three sets: except the usual training one, validation ant test sets were used. In this way, an immediate identification of these phenomena was achieved (so that training was automatically interrupted), and moreover, a test of the models in new "unknown" samples was carried out, so that generalization potentialities were checked. Samples sets were split randomly (as modern bibliography dictates), or were based on DA pre-treatment (a method that has never been used in the past). Moreover, the simplest structures were used: with few parameters (variables, weights) and operating processing units (neurons).

SUBJECT AREA: Chemometrics

KEYWORDS: Multivariate techniques, Classification Trees, Artificial Neural Networks

Στους γονείς μου Γιώργο και Βαρβάρα ,
για ότι είμαι τώρα,
και στο σύζυγό μου Βασίλη,
για την αγάπη του, υποστήριξη και κατανόηση.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΡΟΛΟΓΟΣ

Η διατριβή αυτή πραγματοποιήθηκε στο Εργαστήριο Αναλυτικής Χημείας του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών. Η επιθυμία μου να συνεχίσω τις σπουδές μου σε διδακτορικό επίπεδο, αποτελούσε μια από τις παλαιότερες και σοβαρότερες υποσχέσεις που είχα δώσει στον εαυτό μου. Εξαιτίας όμως οικογενειακών και επαγγελματικών υποχρεώσεων το εγχείρημα αυτό, φάνταζε πραγματικά απρόσιτο. Η ολοκλήρωση λοιπόν αυτής της διατριβής, αποτελεί για μένα τη δικαίωση των κόπων μου αλλά και την εκπλήρωση ενός ονείρου.

Από τη θέση αυτή, θα ήθελα να εκφράσω τις ευχαριστίες μου στον επιβλέποντα της παρούσης διατριβής κ. Κωνσταντίνο Ευσταθίου, καθώς και τον κ. Μιχαήλ Κουππάρη, καθηγητές της Αναλυτικής Χημείας του Πανεπιστημίου Αθηνών, για τη συμμετοχή τους στην εξεταστική επιτροπή της παρούσης εργασίας, για το χρόνο που διέθεσαν για τη διόρθωσή της και τις εύστοχες παρατηρήσεις τους. Θα ήθελα επίσης να ευχαριστήσω θερμά τον κ. Νικόλαο Θωμαΐδη, Επίκουρο καθηγητή, ο οποίος βοήθησε πάρα πολύ ώστε να ολοκληρωθεί αυτή η εργασία. Τον ευχαριστώ πολύ για όλα όσα μου δίδαξε, για το επιστημονικό υλικό που μου προσέφερε, τις συμβουλές του, τη συμπαράστασή του και τις ώρες που μου αφιέρωσε.

Επίσης, θα ήθελα να ευχαριστήσω τους Ν. Θωμαΐδη, Ι. Πασιά, Κ. Μπαρκονίκο, Α. Καστρίτη και Π. Νησιανάκη για την εκτέλεση των προσδιορισμών των μετάλλων/μεταλλοειδών και θρεπτικών συστατικών στα δείγματα των θαλάσσιων ιζημάτων και τις Κ. Μηνιώτη και Ε. Ιωάννου από το Χημικό Εργαστήριο του Γ.Π.Α, για την εκτέλεση των προσδιορισμών REE στα δείγματα ελαιολάδων

Τέλος, ευχαριστώ την οικογένεια μου και ιδιαίτερα το σύζυγό μου για τη βοήθεια που μου προσέφερε στην ηλεκτρονική διαμόρφωση του παρόντος αρχείου.

Η διατριβή συνοδεύεται από ηλεκτρονικό παράρτημα (), το οποίο αποτελείται από δύο μέρη: ΘΕΩΡΗΤΙΚΟ (Θ) και ΠΕΙΡΑΜΑΤΙΚΟ (Π). Στο κείμενο της διατριβής γίνονται παραπομπές στο παράρτημα, οι οποίες αναφέρουν το σχετικό κεφάλαιο αυτού και τα κεφαλαία γράμματα Θ ή Π που αντιστοιχούν στα δύο αναφερθέντα μέρη.

Ελένη Φαρμάκη

Αθήνα, Ιούνιος 2012

ΑΡΚΤΙΚΟΛΕΞΑ

Αρκτικόλεξο/ Βραχυγραφία/ Όροι	**Πλήρης αγγλικός όρος**	**Ελληνική απόδοση**	**Σελίδα ορισμού/αναφοράς**
AD	Applicability domain	περιοχή εφαρμογής	85
AND	---	Συνάρτηση λογικής σύζευξης	65
ANN	Artificial Neural Networks	(Τεχνητά) Νευρωνικά Δίκτυα	59
ANOVA	Analysis of variance	Ανάλυση της διακύμανσης	56
ARE	Average Relative Error, %	Μέσο Εκατοστιαίο Σχετικό Σφάλμα	Θεωρητικό παράρτημα
ART	Adaptive Resonance theory Networks	Δίκτυα Προσαρμόσιμου Συντονισμού	Θεωρητικό παράρτημα
Bias	---	Προκατάληψη	57(CT), 62 (ANN)
BMU	Best Matching Unit	Νικητήρος νευρώνας ή νικητής	122
BP	Back-propagation algorithm	Αλγόριθμος οπισθοδιάδοσης ή κανονικός αλγόριθμος	96
BW	Backward stepwise	Αναδρομική βηματική	39
CA	Cluster analysis	Ανάλυση κατά Συστάδες	51
CART	CART-style Exhaustive search method for univariate splits	Μονοπαραμετρική μέθοδος της Διεξοδικής Σάρωσης	57
CGD	Conjugate Gradient Descent	--	108
COV	Covariance	Συνδιακύμανση	32
CP-ANN	Counter-Propagation ANN	--	144
CV	Cross-validation	Διασταυρούμενη αξιολόγηση	55

Αρκτικόλεξο/ Βραχυγραφία/ Όροι	**Πλήρης αγγλικός όρος**	**Ελληνική απόδοση**	**Σελίδα ορισμού/αναφοράς**
CV cost	Cross-validation cost	Κόστος από τη διασταυρούμενη αξιολόγηση	175
CVAAS	Cold Vapor Atomic Absorption Spectrometry	Φασματοφωτομετρία Ατομικής απορρόφησης, τεχνική ψυχρού ατμού	Πειραματικό παράρτημα
Classic CT	Discriminant-based univariate method	Μονοπαραμετρική Διαχωριστική μέθοδος	56
CT	Classification trees	Δέντρα Ταξινόμησης	53
DA	Discriminant Analysis	Διαχωριστική Ανάλυση	37
DB	Davies-Bouldin index	Davies-Bouldin δείκτης	Θεωρητικό παράρτημα
DBD	Delta-bar-Delta algorithm	Delta-bar-Delta αλγόριθμος	Θεωρητικό παράρτημα
FA	Factor Analysis	Ανάλυση Παραγόντων	47
FN	False negative	Ψευδώς αρνητικά ταξινομημένα σημεία	137
FO	Fraction of objects	Κλάσμα δειγμάτων	169
FN	False negative	Ψευδώς αρνητικά ταξινομημένα δείγματα	137
FP	False positive	Ψευδώς θετικά ταξινομημένα δείγματα	137
FW	Forward stepwise	Εμπρόσθια βηματική	39
GA	Genetic Algorithm	Γενετικός αλγόριθμος	142
GCV	Global cross-validation	Σφαιρική διασταυρούμενη αξιολόγηση	55
ICP-MS	Inductively coupled plasma mass spectrometry	Φασματομετρία Ατομικής Μάζας σε επαγωγικά συζευγμένο πλάσμα Αργού	Πειραματικό παράρτημα
KNN	K-Nearest Neighbor	Κ – κοντινότεροι γείτονες	113
K-S	Kennard-Stone algorithm	K-S αλγόριθμος	85

Αρκτικόλεξο/ Βραχυγραφία/ Όροι	**Πλήρης αγγλικός όρος**	**Ελληνική απόδοση**	**Σελίδα ορισμού/αναφοράς**
(L)DA	(Linear) Discriminant Analysis	Γραμμική Διαχωριστική Ανάλυση	37
(L)DFs	(Linear) Discriminant Functions	Γραμμικές διαχωριστικές συναρτήσεις	37
LCM	Discriminant-based linear combination method	Μέθοδος των γραμμικών συνδυασμών	55
MLP	Multi Layers Perceptron	Perceptron πολλαπλών στιβάδων ή Πολυστιβαδικά Νευρωνικά Δίκτυα	66
LM	Levenberg-Marquardt algorithm	Levenberg-Marquardt αλγόριθμος	Θεωρητικό παράρτημα
MSE	Mean Square Error	Μέσο Τετράγωνο Σφάλμα	93
NSE	Normalized Standard Error	Κανονικοποιημένο Τυπικό Σφάλμα	Θεωρητικό παράρτημα
OR	---	Συνάρτηση λογικής διάζευξης	65
PC	Principal Component	Κύρια Συνιστωσα	45
PCA	Principal Components Analysis	Ανάλυση Κυρίων Συνιστωσών	44
Perceptron	---	Λειτουργικού Νευρωνικό Δίκτυο ή Στοιχειώδης Αισθητήρας	62
QDA	Quadratic Discriminant analysis	Δευτεροβάθμια Διαχωριστική Ανάλυση	44
QP	Quick-propagation algorithm	Quick-propagation αλγόριθμος	Θεωρητικό παράρτημα
RBF	Radial Basis Function	Δίκτυα Ακτινιδικής Βάσης	109
REE	Rare Earth Elements	Στοιχεία σπάνιων γαιών	150, 226
RMS ή RMSE	Root Mean Squared Error	Τετραγωνική ρίζα της μέσης τιμής των τετραγώνων των σφαλμάτων	93

Αρκτικόλεξο/ Βραχυγραφία/ Όροι	**Πλήρης αγγλικός όρος**	**Ελληνική απόδοση**	**Σελίδα ορισμού/αναφοράς**
ROC	Receiving Operating Characteristic ή Relative Operating Characteristic	---	213
SOFM ή SOM	Self-organizing feature maps	Αυτό-οργανούμενη απεικόνιση χαρακτηριστικών	116
SOMDI	Self-Orginizing Map Discrimination Index	---	125
SPXY	Sampling set Partioning based on joint x-y distances algorithm	SPXY αλγόριθμος	85
SSOMs	Supervised Self-Organising Maps	Επιβλεπόμενα δίκτυα Kohonen	129
THGA	Transversely Heated Graphite Atomizer	Ισόθερμος ατομοποιητής	Πειραματικό παράρτημα
Threshold	---	Κατώφλι	53, 60, 62
TN	True negative	Αληθώς αρνητικά ταξινομημένα δείγματα	137
TP	True positive	Αληθώς θετικά ταξινομημένα δείγματα	137
U-matrix	Unified distance matrix	Ενιαίος πίνακας αποστάσεων	123
WTPs	Wastewater Treatments Plants	Μονάδες επεξεργασίας λυμάτων	129
XOR	Exclusive OR	Συνάρτηση αποκλειστικής διάζευξης	66

ΚΕΦ. 1 ΠΟΛΥΠΑΡΑΜΕΤΡΙΚΕΣ ΣΤΑΤΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ

1.1. ΕΙΣΑΓΩΓΗ

Σχεδόν σε όλα τα επίπεδα της ανθρώπινης προσπάθειας, αλλά και ειδικότερα στην επιστήμη, έχουν γίνει τα τελευταία χρόνια δραματικές αλλαγές. Έτσι, από την ιστορία της επιστήμης είναι φανερό, ότι λίγες δεκαετίες πριν, το κύριο πρόβλημα των επιστημόνων ήταν πως να αποκτήσουν δεδομένα [1]. Οι μετρήσεις απαιτούσαν πολύ χρόνο, οι μέθοδοι δεν είχαν μεγάλη ευαισθησία, οι τεχνικές ήταν πανάκριβες, ενώ ήταν διαρκώς απαραίτητη η συνεχής παρουσία ενός πεπειραμένου αναλυτή. Οι επιστήμονες έπρεπε να πραγματοποιήσουν μια καθόλου ευχάριστη εργασία ρουτίνας, μόνο και μόνο για να αποκτήσουν κάποια ελάχιστα αποτελέσματα.

Στις μέρες μας, η αλματώδης τεχνολογική ανάπτυξη και ειδικότερα στον τομέα της ενόργανης ανάλυσης, οδηγεί στη συλλογή πολλών δεδομένων (αναλυτικών παραμέτρων) για ένα μόλις δείγμα. Ο κλάδος της εφαρμοσμένης στατιστικής που έχει ως αντικείμενο μελέτης την ανάλυση τέτοιων δεδομένων, ονομάζεται **Πολυπαραμετρική ανάλυση** [2]. Το κύριο πρόβλημα λοιπόν σήμερα, δεν είναι η συλλογή αποτελεσμάτων, αλλά το πως θα απαλλαγούμε από τα περισσότερα από αυτά! Αυτό συμβαίνει γιατί δυστυχώς, μόνο μια μικρή ποσότητα από τα παραγόμενα δεδομένα σχετίζονται με το πρόβλημα. Η πολύτιμη πληροφορία που οι επιστήμονες ψάχνουν, μπορεί να εξαχθεί δύσκολα από το πλήθος των αποτελεσμάτων που τόσο εύκολα αποκτώνται σήμερα. Η **Χημειομετρία** είναι ο κλάδος αυτός της επιστήμης και της τεχνολογίας, που εκμαιεύει όλων των ειδών τις χρήσιμες πληροφορίες από πολυπαραμετρικά δεδομένα, με τη βοήθεια στατιστικών και μαθηματικών μεθόδων. Οι χημειομετρικές μέθοδοι είναι απαραίτητες για την επίλυση πολλών επιστημονικών και πρακτικών προβλημάτων σε πεδία όπως η χημεία, η προστασία του περιβάλλοντος, η ιατρική, η βιολογία, οι δικανικές επιστήμες, η βιομηχανία κ.α. [3]. Έτσι, τα όρια του ανθρώπου σήμερα, εξαρτώνται όχι μόνο από τις ικανότητές του σε ότι βλέπει ή ακούει, ή την εφαρμοσμένη τεχνολογία (αναλυτική οργανολογία) αλλά και την αποτελεσματικότητα των χημειομετρικών μεθόδων. Ειδικότερα, υπάρχει ένα διαρκώς αυξανόμενο ενδιαφέρον για την εφαρμογή της χημειομετρίας στη χημεία, γιατί μπορεί και λύνει προβλήματα που απαιτούν πολύπλοκους υπολογισμούς και βοηθά στην πλήρη ερμηνεία δεδομένων, αντικειμένων, διαδικασιών και φαινομένων. Υπάρχουν δηλαδή δεδομένα, αντικείμενα, διαδικασίες και φαινόμενα που περιέχουν "κρυφές" πληροφορίες, όχι άμεσα προσιτές. Η πρόσβαση σε αυτές τις πληροφορίες εξαρτάται συνεχώς αυξανόμενα, από τις χημειομετρικές μεθόδους, με τις οποίες μπορεί να γίνει αξιόπιστα η ερμηνεία της μετρούμενης πληροφορίας [3].

Στην εργασία αυτή, θα ασχοληθούμε αποκλειστικά με προβλήματα ομαδοποίησης/ταξινόμησης και επομένως η ανάλυση που ακολουθεί (θεωρητική και πειραματική), αφορά εκτός αν αναφέρεται διαφορετικά, μόνο τέτοιου είδους περιπτώσεις.

1.2. ΣΥΝΔΥΑΣΜΟΣ ΜΕΤΑΒΛΗΤΩΝ

1.2.1. Συνδιακύμανση (Covariance)

Κατά την εφαρμογή των πολυπαραμετρικών στατιστικών τεχνικών, είναι σημαντικό, οι μετρούμενες παράμετροι (μεταβλητές), να μην εξετάζονται ξεχωριστά, αλλά να συνδυάζονται ώστε να παρέχουν και να περιγράφουν όσο το δυνατό ένα πλήρες σύστημα. Μεταβλητές που δεν αλληλεπιδρούν με άλλες, θεωρούνται στατιστικά ανεξάρτητες, δηλαδή αλλαγές που συμβαίνουν σε μια μεταβλητή δεν μπορούν να επηρεάσουν ή να προβλέψουν αλλαγές σε κάποια άλλη μεταβλητή. Ωστόσο ορισμένες φορές, οι μεταβλητές δεν είναι ανεξάρτητες και απαιτείται η εκτίμηση της αλληλεπίδρασής τους ώστε να ερμηνευτούν τα δεδομένα και να χαρακτηριστούν τα δείγματα. Η έκταση της αλληλεπίδρασης μεταξύ δυο μεταβλητών j, k εκτιμάται με τον υπολογισμό της **συνδιακύμανσης** (covariance) σε αντιδιαστολή με τη **διακύμανση ή διασπορά** (variance) που περιγράφει τη μεταβολή και διασπορά μίας μεταβλητής. Η συνδιακύμανση s^2 ή COV (j,k) δίνεται από τη σχέση (1.1):

$$\text{COV (j,k)} = \frac{\sum x_d{}^2}{N-1} = \frac{\sum_{i=1}^{N}(x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k})}{N-1} \qquad (1.1)$$

όπου: $x_d = x_i - \bar{x}$, για κάθε δείγμα i και κάθε j ή k μεταβλητή,

x_{ij} = η τιμή που αντιστοιχεί στο i δείγμα της j μεταβλητής,

x_{ik} = η τιμή που αντιστοιχεί στο i δείγμα της k μεταβλητής, και

N = ο συνολικός αριθμός δειγμάτων.

Η συνδιακύμανση δίνει ουσιαστικά τη διαφορά της κάθε μεταβλητής από το μέσο όρο της σε σχέση με κάποια άλλη. Η μέτρηση αυτή γίνεται πάντα μεταξύ **δύο** μεταβλητών. Αν υπολογιστεί η συνδιακύμανση μεταξύ μιας μεταβλητής και του εαυτού της, θα βρεθεί η διακύμανση αυτής. Έτσι, αν έχουμε δεδομένα τριών (3) μεταβλητών (ή διαστάσεων) x, y, z, θα πρέπει να υπολογίσουμε τη συνδιακύμανση μεταξύ των x και y, y και z, z και x. Με αυτόν τον τρόπο, δημιουργείται πάντα ένας πίνακας δεδομένων (στην περίπτωσή μας 3×3) που ονομάζεται πίνακας **διακυμάνσεων-συνδιακυμάνσεων** (variance-covariance) των μεταβλητών (βλ. § 2.2). Οι τιμές της διαγωνίου εδώ, αντιστοιχούν στις διακυμάνσεις των μεταβλητών, ενώ ο πίνακας είναι συμμετρικός ως προς τη διαγώνιο (COV (x, y) = COV (y, x)).

Η ακριβής τιμή της συνδιακύμανσης δεν έχει τόση σημασία **όσο το πρόσημο αυτής.** Αν αυτό είναι θετικό, σημαίνει ότι οι μεταβλητές αυξάνονται συγχρόνως, ενώ αν αυτό είναι αρνητικό, όταν μια μεταβλητή αυξάνεται, η άλλη μειώνεται. Αν η τιμή της συνδιακύμανσης είναι μηδέν (0), σημαίνει ότι οι δυο μεταβλητές είναι ανεξάρτητες [4].

1.2.2. Γραμμικοί συνδυασμοί μεταβλητών

Η ερμηνεία πολλών πολυπαραμετρικών προβλημάτων μπορεί να απλοποιηθεί λαμβάνοντας υπόψη όχι μόνο τις αρχικές μεταβλητές, αλλά επίσης και ένα γραμμικό συνδυασμό αυτών. Έτσι δημιουργούνται συνδυασμοί των αρχικών μεταβλητών κατάλληλα "ζυγισμένων", ανάλογα με την κρισιμότητα της καθεμιάς από αυτές. Οποιαδήποτε μέθοδος και να χρησιμοποιηθεί γι' αυτό, ο σκοπός είναι να μειωθεί ο αριθμός των μεταβλητών και να ληφθεί μια βελτιωμένη αναπαράσταση των αρχικών δεδομένων.

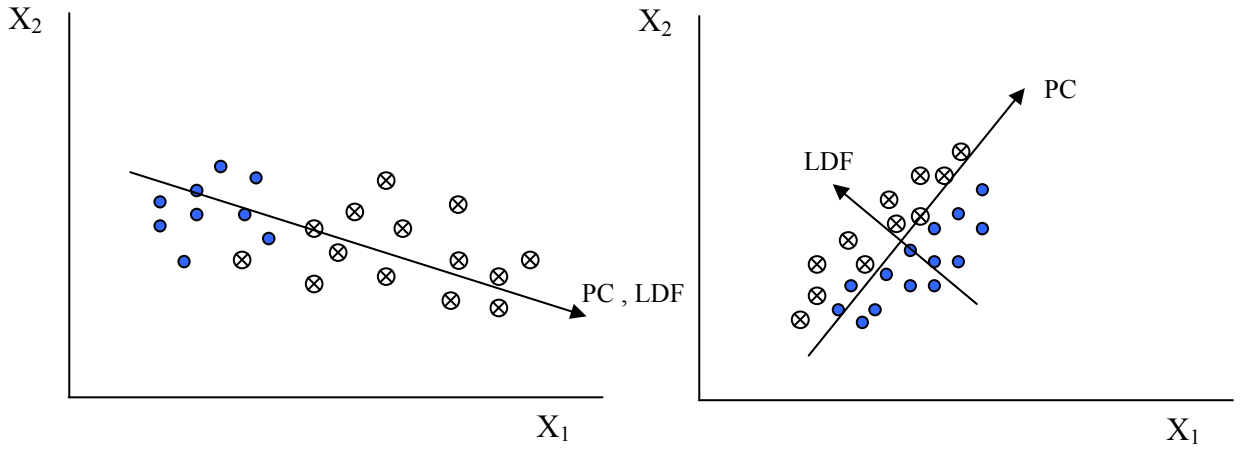
Διαφορετικοί γραμμικοί συνδυασμοί των αρχικών παραμέτρων μπορούν να προκύψουν ανάλογα με τη φύση του προβλήματος που αντιμετωπίζεται. Έτσι:

- ✓ Ο γραμμικός συνδυασμός ο οποίος "αναδεικνύει" τη μέγιστη διαφοροποίηση μεταξύ των ομάδων των δεδομένων (σχ 1.1 (α)), είναι κατάλληλος στην **επιβλεπόμενη** (supervised) **αναγνώριση σχηματομορφής ή προτύπων** (pattern recognition). Ο παραπάνω προβληματισμός θέτει τα θεμέλια για τη **Γραμμική Διαχωριστική Ανάλυση** (Linear Discriminant Analysis, LDA, βλ. § 2.1).
- ✓ Ο γραμμικός συνδυασμός ο οποίος "αναδεικνύει" τη μέγιστη διακύμανση (σχ 1.1 (β)), είναι κατάλληλος όταν ζητείται η μείωση των διαστάσεων του προβλήματος. Ο παραπάνω προβληματισμός θέτει τα θεμέλια για την **Ανάλυση Κυρίων Συνιστωσών** (Principal Components Analysis, PCA, βλ. § 2.2), η οποία εκπροσωπεί τη **μη επιβλεπόμενη** (unsupervised) **αναγνώριση σχηματομορφής ή προτύπων** (pattern recognition).

Η διάκριση μεταξύ επιβλεπόμενων και μη επιβλεπόμενων τεχνικών γίνεται παρακάτω (§ 1.3), αλλά παράλληλα αποκαλύπτεται και μέσα από την περιγραφή της θεωρίας των αναφερόμενων στατιστικών πολυπαραμετρικών τεχνικών. Γενικά, στις επιβλεπόμενες τεχνικές, είναι εκ των προτέρων γνωστές οι ομάδες ταξινόμησης και στη διάρκεια της ανάλυσης, αξιολογείται ο βαθμός επιτυχίας της εύρεσης των ομάδων αυτών με τη χρήση **των ομάδων εκπαίδευσης ή εκμάθησης και ελέγχου** (§ 2.1.1). Στις μη επιβλεπόμενες τεχνικές, η ταξινόμηση γίνεται με τη βοήθεια δειγμάτων (αντικειμένων) για τα οποία δεν είναι εκ των προτέρων γνωστή η ομάδα που ανήκουν.

Επιβλεπόμενες και μη τεχνικές χρησιμοποιούνται παράλληλα και συγκριτικά. Ο συνδυασμός επιβλεπόμενων και μη επιβλεπόμενων τεχνικών στην ίδια βάση δεδομένων μπορεί να

δώσει πρόσθετες πληροφορίες για τη δομή των δεδομένων αυτών, να αποκαλύψει την ύπαρξη κρυφών συσχετίσεων, να χαρακτηρίσει ομάδες δειγμάτων και να βοηθήσει στην εύρεση μαθηματικών μοντέλων με ικανότητες πρόβλεψης ή ομαδοποίησης [5]. Ωστόσο γενικά, οι τεχνικές επιβλεπόμενης ταξινόμησης θεωρούνται ότι επιδεικνύουν ανθεκτικότητα και επιτυγχάνουν ακριβέστερα αποτελέσματα σε σχέση με τις μη επιβλεπόμενες τεχνικές [6].



(α) (β)

Σχήμα 1.1: Παραδείγματα όπου η PCA (PC στο σχήμα) και η DA (LDF στο σχήμα) είναι ακριβώς ίδιες (α) ή ορθογώνιες (β)(βλ. επίσης § 2.1.1, 2.2).

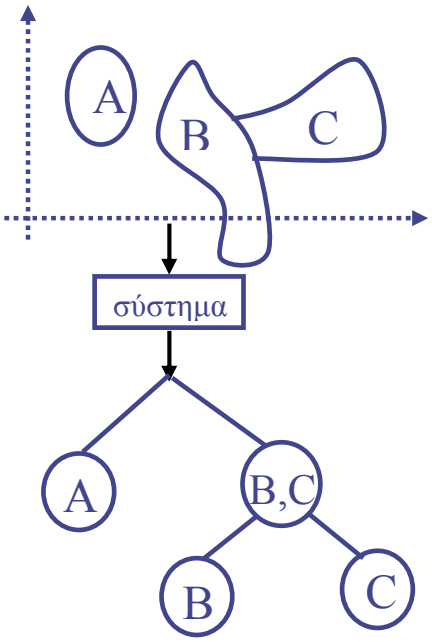
1.3. ΕΠΙΒΛΕΠΟΜΕΝΕΣ ΚΑΙ ΜΗ ΕΠΙΒΛΕΠΟΜΕΝΕΣ ΤΕΧΝΙΚΕΣ ΑΝΑΓΝΩΡΙΣΗΣ ΠΡΟΤΥΠΩΝ

Όπως ήδη αναφέρθηκε, η πορεία μιας ταξινόμησης μπορεί να γίνει με επιβλεπόμενο ή μη επιβλεπόμενο τρόπο.

Στη διάρκεια μιας μη επιβλεπόμενης τεχνικής, το σύστημα εφοδιάζεται με τα δεδομένα και αφήνεται ή όχι να "κατασταλάξει" σε μια σταθερή κατάσταση (σχ. 1.2) [1]. Δομικό χαρακτηριστικό των μη επιβλεπόμενων τεχνικών είναι η βελτιστοποίηση, η οποία χρησιμοποιείται για την αξιολόγηση του αποτελέσματος στο τέλος κάθε περιόδου ή κύκλου. Αυτή αφορά κάποιο γενικότερο κριτήριο, όπως την ελαχιστοποίηση της απόστασης μεταξύ των αντικειμένων ή την επίτευξη ορισμένου αριθμού περιόδων.

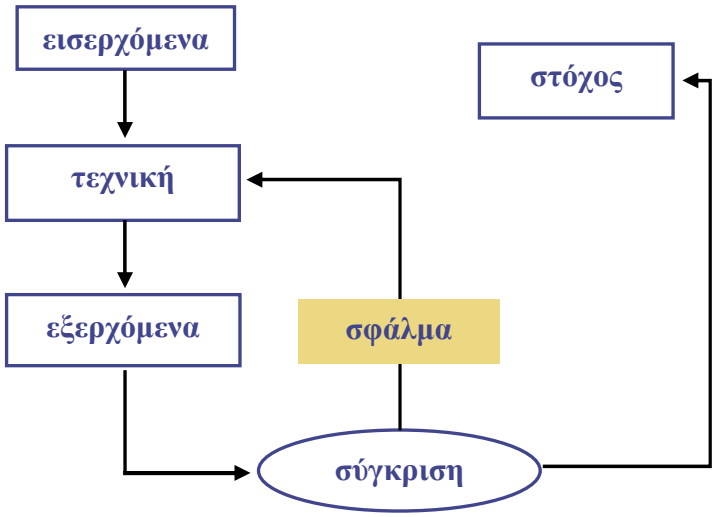
Η εκπαίδευση λοιπόν στις μη επιβλεπόμενες τεχνικές είναι βασικά μια διαδικασία βελτιστοποίησης. Παράδειγμα αποτελεί η μέθοδος της **Ανάλυσης κατά Συστάδες** (Cluster analysis, CA, βλ. § 2.4), όπου δύο είναι τα βασικά κριτήρια: η απόσταση ανάμεσα στα αντικείμενα και η

απόσταση ανάμεσα στις συστάδες των αντικειμένων. Το τελικό αποτέλεσμα οφείλει να είναι **έκπληξη** ή τουλάχιστον να είναι ανεξάρτητο από τις προσδοκίες μας!



Σχήμα 1.2: Μη επιβλεπόμενη τεχνική εκπαίδευσης [1]

Ωστόσο, στην περίπτωση που υπάρχουν κάποια αντικείμενα των οποίων η απόκριση είναι γνωστή, σχηματίζονται "ζεύγη" δεδομένων (εισερχόμενα και στόχοι-ομάδες για την περίπτωση της ταξινόμησης). Εδώ, ο σκοπός των επιβλεπόμενων τεχνικών είναι η **δημιουργία μοντέλων**, που θα συνδυάσει σωστά τα εισερχόμενα με τους στόχους (σχ. 1.3). Κατά μια έννοια οι στόχοι δεν συμμετέχουν στη διαδικασία εκπαίδευσης: απλά εξυπηρετούν ως κριτήριο για το πόσο καλά έχει "εκπαιδευτεί" το σύστημα.



Σχήμα 1.3: Επιβλεπόμενη τεχνική εκπαίδευσης [1]

Κατά την εφαρμογή μιας επιβλεπόμενης τεχνικής εκπαίδευσης, πρέπει να διαχωρίσουμε δυο περιπτώσεις, οι οποίες διαφέρουν μεταξύ τους στον τρόπο που οι στόχοι συνδυάζονται με τα εισερχόμενα, δηλαδή:

- αν υπάρχουν εγγενείς συσχετίσεις μεταξύ αυτών ή
- αν συσχετίζονται αυθαίρετα.

Παράδειγμα αυθαίρετης συσχέτισης, αποτελεί ο αριθμός δύο με το σύμβολο 2 ή 10 (δεκαδικό ή δυαδικό σύστημα αρίθμησης αντίστοιχα). Η χημική δομή ωστόσο, και το IR φάσμα των ενώσεων έχουν εγγενή συγγένεια, εφόσον η δομή "προκαλεί" το φάσμα. Σε αυτή τη δεύτερη περίπτωση, είναι φανερό ότι μπορεί να γίνει "γενίκευση" του μοντέλου και να προβλεφθούν οι αποκρίσεις για αντικείμενα διαφορετικά από αυτά που χρησιμοποιήθηκαν κατά την εκπαίδευση [1].

ΚΕΦ. 2 ΚΛΑΣΙΚΕΣ ΜΕΘΟΔΟΙ

2.1. ΔΙΑΧΩΡΙΣΤΙΚΗ ΑΝΑΛΥΣΗ (DISCRIMINANT ANALYSIS)

2.1.1. Αξιολόγηση μεταβλητών

Η πιο δημοφιλής **παραμετρική** (δηλ. για κανονικές κατανομές) μέθοδος για την επιβλεπόμενη αναγνώριση σχηματομορφής ή προτύπων (supervised pattern recognition) είναι η **Γραμμική Διαχωριστική Ανάλυση** (Linear Discriminant analysis, LDA ή DA). Στην DA μπορούμε να διακρίνουμε δυο διαφορετικούς στόχους: την εύρεση των γραμμικών συναρτήσεων (discriminate predictive analysis, βλ. παρακάτω) και την αξιολόγηση των συναρτήσεων αυτών, ώστε να ταξινομήσουμε τρέχοντα και μελλοντικά δείγματα (discriminate classification analysis, § 2.1.2) [7].

Στη διάρκεια της πρώτης, επιχειρείται να καθοριστούν οι μεταβλητές, που διαχωρίζουν δύο ή παραπάνω φυσικές ομάδες δειγμάτων. Ο σκοπός εδώ δηλαδή, είναι να βρεθούν οι μεταβλητές αυτές που είναι κρίσιμες στην ταξινόμηση των δειγμάτων σε γνωστές (τουλάχιστον αρχικά) εκ των προτέρων ομάδες.

Έτσι, η Διαχωριστική Ανάλυση, χρησιμοποιώντας μια **αρχική ομάδα εκπαίδευσης ή εκμάθησης** ("training" ή "learning set"), αναζητά τις **γραμμικές διαχωριστικές συναρτήσεις** (Linear Discriminant Functions (L)DFs ή Canonical Functions (roots), σχήμα 1.1), Y οι οποίες είναι γραμμικός συνδυασμός των αρχικών μεταβλητών X_1, X_2 ….X_p κ.λπ.:

$$Y = a_1X_1 + a_2X_2 + \ldots..a_pX_p \ (+ C) \qquad (2.1)$$

όπου C (όταν υπάρχει) η διαχωριστική σταθερά.

Οι συντελεστές a_1, a_2, ……, a_p είναι τέτοιοι ώστε να επιτυγχάνεται ο μέγιστος διαχωρισμός μεταξύ των ομάδων και αναλογικά οι μεγαλύτεροι από αυτούς (βλ. παρακάτω) αντιστοιχούν στις μεταβλητές **που συνεισφέρουν περισσότερο στο διαχωρισμό των ομάδων.** Έτσι η Y επιτυγχάνει καλύτερο διαχωρισμό απ' ότι οι αρχικές X_1, X_2….X_p κ.λπ. [8]. Οι συντελεστές αυτοί αποτελούν τα βάρη που εκφράζουν το βαθμό που κάθε ανεξάρτητη μεταβλητή X συνεισφέρει στο διαχωρισμό των ομάδων [9]. Το Y για κάθε δείγμα είναι το σκορ (score), δηλαδή ένας αριθμός που χρησιμοποιείται για την πρόβλεψη της ομάδας που ανήκει [9].

Η συγκεκριμένη σχέση 2.1 εμπεριέχει τους συντελεστές a_1, a_2, ……, a_p στους οποίους δεν έχει γίνει κάποια κανονικοποίηση (unstandardized Canonical discriminant function coefficients). Δεν μπορεί λοιπόν να γίνει σύγκριση ανάμεσά τους, αν οι συντελεστές δεν κανονικοποιηθούν (standardized Canonical discriminant function coefficients). Οι τελευταίοι είναι χρήσιμοι όταν έχουμε ανεξάρτητες μεταβλητές διαφορετικής κλίμακας και δίνουν μια ένδειξη της συνεισφοράς της κάθε μεταβλητής στη διαχωριστική συνάρτηση [10]. Στην περίπτωση αυτή, δεν θα υπάρχει διαχωριστική σταθερά C [9, 11].

Προκειμένου να ταυτοποιηθούν οι μεταβλητές που βοηθούν στη διαφοροποίηση των ομάδων, κάποιος μπορεί ακόμα να ελέγξει τους **συντελεστές δομής** ή **διαχωριστικές φορτίσεις** (structure coefficients/discriminant loadings) όπως ονομάζονται εναλλακτικά. Αυτοί δείχνουν τη σχέση ανάμεσα σε μια ανεξάρτητη αρχική μεταβλητή και τη νέα LDF και μπορούν να χρησιμοποιηθούν για να αξιολογήσουμε πόσο σημαντική είναι κάθε μεταβλητή για τη κατάσκευή της διαχωριστικής συνάρτησης [10]. Όσο μεγαλύτερες είναι αυτές οι τιμές, τόσο καλύτερη η διαχωριστική ικανότητα της μεταβλητής [9, 11]. Οι συντελεστές δομής δεν επηρεάζονται από τη συνδιακύμανση των αρχικών μεταβλητών (βλ. παρακάτω § 2.1.3) και επομένως είναι ασφαλέστεροι για την εξαγωγή συμπερασμάτων που αφορούν την κρισιμότητα τους [12].

Αναλυτικότερα, οι LDF είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών που επιτυγχάνουν να μεγιστοποιήσουν την παράμετρο **F-ratio.** Η F-ratio παράμετρος ορίζεται ως ο λόγος των διακυμάνσεων ανάμεσα στις ομάδες (between groups) και στην ίδια ομάδα (within group). Οι μεταβλητές με τα υψηλότερα F-ratios έχουν την καλύτερη διαχωριστική ικανότητα [13, 14].

Η παράμετρος F-ratio δείχνει για κάθε μεταβλητή τη στατιστική σημασία της στο διαχωρισμό των ομάδων, δηλαδή την έκταση που αυτή συνεισφέρει στην πρόβλεψη των σωστών ομάδων και συνδέεται με τις παραμέτρους Wilks' lambda/model (Λ) και Partial lambda, με βάση την παρακάτω σχέση:

$$\text{F-ratio} = \frac{1-\Lambda}{\Lambda} = \frac{N-p-G}{G-1} \times \frac{1-\text{partial lambda}}{\text{partial lambda}} \qquad (2.2)$$

όπου: G ο αριθμός των ομάδων, N ο αριθμός των δειγμάτων, και p ο αριθμός των μεταβλητών. Η παράμετρος Λ του μοντέλου δίνεται από τη σχέση:

$$\Lambda = \frac{\sum (Y_{ig} - \overline{Y_g})^2}{\sum (Y_{ig} - \overline{Y})^2} = \frac{\text{within}}{\text{total}} \qquad (2.3)$$

όπου: Y τα σκορ (βλ. σχέση 2.1) και ειδικότερα, Y_{ig} το σκορ του δείγματος i που ανήκει στην ομάδα g, $\overline{Y_g}$ το μέσο σκορ των δειγμάτων της ομάδας g και $\overline{Y}$ το μέσο σκορ όλων των δειγμάτων [9]. Η παράμετρος αυτή χρησιμοποιείται για να ελέγξει την κρισιμότητα μιας LDF ως σύνολο (πόσο καλά μια διαχωριστική συνάρτηση διαχωρίζει τα δείγματα σε ομάδες). Όσο μικρότερο είναι το Λ, τόσο σημαντικότερη η διαχωριστική συνάρτηση. Σημαντικό Λ σημαίνει ότι κάποιος μπορεί να απορρίψει τη μηδενική υπόθεση ότι δυο ή περισσότερες ομάδες έχουν τις ίδιες μέσες τιμές σκορ, οπότε το μοντέλο μπορεί να τις διαχωρίσει [15, 16]. Ισούται με το κλάσμα της συνολικής διακύμανσης των σκορ που δεν ερμηνεύεται από τις διαφορές ανάμεσα στις ομάδες [12]. Θα αναφερθούμε σε αυτήν και παρακάτω (§ 2.1.3).

Η παράμετρος Partial lambda παίρνει μεγαλύτερες τιμές για τις μεταβλητές με τη χειρότερη διαχωριστική ικανότητα. Συνδέεται με την παράμετρο Wilks' lambda/variable ή U statistic [16] (περισσότερες λεπτομέρειες αναφέρονται στο ηλεκτρονικό παράρτημα της διατριβής,  ΚΕΦ. 1, Θ).

Στην περίπτωση ύπαρξης μόνο δυο ομάδων δειγμάτων, η DA κατασκευάζει μία μόνο γραμμική συνάρτηση, ενώ στην περίπτωση 3 ομάδων, θα μπορούσαμε να θεωρήσουμε:

1. μία γραμμική διαχωριστική συνάρτηση που να διαχωρίζει την ομάδα 1 από τις 2 και 3 μαζί και
2. μία άλλη συνάρτηση για το διαχωρισμό των ομάδων 2 και 3 μεταξύ τους.

Η πρώτη από τις συναρτήσεις αυτές επιτυγχάνει τον καλύτερο συνολικά διαχωρισμό των ομάδων, η δεύτερη τον επόμενο κ.ο.κ. Οι συναρτήσεις αυτές είναι ανεξάρτητες και ορθογώνιες, δηλαδή η συνεισφορά τους στο διαχωρισμό δεν αλληλεπικαλύπτεται. Ο μέγιστος αριθμός των LDF που υπολογίζονται ισούται με τον αριθμό των ομάδων μείον ένα ή τον αριθμό των μεταβλητών της ανάλυσης (όποιο είναι μικρότερο, βλ. § 3.2.1) [17].

Η DA μπορεί να επιτύχει την εύρεση των πιο κρίσιμων μεταβλητών με τη βοήθεια 3 (τριών) προσεγγίσεων:

- ✓ Την **κλασική προσέγγιση** (standard) που χρησιμοποιεί όλες τις μεταβλητές αξιολογώντας ωστόσο την καθεμιά, με βάση στατιστικά μεγέθη που έχουν ήδη καθοριστεί στην αρχή της ανάλυσης (F εισόδου/αποκλεισμού ή F to enter/remove) και συγκρίνονται με τις τιμές F που προκύπτουν από τις μεταβλητές.
- ✓ Την **εμπρόσθια βηματική ή κλιμακωτή προσέγγιση** (forward stepwise, FW), η οποία "χτίζει" το μοντέλο βήμα-βήμα. Σε κάθε βήμα, ανασκοπούνται οι μεταβλητές και αποφασίζεται ποια συνεισφέρει περισσότερο στο διαχωρισμό μεταξύ των ομάδων, ώστε να παραμείνει για το επόμενο βήμα [17].
- ✓ Την **αναδρομική βηματική προσέγγιση** (backward stepwise, BW), περιλαμβάνει αρχικά στο μοντέλο όλες τις μεταβλητές, και σε κάθε βήμα αποκλείει τη μεταβλητή που συνεισφέρει λιγότερο στην επιτυχημένη πρόβλεψη των ομάδων. Έτσι σε μια επιτυχημένη ανάλυση, τελικά διατηρούνται οι "κρίσιμες" μεταβλητές, δηλαδή αυτές που συνεισφέρουν περισσότερο στο διαχωρισμό [17].

Στις βηματικές προσεγγίσεις οι μεταβλητές εισέρχονται στο μοντέλο όταν η αντίστοιχη τιμή F είναι μεγαλύτερη από την τιμή F εισόδου που καθορίζεται από το χρήστη. Μεταβλητές επίσης που έχουν τιμή του F μικρότερη από το καθορισμένο F αποκλεισμού, απομακρύνονται από το μοντέλο. Η τιμή του F εισόδου είναι πάντα μεγαλύτερη της τιμής του F αποκλεισμού. Αν στην τεχνική FW, θέλουμε να χρησιμοποιηθούν όλες οι μεταβλητές, η τιμή του F εισόδου πρέπει να είναι πολύ μικρή (πχ 0,0001) και η τιμή του F αποκλεισμού 0. Αν στην τεχνική BW,

θέλουμε να απομακρύνουμε όλες οι μεταβλητές, η τιμή του F εισόδου πρέπει να είναι πολύ μεγάλη (πχ 9999) και η τιμή του F αποκλεισμού οριακά μικρότερη (πχ 9998) [17].

2.1.2. Ομαδοποίηση - Ταξινόμηση

Η DA χρησιμοποιείται ακόμα για την πρόβλεψη της ομάδας αντικειμένων (δειγμάτων). Ο σκοπός εδώ, είναι να βρεθούν οι κανόνες εκείνοι που θα μπορούν στο μέλλον να κατατάξουν "άγνωστα" δείγματα στις ομάδες τους χρησιμοποιώντας όμως τις μετρηθείσες παραμέτρους.

Το πρώτο βήμα είναι να ξεκινήσουμε με μια ομάδα αντικειμένων των οποίων η κατάταξη είναι γνωστή, όπως για παράδειγμα δείγματα νερού γνωστής προέλευσης. Αυτή η αρχική ομάδα ονομάζεται **ομάδα εκπαίδευσης.** Ο σκοπός είναι να χρησιμοποιηθούν αυτά τα αντικείμενα, ώστε να βρεθεί ο κανόνας (συνάρτηση) που να κατατάσσει και τα άγνωστα αντικείμενα (π.χ. άγνωστης προέλευσης δείγματα νερού) στην ομάδα που ανήκουν.

Η DA τώρα υπολογίζει αυτόματα τις **συναρτήσεις ταξινόμησης** (Fisher's Classification Functions). Οι συναρτήσεις αυτές δεν πρέπει να συγχέονται με τις γραμμικές διαχωριστικές συναρτήσεις (LDF, § 2.1.1). Οι συναρτήσεις ταξινόμησης χρησιμοποιούνται για τον καθορισμό της ομάδας που ανήκει ένα δείγμα. Είναι τόσες, όσες και οι αντίστοιχες ομάδες. Κάθε τέτοια συνάρτηση μας επιτρέπει τον υπολογισμό των σκορ για κάθε ομάδα, με βάση τη σχέση:

$$Si = c_i + w_{i1}X_1 + w_{i2}X_2 + \ldots.. w_{ip}X_p \qquad (2.4)$$

Στη σχέση αυτή, ο δείκτης i δηλώνει την αντίστοιχη ομάδα, c_i είναι η σταθερά για την i ομάδα, w_{ij} (j = 1…..p) είναι το βάρος για την j μεταβλητή και Xj η παρατηρούμενη τιμή της ίδιας μεταβλητής στο συγκεκριμένο δείγμα. Si είναι το τελικό υπολογιζόμενο αποτέλεσμα (ή Fisher's score). Μπορούμε να χρησιμοποιήσουμε τις συναρτήσεις ταξινόμησης για τον άμεσο υπολογισμό της ομάδας που ανήκει ένα άγνωστο δείγμα: το μεγαλύτερο αποτέλεσμα δηλώνει την ομάδα του δείγματος [17]. Τα μεγαλύτερα βάρη (συντελεστές) w_{ij} αντιστοιχούν στις μεταβλητές που συνεισφέρουν περισσότερο στο διαχωρισμό της κάθε ομάδας.

Κάθε νέο αντικείμενο ταξινομείται με βάση την απόστασή του (Ευκλείδεια, § 2.4.2 [18] ή Mahalanobis [16]) από το κεντροειδές (centroid) δηλαδή τη μέση τιμή των σκορ των LDF κάθε ομάδας. Οι ομάδες αυτές ήδη προβάλλονται σε έναν υποχώρο που καθορίζεται από τις LDF.

2.1.3. Αξιολόγηση - Επικύρωση

Έχοντας βρει τους κανόνες κατάταξης, υπάρχουν διάφορες μέθοδοι [2] για τον έλεγχο της αποτελεσματικότητας των κανόνων αυτών (επικύρωσης του μοντέλου):

- ✓ με τη χρήση της ίδιας της ομάδας εκπαίδευσης,
- ✓ με τη χρήση μιας ομάδας ανεξάρτητων δειγμάτων που δεν συμπεριλαμβάνονται στην αρχική ομάδα εκπαίδευσης,
- ✓ με την "εξαιρουμένου ενός" μέθοδο (leave-one-out method), όπου κάθε φορά χρησιμοποιούνται όλα εκτός από ένα δείγμα ώστε να βρεθούν οι κανόνες ταξινόμησης και το δείγμα αυτό χρησιμοποιείται για τον έλεγχο των κανόνων (βλ. επίσης § 3.1.2, 4.4.2). Αν τα αποτελέσματα της επικύρωσης αυτής είναι ικανοποιητικά, τότε μπορεί να χρησιμοποιηθεί το αρχικό μοντέλο για την ταξινόμηση νέων ομάδων [11].

Η πρώτη μέθοδος (με τη χρήση της ίδιας ομάδας εκπαίδευσης), παρουσιάζει όπως είναι φυσικό, προβλήματα γενίκευσης, εφόσον τα μοντέλα αξιολογούνται από δείγματα που χρησιμοποιήθηκαν στην κατασκευή τους (βλ. για το ίδιο πρόβλημα στα Δέντρα Ταξινόμησης αλλά και στα Νευρωνικά δίκτυα § 3.1.2, 4.3.1).

Επιπλέον, στατιστικές δοκιμές όπως ο έλεγχος των παραμέτρων **Press's Q Statistic** και Wilks' lambda/model που αναφέρθηκε παραπάνω (§ 2.1.1) μπορούν να χρησιμοποιηθούν για την αξιολόγηση των μοντέλων [11, 19]. Ο τύπος που περιγράφει το πρώτο είναι:

$$Q = \frac{(N - ng)^2}{N - (g - 1)} \qquad (2.5)$$

όπου: N ο συνολικός αριθμός των δειγμάτων, n ο αριθμός των δειγμάτων που ταξινομήθηκε σωστά, και g ο αριθμός των ομάδων. Το στατιστικό Q ακολουθεί κατανομή χ^2 (chi-square distribution, με βαθμούς ελευθερίας df = 1 και μηδενική υπόθεση H_0: το μοντέλο δεν έχει ικανότητα πρόβλεψης μεγαλύτερη από την τυχαία ταξινόμηση/the model hit ratio is no better than chance) [9, 11]. Στο παράρτημα της διατριβής αυτής (ΚΕΦ. 2, Θ) αναφέρονται και άλλες στατιστικές δοκιμές για τον έλεγχο της **τυχαίας συσχέτισης** ("chance correlation").

Έλεγχοι για την καταλληλότητα των μοντέλων που σχετίζονται με το δεύτερο (Wilks' lambda/model) είναι επίσης:

1. οι ιδιοτιμές λ (eigenvalues, § 2.2),
2. ο Canonical correlation coefficient R_C.

Οι **ιδιοτιμές των LDF** ή χαρακτηριστικές ρίζες (characteristic roots) αφορούν την κρισιμότητα των διαχωριστικών συναρτήσεων LDF. Υπάρχει μία ιδιοτιμή για κάθε LDF [16]. Όσο μεγαλύτερη είναι η ιδιοτιμή, τόσο καλύτερη είναι η ερμηνευτική ικανότητα των διαχωριστικών συναρτήσεων [9]. Οι ιδιοτιμές είναι χρήσιμες ως δείκτες μέτρησης της διασποράς (διακύμανσης) των κεντροειδών στον αντίστοιχο πολυμεταβλητό χώρο [10]. Για πρόβλημα δυο ομάδων, η γραμμική διαχωριστική συνάρτηση είναι μόνο μία, η οποία και ερμηνεύει το σύνολο της διακύμανσης [11, 16]. Αν υπάρχουν παραπάνω LDF, η πρώτη θα είναι η πιο σημαντική με τη μεγαλύτερη ιδιοτιμή, η δεύτερη η αμέσως σημαντικότερη κ.ο.κ. Οι ιδιοτιμές έχουν ιδιαίτερη

σημασία γιατί αντανακλούν το ποσοστό της διακύμανσης που ερμηνεύεται στην εξαρτημένη μεταβλητή. Έτσι, η σχέση των ιδιοτιμών δείχνει τη σχετική διαχωριστική ικανότητα των LDF. Αν για παράδειγμα, δυο ιδιοτιμές έχουν μια αναλογία ίση με 1,4, αυτό σημαίνει ότι η πρώτη LDF ερμηνεύει 40 % παραπάνω από την between-group διακύμανση σε σχέση με τη δεύτερη [16].

Ο **Canonical correlation coefficiet R_c** (ή eta η), συνδέεται με την ιδιοτιμή λ με βάση τη σχέση:

$$R_c = \sqrt{\frac{\lambda}{1+\lambda}} \qquad (2.6)$$

και αποδίδει τη σχέση ανάμεσα στα σκορ και τις ομάδες [10, 11, 16]. Όταν R_c=0, δεν υπάρχει καμιά σχέση ανάμεσα στην LDF και τις ομάδες. Όταν R_c=1, όλη η μεταβλητότητα των σκορ, μπορεί να περιγραφεί από την LDF. Για δυο ομάδες, ο συντελεστής R_c απεικονίζεται με την Pearsonian συσχέτιση (§ 2.2, σχέση 2.8) ανάμεσα την LDF και την ανεξάρτητη μεταβλητή [16].

Η παράμετρος Wilk's lambda/model (Λ) που αναφέρθηκε παραπάνω (§ 2.1.1) δίνεται τώρα από τη σχέση:

$$\Lambda = (1- R_c^2) = \frac{1}{1+\lambda} \qquad (2.7)$$

και ακολουθεί κατανομή χ^2 (chi-square distribution, με βαθμούς ελευθερίας df = p-1, όπου p ο αριθμός των μεταβλητών και μηδενική υπόθεση H_0 που αφορά των ισότητα των σκορ των ομάδων) [9]. Η τιμή 0 δείχνει πως υπάρχει σημαντική διαφορά ανάμεσα στα κεντροειδή των διαφορετικών ομάδων [11].

Μειονέκτημα για την DA είναι ότι βασίζεται στην εκ των προτέρων γνώση για κανονική κατανομή των δεδομένων [6, 9, 11, 16, 20 - 22]. Ωστόσο στην αναλυτική χημεία, δεν ευσταθεί πάντα ο ισχυρισμός μιας πολυπαραμετρικής κανονικής κατανομής. Παρόλα αυτά, αποκλίσεις από την κανονικότητα, εφόσον αυτές προκαλούνται από ασυμμετρία στην κατανομή και όχι από έκτροπες τιμές (outliers), δεν φαίνεται να είναι "μοιραίες" [9]. Το τελευταίο γίνεται αποδεκτό, όταν υπάρχουν τουλάχιστον 20 δείγματα στη μικρότερη ομάδα και οι ανεξάρτητες μεταβλητές είναι λιγότερες από έξι (6) [16].

Στις περιπτώσεις μη κανονικής κατανομής, μπορούν επίσης να γίνουν μετασχηματισμοί (αλλαγές κλίμακας § 4.4.4) ώστε αυτή να επιτευχθεί.

Μια δεύτερη παραδοχή για την DA, είναι ότι προϋποθέτει ότι οι διακυμάνσεις/συνδιακυμάνσεις των μεταβλητών ανάμεσα στις ομάδες είναι ομοιογενείς [9, 11, 16, 17, 20 - 23]. Άνισες διακυμάνσεις/συνδιακυμάνσεις προκαλούν την ταξινόμηση των δειγμάτων στις ομάδες με τη μεγαλύτερη συνδιακύμανση [11, 21]. Συγκεκριμένα, η διακύμανση της κάθε ανεξάρτητης μεταβλητής πρέπει να είναι παρόμοια στις διάφορες ομάδες (ομοσκεδαστικότητα/homo-

scedasticity). Δηλαδή, οι ανεξάρτητες μεταβλητές μεταξύ τους μπορεί να έχουν (και θα έχουν διαφορετικές διακυμάνσεις), αλλά οι ομάδες που σχηματίζονται πρέπει να έχουν παρόμοιες διακυμάνσεις για κάθε μεταβλητή. Η DA είναι ευαίσθητη σε ανομοιογένεια διακυμάνσεων, όταν το μέγεθος του δείγματος είναι μικρό και οι ομάδες πολύ άνισες. Η απουσία ομοιότητας στις διακυμάνσεις μπορεί να ελεγχθεί με τις γραφικές παραστάσεις (scatterplots) των μεταβλητών.

Επιπλέον, απαιτείται ομοιότητα των συνδιακυμάνσεων ανάμεσα σε δυο ανεξάρτητες μεταβλητές στις διάφορες ομάδες. Έτσι, κάθε ομάδα πρέπει να έχει παρόμοιο πίνακα συσχετίσεων ή διακυμάνσεων (covariance/correlation matrix) [16, 20]. Ο έλεγχος στη βιβλιογραφία γίνεται συνήθως με **τη δοκιμή Box's M** (Box's M test). Εδώ, η μηδενική υπόθεση H_0 αφορά την ισότητα των διακυμάνεσων/συνδιακυμάνσεων των ομάδων [9, 16, 17]. Μικρές αποκλίσεις δεν κρίνονται σημαντικές [17]. Οι ομάδες εξάλλου θεωρούνται ότι περιέχουν τουλάχιστον προσεγγιστικά τον ίδιο αριθμό δειγμάτων. Διαφορετικά, υπάρχουν επιπτώσεις στον υπολογισμό των LDF και στην ταξινόμηση των δειγμάτων [19].

Όταν το μέγεθος του δείγματος είναι μεγάλο, ακόμα και μικρές διαφορές στις διακυμάνσεις/συνδιακυμάνσεις φαίνονται σημαντικές στη δοκιμή Box's M. Στην περίπτωση αυτή, ο ερευνητής πρέπει να ελέγχει τους φυσικούς λογάριθμους των οριζουσών των πινάκων συνδιακύμανσης (log determinants). Αν είναι παρόμοιοι, τότε το αποτέλεσμα του Box's M αγνοείται [16].

Επιπλέον, οι μεταβλητές για την DA πρέπει να είναι ανεξάρτητες (μη-συσχετιζόμενες) μεταξύ τους [9, 11, 21]. Διαφορετικά, οι συντελεστές των ανεξάρτητων μεταβλητών (βλ. 2.1.2, σχέση 2.4) είναι δύσκολο να ερμηνευτούν [21]. Επίσης, αν οι ανεξάρτητες μεταβλητές συσχετίζονται, οι standardized Canonical discriminant function coefficients (§ 2.1.1) δεν απεικονίζουν αξιόπιστα τη σχετική κρισιμότητα των μεταβλητών. Ο έλεγχος γίνεται με το συντελεστή συσχέτισης (§ 2.2, σχέση 2.8) ή με τη χρήση του **συνδυασμένου πίνακα διακυμάνσεων** ("pooled within-groups correlation matrix").

Εδώ θα πρέπει να αναφερθεί ότι η DA θεωρείται γενικά ανθεκτική στην παραβίαση όλων των παραπάνω παραδοχών. Όσο μεγαλύτερο είναι δε το δείγμα, τόσο πιο ανθεκτική είναι η ανάλυση [20]. Γενικά επίσης στη βιβλιογραφία, δεν ερευνώνται οι παραδοχές που απαιτούνται από την εφαρμογή της DA, ενώ τα αποτελέσματα είναι ικανοποιητικά [22, 24].

Η DA επικεντρώνεται στην εύρεση των καλύτερων γραμμικών ορίων μεταξύ των ομάδων· ο στόχος είναι η διαφοροποίηση των ομάδων, αλλά το αποτέλεσμα δεν είναι πάντα το καλύτερο. Αυτό συμβαίνει γιατί η γραμμική διαφοροποίηση των ομάδων δεν είναι πάντα δυνατή, αλλά αντίθετα μερικά προβλήματα μπορούν να λυθούν μόνο με τη δημιουργία νέων μη γραμμικών ορίων [25]. Επιπλέον, στις μεθόδους αυτές (discrimination-oriented), είναι

απαραίτητη η ταξινόμηση των δειγμάτων σε μια από τις δοθείσες τάξεις [25, 26]. Έτσι όταν για παράδειγμα θέλουμε να ταξινομήσουμε κάποιες ποικιλίες δειγμάτων λαδιού ή κρασιού ανάλογα με την προέλευσή τους, όπου απαιτείται ταξινόμηση των δειγμάτων σε δυο ή τρεις γνωστές κατηγορίες (από Καλαμάτα ή Κρήτη), κάποιος θα έπρεπε οπωσδήποτε να κατατάξει σε αυτές και ένα δείγμα που προέρχεται πχ από την Αττική. Επίσης, τα περισσότερα προβλήματα ταξινόμησης δεν είναι τόσο απλά: έτσι συχνά απαιτείται η γνώση του βαθμού της πιθανότητας που κάποιο δείγμα ανήκει σε μια συγκεκριμένη ομάδα. Αυτό συμβαίνει για παράδειγμα, στα κλινικά προβλήματα ταξινόμησης [25].

Παραλλαγή της LDA είναι η **Δευτεροβάθμια Διαχωριστική Ανάλυση** (Quadratic Discriminant analysis, QDA), όπου η συνάρτηση κατάταξης, είναι δευτεροβάθμια και όχι γραμμική.

2.2. ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ (PRINCIPAL COMPONENTS ANALYSIS)

Ένα πρόβλημα με τα πολυπαραμετρικά δεδομένα είναι ότι λόγω του τεράστιου όγκου τους, είναι πολύ δύσκολο να αναγνωρίσουμε σε αυτά πρότυπα και συσχετίσεις [8]. Για παράδειγμα, ένα φάσμα χαρακτηρίζεται από μερικές εκατοντάδες μετρήσεις απορρόφησης, οπότε ο πρωταρχικός σκοπός είναι η μείωση των δεδομένων. Αυτός είναι και ο κύριος στόχος της **Ανάλυσης Κυρίων Συνιστωσών** (Principal Components Analysis, PCA). Η PCA είναι μια μέθοδος η οποία έχει σκοπό να δημιουργήσει γραμμικούς συνδυασμούς των αρχικών μεταβλητών έτσι ώστε οι γραμμικοί αυτοί συνδυασμοί να είναι ανεξάρτητοι μεταξύ τους αλλά να περιέχουν όσο γίνεται μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών [10]. Η PCA δεν είναι χρήσιμη όταν οι μεταβλητές είναι μη συσχετιζόμενες [7, 8, 10]. Έτσι, αν στην ανάλυση χρησιμοποιήσουμε ανεξάρτητες μεταβλητές, αυτές θα ταυτιστούν με κάποια συνιστώσα και επομένως ακυρώνεται η συνολική διαδικασία [10]. Χρησιμοποιείται σε πολύπαραμετρικά δεδομένα για τρεις βασικούς λόγους [2]:

- ✓ Αποκαλύπτει εκείνες τις μεταβλητές ή κάποιο συνδυασμό των μεταβλητών που "περιγράφουν" κάποια πρωταρχική και εσωτερική δομή των δεδομένων και που μπορεί ίσως να απεικονίζεται σε χημικούς ή φυσικοχημικούς όρους. Η μέθοδος μας επιτρέπει να αναγνωρίσουμε τις συνιστώσες παρατηρώντας ποιες από τις αρχικές μεταβλητές έχουν μεγάλη επίδραση σε αυτές. Αυτό είναι πολύ χρήσιμο σε κάποιες επιστήμες καθώς μας επιτρέπουν να ποσοτικοποιήσουμε μη μετρήσιμες ποσότητες, όπως η ευφυΐα, η ικανότητα ενός μπασκετμπολίστα, η εμπορευσιμότητα ενός προϊόντος κλπ αφηρημένες

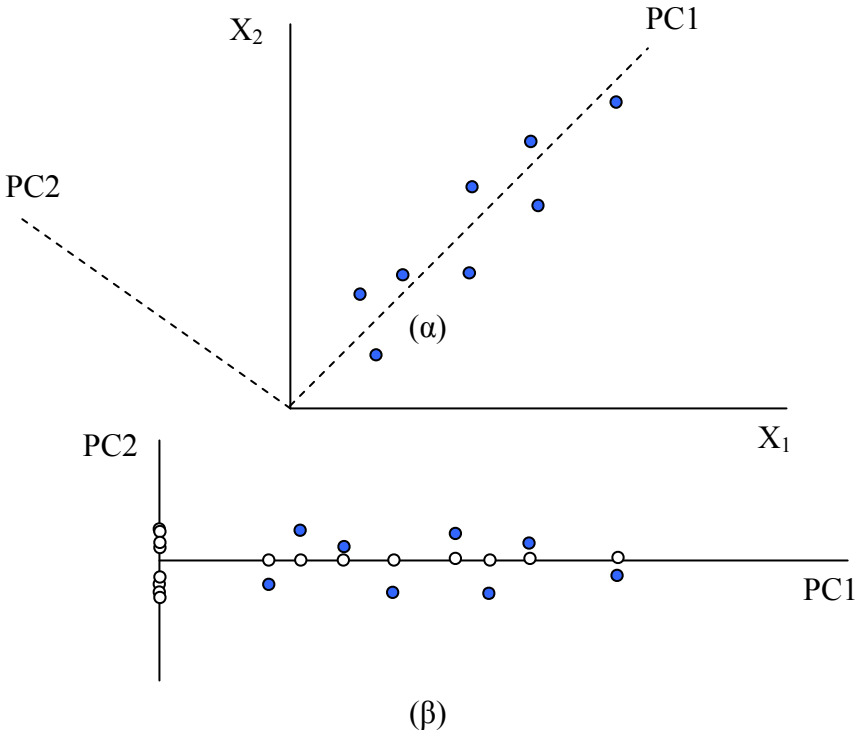
έννοιες. Επιπλέον, μπορούμε να εξετάσουμε τις συσχετίσεις ανάμεσα στις μεταβλητές και να διαπιστώσουμε πόσο οι μεταβλητές συσχετίζονται ή όχι.

- ✓ Περιλαμβάνει μετατροπή και ίσως περιστροφή των αρχικών n-αξόνων που ο καθένας αναπαριστούσε μια μεταβλητή, σε νέους άξονες. Η μετατροπή αυτή γίνεται έτσι, ώστε οι νέοι άξονες που προκύπτουν να είναι "ορθογώνιοι", δηλαδή οι νέες μεταβλητές να είναι μη συσχετιζόμενες. Οι νέες p-μεταβλητές είναι πολύ λιγότερες των n αρχικών.
- ✓ Μειώνει τον αρχικό αριθμό των μεταβλητών, χρησιμοποιώντας κάποιο συνδυασμό αυτών. Σε μερικές εφαρμογές αυτό είναι ζωτικής σημασίας. Για παράδειγμα σε μια τεράστια βάση δεδομένων αντί να χρησιμοποιήσουμε όλες τις μεταβλητές μπορούμε να χρησιμοποιήσουμε μόνο κάποιον αριθμό κυρίων συνιστωσών. Σίγουρα χάνουμε κάποιο μέρος της πληροφορίας αλλά το κέρδος σε χώρο αλλά και ταχύτητα επεξεργασίας μπορεί να είναι τεράστιο. Από την άλλη πλευρά πολλές φορές συμβαίνει να έχουμε λίγες παρατηρήσεις αλλά πολλές μεταβλητές.

Αρχικά ενδιαφερόμαστε για γραμμικούς συνδυασμούς των αρχικών μεταβλητών που απεικονίζουν τη μέγιστη διακύμανση και τους κανονικοποιημένους συντελεστές, ώστε να διατηρείται η αρχική κλίμακα αυτών. Αυτός ο πρώτος γραμμικός συνδυασμός (άξονας) είναι γνωστός ως **πρώτο βασικό συστατικό ή κυρία συνιστώσα** (first principal component). Μετά τον καθορισμό αυτού, αναζητείται ο δεύτερος που είναι επίσης κανονικοποιημένος γραμμικός συνδυασμός των αρχικών n-μεταβλητών, απεικονίζει τη μέγιστη από την απομένουσα διακύμανση και είναι **ορθογώνιος** (ανεξάρτητος) ως προς τον πρώτο. Η πορεία συνεχίζεται ώσπου να υπολογιστούν όλες οι βασικές συνιστώσες. Οι νέες μεταβλητές (Principal Components ή PCs) προκύπτουν πολλαπλασιάζοντας τις αρχικές μεταβλητές με τα φορτίσεις (loadings). Επειδή δε είναι ιεραρχικά οργανωμένες (αρχίζοντας από τη συνιστώσα που περιγράφει τη μέγιστη διακύμανση), η τελική απεικόνιση των δεδομένων παρέχει περισσότερες πληροφορίες από την αρχική με τον ίδιο αριθμό μεταβλητών [27].

Η PCA σε αντιδιαστολή με την DA (§ 2.1), αποσκοπεί στη μεγιστοποίηση της διακύμανσης μεταξύ των αντικειμένων και όχι στη μεγιστοποίηση του λόγου των μεταξύ των τάξεων (between) και εντός των τάξεων (within class) διακυμάνσεων (σχ. 1.1) [28].

Αρχικά n = p, αλλά τελικά επιλέγονται μερικές από τις κύριες συνιστώσες (PCs), συνήθως οι δυο ή τρεις πρώτες προς παραπέρα επεξεργασία και ερμηνεία. Στην πράξη μπορούμε να θεωρήσουμε την PCA ως μια περιστροφή των αξόνων έτσι, ώστε η PC1 να απεικονίζει την κατεύθυνση της μέγιστης διακύμανσης, η PC2 την επόμενη κ.ο.κ. Συνήθως όμως μόνο οι PC1 και PC2 απεικονίζουν σε μόνο δυο διαστάσεις τα αρχικά δεδομένα, ώστε να επιτυγχάνεται τελικά μείωση των αρχικών n-διαστάσεων [8] (σχ. 2.1).



Σχήμα 2.1: (α) Διάγραμμα που εξηγεί τις δυο βασικές συνιστώσες PC1 και PC2 για τις δυο μεταβλητές X1 και X2. (β) Σημεία στους άξονες: τα γαλάζια δηλώνουν τα αρχικά σημεία και τα λευκά τις προβολές τους στους άξονες.

Το πρώτο βήμα για την PCA είναι η εύρεση ενός αρχικού συμμετρικού πίνακα δεδομένων: **διακυμάνσεων/συνδιακυμάνσεων** (variance/covariance ή covariance matrix) ή **συσχετίσεων** (correlation matrix) των μεταβλητών [1, 25]. Οι πίνακες αυτοί περιέχουν αντίστοιχα τις διακυμάνσεις/συνδιακυμάνσεις (§ 1.2.1, σχέση 1.1), ή τον γνωστό συντελεστή συσχέτισης Pearson (correlation coefficient) μεταξύ δύο μεταβλητών x, y:

$$\text{συντελεστής Pearson} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \qquad (2.8)$$

Η επιλογή του πίνακα που τελικά θα χρησιμοποιήσουμε, εξαρτάται από τον σκοπό της ανάλυσης, αλλά και τα αρχικά μας δεδομένα (ΚΕΦ. 1, Θ).

Το δεύτερο βήμα για την PCA, είναι η εξαγωγή των **ιδιοανυσμάτων** (eigenvectors) που ειδικά στην περίπτωση αυτή αποτελούν τις **συνιστώσες** (components). Τα ιδιοανύσματα ενός τετραγωνικού πίνακα δίνουν πληροφορίες για τα δεδομένα αυτού. Απεικονίζουν τη συσχέτιση των δεδομένων του πίνακα και μάλιστα εμπεριέχουν την πληροφορία αυτή κατά σειρά αύξουσας **ιδιοτιμής** (eigenvalue): το ιδιοάνυσμα με τη μεγαλύτερη ιδιοτιμή παρέχει περισσότερη πληροφορία κ.ο.κ.

Περισσότερες λεπτομέρειες που αφορούν τα ιδιοανύσματα (σχήματα και θεωρία που σχετίζεται με αυτά, υπάρχουν στο ηλεκτρονικό παράρτημα της διατριβής (ΚΕΦ. 1, Θ).

Σε κάθε βασική συνιστώσα υπάρχει η ιδιοτιμή, η οποία και δίνει το ποσοστό της συνολικής διακύμανσης που ερμηνεύεται από την κάθε βασική συνιστώσα, καθώς και τον αριθμό των αρχικών μεταβλητών που "περιέχει" κάθε κύρια συνιστώσα ή παράγοντας.

Τα δυο βασικά εργαλεία απεικόνισης για την PCA, είναι τα **διαγράμματα φορτίσεων και συντεταγμένων** (loading και score-plots). Το διάγραμμα φορτίσεων χρησιμοποιεί ως άξονες οποιεσδήποτε δυο (συνήθως τις πρώτες) από τις βασικές συνιστώσες και κάθε σημείο του απεικονίζει τις αρχικές μεταβλητές. Έτσι αναδεικνύεται η συσχέτιση παλαιών και νέων μεταβλητών και οι φορτίσεις των πρώτων που περιέχονται στις δεύτερες. Το διάγραμμα συντεταγμένων χρησιμοποιεί επίσης ως άξονες οποιεσδήποτε δυο (συνήθως τις πρώτες) από τις βασικές συνιστώσες και κάθε σημείο του (που αναπαριστά τώρα το δείγμα), απεικονίζεται κοντά στα περισσότερο με αυτό συσχετιζόμενα. Έτσι με τη χρήση των PC1 & PC2 (που απεικονίζουν το μέγιστο της συνολικής διακύμανσης) διαφαίνονται συνήθως οι αρχικές ομάδες δειγματοληψίας, αφού το συνολικό ποσοστό της ερμηνευόμενης διακύμανσης είναι προφανώς αρκετό για την απεικόνιση της αρχικής δομής.

Το πιο σημαντικό όμως βήμα της ανάλυσης το οποίο δυστυχώς δεν έχει εύκολη και κοινώς αποδεκτή απάντηση είναι η απόφαση για το πόσες συνιστώσες θα αξιοποιηθούν τελικά. Επιλέγοντας προφανώς λιγότερες κύριες συνιστώσες από όσες αρχικές μεταβλητές, χάνουμε πληροφορία. Αυτό είναι το κόστος, για το κέρδος μας να μειώσουμε τις διαστάσεις του προβλήματος. Ωστόσο, συνήθως ενδιαφερόμαστε για το μικρότερο δυνατό αριθμό συνιστωσών [10]. Στη βιβλιογραφία υπάρχουν πολλά κριτήρια για την επιλογή του βέλτιστου αριθμού των συνιστωσών, τα οποία αναφέρονται με λεπτομέρειες στο ηλεκτρονικό παράρτημα της διατριβής, ΚΕΦ. 1, Θ).

Αναλυτικά παραδείγματα κατανόησης της PCA, αναφέρονται επίσης στο ηλεκτρονικό παράρτημα της διατριβής, ΚΕΦ. 1, Θ).

2.3. ΑΝΑΛΥΣΗ ΠΑΡΑΓΟΝΤΩΝ (FACTOR ANALYSIS)

Η PCA είναι σήμερα η πιο ευρέως χρησιμοποιούμενη τεχνική για την εξαγωγή νέων μεταβλητών από ένα αρχικό πίνακα δεδομένων [1, 25]. Μια άλλη τεχνική σχετιζόμενη με την PCA είναι η **Ανάλυση Παραγόντων** (Factor Analysis, FA). Η FA είναι η ανάλυση που γίνεται με σκοπό τη μείωση των αρχικών n μεταβλητών σε συγκεκριμένο αριθμό p λιγότερων γραμμικών συνδυασμών ή παραγόντων (factors). Έτσι επιλέγεται $p < n$ με την προϋπόθεση ότι

τα νέα δεδομένα θα ερμηνεύουν καλύτερα την αρχική κατάσταση. Αν και η PCA είναι μια τεχνική μείωσης των αρχικών διαστάσεων (μεταβλητών) σε ένα μικρότερο αριθμό συνιστωσών, οι οποίες "ευθύνονται" για τη μέγιστη διακύμανση, δεν "υποδηλώνουν" την ύπαρξη ενός προκαλούντος (αιτιολογικού) μοντέλου. Αντίθετα, η FA προϋποθέτει ότι η συνδιακύμανση των μεταβλητών ερμηνεύεται με τη βοήθεια κάποιων παραγόντων που ασκούν αιτιολογική επίδραση στις μεταβλητές αυτές [29].

Το καθοριστικό σημείο διαχωρισμού ανάμεσα στις δυο τεχνικές, είναι ότι στην PCA θεωρούμε ότι όλη η διακύμανση μπορεί να χρησιμοποιηθεί (ερμηνευτεί) στην ανάλυση, ενώ στην FA χρησιμοποιούμε **μόνο τη διακύμανση που είναι κοινή** και μπορεί να ερμηνευτεί σε σχέση με την διακύμανση των άλλων αντικειμένων. Ωστόσο στην πράξη, οι δυο τεχνικές χρησιμοποιούνται αλληλένδετα και η εφαρμογή της FA στα πολυπαραμετρικά δεδομένα (μέσω των στατιστικών πακέτων) χωρίς περιστροφή των αξόνων δίνει τα ίδια αποτελέσματα με την PCA. Θεωρητικά, η PCA συνήθως χρησιμοποιείται ως μέθοδος μείωσης δεδομένων (μεταβλητών), ενώ η FA προτιμάται για την ερμηνεία και ανίχνευση δομών [17].

Έτσι, παρόλο που οι στατιστικολόγοι δεν διαχωρίζουν την PCA από την FA, για την αναλυτική χημεία ίσως και για άλλες επιστήμες [30], οι παράγοντες έχουν μια φυσική σημασία, ενώ οι PCs είναι απλά θεωρητικές [31]. Το πέρασμα από την PCA στην FA γίνεται συνήθως με περιστροφή ή μετασχηματισμό, ώστε οι παράγοντες να αποκτήσουν φυσική σημασία. Το βασικό κίνητρο για την περιστροφή είναι να επιτύχουμε μια απλούστερη και θεωρητικά σημαντικότερη ερμηνεία των παραγόντων που προκύπτουν [25]. Ο αριθμός των παραγόντων που θα επιλεγεί εδώ, εξαρτάται τόσο από το ποσοστό της διακύμανσης που ερμηνεύεται (όσο υψηλότερο, τόσο καλύτερα), όσο και από την κατανόηση του μοντέλου που προκύπτει [29]. Παραδείγματα δείχνουν ότι όταν υπάρχουν πάνω από δυο παράγοντες, η οποιαδήποτε ερμηνεία αυτών είναι αδύνατη και απαιτείται απαραίτητα περιστροφή.

Άλλος ένας λόγος για την περιστροφή των παραγόντων είναι ότι στην περίπτωση μη περιστροφής, με την αφαίρεση κάποιας μεταβλητής οι φορτίσεις μπορεί να αλλάξουν δραστικά, κάτι το οποίο δεν συμβαίνει συνήθως μετά από περιστροφή των αξόνων.

Οι κύριοι λόγοι για τους οποίους γίνεται περιστροφή αξόνων είναι για να επιτύχουμε:

- ✓ πολλές μεταβλητές να έχουν μικρές φορτίσεις (loadings) σε ένα συγκεκριμένο παράγοντα: με αυτόν τον τρόπο ο παράγοντας είναι χαρακτηριστικός για τις άλλες μεταβλητές,
- ✓ λίγες μεταβλητές να έχουν υψηλές φορτίσεις σε αρκετούς παράγοντες.

Μια εκτίμηση για το πόσο καλά μπορεί να δουλεύει ένα μοντέλο FA, παίρνουμε από τις **συμμετοχικότητες [32]** (communalities) των μεταβλητών ( ΚΕΦ. 1, Θ).

2.4. ΑΝΑΛΥΣΗ ΚΑΤΑ ΣΥΣΤΑΔΕΣ (CLUSTER ANALYSIS)

2.4.1. Αλγόριθμοι συσταδοποίησης (clustering algorithms)

Αν και η PCA, μπορεί να αποκαλύψει ομάδες αντικειμένων, δεν αποδεικνύεται πάντα επιτυχής για το σκοπό αυτό [8]. Αντίθετα, **η συσταδοποίηση** ή **ομαδοποίηση** (clustering) είναι μια μη επιβλεπόμενη (unsupervised) μέθοδος (§ 1.3) που "αποκαλύπτει" τις εσωτερικές διαφορές ή απλά υπογραμμίζει στοιχεία στη συμπεριφορά των δεδομένων χωρίς εκ των προτέρων κάποια παραδοχή γι' αυτά, έτσι ώστε να επιτευχθεί ομαδοποίηση των αντικειμένων με βάση τη "γειτονία" ή την ομοιότητα [17].

Η συσταδοποίηση επιχειρεί να οργανώσει μη χαρακτηρισμένες παρατηρήσεις (μεταβλητές, δείγματα, θέσεις) σε συστάδες (clusters), ώστε τελικά αυτά που ανήκουν στην ίδια ομάδα να είναι "πιο όμοια" από εκείνα που δεν ανήκουν [33]. Δηλαδή οι παρατηρήσεις μέσα σε μια ομάδα, πρέπει να είναι όσο το δυνατό πιο ομοιογενείς, ενώ οι παρατηρήσεις διαφορετικών ομάδων να διαφέρουν όσο το δυνατό περισσότερο [10]. Υπάρχουν διάφορες τεχνικές συσταδοποίησης και έτσι η θεωρία που αναπτύσσεται γύρω από αυτές δεν μπορεί να είναι ενιαία.

Στη **σαφή συσταδοποίηση** ("crisp clustering") για παράδειγμα, ένα δείγμα ανήκει μόνο σε μία ομάδα. Στην **ασαφή συσταδοποίηση** ("fuzzy clustering"), ένα αντικείμενο μπορεί και κατατάσσεται σε μια ομάδα, ανάλογα με το βαθμό που υποδεικνύει μια συγκεκριμένη συνάρτηση. Εδώ ελέγχεται διεξοδικά σε πιο βαθμό ένα αντικείμενο "ταιριάζει" στην επιλεγόμενη ομάδα και όχι σε κάποιες άλλες. Στην κατηγορία αυτή συμπεριλαμβάνονται μέθοδοι που όχι μόνο "ψάχνουν" για ομοιότητες μεταξύ των αντικειμένων, αλλά παρέχουν πληροφορίες σχετικά με τη σχέση κάθε αντικειμένου με κάθε ομάδα [17, 33, 34].

Δύο είναι οι πιο δημοφιλείς αλγόριθμοι συσταδοποίησης: **ιεραρχικοί** (hierarchical) και **μεριστικοί** (partitional).

Η πρώτη κατηγορία διακρίνεται σε [33, 34, 35]:

1. **Αθροιστικούς** (agglomerative ή bottom-up), όπου κάθε δείγμα είναι μια ξεχωριστή ομάδα και προοδευτικά συγχωνεύεται σε άλλες ομάδες με βάση μια μετρική απόσταση. Η ομαδοποίηση σταματά όταν όλα τα δείγματα ενσωματωθούν σε μία ομάδα. Οι μέθοδοι αυτές ακολουθούν μια εκ των κάτω προς τα άνω (bottom-up) συγχώνευση. Στην κατηγορία αυτή, ανήκει η **ανάλυση κατά συστάδες** (Cluster Analysis, CA) που θα δούμε σε άλλη παράγραφο (§ 2.4.2).
2. **Διαιρετικούς** (divisive ή top-down), όπου ακολουθείται η αντίστροφη τεχνική. Οι αλγόριθμοι αυτοί αρχίζουν με μια ομάδα που περιέχει όλα τα δείγματα και προοδευτικά χωρίζεται σε άλλες μικρότερες, ώσπου κάθε δείγμα να αποτελεί μία ομάδα, ή ότι είναι επιθυμητό. Οι μέθοδοι αυτές ακολουθούν μια top-down κατάτμηση.

Οι ιεραρχικοί αλγόριθμοι είναι πιο αποτελεσματικοί από τους μεριστικούς στη διαχείριση "θορύβου" (άσχετης πληροφορίας) και έκτροπων τιμών. Ωστόσο εδώ, όταν γίνει ο διαχωρισμός μιας ομάδας, δεν μπορεί να αναθεωρηθεί ή να ανακληθεί [10, 33].

Οι μεριστικοί αλγόριθμοι προϋποθέτουν κάποιες παραδοχές για τα δεδομένα, όπως για παράδειγμα την αναζήτηση σφαιρικών ομάδων [34]. Οι μεριστικοί αλγόριθμοι διαφέρουν από τους ιεραρχικούς, στο ότι κατά τη διάρκεια της εξέλιξής τους, τα δείγματα ενδέχεται να αλλάξουν ομάδα (επανατοποθέτηση/relocation). Παραδέχονται δηλαδή τη δυνατότητα σφάλματος κατά την κατάτμηση των ομάδων. Δεν αντιπροσωπεύουν πάντα ιεραρχικές σχέσεις μεταξύ των δεδομένων και ψάχνουν για την καλύτερη λύση που ελαχιστοποιεί την προκαθορισμένη απόσταση. Πραγματοποιούν περισσότερα από ένα πέρασμα (iteration) από τα δεδομένα και έτσι αποζημιώνουν με το αποτέλεσμα την τυχόν κακή αρχική εκτίμηση [33].

Τυπικός αντιπρόσωπος των αλγορίθμων αυτών, είναι η μέθοδος της **Συσταδοποίησης Κ-μέσων σημείων** (K-means clustering) η οποία με τη βοήθεια ενός αλγορίθμου θα δημιουργήσει ακριβώς Κ συστάδες. Η μέθοδος αυτή βασίζεται στην ταξινόμηση κάθε αντικειμένου με βάση τα κοντινότερα αντικείμενα. Υπολογίζεται η απόσταση του διανύσματος (που περιέχει όλες τις μεταβλητές), ανάμεσα στο κάθε δείγμα και το κεντροειδές (βλ. παρακάτω) καθεμιάς από τις Κ συστάδες. Έπειτα η μικρότερη απόσταση καθορίζει την ομάδα που θα ανήκει το "νέο" δείγμα [17]. Ο αριθμός των συστάδων είναι συνήθως προκαθορισμένος, αλλά μπορεί να προκύπτει μέσω κάποιας συνάρτησης σφάλματος που πρέπει να ελαχιστοποιηθεί:

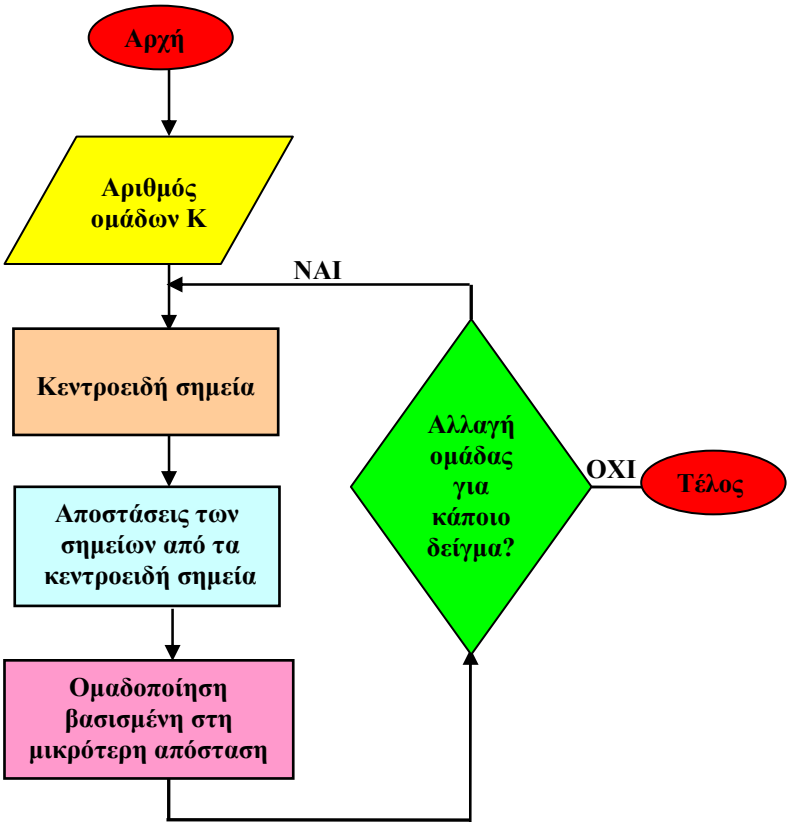
$$E = \sum_{k=1}^{C} \sum_{x \in Q_k} \|\mathbf{x} - \mathbf{C_k}\|^2 \qquad (2.9)$$

όπου, C είναι ο αριθμός των ομάδων, C_k είναι το κεντροειδές της ομάδας k και x είναι ένα δείγμα εκ του συνόλου Q_k που ανήκει στην ομάδα C.

Οποιαδήποτε τυχαία σημεία μπορούν να επιλεγούν αρχικά ως τα Κ κεντροειδή. Τελικά, ο αλγόριθμος K-means θα κάνει τόσες επαναλήψεις τεσσάρων βημάτων, όσες χρειάζεται ώσπου να εκπληρωθεί το κριτήριο προς τη σύγκλιση, όπως φαίνονται στο σχήμα 2.2 [36].

- Βήμα 1^{ο}: Καθορισμός του αριθμού Κ των ομάδων και των αρχικών σημείων ή κεντροειδών Κ (εύρεση Κ συντεταγμένων) τυχαία ή συστηματικά.
- Βήμα 2^{ο}: Υπολογισμός των αποστάσεων των υπολοίπων σημείων από τα κεντροειδή Κ.
- Βήμα 3^{ο}: Τα υπόλοιπα σημεία "αποδίδονται" στις αρχικές ομάδες με βάση την ελάχιστη απόστασή τους από τα κεντροειδή. Μετά την ταξινόμηση, υπολογίζονται εκ νέου τα κεντροειδή σημεία των ομάδων.

- Βήμα 4^{ο}: Επανάληψη του βήματος 3, μέχρι σταθεροποίηση του συστήματος: δηλαδή η ταξινόμηση των δειγμάτων να μην αλλάζει και η συνολική απόσταση να παραμένει σταθερή (ή να μειώνεται) [36].



Σχήμα 2.2: Ο K-means αλγόριθμος [36].

Περισσότερες λεπτομέρειες αναφέρονται στο ηλεκτρονικό παράρτημα της διατριβής, ( ΚΕΦ. 1, Θ).

2.4.2. Ιεραρχική Ανάλυση κατά συστάδες (Hierarchical Cluster Analysis)

Τυπικός αντιπρόσωπος των ιεραρχικών αθροιστικών μεθόδων συσταδοποίησης είναι η **ανάλυση κατά συστάδες** (Cluster Analysis, CA).

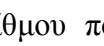
Όπως και στην PCA, στην τεχνική των ιεραρχικών CA μεθόδων, οι ομάδες δεν είναι εκ των προτέρων γνωστές και **δεν γίνεται καμιά παραδοχή για την κατανομή των μεταβλητών.** Η CA ελέγχει τις αποστάσεις μεταξύ δυο αντικειμένων στο n-διάστατο χώρο, λαμβάνοντας υπόψη όλες τις συντεταγμένες (x_1, x_2,……x_n) και (y_1, y_2,……y_n) αυτών. Συνήθως χρησιμοποιείται η Ευκλείδεια απόσταση που δίνεται από τη σχέση 2.11 ή το τετράγωνο αυτής [35]:

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots\ldots\ldots\ldots\ldots + (x_n - y_n)^2} \qquad (2.10)$$

Στη σχέση αυτή, οι συντεταγμένες (x_1, x_2,……x_n) και (y_1, y_2,……y_n) είναι για παράδειγμα, οι μετρηθείσες παράμετροι (τιμές των μεταβλητών) κάθε σημείου. Έτσι αν υπάρχουν 3 μεταβλητές, σχηματίζεται ένας τρισδιάστατος χώρος και η Ευκλείδεια απόσταση σε αυτήν την περίπτωση θα είναι το ίδιο σαν να μετρείται η απόσταση μεταξύ των σημείων με χάρακα [17]. Έτσι, σχηματίζονται διαδοχικές συστάδες, αρχίζοντας με τα πιο όμοια ζεύγη αντικειμένων και σχηματίζοντας υψηλότερη διαβάθμιση, βήμα-βήμα. Αποτέλεσμα είναι η δημιουργία μιας δομής σε σχήμα δέντρου, το δενδρόγραμμα.

Το δενδρόγραμμα δεν περιγράφει μια μοναδική ομαδοποίηση [34]. Αντίθετα διαφορετικές ομαδοποιήσεις επιτυγχάνονται "κόβοντας" το δέντρο σε διάφορα ύψη ( ΚΕΦ. 1, Θ). Γι' αυτό έχουν καθοριστεί κάποια κριτήρια (βλ. παρακάτω).

Συνήθως, οι θέσεις δειγματοληψίας θεωρούνται ως τα αντικείμενα προς ομαδοποίηση, το καθένα από τα οποία καθορίζεται από μια ομάδα μεταβλητών (παραμέτρων). Ωστόσο, είναι επίσης πιθανό, να αναζητούνται σχέσεις μεταξύ των μεταβλητών, οι οποίες γίνονται τώρα τα αντικείμενα προς ταξινόμηση. Τα στάδια που ακολουθούνται είναι τα εξής [37]:

1. Προεπεξεργασία των αρχικών δεδομένων (συνήθως τυποποίηση, βλ. § 4.4.4), ώστε να αποφθεχθεί η επίδραση του διαφορετικού εύρους (διακύμανσης) των χημικών διαστάσεων (μετρήσεων)).
2. Καθορισμός της απόστασης ανάμεσα στα αντικείμενα προς ομαδοποίηση, με την εφαρμογή κάποιου μέτρου σύγκρισης όπως η Ευκλείδεια απόσταση (ή το τετράγωνο αυτής), ή ο συντελεστής συσχέτισης.
3. Υπολογισμού αποστάσεων μεταξύ των αντικειμένων με τη χρήση κατάλληλου αλγορίθμου που καθορίζει τον τρόπο σύνδεσης των ομάδων ( ΚΕΦ. 1, Θ). Αναφέρονται χαρακτηριστικά: η απλή σύνδεση (Single linkage), η Πλήρης σύνδεση (Complete linkage) ή η μέθοδος Ward (Ward's method) [38].
4. Αναπαράσταση των αποτελεσμάτων σαν δενδρόγραμμα.
5. Καθορισμός της σημαντικότητας της ομαδοποίησης με βάση για παράδειγμα τα κριτήρια 1/3 Dmax ή 2/3 Dmax (Sneath index, βλ. § 5.3.3).
6. Ερμηνεία των αποτελεσμάτων για θέσεις και μεταβλητές.

Τα μειονεκτήματα της CA είναι ότι απαιτεί πολύ χρόνο και υπολογιστική ισχύ για μεγάλες βάσεις δεδομένων. Επιπλέον, οι ομάδες που σχηματίζονται σε αρχικά βήματα δεν μπορούν να επανεξεταστούν. Γενικά δημιουργούνται αρκετές ομάδες με πολλές παρατηρήσεις, ενώ κάποιες παρατηρήσεις μπορεί να συνιστούν από μόνες τους μία ομάδα [10].

Ο σημαντικότερος δείκτης που χρησιμοποιείται για τη αξιολόγηση των μεθόδων συσταδοποίησης είναι ο Davies-Bouldin (DB) δείκτης ( ΚΕΦ. 1, Θ).

ΚΕΦ. 3 ΔΕΝΤΡΑ ΤΑΞΙΝΟΜΗΣΗΣ (CLASSIFICATION TREES)

3.1. ΕΙΣΑΓΩΓΗ

3.1.1. Γενικά χαρακτηριστικά

Η μη-παραμετρική τεχνική των **Δέντρων Ταξινόμησης** (Classification trees, CT) περιγράφεται σε ξεχωριστό κεφάλαιο, γιατί δεν ανήκει στις "παραδοσιακές" μεθόδους ταξινόμησης. Η μέθοδος αποσκοπεί στην ταξινόμηση αντικειμένων με τη βοήθεια μιας σειράς μεταβλητών οι οποίες μπορούν να ελέγχονται μία-μία κάθε φορά [17]. Ένα δέντρο ταξινόμησης είναι δηλαδή ένας κανόνας που θεσπίζεται για τη σωστή ταξινόμηση αντικειμένων σε τάξεις, με βάση τις τιμές (παραμέτρους) που τα προσδιορίζουν. Ο σκοπός είναι να προβλεφθούν ή να ερμηνευτούν κατηγορίες με τη χρήση των πιο χρήσιμων από μια σειρά ανεξάρτητων μεταβλητών [39]. Παραδοχές που αφορούν την κατανομή των δεδομένων ή το επίπεδο των τιμών σε αυτές είναι λιγότερο αυστηρές. Τα δεδομένα που περιγράφουν το πραγματικό αρχικό σύστημα, αναπαρίστανται από ένα πίνακα που χρησιμοποιείται **ως μια αρχική ομάδα εκπαίδευσης για την κατασκευή και εκπαίδευση του δέντρου ταξινόμησης** [39, 40].

Αναλυτικότερα, η τεχνική των Δέντρων Ταξινόμησης αποτελεί μια διαδικασία πολυσταδιακών αποφάσεων. Αντί να χρησιμοποιείται μια ομάδα μεταβλητών στη λήψη μιας απόφασης, αξιοποιούνται διαφορετικές υπο-ομάδες αυτών σε κάθε επίπεδο του δέντρου. Ένα δέντρο μπορεί να θεωρηθεί ως μια πολύπλοκη διαδικασία που διευκολύνει τη λήψη μιας απόφασης, υποβαθμίζοντας τη σε μια σειρά απλούστερων αποφάσεων σε κάθε κόμβο (node) του δέντρου. Τα δυαδικά (binary) δέντρα διαμοιράζουν με επιτυχία το χώρο σε δυο τμήματα. Στην περίπτωση αυτή, η διαίρεση (split) γίνεται με υπερ-επίπεδα παράλληλα στους άξονες συντεταγμένων. Ένα δέντρο δεν είναι μοναδικό για μια συγκεκριμένη διαίρεση. Μπορεί επίσης να υπάρχουν πολλά δέντρα ακόμα και με 100 % επιτυχία. Οι αποφάσεις σε κάθε κόμβο μπορεί να μην καθορίζονται από απλά μία και μόνο μεταβλητή σε κάποιο επίπεδο ή κατώφλι δράσης (threshold), αλλά και να περιλαμβάνουν γραμμικό ή μη συνδυασμό πολλών μεταβλητών [41].

Το δέντρο κατασκευάζεται με τη χρήση μιας αρχικής ομάδας εκπαίδευσης, όπου κάθε ομάδα (κάθε δείγμα σε αυτήν), έχει μια ταυτότητα ήδη γνωστή (επιβλεπόμενη αναγνώριση προτύπων) [39]. Κάθε διαχωρισμός αντιπροσωπεύεται από ένα κόμβο (node ή leaf). Οι τεχνικές των CT, κατασκευάζουν το δέντρο από κάτω προς τα πάνω, αρχίζοντας από τη ρίζα και καταλήγοντας στα φύλλα, επιτυγχάνοντας το διαχωρισμό του χώρου. Τα Δέντρα Ταξινόμησης περιλαμβάνουν γενικά τρία στάδια:

1. Επιλογή του κανόνα διαχωρισμού για κάθε κόμβο. Αυτό σημαίνει ότι καθορίζονται οι μεταβλητές και το κατώφλι, που μπορούν θα χρησιμοποιηθούν στο διαχωρισμό.

2. Καθορισμός των τερματικών κόμβων. Αυτό σημαίνει ότι σε κάθε κόμβο, πρέπει να αποφασίσουμε αν θα συνεχιστεί ο διαχωρισμός, ή αυτός θα καθοριστεί ως τερματικός με την "επικόλληση" κάποιας ετικέτας. Αν αυτός διαχωριστεί επιπλέον, μέχρι που κάθε κόμβος να περιέχει μία "καθαρόαιμη" τάξη, θα καταλήξουμε σε ένα μεγάλο δέντρο **που υπερ-προσαρμόζεται** στα δεδομένα εκπαίδευσης και έχει μικρή ακρίβεια σε νέα δείγματα. Είναι γνωστό το παράδειγμα του τζογαδόρου στις ιπποδρομίες, που κατασκεύασε ένα τεράστιο δέντρο με πολλούς κόμβους και περίμενε μάταια ότι θα πλούτιζε στον επόμενο αγώνα. Εναλλακτικά, οι λιγότεροι μη "καθαρόαιμοι" κόμβοι που περιέχουν δείγματα από μικτές ομάδες, οδηγούν σε μειωμένη ακρίβεια. Η διασταυρούμενη αξιολόγηση (βλ. παρακάτω), χρησιμοποιείται συνήθως για το επιλεκτικό **"κλάδεμα"** ("pruning") του δέντρου με αποδεκτό σφάλμα.
3. "Επικόλληση" ετικετών ή επισήμανση ("labeling") δηλαδή χαρακτηρισμός των τερματικών κόμβων [41].

Ένα απλό παράδειγμα κατανόησης των CT και εφαρμογές τους, αναφέρονται στο ηλεκτρονικό παράρτημα της διατριβής ( ΚΕΦ. 1, Θ).

3.1.2. Αξιολόγηση - Επικύρωση

Τα Δέντρα Ταξινόμησης έχουν χρησιμοποιηθεί σε ένα εύρος προβλημάτων μέχρι σήμερα. Στα πλεονεκτήματα τους συμπεριλαμβάνεται το γεγονός ότι μπορούν αποτελεσματικά να ταξινομήσουν νέα δείγματα αποδεικνύοντας καλή "γενίκευση" σε πλήθος προβλημάτων, με τη βοήθεια μιας σειράς διαφορετικών μεθόδων που περιγράφονται παρακάτω. Μειονεκτήματα αποτελούν:

1. Η δυσκολία σχεδίασης του βέλτιστου δέντρου, έτσι ώστε να αποφεύγονται τελικά, μεγάλα δέντρα με μικρή ακρίβεια πρόβλεψης. Το κλάδεμα του δέντρου [26, 40] έχει αποδειχθεί ο σημαντικότερος μηχανισμός για την παρεμπόδιση της υπερ-προσαρμογής, δηλαδή της φτωχής γενίκευσης σε νέα δείγματα (βλ. για το ίδιο πρόβλημα στα Νευρωνικά Δίκτυα, § 4.3.1). Ένα απλό δέντρο γίνεται ευκολότερα κατανοητό, και παρουσιάζει μεγαλύτερη ακρίβεια σε νέα δεδομένα (που δεν συμμετείχαν στην κατασκευή του μοντέλου). Ένα δέντρο που εμφανίζει φαινόμενα υπερ-προσαρμογής ταξινομεί τέλεια τα δεδομένα της ομάδας εκπαίδευσης, αλλά κάνει σημαντικά λάθη στην ομάδα ελέγχου για την οποία υπάρχει πραγματική ανάγκη πρόβλεψης [42]. Το κλάδεμα μπορεί να ενεργοποιηθεί κατά τη διάρκεια της κατασκευής του δέντρου (pre-pruning) ή μετά την ολοκλήρωσή του (post-pruning). Στην πραγματικότητα, ένας προκαθορισμένος ελάχιστος αριθμός δειγμάτων στα "κλαδιά" του δέντρου αποτελεί το

σύνηθες pre-pruning, ενώ κάποιο προκαθορισμένο επίπεδο ακρίβειας (confidence accuracy level) λειτουργεί ως post-pruning [41]. Ο Kim [43] εξάλλου, χρησιμοποιεί το κριτήριο ενός προκαθορισμένου αριθμού δειγμάτων παρόντων στον αρχικό κόμβο, για να ερευνηθεί η διαδικασία του split ως pre-pruning.

2. Πολλές τεχνικές δεν είναι προσαρμόσιμες. Αυτό σημαίνει ότι γενικά χρησιμοποιείται κάποια σταθερή ομάδα εκπαίδευσης, ενώ για την εισαγωγή νέων δεδομένων απαιτείται ανακατασκευή του δέντρου [41].
3. Η ακρίβεια των CT φαίνεται γενικά να επηρεάζεται αρκετά (περισσότερο από ότι τα Νευρωνικά Δίκτυα, § 4.4.1) από την παρουσία θορύβου στα δεδομένα [44].

Η αξιολόγηση του αποτελέσματος μπορεί να επιτευχθεί με τη χρήση μιας σειράς μεθόδων, όπως και στις κλασικές τεχνικές (βλ. § 2.1.3 για την DA), όπως για παράδειγμα:

- ✓ Τη διασταυρούμενη αξιολόγηση με χρήση ομάδας ελέγχου (test sample cross-validation), όπου με τη βοήθεια μιας νέας, (βλ. § 4.4.2) ομάδας δειγμάτων, αξιολογείται ο βαθμός επιτυχίας του "δέντρου" που έχει σχηματιστεί από την ομάδα εκπαίδευσης.
- ✓ Την v-πλάσια διασταυρούμενη αξιολόγηση (v-fold cross-validation, CV), όπου δημιουργούνται v υπο-ομάδες δειγμάτων, από τις οποίες κάθε φορά οι v-1 χρησιμοποιούνται ως ομάδα εκπαίδευσης και μία ως ομάδα ελέγχου.
- ✓ Τη σφαιρική διασταυρούμενη αξιολόγηση (global cross-validation, GCV), όπου η ανάλυση επαναλαμβάνεται εξ΄ολοκλήρου v φορές, ενώ κάθε φορά παραμένουν εκτός v δείγματα που ελέγχονται ως ομάδα ελέγχου όταν τα υπόλοιπα χρησιμοποιούνται ως ομάδα εκπαίδευσης.

Υπάρχουν πολλές μέθοδοι διαχωρισμού, που μπορούν να χρησιμοποιηθούν στη τεχνική των CT. Στις επόμενες παραγράφους εξετάζονται αναλυτικότερα τρεις από αυτές, ενώ παράλληλα συγκρίνονται μεταξύ τους μέσα από τη θεωρία τους αλλά και την πειραματική εφαρμογή τους (ΚΕΦ. 5-7).

3.2. CT ΜΕΘΟΔΟΙ

3.2.1. Μέθοδος των γραμμικών συνδυασμών (Discriminant-based linear combination method, LCM)

Η **μέθοδος των γραμμικών συνδυασμών** (Discriminant-based linear combination method, LCM), ταξινομεί τα αντικείμενα με βάση γραμμικούς συνδυασμούς που σχηματίζει με τις αρχικές συνεχείς μεταβλητές. Τα δέντρα ταξινόμησης που βασίζονται στην παραπάνω τεχνική, είναι συνήθως μικρότερα και ακριβέστερα από τα δέντρα που σχηματίζονται από

"μονοπαραμετρικές" τεχνικές (δηλαδή CT τεχνικές που χρησιμοποιούν μία μεταβλητή για τον κάθε διαχωρισμό). Η LCM είναι παρόμοια με την παραδοσιακή DA, αλλά δεν υπάρχει περιορισμός στον αριθμό των γραμμικών διαχωριστικών συναρτήσεων LDF που δημιουργούνται. Αντίθετα στην DA (βλ. § 2.1.1), ο αριθμός των LDF είναι ο μικρότερος αριθμός από τον αριθμό των μεταβλητών ή των ομάδων ταξινόμησης μείον ένα [17, 27]. Έτσι αν για παράδειγμα, θελήσουμε να ταξινομήσουμε 4 κατηγορίες μεταλλικών νομισμάτων αξίας 2€, 1€, 50 και 20 λεπτών, μόνο με τη βοήθεια δυο μεταβλητών (διάμετρος και ύψος νομίσματος), η κλασική DA, θα εξάγει 2 μόνο LDF. Η LCM ωστόσο, θα μπορούσε να εξάγει πολλές LDF, πραγματοποιώντας πολλούς διαχωρισμούς ακόμα και σε περιπτώσεις πολλών μεταβλητών και δυο μόνο κατηγοριών. Εξαιτίας της περιοδικότητας αυτής (επαναξιολόγηση και επαναξιοποίηση των δεδομένων σε κάθε διαχωρισμό), η LCM αξιοποιεί όλη την παρέχουσα πληροφορία και **δεν αφήνει ανεκμετάλλευτα δεδομένα,** σε αντίθεση με την DA. Συνολικά, τα πλεονεκτήματα της μεθόδου αυτής είναι:

1. κατασκευή μικρών δέντρων με υψηλή ακρίβεια,
2. απεριόριστοι γραμμικοί συνδυασμοί και
3. αξιοποίηση της συνολικής πληροφορίας που παρέχεται από τα δεδομένα, λόγω του περιοδικού χαρακτήρα της τεχνικής [17].

3.2.2. Μονοπαραμετρική Διαχωριστική μέθοδος (Discriminant-based univariate method, Classic CT)

Η **Μονοπαραμετρική Διαχωριστική μέθοδος** (Discriminant-based univariate method, Classic CT), μπορεί να θεωρηθεί ως μια ειδική περίπτωση της προηγούμενης μεθόδου, αν θεωρήσουμε μηδενικούς όλους τους συντελεστές για όλες τις μεταβλητές, εκτός από μίας. Έτσι, η ταξινόμηση που επιτυγχάνεται είναι κάθε φορά "δυαδική", με τη χρήση όμως μίας μόνο μεταβλητής [17].

Το πρώτο βήμα είναι να καθοριστεί ο καλύτερος τερματικός κόμβος που θα διαχωριστεί από το υπόλοιπο δέντρο και στη συνέχεια ποια μεταβλητή θα υλοποιήσει το διαχωρισμό. Σε κάθε τερματικό κόμβο, υπολογίζονται με τη μέθοδο ANOVA (για συνεχείς μεταβλητές) τα p-επίπεδα σημαντικότητας που αφορούν τη σχέση μεταξύ των τιμών των μεταβλητών που παίρνουν μέρος στο διαχωρισμό και τη δυνατότητα να ανήκει ένα δείγμα στην ομάδα του. Η μεταβλητή με το μικρότερο p-επίπεδο επιλέγεται να διαχωρίσει τον αντίστοιχο κόμβο.

Το επόμενο βήμα αφορά τον καθορισμό του επιπέδου στο διαχωρισμό. Εδώ χρησιμοποιούνται διάφοροι αλγόριθμοι, αλλά στην περίπτωση των συνεχών μεταβλητών, προτιμάται η

μέθοδος Συσταδοποίησης K-μέσων σημείων (K-means clustering) (βλ. § 2.4.1). Σε κάθε δυαδικό διαχωρισμό που πραγματοποιείται στην περίπτωση των CT, K = 2.

Η **προκατάληψη** ή **προδιάθεση** ή **πόλωση** (bias) στη μέθοδο των Classic CT, σε σχέση με τη **μονοπαραμετρική μέθοδο της Διεξοδικής Σάρωσης** που θα δούμε παρακάτω (βλ. § 3.2.3), είναι μειωμένη [17].

3.2.3. Μονοπαραμετρική μέθοδος της Διεξοδικής Σάρωσης (CART-style Exhaustive search method for univariate splits, CART)

Στη μονοπαραμετρική μέθοδο της **Διεξοδικής Σάρωσης για Δέντρα Ταξινόμησης και Συσχέτισης** (Classification and Regression Trees-style Exhaustive search method for univariate splits, CART), εξετάζονται όλοι οι δυνατοί συνδυασμοί διαχωρισμών, για κάθε μεταβλητή και σε κάθε κόμβο, ώστε να βρεθεί ο καλύτερος. Για παράδειγμα, για συνεχείς μεταβλητές με k διακριτά επίπεδα, θα υπάρχουν k-1 βήματα μεταξύ των k επιπέδων. Έτσι είναι φανερό, ότι όταν υπάρχει μεγάλος αριθμός μεταβλητών και μάλιστα σε πολλά επίπεδα, ο αριθμός των πιθανών διαχωρισμών που πρέπει να εξεταστεί γίνεται πολύ μεγάλος. Έτσι, αυτή η μέθοδος δείχνει προκατάληψη ως προς τις μεταβλητές με περισσότερα επίπεδα και χρειάζεται περισσότερη ώρα για υπολογισμούς. Ωστόσο, εγγυάται τους καλύτερους διαχωρισμούς και την καλύτερη ταξινόμηση. Η CART μέθοδος αξιολογεί όλες τις μεταβλητές για να καθοριστούν τελικά αυτές που δημιουργούν τους καλύτερους διαχωρισμούς, δηλαδή εκείνους με τις πιο "καθαρόαιμες" τάξεις στους κόμβους με τη βοήθεια του **δείκτη Gini** [45] (ΚΕΦ. 1, Θ). Η μέθοδος Gini θεωρείται καλύτερη από τη μέθοδο του δείκτη εντροπίας [42] η οποία αυξάνεται με την αύξηση της ανομοιογένειας σε ένα κόμβο (§ 5.3.8) [28, 46]. Δηλαδή η εντροπία φτάνει στην ελάχιστη τιμή της, όταν υπάρχει η μέγιστη δυνατή ομοιομορφία σε μια τάξη αντικειμένων [46]. Η CART μέθοδος παρουσιάζει ανοχή σε ελλιπή δεδομένα και θεωρείται ότι μπορεί να λειτουργήσει αποτελεσματικά ακόμα και σε περιπτώσεις όπου η έλλειψη των δεδομένων φτάνει σε ποσοστό μέχρι και 5 % [42].

Συνοψίζοντας, τα πλεονεκτήματα των μονοπαραμετρικών CT μεθόδων είναι:

1. κατασκευή δέντρων με πολύ καλή ακρίβεια,
2. δυνατότητα κατασκευής μικρών δέντρων με επιλογή των σωστών παραμέτρων και
3. άμεση αξιολόγηση της κρισιμότητας των μεταβλητών, λόγω και του μονοπαραμετρικού χαρακτήρα των τεχνικών.

ΚΕΦ. 4 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (ARTIFICIAL NEURAL NETWORKS, ANN)

4.1. ΕΙΣΑΓΩΓΗ

4.1.1. Προέλευση

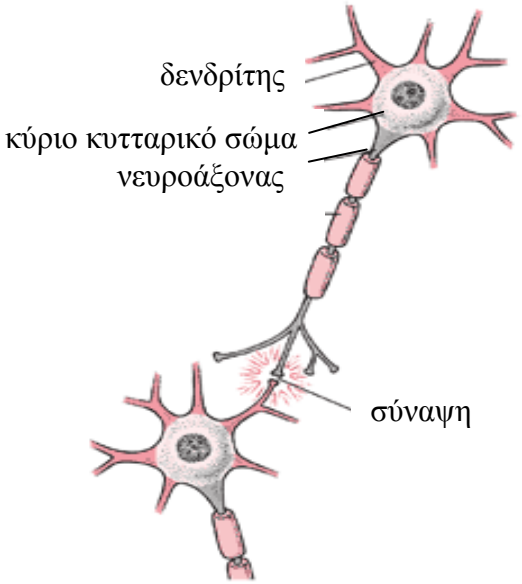
Η δεύτερη λέξη "neural" του όρου **Νευρωνικά Δίκτυα** (Artificial Neural Networks, ANN), δείχνει μια ξεκάθαρη σχέση με το νευρώνα ή νευρικό κύτταρο. Γίνεται μια συσχέτιση με το τμήμα εκείνο του ανθρώπινου σώματος που θεωρείται το πιο θαυμαστό στο ανθρώπινο είδος: τον εγκέφαλο. Ωστόσο, η έμφαση στον όρο "neural networks", δεν πρέπει να δοθεί στο επίθετο **"νευρωνικός"** ("neural"), αλλά στο ουσιαστικό **"δίκτυο"** ("networks") [1]. Η **δικτύωση** (networking) είναι από τα πιο σημαντικά στην οργάνωση οποιουδήποτε σχήματος: από τη διεύθυνση μιας τράπεζας μέχρι το χειρισμό επιστημονικών δεδομένων. Στα κεφάλαια λοιπόν που ακολουθούν, οι λέξεις "δικτύωση"/"δίκτυα" σημαίνουν μια σειρά μικρών μονάδων (τους νευρώνες) οι οποίες πραγματοποιούν το ίδιο "πακέτο" λειτουργιών όλη την ώρα. Ωστόσο, όπως θα αναλυθεί και παρακάτω, το τελικό αποτέλεσμα δεν οφείλεται τόσο σε αυτές τις λειτουργίες, αλλά στον τρόπο που αυτές οι μικρές μονάδες "συνδέονται" μεταξύ τους και αλλάζουν τις εσωτερικές τους παραμέτρους, ώστε να προσαρμόζουν το κάθε ατομικό εξερχόμενο σε ένα συνολικό εξωτερικό έλεγχο ή ανταγωνισμό μεταξύ των νευρώνων.

Αποσπάσματα από τη θεωρία που ακολουθεί δημοσιεύτηκαν πρόσφατα σε σχετικές εργασίες [47, 48].

Υποθέτοντας λοιπόν σωστά εξαρχής, η ανάπτυξη των Νευρωνικών Δικτύων έχει την απαρχή της στα **βιολογικά νευρωνικά δίκτυα**, τα οποία και προσπάθησαν οι πρώτοι ερευνητές να μιμηθούν. Ο ανθρώπινος εγκέφαλος αποτελείται βασικά από ένα μεγάλο αριθμό νευρώνων (περίπου 10^{10}) συνδεδεμένων μεταξύ τους (αντιστοιχούν μερικές χιλιάδες συνδέσεις σε κάθε νευρώνα, αν και ο αριθμός αυτός, γενικά ποικίλει). Κάθε νευρώνας είναι ένα εξειδικευμένο κύτταρο που μπορεί να διαβιβάσει ένα ηλεκτροχημικό σήμα. Ο νευρώνας αποτελείται από μια διακλαδισμένη δομή εισόδου (input), τους **δενδρίτες** (dendrites), ένα **κύριο κυτταρικό σώμα** (cell body) και από μια διακλαδισμένη δομή εξόδου (output), τους **νευροάξονες** (axon, σχήμα 4.1) [17]. Οι δενδρίτες εξαιτίας της εκτεταμένης διακλάδωσης, καταλαμβάνουν μια αρκετά μεγάλη περιοχή (μέχρι 0,25 mm^2). Οι νευροάξονες του ενός κυττάρου συνδέονται με τους δενδρίτες του άλλου, μέσω των **συνάψεων** (synapse): πάνω από 40 % της επιφάνειας ενός νευρώνα καλύπτεται από τις συνάψεις [1]. Οι νευρώνες ως κύτταρα πιστεύεται ότι δεν πολλαπλασιάζονται και δεν αναπαράγονται, ενώ αντίθετα, καθόλη τη διάρκεια της ζωής ενός οργανισμού οι συνάψεις βρίσκονται σε μια δυναμική ισορροπία, δημιουργούνται καινούργιες και

καταστρέφονται παλιές. Η δημιουργία των νέων συνάψεων γίνεται όταν ο εγκέφαλος αποκτά περισσότερες εμπειρίες από το περιβάλλον, μαθαίνει, αναγνωρίζει, κατανοεί [49].

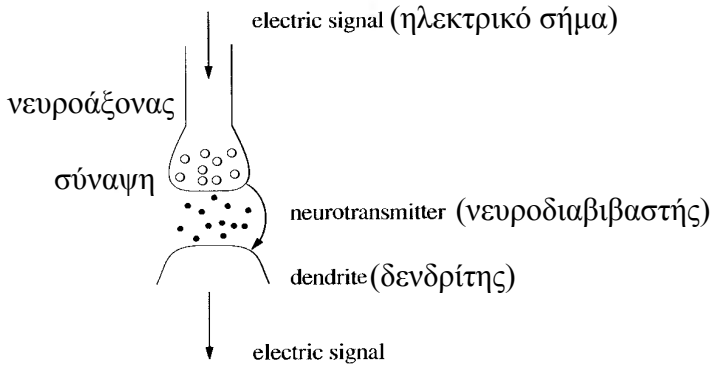
Όταν ένας νευρώνας ενεργοποιείται, "πυροδοτεί" ένα ηλεκτροχημικό σήμα κατά μήκος του νευροάξονα. Αυτό το σήμα "διασχίζει" τη σύναψη προς το επόμενο κύτταρο, το οποίο μπορεί επίσης να "πυροδοτηθεί". Ένας νευρώνας πυροδοτείται μόνο όταν το σήμα που φτάσει από τους δενδρίτες του υπερβαίνει ένα ορισμένο επίπεδο: το κατώφλι πυροδότησης (firing threshold) [17]. Ενδιάμεσες καταστάσεις δεν υπάρχουν. Κατά κάποιο τρόπο δηλαδή, ο βιολογικός νευρώνας είναι ένα δυαδικό στοιχείο [49].



Σχήμα 4.1: Ο νευρώνας [50]

Η διάδοση των σημάτων από τους δενδρίτες μέχρι το νευροάξονα γίνεται ηλεκτρικά, δηλαδή με τη **μεταφορά ιόντων.** Ωστόσο, η διάδοση αυτή κατά μήκος των συνάψεων γίνεται με **χημικές ουσίες.** Το ηλεκτρικό σήμα ελευθερώνει στον άξονα μια χημική ουσία, ένα νευροδιαβιβαστή, η οποία αποθηκεύεται σε κυψελίδες στην προσυναπτική μεμβράνη. Ο νευροδιαβιβαστής διασχίζει το "κενό" των συνάψεων (σχ. 4.2) και εισέρχεται στους δενδρίτες του επόμενου νευρώνα [1].

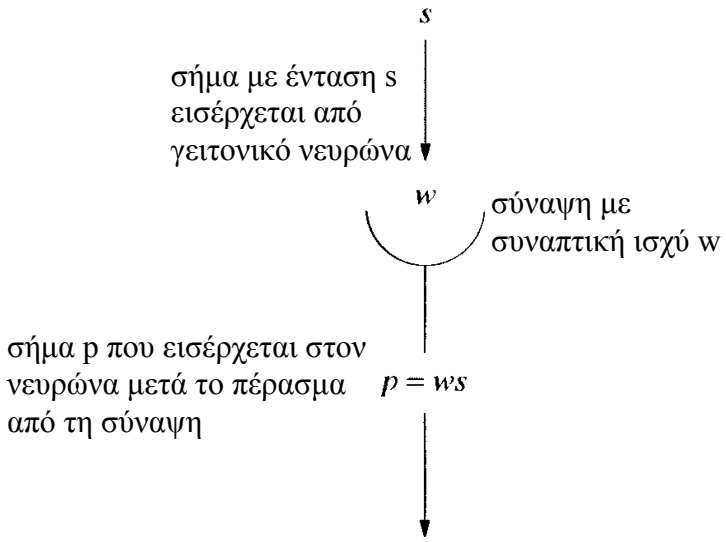
Η σύναψη μπορεί να ελευθερώσει το νευροδιαβιαβαστή προς μία μόνο κατεύθυνση και έτσι δρα ως πύλη ελέγχου. Αυτό είναι πολύ σημαντικό για τη μεταβίβαση της πληροφορίας και βρίσκει αντιστοιχία στα ANN. Στο δενδρίτη, ο νευροδιαβιβαστής δημιουργεί ένα νέο ηλεκτρικό σήμα, το οποίο περνά κατά μήκος του δεύτερου νευρώνα. Τα σήματα που παράγονται στους νευρώνες, ανεξάρτητα από το κύτταρο ή τον οργανισμό είναι παρόμοια μεταξύ τους [1, 49, 51]. Ωστόσο, η ένταση των σημάτων που παράγονται, μπορεί να διαφέρει, εξαρτώμενη από την ένταση του ερεθίσματος [1].



Σχήμα 4.2: Σχηματική αναπαράσταση των συνάψεων [1]

Το τελευταίο αυτό συμπέρασμα είναι πολύ βασικό για την κατανόηση των ANN. Είναι φανερό, ότι η ομοιότητα των σημάτων υποδηλώνει ότι η αληθινή λειτουργία του εγκεφάλου δεν βασίζεται τόσο στο ρόλο ενός μόνο νευρικού κυττάρου, όσο στο **σύνολο** αυτών, δηλαδή τον τρόπο που αυτά συνδέονται μεταξύ τους [1, 51].

Οι συνάψεις μέσω των οποίων σήματα γειτονικών κυττάρων εισέρχονται στο συγκεκριμένο κύτταρο, αντιπροσωπεύουν φράγματα που διαμορφώνουν/προσαρμόζουν το σήμα που φτάνει σε αυτές. Το μέγεθος της αλλαγής που υφίσταται στο σήμα εξαρτάται από τη **συναπτική ισχύ** (synaptic strength), η οποία συνεχώς αλλάζει και προσαρμόζεται (όπως τα βάρη των ANN [52] που θα δούμε παρακάτω, § 4.1.2). Είναι γνωστό το **πείραμα του Pavlov** (πείραμα Pavlovian), όπου το χτύπημα ενός κουδουνιού αποτελούσε το ειδοποιητήριο σήμα για ένα σκύλο ως προς το γεύμα που θα ακολουθούσε. Έτσι κάθε φορά που ο σκύλος άκουγε το κουδούνι, οι σιελογόνοι αδένες του, δημιουργούσαν έντονη ροή σάλιου, ως αποτέλεσμα της ενδυνάμωσης των συνάψεων μεταξύ αυτών και του κατάλληλου τμήματος του ακουστικού πόρου [17]. Έτσι, ο σκύλος **έμαθε** σύντομα πώς συνδέονται το χτύπημα του κουδουνιού με το φαγητό. Στα ANN η συναπτική ισχύς αποτελεί το βάρος της σύνδεσης που θα δούμε αμέσως παρακάτω (σχ. 4.3). Χωρίς να υπεισέλθουμε παραπάνω στη φυσική και χημεία των μεμβρανών, μπορούμε τα πούμε ότι η συναπτική ισχύς καθορίζει την "ποσότητα" του σήματος που εισέρχεται στο κυτταρικό σώμα διαμέσου των δενδριτών. Γρήγορες αλλαγές/προσαρμογές στη συναπτική ισχύ θεωρούνται ζωτικής σημασίας για την ορθή και αποτελεσματική αλλαγή του εγκεφάλου. Η προσαρμογή της συναπτικής ισχύος σε ένα συγκεκριμένο πρόβλημα θεωρείται η ουσία της εκπαίδευσης [1]. Η έννοια αυτή εξάλλου εμπεριέχει τα δυο βασικά χαρακτηριστικά που μιμήθηκαν τα ANN από τα αντίστοιχα βιολογικά: **ανοχή στο λάθος** ("fault tolerance") και **ικανότητα για εκπαίδευση** ("capacity to learn"). Πιο συγκεκριμένα, αν και δεν μπορούμε να προσδιορίσουμε ένα γενικό ποσοστό ανοχής σφάλματος των νευρωνικών δικτύων, η μεγαλύτερη δυνατή ανοχή είναι της τάξης του 10–15 % [49].



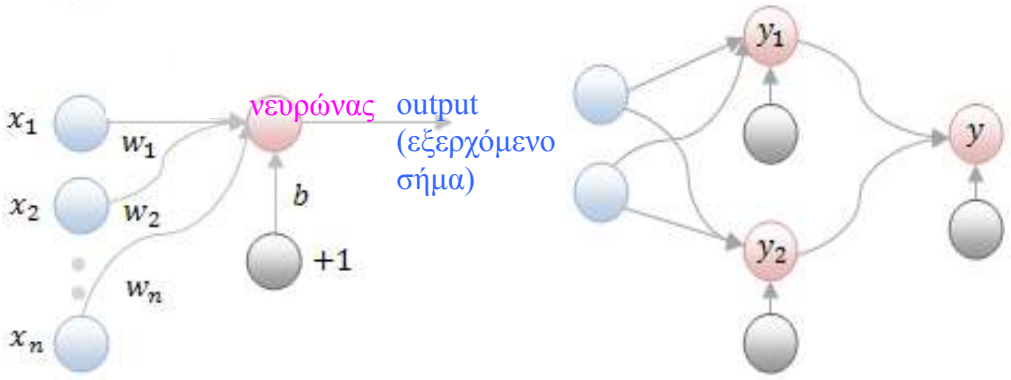
Σχήμα 4.3: Σχηματική αναπαράσταση των συνάψεων [1]

Ένας βιολογικός νευρώνας μπορεί να "πυροδοτηθεί" ξανά μετά από 10^{-3} s (= 1 ms). Εφόσον λοιπόν, ο χρόνος αντίδρασης των περισσότερων σπονδυλωτών είναι 10^{-1} s, υποθέτουμε ότι οποιαδήποτε αντίδραση εκδηλώνεται σε λιγότερο από 100 "επαναλήψεις πυροδότησης" ("firing times"). Αυτό ονομάζεται **"το παράδοξο των εκατό βημάτων"** ("hundred steps paradoxon"). Το εντυπωσιακό συμπέρασμα που καταλήγουμε από τα παραπάνω, είναι ότι ο εγκέφαλος μπορεί να διαχειριστεί έναν ισχυρότατο αλγόριθμο επεξεργασίας σημάτων που μπορεί να διεκπεραιώσει και τις πιο δύσκολες αποστολές σε λιγότερα από 100 βήματα [1]!

4.1.2. Τα βασικά: νευρώνες και βάρη

Η ανάπτυξη των ANN βασίστηκε στην αρχική ιδέα του Rosenblatt [53], ο οποίος εισήγαγε την έννοια της πρωταρχικής μονάδας του **Λειτουργικού Νευρωνικού Δικτύου ή Στοιχειώδη Αισθητήρα** (perceptron) [49]. Το τελευταίο δίνει εξερχόμενα δυαδικά: 0/+1 (binary outputs) ή δίπολα: −1/+1 (bipolar outputs) κατά άλλους ερευνητές, ανάλογα με τους "ζυγισμένους" γραμμικούς συνδυασμούς των εισερχομένων μεταβλητών (inputs).

Η δομή του perceptron είναι πολύ απλή: υπάρχει ένας αριθμός εισερχομένων δεδομένων (x_1, x_2, ….x_n), ένας αριθμός βαρών (weights, w_1, w_2, …..w_n), μια **προκατάληψη** (bias, b) και ένα εξερχόμενο σήμα (σχ. 4.4(α)) [54]. Εναλλακτικά μπορεί να χρησιμοποιηθεί ένα **κατώφλι (threshold, θ)** η έννοια του οποίου θα αναλυθεί παρακάτω (σχ. 4.7(α) και § 4.1.4, 4.2.1, 4.3.4). Η προκατάληψη θεωρείται επίσης βάρος (εναλλακτικά μπορεί να συμβολίζεται με w_0) μιας εισερχόμενης μεταβλητής που η τιμή της είναι ίση με τη μονάδα ($x_0 = 1$). Με τη χρήση της, προστίθεται στο σύστημα ένας παραπάνω βαθμός ελευθερίας (παρέχει στην έξοδο μια βάση δραστηριότητας) και αυξάνεται η ευελιξία του δικτύου (βλ. § 4.2.1).



(α) (β)

Σχήμα 4.4: Το perceptron (α) σε σχέση με το Multi-layers perceptron Network (β) [54]

Για κάθε εισερχόμενη μεταβλητή x_i, εφαρμόζεται ένα βάρος w_i το οποίο πολλαπλασιάζεται με την τιμή αυτής και υπολογίζεται τελικά (με την προσθήκη της προκατάληψης b), το άθροισμα Σ:

$$\Sigma = \sum_{i}^{n} x_i w_i + b \qquad (4.1)$$

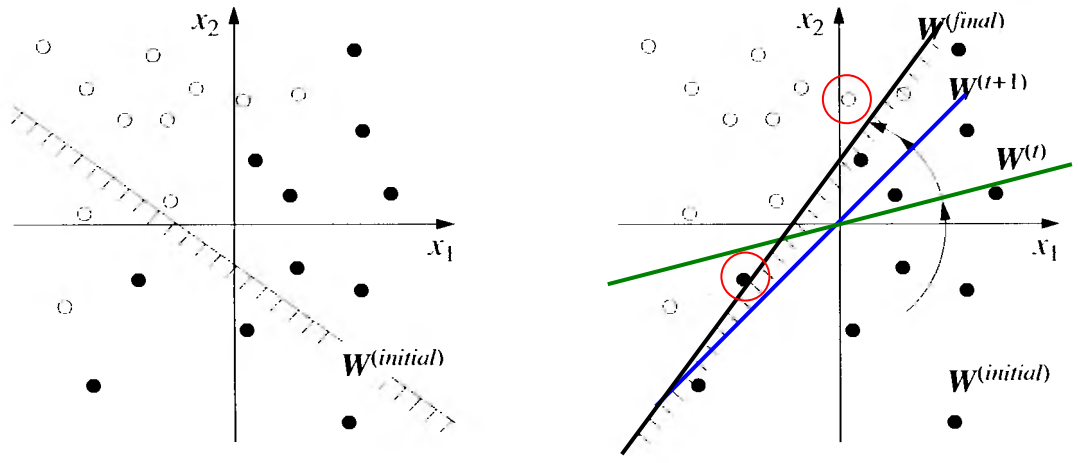
Η παραπάνω σχέση ορίζει μερικώς την έννοια του **νευρώνα** ή **μονάδας ενός δικτύου** (neuron ή processing unit ή element). Η συνάρτηση ενεργοποίησης ή μεταφοράς, την οποία θα δούμε παρακάτω (βλ. § 4.3.4) αποτελεί τη δεύτερη σχέση που συμπληρώνει τον ορισμό του νευρώνα. Ως νευρώνας λοιπόν, ορίζεται μια μονάδα η οποία επεξεργάζεται εισερχόμενα δεδομένα, ώστε να παραχθούν με τη βοήθεια βαρών και ειδικών συναρτήσεων ένα ή περισσότερα εξερχόμενα. Ο νευρώνας συνολικά:

1. μετασχηματίζει τα εισερχόμενα σήματα σε εξερχόμενο σήμα, χρησιμοποιώντας βάρη και μια συνάρτηση ενεργοποίησης,
2. μεταδίδει το σήμα στον επόμενο νευρώνα (σχ. 4.4(α)),
3. μεταδίδει το υπολογιζόμενο σφάλμα που λαμβάνει από τον επόμενο νευρώνα, στον προηγούμενο (βλ. delta-rule, § 4.3.6).

Η παρουσία των βαρών στην παραπάνω σχέση είναι πολύ σημαντική, καθώς ανάλογα και με τη δομή του δικτύου, τα βάρη καθορίζουν τη διάδοση του σήματος μέσα σε αυτό. Είναι πραγματικοί αριθμοί που **καθορίζουν την ισχύ κάθε σύνδεσης** ανάμεσα σε δυο νευρώνες. Κάθε σήμα που μεταφέρεται μεταξύ των συνδέσεων πολλαπλασιάζεται με το αντίστοιχο βάρος της σύνδεσης. Τα βάρη προσαρμόζονται και βελτιστοποιούνται. Τα βάρη μπορούν να είναι θετικά ή αρνητικά (δρουν λοιπόν διεγερτικά αλλά και ανασταλτικά). Γενικά, συνιστώνται χαμηλές τιμές γι' αυτά (−1 έως +1, βλ. § 4.3.4). Καθορίζουν ουσιαστικά τη σχέση μεταξύ των εισερχόμενων δεδομένων και της ή των εξερχόμενων σημάτων ή πληροφοριών. Έτσι, μπορεί

να θεωρηθεί ότι περιέχουν μια συνολική γνώση του δικτύου και αποτελούν τη "μνήμη" ολόκληρου του συστήματος [55].

Η έννοια της παραμετρικής συνάρτησης που αντιπροσωπεύουν οι νευρώνες, γίνεται φανερή με τα βάρη w_i, τα οποία εδώ συνδέονται άμεσα με τα εισερχόμενα δεδομένα (βλ. σχέση (4.1)) [56]. (Η σχέση αυτή όπως θα δούμε παρακάτω, αφορά εκτός του perceptron, και τη δημοφιλέστερη τεχνική ANN: τα perceptrons πολλαπλών στιβάδων ή Πολυστιβαδικά Νευρωνικά Δίκτυα (Multi-layer perceptrons, MLP), βλ. σχήμα 4.4(β) και § 4.1.4, 4.3.3).



(α) (β)

Σχήμα 4.5: Αλλάζοντας το αρχικό διάνυσμα $W^{initial}$ (α) προς μια καλύτερη θέση: το W^{t+1} είναι καλύτερο από το W^t (β) γιατί ταξινομεί λάθος μόνο δυο αντικείμενα, σε σχέση με τα πέντε του W^t. Το τελικό διάνυσμα W^{final} θα ταξινομεί όλα τα αντικείμενα από τις δυο κατηγορίες εκτός από δύο (ένα μαύρο και έναν άδειο κύκλο) [1].

Η αρχική επιλογή των βαρών γίνεται συνήθως τυχαία [57, 58], ξεκινώντας από μια (προφανώς κακή) εικασία αρχική $\mathbf{W}^{\mathbf{initial}}$ (σχ. 4.5(α)). (Τα βάρη **W** σημειώνονται εδώ με έντονη γραφή γιατί αφορούν διανύσματα και όχι μονόμετρα μεγέθη). Το αρχικό βάρος $\mathbf{W}^{\mathbf{initial}}$ βελτιώνεται διαρκώς αναδρομικά. Αυτό σημαίνει ότι το κάθε φορά το επόμενο βάρος $\mathbf{W}^{\mathbf{t+1}}$, προκύπτει από το προηγούμενο $\mathbf{W}^{\mathbf{t}}$ (σχ. 4.5(β)) [1]. Σπανίως αναφέρονται στη βιβλιογραφία και άλλοι τρόποι εκτός της τυχαίας επιλογής των βαρών (βλ. § 4.3.10). Η διαμόρφωση των κατάλληλων βαρών πραγματοποιείται στη διάρκεια της **εκπαίδευσης** (ή "εκμάθησης"), με τη βοήθεια του **"κανόνα εκπαίδευσης"** ("learning rule"). Η εκπαίδευση, βλ. § 4.3.1) του δικτύου είναι το σημαντικότερο στάδιο στην ανάπτυξη των ANN [51].

4.1.3. Εν αρχή: η έννοια του perceptron

Το perceptron μπόρεσε εύκολα να ανταπεξέλθει στα κλασικά προβλήματα των συναρτήσεων της λογική διάζευξης (OR) και λογική σύζευξης (AND) (πίνακας 4.1). Με την προϋπόθεση ότι τα τρία από τα τέσσερα σημεία ((0, 0), (0,1), (1,0) και (1,1)) ανήκουν στην ίδια ομάδα, το πρόβλημα επιλύεται **γραμμικά.** Ο διαχωρισμός των δεδομένων, επιτυγχάνεται γραμμικά σε πολλά τέτοιου είδους προβλήματα, με κατάλληλη επιλογή των βαρών και της προκατάληψης. Στην περίπτωση μάλιστα που έχουμε δύο εισερχόμενες μεταβλητές, τότε ο διαχωρισμός γίνεται από μια ευθεία γραμμή. Αν το πρόβλημά μας είχε τρεις εισερχόμενες μεταβλητές, τότε ο διαχωρισμός θα γινόταν από ένα επίπεδο που θα έτεμνε τον τρισδιάστατο χώρο [49].

Πίνακας 4.1: Διανύσματα εκπαίδευσης **x** για τις 3 συναρτήσεις: δίνονται και οι δυο περιπτώσεις: δυαδικά 0 /+1 (ή δίπολα −1/+1)

Εισερχόμενο διάνυσμα (x_1, x_2)		**Θεωρητική απόκριση, output d(x)**		
		συνάρτηση λογικής διάζευξης, OR	**συνάρτηση λογικής σύζευξης, AND**	**συνάρτηση αποκλειστικής διάζευξης, XOR**
(0, 0)	(−1, −1)	0 (−1)	0 (−1)	0 (−1)
(0, 1)	(−1, 1)	1	0 (−1)	1
(1, 0)	(1, −1)	1	0 (−1)	1
(1, 1)		1	1	0 (−1)

Ο επιβλεπόμενος κανόνας εκπαίδευσης του perceptron περιγράφεται ως εξής [59]:

1. Αρχίζουμε με τυχαία βάρη (w_i) και προκατάληψη (b).
2. Επιλέγουμε ένα εισερχόμενο δείγμα (διάνυσμα) **x** (x_i) από την ομάδα εκπαίδευσης και υπολογίζουμε την απόκριση y.
3. Αν $y \neq d(x)$, όπου d(x) η θεωρητική απόκριση του **x**, αλλάζουμε όλα τα βάρη w_i, ώστε $\Delta w_i = d(x)x_i$ και:

$$\Delta b = \begin{cases} 0 \text{ αν η απόκριση του perceptron είναι η σωστή} \\ d(x) \text{ στην αντίθετη περίπτωση} \end{cases}$$

4. Επιστρέφουμε στο $2^{ο}$ βήμα.

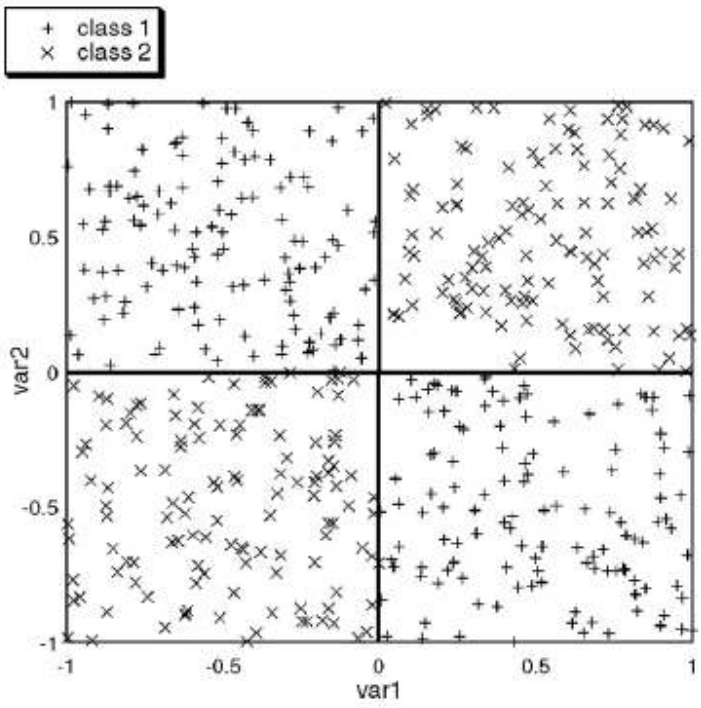
Έτσι, οι σχέσεις που ορίζουν το perceptron του πίνακα 4.1 για δίπολα εισερχόμενα είναι:

$$\Sigma = \sum_i^n x_i w_i + b \quad \text{και} \quad y = F(\Sigma) = \begin{cases} 1 \text{ αν } \Sigma > 0 \\ -1 \text{ αν } \Sigma \leq 0 \end{cases} \qquad (4.2)$$

Ένα απλό παράδειγμα κατανόησης του perceptron αναφέρεται στο ηλεκτρονικό παράρτημα της διατριβής (ΚΕΦ. 2, Θ).

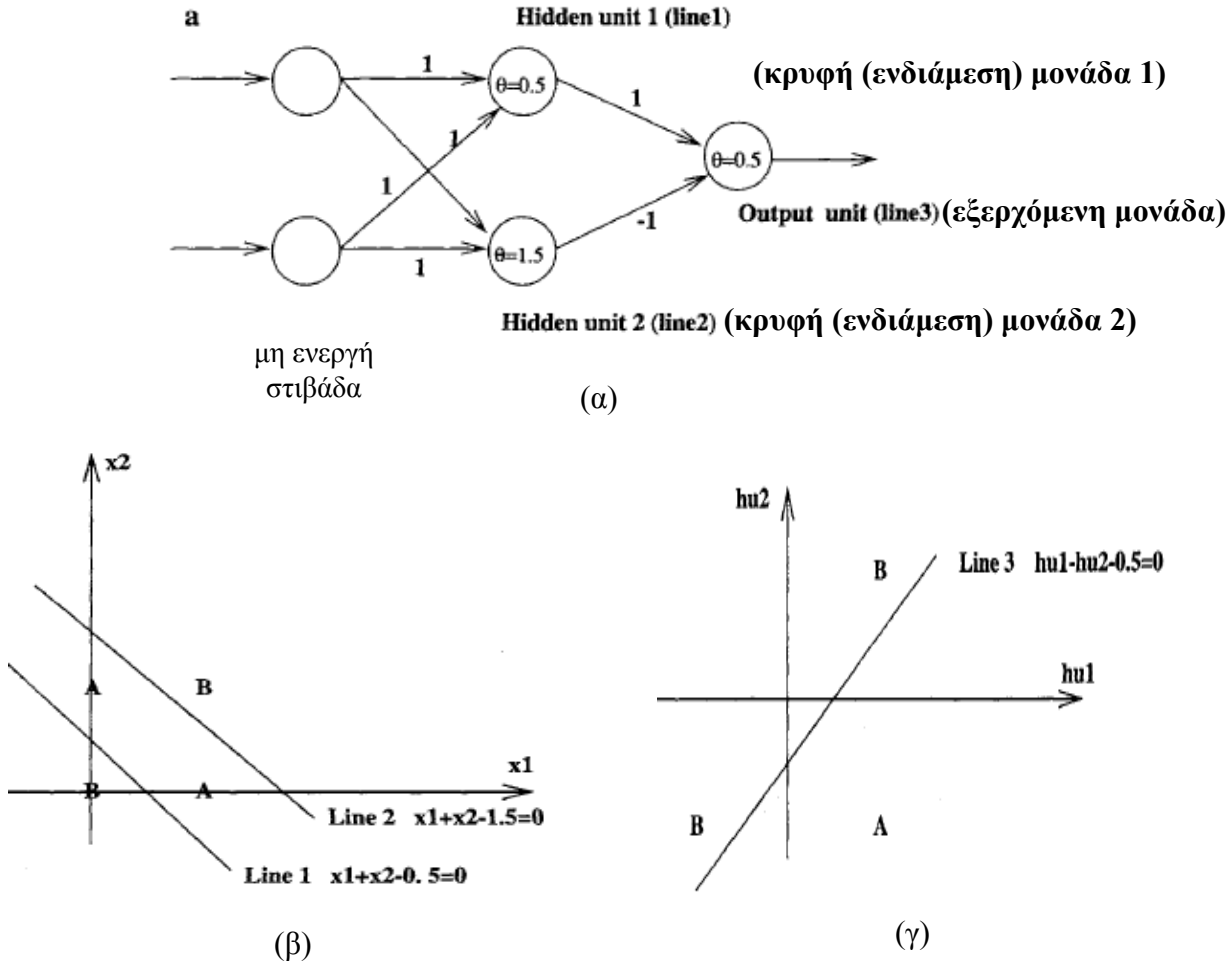
4.1.4. Η επέκταση: Πολυστιβαδικά Νευρωνικά Δίκτυα ή Perceptrons Πολλαπλών Στιβάδων (Multi-layers perceptron, MLP)

Ωστόσο, το perceptron είχε περιορισμένες δυνατότητες: η λύση του προβλήματος της κλασικής **XOR συνάρτησης** (exclusive OR ή αποκλειστικής διάζευξης: ή το ένα ή το άλλο, όχι όμως και τα δύο), (πίνακας 4.1), δεν ήταν δυνατή. Στην XOR συνάρτηση, όταν οι τιμές των δυο μεταβλητών είναι ίδιες, τα δείγματα ανήκουν στην πρώτη ομάδα (αποτέλεσμα 0 ή −1), ενώ αν είναι διαφορετικές στη δεύτερη ομάδα (αποτέλεσμα 1, πίνακας 4.1 και σχ. 4.6) [55].



Σχήμα 4.6: Αναπαράσταση της XOR συνάρτησης [55]

Το μοντέλο της μονής στιβάδας του perceptron (input-output), δεν μπορούσε να διαχωρίσει τα δεδομένα αυτά με μία μόνο γραμμή. Η λύση δίνεται με μια "**ενδιάμεση", "κρυφή"** (hidden) **στιβάδα** (layer) με επιπλέον μονάδες ή νευρώνες (σχ. 4.4(β), 4.7(α)) [51]. Η νέα στιβάδα ονομάζεται κρυφή γιατί δεν μπορούμε (άμεσα τουλάχιστον) να δούμε το εξερχόμενο από αυτήν [60] ή επειδή δεν συνδέεται άμεσα με τον εξωτερικό "ορατό" κόσμο [58]. Στην πραγματικότητα η ενδιάμεση στιβάδα δημιουργεί την εσωτερική αναπαράσταση των σημάτων εισόδου, αλλά δεν "βλέπει" κατευθείαν ούτε την είσοδο ούτε την έξοδο του δικτύου αλλά μόνον το εσωτερικό του [49].



Σχήμα 4.7: (α) Η δομή του δικτύου που λύνει το κλασσικό XOR πρόβλημα. (β) Οι δυο γραμμές, όπως αυτές καθορίζονται από τις ενδιάμεσες μονάδες. (γ) Αναπαράσταση των δειγμάτων, όπως αυτά καθορίζονται από τις ενδιάμεσες μονάδες hu1 και hu2 [51].

Όταν αναφέρουμε τη λέξη στιβάδα, εννοούμε μια ομάδα νευρώνων, οι οποίοι έχουν όλοι τον ίδιο αριθμό βαρών και όλοι λαμβάνουν ταυτόχρονα το ίδιο πολυδιάστατο εισερχόμενο σήμα. Ο παραπάνω ορισμός αναφέρεται στην "ενεργή στιβάδα", ώστε να γίνει διαφοροποίηση από την εισερχόμενη για παράδειγμα στιβάδα, η οποία είναι ανενεργή (σχ. 4.7(α)): δεν περιέχει βάρη ή συνάρτηση ενεργοποίησης (§ 4.2.1, 4.3.4) και συμβάλλει απλώς στη διανομή των σημάτων, χωρίς να τα διαμορφώνει [1] ή να κάνει υπολογισμούς [59, 60]. Η εισερχόμενη στιβάδα θεωρείται ότι δεν συμμετέχει στην καταμέτρηση των στιβάδων ενός δικτύου τουλάχιστον στα "εμπροσθοτροφοδοτούμενα" δίκτυα (§ 4.2.1) [1]. Ωστόσο, συχνά καταμετράται στη βιβλιογραφία [59, 61 - 63], καθώς δεν υπάρχει μια γενικά αποδεκτή άποψη.

Με τη χρήση της ενδιάμεσης στιβάδας, το πρόβλημα διαχωρίζεται σε τρία διαφορετικά προβλήματα, τα οποία λύνονται με απλές γραμμές (σχ. 4.7(β)). Συγκεκριμένα, οι δυο ομάδες των δειγμάτων (A: (0,1) και (1,0) και B: (0,0) και (1,1)), δεν μπορούν να διαχωριστούν με μία απλή γραμμή. Όταν όμως αξιοποιηθούν τα εξερχόμενα από την ενδιάμεση στιβάδα (μονάδες

hu1 και hu2), οι ομάδες αυτές, διαχωρίζονται απλά με μια τρίτη γραμμή (σχ. 4.7(γ)). Παρακάτω περιγράφεται αναλυτικά η πορεία λύσης του XOR προβλήματος [51].

Η συνάρτηση που χρησιμοποιείται στις δυο ενδιάμεσες μονάδες (σχ. 4.7(α)), αλλά και στην εξερχόμενη είναι η δυαδική βηματική (binary step) (βλ. § 4.3.4) του τύπου:

$$f(x) = \begin{cases} 1 \text{ αν } x \geq \theta \\ 0 \text{ αν } x < \theta \end{cases} \qquad (4.3)$$

όπου θ (κατώφλι) = 0,5 / 1,5 και 0,5 αντίστοιχα για τις τρεις μονάδες.

Αναφέρουμε τώρα ένα παράδειγμα υπολογισμού για το πρώτο διάνυσμα (0,0):

Εισερχόμενο στην πρώτη ενδιάμεση μονάδα (hu1): $0 \times 1 + 0 \times 1 = 0 < 0,5$

Εξερχόμενο από την πρώτη ενδιάμεση μονάδα (hu1), μετά την εφαρμογή της συνάρτησης: 0

Εισερχόμενο στη δεύτερη ενδιάμεση μονάδα (hu2): $0 \times 1 + 0 \times 1 = 0 < 1,5$

Εξερχόμενο από τη δεύτερη ενδιάμεση μονάδα (hu2), μετά την εφαρμογή της συνάρτησης: 0

Εισερχόμενο στην τελική μονάδα: $0 \times 1 + 0 \times (-1) = 0 < 0,5$

Εξερχόμενο από την τελική, μετά την εφαρμογή της συνάρτησης: 0.

Τα αποτελέσματα για τα υπόλοιπα διανύσματα φαίνονται στον πίνακα 4.2.

Πίνακας 4.2: Ενδιάμεσα και τελικά αποτελέσματα για τα δυαδικά διανύσματα εκπαίδευσης x της XOR συνάρτησης

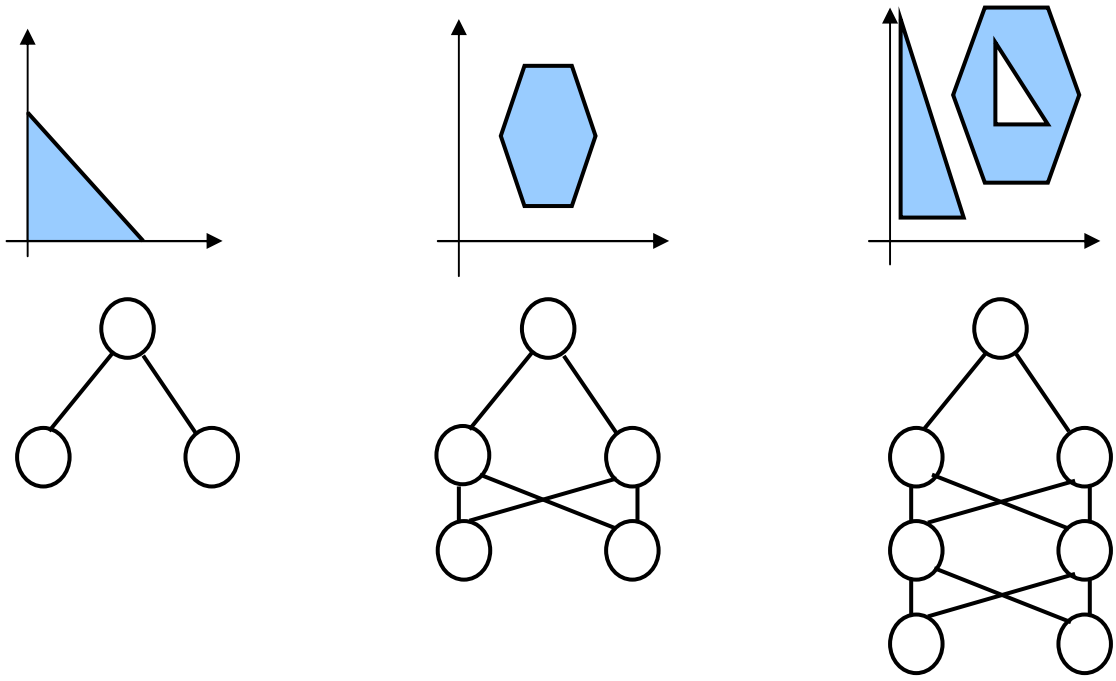
Εισερχόμενο διάνυσμα input (x_1, x_2)	**Εισερχόμενο στην πρώτη ενδιάμεση μονάδα (hu1)**	**Εξερχόμενο από την πρώτη ενδιάμεση μονάδα (hu1)**	**Εισερχόμενο στη δεύτερη ενδιάμεση μονάδα (hu2)**	**Εξερχόμενο από τη δεύτερη ενδιάμεση μονάδα (hu2)**	**Τελικό εξερχόμενο**
(0, 0)	0	**0**	0	**0**	0
(0, 1)	1	**1**	1	**0**	1
(1, 0)	1	**1**	1	**0**	1
(1, 1)	2	**1**	2	**1**	0

Τα δείγματα της ομάδας Α (0,1) και (1,0), δίνουν στην ενδιάμεση στιβάδα το ίδιο αποτέλεσμα (hu1 = 1 και hu2 = 0, πίνακας 4.2 και σχ. 4.7(γ)), ενώ της ομάδας Β (0,0) και (1,1) εξακολουθούν να δίνουν διαφορετικά αποτελέσματα [51]. Έτσι τα αρχικά τέσσερα (4) σημεία

γίνονται τρία (3) και μία απλή γραμμή, αρκεί για να τα διαχωρίσει. Οι υπολογισμοί πραγματοποιούνται στην ενδιάμεση και εξωτερική στιβάδα [64].

Στο παρακάτω σχήμα (4.8) φαίνονται οι δυνατότητες που προσφέρουν οι πολυπλοκότερες ANN αρχιτεκτονικές αρχίζοντας από καμιά μέχρι δυο ενδιάμεσες στιβάδες. Το μοντέλο δύο στιβάδων μπορεί να ξεχωρίσει σημεία που περιλαμβάνονται σε ανοιχτές ή κλειστές κυρτές περιοχές [49]. Παραπάνω ενδιάμεσες στιβάδες φαίνονται να "χαράσσουν" πιο πολύπλοκα όρια.

Με απαρχή λοιπόν το XOR πρόβλημα, αναπτύχθηκαν προοδευτικά τα MLP: η προσθήκη των ενδιάμεσων στιβάδων έδρασε καταλυτικά και μπόρεσε να δοθεί λύση σε ένα πλήθος προβλημάτων με τη βοήθεια των ANN.



(α) (β) (γ)

Σχήμα 4.8: Δυνατότητες των MLP: (α) Η αρχιτεκτονική χωρίς καμιά ενδιάμεση στιβάδα "χαράσσει" γραμμικά όρια. (β) Η ενδιάμεση στιβάδα ενώνει τις γραμμές. (γ) Οι δυο ενδιάμεσες στιβάδες "χαράσσουν" πολύπλοκα όρια [65].

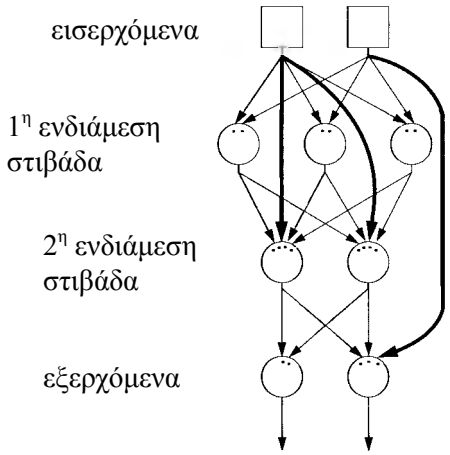
4.2. ΓΕΝΙΚΗ ΠΕΡΙΓΡΑΦΗ ΤΩΝ ANN

4.2.1. Βασικά χαρακτηριστικά

Τα Νευρωνικά Δίκτυα λαμβάνουν ένα αριθμό εισερχομένων δεδομένων στις μονάδες επεξεργασίας (νευρώνες), οι οποίες είναι ικανές να επικοινωνήσουν μεταξύ τους με την αποστολή μηνυμάτων διαμέσου "ζυγισμένων" συνδέσεων.

Όλα τα "τοπολογικά" στοιχεία ενός δικτύου συνιστούν την **"αρχιτεκτονική"** του (architecture ή design)**:**

1. ο αριθμός των εισερχόμενων δεδομένων και εξερχόμενων σημάτων,
2. ο αριθμός των στιβάδων,
3. ο αριθμός των νευρώνων σε κάθε στιβάδα,
4. ο αριθμός των βαρών σε κάθε νευρώνα,
5. ο τρόπος που οι νευρώνες συνδέονται μεταξύ τους μέσα στην ίδια στιβάδα, ή μεταξύ των στιβάδων (μπορεί για παράδειγμα, μερικοί νευρώνες να συνδέονται με άλλους μεθεπόμενης στιβάδας, ή κάποιοι νευρώνες να μην λαμβάνουν καθόλου κάποια σήματα (σχ. 4.9)),
6. ποιοι νευρώνες συμμετέχουν στη διόρθωση των βαρών [1].



Σχήμα 4.9: Απεικόνιση ενός δικτύου δυο ενδιάμεσων στιβάδων, με κάποιους νευρώνες να "προσπερνιόνται" και το σήμα να "στοχεύει" στην αμέσως επόμενη στιβάδα [1].

Η κυριότερη διάκριση μεταξύ των διαφόρων τύπων Νευρωνικών Δικτύων αφορά ακριβώς τη μορφή των συνδέσεων μεταξύ των νευρώνων. Έτσι πρωταρχικά διακρίνουμε:

1. **Εμπροσθοτροφοδοτούμενα δίκτυα** (Feed-forward) (σχ. 4.8), όπου η ροή των δεδομένων γίνεται αυστηρά από την πρώτη προς την τελευταία στιβάδα. Ενδιάμεσες, πολλαπλές ή όχι στιβάδες, μπορεί να συνυπάρχουν, αλλά οποιαδήποτε ανάδραση (feedback), στις συνδέσεις, είναι απαγορευτική. Αυτό σημαίνει ότι δεν υπάρχουν συνδέσεις από τα εξερχόμενα προς τα εισερχόμενα, ή μεταξύ μονάδων της ίδιας στιβάδας [59, 66, 67]. Η πληροφορία που "διαδίδεται" από τα εισερχόμενα προς τα εξερχόμενα, διαμορφώνεται με βάση τα βάρη των συνδέσεων. Τα εισερχόμενα δεν πραγματοποιούν κανένα υπολογισμό, αλλά "κατανέμουν" την πληροφορία στους επόμενους νευρώνες [1, 17, 59, 68].

Τα πιο συνήθη από τα εμπροσθοτροφοδοτούμενα δίκτυα είναι τα πλήρως συνδεδεμένα (fully connected), όπου όλοι οι νευρώνες της πρώτης στιβάδας συνδέονται με όλους τους νευρώνες της επόμενης στιβάδας. Ωστόσο, υπάρχει η δυνατότητα της μερικής σύνδεσης εντός των δικτύων, ή της έμμεσης σύνδεσης του νευρώνα μιας στιβάδας με κάποιο νευρώνα της μεθεπόμενης [69].

2. **Αναδρομικά δίκτυα** (Recurrent), τα οποία περιέχουν και συνδέσεις ανατροφοδότησης (feedback,  ΚΕΦ. 2, Θ).

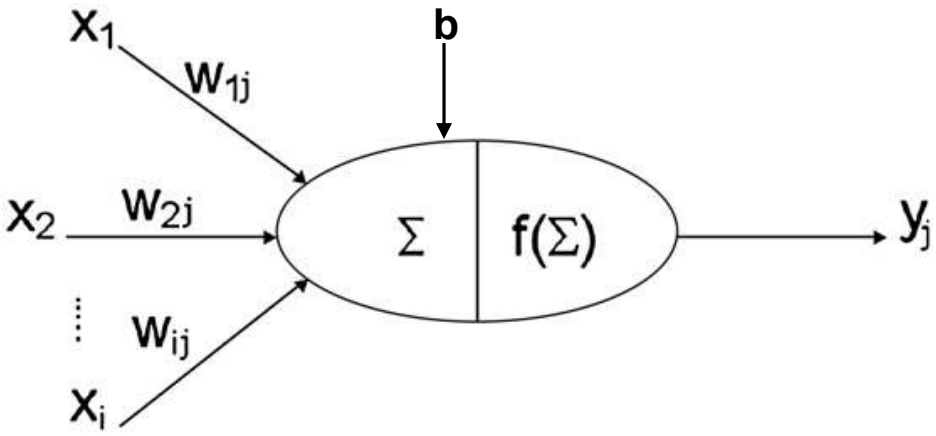
Τα γνωρίσματα των ANN, στα οποία πρέπει να δοθεί εξ' αρχής έμφαση είναι (σχ. 4.9):

1. Μια ομάδα μονάδων επεξεργασίας (νευρώνες)
2. Μια ομάδα **εισερχόμενων δειγμάτων** με τη μορφή διανύσματος ((x_1, x_2, …….x_i), βλ. παρακάτω).
3. Συνδέσεις για κάθε μονάδα (κάθε σύνδεση καθορίζεται από **βάρη** w_{ij} μεταξύ των εισερχομένων μονάδων i και των εξερχομένων ή επόμενων μονάδων/νευρώνων j).
4. Ένα εξωτερικό εισερχόμενο b (προκατάληψη) για κάθε μονάδα.
5. **Εξερχόμενα** δεδομένα y_j, ένα για κάθε μονάδα επεξεργασίας.
6. Ένα **κανόνα "διάδοσης ή εισόδου**", ο οποίος καθορίζει (αποφασίζει/διαχωρίζει) τα πραγματικά εισερχόμενα δεδομένα Σ από τα εξωτερικά εισερχόμενα δεδομένα/ερεθίσματα x_i. Ο κανόνας διάδοσης υπολογίζει την τιμή ενεργοποίησης ("activation value"), η οποία είναι αυτή που θα εισέλθει στον επόμενο νευρώνα (αφού προστεθούν τα γινόμενα βαρών και εισερχόμενων τιμών και αφαιρεθεί τυχόν υπάρχον κατώφλι ενεργοποίησης).
7. Μια **συνάρτηση "ενεργοποίησης"** f (συνήθως σιγμοειδή, για τις ενδιάμεσες στιβάδες ή γραμμική για την εξωτερική [57, 70]), η οποία καθορίζει τη σχέση ανάμεσα στο άθροισμα των εισερχομένων: $\Sigma = \sum_{i}^{n} x_i w_i + b$ και το εξερχόμενο y_j της κάθε μονάδας.
8. Μια **μέθοδο** (αλγόριθμο) για τη διαρκή ενημέρωση του μοντέλου και αξιολόγηση της πληροφορίας [59].

Το σχήμα 4.10 απεικονίζει τα όσα περιγράφηκαν παραπάνω για το νευρώνα ενός ANN.

Όλα τα εισερχόμενα δείγματα πρέπει να "παρουσιάζονται" στο δίκτυο με την ίδια μορφή, δηλαδή ως διανύσματα με διαστάσεις ίσες με τον αριθμό των μεταβλητών που τα χαρακτηρίζουν (πχ για το δείγμα X_n: X_{1n}, X_{2n}, ……..X_{in}). Η κάθε μεταβλητή μπορεί και συνήθως έτσι γίνεται, να "αντιπροσωπεύεται" από τον ίδιο εισερχόμενο νευρώνα (εντοπισμένη αντιπροσώπευση, localized representation) όπως για παράδειγμα στα MLP (§ 4.1.4), στα δίκτυα RBF (§ 4.3.12) και στα δίκτυα Kohonen (§ 4.3.13). Ωστόσο, υφίσταται και η εναλλακτική της παρουσίασης μιας μεταβλητής σε περισσότερους νευρώνες (distributed representation) [52], όπως για

παράδειγμα σε εργασία του Aoyama et al. [71], όπου 6 εισερχόμενες μονάδες αντιπροσωπεύουν τις 6 μεταβλητές και άλλες 6 τα τετράγωνα τους.



Σχήμα 4.10: Απεικόνιση των βασικών χαρακτηριστικών ενός feed-forward ANN.

Η προκατάληψη της κάθε μονάδας, παρέχει ένα σταθερό όρο στο άθροισμα των βαρών, γεγονός που αυξάνει την ευελιξία και προσαρμοστικότητα του δικτύου [1]. Έτσι, διατίθεται ένας επιπλέον βαθμός ελευθερίας όταν γίνεται προσπάθεια ελαχιστοποίησης του σφάλματος μεταξύ πραγματικής και θεωρητικής τιμής των εξερχόμενων. Η χρήση της προκατάληψης b, στην εξίσωση του αθροίσματος των εισερχομένων, είναι ανάλογη με την τομή επί της αρχής (intercept) στη γραμμική συσχέτιση [72]. Επιπλέον, αποφεύγεται η λήψη των ίδιων σημάτων από τους επόμενους νευρώνες όταν τα εισερχόμενα αθροίσματα είναι μηδέν (0) [58]. Ωστόσο, χρειάζεται να γίνει ένας διαχωρισμός της προκατάληψης σε σχέση με το κατώφλι (threshold) που εναλλακτικά χρησιμοποιείται. Το τελευταίο αφορά τον τύπο των συναρτήσεων ενεργοποίησης που χρησιμοποιούνται (συναρτήσεις κατωφλίου ή threshold functions, βλ. § 4.3.4) και έχει σταθερή, μη προσαρμόσιμη τιμή σε αντίθεση με το προκατάληψη [73]. Η προκατάληψη "εκπαιδεύεται", δηλαδή η τιμή της βελτιστοποιείται μαζί με των υπολοίπων βαρών κατά τη διάρκεια της εκπαίδευσης του δικτύου (βλ. § 4.3.1).

Η βασική ιδέα του αλγορίθμου, είναι ότι τα σφάλματα που προκύπτουν (διαφορές δηλαδή μεταξύ πραγματικής και θεωρητικής απόκρισης), "διαδίδονται" (κατανέμονται) προς τα πίσω κατά μήκος των μονάδων (νευρώνων) των ενδιάμεσων (κρυφών) στιβάδων. Αυτός είναι ο σημαντικότερος και ευρύτερα χρησιμοποιούμενος αλγόριθμος (back-propagation, BP κανόνας εκπαίδευσης ή αλγόριθμος) που θα δούμε παρακάτω (§ 4.3.9).

4.2.2. Νευρωνικά Δίκτυα Ταξινόμησης

Τα Νευρωνικά Δίκτυα χρησιμοποιούνται σήμερα σε πολλές επιστήμες (ιατρική, μηχανική, φυσική, χημεία) για ένα πλήθος εφαρμογών (συσχέτιση, μοντελισμός, ταξινόμηση - ομαδοποίηση, δημιουργία χρονοσειρών). Στην παρούσα εργασία ωστόσο, θα εξεταστούν μόνο δεδομένα ομαδοποίησης/ταξινόμησης (clustering/classification) και τελικά η δυνατότητα κατασκευής μοντέλων. Η ταξινόμηση ορίζεται ως η απόδοση ενός δείγματος, σε μια κατηγορία ανάλογα με τις τιμές κάποιων δεικτών (μεταβλητών) που μετρούνται σε αυτό [56]. Ο μοντελισμός εξάλλου, παρότι τυπικά απαιτεί αριθμητικό αποτέλεσμα, δεν φαίνεται να διαφοροποιείται ουσιαστικά από την ταξινόμηση. Τα όρια ανάμεσα στις δυο πορείες δεν είναι αδιαπέραστα [1]: έτσι αν αντί της κατηγορίας των δειγμάτων μπορούμε να αποκομίσουμε αριθμητική τιμή για κάποια ιδιότητα αυτών, το μοντέλο γίνεται άμεσα αριθμητικό.

Σε ένα πρόβλημα ταξινόμησης το επιθυμητό αποτέλεσμα είναι "ονομαστικό" (nominal), δηλαδή μια μεταβλητή κατηγοροποίησης. Για να μετατραπούν ωστόσο, τα αριθμητικά αποτελέσματα σε ονομαστικό αποτέλεσμα, χρησιμοποιούνται συναρτήσεις (post-processing), που συγκρίνουν τις εξερχόμενες τιμές στις εξωτερικές μονάδες (νευρώνες) με κάποια κατώφλια ταξινόμησης (classification thresholds) και έτσι καθορίζουν την ομάδα [17].

Σε ένα πρόβλημα **δυο ομάδων/επιπέδων** (δυαδική ταξινόμηση, binary ή two-state classification), η εξωτερική μονάδα είναι μοναδική (αν και αναφέρονται περιπτώσεις με δυο εξωτερικές μονάδες, χωρίς να υπάρξει ιδιαίτερο όφελος, [1]), με ένα υψηλό εξερχόμενο (τιμή απόκρισης) για τη μια ομάδα και ένα χαμηλό για την άλλη. Μια μεταβλητή δυο ομάδων εύκολα μετατρέπεται σε αριθμητική (πχ Υλίκη = 0, Μόρνος = 1).

Όσον αφορά τα κατώφλια ταξινόμησης, υπάρχουν δυο εναλλακτικές για ένα ερευνητή [17]:

1. Να τεθούν αρχικά δυο κατώφλια ταξινόμησης: **αποδοχής** (accept) και **απόρριψης** (reject). Αν η εξερχόμενη τιμή είναι μεγαλύτερη από το κατώφλι αποδοχής, το σημείο (δείγμα) αποδίδεται στη δεύτερη ομάδα. Αν η εξερχόμενη τιμή είναι μικρότερη από το κατώφλι απόρριψης, το σημείο (δείγμα) αποδίδεται στην πρώτη ομάδα. Αν η εξερχόμενη τιμή κυμαίνεται μεταξύ των δυο κατωφλιών, η πρόβλεψη είναι: **"άγνωστο"** (μη αναγνωρίσιμο), υπονοώντας ότι η ταξινόμηση είναι **αμφίβολη** (doubt classification) και το συγκεκριμένο σημείο βρίσκεται σε περιοχή που οι δυο ομάδες αλληλεπικαλύπτονται.
2. Να δοθεί μία μόνο τιμή για το κατώφλι ταξινόμησης, με την έννοια ότι τα δυο κατώφλια (αποδοχής και απόρριψης) ισούται. Τότε δεν υπάρχει περίπτωση άγνωστων δειγμάτων. Με την αλλαγή του κατωφλίου αυτού εξάλλου (δοκιμές διαφόρων τιμών), μπορεί να επιτευχθεί η βέλτιστη ταξινόμηση με το μικρότερο σφάλμα (misclassification error).

Στην περίπτωση αυτή, γίνεται προσπάθεια εξισορρόπησης της ευαισθησίας και εξειδίκευσης στις δυο ομάδες (βλ. § 4.4.2, 6.3.4).

Η συνήθης διαμόρφωση για τα κατώφλια ταξινόμησης είναι: αποδοχής = 0,95 και απόρριψης = 0,05 (εναλλακτική 1) ή αποδοχής = απόρριψης = 0,5 (εναλλακτική 2). Αυτό σημαίνει 95 % στάθμη εμπιστοσύνης για την αποδοχή μιας τιμής για την πρώτη περίπτωση και κανένα αμφίβολο δείγμα για τη δεύτερη. Και στις δυο περιπτώσεις ωστόσο, είναι φανερό ότι χρησιμοποιείται συνάρτηση με εξερχόμενο στο εύρος 0 – 1 (βλ. § 4.3.4) [17].

Σε ένα πρόβλημα **πολλών ομάδων/επιπέδων** (many-state classification), υπάρχουν συνήθως τόσες εξωτερικές μονάδες, όσες και οι αρχικές ομάδες. Περίπτωση μίας εξωτερικής μονάδας για ένα πρόβλημα n (n > 2) ομάδων έχει αναφερθεί [74], αλλά το δίκτυο τότε φαίνεται να υστερεί καθώς ένα εύρος αποκρίσεων από 0 ως 1 πρέπει να διαιρεθεί σε τόσα διαστήματα (n) όσα και οι ομάδες. Τα διαστήματα αυτά είναι διαδοχικά με αποτέλεσμα δυο ομάδες που αντιπροσωπεύονται από διαδοχικά διαστήματα να θεωρούνται "πιο όμοιες" από δυο άλλες που αντιπροσωπεύονται από τυχαία διαστήματα. Η παραδοχή αυτή δεν ευσταθεί και δεν δικαιολογείται για πραγματικές εφαρμογές [74].

Στην περίπτωση των πολλών εξωτερικών μονάδων, το εισερχόμενο δείγμα αποδίδεται στη μονάδα (άρα και ομάδα) με τη μεγαλύτερη απόκριση [17]. Η απόκριση αυτή μπορεί να ερμηνευτεί και ως την πιθανότητα να ανήκει το δείγμα στη συγκεκριμένη ομάδα. Εξάλλου, "ένα δείγμα πρέπει να αποδίδεται στην ομάδα για την οποία παρουσιάζει τη μέγιστη πιθανότητα" (posterior probability, Bayes' rule) [14]. Αν εναλλακτικά τεθούν κατώφλια ταξινόμησης, το νέο δείγμα επίσης αποδίδεται στην ομάδα με τη μεγαλύτερη απόκριση, αλλά τώρα υπάρχει περίπτωση αμφίβολης ταξινόμησης. Οι μεταβλητές πολλών ομάδων διαχειρίζονται δυσκολότερα από τις αντίστοιχες των δυο ομάδων. Θα μπορούσαν να αναπαρίστανται αριθμητικά (πχ Υλίκη = 0, Μόρνος = 1 και Μαραθώνας = 2), αλλά αυτό υποδηλώνει έμμεσα ότι ο Μόρνος κυμαίνεται κατά κάποιο τρόπο μεταξύ Υλίκης και Μαραθώνα (βλ. παραπάνω). Καλύτερη προσέγγιση είναι η **κωδικοποίηση τύπου ένα-από-N** (one-of-N encoding ή distributed representation) [1, 17, 58]. Με αυτό τον τρόπο, μία ονομαστική μεταβλητή αντιπροσωπεύεται από ένα σύνολο αριθμητικών μεταβλητών. Ο αριθμός των μεταβλητών αυτών ισούται με τον αριθμό των αντίστοιχων ομάδων (πχ Υλίκη = (1, 0, 0), Μόρνος = (0, 1, 0) και Μαραθώνας = (0, 0, 1)). Έτσι, κατά τη διάρκεια της εκπαίδευσης μία από τις εξωτερικές μονάδες είναι ON και οι άλλες OFF [17] ή αλλιώς μία εξωτερική μονάδα δίνει εξερχόμενο σήμα 1 (ή κοντά σε αυτό) και οι υπόλοιπες 0 (ή κοντά σε αυτό) [74].

Η ικανότητα ενός μοντέλου στην ταξινόμηση δειγμάτων μπορεί να αξιολογηθεί με διαφορετικούς τρόπους:

1. Με την **ικανότητα αναγνώρισης** (recognition ability), δηλαδή το ποσοστό των δειγμάτων της ομάδας εκπαίδευσης (βλ. § 4.3.1), που ταξινομούνται σωστά στο στάδιο της δημιουργίας του μοντέλου.
2. Με την **ικανότητα πρόβλεψης** (prediction ability), δηλαδή το ποσοστό των δειγμάτων της ανεξάρτητης ομάδας ελέγχου (βλ. § 4.3.1), που ταξινομούνται σωστά από το μοντέλο.
3. Με την **ικανότητα ταξινόμησης** (classification ability), δηλαδή το ποσοστό των δειγμάτων των ομάδων εκπαίδευσης και ελέγχου, που ταξινομούνται σωστά από το μοντέλο [26].

Τέλος η ικανότητα των δικτύων ταξινόμησης φαίνεται να επηρεάζεται από τον αριθμό των δειγμάτων σε κάθε ομάδα, με τους ερευνητές να επιζητούν ίσο αριθμό δειγμάτων σε κάθε ομάδα για αυξημένη ακρίβεια [75].

4.3. ΘΕΩΡΙΑ

4.3.1. "Εκπαιδεύοντας" τα Νευρωνικά Δίκτυα

Η χρήση των ANN, υπονοεί άμεσα ότι υπάρχει μια σχέση μεταξύ των εισερχομένων και εξερχομένων, της οποίας η φύση τουλάχιστον είναι γνωστή. Αυτή η σχέση αποκαλύπτεται κατά τη διάρκεια μιας πορείας που είναι γνωστή ως **"εκπαίδευση"** ή **"εκμάθηση"** και είναι ανάλογη με την εκτίμηση των παραμέτρων στις παραδοσιακές στατιστικές τεχνικές [76].

Οι πολυπαραμετρικές αυτές τεχνικές (§ 1.3), όπως και τα ANN, χωρίζονται σε δυο βασικές κατηγορίες: **στις επιβλεπόμενες τεχνικές** (supervised) και τις **μη επιβλεπόμενες** (unsupervised) [17].

Στις επιβλεπόμενες τεχνικές ANN, υπάρχει ή ανάγκη ενός "δασκάλου" ο οποίος χρησιμοποιεί μια ομάδα παραδειγμάτων/δειγμάτων (cases ή pattern ή objects), για την "εκπαίδευση" του νέου μοντέλου. Η ομάδα αυτή (εκπαίδευσης ή εκμάθησης), περιέχει εισερχόμενα δεδομένα συνδυασμένα με τις εξερχόμενες αποκρίσεις (outputs), έτσι ώστε να "χτιστεί" μια σχέση μεταξύ τους. Οι εξερχόμενες αποκρίσεις είναι ήδη γνωστές από την αρχή και τα Νευρωνικά Δίκτυα υπολογίζουν και διαμορφώνουν κατάλληλα τα βάρη, στη διάρκεια μιας σειράς **κύκλων ή "εποχών"/"περιόδων"/"προσπαθειών"** (epochs ή iterations), ώστε οι θεωρητικές και οι υπολογιζόμενες αποκρίσεις να είναι όσον το δυνατό πλησιέστερες [52, 77, 78]. Αυτό σημαίνει ότι το σφάλμα πρόβλεψης στην ομάδα εκπαίδευσης ελαχιστοποιείται με την κατάλληλη διαμόρφωση των βαρών. Αν το μοντέλο εκπαιδεύεται σωστά, θεωρείται ότι "γνωρίζει" πλέον τη σχέση μεταξύ εισερχομένων και εξερχομένων και μπορεί να προβλέψει τις εξερ-

χόμενες αποκρίσεις για άγνωστα δείγματα (με δεδομένα τα εισερχόμενα), με τη βοήθεια μιας άγνωστης για μας συνάρτησης. Το πρόβλημα βέβαια είναι ότι κατά τη διάρκεια της εκπαίδευσης, δεν ελαχιστοποιείται το σφάλμα που πραγματικά μας ενδιαφέρει, δηλαδή στην άγνωστη ομάδα δειγμάτων αλλά στην ομάδα εκπαίδευσης [17].

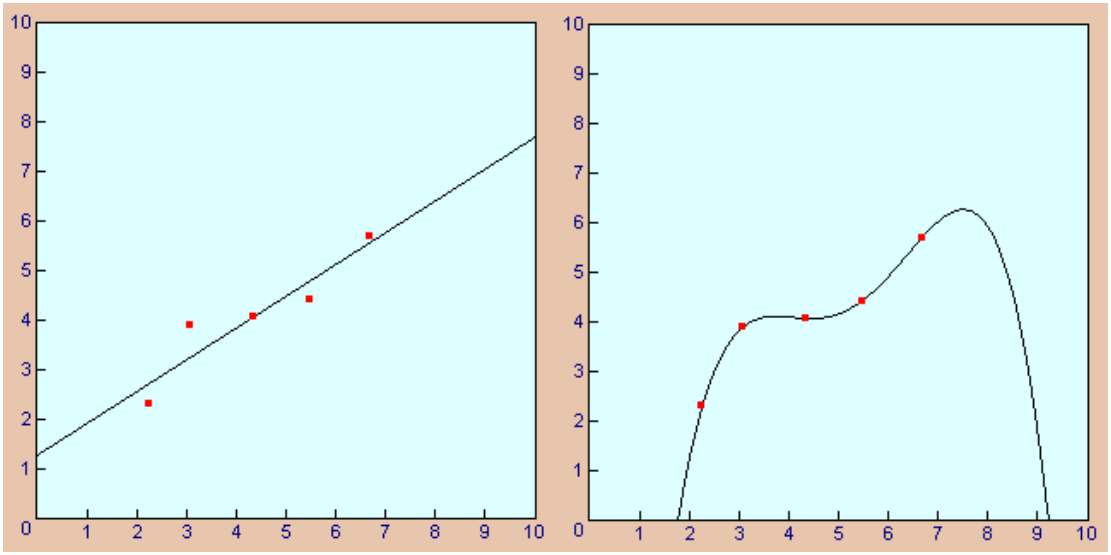
Στις μη επιβλεπόμενες τεχνικές, το μοντέλο "εφοδιάζεται" με μια ομάδα δειγμάτων, και το σύστημα αφήνεται να "κατασταλάξει" ή όχι, χωρίς τη γνώση (ή την καλύτερα τη χρήση της γνώσης) επιθυμητού εξερχομένου [77] ή κάποιων εγγενών ιδιοτήτων [79]. Δηλαδή, στη μη επιβλεπόμενη εκπαίδευση η γνώση της ομάδας **δεν χρησιμοποιείται στην εκπαίδευση του μοντέλου ή τη διόρθωση των βαρών** [52]. Το μοντέλο πειραματίζεται στο να μάθει τη δομή των δεδομένων, αναγνωρίζοντας συστάδες (clusters) σε αυτά (βλ. για παράδειγμα τα δίκτυα Kohonen, § 4.3.13).

Τόσο ένα επιβλεπόμενο όσο και ένα μη επιβλεπόμενο μοντέλο με πολλά βάρη (εισερχόμενες μεταβλητές ή/και ενδιάμεσες μονάδες ή στιβάδες), περισσότερα από όσα είναι αναγκαία στην πραγματικότητα, ή διαφορετικά με μικρό λόγο αριθμού δειγμάτων εκπαίδευσης και βαρών [76, 80, 81] "χτίζει" μια πολύπλοκη συνάρτηση και είναι επιρρεπές στην **υπερ-προσαρμογή** (over-fitting ή over-learning ή "over-training" ή "over-specificity"). Έτσι το μοντέλο μπορεί να "απομνημονεύσει" την ομάδα εκπαίδευσης και σαν αποτέλεσμα να είναι λιγότερο ικανό να γενικεύσει σε άλλα ζεύγη εισερχομένων - εξερχομένων [17].

Το πρόβλημα μπορεί να περιγραφεί με τη βοήθεια μιας πολυωνυμικής συνάρτησης. Για μια δεδομένη ομάδα δειγμάτων, μπορεί να χρειαστούμε μια πολυωνυμική καμπύλη (μοντέλο) για να ερμηνεύσουμε ή να προσομοιάσουμε τα δεδομένα. Προφανώς επίσης, υπάρχει θόρυβος στα δεδομένα και έτσι δεν περιμένουμε απαραίτητα η βέλτιστη καμπύλη να "περάσει" από όλα τα σημεία. Ένα μικρότερης δύναμης πολυώνυμο, μπορεί να μην είναι εξίσου ευέλικτο για να περάσει από όλα τα σημεία, ενώ αντίθετα, ένα μεγαλύτερης δύναμης είναι στην πραγματικότητα τόσο ευέλικτο ή προσαρμοστικό, ώστε "προσεγγίζει" ακριβώς τα σημεία, υιοθετώντας ένα τόσο παράδοξο σχήμα, το οποίο μπορεί τελικά να μη σχετίζεται με την πραγματική συνάρτηση που περιγράφει τη σχέση των δεδομένων [17] (ΚΕΦ. 2, Θ).

Το φαινόμενο της υπερ-προσαρμογής μπορεί να ερμηνευτεί ως μια συνέπεια του **"πλεονασμού παραμέτρων"** (parameters redundancy), δηλαδή το σύστημα έχει πολύ περισσότερες παραμέτρους, από όσες χρειάζεται για τη λύση του προβλήματος. Αυτό για ένα πολυώνυμο σημαίνει ότι έχει πολλούς όρους: έτσι προσαρμόζεται ακριβώς στα δεδομένα, αντί να ομαλοποιεί τις ταλαντώσεις που προέρχονται από το θόρυβο αυτών [1]. Για παράδειγμα, υποθέτουμε ότι προσαρμόζουμε πέντε σημεία (κόκκινους κύκλους στο σχήμα 4.11(β)) στο μοντέλο του 4-βάθμιου πολυωνύμου $ax^4+bx^3+cx^2+dx+e$, αντί της ευθείας ax+b (σχ. 4.11(α)).

Παρόλο που η τελευταία δεν μπορεί να περάσει από όλα τα σημεία αυτά, ωστόσο είναι πιθανόν να κάνει καλύτερη πρόβλεψη για τα νέα σημεία [82].



(α) (β)

Σχήμα 4.11: Στην κατασκευή του μοντέλου, οι περισσότερες παράμετροι δεν σημαίνουν πάντα καλύτερα αποτελέσματα. Η καμπύλη των πέντε σταθερών (a, b, c, d, e) προσαρμόζεται ακριβώς στα αρχικά δεδομένα (κόκκινοι κύκλοι) (β) καλύτερα από τη γραμμική συνάρτηση (α), αλλά η "πρόβλεψη" των νέων σημείων θα είναι χειρότερη [82].

Τα ANN έχουν ακριβώς το ίδιο πρόβλημα: ένα δίκτυο με περισσότερα βάρη, μπορεί να περιγράψει μια πολύπλοκη συνάρτηση (επιπλέον και κάθε άσχετη πληροφορία που εμπεριέχεται στα δεδομένα, αλλά και θόρυβο [76, 83, 84]), και είναι επιρρεπές στην υπερ-προσαρμογή. Η επιλογή είναι ανάμεσα σε ένα επαρκή αριθμών ελεύθερων παραμέτρων (βαρών) ικανών να αντιπροσωπεύσουν τη συνάρτηση και ένα μεγάλο αριθμό ελεύθερων παραμέτρων που μπορεί να οδηγήσουν σε υπερ-προσαρμογή [76]. Αυτό συμβαίνει γιατί τότε τα δείγματα εκπαίδευσης είναι πολύ λίγα για να "εκπαιδεύσουν" επαρκώς, το μεγάλο αριθμό βαρών ή ενδιάμεσων μονάδων [51]. Πραγματικά, ο λόγος του αριθμού των δειγμάτων εκπαίδευσης και των βαρών φαίνεται να παίζει πολύ σημαντικό ρόλο στη δημιουργία ή όχι φαινομένων υπερ-προσαρμογής. Έτσι, πολλοί συγγραφείς (βλ. παρακάτω) ζητούν τα δείγματα εκπαίδευσης να είναι τουλάχιστον περισσότερα από τον αριθμό των βαρών, ενώ όταν ο αριθμός των δειγμάτων, **είναι τουλάχιστον 30 φορές μεγαλύτερος,** δεν φαίνεται να υφίστανται ανάλογα φαινόμενα [76]. Αντίθετα, οι Fernandes και Lona [85] αμφισβητούν έντονα τον παραπάνω κανόνα δηλώνοντας ότι μόνο για τις κλασικές παραδοσιακές μεθόδους απαιτείται ο αριθμός των δειγμάτων εκπαίδευσης να είναι μεγαλύτερος του αριθμού των βαρών. Έτσι, συνιστούν τη χρήση τόσων δειγμάτων εκπαίδευσης ώστε ο αριθμός τους να ισούται με το 20πλάσιο

του γινομένου εισερχομένων και εξερχομένων. Για μεγαλύτερο αριθμό δειγμάτων, η απόδοση του δικτύου αποσταθεροποιείται ("insensitive to too many data"). Χωρίζουν δε τα μοντέλα ANN σε τρεις ομάδες ανάλογα με την αναλογία των αριθμών εισερχομένων και εξερχομένων και προτείνουν πρακτικούς κανόνες για τον αριθμό των ενδιάμεσων στιβάδων και νευρώνων ( ΚΕΦ. 2, Θ). Από την άλλη μεριά, οι Zupan και Gasteiger [1] συμβουλεύουν τη χρήση άλλων τεχνικών όταν δεν υπάρχουν αρκετά δεδομένα προς εκπαίδευση του μοντέλου, ενώ ο Kim [43] αποδεικνύει πειραματικά την καλύτερη απόδοση των ANN ειδικά στην περίπτωση πολυπλοκότερων προβλημάτων (περισσότερα δείγματα, ομάδες ταξινόμησης, ανεξάρτητες μεταβλητές συνεχείς και κατηγοροποίησης). Στην πραγματικότητα, ακόμα και ένα γραμμικό μοντέλο μπορεί να οδηγηθεί στην υπερ-προσαρμογή σε ένα χώρο πολλών διαστάσεων [86]. Στην περίπτωση των λίγων δειγμάτων εκπαίδευσης, το μικρότερο σφάλμα πρόβλεψης που παρατηρείται μπορεί να οφείλεται στην υπερ-προσαρμογή και όχι στην εύρεση του καλύτερου μοντέλου. Η δυνατότητά του για **"γενίκευση**" (generalization) είναι περιορισμένη, με αποτέλεσμα να μην μπορεί να ανταποκριθεί σε νέα δείγματα, εκτός της ομάδας εκπαίδευσης. Αντίθετα, ένα δίκτυο με λιγότερα βάρη, μπορεί να μην είναι ικανό να περιγράψει την υπάρχουσα συνάρτηση. Για παράδειγμα, ένα δίκτυο με μηδενικές ενδιάμεσες στιβάδες μπορεί απλά να απεικονίσει μια γραμμική συνάρτηση (βλ. επίσης σχ. 4.8, § 4.1.4).

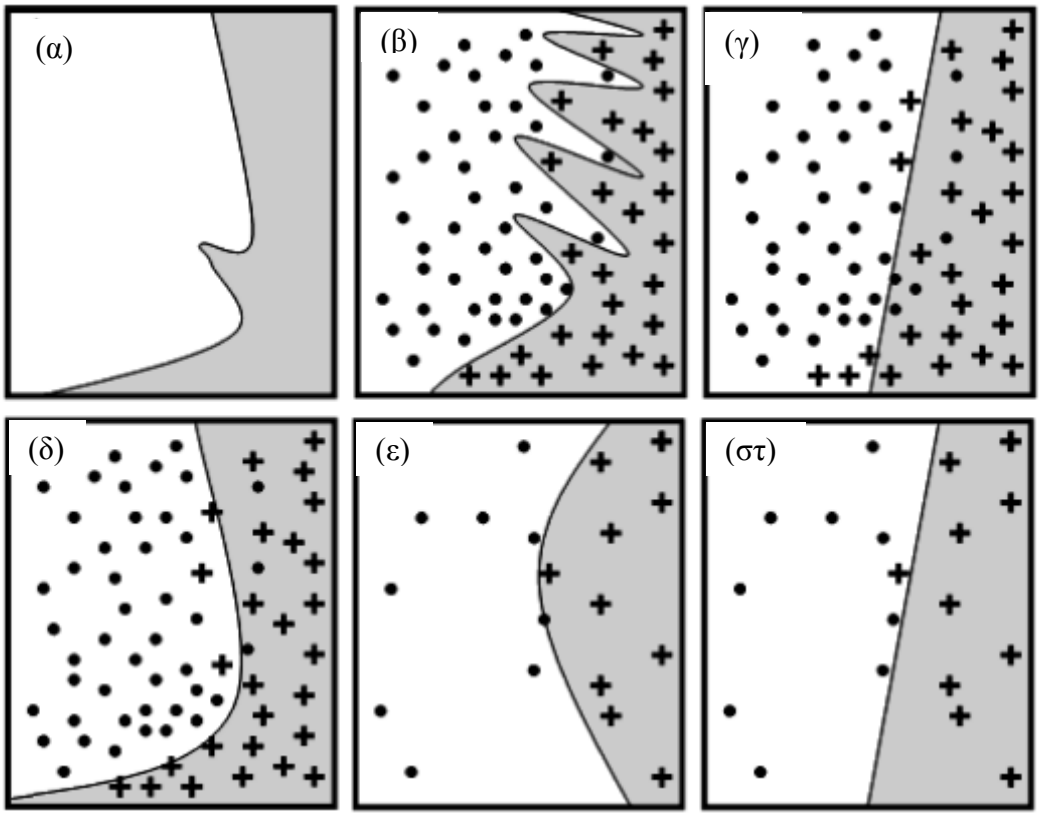
Συμπερασματικά λοιπόν, φαινόμενα υπερ-προσαρμογής εμφανίζονται συχνά σε δίκτυα με λίγα δείγματα. Ωστόσο, δίκτυα με μεγάλη ομάδα εκπαίδευσης είναι επίσης επιρρεπή σε φαινόμενα υπερ-προσαρμογής όταν οι μεταβλητές είναι επίσης πολλές και η πραγματική συνάρτηση πολύπλοκη [87].

Οι μέθοδοι που εφαρμόζονται για την αποφυγή φαινομένων υπερ-προσαρμογής στα ANN μοντέλα περιγράφονται συνολικά με τον όρο **"ομαλοποίηση"** ("regularization" [17, 86]. Ο σκοπός εδώ είναι τα δεδομένα να λαμβάνονται λιγότερο υπόψη από το μοντέλο (μικρότερη διακύμανση, βλ. παρακάτω και § 4.3.16). Περισσότερες λεπτομέρειες για τις μεθόδους ομαλοποίησης, αναφέρονται στο παράρτημα της διατριβής, ( ΚΕΦ. 2, Θ). Εδώ θα αναφερθεί παρακάτω, μόνο η **"μέθοδος του πρώιμου τερματισμού"** ("early stopping method"), που αποτελεί την πιο δημοφιλή και επιπλέον χρησιμοποιείται στην παρούσα εργασία.

Γενικότερα, συγγραφείς και ερευνητές ανά τον κόσμο, ενθαρρύνουν τη χρήση λιγότερων βαρών, νευρώνων (μονάδων) ή μεταβλητών για την αποφυγή προβλημάτων υπερ-προσαρμογής και συνεπώς καλύτερα αποτελέσματα στα νέα άγνωστα δείγματα [1, 17, 26, 57, 76, 86, 88 - 94]. Οι **"συμπαγείς δομές "** (compact ή simple) θεωρούνται σαφώς ότι πλεονεκτούν σε σχέση με τις πολυπλοκότερες [1, 17, 92 - 94]. Υπάρχουν αλγόριθμοι οι οποίοι μετά το τέλος της εκπαίδευσης "κόβουν" τους νευρώνες με πολύ μικρά βάρη (σε αντιδιαστολή με τα μεγάλα βάρη που αναφέρθηκαν παραπάνω), ή νευρώνες που έχουν παρόμοιες ή σχεδόν

σταθερές εξερχόμενες τιμές, καθώς η συνεισφορά τους στο επίπεδο του επόμενου νευρώνα είναι μάλλον ελάχιστη, ώστε τελικά η δομή του δικτύου να γίνει πιο συμπαγής [17, 52, 95]. Γενικά επίσης, προτείνονται μέθοδοι για την αποτελεσματική μείωση των μεταβλητών (βλ. § 4.4.3 και  ΚΕΦ. 2, Θ).

Σε πολλές από τις βιβλιογραφικές προσεγγίσεις, διαφαίνεται η σχέση ανάμεσα στην πολυπλοκότητα του μοντέλου και τα συνολικά διαθέσιμα δείγματα. Έτσι, ο Grzesiak et al. [96] κάνει αναφορές σε μελέτες που θέλουν τον αριθμό των δειγμάτων εκπαίδευσης n_t να δίνεται από τις σχέσεις: $h \times v \leq n_t \leq 2 \times w_n \times (1+\log N)$ ή $n_t = 2^v$ όπου h ο αριθμός των μονάδων της ενδιάμεσης στιβάδας, N, w_n οι συνολικοί αριθμοί μονάδων και βαρών αντίστοιχα και v ο αριθμός των μεταβλητών. Ένα γραφικό παράδειγμα για δυαδική ταξινόμηση φαίνεται στο σχήμα 4.12 [97].



Σχήμα 4.12: Συσχέτιση ανάμεσα στην πολυπλοκότητα του μοντέλου και τον αριθμό των δειγμάτων [97].

Στο σχήμα αυτό φαίνονται οι δυο ομάδες δειγμάτων και τα προτεινόμενα όρια μεταξύ αυτών. Τα πραγματικά όρια φαίνονται στο σχήμα 4.12(α), τα οποία αρχικά είναι άγνωστα. Ο έλεγχος της πολυπλοκότητας είναι βασικός για την αποφυγή φαινομένων υπερ-προσαρμογής (σχ. 4.12(β)) ή αντιστρόφως την κατασκευή υπερ-απλουστευμένων μοντέλων (σχ. 4.12(γ)). Μια καλή προσαρμογή, ακριβής για τη βάση των δειγμάτων που διατίθεται, φαίνεται στο

σχήμα 4.12(δ). Με λιγότερα δείγματα, ο έλεγχος της πολυπλοκότητας θα μπορούσε να είναι παραπλανητικός (σχ. 4.12(ε)), ενώ αντίθετα μια απλή λύση θα ήταν πιο ακριβής ((σχ. 4.12(στ)). Το τελευταίο εξηγείται από το γεγονός ότι για περισσότερα δείγματα, είναι ασφαλέστερο να αναζητούνται πολυπλοκότερα μοντέλα, αλλά και αντίστροφα, αν τα όριο μεταξύ των ομάδων είναι πράγματι πολύπλοκα, δεν είναι δυνατό να βρεθεί ασφαλές μοντέλο χωρίς πολλά δείγματα [97].

Μια γενικότερη προσέγγιση επιχειρεί ο Stathakis [92], ο οποίος διευκρινίζει επιπλέον το καθεστώς **της μίας μόνο ενδιάμεσης στιβάδας** όπως έχει αυτό επικρατήσει. Σύμφωνα λοιπόν με τον Hecht-Nielsen [98], οποιαδήποτε συνεχής συνάρτηση μπορεί να προσεγγιστεί από ANN με μία ενδιάμεση στιβάδα και 2×v+1 μονάδες. Εσφαλμένα ωστόσο, θεωρήθηκε πανάκεια η λύση της μίας στιβάδας με τον αντίστοιχο αριθμό μονάδων, καθώς ο Hecht-Nielsen αναφερόταν σε μια πιο πολύπλοκη συνάρτηση σε σχέση με τη συνήθη σιγμοειδή που χρησιμοποιείται ως συνάρτηση ενεργοποίησης (§ 4.3.4). Δυο ενδιάμεσες στιβάδες ωστόσο, είναι ικανές να εξισορροπήσουν τη διαφορά που προκύπτει από τη χρήση απλούστερων συναρτήσεων με μικρότερο μάλιστα αριθμό μονάδων και μεγαλύτερη ευελιξία [76, 92]. Έτσι, προτείνονται λύσεις όπως η χρήση μίας στιβάδας με σιγμοειδή συνάρτηση και μεγάλο αριθμό μονάδων [6], ίσο για παράδειγμα με τον αριθμό των δειγμάτων εκπαίδευσης n_t, ή εναλλακτικά δυο στιβάδων με αριθμό μονάδων $\sqrt{(m+2)\times n_t}+2\sqrt{n_t/(m+2)}$ και $m\sqrt{n_t/(m+2)}$ για την πρώτη και δεύτερη στιβάδα αντίστοιχα, όπου m ο αριθμός των εξωτερικών νευρώνων. Ωστόσο όπως αναφέρθηκε παραπάνω, εξαιτίας φαινομένων υπερ-προσαρμογής, οι παραπάνω σχέσεις αφορούν το μέγιστο προτεινόμενο αριθμό στιβάδων και όχι την πραγματική λύση [92]. Το ίδιο εξάλλου υποστήριζε και ο Hornik et al. [99], δυο δεκαετίες νωρίτερα αποδεικνύοντας ότι "οι MPL δομές μπορούν με τη χρήση μιας οποιασδήποτε squash συνάρτησης ενεργοποίησης να προσεγγίσουν ουσιαστικά οποιαδήποτε συνάρτηση με τον επιθυμητό βαθμό ακρίβειας, εφόσον υπάρχουν αρκετές διαθέσιμες ενδιάμεσες μονάδες. Αυτό σημαίνει ότι τα MLP μοντέλα είναι **"καθολικοί εκτιμητές"** (universal approximators) [85, 80, 99]. Αποτυχία στην εφαρμογή τους μπορεί μόνο να αποδοθεί στον ανεπαρκή αριθμό ενδιάμεσων ομάδων ή στην παρουσία λογικών και όχι νομοτελειακών (μαθηματικών) σχέσεων μεταξύ των δεδομένων" [99].

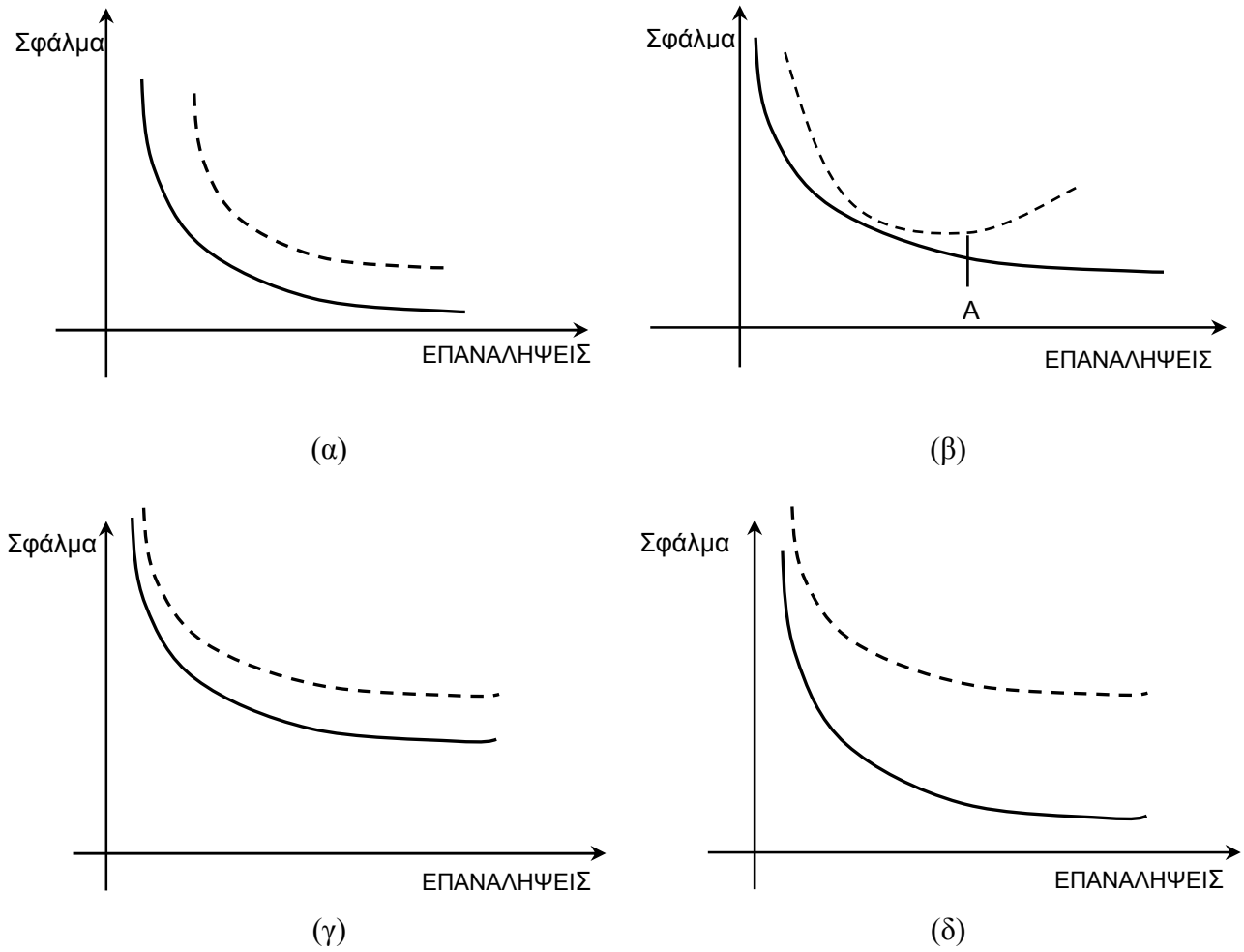
Η καλύτερη κάθε φορά λύση φαίνεται τελικά ότι εξαρτάται από το ίδιο το πρόβλημα και οι παραπάνω κανόνες δεν εξασφαλίζουν τη βέλτιστη πάντα τοπολογία. Γενικότερα, η βέλτιστη γεωμετρία δείχνει να είναι συνυφασμένη με τα μικρότερα δίκτυα που μπορούν να "συλλάβουν" τη σχέση ανάμεσα στα δείγματα της ομάδας εκπαίδευσης. Οι δοκιμές κυρίως ("trial and error" [1, 20, 49, 63, 69, 84, 85]) αλλά και σχετικά προσφάτως αναπτυχθέντες αλγόριθμοι φαίνεται γενικά να επιτυγχάνουν καλές λύσεις [76, 80, 92].

Μια γενικότερη επίσης αποδεκτή λύση στο πρόβλημα της υπερ-προσαρμογής των δικτύων, είναι η χρήση μιας άλλης ανεξάρτητης ομάδας δειγμάτων, **της ομάδας επικύρωσης** ("selection" ή "validation" ή "calibration" ή "control" ή "monitoring" ή "over-fitting test" ή "evaluation" ή "internal" sample set)**.** Έτσι, κάποια δείγματα **δεν χρησιμοποιούνται** στην ομάδα εκπαίδευσης, αλλά "φυλάσσονται" για ένα ανεξάρτητο έλεγχο του μοντέλου (επομένως δεν υπάρχει **"διασταύρωση"** ("crossing"), όπως στις συνήθεις μεθόδους επικύρωσης (βλ. § 4.4.2). Η επιλογή των δυο ομάδων δειγμάτων πρέπει να είναι αντικειμενική, καθώς έχει μεγάλη επίδραση στη συνολική απόδοση του δικτύου και την ικανότητα γενίκευσης [76, 100] (βλ. παρακάτω). Αρχικά τουλάχιστον και όσο η εκπαίδευση του μοντέλου συνεχίζεται, το σφάλμα πρόβλεψης μειώνεται για την ομάδα εκπαίδευσης, αλλά και την ομάδα επικύρωσης (σχ. 4.13(α)). Ωστόσο, αν το σφάλμα στη δεύτερη ομάδα σταματήσει να μειώνεται ή/και αντίθετα αρχίζει να αυξάνεται, αποδεικνύεται η υπερ-προσαρμογή του μοντέλου (σχ. 4.13(β)) [51, 76]. Τότε, συνίσταται η μείωση των ενδιάμεσων στιβάδων ή των μονάδων αυτών, καθώς το μοντέλο φαίνεται να είναι πιο πολύπλοκο από όσο υπαγορεύει η συνάρτηση συσχέτισης των εισερχόμενων/εξερχομένων δεδομένων. Αντίθετα, αν το δίκτυο δεν είναι ικανό να απεικονίσει τη συνάρτηση αυτή, το σφάλμα πρόβλεψης δεν θα μειωθεί σε καμιά από τις δυο ομάδες δειγμάτων σε ικανοποιητικό επίπεδο και δεν πρόκειται προφανώς να συμβεί υπερ-προσαρμογή. Η παραπάνω μέθοδος σύγκρισης των σφαλμάτων πρόβλεψης των ομάδων εκπαίδευσης και επικύρωσης, η οποία και οδηγεί στην παύση της εκπαίδευσης, ονομάζεται **"μέθοδος του πρώιμου τερματισμού"** ("early stopping method") [86, 101].

Τέλος, το δίκτυο ελέγχεται ως προς την ακρίβεια του, από μια νέα (τρίτη) ομάδα δειγμάτων: **της ομάδας ελέγχου** ("test" ή "unknown" sample set). Τα δείγματα αυτά είναι τελείως ανεξάρτητα και δεν συμμετέχουν στη διαμόρφωση των παραμέτρων του δικτύου. Εξάλλου, θα ήταν τελείως άδικο να ελεγχθεί το μοντέλο με μια ομάδα δειγμάτων (την ομάδα επικύρωσης) για την οποία τερματίστηκε η εκπαίδευση με βάση το σφάλμα που μετρήθηκε σε αυτήν. Η εκτίμηση της ακρίβειας θα ήταν εντελώς μεροληπτική [102, 103].

Το σχήμα 4.13(γ) απεικονίζει τη μορφή ενός **"παραλυμένου" δικτύου** (paralyzed ή premature saturated) (βλ. § 4.3.9). Αυτό σημαίνει, ότι η μείωση του σφάλματος πρόβλεψης για τις ομάδες εκπαίδευσης και επικύρωσης σταματά πολύ νωρίς, ώστε αυτό να είναι πολύ υψηλό για να γίνει αποδεκτό.

Το σχήμα 4.13(δ) απεικονίζει τη συμπεριφορά του δικτύου, όταν οι ομάδες εκπαίδευσης και επικύρωσης αντιπροσωπεύουν διαφορετικά δείγματα, ή όταν υπάρχουν έκτροπες τιμές στην ομάδα επικύρωσης και όχι στην ομάδα εκπαίδευσης (βλ. παρακάτω) [51]. Η ύπαρξη σημαντικών εκτρόπων τιμών στην ομάδα εκπαίδευσης εξάλλου, μπορεί να "συμπιέσει" τα υπόλοιπα δείγματα σε ένα πολύ στενό εύρος τιμών και να δυσκολέψει την εκπαίδευση του δικτύου [58].



Σχήμα 4.13: (α) Ιδανική απεικόνιση δικτύου MLP. Η μαύρη γραμμή αναπαριστά το σφάλμα στην ομάδα εκπαίδευσης, ενώ η διακεκομμένη στην ομάδα επικύρωσης. (β) Υπερ-προσαρμογή δικτύου. (γ) "Παραλυμένο" δίκτυο. (δ) Δίκτυο όπου οι ομάδες εκπαίδευσης και επικύρωσης περιέχουν διαφορετικά δείγματα ή η ομάδα επικύρωσης περιέχει έκτροπες τιμές [51].

Ανακεφαλαιώνοντας τα παραπάνω, στις επιβλεπόμενες και μη ANN τεχνικές, χρησιμοποιούνται και πρέπει να γίνεται έτσι 3 ομάδες δειγμάτων:

1. η **ομάδα εκπαίδευσης** που περιέχει την ομάδα των "πρότυπων" δειγμάτων,
2. η **ομάδα επικύρωσης** που καθορίζει το τέλος της εκπαίδευσης [89] και χρησιμοποιείται στη **βελτιστοποίηση των παραμέτρων** [26] και την **αξιολόγηση του μοντέλου στα διάφορα στάδια της εκπαίδευσης** [76] και
3. η **ομάδα ελέγχου** που ελέγχει την επιτευχθείσα **ικανότητα πρόβλεψης** (§ 4.2.2) [89].

4.3.2. Επιλογή και σύνθεση των ομάδων

"Η επιλογή του **σωστού μεγέθους και των σωστών παραδειγμάτων** για την ομάδα εκπαίδευσης, πρέπει να είναι η πιο προσεκτικά σχεδιασμένη προκατεργασία των ANN" [52].

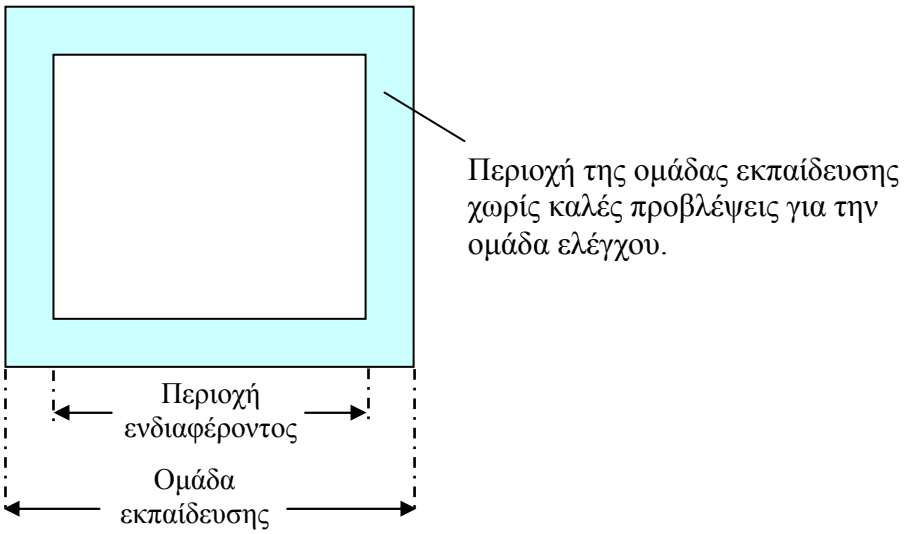
Το μέγεθος της ομάδας εκπαίδευσης είναι πολύ σημαντικό (§ 4.3.10) για την αποτελεσματική εκπαίδευση του δικτύου. Μικρό μέγεθος στην ομάδα εκπαίδευσης καθιστά το δίκτυο ανεπαρκές να αναγνωρίσει τις ομάδες και τα "σύνορα" αυτών, ενώ αντίθετα, ένα μεγάλο δίκτυο οδηγεί σε φαινόμενα υπερ-προσαρμογής και αυξάνει τον υπολογιστικό χρόνο. Γενικά, ο αριθμός των δειγμάτων εκπαίδευσης πρέπει να αντανακλά την πολυπλοκότητα των ομάδων ταξινόμησης [6]. Μερικοί συγγραφείς επιλέγουν τις ομάδες εκπαίδευσης (εδώ μπορεί να συμπεριλαμβάνεται και η ομάδα επικύρωσης) και ελέγχου με βάση κανόνες όπως 90 % (vs 10%), 80 % (vs 20%) ή 70 % (vs 30%) [69]. Ανάλογα ο αριθμός των βαρών μπορεί να κυμαίνεται από μισό εκατομμύριο μέχρι και λιγότερα από 100, γεγονός που επίσης αντανακλά την πολυπλοκότητα του προβλήματος και όχι τελικά την ποιότητα των αποτελεσμάτων [52].

Ο Palani et al. [91] υποστηρίζει ότι η ομάδα επικύρωσης πρέπει να είναι να κυμαίνεται σε ποσοστό 10 – 40 % της ομάδας εκπαίδευσης, η Gramatica θεωρεί ότι τουλάχιστον 20 % των δειγμάτων πρέπει να χρησιμοποιείται για την ομάδα ελέγχου, χωρίς να αποκλείει και ανατροπή των αναλογιών για συγκεκριμένες εφαρμογές (QSAR, Quantitative Structure Activity Relationships) [104], ενώ ο Zupan et al. [74] υποστηρίζει ότι οι ομάδες εκπαίδευσης και ελέγχου πρέπει να έχουν το ίδιο μέγεθος, ή η δεύτερη να είναι ελαφρά μεγαλύτερη (στην πράξη το εφαρμόζει και ο ίδιος χρησιμοποιώντας πολύ μικρές ομάδες εκπαίδευσης: 20-25 % του συνόλου των δειγμάτων [79]). Τέλος ο García-González et al. [105] χρησιμοποιεί την αναλογία 50/25 και 25 % αντίστοιχα για τις ομάδες εκπαίδευσης, επικύρωσης και ελέγχου.

Επιπλέον, η κατανομή των δειγμάτων στις τρεις κατηγορίες δειγμάτων είναι πολύ κρίσιμη στην ανάπτυξη των ANN. Πρέπει να γίνεται με τρόπο αντικειμενικό, καθώς έχει σημαντική επίδραση στην απόδοση του δικτύου [1, 6, 17, 69, 102, 104]. Έτσι, το δίκτυο που μαθαίνει από την ομάδα εκπαίδευσης, θα έχει φτωχή "γενίκευση" αν η ομάδα αυτή δεν είναι αντιπροσωπευτική όλων των δειγμάτων [100], ενώ θα δείχνει προκατάληψη προς την περιοχή που συγκεντρώνονται τα περισσότερα δείγματα [80, 85]. Έτσι, η ομάδα εκπαίδευσης όχι απλά πρέπει να καλύπτει ένα μεγάλο εύρος τιμών, αλλά και να περιέχει όσον το δυνατό διαφορετικές τιμές στις μεταβλητές που την ορίζουν [85]. Μια καλή επιλογή της ομάδας εκπαίδευσης πρέπει να καλύπτει τυπικά αλλά και απρόσμενα σενάρια [106]. Ένα δίκτυο εκπαιδευμένο σε 900 "καλές" και 100 "κακές" περιπτώσεις θα δείξει προκατάληψη στις αποφάσεις υπερ των καλών περιπτώσεων, αφού έτσι θα μειώσει προφανώς το σφάλμα του. Αν ωστόσο, η κατανομή των καλών και κακών στον πραγματικό πληθυσμό είναι διαφορετική, οι αποφάσεις του μοντέλου θα είναι επίσης λανθασμένες. Ένα καλό παράδειγμα είναι η διάγνωση μιας ασθένειας. Ας υποθέσουμε ότι το 90 % των ασθενών που εξετάζονται είναι υγιείς. Το δίκτυο εκπαιδεύεται με ένα ανάλογο δείγμα, αλλά χρησιμοποιείται για ένα ειδικό δείγμα ασθενών με πιθανότητες της ασθένειας 50/50. Το μοντέλο τότε, θα αποτύχει να αναγνωρίσει την ασθένεια σε κάποιους

ασθενείς. Αν αντίθετα, το δίκτυο εκπαιδευτεί στην ειδική αυτή ομάδα ασθενών και δοκιμαστεί σε συνήθη πληθυσμό, το μοντέλο θα "κατασκευάσει" ασθενείς [17]. Για τον ίδιο λόγο ο Zupan et al. [74] προτείνει να υπάρχει ομοιογένεια (ίσος αριθμός δειγμάτων από κάθε ομάδα) στα δείγματα εκπαίδευσης, ενώ οι Grzesiak et al. [96] και ο García-González et al. [105] να υφίσταται η ίδια αναλογία των ομάδων στα δείγματα εκπαίδευσης και ελέγχου. Επίσης, οι Balabin και Lomakina [107] επιλέγουν συνειδητά τα δείγματα με τις μικρότερες ή μεγαλύτερες τιμές εισερχομένων ως ομάδα εκπαίδευσης ώστε να αποκλείσουν την πιθανότητα **συμπερασματικής εξαγωγής με προεκβολή** (extrapolation).

Η δυνατότητα για προεκβολή αποδεικνύεται γενικά επικίνδυνη στην περίπτωση των ANN (βλ. § 4.3.10) [17, 51, 52, 76, 77]. Πράγματι στην περίπτωση αυτή, όταν το μοντέλο αναπόφευκτα κάποια στιγμή "αντιμετωπίσει" τα εξωτερικά δείγματα ελέγχου με απαιτήσεις για καλή απόδοση, θα αποτύχει [100]. Έτσι αρκετές φορές, απαιτείται μια οργανωμένη επιλογή των τριών ομάδων, που να διασφαλίζει την κατανομή των αντικειμένων στο χώρο των δειγμάτων και όχι μια αυθαίρετη, τυχαία "ανάθεση" καθηκόντων [100, 108]. Σχετικά, οι Fernandes και Lona μάλιστα [85], αναφέρουν ότι η ομάδα ελέγχου θα πρέπει να οριοθετείται στο 85 με 95 % της ομάδας εκπαίδευσης για καλύτερα αποτελέσματα (σχ. 4.14).



Σχήμα 4.14: Εύρος της ομάδας εκπαίδευσης [85].

Για αυτόν το σκοπό, μπορεί επίσης να χρησιμοποιηθεί ο αλγόριθμος **Kennard-Stone** (K-S), σύμφωνα με τον οποίο, τα δείγματα της ομάδας εκπαίδευσης, επιλέγονται με βάση το κριτήριο της **ενδο-δειγματικής απόστασης** (inter-distance criterion,  ΚΕΦ. 2, Θ). Τα δείγματα εκπαίδευσης που επιλέγονται από τον K-S αλγόριθμο, "σαρώνουν" όλο το χώρο έτσι ώστε, το δείγμα ελέγχου που θα δοκιμαστεί να υφίσταται ουσιαστικά **παρεμβολή** (inter-

polation) και να πέφτει στην **περιοχή εφαρμογής του αντίστοιχου μοντέλου** (applicability domain, AD) [61, 104, 109]. Ωστόσο, με την εφαρμογή του K-S αλγόριθμου, μπορεί να παρατηρηθεί το φαινόμενο του μεγαλύτερου σφάλματος για την ομάδα εκπαίδευσης, παρά για την άγνωστη ομάδα ελέγχου (recognition ability vs prediction ability, βλ. § 4.2.2, 7.3.5). Αυτό συμβαίνει, γιατί καθώς ο αλγόριθμος ερευνά για τα πιο "απομακρυσμένα" στο χώρο των δεδομένων δείγματα για την ομάδα εκπαίδευσης, τα εναπομείναντα δείγματα ελέγχου χαρακτηρίζονται από τη μέγιστη ομοιομορφία. Έτσι, τα δείγματα αυτά ανήκουν σε ένα υπο-χώρο του χώρου εκπαίδευσης και μπορούν να προβλεφθούν με μεγάλη ακρίβεια. Αντίθετα, η τυχαία επιλογή των δειγμάτων εκπαίδευσης (random selection, R-S) θα οδηγούσε ίσως σε "κοντινά" μεταξύ τους δείγματα για την ομάδα αυτή και σε άγνωστα δείγματα εκτός του εύρους εκπαίδευσης. Το αποτέλεσμα θα ήταν η φτωχή ικανότητα πρόβλεψης [110, 111], καθώς η τυχαία επιλογή δειγμάτων δεν εγγυάται την αντιπροσωπευτική επιλογή αυτών. Έτσι, δεν εξασφαλίζεται ότι τα όρια των εξωτερικών δειγμάτων ελέγχου θα περιλαμβάνονται στα όρια της βαθμονόμησης του μοντέλου. Ωστόσο η τυχαία επιλογή δειγμάτων, υιοθετείται συχνά από τους ερευνητές γιατί αποτελεί την πιο απλή λύση. Μια υποομάδα δειγμάτων τυχαία επιλεγμένη από μια μεγαλύτερη βάση δεδομένων ακολουθεί τη στατιστική κατανομή ολόκληρου του δείγματος [111].

Ο αλγόριθμος **SPXY** (Sampling set Partioning based on joint x-y distances) αποτελεί μια βελτιωμένη εναλλακτική του K-S αλγόριθμου, καθώς λαμβάνει υπόψη του τις αποστάσεις στις ανεξάρτητες **x**, αλλά και την εξαρτημένη y μεταβλητή ( ΚΕΦ. 2, Θ).

Οι Zupan και Gasteiger [1] εφαρμόζουν πειραματικό σχεδιασμό (experimental design) για να επιλέξουν την καλύτερη ομάδα εκπαίδευσης σε μια εφαρμογή για την πρόβλεψη της πολικότητας των δεσμών με τη χρήση νευρωνικών δικτύων. Για τις τέσσερις μεταβλητές ένας τέτοιος σχεδιασμός τριών επιπέδων (χαμηλή πολικότητα, μέτρια, υψηλή) απαιτεί $3^4 = 81$ δεδομένα, τα οποία επιλέγονται προσεχτικά από τη διαθέσιμη ομάδα εκπαίδευσης.

Άλλοι τρόποι εύρεσης για την εύρεση της αντιπροσωπευτικότερης ομάδας δειγμάτων της ομάδας εκπαίδευσης αλλά και την επιλογή των δυο άλλων ομάδων αναφέρονται και αλλού ( ΚΕΦ. 2, Θ). Οι πιο πολλές ωστόσο μέθοδοι αντιπροσωπευτικής επιλογής δειγμάτων για την εκπαίδευση και έλεγχο των μοντέλων, επικεντρώνονται στη συνεχή και τυχαία **επαναδειγματοληψία** ("resampling") της αρχικής ομάδας με τεχνικές όπως η Monte Carlo, bootstrapping και cross-validation (§ 4.3.16, 4.4.2) [17, 86].

4.3.3. Τεχνικές ANN

Η αρχιτεκτονική των ANN περιλαμβάνει όπως ήδη αναφέρθηκε, την εισερχόμενη στιβάδα, τις ενδιάμεσες στιβάδες (αν υπάρχει είναι συνήθως μόνο μία) και την εξερχόμενη. Στις ενδιάμεσες και εξερχόμενες στιβάδες, περιλαμβάνονται νευρώνες που επεξεργάζονται τα εισερχόμενα (από τους νευρώνες των προηγούμενων στιβάδων) αποτελέσματα με τη βοήθεια βαρών και συναρτήσεων ενεργοποίησης.

Ανάλογα με τον τύπο της συνάρτησης που χρησιμοποιείται (βλ. § 4.3.4), τα δίκτυα διαφοροποιούνται σημαντικά και επομένως το χαρακτηριστικό αυτό ορίζει ουσιαστικά την ANN τεχνική:

1. **Γραμμικά δίκτυα** (linear): η συνάρτηση ενεργοποίησης είναι γραμμική. Η ικανότητα μοντελισμού και προσέγγισης της πραγματικής συνάρτησης είναι μειωμένη. Ωστόσο εδώ, είναι δυνατό να προσδιοριστεί το μοντέλο εκείνο που πραγματικά μηδενίζει το σφάλμα (βλ. παρακάτω). Συγκρινόμενα δε, με παραδοσιακές γραμμικές τεχνικές, τα Γραμμικά Νευρωνικά Δίκτυα είναι ελεύθερα παραδοχών και περιορισμών (§ 4.4.1) και παρέχουν τη δυνατότητα αναπαράστασης μέσω των καμπυλών ROC σε δυαδικά προβλήματα (§ 6.3.4) [112].
2. **Πολυστιβαδικά Νευρωνικά Δίκτυα** (Multi-layer perceptron, MLP): έχει επικρατήσει να ονομάζονται τα δίκτυα με συναρτήσεις σιγμοειδείς ή υπερβολές. Εδώ, το τίμημα για την πραγματικά μεγάλη ικανότητα μοντελισμού είναι ότι ενώ μπορούμε να ελαττώσουμε το σφάλμα προσδιορίζοντας το βέλτιστο μοντέλο, ποτέ δεν είμαστε σίγουροι ότι αυτό δεν μπορεί να γίνει ακόμα μικρότερο.
3. **Δίκτυα Ακτινικής Βάσης δίκτυα** (Radial Basis Function, RBF): η συνάρτηση ενεργοποίησης είναι Γκαουσιανή (Gaussian).

Οι αλγόριθμοι που χρησιμοποιούνται στα παραπάνω δίκτυα, ελαχιστοποιούν το σφάλμα, δηλαδή τη διαφορά ανάμεσα στη θεωρητική τιμή και την υπολογιζόμενη από το δίκτυο απόκριση (εξερχόμενο σήμα) σε μια σειρά περιόδων με συνεχείς κυκλικούς υπολογισμούς. Έτσι μπορούμε να φανταστούμε μια επιφάνεια σφάλματος (βλ. § 4.3.5), στην οποία κινούμαστε επιζητώντας το ολικό ελάχιστο. Οι περίοδοι τερματίζονται όταν εκπληρωθούν οι προδιαγραμμένες από την αρχή συνθήκες (βλ. κανόνες τερματισμού § 4.3.7). Παράλληλα αξιολογούνται οι εισερχόμενες μεταβλητές και απορρίπτονται δεδομένα που θα μπορούσαν να οδηγήσουν σε υπερ-προσαρμογή του δικτύου.

Στις παραγράφους που ακολουθούν, εξετάζονται αναλυτικά όλα τα χαρακτηριστικά των ANN αλλά ειδικότερα των MLP.

4.3.4. Συναρτήσεις ενεργοποίησης

Ποικιλία συναρτήσεων μπορούν να χρησιμοποιηθούν ως συναρτήσεις ενεργοποίησης σε ενδιάμεσες ή την εξωτερική στιβάδα των Νευρωνικών Δικτύων. Οι συναρτήσεις αυτές πρέπει έχουν παράγωγο που να ορίζεται σε μια ευρεία περιοχή [51] δηλαδή να είναι παραγωγίσιμες (differentiable) και επιπλέον συνεχείς, φραγμένες, αύξουσες [6, 65, 73, 113]. Ουσιαστικά, η συνάρτηση ενεργοποίησης είναι ο κανόνας μετασχηματισμού των "ζυγισμένων" εισερχομένων σε εξερχόμενα και εισάγει παράλληλα τη μη-γραμμικότητα στο δίκτυο [113].

Επιγραμματικά θα αναφέρουμε μερικές μόνο συνήθεις (γραμμικές ή μη) συναρτήσεις ενεργοποίησης [73]. Περισσότερες λεπτομέρειες είναι διαθέσιμες στο ηλεκτρονικό παράρτημα της διατριβής (ΚΕΦ. 2, Θ).

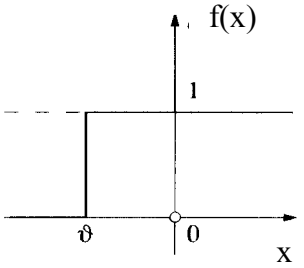
1. Ταυτοτική συνάρτηση (Identity function):

$$f(x) = x \qquad (4.4)$$

2. Δυαδική βηματική συνάρτηση ή συνάρτηση κατωφλίου θ (binary step function ή threshold function (σχ. 4.15):

$$f(x) = \begin{cases} 1 \text{ αν } x \geq \theta \\ 0 \text{ αν } x < \theta \end{cases} \qquad (4.5\alpha)$$

Η συνάρτηση αυτή μπορεί να πάρει μόνο δυο τιμές: 0 ή 1. Η σημαντική παράμετρος είναι το κατώφλι θ. Το εξερχόμενο της συνάρτησης εξαρτάται από αυτό, δηλαδή η τιμή του θ καθορίζει αν ο νευρώνας θα ενεργοποιηθεί ή όχι. Η δυαδική βηματική συνάρτηση δίνει καθορισμένες σαφείς απαντήσεις: **ναι** ή **όχι**. Έτσι, χρησιμοποιείται σε τελικά στάδια (final outputs), όταν απαιτούνται ξεκάθαρες απαντήσεις [1].



Σχήμα 4.15: Δυαδική βηματική συνάρτηση [1].

Εναλλακτική της παραπάνω συνάρτησης αποτελεί η συνάρτηση λογική κατωφλίου (threshold logic) [52]:

$$f(x) = \max [0, \min(x, 1)] \qquad (4.5\beta)$$

η οποία τελικά φαίνεται να παίρνει τρεις τιμές: 0, x, 1.

3. Δυαδική σιγμοειδής συνάρτηση (binary sigmoid function)

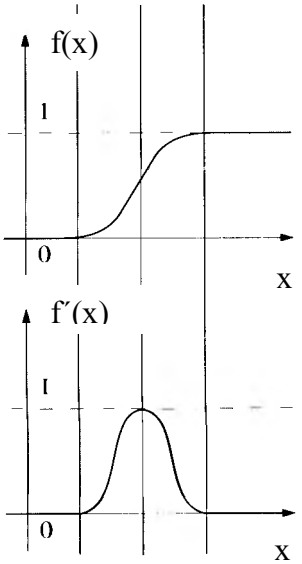
$$f(x) = \frac{1}{1+e^{-(\beta x+\theta)}} \qquad (4.6)$$

4. Γκαουσιανή συνάρτηση (Gaussian function)

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{t=-z}^{t=z} e^{-t^2} dt \ [1] \qquad (4.7)$$

Η σιγμοειδής συνάρτηση (σχέση 4.6) είναι η πιο σημαντική για τα MLP δίκτυα. Πλεονεκτεί έναντι των πολυωνυμικών συναρτήσεων όταν τα δεδομένα περιέχουν θόρυβο ή υποκρύπτονται μη γραμμικές σχέσεις [76]. Λειτουργεί ακριβώς το αντίθετο από τη δυαδική βηματική (ναι/όχι ή αληθώς/ψευδώς) συνάρτηση που είδαμε παραπάνω (σχέση 4.5α). Η σιγμοειδής συνάρτηση μπορεί να "συντηρήσει" όλες τις ενδιάμεσες καταστάσεις μεταξύ των ακραίων "yes/no" ή "true/false". Αποτελεί ίσως ένα συνδετικό κρίκο μεταξύ των ANN και των "ασαφών" (fuzzy) μεθόδων (§ 2.4.1). Η τιμή της περιορίζεται πάντοτε στο διάστημα $0 < f(x) < 1$, για οποιαδήποτε τιμή της εισερχόμενης τιμής x. Αυτό είναι πολύ σημαντικό, διότι έτσι είμαστε βέβαιοι ότι δεν θα υπάρχουν περιπτώσεις που η έξοδος παίρνει μεγάλες τιμές ή αποκτά άπειρη τιμή [49].

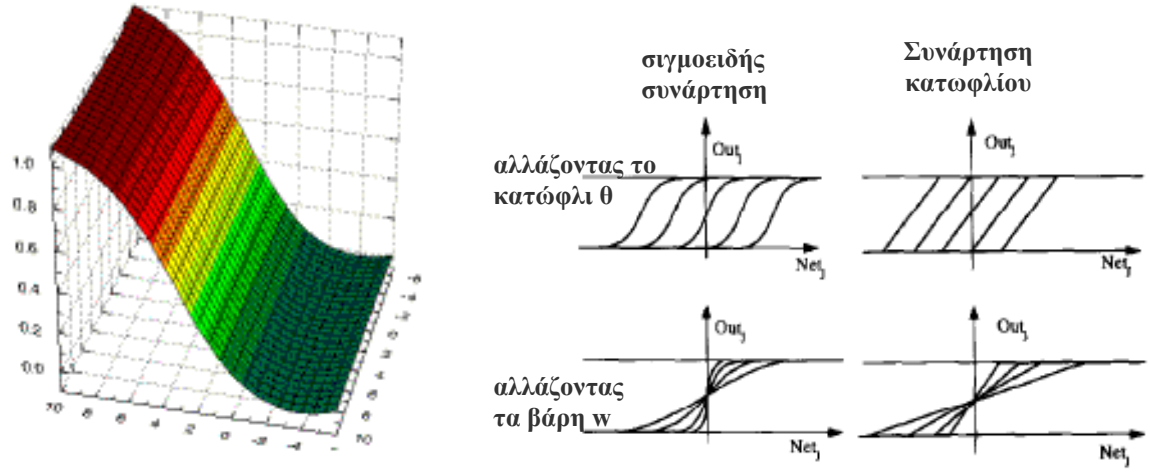
Επιπλέον, η πρώτη παράγωγος της σιγμοειδούς συνάρτησης στις "επίπεδες" περιοχές είναι μηδέν (0) (σχ. 4.16), γεγονός πολύ σημαντικό, όπως θα φανεί παρακάτω, όταν εξετάζουμε τις περιοχές όπου τα μοντέλα μαθαίνουν "ευκολότερα" (βλ. § 4.3.10).



Σχήμα 4.16: Η παράγωγος της σιγμοειδούς συνάρτησης [1].

Εδώ θα πρέπει επίσης να τονιστεί ότι **οι υψηλές τιμές βαρών μπορούν εύκολα να μετατρέψουν τη σιγμοειδή συνάρτηση σε δυαδική βηματική** ενώ αντίθετα **μικρές τιμές αυτών, ενθαρρύνουν το δίκτυο να "αποκρίνεται" διαρκώς** [1]. Το πρόβλημα των υψηλών τιμών βαρών αλλά και της υψηλής τιμής των εισερχομένων (x) στη σιγμοειδή συνάρτηση, γίνεται αμέσως κατανοητό από τη μορφή της συνάρτησης αυτής. Βοηθητικά παραδείγματα για την κατανόηση του παραπάνω αναφέρονται στο ηλεκτρονικό παράρτημα της διατριβής (ΚΕΦ. 2, Θ).

Στο σχήμα 4.17(α) δίνεται επίσης, η επιφάνεια απόκρισης (με χρήση σιγμοειδούς συνάρτησης) ενός νευρώνα σε ένα MLP δίκτυο με δυο εισερχόμενες μεταβλητές. Το σχήμα 4.17(β) αποδίδει σε δυο διαστάσεις το ίδιο φαινόμενο. Η αλλαγή των βαρών και του κατωφλίου οδηγεί σε αντίστοιχη προσαρμογή της επιφάνειας αυτής [17, 52].



(α) (β)

Σχήμα 4.17: (α) Επιφάνεια απόκρισης για μία μονάδα και δυο εισερχόμενες μεταβλητές ενός MLP δικτύου [17]. (β) Η επίδραση των τιμών των βαρών w (κάτω δεξιά διαγράμματα) και του κατωφλίου θ (πάνω δεξιά διαγράμματα). Τα βάρη w αλλάζουν την κλίση της συνάρτησης ενεργοποίησης, ενώ η παράμετρος θ την "μετακινεί" αριστερά ή δεξιά [52].

Μια μεγάλη κλίση αντανακλά μεγάλες τιμές βαρών και επομένως "ισοπέδωση" της σιγμοειδούς συνάρτησης (βλ. παρακάτω). Με ανάλογο τρόπο, η απαρχή της επιφάνειας απόκρισης εξαρτάται από το κατώφλι της συνάρτησης [17, 52]. (Η σύγκριση με δυο απλά διώνυμα μπορεί να βοηθήσει στην κατανόηση του παραπάνω. Τα διώνυμα 3x + 2 και 3x + 5, έχουν την ίδια κλίση (= 3) που εξαρτάται από το "βάρος" (= 3) του x. Αντίθετα, οι απαρχές τους (τομές στον άξονα y) είναι διαφορετικές και εξαρτώνται από το κατώφλι (= 2 και 5)).

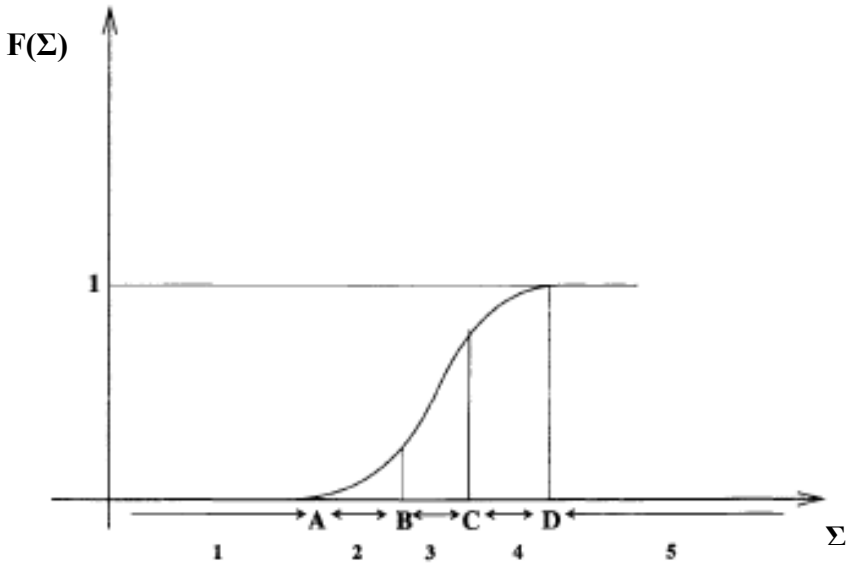
Η συνάρτηση ενεργοποίησης παίζει γενικά δυο σημαντικούς ρόλους στη διάδοση του σήματος: "συμπιέζει" τις εξερχόμενες τιμές στο εύρος μεταξύ 0 – 1, ώστε να μην πάρουν πολύ

υψηλές τιμές και επιπλέον επιτυγχάνουν **τα μη γραμμικά μοντέλα** που είναι και το κύριο πλεονέκτημα των ANN [51]. Για την εξωτερική στιβάδα, προτιμάται γενικά η σιγμοειδής συνάρτηση σε προβλήματα ταξινόμησης (εξαιτίας της "συμπίεσης" τιμών που αναφέρθηκε), ενώ στις συσχετίσεις ή χρονοσειρές συνίσταται συνεχείς συναρτήσεις [69, 76]. Έτσι, μπορεί να επιτευχθεί και προεκβολή όπου είναι ασφαλής και αναγκαία [76]. Εδώ η "συμπίεση" όχι μόνο δεν είναι απαραίτητη, αλλά πολλές φορές είναι και επιζήμια. Για την ενδιάμεση στιβάδα, συνίσταται η σιγμοειδής συνάρτηση και τα δυο είδη προβλημάτων. Η παράμετρος β στη σχέση 4.6 συνήθως παραλείπεται (β=1), καθώς αυτή καθορίζει την κλίση της συνάρτησης, η οποία ρυθμίζεται επαγωγικά από τα βάρη. Ομοίως συνήθως θ = 0.

Στο σχήμα 4.18, φαίνεται ότι υπάρχουν πέντε διαφορετικές περιοχές για τη συνάρτηση αυτή [51]:

1. Σ (εισερχόμενο άθροισμα δικτύου) < A: η απόκριση προσεγγίζει το μηδέν (0),
2. $A \leq \Sigma < B$: η απόκριση είναι μη γραμμική, κυρτή,
3. $B \leq \Sigma < C$: η απόκριση είναι σχεδόν γραμμική,
4. $C \leq \Sigma < D$: η απόκριση είναι μη γραμμική, κοίλη,
5. $\Sigma \geq D$: η απόκριση είναι μέγιστη και προσεγγίζει το 1.

Η περιοχή από A ως D (2 – 4) συνιστά τη δυναμική περιοχή. Οι περιοχές 2 και 4 είναι οι πιο σημαντικές και συνιστούν **τη μεγαλύτερη διαφορά ως προς την κλασική γραμμική συνάρτηση του perceptron**: είναι υπεύθυνες για τη μη γραμμική συμπεριφορά των μοντέλων ANN [1, 51].



Σχήμα 4.18: Διαφορετικές περιοχές απόκρισης για τη σιγμοειδή συνάρτηση [51].

4.3.5. Επιφάνεια σφάλματος

Μια βασική έννοια για τα ANN, είναι η **επιφάνεια σφάλματος** (error surface). Τα N βάρη (συμπεριλαμβανομένου και της προκατάληψης), θεωρούνται διαστάσεις σε ένα χώρο N+1 διαστάσεων. Η τελευταία διάσταση, είναι το σφάλμα του δικτύου. Για κάθε λύση που δίνεται, πρέπει να υπολογίζεται σε κάθε περίοδο το σφάλμα ώστε να βρεθεί τελικά το ελάχιστο σε αυτήν την επιφάνεια. Η μέθοδος που συχνά χρησιμοποιείται γι' αυτό το σκοπό, είναι η **μέθοδος της πιο απότομης κατάβασης** (steepest–descent minimization method). Εδώ χρησιμοποιείται το βαθμωτό διάνυσμα (gradient vector) της επιφάνειας σφάλματος. Το διάνυσμα αυτό έχει την κατεύθυνση της πιο απότομης κατάβασης και καθώς κινούμαστε κατά μήκος αυτού, ελαττώνουμε το σφάλμα [17, 114]. Μια σειρά τέτοιων κινήσεων (με πιο αργά βήματα καθώς πλησιάζουμε προς το τέλος), θα καταλήξει τελικά στο ελάχιστο [17].

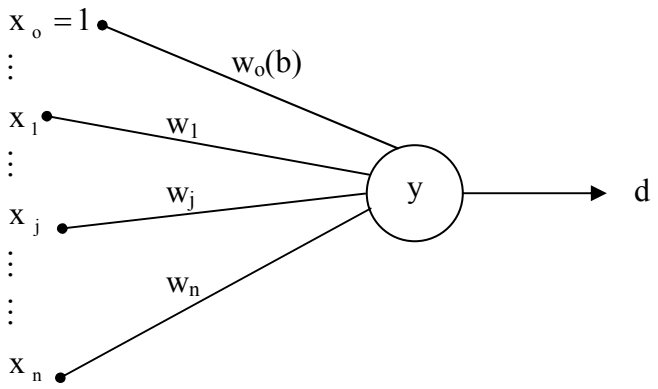
Η επιφάνεια σφάλματος μπορεί να περιέχει λόφους και κοιλάδες. Εξαιτίας δε της απότομης κατάβασης, το δίκτυο μπορεί να παγιδευτεί σε κάποιο τοπικό ελάχιστο, ενώ ένα ακόμα μεγαλύτερο ελάχιστο υπάρχει λίγο παρακάτω. Μέθοδοι πιθανοτήτων (probabilistic methods), μπορούν να βοηθήσουν ώστε ν' αποφθεχθεί μια τέτοια παγίδα, αλλά γενικά αποδεικνύονται εξαιρετικά αργές. Μια άλλη δυνατότητα είναι να αυξηθεί ο αριθμός των μονάδων της ενδιάμεσης στιβάδας. Έτσι αυξάνονται οι διαστάσεις της επιφάνειας σφάλματος και η πιθανότητα παγίδευσης σε τοπικό ελάχιστο γίνεται μικρότερη. Ωστόσο, φαίνεται ότι υπάρχει ένα ανώτατο όριο το οποίο όταν ξεπεραστεί, το σύστημα φαίνεται να παγιδεύεται πάλι σε τοπικά ελάχιστα [59].

4.3.6. Κανόνας Δέλτα (Delta-rule)

Ο κανόνας Δέλτα (Delta-rule) είναι ένας κανόνας εκπαίδευσης για Νευρωνικά Δίκτυα απλής στιβάδας (single-layer neural networks), χωρίς δηλαδή ενδιάμεση στιβάδα. Αποτελεί τη βάση του πιο βασικού αλγόριθμου (back-propagation, BP) που αναφέρεται παρακάτω (§ 4.3.9), και περιγράφεται με λεπτομέρειες στο ηλεκτρονικό παράρτημα της διατριβής (ΚΕΦ. 2, Θ). Γενικά, ο κανόνας Δέλτα πρεσβεύει τη βελτίωση του διανύσματος **w** του βάρους με τη διόρθωση **Δw** να είναι ανάλογη μιας παραμέτρου δ (η οποία είναι ανάλογη του σφάλματος), και του εισερχόμενου διανύσματος **x** για το οποίο λήφθηκε η λανθασμένη απάντηση [1]. Οι διορθώσεις λοιπόν των βαρών είναι τοπικές, με την έννοια ότι η αλλαγή στο βάρος w_j "ελέγχεται" από το εισερχόμενο x_j που "διατρέχει" ολόκληρη τη σύνδεση και το σφάλμα δ^q στον εξερχόμενο νευρώνα y [65]:

$$\Delta^q w_j = n\delta^q x_j \qquad \text{με} \quad \delta^q = d^q - y^q \qquad (4.8)$$

όπου ο εκθέτης q (=1,……N) αναφέρεται στο δείγμα q από ολόκληρο το πλέγμα των δειγμάτων, ο δείκτης j στην αντίστοιχη μεταβλητή, y^q και d^q αντίστοιχα το εξερχόμενο και η θεωρητική απόκριση (σχ. 4.19) και n>0 **ο ρυθμός εκπαίδευσης** (learning rate).



Σχήμα 4.19: Δίκτυο απλής στιβάδας με βάρη w_j

Έτσι, ο κανόνας Δέλτα διαμορφώνει τα βάρη, ελέγχοντας τα πραγματικά και θεωρητικά εξερχόμενα του δικτύου. Επιπλέον, μπορεί να εφαρμοστεί σε συνεχή ή δυαδικά εισερχόμενα και εξερχόμενα δεδομένα [59].

Ο ρυθμός εκπαίδευσης (σχέση 4.8 καθορίζει σε ποιο βαθμό θα διορθωθούν τα βάρη [1, 69, 86, 115] και καθορίζει ουσιαστικά το βήμα (μαζί με μια σειρά άλλων παραμέτρων, όπως την ορμή (βλ. § 4.3.9), το χρόνο της περιόδου και τη συνάρτηση ενεργοποίησης) που γίνεται στο χώρο των βαρών [76]. Μικρές τιμές αυτού εξαναγκάζουν το μοντέλο σε αργή σύγκλιση, ενώ υπάρχει κίνδυνος παγίδευσης αυτού σε τοπικά ελάχιστα [69, 76, 116]. Αντίθετα σε περίπτωση υψηλών τιμών, το σύστημα μπορεί να είναι ασταθές και να ταλαντώνεται (βλ. § 4.3.9) [51, 69, 76, 84, 86, 116]. Αρκετές φορές συνίσταται μεταβαλλόμενος ρυθμός εκπαίδευσης, με την τιμή αυτού να προσαρμόζεται κατά τη διάρκεια της εκπαίδευσης ανάλογα με το αν το σφάλμα αυξάνεται ή ελαττώνεται (βλ. § 4.3.11) [76, 117]. Αναφέρονται δε συναρτήσεις που καθορίζουν τη μεταβολή του ρυθμού εκπαίδευσης [118] (ΚΕΦ. 2, Θ). Οι τιμές του κυμαίνονται μεταξύ 0,5 - 1 για τις σιγμοειδείς συναρτήσεις και 0,001 - 0,1 για τις γραμμικές [51].

Η "μεταδοτική" λειτουργία του κανόνα Δέλτα θα επαληθευτεί στη γενικευμένη μορφή του, κατά την περιγραφή του αλγορίθμου BP (βλ. § 4.3.9). Ο αλγόριθμος αυτός χρησιμοποιείται ευρύτατα στα ΑΝΝ και θεωρείται ως μια γενίκευση του κανόνα Δέλτα για μη γραμμικές συναρτήσεις ενεργοποίησης σε Νευρωνικά Δίκτυα πολλαπλών στιβάδων [59].

4.3.7. Κανόνες τερματισμού / εκτίμησης και σύγκρισης μοντέλων

Η εκπαίδευση των MLP αλλά και άλλων τύπων δικτύων, ανεξάρτητα του αλγορίθμου που χρησιμοποιείται, έχει ανάγκη από κάποιον κανόνα τερματισμού (έλεγχο μιας συνάρτησης σφάλματος) που να "σφραγίζει" τον τερματισμό της διαδικασίας, ανάλογο με το μηχανισμό του "κλαδέματος" που περιγράφηκε για τα CT (§ 3.1.2). Οι κανόνες αυτοί χρησιμοποιούνται συγκριτικά και για την αξιολόγηση μοντέλων της ίδιας ή διαφορετικών τεχνικών.

Το εξερχόμενο σήμα συγκρίνεται με την επιθυμητή απόκριση, ώστε να υπολογιστεί το σφάλμα. Συνήθως παρακολουθούνται διάφοροι παράμετροι σφάλματος με γενική απαίτηση την ελάχιστη τιμή αυτών. Παρακάτω αναφέρονται μερικές από αυτές:

1. Το **Μέσο Τετράγωνο Σφάλμα** (Mean Square Error, MSE) που ισούται με το άθροισμα των τετραγώνων των διαφορών ανάμεσα στο εξερχόμενο σήμα y και τη θεωρητική απόκριση d για κάθε δείγμα και νευρώνα στην εξωτερική στιβάδα. Για κάθε δείγμα ισχύει:

$$\mathrm{MSE} = \frac{1}{K} \sum_{k=1}^{K} (y_k - d_k)^2 \qquad (4.9)$$

   όπου K είναι ο αριθμός των νευρώνων στην εξωτερική στιβάδα [66, 90] ή των όρων σφάλματος [69].

   Το MSE είναι η πιο συχνά χρησιμοποιούμενη συνάρτηση τερματισμού. Ωστόσο, αμφισβητείται η αξία της, όταν χρειάζεται να συγκρίνουμε δίκτυα διαφορετικών δεδομένων. Επιπλέον, αγνοεί τη σημαντική πληροφορία του αριθμού των βαρών. Αυτό σημαίνει ότι καθώς ο αριθμός των εκτιμούμενων παραμέτρων μεγαλώνει, ο αριθμός των βαθμών ελευθερίας μικραίνει και αυξάνεται η πιθανότητα υπερ-προσαρμογής του δικτύου. Ένα βελτιωμένο MSE θα περιελάμβανε στον παρονομαστή τους βαθμούς ελευθερίας, δηλαδή το συνολικό αριθμό των παρατηρήσεων αφού αφαιρεθούν ο αριθμός των βαρών και οι προκαταλήψεις του δικτύου [69].

2. Η **τετραγωνική ρίζα της μέσης τιμής των τετραγώνων των σφαλμάτων** (Root Mean Squared Error, RMS ή RMSE) είναι λιγότερο κατανοητό από την απόδοση (βλ. παρακάτω), αλλά αφορά την παράμετρο που πραγματικά ελαχιστοποιείται κατά τη διάρκεια της εκπαίδευσης. Αφορά το RMS των σφαλμάτων του δικτύου, όπου το κάθε σφάλμα (για κάθε δείγμα), υπολογίζεται με βάση τη θεωρητική και την πραγματική απόκριση [17, 61, 75, 115]:

$$\mathrm{RMS} = \sqrt{\frac{\sum_{p=1}^{P} (y_p - d_p)^2}{P}} \qquad (4.10)$$

όπου P ο αριθμός των δειγμάτων (στην ομάδα εκπαίδευσης ή επικύρωσης). Γενικά, το RMS θα πρέπει να ελέγχεται για όλες τις ομάδες δειγμάτων (εκπαίδευσης, επικύρωσης και ελέγχου, § 4.3.1), καθώς παρόμοιες τιμές δηλώνουν καλή ικανότητα πρόβλεψης αλλά και γενίκευσης του μοντέλου [104].

Εδώ πρέπει να σημειωθεί, ότι στις παραπάνω δυο παραμέτρους, αν απλά αλλάξουμε την κλίμακα των δεδομένων, τα μεγέθη του σφάλματος μπορεί δείχνουν να ελαττώνονται, χωρίς την πραγματικότητα να ισχύει αυτό [119]. Επίσης, αρκετές φορές στη βιβλιογραφία παρατηρείται σύγχυση των δυο παραπάνω μεγεθών με το MSE να αναφέρεται ως το τετράγωνο του RMS [75].

3. Ο **συντελεστής συσχέτισης** (correlation coefficient) μεταξύ των θεωρητικών και πραγματικών αποκρίσεων (βλ. επίσης, § 2.2, σχέση 2.8), λύνει το πρόβλημα αυτό. Ο συντελεστής αυτός κινείται στο διάστημα [-1, 1] και περιγράφει τη συνδιακύμανση ή όχι των δυο μεγεθών [119].
4. Η **Απόδοση του Δικτύου** (Performance [17] ή Non-error rate [120] ή Correctness rate [119]) που μπορεί να αφορά χωριστά (Training/Selection/Test performance) την κάθε ομάδα δειγμάτων και εξαρτάται από τον τύπο των εξερχόμενων μεταβλητών. Αν αυτές είναι συνεχείς (δηλαδή αναφερόμαστε σε δίκτυα συσχέτισης, regression networks), η απόδοση μετράται ως **ο Λόγος των Τυπικών Αποκλίσεων** (Standard Deviation Ratio) των αντίστοιχων τυπικών αποκλίσεων των προβλεπόμενων τιμών και των πραγματικών τιμών των ομάδων (εκπαίδευσης/επικύρωσης). Ένας τέτοιος λόγος μικρότερος από 1, δείχνει ότι το μοντέλο κάνει απλά μια καλή εκτίμηση των δεδομένων. Στην πράξη, τιμές ίσες με 0,1 ή μικρότερες αντιπροσωπεύουν πολύ καλή **"regression performance".**

   Αν πρόκειται για μεταβλητές κατηγοροποίησης (categorical), η απόδοση (ακρίβεια ή ικανότητα) υπολογίζεται ως το ποσοστό των σωστά ταξινομημένων δειγμάτων. Αυτή η προσέγγιση δεν λαμβάνει υπόψη τις αμφίβολες περιπτώσεις και έτσι ένα δίκτυο με συντηρητικά επίπεδα αποδοχής (thresholds ή confidence limits) μπορεί να παρουσιάζει χαμηλή φαινομενικά απόδοση [17].
5. Το **σφάλμα Kohonen** είναι η απόσταση ανάμεσα στο εισερχόμενα διάνυσμα (input) και τον κοντινότερο από τους νευρώνες του δικτύου Kohonen (βλ. § 4.3.13). Χρησιμοποιείται μόνο στα δίκτυα Kohonen [17].

Εκτός από τα παραπάνω κριτήρια τερματισμού, διάφοροι άλλοι δείκτες σύγκρισης των μοντέλων χρησιμοποιούνται από τους ερευνητές, όπως για παράδειγμα, ο αριθμός των περιόδων [121], άλλοι συντελεστές συσχέτισης εκτός του Pearson μεταξύ των προβλεπόμενων και πραγματικών τιμών [93, 122], ή η κανονική κατανομή του σφάλματος (error residuals)

[123]. Περισσότερες λεπτομέρειες αναφέρονται στο ηλεκτρονικό παράρτημα της διατριβής ( ΚΕΦ. 2, Θ).

4.3.8. Ανάλυση ευαισθησίας

Κατά την **ανάλυση ευαισθησίας** (sensitivity analysis) γίνεται "εκτίμηση" της ευαισθησίας που έχουν τα εξερχόμενα στις αλλαγές των εισερχομένων (μεταβλητών). Εδώ ακριβώς οφείλεται και το όνομά της [124].

Όταν πραγματοποιείται στα δίκτυα MLP, κατατάσσει τις μεταβλητές σε σχέση με το βαθμό της υποβάθμισης στην απόδοση του μοντέλου, στην περίπτωση που μια μεταβλητή αφαιρείται από αυτό. Καθορίζει έτσι, τη σχετική επίδραση της κάθε μεταβλητής στο εξερχόμενο αποτέλεσμα [58, 67]. Ωστόσο, επειδή οι μεταβλητές **δεν είναι ανεξάρτητες αλλά υφίστανται αλληλεξαρτήσεις,** μια απλή τιμή κατάταξης που αποδίδεται σε αυτές μέσα από την ανάλυση ευαισθησίας, δεν μπορεί να αντανακλά πλήρως τη σύνθετη εσωτερική δομή του δικτύου [96, 115]**.** Έτσι, χρειάζεται προσοχή στην εξαγωγή αποτελεσμάτων που αφορούν στην κρισιμότητα ή όχι των μεταβλητών.

Για παράδειγμα, στην περίπτωση που δυο μεταβλητές "μεταφέρουν" την ίδια πληροφορία (είναι δηλαδή συσχετιζόμενες σε υψηλό βαθμό), μπορεί να συμβαίνει κάτι από τα παρακάτω:

- ✓ το μοντέλο MLP εξαρτάται αποκλειστικά από τη μία από τις δυο μεταβλητές, ή
- ✓ το μοντέλο MLP εξαρτάται από κάποιο αυθαίρετο συνδυασμό των παραπάνω δυο μεταβλητών.

Αν κάποια από τις δυο μεταβλητές αφαιρεθεί, το μοντέλο εξακολουθεί να ανταποκρίνεται επαρκώς, εφόσον η άλλη μεταβλητή παρέχει ακόμα την κρίσιμη πληροφορία. Έτσι η ανάλυση ευαισθησίας θα προσδίδει στις δυο συγκεκριμένες μεταβλητές χαμηλή τιμή κρισιμότητας. Αντίθετα, μία μόνο μεταβλητή μεταφέρουσα μάλλον ασήμαντη πληροφορία, μπορεί να αποκομίσει υψηλότερη τιμή από τις δυο προαναφερθείσες που αμοιβαία μεταφέρουν πιο κρίσιμη πληροφορία, με την αφορμή ότι είναι η μόνη διαθέσιμη για αυτό το σκοπό.

Επιπλέον, σε κάποιες περιπτώσεις, η ανάλυση ευαισθησίας θεωρεί ότι οι μεταβλητές είναι αλληλεξαρτώμενες μεγάλης σημασίας, μόνο επειδή χρησιμοποιούνται στο μοντέλο ως σύνολο. Διαφορετικά, θα ήταν μηδενικής σημασίας, γιατί η πληροφορία που μεταφέρουν είναι μη αναγνωρίσιμη. Η βασική παράμετρος στην ανάλυση ευαισθησίας είναι το ποσοστό (ratio) του σφάλματος. Αυτό είναι μεγάλο, όταν το μοντέλο είναι ευαίσθητο σε μια μεταβλητή με τη λογική ότι αν αυτή αφαιρεθεί, η υποβάθμιση στην απόδοση του μοντέλου είναι μεγάλη. Όταν

το ποσοστό είναι 1 ή μικρότερο, αφαιρώντας τη συγκεκριμένη μεταβλητή δεν υπάρχει καμιά επίδραση στην απόδοση του μοντέλου, αντίθετα μπορεί αυτή να βελτιωθεί [20, 115].

4.3.9. Ο back-propagation αλγόριθμος

Από την εφαρμογή της απλής συνάρτησης XOR (§ 4.1.3, 4.1.4), έγινε φανερό, ότι στις περισσότερες των περιπτώσεων χρειάζονται Νευρωνικά Δίκτυα πολλαπλών στιβάδων για να προσεγγιστούν γενικότερες συσχετίσεις. Έτσι απαιτούνται αλγόριθμοι εκπαίδευσης για πιο πολύπλοκα ANN [60].

Ο **κανονικός αλγόριθμος** ή **αλγόριθμος οπισθοδιάδοσης** (BP, back-propagation), δημοσιεύτηκε πρώτα από τον Rumelhart [125] και αποτελεί μια γενίκευση του κανόνα Δέλτα που περιγράφηκε παραπάνω (§ 4.3.6). Κάθε φορά που ένα δείγμα εισάγεται στο δίκτυο, τιμές (διαμέσου των βαρών, της προκατάληψης και της συνάρτησης ενεργοποίησης), "φτάνουν" στις εξωτερικές μονάδες. Υπολογίζονται οι τετραγωνισμένες διαφορές ανάμεσα στις ευρισκόμενες και τις θεωρητικές τιμές και αν αυτές είναι διαφορετικές από το μηδέν, η μέθοδος πρέπει να είναι **"αμείλικτη"** στη διόρθωση των βαρών [59]. Η καινοτομία ωστόσο που εισάγεται εδώ είναι ότι μπορούμε να επιφέρουμε τις κατάλληλες μεταβολές στα βάρη και στις ενδιάμεσες στιβάδες, εκεί δηλαδή που δεν υπάρχει στόχος και άρα δεν μπορεί να χρησιμοποιηθεί μια απλή τεχνική, όπως ο κανόνας Δέλτα (§ 4.3.6) [49].

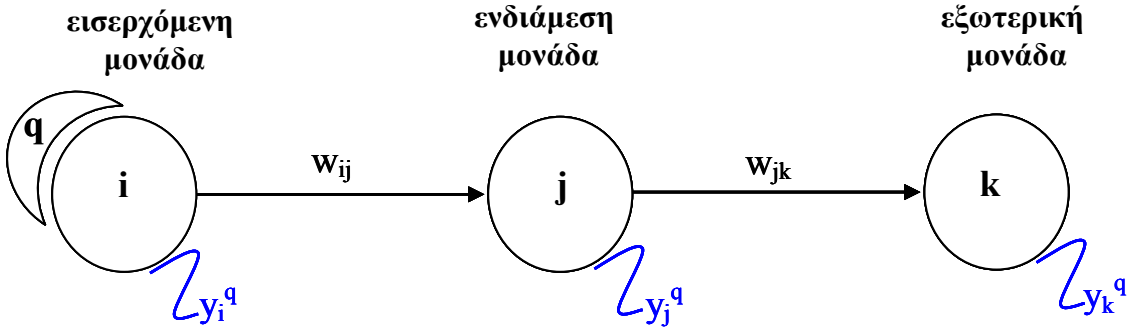
Έτσι, ο αλγόριθμος BP περιλαμβάνει δυο φάσεις:

1. Εμπρόσθια (Forward phase, FW): με τη μετάδοση του σήματος από τα εισερχόμενα προς τα εξερχόμενα και
2. Αναδρομική (Backward phase, BW): με τη μετάδοση του σφάλματος (διά της διόρθωσης των βαρών), ξεκινώντας από τα εξερχόμενα προς τα εισερχόμενα.

Η διόρθωση ενός βάρους είναι ανάλογη του **όρου σφάλματος δ** που "διαπραγματεύεται" δεδομένα της μονάδας που **λαμβάνει το σήμα** αλλά και της μονάδας από την οποία **φεύγει το σήμα.** Δηλαδή, η διόρθωση Δw_{jk} του βάρους w_{jk} που αντιστοιχεί στη σύνδεση της ενδιάμεσης μονάδας j και της εξωτερικής k (σχ. 4.20) για την παρατήρηση (δείγμα) q (μίας από το συνολικό πλέγμα των δειγμάτων), δίνεται από τη σχέση (βλ. επίσης κανόνα Δέλτα, § 4.3.6):

$$\Delta w_{jk} = n\delta_k^q y_j^q \qquad (4.11)$$

όπου n>0 είναι ο ρυθμός εκπαίδευσης (learning rate) και y_j^q το εξερχόμενο από την ενδιάμεση μονάδα j προς την εξωτερική k.



Σχήμα 4.20: Μονάδες διαδοχικών στιβάδων δικτύου με τα αντίστοιχα βάρη w και εξερχόμενα y για το εισερχόμενο δείγμα q.

Ξεκινώντας από την παραπάνω σχέση και με την προϋπόθεση ότι η συνάρτηση f που εφαρμόζεται στην ενδιάμεση και εξωτερική στιβάδα είναι σιγμοειδής, με μια σειρά βασικών πράξεων (ΚΕΦ. 2, Θ), καταλήγουμε στη σχέση (4.12) που περιγράφει τις διορθώσεις Δw_{ij} που γίνονται στα βάρη ανάμεσα στην εισερχόμενη στιβάδα και την ενδιάμεση:

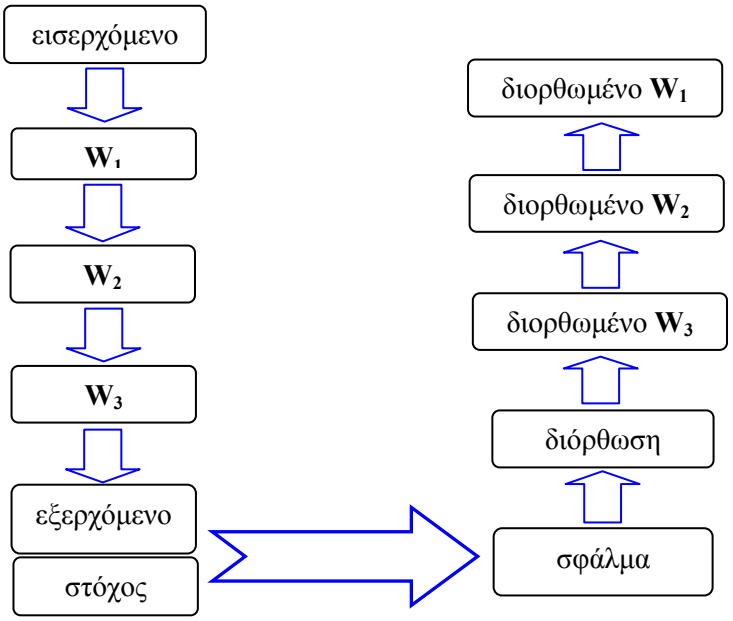
$$\Delta w_{ij}(t) = n\delta_j^q y_i^q + a\,\Delta w_{ij}(t-1) \qquad (4.12)$$

όπου Δw_{ij} είναι η "ενημέρωση" ή διόρθωση που γίνεται στα βάρη ανάμεσα στη μονάδα j της ενδιάμεσης στιβάδας και τη μονάδα i της αρχικής στιβάδας και y_i η απόκριση (εξερχόμενο) της μονάδας i της εισερχόμενης στιβάδας προς την ενδιάμεση j. Επιπλέον, n>0 είναι ο ρυθμός εκπαίδευσης και **a** η **ορμή** (momentum, βλ. παρακάτω). Οι όροι t και t-1, αναφέρονται όπως και στον κανόνα Δέλτα, στην παρούσα και την προηγούμενη περίοδο [126].

Ο όρος σφάλματος δ_j^q για την παρατήρηση q και τη μονάδα (νευρώνα) j της ενδιάμεσης στιβάδας είναι:

$$\delta_j^q = y_j^q (1-y_j^q) \sum_{k=1}^{K} \delta_k^q w_{jk} \qquad (4.13)$$

όπου w_{jk} το βάρος της σύνδεσης μεταξύ των μονάδων j και k, y_j η απόκριση (εξερχόμενο) της μονάδας j της ενδιάμεσης στιβάδας προς τον εξωτερικό νευρώνα k και K το σύνολο των εξωτερικών νευρώνων [59].



Σχήμα 4.21: Σχηματική αναπαράσταση της διόρθωσης των βαρών με τη βοήθεια του back-propagation αλγορίθμου [1].

Έτσι το τελικό σφάλμα από την εξωτερική στιβάδα, "διαδίδεται" στην ενδιάμεση "κρυφή" μέσου του όρου δ_k^q (σχέση 4.11). Αυτό είναι το πρώτο βήμα. Με αυτόν τον τρόπο ωστόσο, δεν αλλάζουν όλα τα βάρη. Το επόμενο βήμα όμως, δίνει λύση στη "μετατόπιση" του σφάλματος και **στην πρώτη στιβάδα.** Διαδοχικά λοιπόν, το σφάλμα "μετατοπίζεται" ή διανέμεται ακόμα πιο πίσω, στην πρώτη στιβάδα, με τις κατάλληλες ρυθμίσεις που γίνονται στα βάρη ανάμεσα σε αυτές και τις ενδιάμεσες. Με τον τρόπο αυτό, δικαιολογείται το όνομα του αλγορίθμου καθώς το σφάλμα **"διαδίδεται"** (is propagated) προς τα πίσω (σχ. 4.21).

Συνδυάζοντας τις σχέσεις (4.12) και (4.13), είναι φανερό ότι τιμές από **τρεις στιβάδες** επηρεάζουν τη διόρθωση των βαρών σε μια στιβάδα:

- τιμές από την "τρέχουσα" στιβάδα (εδώ ενδιάμεση μέσω του βάρους $\Delta w_{ij}(t\text{-}1)$),
- τιμές από την προηγούμενη στιβάδα (εδώ εισερχόμενη μέσω της απόκρισης y_i που προέρχεται από το νευρώνα i) και
- τιμές από την επόμενη στιβάδα (εδώ εξωτερική μέσω του όρου $\sum_{k=1}^{K} \delta_k^q w_{jk}$) [1].

Οι σχέσεις (4.11) έως (4.13), δικαιολογούν απόλυτα την "μεταδοτική" λειτουργία του αλγορίθμου BP: αρχίζοντας από το υπολογιζόμενο σφάλμα $d_k^q - y_k^q$ στην εξωτερική στιβάδα, η διόρθωση (ρύθμιση) "περνά" στα βάρη w_{jk} της ενδιάμεσης στιβάδας και καταλήγει στην πρώτη μέσα από τα βάρη w_{ij}. Μπορούμε να φανταστούμε ένα ντόμινο το οποίο κινείται εμπρός, από τα εισερχόμενα (μονάδες της πρώτης στιβάδας) προς τα εξερχόμενα, (μονάδες της εξωτερικής

στιβάδας) αλλά μετά τελείως απροσδόκητα, γυρίζει πίσω και μεταδίδει την κίνηση πάλι στις πρώτες στιβάδες [48].

Η τελική σχέση 4.13 αποτελείται από δυο προσθετέους οι οποίοι και "παρασύρουν" τη διόρθωση των βαρών προς δυο διαφορετικές κατευθύνσεις. Ο πρώτος όρος (που περιέχει το ρυθμό εκπαίδευσης) στοχεύει σε μια σύγκλιση απότομης κατάβασης (§ 4.3.5), ενώ ο δεύτερος όρος (που περιέχει την ορμή) παρεμποδίζει την παγίδευση σε τοπικά ελάχιστα (βλ. παρακάτω). Το μέγεθος των δυο αυτών σταθερών καθορίζει το βαθμό επίδρασης του κάθε όρου [1].

Η ορμή είναι ένα συντελεστής που χρησιμοποιείται για να καθορίσει **σε ποιο βαθμό, η κάθε περίοδος εξαρτάται από την προηγούμενη** [1, 51, 59, 69, 86, 115, 116, 127]. Έτσι, αναστέλλονται ξαφνικές αλλαγές στην κατεύθυνση που γίνονται οι διορθώσεις [1, 115], το δίκτυο ανταποκρίνεται καλύτερα στις "τάσεις" της επιφάνειας σφάλματος [127], ενώ αποφεύγεται η προσκόλληση σε τοπικά ελάχιστα [108, 116, 127]. Περισσότερες λεπτομέρειες για την ορμή, αναφέρονται στο ηλεκτρονικό παράρτημα της διατριβής ( ΚΕΦ. 2, Θ).

Η εκπαίδευση μέσω του αλγορίθμου BP, μπορεί να γίνει με δυο τρόπους:

1. **Pattern mode** (ή case-by-case ή on-line mode ή immediate correction ή incremental approach).
2. **Batch mode** (ή deferred correction).

Στην πρώτη περίπτωση, οι υπολογισμοί γίνονται μετά από κάθε δείγμα ή παρατήρηση (pattern ή case), ενώ στη δεύτερη περίπτωση, οι υπολογισμοί και συνεπώς η διόρθωση των βαρών, γίνονται μετά την παρουσίαση ολόκληρου του πλέγματος των δειγμάτων. Θεωρητικά με τη χρήση του pattern mode, αποφεύγονται καλύτερα τα τοπικά ελάχιστα, καθώς τα δείγματα εισέρχονται τυχαία στο νευρωνικό μοντέλο, αλλά με χρήση της batch mode μεθόδου, υπολογίζεται ακριβέστερα το βαθμωτό διάνυσμα (βλ. § 4.3.5) [76, 77] ( ΚΕΦ. 2, Θ).

Ο BP έχει λιγότερες απαιτήσεις μνήμης από άλλους αλγορίθμους. Αποτελεί συνεπώς μια καλή λύση όταν υπάρχουν πολλά δεδομένα ή όταν υπάρχει περίσσεια δεδομένων (πολλές επαναλαμβανόμενες περιπτώσεις). Έτσι, δουλεύοντας σε pattern mode και διπλασιάζοντας τα δείγματα, ο χρόνος σύγκλισης των δεδομένων θα είναι μεγαλύτερος, αλλά τα αποτελέσματα θα είναι τα ίδια, χωρίς αλλοιώσεις. Αλλά ο BP είναι εξίσου καλός και για μικρό αριθμό δεδομένων: αντίθετα ένας πιο πολύπλοκος αλγόριθμος θα ήταν αμείλικτος για το μέγεθος της ομάδας εκπαίδευσης ή της ομάδας επικύρωσης. Επιπλέον, ο αλγόριθμος BP μπορεί να τροποποιηθεί με τη χρήση του όρου της ορμής, ο οποίος επιταχύνει την "κατάβαση" όταν γίνονται αρκετά βήματα προς την ίδια κατεύθυνση. Το αποτέλεσμα, είναι γρηγορότερη σύγκλιση προς το ελάχιστο της επιφάνειας σφάλματος και αποφυγή των τοπικών ελαχίστων [17].

Γενικότερα, ο αλγόριθμος BP θεωρείται ότι επιτυγχάνει τα καλύτερα αποτελέσματα σε προβλήματα ταξινόμησης όταν μάλιστα συνδυάζεται με εμπροσθοτροφοδοτούμενο δίκτυο (§ 4.2.1) [119].

Παρά τα πλεονεκτήματα που περιγράφηκαν ωστόσο, και την επιτυχία που γνωρίζει ο BP (προφανής από τις αναρίθμητες εφαρμογές), έχουν επίσης εντοπισθεί κάποια μειονεκτήματα, τα οποία και οδήγησαν σε βελτιωμένες εκδόσεις του από πολλούς ερευνητές [59]:

1. Το δίκτυο "παραλύει" (σχ. 4.13(γ)). Αυτό σημαίνει ότι κατά τη διάρκεια της εκπαίδευσης, τα βάρη μπορεί να διαμορφωθούν **σε πολύ ψηλές τιμές** και εξαιτίας της συνήθους συνάρτησης ενεργοποίησης που είναι σιγμοειδής, το αποτέλεσμα να είναι κοντά στο 0 ή το 1. Τότε, παρατηρώντας τη συνάρτηση f:

$$y = f(x) = \frac{1}{1+e^{-\beta x}} \rightarrow f'(x) = y\,(1-y) \qquad (4.14)$$

   όπου β σταθερά, είναι φανερό ότι το σύστημα μπορεί πραγματικά να παραμείνει στάσιμο [49, 65, 76, 77] ( ΚΕΦ. 2, Θ).
2. Ο αλγόριθμος μπορεί να "παγιδευτεί" σε τοπικά ελάχιστα, όταν υπάρχει χαμηλότερη περιοχή κοντά [49, 76, 77, 128]. Εδώ, συνίσταται επίσης η επανεκπαίδευση του δικτύου με νέα αρχικά βάρη [51, 76], η χρήση του pattern mode (βλ. παραπάνω) ή η αύξηση των νευρώνων της ενδιάμεσης στιβάδας (§ 4.3.5).
3. Μερικές εφαρμογές συγκλίνουν αργά και ο χρόνος εκπαίδευσης, είναι πραγματικά υψηλός [1, 49, 77, 128, 129]. Εδώ, σημαντικό ρόλο παίζουν και οι παράμετροι που βελτιστοποιούνται ώστε να αποκτηθεί το καλύτερο μοντέλο BP (βλ. § 5.3.8) [129].
4. Γενικότερα, θεωρείται ότι ο αλγόριθμος BP είναι πιο επιτυχής σε προβλήματα ταξινόμησης (§ 4.2.2). Για παράδειγμα, σε δυαδικές εφαρμογές είναι αρκετό η εξερχόμενη απόκριση να κυμαίνεται από 1,0 ως 0,6 για το "1" και από 0,4 ως 0,0 για το "0". Αντίθετα, τα προβλήματα συσχέτισης απαιτούν μεγάλη ακρίβεια [1].

Παράδειγμα κατανόησης του BP αλγορίθμου, αναφέρεται στο ηλεκτρονικό παράρτημα της διατριβής ( ΚΕΦ. 2, Θ).

4.3.10. Αποτελεσματικότητα των Multi-layers perceptron

Η προσέγγιση των MLP στα αρχικά δεδομένα δεν είναι τέλεια. Η αποτελεσματικότητά τους (ακρίβεια αλλά και χρόνος "εκτέλεσης" των αγνώστων δειγμάτων) εξαρτάται από παράγοντες όπως:

1. Τον **αλγόριθμο** που χρησιμοποιείται και **τον αριθμό των περιόδων.** Αυτό καθορίζει την ελαχιστοποίηση του σφάλματος (αποφυγή τοπικών ελαχίστων) στην προσέγγιση των δεδομένων [59, 76].
2. Το **μέγεθος της ομάδας εκπαίδευσης** (§ 4.3.2). Αυτό καθορίζει πόσο καλά προσεγγίζεται η πραγματική συνάρτηση.
3. Τον **αριθμό των μονάδων στην ενδιάμεση στιβάδα**. Ο βέλτιστος αριθμός αυτών έχει σχέση με τον αριθμό των εισερχόμενων και εξερχόμενων μεταβλητών, τον αριθμό των δειγμάτων εκπαίδευσης, την πολυπλοκότητα του προβλήματος, το επίπεδο του θορύβου που περικλείουν τα δεδομένα και την ομοιογένεια αυτών [6]. Απλές "ομαλές" συναρτήσεις χρειάζονται μικρό αριθμό ενδιάμεσων μονάδων, ενώ συναρτήσεις με μεγάλο εύρος διακύμανσης απαιτούν πολλές μονάδες στην ενδιάμεση στιβάδα [59].
4. Την **αρχική επιλογή των βαρών** [73].
5. Τη **συνάρτηση ενεργοποίησης** που χρησιμοποιείται σε κάθε στιβάδα.
6. Τον **αριθμό των μεταβλητών** που χρησιμοποιούνται [115].

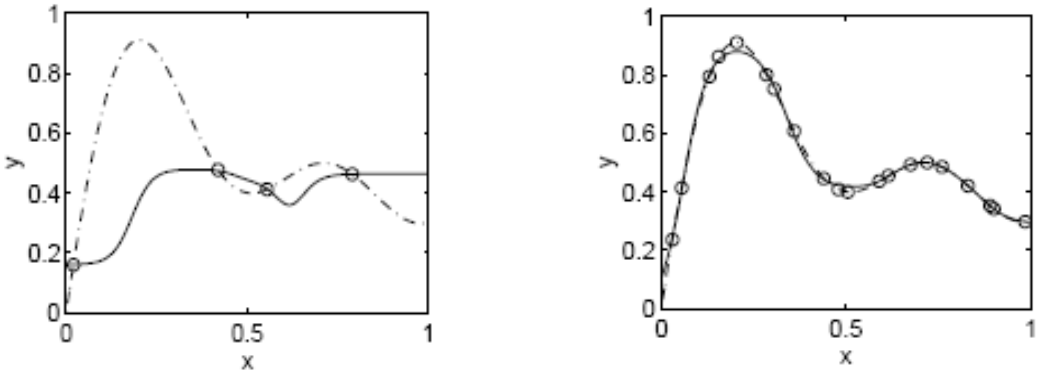
Ο BP είναι σίγουρα ο πιο συχνά χρησιμοποιούμενος **αλγόριθμος** στη βιβλιογραφία. Η επιλογή του κατάλληλου αλγορίθμου, παίζει πρωτεύοντα ρόλο, καθώς επηρεάζονται άμεσα η απόδοση του δικτύου και ο χρόνος σύγκλισης. Ο τελευταίος εξαρτάται από ένα πλήθος παραγόντων, όπως από το μέγεθος της ομάδας εκπαίδευσης (βλ. παρακάτω), από παραμέτρους σχετιζόμενες με τη βελτιστοποίηση των μοντέλων (§ 5.3.8) και το μέγεθος του δικτύου. Μεγάλα δίκτυα απαιτούν τη διόρθωση λιγότερων βαρών για την εύρεση αποδεκτής λύσης (μαθαίνουν γρηγορότερα, δηλαδή χρειάζονται λιγότερες περιόδους) αλλά παρόλα αυτά, ο χρόνος που χρειάζεται για τη διόρθωση ενός βάρους είναι μεγαλύτερος. Επιπλέον, αποφεύγουν ευκολότερα τα τοπικά ελάχιστα στην επιφάνεια σφάλματος και σχηματίζουν πολλαπλά επίπεδα αποφάσεων. Αντίθετα τα μικρά δίκτυα έχουν καλύτερη ικανότητα γενίκευσης, απαιτούν λιγότερες hardware δυνατότητες (πχ υπολογιστική μνήμη) και λιγότερο χρόνο εκπαίδευσης και εκτέλεσης για άγνωστα δείγματα [76].

Το μέγεθος των δειγμάτων συνδέεται άμεσα με την ακρίβεια (απόδοση) του μοντέλου. Περισσότερα δείγματα σημαίνει ακριβέστερο μοντέλο. Για μια δεδομένη υψηλή απόδοση του μοντέλου, απαιτείται ένα μεγαλύτερο δείγμα παρατηρήσεων καθώς η πραγματική συνάρτηση που περιγράφει τη δομή τους είναι πολύπλοκη και ο θόρυβος υψηλότερος. Στην πραγματικότητα ωστόσο, το μέγεθος του δείγματος περιορίζεται από τη διαθεσιμότητα των δεδομένων [69]. Επίσης, είναι γνωστό ότι η ακρίβεια των πολυπαραμετρικών αναλύσεων αυξάνεται με το μέγεθος του δείγματος, αλλά "κορεννύεται" για πολύ μεγάλο αριθμό δειγμάτων [130]. Με ένα αρκετά μεγάλο δείγμα, τα ANN μπορούν να μοντελοποιήσουν οποιαδήποτε πολύπλοκη δομή. Τα νευρωνικά δίκτυα φαίνεται να "κερδίζουν" από τις περιπτώσεις πολλών δειγμάτων περισ-

σότερο από τα γραμμικά μοντέλα. Εξάλλου, ότι τα ANN δεν απαιτούν απαραίτητα ένα μεγάλο δείγμα παρατηρήσεων για να έχουν υψηλές αποδόσεις [69].

Γενικά επίσης, αυξάνοντας **τον αριθμό των περιόδων εκπαίδευσης** βελτιώνεται η απόδοση του μοντέλου, αλλά αυξάνεται ο κίνδυνος υπερ-προσαρμογής και μειώνεται η δυνατότητα "γενίκευσης" [68, 90]. Έτσι συνίσταται η θέσπιση κριτηρίων για τον πρώιμο τερματισμό της διαδικασίας της εκπαίδευσης (§ 4.3.1) (πριν την ολοκλήρωση όλων των περιόδων), στην περίπτωση ανίχνευσης υπερ-προσαρμογής [90].

Για την επίδραση **της ομάδας εκπαίδευσης** στην αποτελεσματικότητα των MLP, ας θεωρήσουμε ότι μια συνάρτηση y = f(x) προσεγγίζεται με ένα εμπροσθοτροφοδοτούμενο δίκτυο. Αυτό αποτελείται από κάποιους νευρώνες στην πρώτη στιβάδα, 5 ενδιάμεσες μονάδες και μια εξωτερική. Αν έχουμε π.χ. μόνο 4 δείγματα που αποτελούν την ομάδα εκπαίδευσης, το δίκτυο εκπαιδεύεται με αυτά και η εκπαίδευση σταματά, όταν το σφάλμα δεν μειώνεται παραπέρα. Η αρχική (επιθυμητή) συνάρτηση φαίνεται στο σχήμα 4.22(α) (διακεκομμένη γραμμή). Η προσέγγιση που τελικά επιτεύχθηκε από το δίκτυο (συνεχής γραμμή), φαίνεται επίσης στο ίδιο σχήμα. Η γραμμή του δικτύου περνά από την ομάδα εκπαίδευσης, αλλά αυτή είναι τόσο ανεπαρκής που το σφάλμα του δικτύου είναι μεγάλο. Η προσέγγιση που επιτυγχάνεται με ομάδα εκπαίδευσης 20 δειγμάτων φαίνεται στο σχήμα 4.22(β): το σφάλμα είναι πολύ μικρότερο, αλλά η ομάδα των δειγμάτων (κύκλοι στο σχήμα), πολύ μεγαλύτερη.

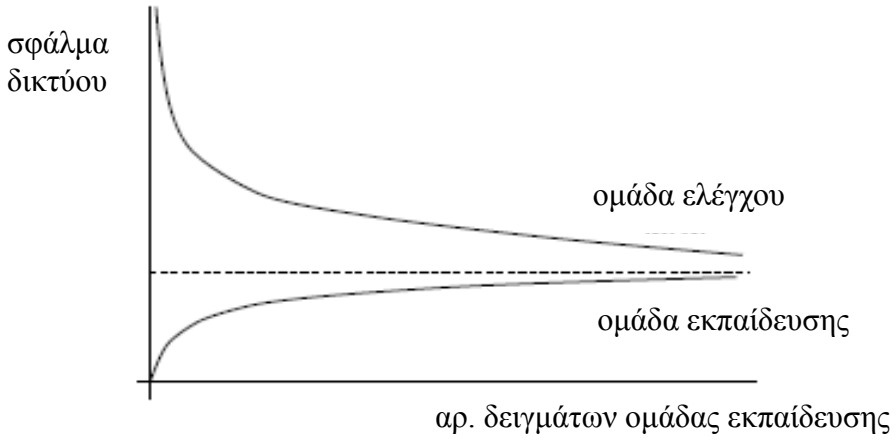


(α) (β)

Σχήμα 4.22: Επίδραση του μεγέθους της ομάδας εκπαίδευσης στην "γενίκευση" του δικτύου. Η διακεκομμένη γραμμή δηλώνει την αρχική συνάρτηση, ενώ η συνεχής την προσέγγιση που επιτεύχθηκε για (α) 4 δείγματα εκπαίδευσης και (β) 20 δείγματα εκπαίδευσης [59].

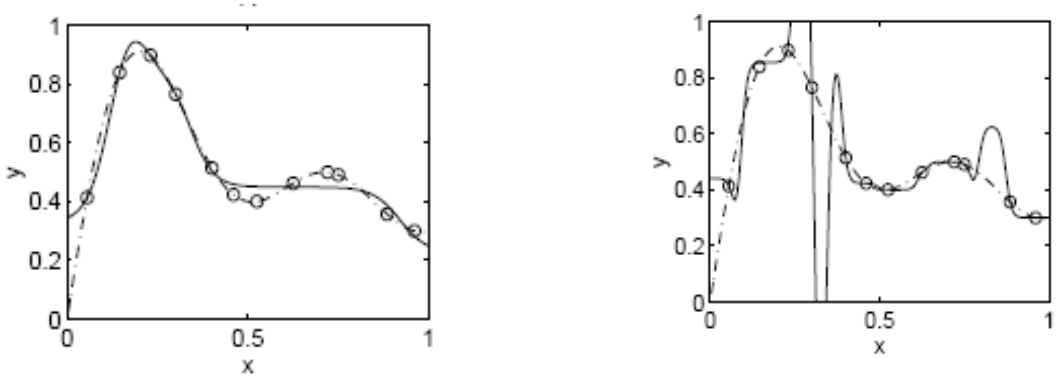
Το ίδιο πείραμα επαναλήφθηκε και για άλλα μεγέθη ομάδων εκπαίδευσης (10 φορές το καθένα). Οι μέσοι όροι για το **σφάλμα εκπαίδευσης** και το **σφάλμα ελέγχου** δίνονται στο σχήμα 4.23. Το σφάλμα εκπαίδευσης αυξάνεται με την αύξηση του μεγέθους της ομάδας εκπαίδευσης, ενώ αντίθετα το σφάλμα ελέγχου μειώνεται. Ένα μικρό σφάλμα εκπαίδευσης λοιπόν,

δεν εγγυάται πάντα καλή απόδοση για το μοντέλο. Με την αύξηση της ομάδας εκπαίδευσης, τα δυο σφάλματα συγκλίνουν στην ίδια τιμή.



Σχήμα 4.23: Επίδραση του μεγέθους της ομάδας εκπαίδευσης στην απόδοση του δικτύου. Τα σφάλματα στην ομάδα εκπαίδευσης και ελέγχου συγκλίνουν στην ίδια τιμή [59].

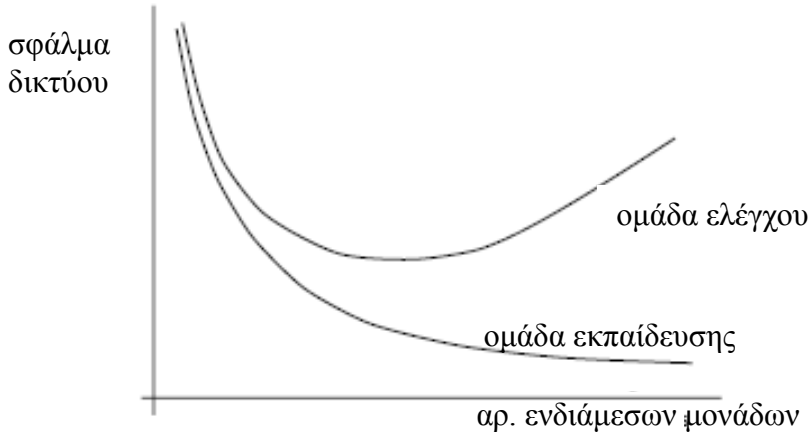
Το ίδιο παράδειγμα, μπορεί να δείξει και την επίδραση **του αριθμού των ενδιάμεσων ομάδων**. Η αρχική συνάρτηση προσεγγίζεται με 5 και 20 μονάδες αντίστοιχα στην ενδιάμεση στιβάδα (και 12 δείγματα εκπαίδευσης) στα σχήματα 4.24(α) και 4.24(β).



Σχήμα 4.24: Επίδραση του αριθμού των μονάδων ενδιάμεσης στιβάδας στην απόδοση του δικτύου. (α) 5 ενδιάμεσες μονάδες, (β) 20 ενδιάμεσες μονάδες [59].

Το σχήμα 4.24(β) απεικονίζει το φαινόμενο της υπερ-προσαρμογής που συζητήσαμε παραπάνω. Το δίκτυο προσεγγίζει ακριβώς την ομάδα εκπαίδευσης, αλλά επειδή οι ενδιάμεσες μονάδες είναι τόσες πολλές, η τελική συνάρτηση του μοντέλου απέχει πολύ της επιθυμητής. Στην πραγματικότητα, το δίκτυο προσεγγίζει και το θόρυβο (fits the noise) των δειγμάτων εκπαίδευσης, αντί να "χτίζει" μια ομαλή συνάρτηση. Με την προσθήκη ενδιάμεσων στιβάδων,

μειώνεται το σφάλμα εκπαίδευσης, αλλά το σφάλμα ελέγχου ενώ μειώνεται αρχικά, στη συνέχεια αυξάνεται (peaking effect, σχ. 4.25) [59]. Καταλήγοντας, η υπερ-προσαρμογή του δικτύου είναι αναπόφευκτη όταν χρησιμοποιείται μεγάλος αριθμός ενδιάμεσων μονάδων, αλλά ο μικρός αριθμός αυτών, δεν δίνει την ευκαιρία στο δίκτυο να "αιχμαλωτίσει" τις εγγενείς συσχετίσεις. Δεν έχει τη δυνατότητα να "γενικεύσει" και είναι ασταθές [26, 59, 66].



Σχήμα 4.25: Τα σφάλματα στην ομάδα εκπαίδευσης και ελέγχου συναρτήσει του αριθμού των ενδιάμεσων μονάδων [59].

Εκτός όμως από την ακρίβεια του μοντέλου, ο αριθμός των ενδιάμεσων μονάδων, επηρεάζει άμεσα και την εκτέλεση του δικτύου σε άγνωστα δείγματα. Για την αύξηση της ταχύτητας εκτέλεσης, είναι επιθυμητό να διατηρείται μικρός ο αριθμός των μονάδων αυτών, άρα και των αντιστοίχων βαρών. Επιπλέον οι συνδέσεις των νευρώνων συνίσταται να είναι όσο το δυνατό απλούστερες [76].

Η επιλογή **των βαρών** παίζει επίσης σημαντικό ρόλο στην αποτελεσματικότητα ενός δικτύου. Μπορεί να γίνει με τους παρακάτω τρόπους:

1. **Τυχαία επιλογή.** Η επιλογή των αρχικών βαρών θα επηρεάσει την κατάληξη του δικτύου σε ολικό ή τοπικό ελάχιστο του σφάλματος (βλ. § 4.3.5), καθώς και αν το δίκτυο τελικά συγκλίνει ή όχι. Η διαμόρφωση των βαρών ανάμεσα σε δυο νευρώνες εξαρτάται από την παράγωγο της συνάρτησης ενεργοποίησης του δεύτερου νευρώνα και το εξερχόμενο του πρώτου νευρώνα (ΚΕΦ. 2, Θ). Πρέπει λοιπόν να αποφθεχθούν τιμές, που θα οδηγήσουν σε μηδενισμό των παραγώγων ή των εξερχομένων (§ 4.3.4 και § 4.3.9, σχέση 4.12). Οι τιμές των αρχικών βαρών δεν πρέπει να είναι πολύ μεγάλες, γιατί τότε τα αρχικά εισερχόμενα στην κάθε ενδιάμεση ή εξωτερική μονάδα, θα πέσουν στην περιοχή όπου η παράγωγος της σιγμοειδούς συνάρτησης έχει πολύ μικρή τιμή (§ 4.3.9, σχέση 4.13). Αντίθετα, όταν τα αρχικά βάρη είναι πολύ μικρά,

το εισερχόμενο σε μια ενδιάμεση ή εξωτερική μονάδα θα είναι τόσο κοντά στο μηδέν, ώστε η εκπαίδευση του δικτύου γίνεται με πολύ αργά βήματα [60, 73].

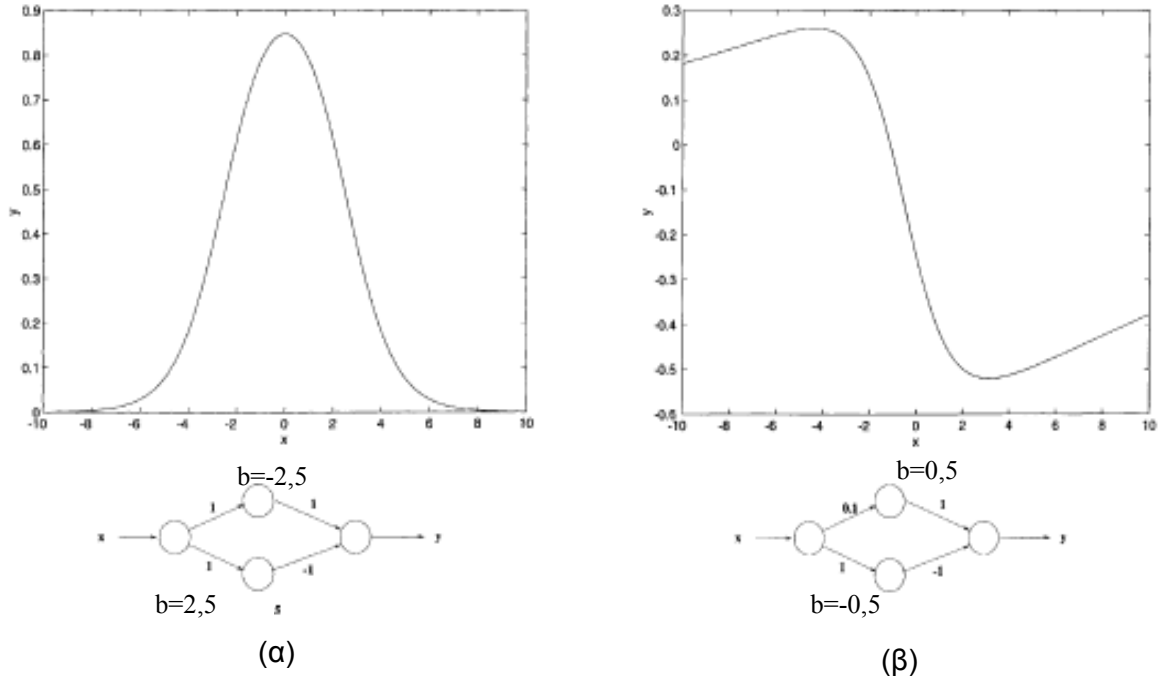
Στην εργασία τους ανασκόπησης οι Maier και Dandy [76] αναφέρουν πρόταση ερευνητών για αρχική τοποθέτηση των βαρών στην τιμή $1/\sqrt{f_i}$ όπου f_i ο αριθμός των εισερχομένων στο νευρώνα i, χωρίς ωστόσο να υιοθετούν τη χρήση του κανόνα αυτού.

Συνήθως, τα βάρη (και η προκατάληψη) ξεκινούν από τιμές μεταξύ -0,5 και +0,5 (ή μεταξύ -1 και +1). Μπορεί επίσης να επιλεγεί κανονική κατανομή αυτών ανάμεσα στις επιλεχθείσες τιμές [131], για καλύτερη κάλυψη του εύρους των τιμών.

2. **Nguyen-Widrow έναρξη.** Πρόκειται για μια τροποποίηση της τυχαίας επιλογής βαρών, που όμως οδηγεί σε γρηγορότερη "εκπαίδευση". Τα βάρη από τις εισερχόμενες μονάδες προς τις μονάδες της ενδιάμεσης στιβάδας, διαμορφώνονται έτσι ώστε να βελτιώνεται η ικανότητα των ενδιάμεσων μονάδων να εκπαιδεύονται. Αυτό επιτυγχάνεται με την κατανομή των αρχικών βαρών και της προκατάληψης έτσι ώστε, για κάθε παρατήρηση (δείγμα), το εισερχόμενο του δικτύου σε μια ενδιάμεση μονάδα να είναι στο εύρος που αυτό "μαθαίνει" γρηγορότερα.

   Τα βάρη μεταξύ των ενδιάμεσων και εξωτερικών μονάδων ξεκινούν και στην περίπτωση αυτή από τυχαίες τιμές μεταξύ -0,5 και +0,5 [73].

Μια μη γραμμική **συνάρτηση ενεργοποίησης**, όσο και αν φαίνεται περίεργο, επαρκεί για την "μοντελοποίηση" των πιο παράξενων και πολύπλοκων συναρτήσεων. Στο σχήμα 4.26 απεικονίζονται τα αποτελέσματα του συνδυασμού δυο σιγμοειδών συναρτήσεων (για δυο ενδιάμεσες στιβάδες). Με αλλαγή των βαρών, η συνάρτηση μετασχηματίζεται από Γκαουσιανή (σχ. 4.26(α)) σε σιγμοειδή (σχ. 4.26(β)). Με το σωστό συνδυασμό βαρών επομένως, μπορούν να προσεγγιστούν πολύπλοκες σχέσεις με αύξηση ενδεχομένως των σιγμοειδών συναρτήσεων που χρησιμοποιούνται [51].

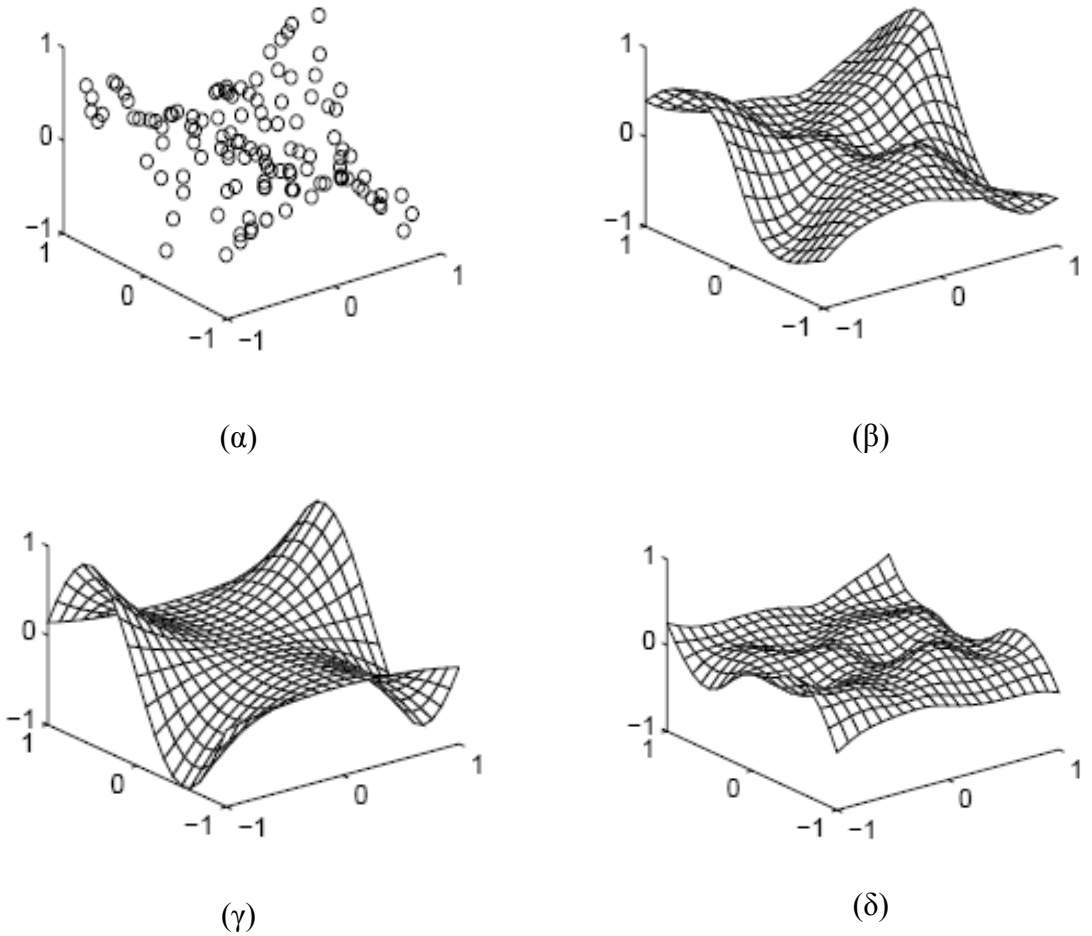


(α) (β)

Σχήμα 4.26: Εξερχόμενες τιμές y, σε σχέση με τις εισερχόμενες (x), για διαφορετικά βάρη (και προκατάληψη b). Η συνάρτηση από Γκαουσιανή (α) μετασχηματίζεται σε σιγμοειδή (β)[51].

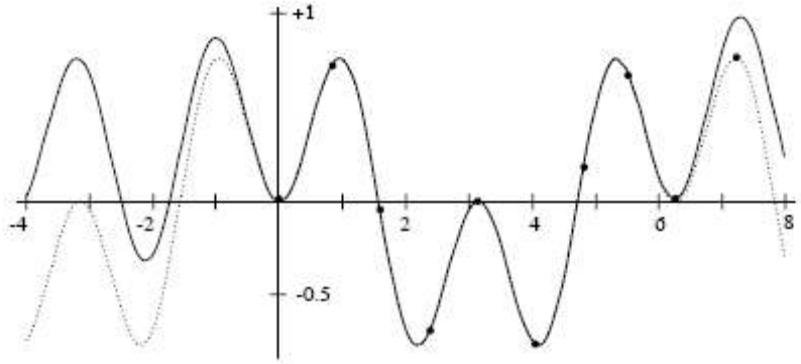
Ένα παράδειγμα θα βοηθήσει στην κατανόηση της σημασίας της συνάρτησης ενεργοποίησης. Ας θεωρήσουμε λοιπόν, ότι έχουμε ένα σύστημα με δυο εισερχόμενες μεταβλητές (δισδιάστατο εισερχόμενο διάνυσμα **x**) και μια εξερχόμενη (d). Θέλουμε να εκτιμήσουμε την πραγματική σχέση (συνάρτηση) d = f(x), για 80 δείγματα (σχήμα 4.27(α)).

Το δίκτυο που χρησιμοποιείται, περιλαμβάνει 10 μονάδες σε μια ενδιάμεση στιβάδα (με σιγμοειδή συνάρτηση ενεργοποίησης) και μια εξωτερική (με γραμμική συνάρτηση ενεργοποίησης). Ξεκινώντας από μικρά τυχαία βάρη, αλγόριθμο BP και 5000 περιόδους η πραγματική συνάρτηση f (σχ. 4.27(γ)), προσεγγίζεται από τη συνάρτηση της ομάδας εκπαίδευσης που απεικονίζεται στο σχήμα 4.27(β)). Το σφάλμα (διαφορά των δυο συναρτήσεων), απεικονίζεται στο σχήμα 4.26(δ)), από το οποίο γίνεται φανερό, ο σημαντικός ρόλος της συνάρτησης ενεργοποίησης στο τελικό αποτέλεσμα. Επιπλέον, με απλή παρατήρηση είναι δυνατό να φανεί ότι **το σφάλμα είναι μεγαλύτερο στις άκρες της επιφάνειας** (στα όρια δηλαδή του εύρους λειτουργίας του δικτύου). Το δίκτυο εκπαιδεύτηκε για κάποια ομάδα δειγμάτων με συγκεκριμένες διαστάσεις (παραμέτρους που περιγράφουν τα εισερχόμενα δείγματα) με αποτέλεσμα να "λειτουργεί" ακριβέστερα εντός του χώρου αυτού. **Το δίκτυο λοιπόν, αποδίδει πολύ καλύτερα με παρεμβολή, παρά με προεκβολή** [59].

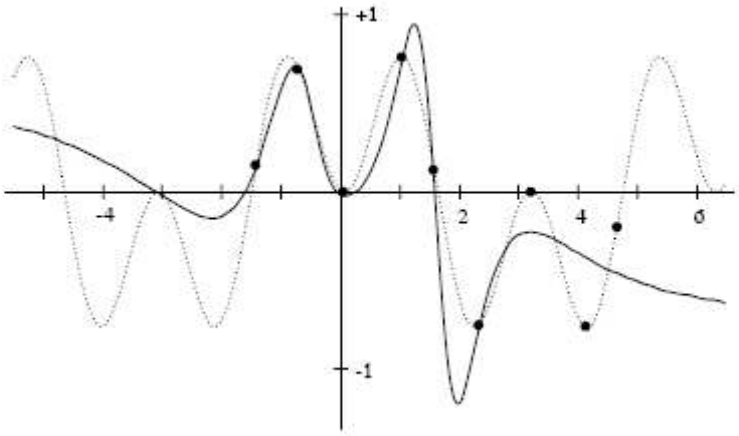


Σχήμα 4.27: Παράδειγμα της συνάρτησης προσέγγισης ενός δικτύου. (α) Η αρχική ομάδα εκπαίδευσης. (β) Η προσέγγιση που επιτυγχάνεται από το δίκτυο. (γ) Η πραγματική συνάρτηση που περιγράφει την ομάδα εκπαίδευσης. (δ) Το σφάλμα προσέγγισης [59].

Η σημασία της συνάρτησης ενεργοποίησης φαίνεται και στο παρακάτω παράδειγμα. Ας υποθέσουμε ότι έχουμε εκπαιδεύσει ένα εμπροσθοτροφοδοτούμενο δίκτυο με τέσσερις ενδιάμεσες μονάδες με ημιτονοειδή συνάρτηση ενεργοποίησης με 10 δείγματα που υπακούουν στη συνάρτηση f με f(x) = sin(2x)sin(x). Το αποτέλεσμα φαίνεται στο σχήμα 4.28(α). Η ίδια συνάρτηση προσεγγίζεται από ένα δίκτυο με διπλάσιες (8) σιγμοειδείς ενδιάμεσες μονάδες (σχ. 4.28(β)). Είναι φανερό ότι η σημασία της συνάρτησης ενεργοποίησης αλλά και βοήθεια που μπορεί να προσφέρει η εξ΄αρχής γνώση του προβλήματος.



(α)



(β)

Σχήμα 4.28: Η περιοδική συνάρτηση f(x) = sin(2x)sin(x) προσεγγίζεται από μια ημιτονοειδή συνάρτηση (α) ή μια σιγμοειδή. Η διακεκομμένη γραμμή απεικονίζει την αρχική συνάρτηση και η συνεχής τη συνάρτηση ενεργοποίησης [59].

Η επίδραση **του αριθμού των εισερχόμενων μεταβλητών**, μελετήθηκε ήδη θεωρητικά (βλ. § 4.3.1 και 4.3.8), αλλά θα εξεταστεί διεξοδικά και παρακάτω (§ 5.3.8), στην πειραματική εφαρμογή των δικτύων MLP.

4.3.11. Άλλοι αλγόριθμοι

Υπάρχουν δυο βασικές τροποποιήσεις του BP αλγορίθμου: ο quick-propagation (QP) και ο Delta-bar-Delta (DBD). Επιπλέον, νέοι προηγμένοι αλγόριθμοι έχουν αναπτυχθεί που φιλοδοξούν να ξεπεράσουν τα μειονεκτήματα του αλγορίθμου BP που αναφέρθηκαν παραπάνω.

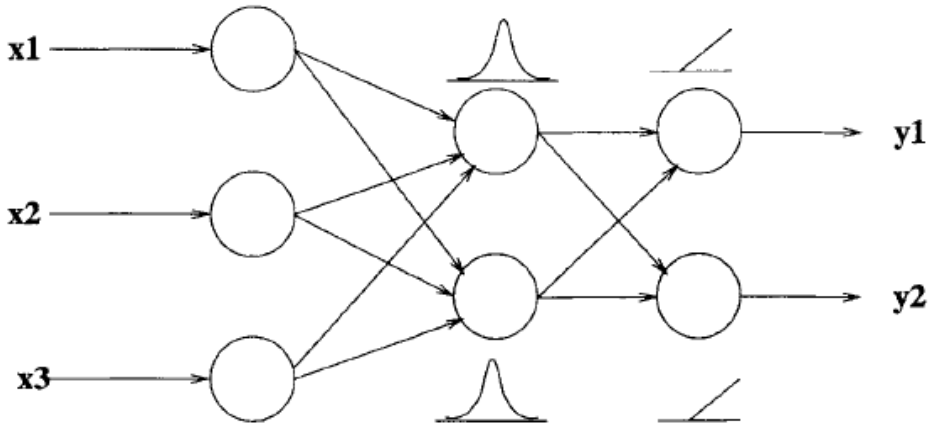
Εδώ θα αναφερθεί μόνο ο **Conjugate Gradient Descent** (CGD) που χρησιμοποιείται στο πειραματικό μέρος της διατριβής. Άλλοι αλγόριθμοι αναφέρονται στο παράρτημα ( ΚΕΦ. 2, Θ).

Ο CGD είναι ένα προηγμένος αλγόριθμος, ο οποίος μπορεί να χρησιμοποιηθεί όπου και ο BP. Είναι επίσης batch mode αλγόριθμος και συνίσταται για κάθε δίκτυο με μεγάλο αριθμό βαρών ή/και πολλαπλών εξωτερικών στιβάδων. Ο CGD αρχίζει με την κατεύθυνση της πιο απότομης κατάβασης όπως και ο BP, αλλά συνεχίζει προβάλλοντας μια ευθεία γραμμή κατά μήκος αυτής της διεύθυνσης και εντοπίζοντας το ελάχιστο πάνω σε αυτήν.

Η πορεία θεωρείται πιο σύντομη, καθώς περιλαμβάνει την εύρεση ελαχίστου σε μία μόνο διάσταση. Οι διευθύνσεις των γραμμών έρευνας επιλέγονται έτσι ώστε, να διασφαλίζεται ότι παραμένει η ελαχιστοποίηση γι' αυτές που έχουν ήδη ερευνηθεί. Αν ο αλγόριθμος ανακαλύψει ότι η τρέχουσα ευθεία δεν υποδεικνύει το ελάχιστο της επιφάνειας, επαναϋπολογίζει την κατεύθυνση της απότομης κατάβασης και ξαναρχίζει από την αρχή. Έτσι, κάθε περίοδος του CGD αλγορίθμου μπορεί να περιλαμβάνει ένα υπολογισμό και πολλές λάθος αξιολογήσεις και συνεπώς ο χρόνος σύγκλισης αυξάνεται [17, 132]. Κάθε περίοδος μπορεί να διαρκεί 3 με 10 φορές περισσότερο από μια περίοδο BP. Ωστόσο, ο CGD αλγόριθμος έχει μικρές απαιτήσεις μνήμης, ανάλογες με τον αριθμό των βαρών [17].

4.3.12. Radial Basis Function

Όπως και τα δίκτυα MLP-BP, ένα **Radial Basis Function** (RBF) δίκτυο, αποτελείται από τρεις στιβάδες: την εισερχόμενη (input), την ενδιάμεση (hidden) και την εξερχόμενη (output). Κάθε εισερχόμενη μονάδα συνδέεται με όλες τις ενδιάμεσες και αυτές με τις εξερχόμενες, διαμέσου μιας σειράς βαρών [133]. Οι ενδιάμεσες μονάδες διαμορφώνουν τα δεδομένα των εισερχομένων, χρησιμοποιώντας μια συνάρτηση που συνήθως είναι η Γκαουσιανή (σχ. 4.29).

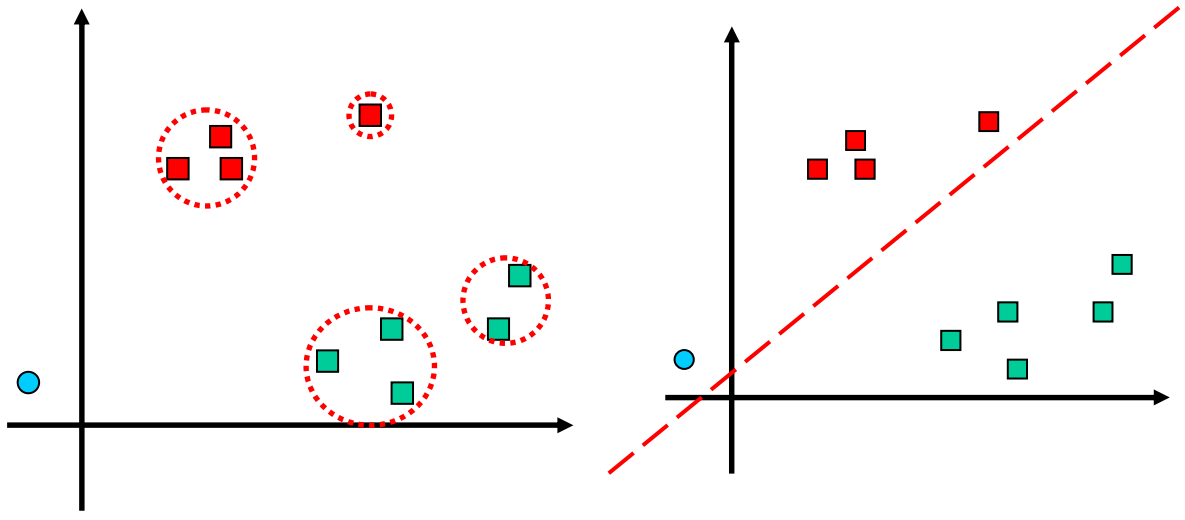


Σχήμα 4.29: Παράδειγμα RBF δικτύου [51].

Αυτή η συμμετρική συνάρτηση έχει ένα **κεντροειδές** (center) C και μια **διασπορά** (spread ή peak width ή width factor) **σ.** Η διασπορά η οποία είναι η ακτινωτή (radial) απόσταση από το κέντρο έχει τιμή διαφορετική του μηδενός. Η περιοχή που αυτή ορίζει, αποτελεί το αντίστοιχο πεδίο (field, c ± σ) της κάθε μονάδας ή νευρώνα. Έτσι, κάθε εισερχόμενο δείγμα που θα πέσει στο πεδίο, δίνει μια σημαντική απόκριση. Για κάθε δείγμα ωστόσο που είναι "απομακρυσμένο" από το κέντρο, δηλαδή έχει απόσταση μεγαλύτερη από τη διασπορά, (με βάση συνήθως την Ευκλείδεια απόσταση [113, 134, 135]), η Γκαουσιανή συνάρτηση εκφυλίζει το αποτέλεσμα σε μηδέν [133]. Το εξερχόμενο σήμα h_j από την ενδιάμεση μονάδα j δίνεται από τη σχέση:

$$h_j = e^{-(\frac{|x - Cj|}{\sigma_j})^2} \qquad (4.15)$$

όπου όρος $|x - C_j|$ απεικονίζει την Ευκλείδεια απόσταση ανάμεσα στο εισερχόμενο διάνυσμα **x** και το κεντροειδές $\mathbf{C_j}$ της Γκαουσιανής συνάρτησης [51]. Η συνάρτηση ανταποκρίνεται μόνο σε ένα μικρό χώρο των εισερχομένων, γύρω από το καθορισμένο κεντροειδές της. Το κλειδί λοιπόν για ένα επιτυχημένο δίκτυο RBF αποτελεί η σωστή επιλογή των κεντροειδών και της διασποράς [51, 136]. Με τη βοήθεια αυτών εξάλλου, τα δίκτυα RBF δημιουργούν υπερ-σφαίρες (hyperspheres) για να χωρίσουν το χώρο (σχ. 4.30(α)), ενώ αντίθετα τα δίκτυα MLP χρησιμοποιούν υπερ-επίπεδα (hyperplanes, σχ. 4.30(β)) [17, 51].

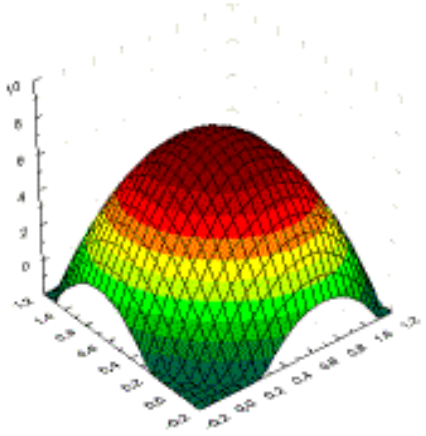


(α) (β)

Σχήμα 4.30: Διαχωρισμός του χώρου για τα δίκτυα RBF (α) και τα MLP (β).

Η επιφάνεια απόκρισης (§ 4.3.4, σχ. 4.17(α)), για μία μονάδα (νευρώνα) ενός δικτύου RBF είναι επίσης Γκαουσιανή (σχ. 4.31) με κορυφή στο κέντρο και "κατηφορική" προς τα έξω.

Όπως όμως μπορεί να αλλάξει η κλίση της σιγμοειδούς συνάρτησης των δικτύων MLP, το ίδιο μπορεί να γίνει και με την κλίση της Γκαουσιανής μονάδας [17].



Σχήμα 4.31: Επιφάνεια απόκρισης για μία μονάδα και δυο εισερχόμενες μεταβλητές ενός δικτύου RBF [17].

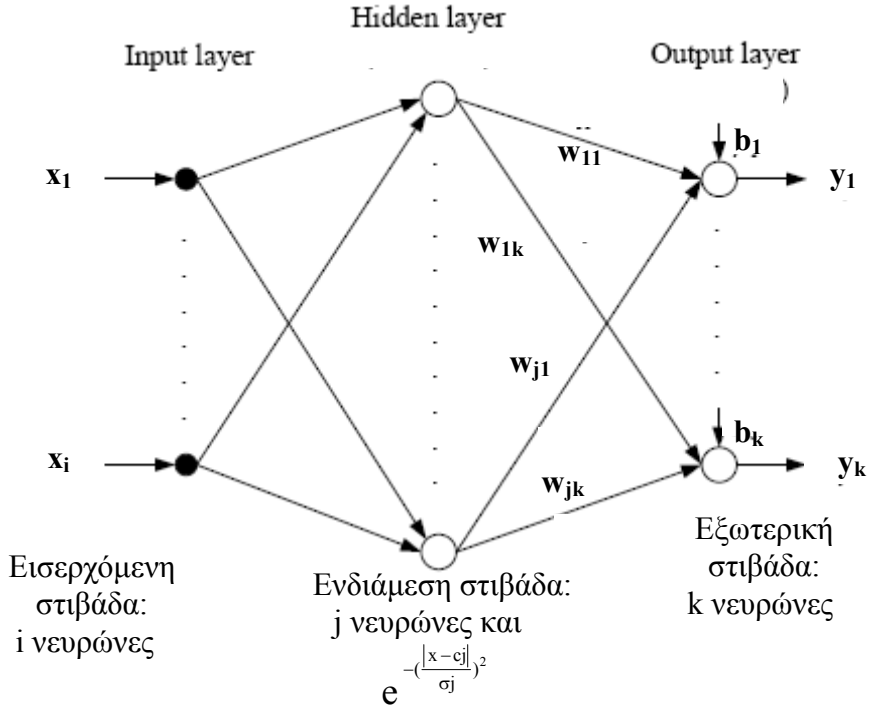
Οι μονάδες (νευρώνες) MLP καθορίζονται από τα βάρη και το κατώφλι. Πριν την εφαρμογή της σιγμοειδούς συνάρτησης, το εισερχόμενο στην επόμενη μονάδα είναι ένα "ζυγισμένο άθροισμα", το οποίο μαθηματικά είναι ένα γινόμενο διανυσμάτων (dot product) των εισερχομένων διαστάσεων και των βαρών. Αντίθετα στα δίκτυα RBF, η μονάδα καθορίζεται από το κεντροειδές και τη διασπορά. Εδώ το κεντροειδές αντιστοιχεί στα βάρη και η διασπορά (ή ακτίνα) στο κατώφλι. Ωστόσο, τα "βάρη" των δικτύων RBF (στην πραγματικότητα διάνυσμα του νευρώνα [137]) σχηματίζουν ένα σημείο στο χώρο (dot product unit) και το "κατώφλι" είναι στην πράξη μια απόκλιση [17].

Τελικά, κάθε μονάδα στην εξωτερική στιβάδα, μετασχηματίζει με τη βοήθεια γραμμικής συνάρτησης τα δεδομένα της ενδιάμεσης στιβάδας (σχέση 4.16):

$$y_k = \sum w_{jk} h_j \qquad (4.16)$$

όπου w_{jk} τα βάρη ανάμεσα στην ενδιάμεση μονάδα j και την εξωτερική k [51]. Συνολικά λοιπόν, η πορεία σχεδιασμού ενός μοντέλου RBF (ή διαφορετικά η εκπαίδευσή του, περιλαμβάνει την εύρεση του σωστού αριθμού μονάδων (νευρώνων) στην ενδιάμεση στιβάδα και τον προσδιορισμό των παρακάτω παραμέτρων (σχ. 4.32) [64, 135]:

- ✓ σ_j (η διασπορά της Γκαουσιανής συνάρτησης),
- ✓ c_j (η θέση των μονάδων),
- ✓ w_{jk} (τα βάρη της εξωτερικής στιβάδας),
- ✓ b (οι πιθανές προκαταλήψεις για την εξωτερική στιβάδα).



Σχήμα 4.32: Δίκτυο RBF. Επισημαίνονται οι προσδιοριστέες παράμετροι [135].

Οι παράμετροι αυτοί βελτιστοποιούνται μέσα από την εκπαίδευση των δικτύων RBF η οποία περιλαμβάνει γενικά δυο φάσεις:

1. τον καθορισμό των κεντροειδών (αριθμός/θέση αυτών) και της διασποράς τους [135],
2. τη βελτιστοποίηση της εξωτερικής γραμμικής στιβάδας.

Οι **θέσεις** των κεντροειδών $\mathbf{C_j}$ της Γκαουσιανής συνάρτησης καθορίζονται με την έναρξη της κατασκευής του μοντέλου. Για την καταλληλότερη κατανομή των θέσεων αυτών, συστήνονται τα παρακάτω [17, 51]:

1. τυχαία επιλογή στη σειρά εισροής των δειγμάτων εκπαίδευσης,
2. συνολική "κάλυψη" του ενδιαφερόμενου εύρους, τυχαία (στατιστικά θα υπάρχει αντιπροσωπευτική κάλυψη) ή ίσως με τη χρήση του αλγορίθμου K-S (§ 4.3.2),
3. επιλογή αντιπροσωπευτικών δειγμάτων εκπαίδευσης, ίσως με τη χρήση της K-means μεθόδου ταξινόμησης (§ 2.4.1). Ο αλγόριθμος αυτός βρίσκει τη βέλτιστη ομάδα σημείων από την ομάδα εκπαίδευσης που τοποθετούνται στα κεντροειδή. Για K μονάδες ενδιάμεσης στιβάδας, ο αλγόριθμος εξασφαλίζει ότι κάθε δείγμα εκπαίδευσης θα ανήκει στην ομάδα της οποίας το κεντροειδές είναι κοντύτερα σε αυτό από οποιοδήποτε άλλο και κάθε κεντροειδές (από τα K) αντιπροσωπεύει τα δείγματα εκπαίδευσης που ανήκουν στην ομάδα του.

Ο **αριθμός των μονάδων** (κεντροειδών) στην ενδιάμεση στιβάδα αξιολογείται και αποφασίζεται στη διάρκεια της εκπαίδευσης: στην αρχή μπορεί να είναι μηδέν και συνεχώς

προστίθενται μονάδες μέχρι να εκπληρωθούν οι κανόνες τερματισμού που έχουν τεθεί αρχικά. Αν ο αριθμός των μονάδων είναι μικρός, το μοντέλο δεν μπορεί να εκπαιδευτεί σωστά, ενώ από την άλλη πλευρά, αν ο αριθμός αυτός είναι μεγάλος, οδηγούμαστε σε φαινόμενα υπερ-προσαρμογής και πολύπλοκα δίκτυα χωρίς ικανότητα γενίκευσης [135, 137]. Όταν ο αριθμός των μονάδων είναι ίδιος με τον αριθμό των δειγμάτων εκπαίδευσης, το σφάλμα (διαφορά μεταξύ θεωρητικής και πραγματικής απόκρισης) είναι μηδέν. Το δίκτυο RBF που μόλις περιγράφηκε, ονομάζεται **ακριβές** ("exact") RBF [64].

Η διαδικασία επιλογής των θέσεων των κεντροειδών της Γκαουσιανής συνάρτησης με τον K-means αλγόριθμο είναι μη επιβλεπόμενη. Αρχικά, επιλέγονται Κ σημεία με αυθαίρετες θέσεις (βλ. περισσότερα § 2.4.1). Στη συνέχεια υπολογίζεται η απόσταση για κάθε δείγμα από τα υπόλοιπα, από καθένα από τα Κ σημεία. Επιλέγεται το κοντινότερο σημείο για κάθε δείγμα, ώστε όλα να "αποδοθούν" σε κάποιο από τα αρχικά σημεία. Μετά, από όλα τα σημεία που έχουν ταξινομηθεί στην ομάδα 1 (ή 2, 3……..Κ), υπολογίζεται ο μέσος όρος των συντεταγμένων. Αυτός αποτελεί τις νέες συντεταγμένες θέσεις για το κεντροειδές σημείο που αντιστοιχεί στην ομάδα. Η διαδικασία συνεχίζεται μέχρι σταθεροποίησης του συστήματος [17, 36].

Η διασπορά διαμορφώνεται μετά την επιλογή των κεντροειδών και κατά τη διάρκεια της εκπαίδευσης με τους εξής τρόπους:

1. Με τη χρήση του αλγόριθμου **Κ – κοντινότεροι γείτονες** (K-Nearest Neighbor, KNN). Η μέθοδος αυτή είναι γενικά μια απλή μη γραμμική μέθοδος ταξινόμησης με πολύ καλά αποτελέσματα [5], όπου οι Κ κοντινότεροι γείτονες ενός δείγματος χρησιμοποιούνται για τον καθορισμό της ομάδας που θα χαρακτηρίσει το δείγμα αυτό ("voted" KNN) [62, 130, 138]. Ο αλγόριθμος KNN παραπέμπει σε μια επιβλεπόμενη/μη-παραμετρική μέθοδο (η μόνη παραδοχή είναι ότι κάθε ομάδα περιέχει παρόμοιο αριθμό δειγμάτων) [139] και εκτελείται ως εξής:
   - ✓ βρίσκουμε την παράμετρο Κ που καθορίζει τον αριθμό των κοντινότερων γειτόνων,
   - ✓ υπολογίζουμε την απόσταση ανάμεσα σε κάθε νέο δείγμα και όλα τα γνωστά δείγματα της ομάδας εκπαίδευσης,
   - ✓ ταξινομούμε τις αποστάσεις και βρίσκουμε τους Κ κοντινότερους γείτονες του κάθε δείγματος,
   - ✓ βρίσκουμε την ομάδα που ανήκουν οι πιο πολλοί από τους Κ γείτονες, έτσι ώστε η ίδια ομάδα να αποδοθεί και στο νέο δείγμα [140, 141].

   Η ταξινόμηση μπορεί να αλλάξει για τις διάφορες τιμές του Κ (η τιμή του είναι συχνά ίση με $\sqrt{N}$, όπου N το σύνολο των δειγμάτων εκπαίδευσης [138]) και γι' αυτό ο προσδιορισμός της τιμής του θεωρείται ιδιαίτερα κρίσιμος. Η βέλτιστη τιμή του μπορεί να

βρεθεί με κάποια τεχνική επικύρωσης (όπως cross-validation, § 3.1.2, 4.4.2) [142]. Η ομάδα εκπαίδευσης δεν μπορεί να είναι πολύ μικρή. Ωστόσο, η θεωρία του αλγορίθμου είναι πολύ απλή και μπορεί να εφαρμοστεί χωρίς κάποια a-priori γνώση της δομής των δεδομένων [22, 130, 140] και χωρίς ιδιαίτερες υπολογιστικές απαιτήσεις [130].

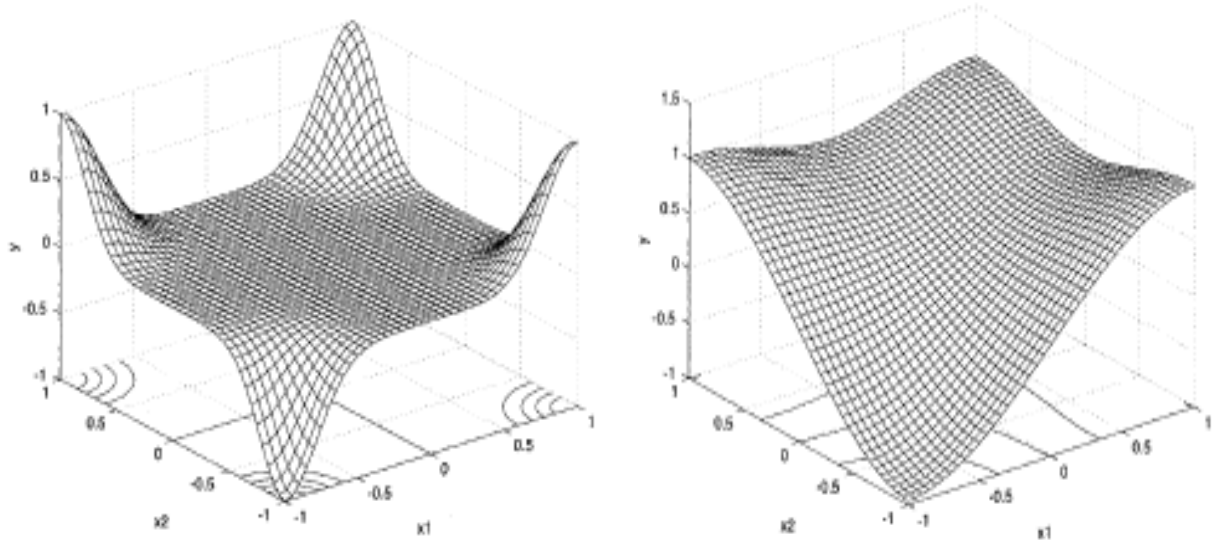
Στη συμβατική KNN τεχνική αποδίδεται ίση βαρύτητα σε κάθε γείτονα. Η ταξινόμηση εξαρτάται από την πλειονότητα των γειτόνων και όχι τη σχετική θέση του καθενός [140]. Σε κάποιες εναλλακτικές μεθόδους ωστόσο, μπορεί επιπλέον να χρησιμοποιηθούν "ζυγισμένοι" ("weighted") γείτονες, ανάλογα με την απόστασή τους από το άγνωστο δείγμα ("weighted" KNN) [23, 62, 130, 138, 140]. Διάφορες συναρτήσεις ("functions for weighting") μπορούν να χρησιμοποιηθούν εδώ [130].

Στα δίκτυα RBF, η διασπορά κάθε μονάδας τίθεται απλά στη μέση της απόστασης των K κοντινότερων γειτόνων.

2. Με την επιλογή σταθερών εξ' αρχής τιμών, βασιζόμενοι στην εμπειρία ή την εκ των προτέρων γνώση της δομής των δεδομένων.

Το μέγεθος της διασποράς καθορίζει πόσο "οξεία" είναι η Γκαουσιανή συνάρτηση που χρησιμοποιείται. Αν αυτή είναι πολύ μικρή, η συνάρτηση είναι οξεία και το δίκτυο δεν μπορεί να γνωμοδοτήσει (interpolate) ούτε μεταξύ γνωστών σημείων, δηλαδή ουσιαστικά χάνεται η δυνατότητα γενίκευσης (βλ. παρακάτω) [17, 137]. Στην πραγματικότητα εδώ βλέπουμε μια άλλη όψη του φαινομένου της υπερ-προσαρμογής (§ 4.3.1). Αν η διασπορά είναι μεγάλη, το δίκτυο είναι μικρότερο, έχει γρηγορότερη απόκριση, αλλά η συνάρτηση είναι πολύ ευρεία και το δίκτυο χάνει τις λεπτομέρειες. Τώρα, οι νευρώνες γνωμοδοτούν το ίδιο και εξαιτίας των επικαλύψεων, το δίκτυο δεν μπορεί να σχεδιαστεί σωστά [137]. Γενικά, η διασπορά των δικτύων RBF πρέπει να επιλέγεται έτσι ώστε, οι Γκαουσιανές συναρτήσεις να επικαλύπτονται με λίγα κοντινά κεντροειδή [17].

Ένα παράδειγμα (της κλασικής OR συνάρτησης) μπορεί να απεικονίσει την ιδιαίτερη σημασία της διασποράς στη διαμόρφωση του μοντέλου RBF (σχ. 4.33). Χρησιμοποιούνται τα δίπολα διανύσματα (πίνακας 4.1, § 4.1.3). Είναι φανερό ότι και στις δυο περιπτώσεις, το εξερχόμενο σήμα είναι υψηλό (1) ή χαμηλό (-1) στα διανύσματα που πρέπει. Ωστόσο, στην πρώτη περίπτωση (σχ. 4.33(α)), το μοντέλο δίνει μηδέν, για ενδιάμεσες τιμές των εισερχόμενων. Στη δεύτερη περίπτωση (σχ. 4.33(β)), το μοντέλο δίνει ομαλή παρεμβολή εξαιτίας μεγαλύτερης διασποράς. Όσο μεγαλύτερη είναι η διασπορά, τόσο πιο επίπεδη και ομαλότερη θα είναι η Γκαουσιανή συνάρτηση που προσεγγίζει τα δεδομένα. Μεγάλη διασπορά σημαίνει ότι θα χρειαστούν πολλοί νευρώνες για μια γρήγορα μεταβαλλόμενη συνάρτηση. Αντίθετα, μικρή διασπορά σημαίνει ότι θα χρειαστούν πολλοί νευρώνες για μια ομαλή συνάρτηση και το δίκτυο δεν θα έχει καλή δυνατότητα γενίκευσης [83].



(α) (β)

Σχήμα 4.33: (α) Συνάρτηση OR (β) Η ίδια συνάρτηση με μεγαλύτερη διασπορά [51].

Τα βάρη στα μοντέλα RBF, διαμορφώνονται γενικά με την κατανομή του λάθους από τις επόμενες προς τις προηγούμενες στιβάδες (BP αλγόριθμος) [51]. Ωστόσο, η βελτιστοποίηση δεν αφορά εδώ τις συνήθεις BP παραμέτρους (§ 4.3.9, 5.3.8), με αποτέλεσμα μικρότερο χρόνο εκπαίδευσης και ανάπτυξης του μοντέλου (βλ. παρακάτω) [129].

Τα δίκτυα RBF παρουσιάζουν κάποια πλεονεκτήματα σε σχέση με τα MLP, ώστε αρκετές φορές να θεωρούνται ελκυστικότερα αυτών, όπως:

1. Είναι γενικά πιο ανθεκτικά, αξιόπιστα και ευαίσθητα σε δεδομένα περιέχοντα θόρυβο σε σχέση με τα μοντέλα BP [136]. Επιπλέον, είναι πιο αναπαραγώγιμα κατά την εκτέλεση των επαναλήψεων στην πορεία της εκπαίδευσης (βλ. § 5.3.9) [137].
2. Η ρύθμιση των βαρών της εξωτερικής στιβάδας θεωρείται σχετικά απλή λόγω της γραμμικής συνάρτησης και η πορεία εκπαίδευσης εγγυάται σίγουρη σύγκλιση [17, 136].
3. Η σύγκλιση και εκπαίδευση των δικτύων γίνεται γρηγορότερα [64, 137, 143] καθώς η βελτιστοποίηση αφορά μικρότερο αριθμό παραμέτρων [129] και δεν υποφέρουν από τοπικά ελάχιστα [143] τουλάχιστον στην εξωτερική στιβάδα [17, 100, 128, 135].
4. Παρέχουν καλή "γενίκευση" με μικρό αριθμό δειγμάτων εκπαίδευσης [129] και ελάχιστο αριθμό νευρώνων [100].
5. Μπορούν να προσεγγίσουν οποιαδήποτε μη γραμμική συνάρτηση [135] με μία μόνο ενδιάμεση στιβάδα: μερικές μονάδες στη μία και μοναδική στιβάδα επαρκούν [17].
6. Κατά τον Marini [56] και τον Gulbag et al. [143] πρέπει να προτιμώνται σε περιπτώσεις πολλών μεταβλητών και ομάδων ταξινόμησης (high-dimensional, multi-class problems), καθώς απαιτούν μικρότερο λόγο αριθμού δειγμάτων προς μεταβλητές.

Ωστόσο, η εύρεση των κατάλληλων κεντροειδών είναι κρίσιμη στο σχεδιασμό των δικτύων RBF και αποτελεί τροχοπέδη στην επιλογή του [17, 100, 135, 143]. Επιπλέον, τα δίκτυα RBF είναι πιο ευαίσθητα στα προβλήματα που δημιουργούν οι πολλές μεταβλητές (βλ. § 4.4.3) και τελικά να απαιτούν πολλές ενδιάμεσες μονάδες (περισσότερες από τα MLP) για να προσεγγίσουν επαρκώς μια συνάρτηση [17, 143]. Αυτό συμβαίνει γιατί ο αριθμός των νευρώνων στην ενδιάμεση στιβάδα των δικτύων RBF τείνει να αυξάνει με την αύξηση του αριθμού των δειγμάτων εκπαίδευσης [143]. Έτσι η "εκτέλεση" των δικτύων RBF γενικά απαιτεί περισσότερο χρόνο από τα MLP. Η προεκβολή είναι ακόμα πιο απαγορευτική στην περίπτωση των RBF, εφόσον η απόκριση στα σημεία αυτά "πέφτει" στο μηδέν [17].

Γενικότερα, τα δίκτυα MLP θεωρούνται **"καθολικά" ή "σφαιρικά"** ("global") με την έννοια ότι όλα τα εισερχόμενα δείγματα δίνουν ένα εξερχόμενο. Αντίθετα τα δίκτυα RBF είναι **"τοπικά"** ("local"), καθώς μόνο δείγματα εντός του πεδίου, θα δώσουν κάποια απόκριση. Τα εκτός πεδίου δείγματα θα δώσουν μηδενική απόκριση.

Ένα παράδειγμα κατανόησης για τα δίκτυα RBF, αναφέρεται στο παράρτημα της διατριβής (ΚΕΦ. 2, Θ).

4.3.13. Δίκτυα Kohonen

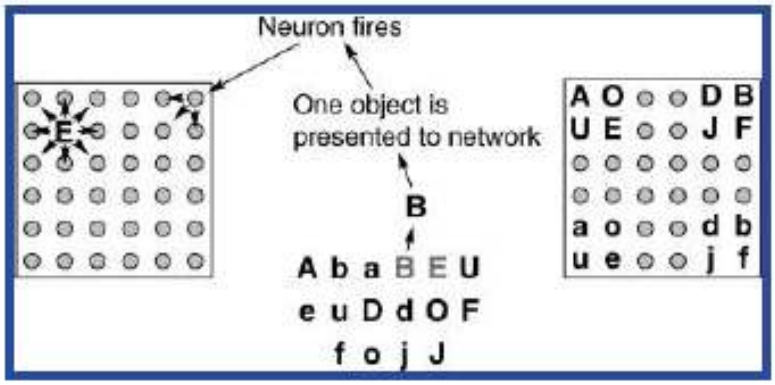
Αποσπάσματα της θεωρίας Kohonen δημοσιεύτηκε σε σχετικές εργασίες [48, 144].

Τα **Νευρωνικά Δίκτυα Kohonen,** γνωστά και ως **αυτό-οργανούμενη απεικόνιση χαρακτηριστικών** [49] (self-organizing feature maps, SOFM ή SOM), παρομοιάζουν περισσότερο από όλες τις δομές, τα βιολογικά νευρωνικά δίκτυα [1]. Το πιο χαρακτηριστικό γνώρισμα των δικτύων Kohonen που το προσομοιάζει με τα τελευταία, είναι ο τρόπος εφαρμογής των διορθώσεων. Οι διορθώσεις αυτές δεν καλύπτουν ολόκληρο το δίκτυο, ούτε καν τον ίδιο αριθμό νευρώνων στα διάφορα στάδια της εκπαίδευσης. Ο αριθμός και η έκταση των διορθώσεων αλλάζουν στη διάρκεια της εκπαίδευσης. Η εκπαίδευση είναι μια βασικά **"τοπική" διαδικασία** (local feedback): δεν επηρεάζει όλους τους νευρώνες του δικτύου, αλλά μόνο μερικούς από αυτούς (εκείνους που συνορεύουν με τον **"νικητή"**) [79].

Ο εγκεφαλικός φλοιός είναι στην πραγματικότητα ένας μεγάλος επίπεδος χάρτης (περίπου 0,5 m^2), παρόμοιος με αυτούς που θα δούμε παρακάτω στα δίκτυα Kohonen, ο οποίος αναδιπλώνεται για λόγους μειωμένου χώρου στο κρανίο, με γνωστές "τοπολογικές" ιδιότητες. Έτσι για παράδειγμα, η περιοχή που ανταποκρίνεται στην παλάμη είναι δίπλα στην περιοχή που ανταποκρίνεται στο μπράτσο [17]. Αυτό σημαίνει ότι η εγγύτητα ανάμεσα σε δυο εγκεφαλικούς νευρώνες **αντανακλά κάποιο είδος ομοιότητας στα σήματα που τους ενεργοποιεί.** Το ίδιο ωστόσο συμβαίνει και στα δίκτυα Kohonen, όπου η ενεργοποίηση ενός νευρώνα προκαλεί την

ενεργοποίηση (αλλά και την άμεση διόρθωση των βαρών) **και στους γείτονές του,** δημιουργώντας έτσι ομάδες. Οι τελευταίες όπως και στα βιολογικά δίκτυα "περικλείουν" κάποιες ομοιότητες με τα σήματα που διεγείρουν τους νευρώνες [145].

Τα δίκτυα Kohonen ανήκουν στις **μη γραμμικές μη επιβλεπόμενες** (unsupervised) τεχνικές και χρησιμοποιούνται καταρχήν για την αναγνώριση συστάδων (clusters) μέσα στην ομάδα εκπαίδευσης. Είναι μια πορεία εκπαίδευσης που χωρίζει τα εισερχόμενα δεδομένα σε ομάδες που ήδη προϋπήρχαν σε αυτά [59]. Σκοπός των δικτύων Kohonen είναι να "χαρτογραφήσουν" παρόμοια ή ίδια σήματα σε νευρώνες του ίδιου χώρου [1], καταφέρνουν δηλαδή να αναδείξουν κοντινά δείγματα και κοντινές ομάδες (σχ. 4.34) [17]. Η εκπαίδευση θεωρείται επιτυχής, όταν οι νευρώνες που σχηματίζουν αυτές τις ομάδες έχουν παρόμοια βάρη [145]. Τα εισερχόμενα (από ένα υψηλών διαστάσεων χώρο) περιορίζονται σε ένα υπο-χώρο (συνήθως σε ένα δισδιάστατο πλέγμα νευρώνων), έτσι ώστε τελικά να επιτυγχάνεται μείωση των διαστάσεων με τη βοήθεια του δικτύου Kohonen. Η παραπάνω πορεία περιγράφεται ως **"χαρτογράφηση"** (mapping) και αποσκοπεί στην αποκομιδή ενός δισδιάστατου χάρτη, σε αντιδιαστολή με την ταξινόμηση (classification, § 4.2.2) όπου θέλουμε να ταυτοποιήσουμε την ομάδα που ανήκει ένα αντικείμενο. Η διαφορά ωστόσο, δεν βρίσκεται τόσο στη μέθοδο, όσο στην ερμηνεία των αποτελεσμάτων [1].



Σχήμα 4.34: Ο νευρώνας καταφέρνει να "αιχμαλωτίσει" το δείγμα (γράμμα Β) στη δεξιά πάνω γωνία του πρώτου χάρτη. Αριστερά στον πρώτο χάρτη, το γράμμα Ε έχει ήδη αιχμαλωτιστεί. Στο δεξιό χάρτη, τα γράμματα έχουν ήδη οργανωθεί σε ομάδες: τα μικρά και τα κεφαλαία γράμματα ανήκουν σε διαφορετικές περιοχές του χάρτη [146].

Επιπλέον, η "συμπίεση" (compression) των δεδομένων είναι συνήθης πρακτική στην "διαχείριση" περίσσειας δεδομένων. Προφανώς ωστόσο, επιζητείται η μέγιστη συμπίεση με την ελάχιστη απώλεια της πληροφορίας. Έτσι η προσαρμογή στο νέο χώρο, γίνεται με τέτοιο

τρόπο, ώστε νευρώνες που φυσικά τοποθετούνται κοντά μεταξύ τους στο δίκτυο Kohonen, να περιέχουν παρόμοια εισερχόμενα ερεθίσματα [100].

Κάθε νευρώνας χαρακτηρίζεται από ένα **διάνυσμα βαρών** (reference ή codebook vector ή prototype vector) το οποίο και συνδυάζεται (μέσω της υπολογιζόμενης μεταξύ τους Ευκλείδειας απόστασης) με το αντίστοιχο διάνυσμα των δειγμάτων (παρατηρήσεων ή αντικειμένων). Έτσι, κάθε νευρώνας έχει τον ίδιο αριθμό βαρών με τις διαστάσεις του εισερχόμενου δείγματος (αν και κάποιες φορές μπορεί οι διαστάσεις των νευρώνων να είναι λιγότερες, ώστε να επιτυγχάνεται μείωση αυτών [33]). Ο νευρώνας που βρίσκεται πλησιέστερα στο εισερχόμενο δείγμα **x** (x_i, όπου i οι παράμετροι που το χαρακτηρίζουν), θεωρείται νικητής και τα βάρη του διορθώνονται, ενώ το δείγμα θεωρείται ότι "ανήκει" στο νευρώνα ή την ομάδα του. Η έννοια του **"πλησιέστερα",** έχει σχέση με δυο βασικά κριτήρια. Ο νικητής θα πρέπει:

1. είτε να δίνει το υψηλότερο εξερχόμενο σε ολόκληρο το δίκτυο :

$$y\ (\text{winner}) = \max\ [y(j)] = \max\ (\sum_i w_{ij}\ x_i\ ) \qquad (4.17)$$

2. είτε να έχει τη μικρότερη απόσταση από το εισερχόμενο:

$$D(j) = \sum_i (w_{ij} - x_i)^2 \qquad (4.18)$$

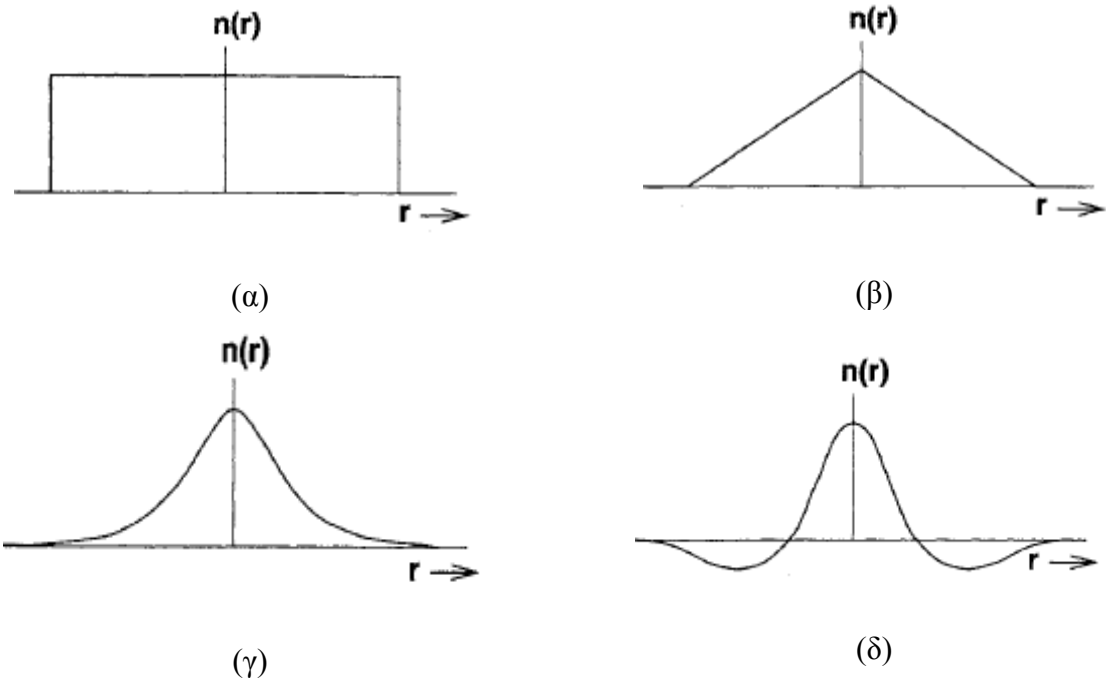
όπου οι δείκτες i (i = 1,2,……m) και j (j = 1,2…….n) αναφέρονται στις διαστάσεις του εισερχόμενου δείγματος (διανύσματος) **x** και στο συγκεκριμένο νευρώνα αντίστοιχα [1, 52].

Η διόρθωση των βαρών (ώστε το εξερχόμενο να γίνει μεγαλύτερο ή κοντύτερα στο εισερχόμενο), γίνεται με βάση τη σχέση:

$$\mathbf{w}(t+1) = \mathbf{w}(t) + n\ N(t,R)(\mathbf{x} - \mathbf{w_j}(t)) \qquad (4.19)$$

όπου n είναι ο ρυθμός εκπαίδευσης, **w**(t), **w**(t+1) είναι τα διανύσματα βαρών των νευρώνων, **x** το διάνυσμα βαρών του εισερχόμενου δείγματος και N(t,R) **συνάρτηση γειτνίασης** που περιλαμβάνει τις παραμέτρους των t (αριθμός περιόδων) και R (ακτίνα γειτνίασης). Παράλληλα λοιπόν με τον "νικητή", διορθώνονται τα βάρη των κοντινότερων (με βάση τη συνάρτηση γειτνίασης) γειτόνων του. Οι συνήθεις συναρτήσεις γειτνίασης φαίνονται στο σχήμα 4.35 [51].

Η Mexican-hat συνάρτηση (σχ. 4.35(δ)) επιτείνει τη διαφορά στα "σύνορα" των νευρώνων. Αυτό γενικά είναι πολύ χρήσιμο, αλλά αν η διαφορά αυτή μεγαλώσει υπερβολικά, εμφανίζονται "κενά", δηλαδή μη αναγνωρίσιμοι νευρώνες (§ 5.3.10) στα σύνορα αυτών [1, 52].



Σχήμα 4.35: Μερικά παραδείγματα συναρτήσεων γειτνίασης. (α) ορθογώνια (block) συνάρτηση (β) τριγωνική (triangular) (γ) Γκαουσιανή (δ) Mexican-hat συνάρτηση [51].

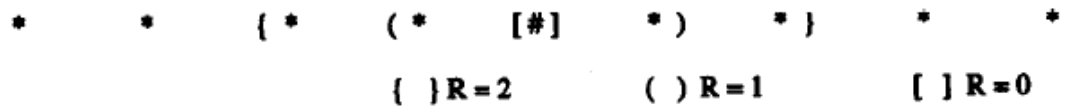
Υπάρχουν πολλές εναλλακτικές αρχιτεκτονικές δομές ("maps" ή "χάρτες" ή "τοπολογίες") για τα δίκτυα Kohonen όπως:

- ✓ χάρτες χωρίς δομή,
- ✓ μονοδιάστατες ή γραμμικές (strings σχ. 4.36(α)), ή
- ✓ δισδιάστατες (σχ. 4.36(β)) για καλύτερη οπτικοποίηση [73, 147].

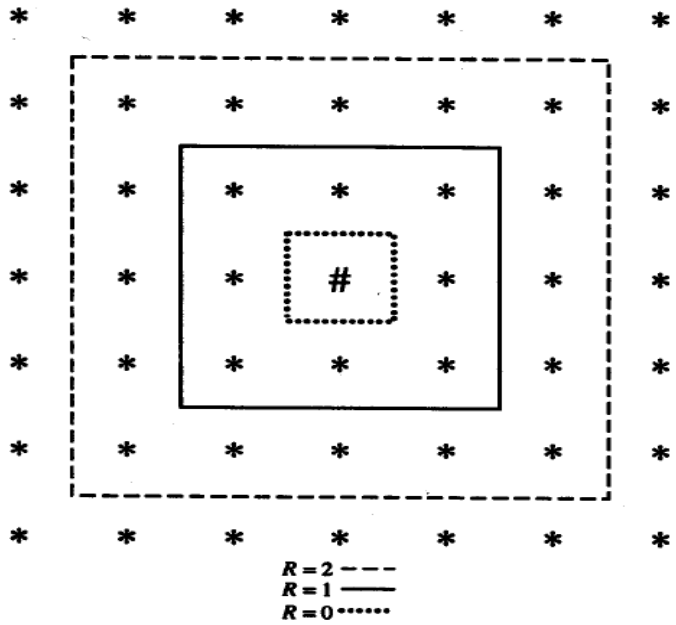
Μεγαλύτερες διαστάσεις είναι επίσης πιθανές, αλλά τέτοιες δομές δεν αποδίδονται καλά οπτικά και επομένως γενικά δεν συνηθίζονται.

Για τη δισδιάστατη δομή, μπορούν να επιλεχθούν τετράγωνες ή εξαγωνικές δομές, αλλά οι τελευταίες κατά τον Kalteh et al. [100] προσφέρονται για καλύτερη οπτικοποίηση των δεδομένων. Ο Marini et. al [102] ωστόσο δηλώνει, ότι η μορφή της δισδιάστατης δομής δεν ασκεί καμιά επίδραση στην απόδοση του δικτύου.

Έτσι, για τη δυσδιάστατη "γειτνίαση" ενός νευρώνα που ορίζεται συνήθως σε τετράγωνα ή εξάγωνα, σημαίνει ότι κάθε νευρώνας μπορεί να έχει 4 ή 6 αντίστοιχα κοντινότερους γείτονες (σχ. 4.37). Βέβαια, η γειτνίαση σε τετράγωνη δομή μπορεί να δώσει 8 κοντινότερους γείτονες και όχι 4, με την έννοια ότι ενδιαφερόμαστε για την τοπολογία (τις άμεσες συνδέσεις) και όχι την πραγματική γεωμετρική απόσταση [1]. Η απόδοση ενός δικτύου Kohonen δεν φαίνεται τελικά να επηρεάζεται από το ακριβές σχήμα της γειτνίασης [148].



(α)



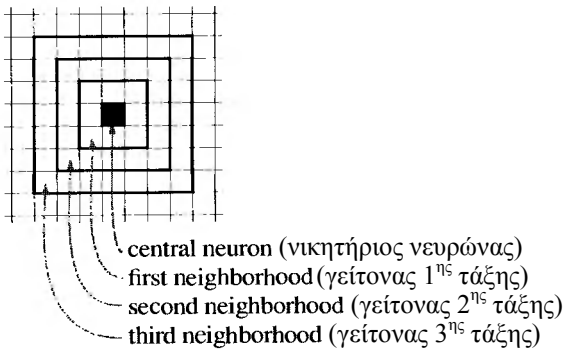
(β)

*Σχήμα 4.36: (α) Μονοδιάστατη ή γραμμική διάταξη των ομάδων και (β) δισδιάστατη διάταξη των ομάδων. Με # συμβολίζεται η "νικήτρια" ομάδα και με * οι υπόλοιπες [73].*

(α) (β)

Σχήμα 4.37: (α) Τετραγωνική και (β) εξαγωνική δομή των νευρώνων [1].

Τοπολογικά, θα πρέπει επίσης να σιγουρευτούμε ότι κάθε νευρώνας, έχει τον ίδιο αριθμό κοντινότερων γειτόνων (ή γειτόνων $1^{ης}$ τάξης, γειτόνων $2^{ης}$ τάξης κ.ο.κ.) (σχ. 4.38), αλλά το δίκτυο είναι **περιορισμένων διαστάσεων με καθορισμένα άκρα** [1]. Έτσι, οι νευρώνες που βρίσκονται στα αντίθετα άκρα, θεωρούνται απομακρυσμένοι και δεν συνδέονται μεταξύ τους [149], ενώ έχουν λιγότερους κοντινότερους γείτονες.



Σχήμα 4.38: Στην τετραγωνική δομή των νευρώνων έχουμε 8, 16, 24 κλπ 1ης, 2ης, 3ης τάξης γείτονες [1, 74].

Ένα επίπεδο μπορεί να "διπλωθεί" σε μια δακτυλιοειδή καμπύλη (torus ή toroid) ώστε να δημιουργηθούν "ισοδύναμοι" (ισάριθμοι) γείτονες για κάθε νευρώνα, χωρίς βέβαια να παραποιείται το δίκτυο (ΚΕΦ. 2, Θ).

Στην πράξη, οι toroid δομές των δικτύων Kohonen, ελαττώνουν τη διαθέσιμη περιοχή του χάρτη κατά τέσσερις (4) φορές [79]. Έτσι, ενώ η μέγιστη απόσταση μεταξύ δυο νευρώνων σε ένα toroid δίκτυο N×N είναι N/2, σε ένα non-toroid είναι N. Σε περιπτώσεις λοιπόν που χρειάζεται μεγάλη περιοχή χαρτογράφησης, το πλεονέκτημα της διπλάσιας απόστασης μεταξύ των πιο απομακρυσμένων νευρώνων μπορεί να είναι δελεαστικό για την επιλογή ενός non-toroid δικτύου. Μεγαλύτερες αποστάσεις μεταξύ των νευρώνων προσφέρουν μεγαλύτερη δυνατότητα διαχωρισμού των ομάδων σε ένα non-toroid δίκτυο σε σχέση με το αντίστοιχο toroid του ίδιο μεγέθους. Επιπλέον, οι non-toroid δομές έχουν τη δυνατότητα της αναπαράστασης των έκτροπων τιμών στις άκρες του χάρτη (πλεονέκτημα), ή διαφορετικά κάποια δείγματα θα πρέπει να βρίσκονται πάντα στη μέση ή στις άκρες αυτού (μειονέκτημα) [150].

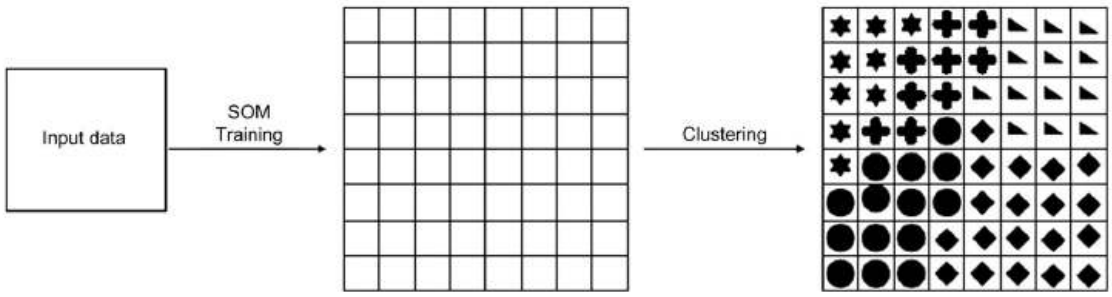
Η μέθοδος Kohonen ανήκει στις **ανταγωνιστικές** (competitive) τεχνικές, εφόσον οι νευρώνες συναγωνίζονται μεταξύ τους για την "κατάκτηση" του δείγματος [100]. Ωστόσο δε είναι μια κλασική "winner takes all" μέθοδος, εφόσον ο αλγόριθμος Kohonen διορθώνει όχι μόνο τα βάρη του νικητή, αλλά και των γειτόνων του [151].

Η πορεία εφαρμογής του Kohonen αλγορίθμου περιλαμβάνει τρία βασικά στάδια [100]:

1. **Συλλογή δεδομένων και αλλαγή κλίμακας.** Το τελευταίο πραγματοποιείται όταν υφίσταται μεγάλη διαφορά στην κλίμακα των μεταβλητών. Τότε αυτή περιορίζεται στο εύρος 0 - 1 και διασφαλίζεται η ισότιμη συνεισφορά των μεταβλητών στη δημιουργία του δικτύου.
2. **Εκπαίδευση.** Μετά την όποια προκατεργασία των δεδομένων, το κάθε δείγμα (που αντιπροσωπεύεται από ένα διάνυσμα), εισέρχεται στον κυκλικό αλγόριθμο της πορείας

εκπαίδευσης (βλ. παρακάτω) για να "χτιστεί" το δίκτυο. Συνίσταται ο αριθμός των περιόδων, να είναι τουλάχιστον 500 φορές του αριθμού των νευρώνων της εξωτερικής στιβάδας. Το εισερχόμενο διάνυσμα συγκρίνεται με το διάνυσμα βαρών του κάθε νευρώνα. Ο νευρώνας που παρουσιάζει τη μεγαλύτερη ταύτιση, ονομάζεται "νικητής" (winner ή best matching unit, BMU). Τα διανύσματα βαρών του BMU και των "γειτόνων" αυτού, διορθώνονται ώστε να αναπαράγουν το εισερχόμενο διάνυσμα.

3. **Αποκόμιση πληροφοριών από το δίκτυο Kohonen**. Μόλις το δίκτυο εκπαιδευτεί, μπορεί να δημιουργηθεί ο **τοπολογικός χάρτης** (topological ή top map, ή Kohonen map, σχ. 4.39). Αυτός επιτυγχάνει:
   - ✓ την ταξινόμηση των αρχικών δειγμάτων (clustering), λειτουργώντας ως κλασική μη επιβλεπόμενη μέθοδος,
   - ✓ την ετικετοποίηση (χαρακτηρισμό, label) των ομάδων, εφόσον βέβαια παρέχονται δεδομένα (outputs) γι΄ αυτό [100],
   - ✓ την κατάταξη (classification) νέων δειγμάτων,
   - ✓ την αναγνώριση ακόμα και δειγμάτων που δεν ανήκουν στις γνωστές ομάδες (novelty detection) [17] με τη βοήθεια των κενών νευρώνων (§ 5.3.10).
   - ✓ τη γρήγορη οπτικοποίηση των δεδομένων για άμεση παροχή πληροφοριών [79].



Σχήμα 4.39: Διάγραμμα δικτύου Kohonen δυο επιπέδων. Διαφορετικά σύμβολα αντιπροσωπεύουν διαφορετικές ομάδες [100].

Στο σχήμα 4.40(α) βλέπουμε ένα πίνακα καταμέτρησης (counting map), ο οποίος απεικονίζει τον αριθμό των δειγμάτων εκπαίδευσης που αντιστοιχούν δε κάθε νευρώνα. Ο σκουρότερος νευρώνας έχει δεχθεί τα περισσότερα "χτυπήματα" ("hits") από τα δείγματα αυτά. Ο χάρτης αυτός παρέχει μια "ματιά" στο εσωτερικό της αρχικής βάσης δεδομένων. Όταν για παράδειγμα, όλα τα εισερχόμενα δείγματα αντιστοιχίζονται σε δυο διακριτές περιοχές του χάρτη, μπορούμε να συμπεράνουμε ότι υπάρχουν δυο ομάδες στα δεδομένα.

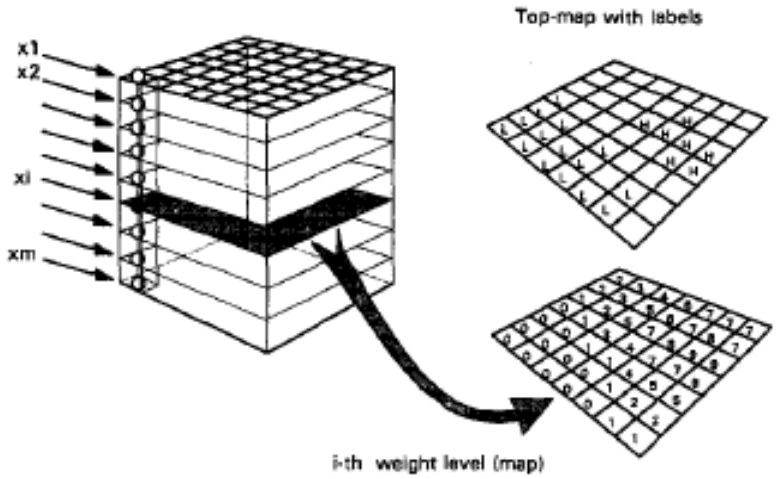
(α) (β)

Σχήμα 4.40: Δίκτυο Kohonen (α) counting map (β) output-activity map [51].

Για κάθε δείγμα εξάλλου που εισέρχεται στο μοντέλο, μπορεί να δημιουργηθεί ένας **χάρτης ενεργοποίησης** (output-activity map, σχ. 4.40(β)) που απεικονίζει το "νικητήριο" γι' αυτό νευρώνα, αλλά και τις συσχετίσεις του με τους άλλους νευρώνες. Είναι άλλωστε φανερό, ότι ένα αντικείμενο μπορεί να παρουσιάζει ομοιότητες και με άλλους νευρώνες, εκτός του νικητή. Για κάθε εισερχόμενο δείγμα, δημιουργείται ένας νέος χάρτης ενεργοποίησης [51].

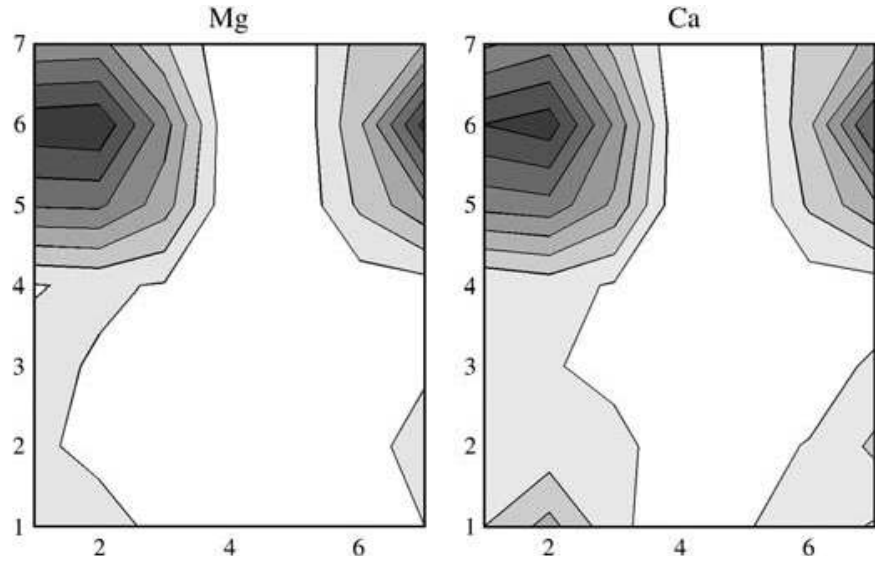
Ιδιαίτερα δημοφιλής ωστόσο, είναι ο **ενιαίος πίνακας αποστάσεων** (U-matrix, unified distance matrix). Ο πίνακας αυτός, οπτικοποιεί τις αποστάσεις μεταξύ των γειτονικών νευρώνων [33, 152 - 154]. Περισσότερες λεπτομέρειες αναγράφονται στο παράρτημα αυτής της διατριβής (ΚΕΦ. 2, Θ).

Ένα ενδιαφέρον γνώρισμα του δικτύου Kohonen, είναι ότι κάθε νευρώνας έχει τον ίδιο αριθμό βαρών, ενώ σε κάθε επίπεδο βαρών, "διαχειρίζονται" δεδομένα μόνο από μια συγκεκριμένη μεταβλητή (localized presentation, § 4.2.1). Σε κάθε βάρος, σε μια καθορισμένη και σταθερή θέση στο νευρώνα, θα "περνά" πάντα η ίδια μεταβλητή. Έτσι πχ, το πρώτο βάρος w_{1j} του νευρώνα j, "χειρίζεται" μόνο την πρώτη μεταβλητή, το δεύτερο βάρος του ίδιου νευρώνα w_{2j} χειρίζεται μόνο τη δεύτερη κλπ. [79]. Η ίδια διαδικασία επαναλαμβάνεται για όλους τους νευρώνες. Στο τέλος λοιπόν της εκπαίδευσης, σε κάθε επίπεδο του χάρτη, απεικονίζεται η κατανομή μίας και μόνο μεταβλητής (σχ. 4.41).



Σχήμα 4.41: Σύγκριση του i-επιπέδου (χάρτης βάρους) με τον τοπολογικό χάρτη (top-map) του δικτύου Kohonen [79].

Ενώ δηλαδή συγκεντρωτικά, με την αξιοποίηση όλων των μεταβλητών, παίρνουμε τον τοπολογικό χάρτη που απεικονίζει την ομαδοποίηση/ταξινόμηση όλων των δειγμάτων, σε κάθε επιμέρους επίπεδο μπορούμε να δούμε την ομαδοποίηση/ταξινόμηση που επιτυγχάνει ή όχι η κάθε μεταβλητή, της οποίας "ανήκει" το επίπεδο. Τα επίπεδα αυτά ονομάζονται **χάρτες βαρών** (weight maps ή component planes) και απεικονίζονται ως ισοϋψή διαγράμματα (contour plots, σχ. 4.42) [33, 155, 156, 157].



Σχήμα 4.42: Ισοϋψή διαγράμματα Kohonen για τα βάρη των Mg και Ca (οι σκουρότερες περιοχές αντιστοιχούν σε υψηλότερα βάρη). Οι άξονες x και y αντιστοιχούν στις διαστάσεις του τοπολογικού χάρτη (7 x 7 νευρώνες) [156].

Οι χάρτες βαρών είναι πολύ χρήσιμοι καθώς οπτικοποιούν τη διακύμανση της κάθε μεταβλητής (τα σκουρόχρωμα τμήματα δείχνουν υψηλές τιμές των μεταβλητών και τα ανοιχτόχρωμα, χαμηλές). Επιπλέον, συγκρίνοντας τους χάρτες βαρών μεταξύ τους, εξάγονται συμπεράσματα για τις συσχετίσεις των μεταβλητών [158, 159]. Οι χάρτες βαρών μπορούν να κατασκευαστούν και με βάση τα U-matrices [152].

Συγκεκριμένα, από τους χάρτες βαρών, μπορούν να εξαχθούν συμπεράσματα για την κρισιμότητα των μεταβλητών και τη διαχωριστική τους ικανότητα, όπως για παράδειγμα αν μια μεταβλητή (πχ στο i-επίπεδο) ανταποκρίνεται καλά ή όχι στην ομαδοποίηση των δειγμάτων του τοπολογικού χάρτη [79]. Έτσι, οι χάρτες βαρών συγκρίνονται οπτικά (ή με τη βοήθεια δεικτών όπως ο Self-Orginizing Map Discrimination Index, SOMDI [150, 160, 161]) ώστε αυτοί να ανταποκρίνονται όσο το δυνατό καλύτερα στις ομάδες του τοπολογικού χάρτη [150]. Με τον τρόπο αυτό, αίρονται τουλάχιστον για τα δίκτυα Kohonen, ενστάσεις που αφορούν ένα εκπαιδευμένο μοντέλο ANN και τη θεώρηση αυτού ως "black box" (§ 4.4.1) [74].

Όπως ήδη αναφέρθηκε (σχέση 4.19), τα νέα βάρη υπολογίζονται και διορθώνονται με βάση τα παλιά μέσω ενός ρυθμού εκπαίδευσης, καθορισμένου από την αρχή. Ο ίδιος ο ρυθμός εκπαίδευσης διορθώνεται σε κάθε περίοδο (συνήθως μειώνεται γεωμετρικά με βάση τον προηγούμενο). Επομένως ο αριθμός και η έκταση των διορθώσεων αλλάζει με το χρόνο [155]. Επιπλέον, έχοντας από την αρχή καθορίσει τους γείτονες της κάθε ομάδας (μέσω της ακτίνας R), η νικήτρια ομάδα συμπαρασύρει και τις γειτονικές, των οποίων τα βάρη διορθώνονται ομοίως. Όσο μεγαλύτερη είναι η ακτίνα R, τόσο περισσότεροι γείτονες διορθώνουν τα βάρη τους. Η τιμή R ελαττώνεται μετά από κάθε διόρθωση του ρυθμού εκπαίδευσης. Ο αλγόριθμος εκπαίδευσης θα μπορούσε να συνοψιστεί σε επτά βήματα:

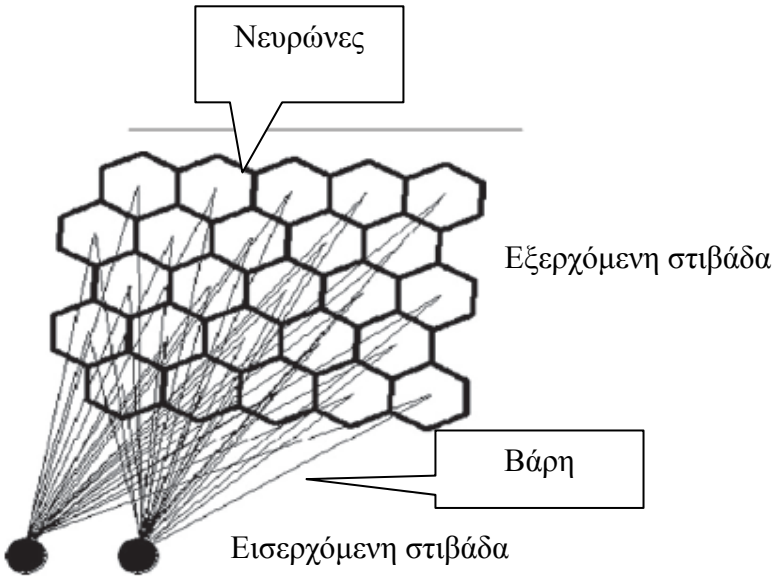
1. Επιλέγουμε βάρη w_{ij} (κυρίως τυχαία) για κάθε νευρώνα (ομάδα) j, τιμή για την ακτίνα R και ρυθμό εκπαίδευσης n.
2. Για κάθε εισερχόμενο διάνυσμα **x** (x_i), εκτελούμε τα βήματα 3 ως 5.
3. Για κάθε νευρώνα j υπολογίζουμε την απόσταση $D(j) = \sum_i (w_{ij} - x_i)^2$ (αν το κριτήριο είναι αυτό).
4. Βρίσκουμε το νευρώνα (νικητή) που έχει τη μικρότερη απόσταση για το συγκεκριμένο δείγμα (διάνυσμα).
5. Για το νικητή j και τους γείτονες του (που καθορίζονται από το R), διορθώνουμε τα βάρη, με βάση τη σχέση: w_{ij} (new) = w_{ij} (old) + n N(t,R) [x_i - w_{ij} (old)].
6. Αφού ελεγχθούν όλα τα δείγματα, διορθώνουμε (ελαττώνουμε) το ρυθμό εκπαίδευσης και την ακτίνα.

7. Ελέγχουμε αν πληρούται ο κανόνας τερματισμού, ο οποίος αφορά συνήθως τον αριθμό των περιόδων [1, 56].

Όπως και στον αλγόριθμο BP (§ 4.3.9), η εκπαίδευση των δικτύων Kohonen μπορεί να γίνει και σε batch mode με ταυτόχρονη παρουσίαση ολόκληρου του πλέγματος των δειγμάτων [148]. Στην περίπτωση αυτή, τα βάρη διορθώνονται με βάση τη μέση τιμή όλων των εισερχομένων για τα οποία ο νευρώνας είναι νικητής, ή ανήκει στη γειτονιά του νικητή.

Κατά τη διάρκεια της εκπαίδευσης, τα δίκτυα Kohonen συμπεριφέρονται ως ένα ελαστικό δίχτυ που "αγκαλιάζει" το "νέφος" που καλύπτει τα εισερχόμενα δείγματα. Καθώς η εκπαίδευση εξελίσσεται, το δίχτυ παίρνει το ακριβές σχήμα του νέφους και όμοια δείγματα αντιστοιχίζονται σε γειτονικούς νευρώνες. [33, 34, 152].

Προσομοιάζοντας με τις αρχιτεκτονικές δομές δικτύων που ήδη αναφέρθηκαν (ώστε να γίνουν καλύτερα κατανοητά), η τυπική δομή ενός δικτύου Kohonen, αποτελείται επίσης από δυο στιβάδες (σχ. 4.43): μια εισερχόμενη (input) και μια εξερχόμενη ενεργή (output ή Kohonen). Η εισερχόμενη περιέχει ένα νευρώνα (μονάδα) για κάθε μεταβλητή. Οι νευρώνες της εξερχόμενης στιβάδας, συνδέονται με προσαρμόσιμα βάρη με τους νευρώνες της εισερχόμενης. Τα βάρη αυτά **αναπαριστούν την κατανομή** [100] ή **διαφορετικά τους ζυγισμένους μέσους [159] των εισερχόμενων διανυσμάτων.** Πιο θεωρητικά, ο Zupan et al. [79] αναφέρει ότι "αν η εκπαίδευση των Kohonen δικτύων είναι αρκετά μεγάλη, τα βάρη **Wj** κάθε νευρώνα j, ισούνται με τη μέση τιμή των m μεταβλητών των s αντικειμένων που ενεργοποίησαν το συγκεκριμένο j νευρώνα ($Wj = \sum x_{1j}/s,\ \sum x_{2j}/s, \ldots\ldots \sum x_{mj}/s$)". Η θέση αυτή, αποδεικνύεται και πειραματικά (ΚΕΦ. 2, Θ).



Σχήμα 4.43: Δισδιάστατη δομή 5x5 ενός δικτύου Kohonen [100].

Η εκπαίδευση γίνεται συνήθως σε δυο φάσεις: αρχικά υπάρχει μια γενική (rough ή ordering [162]), με μεγάλη ακτίνα γειτνίασης και ρυθμό εκπαίδευσης και μετά ακολουθεί μια πιο λεπτομερής (fine tuning), με μικρή ακτίνα και αντίστοιχο ρυθμό εκπαίδευσης. Αυτή η πορεία οδηγεί στη ταξινόμηση των δεδομένων στις ομάδες με τις οποίες έχουν όσο το δυνατό πιο όμοια χαρακτηριστικά διανύσματα (μεταβλητές για τα δεδομένα – βάρη για τους νευρώνες/ομάδες) [163, 164]. Ο Kohonen [165, 166] επίσης διαχωρίζει τις δυο προαναφερθείσες φάσεις ως:

- ✓ τον αρχικό σχηματισμό της σωστής σειράς των νευρώνων και
- ✓ την τελική σύγκλιση.

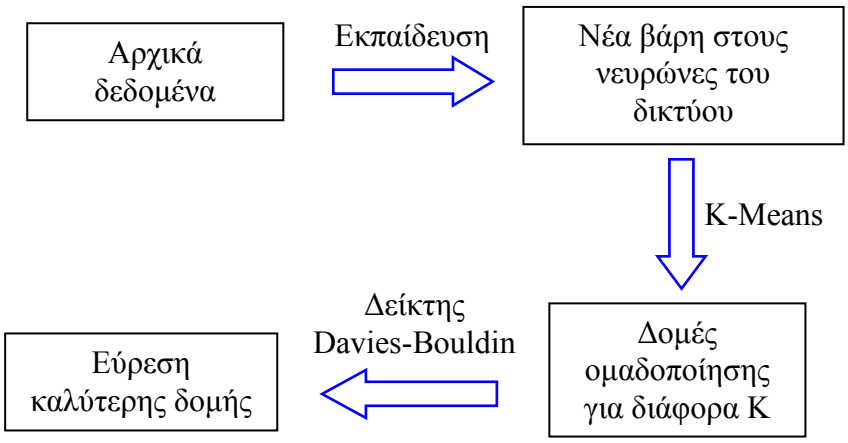
Η δεύτερη φάση διαρκεί πολύ περισσότερο και απαιτεί μικρό ρυθμό εκπαίδευσης. Σε μερικές εφαρμογές δε, απαιτείται ιδιαίτερα υψηλός αριθμός περιόδων.

Η αξιολόγηση των δικτύων Kohonen μπορεί να γίνει με δυο δείκτες που εκφράζουν ουσιαστικά την ελαχιστοποίηση των σφαλμάτων [33, 154, 158, 159, 164, 167]: **τοπογραφικό** (topographic error) και **κβαντικό** (resolution ή quantization error) ( ΚΕΦ. 2, Θ).

4.3.14. Ομαδοποίηση του τοπολογικού χάρτη

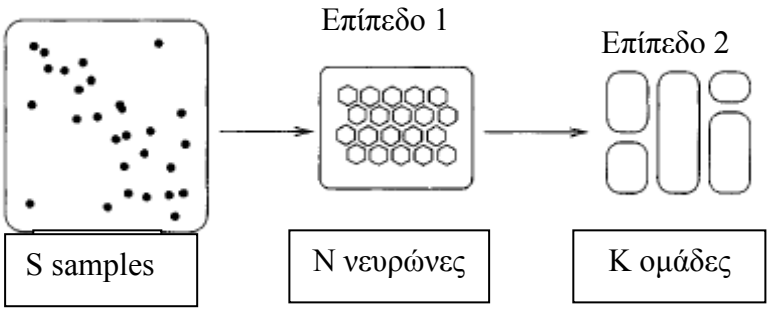
Μέχρι τώρα αναφέρθηκε ότι η ομαδοποίηση στον τοπογραφικό χάρτη Kohonen γίνεται οπτικά με το χαρακτηρισμό των νευρώνων με βάση τον αριθμό των δειγμάτων σε κάθε νευρώνα ή με τον U-matrix που αναπαριστά τις αποστάσεις μεταξύ των νευρώνων (§ 4.3.13, [34]).

Ωστόσο, η διαδικασία αυτή μπορεί να γίνει και με άλλους τρόπους, όπως με τη βοήθεια απλών γνωστών και εξειδικευμένων αλγορίθμων ή μεθόδων. Η συνολική διαδικασία αναφέρεται ως Προσέγγιση Δυο-Επιπέδων ("Two-Level Approach" [33, 34]) και χρησιμοποιεί τα διανύσματα βαρών των νευρώνων. Συνήθως χρησιμοποιείται ο K-means αλγόριθμος (έπεται συνήθως η εφαρμογή του DB δείκτη ( ΚΕΦ. 1, Θ) για την εύρεση του βέλτιστου αριθμού K, σχ. 4.44) [33, 34, 168, 169] ή η CA [168].



Σχήμα 4.44: Διαδικασία εύρεσης της καλύτερης ομαδοποίησης με τον αλγόριθμο K-means [33].

Η πρώτη περίπτωση (χρήση του αλγορίθμου K-means) είναι η πιο συνήθης για τιμές K από 2 μέχρι $\sqrt{S}$ όπου S ο συνολικός αριθμός των δειγμάτων [34], ή N, όπου N ο αριθμός των νευρώνων [168] (σχ. 4.45).



Σχήμα 4.45: Προσέγγιση Δυο-Επιπέδων για την ομαδοποίηση του χάρτη Kohonen [34].

Η ομαδοποίηση του χάρτη Kohonen με τη βοήθεια αλγορίθμων (σε σχέση με την εξ' αρχής εφαρμογή των μεθόδων αυτών στα δεδομένα) έχει πολλά πλεονεκτήματα όπως η μείωση του υπολογιστικού χρόνου, του θορύβου [34], αλλά και του αριθμού των δειγμάτων [169]. Πράγματι τα διανύσματα των βαρών που αντιστοιχούν στους νευρώνες είναι λιγότερο ευαίσθητα σε τυχαίες διακυμάνσεις σε σχέση με τα αρχικά δεδομένα.

4.3.15. Εφαρμογές δικτύων Kohonen

Μέχρι τώρα είδαμε τη χρήση των δικτύων Kohonen για την ομαδοποίηση/ταξινόμηση δειγμάτων στις ομάδες τους αλλά και για την επιλογή των κρισιμότερων μεταβλητών με την αξιοποίηση των χαρτών βαρών (§ 4.3.13).

Ωστόσο, τα δίκτυα Kohonen μπορεί να βρουν και άλλες εφαρμογές όπως για παράδειγμα στην επιλογή της κατάλληλης ομάδας εκπαίδευσης. Χαρακτηριστικά είναι τα παραδείγματα που αναφέρει ο Zupan et al. [79], ο οποίος "προβάλει" τα συνολικά δείγματα σε ένα χάρτη Kohonen και στη συνέχεια επιλέγει ένα δείγμα από κάθε ενεργοποιημένο νευρώνα ώστε να σχηματίσει μια αντιπροσωπευτική ομάδα εκπαίδευσης. Τα υπόλοιπα δείγματα χρησιμοποιούνται στην ομάδα ελέγχου.

Στην ίδια εργασία, με μια μικρή αλλά βασική αλλαγή στη σχέση (4.18) που διέπει την εύρεση του νικητήριου νευρώνα δίνεται η δυνατότητα στα δίκτυα Kohonen να διαχειρίζονται ελλιπή δεδομένα. Συγκεκριμένα, οι αποστάσεις ανάμεσα στα εισερχόμενα δείγματα και τους νευρώνες "κανονικοποιούνται" και η παραπάνω σχέση γίνεται:

$$D(j) = \frac{1}{m-k} \sum_{i=1}^{m-k} (w_{ij} - x_i)^2 \qquad (4.20)$$

όπου οι δείκτες i (i = 1,2,……m) και j (j = 1,2…….n) αναφέρονται στις διαστάσεις του εισερχόμενου δείγματος (διανύσματος) **x** και στο συγκεκριμένο νευρώνα αντίστοιχα, ενώ k είναι ο αριθμός των μεταβλητών που λείπουν σε κάθε δείγμα [79, 170].

Σε άλλη αναφορά ο Lamrini et al. [164], χρησιμοποιεί δίκτυο Kohonen για την "ανακατασκευή" (reconstruction) ελλιπών δεδομένων. Έτσι, ταυτοποιεί αρχικά τα "μη τυπικά" δείγματα με τη μέτρηση των αποστάσεων κάθε εισερχόμενου από το κοντινότερο νευρώνα. Στη συνέχεια, εκτιμάται η τιμή που λείπει ή θεωρείται εσφαλμένη με τη βοήθεια ενός συνδυασμού των βαρών του νικητή-νευρώνα και των k κοντινότερων γειτόνων του.

Ενδιαφέρουσα είναι επίσης μια νεώτερη αναφορά [146], όπου ερευνάται η ικανότητα των δικτύων Kohonen να ταξινομήσουν τα χημικά στοιχεία με δεδομένη την περιορισμένη πληροφορία που είχε ο Mendeleev όταν πρωτοεπινόησε τον περιοδικό πίνακα. Το δίκτυο επιτυγχάνει γενικά το διαχωρισμό των στοιχείων με μικρές αποκλίσεις.

Νεώτερες εργασίες επίσης, χρησιμοποιούν τα δίκτυα Kohonen για την παρακολούθηση και αξιολόγηση μονάδων επεξεργασίας λυμάτων (Wastewater Treatments Plants, WTPs) [33, 169], ή πόσιμου νερού [171] αναλύοντας δεδομένα από τα διάφορα στάδια της διαδικασίας που ακολουθείται.

Τέλος, αναφέρονται εδώ πρόσφατες δημοσιεύσεις που αφορούν **επιβλεπόμενα δίκτυα Kohonen** (Supervised Self-Organising Maps, SSOMs), αλλά δεν θα χρησιμοποιηθούν σε αυτήν την εργασία [151, 160, 161, 167, 172].

4.3.16. Προκατάληψη έναντι διακύμανσης

Έχουμε ήδη συζητήσει τα προβλήματα της υπερ-προσαρμογής (§ 4.3.1), η οποία μπορεί να περιορίσει την ικανότητα των δικτύων για γενίκευση σε νέα δεδομένα. Μια καινοτομική προσέγγιση ωστόσο, που μπορεί να βελτιώσει την απόδοση των δικτύων είναι η δημιουργία **υβριδικών δικτύων** (hybrid [103]) ή stacked [85] ή ensembles [17, 173, 174]). Εδώ, οι προβλέψεις των μεμονωμένων δικτύων συνδυάζονται για να δώσουν μια συνολική πρόβλεψη με τη διαδικασία του μέσου όρου για δίκτυα συσχέτισης (regression) ή με απλή πλειοψηφία για δίκτυα ταξινόμησης (classification) [173, 174]. Συχνά εξάλλου, η δημιουργία των υβριδικών δικτύων συμπληρώνεται με την **επαναδειγματοληψία** (resampling) των δειγμάτων (§ 4.3.2). Αυτή η συνδυασμένη τεχνική μπορεί να βελτιώσει την ικανότητα γενίκευσης των δικτύων.

Πραγματικά, τα σφάλματα που μπορεί να επηρεάσουν την ικανότητα πρόβλεψης των δικτύων είναι δυο ειδών [17, 173]:

1. Αρχικά, ακόμα και ένα τέλειο δίκτυο που μπορεί να προσεγγίσει με τον καλύτερο τρόπο τη συνάρτηση των δεδομένων, μπορεί να σφάλει εξαιτίας του θορύβου.
2. Επιπλέον, υπάρχει το σφάλμα που προέρχεται από το γεγονός ότι πρέπει ένα νευρωνικό μοντέλο να προσαρμοστεί σε μια "πεπερασμένη" ("finite") βάση δεδομένων.

Το τελευταίο σφάλμα διακρίνεται σε δυο συστατικά: την **ακρίβεια** ή **"προκατάληψη"** (bias) του μοντέλου και την **"αμφιβολία"** (ambiguity [174]) ή **"διακύμανση"** (variance [17, 173]). Η προκατάληψη είναι ο μέσος όρος του σφάλματος που θα γίνει στη διάρκεια μιας πορείας εκπαίδευσης με διαφορετικές ομάδες εκπαίδευσης. Η διακύμανση απεικονίζει την επίδραση που έχει στην πορεία εκπαίδευσης η επιλογή μιας συγκεκριμένης ομάδας εκπαίδευσης (βλ. επίσης § 4.3.1) [17, 173].

Μπορούμε να "εξισορροπήσουμε" προκατάληψη και διακύμανση. Έτσι, αν σε μια ακραία περίπτωση επιλεγεί μια συνάρτηση που αγνοεί πλήρως τα δεδομένα, το μοντέλο θα έχει προφανώς μηδενική διακύμανση, αλλά πολύ υψηλή προκατάληψη. Αυτό συμβαίνει γιατί δεν λαμβάνονται καθόλου υπόψη τα δεδομένα της βάσης. Στο αντίθετο άκρο, όταν επιλεγεί μια πολύπλοκη συνάρτηση που προσαρμόζεται πλήρως στα δεδομένα, το μοντέλο θα δώσει μηδενική προκατάληψη. Αντίθετα, η διακύμανση του μοντέλου είναι πολύ υψηλή καθώς η πολύπλοκη συνάρτηση θα αλλάζει δραστικά σχήμα για να απεικονίζει κάθε φορά τα ακριβή σημεία της βάσης δεδομένων. Υψηλή προκατάληψη και χαμηλή διακύμανση δίνουν τα απλά, μικρής πολυπλοκότητας μοντέλα (γραμμικά ή/και μοντέλα με μικρό αριθμό νευρώνων). Αντίθετα, χαμηλή προκατάληψη και υψηλή διακύμανση δίνουν τα πολύπλοκα μοντέλα [17].

Η παραπάνω ανάλυση σχετίζεται με τα υβριδικά δίκτυα και την επαναδειγματοληψία (λεπτομέρειες περιγράφονται στο ηλεκτρονικό παράρτημα της διατριβής,  ΚΕΦ. 2, Θ).

4.4. ΣΥΝΟΛΙΚΗ ΘΕΩΡΗΤΙΚΗ ΑΠΟΤΙΜΗΣΗ

4.4.1. Συμβατικές τεχνικές και ΑΝΝ

Στο σημείο αυτό, θα γίνει μια θεωρητική αποτίμηση των ΑΝΝ μέσα από τη βιβλιογραφία, εφόσον πειραματικά θα δοκιμαστούν και θα συγκριθούν με τις παραδοσιακές πολυπαραμετρικές τεχνικές στα επόμενα κεφάλαια.

Μερικά από τα πλεονεκτήματα των ΑΝΝ σε σχέση με τις παραδοσιακές τεχνικές είναι:

1. Δεν απαιτείται καμιά (a priory) γνώση για τις συσχετίσεις (είδος συνάρτησης, αριθμός και τιμή των παραμέτρων) ανάμεσα στις εισερχόμενες και εξερχόμενες μεταβλητές. Οι σχέσεις αφομοιώνονται με διαρκή, επαναλαμβανόμενη εκπαίδευση [1, 26, 70, 77 – 79, 134, 137, 175 - 178]. Αντίθετα, οι κλασικές συμβατικές μέθοδοι εύρεσης μοντέλων απαιτούν την υιοθέτηση μιας υποθετικής συνάρτησης (πχ κάποιο δευτεροβάθμιο πολυώνυμο και τον επίπονο ίσως υπολογισμό όλων των παραμέτρων που περιέχονται σε αυτό) [1].
2. Η συνολική διαδικασία εύρεσης του βέλτιστου μοντέλου "αφομοιώνει" όσο το δυνατό λιγότερες μεταβλητές με τη μέγιστη δυνατή πληροφορία. Εξερευνούνται και αξιοποιούνται μόνο οι μεταβλητές που οδηγούν στην ακριβέστερη λύση του προβλήματος [91, 134]. Έτσι, ελέγχεται αποτελεσματικά το πρόβλημα της πληθώρας των μεταβλητών **"curse of dimensionality"** (βλ. § 4.4.3) [22, 78]. Συγκεκριμένα, τα νευρωνικά δίκτυα (όπως και τα Δέντρα Ταξινόμησης [94]) ανήκουν στις **"data driven approaches"**, σε αντιδιαστολή με τις συμβατικές στατιστικές μεθόδους που είναι **"model driven".** Αυτό σημαίνει, ότι στις τελευταίες πρώτα καθορίζεται η δομή του μοντέλου, με τη βοήθεια εμπειρικών ή αναλυτικών μεθόδων και μετά αξιολογούνται οι άγνωστοι παράμετροι. Οι data driven approaches από την άλλη πλευρά, έχουν την ικανότητα να καθορίζουν ποιες μεταβλητές είναι κρίσιμες [76]. Εδώ, οι a priori παραδοχές για το μοντέλο που αναζητείται είναι ελάχιστες και η προσέγγιση πρακτικών προβλημάτων ευκολότερη, καθώς δεν απαιτούνται θεωρητικές εικασίες για τη δομή των δεδομένων που ερευνούνται. Το μόνο πρόβλημα για τις data driven approaches είναι η παρουσία θορύβου που μπορεί να "θολώνει" τα δεδομένα, αλλά ακόμα και τότε ίσως είναι οι μόνες μέθοδοι που μπορούν να αντιμετωπίσουν πραγματικά προβλήματα [69].
3. Τα ΑΝΝ επιτρέπουν και ενθαρρύνουν μη γραμμικές σχέσεις μεταξύ των δεδομένων, με αποτέλεσμα την καλύτερη προσαρμογή σε αυτά, αποδοτικότερη προσέγγιση της επιθυμητής συνάρτησης και ευελιξία [1, 26, 17, 56, 69, 77, 86, 91, 117, 134, 147, 175, 176, 178 - 182]. Τα νευρωνικά δίκτυα εισήγαγαν τα πολύπλοκα μη γραμμικά μοντέλα, που μέχρι τότε χρήστες ανά τον κόσμο αρνούνταν εξαιτίας των αυστηρών περιορισμών

να χρησιμοποιήσουν ως κοινές και άμεσες πρακτικές [76]. Τα γραμμικά μοντέλα πλεονεκτούν γιατί μπορούν εύκολα να εφαρμοστούν, ερμηνευτούν και να γίνουν κατανοητά. Ωστόσο, είναι ακατάλληλα για τον πραγματικό κόσμο, όπου οι θεμελιώδεις μηχανισμοί περιγράφονται από μη γραμμικές συναρτήσεις [69].

4. Παρουσιάζουν ανθεκτικότητα στο θόρυβο, παρέχοντας ακριβείς προβλέψεις ανεξάρτητα από ελλιπή δεδομένα ή την παρουσία αβέβαιων δεδομένων και σφαλμάτων μέτρησης [1, 6, 17, 26, 77, 86, 90, 103, 126, 177, 183-185]. Η παραπάνω ικανότητα των ANN συνοψίζεται στον όρο **"association"** και αφορά ελλιπή και ελαφρά παραποιημένα δεδομένα, κάτι το οποίο είναι σημαντικό για καλή μοντελοποίηση εκτός των περιοχών εκπαίδευσης [1].
5. Προσφέρουν υψηλό βαθμό "παραλληλίας", το οποίο συνεπάγεται γρήγορη επεξεργασία αποτελεσμάτων, και μεγαλύτερη ανοχή σε προβληματική hardware υποστήριξη [26, 177, 183, 184].
6. Η εκπαίδευση, η δυνατότητα της ανάδρασης με το συσχετισμό εισερχομένων/εξερχομένων, η ευκολία στη χρήση και η προσαρμοστικότητα των δικτύων επιτρέπουν την αναβάθμιση των μοντέλων, αλλάζοντας την εσωτερική τους δομή, ανάλογα με τις αλλαγές του "περιβάλλοντος" [17, 26 56, 78, 79, 90, 134, 177, 180, 183]. Εδώ αξίζει να αναφερθεί η πρόσφατη εισαγωγή της μεθόδου **"δυναμική εκμάθηση"** (online learning mode) από τους Prieto και Allen [145], οι οποίοι ενσωμάτωναν τα κάθε φορά εισερχόμενα άγνωστα δείγματα στην ομάδα εκπαίδευσης μετά την επιτυχή ταξινόμησή τους. Συγκεκριμένα, τα δίκτυα Kohonen που οι Prieto και Allen χρησιμοποίησαν για την ανίχνευση και αναγνώριση των σημάτων οδικής κυκλοφορίας επαναεκπαιδεύονταν μετά την εισαγωγή νέων δειγμάτων ("training during execution") και νέα βάρη επαναϋπολογίζονταν για το νικητή νευρώνα και τους γείτονές του. Το σύστημα μπορούσε να προσαρμόζεται σε μικρές αλλαγές που αφορούσαν εξωτερικά γνωρίσματα των σημάτων**.**
7. Η δυνατότητα "γενίκευσης", επιτρέπει την εφαρμογή αγνώστων δεδομένων στο ήδη εκπαιδευμένο μοντέλο [17, 26, 69, 77, 86, 90, 93, 103, 134, 184].
8. Τα Νευρωνικά δίκτυα θεωρητικά μπορούν προσεγγίσουν οποιαδήποτε συνεχή συνάρτηση [69, 78, 92, 93, 99, 107, 134, 143, 176, 186], και να ανιχνεύσουν και τις πιο πολύπλοκες συσχετίσεις [78, 187].
9. Τα Νευρωνικά δίκτυα μπορούν να χρησιμοποιήσουν συνεχείς, αλλά και διακριτές (discrete) μεταβλητές [1, 17].
10. Η λειτουργία τους βασίζεται σε λιγότερες παραδοχές και περιορισμούς για την κατανομή των δεδομένων, από ότι οι παραδοσιακές στατιστικές τεχνικές [6, 56, 76, 79, 86, 93, 176]. Τα Νευρωνικά δίκτυα είναι μη παραμετρικές τεχνικές [102], τα οποία

επιπλέον "ανέχονται" ή "παραβλέπουν" συσχετίσεις μεταξύ των μεταβλητών [51], ενώ παραδοσιακές τεχνικές όπως η PCA "αχρηστεύονται" στην περίπτωση μη συσχετιζόμενων μεταβλητών [8].

11. Η ποικιλία μοντέλων (αρχιτεκτονικών δομών και αλγορίθμων) προσφέρουν πολλές δυνατότητες επίλυσης σε πληθώρα προβλημάτων. Μπορεί εύκολα να αλλάξει για παράδειγμα η πολυπλοκότητα ενός μοντέλου με την αλλαγή της συνάρτησης ενεργοποίησης ή την αρχιτεκτονική του δικτύου [76]. Εξάλλου, οι κλασικές χημειομετρικές μέθοδοι αφορούν στο σύνολό τους ένα μοντέλο μοναδικό και αναπαραγώγιμο (κάτω από σταθερές συνθήκες κατασκευής). Τα ANN παρέχουν ποικίλες εναλλακτικές (ανάλογα με τις αρχικές συνθήκες, βλ. § 5.3.8) [160].
12. Πολλές τεχνικές των νευρωνικών δικτύων μπορούν να "λειτουργήσουν" ανεξάρτητα ("stand-alone executable systems") [86], αξιοποιώντας και επιλέγοντας μεταβλητές, "κατασκευάζοντας" το μοντέλο και επικυρώνοντας αυτό σε επόμενο στάδιο, χωρίς τη βοήθεια άλλων μεθόδων.
13. Τα νευρωνικά δίκτυα μπορούν πολύ εύκολα να χρησιμοποιηθούν σε μονοπαραμετρικά, αλλά και πολυπαραμετρικά συστήματα [76].
14. Τα κλασικά μοντέλα (αριθμητικής απόκρισης) απαιτούν μια ξεχωριστή αναλυτική συνάρτηση για κάθε απόκριση, ενώ τα νευρωνικά δίκτυα μπορούν να κατασκευάσουν ένα πολυδιάστατο διάνυσμα απόκρισης [1].

Ειδικότερα, τα μη γραμμικά δίκτυα Kohonen υπερτερούν από τις παραδοσιακές τεχνικές ομαδοποίησης και μείωσης διαστάσεων. Έχουν την ικανότητα να χειρίζονται από πολύ μικρές βάσεις δεδομένων [188, 189] ως πολύ μεγάλες [152], χωρίς ιδιαίτερες απαιτήσεις για την κατανομή των δεδομένων [188]. Μπορούν πάντα να προβάλλουν μεταβλητές και αντικείμενα (δείγματα) ταυτόχρονα σε ένα δισδιάστατο χώρο [188], διατηρώντας την αρχική δομή [100], ενώ στην περίπτωση της γραμμικής PCA, οι διαστάσεις δεν μπορούν πάντα να μειωθούν με επιτυχία σε δύο. Η PCA οπτικοποιεί απλά τις τάσεις των δεδομένων και παρέχει μια πρώτη εκτίμηση της διαχωριστικής ικανότητας των μεταβλητών [181]. Η δυνατότητα οπτικοποίησης της PCA εκφυλίζεται όταν 3 ή περισσότερες συνιστώσες καλούνται να ερμηνεύσουν ένα σημαντικό ποσοστό διακύμανσης [158]. Επιπλέον, τα μοντέλα PCA είναι γραμμικά και επηρεάζονται εύκολα από έκτροπες τιμές [160], ενώ μπορεί πάντα να υπάρχει ένα κενός "white space" που δεν χρησιμοποιείται σε ένα διάγραμμα συντεταγμένων. Αντίθετα, ο χάρτης Kohonen χρησιμοποιεί όλο το διαθέσιμο χώρο [161].

Η θεωρεία των δικτύων Kohonen εξάλλου, είναι απλή και γενικά κατανοητή, ενώ προσφέρουν επιπλέον πολύτιμη πληροφορία [74].

Έτσι, οι χάρτες βαρών (weights maps), είναι μερικές φορές ανώτεροι των φορτίσεων (loadings) της PCA. Αυτό συμβαίνει γιατί γίνονται καλύτερα αντιληπτές οι συσχετίσεις ανάμεσα στις ομάδες και τις αρχικές μεταβλητές. Τα δίκτυα Kohonen έχουν άριστες δυνατότητες "οπτικοποίησης" των μεταβλητών και ερμηνεύουν καλύτερα την αρχική πληροφορία [155, 160, 190]. Επιπλέον, το δίκτυο Kohonen έχει τη δυνατότητα να λειτουργήσει ως μια "συσκευή μοντελισμού" (με την ευρύτερη έννοια αυτού, βλ. § 4.2.2), με προϋπόθεση τη χορήγηση επαρκούς αριθμού δειγμάτων για τη λειτουργία της εκπαίδευσης [155].

Σε σχέση με την CA εξάλλου, το Kohonen δίκτυο έχει τη δυνατότητα (βλ. § 5.3.10), να κατατάξει ένα νέο άγνωστο δείγμα σε μια περιοχή του ήδη σχηματισμένου χάρτη ή να αναγνωρίσει ότι δεν ανήκει σε καμιά από τις γνωστές ομάδες [17], και να αποκαλύψει πολύτιμες πληροφορίες για τις μεταβλητές [158]. Αυτές οι ιδιότητες αποτελούν σημαντική διαφορά από τη κλασική CA παρότι γενικά παράγουν τις ίδιες ομάδες [191].

Σε αντιδιαστολή με όλα τα παραπάνω πλεονεκτήματα που αναφέρθηκαν σε σχέση με τις παραδοσιακές μη επιβλεπόμενες τεχνικές PCA και CA, τα δίκτυα Kohonen απαιτούν επικύρωση των αποτελεσμάτων με κάποια συνήθως εξωτερικά ανεξάρτητα δείγματα. Ο αλγόριθμος απαιτεί εκπαίδευση και βελτιστοποίηση [157]. Ωστόσο, μετά από την επιτυχή έκβαση της διαδικασίας αυτής, τα αποτελέσματα μπορούν να θεωρηθούν αξιόπιστα καθώς επιβεβαιώνονται με ανεξάρτητες πορείες.

Στα μειονεκτήματα των ANN συγκαταλέγονται τα εξής:

1. Φαινόμενα υπερ-προσαρμογής (βλ. § 4.3.1), συχνά "ταλαιπωρούν" τα ANN μοντέλα με αποτέλεσμα τη φτωχή γενίκευση σε νέα δείγματα [24, 69, 86, 93, 182, 192, 193]. Στις περιπτώσεις που τα ANN χρησιμοποιούνται στη λύση προβλημάτων απλούστερης δομής το φαινόμενο της υπερ-προσαρμογής είναι εντονότερο [55, 78].
2. Η τελική λύση εξαρτάται από τις αρχικές συνθήκες [92, 192], (όπως για παράδειγμα την αρχική διευθέτηση βαρών) του δικτύου, και έτσι το τελικό αποτέλεσμα δεν είναι επαναλήψιμο ή μοναδικό [103, 192]. Το φαινόμενο ονομάζεται **"noisy fitness evaluation problem"** [92, 186] και συνίσταται στη λήψη διαφορετικών αποτελεσμάτων λόγω διαφορετικών αρχικών συνθηκών (βαρών), ακόμα και αν οι υπόλοιπες συνθήκες διατηρούνται σταθερές. Γι' αυτό, ένα και μοναδικό πείραμα δεν είναι ποτέ αρκετό για να αξιολογηθεί μια αρχιτεκτονική δικτύου. Χρειάζονται πάντα πολλαπλά πειράματα και μάλιστα με την αύξηση του αριθμού των δειγμάτων, μειώνονται οι συνέπειες του noisy fitness evaluation problem [92].

   Αν λοιπόν, τα αρχικά βάρη δεν είναι τα κατάλληλα, το δίκτυο δεν θα φτάσει ποτέ στο επιθυμητό αποτέλεσμα, ανεξάρτητα από το πλήθος των περιόδων. Οι διαδικασίες εκπαίδευσης και επικύρωσης φαίνεται να εξαρτώνται πλήρως από τα αρχικά βάρη [134].

Το ίδιο συμβαίνει και για τα δίκτυα Kohonen, των οποίων τα αποτελέσματα φαίνονται πραγματικά να εξαρτώνται από τις αρχικές συνθήκες (αρχικά βάρη και δείγματα εκπαίδευσης) [147].

Γενικότερα, καθώς τα ANN είναι "data driven" τεχνικές και "model-free" είναι πολύ εξαρτώμενα και από τα δείγματα που χρησιμοποιούνται: παρουσιάζουν μεγάλη διακύμανση ανάλογα με την σύνθεση των ομάδων των δειγμάτων (§ 4.3.1, 4.3.16) [69].

3. Χρειάζεται μεγάλη προσοχή στην επιλογή των κατάλληλων εισερχόμενων μεταβλητών, όσον αφορά τον αριθμό αυτών, αλλά και την κρισιμότητα/συσχέτιση μεταξύ τους και με τα εξερχόμενα (βλ. § 4.4.3). Η χρήση πολλών μεταβλητών μπορεί να κάνει τα data driven μοντέλα (βλ. παραπάνω) ασαφή και πολύπλοκα. Έτσι, κάθε ερευνητής πρέπει να αντλήσει χρήσιμη πληροφορία για την κρισιμότητα των μεταβλητών από την ανάλυση ευαισθησίας (§ 4.3.8) των νευρωνικών δικτύων ή να κατασκευάσει μικρότερα δίκτυα (αποφεύγοντας φαινόμενα υπερ-προσαρμογής) με την επιλογή των κατάλληλων μεταβλητών [94].
4. Χρειάζεται επίσης μεγάλη προσοχή στην επιλογή των κατάλληλων εισερχόμενων δειγμάτων (§ 4.3.2). Το παλιό γνωμικό **"garbage in, garbage out"** μπορεί απόλυτα να εφαρμοστεί στη περίπτωση των ANN. Αν τα αρχικά δεδομένα της ομάδας εκπαίδευσης δεν είναι αντιπροσωπευτικά και δεν απεικονίζουν την πραγματική κατάσταση ("real world scenario"), το μοντέλο απλά "συμβιβάζεται" [86].
5. Η διαδικασία βελτιστοποίησης των ANN μοντέλων είναι συχνά χρονοβόρα [86, 176, 192], δεν υπάρχει κάποια πρότυπη μέθοδος εύρεσης της βέλτιστης δομής και πορείες trial and error χρησιμοποιούνται συχνά [69, 193]. Η σύγκλιση κατά την εκπαίδευση μπορεί να είναι αργή [93], και πολλοί παράμετροι πρέπει να εξεταστούν για την εύρεση του καλύτερου μοντέλου (βλ. § 5.3.8).
6. Είναι σημαντικό να γνωρίζει κανείς τη σχετική σημασία των εισερχόμενων μεταβλητών. Αρκετοί συνδυασμοί αυτών πρέπει να δοκιμαστούν, εφόσον η πληθώρα μεταβλητών μπορεί εύκολα να οδηγήσει σε υπερ-προσαρμογή του μοντέλου [115, 134].
7. Το τελικό αποτέλεσμα είναι σε μικρότερο ή μεγαλύτερο βαθμό ένα **"black-box"** [6, 24, 69, 86, 93, 103, 184], αν και το γεγονός αυτό θεωρείται από ορισμένους χρήστες ως πλεονέκτημα [76]. Ο χρήστης πρέπει πραγματικά να μαντέψει τι υπάρχει μέσα σε ένα επιτυχές μοντέλο [179], χωρίς να μπορεί να εξάγει θεωρητική πληροφορία από αυτό ή να ερμηνεύσει τα βάρη που τελικά επιλέγονται [6, 77, 176]. Έτσι τελικά, ο χρήστης δεν ξέρει πραγματικά, αν η ομάδα των δειγμάτων εκπαίδευσης, επικύρωσης και ελέγχου είναι αντιπροσωπευτικές [51].

8. Συνήθως υπάρχουν πολλά τοπικά ελάχιστα στην επιφάνεια σφάλματος. Είναι απίθανο να εγγυηθεί κάποιος, ότι έχει απόλυτα προσεγγίσει το συνολικό ελάχιστο [51, 193].
9. Χρειάζεται συνήθως μεγάλος αριθμός δειγμάτων για την "εκπαίδευση" του δικτύου [1, 51, 91, 192]. Οι Zupan και Gasteiger [1] μάλιστα συμβουλεύουν τον αναγνώστη τους να στραφεί προς πιο παραδοσιακές τεχνικές στην περίπτωση που δεν διαθέτει αρκετά δεδομένα.
10. Η προεκβολή είναι απαγορευτική (§ 4.3.10) [76]. Όταν η ομάδα εκπαίδευσης δεν είναι αντιπροσωπευτική (ευρεία κατανομή των δειγμάτων σε όλα τα επίπεδα), ακόμα και η παρεμβολή μπορεί να δημιουργήσει προβλήματα και πρέπει να γίνεται προσεχτικά [51].
11. Δεν είναι προσαρμόσιμα σε νέα δεδομένα και συνήθως σε τέτοιες περιπτώσεις, απαιτείται η επανεκπαίδευση του δικτύου.

4.4.2. Μέθοδοι επικύρωσης (validation of the models)

Η σύγκριση των παραπάνω μεθόδων (παραδοσιακών πολυπαραμετρικών τεχνικών) και νεώτερων θα επιχειρηθεί στα παρακάτω κεφάλαια με βάση μια σειρά μεθόδων επικύρωσης (models' validation). Ακόμα και μοντέλα προκύπτοντα από την ίδια βασική τεχνική (πχ ANN), θα συγκριθούν μεταξύ τους. Θεωρήθηκε λοιπόν αναγκαίο στο σημείο αυτό, να περιγραφούν οι μέθοδοι αυτές, παρότι έγινε (μια λιγότερο λεπτομερής) αναφορά σε προηγούμενα κεφάλαια (§ 2.1.3, 3.1.2).

Μια από τις σημαντικότερες αρχές των μεθόδων αναγνώρισης προτύπων, είναι η επικύρωση των μοντέλων που "κατασκευάζονται" από τις επιβλεπόμενες τεχνικές. Επικύρωση των μοντέλων εννοείται η αξιολόγηση:

- ✓ του αριθμού των κρίσιμων μεταβλητών ή συστατικών (συνιστωσών) που χρειάζονται για να χαρακτηρίσουν τα αρχικά δεδομένα,
- ✓ του αντιπροσωπευτικού χαρακτήρα των δεδομένων που χρησιμοποιούνται για την "κατασκευή" του μοντέλου και
- ✓ της ικανότητας του μοντέλου για ταξινόμηση αγνώστων δειγμάτων, [26].

Συνήθως ωστόσο, η επικύρωση μιας επιβλεπόμενης στατιστικής τεχνικής επικεντρώνεται στο τελευταίο από τα παραπάνω σημεία. Η ικανότητα ταξινόμησης ενός μοντέλου, αξιολογείται με βάση τις **ικανότητες αναγνώρισης και πρόβλεψης**. Τα μοντέλα μπορούν επίσης να αξιολογηθούν ως προς **την ευαισθησία τους** (sensitivity ή sensibility [7, 20]) ή την εξειδίκευση τους (specificity). Η ευαισθησία ενός μοντέλου ταξινόμησης ορίζεται ως το ποσοστό των αντικειμένων (δειγμάτων) που σωστά ταξινομήθηκαν σε μια ομάδα, και η εξειδίκευση

ως το ποσοστό των αντικειμένων (δειγμάτων) που σωστά δεν συμπεριλήφθησαν σε μια ομάδα (βλ. § 6.3.4) [15, 26, 109, 130, 194 - 197]:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \ \% \qquad (4.21)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \ \% \qquad (4.22)$$

όπου TP (True Positive) και TN (True Negative) **τα αληθώς θετικά και αρνητικά ταξινομημένα δείγματα** αντίστοιχα και FP (False Positive) και FN (False Negative) **τα ψευδώς θετικά και αρνητικά ταξινομημένα δείγματα** αντίστοιχα.

Εδώ πρέπει να σημειωθεί, ότι όταν η αναλογία των σωστά ταξινομημένων δειγμάτων σε σχέση με το σύνολο των δειγμάτων αναφέρεται στο διαθέσιμο δείγμα (και όχι σε μια ομάδα) χρησιμοποιείται γενικά στη βιβλιογραφία η έννοια **της ακρίβειας ταξινόμησης** αντί της ευαισθησίας (βλ. για παράδειγμα, αναφορά [194]). Σε αυτήν την εργασία, οι δυο όροι χρησιμοποιούνται για να περιγράψουν το ίδιο ακριβώς κλάσμα.

Η ιδανική κατάσταση είναι όταν υπάρχουν αρκετά δείγματα, ώστε αυτά να χωριστούν σε ομάδες εκπαίδευσης, επικύρωσης και ελέγχου, με καθεμιά από αυτές να περιέχει αντιπροσωπευτικά αντικείμενα για κάθε τάξη. Η επικύρωση αυτής της μορφής, ορίζεται ως **εξωτερική** (external validation). Η ομάδα ελέγχου εδώ, είναι τελείως ανεξάρτητη από τη διαδικασία ανάπτυξης του μοντέλου (επιλογή μεταβλητών, εκτίμηση παραμέτρων). Η εξωτερική επικύρωση ενός μοντέλου αποτελεί "φθηνή" ("cheap") μέθοδο με την έννοια των ελάχιστων υπολογισμών που απαιτούνται για την εύρεση του σφάλματος. Ωστόσο, θεωρείται μέθοδος με υψηλή διακύμανση, εφόσον αν δεν διαθέτουμε αρκετά δεδομένα η διαφοροποίηση της ομάδας ελέγχου, μπορεί να οδηγήσει σε υπερ-αισιόδοξα/απαισιόδοξα αποτελέσματα (§ 4.3.1, 4.3.16) [198].

Η **διασταυρούμενη επικύρωση** (test sample cross-validation) είναι μέθοδος **εσωτερικής αξιολόγησης** (internal validation) και χρησιμοποιείται όταν ο αριθμός των δειγμάτων δεν είναι αρκετός για να καλύψει τις ομάδες εκπαίδευσης και ελέγχου. Τότε, η ικανότητα πρόβλεψης του μοντέλου, ελέγχεται με τη χρήση μιας υπο-ομάδας από τα αρχικά δεδομένα ως ομάδα εκπαίδευσης και μιας άλλης ως ομάδα ελέγχου. Και οι δυο υπο-ομάδες περιέχουν αντιπροσωπευτικά δείγματα από κάθε τάξη. Η διαδικασία επαναλαμβάνεται πολλές φορές, έτσι ώστε, τα ίδια δείγματα να έχουν την πιθανότητα να μετέχουν ως αντικείμενα εκπαίδευσης αλλά και ελέγχου [26]. Η εκτίμηση του σφάλματος γίνεται με τον υπολογισμό του μέσου όρου όλων των συνιστωσών αυτού από τη συνολική διαδικασία εκπαίδευσης. Σε τεχνικές όπως τα ANN, η διασταυρούμενη επικύρωση θα μπορούσε να οδηγήσει σε μια υπεραισιόδοξη εκτίμηση του σφάλματος καθώς δείγματα που μετέχουν στην κατασκευή του μοντέλου, συμμετέχουν και

στην αξιολόγησή του [13]. Παραλλαγές της διασταυρούμενης επικύρωσης αποτελούν και οι παρακάτω μέθοδοι:

- ✓ Η **"εξαιρουμένου ενός" μέθοδος** (leave-one-out ή jack knife method), που περιγράφηκε σε προηγούμενο κεφάλαιο (§ 2.1.3). Η μέθοδος είναι χρήσιμη για μικρές βάσεις δεδομένων, παρέχοντας τη δυνατότητα μιας ικανοποιητικής διευθέτησης των δειγμάτων [26, 104, 197]. Ωστόσο, μπορεί να οδηγήσει σε φαινόμενα υπερ-προσαρμογής, όταν τα δεδομένα δεν είναι αρκετά. Μπορεί να έχει πολύ καλή απόδοση στην εκτίμηση του σφάλματος συνεχών συναρτήσεων όπως τα MSE ή RMS (§ 4.3.7). Αντίθετα, παρουσιάζει προβλήματα για ασυνεχείς συναρτήσεις σφάλματος όπως ο αριθμός των εσφαλμένα ταξινομημένων δειγμάτων [14]. Θεωρείται "ακριβή" ("expensive") μέθοδος με πολλούς υπολογισμούς, αλλά δεν "ξοδεύει" πολλά δείγματα στην περίπτωση μικρών βάσεων [198].
- ✓ Η **"εξαιρουμένου πολλαπλών" μέθοδος** (leave-multiple-out ή leave-n-out method), που αποτελεί βελτίωση της προηγούμενης και περιλαμβάνει την τυχαία διαίρεση των δειγμάτων σε υπο-ομάδες εκπαίδευσης και ελέγχου με τέτοιο τρόπο, ώστε η τελευταία υπο-ομάδα να περιέχει το 40 – 60 % των δειγμάτων (n δείγματα). Η ικανότητα πρόβλεψης υπολογίζεται έτσι, από ένα μεγάλο αριθμό διαφορετικών διαιρέσεων σε υπο-ομάδες εκπαίδευσης και ελέγχου (περιλαμβάνονται πολλαπλοί συνδυασμοί των n δειγμάτων που κάθε φορά εξαιρούνται). Με αυτόν τον τρόπο, υπάρχουν ίσως λιγότερα δεδομένα για την κατασκευή του μοντέλου, αλλά περισσότερα για την εκτίμηση της ποιότητας του. Η διαδικασία επικεντρώνεται έτσι σε γενικότερα "πρότυπα" δείγματα και μειώνεται η πιθανότητα υπερ-προσαρμογής [26]. Η μέθοδος θεωρείται επίπονη (ακριβή σύμφωνα με τα παραπάνω) και "ξοδεύει" πολλά δείγματα στις ομάδες εκτός της εκπαίδευσης [104].
- ✓ Η **ν-πλάσια διασταυρούμενη αξιολόγηση 1** (v-fold cross-validation type 1), όπου δημιουργούνται ν υπο-ομάδες δειγμάτων, από τις οποίες κάθε φορά οι ν-1 χρησιμοποιούνται ως ομάδα εκπαίδευσης και μια ως ομάδα ελέγχου. Το πλεονέκτημα είναι ότι αυτό επαναλαμβάνεται κυκλικά, ώσπου κάθε υπο-ομάδα να παραμείνει ως ομάδα ελέγχου μια φορά [14, 26]. Αν ωστόσο, το ν είναι πολύ μικρό, το σφάλμα εκτιμάται σε μεγαλύτερες τιμές, καθώς υπάρχει μεγάλη διαφορά ανάμεσα στην ομάδα εκπαίδευσης της διασταυρούμενης αξιολόγησης και της αρχικής ομάδας εκπαίδευσης [14]. Γενικά, όσο μεγαλύτερος είναι ο αριθμός των υπο-ομάδων ν, τόσο λιγότερα δείγματα ξοδεύονται (δεν χρησιμοποιούνται στην ομάδα εκπαίδευσης), αλλά τόσο πιο ακριβή είναι η μέθοδος επικύρωσης. Όταν ο αριθμός των υπο-ομάδων ν ισούται με το συνολικό αριθμό των δειγμάτων, η μέθοδος συμπίπτει με την "εξαιρουμένου ενός" μέθοδο [198].

- ✓ Η **ν-πλάσια διασταυρούμενη αξιολόγηση 2** (v-fold cross-validation type 2), όπου δημιουργούνται 2 υπο-ομάδες δειγμάτων, από τις οποίες η ομάδα ελέγχου περιέχει κάθε φορά 1/ν δείγματα και η ομάδα εκπαίδευσης τα υπόλοιπα. Έτσι, η τελευταία είναι αρκετά μεγάλη, ώστε να περιέχει αντιπροσωπευτικό αριθμό δειγμάτων και η πρώτη επαρκή. Η διαδικασία επαναλαμβάνεται ν φορές με διαφορετικά μέλη σε κάθε υπο-ομάδα, έτσι ώστε όλα τα δείγματα να μετέχουν στην ομάδα ελέγχου τουλάχιστον μια φορά [26].

Λιγότερο συνηθισμένες μέθοδοι επικύρωσης των ANN μοντέλων αναφέρονται στο ηλεκτρονικό παράρτημα της διατριβής ( ΚΕΦ. 2, Θ).

Γενικότερα, η ικανότητα αναγνώρισης ενός μοντέλου, είναι καλύτερη από την ικανότητα πρόβλεψης. Ωστόσο, αν αυτές είναι πολύ διαφορετικές μεταξύ τους, αυτό σημαίνει ότι το μοντέλο εξαρτάται πολύ περισσότερο από τα αντικείμενα (δείγματα) της ομάδας εκπαίδευσης και έτσι η λύση που επιτυγχάνεται στο πρόβλημα που αντιμετωπίζουμε δόθηκε τυχαία και επομένως είναι αναξιόπιστη [26].

4.4.3. Μείωση μεταβλητών

Η **επιλογή των μεταβλητών** (“feature selection”) μπορεί να οριστεί ως το πρόβλημα της εύρεσης του υποσυνόλου των μεταβλητών αυτών που ικανοποιούν επαρκώς την επίτευξη του στόχου, αν υποθέσουμε ότι διαθέτουμε όλες τις μεταβλητές που είναι απαραίτητες [44]. Κατά την επιλογή των μεταβλητών, πρέπει να αποφθεχθεί η χρήση περισσοτέρων ή πολύ λίγων μεταβλητών. Αν οι μεταβλητές είναι λίγες, η παροχή της πληροφορίας μπορεί να είναι ανεπαρκής. Αν αντίθετα οι μεταβλητές είναι πολλές, μερικές από αυτές μπορεί να παρέχουν άσχετη πληροφορία. Συνεπώς προσδίδουν θόρυβο στα δεδομένα ο οποίος καλύπτει τη χρήσιμη πληροφορία. Έτσι, η σημαντικότερη πρόκληση που προκύπτει στο θέμα της επιλογής των μεταβλητών είναι η εξισορρόπηση ανάμεσα στο θόρυβο και τη χρήσιμη πληροφορία. Ο συνδυασμός των μεταβλητών αποτελεί επίσης ένα κρίσιμο θέμα (βλ. επίσης § 4.3.8). Κάποιες μεταβλητές μπορεί να παρέχουν σημαντικότερη πληροφορία όταν συνδυάζονται με κάποιες άλλες. Έτσι, οι επιλεχθείσες μεταβλητές πρέπει να θεωρούνται μέρος του συνόλου των μεταβλητών που χρησιμοποιούνται τελικά και σχηματίζουν τη βέλτιστη ομάδα αυτών (“best set”). Όλοι λοιπόν οι δυνατοί συνδυασμοί μεταβλητών πρέπει να εξετάζονται, ώστε να επιτευχθεί το βέλτιστο αποτέλεσμα.

Ειδικότερα, η μείωση των μεταβλητών θεωρείται μέγιστης σημασίας για την ανάπτυξη των ANN μοντέλων. Η παρουσίαση και χρήση όλων των δυνατών μεταβλητών στην κατάσκευή των δικτύων, μπορεί να δημιουργήσει μεγάλα προβλήματα ακόμα και αν η επιλογή των

πιο κρίσιμων από αυτές, ανατεθεί αποκλειστικά στο ίδιο το δίκτυο [100]. Υπάρχουν μεταβλητές οι οποίες σίγουρα δεν συνεισφέρουν στην "δράση" των νευρώνων του δικτύου είτε γιατί είναι αρκετά σταθερές, είτε τα βάρη που τους αποδόθηκαν κατά τη διάρκεια της εκπαίδευσης είναι πολύ μικρά και επομένως μη σημαντικά [58]. Αποτέλεσμα είναι η εκδήλωση φαινομένων υπερ-προσαρμογής για την αποφυγή των οποίων, πρέπει να ακολουθείται η αρχή της **"φειδωλότητας"** ("parsimony principle" ή "Occam razor") [69, 70]. Τα **"φειδωλά" μοντέλα** όχι μόνο έχουν μεγαλύτερη δυνατότητα αναγνώρισης (§ 4.2.2) αλλά και γενίκευσης (§ 4.3.1) [69]. Έτσι, ανάμεσα σε ισοδύναμα, πρέπει να επιλέγεται το μοντέλο με το λιγότερο αριθμό μεταβλητών [26, 70, 91]. Ενδεικτικά θα μπορούσαν να αναφερθούν μερικά μόνο από τα μειονεκτήματα των μοντέλων με πολλές μεταβλητές όπως:

- ✓ μεγάλη computational πολυπλοκότητα (λίγο μας απασχολεί στις μέρες μας και περισσότερο σχετίζεται με τη δυσκολία "κτίσιμου" του μοντέλου και όχι εφαρμογής του [62]) και απαιτήσεις μνήμης [44, 94, 192, 199 - 201],
- ✓ αυξημένες απαιτήσεις αριθμού δειγμάτων για να ανταποκριθούν στις πολλές μεταβλητές (curse of dimensionality) [17, 76, 101, 201],
- ✓ δυσκολία στην εκπαίδευση [199, 200],
- ✓ κακή σύγκλιση και φτωχή απόδοση των μοντέλων [155, 193, 199, 200, 202],
- ✓ αυξημένη πολυπλοκότητα του μοντέλων και συνεπώς [199, 200, 202],
- ✓ δυσκολία στην κατανόηση τους [44, 193, 199, 200] εξαιτίας και της μη δυνατότητας οπτικοποίησης των δεδομένων [155], και
- ✓ μεγάλος χρόνος εκτέλεσης του μοντέλου για τα άγνωστα δείγματα [76],
- ✓ αυξημένος θόρυβος λόγω του συνυπολογισμού μη σημαντικών μεταβλητών [44, 88, 94, 199, 200] και επομένως
- ✓ κίνδυνος υπερ-προσαρμογής του μοντέλου [88, 201, 203],
- ✓ αύξηση της πιθανότητας για τυχαία συσχέτιση (§ 4.4.2) [26, 88],
- ✓ παρεμπόδιση στην εύρεση του βέλτιστου μοντέλου [88].

Ο Bowden et al. [199, 200] προτείνει δυο μεθοδολογίες για την επιλογή των λιγότερων μεταβλητών για ένα **"φειδωλό"** (parsimonious) μοντέλο:

1. Το πρώτο αναφέρεται σε προβλήματα πρόβλεψης και χρησιμοποιεί ένα κριτήριο μέτρησης της **"αμοιβαίας πληροφορίας"** (mutual information), ώστε να βρεθούν οι μεταβλητές με τη μεγαλύτερη συσχέτιση με την εξερχόμενη (προβλεφθείσα) μεταβλητή.
2. Το δεύτερο χρησιμοποιεί ένα μοντέλο Kohonen για την ομαδοποίηση των συσχετιζόμενων μεταβλητών. Στη συνέχεια επιλέγεται μία μεταβλητή από κάθε ομάδα.

Η συνήθης πρακτική ωστόσο, που προτείνεται από πολλούς συγγραφείς και ερευνητές στον κόσμο για τη μείωση των μεταβλητών, είναι η χρήση της PCA ως τεχνική για την προ-

επεξεργασία των δειγμάτων [26, 83, 88, 122, 126, 127, 137, 155, 178, 185, 204 - 207]. Στις περιπτώσεις αυτές, τα scores από την PCA χρησιμοποιούνται ως τις νέες εισερχόμενες "ορθογώνιες" μεταβλητές. Όταν η PCA χρησιμοποιηθεί για την ομαδοποίηση των μεταβλητών, μπορεί να επιλεχθεί μία μόνο μεταβλητή από κάθε ομάδα, η οποία και αντιπροσωπεύει τις υπόλοιπες [162]. Οι συνήθεις αντιρρήσεις που αφορούν τη χρήση της PCA εδώ, αφορούν την πιθανή απώλεια της αρχικής ταυτότητας των μεταβλητών. Έτσι, καθίσταται δύσκολη η ερμηνεία των αποτελεσμάτων και η εξαγωγή συμπερασμάτων από τα ANN μοντέλα. Επιπλέον, οι νέες μεταβλητές (κάποιες PC συνιστώσες) ενδέχεται να "περιέχουν" μειωμένη πληροφορία σε σχέση με τις αρχικές μεταβλητές [58].

Η τεχνική της DA, σπάνια έχει χρησιμοποιηθεί ως μέθοδος επιλογής και επομένως μείωσης των μεταβλητών. Συγκεκριμένα, αναφέρεται μια περίπτωση [195], όπου χρησιμοποιείται η FW προσέγγιση (DA), για την επιλογή των μεταβλητών εκείνων (μέταλλα – μεταλλοειδή) που θα χρησιμοποιηθούν για τη διαφοροποίηση κρασιών με MLP και BP αλγόριθμο.

Εξίσου σπάνια, αναφέρονται στη βιβλιογραφία η χρήση δεικτών όπως ο συντελεστής Pearson για την εύρεση των συσχετιζόμενων μεταβλητών. Έτσι, η μία από αυτές θα είναι η εξερχόμενη μεταβλητή (στην περίπτωση προβλημάτων πρόβλεψης) και οι υπόλοιπες εισερχόμενες [91].

Αναφορά θα πρέπει επίσης να γίνει και σε άλλες τεχνικές εύρεσης των σημαντικότερων μεταβλητών. Για παράδειγμα ο Lei et al. [201], επιλέγει τις μεταβλητές εκείνες που κάνουν την απόσταση μεταξύ των δειγμάτων της ίδιας ομάδας μικρότερη, ενώ την αντίστοιχη ανάμεσα σε διαφορετικές ομάδες μεγαλύτερη. Ο Faisal et al. [208] χρησιμοποιεί δίκτυα Kohonen για την εύρεση των πιο σημαντικών μεταβλητών. Εξαρτημένες και ανεξάρτητες μεταβλητές χρησιμοποιούνται για την κατασκευή του χάρτη Kohonen και όποιες από τις πρώτες φαίνονται να παίζουν σημαντικό ρόλο στην ομαδοποίηση φυλάσσονται για την "τροφοδότηση" του μοντέλου MLP που ακολουθεί.

Βηματικές προσεγγίσεις ανάλογες της DA (βλ. § 5.2.2) μπορούν επίσης να εφαρμοστούν και για τα ANN. Ειδικότερα, η FW προσέγγιση θεωρείται ότι είναι γρηγορότερη, αλλά μπορεί να παραβλέψει μεταβλητές που αλληλοεξαρτώνται, ενώ αντίθετα η BW πλεονεκτεί και εφαρμόζεται συχνότερα, ειδικά σε περιπτώσεις λίγων μεταβλητών [17, 203]. Μια διαφορετική βηματική προσέγγιση περιλαμβάνει την εκπαίδευση διαφορετικών μοντέλων με καθεμιά μεταβλητή και στη συνέχεια τη διαδοχική προσθήκη μεταβλητών στο βέλτιστο μοντέλο. Η μέθοδος είναι χρονοβόρα και ανεπαρκής να "συλλάβει" τη σημαντικότητα συνδυασμών κάποιων μεταβλητών, που από μόνες τους μπορεί να δείχνουν ασήμαντες (§ 4.3.8) [76]. Κριτήρια που αφορούν κυρίως την απόδοση του δικτύου (§ 4.3.7) χρησιμοποιούνται τελικά για την εύρεση των κρισιμότερων μεταβλητών [120].

Ο **γενετικός αλγόριθμος** (genetic algorithm, GA) είναι μια άλλη μέθοδος για την επιλογή των βέλτιστων μεταβλητών. Τέτοιου είδους αλγόριθμοι θεωρούνται ότι προσομοιώνουν τη βιολογική εξέλιξη των ειδών και στο γεγονός αυτό οφείλεται και το όνομά τους. Πράγματι, η βιολογική εξέλιξη η οποία έχει αποδειχθεί ιδιαίτερα επιτυχής διαμέσου των αιώνων, στηρίζεται στους παρακάτω μηχανισμούς:

1. τους **κλασικούς κανόνες του Darwin** για τον αγώνα για επιβίωση (competition rule) και την επικράτηση του καλύτερου (selection rule),
2. εφαρμόζεται **στην "κωδικοποιημένη" μορφή ζωής** δηλαδή τα χρωμοσώματα. Εισάγονται δε τυχαίες αλλαγές μέσω της φυσικής διαφοροποίησης ("natural mutation").

Κατά την προσομοίωση της βιολογικής εξέλιξης σε αλγορίθμους βελτιστοποίησης, πρέπει ανάλογα να προβλεφθούν:

1. μια τεχνική κωδικοποίησης για τις προτεινόμενες λύσεις του προβλήματος,
2. ανταγωνισμός (competition) ανάμεσα στις προτεινόμενες λύσεις,
3. συνδυασμοί μεταξύ των επιβιωσάντων λύσεων, ώστε να προκύψουν νέες **γενιές** ("generations") καλύτερων λύσεων,
4. τυχαίες αλλαγές [209].

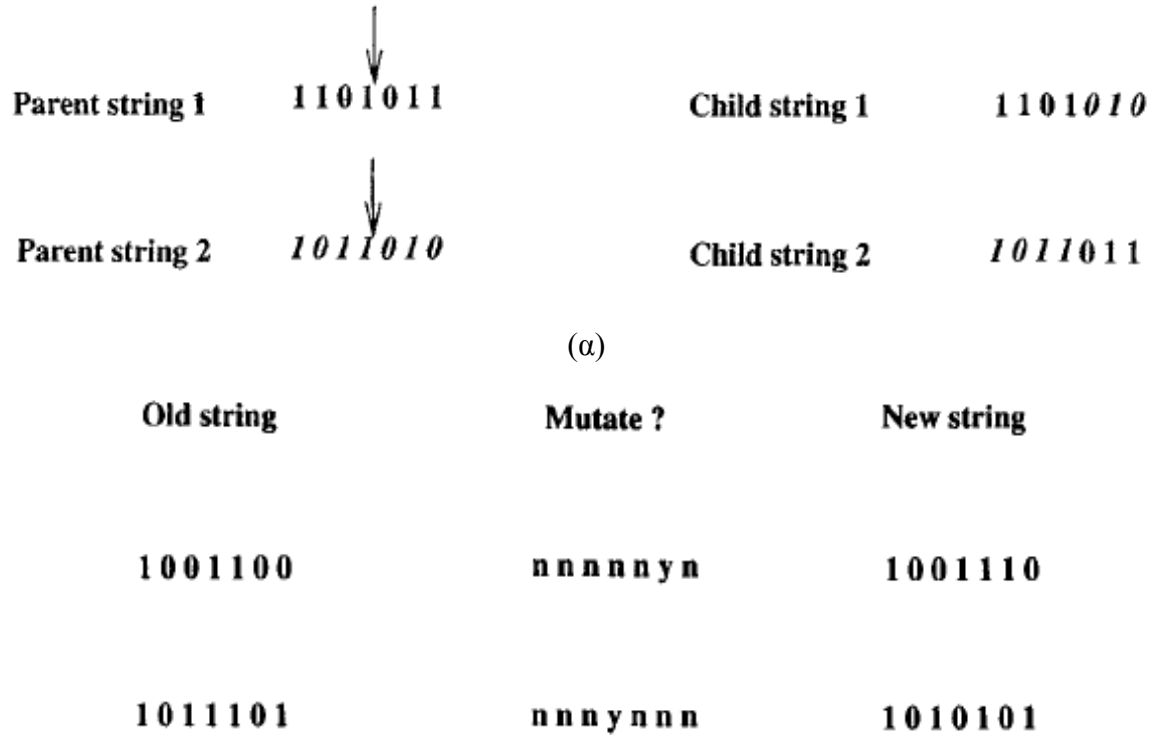
Όσον αφορά την τεχνική κωδικοποίησης, ο GA ερευνά για **δυαδικές αλληλουχίες ή χρωμοσώματα** (strings ή chromosomes) που αντιπροσωπεύουν τις μεταβλητές που θα μπορούσαν δυνητικά να χρησιμοποιηθούν για την κατασκευή του μοντέλου. Έτσι με "0" αναπαρίστανται οι μεταβλητές που δεν μπορούν να χρησιμοποιηθούν, ενώ με "1", εκείνες που πρέπει να συμμετέχουν στο μοντέλο. Για παράδειγμα, ένα string "001101" δείχνει ότι από τις έξι διαθέσιμες μεταβλητές, η πρώτη, δεύτερη και πέμπτη πρέπει να απορριφθούν, ενώ η τρίτη, τέταρτη και έκτη να παραμείνουν. Ο GA δημιουργεί τυχαία μια σειρά από τέτοια strings ("first generation") και αφού δοκιμαστούν επιλέγονται τα καλύτερα (ανταγωνισμός ανάμεσα στις προτεινόμενες λύσεις) [17].

Τα τελευταία υπόκεινται σε παραπέρα bit-by-bit βελτίωση. Η διαδικασία αυτή περιλαμβάνει **συνδυασμούς** ("cross-over" ή "recombination") των προτεινόμενων λύσεων και **μεταλλάξεις** ("mutation") [140, 186, 209, 210]. Η διαφορά των παραπάνω φαίνεται στο σχήμα 4.46.

Κατά την cross-over διαδικασία (σχ. 4.46(α)), οι δυο προτεινόμενες λύσεις (parent strings) συνδυάζονται έτσι ώστε να προκύψουν δυο νέες λύσεις (child strings). Έτσι, σημαντική πληροφορία που περιέχουν οι αρχικές λύσεις μεταφέρεται στις νεώτερες.

Όταν η παραπάνω διαδικασία ωστόσο, δεν προσφέρει την απαιτούμενη διαφοροποίηση, χρησιμοποιείται η εναλλακτική πορεία mutation (σχ. 4.46(β)). Για παράδειγμα, όταν ένα bit είναι το ίδιο σε όλα τα strings, αυτό δεν μπορεί να αλλάξει με την cross-over διαδικασία. Τότε, ο αλγόριθμος παγιδεύεται σε τοπικά ελάχιστα. Η mutation αντίθετα, εξασφαλίζει τη διαφορο-

ποίηση του πληθυσμού και εμποδίζει την πρώιμη σύγκλιση με την εναλλαγή του 0 σε 1 και αντίθετα [209, 211].



Σχήμα 4.46: Παράδειγμα βελτίωσης των προτεινόμενων λύσεων με διαδικασίες cross-over (α) και mutation (β) [209].

Οι GA είναι πολύ αποτελεσματικοί στην επιλογή μεταβλητών, καθώς αναγνωρίζουν αλληλοσχετιζόμενα bits. Γενικά όμως, είναι χρονοβόρες μέθοδοι, αλλά ο απαιτούμενος χρόνος εξαρτάται από τον αριθμό των μεταβλητών, ενώ αντίθετα οι βηματικές μέθοδοι που προαναφέρθηκαν απαιτούν χρόνο ανάλογο του τετραγώνου του αριθμού των μεταβλητών. Συνήθως χρησιμοποιούνται όταν ο αριθμός των διαθέσιμων μεταβλητών είναι μεγάλος (μεγαλύτερος από πενήντα), ή αποτελούν συμβουλευτική εναλλακτική στην περίπτωση λίγων μεταβλητών. Οι GA είναι ιδιαίτερα αποτελεσματικοί στην ανίχνευση των αλληλοσυσχετίσεων μεταξύ μεταβλητών που τοποθετούνται γειτονικά στο string. Έτσι, στην περίπτωση υποψίας συσχετιζόμενων μεταβλητών, είναι καλύτερα να τοποθετηθούν αυτές σε γειτονικές στήλες πριν την έναρξη του αλγορίθμου [17]. Μειονέκτημα των GA αποτελεί επίσης η ανάγκη βελτιστοποίησης από το χρήστη μιας ομάδας παραμέτρων όπως ο ρυθμός μετάλλαξης (mutation rate), η μορφή των συνδυασμών (cross-over scheme), ο αρχικός πληθυσμός (initial population) [202].

Οι GA έχουν επιτυχώς χρησιμοποιηθεί και σε άλλες εφαρμογές, εκτός της επιλογής των κρισιμότερων μεταβλητών [212]. Ενδιαφέρον έχει η πρόσφατη εργασία του Ballabio et al.

[210], όπου χρησιμοποιούνται GA για την εύρεση του βέλτιστου αριθμού των περιόδων εκπαίδευσης αλλά και του αριθμού των νευρώνων σε CP-ANN (Counter-Propagation ANN) μοντέλα. Εδώ, κάθε string είναι δυαδική απεικόνιση των περιόδων αλλά και των νευρώνων του μοντέλου. Ο Feng et al. [213] επίσης, χρησιμοποιεί GA για τη βελτιστοποίηση βαρών και της προκατάληψης στα BP-ANN μοντέλα που χρησιμοποιεί.

4.4.4. Αλλαγή κλίμακας (scaling) των δεδομένων

Όπως και με τις παραδοσιακές πολυπαραμετρικές στατιστικές τεχνικές, έτσι και πριν την εφαρμογή των ANN, είναι μερικές φορές απαραίτητο η αλλαγή της κλίμακας (scaling) των αρχικών δεδομένων. Εφαρμόζεται κυρίως όταν υπάρχει μεγάλη διαφορά στο εύρος των μεταβλητών. Τότε δεν μπορεί να εξασφαλιστεί η συμπεριφορά του μοντέλου απέναντι στις μεταβλητές, αλλά αντίθετα οι "μεγάλες" μεταβλητές κυριαρχούν στην κατασκευή του. Επιπλέον οι μεγάλες μεταβλητές σε ένα νευρώνα, συνδυασμένες με μεγάλα βάρη μπορεί να οδηγήσουν το δίκτυο σε παράλυση (βλ. § 4.3.9). Η παράγωγος της σιγμοειδούς συνάρτησης θα είναι τότε πολύ μικρή και η διόρθωση των βαρών ελάχιστη (σχέση 4.14). Όταν λοιπόν, οι τιμές της συνάρτησης κινούνται από 0 ως +1, τα δεδομένα προσαρμόζονται συνήθως στο εύρος 0,1 – 0,9 ή 0,2 – 0,8 (ώστε να βρίσκονται στη γραμμική περιοχή της σιγμοειδούς συνάρτησης (§ 4.3.4) [1]. Επιπλέον, αν οι αλλαγές αυτές γίνουν μέχρι τα πιο ακραία όρια της συνάρτησης ενεργοποίησης, οι διορθώσεις των βαρών είναι ελάχιστες [76]. Η αλλαγή της κλίμακας, επαναφέρει τις μεταβλητές στο κατάλληλο εύρος της σιγμοειδούς συνάρτησης (βλ. § 4.3.4) [51]. Εδώ πρέπει να τονιστεί ότι αν η συνάρτηση είναι μη φραγμένη (πχ γραμμική), η αλλαγή κλίμακας δεν απαιτείται αυστηρά, αλλά ωστόσο συνίσταται η χρήση ομοιόμορφων μονάδων [76]. Ο Zhang et. al. [69] τέλος, υποστηρίζει ότι η αλλαγή στην κλίμακα των δεδομένων πρέπει να γίνεται ακόμα και σε περίπτωση γραμμικής συνάρτησης, ώστε να αποφεύγονται υπολογιστικά προβλήματα, να εκπληρώνονται τυχόν απαιτήσεις των αλγορίθμων και να διευκολύνεται η εκπαίδευση του δικτύου. Τα πλεονεκτήματα από την αλλαγή της κλίμακας των δεδομένων ωστόσο, μειώνονται όσο το δίκτυο μεγαλώνει.

Οι πιο συνήθεις μέθοδοι αλλαγής κλίμακας είναι [26]:

1. **Mean-centering:** η μέση τιμή αφαιρείται από κάθε μεταβλητή.
2. **Τυποποίηση** (standardization ή auto-scaling): η μέση τιμή αφαιρείται σε κάθε μεταβλητή (mean-centered), μετά διαιρείται με την τυπική απόκλιση.
3. **Κανονικοποίηση** (normalization): οι μεταβλητές διαιρούνται με την τετραγωνική ρίζα του αθροίσματος όλων των μεταβλητών για κάθε δείγμα.

Άλλες μέθοδοι αναφέρονται στο ηλεκτρονικό παράρτημα της διατριβής (ΚΕΦ. 2, Θ). Αρκετές φορές επίσης, η προκατεργασία των τιμών των μεταβλητών, είναι μέρος της χημειομετρικής τεχνικής (πχ PCA) [26].

4.5. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

Η Χημειομετρική ανάλυση δεδομένων προσφέρει ένα πανίσχυρο εργαλείο στην ανάλυση και ερμηνεία μεγάλου αριθμού περιβαλλοντικών και όχι μόνο πολυπαραμετρικών δεδομένων. Έτσι, όσον αφορά τις πιο συμβατικές πολυπαραμετρικές τεχνικές:

- ✓ στην ανίχνευση και ερμηνεία συσχετίσεων, στην ταυτοποίηση πηγών ρύπανσης, στην μελέτη γεωγραφικών ή εποχιακών κατανομών τους χρησιμοποιούνται κατά κύριο λόγο οι "παραδοσιακές" PCA/FA.
- ✓ στην ταξινόμηση δειγμάτων, θέσεων (σημείων δειγματοληψίας), αλλά και μεταβλητών, χρησιμοποιούνται οι PCA, CA, ενώ
- ✓ στην αξιολόγηση μεταβλητών, στην εξαγωγή ενός "μοντέλου" ταξινόμησης ή στην αναζήτηση γεωγραφικών και εποχιακών διακυμάνσεων χρησιμοποιείται η DA [38].

Στην παρούσα εργασία, επιχειρήθηκε η σύγκριση των παραπάνω κλασικών χημειομετρικών τεχνικών με νέες πολυπαραμετρικές μεθόδους όπως τα Νευρωνικά Δίκτυα (ANN). Παράλληλα, μελετήθηκε σε μεγαλύτερο βαθμό η τεχνική των CT (Δέντρων ταξινόμησης), με σύγχρονη εφαρμογή τριών προσεγγίσεων αυτής της μεθόδου. Τα πεδία σύγκρισης ήταν μεγάλες βάσεις δεδομένων όπως:

1. Μέταλλα-μεταλλοειδή από τους τρεις ταμιευτήρες που χρησιμοποιούνται για την ύδρευση της πρωτεύουσας (Υλίκη, Μόρνο και Μαραθώνα) (ΚΕΦ. 5).
2. Μέταλλα-μεταλλοειδή, ανόργανα στοιχεία από δείγματα ιζημάτων σε μεγάλες ιχθυοκαλλιέργειες της χώρας (ΚΕΦ. 6).
3. Σπάνιες γαίες (λανθανίδες) από δείγματα εδώδιμων λαδιών (ελαιολάδων) από διάφορες περιοχές της χώρας (ΚΕΦ. 7).

Για την πρώτη βάση δεδομένων που μελετήθηκε, θα μπορούσε κανείς να αναφέρει ένα πλήθος άρθρων που αφορούν την εφαρμογή ANN στην ανάλυση του νερού. Έτσι, βελτιστοποιημένα μοντέλα συνήθως MLP, RBF και Kohonen χρησιμοποιούνται συχνά για τη συσχέτιση δεδομένων [57, 66, 70, 72, 133, 134, 214], την εκτίμηση της ποιότητας του νερού [154, 188, 189, 214, 215], την ταξινόμηση δειγμάτων [190, 215], ή τη διερεύνηση της κρισιμότητας και συσχετίσεων μεταβλητών [190, 214, 215]. Συχνά τα ANN συγκρίνονται με πιο συμβατικές τεχνικές, με τα πρώτα να αποδεικνύονται πολύ αποτελεσματικότερα [154, 214].

Ειδικότερες αναφορές γίνονται στην εισαγωγή του σχετικού κεφαλαίου (§ 5.1). Παρά τον αριθμό των δημοσιευμένων εργασιών ωστόσο, είναι η πρώτη φορά που τόσες πολλές πολυπαραμετρικές τεχνικές (επιβλεπόμενες και μη), συγκρίθηκαν μεταξύ τους. Συγκεκριμένα, τέσσερις μη επιβλεπόμενες τεχνικές (PCA/FA, CA, Kohonen) και τέσσερις επιβλεπόμενες (DA, CT, MLP και RBF) εφαρμόστηκαν και αξιολογήθηκαν για την ίδια βάση δεδομένων. Κλασικές και νεώτερες μέθοδοι αξιοποιήθηκαν σε μια μεγάλη και "εύφορη" βάση δεδομένων και τα αποτελέσματά τους (ποιοτικά και ποσοτικά) ερμηνεύτηκαν διεξοδικά.

Για τη δεύτερη βάση δεδομένων που χρησιμοποιήθηκε (δείγματα θαλασσίων ιζημάτων από ιχθυοκαλλιέργειες), επιχειρήθηκε παράλληλα με τη σύγκριση πολυπαραμετρικών μεθόδων που αναφέρθηκε παραπάνω, η εξαγωγή συμπερασμάτων για την επίπτωση που έχει η λειτουργία μιας μονάδας ιχθυοκαλλιέργειας στο φυσικό περιβάλλον που εγκαθίσταται. Ανόργανα συστατικά (N, P, C) που περιέχονται στις τεχνητά παραγόμενες και εξωτερικά προστιθέμενες ιχθυοτροφές και περιττώματα ψαριών, φαίνεται ότι επηρεάζουν τη σύσταση του θαλάσσιου πυθμένα που γειτονεύει με τέτοιες εγκαταστάσεις [216].

Πολλές εργασίες που καταγράφουν την επίδραση των ιχθυοκαλλιεργιών στον πυθμένα του θαλάσσιου περιβάλλοντος, έχουν γενικά καταγραφεί στη διάρκεια των τελευταίων δυο δεκαετιών. Ο P φαίνεται να είναι το κυρίαρχο ανόργανο συστατικό στις μελέτες αυτές [216, 217]. Αυξημένα επίπεδα βαρέων μετάλλων όπως Cu, Zn, Cd καταγράφονται σε άλλες εργασίες [218, 219, 220 - 222]. Ωστόσο, λίγες από αυτές τις εργασίες, χρησιμοποιούν στατιστικές τεχνικές για την αξιολόγηση των αποτελεσμάτων. Πολύ βασικές μέθοδοι και δοκιμές χρησιμοποιούνται για τον σκοπό αυτό [219, 223 - 225]. Ειδικότερες αναφορές γίνονται στην εισαγωγή του σχετικού κεφαλαίου (§ 6.1).

Στη διατριβή αυτή, μελετήθηκαν διεξοδικά τα περιβαλλοντικά δεδομένα του προβλήματος, ενώ παράλληλα χρησιμοποιήθηκαν μια πληθώρα από επιβλεπόμενες και μη, πολυπαραμετρικές τεχνικές (PCA/FA, DA, CT, ANN) για την ερμηνεία και αξιολόγηση των αποτελεσμάτων.

Για την τρίτη βάση δεδομένων που αφορά την προέλευση εδώδιμων λαδιών, τα αποτελέσματα της επιτυχούς κατάταξής τους, θα μπορούσαν να φανούν χρήσιμα για μια μελλοντική ταυτοποίηση αγνώστων ή αμφιβόλων δειγμάτων. Γενικότερα, συγκεκριμένες παράμετροι μπορούν να χρησιμοποιηθούν ως ένα δικανικό (forensic) εργαλείο για να ανιχνευθούν δυνητικές προσπάθειες νοθείας στην αγορά των τροφίμων και όχι μόνο. Έτσι, ANN έχουν συχνά χρησιμοποιηθεί για την ταυτοποίηση της γεωγραφικής/βοτανικής προέλευσης και ποικιλίας:

1. δειγμάτων κρασιών με βάση τη μεταλλική τους σύσταση [15, 195], των φαινολικών ενώσεων που περιέχουν [22, 24], ή τα πτητικά οργανικά συστατικά τους [207],

2. ποικιλιών Ιταλικού ρυζιού με βάση κοινά χαρακτηριστικά που αφορούν το μέγεθος του κόκκου ή ποιοτικούς δείκτες της παραγωγικής διαδικασίας [110],
3. ειδών αλκοολούχων ροφημάτων με βάση μέταλλα: Cu, Fe και Zn [23, 129],
4. δειγμάτων βενζίνης με βάση το NIR φάσμα [62],
5. δειγμάτων εμφιαλωμένου μεταλλικού νερού με βάση χημικά και γεωλογικά χαρακτηριστικά τους [226],
6. δειγμάτων μελιού με βάση πτητικά συστατικά τους [187], τη μεταλλική τους σύσταση [227], ή διάφορες εύκολα μετρούμενες χημικές παραμέτρους (ολική οξύτητα και ειδική στροφική ικανότητα) [228],
7. δειγμάτων γάλακτος (από διάφορα θηλαστικά) με βάση το φάσμα μάζας τους [28],
8. δειγμάτων μπύρας με βάση GC profiles των πτητικών συστατικών τους [229].

Ειδικότερες αναφορές για το ελαιόλαδο και τις στατιστικές τεχνικές που έχουν χρησιμοποιηθεί, γίνονται στην εισαγωγή του σχετικού κεφαλαίου (§ 7.1). Ωστόσο είναι η πρώτη φορά, που έγινε προσπάθεια να επιτευχθεί ο γεωγραφικός χαρακτηρισμός των ελαιολάδων με βάση την περιεκτικότητά τους σε σπάνιες γαίες. Επιπλέον μια καινοτομική προσέγγιση επιχειρήθηκε για την αντιπροσωπευτικότερη επιλογή των ομάδων εκπαίδευσης και ελέγχων των μοντέλων CT και ANN που χρησιμοποιήθηκαν.

ΣΚΟΠΟΣ ΕΡΓΑΣΙΑΣ

Σκοπός της εργασίας αυτής, είναι αρχικά η παρουσίαση των νέων πολυπαραμετρικών τεχνικών: Δέντρων Ταξινόμησης (Classification Trees, CT) και Νευρωνικών Δικτύων (Artficial Neural Networks, ANN), η εφαρμογή τους σε μεγάλες βάσεις δεδομένων και η αξιολόγηση των αποτελεσμάτων, όσον αφορά

- την ομαδοποίηση/ταξινόμηση των αντικειμένων (δειγμάτων) και
- την αξιολόγηση μεταβλητών.

Παράλληλα επιχειρήθηκε μια σύγκριση των παραδοσιακών και νεώτερων τεχνικών. Συγκεκριμένα, χρησιμοποιήθηκαν διαφορετικές μέθοδοι CT και διαφορετικά μοντέλα ANN για την ανίχνευση και επιβεβαίωση ομάδων (την πρόβλεψη της πηγής προέλευσης "αγνώστων" δειγμάτων) και την εύρεση των κρισιμότερων μεταβλητών. Τα αποτελέσματα των δοκιμών αυτών συγκρίθηκαν με τις παραδοσιακές: Διαχωριστική Ανάλυση (Discriminant Amalysis, DA), Ανάλυση Κυρίων Συνιστωσών (Principal Components Analysis, PCA) και Ανάλυση κατά συστάδες (Cluster Analysis, CA).

Έτσι, νέοι ερευνητές με δυνητικό ενδιαφέρον για τις τεχνικές αυτές θα μπορούν να τις εφαρμόσουν στις δικές τους βάσεις δεδομένων, βοηθούμενοι όσον είναι αυτό εφικτό, ώστε να ελαχιστοποιηθεί η προσπάθεια και το σφάλμα και να μεγιστοποιηθεί η ακρίβεια του αποτελέσματος και η πληροφορία που τελικά εκμαιεύεται από τη συνολική ανάλυση.

Οι ιδιαιτερότητες της κάθε τεχνικής, οι δυνατότητες και οι περιορισμοί τους εξετάστηκαν θεωρητικά, αλλά και μέσα από τις βάσεις δεδομένων που μελετώνται, και αναπτύχθηκαν τρόποι αντιμετώπισης των αδυναμιών τους.

Τα CT μελετήθηκαν για πρώτη φορά μέσα από τρεις διαφορετικές μεθόδους/παραλλαγές τους (Classic CT, LCM και CART) και ταυτόχρονα επικυρώθηκαν με νέα δείγματα (ελέγχου) και τεχνικές Cross-Validation (CV). Οι χρησιμοποιούμενες μεταβλητές αξιολογήθηκαν και χρησιμοποιήθηκαν οι κρισιμότερες από αυτές.

Τα ANN χρησιμοποιήθηκαν στις ίδιες βάσεις ταξινόμησης με διάφορες αρχιτεκτονικές και αλγορίθμους (επιβλεπόμενους και μη). Προβλήματα όπως η υπερ-προσαρμογή των μοντέλων, η πληθώρα των παραμέτρων, ο αντιπροσωπευτικός διαχωρισμός των δειγμάτων σε ομάδες εκπαίδευσης, επικύρωσης και ελέγχου έπρεπε να αντιμετωπιστούν.

Ειδικότερα, όσον αφορά την πρώτη βάση δεδομένων, είναι η πρώτη φορά που εφαρμόστηκαν πολυπαραμετρικές τεχνικές σε τέτοια έκταση για την εκτίμηση της ποιότητας του νερού (σε μέταλλα/μεταλλοειδή) στους τρεις ταμιευτήρες που χρησιμοποιούνται για την ύδρευση της Αθήνας. Εφαρμόστηκε έτσι, μια γενική στρατηγική για την εκτίμηση της ποιο-

τητας του νερού και τη διαχείριση των υδάτινων πόρων με τη βοήθεια επιβλεπόμενων και μη τεχνικών με σκοπό:

- την ταξινόμηση των δειγμάτων,
- την εύρεση των πιο κρίσιμων μεταβλητών, υπεύθυνων για τη διαφοροποίηση των δειγμάτων,
- την εφαρμογή και σύγκριση τεχνικών CT και ANN μέσω της ικανότητάς τους στην πρόβλεψη των νέων δειγμάτων,
- την κατασκευή μοντέλων για κάθε υδάτινο ταμιευτήρα,
- την ταυτοποίηση διαφορών και ομοιοτήτων μεταξύ των σημείων δειγματοληψίας.

Για τη δεύτερη μεγάλη βάση δεδομένων των ιζημάτων, μόνο βασικές στατιστικές μέθοδοι έχουν χρησιμοποιηθεί ως τώρα στη βιβλιογραφία για την εκτίμηση της επιβάρυνσης εξαιτίας των παρακείμενων ιχθυοκαλλιεργειών. Έτσι στη διατριβή αυτή, με τη βοήθεια προηγμένων πολυπαραμετρικών τεχνικών, εκπληρώθηκαν πέντε βασικοί στόχοι:

- η μελέτη της επίδρασης των ιχθυοκαλλιεργειών στα παρακείμενα θαλάσσια ιζήματα,
- η ταυτοποίηση των πιο σημαντικών ρυπογόνων παραγόντων από το σύνολο των μετάλλων/μεταλλοειδών και ανόργανων στοιχείων που προσδιορίστηκαν,
- η σύγκριση μεταξύ των διαφορετικών μοντέλων CT και ANN πάνω στην ίδια βάση, με κριτήριο τα σωστά ποσοστά ταξινόμησης στα δείγματα εκπαίδευσης και ελέγχου.

Η επιλογή των δειγμάτων στη βάση αυτή ήταν τυχαία, ώστε να χρησιμοποιούνται δείγματα από όλες τις μονάδες ιχθυοκαλλιέργειας. Όλα τα μοντέλα ωστόσο, ανταποκρίθηκαν στην πρόκληση και έδωσαν πολύ υψηλά ποσοστά, επιβεβαιώνοντας την ομοιογένεια των δειγμάτων σε σχέση με τους ρυπαντές που μελετήθηκαν.

Στην τρίτη βάση δεδομένων που περιελάμβανε ελαιόλαδα από διαφορετικές περιοχές της χώρας, δεδομένα σπανίων γαιών (Rare Earth Elements, REE) χρησιμοποιήθηκαν για πρώτη φορά για τον γεωγραφικό χαρακτηρισμό τους. Η επιλογή των δειγμάτων για την εκπαίδευση και έλεγχο των μοντέλων ANN και CT που κατασκευάστηκαν έγινε με τη χρήση των DA roots, ώστε να επιτευχθεί αντιπροσωπευτική επιλογή τους και κατασκευή των βέλτιστων μοντέλων, απαραίτητων σε μια τόσο απαιτητική βάση.

Οδηγός σε όλα τα παραπάνω, αποτέλεσε η πεποίθηση ότι τα αποτελέσματα μεθόδων της ίδιας κατηγορίας (ομαδοποίησης ή συσχέτισης), πρέπει να είναι παρόμοια ή τουλάχιστον συμπληρωματικά σε κάποιες περιπτώσεις. Αν κάποια/ες από αυτές τις μεθόδους προβάλλουν για παράδειγμα σημαντικά διαφορετική ομαδοποίηση δεδομένων από τις άλλες, τότε έχουμε μια σοβαρή προειδοποίηση ότι τα δεδομένα δεν "περιέχουν" αρκετή πληροφορία ώστε να "συντηρήσουν" ένα μοντέλο [26]. Από την άλλη μεριά, είναι σημαντικό να αναφέρονται ακό-

μα και προσπάθειες που κατέληξαν σε απογοητευτικά αποτελέσματα, καθώς αρκετές φορές νέες ιδέες "κατακρημνίζονται" σε τέτοιες προσπάθειες [52].

Όλοι οι υπολογισμοί, διαγράμματα, πίνακες που αφορούν την εφαρμογή των πολύ-παραμετρικών τεχνικών, έγιναν με τη χρήση του λογισμικού Statistica [17].

ΚΕΦ. 5 ΤΑΜΙΕΥΤΗΡΕΣ ΠΟΣΙΜΟΥ ΥΔΑΤΟΣ ΑΤΤΙΚΗΣ

5.1. ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο αυτό εφαρμόστηκαν οι χημειομετρικές τεχνικές που αναφέρθηκαν στα θεωρητικά κεφάλαια της διατριβής αυτής, σε μια μεγάλη βάση δεδομένων που αφορούσε αποτελέσματα προσδιορισμών μετάλλων στους ταμιευτήρες ύδρευσης της Αθήνας στη διάρκεια έξι (6) μηνών (Νοέμβριος 2006 - Απρίλιος 2007). Λεπτομέρειες που αφορούν τη δειγματοληψία, τις μεθόδους ανάλυσης, τα αναλυτικά αποτελέσματα και τη βασική στατιστική επεξεργασία, αναφέρονται στην Ερευνητική Εργασία μου Διπλώματος ειδίκευσης [38].

Μελέτες (συγκριτικές ή όχι) ANN και άλλων πολυπαραμετρικών τεχνικών για την ταξινόμηση δειγμάτων νερού, έχουν αρκετές φορές καταγραφεί στο παρελθόν. Έτσι:

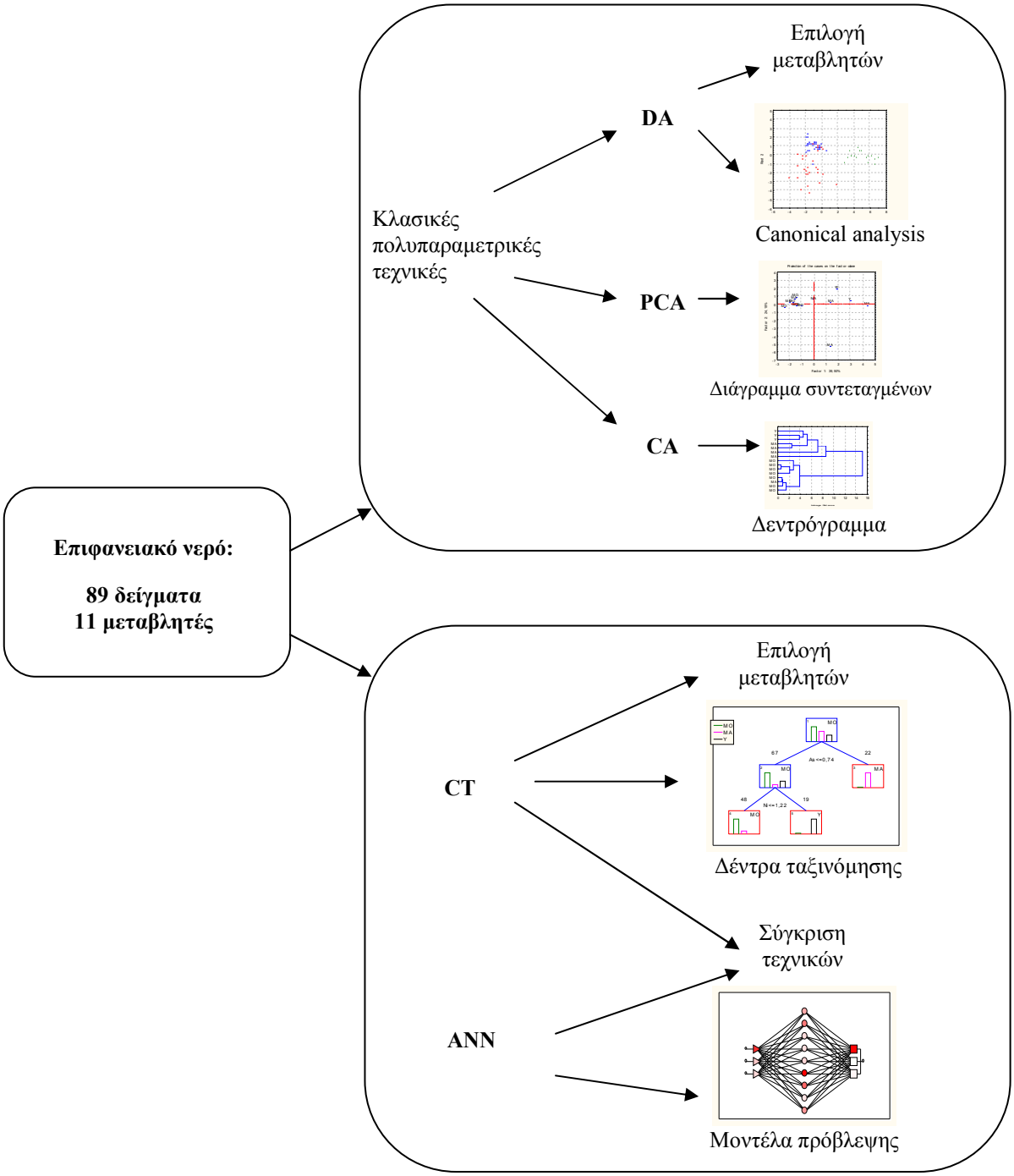
1. PCA, δίκτυα Kohonen και CP-ANN εφαρμόστηκαν για την ομαδοποίηση και ταξινόμηση δειγμάτων επιφανειακού νερού με βάση το βαθμό μόλυνσης αυτών [214].
2. CA, PCA και δίκτυα Kohonen χρησιμοποιήθηκαν για την ταξινόμηση των δειγμάτων ενός ποταμού με βάση 14 χημικές παραμέτρους [215].
3. N-way PCA και δίκτυα Kohonen εφαρμόστηκαν επίσης από την ίδια ερευνητική ομάδα με σκοπό την ερμηνεία και παρακολούθηση μιας μεγάλης βάσης δεδομένων που αφορούν την ποιότητα του νερού [230].
4. Μόνο CP-ANN εκμεταλλεύτηκε ο Grošelj et al. για την ταξινόμηση δειγμάτων εμφιαλωμένου νερού, με βάση τη γεωγραφική τους προέλευση [226].

Ειδικότερα, ανασκόπηση της χρήσης των ANN στην ανάλυση του νερού με αναφορά σε εργασίες που αφορούν συσχέτιση και πρόβλεψη παραμέτρων, γενικότερη εκτίμηση της συνολικής ποιότητας του νερού, αλλά και ταξινόμηση δειγμάτων έγινε πρόσφατα σε σχετική δημοσίευση [48].

Σκοπός του κεφαλαίου αυτού, είναι αρχικά η συγκριτική παρουσίαση κλασικών αλλά και νεώτερων πολυπαραμετρικών τεχνικών όπως CT και ANN μέσω της εφαρμογής τους σε μια μεγάλη βάση δεδομένων. Η αξιολόγηση τους έγινε αναφορικά με την ικανότητα ομαδοποίησης/ταξινόμησης των δειγμάτων αλλά και τη δυνατότητα αξιολόγησης των μεταβλητών. Το σχήμα 5.1 απεικονίζει διαγραμματικά τη μεθοδολογία που ακολουθήθηκε σε αυτό το κεφάλαιο και πρόσφατα δημοσιεύτηκε [231]. Επιγραμματικά, τέσσερις είναι οι στόχοι της σχηματικής επεξεργασίας των δεδομένων που απεικονίζεται στο σχήμα αυτό:

1. να αποκαλυφθούν οι κρίσιμες μεταβλητές που διαφοροποιούν μεταξύ τους, τους τρεις ταμιευτήρες,
2. να κατασκευαστούν μοντέλα που να περιγράφουν την κάθε λίμνη και να αξιοποιούν τις κρίσιμες αυτές μεταβλητές που καθορίζουν την ποιότητα του νερού,

3. να εφαρμοστούν και να συγκριθούν τα διαφορετικά μοντέλα DA, CT και ANN μεταξύ τους, με βάση την ικανότητά τους στην πρόβλεψη άγνωστων δειγμάτων,
4. αποκαλυφθούν ομοιότητες και διαφορές μεταξύ των σημείων δειγματοληψίας και/ή των Αθηναϊκών ταμιευτήρων.



Σχήμα 5.1: Σχηματική αναπαράσταση της μεθοδολογίας που ακολουθείται σε αυτό το κεφάλαιο.

5.2. ΜΕΘΟΔΟΛΟΓΙΑ

5.2.1. Αποτελέσματα

Αναλυτικά τα αποτελέσματα των μετρήσεων παρατίθενται στην Ερευνητική Εργασία μου Διπλώματος ειδίκευσης [38].

Παρακάτω ωστόσο, παρατίθεται βοηθητικά ο πίνακας 5.1 με τα σημεία δειγματοληψίας για την καλύτερη ερμηνεία των αποτελεσμάτων.

Πίνακας 5.1: Κωδικοποίηση των σημείων δειγματοληψίας των ταμιευτήρων πόσιμου ύδατος Αττικής.

Α/Α	**Κωδικός**	**Λίμνη***	**Περιγραφή θέσεως δειγματοληψίας**
1	9501	Υ	Μπούκα: σημείο εισροής από σήραγγα Καρδίτσας (Βοιωτικός Κηφισσός)
2	9502	Υ	Κέντρο λίμνης (παράκτιο σημείο)
3	9503	Υ	Αντλιοστάσιο Μουρικίου
4	9601	ΜΟ	Εκβολή ποταμού Μόρνου
5	9602	ΜΟ	Εκβολή ποταμού Άβορου
6	9603	ΜΟ	Κέντρο λίμνης (παράκτιο σημείο)
7	9604	ΜΟ	Λύματα Λιδορικίου
8	9606	ΜΟ	Πύργος υδροληψίας: παραλιακά κοντά στο φράγμα σε βάθος
9	9607	ΜΟ	Κατάντι (στάσιμο νερό)
10	9608	ΜΟ	Εκβολή ποταμού Κόκκινου
11	9701	ΜΑ	Πύργος υδροληψίας: παραλιακά κοντά στο φράγμα σε βάθος (Βεντούρι)
12	9702	ΜΑ	Σημείο εισροής ρεύματος Κιούρκων (κανάλι Μόρνου)
13	9703	ΜΑ	Σημείο εισροής χειμάρρου Βαρνάβα (συνεχής ροή)
14	9705	ΜΑ	Σημείο εισροής χειμάρρου Σταμάτας (συνεχής ροή)
15	9706	ΜΑ	Σημείο εισροής χειμάρρου Αγ. Στεφάνου (συνεχής ροή το χειμώνα)

*Υ: Υλίκη, ΜΟ: Μόρνος, ΜΑ: Μαραθώνας

Χρησιμοποιήθηκαν τελικά 11 μεταβλητές (βλ. Πίνακα 5.2 και  ΚΕΦ. 1, Π), ενώ ο Pb αποκλείστηκε από την παρακάτω ανάλυση, καθώς οι περισσότερες τιμές του, βρίσκονται κοντά στο LOD. Ο αριθμός των δειγμάτων ήταν 89 (18 Υλίκη, Υ, 42 Μόρνος, ΜΟ και 29 Μαραθώνας, ΜΑ). Η κωδικοποίηση των δειγμάτων έγινε με βάση τη λίμνη προέλευσή τους (τρίτη στήλη του πίνακα 5.1) και όχι τους μεμονωμένους κωδικούς (δεύτερη στήλη του πίνακα 5.1). Η εφαρμογή αυτή παρακάτω αναφέρεται σαν "ενοποιημένα" δεδομένα.

5.2.2. Διαχωριστική Ανάλυση (DA)

Στο κεφάλαιο αυτό, καθώς είναι το πρώτο από τα πειραματικά κεφάλαια που ακολουθούν, εξετάστηκαν αν ισχύουν για τις μεταβλητές κάποιες παραδοχές που απαιτεί η εφαρμογή της DA (§ 2.1.3). Έτσι, μελετήθηκαν οι παράμετροι που αναφέρθηκαν στο θεωρητικό τμήμα της διατριβής αυτής.

Η DA χρησιμοποιήθηκε στη βάση αυτή για την εύρεση των κρίσιμων μεταβλητών, αλλά και την ομαδοποίηση/ταξινόμηση των σημείων (κατασκευή μοντέλου). Εφαρμόστηκε στα "ενοποιημένα" συνολικά δεδομένα όλων των ταμιευτήρων (όχι στους μέσους όρους). Η αξιολόγηση των μοντέλων DA επιτεύχθηκε με μια σειρά δειγμάτων που ελήφθησαν από μεταγενέστερη χρονική περίοδο (11-12/2007) από τους ίδιους ταμιευτήρες.

5.2.3. Ανάλυση Κυρίων Συνιστωσών (PCA) και Ανάλυση Παραγόντων (FA)

Η FA που εφαρμόστηκε σε συνδυασμό με την PCA, ταξινόμησε και συσχέτισε τις μεταβλητές, αναδεικνύοντας ταυτόχρονα τις κυριότερες πηγές ρύπανσης. Επιτεύχθηκε επίσης ομαδοποίηση και των θέσεων δειγματοληψίας, με κύριο στόχο τη μελλοντική αποτελεσματική και αξιόπιστη μείωση αυτών [38]. Οι PCA/FA εφαρμόστηκαν στα "ενοποιημένα" δείγματα των ταμιευτήρων (μέσους όρους ανά σημείο).

5.2.4. Ανάλυση κατά Συστάδες (CA)

Η CA είχε εφαρμοστεί [38] στους μέσους όρους του διαλυτού κλάσματος των μετάλλων (τυποποιημένα δεδομένα, § 4.4.4) σε όλες τις θέσεις δειγματοληψίας. Ωστόσο, στη διατριβή αυτή, εφαρμόστηκε στα "ενοποιημένα" δείγματα των λιμνών (αφού έγινε τυποποίηση των δεδομένων, § 2.4.2, 4.4.4), όπως αυτά περιγράφηκαν παραπάνω (μέσους όρους ανά σημείο), ώστε να μπορεί γίνει σύγκριση με τις άλλες τεχνικές.

5.2.5. Δέντρα Ταξινόμησης (CT)

Η τεχνική των CT, παρουσίασε ιδιαίτερο ενδιαφέρον όταν πρωτοεφαρμόστηκε [38] στο αρχείο της DA, δηλαδή στα "ενοποιημένα" συνολικά δεδομένα όλων των ταμιευτήρων. Στη διατριβή αυτή χρησιμοποιήθηκαν και οι τρεις μέθοδοι που παρουσιάστηκαν παραπάνω (§ 3.2.1 - 3.2.3) σε ένα σύνολο 89 δειγμάτων.

Παράλληλα, η τεχνική των CT (οι τρεις διαφορετικές μέθοδοι) χρησιμοποιήθηκε για να εξαχθούν αποτελέσματα για την κρισιμότητα των μεταβλητών. Έτσι μπόρεσε να επιτευχθεί σύγκριση των πολυπαραμετρικών τεχνικών (συμβατικών και μη) σε επίπεδο δειγμάτων (θέσεων δειγματοληψίας) και μεταβλητών (μετάλλων/μεταλλοειδών). Η αξιολόγηση των μοντέλων επιτεύχθηκε και εδώ, με μια σειρά δειγμάτων που ελήφθησαν από μεταγενέστερη χρονική περίοδο (11-12/2007) από τους ίδιους ταμιευτήρες.

5.2.6. Νευρωνικά Δίκτυα (ANN)

Εξετάστηκε η εφαρμογή των ANN, στα ίδια δεδομένα ("ενοποιημένα" συνολικά δεδομένα) που προέκυψαν από τους προσδιορισμούς μετάλλων/μεταλλοειδών στους ταμιευτήρες πόσιμου ύδατος της Αττικής και που μελετήθηκαν παραπάνω με τις "παραδοσιακές" χημειομετρικές τεχνικές.

Όπως ήδη αναφέρθηκε, υπάρχουν διάφορες εναλλακτικές μέθοδοι και αλγόριθμοι. Το πρώτο βήμα λοιπόν ήταν η βελτιστοποίηση των μοντέλων μέσω των παραμέτρων τους, ενώ τελικά έγινε σύγκριση μεταξύ των ANN μοντέλων, αλλά και των άλλων τεχνικών.

5.3. ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΣΥΖΗΤΗΣΗ

5.3.1. Εφαρμογή της Διαχωριστικής Ανάλυσης

Στον πίνακα 5.2 έγινε μια πρώτη στατιστική επεξεργασία και απεικονίζονται κάποια βασικά χαρακτηριστικά για τις μεταβλητές.

Υπενθυμίζεται ότι αν η ασυμμετρία (skewness) η οποία μετρά την απόκλιση της κατανομής από τη συμμετρία είναι πολύ διαφορετική από το μηδέν, η κατανομή θεωρείται ασύμμετρη, ενώ οι κανονικές κατανομές θεωρούνται απολύτως συμμετρικές. Επίσης, αν η κύρτωση (kurtosis), η οποία μετρά την οξύτητα της κορυφής της κατανομής είναι πολύ διαφορετική από το μηδέν, η κατανομή είναι πιο "επίπεδη" ή πιο "οξεία" από την κανονική κατανομή, ενώ οι κανονικές κατανομές θεωρούνται μεσόκυρτες [17]. Στην κανονική κατανομή επίσης η μέση τιμή (mean) ισούται με τη διάμεση τιμή (median). Στη θετική συμμετρία για

παράδειγμα ισχύει: μέση τιμή > διάμεση τιμή. Η μεταβλητή του Cr προσεγγίζει καλύτερα τις παραπάνω απαιτήσεις.

Πίνακας 5.2: Βασικά περιγραφικά χαρακτηριστικά για τις 11 μεταβλητές.

Μεταβλητή	Μέση τιμή (μg/L)	Διάμεση τιμή (μg/L)	Ασυμμετρία (Skewness)	Κύρτωση (Kurtosis)
Fe	5,90	1,97	5,28	36,0
B	34,8	24,6	3,54	13,7
Al	3,54	1,76	4,08	16,8
V	0,46	0,35	1,76	3,07
Cr	0,47	0,36	1,01	0,26
Mn	26,0	0,63	5,05	25,5
Ni	1,74	0,81	1,56	1,53
Cu	0,56	0,41	3,09	10,9
Zn	16,0	15,2	1,35	2,56
As	0,64	0,31	1,80	2,43
Ba	63,3	58,9	1,97	5,30

Ο έλεγχος των συσχετίσεων των μεταβλητών (υπενθυμίζουμε ότι η DA απαιτεί ανεξάρτητες μεταβλητές, § 2.1.3) έγινε με τη βοήθεια των συντελεστών Pearson (§ 2.2, σχέση 2.8) και Spearman. Μικρές ως μέτριες συσχετίσεις καταγράφτηκαν (για τα αποτελέσματα βλ.  ΚΕΦ. 1, Π). Η συσχέτιση των μεταβλητών συζητείται και παρακάτω.

Ο έλεγχος των διακυμάνσεων/συνδιακυμάνσεων, πραγματοποιήθηκε με βάση τη δοκιμή Box's M (§ 2.1.3). Η μηδενική υπόθεση απορρίφθηκε και η ισότητα των διακυμάνσεων /συνδιακυμάνσεων των ομάδων τέθηκε υπό αμφισβήτηση. Τα log determinants (§ 2.1.3) ωστόσο, είναι παρόμοια (Y: 115, MO: 103 και MA: 131) και το αποτέλεσμα του Box's M μπορεί να αμφισβητηθεί.

Εφαρμόστηκε λοιπόν αρχικά η κλασική DA προσέγγιση. Ο παρακάτω πίνακας 5.3 αφορά τον έλεγχο και της παραμέτρου partial lambda (§ 2.1.1).

Πίνακας 5.3: Κρισιμότητα των 11 μεταβλητών (κλασική DA).

Μεταβλητή	Partial lambda	F-ratio	p-level	Ανοχή	Έλεγχος της H_0 για ισότητα των μέσων τιμών*
Fe	0,911	3,707	0,029	0,593	X
B	0,705	15,907	0,000	0,302	X
Al	0,968	1,257	0,290	0,695	✓
V	0,388	59,958	0,000	0,266	X
Cr	0,851	6,631	0,002	0,457	X
Mn	0,822	8,235	0,001	0,471	X
Ni	0,482	40,775	0,000	0,292	X
Cu	0,814	8,692	0,000	0,462	X
Zn	0,976	0,941	0,395	0,469	✓
As	0,650	20,447	0,000	0,447	X
Ba	0,945	2,229	0,115	0,507	✓

* Η μηδενική υπόθεση H_0 αφορά την ισότητα των μέσων τιμών της μεταβλητής στις τρεις ομάδες (Υ, ΜΟ, ΜΑ). Ο έλεγχος γίνεται σε επίπεδο εμπιστοσύνης 95 %.

Η παράμετρος **"ανοχή"** (tolerance) υπολογίζεται ως τη διαφορά $1\text{-}R^2$ (όπου R η συσχέτιση της κάθε μεταβλητής με τις υπόλοιπες του μοντέλου). Είναι δηλαδή το κλάσμα της διακύμανσης που είναι μοναδική για την αντίστοιχη μεταβλητή. Έτσι, όταν μια μεταβλητή είναι πλεονάζουσα, η τιμή του δείκτη ανοχής είναι 0 [17]. Στον παραπάνω πίνακα η τιμή αυτή κυμαίνεται από χαμηλότερες (0,27 / 0,29 / 0,30) ως υψηλότερες τιμές. Δεν υπάρχουν ωστόσο πολύ χαμηλές ή πολύ υψηλές τιμές. Η συσχέτιση των μεταβλητών μεταξύ τους (που μπορεί να εκτιμηθεί με τη διαφορά 1- tolerance) είναι μέτρια [7, 17]. Αυτό σημαίνει ότι η πληροφορία για τη συγκεκριμένη βάση δεδομένων διαμοιράζεται σε κάποιες μερικώς συσχετιζόμενες μεταβλητές. Το γεγονός αυτό επιβεβαιώθηκε και παραπάνω κατά τον υπολογισμό των συντελεστών Pearson και Spearman (επιπλέον αναφορά γίνεται και κατά την εφαρμογή της PCA, § 5.3.2).

Για τις μεταβλητές Ba, Al, Zn, η μηδενική υπόθεση της ισότητας των μέσων τιμών στις τρεις ομάδες γίνεται αποδεκτή και άρα αυτές κρίνονται ως **μη σημαντικές.**

Ο πίνακας 5.4 απεικονίζει τους τυποποιημένους συντελεστές των διαχωριστικών συναρτήσεων LDF (standardized Canonical discriminant function coefficients) και τους συντε-

λεστές δομής (§ 2.1.1). Οι μεγαλύτεροι κατά απόλυτη τιμή συντελεστές δηλώνουν τις σημαντικότερες μεταβλητές: V, Ni, As, B.

Πίνακας 5.4: Τυποποιημένοι συντελεστές των LDF και συντελεστές δομής (κλασική DA).

Μεταβλητή	**Τυποποιημένοι συντελεστές των LDF**		**Συντελεστές δομής**	
	Root 1	**Root 2**	**Root 1**	**Root 2**
Fe	0,400	-0,153	0,149	-0,252
B	-1,067	0,031	-0,054	-0,416
Al	0,139	-0,236	0,092	-0,110
V	1,640	0,030	0,277	-0,214
Cr	-0,612	0,089	-0,003	-0,090
Mn	-0,662	0,068	-0,055	-0,284
Ni	1,437	-0,109	0,239	-0,575
Cu	-0,684	0,060	0,078	-0,278
Zn	0,238	-0,074	0,054	-0,100
As	-0,523	-1,015	-0,114	-0,868
Ba	-0,126	0,424	-0,114	0,112

Οι αντίστοιχες μέσες τιμές των σκορ (για τα κεντροειδή κάθε ομάδας) φαίνονται στον πίνακα 5.5. Είναι φανερή η διαφοροποίηση της Υλίκης σε σχέση με τις άλλες δυο λίμνες που παρουσιάζουν ομοιότητες στις τιμές των συντεταγμένων.

Πίνακας 5.5: Μέσες τιμές των σκορ (Means of Canonical Variables, κλασική DA).

Ομάδα	**Root 1**	**Root 2**
Υλίκη (Υ)	4,758	-0,165
Μόρνος (ΜΟ)	-1,029	1,013
Μαραθώνας (ΜΑ)	-1,463	-1,365

Τέλος, ενδιαφέρον παρουσιάζει ο πίνακας 5.6 με τις ιδιοτιμές (eigenvalues) και την παράμετρο Wilks' lambda/model που απεικονίζει τη συνολική ποιότητα του μοντέλου

(§ 2.1.3). Συγκεκριμένα, επειδή η ιδιοτιμή μπορεί να ερμηνευτεί ως το μέτρο της διασποράς των κεντροειδών (§ 2.1.3), μπορούμε να πούμε ότι η πρώτη LDF εξηγεί το: 5,98 / (5,98 + 1,14) × 100 = 84,0 % της συνολικής διακύμανσης. Επιπλέον, μετά την αφαίρεση της πρώτης συνάρτησης LDF, η δεύτερη συνεχίζει να παραμένει σημαντική και άρα πρέπει να διατηρήσουμε και τις δυο συναρτήσεις. Πράγματι, ο παρακάτω πίνακας ελέγχει την υπόθεση της ισότητας των μέσων τιμών των LDF (τα κεντροειδών), δηλαδή ελέγχει αν υπάρχει περίπτωση κακού διαχωρισμού. Και στις δυο περιπτώσεις απορρίπτεται η υπόθεση.

Πίνακας 5.6: χ^2 δοκιμή για τη διατήρηση των δυο LDF

LDF	**Eigenvalues**	**Canonical correlation R_c**	**Wilks' lambda/model**	**χ^2**	**df**	**p-level**	**Έλεγχος της H_0 για ισότητα των κεντροειδών***
1	5,98	0,926	0,067	218,8	22	0,000	X
2	1,14	0,729	0,468	61,47	10	0,000	X

* Η μηδενική υπόθεση H_0 αφορά την ισότητα των μέσων τιμών των σκορ (κεντροειδών) των ομάδων (Y, MO, MA). Ο έλεγχος γίνεται σε επίπεδο εμπιστοσύνης 95 %.

Η εργασία αξιολόγησης της DA, συνεχίστηκε με την ταξινόμηση των δειγμάτων εκπαίδευσης για όλες τις τεχνικές (η παραπάνω εργασία επαναλήφθηκε και για τις εναλλακτικές προσεγγίσεις: FW και BW). Ο πίνακας 5.7 απεικονίζει τα ποσοστά επιτυχίας για την κάθε προσέγγιση DA για την ομάδα εκπαίδευσης. Υπολογίστηκε επίσης, το Press's Q Statistic για τον έλεγχο της τυχαίας ταξινόμησης (§ 2.1.3), ενώ στον 5.8 φαίνονται αναλυτικά οι προβλέψεις (predicted vs observed classes), για κάθε λίμνη (πίνακας ταξινόμησης, classification matrix).

Πίνακας 5.7: Αποτελέσματα με βάση την ομάδα εκπαίδευσης

(3 τεχνικές DA: 11 μεταβλητές και 89 δείγματα).

Θέση	**% Ποσοστά επιτυχίας**		
	κλασική	**FW**	**BW**
Υλίκη (Υ)	100,0	100,0	94,4
Μόρνος (ΜΟ)	97,6	97,6	97,6
Μαραθώνας (ΜΑ)	72,4	72,4	65,5
Press's Q Statistic*	262	262	232
Συνολικά	89,9	89,9	86,5

*Στατιστικό Q = 3,84 για 1 βαθμό ελευθερίας σε επίπεδο εμπιστοσύνης 95 %.

Τα αποτελέσματα για το Q είναι μεγαλύτερα της κρίσιμης τιμής, δηλαδή η ταξινόμηση DA των δειγμάτων επιτυγχάνει καλύτερα αποτελέσματα από την τυχαία ταξινόμηση.

Πίνακας 5.8: Πίνακας ταξινόμησης της ομάδας εκπαίδευσης: Παρατηρούμενες θέσεις (σειρές) έναντι προβλεπόμενων (στήλες)

Τεχνική	**Κλασική / FW**			**BW**			**Συνολικός αρ. δειγμάτων**
Πρόβλεψη / Θέση	**Υ**	**ΜΟ**	**ΜΑ**	**Υ**	**ΜΟ**	**ΜΑ**	
Υλίκη (Υ)	18	0	0	17	**1**	0	18
Μόρνος (ΜΟ)	0	41	**1**	0	41	**1**	42
Μαραθώνας (ΜΑ)	0	**8**	21	**1**	**9**	19	29

Είναι φανερό, ότι κάθε λίμνη φαίνεται να έχει ιδιαίτερα χαρακτηριστικά που τη διαχωρίζουν από τις υπόλοιπες. Έτσι για παράδειγμα, από ένα σύνολο 29 δειγμάτων της ομάδας εκπαίδευσης του Μαραθώνα, μόνο 8 (για τις τεχνικές κλασική και FW) αποδόθηκαν στο Μόρνο. Αυτό συνέβηκε γιατί το κανάλι του Μόρνου (θέση με Α/Α 12: 9702 στον πίνακα 5.1) εκρέει στο Μαραθώνα.

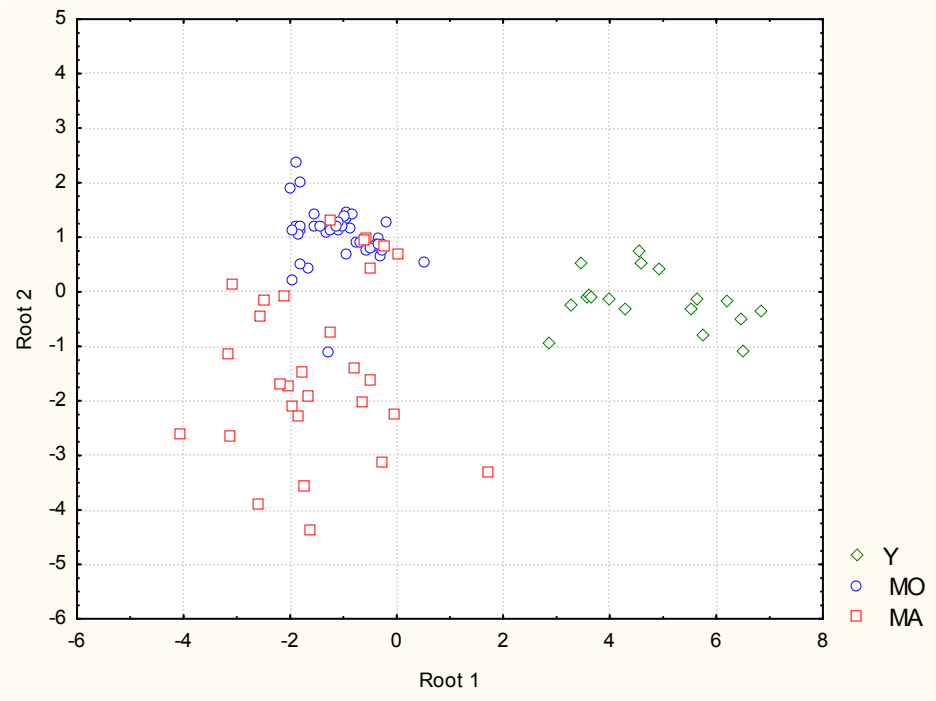
Ο πίνακας 5.9 απεικονίζει τις μεταβλητές που τελικά χρησιμοποιήθηκαν ή αποκλείστηκαν (με σειρά κρισιμότητας) από την κάθε τεχνική.

Πίνακας 5.9: "Κρίσιμες" και μη μεταβλητές για τις τρεις DA προσεγγίσεις (εντός παρενθέσεων αναγράφεται η παράμετρος partial lambda (§ 2.1.1) για τις σημαντικότερες μεταβλητές)

Τεχνική	**Μεταβλητές εντός του μοντέλου**	**Μεταβλητές εκτός του μοντέλου**
κλασική	V(0,39)>Ni(0,48)>As(0,65)>B(0,70)>Cu> Mn>Cr> Fe>Ba>Al >Zn	----
FW	V(0,40)> Ni(0,40)>As(0,65)>B(0,70)>Cu> Mn>Cr> Fe>Ba>Al	Zn
BW	Ni(0,46)>V(0,47)>As(0,56)>B(0,76)	Mn>Ba>Al>Fe>Cu>Cr>Zn

Η κλασική προσέγγιση χρησιμοποίησε το σύνολο των μεταβλητών (11), η FW τις 10 από αυτές και η BW μόλις 4. Είναι φανερό ότι οι δυο πρώτες τεχνικές χρησιμοποιώντας διαφορετικό αριθμό μεταβλητών (11 και 10, αντίστοιχα), επιτυγχάνουν τα ίδια ποσοστά ορθής πρόβλεψης. Η BW με μόνο 4 μεταβλητές επιτυγχάνει συνολικό ποσοστό υψηλό (>85 %). Συμπερασματικά μπορούμε να πούμε ότι με κάποιες λίγες κρίσιμες μεταβλητές μπορούμε να διαχωρίσουμε τα δείγματα των τριών λιμνών με επιτυχία, χωρίς οι υπόλοιπες μεταβλητές να συνεισφέρουν δραστικά.

Στο σχήμα 5.2 απεικονίζεται ο διαχωρισμός των τριών λιμνών όπως αυτός επιτυγχάνεται από τις γραμμικές διαχωριστικές συναρτήσεις (LDF) και την DA κλασική τεχνική. Είναι φανερός ο επιτυχής γραμμικός διαχωρισμός των δειγμάτων Υλίκης (Υ) από τα δείγματα Μόρνου (ΜΟ) και Μαραθώνα (ΜΑ). Ωστόσο, ο Μαραθώνας (ΜΑ) (λόγω του καναλιού στη θέση 9702) φαίνεται να "αναμιγνύεται" με τα δείγματα του Μόρνου.



Σχήμα 5.2: Διαχωρισμός των τριών λιμνών (Canonical plot) από την κλασική τεχνική DA.

Η ακρίβεια του νέου μοντέλου, μπορεί να επιβεβαιωθεί με μια σειρά δειγμάτων που ελήφθησαν από μεταγενέστερη χρονική περίοδο (11-12/2007) από τους ίδιους ταμιευτήρες. Τα δείγματα αυτά (που αποτελούν την ομάδα "ελέγχου"), χρησιμοποιήθηκαν ως **εξωτερικά "άγνωστα" δείγματα** που επικυρώσουν την ακρίβεια ή όχι των μοντέλων. Για τις τρεις τεχνικές DA, χρησιμοποιήθηκαν κάθε φορά εκείνες οι μεταβλητές που απαιτήθηκαν από το αντίστοιχο μοντέλο (βλ. πίνακα 5.9).

Οι πίνακες 5.10 – 5.12 συνοψίζουν τα αποτελέσματα, τα οποία είναι ανάλογα των αποτελεσμάτων της ομάδας εκπαίδευσης. Ο Μόρνος εξακολουθεί να δίνει υψηλά ποσοστά επιτυχίας, αλλά η απόδοση των δειγμάτων της Υλίκης φαίνεται να έχει λιγότερη επιτυχία (εξαίρεση αποτελεί η τεχνική BW με πολύ καλά αποτελέσματα). Έτσι, από ένα σύνολο 9 δειγμάτων του Μαραθώνα, μόνο 2 (για όλα τα μοντέλα) αποδόθηκαν στην Υλίκη ή το Μαραθώνα (77,8 % ποσοστό επιτυχίας). Τα δείγματα ωστόσο του Μαραθώνα εδώ αποδίδονται και στην Υλίκη και όχι μόνο στο Μαραθώνα. Αυτό μπορεί εύκολα να εξηγηθεί: μετά τον Οκτώβριο του 2007, η παροχή Μόρνου συμπληρώθηκε κατά το ήμισυ με νερό Υλίκης, με αποτέλεσμα το αμφιλεγόμενο σημείο (θέση με A/A 12: 9702 στον πίνακα 5.1), **να "αντιστοιχεί" εξίσου σε Υλίκη και Μόρνο.** Η "σύγχυση" λοιπόν των δειγμάτων Μαραθώνα–Υλίκη είναι αναμενόμενη. Αντίστροφα, η "απόδοση" του συγκεκριμένου δείγματος στην Υλίκη ή το Μόρνο μπορεί να οδηγήσει σε συμπεράσματα για τη σύσταση του νερού στο κανάλι.

Η σύγκριση των τριών τεχνικών DA για την ομάδα ελέγχου, αποδεικνύει την υπεροχή της τεχνικής BW, η οποία χρησιμοποιώντας μόνο 4 μεταβλητές δίνει υψηλά ποσοστά επιτυχίας.

Επιπλέον, οι πίνακες 5.10 – 5.12 απεικονίζουν την εξειδίκευση (βλ. § 4.4.2) των μοντέλων DA. Τα συνολικά ποσοστά είναι πραγματικά υψηλά (> 86 %), ενώ το μοντέλο BW εντυπωσίασε για μια ακόμα φορά.

Πίνακας 5.10: Αποτελέσματα για την ομάδα ελέγχου (τεχνική: κλασική)

Θέση	**Συνολικός αρ. δειγμάτων**	**% Ποσοστά επιτυχίας**	**% Εξειδίκευση**	**Αρ. προβλεπόμενων δειγμάτων**		
				Υ	**ΜΟ**	**ΜΑ**
Υλίκη	6	50,0	85,2	3	3	0
Μόρνος	14	78,6	78,9	3	11	0
Μαραθώνας	9	77,8	100,0	**1**	1	7
Συνολικά	29	72,6	86,8			

Πίνακας 5.11: Αποτελέσματα για την ομάδα ελέγχου (τεχνική: FW)

Θέση	**Συνολικός αρ. δειγμάτων**	**% Ποσοστά επιτυχίας**	**% Εξειδίκευση**	**Αρ. προβλεπόμενων δειγμάτων**		
				Υ	**ΜΟ**	**ΜΑ**
Υλίκη	6	33,3	95,8	2	4	0
Μόρνος	14	100,0	78,9	0	14	0
Μαραθώνας	9	77,8	100,0	**1**	1	7
Συνολικά	29	79,3	88,9			

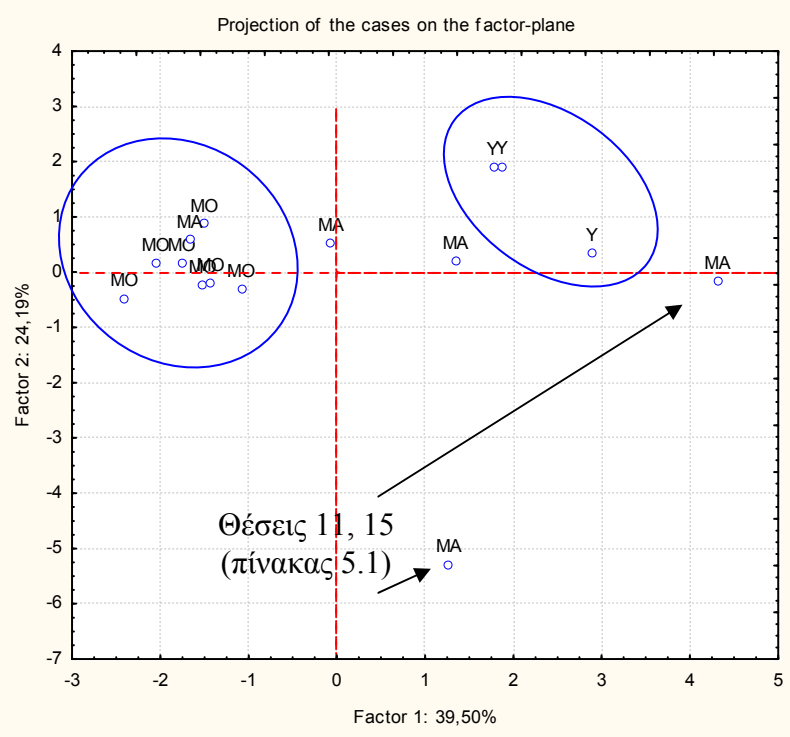
Πίνακας 5.12: Αποτελέσματα για την ομάδα ελέγχου (τεχνική: BW)

Θέση	**Συνολικός αρ. δειγμάτων**	**% Ποσοστά επιτυχίας**	**% Εξειδίκευση**	**Αρ. προβλεπόμενων δειγμάτων**		
				Υ	**ΜΟ**	**ΜΑ**
Υλίκη	6	100,0	85,2	6	0	0
Μόρνος	14	85,7	100,0	2	12	0
Μαραθώνας	9	77,8	100,0	**2**	0	7
Συνολικά	29	86,5	96,9			

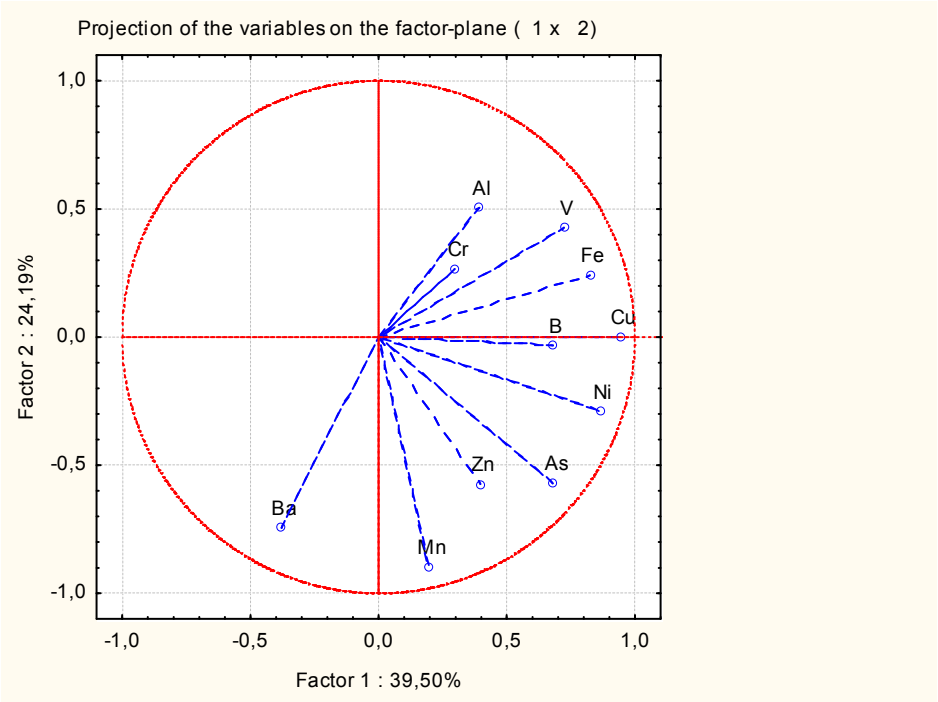
5.3.2. Εφαρμογή της Ανάλυσης Κυρίων Συνιστωσών (PCA) και της Ανάλυσης Παραγόντων (FA)

Η ανάλυση που πραγματοποιήθηκε και εδώ στα ενοποιημένα δείγματα (μέσοι όροι και χρήση του πίνακα συσχετίσεων), αποδείχθηκε επιτυχής μέσα από το αντίστοιχο διάγραμμα συντεταγμένων (σχ. 5.3(α)). Εδώ, επαληθεύτηκαν τα παραπάνω αποτελέσματα της καλής ταξινόμησης για Υλίκη και Μόρνο. Ο Μαραθώνας "συγχέεται" εκ νέου με τις θέσεις του Μόρνου (βλ. συγκριτικά πίνακα 5.8), ενώ εμφανίστηκαν δυο έκτροπες τιμές αυτού (θέσεις 11, 15 από πίνακα 5.1). Παράλληλα το διάγραμμα των φορτίσεων (σχ. 5.3(β)) επιβεβαίωσε τη μερική συσχέτιση των μεταβλητών που συζητήθηκε παραπάνω (§ 5.3.1).

Ο διαχωρισμός των ομάδων επιτυγχάνεται στον άξονα x (factor 1, σχ. 5.3(α)), ενώ στον ίδιο άξονα οι δεσπόζουσες (κρίσιμες) μεταβλητές είναι οι Ni, V, As, B, Cu και Fe (σχ. 5.3(β)). Οι ίδιες μεταβλητές επιλέχθηκαν από τις DA τεχνικές (πίνακας 5.9).



(α)



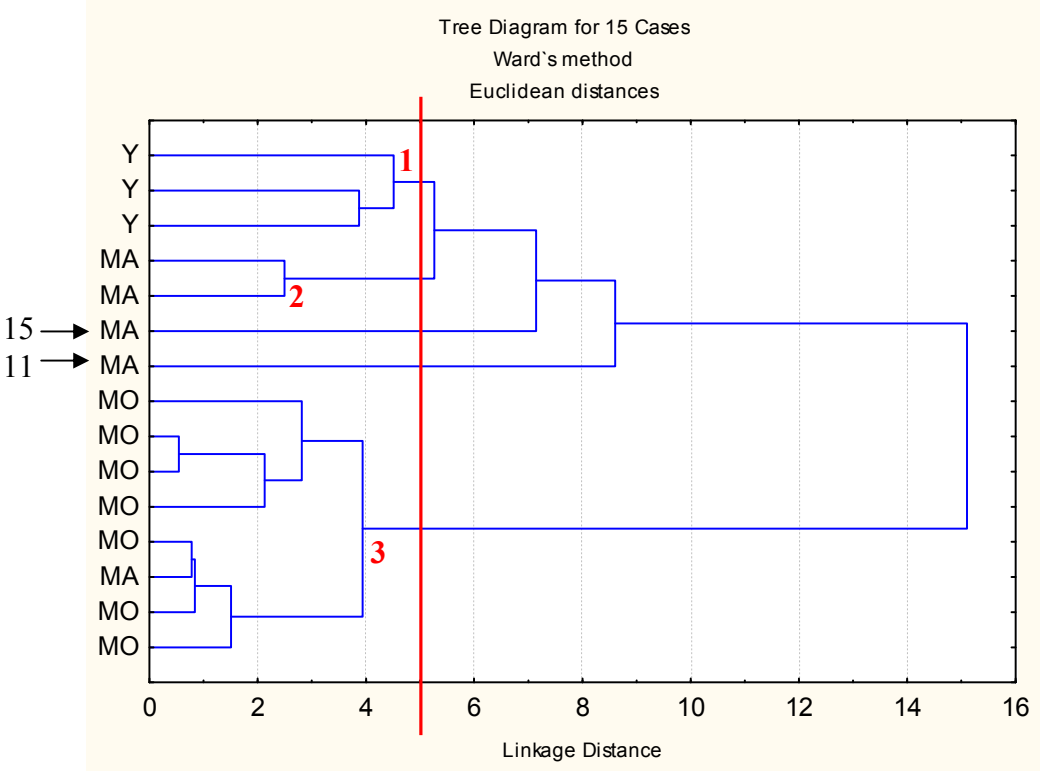
(β)

Σχήμα 5.3: Διάγραμμα συντεταγμένων (α) και διάγραμμα φορτίσεων (β) από την PCA ανάλυση (ενοποιημένοι μέσοι όροι αποτελεσμάτων).

5.3.3. Εφαρμογή της Ανάλυσης κατά Συστάδες (CA)

Πραγματοποιήθηκε CA στις θέσεις Υ (Υλίκη), ΜΟ (Μόρνος) και ΜΑ (Μαραθώνας) και τα αποτελέσματα παρουσιάζονται παρακάτω.

Στο σχήμα 5.4, απεικονίζεται το δενδρόγραμμα με χρησιμοποιούμενη τεχνική: **Ιεραρχική/Ward** (Hierarchical/Ward's method, § 2.4.2) και μέτρο σύγκρισης την **Ευκλείδεια απόσταση.** Η κόκκινη διαχωριστική απεικονίζει τις ομάδες που διαφοροποιούνται με βάση το κριτήριο 1/3 Dmax (όπου Dmax, η μέγιστη απόσταση μεταξύ των ομάδων), αυστηρότερου του γνωστού δείκτη σημαντικότητας Sneath για 2/3 Dmax [215].



Σχήμα 5.4: Ανάλυση κατά συστάδες στις θέσεις. Χρησιμοποιήθηκαν ενοποιημένοι μέσοι όροι αποτελεσμάτων και μέθοδος Ward.

Σχηματίζονται 3 συστάδες δειγμάτων. Η συστάδα 1 περιέχει αποκλειστικά δείγματα Υλίκης. Η συστάδα 2 περιέχει δυο μόνο δείγματα Μαραθώνα (θέσεις 13, 14 από πίνακα 5.1), ενώ η συστάδα 3 περιέχει υπο-ομάδες από δείγματα Μόρνου και Μαραθώνα. Είναι εμφανής η συμφωνία με τα αποτελέσματα από τις προηγούμενες μεθόδους: οι έκτροπες τιμές από τις θέσεις 11, 15 του πίνακα 5.1 (βλ. επίσης σχ. 5.3(α)), διαφοροποιήθηκαν και στο διάγραμμα αυτό. Τα σημεία Υ της Υλίκης διαχωρίζονται επίσης άμεσα, ενώ οι λίμνες Μόρνου και του Μαραθώνα χάνουν την ακεραιότητά τους και "συγχέονται" μεταξύ τους.

Οι κρίσιμες μεταβλητές που είναι υπεύθυνες για τον διαχωρισμό των ομάδων, μπορούν να βρεθούν από τον υπολογισμό των αντίστοιχων μέσων τιμών. Το σχετικό διάγραμμα φαίνεται στο ηλεκτρονικό παράρτημα της διατριβής ( ΚΕΦ. 1, Π). Είναι φανερό, ότι η Υλίκη (συστάδα 1) χαρακτηρίζεται από υψηλές τιμές Al, V, Ni και Zn, ενώ οι θέσεις 13 και 14 (συστάδα 2) αντιπροσωπεύονται από B και Mn. Η συστάδα 3 (βασικά θέσεις Μόρνου) χαρακτηρίζονται από την απουσία όλων των μετάλλων/μεταλλοειδών, εκτός των Ba και B.

5.3.4. Εφαρμογή των Δέντρων Ταξινόμησης (CT) - Μέθοδος των γραμμικών συνδυασμών (LCM)

Αρχικά χρησιμοποιήθηκε η LCM (πολυπαραμετρική μέθοδος Discriminant-based linear combination method) και τα αποτελέσματα για **κλάσμα δειγμάτων** (fraction of objects, FO = 0,1), φαίνονται στον πίνακα 5.13. Εφαρμόστηκε δηλαδή pre-pruning (§ 3.1.2), με καθορισμό της παραμέτρου FO στην τιμή 0,1. Αυτή η επιλογή καθορίζει τον ελάχιστο αριθμό των δειγμάτων (αντικειμένων) σε κάθε τάξη, σε σχέση με το συνολικό (αρχικό) αριθμό δειγμάτων. Ο διαχωρισμός σταματά όταν όλοι οι τερματικοί κόμβοι που περιέχουν πάνω από μία τάξη, δεν έχουν παραπάνω δείγματα από αυτά που καθορίζονται από την τιμή της παραμέτρου που αναφέρθηκε: έτσι ο κάθε τερματικός κόμβος δεν πρέπει να περιέχει πάνω από το 10 % του συνόλου των δειγμάτων μιας τάξης.

Το αντίστοιχο δέντρο απεικονίζεται στο σχήμα 5.5.

Πίνακας 5.13: Αποτελέσματα για την ομάδα εκπαίδευσης: Παρατηρούμενες θέσεις (στήλες) έναντι προβλεπόμενων (σειρές) / το δείγμα περιέχει 89 αντικείμενα.

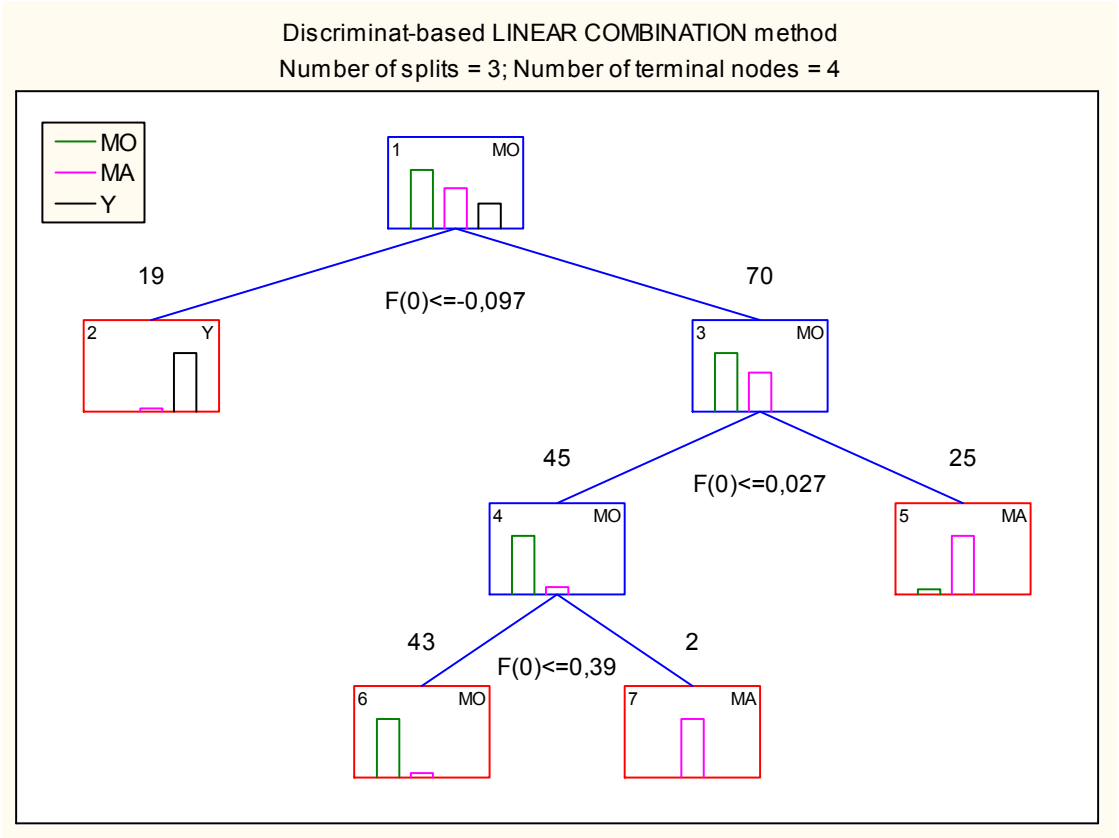
Θέση / Πρόβλεψη	Υ	ΜΟ	ΜΑ
Υλίκη (Υ)	18	0	**1**
Μόρνος (ΜΟ)	0	40	**3**
Μαραθώνας (ΜΑ)	0	**2**	25
Συνολικά	18	42	29
% Ποσοστά επιτυχίας	100	95,2	86,2

Το δέντρο περιλαμβάνει 3 διαχωρισμούς με βάση γραμμικούς συνδυασμούς των τιμών των μεταβλητών και 4 τερματικούς κόμβους. Οι αρχικοί κόμβοι παριστάνονται με μπλε πλαίσια και οι τερματικοί με κόκκινα. Ο χαρακτηρισμός μέσα σε κάθε πλαίσιο δείχνει την

πρόβλεψη γι' αυτό, ενώ ο αριθμός πάνω στο πλαίσιο, δείχνει τον αριθμό των αντικειμένων κάθε πλαισίου. Οι ράβδοι μέσα σε κάθε πλαίσιο, απεικονίζουν το ποσοστό αντικειμένων που αντιστοιχεί σε κάθε λίμνη. Ο πρώτος τερματικός κόμβος (αριθμός 2 στο σχήμα), "ονομάζεται" Υ (Υλίκη) γιατί περιέχει 18 δείγματα Υλίκης και 1 Μαραθώνα. Ο δεύτερος (αριθμός 5 στο σχήμα), χαρακτηρίζεται ως Μαραθώνας, αλλά περιέχει 23 δείγματα Μαραθώνα και 2 Μόρνου, ενώ ο τρίτος (αριθμός 6 στο σχήμα), περιέχει 40 δείγματα Μόρνου και 3 Μαραθώνα. Τέλος, ο τέταρτος τερματικός κόμβος (αριθμός 7 στο σχήμα), περιέχει μόνο 2 δείγματα Μαραθώνα.

Η κατασκευή (tree structure) του "δέντρου" που σχηματίστηκε, απεικονίζεται στον πίνακα 5.14. Το αριστερό τμήμα κάθε κόμβου (left branch) **είναι αυτό που κάθε φορά ικανοποιεί τη συνθήκη που περιγράφεται από τη σταθερά διαχωρισμού** (split constant). Ο κόμβος 1 για παράδειγμα, καθορίζει ότι τα δείγματα για τα οποία η συνάρτηση $0,097 + C_{Fe} \times(-0,0015) + C_B \times(0,0014) + \ldots\ldots + C_{Ba} \times(0,00020)$ έχει τιμή μηδενική ή μικρότερη θα "σταλούν" αριστερά (C_x: η συγκέντρωση του μετάλλου Χ σε μg/L).

Ο πίνακας 5.14 αναδεικνύει επίσης, τους γραμμικούς συνδυασμούς των μεταβλητών που χρησιμοποιήθηκαν (τελευταίες στήλες) για το διαχωρισμό που προτείνεται. Το μέγεθος των συντελεστών που χρησιμοποιούνται για κάθε μεταβλητή αποτελεί μια ένδειξη για τη σημαντικότητα της καθεμιάς για το συγκεκριμένο διαχωρισμό. Κρισιμότερες αναδεικνύονται οι παράμετροι V, Ni As, Cr, και Cu.



Σχήμα 5.5: Δέντρο ταξινόμησης για τα ενοποιημένα δείγματα των ταμιευτήρων Αττικής. Μέθοδος: LCM

Πίνακας 5.14: Κατασκευή δέντρου, κόμβοι που δημιουργούνται, παρατηρούμενες (στήλες) έναντι προβλεπόμενων (σειρές) θέσεων, σταθερές και μεταβλητές διαχωρισμού.

Κόμβοι	**Αριστ. κλάδος**	**Δεξιός κλάδος**	**ΜΟ**	**ΜΑ**	**Υ**	**Πρόβλεψη**	**Σταθερά διαχ/μου**	**Fe**	**B**	**Al**	**V**	**Cr**	**Mn**	**Ni**	**Cu**	**Zn**	**As**	**Ba**
1	2	3	42	29	18	ΜΟ	0,097	-0,0015	0,0014	-0,00084	**-0,21**	**0,086**	0,00025	**-0,038**	**0,042**	-0,00068	**0,036**	0,00020
2*			0	1	18	Υ												
3	4	5	42	28	0	ΜΟ	-0,027	0,0040	0,00069	0,0029	**-0,18**	**-0,027**	-0,000009	**-0,0081**	**0,087**	0,0013	**0,10**	-0,0011
4	6	7	40	5	0	ΜΟ	-0,39	0,019	-0,0092	0,059	**0,093**	**-0,68**	0,047	**0,46**	**-0,062**	-0,00056	**-0,51**	0,0047
5*			2	23	0	ΜΑ												
6*			40	3	0	ΜΟ												
7*			0	2	0	ΜΑ												

* Τερματικοί κόμβοι

Συμπερασματικά, η LCM έδωσε συνολικό ποσοστό επιτυχίας για την ομάδα εκπαίδευσης των τριών λιμνών 93,2 %.

Η ακρίβεια του νέου μοντέλου, μπορεί να επιβεβαιωθεί και εδώ με μια σειρά αποτελεσμάτων που ελήφθησαν από μεταγενέστερη χρονική περίοδο (11-12/2007) από τους ίδιους ταμιευτήρες. Η δοκιμή πραγματοποιήθηκε μετά την πειραματική εφαρμογή και των τριών CT μεθόδων (§ 5.3.7), ώστε να μπορεί να γίνει σύγκριση μεταξύ τους.

5.3.5. Μονοπαραμετρική κλασική μέθοδος (Classic CT)

Επόμενη τεχνική που χρησιμοποιήθηκε είναι η Classic CT **(**μονοπαραμετρική τεχνική Discriminant-based univariate method), και τα αποτελέσματα για το ίδιο FO = 0,1 φαίνονται στον πίνακα 5.15. Το αντίστοιχο δέντρο απεικονίζεται στο σχήμα 5.6.

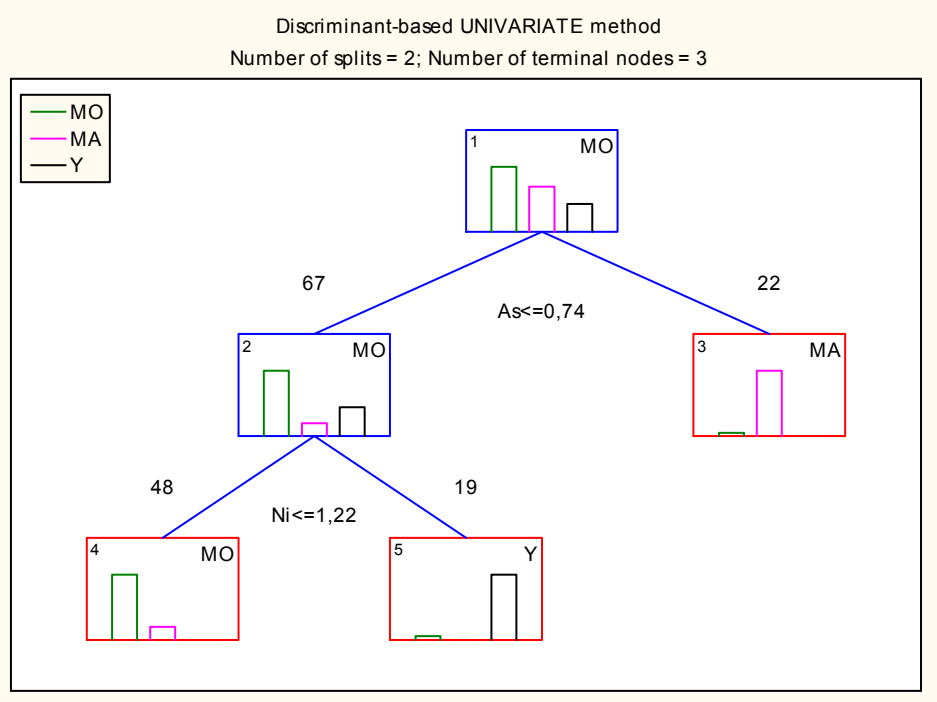
Πίνακας 5.15: Αποτελέσματα για την ομάδα εκπαίδευσης.

Θέση / Πρόβλεψη	**Υ**	**ΜΟ**	**ΜΑ**
Υλίκη (Υ)	18	**1**	0
Μόρνος (ΜΟ)	0	40	**8**
Μαραθώνας (ΜΑ)	0	**1**	21
Συνολικά	18	42	29
% Ποσοστά επιτυχίας	100	95,2	72,4

Το δέντρο περιλαμβάνει 2 "δυαδικούς" διαχωρισμούς (ο πρώτος με βάση την τιμή του As και ο δεύτερος του Ni) και 3 τερματικούς κόμβους. Ο αρχικός κόμβος "ονομάζεται" ΜΟ (Μόρνος) γιατί περιέχει περισσότερα δείγματα προερχόμενα από το Μόρνο. Στη συνέχεια χωρίζεται σε δυο κλάδους από τους οποίους, ο δεξιός (αριθμός 3 στο σχήμα) είναι τερματικός και χαρακτηρίζεται ορθά ως ΜΑ (Μαραθώνας) και περιέχει 22 δείγματα. Ο αριστερός κλάδος χωρίζεται σε δυο τερματικούς. Ο δεξιός από αυτούς χαρακτηρίζεται ως Υ (Υλίκη) και περιέχει 19 δείγματα. Προφανώς (πίνακας 5.16), τα 1 από αυτά, ανήκει στο Μόρνο. Ο αριστερός "ονομάζεται" ΜΟ (Μόρνος) και περιέχει 48 δείγματα. Προφανώς (πίνακας 5.15) τα 8 από αυτά, ανήκουν στο Μαραθώνα.

Η κατασκευή του "δέντρου" που σχηματίστηκε, απεικονίζεται στον πίνακα 5.16. Ο διαχωρισμός γίνεται κάθε φορά, με βάση τη μεταβλητή που διαφοροποιεί τα αντικείμενα μεταξύ τους. Στους τερματικούς κόμβους αυτούς συνοψίζεται η επιτυχία ή όχι του διαχωρισμού.

Έτσι για παράδειγμα, ο κόμβος 3 είναι τερματικός: 21 σημεία της λίμνης ΜΑ και 1 (ένα) σημείο της λίμνης ΜΟ αποδίδονται στο ΜΑ (βλ. στήλη: πρόβλεψη).



Σχήμα 5.6: Δέντρο ταξινόμησης για τα ενοποιημένα δείγματα των ταμιευτήρων Αττικής. Μέθοδος: Classic CT.

Πίνακας 5.16: Στοιχεία της κατασκευής του δέντρου.

Κόμβοι	**Αριστ. κλάδος**	**Δεξιός κλάδος**	**Y**	**ΜΟ**	**ΜΑ**	**Πρόβλεψη**	**Σταθερά διαχ/μου**	**Μεταβλητή διαχ/μου**
1	2	3	18	42	29	ΜΟ	-0,74	As
2	4	5	18	41	8	ΜΟ	-1,22	Ni
3*			0	1	21	ΜΑ		
4*			0	40	8	ΜΟ		
5*			18	1	0	Y		

* Τερματικοί κόμβοι

Τέλος ο πίνακας 5.16, αναδεικνύει τις πιο κρίσιμες μεταβλητές που χρησιμοποιήθηκαν (τελευταία στήλη) για το διαχωρισμό που προτείνει. Οι μεταβλητές αυτές είναι As και Ni.

Συμπερασματικά, από το σύνολο των 42 σημείων του Μόρνου, 1 σημείο αυτού "αποδόθηκε" εσφαλμένα στην Υλίκη και 1 στο Μαραθώνα, ενώ από το σύνολο των 29 δειγμάτων του Μαραθώνα, "αποδόθηκαν" εσφαλμένα 8 στο Μόρνο.

Η Classic CT μέθοδος έδωσε συνολικό ποσοστό επιτυχίας για την ομάδα εκπαίδευσης των τριών λιμνών 88,8 %.

5.3.6. Μονοπαραμετρική μέθοδος της Διεξοδικής Σάρωσης (CART)

Επόμενη που χρησιμοποιήθηκε είναι η CART μέθοδος (επίσης με μονοπαραμετρικούς διαχωρισμούς). Εδώ, χρησιμοποιήθηκε η μέθοδος **"ελαχίστου κόστους-πολυπλοκότητας διασταυρούμενη επικύρωση"** (minimal cost-complexity cross-validation) για την εύρεση του καλύτερου δέντρου, με pre-pruning: ελάχιστο αριθμό πέντε (5) δειγμάτων σε ένα κόμβο.

Με αυτόν τον τρόπο, ενεργοποιείται η επιλογή της αυτόματης κατασκευής δέντρου με βάση **το μέγεθος αυτού,** αλλά και **το "κόστος" από τη διασταυρούμενη επικύρωση** (cross-validation, CV cost). Το κόστος αυτό όσον αφορά τα μοντέλα ταξινόμησης, υπολογίζεται με βάση το ποσοστό των λαθεμένων προβλέψεων στο σύνολο των δειγμάτων ελέγχου. Είναι δε τόσο μικρότερο όσο μεγαλύτερο είναι το ποσοστό των αληθινών προβλέψεων. Αν επιπλέον, το κόστος της ομάδας ελέγχου είναι μεγαλύτερο από το αντίστοιχο για την ομάδα εκπαίδευσης (βλ. παρακάτω), φαίνεται ότι το μοντέλο έχει μικρή ικανότητα πρόβλεψης για νέα δείγματα. Το δέντρο **"καταλλήλου μεγέθους"** ("right-sized") που τελικά επιλέγεται, είναι το μικρότερο δυνατό του οποίου το CV cost δεν διαφέρει σημαντικά από το ελάχιστο CV cost που μπορεί να επιτευχθεί με συγκεκριμένους περιορισμούς (και προφανώς πολυπλοκότερο δέντρο). Ο υπολογισμός των κόστων δηλαδή, είναι απαραίτητος ώστε να εφαρμοστεί η διαδικασία **"cost-complexity pruning"** καθώς το δέντρο μεγαλώνει από το πρώτο κόμβο προς το μέγιστο δέντρο που καθορίζεται εξ΄αρχής από τον ελάχιστο αριθμό δειγμάτων σε κάθε τερματικό κόμβο. Το κόστος για την ομάδα εκπαίδευσης υπολογίζεται σε κάθε κόμβο και μειώνεται όσο μεγαλώνει το δέντρο. Ονομάζεται δε, **Resubstitution cost,** ώστε να διαχωρίζεται από το αντίστοιχο για την ομάδα ελέγχου, καθώς πραγματοποιείται επιπλέον v-fold CV σε κάθε διαχωρισμό (§ 3.1.2).

Το μέγεθος του δέντρου εκτιμάται από τον αριθμό των τερματικών κόμβων, καθώς ο πρώτος αρχικός κόμβος θεωρείται ότι έχει μέγεθος 1. Με τη βηματική κατασκευή του δέντρου, υπολογίζεται κάθε φορά μια συνάρτηση που περιλαμβάνει τα κόστη αλλά και το μέγεθος του δέντρου. Εξελίσσεται έτσι ένας αλγόριθμος που "κλαδεύει" τα μεγάλα δέντρα, αλλά και τα μικρά με μεγάλο κόστος.

Αναλυτικότερα η διαδικασία επιλογής του δέντρου καταλλήλου μεγέθους, περιγράφεται από τα παρακάτω βήματα:

1. Βρίσκεται το δέντρο με το μικρότερο CV cost (minCV) και ονομάζεται minSt error, το τυπικό σφάλμα (Standard error) αυτού (προκύπτει από τις επαναλήψεις κατά την CV διαδικασία).

2. Επιλέγεται το δέντρο καταλλήλου μεγέθους, ως το μικρότερο δέντρο με CV cost μικρότερο του αθροίσματος: minCV + minSt error.

Μικρή τιμή του minSt error οδηγεί γενικά στην επιλογή ενός δέντρου καταλλήλου μεγέθους, ελαφρά απλούστερου από το αντίστοιχο με το ελάχιστο CV cost. Αντίθετα μια μεγάλη τιμή του minSt error, οδηγεί στην επιλογή ενός right-sized δέντρου πολύ απλούστερου. Έτσι, η εφαρμογή της διαδικασίας "cost-complexity pruning" εγγυάται τελικά δυο βασικές επιστημονικές αρχές της στατιστικής: **φειδωλότητα** (parsimony, § 4.4.3) και **εκτίμηση της διακύμανσης στην επανάληψη ενός πειράματος** (replication, § 5.3.8). Επιλέγεται λοιπόν το απλούστερο δέντρο (με το μικρότερο μέγεθος ή τους λιγότερους τερματικούς κόμβους) που είναι συνεπές (δεν έχει μεγαλύτερο CV cost από το άθροισμα minCV + minSt error) με την πορεία εύρεσης του καλύτερου δέντρου από ανεξάρτητες δοκιμές (μικρότερο CV cost) [17].

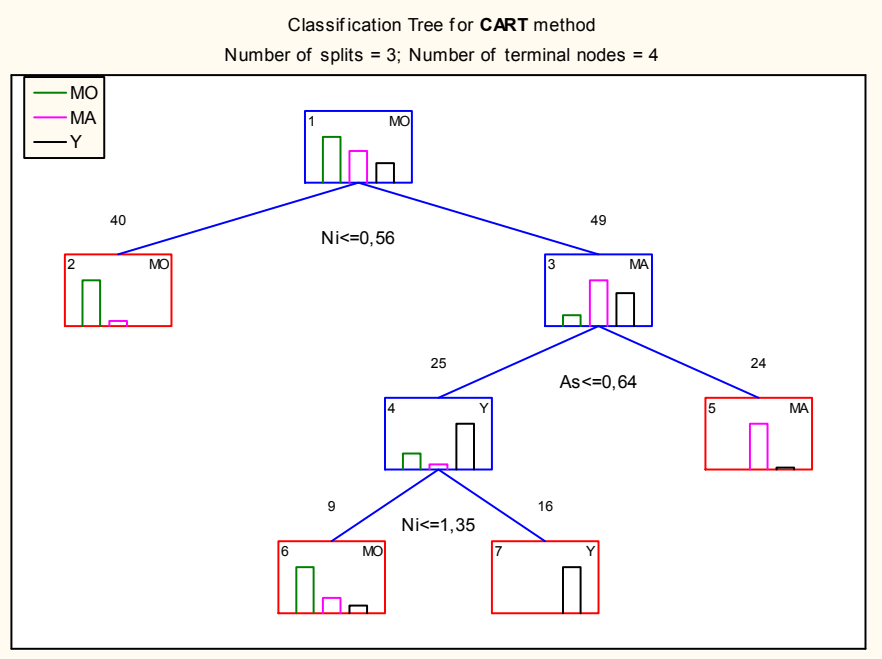
Τα αποτελέσματα ταξινόμησης για το καλύτερο δέντρο και την ομάδα εκπαίδευσης φαίνονται στον πίνακα 5.17. Το αντίστοιχο δέντρο απεικονίζεται στο σχήμα 5.7.

Πίνακας 5.17: Αποτελέσματα για την ομάδα εκπαίδευσης.

Πρόβλεψη \ Θέση	**Υ**	**ΜΟ**	**ΜΑ**
Υλίκη (Υ)	16	0	0
Μόρνος (ΜΟ)	**1**	42	**6**
Μαραθώνας (ΜΑ)	**1**	0	23
Συνολικά	18	42	29
% Ποσοστό επιτυχίας	88,9	100	79,3

Το δέντρο περιλαμβάνει 3 "δυαδικούς" διαχωρισμούς (ο πρώτος με βάση την τιμή του Ni, ο δεύτερος βάση την τιμή του As και ο τρίτος βάση την τιμή του Ni) και 4 τερματικούς κόμβους. Ο πρώτος τερματικός κόμβος "ονομάζεται" ΜΟ (Μόρνος) γιατί περιέχει περισσότερα δείγματα από το Μόρνο και λιγότερα από το Μαραθώνα. Ως Μαραθώνας χαρακτηρίζεται ο δεύτερος τερματικός κόμβος, αλλά περιέχει και 1 δείγμα Υλίκης. Ο επόμενος χαρακτηρίζεται ως Μόρνος, αλλά περιέχει και 2 δείγματα Μαραθώνα και 1 Υλίκη και ο τελευταίος περιέχει 16 δείγματα Υλίκης.

Η κατασκευή του "δέντρου" που σχηματίζεται, απεικονίζεται στον πίνακα 5.18. Ο πίνακας αυτός αναδεικνύει επιπλέον, τις πιο κρίσιμες μεταβλητές που χρησιμοποιήθηκαν για το διαχωρισμό που προτείνει. Οι μεταβλητές αυτές είναι Ni και As.



Σχήμα 5.7: Δέντρο ταξινόμησης για τα ενοποιημένα δείγματα των ταμιευτήρων Αττικής. Μέθοδος: CART.

Πίνακας 5.18: Στοιχεία της κατασκευής του δέντρου.

Κόμβοι	**Αριστ. κλάδος**	**Δεξιός κλάδος**	**MO**	**MA**	**Y**	**Πρόβλεψη**	**Σταθερά διαχ/μου**	**Μεταβλητή διαχ/μου**
1	2	3	42	29	18	MO	-0,5575	Ni
2*			36	4	0	MO		
3	4	5	6	25	18	MA	-0,6438	As
4	6	7	6	2	17	Y	-1,350	Ni
5*			0	23	1	MA		
6*			6	2	1	MO		
7*			0	0	16	Y		

* Τερματικοί κόμβοι

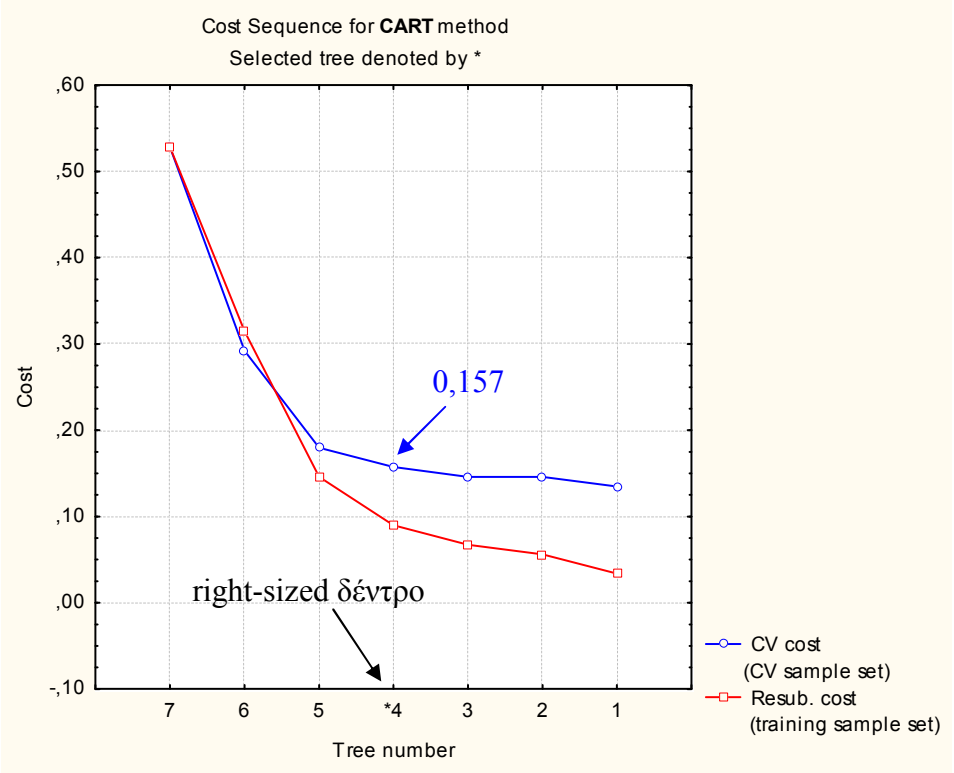
Το παραπάνω δέντρο επιλέχθηκε από μια σειρά δέντρων όπως φαίνεται στον παρακάτω πίνακα 5.19. Το δέντρο καταλλήλου μεγέθους, (με A/A 4 στον πίνακα 5.19), έχει ένα CV cost = 0,157 με τυπικό σφάλμα = 0,039. Το μικρότερο CV cost που απεικονίζεται στον πίνακα είναι 0,135 (A/A 1) με τυπικό σφάλμα = 0,036. Η πολυπλοκότητα του δέντρου αυτού όμως είναι μεγάλη (διαθέτει 9 τερματικούς κόμβους). Αντίθετα, το επιλεγμένο δέντρο με 4 μόνο τερμα-

τικούς κόμβους δίνει CV cost που δεν υπερβαίνει το άθροισμα 0,135+0,036 = 0,171 που αντιστοιχεί στο δέντρο 1.

Πίνακας 5.19: Στατιστικά στοιχεία για τα δέντρα της CART μεθόδου. Το επιλεγμένο δέντρο καταλλήλου μεγέθους, επισημαίνεται με *.

Α/Α δέντρου	**Αρ. τερματικών κόμβων**	**CV cost**	**Τυπικό σφάλμα**
1	9	0,135 (minCV)	0,036 (minSt error)
2	6	0,146	0,037
3	5	0,146	0,037
4*	4	0,157	0,039
5	3	0,180	0,041
6	2	0,292	0,048
7	1	0,528	0,053

Παραστατικά η παραπάνω εικόνα δίνεται στο σχήμα 5.8.



Σχήμα 5.8: Πορεία σχηματισμού και επιλογής δέντρων για τη μέθοδο CART.

Συμπερασματικά, η CART μέθοδος έδωσε συνολικό ποσοστό επιτυχίας για την ομάδα εκπαίδευσης των τριών λιμνών 91,0 %.

5.3.7. Σύγκριση μεθόδων - αποτελέσματα

Ενδιαφέρον παρουσιάζει η σύγκριση των παραπάνω τεχνικών CT (για την ομάδα εκπαίδευσης). Παρουσιάζουν ομοιότητες όπως:

- ✓ τα σημεία της Υλίκης διαφοροποιούνται άμεσα από τα υπόλοιπα,
- ✓ τα σημεία του Μόρνου ταξινομούνται γενικά επιτυχώς, με μερικά από αυτά να "αποδίδονται" στο Μαραθώνα ή στην Υλίκη,
- ✓ τα σημεία του Μαραθώνα ταξινομούνται επίσης επιτυχώς, με μερικά από αυτά να "αποδίδονται" στο Μόρνο και λιγότερο στην Υλίκη.

Ωστόσο, από τις τρεις μεθόδους, η LCM έδωσε το μεγαλύτερο ποσοστό επιτυχίας (93,2 %), σχηματίζοντας ένα μικρό και ακριβές δέντρο. Η δεύτερη προσέγγιση (της Classic CT), έδωσε τα χαμηλότερα ποσοστά (συνολικά 88,8 %). Συγκρινόμενες οι δυο βασικές μέθοδοι ταξινόμησης (DA και CT), δίνουν παρόμοια ποσοστά με μόνη διαφοροποίηση, τη μεγαλύτερη επιτυχία της LCM μεθόδου. Η τρίτη μέθοδος (CART) έδωσε ενδιάμεσα ποσοστά (συνολικό 91,0 %) και είναι η μόνη μέθοδος που δεν διαχώρισε πλήρως τα σαφώς διακεκριμένα σημεία της Υλίκης, αποδίδοντας 2 από αυτά, σε Μαραθώνα και Μόρνο.

Στη συνέχεια, ελέχθησαν συνολικά τα τρία μοντέλα CT όσον αφορά την ακρίβεια (ευαισθησία) και εξειδίκευσή τους (βλ. § 4.4.2) σε εξωτερικά μεταγενέστερα δείγματα (11 – 12/2007) των ίδιων ταμιευτήρων. Η αναλογία στον αριθμό των δειγμάτων ανά λίμνη είναι όμοια με την ομάδα εκπαίδευσης.

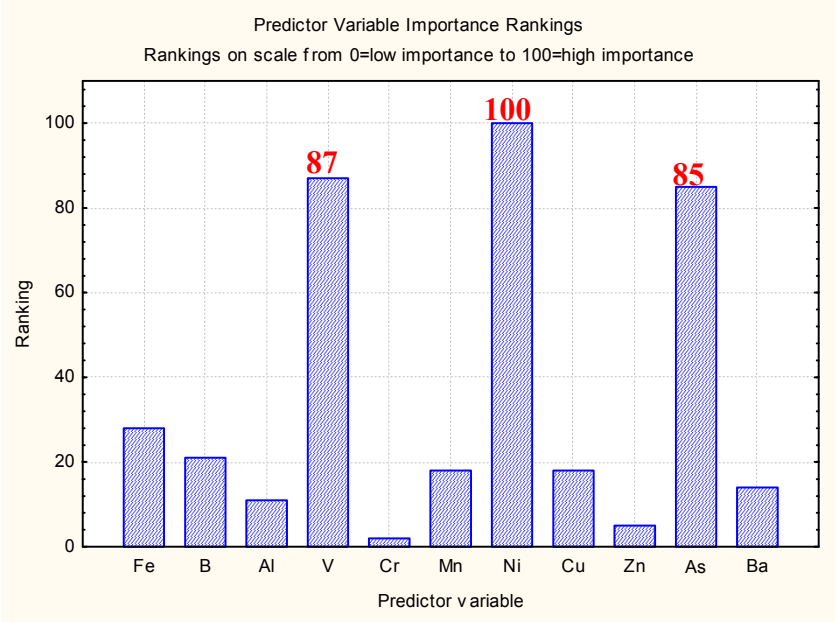
Στο πίνακα 5.20 που ακολουθεί απεικονίζονται συγκριτικά τα αποτελέσματα.

Πίνακας 5.20: Αποτελέσματα για την ομάδα ελέγχου

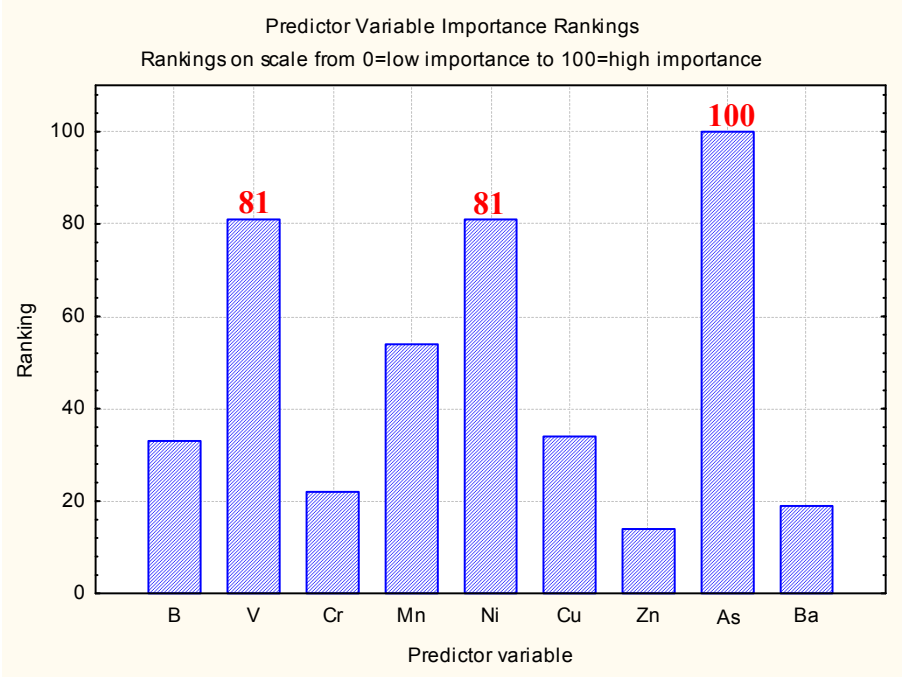
Μέθοδος / **Λίμνη (αρ. δειγμάτων)**	**LCM**		**Classic CT**		**CART**	
	Ακρίβεια	**Εξειδίκευση**	**Ακρίβεια**	**Εξειδίκευση**	**Ακρίβεια**	**Εξειδίκευση**
Υλίκη (6)	66,7	76,7	83,3	88,5	66,7	92,0
Μόρνος (14)	50,0	93,8	85,7	100,0	88,9	95,2
Μαραθώνας (9)	66,7	83,3	88,9	95,2	92,8	93,8
Συνολικά	58,6	87,0	86,5	96,2	86,2	93,8

Παρατηρούνται τα παρακάτω:

1. Ο Μαραθώνας έδωσε γενικότερα τα καλύτερα αποτελέσματα όσον αφορά την ακρίβεια και την εξειδίκευση. Η LCM έδωσε παραδόξως (σε σχέση με την ομάδα εκπαίδευσης) τα χαμηλότερα αποτελέσματα για το Μαραθώνα.
2. Η Classic CT μέθοδος έδωσε τα υψηλότερα αποτελέσματα από όλες τις CT μεθόδους, και όμοια μόνο με την τεχνική BW-DA (βλ. πίνακες 5.10 – 5.12).



(α)



(β)

Σχήμα 5.9: Κρισιμότητα μεταβλητών για τις τρεις CT μεθόδους: (α) Classic CT, (β) CART.

Στα παραπάνω διαγράμματα (σχ. 5.9) απεικονίζεται η σχετική ταξινόμηση των μεταβλητών για τις μεθόδους Classic CT και CART. Από τα ιστογράμματα φαίνεται ότι οι τρεις μεταβλητές V, Ni και As, παρουσιάζουν τα μεγαλύτερα ποσοστά κρισιμότητας. Η αντίστοιχη ανάλυση για την τρίτη μέθοδο LCM απεικονίζεται στον πίνακα 5.14 όπου από το μέγεθος των συντελεστών, διαφαίνεται η κρισιμότητα των V, Ni As, Cr και Cu.

Καταλήγοντας, παρόλο που φαίνεται να υπερέχει κάποια από τις τρεις CT μεθόδους (η LCM για την ομάδα εκπαίδευσης ή η Classic CT για την ομάδα ελέγχου), δίνοντας η καθεμιά, μεγαλύτερα ποσοστά για τη συγκεκριμένη ομάδα δειγμάτων, τελικά αλληλοσυμπληρώνονται και αλληλοεπιβεβαιώνονται καταλήγοντας στα ίδια συμπεράσματα σε επίπεδο θέσεων αλλά και μεταβλητών.

5.3.8. Εφαρμογή των Multi-layer perceptron (MLP)

Μέσα από τη βιβλιογραφία, προτείνονται γενικά οι παρακάτω παράμετροι που επηρεάζουν την ποιότητα ενός μοντέλου MLP:

- ✓ το μέγεθος της ομάδας εκπαίδευσης [59, 72, 134, 178],
- ✓ ο αριθμός των εισερχόμενων μεταβλητών [122, 126, 137, 142, 178, 203]
- ✓ ο αριθμός των μονάδων των ενδιάμεσων στιβάδων [13, 14, 51, 58, 59, 66 - 70, 72, 81, 90 - 93, 101, 103, 108, 115, 116, 122, 126, 131, 133, 134, 137, 142, 175, 177, 178, 180, 205, 206, 208, 232 - 234],
- ✓ ο αριθμός των ενδιάμεσων στιβάδων [6, 13, 14, 43, 69, 81, 93, 175, 234],
- ✓ το πλήθος των περιόδων [68, 90, 103, 108, 137, 178, 208],
- ✓ ο τύπος της συνάρτησης [68, 69, 137, 142, 175, 205, 234],
- ✓ ο τύπος του χρησιμοποιηθέντος αλγορίθμου [69, 70, 93, 101, 116, 203],
- ✓ ο ρυθμός εκπαίδευσης [13, 14, 68, 72, 91, 103, 108, 126, 131, 133, 137, 142, 175, 177, 178, 180, 205, 208, 233],
- ✓ η ορμή [13, 14, 68, 91, 103, 108, 126, 131, 137, 142, 180, 208, 233],
- ✓ η προ-επεξεργασία των δεδομένων [68, 69, 101].

Η επιλογή της αρχιτεκτονικής του δικτύου (αριθμός των εισερχόμενων, ενδιάμεσων, εξερχόμενων νευρώνων και των ενδιάμεσων στιβάδων, § 4.2.1) εξαρτάται από το πρόβλημα που μελετάμε (problem-dependent). Πιο συγκεκριμένα ο αριθμός των ενδιάμεσων στιβάδων και νευρώνων αποτελούν τις πιο κρίσιμες παραμέτρους για ένα δίκτυο ANN [80], αλλά τελικά, δεν φαίνεται να υπάρχει κάποια ξεκάθαρη μέθοδος για την εύρεση του καλύτερου συνδυασμού. Και όπως χαρακτηριστικά γράφει ο Zhang et al., "ο σχεδιασμός ενός νευρωνικού δικτύου είναι περισσότερη τέχνη, παρά επιστήμη" [69].

Οι πιο συνήθεις από τις παραπάνω παραμέτρους χρησιμοποιήθηκαν για την εύρεση του βέλτιστου δικτύου στην εργασία αυτή.

Συγκεκριμένα, σε προηγούμενες παραγράφους (§ 4.3.1, 4.3.10) αναφέρθηκε η ιδιαίτερη σημασία των πέντε πρώτων παραμέτρων. Ειδικότερα, όσον αφορά τη συνάρτηση ενεργοποίησης, η δυαδική σιγμοειδής συνάρτηση (§ 4.3.4) θεωρείται ότι μπορεί να περιγράψει πολύ καλά μη γραμμικές σχέσεις [235] και σπάνια χρησιμοποιείται οποιαδήποτε άλλη. Επιπλέον, ο αριθμός των ενδιάμεσων στιβάδων που θα μπορούσε επίσης να εξεταστεί, έχει αποκλειστεί από τους περισσότερους ερευνητές σε απλά προβλήματα [6, 17, 26, 58, 73, 77, 177, 205, 236], εφόσον η συνάρτηση ενεργοποίησης είναι συνεχής, φραγμένη και αύξουσα [6, 73]. Αυτό σημαίνει, ότι γενικά υποστηρίζεται η χρήση μίας μοναδικής ενδιάμεσης στιβάδας και περισσότερο σε απλά προβλήματα [60]. Δεν αναφέρονται συχνά στη βιβλιογραφία απόπειρες για κατασκευή μοντέλων με δυο ενδιάμεσες στιβάδες [5, 13, 14, 19, 20, 43, 64, 66, 81, 96, 113, 114, 116, 134, 185, 234], ενώ σε τέσσερις και μία μόνο περιπτώσεις αντίστοιχα έχει γίνει αναφορά για τρεις [85, 175, 185, 234] ή τέσσερις ενδιάμεσες στιβάδες [234]! Η προσθήκη στιβάδων γενικά δεν φαίνεται να βελτιώνει τα αποτελέσματα, αλλά αντιθέτως αυξάνει το χρόνο σύγκλισης (λόγω περισσότερων βαρών και νευρώνων), αλλά και την πιθανότητα εγκλωβισμού σε τοπικά ελάχιστα.

Επιπλέον, σχετικά με τη βελτιστοποίηση στον αριθμό των περιόδων, η διαδικασία εκπαίδευσης έχει προγραμματιστεί, ώστε να σταματά όταν ανιχνεύεται υπερ-προσαρμογή (over-fitting, § 4.3.1). Δηλαδή στην περίπτωση που το σφάλμα μειώνεται για την ομάδα εκπαίδευσης, ενώ αυξάνεται για την ομάδα επικύρωσης, η διαδικασία εκπαίδευσης του δικτύου σταματά αυτόματα (μέθοδος του πρώιμου τερματισμού, § 4.3.1).

Τελικά, στην περίπτωση των δεδομένων που θα εξετάσουμε, έγινε βελτιστοποίηση, μέσω των παρακάτω παραμέτρων:

- ✓ τον αριθμό των μονάδων της ενδιάμεσης στιβάδας,
- ✓ τον αριθμό των εισερχόμενων μεταβλητών,
- ✓ το ρυθμό εκπαίδευσης,
- ✓ το μέγεθος της ομάδας εκπαίδευσης.

Το κριτήριο που χρησιμοποιήθηκε για την αξιολόγηση των παραμέτρων είναι το **σφάλμα στην πρόβλεψη της ομάδας επικύρωσης** (selection error, RMS, § 4.3.7). Σε αντιδιαστολή, συγκρίνεται εκλεκτικά με το κριτήριο **της τελικής απόδοσης για την ομάδα επικύρωσης** (selection performance). Η συνάρτηση ενεργοποίησης στην εξωτερική στιβάδα είναι η softmax, η οποία μετατρέπει την τιμή που δέχεται ώστε να περιέχεται στο διάστημα (0, 1). Επιπλέον το άθροισμα των τιμών είναι ίσο με τη μονάδα. Έτσι, τα αποτελέσματα μπορούν να ερμηνευτούν σαν πιθανότητες ώστε να ανήκει το αντικείμενο σε μια ομάδα [81]. Η

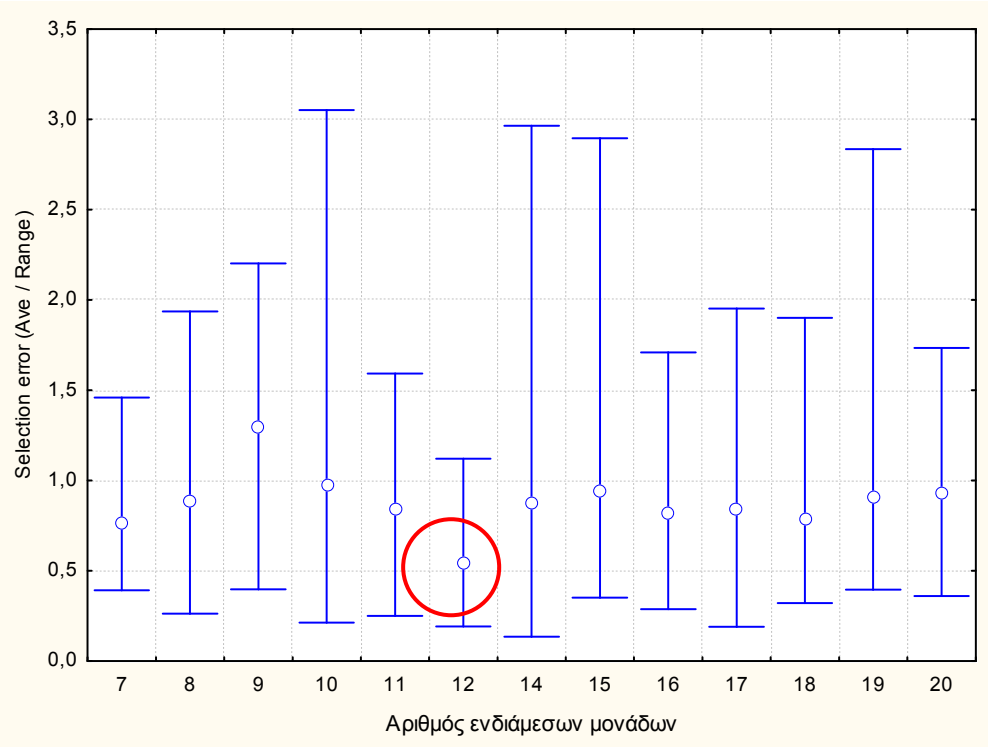
κωδικοποίηση των εξερχομένων γίνεται με τη χρήση συναρτήσεων εντροπίας (cross-entropy) ως συναρτήσεις σφάλματος οι οποίες παρέχουν τη δυνατότητα βελτιστοποίησης με μέγιστες πιθανότητες [17].

Χρησιμοποιήθηκαν αρχικά 4 εισερχόμενες μεταβλητές (V, Ni, As, B), οι πιο κρίσιμες δηλαδή, όπως αυτές προέκυψαν από την ανάλυση DA (πίνακας 5.9). Οι μεταβλητές αυτές προσαρμόστηκαν αυτόματα στην κατάλληλη κλίμακα μέσω του λογισμικού (§ 4.4.4).

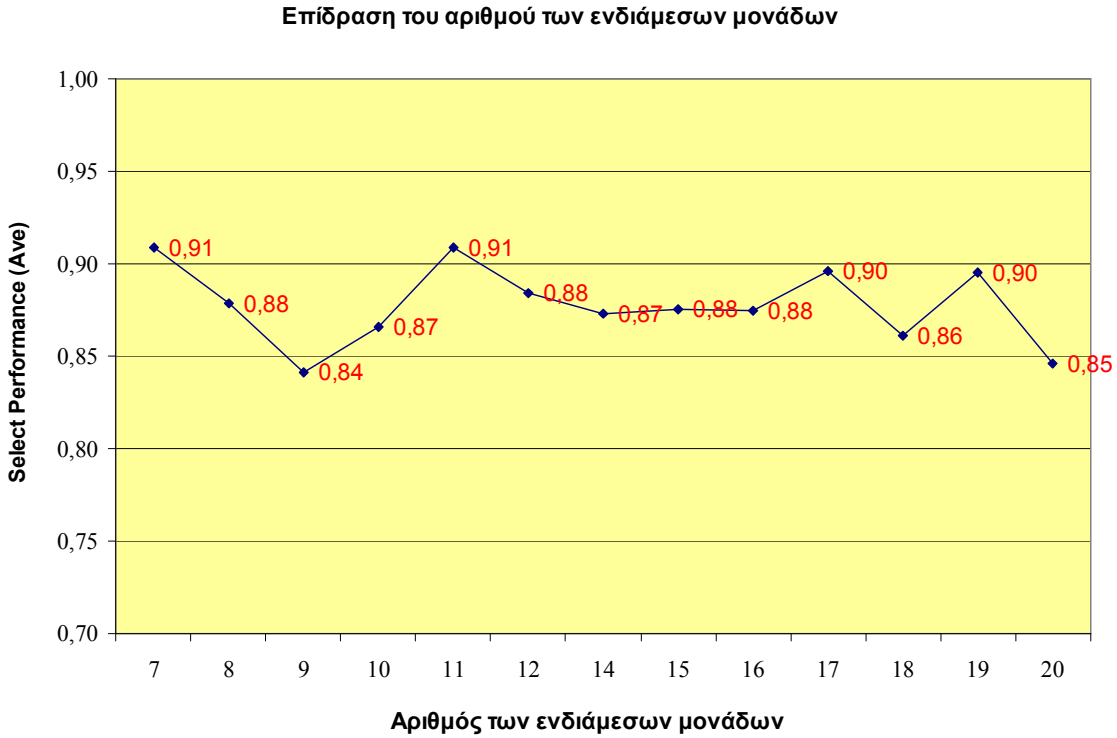
Για κάθε εύρος αριθμού μονάδων της ενδιάμεσης στιβάδας ελέχθησαν 20 διαφορετικά δίκτυα με διαφορετικές αρχικές συνθήκες (§ 4.3.10, 4.4.1), όπως αυτά συντίθενται βαθμιαία μέσω του λογισμικού Statistica. Οι ομάδες εκπαίδευσης, επικύρωσης και ελέγχου μπορούν να διατηρηθούν σταθερές σε όλες τις δοκιμές ως προς τον αριθμό αλλά και τα επιλεγμένα δείγματα, ώστε να μπορεί να γίνει σύγκριση μεταξύ των μοντέλων. Μετά τη διαμόρφωση του τελικού μοντέλου, μπορεί η επιλογή των δειγμάτων να γίνεται κάθε φορά τυχαία, ώστε να επιτευχθεί ο καλύτερος/αντιπροσωπευτικότερος διαχωρισμός αυτών. Εναλλακτικά, η επιλογή των δειγμάτων μπορεί να είναι τυχαία εξαρχής, και πραγματοποιούνται πολλές επαναλήψεις σε κάθε μοντέλο [17]. Στην εργασία αυτή, χρησιμοποιείται η εναλλακτική πορεία. Επιλέγονται διάφορα εύρη στον αριθμό των μονάδων σε κάθε δοκιμή (όπως 2-8, 6-12, 10-16, 14-20, 2-10, ….13-20,….6-10, 8-18) ώστε να επικυρωθεί το αποτέλεσμα. Έτσι, ένα μεγάλο εύρος βαρών και της προκατάληψης εξερευνώνται [123], και οι συνθήκες έναρξης για την "κατασκευή" των μοντέλων είναι ανεξάρτητες μεταξύ τους. Με τον τρόπο αυτό, επικυρώνεται **η μη τυχαιότητα του αποτελέσματος** [137], αποφεύγονται τοπικά ελάχιστα (πλήρης διαδικασία εκπαίδευσης) και η "παραλυσία" του δικτύου [51, 58, 59, 68, 92, 93, 101, 114, 206], ενώ ελέγχεται και **η σταθερότητα** ("robustness") των δικτύων [51, 135]. Εξάλλου, η στατιστική στηρίζεται στην αξία της επανάληψης για την εκτίμηση της διακύμανσης (replication, § 5.3.6).

Το σχετικό διάγραμμα δοκιμής για τον βέλτιστο αριθμό μονάδων της ενδιάμεσης στιβάδας, φαίνεται στο σχήμα 5.10. Η αρχιτεκτονική με 12 μονάδες, αποδείχθηκε η πιο αποτελεσματική με μέσο όρο για το **RMS = 0,54,** αλλά και η πιο σταθερή με τη μικρότερη διακύμανση σφάλματος στις επαναλαμβανόμενες δοκιμές. Είναι φανερό, ότι δομές πολύπλοκων δικτύων με πολλές ενδιάμεσες μονάδες δεν δίνουν πάντα την καλύτερη απόδοση.

Μπορεί επίσης να γίνει σύγκριση με το κριτήριο της απόδοσης στην ομάδα επικύρωσης (selection performance, σχ. 5.11). Η διαφοροποίηση για όλες τις τιμές των μονάδων (από 7 - 20) είναι ελάχιστη και δεν μπορεί να αξιολογηθεί η βέλτιστη επιλογή.



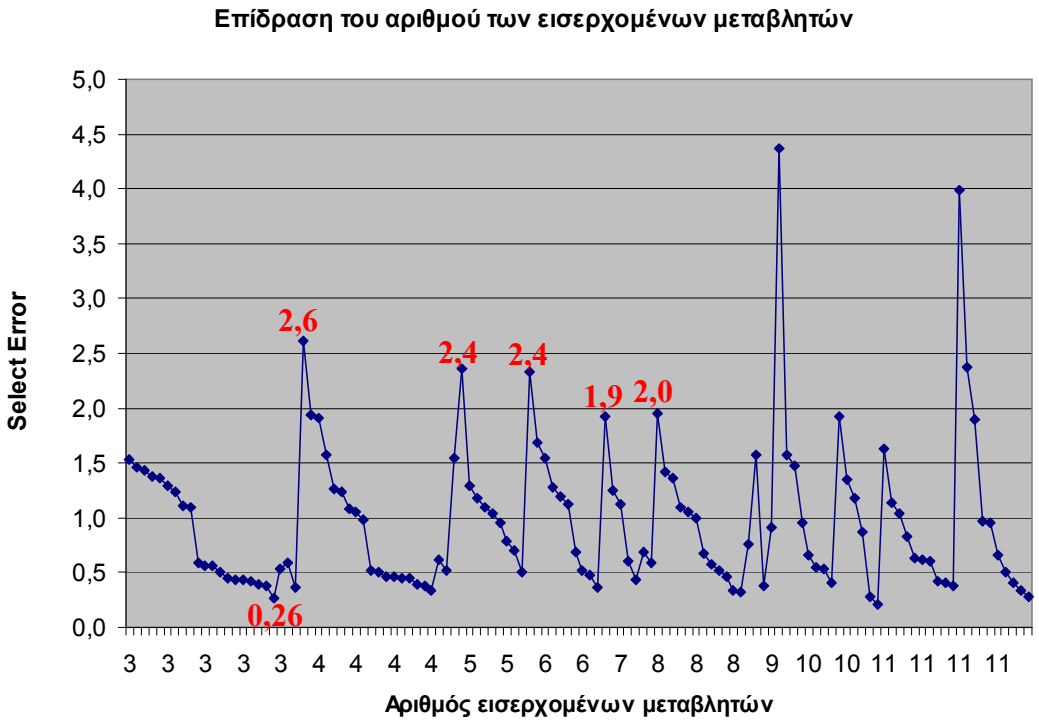
Σχήμα 5.10: Διάγραμμα της επίδρασης του αριθμού των μονάδων της ενδιάμεσης στιβάδας στην αποτελεσματικότητα του δικτύου (κριτήριο: RMS).



Σχήμα 5.11: Διάγραμμα της επίδρασης του αριθμού των μονάδων της ενδιάμεσης στιβάδας στην αποτελεσματικότητα του δικτύου (κριτήριο: Selection Performance).

Για τη βελτιστοποίηση του αριθμού των εισερχόμενων μεταβλητών, χρησιμοποιήθηκαν εκ νέου τα δεδομένα της DA (§ 5.3.1). Επιλέχθηκαν έτσι να εξεταστούν οι περιπτώσεις των 3 μεταβλητών (V, Ni, As), 4 (V, Ni, As, B), 6 (V, Ni, As, B, Cu, Mn), 8 (V, Ni, As, B, Cu, Mn, Cr, Fe) και 11 (V, Ni, As, B, Cu, Mn, Cr, Fe, Ba, Al, Zn). Εδώ αξίζει να σημειωθεί ότι επιπλέον χρησιμοποιήθηκαν συμβουλευτικά οι βηματικές προσεγγίσεις των ANN (FW και BW) αλλά και GA (§ 4.4.3). Τα αποτελέσματα ήταν απολύτως ενθαρρυντικά, αφού από τις πολλαπλές δοκιμές που έγιναν, οι μόνες μεταβλητές που δεν απορρίφθηκαν αλλά αντιθέτως θεωρήθηκαν κρίσιμες για την κατασκευή των μοντέλων ήταν οι: B, V, Ni, και As.

Κατά τη διάρκεια των δοκιμών βελτιστοποίησης, το στατιστικό πρόγραμμα χρησιμοποιεί όσες από τις μεταβλητές χρειάζονται με μέγιστο τον αριθμό που καθορίζεται κάθε φορά από το χρήστη. Με βάση την προηγούμενη δοκιμή βελτιστοποίησης, χρησιμοποιήθηκαν σε όλες τις δοκιμές 12 ενδιάμεσες μονάδες και πραγματοποιήθηκαν 20 επαναλήψεις για κάθε επιλογή αριθμού εισερχομένων μεταβλητών. Τα αποτελέσματα φαίνονται στο διάγραμμα του σχήματος 5.12.

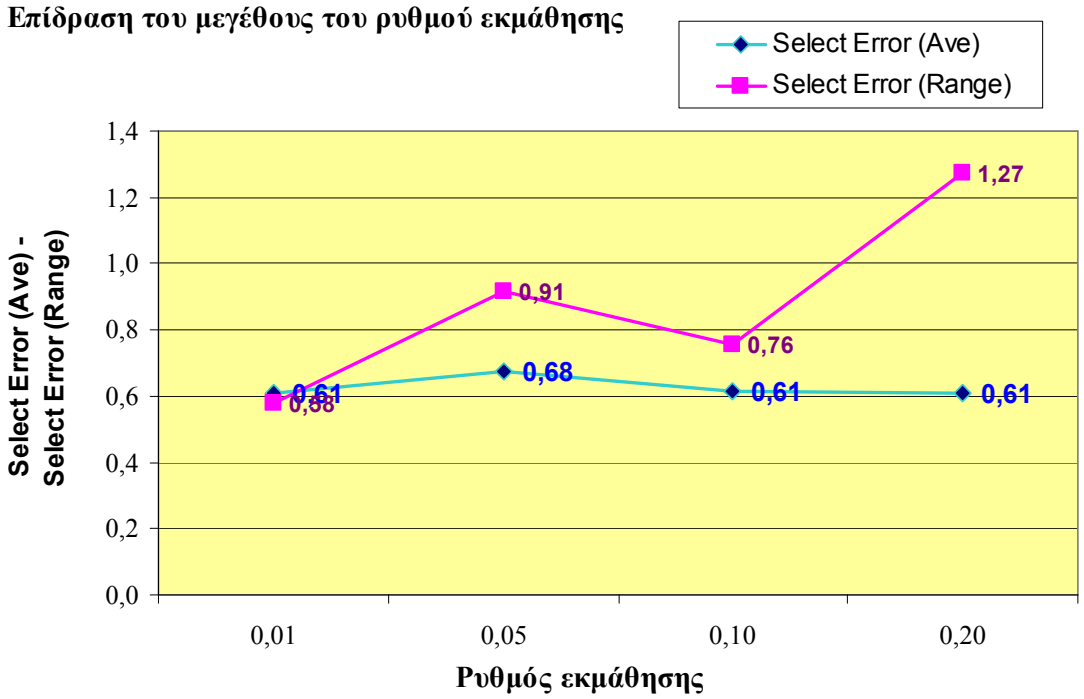


Σχήμα 5.12: Διάγραμμα της επίδρασης του αριθμού των εισερχομένων μεταβλητών στην αποτελεσματικότητα του δικτύου (κριτήριο: RMS).

Η επιλογή των 3 αρχικών μεταβλητών, έδωσε σταθερότερα χαμηλά αποτελέσματα.

Για κάθε επιλεγόμενη τιμή του ρυθμού εκπαίδευσης, (learning rate), ελέχθησαν 10 διαφορετικά δίκτυα, με σταθερές τις ήδη επιλεγμένες βέλτιστες παραμέτρους: 12 μονάδες στην

ενδιάμεση στιβάδα και 3 εισερχόμενες μεταβλητές. Οι τιμές που εξετάστηκαν είναι 0,01 (default), 0,05, 0,10 και 0,20. Τα σχετικό διάγραμμα που περιλαμβάνει τη μέση τιμή (ave) και το εύρος (range) για το RMS των 10 δοκιμών, φαίνεται στο σχήμα 5.13. Ο αρχικός ρυθμός εκπαίδευσης (0,01) έδωσε τα καλύτερα αποτελέσματα: ακριβή και επαναλήψιμα.

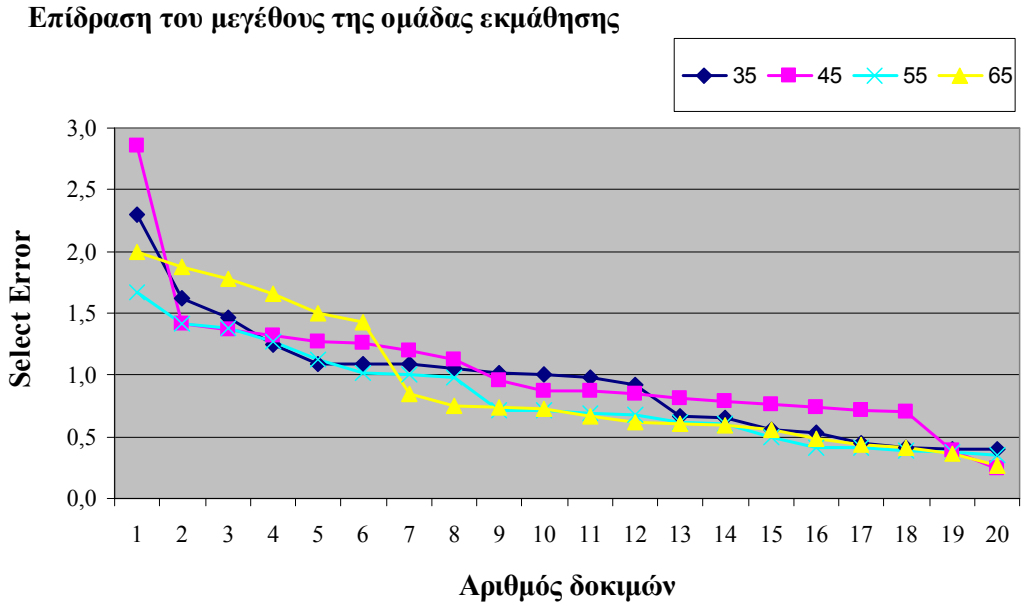


Σχήμα 5.13: Διάγραμμα της επίδρασης του ρυθμού εκπαίδευσης (learning rate) στην αποτελεσματικότητα του δικτύου (κριτήριο: RMS).

Τέλος, εξετάστηκε η επίδραση του μεγέθους της ομάδας εκπαίδευσης στην αποτελεσματικότητα του δικτύου. Η τριπλέτα του διαχωρισμού των δειγμάτων που χρησιμοποιήθηκε μέχρι τώρα (45/22/22), διατηρήθηκε ως προς το μεσαίο τμήμα της (τον αριθμό των δειγμάτων επικύρωσης = 22), εφόσον ως κριτήριο επιλογής παρέμεινε το σφάλμα σε αυτήν την ομάδα και έπρεπε να διασφαλιστεί η συνέπεια στους υπολογισμούς. Έγιναν 4 δοκιμές (20 πειράματα ανά δοκιμή) και ο πίνακας δειγματοληψίας φαίνεται παρακάτω (πίνακας 5.21), ενώ τα αποτελέσματα φαίνονται στο σχήμα 5.14. Η δοκιμή με μέγεθος 65 για την ομάδα εκπαίδευσης, φαίνεται να παρέχει σταθερότερα αποτελέσματα με πολύ καλή ακρίβεια (RMS = 0,26), χωρίς ωστόσο μεγάλες διαφορές μεταξύ των δοκιμών.

Πίνακας 5.21: Αριθμός δειγμάτων εκπαίδευσης, επικύρωσης και ελέγχου στις δοκιμές βελτιστοποίησης του δικτύου MLP.

Α/Α Δοκιμής	**Αριθμός δειγμάτων ανά ομάδα**			**Σύνολο**
	Εκπαίδευσης	**Επικύρωσης**	**Ελέγχου**	
1	35	22	32	89
2	45	22	22	89
3	55	22	12	89
4	65	22	2	89

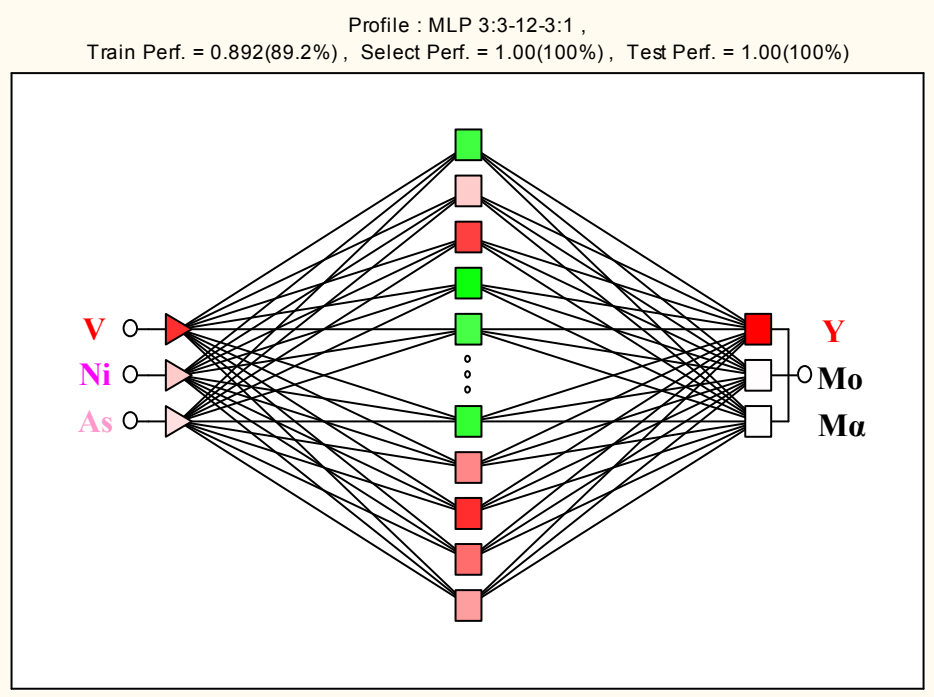


Σχήμα 5.14: Διάγραμμα της επίδρασης του μεγέθους της ομάδας εκπαίδευσης) στην αποτελεσματικότητα του δικτύου (κριτήριο: RMS).

Συμπερασματικά, μπορούμε να καταλήξουμε στη βέλτιστη (RMS = 0,26) MLP αρχιτεκτονική δομή για τα συγκεκριμένα δεδομένα που περιλαμβάνει:

- ✓ Perceptron δυο στιβάδων (μίας ενδιάμεσης) με
- ✓ 12 μονάδες (neurons) σε αυτήν και
- ✓ 3 εισερχόμενες μεταβλητές (inputs: V, Ni, As).
- ✓ Αλγόριθμοι: BP (back-propagation) και CGD (conjugate gradient descent). Ο πρώτος αλγόριθμος ανιχνεύει κατά προσέγγιση τη θέση του ελαχίστου και ο δεύτερος τερματίζει την εκπαίδευση όταν εμφανίζονται φαινόμενα υπερ-προσαρμογής [93].

Σχηματική παράσταση φαίνεται στο σχήμα 5.15. Η πρώτη εξερχόμενη μονάδα αντιστοιχεί στα δείγματα της Υλίκης. Οι νευρώνες της ενδιάμεσης στιβάδας που φαίνονται να συμμετέχουν περισσότερο σε αυτά είναι ο τρίτος από την αρχή και ο αντίστοιχος από το τέλος (συμβολίζονται με το ίδιο κόκκινο χρώμα στο σχήμα 5.15). Η αντίστοιχη μεταβλητή που συμβάλει στο χαρακτηρισμό των δειγμάτων της Υλίκης είναι το V (βλ. § 5.3.10) [17, 237]. Η συγκεκριμένη αρχιτεκτονική με 12 μονάδες στη μοναδική ενδιάμεση στιβάδα συμφωνεί απόλυτα με τους κανόνες που ανέπτυξαν οι Fernandes και Lona [80, 85] ως οδηγό για την πιο αποτελεσματική trial and error πορεία με βάση την αναλογία εισερχομένων-εξερχομένων (§ 4.3.1). Συγκεκριμένα, το μοντέλο ανήκει στην ομάδα I (CLASS I) με αριθμό εισερχομένων μεγαλύτερο των εξερχομένων και προτεινόμενη τη δομή της μιας ενδιάμεσης στιβάδας με 8-20 νευρώνες (
ΚΕΦ. 2, Θ).



Σχήμα 5.15: Αρχιτεκτονική δομή του τελικού MLP δικτύου (3:3-12-3:1).

Η χρησιμότητα του μοντέλου αναδεικνύεται από το γεγονός ότι αρκούν τρεις (3) μόνο μεταβλητές (V, Ni και As) για την εξαγωγή συμπερασμάτων. Η ακρίβεια του νέου μοντέλου, επιβεβαιώθηκε με μια σειρά αποτελεσμάτων που ελήφθησαν από μεταγενέστερη χρονική περίοδο (12/2007) από τους ίδιους ταμιευτήρες. Ο πίνακας 5.22 συνοψίζει τα αποτελέσματα. Παρατηρείται μία μόνο λαθεμένη πρόβλεψη σε σύνολο 14 δειγμάτων. Συγκεκριμένα, το σημείο αυτό αντιστοιχεί στη θέση 9702 του Μαραθώνα (θέση με Α/Α 12 στον πίνακα 5.1) και περιέχει νερό από την υπερχείλιση του καναλιού που έρχεται από το Μόρνο (μέχρι τον Οκτώβριο του 2007) ή την Υλίκη (όταν έγινε δειγματοληψία για τα δείγματα της ομάδας ελέγχου, βλ. § 5.3.1).

Το σημείο αυτό "αντιστοιχεί" εξίσου σε Υλίκη και Μόρνο. Η "σύγχυση" λοιπόν των δειγμάτων είναι αναμενόμενη. Σκεπτόμενοι δε αντιστρόφως, θα μπορούσε η "απόδοση" του δείγματος στην Υλίκη ή Μόρνο να οδηγήσει σε συμπεράσματα για τη σύσταση του νερού στο κανάλι.

Είναι λοιπόν παρήγορο να ξέρουμε ότι πολύτιμη πληροφορία μπορεί να εξαχθεί ακόμα και από λάθος προβλέψεις [1].

Πίνακας 5.22: Προβλέψεις σε νέα δείγματα με βάση τα βέλτιστα μοντέλα MLP, RBF και Kohonen 3:3-8-3:1 (βλ § 5.3.9, 5.3.10). Οι τιμές των V, Ni και As αναφέρονται σε μg/L.

Μόρνος (ΜΟ)			**Μαραθώνας (ΜΑ)**			**Υλίκη (Υ)**			**Πρόβλεψη**	
V	**Ni**	**As**	**V**	**Ni**	**As**	**V**	**Ni**	**As**		
0,42	1,16	0,25							ΜΟ	√
0,20	0,37	0,16							ΜΟ	√
0,70	0,43	0,41							ΜΟ	√
0,78	0,41	0,42							ΜΟ	√
0,66	0,33	0,33							ΜΟ	√
0,42	0,35	0,19							ΜΟ	√
0,39	0,33	0,22							ΜΟ	√
			1,22	2,02	2,46				ΜΑ	√
			1,13	3,07	1,04				Υ	X
			0,56	0,73	1,96				ΜΑ	√
			0,61	1,87	2,34				ΜΑ	√
						1,41	2,38	0,62	Υ	√
						0,91	3,38	0,80	Υ	√
						0,67	6,83	0,59	Υ	√

5.3.9. Εφαρμογή της Radial Basis Function (RBF) αρχιτεκτονικής

Τα επόμενα δίκτυα ANN που εξετάστηκαν, ήταν τα Radial Basis Function, RBF.

Μέσα από τη βιβλιογραφία, προτείνονται γενικά οι παρακάτω παράμετροι που επηρεάζουν την ποιότητα του μοντέλου:

- ✓ η ανοχή (tolerance), η οποία καθορίζει τον αριθμό των ενδιάμεσων μονάδων [51, 83, 100, 133, 135, 136, 137],
- ✓ η διασπορά (spread) σ [51, 133, 137],
- ✓ και σπάνια ο αριθμός των περιόδων [137].

Ο αριθμός των εισερχόμενων μεταβλητών παρέμεινε σταθερός, όπως προέκυψε από την προηγούμενη αξιολόγηση (3: V, Ni και As), ώστε να γίνει σύγκριση των δυο μοντέλων. Έτσι λοιπόν η βελτιστοποίηση αφορούσε τις παρακάτω παραμέτρους:

- ✓ τον αριθμό των μονάδων της ενδιάμεσης στιβάδας, και
- ✓ τη διασπορά.

Η ανοχή υπολογιστικά εκφράζεται με το άθροισμα των τετραγώνων του σφάλματος (sum of squared error) και καθορίζει τον αριθμό των μονάδων της ενδιάμεσης στιβάδας και επομένως την πολυπλοκότητα του δικτύου: μεγαλύτερος αριθμός μονάδων της ενδιάμεσης στιβάδας σημαίνει πολύπλοκη δομή για το δίκτυο και άρα μεγαλύτεροι χρόνοι υπολογισμού. Η διασπορά σε ένα μοντέλο RBF, πρέπει αν είναι αρκετά μεγάλη, ώστε να καλύπτει κενές περιοχές, αλλά όχι τόσο μεγάλη, ώστε όλοι οι νευρώνες να αντιστοιχούν στο ίδιο δείγμα. Η βέλτιστη διασπορά, καθορίζει τη δυνατότητα του δικτύου για καλή "γενίκευση" [133].

Το κριτήριο που χρησιμοποιήθηκε και εδώ για την αξιολόγηση των παραμέτρων είναι το σφάλμα στην πρόβλεψη της ομάδας επικύρωσης (RMS).

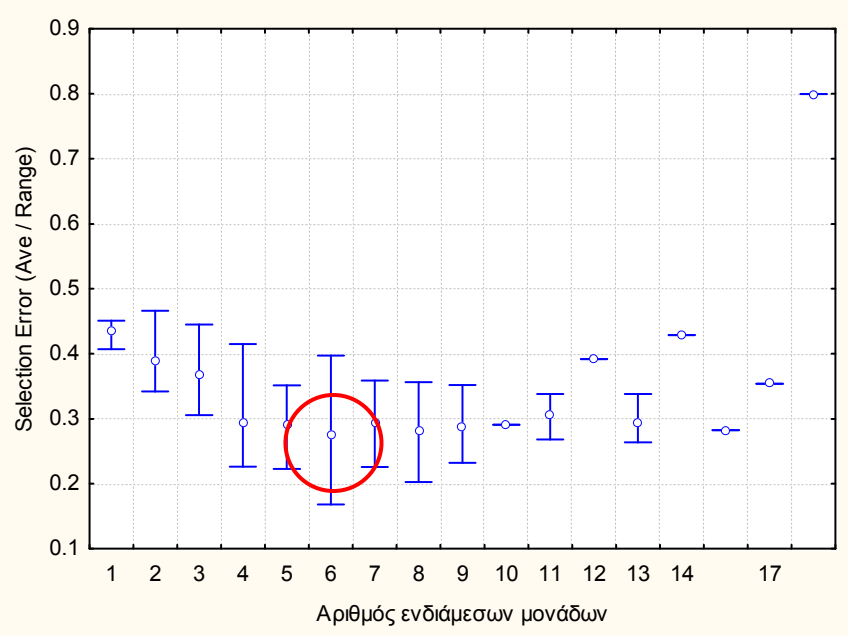
Στη βιβλιογραφία συχνά αναφέρεται ότι ο αριθμός των μονάδων της ενδιάμεσης κρυφής στιβάδας πρέπει να είναι ίσος [129, 143, 207] ή μικρότερος [137] από τον αριθμό των δειγμάτων εκπαίδευσης. Εδώ, οι μονάδες της ενδιάμεσης στιβάδας προστίθενται μία-μία μέχρι το δίκτυο να φτάσει σε αποδεκτή ακρίβεια. Για κάθε περίοδο, υπολογίζεται το άθροισμα των τετραγώνων του σφάλματος και αν αυτό είναι χαμηλότερο από το προβλεπόμενο (εφόσον καθορίζεται), η εκπαίδευση σταματά και ο αριθμός των νευρώνων στην ενδιάμεση στιβάδα θεωρείται ο βέλτιστος. Διαφορετικά, η πορεία συνεχίζεται μέχρι το σφάλμα να πέσει όσον το δυνατό χαμηλότερα [133]. Τα δίκτυα με το χαμηλότερο σφάλμα ανακτώνται αυτόματα, ενώ βελτιστοποιούνται συγχρόνως οι τιμές της διασποράς.

Συνολικά ακολουθήθηκαν δυο διαφορετικές προσεγγίσεις:

1. Μέσω του λογισμικού, συντίθενται βαθμιαία περίπου 20 διαφορετικά δίκτυα με τον αριθμό των μονάδων της ενδιάμεσης στιβάδας να κυμαίνονται σε εύρος 1 – 22 (default). Το πείραμα επαναλαμβάνεται 5 φορές (δηλαδή συνολικά ελέγχονται περίπου 100 δίκτυα), ώστε να επικυρωθεί το αποτέλεσμα.
2. Επιλέγονται διάφορα εύρη στον αριθμό των μονάδων σε κάθε δοκιμή (όπως 1-5, 4-10, 5-11, 8-15, 12-18, 15-22, 1-8, 4-12, 8-18). Για κάθε εύρος μονάδων της ενδιάμεσης

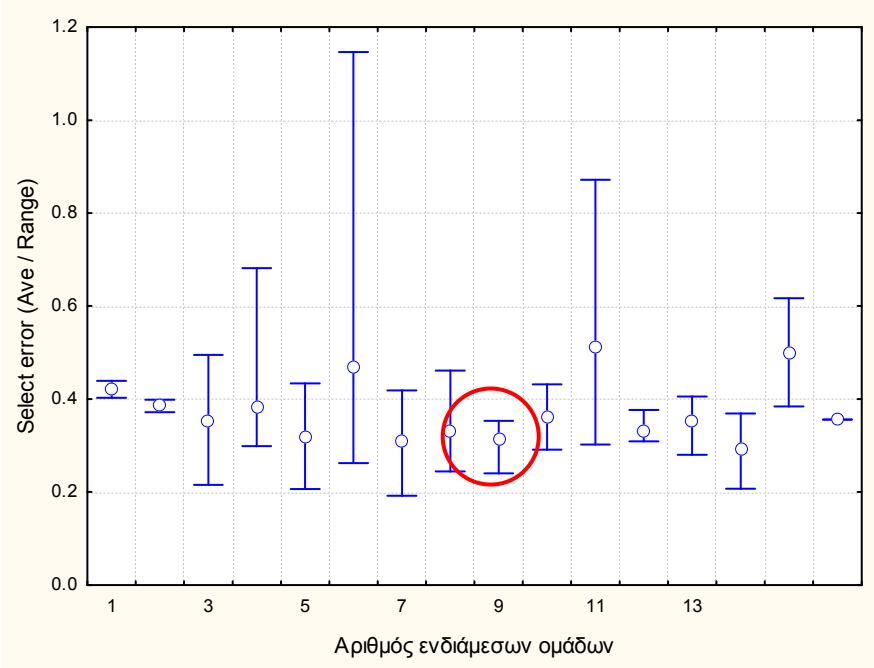
στιβάδας ελέγχονται από 12-20 διαφορετικά δίκτυα, όπως αυτά συντίθενται βαθμιαία μέσω του λογισμικού.

Τα σχετικά διαγράμματα φαίνονται στα σχήματα 5.16 και 5.17 αντίστοιχα. Στην πρώτη περίπτωση η αρχιτεκτονική με 6 μονάδες στην ενδιάμεση στιβάδα, αποδείχθηκε η πιο αποτελεσματική με μέσο όρο για το RMS = 0,28, ενώ η δομή των 9 μονάδων έδωσε επίσης χαμηλό σφάλμα (μέσο όρο για το RMS = 0,29) και μικρότερη διακύμανση.



Σχήμα 5.16: Διάγραμμα της επίδρασης του αριθμού των μονάδων της ενδιάμεσης στιβάδας στην αποτελεσματικότητα του δικτύου (κριτήριο: RMS, προσέγγιση 1).

Στη δεύτερη περίπτωση (σχ. 5.17) η αρχιτεκτονική με 9 μονάδες στην ενδιάμεση στιβάδα, έδωσε επίσης το χαμηλότερο σφάλμα (μέσος όρος για RMS = 0,31), αλλά και τη μικρότερη διακύμανση. Είναι φανερό λοιπόν, ότι καθώς αυξάνεται ο αριθμός των μονάδων, δεν βελτιώνεται ανάλογα και η απόδοση των μοντέλων [135]. Επιπλέον, συγκριτικά με τα δίκτυα MLP (σχ. 5.10), είναι φανερό το μικρότερο εύρος των επαναλήψεων στα δίκτυα RBF, τα οποία δείχνουν να είναι πιο επαναλήψιμα (§ 4.3.12).

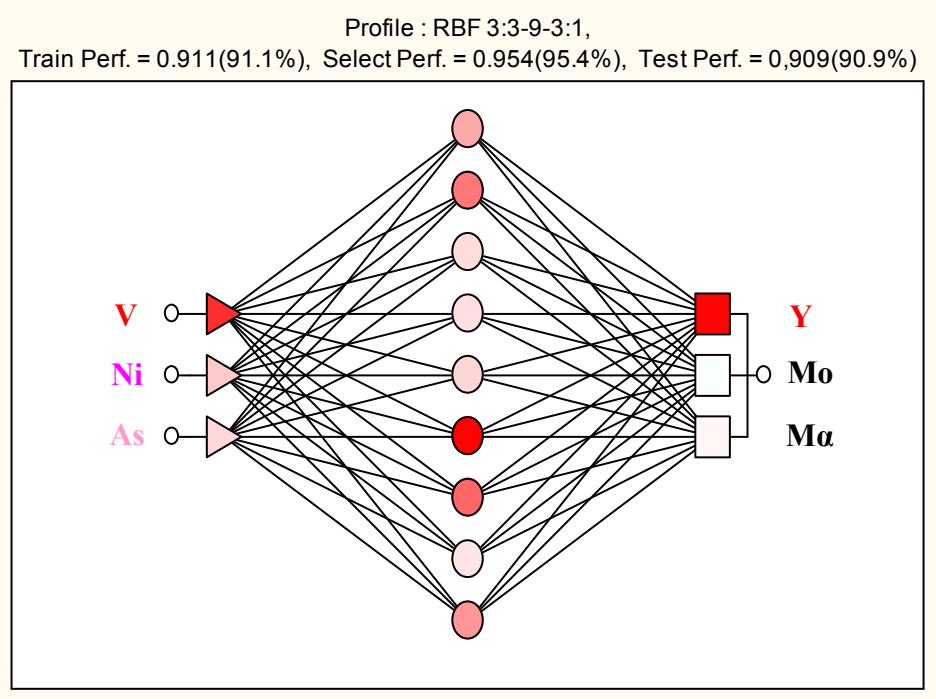


Σχήμα 5.17: Διάγραμμα της επίδρασης του αριθμού των μονάδων της ενδιάμεσης στιβάδας στην αποτελεσματικότητα του δικτύου (κριτήριο: RMS, προσέγγιση 2).

Η προηγούμενη εργασία περιελάμβανε και τη βελτιστοποίηση της διασποράς. Μπορεί λοιπόν τώρα να επιλεχθεί το βέλτιστο δίκτυο RBF (με RMS = 0,23), για τα συγκεκριμένα δεδομένα που περιλαμβάνει:

- ✓ 3 εισερχόμενες μεταβλητές (inputs V, Ni, As),
- ✓ εννιά (9) μονάδες στην ενδιάμεση στιβάδα,
- ✓ αλγορίθμους: KM (K-means), KNN (K-Nearest Neighbor) όπως περιγράφηκαν παραπάνω (§ 4.3.12) και PI (Pseudo-inverse) για τη βελτιστοποίηση της γραμμικής συνάρτησης.

Σχηματική παράσταση του δικτύου φαίνεται στο σχήμα 5.18.



Σχήμα 5.18: Αρχιτεκτονική δομή του τελικού RBF δικτύου (3:3-9-3:1).

Η ακρίβεια του νέου μοντέλου, επιβεβαιώθηκε με την ίδια σειρά αποτελεσμάτων των τριών ταμιευτήρων, που χρησιμοποιήθηκε και για την αξιολόγηση του μοντέλου MLP (12/2007). Ο πίνακας 5.22 συνοψίζει και σε αυτήν την περίπτωση τα αποτελέσματα. Παρατηρείται 1 μόνο λαθεμένη πρόβλεψη σε σύνολο 14 δειγμάτων που αφορά το γνωστό δείγμα του Μαραθώνα (§ 5.3.8).

5.3.10. Εφαρμογή της τεχνικής Kohonen

Η μη επιβλεπόμενη τεχνική Kohonen, εφαρμόστηκε ξεκινώντας από την ήδη επικυρωμένη χρήση των τριών μόνο μεταβλητών (V, Ni και As), ώστε να συγκριθούν δυνατότητες και αποτελέσματα των μοντέλων. Η συνήθης προσέγγιση της βιβλιογραφίας για τη βελτιστοποίηση των δικτύων αυτού του τύπου, αφορά τον αριθμό των νευρώνων (ή εμβαδόν του τοπογραφικού χάρτη) [51, 74, 102, 108, 115, 155, 162, 168, 181, 238, 239], την ακτίνα γειτνίασης [115, 162], τον αριθμό των περιόδων και το ρυθμό εκπαίδευσης [108, 115, 162, 239]. Σπανιότερα αναφέρεται η τοπολογική δομή (π.χ. τετράγωνη ή εξαγωνική, § 4.3.13) [102, 162], το μήκος των διαστάσεων του χάρτη, ο τρόπος αλλαγής της κλίμακας των δεδομένων [168], ή ο μαθηματικός τρόπος υπολογισμού της απόστασης (π.χ. Ευκλείδεια ή Manhattan) [162].

Σημαντικό επίσης στους τοπολογικούς χάρτες, είναι η ταξινόμηση των ομάδων και η αναγνώριση όλων των νευρώνων, ώστε να μην υπάρχουν δηλαδή σημαντικά κενά (μη αναγνωρισμένοι/ενεργοποιημένοι νευρώνες) [17, 74, 115] ή διχογνωμίες (conflicts) στην

αναγνώριση των νευρώνων (βλ. παρακάτω) [74]. Παράλληλα, είναι λογικό ότι ο αριθμός των κενών νευρώνων εξαρτάται από το μέγεθος του δικτύου και μάλιστα αυξάνεται καθώς αυτό μεγαλώνει. Θεωρείται δε ότι η σχέση αυτή είναι σχεδόν γραμμική τουλάχιστον μέχρι ορισμένου εύρους [1] και ένα καλό μέγεθος δικτύου Kohonen, είναι αυτό για το οποίο η σχέση του με τον αριθμό των κενών νευρώνων είναι γραμμική [74]. Αν ο αριθμός των νευρώνων είναι μεγάλος, υπάρχει κίνδυνος υπερ-προσαρμογής του δικτύου [158, 159], η διασπορά του δικτύου θα είναι μεγάλη και η δυνατότητα τοπογραφικής εγγύτητας μειωμένη [168], ενώ θα υπάρχει σημαντικός αριθμός κενών και ελάχιστα δείγματα θα ενεργοποιούν κάθε νευρώνα. Αντίθετα, όταν χρησιμοποιηθούν λίγοι νευρώνες, διαφορετικές ομάδες εισερχομένων θα ενεργοποιήσουν/"πέσουν" στους ίδιους νευρώνες [102] και μικρές διαφορές δεν θα αναγνωρίζονται [159] (γεγονός που συνεπάγεται και τη μειωμένη ακρίβεια του δικτύου [168]). Στα μικρά δίκτυα, αυξάνεται ο αριθμός των νευρώνων που ενεργοποιούνται από διαφορετικές ομάδες δειγμάτων (conflicts) και έτσι δεν αναγνωρίζονται καθόλου ("unknown", βλ. παρακάτω). Θα πρέπει λοιπόν, να βρεθεί μια μέση οδός μεταξύ του αριθμού των κενών και των μη αναγνωρισμένων νευρώνων. Βέβαια, οι κενοί νευρώνες δεν θεωρούνται πάντα "χαμένοι" νευρώνες, καθώς εξυπηρετούν τη διευθέτηση των δειγμάτων ελέγχου που δεν έχουν καμιά σχέση με την ομάδα εκπαίδευσης (novelty detection, § 4.3.13) [74].

Έτσι, προτείνεται ο αριθμός των νευρώνων να είναι ο διπλάσιος από τις αναμενόμενες ομάδες [51] ή διαστάσεων 2N × 2N, όπου N ο αριθμός των ομάδων [179]. Οι Jin et al. [168], Astel et al., [215, 230] Garcia και González [33] και Lamrini et al. [164] επίσης, χρησιμοποιούν τη σχέση:

$$\text{αριθμός νευρώνων} = 5 \times \sqrt{\text{αριθμός δειγμάτων εκμάθησης}},$$

για να υπολογίσουν τον καταλληλότερο αριθμό των νευρώνων για το δίκτυο Kohonen που θα εφαρμόσουν. Οι Alvarez-Guerra et al. [159] και Jin et al. [168] εκμεταλλεύονται την παραπάνω σχέση χρησιμοποιώντας παράλληλα τον αλγόριθμο του Alhoniemi ο οποίος αξιοποιεί το λόγο των δυο μεγαλύτερων ιδιοτιμών των εισερχομένων. Από την άλλη πλευρά, ο Zhang et al. [240], χρησιμοποιεί δίκτυα Kohonen στα οποία ο αριθμός νευρώνων είναι περίπου ίδιος με τον αριθμό των δειγμάτων. Το ίδιο συνιστά και ο Lischeid [166], ενώ παράλληλα προτρέπει στη χρήση μη συμμετρικών δομών Kohonen (όχι τετράγωνους χάρτες, εφόσον αυτοί πρέπει να αντανακλούν διαφορετικές διακυμάνσεις κατά μήκος των δυο αξόνων του χάρτη).

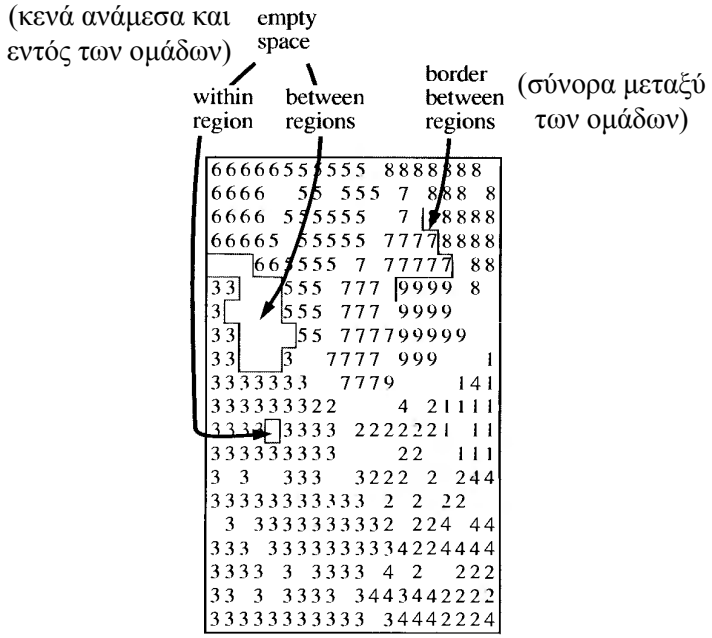
Αρχίζοντας λοιπόν από την πιο απλή δομή (λιγότεροι νευρώνες) και συνεχίζοντας προς ανώτερες δομές, εξετάστηκαν οι παρακάτω πιθανότητες:

- ✓ μονοδιάστατες δομές με 8, 9, 12, 15, ….50 νευρώνες,

- ✓ δισδιάστατες δομές με 4x2 (8) νευρώνες, 4x3 (12) νευρώνες, 4x4 (16) νευρώνες, 5x4 (20) νευρώνες, 5x5 (25) νευρώνες, 5x6 (30) νευρώνες, 6x6 (36) νευρώνες, 7x6 (42) νευρώνες.

Για λόγους που ήδη έχουν αναφερθεί, η βάση των 89 δειγμάτων δεν επιτρέπει διάταξη με περισσότερους νευρώνες.

Όλες δοκιμές, έγιναν σε δυο φάσεις με τα (default) χαρακτηριστικά που φαίνονται στο παράρτημα της διατριβής, (ΚΕΦ. 1, Π). Εξάλλου, η απόδοση των δικτύων Kohonen σε αντιδιαστολή με άλλα μοντέλα ANN, δεν φαίνεται να επηρεάζεται από τις παραμέτρους της πορείας εκπαίδευσης [166].



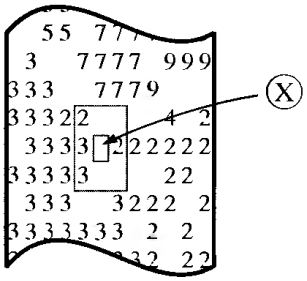
Σχήμα 5.19: Σύνορα ανάμεσα σε ομάδες, και κενά ανάμεσα και εντός των ομάδων [1, 74].

Τα δίκτυα Kohonen είναι μη επιβλεπόμενη τεχνική, αλλά στην περίπτωση των δεδομένων που εξετάζουμε, οι ομάδες των δειγμάτων είναι γνωστές (οι τρεις ταμιευτήρες της Αττικής). Επομένως, η μέθοδος μπορεί να αξιολογηθεί ως επιβλεπόμενη με βάση τα αποτελέσματα κατάταξης, μετά τον χαρακτηρισμό (ετικετοποίηση) των νευρώνων. Κάθε νευρώνας αποκτά μια "ετικέτα" (χαρακτηρισμό) με βάση την ομάδα των δειγμάτων που ενεργοποιούν το νευρώνα. Είναι επιθυμητό, εφόσον περισσότερα από ένα δείγματα ενεργοποιούν ένα νευρώνα, αυτά να προέρχονται από την ίδια ομάδα, αλλά ωστόσο αναμένεται ότι το μεγαλύτερο ποσοστό (και όχι όλα) από αυτά, θα μπορούν να πληρούν τη συνθήκη αυτή. Η περιοχή της οποίας οι νευρώνες ενεργοποιούνται από την ίδια ομάδα δειγμάτων μπορούν να σχηματίσουν "σύνορα" με τις άλλες ομάδες ή να διαχωρίζονται από κενά που ανταποκρίνονται σε νευρώνες που δεν

ενεργοποιήθηκαν ποτέ. Κενά μπορεί να εμφανίζονται και μεταξύ των νευρώνων της ίδιας ομάδας (σχ. 5.19) [1, 74].

Αρκετές φορές επίσης, μπορεί ο ίδιος νευρώνας να ενεργοποιηθεί από δείγματα διαφορετικών ομάδων (το φαινόμενο αναφέρεται ως **" διχογνωμία"** (conflict) [74])**.** Στην περίπτωση αυτή ο νευρώνας μπορεί να χαρακτηριστεί "**μη αναγνωρισμένος"** (unknown) ή μπορούν να χρησιμοποιηθούν διάφορες μέθοδοι για το χαρακτηρισμό του.

Μια μέθοδος που μπορεί να χρησιμοποιηθεί είναι η **KL-N Κοντινότερες περιπτώσεις** (KL-Nearest cases**),** η οποία σε αναλογία με την KNN (§ 4.3.12), καθορίζει την ομάδα που θα ονομάσει το νευρώνα από ένα συγκεκριμένο αριθμό γειτόνων K. Η πλειονότητα αυτών L, "αποφασίζει" την ομάδα του κεντρικού νευρώνα. Αν λιγότεροι από L νευρώνες χαρακτηρίζονται από την ίδια ομάδα, ή υπάρχει "ισοπαλία", ο νευρώνας δεν αναγνωρίζεται (σχ. 5.20) [1].

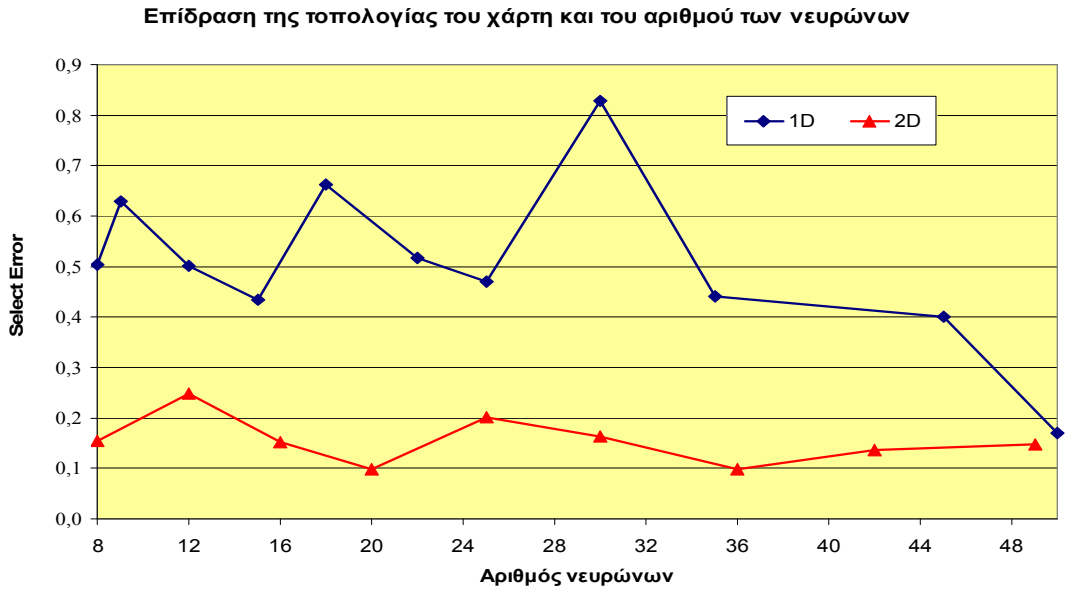


Σχήμα 5.20: Από τους οκτώ (K = 8) γείτονες ενός μη αναγνωρισμένου νευρώνα X, δύο ανήκουν στην ομάδα 2, δύο στην 3 και δύο παραμένουν κενοί. Ο νευρώνας X δεν αναγνωρίζεται [1, 74].

Στην εργασία αυτή, χρησιμοποιήθηκε η **μέθοδος Voronoi** με minimum proportion = 0,8. Αυτό σημαίνει ότι με την προϋπόθεση, ότι τουλάχιστον το 80 % των δειγμάτων που έχουν "πέσει" πάνω στο νευρώνα ή τους άμεσους γείτονές του, είναι της ίδιας ομάδας, αυτά χρησιμοποιούνται για να ονομάσουν το νευρώνα [17].

Πραγματοποιήθηκε μία δοκιμή σε κάθε δομή (το σφάλμα διατηρείται γενικά σε χαμηλά επίπεδα) και τα αποτελέσματα αξιολογήθηκαν με βάση το κριτήριο του χαμηλότερου RMS στην ομάδα επικύρωσης.

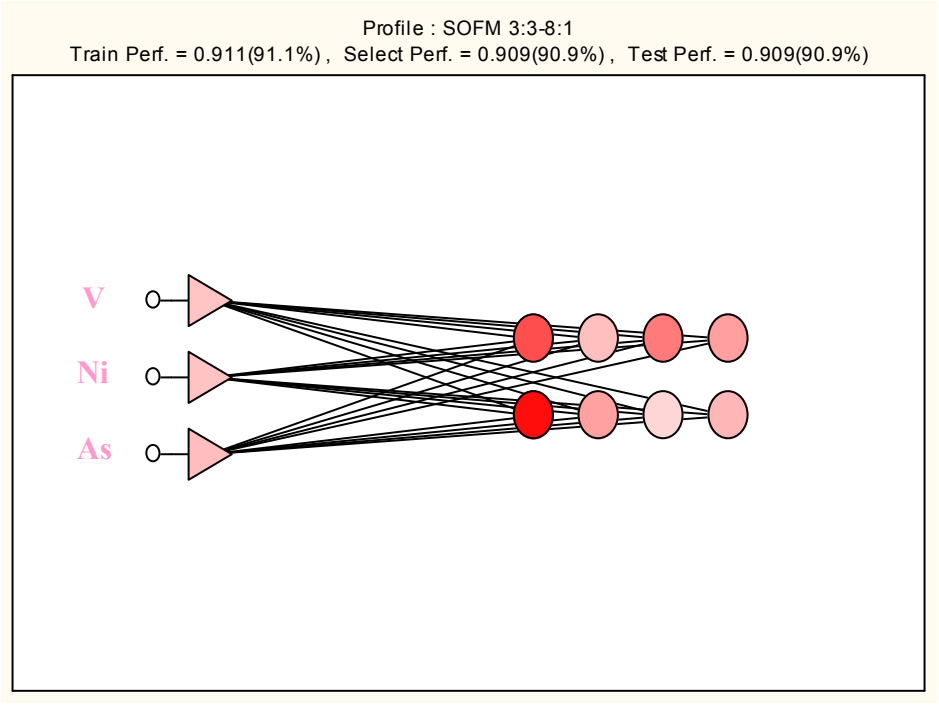
Το διάγραμμα 5.21 απεικονίζει συγκριτικά τα αποτελέσματα για τις δυο κατηγορίες των δομών που εξετάστηκαν (μονοδιάστατες 1D και δισδιάστατες 2D). Οι δισδιάστατες δομές με 20 (5x4) και 36 (6x6) νευρώνες, αποδείχθηκαν οι πιο αποτελεσματικές με RMS = 0,098. Ωστόσο, η απλούστερη δισδιάστατη δομή με 8 (4x2) νευρώνες, έδωσε επίσης πολύ χαμηλό RMS = 0,15, με υψηλές αποδόσεις για τις ομάδες εκπαίδευσης και ελέγχου (training/test performance). Παράλληλα, η δομή είναι λιγότερο πολύπλοκη και επομένως επιρρεπής σε φαινόμενα υπερπροσαρμογής, ενώ ο αντίστοιχος τοπογραφικός χάρτης δεν παρουσιάζει κενά (βλ. παρακάτω).



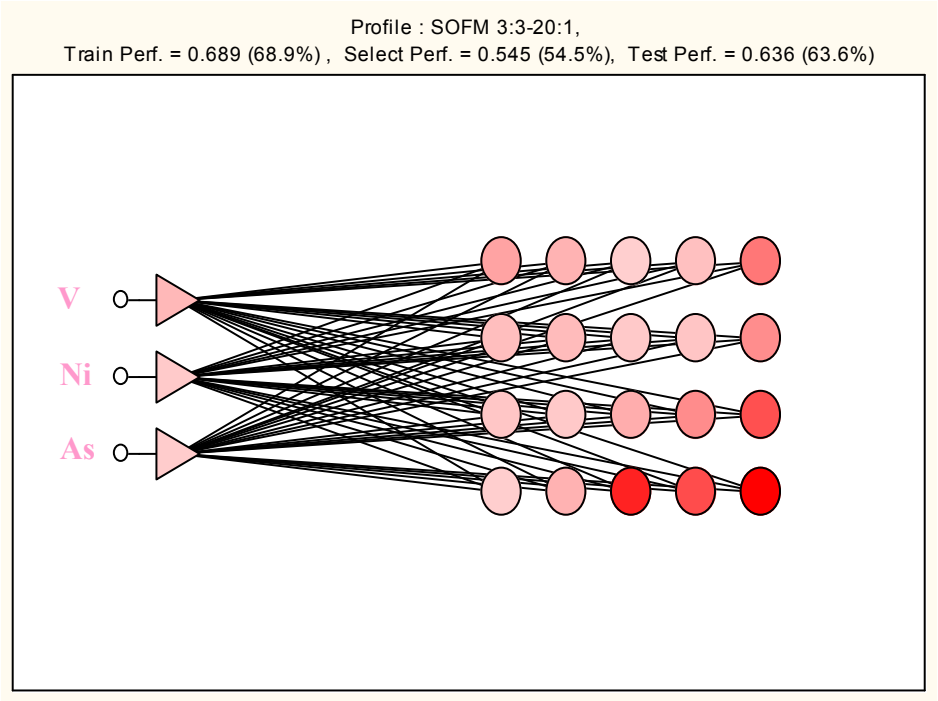
Σχήμα 5.21: Διάγραμμα της επίδρασης της δομής του χάρτη Kohonen και του αριθμού των νευρώνων στην αποτελεσματικότητα του δικτύου (κριτήριο: RMS).

Και οι δυο λοιπόν δομές (δισδιάστατες με 8 και 20 νευρώνες) εξετάστηκαν και αξιολογήθηκαν παρακάτω. Η σχηματική παράσταση της πορείας Kohonen (για την 4x2 δομή) δίνεται στο παράρτημα της διατριβής,  ΚΕΦ. 1, Π).

Σχηματική παράσταση των βέλτιστων δομών με τους αντίστοιχους χάρτες, φαίνεται στα σχήματα 5.22, 5.23.

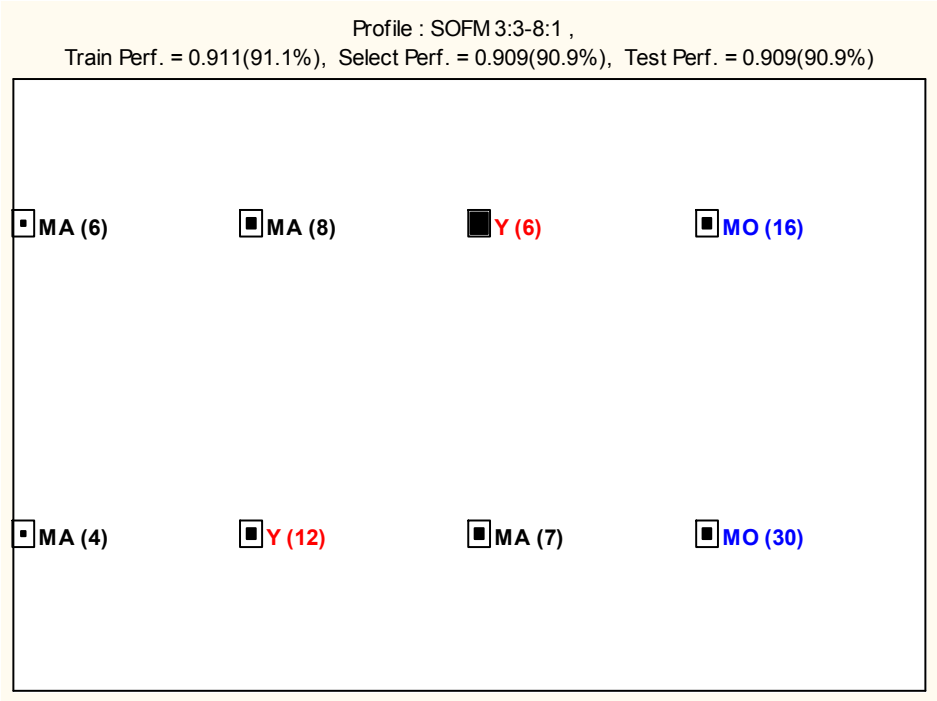


(α)

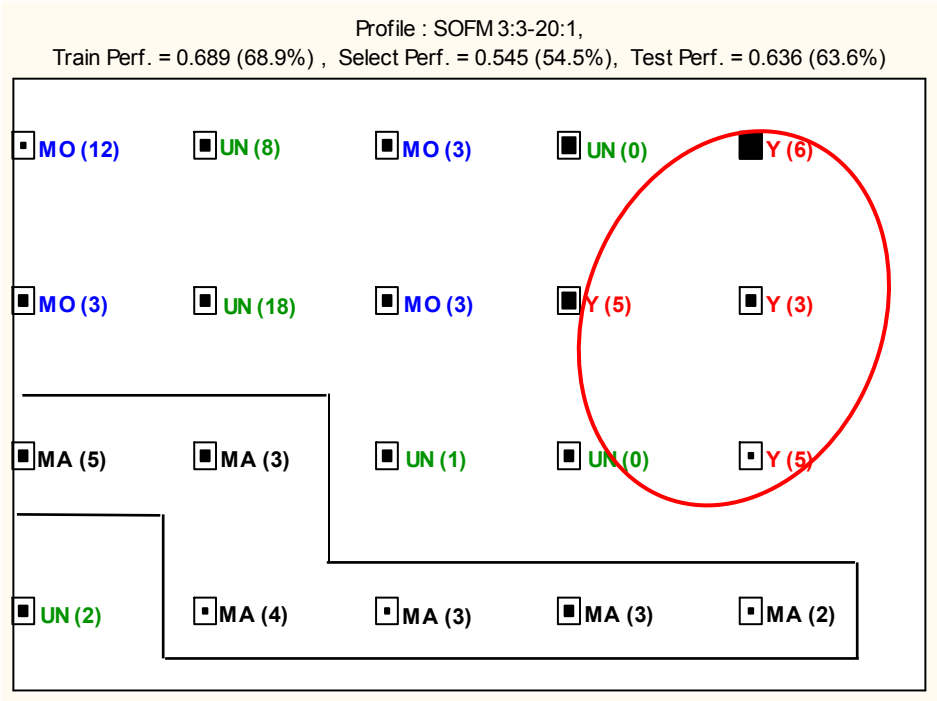


(β)

Σχήμα 5.22: Δομή των βέλτιστων δικτύων Kohonen (α) 3:3-8:1 και (β) 3:3-20:1



(α)



(β)

Σχήμα 5.23: Τοπογραφικός χάρτης των βέλτιστων δικτύων Kohonen (α) 3:3-8:1 και (β) 3:3-20:1

Οι αριθμοί εντός των παρενθέσεων, στο σχήμα 5.23 αναπαριστούν τις **"συχνότητες επιτυχίας"** (win frequencies), δηλαδή τις αντίστοιχες φορές που ο κάθε νευρώνας ενεργοποιήθηκε και ανακηρύχτηκε νικητής στη σύγκρισή του με το αντίστοιχο δείγμα. Το άθροισμα λοιπόν των συχνοτήτων και στους δυο χάρτες είναι 89. Οι νευρώνες με την ένδειξη UN (UNKNOWN, μη αναγνωρισμένοι) στο σχήμα 5.23(β), δεν έχουν ανακηρυχθεί ποτέ νικητές (συχνότητα επιτυχίας

= 0) ή ανακηρύχθηκαν νικητές από μη ικανό ποσοστό δειγμάτων, διαφορετικής ομάδας (μικρότερο του 80 % αντιστοιχούν σε κάθε ομάδα). Ο τοπογραφικός χάρτης της πρώτης δομής 3:3-8:1 (σχ. 5.23(α)) δεν περιέχει μη αναγνωρισμένους νευρώνες. Ωστόσο, η ταξινόμηση των ομάδων στο συγκεκριμένο χάρτη φαίνεται "συγκεχυμένη". Ο Μόρνος είναι ο μοναδικός που παρουσιάζει μια σαφή διάκριση σε σχέση με τις άλλες ομάδες, ενώ αντίθετα τα σημεία της Υλίκης και του Μαραθώνα δείχνουν σαφείς ομοιότητες, όπως υποδηλώνεται από τους γειτονικούς νευρώνες που τα εκπροσωπούν. Εξάλλου, τα δίκτυα Kohonen μπορούν τα "τοποθετήσουν" ομοειδή δείγματα στους ίδιους νευρώνες, αλλά κι παρόμοια δείγματα σε γειτονικούς νευρώνες (§ 4.3.13) [1]. Στο χάρτη του μοντέλου 3:3-20:1 (σχ. 5.23(β)) οι ομάδες είναι πιο εμφανείς. Οι λίμνες της Υλίκης και του Μαραθώνα φαίνεται να "καταλαμβάνουν" διακριτές περιοχές οι οποίες διαχωρίζονται από τις υπόλοιπες από σειρά μη αναγνωρισμένων νευρώνων.

Η ακρίβεια ωστόσο του απλούστερου από τα νέα μοντέλα, δεν "συμβαδίζει" με τη συγκεχυμένη εικόνα του πρώτου χάρτη (σχ. 5.23(α)), η οποία είναι στο επίπεδο του μοντέλου MLP. Αυτό επιβεβαιώθηκε με την ίδια σειρά αποτελεσμάτων των τριών ταμιευτήρων (12/2007), που χρησιμοποιήθηκε και για την αξιολόγηση των προηγούμενων μοντέλων. Ο πίνακας 5.22 συνοψίζει τα αποτελέσματα των βέλτιστων μοντέλων MLP, RBF και Kohonen 3:3-8:1, ενώ ο πίνακας 5.23 του Kohonen 3:3-20:1. Παρατηρήθηκε μία (1) μόνο λαθεμένη πρόβλεψη σε σύνολο 14 δειγμάτων και για τα δυο μοντέλα (στο δείγμα 9702), ενώ επιπλέον, το δεύτερο μοντέλο, παρουσιάζει τέσσερα (4) μη αναγνωρισμένα δείγματα.

Πίνακας 5.23: Προβλέψεις σε νέα δείγματα με βάση το επιλεγμένο μοντέλο Kohonen 3:3-20:1. Οι τιμές των V, Ni και As αναφέρονται σε μg/L.

Μόρνος (MO)			**Μαραθώνας (MA)**			**Υλίκη (Y)**			**Πρόβλεψη**	
V	**Ni**	**As**	**V**	**Ni**	**As**	**V**	**Ni**	**As**		
0,42	1,16	0,25							MO	√
0,20	0,37	0,16							UN	
0,70	0,43	0,41							MO	√
0,78	0,41	0,42							MO	√
0,66	0,33	0,33							MO	√
0,42	0,35	0,19							UN	
0,39	0,33	0,22							UN	
			1,22	2,02	2,46				MA	√
			1,13	3,07	1,04				Y	X
			0,56	0,73	1,96				MA	√
			0,61	1,87	2,34				UN	
						1,41	2,38	0,62	Y	√
						0,91	3,38	0,80	Y	√
						0,67	6,83	0,59	Y	√

Ενδιαφέρον έχει επίσης η σύγκριση των δυο μοντέλων στο σύνολο των δειγμάτων (εκπαίδευσης, επικύρωσης και ελέγχου). Ο πίνακας 5.24 περιλαμβάνει και τα δυο μοντέλα. Οι αριθμοί σε παρένθεση αντιπροσωπεύουν το 3:3-20:1, εφόσον υπάρχει διαφορά από το 3:3-8:1. Οι λανθασμένες προβλέψεις (Confusion matrix) απεικονίζεται στον πίνακα 5.25.

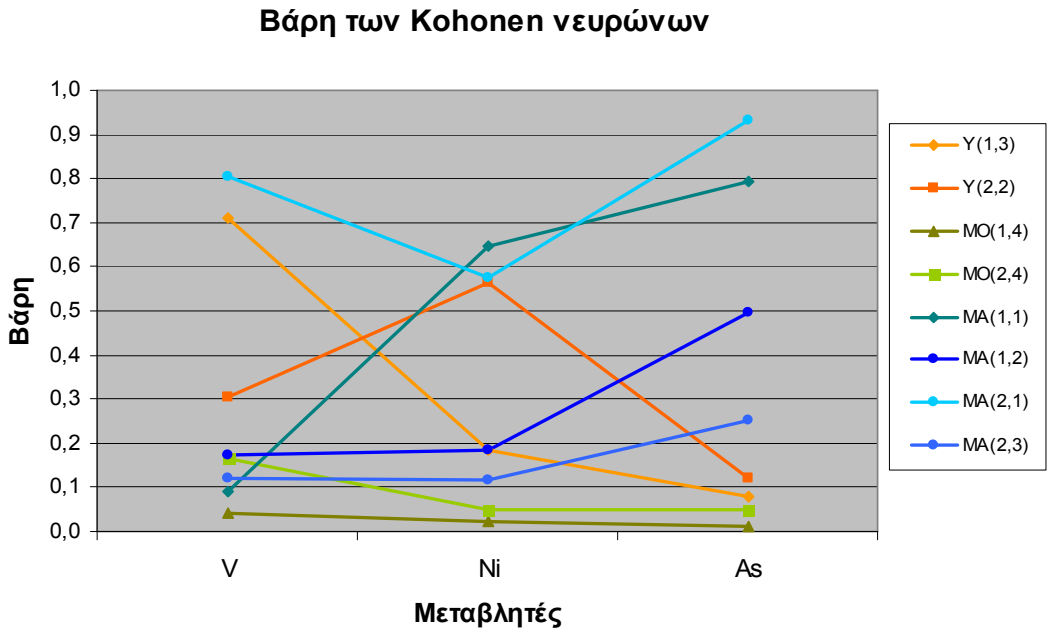
Πίνακας 5.24: Αποτελέσματα από τα μοντέλα Kohonen. Οι αριθμοί σε παρένθεση αντιπροσωπεύουν το 3:3-20:1, εφόσον υπάρχει διαφορά από το 3:3-8:1.

	Υ	**ΜΟ**	**ΜΑ**
Συνολικά	18	42	29
Σωστά	18	40 (19)	23 (20)
Λάθος	0	2 (0)	6 (3)
Μη αναγνωρισμένα	0	0 (23)	0 (6)
Σωστά %	100,0	95,2 (45,2)	79,3 (69,0)
Λάθος %	0,0	4,8 (0,0)	20,7 (10,3)
Μη αναγνωρισμένα %	0,0	0,0 (54,8)	0,0 (20,7)

Πίνακας 5.25: Λανθασμένες προβλέψεις με παρατηρούμενες (στήλες) έναντι προβλεπόμενων (σειρές) θέσεων. Οι αριθμοί σε παρένθεση αντιπροσωπεύουν το 3:3-20:1, εφόσον υπάρχει διαφορά από το 3:3-8:1.

	Υ	**ΜΟ**	**ΜΑ**
Υλίκη (Υ)		0	0 (1)
Μόρνος (ΜΟ)	0		6 (2)
Μαραθώνας (ΜΑ)	0	2 (0)	
Συνολικά	0	2 (0)	6 (3)

Για το δεύτερο μοντέλο Kohonen, τα μικρά ποσοστά επιτυχίας επικεντρώνονται στην ύπαρξη μη αναγνωρισμένων νευρώνων και όχι στο μεγάλο ποσοστό λαθεμένων προβλέψεων. Πραγματικά ο αριθμός των λαθεμένων προβλέψεων για το μοντέλο 3:3-20:1, είναι μικρότερος από τον αντίστοιχο του μοντέλου 3:3-8:1 (βλ. πίνακα 5.25). Το πρώτο μοντέλο λοιπόν, υπερέχει σε μεγάλο βαθμό, επικυρώνοντας τη σημασία της απουσίας κενών νευρώνων.



Σχήμα 5.24: "Ματιά" εντός του 3:3-8:1 μοντέλου Kohonen: βάρη των νευρώνων που αντιστοιχούν στις τρεις λίμνες. Τα ζεύγη των αριθμών στην ετικέτα του διαγράμματος αντιστοιχούν στις θέσεις των αντίστοιχων νευρώνων (σχ. 5.22(α)).

Στο διάγραμμα του σχήματος 5.24 φαίνονται τα βάρη των 8 νευρώνων του επιλεχθέντος 3:3-8:1 μοντέλου για τις τρεις μεταβλητές (V, Ni, As). Οι νευρώνες με τα μικρότερα βάρη για τις μεταβλητές ανήκουν στο Μόρνο (με τη μικρότερη "μεταλλική "επιβάρυνση). Στους νευρώνες του Μαραθώνα κυριαρχεί το As, ενώ το V χαρακτηρίζει Υλίκη και Μαραθώνα όπως είχε ήδη επισημανθεί και παλαιότερα [38].

Παραπέρα αξιολόγηση των βαρών Kohonen με τις τεχνικές PCA και DA είναι διαθέσιμη στο ηλεκτρονικό παράρτημα της διατριβής (ΚΕΦ. 1, Π).

5.4. ΣΥΓΚΡΙΣΗ ΤΕΧΝΙΚΩΝ

5.4.1. Ομαδοποίηση και ταξινόμηση θέσεων

Όλες οι παραδοσιακές τεχνικές που μελετήθηκαν σε αυτήν την εργασία, δηλαδή οι DA, PCA, CA μπόρεσαν να ομαδοποιήσουν επιτυχώς τις θέσεις δειγματοληψίας της βάσης των ταμιευτήρων της Αθήνας που εξετάστηκε. Παρόλα αυτά, οι τεχνικές των PCA, CA δεν μπορούν να ποσοτικοποιήσουν τα αποτελέσματα, ώστε να γίνουν άμεσες συγκρίσεις μεταξύ των διαφορετικών προσεγγίσεων.

Οι τεχνικές των DA και CT έδωσαν ωστόσο, αριθμητικά αποτελέσματα. Η τεχνική των CT έδωσε υψηλά ποσοστά επιτυχίας στην ομάδα εκπαίδευσης με τις τρεις χρησιμοποιούμενες

μεθόδους (LCM, Classic CT και CART). Η πρώτη από αυτές έδωσε τα υψηλότερα αποτελέσματα σε σχέση με τις υπόλοιπες αλλά κα τις τρεις προσεγγίσεις της τεχνικής DA.

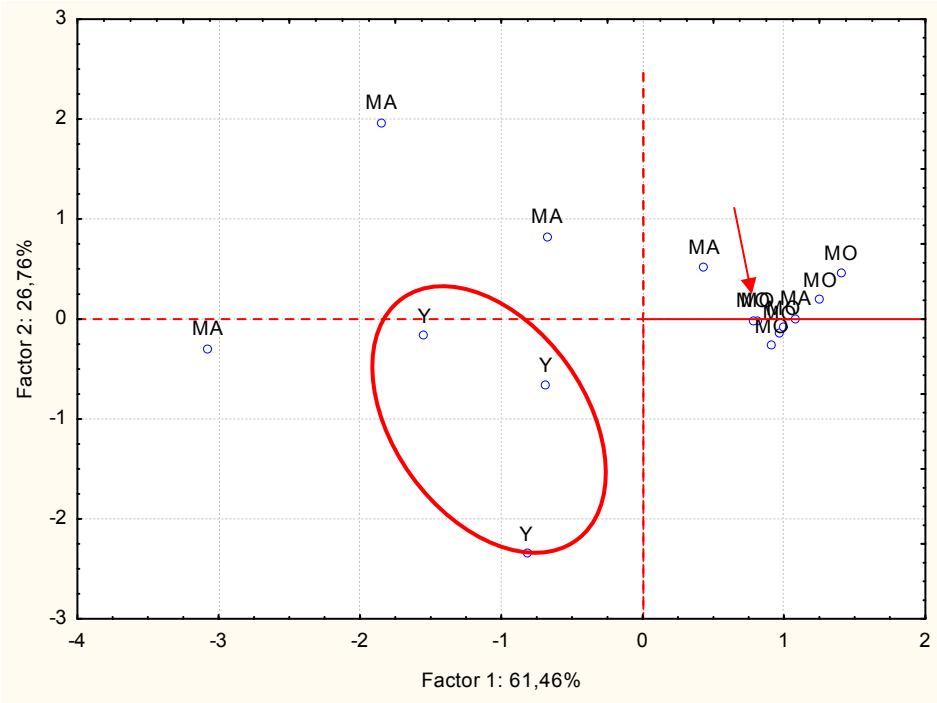
Για τα άγνωστα δείγματα της ομάδας ελέγχου ωστόσο, η Classic CT μέθοδος έδωσε τα καλύτερα αποτελέσματα.

Ακόμα περισσότερο εντυπωσιακά αποτελέσματα εξάλλου, έδωσαν τα βελτιστοποιημένα μοντέλα ANN. Χρησιμοποιήθηκαν 3 μόνο μεταβλητές (επιλεγμένες με τη βοήθεια της DA) και τα ποσοστά επιτυχίας (σε ομάδες εκπαίδευσης και ελέγχου) κυμαίνονταν σε υψηλά επίπεδα σε σχέση με τις υπόλοιπες τεχνικές με τη δυνατότητα ποσοτικοποίησης (DA και CT) που χρησιμοποίησαν το σύνολο των μεταβλητών. Τα οφέλη σε εργαστηριακό κόστος και χρόνο στην περίπτωση αυτή είναι φανερά. Παρόμοια αποτελέσματα που αφορούν στη σύγκριση των επιβλεπόμενων τεχνικών DA και ANN σε σχέση με τον αριθμό των απαιτούμενων κάθε φορά μεταβλητών, έχουν καταγραφεί στο παρελθόν και από άλλους ερευνητές [13, 14].

5.4.2. Αξιολόγηση μεταβλητών

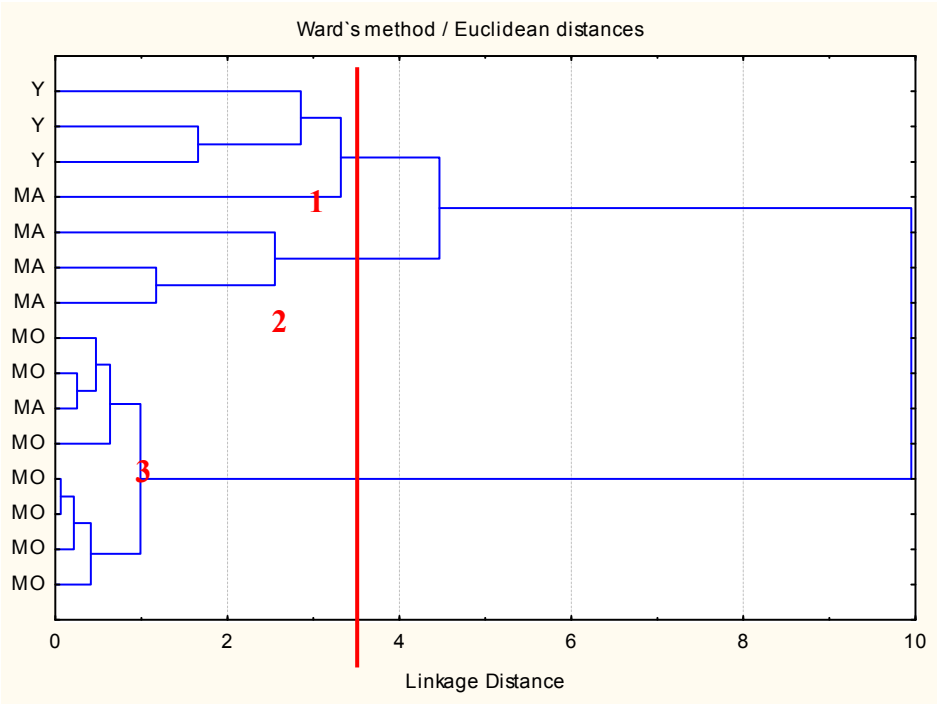
Στο σημείο αυτό, γίνεται σύγκριση των βέλτιστων μοντέλων ANN με τις παραδοσιακές στατιστικές τεχνικές DA, PCA και CA, σε επίπεδο μεταβλητών. Η τεχνική της DA ανέδειξε πρώτη τη σημασία των τριών μεταβλητών που ήδη χρησιμοποιήθηκαν (§ 5.3.1, πίνακας 5.9). Οι τεχνικές PCA και CA δεν παρέχουν άμεσους μηχανισμούς για την ανάδειξη των κρίσιμων μεταβλητών. Θα μπορούσαν λοιπόν οι τεχνικές αυτές να διαχωρίσουν το ίδιο αποτελεσματικά τις ομάδες δειγμάτων μόνο με τη βοήθεια των τριών μεταβλητών V, Ni, As;

Για να εξακριβωθεί αυτό, οι προαναφερθείσες τεχνικές εφαρμόστηκαν στα δεδομένα με τη χρήση όμως μόνο των τριών μεταβλητών. Τα αποτελέσματα είναι εντυπωσιακά (πολύ καλός διαχωρισμός για τρεις μόνο μεταβλητές) και απεικονίζονται στα σχήματα 5.25 και 5.26.



Σχήμα 5.25: Διάγραμμα συντεταγμένων από την PCA ανάλυση της βάσης δεδομένων των ταμιευτήρων της Αττικής. Χρησιμοποιήθηκαν μόνο 3 μεταβλητές: V, Ni, As.

Οι δυο λοιπόν παραδοσιακές τεχνικές επιβεβαιώνουν τα αποτελέσματα που προκύπτουν από τα μοντέλα ANN, χωρίς όμως να μπορούν άμεσα να βοηθήσουν στην εξαγωγή συμπερασμάτων για την κρισιμότητα των μεταβλητών.



Σχήμα 5.26: CA για τη βάσης των ταμιευτήρων της Αττικής. Χρησιμοποιήθηκαν μόνο 3 μεταβλητές: V, Ni, As.

Η τεχνική των CT μπορεί ωστόσο να εξάγει συμπεράσματα για την ιδιαίτερη ή όχι σημασία των μεταβλητών. Στα ιστογράμματα (§ 5.3.7, σχ. 5.9) απεικονίστηκε η σχετική ταξινόμηση των μεταβλητών για τις μονοπαραμετρικές CT μεθόδους (Classic CT και CART). Οι τρεις μεταβλητές V, Ni και As, παρουσίασαν τα μεγαλύτερα ποσοστά κρισιμότητας. Η αντίστοιχη ανάλυση για την τρίτη μέθοδο LCM εξάλλου (§ 5.3.4, πίνακας 5.14), επιβεβαίωσε αυτήν την επιλογή.

Τέλος, τα χρησιμοποιηθέντα μοντέλα ANN "εκμεταλλεύτηκαν" την ταξινόμηση των μεταβλητών όπως αυτή προτάθηκε από την DA και έδωσαν πολύ υψηλά ποσοστά πρόβλεψης, με πολύ λιγότερες μεταβλητές.

Επομένως, παρόλο που δεν εξάγονται άμεσα αποτελέσματα από όλες τις τεχνικές, αλληλοσυμπληρώνονται και καταλήγουν τουλάχιστον έμμεσα, στα ίδια συμπεράσματα σε επίπεδο θέσεων αλλά και μεταβλητών.

ΚΕΦ. 6 ΠΟΛΥΠΑΡΑΜΕΤΡΙΚΕΣ ΤΕΧΝΙΚΕΣ ΣΤΗΝ ΜΕΛΕΤΗ ΤΗΣ ΕΠΙΔΡΑΣΗΣ ΤΩΝ ΙΧΘΥΟΚΑΛΛΙΕΡΓΕΙΩΝ ΣΤΑ ΘΑΛΑΣΣΙΑ ΙΖΗΜΑΤΑ

6.1. ΕΙΣΑΓΩΓΗ

Στη διάρκεια των τελευταίων χρόνων, εκτεταμένες ιχθυοκαλλιέργειες έχουν αναπτυχθεί κατά μήκος των ηπειρωτικών και νησιωτικών ελληνικών ακτών. Αυτή η ανάπτυξη απαιτεί περαιτέρω μελέτη όχι τόσο στην υδάτινη στήλη, αλλά στα θαλάσσια ιζήματα γύρω από τους κλωβούς των ψαριών. Εξάλλου, ιζήματα και χώματα αντιπροσωπεύουν δείγματα που μπορούν να χρησιμοποιηθούν για την έρευνα ρύπων με μακροχρόνιες επιπτώσεις (ο όρος που χρησιμοποιείται στην βιβλιογραφία για το είδος αυτών των δειγμάτων είναι **"conservative"** [30]) σε αντιδιαστολή με τον αέρα και το νερό που αποτελούν κατάλληλα δείγματα για την αξιολόγηση βραχυχρόνιων επιπτώσεων και διακυμάνσεων [189]. Πραγματικά, οι ιχθυοκαλλιέργειες φαίνονται ότι επηρεάζουν το προφίλ των συγκεντρώσεων της οργανικής ύλης, θρεπτικών συστατικών [241, 242] και μετάλλων [218, 243, 244]. Εργασίες συντήρησης, ιχθυοτροφές, μεταβολικά προϊόντα και περιττώματα συνεισφέρουν επίσης στην επιμόλυνση της περιοχής.

Εργασίες που καταγράφουν την επίδραση των ιχθυοκαλλιεργειών στον πυθμένα του θαλάσσιου περιβάλλοντος, έχουν γενικά καταγραφεί στη διάρκεια των τελευταίων δυο δεκαετιών. Συγκεκριμένα:

1. Μελετάται η επίδραση ιχθυοκαλλιέργειας σε θαλάσσια ιζήματα στη Μεσόγειο (Σούνιο/Ελλάδα, Alicante/Ισπανία, Sicily/Ιταλία) με χρήση βασικών στατιστικών εργαλείων (ANOVA). Η συγκέντρωση του ολικού φωσφόρου P βρέθηκε σημαντικά αυξημένη στα παρακείμενα των κλωβών ιζήματα [216].
2. Εκτιμάται η επιφόρτιση θαλασσίων ιζημάτων και οργανισμών σε μέταλλα εξαιτίας παρακείμενων ιχθυοκαλλιεργειών [244]. Σε κάποιες από τις εργασίες αυτές PCA και CA χρησιμοποιούνται για την ομαδοποίηση των δειγμάτων και την επιλογή των κρισιμότερων μεταβλητών [219, 223]. Αλλού καταγράφονται υψηλές τιμές Cu, Zn, Fe, οργανικής ύλης και σωματιδίων < 63 μm [218, 219].
3. Ανασκοπούνται σχετικές εργασίες ώστε να καταγραφούν οι κίνδυνοι που προκύπτουν από πρόσθετα σε μοντέρνες εγκαταστάσεις ιχθυοκαλλιέργειας και σχετίζονται με τη δημόσια υγεία [242, 245, 246].
4. Διαφοροποιούνται δείγματα θαλασσίων ιζημάτων με βάση μέταλλα όπως τα Zn, Cu και τεχνικές κανονικοποίησης (lithium-normalization technique) [220]. Αυξημένα επίπεδα βαρέων μετάλλων όπως Cu, Zn, Cd και Pb καταγράφονται και αλλού [221, 222].

5. Ερευνάται η επίδραση της εντατικής ιχθυοκαλλιέργειας στην κατανομή του P στα θαλάσσια ιζήματα [217].
6. Ερευνώνται εποχιακές διακυμάνσεις στην ποιότητα των θαλασσίων ιζημάτων που συνορεύουν με μονάδες ιχθυοκαλλιέργειας. Βασικές στατιστικές δοκιμές (Kruskal-Wallis και U Mann-Whitney tests) και μέθοδοι (ANOVA) χρησιμοποιούνται για τον σκοπό αυτό [224, 225].

Σε όλες τις παραπάνω εργασίες ωστόσο, δεν πραγματοποιήθηκε καμιά συστηματική στατιστική επεξεργασία που να αφορά την ταξινόμηση των δειγμάτων ή τη συσχέτιση και κρισιμότητα των μεταβλητών.

Έτσι, στη διατριβή αυτή ασχοληθήκαμε με τη στατιστική επεξεργασία και την εφαρμογή χημειομετρικών τεχνικών (κλασικών και μη), σε μια μεγάλη βάση δεδομένων που αφορά τον προσδιορισμό στοιχείων σε δείγματα ιζημάτων (λάσπης) σε τρεις μονάδες ιχθυοκαλλιέργειας της χώρας: στη Ναύπακτο (NA), στη Χίο-Οινούσσες (CH) και στον Αστακό (AS). Συλλέχθηκαν δείγματα κοντά στους κλωβούς (μηδενική απόσταση) ή μακρύτερα από αυτούς (σε 50 και 100 m απόσταση), ενώ προσδιορίστηκαν μέταλλα, As, P, N και C. Είναι η πρώτη φορά που χρησιμοποιούνται τόσες επιβλεπόμενες και μη τεχνικές για τη διαφοροποίηση θέσεων μέσα στην ίδια ή διαφορετικών μονάδων ιχθυοκαλλιέργειας. Ειδικότερα, τεχνικές όπως τα Δέντρα Ταξινόμησης (CT) και τα Νευρωνικά Δίκτυα (ANN) εφαρμόστηκαν για πρώτη φορά επιτυγχάνοντας τη διαφοροποίηση σημείων και την αξιολόγηση μεταβλητών.

Γενικά, ακολουθήθηκαν τα παρακάτω βήματα:

1. Αρχικά έγινε καταγραφή, μελέτη και αξιολόγηση των αποτελεσμάτων. Περιβαλλοντικά, τοπολογικά δεδομένα (ακριβείς θέσεις δειγματοληψίας, χρονική διάρκεια λειτουργίας και δυναμικότητα των εγκαταστάσεων, κατεύθυνση και ταχύτητα ανέμων, τύπος του θαλάσσιου ιζήματος) ήταν απαραίτητα για το σκοπό αυτό (ΚΕΦ. 2, Π).
2. Βασικές παραμετρικές (t-test) και μη παραμετρικές (Kruskal-Wallis και U Mann-Whitney tests) δοκιμές χρησιμοποιήθηκαν για την αρχική αξιολόγηση των αποτελεσμάτων (ΚΕΦ. 2, Π).
3. Εφαρμόζοντας κλασικές μεθόδους (PCA/FA, DA) και χρησιμοποιώντας τα δεδομένα συνολικά από τις τρεις ιχθυοκαλλιέργειες, έγινε προσπάθεια αξιολόγησης και εύρεσης των συσχετίσεων των μεταβλητών, ταυτοποίησης των παραγόντων που συνεισφέρουν στη μόλυνση των υπό μελέτη περιοχών και διαφοροποίησης των μονάδων βασιζόμενοι προφανώς στις "γηγενείς" μεταβλητές. Τέτοια είναι τα στοιχεία (μέταλλα και ανόργανα) που προϋπάρχουν στο θαλάσσιο πυθμένα και δεν προέρχονται από εξωτερικούς παράγοντες. Τα σημεία που χρησιμοποιήθηκαν εδώ είναι αυτά που βρίσκονται σε απόσταση μεγαλύτερη ή ίση των 50 m από τους κλωβούς (ΚΕΦ. 2, Π).

Στο ίδιο κεφάλαιο, έγινε επίσης μια προσπάθεια ανίχνευσης εποχιακών διαφορών που τυχόν υφίστανται στις συγκεντρώσεις των στοιχείων για όλες τις μονάδες ιχθυοκαλλιέργειας. Εδώ χρησιμοποιήθηκαν μόνο τα σημεία κοντά στους κλωβούς.

4. Χρησιμοποιώντας τα αποτελέσματα από όλες τις ιχθυοκαλλιέργειες (CH, NA, AS), επιχειρήθηκε η ομαδοποίηση/ταξινόμηση των σημείων δειγματοληψίας με βάση την απόσταση αυτών από την κλωβό. Εδώ χρησιμοποιήθηκαν διάφορες πολυπαραμετρικές τεχνικές (συμπεριλαμβανομένων των KNN, DA, CT και ANN). Επισημάνθηκαν δυο κατηγορίες σημείων: DI (DISTANT) με απόσταση μεγαλύτερη ή ίση των 50 m από τον κλωβό και ZE (ZERO) με μηδενική απόσταση από τον κλωβό ιχθυοκαλλιέργειας. Με τον τρόπο αυτό, ταυτοποιήθηκαν παράμετροι που τυχόν επιφορτίζουν/επιμολύνουν το φυσικό θαλάσσιο περιβάλλον εξαιτίας της λειτουργίας των εγκαταστάσεων ιχθυοκαλλιέργειας (ΚΕΦ. 6).
5. Τα μοντέλα που προέκυψαν επικυρώθηκαν είτε με κάποια ανεξάρτητη ομάδα ελέγχου (τμήμα της αρχικής ομάδας δειγμάτων που δεν χρησιμοποιήθηκε για την εκπαίδευση αυτών) είτε με CV (§ 3.1.2, 4.4.2), ώστε:
   - να επιβεβαιωθεί η καταλληλότητα των μοντέλων από δείγματα που δεν συμμετέχουν στη διαμόρφωση αυτών,
   - να ελεγχθεί αν παράγοντες που επηρεάζουν ή φαίνονται να επιμολύνουν τη μια μονάδα ιχθυοκαλλιέργειας χαρακτηρίζουν και τις άλλες (όλες οι ομάδες δειγμάτων επιλέγονται τυχαία) ώστε να επιβεβαιωθεί η μη τυχαιότητα του αποτελέσματος (ΚΕΦ. 6).

Τέλος, αξιολογήθηκαν συνολικά οι επιπτώσεις μιας μόνιμης εγκατάστασης ιχθυοκαλλιέργειας στον περιβάλλοντα θαλάσσιο χώρο με τη συνολική στατιστική διαδικασία να επικυρώνει τα αποτελέσματα.

Συγκεκριμένα, στα δυο αυτά κεφάλαια αυτά πέντε (5) στόχοι επιτεύχθηκαν:

1. να παρουσιαστούν οι επιπτώσεις των εγκαταστάσεων ιχθυοκαλλιέργειας στην ποιότητα των παρακειμένων θαλασσίων ιζημάτων,
2. να ανιχνευτούν οι πιο σημαντικοί από τους παράγοντες επιμόλυνσης στο σύνολο των μετάλλων και θρεπτικών συστατικών που προσδιορίζονται,
3. να δοθεί επιστημονική βοήθεια για τη διαμόρφωση μιας στρατηγικής παρακολούθησης και πρακτικών καλλιέργειας που να "ανακουφίζουν" το περιβάλλον,
4. να συγκριθούν διαφορετικές αρχιτεκτονικές ANN με κριτήριο την ικανότητα πρόβλεψης των μοντέλων σε άγνωστα δείγματα και τέλος,
5. να συγκριθούν οι δυνατότητες των ANN με άλλες πολυπαραμετρικές τεχνικές (CT).

6.2. ΜΕΘΟΔΟΛΟΓΙΑ

6.2.1. Σημεία και περίοδοι δειγματοληψίας

Συνολικά προσδιοριστήκαν 12 στοιχεία: Cu, Cd, Pb, Hg, As, Fe, Mn, Zn, Ni, C, N, P σε 3 ιχθυοκαλλιέργειες της χώρας: στη Χίο - Οινούσσες (CH), στη Ναύπακτο (NA), και στον Αστακό (AS). Τα σημεία δειγματοληψίας ποίκιλαν σε αριθμό σε κάθε ιχθυοκαλλιέργεια, αλλά γενικά στην CH, πραγματοποιήθηκαν οι περισσότερες δειγματοληψίες.

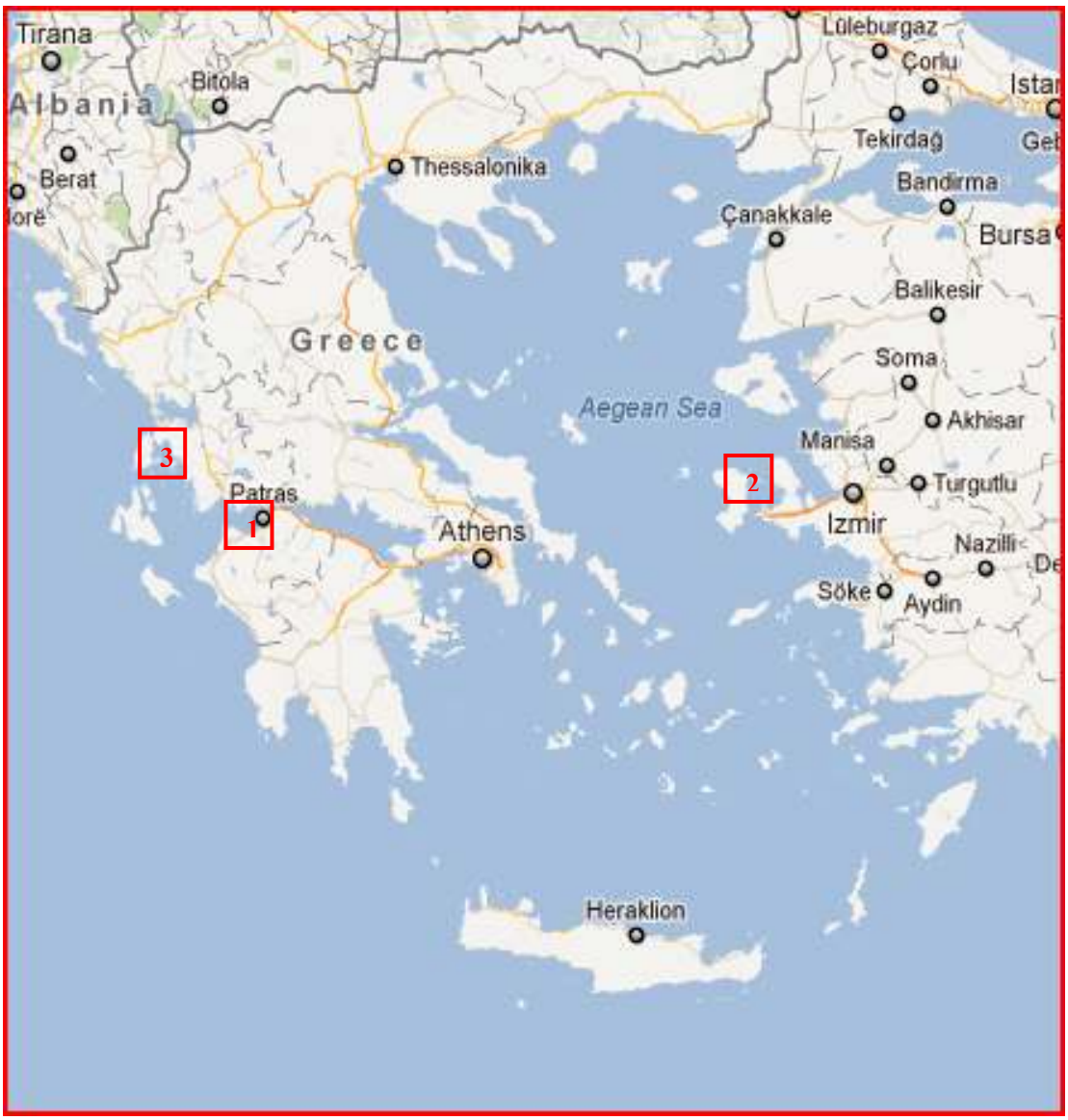
Ειδικότερα έγιναν δειγματοληψίες σε τέσσερις περιόδους: Δεκέμβριος 2005, Ιούνιος 2006, Δεκέμβριος 2006 και Δεκέμβριος 2007 σε ιζήματα (25 – 42 m βάθος) κοντά στον κλωβό (μηδενική απόσταση από αυτόν) και σε αποστάσεις 50 m και 100 m από αυτόν, προς την κατεύθυνση των ρευμάτων. Παράλληλα συλλέχθηκαν "δείγματα-μάρτυρες", δηλαδή δείγματα καθαρά (ή "ελέγχου") στα οποία δεν αναμενόταν υψηλή συγκέντρωση μετάλλων ή θρεπτικών ανόργανων στοιχείων. Συνολικά συλλέχτηκαν 99 δείγματα (25 από τα οποία ήταν ελέγχου). Δύτες συνέλεξαν τα δείγματα σε πλαστικά μπουκάλια 1 L, τα οποία καλύφθηκαν άμεσα με πάγο και αποστάλησαν στο εργαστήριο, όπου φυλάχτηκαν στους -20 ºC μέχρι την ανάλυση.

Οι ιχθυοκαλλιέργειες που επιλέχθηκαν, αποτελούν ένα μικρό ποσοστό των μονάδων που υπάρχουν στη χώρα (σχ. 6.1) αλλά αρκετά αντιπροσωπευτικές, ώστε να μπορούν να εξαχθούν ασφαλή συμπεράσματα για την "δράση" και επίδραση των εγκαταστάσεων ιχθυοκαλλιέργειας στο φυσικό θαλάσσιο περιβάλλον.

6.2.2. Μέθοδοι ανάλυσης

Λεπτομέρειες που αφορούν την προκατεργασία των δειγμάτων, τις μεθόδους ανάλυσης, και τον ποιοτικό έλεγχο, περιγράφονται στο παράρτημα της διατριβής (ΚΕΦ. 2, Π). Στο ίδιο κεφάλαιο έγιναν οι πρώτες περιβαλλοντικές μελέτες όπως περιγράφηκε παραπάνω (§ 6.1).

Από το σύνολο των δειγμάτων (99), χρησιμοποιήθηκαν τα 74 που φορούσαν δειγματοληψία κοντά ή μακριά από τους κλωβούς (σε απόσταση μέχρι 100 m), ενώ αγνοήθηκαν τα 25 δείγματα ελέγχου.



Σχήμα 6.1: Σημεία δειγματοληψίας για τον προσδιορισμό 12 στοιχείων σε δείγματα ιζημάτων σε τρεις εγκαταστάσεις ιχθυοκαλλιέργειας της Ελλάδας: (1: ΝΑ (Ναύπακτος), 2: CH (Χίος - Οινούσσες), 3: AS (Αστακός)).

6.2.3. Κ – κοντινότεροι γείτονες (KNN)

Η μέθοδος KNN χρησιμοποιήθηκε στην ανάλυση αυτή λόγω της απλότητας και της αποτελεσματικότητας της. Πραγματικά αρκετές φορές στη βιβλιογραφία [5] έχει αποδειχτεί ότι παρόλο που περιγράφεται από ένα απλό αλγόριθμο (§ 4.3.12), δίνει εφάμιλλα αποτελέσματα με πολύπλοκες τεχνικές.

Η τεχνική εφαρμόστηκε σε μια αρχική ομάδα 74 συνολικά δειγμάτων από τα οποία, ποσοστό ίσο με το 80 % (= 59) τυχαία επιλεγμένα, χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου, ενώ τα υπόλοιπα (15) για έναν ανεξάρτητο έλεγχο αυτού. Επιπλέον, για την επικύρωση του μοντέλου, εφαρμόστηκε η v-fold CV μέθοδος (§ 3.1.2, 4.4.2). Δημιουργήθηκαν v = 10

ομάδες, από τις οποίες οι 9 χρησιμοποιήθηκαν ως ομάδα εκπαίδευσης και η μία ως ομάδα ελέγχου.

6.2.4. Διαχωριστική Ανάλυση (DA)

Η Διαχωριστική Ανάλυση (DA) εφαρμόστηκε στο σύνολο των δειγμάτων (74), ενώ ένα μέρος αυτών (14) τυχαία επιλεγμένα, επιλέχθηκαν για την επικύρωση των μοντέλων (CV "εξαιρουμένου πολλαπλών" μέθοδος, § 4.4.2). Η DA εξήγε συναρτήσεις ταξινόμησης για κάθε ομάδα δειγματοληψίας (DI ή ZE), χρησιμοποιώντας τις τιμές των προσδιοριζόμενων στοιχείων στα 74 αρχικά δείγματα. Παράλληλα, αξιολογήθηκε το σύνολο των μεταβλητών και βρέθηκαν οι πιο κρίσιμες από αυτές, που είναι υπεύθυνες για το διαχωρισμό των σημείων αυτών.

Στη συνέχεια η κάθε τεχνική DA (κλασική, FW, BW) αξιολογήθηκε ελέγχοντας την επιτυχία του μοντέλου για τα δείγματα της ομάδας CV. Εδώ πρέπει να επισημανθεί ότι η εφαρμογή της DA που έγινε στο ΚΕΦ. 2, Π του παραρτήματος () και αφορούσε τα ίδια δείγματα, έγινε χωρίς επικύρωση του μοντέλου και μόνο με την εφαρμογή της κλασικής προσέγγισης, καθώς διέφερε ο σκοπός της όλης ανάλυσης.

6.2.5. Δέντρα Ταξινόμησης (CT)

Εφαρμόστηκαν και οι τρεις μέθοδοι των Δέντρων Ταξινόμησης: Discriminant-based linear combination method (LCM), Discriminant-based univariate method (Classic CT) και CART-style Exhaustive search method for univariate splits (CART). Η τεχνική εφαρμόστηκε σε μια ομάδα 60 δειγμάτων, ενώ 14 τυχαία επιλεγμένα δείγματα, επιλέχθηκαν για τον ανεξάρτητο έλεγχο των μοντέλων.

6.2.6. Νευρωνικά Δίκτυα (ANN)

Το πρώτο βήμα στην εφαρμογή των ANN στα δεδομένα του κεφαλαίου αυτού, ήταν η αξιοποίηση βηματικών προσεγγίσεων των ANN (FW και BW) αλλά και ο GA (§ 4.4.3), για την επιλογή των πιο κρίσιμων μεταβλητών.

Στη συνέχεια, δοκιμάστηκαν και βελτιστοποιήθηκαν πολλές αρχιτεκτονικές (γραμμικά δίκτυα, MLP, RBF). Η αρχική ομάδα δειγμάτων (74) διαχωρίστηκε τυχαία σε ομάδα εκπαίδευσης (44 δείγματα), επικύρωσης (15) και ελέγχου (15). Τα κριτήρια που χρησιμοποιήθηκαν ήταν οι ικανότητες αναγνώρισης και πρόβλεψης (§ 4.2.2), το σφάλμα (RMS) στην ομάδα επικύ-

ρωσης, αλλά και η περιοχή AUC κάτω από την ROC καμπύλη (βλ. παρακάτω, § 6.2.7). Τελικά, έγινε σύγκριση μεταξύ των μοντέλων ANN, αλλά και των άλλων τεχνικών.

6.2.7. ROC (Receiving Operating Characteristic) καμπύλες

Οι καμπύλες **ROC** (Receiving Operating Characteristic) αποτελούν ένα **μη-παραμετρικό δείκτη σύγκρισης μοντέλων** [87] και διευκολύνουν στην επιλογή των καλύτερων από αυτά [247] και στην απόρριψη των λιγότερων καλών, παρέχοντας άμεσα τα αποτελέσματα της πρόβλεψης σε μια δυαδική ταξινόμηση (§ 4.2.2). Οι καμπύλες ROC πρωτοεμφανίστηκαν κατά τη διάρκεια του Β΄ Παγκοσμίου Πολέμου για την ανάλυση των σημάτων radar που χρησιμοποιούνταν στην ανίχνευση των εχθρικών αεροπλάνων [248]. Εδώ αντικατοπτρίζεται ίσως, η χρήση του πρώτου συνθετικού (receiving για τη λήψη σημάτων), στο ακρώνυμο της λέξης ROC. Οι καμπύλες εξάλλου, είναι επίσης γνωστές ως Relative Operating Characteristic curves, επειδή συγκρίνουν βασικά λειτουργικά χαρακτηριστικά (TP vs FN, βλ. παρακάτω).

Οι ROC καμπύλες συνοψίζουν την απόδοση ενός δικτύου ταξινόμησης δυο ομάδων, κατά την αλλαγή στις τιμές των κατωφλιών ταξινόμησης (classification thresholds, § 4.2.2). Ο άξονας Y απεικονίζει την ευαισθησία (sensitivity, § 4.4.2, σχέση 4.21) του δικτύου, δηλαδή την αναλογία **των αληθώς θετικά ταξινομημένων σημείων** (true positive, TP) της δεύτερης ομάδας (βλ. παρακάτω, σχ. 6.4). Ο άξονας X απεικονίζει συνολικά τη σχέση: 1-εξειδίκευση (1-specificity, § 4.4.2, σχέση 4.22), δηλαδή την αναλογία των ψευδώς αρνητικά ταξινομημένων (false negative, FN) της πρώτης ομάδας ή **των ψευδώς θετικά ταξινομημένων σημείων** (false positive, FP) της δεύτερης ομάδας [17, 58, 249].

Με αλλαγή των κατωφλιών ταξινόμησης, κατασκευάζονται δίκτυα με εναλλαγές στην αναλογία των ψευδώς θετικά ταξινομημένων και των ψευδώς αρνητικά ταξινομημένων. Με ένα κατώφλι ταξινόμησης ίσο με το 0, όλα τα σημεία θα κατατάσσονταν στη δεύτερη ομάδα και έτσι θα αποδίδονταν τα περισσότερα εσφαλμένα θετικά (για τη δεύτερη ομάδα) και καθόλου εσφαλμένα αρνητικά (για την πρώτη ομάδα). Το αντίθετο θα συνέβαινε με κατώφλι ταξινόμησης ίσο με το 1. Καθώς το κατώφλι ταξινόμησης αυξάνεται, η καμπύλη "μετακινείται" από αριστερά προς τα δεξιά.

Όσο μεγαλύτερη είναι η περιοχή κάτω από την καμπύλη (AUC, Area Under Curve), τόσο μεγαλύτερη είναι η ικανότητα ταξινόμησης του μοντέλου που αντιπροσωπεύει [184], καθώς η AUC ερμηνεύεται ως την πιθανότητα της σωστής ταξινόμησης [81]. Μια καμπύλη ROC ιδανικού δικτύου προσεγγίζει την πάνω αριστερή καμπύλη του διαγράμματος και η AUC που καλύπτει είναι 1,0. Το σημείο (0,1) (σχ. 6.4) ονομάζεται **το τέλειο σημείο ταξινόμησης**

(perfect classification, [248]), και αντιπροσωπεύει 100 % ευαισθησία (μηδενικά ψευδώς αρνητικά) και 100 % εξειδίκευση (μηδενικά ψευδώς θετικά). Αυτό σημαίνει ότι το μοντέλο θα έχει 100 % επιτυχία στην εύρεση της σωστής ομάδας [81]. Κάποιες μαθηματικές σχέσεις που περιγράφουν τα ποιοτικά χαρακτηριστικά ενός μοντέλου με βάση τους παραπάνω ορισμούς, περιγράφονται στο παράρτημα αυτής της διατριβής (ΚΕΦ. 3, Π).

Η ιδανική καμπύλη ROC επιτυγχάνεται μόνο για τις πολύ καλά διαφοροποιημένες ομάδες, μερικές φορές ωστόσο, απλά θεωρείται σημαντικότερο να αποφθεχθούν λάθη στη μια ομάδα, από όσο στη δεύτερη. Αυτό συμβαίνει για παράδειγμα, σε ιατρικές εφαρμογές.

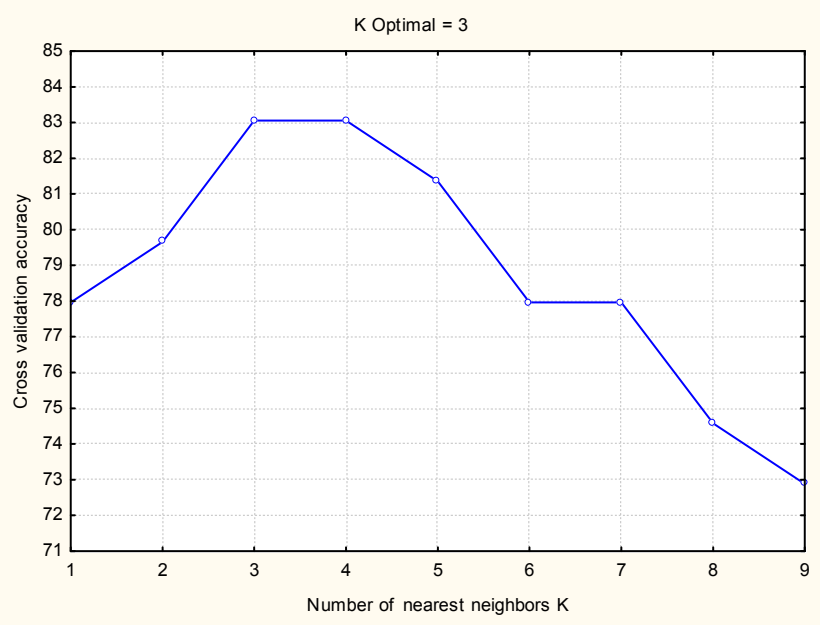
Ένα μοντέλο με AUC 0,7 για την καμπύλη ROC θεωρείται ότι επιτυγχάνει ικανοποιητικό διαχωρισμό των ομάδων, με AUC 0,8 καλό διαχωρισμό και με AUC 0,9 πολύ καλό [94]. Ένα πείραμα τυχαίας πρόβλεψης (random classifier [17]) δίνει AUC περίπου 0,5. Ένα τέτοιο σημείο (random guess [248], που θα λαμβάναμε για παράδειγμα με το γνωστό παιγνίδι του κορώνα-γράμματα σε ένα νόμισμα), θα βρισκόταν πάνω στη διαγώνια κόκκινη γραμμή (line of no-discrimination, σχ. 6.4) που ενώνει την κάτω αριστερή γωνία του διαγράμματος με την πάνω δεξιά. Η διαγώνια αυτή χωρίζει το χώρο της καμπύλης ROC, σε περιοχές καλής και κακής ταξινόμησης. Σημεία πάνω από τη γραμμή δηλώνουν ορθά αποτελέσματα, ενώ σημεία κάτω από αυτή, εσφαλμένα. Εδώ επισημαίνεται, ότι με την "αντιστροφή" της μεθόδου πρόβλεψης (δηλαδή αντιστροφή των αποφάσεων της μεθόδου, όλα τα εσφαλμένα αποτελέσματα μετατρέπονται σε ορθά).

6.3. ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΣΥΖΗΤΗΣΗ

6.3.1. Εφαρμογή της μεθόδου Κ – κοντινότεροι γείτονες (ΚΝΝ)

Σαν απόσταση μεταξύ των δειγμάτων χρησιμοποιήθηκε το τετράγωνο της Ευκλείδειας απόστασης (Euclidean squared), ενώ για την εύρεση του βέλτιστου αριθμού Κ των ομάδων, επιλέχθηκε το κριτήριο του σφάλματος στην CV ομάδα δειγμάτων. Χρησιμοποιήθηκαν όλες οι μεταβλητές. Για τον αριθμό Κ επιλέχθηκε εύρος τιμών από 1 ως 9, ώστε να καλύπτεται η τιμή $\sqrt{N}$, όπου N το σύνολο των δειγμάτων εκπαίδευσης [138] (§ 4.3.12).

Το σχήμα 6.2 απεικονίζει την διαδικασία εύρεσης του βέλτιστου Κ, ενώ στον πίνακα 6.1 φαίνονται τα αποτελέσματα επικύρωσης για την ομάδα ελέγχου. Για την CV μέθοδο επικύρωσης, το ποσοστό επιτυχίας ήταν 83,0 %.



Σχήμα 6.2: Διαδικασία εύρεσης βέλτιστου K (=3) με κριτήριο το ελάχιστο σφάλμα ή τη μέγιστη ακρίβεια στην CV ομάδα.

Πίνακας 6.1: Αποτελέσματα από την επικύρωση του μοντέλου KNN με βάση την ομάδα ελέγχου.

Ομάδα	**Συνολικός αρ. δειγμάτων**	**Αρ. προβλεπόμενων δειγμάτων**		**% Ποσοστά επιτυχίας**	
		Σωστά	**Λάθος**	**Σωστά**	**Λάθος**
DI	11	9	2	81,8	18,2
ZE	4	4	0	100,0	0
Συνολικά	15	13	2	86,7	13,3

6.3.2. Εφαρμογή της Διαχωριστικής Ανάλυσης (DA)

Τα αποτελέσματα της κάθε τεχνικής (κλασική, FW, BW) για το σύνολο των δειγμάτων (τελικό αποδεχθέν μοντέλο), αλλά και την CV ομάδα φαίνεται στον πίνακα 6.2. Επιπλέον απεικονίζονται οι χρησιμοποιηθείσες μεταβλητές.

Πίνακας 6.2: Αποτελέσματα για την αρχική και CV ομάδες δειγμάτων (DA μέθοδος)

	Κλασική (12*)		**FW (5) / BW (5)**		**FW (1) / BW (1)**	
Ομάδες δειγμάτων	Σύνολο δειγμάτων	CV	Σύνολο δειγμάτων	CV	Σύνολο δειγμάτων	CV
% Ποσοστά επιτυχίας	86,5	92,8	83,8	85,7	82,4	85,7
Μεταβλητές	Όλες		P, N, Pb, Hg, Cu	Cu, Cd, N, Hg, P	P	

* Στην παρένθεση, ο αριθμός των χρησιμοποιηθέντων μεταβλητών.

Οι τεχνικές FW (5) και BW (5) αντιστοιχούν σε επιλεγμένες τιμές F εισόδου/αποκλεισμού (§ 2.1.1), έτσι ώστε να συμπεριληφθούν περισσότερες μεταβλητές στα αντίστοιχα μοντέλα. Δίνεται λοιπόν η ευελιξία στα μοντέλα να χρησιμοποιήσουν περισσότερες ή λιγότερες μεταβλητές. Παρόλα αυτά, είναι φανερό ότι για τη συγκεκριμένη βάση, μικρές βελτιώσεις επιτυγχάνονται στα ποσοστά επιτυχίας με την αύξηση των συμμετεχουσών μεταβλητών. Μία μόνο μεταβλητή (ο P) είναι ικανή για μια επιτυχή ταξινόμηση των δειγμάτων. Τα ποσοστά επιτυχίας είναι ανάλογα της μεθόδου KNN (πίνακας 6.1).

Συμπερασματικά, μπορούμε να πούμε ότι ο P μπορεί να διαχωρίσει τα δείγματα των ιχθυοκαλλιεργειών με επιτυχία, χωρίς οι υπόλοιπες μεταβλητές να συνεισφέρουν δραστικά. Φαίνεται λοιπόν ότι μια μονάδα ιχθυοκαλλιέργειας, επιδρά δραστικά και αδιαμφισβήτητα στο επίπεδο P του θαλάσσιου ιζήματος που εγκαθίσταται [216]. Εξάλλου, ο P έχει πολλάκις χρησιμοποιηθεί στη βιβλιογραφία ως δείκτης της επίδρασης των ιχθυοκαλλιεργειών στην ποιότητα των θαλασσίων ιζημάτων [217, 250, 251].

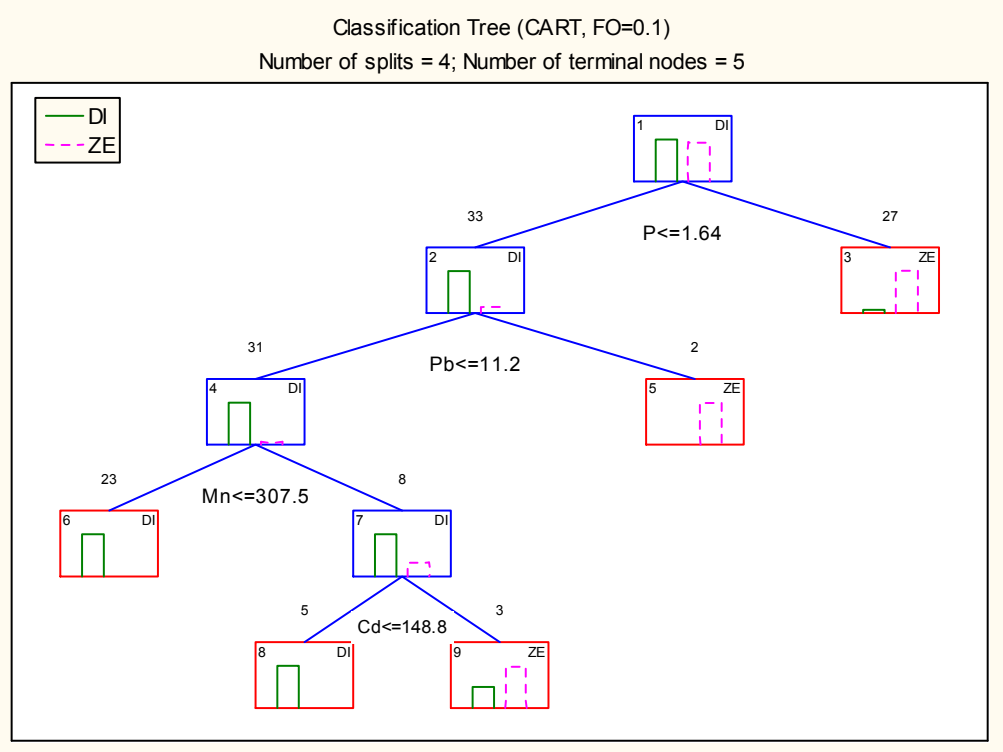
6.3.3. Εφαρμογή των Δέντρων Ταξινόμησης (CT)

Τα αποτελέσματα της κάθε μεθόδου για τις ομάδες εκπαίδευσης (60 δείγματα) και ελέγχου (14) φαίνονται στον πίνακα 6.3. Επιπλέον απεικονίζονται οι χρησιμοποιηθείσες μεταβλητές και η σειρά κρισιμότητας αυτών. Για τις μεθόδους LCM και CART, εφαρμόστηκαν pre-pruning, με καθορισμό της παραμέτρου FO = 0,1 (§ 5.3.4), ενώ για την Classic CT, η "ελαχίστου κόστους-πολυπλοκότητας διασταυρούμενη επικύρωση" για την εύρεση του καλύτερου δέντρου (§ 5.3.6). Η καλύτερη μέθοδος, με ένα μικρό και ακριβές δέντρο είναι η CART (με FO = 0,1). Συγκριτικά σε παρενθέσεις δίνονται τα αποτελέσματα για τις μεθόδους LCM και CART με FO = 0,05. Η σύγκριση αποδεικνύει την ευελιξία των μεθόδων CT και την ανταγωνιστικότητα ανάμεσα στο κόστος (ακρίβεια της μεθόδου) και το μέγεθος του δέντρου

(μεγαλύτερο δέντρο υποδηλώνει εντονότερα φαινόμενα υπερ-προσαρμογής). Το καλύτερο δέντρο απεικονίζεται στο σχήμα 6.3.

Πίνακας 6.3: Αποτελέσματα για τις ομάδες εκπαίδευσης και ελέγχου/χαρακτηριστικά των CT μοντέλων

Μέθοδοι		**Classic CT**	**LCM**	**CART**
Τερματικοί κόμβοι		**2**	**3 (8)**	**5 (9)**
% Ποσοστά επιτυχίας	Ομάδα εκπαίδευσης	86,7	93,3 (100,0)	95,0 (100,0)
	Ομάδα ελέγχου	85,7	85,7 (85,7)	92,8 (85,7)
Μεταβλητές	Αριθμός	1	12 (12)	4 (6)
	Σειρά κρισιμότητας	**P**>N>C>Zn	---	Zn>**Cd**>**P**>**Mn**>Cu>**Pb** (Zn>**Cd**>**P**>**N**>**C**=**Mn**>**Pb**)



Σχήμα 6.3:Βέλτιστο δέντρο ταξινόμησης για τα δείγματα των ιχθυοκαλλιεργειών. Μέθοδος: CART

Το δέντρο περιλαμβάνει 4 διαχωρισμούς με βάση τις μεταβλητές P, Pb, Mn και Cd και 5 τερματικούς κόμβους. Ο πρώτος τερματικός κόμβος (αριθμός 3 στο σχήμα), ονομάζεται ZE (ZERO) γιατί περιέχει 25 δείγματα ZE και 2 δείγματα DI. Ο δεύτερος (αριθμός 5 στο σχήμα),

χαρακτηρίζεται επίσης ως ΖΕ, και περιέχει μόνο 2 δείγματα ΖΕ. Ο τρίτος (αριθμός 6 στο σχήμα), περιέχει 23 δείγματα DI. Οι τέταρτος και πέμπτος τερματικός κόμβος (αριθμοί 8 και 9 αντίστοιχα στο σχήμα) περιέχουν 5 δείγματα DI και 1 DI και 2 ΖΕ. Η κατασκευή (tree structure) του δέντρου που σχηματίζεται, απεικονίζεται στο παράρτημα ( ΚΕΦ. 3, Π).

6.3.4. Εφαρμογή των Νευρωνικών Δικτύων (ANN)

Η αρχική εφαρμογή των βηματικών προσεγγίσεων (FW, BW) αλλά κυρίως του GA, έδειξε τις σημαντικότερες μεταβλητές για τον διαχωρισμό των DI και ΖΕ σημείων για τις τρεις μονάδες ιχθυοκαλλιέργειας: N>Zn>P>C>Cu>Cd. Με σκοπό λοιπόν να αποφθεχθούν φαινόμενα υπερ-προσαρμογής (πολλά βάρη, μεταβλητές και ενδιάμεσες μονάδες στα δίκτυα), οι πρώτες δοκιμές έγιναν με τη χρήση των έξη αυτών κρισιμότερων μεταβλητών.

Η βελτιστοποίηση αφορούσε γενικά:

- ✓ τον αριθμό των μονάδων της ενδιάμεσης στιβάδας και
- ✓ τον αριθμό των εισερχόμενων μεταβλητών.

Κατά την διάρκεια των πρώτων δοκιμών, η διαίρεση των δειγμάτων στις τρεις βασικές τους ομάδες (εκπαίδευσης, επικύρωσης και ελέγχου), γινόταν εκ νέου σε κάθε προσπάθεια κατασκευής ενός νέου μοντέλου. Ο αριθμός των δειγμάτων κάθε ομάδας παρέμενε σταθερός, όπως αρχικά είχε καθοριστεί (§ 6.2.6). Με τον τρόπο αυτό, δοκιμάστηκαν πολλοί διαφορετικοί συνδυασμοί δειγμάτων σε κάθε αρχιτεκτονική, ώστε να μην ευνοηθούν ή αδικηθούν μοντέλα εξαιτίας μιας τυχαίας επιτυχούς ή ατυχούς επιλογής ομάδων δειγμάτων αντίστοιχα.

Στη συνέχεια, και μετά την επιλογή των βέλτιστων αρχιτεκτονικών, η σύσταση των ομάδων παρέμεινε σταθερή και νέες προσπάθειες έγιναν για την κατασκευή ενός αξιόλογου μοντέλου.

Τα καλύτερα αποτελέσματα περιλαμβάνουν τρεις αρχιτεκτονικές:

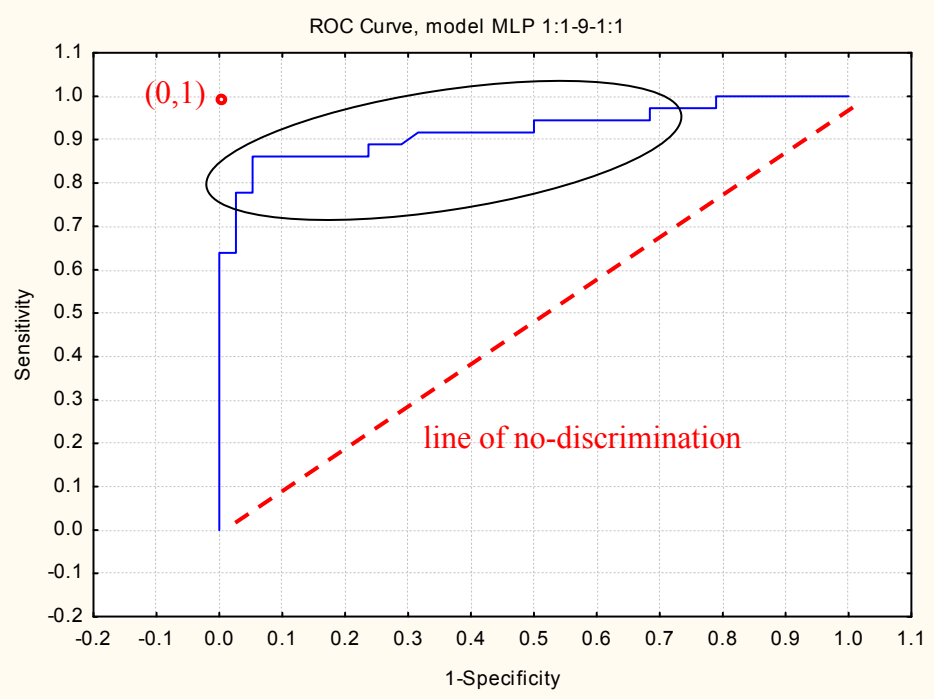
- ➢ Γραμμικό δίκτυο με μία μεταβλητή (Linear 1:1-1:1),
- ➢ δίκτυα MLP μίας ενδιάμεσης στιβάδας με μια ή τρεις μεταβλητές και 9 ή 7 ενδιάμεσες μονάδες αντίστοιχα (MLP 1:1-9-1:1, MLP 3:3-7-1:1,),
- ➢ δίκτυο RBF με μια ή δύο μεταβλητές και πέντε ή τέσσερις ενδιάμεσες μονάδες αντίστοιχα (RBF 1:1-5-1:1, RBF 2:2-4-1:1).

Τα χαρακτηριστικά και αποτελέσματα των βέλτιστων δικτύων φαίνονται στον πίνακα 6.4 που ακολουθεί.

Πίνακας 6.4: Χαρακτηριστικά και αποτελέσματα των βέλτιστων ANN μοντέλων

Μοντέλο	**Χρησιμοποιηθείσες μεταβλητές**	**RMS**	**AUC**	**Ικανότητα (απόδοση) μοντέλου**				
				Ανά ομάδα δειγμάτων			**Ανά θέση**	
				Εκπαίδευσης	**Επικύρωσης**	**Ελέγχου**	**DI**	**ZE**
Linear 1:1-1:1	P	0,38	0,922	90,9	80,0	93,3	92,1	86,1
MLP 1:1-9-1:1	P	0,50	0,922	86,4	86,7	86,7	86,8	86,1
MLP 3:3-7-1:1	P>N>Cu	0,22	0,948	86,4	100,0	80,0	89,5	86,1
RBF 1:1-5-1:1	P	0,35	0,923	90,9	86,7	86,7	94,7	83,3
RBF 2:2-4-1:1	P>Cd	0,36	0,930	93,2	86,7	86,7	94,7	86,1

Η καμπύλη ROC (για το MLP 1:1-9-1:1) φαίνεται στο σχήμα 6.4.

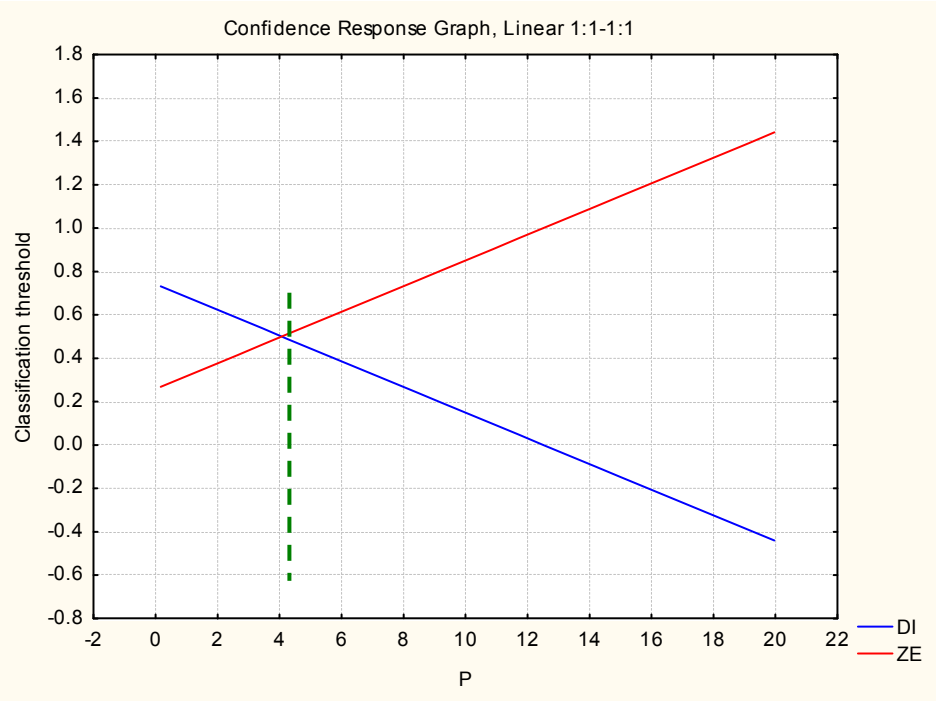


Σχήμα 6.4: Καμπύλη ROC για το τελικό μοντέλο MLP (1:1-9-1:1). Δίνεται η AUC της καμπύλης = 0,92.

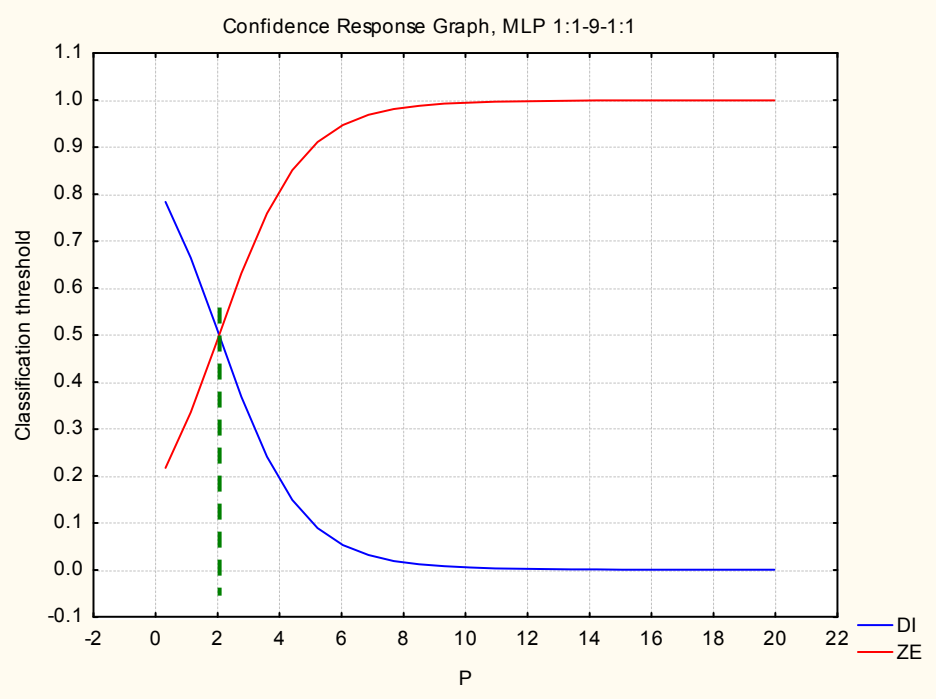
Το συγκεκριμένο βελτιστοποιημένο μοντέλο δίνει μια περιοχή 0,92 (πολύ κοντά στο 1) ενώ φαίνεται να διατηρεί σε πολύ καλές τιμές ευαισθησία και εξειδίκευση (εντός του μαύρου κύκλου) για τις διάφορες τιμές του κατωφλίου ταξινόμησης (τα δίκτυα με πολύ υψηλές τιμές ευαισθησίας και χαμηλές εξειδίκευσης και αντιστρόφως συνήθως δεν μας ενδιαφέρουν και σίγουρα όχι για τα συγκεκριμένα δεδομένα).

Η επιλογή της μεταβλητής P (σε συνδυασμό ή όχι με άλλες) αποτελεί μονόδρομο για τα δίκτυα ANN: δίνει σταθερά τα καλύτερα αποτελέσματα.

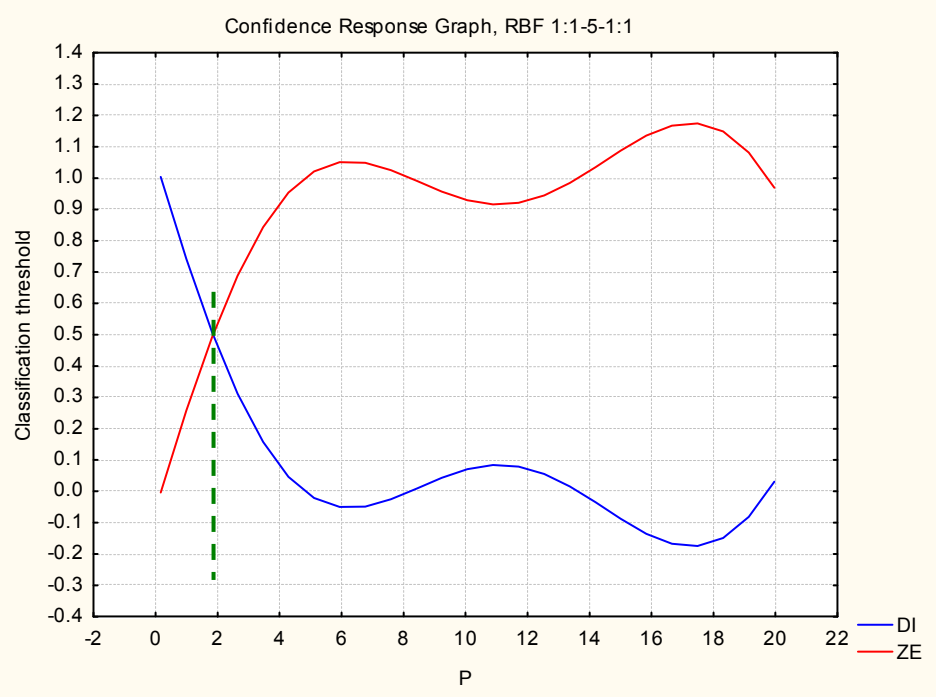
Ενδιαφέρον παρουσιάζουν τα διαγράμματα απόκρισης (confidence response graph, σχ. 6.5) για τα μονοπαραμετρικά βελτιστοποιημένα μοντέλα και τη μοναδική μεταβλητή που χρησιμοποιείται. Έτσι, για τιμές P μεγαλύτερες του 4,3 / 2,0 και 1,8 mg/g αντίστοιχα για τα τρία μοντέλα (Linear, MLP, RBF), το δείγμα ταξινομείται ως ZE, ενώ για τιμές μικρότερες του ορίου αυτού, ως DI. Η επίδραση της εγκατάστασης του ιχθυοτροφείου στο θαλάσσιο πυθμένα μέσω των εξωτερικά προστιθέμενων τροφών (πλούσιων σε ανόργανα συστατικά) ή/και των περιττωμάτων των ψαριών, είναι για άλλη μια φορά εμφανής.



(α)



(β)



(γ)

Σχήμα 6.5: Διαγράμματα απόκρισης των μονοπαραμετρικών βελτιστοποιημένων μοντέλων (α) Liner 1:1-1:1, (β)MLP 1:1-9-1:1, (γ) RBF 1:1-5-1:1.

6.4. ΣΥΓΚΡΙΣΗ ΤΕΧΝΙΚΩΝ - ΣΥΜΠΕΡΑΣΜΑΤΑ

6.4.1. Αξιολόγηση μεταβλητών

Ενδιαφέρον παρουσιάζει η σύγκριση όλων των παραπάνω τεχνικών KNN, DA, CT και ANN. Όσον αφορά **την αξιολόγηση των μεταβλητών:**

- ✓ η κύρια μεταβλητή που "δεσπόζει" στους διαχωρισμούς για τις μεθόδους που έχουν τη δυνατότητα εκτίμησης των μεταβλητών, είναι ο P,
- ✓ επόμενες κρισιμότερες μεταβλητές αναδεικνύονται τα N, Cu, Cd, C, Zn και λιγότερο τα Mn και Pb.

Το γεγονός ότι μοντέλα (DA, CT, ANN) με τη χρήση **μίας μόνο μεταβλητής** (single-element models) μπορούν να επιτύχουν τόσο καλές προβλέψεις είναι ιδιαίτερα ελκυστικό. Αυτόματα σημαίνει μειωμένο κόστος, χρόνος ανάλυσης και λιγότερη εργαστηριακή εργασία [129].

Έτσι, γενικότερο συμπέρασμα αποτελεί η επιβάρυνση του θαλάσσιου πυθμένα με ανόργανα θρεπτικά συστατικά που περιέχονται στις τεχνητές τροφές με τις οποίες εκτρέφονται τα ψάρια. Το ίδιο επισημαίνεται για τα μέταλλα Cd (ιδιαίτερα τοξικό) και Zn. Τεχνητά παρασκευασμένες ιχθυοτροφές/ιχθυάλευρα και συμπληρώματα φαίνεται να επιβαρύνουν το θαλάσσιο πυθμένα, με ανόργανα και μεταλλικά στοιχεία [245] που ανακυκλώνονται και βιοσυσσωρεύονται μεταφέροντας το πρόβλημα.

6.4.2. Ομαδοποίηση και ταξινόμηση θέσεων

Σε σχέση με την αποτελεσματικότητα των μοντέλων για τις ομάδες εκπαίδευσης και ελέγχου, η σύγκριση γίνεται μέσω του πίνακα 6.5.

Τα ANN με μικρότερη ομάδα εκπαίδευσης, έδωσαν εφάμιλλα των άλλων μεθόδων αποτελέσματα. Παρότι η μέθοδος είναι "ακριβή" (§ 4.4.2, χρειάζεται πολλά δείγματα για την κατασκευή και αξιολόγηση των μοντέλων), τα αποτελέσματα είναι πολύ αισιόδοξα. Εντυπωσιακό είναι το γραμμικό ANN δίκτυο, που με μία μόνο μεταβλητή και ένα απλό αλγόριθμο δίνει τα υψηλότερα ποσοστά.

Όλες οι τεχνικές έχουν δυνατότητα ευελιξίας και ανάλογα με τον σκοπό της μελέτης, τις διαθέσιμες μεταβλητές και τα δείγματα, μπορούν να επιτύχουν αποτελεσματικά, ανθεκτικά μοντέλα με δυνατότητες γενίκευσης.

Πίνακας 6.5: Χαρακτηριστικά και αποτελέσματα των βέλτιστων μοντέλων

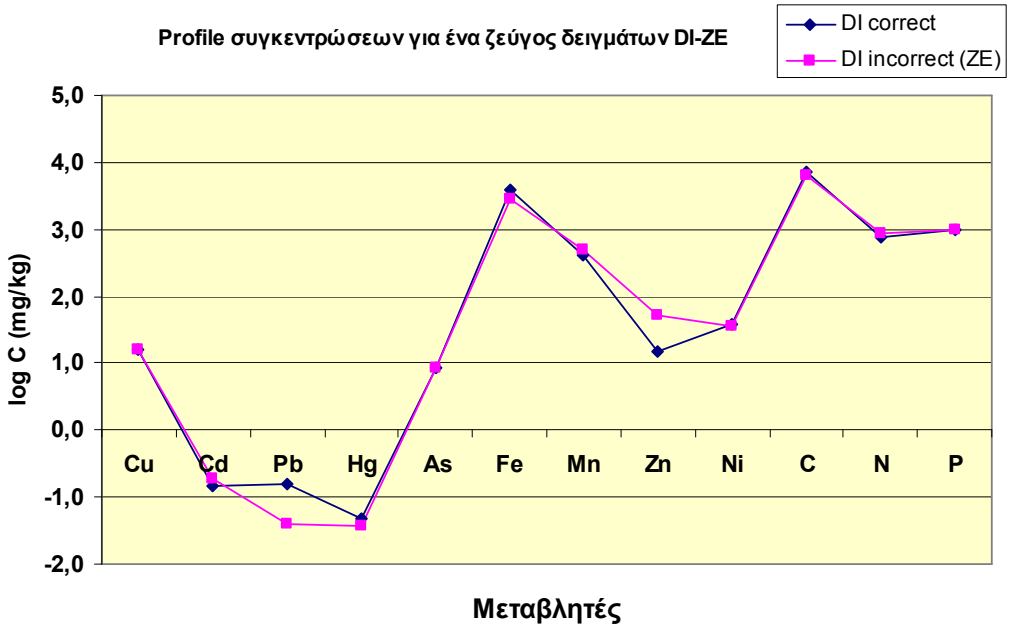
Μοντέλο	**Αρ. δειγμάτων στην ομάδα εκπαίδευσης**	**Αρ. μεταβλητών**	**% Ποσοστά επιτυχίας**	
			Ομάδα εκπαίδευσης	**Ομάδα ελέγχου**
KNN	59	12	---	86,7
DA κλασική	74	12	86,5	92,8
DA FW / BW *	74	5	83,8	85,7
DA FW / BW *	74	1	82,4	85,7
Classic CT	60	1	86,7	85,7
LCM	60	12	93,3	85,7
CART	60	4	95,0	92,8
ANN Linear	44	1	90,9	93,3
ANN MLP *	44	1	86,4	86,7
ANN MLP *	44	3	86,4	80,0
ANN RBF *	44	1	90,9	86,7
ANN RBF *	44	2	93,2	86,7

* Οι διαφορές των μοντέλων αυτών, αφορά τον αριθμό των μεταβλητών.

Στο σημείο αυτό, θα πρέπει να αναδειχθεί επίσης μια αδυναμία της συγκεκριμένης βάσης που επηρέασε την ικανότητα των μοντέλων. Από τους πίνακες των αρχικών αποτελεσμάτων ( ΚΕΦ. 2, Π), είναι φανερό ότι υπάρχουν δείγματα που έχουν ίδιες τιμές για πολλές μεταβλητές (μέταλλα/θρεπτικά συστατικά), ενώ ανήκουν σε διαφορετικές ομάδες (DI ή ZE). Θα ήταν δηλαδή παράλογο να "αξιώσουμε" από το νέο μοντέλο να προβλέψει διαφορετική ομάδα για την ίδια εισερχόμενη τιμή. Ο Alvarez-Guerra et al. [249] αναφέρει το παραπάνω πρόβλημα στην εργασία του και ερευνά για το **"δείγμα της αντίθετης ομάδας"** (opposite class). Ως τέτοιο ορίζει το δείγμα εκείνο με τις πλησιέστερες συγκεντρώσεις των εξεταζόμενων μεταβλητών, που όμως ανήκει στην αντίθετη ομάδα ταξινόμησης από κάποιο άλλο. Το ίδιο πρόβλημα αναφέρεται και αλλού [234].

Το αντίστοιχο διάγραμμα για τα δείγματα DI και ZE της Ναυπάκτου (NA) με 0,99 mg/g P, φαίνεται στο σχήμα 6.6. Είναι φανερή η σύμπτωση των σημείων. Προφανώς κανένα μαθη-

ματικό μοντέλο δεν μπορεί να προβλέψει σωστά δείγματα διαφορετικών ομάδων που έχουν όμοιες μεταβλητές [249].



Σχήμα 6.6: Profile συγκεντρώσεων για ένα ζεύγος δειγμάτων "opposite class", δηλαδή δειγμάτων που ανήκουν σε διαφορετικές ομάδες, αλλά ταξινομούνται στην ίδια (DI).

ΚΕΦ. 7 ΤΑΞΙΝΟΜΗΣΗ ΕΛΑΙΟΛΑΔΩΝ ΜΕ ΒΑΣΗ ΤΗ ΓΕΩΓΡΑΦΙΚΗ ΤΟΥΣ ΠΡΟΕΛΕΥΣΗ

7.1. ΕΙΣΑΓΩΓΗ

Σήμερα το ελαιόλαδο θεωρείται ένα προϊόν διατροφής που συνδυάζει θρεπτικές και αισθητήριες αξίες. Τα κέρδη που αποκομίζει κανείς με την κατανάλωση ελαιολάδου εξαιτίας των μόνο- και πολύ- ακόρεστων οξέων που περιέχει, είναι αδιαμφισβήτητα [252, 253]. Ως αποτέλεσμα, η κατανάλωσή του έχει αυξηθεί την τελευταία δεκαετία, καθώς το ελαιόλαδο θεωρείται πλέον από πολλούς καταναλωτές σαν ένα καθημερινό λάδι μαγειρικής και όχι ένα gourmet υλικό για επιλεγμένα μόνο γεύματα [252].

Επιπλέον, πολλά ελαιόλαδα είναι σήμερα πιστοποιημένα ως ΠΟΠ (Προστατευμένη Ονομασία Προέλευσης) ή ΠΓΕ (Προστατευμένη Γεωγραφική Ένδειξη), το οποίο αυτόματα σημαίνει υψηλότερες τιμές για τον καταναλωτή που το επιλέγει αλλά και ευκαιρίες νοθείας. Έτσι, συχνά εγείρονται προβλήματα αυθεντικότητας που έχουν σχέση με τη γεωγραφική προέλευση ή την ποικιλία των ελαιολάδων, τα οποία διέπονται σήμερα από ένα αυστηρό νομοθετικό πλαίσιο που περιγράφει ένα πλήθος παραμέτρων που μπορούν να εξασφαλίσουν την αυθεντικότητά τους.

Μια σειρά αναλυτικών μεθόδων (κυρίως χρωματογραφικές και φωτομετρικές) χρησιμοποιούνται για αυτόν το σκοπό, ενώ παράλληλα διαφορετικές χημειομετρικές μέθοδοι επιστρατεύονται για τη διαφοροποίηση δειγμάτων ελαιολάδου.

Τα λιπαρά οξέα και οι στερόλες είναι οι κυρίαρχες χημικές ενώσεις που προσδιορίζονται [74, 103, 105, 149, 254, 255, 256, 257]. Άλλες επίσης μέθοδοι όπως NIR και NMR ή MS φασματοσκοπία, φθορισμομετρία έχουν ευρέως χρησιμοποιηθεί, ενώ έχουν προσδιοριστεί παράμετροι όπως δείκτες UV και πτητικές ουσίες (εστέρες και αλκοόλες) [28, 105, 197, 252, 255, 258, 259, 260, 261].

Όσον αφορά τις χημειομετρικές τεχνικές, τα ANN έχουν ευρέως χρησιμοποιηθεί [74, 103, 105, 149, 197, 252, 254, 259, 261]. Επίσης, η μέθοδος SIMCA (Soft Independnt Modelling of Class Analogy) έχει εφαρμοστεί σε πολλές περιπτώσεις [197, 256, 257, 258]. Η DA εξάλλου, χρησιμοποιήθηκε σε διάφορες παραλλαγές (LDA, UNEQ-DA, PLS-DA) [28, 103, 197, 252, 254, 255, 256, 257, 258, 261]. Τα CT (CART) έχουν χρησιμοποιηθεί μία μόνο φορά για την ταξινόμηση δειγμάτων ελαιολάδου από έξι (6) Μεσογειακές χώρες με βάση τα φάσματα NMR του μη σαπωνοποιημένου κλάσματος (unsaponifiable matter) [197].

Ωστόσο, η αυθεντικότητα που αφορά γεωγραφικό χαρακτηρισμό ή ποικιλία, δεν μπορεί να αποδειχθεί με χημικές ενώσεις των οποίων η σταθερότητα επηρεάζεται από το χρόνο [253]. Έτσι ο εναλλακτικός τρόπος που παρουσιάζεται στο κεφάλαιο αυτό, είναι ο προσδιορισμός

στοιχείων σπάνιων γαιών (Rare Earth Elements, REE), των οποίων η συγκέντρωση δεν επηρεάζεται από το πέρασμα του χρόνου, αλλά περισσότερο από το καλλιεργήσιμο έδαφος [262]. Ποτέ στο παρελθόν δεν έχει χρησιμοποιηθεί η περιεκτικότητα των ελαιολάδων σε REE για τη γεωγραφική ταυτοποίησή τους. Είναι πρώτη φορά που συγκεντρώθηκαν και αξιολογήθηκαν στατιστικά δεδομένα τέτοιου είδους. Η συγκεκριμένη εργασία δημοσιεύτηκε πρόσφατα [263].

Έτσι στο κεφάλαιο που ακολουθεί, δεδομένα που αφορούν τον προσδιορισμό στοιχείων σπάνιων γαιών σε δείγματα ελαιολάδων από διάφορες περιοχές της Ελλάδας επεξεργάστηκαν με διαφορετικές στατιστικές μεθόδους: DA, CT και ANN. Η μελέτη αποσκοπεί στη γεωγραφική ταξινόμηση των δειγμάτων βασιζόμενοι αποκλειστικά στη σύστασή τους σε στοιχεία σπάνιων γαιών. Αποτελέσματα της επιτυχούς κατάταξής τους, θα μπορούσαν να φανούν χρήσιμα για μια μελλοντική ταυτοποίηση αγνώστων ή αμφιβόλων δειγμάτων.

Η κρισιμότητα της σωστής επιλογής της ομάδας εκπαίδευσης τουλάχιστον για τα ANN μοντέλα, έχει ήδη συζητηθεί σε προηγούμενες παραγράφους (§ 4.3.1, 4.3.2, 4.4.1, 4.4.2). Παρακινούμενοι λοιπόν από τον προβληματισμό αυτό, χρησιμοποιήθηκαν τα αποτελέσματα της DA (βλ. παρακάτω, § 7.2.2, 7.3.2), ώστε να αξιοποιηθούν όλες οι διαθέσιμες αρχικές μεταβλητές και να επιλεχθεί η πιο αντιπροσωπευτική ομάδα εκπαίδευσης. Η ομάδα αυτή χρησιμοποιήθηκε συγκριτικά για τις επιβλεπόμενες τεχνικές που ακολούθησαν (CT και ANN). Ειδικότερα, μια καινοτόμος προσέγγιση επιχειρήθηκε για την επιλογή των πιο αντιπροσωπευτικών ομάδων εκπαίδευσης και ελέγχου: αντί των αρχικών μεταβλητών (REE) χρησιμοποιήθηκαν οι γραμμικοί συνδυασμοί τους (LDF ή roots, § 2.1.1). Με τον τρόπο αυτό, επιλέγοντας τα δείγματα με τις πιο **"ακραίες" τιμές των roots,** υπάρχει η δυνατότητα "εκμετάλλευσης" όλων των αρχικών μεταβλητών για την επιλογή των πιο αντιπροσωπευτικών ομάδων εκπαίδευσης των μοντέλων CT και ANN.

Τα δεδομένα λοιπόν επεξεργάστηκαν με την κλασική επιβλεπόμενη πολυπαραμετρική DA τεχνική (εξαγωγή των roots) αλλά και με CT και ANN (κατασκευή μοντέλων). Ενενήνταεπτά (97) δείγματα από τέσσερις (4) περιοχές της Ελλάδας αναλύθηκαν για δέκα (10) REE και μοντέλα CT και ANN κατασκευάστηκαν και συγκρίθηκαν μεταξύ τους ως προς την ικανότητα ταξινόμησης.

7.2. ΜΕΘΟΔΟΛΟΓΙΑ

7.2.1. Συλλογή και ανάλυση των δειγμάτων

Συνολικά συλλέχτηκαν δείγματα από έντεκα (11) γεωγραφικές περιοχές της Ελλάδας: Αρκαδία, Εύβοια, Ζάκυνθος, Ηράκλειο, Λακωνία, Λέσβος, Μεσσηνία, Πιερία, Ρέθυμνο, Χαλ-

κιδική και Χανιά. Οι περισσότερες περιοχές ωστόσο, (όπως Αρκαδία, Εύβοια, Λέσβος, Πιερία, Ρέθυμνο, Χαλκιδική και Χανιά) παραλείφτηκαν από τη στατιστική ανάλυση που ακολουθεί, καθώς αντιπροσωπεύονταν από πολύ μικρό αριθμό δειγμάτων (από 1 ως 9). Έτσι, οι ομάδες ταξινόμησης που τελικά επιλέχθηκαν λόγω του μεγαλύτερου αριθμού δειγμάτων ήταν το **Ηράκλειο** (I, n=22), **η Λακωνία** (LA, n=17), **η Μεσσηνία** (ME, n=34) **και η Ζάκυνθος** (ZA, n=24). Συνολικά αξιοποιήθηκαν 97 δείγματα.

Παράλληλα, οι παράμετροι που μετρήθηκαν στα δείγματα ήταν 14 σπάνιες γαίες: Y (Ύττριο), La (Λανθάνιο), Ce (Δημήτριο), Pr (Πρασινοδύμιο), Nd (Νεοδύμιο), Sm (Σαμάριο), Gd (Γαδολίνιο), Tb (Τέρβιο), Dy (Δυσπρόσιο), Ho (Όλμιο), Er (Έρβιο), Tm (Θούλιο), Yb (Υττέρβιο) και Th (Θόριο). Από αυτές, κάποιες: Ce, Nd, Tb, Th (σημειώνονται με έντονη γραφή στον πίνακα των αποτελεσμάτων ΚΕΦ. 4, Π) απορρίφθηκαν εξ' αρχής καθώς πολλές από τις τιμές τους ήταν κάτω του ορίου ανίχνευσης.

Η ανάλυση των δειγμάτων πραγματοποιήθηκε από τις Κ. Μηνιώτη και Ε. Ιωάννου στο Χημικό Εργαστήριο του Γ.Π.Α σε διαφορετικές χρονικές περιόδους. Προηγήθηκε χώνευση των δειγμάτων σε φούρνο μικροκυμάτων (CEM microwave-assisted system, Mars, USA) όπως περιγράφεται παρακάτω: 0,500 g κάθε δείγματος ελαιολάδου ζυγίστηκε σε κατάλληλο δοχείο και προστέθηκαν σε αυτό 5 ml π. HNO_3. Τα δείγματα αφέθηκαν για 30 min και ακολουθήθηκε συγκεκριμένο πρόγραμμα χώνευσης. Τα τελικά δείγματα αραιώθηκαν σε τελικό όγκο 20 ml με υπερκάθαρο νερό.

Για τον προσδιορισμό των REE χρησιμοποιήθηκε η τεχνική ICP-MS (PE SCIEX, Canada). Οι παράμετροι λειτουργίας του οργάνου δίνονται στο παράρτημα, ΚΕΦ. 4, Π).

Η μέθοδος επικυρώθηκε και τα δεδομένα ακρίβειας (ανακτησιμότητα: 83 – 123 %, μέση τιμή = 97 %) και επαναληψιμότητας φαίνονται στο παράρτημα, ΚΕΦ. 4, Π). Κατάλληλα πρότυπα διαλύματα χρησιμοποιήθηκαν για την κατασκευή των καμπυλών αναφοράς, ενώ προστέθηκε In (Ίνδιο) ως εσωτερικό πρότυπο. Όλες οι μετρήσεις έγιναν εις τριπλούν, ενώ υπολογίστηκαν τα όρια ανίχνευσης (LOD) (ΚΕΦ. 4, Π) και ποσοτικοποίησης (LOQ).

7.2.2. Προκατεργασία με χρήση της Διαχωριστικής Ανάλυσης (DA)

Τα συνολικά αποτελέσματα μαζί με κάποιες βασικές στατιστικές παραμέτρους, φαίνονται στον πίνακα των αποτελεσμάτων (ΚΕΦ. 4, Π). Οι συγκεντρώσεις σε REE που βρέθηκαν χαμηλότερες από το όριο ανίχνευσης της μεθόδου (LOD), αντικαταστάθηκαν από το μισό αυτού (LOD/2), στη στατιστική επεξεργασία που ακολουθήθηκε.

Επιπλέον, χρησιμοποιήθηκε η DA για την προκατεργασία των αποτελεσμάτων και την εύρεση της πιο αντιπροσωπευτικής ομάδας για την εκπαίδευση των μοντέλων.

Οι Zupan και Gasteiger στο βιβλίο τους "Νευρωνικά Δίκτυα για Χημικούς" [1], εφαρμόζουν μια τεχνική πειραματικού σχεδιασμού (experimental design technique), θέλοντας να τεκμηριώσουν την αξία της σωστής επιλογής μιας αντιπροσωπευτικής ομάδας εκπαίδευσης για τα ANN. Η τεχνική τους βασίζεται στη διαφοροποίηση τριών επιπέδων για τις συνεχείς μεταβλητές που χρησιμοποιούν για την πρόβλεψη της δραστικότητας των δεσμών οργανικών ενώσεων (βλ. επίσης § 4.3.2).

Επιλέγονται λοιπόν αρχικά 4 μεταβλητές που κρίνονται σημαντικές και διαχωρίζονται 3 επίπεδα σε καθεμιά από αυτές: χαμηλό, μεσαίο και υψηλό. Έτσι, δημιουργούνται $3^4 = 81$ "κελιά" που πρέπει να γεμίσουν με αντίστοιχα δείγματα εκπαίδευσης. Αυτά, θεωρητικά αλλά και πειραματικά όπως αποδεικνύεται, καλύπτουν την ενδιαφερόμενη περιοχή και δίνουν καλύτερα ποσοστά εκπαίδευσης και ελέγχου στα μοντέλα που θα εκπαιδευτούν.

Από την άλλη πλευρά, η DA (§ 2.1.1) στηρίζεται στην ανίχνευση ομάδων οι οποίες βασικά διαφέρουν στο μέσο όρο των μεταβλητών [17]. Οι νέες συναρτήσεις (LDF) που προκύπτουν με την εφαρμογή της DA αφορούν τη μέγιστη διαφοροποίηση των ομάδων και σχετίζονται άμεσα με τα επίπεδα των μεταβλητών για κάθε δείγμα. Βασιζόμενοι λοιπόν στην ιδέα της τεχνικής Zupan και Gasteiger που αναφέρθηκε παραπάνω, μπορούμε να χρησιμοποιήσουμε τις σημαντικότερες LDF (νέες συντεταγμένες) που προκύπτουν από τα αρχικά δεδομένα και οι οποίες ουσιαστικά περιγράφουν τα δείγματα (ή αλλιώς τα επίπεδα των μεταβλητών σε αυτά). Χρησιμοποιήθηκε λοιπόν αρχικά η κλασική DA προσέγγιση ώστε:

1. να εξαχθούν συμπεράσματα για την κρισιμότητα των 10 μεταβλητών,
2. να γίνει μια προσπάθεια αρχικής αξιολόγησης των δεδομένων (κατασκευή του μοντέλου και εκτίμηση των δυνατοτήτων που υπάρχουν),
3. να υπολογιστούν οι νέες συντεταγμένες (roots), που αντικαταστούν τις αρχικές 10 μεταβλητές (REE) και θα βοηθήσουν στη διαμόρφωση μιας αντιπροσωπευτικής ομάδας εκπαίδευσης για τα CT και ANN μοντέλα,
4. να "καθοριστούν" τα όρια των μεταβλητών αυτών και να βρεθούν τα "ακραία" αυτά δείγματα που μπορούν να ενταχθούν στην ομάδα εκπαίδευσης.

   Παράλληλα οι DA τεχνικές FW και BW χρησιμοποιήθηκαν για λόγους σύγκρισης.

7.2.3. Δέντρα Ταξινόμησης (CT)

Από την τεχνική των CT, χρησιμοποιήθηκαν και οι τρεις ήδη δοκιμασμένες σε μέθοδοι: LCM (Discriminat-based linear combination method), Classic CT (Discriminat-based univariate method) και CART (CART-style Exhaustive search method for univariate splits). Στο κεφάλαιο αυτό όμως, αναφέρεται αναλυτικά η πρώτη μόνο μέθοδος (LCM), καθώς έδωσε τα καλύτερα

αποτελέσματα. Λεπτομέρειες για τις άλλες μεθόδους καταγράφονται στο παράρτημα ( ΚΕΦ. 4, Π). Σύγκριση των μεθόδων παρατίθεται παρακάτω (§ 7.3.4).

Δυο προσεγγίσεις δοκιμάστηκαν:

1η προσέγγιση: Χρήση των roots 1, 2 για μια προσεκτική επιλογή της ομάδας εκπαίδευσης (49 δείγματα), βάση των ακραίων τιμών αυτών. Στη συνέχεια, όλες οι μεταβλητές (REE) χρησιμοποιήθηκαν ως εισερχόμενα στα CT μοντέλα. Συγκεκριμένα, το διάγραμμα από την Canonical analysis (§ 7.3.2) που προκύπτει από την DA χρησιμοποιήθηκε για να "ανιχνεύσει" τα πιο ακραία δείγματα από το σύνολο των διαθέσιμων. Αυτά χρησιμοποιήθηκαν για τη σύσταση της ομάδας εκπαίδευσης ώστε να καλύπτουν μια ευρεία περιοχή.

2η προσέγγιση: Χρήση όλων των αρχικών μεταβλητών (REE) και τυχαία επιλογή των δειγμάτων εκπαίδευσης (49 δείγματα), αλλά και των δειγμάτων ελέγχου (εναπομείναντα 48).

Η τελική σύγκριση που έγινε, αφορούσε τις μεθόδους μεταξύ τους, αλλά και τις δυο προσεγγίσεις στο σύνολό τους. Τα μοντέλα που κατασκευάστηκαν, συγκρίθηκαν όσον αφορά τις ικανότητες αναγνώρισης και πρόβλεψης (§ 4.2.2). Εναλλακτικά, η σύγκριση των μοντέλων μπορεί να γίνει με τη χρήση των CV και resubstitution costs (§ 5.3.6).

Για τις τρεις μεθόδους CT οι παράμετροι που χρησιμοποιούνται προέρχονται μετά από συνεχείς δοκιμές και βελτιστοποιήσεις. Έτσι για παράδειγμα, όλες οι δυνατές επιλογές για την τιμή του FO (§ 5.3.4) έχουν δοκιμαστεί, πριν οριστικοποιηθεί η τελική τιμή της παραμέτρου.

7.2.4. Νευρωνικά Δίκτυα (ANN)

Λόγω της ιδιαιτερότητας της βάσης δεδομένων (βλ. παρακάτω, § 7.3.1), εξετάστηκαν γραμμικά δίκτυα, MLP και RBF, καθώς τα δίκτυα Kohonen απέτυχαν σε οποιαδήποτε προσπάθεια ταξινόμησης των δειγμάτων.

Εφαρμόστηκαν δυο προσεγγίσεις για την κατασκευή/εκπαίδευση των μοντέλων. Για να συγκριθούν δε καλύτερα οι προσεγγίσεις αυτές και να καλυφθούν οι ανάγκες των ANN για ανεξάρτητες ομάδες εκπαίδευσης και ελέγχου χρησιμοποιήθηκαν ισάριθμες ομάδες και για τις δυο προσεγγίσεις. Έτσι, στην παράγραφο αυτή δουλέψαμε ως εξής:

1η προσέγγιση: Εισερχόμενες μεταβλητές: οι επτά (7) πιο κρίσιμες μεταβλητές όπως προέκυψαν από την DA κλασική τεχνική, καθώς τα Νευρωνικά Δίκτυα είναι επιρρεπή σε φαινόμενα υπερ-προσαρμογής.

Ομάδα εκπαίδευσης (49 δείγματα): προσεχτικά επιλεγμένη ώστε να καλύπτει όλο το εύρος των τιμών των μεταβλητών.

Ομάδα ελέγχου (48 δείγματα): όσα δείγματα απομένουν εκτός της ομάδας εκπαίδευσης (δεν χρησιμοποιείται ομάδα επικύρωσης).

$2^{η}$ προσέγγιση: Εισερχόμενες μεταβλητές: οι επτά (7) πιο κρίσιμες μεταβλητές όπως προέκυψαν από την DA κλασική τεχνική.

Ομάδα εκπαίδευσης (49 δείγματα): τυχαία επιλεγμένη.

Ομάδα ελέγχου (48 δείγματα): όσα δείγματα απομένουν εκτός της ομάδας εκπαίδευσης (δεν χρησιμοποιείται ομάδα επικύρωσης).

Εδώ πρέπει να αναφερθεί ότι η μη χρήση της ομάδας επικύρωσης έχει αναφερθεί παλαιότερα στη βιβλιογραφία [79], αλλά και σε πιο σύγχρονες εργασίες, εφόσον πραγματοποιείται αξιολόγηση των μοντέλων (διασταυρούμενη ή εξωτερική) και ελέγχονται κριτήρια τερματισμού (§ 4.3.7) για την ελάχιστη τιμή τους [23, 61, 62, 75, 119, 139, 202, 264]. Εκτός όμως των παραπάνω, στην παρούσα εφαρμογή, γίνεται μια προσπάθεια αξιολόγησης της χρησιμοποιούμενης νέας τεχνικής για την επιλογή των ομάδων εκπαίδευσης και ελέγχου. Έτσι, ο σκοπός της μελέτης, αφορά τη σύγκριση των αποτελεσμάτων για τις δυο προσεγγίσεις και όχι τόσο τη δημιουργία του καλύτερου μοντέλου. Το τελευταίο μπορεί να επιτευχθεί σε μια βάση δεδομένων με καλύτερες προδιαγραφές (βλ. § 7.3.1) η οποία θα παρέχει και τη δυνατότητα τριχοτόμησης των δεδομένων σε ομάδες εκπαίδευσης, επικύρωσης και ελέγχου.

Αρχικά, έγιναν πολλαπλές επαναλήψεις για την κατασκευή των μοντέλων, ώστε να αποφθεχθούν τοπικά ελάχιστα και να βελτιστοποιηθούν οι αρχικές συνθήκες (§ 5.3.8). Στη συνέχεια, τα βέλτιστα μοντέλα ANN συγκρίθηκαν μεταξύ τους, αλλά και με τα αντίστοιχα CT.

7.3. ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΣΥΖΗΤΗΣΗ

7.3.1. Προσδιορισμός των REE και συσχετίσεις Spearman

Η συγκεκριμένη βάση ήταν η δυσκολότερη από όσες μελετήθηκαν στην εργασία αυτή, καθώς περιείχε λίγα δείγματα (125 αρχικά) που έπρεπε να ταξινομηθούν σε πολλές ομάδες (11 αρχικά) με βάση ένα πλήθος παραμέτρων (14) με αντικρουόμενα αποτελέσματα για την κρισιμότητά τους. Έτσι, οι μη επιβλεπόμενες τεχνικές (PCA και CA, αλλά και τα δίκτυα Kohonen, βλ. § 7.2.4) λόγω της δυσκολίας της βάσης, δεν επέτρεψαν ικανοποιητική ομαδοποίηση των περιοχών προέλευσης ή την εξαγωγή ασφαλών συμπερασμάτων για τις μεταβλητές.

Επιπλέον, οι χρησιμοποιούμενες τελικά επιβλεπόμενες τεχνικές μπορούν να δώσουν ποσοτικά αποτελέσματα για την ομάδα εκπαίδευσης ή την ομάδα ελέγχου, με αποτέλεσμα να μπορούν να γίνουν συγκρίσεις μεταξύ τους.

Από τον πίνακα των αποτελεσμάτων ( ΚΕΦ. 4, Π) και τον πίνακα 7.1 των συσχετίσεων Spearman που ακολουθεί, μπορούν να εξαχθούν τα πρώτα συμπεράσματα σχετικά με τη συσχέτιση των μεταβλητών αλλά και τη διαφοροποίηση των τεσσάρων περιοχών. Οι μεταβλητές δείχνουν γενικά **μικρές συσχετίσεις μεταξύ τους,** ενώ γενικά δεν παρατηρήθηκαν διαφορές των μεταβλητών για τις τέσσερις επιλεγμένες περιοχές.

Πίνακας 7.1: Spearman συσχετίσεις των μεταβλητών. Σημειώνονται με έντονη γραφή οι συσχετίσεις με συντελεστή Spearman ≥ 0,60.

	Y	**La**	**Pr**	**Sm**	**Gd**	**Dy**	**Ho**	**Er**	**Tm**	**Yb**
Y	**1,00**									
La	**0,61**	**1,00**								
Pr	**0,79**	**0,79**	**1,00**							
Sm	**0,66**	0,45	**0,62**	**1,00**						
Gd	**0,67**	**0,63**	**0,65**	**0,60**	**1,00**					
Dy	**0,75**	0,31	0,58	0,59	0,45	**1,00**				
Ho	0,06	0,03	-0,05	0,21	0,28	0,02	**1,00**			
Er	0,47	0,31	0,37	0,51	0,49	0,39	0,41	**1,00**		
Tm	0,01	0,16	0,02	0,16	0,17	-0,10	**0,61**	0,52	**1,00**	
Yb	0,26	0,32	0,19	0,41	0,50	0,11	**0,66**	**0,61**	**0,70**	**1,00**

Τα παραπάνω επιβεβαιώθηκαν στο σχήμα 4.1 ( ΚΕΦ. 4, Π), όπου απεικονίζονται τα διαγράμματα μερικών από τις μεταβλητές ανά περιοχή. Είναι φανερή η συσχέτιση των μεταβλητών Y-Pr-Sm ή Yb-Tm, αλλά και η διαφοροποίηση των Gd-Dy-Ho. Τέλος, μερικά σημεία LA (Λακωνία) και ΜΕ (Μεσσηνία) φαίνονται να διαφοροποιούνται από τα υπόλοιπα της κατηγορίας τους.

7.3.2. Εφαρμογή της Διαχωριστικής Ανάλυσης (DA)

Με την εφαρμογή της DA, οι νέες μεταβλητές που χαρακτηρίζουν τώρα το κάθε δείγμα είναι 3 (4 ομάδες ταξινόμησης μείον ένα, § 2.1.1). Ο πίνακας που καταγράφει τις τιμές αυτές για τα 97 δείγματα συμπεριλαμβάνεται στο παράρτημα της εργασίας αυτής (ΚΕΦ. 4, Π).

Τα αποτελέσματα των τριών προσεγγίσεων DA φαίνονται στον πίνακα 7.2. Συμφωνούν αρκετά μεταξύ τους σε επίπεδο ποσοστών επιτυχίας (οι κλασική και FW τεχνικές) αλλά και στην επιλογή των κρισιμότερων μεταβλητών. Καλύτερα διαχωριζόμενες (με υψηλά ποσοστά ακρίβειας) είναι οι περιοχές της Μεσσηνίας (ΜΕ) και Λακωνίας (LA). Τα Gd, Sm, Y, Dy, Pr, Tm και Er φαίνονται να είναι οι πιο κρίσιμες μεταβλητές. Η κλασική τεχνική DA που "εκμεταλλεύτηκε" όλες τις παραμέτρους έδωσε τα υψηλότερα ποσοστά για την ομάδα εκπαίδευσης: 76,3 %. Κρισιμότερες μεταβλητές αναδείχθηκαν και εδώ οι: Gd, Sm, Y, Dy, Pr, Tm και Er. Αντίθετα, η BW με χρήση μόλις 5 μεταβλητών πέτυχε ένα ποσοστό ακρίβειας περίπου 63 %.

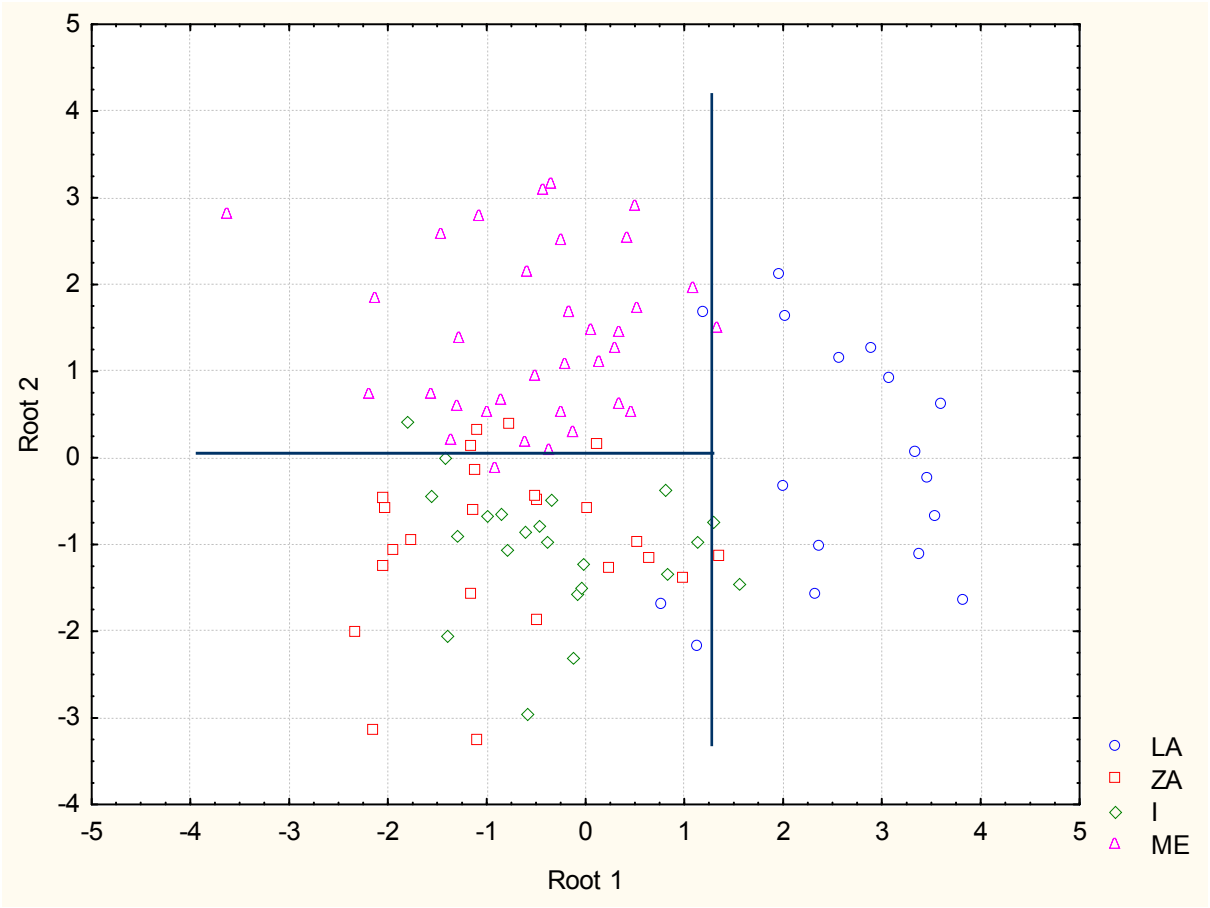
Πίνακας 7.2: Αποτελέσματα για την ομάδα εκπαίδευσης (3 τεχνικές DA: 10 μεταβλητές και 97 δείγματα).

Περιοχή	**Συνολικός αρ. δειγμάτων**	**Κλασική**		**FW**		**BW**	
		TP*	**% Ποσοστά επιτυχίας**	**TP**	**% Ποσοστά επιτυχίας**	**TP**	**% Ποσοστά επιτυχίας**
I	22	11	50,0	8	36,4	7	31,8
LA	17	14	82,4	14	82,4	14	82,4
ΜΕ	34	33	97,0	32	94,1	28	82,4
ZA	24	16	66,7	15	62,5	12	50,0
Συνολικά	97	74	**76,3**	69	**71,1**	61	**62,9**
Σειρά κρισιμότητας των μεταβλητών		Gd>Sm>Y>Dy>Pr>Tm >Er>La>Yb>Ho		Gd>Sm>Tm>Pr> Dy>Y		Gd>Tm>Pr>Sm>Dy	

*TP: True positive

Στο σχήμα 7.1 απεικονίζεται ο διαχωρισμός των τεσσάρων περιοχών (Canonical analysis plot) όπως αυτός επιτυγχάνεται από τις γραμμικές διαχωριστικές συναρτήσεις (LDF) και την κλασική τεχνική DA. Χαρακτηριστική είναι η "ανάμιξη" των περιοχών Ηρακλείου (Ι) και Ζακύνθου (ZA) και της Μεσσηνίας (ΜΕ) σε μικρότερο ποσοστό. Η περιοχή της Λακωνίας (LA)

διαχωρίζεται ικανοποιητικά. Ειδικότερα, οι περιοχές αυτές (Μεσσηνία και Λακωνία) φαίνεται ότι θα μπορούσαν να διαχωριστούν γραμμικά.



Σχήμα 7.1: Διαχωρισμός των τεσσάρων περιοχών από την DA κλασική τεχνική.

Η ίδια εντύπωση αποκομίζεται και από τον πίνακα ταξινόμησης όπως αυτός απεικονίζεται στο παράρτημα (ΚΕΦ. 4, Π), όπου σημειώνονται με έντονη γραφή οι λαθεμένες εκτιμήσεις για τα δείγματα Ηρακλείου (Ι) και Ζακύνθου (ΖΑ).

Ανεξάρτητη αξιολόγηση των τριών μοντέλων DA (με ομάδα ελέγχου) δεν είναι δυνατή, (λόγω του μικρού αριθμού δειγμάτων), αλλά επιπλέον είναι εκτός του σκοπού της ανάλυσης.

Το σχήμα 7.1 και οι συντεταγμένες (roots 1, 2) που απεικονίζει για την κλασική τεχνική, θα χρησιμοποιήθηκαν παραπέρα για την επιλογή των δειγμάτων εκπαίδευσης. Δηλαδή, τα πιο ακραία από αυτά με τις "οριακές" συντεταγμένες επιλέχθηκαν ως δείγματα εκπαίδευσης, ενώ τα "εσωτερικά" χρησιμοποιήθηκαν ως δείγματα ελέγχου. Τα δείγματα εκπαίδευσης χρησιμοποιήθηκαν στην κατασκευή των μοντέλων CT (§ 7.3.3) και ANN (§ 7.3.5, 7.3.6).

Ένα παράδειγμα επιλογής των αντιπροσωπευτικότερων δειγμάτων, φαίνεται στον πίνακα 7.3 για τα δείγματα Ηρακλείου (Ι). Τα δείγματα κατατάσσονται κατά αύξουσα σειρά root 1 (πρώτες στήλες του πίνακα) και root 2 (τελευταίες στήλες του πίνακα). Τα δείγματα με τις

ακραίες τιμές των roots 1,2 με τη βοήθεια πάντα του σχήματος 7.1 επιλέχθηκαν ως ομάδα εκπαίδευσης (έγχρωμα κελιά στον πίνακα 7.3).

Πίνακας 7.3: Παράδειγμα επιλογής της ομάδας εκπαίδευσης για τα δείγματα Ηρακλείου.

Α/Α Δείγματος	**Root 1**	**Root 2**	**Α/Α Δείγματος**	**Root 1**	**Root 2**
14	-1,7994	0,4086	**4**	-0,5853	-2,9595
22	-1,5671	-0,4476	**19**	-0,1142	-2,3165
13	-1,4170	-0,0116	**8**	-1,4040	-2,0593
8	-1,4040	-2,0593	**11**	-0,0835	-1,5718
15	-1,2846	-0,8994	**20**	-0,0485	-1,5105
2	-1,0001	-0,6653	**16**	1,5590	-1,4578
17	-0,8427	-0,6601	**7**	0,8300	-1,3454
9	-0,7945	-1,0644	**1**	-0,0263	-1,2240
3	-0,5983	-0,8692	**9**	-0,7945	-1,0644
4	-0,5853	-2,9595	**21**	1,1266	-0,9690
18	-0,4672	-0,7790	**12**	-0,3793	-0,9685
12	-0,3793	-0,9685	**15**	-1,2846	-0,8994
6	-0,3404	-0,4874	**3**	-0,5983	-0,8692
19	-0,1142	-2,3165	**18**	-0,4672	-0,7790
11	-0,0835	-1,5718	**5**	1,2875	-0,7355
20	-0,0485	-1,5105	**2**	-1,0001	-0,6653
1	-0,0263	-1,2240	**17**	-0,8427	-0,6601
10	0,8133	-0,3870	**6**	-0,3404	-0,4874
7	0,8300	-1,3454	**22**	-1,5671	-0,4476
21	1,1266	-0,9690	**10**	0,8133	-0,3870
5	1,2875	-0,7355	**13**	-1,4170	-0,0116
16	1,5590	-1,4578	**14**	-1,799	0,409

7.3.3. Εφαρμογή των Δέντρων Ταξινόμησης (CT) - Μέθοδος των γραμμικών συνδυασμών (LCM)

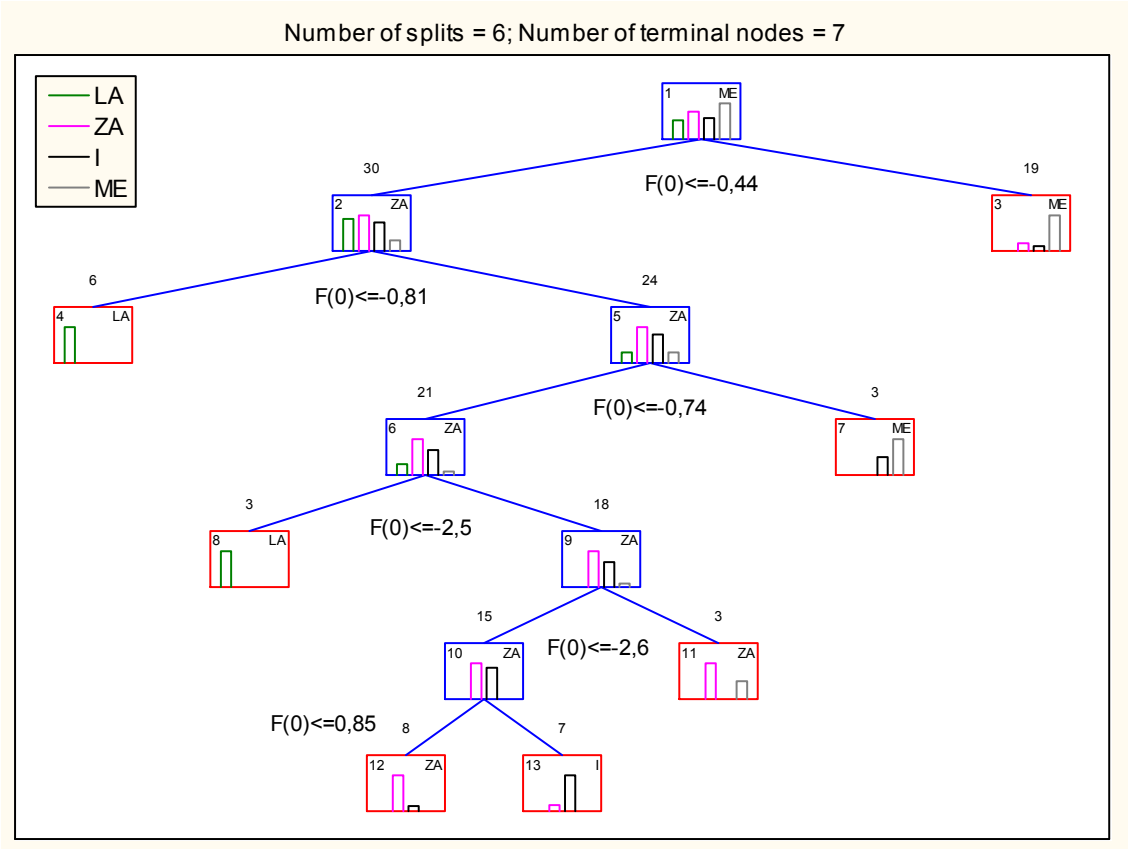
Με βάση την πρώτη προσέγγιση (αρχείο των DA roots και επιλεγμένες ομάδες εκπαίδευσης και ελέγχου), τα καλύτερα αποτελέσματα ελήφθησαν για FO = 0,3 (§ 5.3.4). Ο αντίστοιχος πίνακας ταξινόμησης για την ομάδα εκπαίδευσης (σύνολο 49 δειγμάτων) φαίνεται στον πίνακα 7.4. Το σχετικό δέντρο απεικονίζεται στο σχήμα 7.2.

Πίνακας 7.4: Αποτελέσματα για την ομάδα εκπαίδευσης (μέθοδος LCM): Παρατηρούμενες θέσεις (στήλες) έναντι προβλεπόμενων (σειρές)

Παρατηρήσεις / **Προβλέψεις**	**LA**	**ZA**	**I**	**ME**	**Συνολικά**
LA	9	0	0	0	
ZA	0	9	1	1	
I	0	1	6	0	
ME	0	**3**	**3**	16	
Συνολικός αρ. δειγμάτων	9/9	9/13	6/10	16/17	40/49
% Ποσοστά επιτυχίας	100,0	69,2	60,0	94,1	81,6

Είναι φανερό, ότι αρκετά δείγματα ZA (3) και I (9) αποδίδονται εσφαλμένα στη Μεσσηνία (ΜΕ). Η "σύγχυση" των δειγμάτων αυτών έχει ήδη επισημανθεί στο σχήμα 7.1. Η Λακωνία (LA) δίνει τα υψηλότερα ποσοστά. Αντίθετα οι περιοχές των Ζακύνθου (ZA) και Ηρακλείου (I) δίνουν τα χαμηλότερα ποσοστά, αλλά υψηλότερα από τα αντίστοιχα της DA (πίνακας 7.2, § 7.3.2).

Το δέντρο (σχ. 7.2) περιλαμβάνει 6 διαχωρισμούς με βάση γραμμικούς συνδυασμούς των τιμών των μεταβλητών και 7 τερματικούς κόμβους. Η κατασκευή (tree structure) του "δέντρου" που σχηματίζεται, απεικονίζεται στο παράρτημα ( ΚΕΦ. 4, Π). Στις τελευταίες στήλες του πίνακα αυτού, μπορεί κανείς να έχει μια ένδειξη (ουσιαστικό ρόλο παίζει και το μέγεθος των τιμών) για τη συνεισφορά των μεταβλητών στο διαχωρισμό των ομάδων: Gd, Sm, Dy είναι οι πιο σημαντικές μεταβλητές.



Σχήμα 7.2: Δέντρο ταξινόμησης για τα ελαιόλαδα των τεσσάρων περιοχών. Μέθοδος: Discriminant-base linear combination method, LCM.

Για την ομάδα ελέγχου (48 δείγματα), τα αποτελέσματα απεικονίζονται στον πίνακα 7.5.

Πίνακας 7.5: Αποτελέσματα για την ομάδα ελέγχου (μέθοδος LCM).

Παρατηρήσεις / Προβλέψεις	**LA**	**ZA**	**I**	**ME**	**Συνολικά**
LA	5	0	0	0	
ZA	1	5	**4**	0	
I	0	0	2	0	
ME	2	6	**6**	17	
Συνολικός αρ. δειγμάτων	5/8	5/11	2/12	17/17	29/48
% Ποσοστά επιτυχίας	62,5	45,5	16,7	100,0	60,4

Ένα μεγάλο ποσοστό δειγμάτων Ηρακλείου (Ι) αποδόθηκε στη Ζάκυνθο (ΖΑ) και τη Μεσσηνία (ΜΕ). Το ίδιο συμβαίνει και για δείγματα της Ζακύνθου (ΖΑ) που αποδόθηκαν στη Μεσσηνία (ΜΕ). Γενικά, παρά την προσεκτική επιλογή των δειγμάτων, τα ποσοστά για την ομάδα εκπαίδευσης (πίνακας 7.4) είναι αρκετά υψηλότερα από τα αντίστοιχα για την ομάδα ελέγχου (πίνακας 7.5). Παρατηρήθηκαν λοιπόν φαινόμενα υπερ-προσαρμογής του μοντέλου. Τα αντίστοιχα κόστη (§ 5.3.6) ήταν: Resub. Cost = 0,18 (ομάδα εκπαίδευσης) και CV cost = 0,40 (ομάδα ελέγχου).

7.3.4. Σύγκριση μεθόδων - αποτελέσματα

Τα αποτελέσματα των άλλων CT μεθόδων (1η προσέγγιση) καταγράφονται στο παράρτημα ( ΚΕΦ. 4, Π). Η σύγκριση των μεθόδων CT έγινε βάση των ορθών ποσοστών ταξινόμησης για τις ομάδες εκπαίδευσης και ελέγχου. Παράλληλα, επιχειρήθηκε σύγκριση με τα αποτελέσματα που θα είχαμε στην περίπτωση που χρησιμοποιούσαμε το σύνολο των αρχικών μεταβλητών (REE) **και τυχαίο διαχωρισμό των δειγμάτων σε ομάδες εκπαίδευσης και ελέγχου** (2η προσέγγιση, § 7.2.3).

Τα αποτελέσματα ταξινόμησης για την πρώτη προσέγγιση (χρήση των roots 1, 2 και επιλεγμένων ομάδων εκπαίδευσης/ελέγχου) φαίνονται αναλυτικά στον παρακάτω πίνακα 7.6.

Πίνακας 7.6: Σύγκριση CT μοντέλων/ποσοστά επιτυχίας (%) για τις ομάδες εκπαίδευσης και ελέγχου (1η προσέγγιση)

Discriminant-based linear combination (LCM)		**Discriminant-based univariate (Classic CT)**		**CART**	
Ομάδα εκπαίδευσης	**Ομάδα ελέγχου**	**Ομάδα εκπαίδευσης**	**Ομάδα ελέγχου**	**Ομάδα εκπαίδευσης**	**Ομάδα ελέγχου**
81,6	60,4	57,1	41,7	79,6	37,5

Η πρώτη μέθοδος (LCM) παρουσίασε τα καλύτερα ποσοστά στις δυο ομάδες δειγμάτων. Οι άλλες δυο μέθοδοι δίνουν χαμηλότερα ποσοστά ή έντονα φαινόμενα υπερ-προσαρμογής (CART).

Η χρήση όλων των αρχικών μεταβλητών με τυχαία επιλογή των δειγμάτων οδηγεί στα αποτελέσματα του πίνακα 7.7.

Η καλύτερη μέθοδος (LCM) έδωσε ένα σωστό ποσοστό 91,8 % για την ομάδα εκπαίδευσης. Το κόστος για την ίδια ομάδα (Resub. cost) είναι όπως αναμενόταν πολύ χαμηλότερο

του CV cost που αντιστοιχεί στην ομάδα ελέγχου. Το γεγονός αυτό δηλώνει έντονα φαινόμενα υπερ-προσαρμογής. Ωστόσο, από τη σύγκριση των δυο προσεγγίσεων για τα CT μοντέλα (πίνακες 7.6, 7.7), προκύπτει ότι στη δεύτερη τα ποσοστά επιτυχίας για τις ομάδες εκπαίδευσης και ελέγχου είναι υψηλότερα (91,8/68,8 αντί των 81,6/60,4 %). Έντονα φαινόμενα υπερ-προσαρμογής παρατηρούνται και στις δυο περιπτώσεις.

Η αξιολόγηση των μεταβλητών οδήγησε σε αποτελέσματα μερικώς συγκρίσιμα με τις DA τεχνικές (§ 7.3.2, πίνακας 7.2): Sm, Gd, Dy, Tm, Er φαίνονται να αναδεικνύονται οι σημαντικότερες μεταβλητές.

Πίνακας 7.7: Σύγκριση CT μοντέλων ($2^{η}$ προσέγγιση).

Discriminant-based linear combination (LCM)				**Discriminant-based univariate (Classic CT)**				**CART**			
Ομάδα εκπαίδευσης		**Ομάδα ελέγχου CVcost****	**σημαντικότερες μεταβλητές**	**Ομάδα εκπαίδευσης**		**Ομάδα ελέγχου CVcost**	**σημαντικότερες μεταβλητές**	**Ομάδα εκπαίδευσης**		**Ομάδα ελέγχου CVcost**	**σημαντικότερες μεταβλητές**
% ποσοστό επιτυχίας	**Resub. cost***			**% ποσοστό επιτυχίας**	**Resub. cost**			**% ποσοστό επιτυχίας**	**Resub. cost**		
91,8	0,082	0,31	Sm, Gd, Tm, Dy	57,1	0,43	0,52	Gd> Sm>> Er>Y> Yb	69,4	0,31	0,64	Y>Pr> Gd=Sm >La

* Resub. cost: κόστος για την ομάδα εκπαίδευσης (§ 5.3.6).

** CV cost: κόστος προκύπτον για την ομάδα ελέγχου (§ 5.3.6).

7.3.5. Εφαρμογή των ANN (1η προσέγγιση)

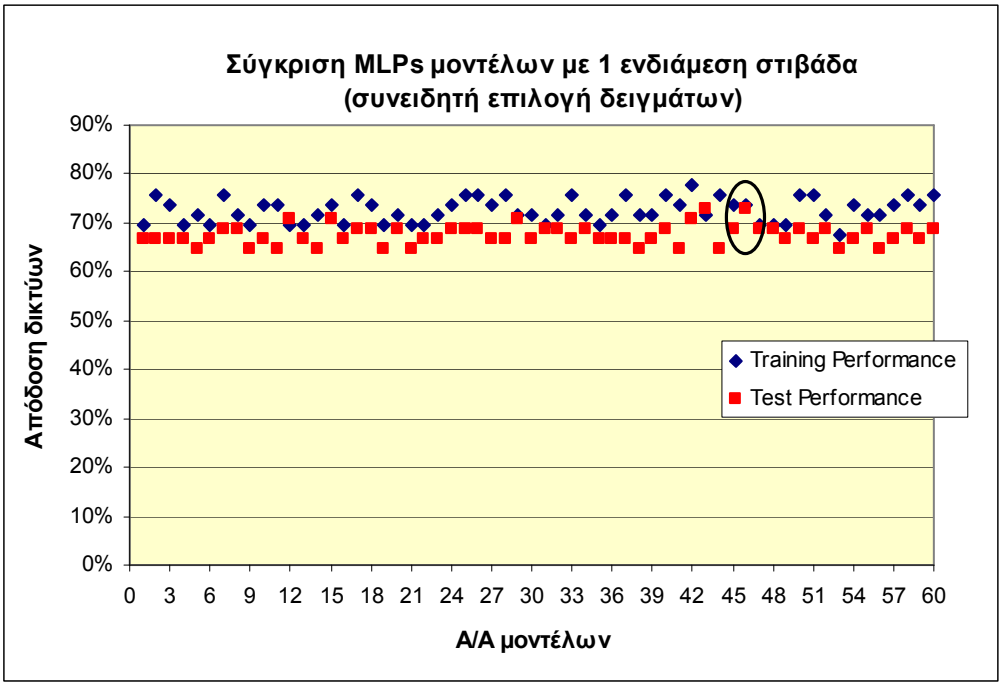
Τρεις δυνατότητες αρχιτεκτονικών ANN δοκιμάστηκαν:

1. μοντέλα MLP με μια ενδιάμεση στιβάδα,
2. μοντέλα MLP με δυο ενδιάμεσες στιβάδες,
3. RBF με μια ενδιάμεση στιβάδα.

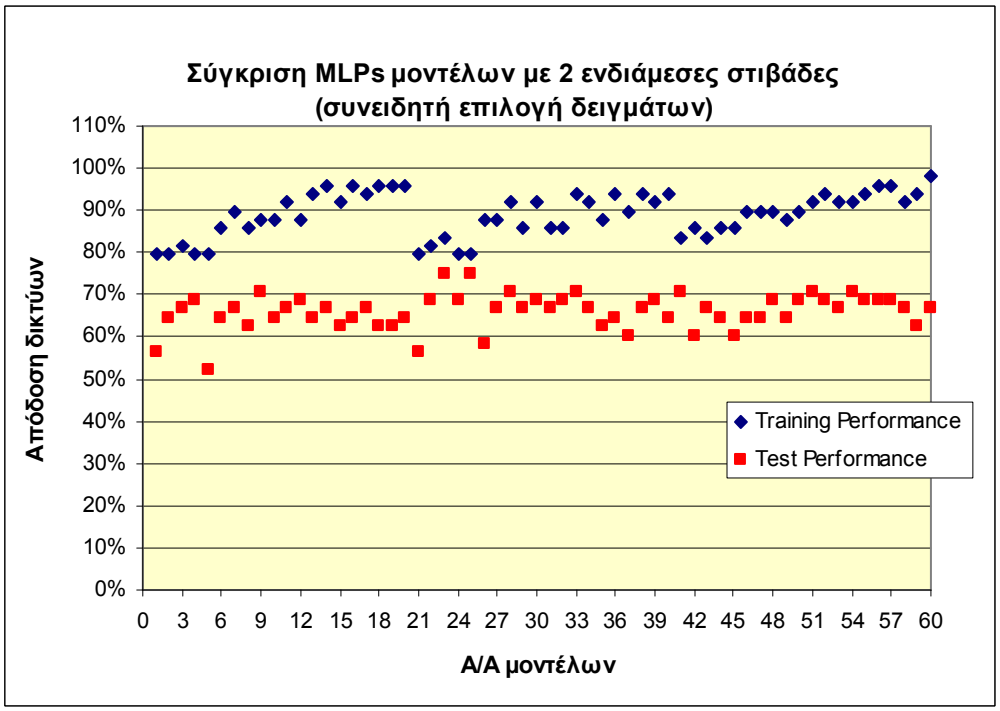
Η σύγκριση των δικτύων έγινε με βάση την απόδοση στις ομάδες εκπαίδευσης και ελέγχου (§ 4.3.7), ώστε να γίνουν άμεσα κατανοητά τα αποτελέσματα.

Από τις δέκα (10) αρχικές μεταβλητές (REE), χρησιμοποιήθηκαν μόνο επτά (7), ώστε να αποφθεχθούν φαινόμενα υπερ-προσαρμογής που γενικά ταλαιπωρούν τα ANN (§ 4.4.1). Αυτές είναι οι πιο κρίσιμες ήδη επιλεγμένες από την DA (§ 7.3.2), εκτός του Er που αντικαταστάθηκε από το La. Αυτό συνέβηκε γιατί αρχικές δοκιμές με τη βοήθεια γενετικού αλγόριθμου (GA, § 4.4.3) έδειξαν την κρισιμότητα του La έναντι του Er.

Παρακάτω (σχ. 7.3, 7.4) απεικονίζονται τα διαγράμματα σύγκρισης των αποδόσεων για όλα τα μοντέλα ANN.

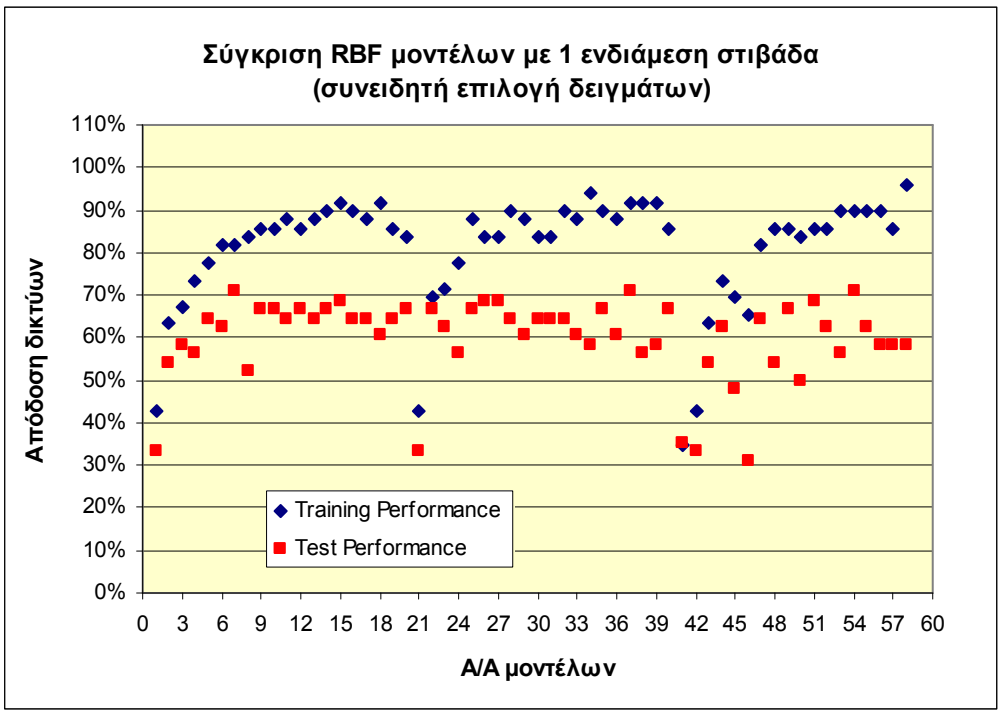


(α)

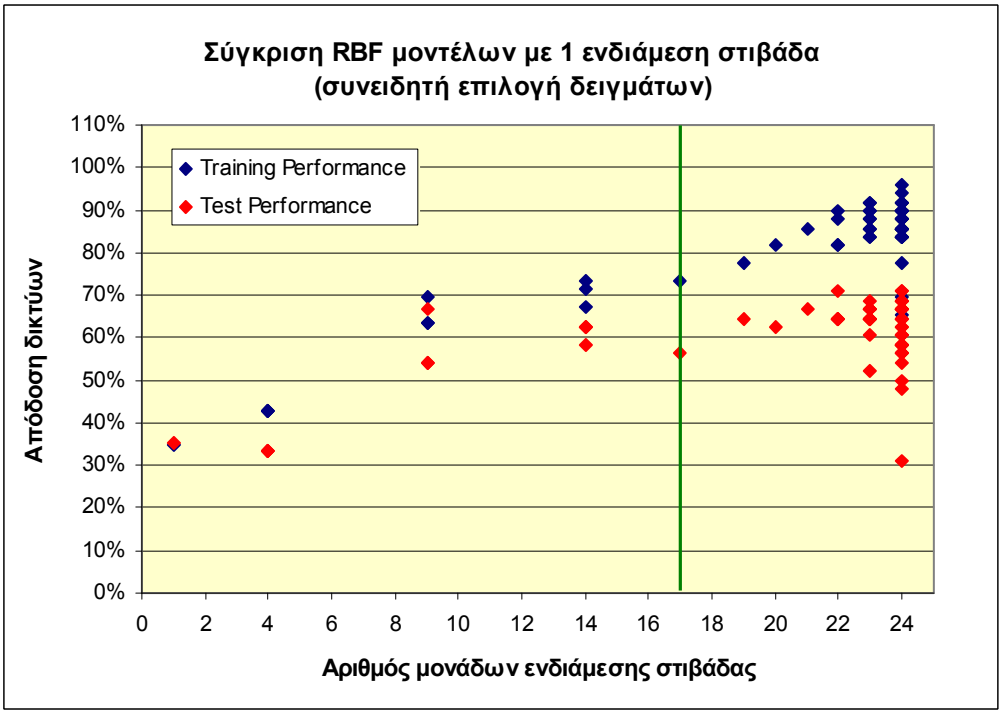


(β)

Σχήμα 7.3: Διάγραμμα σύγκρισης των αποδόσεων των ομάδων εκπαίδευσης και ελέγχου για τα μοντέλα MLP με μια (α) ή δυο (β) ενδιάμεσες στιβάδες ($1^η$ προσέγγιση).



(α)



(β)

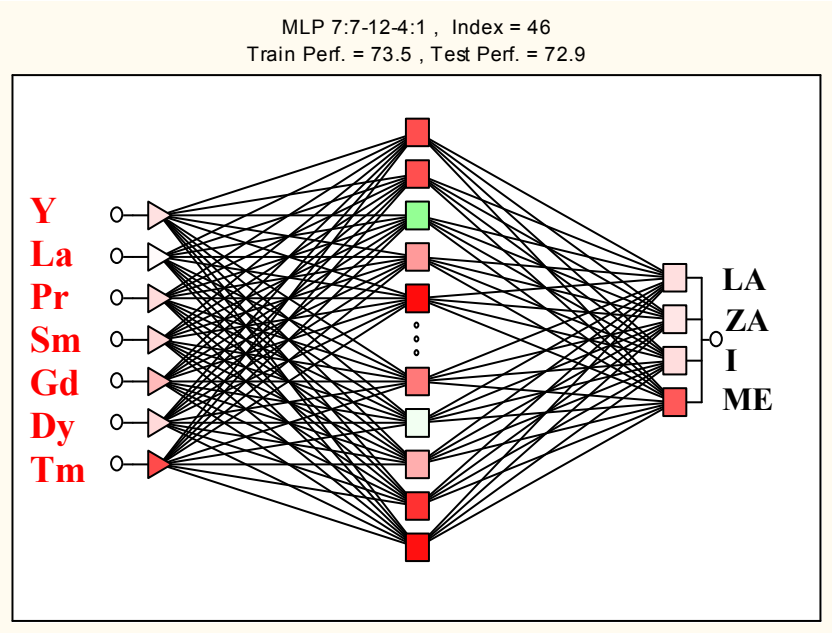
Σχήμα 7.4: Διάγραμμα σύγκρισης των αποδόσεων των ομάδων εκπαίδευσης και ελέγχου για τα μοντέλα RBF. Στον οριζόντιο άξονα απεικονίζεται (α) ο A/A των μοντέλων ή (β) ο αριθμός των μονάδων της ενδιάμεσης στιβάδας ($1^{η}$ προσέγγιση).

Για τα μοντέλα MLP με μια (σχ. 7.3(α)) ή δυο στιβάδες (σχ. 7.3(β)), χρησιμοποιήθηκαν πάντα όλες οι μεταβλητές (7 REE). Τα ποσοστά για τα δείγματα εκπαίδευσης (training

performance) είναι πάντα ελαφρά υψηλότερα από τα αντίστοιχα για την ομάδα ελέγχου (test performance) ή κυμαίνονται στα ίδια επίπεδα, για τα μοντέλα MLP με μια ενδιάμεση στιβάδα. Αυτό συμβαίνει γιατί η πορεία που ακολουθήθηκε για την επιλογή της ομάδας εκπαίδευσης έδωσε την πιο αντιπροσωπευτική ομάδα δειγμάτων. Έτσι, η ομάδα ελέγχου που απέμεινε υπακούει ουσιαστικά στον κανόνα της παρεμβολής. Τα μοντέλα είναι γενικά πολύ καλά με υψηλά ποσοστά επιτυχίας (training performance > 67 % και test performance > 64 %).

Τα μοντέλα MLP με δυο στιβάδες έδωσαν γενικά υψηλότερα ποσοστά από τα αντίστοιχα με μια μόνο στιβάδα για την ομάδα εκπαίδευσης (training performance). Ωστόσο, επειδή γενικά είναι πολύπλοκα στη δομή τους δεν έδωσαν υψηλά ποσοστά για την ομάδα ελέγχου (test performance). Έτσι, **παρουσιάζουν έντονα φαινόμενα υπερ-προσαρμογής.** Τα ποσοστά επιτυχίας είναι μεγαλύτερα από 79 % και 52 % για την ομάδα εκπαίδευσης και ελέγχου αντίστοιχα.

Τα δίκτυα RBF (σχ 7.4) ήταν μάλλον απογοητευτικά και δεν φαίνεται να υπακούουν στον κανόνα της παρεμβολής. Έτσι, **παρουσιάζουν έντονα φαινόμενα υπερ-προσαρμογής** και δίνουν χαμηλές αποδόσεις για τις ομάδες εκπαίδευσης και ελέγχου σε πολλές περιπτώσεις. Δεν χρησιμοποίησαν όλες τις διαθέσιμες μεταβλητές και φαίνονται να χρειάζονται περισσότερες ενδιάμεσες μονάδες για να προσεγγίσουν την πραγματική συνάρτηση (σχ 7.4(β)). Στην πραγματικότητα, μια ομάδα μοντέλων με αριθμό ενδιάμεσων μονάδων > 17 (στο σχ 7.4(β) αυτά βρίσκονται δεξιά της έγχρωμης γραμμής), δίνουν χαμηλά ποσοστά για την ομάδα ελέγχου και πολύ υψηλά για την ομάδα εκπαίδευσης. Τα μοντέλα αυτά παρουσιάζουν έντονα φαινόμενα υπερ-προσαρμογής. Τα υπόλοιπα μοντέλα RBF έχουν 1 – 14 ενδιάμεσες μονάδες, υπακούουν στον κανόνα της παρεμβολής, αλλά δίνουν πολύ χαμηλά ποσοστά.



Σχήμα 7.5: Αρχιτεκτονική δομή του τελικού δικτύου MLP (7:7-12-4:1).

Το καλύτερο μοντέλο αναδείχθηκε το MLP με μια ενδιάμεση στιβάδα και Α/Α 46 (σχ. 7.3(α)): επέτυχε ακρίβεια 73,5 και 72,9 % για τις ομάδες εκπαίδευσης και ελέγχου αντίστοιχα. Σχηματική παράσταση φαίνεται στο σχήμα 7.5.

7.3.6. Εφαρμογή των ΑΝΝ ($2^{η}$ προσέγγιση)

Οι ίδιες επτά αρχικές μεταβλητές επιλέχτηκαν για τη μελέτη που περιγράφεται και σε αυτήν την παράγραφο. Τρεις δυνατότητες αρχιτεκτονικών ΑΝΝ δοκιμάστηκαν και εδώ:

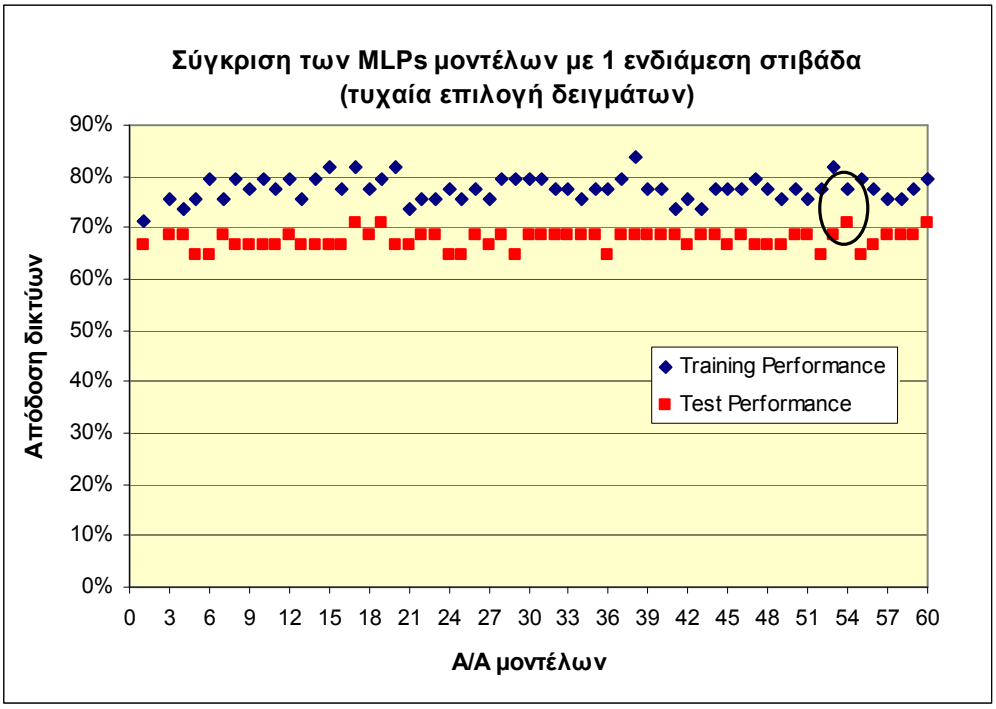
1. μοντέλα MLP με μια ενδιάμεση στιβάδα,
2. μοντέλα MLP με δυο ενδιάμεσες στιβάδες,
3. RBF με μια ενδιάμεση στιβάδα.

Η σύγκριση τους έγινε με βάση την απόδοση στις ομάδες εκπαίδευσης και ελέγχου. Τα διαγράμματα σύγκρισης των αποδόσεων απεικονίζονται παρακάτω (σχ. 7.6, 7.7).

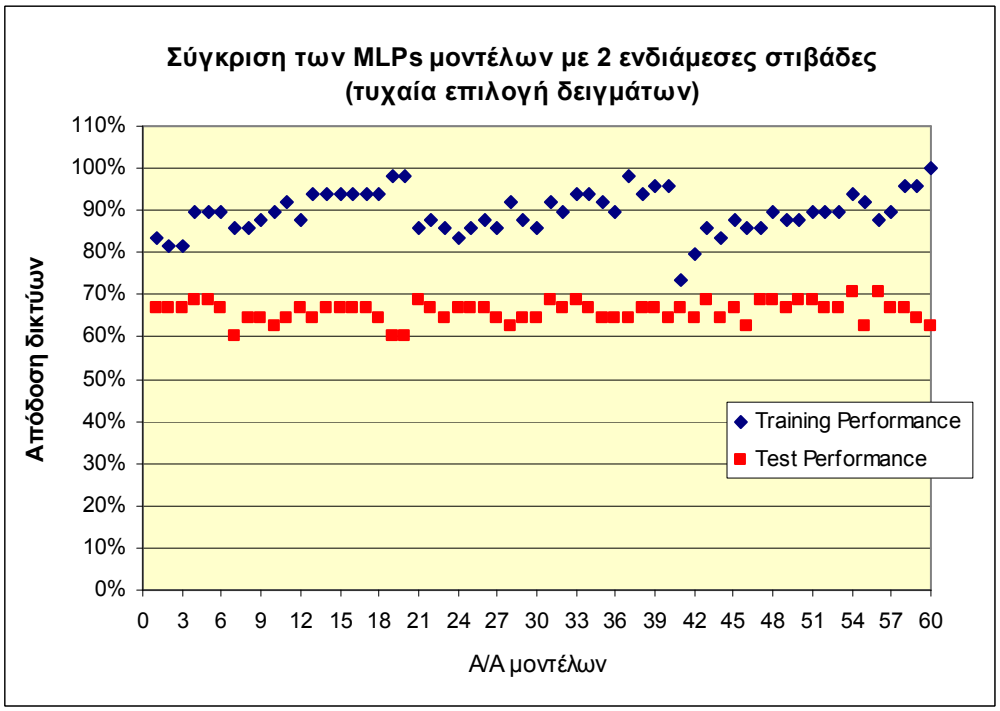
Τα μοντέλα MLP με μια (σχ. 7.6(α)) ή δυο (σχ. 7.6(β)) ενδιάμεσες στιβάδες, χρησιμοποίησαν όλες τις διαθέσιμες μεταβλητές (7), με **έντονα ωστόσο φαινόμενα υπερ-προσαρμογής.** Αυτά φαίνεται να είναι εντονότερα στις πιο πολύπλοκες δομές (μοντέλα με δυο ενδιάμεσες στιβάδες και πολλές μονάδες).

Τα μοντέλα RBF (σχ 7.7) φαίνονται να παρουσιάζουν τα εντονότερα φαινόμενα υπερ-προσαρμογής (ειδικότερα εκείνα με πάνω από 10 ενδιάμεσες μονάδες). Τα πιο απλά μοντέλα δεν επιτυγχάνουν υψηλά ποσοστά.

Τελικά από την ανάλυση αυτή ($2^{η}$ προσέγγιση), το καλύτερο μοντέλο αναδείχθηκε το MLP με μια ενδιάμεση στιβάδα και Α/Α 54 (σχ. 7.6(α)): επέτυχε ακρίβεια 77,6 και 70,8 % για τις ομάδες εκπαίδευσης και ελέγχου αντίστοιχα.

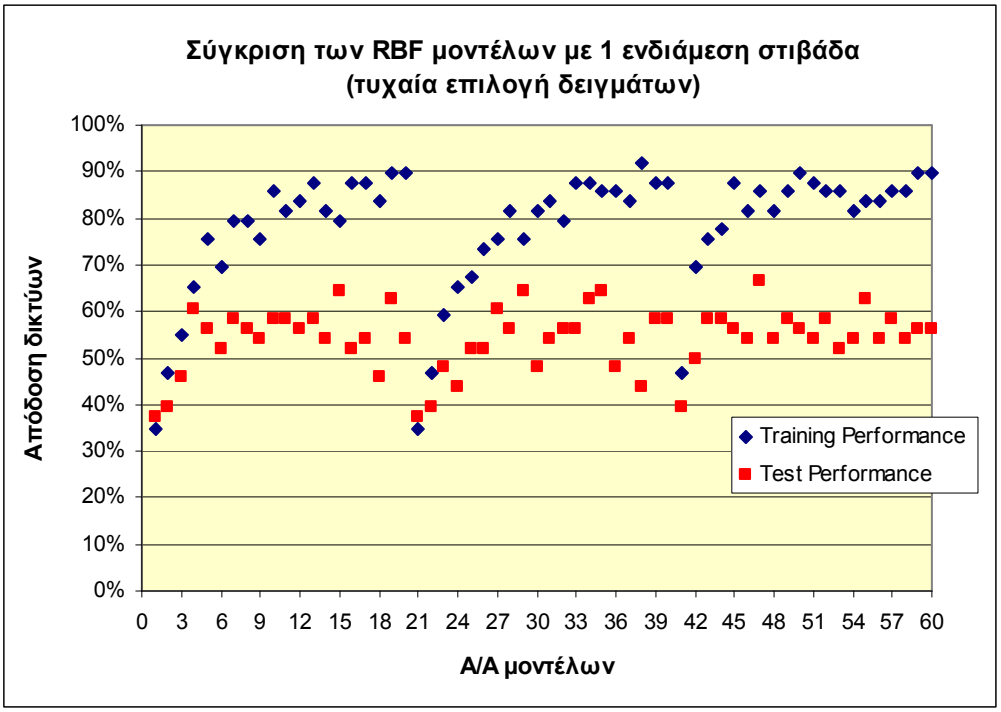


(α)

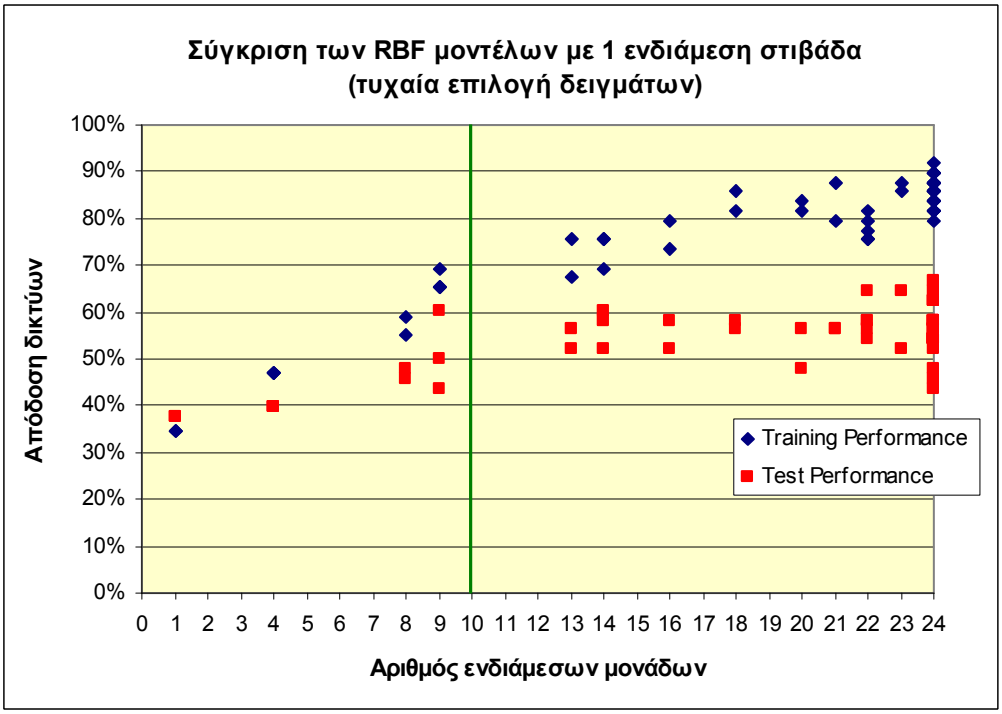


(β)

Σχήμα 7.6: Διάγραμμα σύγκρισης των αποδόσεων των ομάδων εκπαίδευσης και ελέγχου για τα μοντέλα MLP με μια (α) ή δυο (β) ενδιάμεσες στιβάδες ($2^{η}$ προσέγγιση).



(α)



(β)

Σχήμα 7.7: Διάγραμμα σύγκρισης των αποδόσεων των ομάδων εκπαίδευσης και ελέγχου για τα μοντέλα RBF. Στον οριζόντιο άξονα απεικονίζεται (α) ο Α/Α των μοντέλων ή (β) ο αριθμός των μονάδων της ενδιάμεσης στιβάδας ($2^η$ προσέγγιση).

Οι διαφορές στα παραπάνω διαγράμματα (σχ. 7.6, 7.7) σε σχέση με τα αντίστοιχα που αφορούσαν την $1^η$ προσέγγιση (σχ. 7.3, 7.4) εστιάζονται στον κανόνα της παρεμβολής που συ-

ζητήθηκε παραπάνω. Στην 2η προσέγγιση η απόδοση των δειγμάτων εκπαίδευσης (training performance) ήταν γενικά πολύ υψηλότερη από την αντίστοιχη για την ομάδα ελέγχου (test performance). Αυτό σημαίνει ότι εμφανίστηκαν **φαινόμενα υπερ-προσαρμογής, γεγονός που δικαιολογείται και από την απουσία της ομάδας επικύρωσης [48].** Ωστόσο, ακόμα και σε αυτήν την περίπτωση, τα MLP μοντέλα της 1ης προσέγγισης, δεν εμφάνισαν τόσο έντονα φαινόμενα υπερ-προσαρμογής.

7.3.7. Σύγκριση μεθόδων - Αποτελέσματα

Η σύγκριση των παραπάνω μεθόδων ANN έγινε με βάση τα ποσοστά επιτυχίας σε ομάδες εκπαίδευσης και ελέγχου.

Τα αποτελέσματα ταξινόμησης για τις δυο προσεγγίσεις (χρήση των 7 πιο κρίσιμων μεταβλητών και συνειδητή επιλογή δειγμάτων εκπαίδευσης/ελέγχου ή τυχαία επιλογή αυτών) φαίνονται αναλυτικά στον παρακάτω πίνακα 7.8.

Ο παραπάνω πίνακας περιέχει μόνο τα βέλτιστα μοντέλα από τις δυο προσεγγίσεις. Ωστόσο, αντανακλά τη γενικότερη τάση των μοντέλων για όλες τις προσπάθειες που έγιναν.

Πίνακας 7.8: Σύγκριση των βέλτιστων ANN μοντέλων

MLP μοντέλα με 1 ενδιάμεση στιβάδα			
1η προσέγγιση 7:7-12-4:1		**2η προσέγγιση 7:7-11-4:1**	
Ομάδα εκπαίδευσης	**Ομάδα ελέγχου**	**Ομάδα εκπαίδευσης**	**Ομάδα ελέγχου**
73,5	72,9	77,6	70,8

Έτσι, από τη σύγκριση των δυο προσεγγίσεων που χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων ANN, η πρώτη αναδεικνύεται σαφέστατα καλύτερη: ακόμα και αν το ποσοστό επιτυχίας για την ομάδα εκπαίδευσης είναι χαμηλότερο, αποφεύχθησαν φαινόμενα υπερ-προσαρμογής (βλ. συγκριτικά σχ. 7.3(α), 7.6(α)).

Τέλος, σε σχέση με τα αντίστοιχα μοντέλα CT (πίνακες 7.6, 7.7) τα ANN αναδείχθηκαν σαφώς ανώτερα και πιο αποτελεσματικά, ακόμα και σε δύσκολες βάσεις δεδομένων.

ΚΕΦ. 8 ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία, επιχειρήθηκε η σύγκριση παραδοσιακών χημειομετρικών τεχνικών όπως DA (Discriminant Analysis), PCA (Principal Components Analysis), CA (Cluster Analysis) με νέες πολυπαραμετρικές μεθόδους όπως τα Νευρωνικά δίκτυα (ANN) και τα Δέντρα ταξινόμησης (CT). Το πεδίο σύγκρισης ήταν τρεις μεγάλες βάσεις δεδομένων που προέκυψαν από τον προσδιορισμό:

1. Μετάλλων-μεταλλοειδών στους τρεις ταμιευτήρες που χρησιμοποιούνται για την ύδρευση της πρωτεύουσας (Υλίκη, Μόρνο και Μαραθώνα).
2. Μετάλλων-μεταλλοειδών, ανόργανων στοιχείων σε δείγματα ιζημάτων από τρεις ιχθυοκαλλιέργειες της χώρας (Ναύπακτο, Χίο και Αστακό).
3. Σπανίων γαιών σε δείγματα ελαιολάδων από τέσσερις περιοχές της Ελλάδας.

Στην πρώτη βάση δεδομένων, από τις τρεις διαφορετικές μεθόδους της τεχνικής CT, η Discriminant-based linear combination method (LCM) έδωσε τα μεγαλύτερα ποσοστά επιτυχίας στην πρόβλεψη της ομάδας εκπαίδευσης των δειγμάτων. Συγκρινόμενη δε και με τις τρεις προσεγγίσεις της παραδοσιακής τεχνικής DA (standard, forward και backward), η μέθοδος αυτή, έδωσε υψηλοτέρα ποσοστά. Αντίθετα οι μέθοδοι των Discriminant-based univariate (Classic CT) και CART (μονοπαραμετρικοί διαχωρισμοί) έδωσαν παρόμοια αποτελέσματα με την DA. Για τα "άγνωστα" δείγματα της ομάδας ελέγχου (που αφορούσαν μεταγενέστερη δειγματοληψία από τα ίδια σημεία), η Classic CT μέθοδος έδωσε τα καλύτερα αποτελέσματα.

Η αξιολόγηση των μεταβλητών από τις μεθόδους CT, ανέδειξε τις παραμέτρους V, Ni και As ως τις πιο κρίσιμες για το διαχωρισμό των ομάδων, επιβεβαιώνοντας τα αποτελέσματα της DA. Οι μέθοδοι PCA και CA δεν διαθέτουν μηχανισμούς ανάδειξης των κρίσιμων μεταβλητών.

Παράλληλα, διαφορετικά μοντέλα ANN χρησιμοποιήθηκαν για την ανίχνευση και επιβεβαίωση ομάδων. Τα αποτελέσματα των δοκιμών αυτών (σε επίπεδο θέσεων δειγματοληψίας αλλά και μεταβλητών) συγκρίθηκαν με τις "παραδοσιακές" DA, PCA, CA αλλά και την τεχνική των CT. Τα αποτελέσματα των βέλτιστων μοντέλων ANN ήταν συγκρίσιμα ή καλύτερα των άλλων μεθόδων. Τα υψηλά ποσοστά στη διάκριση των ομάδων επιβεβαιώθηκαν για τα παλαιότερα δείγματα (11/2006 – 4/2007), αλλά και σε νεώτερα (12/2007), παρόλο τις πιθανές αλλαγές σύστασης που αυτά είχαν υποστεί. Εδώ πρέπει να τονιστεί ότι οι μέθοδοι PCA και CA δεν έχουν δυνατότητα μοντελισμού.

Οι μεταβλητές που χρησιμοποιήθηκαν στα παραπάνω μοντέλα ήταν μόνο τρεις (αναδεικνυόμενες από τις προσεγγίσεις DA). Επέτυχαν τα υψηλότερα ποσοστά στην ομαδοποίηση/ταξινόμηση των δειγμάτων και επιβεβαίωσαν με αυτόν τον τρόπο την ιδιαίτερη σημασία τους.

Συνοπτικά, οι τεχνικές ANN υπερείχαν των υπολοίπων, εκπληρώνοντας τρεις βασικούς στόχους:

1. Επιτυχή ομαδοποίηση των θέσεων δειγματοληψίας.
2. Επιτυχή ταξινόμηση νέων δειγμάτων και επομένως δημιουργία ενός ανθεκτικού και με ακρίβεια μοντέλου.
3. Επιτυχή αξιολόγηση των κρισίμων μεταβλητών. Οι διαδικασίες βελτιστοποίησης των μοντέλων ANN αξιολογούν τις μεταβλητές και απορρίπτουν αυτές που βρίσκονται σε "περίσσεια".

Ιδιαίτερα εντυπωσιακή υπήρξε η τεχνική Kohonen, η οποία παρουσιάζει δυο δυνατότητες: λειτουργεί ταυτόχρονα ως κλασική μη επιβλεπόμενη τεχνική (επιτυγχάνοντας υψηλά ποσοστά ταξινόμησης), αλλά επιπλέον διαθέτει δυνατότητα μοντελισμού και ανάδειξης των κρίσιμων μεταβλητών (υπερέχοντας για παράδειγμα της κλασικής CA).

Για τη δεύτερη βάση δεδομένων (μέταλλα-μεταλλοειδή, ανόργανα στοιχεία σε δείγματα ιζημάτων από τρεις ιχθυοκαλλιέργειες της χώρας), οι απαιτήσεις αφορούσαν όχι μόνο τη διάκριση των σημείων, τα οποία βρίσκονταν κοντά στον κλωβό εκτροφής των ψαριών ή μακρύτερα από αυτόν, αλλά και μια περιβαλλοντική μελέτη των αποτελεσμάτων που περιγράφεται αναλυτικά στο ηλεκτρονικό παράρτημα αυτής της διατριβής (ΚΕΦ. 2, Π).

Χρησιμοποιήθηκαν κλασικές τεχνικές (KNN, DA), αλλά και CT, ANN. Η κύρια μεταβλητή που κυριάρχησε σε όλους τους διαχωρισμούς μεταξύ των σημείων που βρίσκονταν κοντά ή μακριά από τους κλωβούς, ήταν ο P. Κρίσιμες αναδείχθηκαν επίσης οι μεταβλητές N, Cu, Cd, C, Zn και λιγότερο τα Mn και Pb. Όλα τα μοντέλα (DA, CT, ANN) με τη χρήση μίας μεταβλητής μπόρεσαν να δώσουν ιδιαίτερα υψηλά ποσοστά προβλέψεων. Τα ANN με μικρότερη ομάδα εκπαίδευσης, έδωσαν εφάμιλλα των άλλων μεθόδων αποτελέσματα. Ακόμα και το γραμμικό μοντέλο ANN με μία μόνο μεταβλητή και ένα απλό αλγόριθμο έδωσε υψηλά ποσοστά.

Γενικότερο συμπέρασμα αποτελεί η επιβάρυνση του θαλάσσιου πυθμένα με ανόργανα θρεπτικά συστατικά που περιέχονται στις τεχνητές τροφές με τις οποίες εκτρέφονται τα ψάρια. Το ίδιο επισημαίνεται για τα μέταλλα Cd (ιδιαίτερα τοξικό) και Zn. Τεχνητά παρασκευασμένες ιχθυοτροφές /ιχθυάλευρα και συμπληρώματα φαίνεται να επιβαρύνουν το θαλάσσιο πυθμένα, με ανόργανα και μεταλλικά στοιχεία [245] που ανακυκλώνονται και βιοσυσσωρεύονται μεταφέροντας το πρόβλημα.

Για την τρίτη βάση δεδομένων, χρησιμοποιήθηκαν δυο προσεγγίσεις για την ταξινόμηση των ελαιολάδων με βάση τη γεωγραφική τους προέλευση:

- ✓ συνειδητή/προσεχτική επιλογή των ομάδων εκπαίδευσης και ελέγχου με βάση τις DA roots των δεδομένων και,

- ✓ τυχαία επιλογή των αντίστοιχων ομάδων των δειγμάτων.

Επιβλεπόμενες τεχνικές όπως CT (3 μέθοδοι) και ANN (μοντέλα MLP, RBF) χρησιμοποιήθηκαν και στις δυο προσεγγίσεις. Η χρήση της πρώτης προσέγγισης έδωσε γενικά καλά ποσοστά ακρίβειας στις ομάδες εκπαίδευσης και ελέγχου. Αντίθετα, στη δεύτερη κλασική προσέγγιση παρατηρήθηκαν γενικά φαινόμενα υπερ-προσαρμογής των μοντέλων, προφανώς λόγω του μεγάλου αριθμού μεταβλητών (10) αναλογικά με τον αριθμό των δειγμάτων (97) και τις ομάδες ταξινόμησης (4). Κατά την αξιολόγηση των αρχικών μεταβλητών κατά την οποία δεν υπήρχε μεγάλη ομοφωνία των τεχνικών που χρησιμοποιήθηκαν, τα Gd, Sm, Tm, Dy Pr, αναδείχθηκαν οι κρισιμότερες από αυτές.

Ειδικότερα, τα ANN έδωσαν τα καλύτερα αποτελέσματα συγκρινόμενα με άλλες επιβλεπόμενες τεχνικές. Το βέλτιστο μοντέλο επέτυχε ακρίβεια 73,5 και 72,9 % για τις ομάδες εκπαίδευσης και ελέγχου αντίστοιχα, όταν DA και CT (LCM) δίνουν 76,3 % (για την ομάδα εκπαίδευσης) και 81,6 / 60,4 % (τις ομάδες εκπαίδευσης και ελέγχου) αντίστοιχα.

Καταλήγοντας, οι ANN αρχιτεκτονικές και αλγόριθμοι φάνηκε να έχουν την ευελιξία αλλά και την αυτονομία να επιτύχουν σε προβλήματα ομαδοποίησης, ταξινόμησης, μοντελισμού περισσότερο από κάθε άλλη τεχνική. Η εφαρμογή τους σε περισσότερες βάσεις δεδομένων θα το αποδείξει.

Στο σημείο αυτό ωστόσο, πρέπει να τονιστούν ορισμένα κρίσιμα σημεία σχετικά με την ορθή χρήση των παραδοσιακών στατιστικών τεχνικών και τη σύγκριση τους με τις ANN αρχιτεκτονικές:

1. Στη βιβλιογραφία αναφέρονται περιπτώσεις όπου τα ANN αποδίδουν χειρότερα από τις γραμμικές παραδοσιακές τεχνικές. Ο λόγος εδώ είναι ότι οι συσχετίσεις είναι πραγματικά γραμμικές χωρίς μεγάλες αποκλίσεις. Έτσι δεν μπορούμε να αναμένουμε από τα νευρωνικά δίκτυα να "λειτουργούν" καλύτερα από τα γραμμικά μοντέλα σε γραμμικά συσχετιζόμενα δεδομένα. Με άλλα λόγια, δεν χρησιμοποιείται η κατάλληλη συνάρτηση για να περιγράψει τα δεδομένα [69].
2. Τα μοντέλα ANN στην παρούσα εργασία "δοκιμάστηκαν" σε νεώτερα ή εξωτερικά δείγματα που δεν μετείχαν στη διαμόρφωσή τους. Αυτό έγινε και στις παραδοσιακές μεθόδους (όπως για παράδειγμα με τη χρήση της CV ή ανεξάρτητης ομάδας ελέγχου). Δυστυχώς όμως, στις παραδοσιακές τεχνικές η επικύρωση δεν αποτελεί "default" διαδικασία, με αποτέλεσμα η απουσία τέτοιας επικύρωσης/αξιολόγησης τελικά να ευνοεί τα αποτελέσματα των τελευταίων.
3. Από την άλλη πλευρά, η σημασία των παραδοχών για τις παλαιότερες στατιστικές μεθόδους είναι μεγάλη, αλλά το γεγονός αυτό δεν φαίνεται γενικά να λαμβάνεται σοβαρά υπόψη από συγγραφείς και ερευνητές σε όλο τον κόσμο [176]. Για παράδειγμα,

σύμφωνα με τον Ainsworth [265], μόνο αν υπάρχουν πάνω από 20 δείγματα σε κάθε ομάδα η DA μπορεί να θεωρηθεί "ανθεκτική" σε περιπτώσεις μη κανονικής κατανομής. Έτσι, η σοβαρότερη θεώρηση αυτών των παραδοχών θα μπορούσε ίσως να οδηγήσει σε καλύτερα αποτελέσματα σε κάποιες από αυτές τις μεθόδους.

4. Στην εργασία αυτή αλλά και γενικότερα, τα τελικά μοντέλα ANN είναι βελτιστοποιημένα. Αντίθετα τέτοιες προσπάθειες σπάνια γίνονται για τις παλαιότερες στατιστικές μεθόδους [176].
5. Οι παραδοσιακές στατιστικές μέθοδοι στηρίζονται γενικά σε γραμμικά μοντέλα. Ωστόσο, τα μη γραμμικά μοντέλα αποδεικνύονται γενικά ισχυρότερα σε αυτό-εκπαιδευόμενες τεχνικές [122]. Έτσι, η σύγκριση μεταξύ των ANN και άλλων πιο συμβατικών τεχνικών ανάγεται τελικά σε αντιπαράθεση μεταξύ μη γραμμικών και γραμμικών μοντέλων, η οποία δεν θα μπορούσε με κανένα τρόπο να ευνοήσει τα τελευταία.
6. Όπως τα γραμμικά μοντέλα, έτσι και τα ANN έχουν αδυναμίες αλλά και μεγάλες δυνατότητες. Έχουν χαρακτηριστικά που τα κάνουν κατάλληλα για μια μεγάλη ποικιλία προβλημάτων, αλλά σίγουρα δεν αποτελούν πανάκεια λύση. Πιστεύεται ότι γενικά είναι κατάλληλα για μεγάλες βάσεις δεδομένων με μη γραμμική δομή [69].

Εξάλλου, δεν υπάρχει "πανάκεια" μέθοδος ικανή να αντιμετωπίσει όλους τους τύπους των προβλημάτων. Συμπληρωματικές πληροφορίες μπορεί να "αντλούνται" συνολικά από παλαιότερες και σύγχρονες μεθόδους όπως τα ANN. Τα τελευταία, (οι πιο σύνθετες από μαθηματικής άποψης τεχνικές), δεν είναι πάντα η καλύτερη επιλογή, καθώς για παράδειγμα απλές προσεγγίσεις όπως η KNN δεν θα πρέπει να απορρίπτονται a priori: το απλό δεν είναι πάντα κακό [5]. Έτσι, τα ANN θα πρέπει να θεωρούνται εναλλακτικές τεχνικές στις πιο παραδοσιακές και όχι ως η λύση σε προηγούμενες αποτυχίες ή διαφορετικά, "όταν η σκόνη έχει κατακάτσει, συνήθως ανακαλύπτουμε ότι η νέα τεχνική δεν είναι μια θαυματουργή γιατρειά για όλα, ούτε μια ολοκληρωτική καταστροφή, αλλά ένα νέο εφόδιο στην εργαλειοθήκη μας το οποίο δουλεύει καλά σε κάποιες περιπτώσεις και άλλες όχι" [76, 266]. Δεν υπάρχει μέθοδος, που θα μπορούσαμε να ισχυριστούμε ότι είναι ανώτερη στην ομαδοποίηση, ταξινόμηση ή "μοντελοποίηση". Σε κάθε εφαρμογή, η επιλογή της καλύτερης μεθόδου, δεν μπορεί να επιτευχθεί μέσω εικασιών ή υποθέσεων, χωρίς πρότερη έρευνα, μελέτη και κατεργασία των δεδομένων. Έτσι οι δυνητικοί χρήστες των ANN, θα πρέπει πάντα να προσφεύγουν και στις παλαιότερες μεθόδους πριν αποκτήσουν επαρκή γνώση των δεδομένων. Επιπλέον, πρέπει ο καθένας να μην ενθουσιάζεται με οποιοδήποτε νέο "εργαλείο", μόνο από το γεγονός ότι απλά είναι νέο. Οποιαδήποτε μέθοδος, όσο και αν φαίνεται αποτελεσματική, μπορεί πολύ εύκολα να αποτύχει αν τα δεδομένα δεν συσχετίζονται ή δεν αντιπροσωπεύουν επαρκώς την πληροφορία που "κυνηγούμε", ή ο χρήστης δεν ξέρει ακριβώς τι ζητά ή δεν έχουν δοκιμαστεί και άλλες μέθοδοι. Η χρήση των ANN, όπως

και των άλλων μεθόδων, απαιτεί καλή θεωρητική γνώση και σοβαρή πειραματική εργασία, ώστε να εξαχθεί πραγματικά η πληροφορία που “κρύβουν” τα δεδομένα [79].

Τελικά, η εφαρμογή στατιστικών μεθόδων σε μεγάλες ή μικρές βάσεις δεδομένων φαίνεται να απαιτεί πολλαπλές δοκιμές, εμπειρία σε παρόμοιες διαδικασίες, χρήση των κατάλληλων κριτηρίων σύγκρισης, υπεύθυνη και αντικειμενική ερμηνεία των αποτελεσμάτων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Zupan, J., Gasteiger, J., Neural Networks for Chemists; An Introduction. VCH Verlagsgescllschaft, Weinhcim, Germany and VCH Publishers, New York, USA, 1993.
2. Adams, M.J., Chemometrics in Analytical spectroscopy. RSC Analytical Spectroscopy monographs, 2nd ed., Neil W. Barnett, The Royal Society of Chemistry, U.K., 2004.
3. Astel. A., Glosińka, G., Sobczyński, T., Boszke, L., Simeonov, V., Siepak, J., Chemometrics in the assessment of the sustainable development rule implementation, Cent. Eur. J. Chem., 2006, **4(3),** 543-564.
4. Smith, L.I., A tutorial on Principal Components analysis, February 26, 2002, http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf (τελευταία επίσκεψη 5/2/2010).
5. Fdez-Ortiz de Vallejuelo, S., Arana, G., de Diego, A., Madariaga, J.M., Pattern recognition and classification of sediments according to their metal content using chemometric tools. A case study: The estuary of Nerbioi-Ibaizabal River (Bilbao, Basque Country), Chemosphere, 2011, **85,** 1347-1352.
6. Kavzoglu, T., Increasing the accuracy of neural network classification using refined training data, Environ. Modell. Softw., 2009, **24,** 850-858.
7. López, A., García, P., Garrido, A., Multivariate characterization of table olives according to their mineral nutrient composition, Food Chem., 2008, **106,** 369-378.
8. Miller, J.N., and Miller, J.C., Statistics and Chemometrics for Analytical Chemistry, 5th ed., Pearson, England, 2005.
9. Friel, C.M., Notes on Discriminat Analysis, Criminal Justice Center, Sam Houston State University.
10. Ντζούφρας, Ι., Σημειώσεις για το μάθημα Ανάλυση Δεδομένων Ι, Στοιχεία Πολυμεταβλητής Ανάλυσης Δεδομένων, Τμήμα Διοίκησης Επιχειρήσεων, Πανεπιστήμιο Αιγαίου, 2001.
11. Chan, Y.H., Biostatistics 303. Discriminant analysis, Singapore Med. J., 2005, **46(2),** 54-62.
12. http://127.0.0.1:49312/help/index.jsp?topic=/com.ibm.spss.statistics.cs/discriminant_table.htm (τελευταία επίσκεψη 15/12/2010)
13. Marini, F., Magrì, A.L., Marini, D., Balestrieri, F., Characterization of the lipid fraction of Niger seeds (Guizotia abyssinica cass.) from different regions of Ethiopia and India and chemometric authentication of their geographical origin, Eur. J. Lipid Sci. Technol., 2003, **105,** 697-704.

14. Marini, F., Balestrieri, F., Bucci, R., Magrì, A.L., Marini, D., Supervised pattern recognition to discriminate the geographical origin of rice bran oils: a first study, Microchem. J., 2003, **74,** 239-248.
15. Moreno, I.M., González-Weller, D., Gutierrez, V., Marino, M., Cameán, A.M., González, A.G., Hardisson, A., Differentiation of two Canary DO red wines according to their metal content from inductively coupled plasma optical emission spectrometry and graphite furnace atomic absorption spectrometry by using Probabilistic Neural Networks, Talanta, 2007, **72,** 263-268.
16. http://www.scribd.com/doc/25044857/discriminant-Function-Analysis (τελευταία επίσκεψη 12/12/2010).
17. STATISTICA 7th edition, software, StatSoft, Inc., 2004.
18. Ghasemi-Varnamkhasti, M., Mohtasebi, S.S., Siadat, M., Lozano, J., Ahmadi, H., Razavi, S.H., Dicko, A., Aging fingerprint characterization of beer using electronic nose, Sensor. Actuat. B-Chem., 2011, **159,** 51-59.
19. Hauser-Davis, R.A., Oliveira, T.F., Silveira, A.M., Silva, T.B., Ziolli, R.L., Case study: Comparing the use of nonlinear discriminating analysis and Artificial Neural Networks in the classification of three fish species: acaras (Geophagus brasiliensis), tilapias (Tilapia rendalli) and mullets (Mugil liza), Ecol. Inform., 2010, **5,** 474-478.
20. Peres, A.M., Baptista, P., Malheiro, R., Dias, L.G., Bento, A., Pereira, J.A., Chemometric classification of several olive cultivars from Trás-os-Montes region (northeast of Portugal) using artificial neural networks, Chemometr. Intell. Lab., 2011, **105,** 65-73.
21. Morrison, D.G., On the interpretation of Discriminant Analysis, J. Mark. Res., 1969, **6(2),** 156-163.
22. Beltrán, N.H., Duarte-Mermoud, M.A., Bustos, M.A., Salah, S.A., Loyola, E.A., Peña-Neira, A.I., Jalocha, J.W., Feature extraction and classification of Chilean wines, J. Food Eng., 2006, **75,** 1-10.
23. Rodríguez, R. I., Delgado, M.F., García, J.B., Crecente, R.M.P., Martín, S.G., Latorre, C.H., Comparison of several chemometric techniques for the classification of orujo distillate alcoholic samples from Galicia (northwest Spain) according to their certified brand of origin, Anal. Bioanal. Chem., 2010, **397,** 2603-2614.
24. Pérez-Magariño, S., Ortega-Heras, M., González-San José M.L., Boger, Z., Comparative study of artificial neural network and multivariate methods to classify Spanish DO rose wines, Talanta, 2004, **62,** 983-990.
25. Massart, D.L., Vandeginste, B.G.M., Deming, S.N., Michotte, Y. and Kaufman, L, Chemometrics: a textbook, Elsevier, 1988.

26. Berrueta, L.A., Alonso-Salces, R.M., Héberger, K., Supervised pattern recognition in food analysis, J. Chromatogr. A, 2007, **1158,** 196-214.
27. Kraic, F., Mocák, J., Fiket, Z., Kniewald, G., ICP MS analysis and classification of potable, spring, and mineral waters, Chem. Pap., 2008, **62,** 445-450.
28. Alsberg, B.K., Goodacre, R., Rowland, J.J., Kell, D.B., Classification of pyrolysis mass spectra by fuzzy multivariate rule induction-comparison with regression, K-nearest neighbour, neural and decision-tree methods, Anal. Chim. Acta, 1997, **348,** 389-407.
29. Chan, Y.H., Biostatistics 302. Principal component and factor analysis, Singapore Med. J., 2004, **45(12),** 558-565.
30. Spanos, T., Simeonov, V., Simeonova, P., Apostolidou, E., Stratis, J., Environmetrics to evaluate marine environment quality, Environ. Monit. Assess., 2008, **143,** 215-225.
31. Brereton, R.G., Chemometrics: Data Analysis for the Laboratory and Chemical Plant, John Wiley and Sons, Ltd, England 2003.
32. Παπαγρηγορίου Ν., Πτυχιακή εργασία, Τμήμα Τεχνολογίας Τροφίμων, Σχολής Τεχνολογίας Τροφίμων και Διατροφής, ΤΕΙ Θεσσαλονίκης, 2001.
33. Garcia, H.L., González, I.M., Self-organizing map and clustering for wastewater treatment monitoring, Artif. Intell., 2004, **17,** 215-225.
34. Vesanto, J., Alhoniemi, E., Clustering of the Self-Organizing Map, IEEE Transactions on Neural Networks, 2000, **11,** 586-600.
35. Chan, Y.H., Biostatistics 304. Cluster analysis, Singapore Med. J., 2005, **46(4),** 153-160.
36. http:// people.revoledu.com/kardi/tutorial/kMean/WhatIs.htm
(τελευταία επίσκεψη Απρίλιος 2011).
37. Simeonova, P., Simeonov V., Andreev G., Water Quality Study of the Struma River Basin, Bulgaria (1989-1998), Cent. Eur. J. Chem., 2003, **2,** 121-136.
38. Φαρμάκη, Ε., Ερευνητική εργασία διπλώματος ειδίκευσης, Τμήμα Αναλυτικής Χημείας, ΕΚΠΑ, 2007.
39. Loh, W-Y., Shih, Y-S., Split Selection Methods for Classification Trees, Stat. Sinica, 1997, **7**, 815-840.
40. Debeljak M., Džeroski S., Jerina K., Kobler A., Adamič M., Habitat suitability modelling for red deer (Cervus elaphus L.) in South-central Slovenia with classification trees, Ecol. Model., 2001, **138,** 321-330.
41. Webb, A.R., Statistical Pattern Recognition, 2nd ed., John Wiley & Sons, Chichester, 2002.
42. Cheng, W., Zhang X., Wang, K., Dai, X., Integrating classification and regression tree (CART) with GIS for assessment of heavy metals pollution, Environ. Monit. Assess., 2009, **158,** 419-431.

43. Kim, Y.S., Performance evaluation for classification methods: A comparative simulation study, Expert Syst. Appl., 2010, **37,** 2292-2306.
44. Tirelli, T., Pessani, D., Importance of feature selection in decision-tree and artificial-neural-network ecological applications. Alburnus alburnus alborella: A practical example, Ecological Informatics, 2011, **6,** 309-315.
45. Smeti, E.M., Thanasoulias, N.C., Lytras, E.S., Tzoumerkas, P.C., Golfinopoulos, S.K., Treated water quality assurance and description of distribution networks by multivariate chemometrics, Water Res., 2009, **43,** 4676-4684.
46. Simeonova, P., Simeonov, V., Chemometrics to evaluate the quality of water sources for human consumption, Microchim. Acta, 2007, **156**, 315-320.
47. Φαρμάκη Ε., Χημικά Χρονικά, Πολυστιβαδικά Νευρωνικά Δίκτυα (MLPs): η θεωρία τους, 2011, **73(4),** 12-16.
48. Farmaki, E.G., Thomaidis, N.S., Efstathiou, C.E., Artificial Neural Networks in water analysis: Theory and applications, Int. J. Environ. An. Ch., 2010, **90,** 85-105.
49. Αργυράκης Π., Νευρωνικά Δίκτυα και Εφαρμογές, Ελληνικό Ανοιχτό Πανεπιστήμιο, Πάτρα 2001.
50. http://www.mirrorservice.org/sites/home.ubalt.edu/ntsbarsh/Business-stat/opre/neurons.gif (τελευταία επίσκεψη Μάιος 2008).
51. Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., De Jong, S., Lewi, P.J., Smeyers-Verbeke, J., Handbook of Chemometrics and Qualimetrics: Part B., Elsevier, Amsterdam, 1998.
52. Zupan, J., Gasteiger, J., Neural networks: A new method for solving chemical problems or just a passing phase? Anal. Chim. Acta, 1991, **248**, 1-30.
53. Rosenblatt, F., The Perceptron: A probabilistic model for information storage and organization in the brain, Psychol. Rev., 1958, **65,** 386-408.
54. http://www.generation5.org/content/1999/perceptron.asp (τελευταία επίσκεψη Φεβρουάριος 2008).
55. Marini, F., Magrì, A.L., Bucci, R., Multilayer feed-forward artificial neural networks for class modeling, Chemometr. Intell. Lab., 2007, **88,** 118-124.
56. Marini, F., Artificial neural networks in foodstuff analyses: Trends and perspectives A review, Anal. Chim. Acta, 2009, **635**, 121-131.
57. Huang, W., Foo, S., Neural network modeling of salinity variation in Apalachicola River, Water Res., 2002, **36,** 356-362.
58. Alfassi, Z.B., Boger, Z., Ronen, Y., Statistical Treatment of Analytical Data, Blackwell Science Ltd, Oxford, UK 2005.

59. Kröse, B., Van der Smagt, P., An introduction to Neural Networks, The University of Amsterdam, Amsterdam 1996.
60. Nguyen, H.T., Prasad, N.R., Walker, C.L., Walker, E.A., A First Course in Fuzzy and Neural Control, Chapman & Hall/CRC, Boca Raton, Florida, USA 2003.
61. Darnag, R., Schmitzer, A., Belmiloud, Y., Villemin, D., Jarid, A., Chat, A., Mazouz, E., Cherqaoui, D., Quantitative structure-activity relationship studies of TIBO derivatives using support vector machines, SAR QSAR Environ. Res., 2010, **21,** 231-246.
62. Balabin, R.M., Safieva, R.Z., Lomakina, E.I., Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques, Anal. Chim. Acta, 2010, **671,** 27-35.
63. LeBouf, R.F., Schuckers, S.A., Rossner, A., Preliminary assessment of a model to predict mold contamination based on microbial volatile organic compound profiles, Sci. Total Environ., 2010, **408,** 3648-3653.
64. Kaftan, I., Salk, M., Senol, Y., Evaluation of gravity data by using artificial neural networks case study: Seferihisar geothermal area (Western Turkey), J. Appl. Geophys., 2011, **75,** 711-718.
65. http://www.sussex.ac.uk/Users/andrewop/Courses/NN/NNs5_6_MLP.ppt (τελευταία επίσκεψη Σεπτέμβριος 2010).
66. Elhatip, H., Kömür, M. A., Evaluation of water quality parameters for the Mamasin dam in Aksaray City in the central Anatolian part of Turkey by means of artificial neural networks, Environ. Geol., 2008, **53,** 1157-1164.
67. Dogan, E., Sengorur, B., Koklu, R., Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique, J. Environ. Manage., 2009, **90,** 1229-1235.
68. Carlucci, G., D' Archivio, A.A., Maggi, M. A., Mazzeo, P., Ruggieri, F., Investigation of retention behaviour of non-steroidal anti-inflammatory drugs in high-erformance liquid chromatography by using quantitative structure–retention relationships, Anal. Chim. Acta, 2007, **601,** 68-76.
69. Zhang, G., Patuwo, B.E., Hu, M.Y., Forecasting with artificial neural networks: The state of the art, Int. J. Forecasting, 1998, **14,** 35-62.
70. Fernández-Sánchez, J.F., Carretero, A.S., Benítez-Sánchez, J.M., Cruses-Balnco, C., Fernández-Gutiérrez, A., Fluorescence optosensor using an artificial neural network for screening of polycyclic aromatic hydrocarbons, Anal. Chim. Acta, 2004, **510,** 183-187.
71. Aoyama, T., Suzuki, Y., Ichikawa, H., Neural networks applied to quantitative structure-activity relationship analysis, J. Med. Chem., 1990, **33,** 2583-2590.

72. Rene, E. R., Saidutta, M. B., Prediction of water quality indices by regression analysis and artificial neural networks, Int. J. Environ. Res., 2008, **2(2),** 183-188.
73. Fausett, L., Fundamentals of Neural Networks–Architectures, Algorithms and Applications, Prentice-Hall, Inc. Upper Saddle River, NJ, USA 1994.
74. Zupan, J., Novič, M., Li, X., Gasteiger, J., Classification of multicomponent analytical data of olive oils using different neural networks, Anal. Chim. Acta, 1994, **292,** 219-234.
75. Panagou, E. Z., Mohareb, F. R., Argyri, A. A., Bessant, C. M., Nychas, G-J. E., A comparison of artificial neural networks and partial least squares modelling for the rapid detection of the microbial spoilage of beef fillets based on Fourier transform infrared spectral fingerprints, Food Microbiol., 2011, **28,** 782-790.
76. Maier, H.R., Dandy, G.C. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications, Environ. Modell. Softw., 2000, **15,** 101-124.
77. Svozil, D., Kvasnička, V., Pospíchal, J., Introduction to multi-layer feed-forward neural networks, Chemometr. Intell. Lab., 1997, **39,** 43-62.
78. Marini, F, Bucci, R., Magrì, A. L., Magrì, A. D., Artificial neural networks in chemometrics: History, examples and perspectives, Microchem. J., 2008, **88,** 178-185.
79. Zupan, J., Novič, M., Ruisánchez, I., Kohonen and counterpropagation artificial neural networks in analytical chemistry, Chemometr. Intell. Lab., 1997, **38,** 1-23.
80. Curteanu, S., Cartwright, H., Neural networks applied in chemistry. I. Determination of the optimal topology of multilayer perceptron neural networks, J. Chemometr., 2011, **25,** 527-549.
81. Román, R.C., Hernández, O.G., Urtubia, U.A., Prediction of problematic wine fermentations using artificial neural networks, Bioproc. Biosyst. Eng., 2011, **34,** 1057-1065.
82. http://www.chem.uoa.gr/applets/AppletPoly/Appl_Poly1.html
(τελευταία επίσκεψη 3/2012).
83. Ni, Y., Ping, Q., Kokot, S., Simultaneous voltammetric determination of four carbamate pesticides with the use of chemometrics, Anal. Chim. Acta, 2005, **537,** 321-330.
84. Tagluk, M.E., Akin, M., Sezgin, N., Classıfication of sleep apnea by using wavelet transform and artificial neural networks, Expert Syst. Appl., 2010, **37,** 1600-1607.
85. Fernandes, F.A.N., Lona, L.M.F., Neural network application in polymerization processes, Braz. J. Chem. Eng. 2005, **22**, 401–418.
86. Lancashire, L.J., Lemetre, C., Ball, G.R., An introduction to artificial neural networks in bioinformatics application to complex microarray and mass spectrometry datasets in cancer studies, Briefings in Bioinformatics, 2009, **10,** 315-329.

87. Zur, R.M., Jiang, Y., Pesce, L.L.,Drukker, K., Noise injection for training artificial neural networks: A comparison with weight decay and early stopping, Med. Phys., 2009, **36,** 4810-4818.
88. Zhang, Y. X., Artificial neural networks based on principal component analysis input selection for clinical pattern recognition analysis, Talanta, 2007, **73,** 68-75.
89. Mutihac, L., Mutihac, R., Mining in chemometrics, Anal. Chim. Acta, 2008, **612,** 1-18.
90. Ramil, A., López, A. J., Yáñez, A., Application of artificial neural networks for the rapid classification of archaeological ceramics by means of laser induced breakdown spectroscopy (LIBS), Appl. Phys. A, 2008, **92,** 197-202.
91. Palani, S., Liong, S-Y., Tkalich, P., An ANN application for water quality forecasting, Mar. Pollut. Bull., 2008, **56,** 1586-1597.
92. Stathakis, D., How many hidden layers and nodes? Int. J. Remote Sens., 2009, **30**, 2133-2147.
93. Jutras, P., Prasher, S.O., Mehuys, G.R., Prediction of street tree morphological parameters using artificial neural networks, Comput. Electron. Agr. 2009, **67,** 9-17.
94. Tirelli, T., Pessani, D., Use of Decision tree and Artificial neural network approaches to model presence/absence of Teleste Muticellus in Piedmont (north-western Italy), River Res. Appl., 2009, **25,** 1001-1012.
95. Sietsma, J., Dow, R.J.F., Creating artificial neural networks that generalize, Neural Networks, 1991, **4,** 67-79.
96. Grzesiak, W., Zaborski, D., Sablik, P., Zukiewicz, A., Dybus, A., Szatkowska, I., Detection of cows with insemination problems using selected classification models, Comp. Electron. Agr. 2010, **74,** 265-273.
97. Xu, Y., Zomer, S., Brereton. R., Support Vector Machines: A Recent Method for Classification in Chemometrics, Crit. Rev. Anal. Chem., 2006, **36,** 177-188.
98. Hecht-Nielsen, R., Kolmogorov's mapping neural network existence theorem. In IEEE First Annual International Conference on Neural Networks, 1987, **3,** 11-13.
99. Hornik, K., Stinchcombe, M., White, H., Multilayer Feedforward Networks are Universal Approximators, Neural Networks, 1989, **2,** 359-366.
100. Kalteh, A.M., Hjorth, P., Berndtsson, R., Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application, Environ. Modell. Softw., 2008, **23,** 835-845.
101. Kutlu, Y., Kuntalp, M., Kuntalp, D., Optimizing the performance of an MLP classifier for the automatic detection of epileptic spikes, Expert Syst., Appl., 2009, **36,** 7567-7575.

102. Marini, F. Magrì, A.L., Bucci, R., Magrì, A.D., Use of different artificial neural networks to resolve binary blends of monocultivar Italian olive oils, Anal. Chim. Acta, 2007, **599,** 232-240.
103. Marini, F. Balestieri F., Bucci, R., Magrì, A.D., Magrì, A.L., Marini, D., Supervised pattern recognition to authenticate Italian extra virgin olive oil varieties, Chemometr. Intell. Lab., 2004, **73,** 85-93.
104. Gramatica, P., Principles of QSAR models validation: internal and external, QSAR Comb. Sci., 2007, **26,** 694-701.
105. García-González, D.L., Luna, G., Morales, M.T., Aparicio. R., Stepwise geographical traceability of virgin olive oils by chemical profiles using artificial neural network models, Eur. J. Lipid Sci. Tech.., 2009, **111**, 1003-1013.
106. Logeswaran, R., J. Cholangiocarcinoma—An Automated Preliminary Detection System Using MLP, Med. Syst., 2009, **33,** 413-421.
107. Balabin, R.M., Lomakina, E.I., Neural network approach to quantum-chemistry data: Accurate prediction of density functional theory energies, J. Chem., Phys., **131,** 074104.
108. Marini, F., Bucci, R., Magrì, A. L., Magrì, A. D., Acquistucci, R., Francisci R., Classification of 6 durum wheat cultivars from Sicily (Italy) using artificial neural networks, Chemometr. Intell. Lab., 2008, **90,** 1-7.
109. Rao, H., Yang, G., Tan, N., Li, P., Li, Z., Li, X., Prediction of HIV-1 Protease Inhibitors Using Machine Learning Approaches, QSAR Comb. Sci., 2009, **28,** 1346-1357.
110. Marini, F., Zupan, J., Magrì, A.L., On the use of counterpropagation artificial neural networks to characterize Italian rice varieties, Anal. Chim. Acta, 2004, **510,** 231-240.
111. Galvão, R.K.H., Araujo, M.C.U., José, G.E., Pontes, M.J.C., Silva, E.C., Saldanha, T.C.B., A method for calibration and validation subset partitioning, Talanta, 2005, **67,** 736-740.
112. Concu, R., Podda, G., Uriarte, E., González-Díaz, H., Computational Chemistry Study of 3D-Structure-Function Relationships for Enzymes Based on Markov Models for Protein Electrostatic, HINT, and van der Waals Potentials, J. Comput. Chem., 2009, **30,** 1510-1520.
113. Lee, S.C., Lin, H.T., Yang, T.Y., Artificial neural network analysis for reliability prediction of regional runoff utilization, Environ. Monit. Assess., 2010, **161,** 315-326.
114. Purkait, B., Kadam, S.S., Das, S.K., Application of Artificial Neural Network Model to Study Arsenic Contamination in Groundwater of Malda District, Eastern India, J. Environ. Inform., 2008, **12,** 140-149.
115. Wesolowski, M., Suchacz, B., Halkiewicz, J., The analysis of seasonal air pollution pattern with application of neural networks, Anal. Bioanal. Chem.., 2006, **384,** 458-467.

116. Yesilnacar, M.I., Sahinkaya, E., Naz, M., Ozkaya, B., Neural network prediction of nitrate in groundwater of Harran Plain, Turkey, Environ. Geol., 2008, **56,** 19-25.
117. Kuo, J-T., Hsieh, M-H., Lung, W-S., She, N., Using artificial neural network for reservoir eutrophication prediction, Ecol. Model., 2007, **200,** 171-177.
118. Kuzmanovski, I., Novič, M., Counter-propagation neural networks in Matlab, Chemometr. Intell. Lab., 2008, **90,** 84-91.
119. Hanafizadeh, P., Ravasan, A.Z., Khaki, H.R., An expert system for perfume selection using artificial neural network, Expert Syst. Appl., 2010, **37,** 8879-8887.
120. Alvarez-Guerra, M., Ballabio, D., Amigo, J.M., Viguri, J.R., Bro, R., A chemometric approach to the environmental problem of predicting toxicity in contaminated sediments, J. Chemometr., 2009, **24,** 379-386.
121. Otto, M., Chemometrics, Statistics and Computer Application in Analytical Chemistry, Wiley-VCH, Weinheim, Germany 1999.
122. Lin, H., Chen, Q., Zhao, J., Zhou, P., Determination of free amino acid content in Radix Pseudostellariae using near infrared (NIR) spectroscopy and different multivariate calibrations, J. Pharmaceut. Biomed., 2009, **50,** 803-808.
123. Williams Jr, D.K., Kovach, A.L., Muddiman, D.C., Hanck, K.W., Utilizing Artificial Neural Networks in MATLAB to Achieve Parts-Per-Billion Mass Measurement Accuracy with a Fourier Transform Ion Cyclotron Resonance Mass Spectrometer, J. Am. Soc. Mass Spectr., 2009, **20,** 1303-1310.
124. Mancia, A., Warr, G. W., Almeida, J. S., Veloso, A., Wells, R. S., Chapman, R. W., Transcriptome Profiles: Diagnostic Signature of Dolphin Populations, Estuar. Coasts, 2010, **33,** 919-929.
125. Rumelhart, D.E., Hinton, G.E., Williams, R.J., Learning representations by back-propagating errors, Nature, 1986, **323,** 533-536.
126. Kompany-Zareh, M., Massoumi, A., Pezeshk-Zadeh, Sh., Simultaneous spectrophotometric determination of Fe and Ni with xylenol orange using principal component analysis and artificial neural networks in some industrial samples, Talanta, 1999, **48,** 283-292.
127. Ramadan, Z., Hopke, P.K., Johnson M.J., Scow K.M., Application of PLS and Back-Propagation Neural Networks for the estimation of soil properties, Chemometr., Intell. Lab., 2005, **75,** 23-30.
128. Luengo, J., García, S., Herrera, F., A study on the use of statistical tests for experimentation with neural networks: Analysis of parametric test conditions and non-parametric tests, Expert Syst. Appl., 2009, **36,** 7798-7808.

129. Hernández-Caraballo, E.A., Ávila de Hernández, R.M., Rivas-Echeverría, F., Capote-Luna, T., Discrimination of Venezuelan spirituous beverages by a trace element-radial basis neural network approach, Talanta, 2008, **74,** 871-878.
130. Balabin, R.M., Safieva, R.Z., Biodiesel classification by base stock type (vegetable oil) using near infrared spectroscopy data, Anal. Chim. Acta, 2011, **689,** 190-197.
131. Fatemi, M.H., Abraham, M.H., Poole, C.F., Combination of artificial neural network technique and linear free energy relationship parameters in the prediction of gradient retention times in liquid chromatography, J. Chromatogr. A., 2008, **1190,** 241-252.
132. Sengorur, B., Dogan, E., Koklu, R., Samandar, A., Dissolved oxygen estimation, using Artificial neural network for water quality control, Fresen. Environ. Bull., 2006, **15,** 1064-1067.
133. Sharma, V., Negi, S.C., Rudra, R.P., Yang, S., Neural networks for predicting nitrate-nitrogen in drainage water, Agr. Water Manage., 2003, **63,** 169-183.
134. Sahoo, G.B., Ray, C., Wade, H.F., Pesticide prediction in ground water in North Carolina domestic wells using artificial neural networks, Ecol. Model., 2005, **183,** 29-46.
135. Kurban, T., Beşdok, A Comparison of RBF Neural Network Training Algorithms for Inertial Sensor Based Terrain Classification, Sensors, 2009, **9,** 6312-6329.
136. Zhang, G., Ni, Y., Churchill, J., Kokot, S., Authentication of vegetable oils on the basis of their physico-chemical properties with the aid of chemometrics, Talanta, 2006, **70,** 293-300.
137. Hasani, M., Emami, F., Evaluation of feed-forward back propagation and radial basis function neural networks in simultaneous kinetic spectrophotometric determination of nitroaniline isomers, Talanta, 2008, **75,** 116-126.
138. Balabin, R.M., Safieva, R.Z., Lomakina, E.I., Near-infrared (NIR) spectroscopy for motor oil classification: From discriminant analysis to support vector machines, Microchem. J., 2011, **98,** 121-128.
139. Ni, Y., Liu, Y., Kokot, S., Two-dimensional fingerprinting approach for comparison of complex substances analysed by HPLC-UV and fluorescence detection, Analyst, 2011, **136,** 550-559.
140. Nguyen, N.T., Lee, H.H., Induction motor fault diagnosis based on the k-NN and optimal feature selection, Int. J. Electron., 2010, **97,** 1071-1081.
141. http://people.revoledu.com/kardi/tutorial/KNN/What-is-K-Nearest-Neighbor-Algorithm.htm (τελευταία επίσκεψη 4/2011).
142. Chen, Q., Cai, J., Wan, X., Zhao, J., Application of linear/non-linear classification algorithms in discrimination of pork storage time using Fourier transform near infrared (FT-NIR) spectroscopy, LWT-Food Sci. Technol., 2001, **44,** 2053-2058.

143. Gulbag, A., Temurtas, F., Tasaltin, C., Öztürk, Z.Z., A study on radial basis function neural network size reduction for quantitative identification of individual gas concentrations in their gas mixtures, Sensor. Actuat. B-Chem., 2007, **124,** 383-392.
144. Φαρμάκη Ε., Δίκτυα Kohonen, Χημικά Χρονικά, 2011, **73(6),** 18-22.
145. Prieto, M.S., Allen, A.R., Using self-organising maps in the detection and recognition of road signs, Image Vision Comput., 2009, **27,** 673-683.
146. Lemes, M.R., Dal Pino, A., Periodic Table of the Elements in the Perspective of Artificial Neural Networks, J. Chem. Educ., 2011, **88,** 1511-1514.
147. Bodt, E., Cottrell, M., Verleysen, M., Statistical tools to assess the reliability of self-organizing maps, Neural Networks, 2002, **15,** 967-978.
148. http://www.mathworks.com/help/toolbox/nnet/ug/bss4b_l-1.html (τελευταία επίσκεψη 17/11/2011).
149. Ballabio, D., Consonni, V., Todeschini, R., The Kohonen and CP-ANN toolbox: A collection of MATLAB modules for Self Organizing Maps and Counterpropagation Artificial Neural Networks, Chemometr. Intell. Lab., 2009, **98,** 115-122.
150. Sim, S.F., Sági-Kiss, V., Brereton, R.G., Self-Organizing Maps and Support Vector Regression as aids to coupled chromatography: Illustrated by predicting spoilage in apples using volatile organic compounds, Talanta, 2011, **83,** 1269-1278.
151. Bayram, E., Santago, P., Harris, R., Xiao, Y.-D., Clauset, A.J., Schmitt, J.D., Genetic algorithms and self-organizing maps: a powerful combination for modeling complex QSAR and QSPR problems, J. Comput. Aid. Mol. Des., 2004, **18,** 483-493.
152. Stavrou, E.T., Charalambous, C., Spiliotis, S., Human resource management and performance: A neural network analysis, Eur. J. Oper. Res., 2007, **181,** 453-467.
153. http://www.ivie.es/downloads/ws/bf/2003/05/08/ponencia05.pdf (τελευταία επίσκεψη 17/11/2010).
154. Bieroza, M., Baker, A., Bridgeman, J., Classification and calibration of organic matter fluorescence data with multiway analysis methods and artificial neural networks: an operational tool for improved drinking water treatment, Environmetrics, 2011, **22,** 256-270.
155. Verdini, R.A., Zorrrilla, S.E., Rubiolo, A.C., Nakai, S., Multivariate statistical methods for Port Salut Argentino cheese analysis based on ripening time, storage conditions, and sampling sites, Chemometr. Intell. Lab., 2007, **86,** 60-67.
156. Ballabio, D., Kokkinofta, R., Todeschini, R., Theocharis, C.R., Characterization of the traditional Cypriot spirit Zivania by means of Counterpropagation Artificial Neural Networks, Chemometr. Intell. Lab., 2007, **87,** 52-58.

157. Wan, N.S. M-D., Dzulkiflee, I., Niamh, N., Classification and Source Determination of Medium Petroleum Distillates by Chemometric and Artificial Neural Networks: A Self Organizing Feature Approach, Anal. Chem., 2011, **83,** 7745-7754.
158. Alvarez-Guerra, M., González-Piñuela, C., Andrés, A., Galán, B., Viguri, J.R., Assessment of Self-Organizing Map artificial neural networks for the classification of sediment quality, Environ. Int., 2008, **34,** 782-790.
159. Alvarez-Guerra, E., Molina, A., Viguri, J., Alvarez-Guerra, M., A SOM-Based Methodology for Classifying Air Quality Monitoring Stations, Environ. Prog. Sust. Energy, 2011, **30(3),** 424-438.
160. Wongravee, K., Lloyd, G.R., Silwood, C.J., Grootveld, M., Brereton, R.G., Supervised Self Organizing Maps for Classification and Determination of Potentially Discriminatory Variables: Illustrated by Application to Nuclear Magnetic Resonance Metabolomic Profiling, Anal. Chem., 2010, **82(2),** 628-638.
161. Lloyd, G.R., Wongravee, K., Silwood, C.J.L., Grootveld, M., Brereton, R.G., Self Organising Maps for variable selection: Application to human saliva analysed by nuclear magnetic resonance spectroscopy to investigate the effect of an oral healthcare product, Chemometr. Intell. Lab., 2009, **98,** 149-161.
162. Torrecilla, J.S., Rojo, E., Oliet, M., Domínguez, J.C., Rodríguez, F., Self-Organizing Maps and Learning Vector Quantization Networks As Tools to Identify Vegetable Oils, Agri. Food Chem., 2009, **57,** 2763-2769.
163. Park, Y-S., Céréghino, R., Compin, A., Lek, S., Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters, Ecol. Model., 2003, **160,** 265-280.
164. Lamrini, B., Lakhal, El-K., Le Lann, M-V., Data validation and missing data reconstruction using self-organizing map for water treatment, Neural Comput. Appl., 2011, **20,** 575-588.
165. Kohonen, T., Self-organization and Associative Memory, 3rd ed., Springer-Verlag, Berlin, Germany, 1989.
166. Lischeid, G., Non-linear visualization and analysis of large water quality data sets: a model-free basis for efficient monitoring and risk assessment, Stoch. Env. Res. Risk A., 2009, **23,** 977-990.
167. Xiao, Y.-D., Clauset, A., Harris, R., Bayram, E., Santago, P., Schmitt, J.D., Supervised Self-Organizing Maps in Drug Discovery. 1. Robust Behavior with Overdetermined Data Sets, J. Chem. Inf. Model., 2005, **45,** 1749-1758.

168. Jin, Y.-H., Kawamura, A., Park, S.-C., Nakagawa, N., Amaguchi, H., Olsson, J., Spatiotemporal classification of environmental monitoring data in the Yeongsan River basin, Korea, using self-organizing maps, J. Environ. Monit., 2011, **13,** 2886-2894.
169. Heikkinen, M., Poutiainen, H., Liukkonen, M., Heikkinen, T., Hiltunen, Y., Subtraction analysis based on self-organizing maps for an industrial wastewater treatment process, Math. Comput. Simulat., 2011, **82,** 450-459.
170. Magallanes, J., García-Reiriz, A., Liberman, S, Zupan, J., Kohonen classification applying 'missing variables' criterion to evaluate the p-boronophenylalanine human-bodyconcentration decreasing profile of boron neutron capture therapy patients, J. Chemometr., 2011, **25,** 340-347.
171. Bieroza. M., Baker, A., Bridgeman, J., New data mining and calibration approaches to the assessment of water treatment efficiency, Adv. Eng. Softw., 2012, **44,** 126-135.
172. Sim, S.F., Sági-Kiss, V., Multiple Self Organising Maps (mSOMs) for simultaneous classification and prediction: Illustrated by spoilage in apples using volatile organic profiles, Chemometr. Intell. Lab., 2011, **109,** 57-64.
173. Zhou, Z-H., Wu, J., Tang, W., Ensembling neural networks: Many could be better than all, Artif. Intell., 2002, **137,** 239–263.
174. Krogh, A., Vedelsby, J., Neural Network Ensembles, Cross Validation, and Active Learning, in: Tesauro, G., Touretzky, D.S., Leen, T.K., (Eds.), Advances in Neural Information Processing Systems 7, 231-238, MIT Press, Cambridge MA, 1995.
175. Zhang, H., Zhou, Y., Cheng, P., Deng, S., Cui, X., Wang, H., Multi-objective simultaneous prediction of waterborne coating properties J. Math. Chem., 2009, **46,** 1050-1059.
176. Paliwal, M., Kumar,U.A., Neural networks and statistical techniques: A review of applications, Expert Syst. Appl., 2009, **36(1),** 2-17.
177. Torrecilla, J.S., Mena, M.L., Yáñez-Sedeño, P., García, J., J. Food Eng., 2007, **81,** 544-552.
178. Safavi, A., Abdollahi, H., Hormozi Nezhad, M.R., Artificial neural networks for simultaneous spectrophotometric differential kinetic determination of Co(II) and V(IV), Talanta, 2003, **59,** 515-523.
179. Melssen, W., Wehrens, R., Buydens, L., Supervised Kohonen networks for classification problems, Chemometr. Intell. Lab., 2006, **83,** 99-113.
180. Gautam, M.R., Watanabe, K., Saegusa, H., Runoff analysis in humid forest catchment with artificial neural network, J. Hydrol., 2000, **235,** 117-136.
181. Ceballos-Magaña, S.G., Jurado, J.M., Martín, M.J., Pablos, F., Quantitation of Twelve Metals in Tequila and Mezcal Spirits as Authenticity Parameters, Agr. Food Chem., 2009, **57,** 1372-1376.

182. Chen, Q., Zhao, J., Lin, H., Study on discrimination of Roast green tea (Camellia sinensis L.) according to geographical origin by FT-NIR spectroscopy and supervised pattern recognition, Spectrochim. Acta A, 2009, **72,** 845-850.
183. Feng, L., Hong, W., Classification error of multilayer perceptron neural networks, Neural Comput. Appl., 2009, **18,** 377-380.
184. Tortajada, S., García-Gómez, J.M., Vicente, J., Sanjuán, J., de frutos, R., Martín-Santos, R., García-Esteve, L., Gornemann, I., Gutiérrez-Zotes, A., Canellas, F., Carracedo, Á., Gratacos, M., Guillamat, R., Baca-García, E., Robles, M., Prediction of Postpartum Depression Using Multilayer Perceptrons and Pruning, Methods Inf., Med., 2009, **48,** 291-298.
185. Chakrabarti, S., Svojanovsky, S.R., Slavik, R., Georg, G.I., Wilson, G.S., Smith, P.G., Active Compounds Artificial Neural Network_Based Analysis of High-Throughput Screening Data for Improved Prediction of active compounds, J. Biomol. Screen., 2009, **14,** 1236-1244.
186. Tan, S.C., Lim, C.P., Integration of supervised ART-based neural networks with a hybrid genetic algorithm, Soft Comput., 2011, **15,** 205-219.
187. Cajka, T., Hajslova, J., Pudil, F., Riddelova, K., Traceability of honey origin based on volatiles pattern processing by artificial neural networks, J. Chromatogr. A, 2009, **1216,** 1458-1462.
188. Tsakovski, S., Tobiszewski, M., Simeonov, V., Polkowska, Z., Namieśnik, Chemical composition of water from roofs in Gdansk, Poland, J., Environ. Pollut., 2010, **158,** 84-91.
189. Tobiszewski, M., Tsakovski, S., Simeonov, V., Namieśnik, J., Surface water quality assessment by the use of combination of multivariate statistical classification and expert information, Chemosphere, 2010, **80,** 740-746.
190. Çinar, Ö., Merdun, H., Application of an unsupervised artificial neural network technique to multivariant surface water quality data, Ecol. Res., 2009, **24,** 163-173.
191. Chon, T.S., Park, Y.S., Moon, K.H., Cha, E.Y., Patternizing communities by using an artificial neural network, Ecol. Model., 1996, **90,** 69-78.
192. Balabin, R.M., Lomakina, E.I., Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data, Analyst, 2011, **136,** 1703-1712.
193. Boger, Z., Selection of quasi-optimal inputs in chemometrics modeling by artificial neural network analysis, Anal. Chim. Acta, 2003, **490,** 31-40.
194. Das, R., Sengur, A., Evaluation of ensemble methods for diagnosing of valvular heart disease, Expert Syst., Appl., 2010, **37,** 5110-5115.

195. Álvarez, M., Moreno, I.M., Ángeles, J., Cameán, A.M., González, A.G., Differentiation of 'two Andalusian DO 'fino' wines according to their metal content from ICP-OES by using supervised pattern recognition methods, Microchem. J., 2007, **87,** 72-76.
196. Kahramanli, H., Allahverdi, N., Extracting rules for classification problems: AIS based approach, Expert Syst. Appl., 2009, **36,** 10494-10502.
197. Alonso-Salces, R.M., Héberger, K., Holland, M.V., Moreno-Rojas, J.M., Mariani, C., Bellan, G., Reniero, F., Guillou, C., Multivariate analysis of NMR fingerprint of the unsaponifiable fraction of virgin olive oils for authentication purposes, Food Chem., 2010, **118,** 956-965.
198. Moore, A., Cross-validation for detecting and preventing overfitting. http://www.cs.cmu.edu/~awm/tutorials
(τελευταία επίσκεψη 7/2010).
199. Bowden, G.J., Dandy, G.C., Maier, H.R., Input determination for neural network models in water resources applications: Part 1 - Background and methodology, J. Hydrol., 2005, **301,** 75-92.
200. Bowden, G.J., Dandy, G.C., Maier, H.R., Input determination for neural network models in water resources applications: Part 2 - Case study: Forecasting salinity in a river, J. Hydrol., 2005, **301,** 93-107.
201. Lei, Y., He, Z., Zi, Y., Application of an intelligent classification method to mechanical fault diagnosis, Expert Syst. Appl., 2009, **36,** 9941-9948.
202. Balabin, R.M., Smirnov, S.V., Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data, Anal. Chim. Acta, 2011, **692,** 63-72.
203. Kruzlicova, D., Mocak, J., Balla, B., Petka J., Farkova, M., Havel, J., Classification of Slovak white wines using artificial neural networks and discriminant techniques, Food Chem., 2009, **112,** 1046-1052.
204. Tran, L.T., Knight, C.G., O'Neil, R.V., Smith, E.R., O'Connell, M., Self-organizing maps for integrated assessment of the mid-Atlantic region, Environ. Manage., 2003, **31(6),** 822-835.
205. Ferrer, R., Guiteras, J., Beltrán, J.L., Artifcial neural networks (ANNs) in the analysis of polycyclic aromatic hydrocarbons in water samples by synchronous fluorescence, Anal. Chim. Acta, 1999, **384,** 261-269.
206. Hernández-Caraballo, E.A., Rivas, F., Pérez, A.G., Marcó-Parra, L.M., Evaluation of chemometric techniques and artificial neural networks for cancer screening using Cu, Fe, Se and Zn concentrations in blood serum, Anal. Chim. Acta, 2005, **533,** 161-168.

207. Aleixandre, M., Lozano, J., Gutiérrez, J., Sayago, I., Fernández, M.J., Horrillo, M.C., Portable e-nose to classify different kinds of wine, Sensor. Actuat. B-Chem., 2008, **131,** 71-76.
208. Faisal, T., Ibrahim, F., Taib, M.N., A noninvasive intelligent approach for predicting the risk in dengue patients, Expert Syst. Appl., 2010, **37,** 2175-2181.
209. Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., De Jong, S., Lewi, P.J., Smeyers-Verbeke, J., Handbook of Chemometrics and Qualimetrics: Part A., Elsevier, Amsterdam, 1998.
210. Ballabio, D., Vasighi, M., Consonni, V., Komapany-Zareh, M., Genetic Algorithms for architecture optimisation of Counter-Propagation Artificial Neural Networks, Chemometr. Intell. Lab., 2011, **105,** 56-64.
211. Kuzmanovski, I., Novič, M., Trpkovska, M., Automatic adjustment of the relative importance of different input variables for optimization of counter-propagation artificial neural networks, Anal. Chim. Acta, 2009, **642,** 142-147.
212. Mantzaris, D., Anastassopoulos, G., Adamopoulos, A., Genetic algorithm pruning of probabilistic neural networks in medical disease estimation, Neural Networks, 2011, **24,** 831-835.
213. Feng, Y., Zhang, W., Sun, D., Zhang, L., Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification, Atmos. Environ., 2011, **45,** 1979-1985.
214. Brodnjak-Voncina, D., Dobcnik, D., Novic, M., Zupan, J., Chemometrics characterisation of the quality of river water, Anal. Chim. Acta, 2002, **462**, 87-100.
215. Astel, A., Tsakovski, S., Barbieri, P., Simeonov, V., Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets, Water Res., 2007, **41,** 4566-4578.
216. Apostolaki, E.T., Tsagaraki, T., Tsapakis, M., Karakassis, I., Fish farming impact on sediments and macrofauna associated with seagrass meadows in the Mediterranean, Estuar. Coast. Shelf S., 2007, **75,** 408-416.
217. Matijević, S., Kušpilić, G., Kljaković-Gašpić, Z., Bogner, D., Impact of fish farming on the distribution of phosphorus in sediments in the middle Adriatic area, Mar. Pollut. Bull., 2008, **56,** 535–548.
218. Tovar, A., Moreno, C., Mánuel-Vez, M.P., García-Vargas, M., Environmental impacts of intensive aquaculture in marine waters, Water Res., 2000, **34(1),** 334-342.

219. Chou, C.L., Haya, K., Paon, L.A., Burridge, L., Moffatt J.D., Aquaculture-related trace metals in sediments and lobsters and relevance to environmental monitoring program ratings for near-field effects, Mar. Pollut. Bull., 2002, 44, 1259-1268.
220. Sutherland, T.F., Petersen, S.A., Levings, C.D., Martin, A.J., Distinguishing between natural and aquaculture-derived sediment concentrations of heavy metals in the Broughton Archipelago, British Columbia, Mar. Pollut. Bull., 2007, **54**, 1451–1460.
221. Mendiguchía, C., Moreno, C., Mánuel-Vez, M.P., García-Vargas, M. Preliminary investigation on the enrichment of heavy metals in marine sediments originated from intensive aquaculture effluents, Aquaculture, 2006, **254,** 317–325.
222. Dalman, O., Demiral, A. Balci, A., Determination of heavy metals (Cd, Pb) and trace elements (Cu, Zn) in sediments and fish of the Southeastern Aegean, Food chem., 2006, **95,** 157-162.
223. Chou, C.L., Haya, K., Paon, L.A., Moffatt J.D., A regression model using sediment chemistry for the evaluation of marine environmental impacts associated with salmon aquaculture cage wastes, Mar. Pollut. Bull., 2004, **49,** 465-472.
224. Barasan, A.K., Aksu, M., Egemen, O., Impacts of the fish farms on the water column nutrient concentrations and accumulation of heavy metals in the sediments in the eastern Aegean Sea (Turkey), Environ. Monit. Assess., 2010, **162,** 439–451.
225. Neofitou, N., Vafidis, D., Klaoudatos, S., Spatial and temporal effects of fish farming on benthic community structure in a semi-enclosed gulf of the Eastern Mediterranean, Aquacult. Environ. Interact., 2010, **1,** 95-105.
226. Grošelj, N., Van der Veer, G., Tušar, M., Vračko, M., Novič, M., Verification of the geological origin of bottled mineral water using artificial neural networks, Food Chem., 2010, **118,** 941-947.
227. Novič, M., Grošelj, N., Bottle-neck type of neural network as a mapping device towards food specifications, Anal. Chim. Acta, 2009, **649,** 68-74.
228. Marini, F., Magrì, A.L., Balestrieri, F., Fabretti, F., Marini, D., Supervised pattern recognition applied to the discrimination of the floral origin of six types of Italian honey samples, Anal. Chim. Acta, 2004, **515,** 117-125.
229. Cajka, T., Riddellova, K., Tomaniova, M., Hajslova J., Recognition of beer brand based on multivariate analysis of volatile fingerprint, J. Chromatogr. A, 2010, **1217,** 4195-4203.
230. Astel, A., Tsakovski, S., Simeonov, V., Reisenhofer, E., Piselli, S., Barbieri, P., Multiway classification and modeling in surface water pollution estimation, Anal. Bioanal. Chem., 2008, **390,** 1283-1292.

231. Farmaki, E. G., Thomaidis, N.S., Simeonov, V., Efstathiou, C.E., A comparative chemometric study for water quality expertise of the Athenian water reservoirs, Environ. Monit. Assess., 2012, IN PRESS.
232. Spanilá, M., Pazourek, J., Farková, M., Havel, J., Optimization of solid-phase extraction using artificial neural networks in combination with experimental design for determination of resveratrol by capillary zone electrophoresis in wines, J. Chromatogr. A., 2005, **1084,** 180-185.
233. May, D. B., Sivakumar, M., Envirion. Prediction of urban stormwater quality using artificial neural networks, Modell. Softw., 2009, **24,** 296-302.
234. Polak, S., Wiśniowska, B., Ahamadi, M., Mendyk, A., Prediction of the hERG Potassium Channel Inhibition Potential with Use of the Artificial Neural Networks, Adv. Soft Comput., 2010, **75,** 91-99.
235. Melssen, W., Buydens, L., Aspects of multi-layer feed-forward neural networks influencing the quality of the fit of univariate non-linear relationships, Anal. Proc., 1995, **32,** 53-56.
236. Kamath, S.D., D'souza, C.S., Mathew, S., George, S.D., Santhosh, C., Mahato, K.K., A pilot study on colonic mucosal tissues by fluorescence spectroscopy technique: Discrimination by principal component analysis (PCA) and artificial neural network (ANN) analysis, J. Chemometr., 2008, **22,** 408-416.
237. Galão, O.F., Borsato, D., Pinto, J.P., Visentainer, J.V., Carrão-Panizzi, M.C., Artificial Neural Networks in the Classification and Identification of Soybean Cultivars by Planting Region, J. Brazil Chem. Soc., 2011, **22(1),** 142-147.
238. Hsu, K., Gupta, H.V., Gao, X., Sorooshian, S., Imam, B., Self-organizing Linear output map (SOLO): An Artificial Neural Network Suitable for Hydrologic Modeling and Analysis, Water Resour. Res., 2002, **38(12),** 1302-1318.
239. Marini, F., Zupan, J., Magrì, A.L., Class-modeling using Kohonen artificial neural networks, Anal. Chim. Acta, 2005, **544,** 306-314.
240. Zhang, J., Dong, Y., Yueiang, X., A comparison of SOFM ordination with DCA and PCA in gradient analysis of plant communities in the midst of Taihang Mountains, China, Ecol. Informatics, 2008, **3,** 367-374.
241. Banas, D., Masson, G., Leglize, L., Usseglio-Polatera, P., Boyd, C.E., Assessment of sediment concentration and nutrient loads in effluents drained from extensively managed fishponds in France, Environ. Pollut., 2008, **152,** 679-685.
242. Mente, E., Graham, J.P., Santos, M.B., Neofitou, C., Effect of feed and feeding in the culture of salmonids on the marine aquatic environment: a synthesis for European aquaculture, Aquacult. Int., 2006, **14,** 499-522.

243. Li, Q., Wu, Z., Chu, B., Zhang, N., Cai, S., Fang, J., Heavy metals in coastal wetland sediments of the Pearl River Estuary, China, Environ. Pollut., 2007, **149,** 158-164.
244. Nghia, N.D., Lunestad, B.T., Trung, T.S., Son, N.T., Maage, A., Heavy Metals in the Farming Environment and in some Selected Aquaculture Species in the Van Phong Bay and Nha Trang Bay of the Khanh Hoa Province in Vietnam, Bull. Environ. Contam. Toxicol., 2009, **82,** 75–79.
245. Sapkota A., Sapkota, A. R., Kucharski, M., Burke, J., McKenzie, S., Walker, P., Lawrence, R., Aquaculture practices and potential human health risks: Current knowledge and future priorities, Environ. Int., 2008, **34,** 1215-1226.
246. Cao, L., Wang, W., Yang, Y., Yang, C., Yuan, Z., Xiong, S., Diana J., Environmental Impact of Aquaculture and Countermeasures to Aquaculture Pollution in China, Env. Sci. Pollut. Res., 2007, **14,** 452–462.
247. Paz Suárez Araujo, C., García Báez, P., Sánchez Rodríguez, Á., Juan Santana Rodríguez, J., HUMANN-based system to identify benzimidazole fungicides using multi-synchronous fluorescence spectra: An ensemble approach, Anal. Bioanal. Chem., 2009, **394,** 1059-1072.
248. http://en.wikipedia.org/wiki/ROC_curve
(τελευταία επίσκεψη 9/1/2009).
249. Alvarez-Guerra, M., Ballabio, D., Amigo, J.N., Bro, R., Viguri, J.R., Development of models for predicting toxicity from sediment chemistry by partial least squares-discriminant analysis and counter-propagation artificial neural networks, Environ. Pollut., 2010, **158,** 607-614.
250. Díaz-Almela, E., Marbá, N., Álvarez, E., Santiago, R., Holmer, M., Grau, A., Mirto, S., Danovaro, R., Petrou, A., Argyrou, M., Karakassis, I., Duarte, C. M., Benthic input rates predict seagrass (Posidonia oceanica) fish farm-induced decline, Mar. Pollut. Bull., 2008, **56,** 1332–1342.
251. Holmer, M., Argyrou, M., Dalsgaard, T., Danovaro, R., Diaz-Almela, E., Duarte, C. M., Frederiksen, M., Grau, A., Karakassis, I., Marbá, N., Mirto, S., Pérez, M., Pusceddu, A., Tsapakis, M., Effects of fish farm waste on Posidonia oceanica meadows: Synthesis and provision of monitoring and management tools, Mar. Pollut. Bull., 2008, **56,** 1618–1629.
252. Cajka, T., Riddellova, K., Klimankova, E., Cerna, M., Pudil F., Hajslova, Traceability of olive oil based on volatiles pattern and multivariate analysis, Food Chem., 2010, **121,** 282-289.
253. García-González, D.L., Aparicio, R., Research in Olive Oil: Challenges for the Near Future, J. Agr. Food Chem., 2010, **58**, 12569-12577.

254. Bucci, R., Magrì, A.D., Magrì, A.L., Marini, D., Marini, F., J. Agric., Chemical Authentication of Extra Virgin Olive Oil Varieties by Supervised Chemometric Procedures, Food Chem. Acta, 2002, **50,** 413-418.
255. Ilyasoglu, H., Ozcelik, B., Van Hoed, V., Verhe, R, Characterization of Aegean Olive Oils by Their Minor Compounds, J. Am. Oil Chem. Soc., 2010, **87**, 627-636.
256. Lanteri, S., Armaninoa, C., Perrib, E., Palopolib, A., Study of oils from Calabrian olive cultivars by chemometric methods, Food Chem., 2002, **76,** 501-507.
257. Marini, F., Magrì, A.L., Bucci, R, Balestrieri, F., Marini, D., Class-modeling techniques in the authentication of Italian oils from Sicily with a Protected Denomination of Origin (PDO), Chemometr. Intell. Lab., 2006, **80**, 140-149.
258. Casale, M., Casolino, C., Oliveri, P., Forina M., The potential of coupling information using three analytical techniques for identifying the geographical origin of Liguria extra virgin olive oil, Food Chem. 2010, **118**, 163-170.
259. Cosio, M.S, Ballabio, D., Benedetti, S., Gigliotti, C., Geographical origin and authentication of extra virgin olive oils by an electronic nose in combination with artificial neural networks, Anal. Chim. Acta, 2006, **567**, 202-210.
260. Dupuy, N., Le Dreau, Y., Ollivier, D., Artaud, J., Pinatel, C., Kister, J, Origin of French Virgin Olive Oil Registered Designation of Origins Predicted by Chemometric Analysis of Synchronous Excitation-Emission Fluorescence Spectra, J. Agr. Food Chem., 2005, **53**, 9631–9638.
261. Rezzi, S., Axelson, D.E., Heberger, K., Reniero, F., Mariani, C., Guillou C., Classification of olive oils using high throughput flow 1H NMR fingerprinting with principal component analysis, linear discriminant analysis and probabilistic neural networks, Anal. Chim. Acta, 2005, **552**, 13–24.
262. Ichihashi, H., Morita, H., Tatsukawa, R., Rare earth elements (REEs) in naturally grown plants in relation to their variation in soils, Eur. J. Lipid Sci. Tech., 1992, **111,** 1003-1013.
263. Farmaki, E.G., Thomaidis, N.S., Minioti, K.S., Ioannou, E., Georgiou, C.A., Efstathiou, C.E., Geographical Characterization of Greek Olive Oils Using Rare Earth Elements Content and Supervised Chemometric Techniques, Anal. Lett., 2012, **45,** 920-932.
264. Cimpoiu, C., Cristea, V-M., Hosu, A., Sandru, M., Seserman, L., Antioxidant activity prediction and classification of some teas using artificial neural networks, Food Chem., 2011, **127,** 1323-1328.
265. Ainsworth, A., Discriminant Function Analysis. Applied Multivariate Statistics and Laboratory, California State University, Northridge. http://www.csun.edu, 2004.

266. Chatfield, C., Neural networks: Forecastine breakthrough or passing fad? Int. J. Forecasting, 1993, **9,** 1-3.

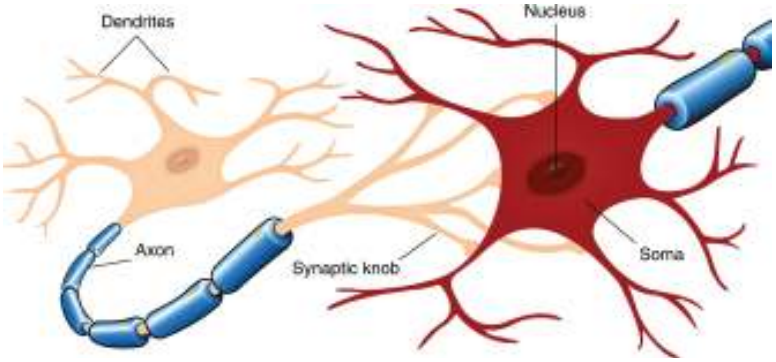
ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΧΗΜΕΙΑΣ

ΘΕΩΡΗΤΙΚΟ ΜΕΡΟΣ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ

ΕΦΑΡΜΟΓΕΣ ΠΟΛΥΠΑΡΑΜΕΤΡΙΚΩΝ ΣΤΑΤΙΣΤΙΚΩΝ ΤΕΧΝΙΚΩΝ ΣΤΗ ΧΗΜΙΚΗ ΑΝΑΛΥΣΗ



ΕΛΕΝΗ ΦΑΡΜΑΚΗ

ΧΗΜΙΚΟΣ

ΑΘΗΝΑ

ΙΟΥΝΙΟΣ 2012

ΠΑΡΑΡΤΗΜΑ (ΘΕΩΡΗΤΙΚΟ ΜΕΡΟΣ)

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΕΦ. 1 ΚΛΑΣΙΚΕΣ ΜΕΘΟΔΟΙ

1.1. ΔΙΑΧΩΡΙΣΤΙΚΗ ΑΝΑΛΥΣΗ (DISCRIMINANT ANALYSIS)

Η F-ratio παράμετρος ορίζεται ως ο λόγος των between groups (ανάμεσα στις ομάδες) και των within group (στην ίδια ομάδα) διακυμάνσεων (βλ. σχέση 1.2 παρακάτω). Οι μεταβλητές με τα υψηλότερα F-ratios έχουν την καλύτερη διαχωριστική ικανότητα [1, 2]. Η παράμετρος λ_j δε, που ονομάζεται **Wilks' lambda/variable,** ορίζεται για κάθε μεταβλητή j και είναι ο λόγος ανάμεσα στο άθροισμα των τετραγώνων των διακυμάνσεων για τα within group δείγματα σ_{jg} και το τετράγωνο της συνολικής διακύμανσης σ_j της μεταβλητής αυτής:

$$\lambda_j = \frac{\sum_{g=1}^{G} \sigma_{jg}^2 (n_g - 1)}{\sigma_j^2 (N - 1)} \qquad (1.1)$$

όπου: G ο αριθμός των ομάδων, N ο συνολικός αριθμός των δειγμάτων και n_g ο αριθμός των δειγμάτων στην ομάδα g [3]. Είναι δηλαδή το ποσοστό της διακύμανσης το οποίο δεν εξηγείται από το μοντέλο της ανάλυσης διακύμανσης για μια μεταβλητή [4]. Η παράμετρος Wilks' lambda/variable κυμαίνεται σε ένα εύρος 0 – 1 για τέλεια και καμιά διαχωριστική ικανότητα αντίστοιχα [3, 4, 5]. Η τιμή 0 σημαίνει ότι οι ομάδες διαφέρουν μεταξύ τους (ως προς τη συγκεκριμένη μεταβλητή), ενώ η τιμή 1, ότι όλες οι ομάδες είναι το ίδιο. Η παράμετρος Wilks' lambda/variable ορισμένες φορές ονομάζεται **U statistic [6].** Η τιμή της δίνει το συνολικό Wilks' lambda/variable μετά την απομάκρυνση της αντίστοιχης μεταβλητής [7].

Το πηλίκο lambda(after)/lambda(before) (όπου το after (μετά), αφορά την παράμετρο Λ (Wilks' lambda/model) μετά την προσθήκη μιας μεταβλητής σε σχέση με την προηγούμενη/πριν (before) τιμή ονομάζεται **Partial lambda,** και σχετίζεται με το F-ratio με βάση τη σχέση:

$$\text{F-ratio} = \frac{N - p - G}{G - 1} \times \frac{1 - \text{partial lambda}}{\text{partial lambda}} = \frac{1 - \Lambda}{\Lambda} \qquad (1.2)$$

όπου: G ο αριθμός των ομάδων N ο αριθμός των δειγμάτων, και p ο αριθμός των μεταβλητών.

Η παράμετρος Λ (Wilks' lambda/model) του μοντέλου δίνεται από τη σχέση:

$$\Lambda = \frac{\sum (Y_{ig} - \overline{Y_g})^2}{\sum (Y_{ig} - \overline{Y})^2} = \frac{\text{within}}{\text{total}} \qquad (1.3)$$

όπου: Y τα σκορ (βλ. σχέση 2.1) και ειδικότερα, Y_{ig} το σκορ του δείγματος i που ανήκει στην ομάδα g, $\overline{Y_g}$ το μέσο σκορ των δειγμάτων της ομάδας g και $\overline{Y}$ το μέσο σκορ όλων των δειγμάτων [8]. Η παράμετρος Wilks' lambda/model χρησιμοποιείται για να ελέγξει την κρισιμότητα μιας LDF ως σύνολο (πόσο καλά μια διαχωριστική συνάρτηση διαχωρίζει τα δείγματα σε ομάδες). Όσο μικρότερο είναι το Λ, τόσο σημαντικότερη η διαχωριστική συνάρτηση. Σημαντικό Λ σημαίνει ότι κάποιος μπορεί να απορρίψει τη μηδενική υπόθεση ότι δυο ή περισσότερες ομάδες έχουν τις ίδιες μέσες τιμές σκορ, οπότε το μοντέλο μπορεί να τις διαχωρίσει [6]. Ισούται με το κλάσμα της συνολικής διακύμανσης των σκορ που δεν ερμηνεύεται από τις διαφορές ανάμεσα στις ομάδες [9].

Η παράμετρος partial lambda αφορά τη συνεισφορά της κάθε μεταβλητής στο μοντέλο για την επίτευξη του διαχωρισμού των ομάδων [7]. Η διαχωριστική ικανότητα της κάθε μεταβλητής, είναι τόσο καλύτερη όσο μικρότερη είναι η τιμή του partial lambda (τιμές 0,0 και 1,0 δηλώνουν και εδώ τέλεια και καμιά διαχωριστική ικανότητα της μεταβλητής αντίστοιχα) [5, 7].

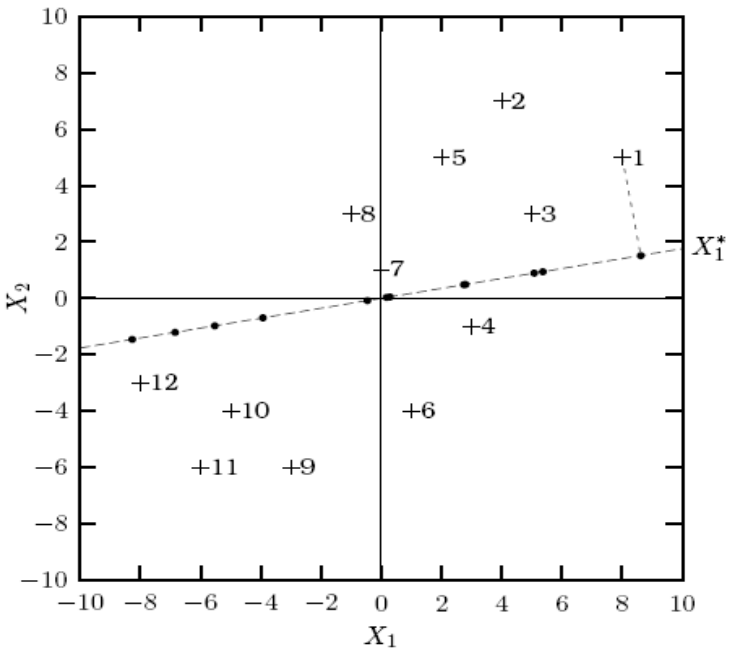
Υπολογιστικά λοιπόν η LDA, φαίνεται να έχει μεγάλες ομοιότητες με την **Ανάλυση της διακύμανσης** (Analysis of variance, ANOVA). Ας αναφέρουμε εδώ ένα παράδειγμα: ας υποθέσουμε ότι μετράμε το ύψος σε ένα τυχαίο πληθυσμό που αποτελείται από 50 άντρες και 50 γυναίκες. Γενικά (κατά μέσο όρο), οι γυναίκες είναι κοντύτερες των αντρών και αυτό θα αντανακλάται στη μέση τιμή της μεταβλητής (ύψος). Έτσι, το ύψος μας επιτρέπει να διαχωρίσουμε τους άντρες από τις γυναίκες με μια καλύτερη πιθανότητα από ένα τυχαίο διαχωρισμό: αν κάποιο άτομο είναι ψηλό, είναι πιθανότατα άντρας, αν είναι κοντύτερο, είναι γυναίκα [7].

1.2. ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ (PRINCIPAL COMPONENTS ANALYSIS)

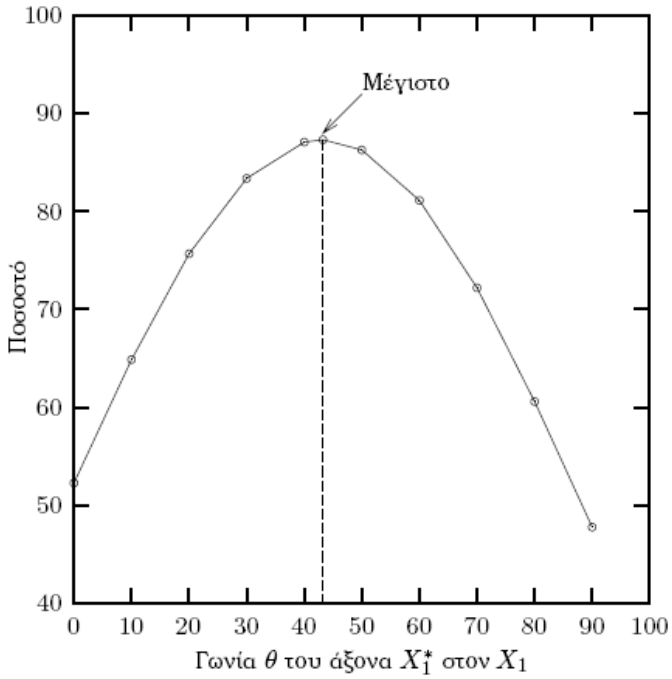
1.2.1. Θεωρία

Μπορούμε να θεωρήσουμε την PCA ως μια περιστροφή των αξόνων έτσι, ώστε η PC1 να απεικονίζει την κατεύθυνση της μέγιστης διακύμανσης, η PC2 την επόμενη κ.ο.κ. Συνήθως όμως μόνο οι PC1 και PC2 απεικονίζουν σε μόνο δυο διαστάσεις τα αρχικά δεδομένα, ώστε να επιτυγχάνεται τελικά μείωση των αρχικών n-διαστάσεων [10].

Μπορούμε λοιπόν να φανταστούμε ένα σύστημα δυο αξόνων X_1, X_2 (με μερικά σημεία να απεικονίζονται σε αυτό, σχ. 1.1(α)) που να περιστρέφεται διαρκώς, ώσπου να βρεθεί η καλύτερη γωνία θ ανάμεσα στον αρχικό οριζόντιο άξονα X_1 και το νέο X_1* (σχ. 1.1(β)). Η γωνία αυτή εξασφαλίζει την περιγραφή της μέγιστης διακύμανσης από το νέο άξονα.



(α)



(β)

Σχήμα 1.1: Νέος άξονας X1 και προβολή των αρχικών σημείων σε αυτόν (α). Ποσοστό της ολικής διακύμανσης που απεικονίζει ο άξονας X1* για κάθε γωνία θ (β)* *[11].*

Το πρώτο βήμα για την PCA είναι η εύρεση ενός αρχικού συμμετρικού πίνακα δεδομένων: **διακυμάνσεων-συνδιακυμάνσεων** (variance-covariance ή covariance matrix)

ή **συσχετίσεων** (correlation matrix) των μεταβλητών [12, 13]. Οι πίνακες αυτοί περιέχουν αντίστοιχα τις διακυμάνσεις/συνδιακυμάνσεις ή τους γνωστούς συντελεστές συσχέτισης Pearson (correlation coefficient) δυο μεταβλητών x, y:

$$\text{συντελεστής Pearson} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \qquad (1.4)$$

Όταν χρησιμοποιείται ο πίνακας διακυμάνσεων, η ανάλυση επηρεάζεται από τις διαφορές στο μέγεθος των τιμών διακυμάνσεων των μεταβλητών. Έτσι, μια ανάλυση βασισμένη στο πίνακα διακυμάνσεων είναι κατάλληλη μόνο όταν η ανίχνευση τέτοιων διαφορών σχετίζεται με το πρόβλημα που μελετάται [7]. Ενδέχεται δηλαδή, η διαφορά στις διακυμάνσεις να περιέχει πληροφορία πολύτιμη για το θέμα που εξετάζουμε. Ίσως λοιπόν κάποιες μεταβλητές να πρέπει να θεωρηθούν πως έχουν μεγαλύτερο βάρος εξαιτίας της διακύμανσής τους και επομένως θέτοντας όλες τις μεταβλητές να έχουν το ίδιο βάρος χάνουμε χρήσιμη πληροφορία [4].

Στις περισσότερες των περιπτώσεων ωστόσο, αυτές οι διαφορές δεν ενδιαφέρουν γιατί αυτές σχετίζονται με διαφορές στις μονάδες μέτρησης. Για παράδειγμα, ας υποθέσουμε ότι χρησιμοποιούνται δυο διαφορετικές σειρές μετρήσεων θερμοκρασίας, σε μονάδες Celcius και Fahrenheit. Η πρώτη PCA συνιστώσα που ανιχνεύει τη μέγιστη διακύμανση μεταξύ των μεταβλητών, θα αντανακλούσε τότε τη διαφορά των μονάδων στις δυο σειρές μετρήσεων. Αντίθετα, αν η ανάλυση γίνει με τη χρήση του πίνακα συσχετίσεων, αποφεύγονται τέτοια προβλήματα. Ο πίνακας συσχετίσεων δεν είναι τίποτα άλλο από τον πίνακα διακυμάνσεων των κανονικοποιημένων (standardized) μεταβλητών [7].

Στην πράξη ωστόσο, είναι ασαφές ποιο από τους δύο πίνακες πρέπει να χρησιμοποιούμε. Μια καλή στρατηγική είναι να αποφεύγουμε τον πίνακα διακύμανσης όταν υπάρχουν κάποιες μεταβλητές με πολύ μεγαλύτερη διακύμανση από ότι οι υπόλοιπες. Αν όμως οι διακυμάνσεις διαφέρουν, αλλά είναι συγκρίσιμες (π.χ. αναφέρονται σε ίδιες μονάδες), θα ήταν δόκιμο να χρησιμοποιούμε αυτή την πληροφορία. Εναλλακτικά θα μπορούσε κανείς να μετασχηματίσει τα δεδομένα του ώστε να γίνουν συγκρίσιμες οι διακυμάνσεις [4].

Σύμφωνα με τα παραπάνω, τα αποτελέσματα που θα εξαχθούν από την PCA δεν είναι ανεξάρτητα από τις μονάδες μέτρησης που χρησιμοποιούνται για κάθε μεταβλητή [13]. Έτσι, οι μεταβλητές πρέπει να μετρούνται όλες με την ίδια κλίμακα. Όταν δεν συμβαίνει αυτό, συνήθως γίνεται μετατροπή των αρχικών μεταβλητών με:

- ✓ αφαίρεση της μέσης τιμής (mean-centering) από κάθε τιμή και διαίρεση με την τυπική απόκλιση (standardization) (§ 2.2.9).

Η πρώτη επεξεργασία (mean-centering) γίνεται ουσιαστικά (με βάση τους ορισμούς διακύμανσης/συνδιακύμανσης), ακόμα και για τη συνήθη διαδικασία της εύρεσης του πίνακα δεδομένων που περιλαμβάνει τις συναρτήσεις αυτές (variance-covariance matrix), ενώ σε συνδυασμό με τη δεύτερη (standardization) αποτελούν τη διαδικασία για την εύρεση του πίνακα δεδομένων που περιλαμβάνει τις συσχετίσεις (correlation matrix), και ακολουθείται στις περιπτώσεις διαφορετικής κλίμακας των μεταβλητών.

Το δεύτερο βήμα για την PCA, είναι η εξαγωγή των **ιδιοανυσμάτων** (eigenvectors) που ειδικά στην περίπτωση αυτή αποτελούν τις **συνιστώσες** (components). Τα ιδιοανύσματα αποτελούν μια ειδική περίπτωση στον πολλαπλασιασμό πινάκων. Αν ένας τετραγωνικός πίνακας πολλαπλασιαστεί από δεξιά με ένα διάνυσμα, χωρίς να αλλάζει τη διεύθυνση αυτού, το διάνυσμα ονομάζεται ιδιοάνυσμα (eigenvector) του πίνακα αυτού. Το πρόθεμα **eigen** προέρχεται από την αντίστοιχη γερμανική λέξη που σημαίνει έμφυτος (ενδογενής). Ένα παράδειγμα φαίνεται παρακάτω:

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 11 \\ 5 \end{bmatrix} \qquad (1.5)$$

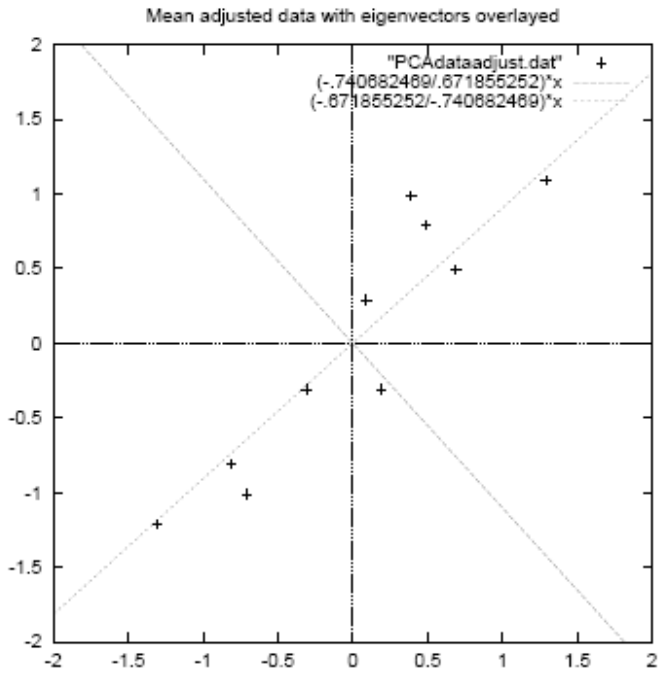
$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \times \begin{bmatrix} 3 \\ 2 \end{bmatrix} \qquad (1.6)$$

Στην πρώτη σχέση (1.5) το διάνυσμα (1, 3) δεν είναι ιδιοάνυσμα, γιατί όταν πολλαπλασιάζεται με τον τετραγωνικό πίνακα, δεν προκύπτει ακέραιο πολλαπλάσιό του. Αντίθετα, στη δεύτερη περίπτωση (σχέση 1.6), το διάνυσμα (3, 2) είναι ιδιοάνυσμα, γιατί κατά τον πολλαπλασιασμό, προκύπτει διάνυσμα 4 φορές πολλαπλάσιο του εαυτού του. Άρα η διεύθυνση του διανύσματος δεν αλλάζει. Η τιμή 4 αποτελεί την **ιδιοτιμή** (eigenvalue), του ιδιοανύσματος. Άρα, ιδιοανύσματα και ιδιοτιμές βρίσκονται πάντα σε ζευγάρια.

Τα ιδιοανύσματα δεν υπάρχουν σε όλους τους τετραγωνικούς πίνακες. Σε όσους n×n πίνακες υπάρχουν ωστόσο, είναι πάντα n [14]. (Στην πραγματικότητα, ο μέγιστος αριθμός των ιδιοανυσμάτων/ιδιοτιμών είναι n: ο αριθμός μπορεί να είναι μικρότερος με την έννοια ότι κάποιες τιμές είναι πολλαπλές λύσεις [15]). Τα ιδιοανύσματα είναι πάντα ανεξάρτητα (ορθογώνια) μεταξύ τους και κατά την εξαγωγή τους συνήθως επιλέγονται

από μαθηματικούς και λογισμικά του εμπορίου, εκείνα που έχουν άθροισμα 1 (για παράδειγμα το διάνυσμα (0,8, 0,6), με $\sqrt{0{,}8^2 + 0{,}6^2} = 1$). Έτσι, διευκολύνονται οι υπολογισμοί χωρίς να αλλάζει η διεύθυνση αυτών.

Τα ιδιοανύσματα ενός τετραγωνικού πίνακα δίνουν πληροφορίες για τα δεδομένα αυτού. Απεικονίζουν τη συσχέτιση των δεδομένων του πίνακα και μάλιστα εμπεριέχουν την πληροφορία αυτή κατά σειρά αύξουσας ιδιοτιμής: το ιδιοάνυσμα με τη μεγαλύτερη ιδιοτιμή παρέχει περισσότερη πληροφορία κ.ο.κ. Το σχήμα 1.2 που ακολουθεί δίνει μια εικόνα των παραπάνω: το πρώτο από τα ιδιοανύσματα (διακεκομμένη διαγώνιος από κάτω αριστερά προς τα πάνω δεξιά) περιγράφει τη μέγιστη πυκνότητα των σημείων, σαν να χαρασσόταν η best of fit ευθεία. Το δεύτερο δίνει την επόμενη λιγότερη σημαντική διασπορά των σημείων [14].



Σχήμα 1.2: Διάγραμμα δεδομένων (σημειωμένα με +) και των ιδιοανυσμάτων του πίνακα διακύμανσης αυτών (διαγώνιες διακεκομμένες γραμμές) [14].

Το πιο σημαντικό όμως βήμα της ανάλυσης το οποίο δυστυχώς δεν έχει εύκολη και κοινώς αποδεκτή απάντηση είναι η απόφαση για το πόσες συνιστώσες θα αξιοποιηθούν τελικά. Επιλέγοντας προφανώς λιγότερες κύριες συνιστώσες από όσες αρχικές μεταβλητές, χάνουμε πληροφορία. Αυτό είναι το κόστος, για το κέρδος μας να μειώσουμε τις

διαστάσεις του προβλήματος. Ωστόσο, συνήθως ενδιαφερόμαστε για κάποιο μικρότερο αριθμό συνιστωσών [4].

Στη βιβλιογραφία υπάρχουν πολλά κριτήρια τα οποία θα προσπαθήσουμε να περιγράψουμε. Αυτά είναι:

- ✓ Το **κριτήριο Kaiser,** σύμφωνα με το οποίο επιλέγονται οι παράγοντες για τους οποίους οι ιδιοτιμές είναι τουλάχιστον 1. Έτσι, έστω λj οι ιδιοτιμές. Το κριτήριο αυτό λέει να πάρουμε τόσες ιδιοτιμές όσες είναι μεγαλύτερες από:

$$\bar{\lambda} = \sum_{j} \lambda_j$$

  δηλαδή μεγαλύτερες από τη μέση τιμή των ιδιοτιμών. Στην περίπτωση που δουλεύουμε με πίνακα συσχετίσεων, ισχύει $\bar{\lambda}$ =1 και επομένως επιλέγουμε τόσες συνιστώσες όσες και οι ιδιοτιμές μεγαλύτερες της μονάδας.

  Το κριτήριο συνήθως υπερεκτιμά τον αριθμό των συνιστωσών που χρειάζονται [4].
- ✓ Το **κριτήριο Cattell** (διάγραμμα διαλογής ή scree plot). Το διάγραμμα διαλογής είναι ένα γράφημα που έχει στον οριζόντιο άξονα τη σειρά των ιδιοτιμών και στον κάθετο άξονα την τιμή της κάθε ιδιοτιμής. Σύμφωνα με το κριτήριο αυτό, πρέπει να βρεθεί το σημείο που υπάρχει ομαλή μείωση των ιδιοτιμών προς το σημείο μηδέν (επιλέγονται οι παράγοντες πριν το σημείο απότομης μείωσης της κλίσης της καμπύλης [11]) ή αλλιώς να πάρουμε τόσες συνιστώσες μέχρι το γράφημα να αρχίσει να γίνεται περίπου επίπεδο [4].

  Το κριτήριο είναι αρκετά υποκειμενικό (δύσκολο να βρεθεί που αλλάζει η κλίση του διαγράμματος).
- ✓ **Ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες:** επιλέγονται τόσοι παράγοντες ώστε να ερμηνεύεται τουλάχιστον το **80 ή 90 %** της συνολικής διακύμανσης [4, 13, 16]. Το κριτήριο αυτό στην πράξη δεν δίνει τα καλύτερα αποτελέσματα, ιδίως αν ο στόχος είναι αρκετά υψηλός. Επίσης δεν είναι ξεκάθαρο ποιο ποσοστό της διακύμανσης πρέπει να βάλουμε ως στόχο [4].
- ✓ **Ποσοστό της διακύμανσης των αρχικών μεταβλητών που ερμηνεύεται.** Το κριτήριο αυτό διαλέγει τόσες συνιστώσες ώστε να ερμηνεύεται για κάθε μεταβλητή ένα προκαθορισμένο υψηλό ποσοστό. Ομοίως, ποιο είναι αυτό το ποσοστό είναι υποκειμενικό. Επίσης μπορεί κάποια μεταβλητή να μην ερμηνεύεται σωστά και αυτό να οδηγήσει σε μεγάλο αριθμό συνιστωσών [4].

1.2.2. Παράδειγμα κατανόησης 1 (PCA)

Παρακάτω περιγράφεται ένα παράδειγμα κατανόησης της PCA με χρήση της άλγεβρας πινάκων [11]. Οι υπολογισμοί που ακολουθούν είναι δύσκολο να γίνουν χωρίς τη χρήση ηλεκτρονικού υπολογιστή, γι' αυτό θα χρησιμοποιηθούν απλά δεδομένα, όπως αυτά που απεικονίζονται στον πίνακα 1.1. Οι αρχικές μεταβλητές X_1, X_2 με διακυμάνσεις 2,12 και 1,58 αντίστοιχα (πίνακας 1.1) πρέπει να μετασχηματιστούν σε άλλες Z_1, Z_2, έτσι ώστε η ίδια συνολική διακύμανση να "κατανεμηθεί" διαφορετικά και η Z_1 να εκφράζει όσο το δυνατό μεγαλύτερο ποσοστό. Επιπλέον, οι νέες μεταβλητές πρέπει να είναι γραμμικοί συνδυασμοί των αρχικών, με το άθροισμα των τετραγώνων των συντελεστών εξάρτησης (φορτίσεις) ίσο με τη μονάδα και ανεξάρτητες μεταξύ τους. Οι τρεις αυτές συνθήκες εκφράζονται με τις παρακάτω σχέσεις 1.7:

$$\left.\begin{array}{cc} Z_1 = x_{11} X_1 + x_{12} X_2 & Z_2 = x_{21} X_1 + x_{22} X_2 \\ x_{11}^2 + x_{12}^2 = 1 & x_{21}^2 + x_{22}^2 = 1 \\ x_{11}x_{21} + x_{12}x_{22} = 0 & \end{array}\right\} \quad (1.7)$$

Πίνακας 1.1: Αρχικές τιμές υποθετικών μεταβλητών και standardization αυτών.

Α/Α	**X_1**		**X_2**	
	Αρχική μεταβλητή	**Μετά από standardization**	**Αρχική μεταβλητή**	**Μετά από standardization**
1	12	-0,471	10	-1,265
2	15	0,943	13	0,632
3	15	0,943	14	1,265
4	13	0,000	12	0,000
5	10	-1,414	11	-0,632
Μέση τιμή	13	0,000	12	0,000
Τυπική απόκλιση	**2,121**	1,000	**1,581**	1,000
Διακύμανση	4,500	1,000	2,500	1,000

Οι νέες τιμές (μετά από standardization) των X_1, X_2 μπορούν να απεικονιστούν με τη μορφή αλγεβρικού πίνακα:

$$A = \begin{bmatrix} -0{,}471 & -1{,}265 \\ 0{,}943 & 0{,}632 \\ 0{,}943 & 1{,}265 \\ 0{,}000 & 0{,}000 \\ 1{,}414 & -0{,}632 \end{bmatrix} \qquad (1.8)$$

Στη συνέχεια δημιουργείται ο πίνακας συσχέτισης R:

$$R = A \times A' =$$

$$\begin{bmatrix} -0{,}471 & -1{,}265 \\ 0{,}943 & 0{,}632 \\ 0{,}943 & 1{,}265 \\ 0{,}000 & 0{,}000 \\ 1{,}414 & -0{,}632 \end{bmatrix} \times \begin{bmatrix} -0{,}471 & 0{,}943 & 0{,}943 & 0{,}000 & 1{,}414 \\ -1{,}265 & 0{,}632 & 1{,}265 & 0{,}000 & -0{,}632 \end{bmatrix} = \begin{bmatrix} 1{,}000 & 0{,}922 \\ 0{,}922 & 1{,}000 \end{bmatrix}$$

Υπολογίζουμε τις ιδιοτιμές και τα ιδιοανύσματα:

$$\left|R - \lambda I\right| = 0 \Leftrightarrow \left\| \begin{bmatrix} 1{,}000 & 0{,}922 \\ 0{,}922 & 1{,}000 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\| = 0 \Leftrightarrow \text{…….....}$$

$$\Leftrightarrow \lambda_1 = 1{,}922 \text{ και } \lambda_2 = 0{,}078 \qquad (1.9)$$

Το άθροισμα των ιδιοτιμών είναι ίσο με το ίχνος (trace) του αρχικού πίνακα R (όταν αυτός αφορά αποκλειστικά συσχετίσεις και όχι διακυμάνσεις):

$$1{,}922 + 0{,}078 = 1{,}000 + 1{,}000 = 2$$

Η λ_1 είναι η διακύμανση της πρώτης κύριας συνιστώσας και η λ_2 της δεύτερης. Το ποσοστό της διακύμανσης που εκφράζει κάθε συνιστώσα είναι:

$$\text{Διακύμανση } (\lambda_1) = \frac{1{,}922}{2} \times 100 = 96{,}1\ \%$$

$$\text{Διακύμανση } (\lambda_2) = \frac{0{,}078}{2} \times 100 = 3{,}9\ \%$$

Είναι φανερή η "ανακατανομή" στα ποσοστά των διακυμάνσεων καθώς οι διακυμάνσεις των αρχικών μεταβλητών X_1, X_2 αντιπροσώπευαν ποσοστά 64,3 και 35,7 αντίστοιχα (βλ. πίνακα 1.1).

Το ιδιοάνυσμα Y υπολογίζεται για κάθε eigenvalue από την εξίσωση:

$$R \times Y = \lambda Y.$$

Έτσι αντίστοιχα για τα λ_1, λ_2 έχουμε:

$$\begin{bmatrix} 1{,}000 & 0{,}922 \\ 0{,}922 & 1{,}000 \end{bmatrix} \begin{bmatrix} y_{11} \\ y_{21} \end{bmatrix} = 1{,}922 \begin{bmatrix} y_{11} \\ y_{21} \end{bmatrix}$$

$$\begin{bmatrix} 1,000 & 0,922 \\ 0,922 & 1,000 \end{bmatrix} \begin{bmatrix} y_{12} \\ y_{22} \end{bmatrix} = 0,078 \begin{bmatrix} y_{12} \\ y_{22} \end{bmatrix}$$

και τελικά αν θέσουμε $y_{11} = 1$ και $y_{12} = -1$:

$$Y_1 = \begin{bmatrix} 1,0 \\ 1,0 \end{bmatrix} \text{ και } Y_2 = \begin{bmatrix} -1,0 \\ 1,0 \end{bmatrix}$$

Τα τελευταία πρέπει να μετασχηματιστούν ώστε το άθροισμα των τετραγώνων των στοιχείων τους να είναι ίσο με τη μονάδα. Αυτό γίνεται υπολογίζοντας ένα συντελεστή k ο οποίος ισούται με το αντίστροφο της τετραγωνικής ρίζας του αθροίσματος των τετραγώνων των στοιχείων του πίνακα:

$$k = \frac{1}{\sqrt{\sum y_i^2}} \qquad (1.10)$$

Επομένως βάση της σχέσης 1.10:

$$k_1 = \frac{1}{\sqrt{1^2 + 1^2}} = 0,707 \text{ και } k_2 = \frac{1}{\sqrt{(-1)^2 + 1^2}} = 0,707$$

και τελικά:

$$Y_1 = 0,707 \begin{bmatrix} 1,0 \\ 1,0 \end{bmatrix} = \begin{bmatrix} 0,707 \\ 0,707 \end{bmatrix} \quad \text{και} \quad Y_2 = 0,707 \begin{bmatrix} -1,0 \\ 1,0 \end{bmatrix} = \begin{bmatrix} -0,707 \\ 0,707 \end{bmatrix}$$

Για επαλήθευση μπορούμε να δούμε ότι:

$$\begin{bmatrix} 1,000 & 0,922 \\ 0,922 & 1,000 \end{bmatrix} \times \begin{bmatrix} 0,707 \\ 0,707 \end{bmatrix} = \begin{bmatrix} 1,359 \\ 1,359 \end{bmatrix} = 1,922 \begin{bmatrix} 0,707 \\ 0,707 \end{bmatrix} \text{ και}$$

$$\begin{bmatrix} 1,000 & 0,922 \\ 0,922 & 1,000 \end{bmatrix} \times \begin{bmatrix} -0,707 \\ 0,707 \end{bmatrix} = \begin{bmatrix} -0,0551 \\ 0,0551 \end{bmatrix} = 0,078 \begin{bmatrix} -0,707 \\ 0,707 \end{bmatrix}$$

Άρα οι τιμές 1,922 και 0,078 αποτελούν τις eigenvalues των αντίστοιχων ιδιοανυσμάτων. Τα ιδιοανύσματα αυτά μπορούν τώρα να παρουσιαστούν με τη μορφή του πίνακα U:

$$U_{2x2} = \begin{bmatrix} 0,707 & -0,707 \\ 0,707 & 0,707 \end{bmatrix} \qquad (1.11)$$

Υπολογίζουμε τώρα τις κύριες συνιστώσες Z_1, Z_2 με βάση την αρχική σχέση 1.7 και τον πίνακα 1.1. Έτσι για την PC1 έχουμε:

$$Z_1 = 0,707\ X_1 + 0,707\ X_2 \quad \text{και:}$$

$$z_1 = 0,707 \times (-0,471) + 0,707 \times (-1,265) = -1,23$$

……

$$z_5 = 0,707 \times (-1,414) + 0,707 \times (-0,632) = -1,45$$

και για την PC2:

$$Z_2 = -0{,}707\ X_1 + 0{,}707\ X_2 \quad \text{και:}$$

$$z_1 = -0{,}707 \times (-0{,}471) + 0{,}707 \times (-1{,}265) = -0{,}56$$

……

$$z_5 = -0{,}707 \times (-1{,}414) + 0{,}707 \times (-0{,}632) = 0{,}55$$

Όλες οι παραπάνω εξισώσεις μπορούν να απεικονιστούν με τη μορφή πινάκων:

$$Z = A \times U = \begin{bmatrix} -0{,}471 & -1{,}265 \\ 0{,}943 & 0{,}632 \\ 0{,}943 & 1{,}265 \\ 0{,}000 & 0{,}000 \\ 1{,}414 & -1{,}265 \end{bmatrix} \times \begin{bmatrix} 0{,}707 & -0{,}707 \\ 0{,}707 & 0{,}707 \end{bmatrix} \qquad (1.12)$$

1.2.3. Παράδειγμα κατανόησης 2 (PCA)

Ας θεωρήσουμε τώρα για παράδειγμα 12 ενώσεις (A, B, C ….. ως L), των οποίων μετράμε την ένταση φθορισμού σε τέσσερα (4) μήκη κύματος (300, 350, 400 και 450 nm) [10]. Ο πίνακας 1.2 απεικονίζει τις μετρήσεις.

Πίνακας 1.2: Ένταση φθορισμού σε 4 διαφορετικά μήκη κύματος σε 12 ενώσεις

Ένωση	**Μήκος κύματος (nm)**			
	300	**350**	**400**	**450**
A	16	62	67	27
B	15	60	69	31
C	14	59	68	31
D	15	61	71	31
E	14	60	70	30
F	14	59	69	30
G	17	63	68	29
H	16	62	69	28
I	15	60	72	30
J	17	63	69	27
K	18	62	68	28
L	18	64	67	29

Οι πίνακες 1.3 και 1.4 αποτελούν αντίστοιχα τους πίνακες διακυμάνσεων και συσχετίσεων της ανάλυσης. Τα έγχρωμα κελιά στον πίνακα 1.3 απεικονίζουν διακυμάνσεις και όχι συνδιακυμάνσεις, εφόσον αναφέρονται στην ίδια μεταβλητή και όχι σε ζεύγος διαφορετικών μεταβλητών. Επιπλέον, στον πίνακα 1.4, τα αντίστοιχα κελιά περιέχουν την τιμή 1,00, εφόσον περιγράφουν τη συσχέτιση της μεταβλητής με τον εαυτό της! Ο ίδιος πίνακας προκύπτει αν υπολογίσουμε τον πίνακα διακύμανσης των κανονικοποιημένων μεταβλητών.

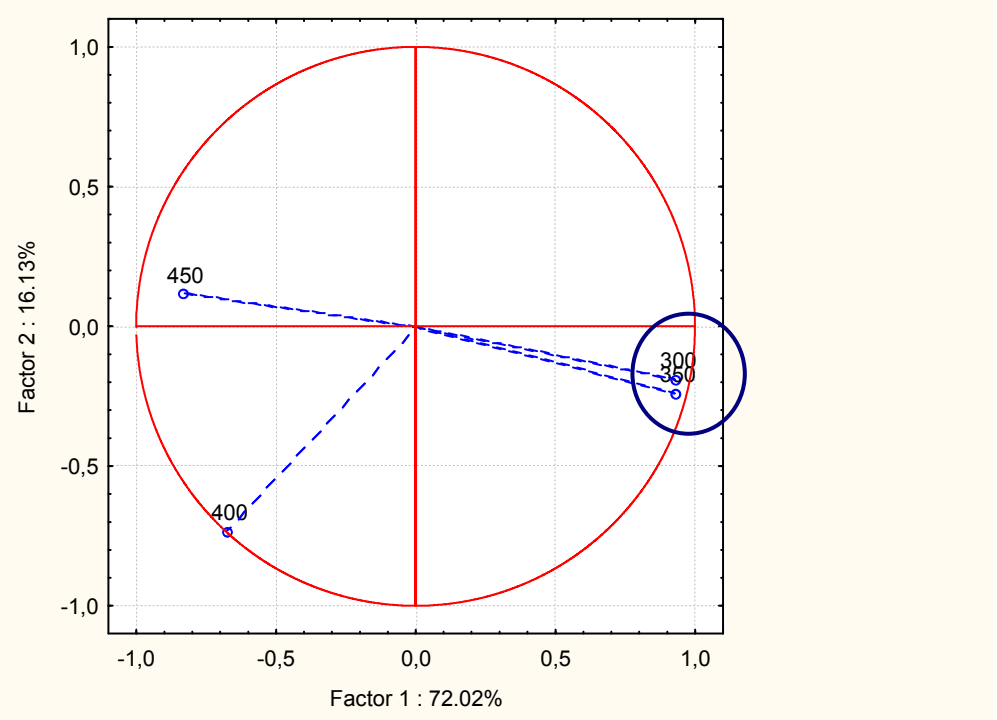
Πίνακας 1.3: Πίνακας διακυμάνσεων των δεδομένων του πίνακα 1.1

	300	**350**	**400**	**450**
300	2,20			
350	2,25	2,75		
400	-1,11	-1,16	2,26	
450	-1,48	-1,70	1,02	2,20

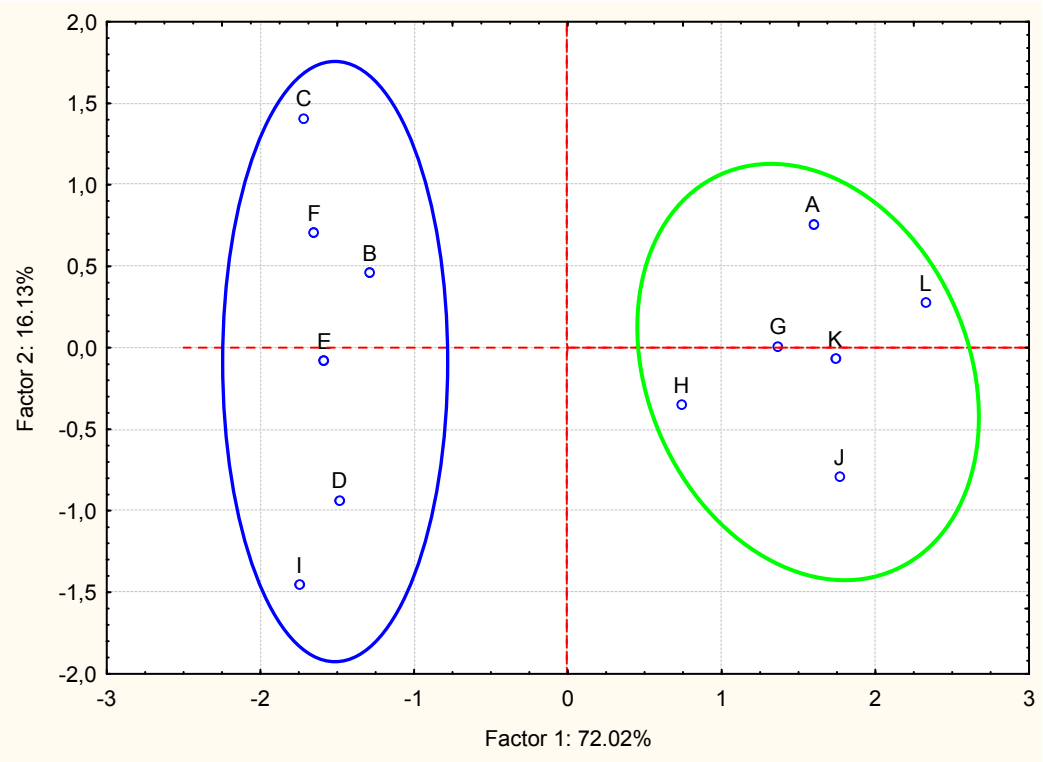
Πίνακας 1.4: Πίνακας συσχετίσεων των δεδομένων του πίνακα 1.1

	300	**350**	**400**	**450**
300	1,00			
350	0,914	1,00		
400	-0,498	-0,464	1,00	
450	-0,670	-0,692	0,458	1,00

Τα αντίστοιχα διαγράμματα φορτίσεων και σκορ φαίνονται στα σχήματα 1.3 και 1.4. Στο πρώτο σχήμα φαίνεται η συσχέτιση των αρχικών μεταβλητών (μηκών κύματος) και στο δεύτερο η συσχέτιση των αντικειμένων (χημικών ενώσεων). Έτσι γίνεται φανερό ότι από τα δυο πρώτα μήκη κύματος (300 και 350 nm), κάποιο θα μπορούσε να παραληφθεί γιατί “μεταφέρουν” την ίδια πληροφορία. Επίσης οι χημικές ενώσεις διαχωρίζονται φανερά σε δυο ομάδες: C, F, B, E, D, I και A, L, G, K, H και J.



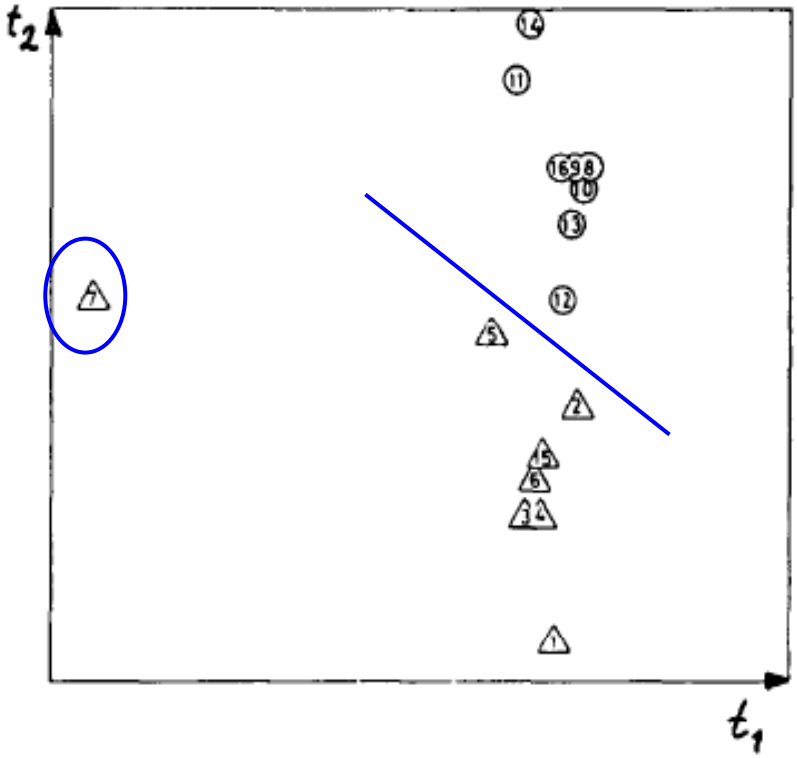
Σχήμα 1.3: Διάγραμμα φορτίσεων (αφορά τις 4 μεταβλητές)



Σχήμα 1.4: Διάγραμμα σκορ (αφορά τις 12 ενώσεις)

Με τον τρόπο αυτό, όχι μόνο επιτυγχάνεται η επιθυμητή κατάταξη, αλλά επιπλέον ανιχνεύονται έκτροπες τιμές, όπως για παράδειγμα στην περίπτωση φρέσκων και κατε-

ψυγμένων χυμών φρούτων με τον προσδιορισμό μιας σειράς παραμέτρων που τους χαρακτηρίζουν [17]. Οι δυο ομάδες διαχωρίζονται σαφώς (σχ. 1.5), ενώ ανιχνεύεται με επιτυχία το σημείο 7 ως έκτροπη τιμή.



Σχήμα 1.5: Διάγραμμα σκορ των δυο πρώτων κύριων συνιστωσών για το παράδειγμα των φρέσκων και κατεψυγμένων χυμών. Τα σημεία 1-7 και 15 είναι φρέσκα δείγματα, ενώ τα σημεία 8-14 και 16 είναι κατεψυγμένα [17].

1.3. ΑΝΑΛΥΣΗ ΠΑΡΑΓΟΝΤΩΝ (FACTOR ANALYSIS)

Μια εκτίμηση για το πόσο καλά μπορεί να δουλεύει ένα FA (Factor Analysis) μοντέλο, παίρνουμε από τις **συμμετοχικότητες [11]** (communalities) των μεταβλητών. Αυτές υπολογίζονται παίρνοντας το άθροισμα των τετραγώνων των φορτίσεων για κάθε μια μεταβλητή. Η συμμετοχικότητα για κάθε μεταβλητή δίνει το ποσοστό της διακύμανσης αυτής που είναι κοινό με τις άλλες μεταβλητές και ερμηνεύεται από το σύνολο των παραγόντων που έχουν επιλεχθεί [18]. Με άλλα λόγια, αν για μια μεταβλητή Var1, έχουμε:

communality Var1= 0,79,

σημαίνει ότι το 79 % της διακύμανσης της Var1, ερμηνεύεται από το συγκεκριμένο FA μοντέλο. Μπορούμε να σκεφτούμε ότι έχουμε ένα μοντέλο πολλαπλής παλινδρόμησης (multiple regression) με συντελεστή R^2 = 0,79 ανάμεσα στη μεταβλητή Var1 και τους επιλεγόμενους παράγοντες.

Για την αξιολόγηση του μοντέλου, απαιτείται οι τιμές των συμμετοχικοτήτων να είναι όσον το δυνατό κοντύτερα στη μονάδα. Αυτό θα σήμαινε ότι το μοντέλο ερμηνεύει τη μέγιστη διακύμανση αυτών των μεταβλητών [19]. Μικρή τιμή συμμετοχικότητας για κάποια μεταβλητή, δείχνει ότι προφανώς το μοντέλο δεν περιγράφει επαρκώς τη συγκεκριμένη μεταβλητή και θα έπρεπε αυτή να αποσυρθεί από την ανάλυση [20].

1.4. ΑΝΑΛΥΣΗ ΚΑΤΑ ΣΥΣΤΑΔΕΣ (CLUSTER ANALYSIS)

1.4.1. Συσταδοποίηση Κ-μέσων σημείων (K-means clustering)

Ο K-means αλγόριθμος μπορεί να θεωρηθεί μη επιβλεπόμενος με την έννοια ότι ταξινομεί κάθε δείγμα αυτόματα με βάση το κριτήριο της ελάχιστης απόστασης από κάποιο κεντροειδές σημείο. Δεν χρειάζεται να επιβλέπουμε το σύστημα ελέγχοντας αν η ταξινόμηση είναι σωστή ή όχι. Παρόλα αυτά υπάρχουν δυο πορείες στη πορεία εκπαίδευσης:

1. Συνεχής εκπαίδευση (infinite training): κάθε δείγμα που εισέρχεται θεωρείται μέρος της διαδικασίας εκπαίδευσης, έτσι ώστε να ορίζονται νέα κεντρικά σημεία.
2. Πεπερασμένη εκπαίδευση (finite training): η διαδικασία εκπαίδευσης θεωρείται ολοκληρωμένη μετά τα πρώτα δείγματα και τα νέα εισερχόμενα ταξινομούνται στις υπάρχουσες ομάδες. Τα κεντρικά σημεία είναι σταθερά [21].

Η τελική ομαδοποίηση (ειδικά για μεγάλο αριθμό δειγμάτων [21]) εξαρτάται από τα αρχικώς υποδεικνυόμενα ως κεντροειδή των ομάδων αλλά και την τιμή του K. Το τελευταίο είναι πολύ σημαντικό και απαιτεί γνώση του αριθμού των ομάδων των δεδομένων, γεγονός που θεωρείται μάλλον απίθανο [4, 22, 23]. Επιπλέον, ο αριθμητικός μέσος που χρησιμοποιείται για τους υπολογισμούς των συντεταγμένων των κεντρικών σημείων είναι ευαίσθητος σε έκτροπες τιμές [21]. Ωστόσο, ο K-means αλγόριθμος δουλεύει πολύ καλά με μεγάλες βάσεις δεδομένων, χωρίς να απαιτείται πολύς χρόνος ή υπολογιστική (computational) δύναμη. Επιπλέον δημιουργεί ομοιόμορφες στο μέγεθος ομάδες [4].

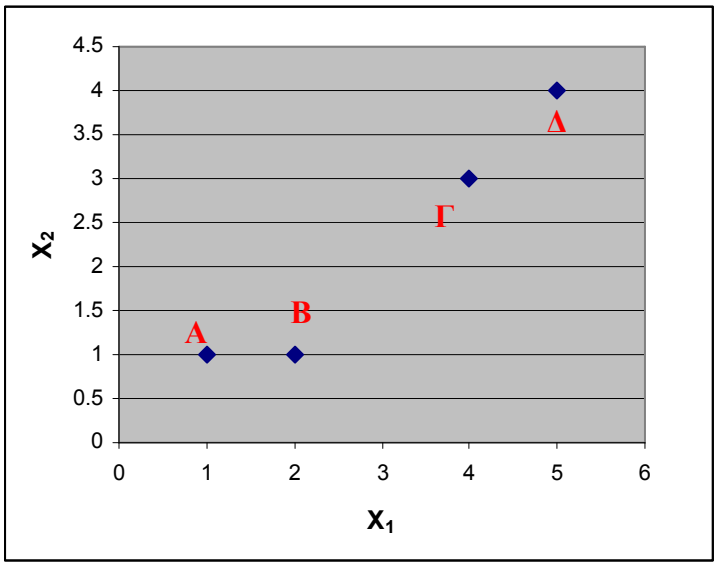
1.4.2. Παράδειγμα κατανόησης (K-means αλγόριθμος)

Θα αναφερθούμε τώρα σε ένα παράδειγμα για την πλήρη κατανόηση του K-means αλγορίθμου. Ας υποθέσουμε ότι έχουμε 4 δείγματα (Α, Β, Γ, Δ) που περιγράφονται από δυο μεταβλητές (X_1, X_2) το καθένα και πρέπει να ταξινομηθούν σε δυο ομάδες (πίνακας 1.5) [21].

Πίνακας 1.5: Δείγματα προς ταξινόμηση

	X_1	X_2
Α	1	1
Β	2	1
Γ	4	3
Δ	5	4

Κάθε δείγμα αντιπροσωπεύει ένα σημείο στο παρακάτω σχήμα (1.6).



Σχήμα 1.6: Παράδειγμα κατανόησης του K-means αλγορίθμου [21].

Ας θεωρήσουμε τα σημεία Α, Β τα πρώτα κεντρικά. Υπολογίζουμε την Ευκλείδεια απόσταση κάθε σημείου από αυτά (πρώτο μισό του πίνακα 1.6). Με βάση την ελάχιστη απόσταση, τα σημεία ταξινομούνται στη σωστή ομάδα (δεύτερο μισό του πίνακα 1.6).

Έτσι, το σημείο Α ανήκει στην $1^{η}$ ομάδα (δηλώνεται από τον αριθμό 1 στο αντίστοιχο κελί) και τα υπόλοιπα στην $2^{η}$ ομάδα.

Πίνακας 1.6: Υπολογισμός αποστάσεων και ταξινόμηση (iteration 0)

	Α	**Β**	**Γ**	**Δ**
αποστάσεις				
Α	0	1	3,61	5
Β	1	0	2,83	4,24
ταξινόμηση				
Α	1	0	0	0
Β	0	1	1	1

Υπολογίζονται τώρα οι συντεταγμένες των νέων κεντρικών σημείων:

$1^{η}$ ομάδα: (1, 1) και

$$2^{η} \text{ ομάδα: } (\frac{2+4+5}{3}, \frac{1+3+4}{3}) = (\frac{11}{3}, \frac{8}{3})$$

Επαναλαμβάνουμε την παραπάνω διαδικασία υπολογίζοντας τις νέες αποστάσεις των σημείων από τα κεντρικά και ταξινομώντας εκ νέου τα δείγματα (πίνακας 1.7).

Πίνακας 1.7: Υπολογισμός αποστάσεων και ταξινόμηση ($1^{ος}$ κύκλος)

	Α	**Β**	**Γ**	**Δ**
αποστάσεις				
Α	0	1	3,61	5
Β	3,14	2,36	0,47	1,89
ταξινόμηση				
Α	1	1	0	0
Β	0	0	1	1

Υπολογίζονται εκ νέου οι συντεταγμένες των νέων κεντρικών σημείων:

$$1^{η} \text{ ομάδα: } (\frac{1+2}{2}, \frac{1+1}{2}) = (1{,}5,\ 1) \text{ και}$$

$$2^{η} \text{ ομάδα: } (\frac{4+5}{2}, \frac{3+4}{2}) = (4{,}5,\ 3{,}5)$$

Επαναλαμβάνουμε τη διαδικασία υπολογίζοντας τις νέες αποστάσεις των σημείων από τα κεντρικά και ταξινομώντας εκ νέου τα δείγματα (πίνακας 1.8).

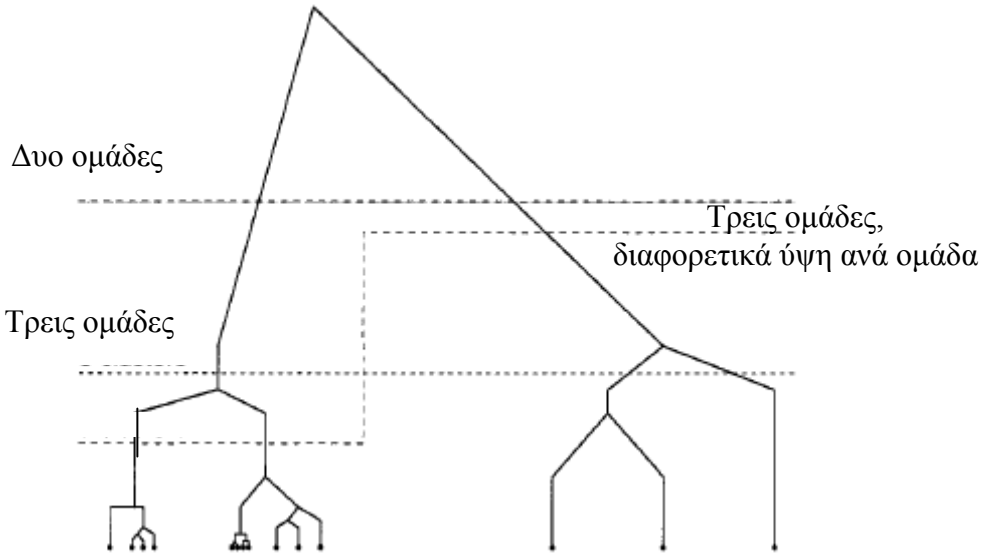
Πίνακας 1.8: Υπολογισμός αποστάσεων και ταξινόμηση ($2^{ος}$ κύκλος)

	Α	**Β**	**Γ**	**Δ**
αποστάσεις				
Α	0,5	0,5	3,20	4,61
Β	4,30	3,54	0,71	0,71
ταξινόμηση				
Α	1	1	0	0
Β	0	0	1	1

Η ταξινόμηση δεν αλλάζει και έτσι παίρνουμε τα τελικά αποτελέσματα: τα σημεία Α, Β ανήκουν στην πρώτη ομάδα και τα σημεία Γ, Δ στη δεύτερη.

1.4.3. Ιεραρχική Ανάλυση κατά συστάδες (Hierarchical Cluster Analysis)

Το δενδρόγραμμα στην Cluster Analysis (CA) δεν περιγράφει μια μοναδική ομαδοποίηση [24]. Αντίθετα υπάρχουν διάφοροι τρόποι (σχ. 1.7) για να "κοπεί" και διαφορετικές ομαδοποιήσεις επιτυγχάνονται "κόβοντας" το δέντρο σε διάφορα ύψη (σχ. 1.7). Το πιο ευνόητο ωστόσο, είναι να κοπεί το δέντρο σε σημείο που να υπάρχει αρκετή απόσταση από τις δυο ομάδες που συγχωνεύονται. Παρόλα αυτά, η ομαδοποίηση αυτή αγνοεί το γεγονός της διαφορετικής απόστασης μεταξύ των υπο-ομάδων που μπορεί να περιέχει μια ομάδα. Έτσι, θα μπορούσαν οι δυο κλάδοι ενός δέντρου να κοπούν σε διαφορετικά ύψη (σχ. 1.7).

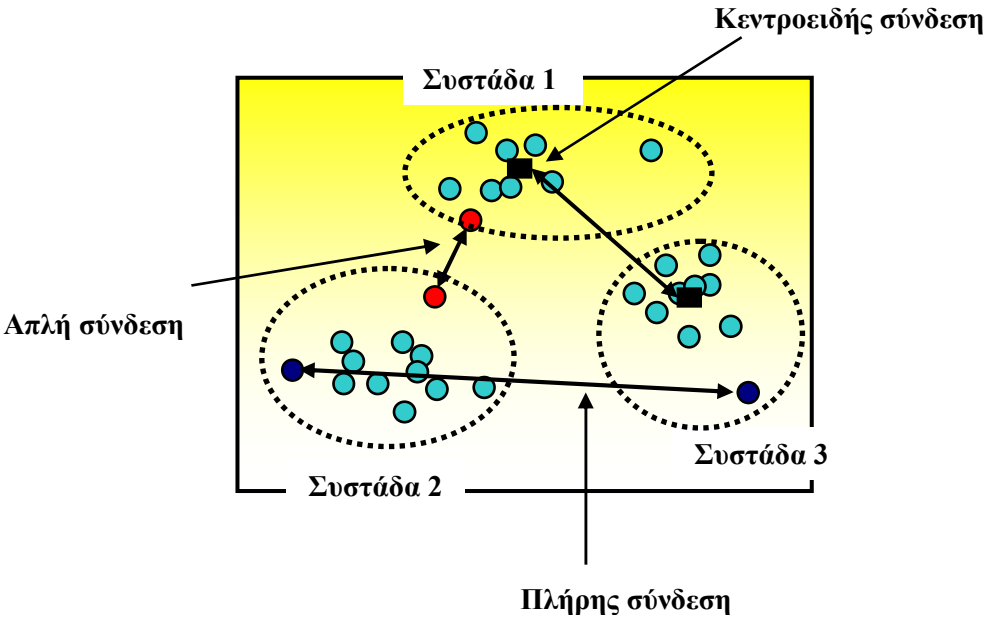


Σχήμα 1.7: Δενδρόγραμμα 14 δειγμάτων. Το δέντρο μπορεί να κοπεί σε διάφορα ύψη [24].

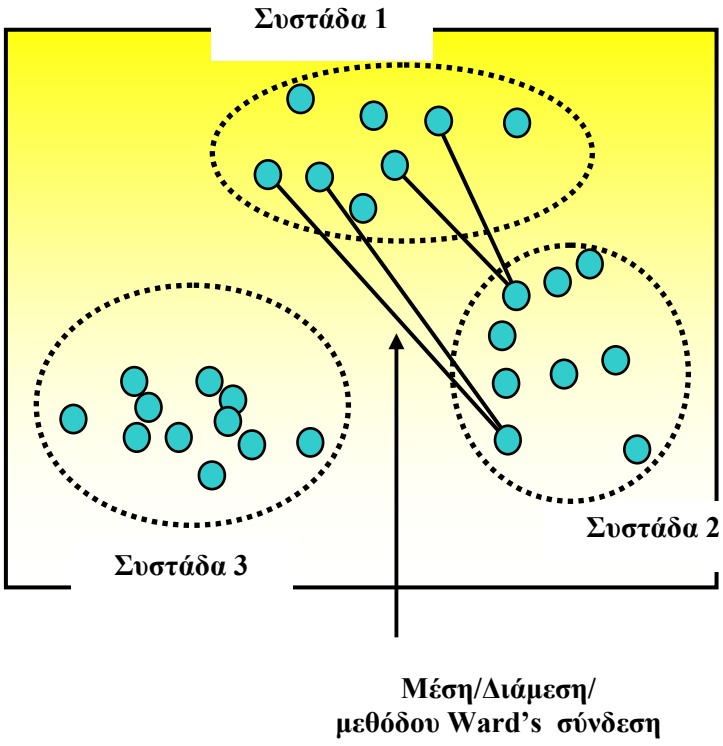
Ο υπολογισμός των αποστάσεων μεταξύ των αντικειμένων γίνεται με την βοήθεια διαφόρων αλγορίθμων που καθορίζουν τον τρόπο σύνδεσης των ομάδων [16].

Μερικοί από τους αλγόριθμους αυτούς είναι [25]:

1. **Απλή σύνδεση ή κοντινότερων γειτόνων** (single linkage ή nearest neighbors): η απόσταση μεταξύ δύο ομάδων, είναι η απόσταση μεταξύ των δύο κοντινότερων μελών των ομάδων (σχ. 1.8)
2. **Πλήρης σύνδεση ή μακρινότερων γειτόνων** (complete linkage ή furthest neighbors): η απόσταση μεταξύ δύο ομάδων, είναι η απόσταση μεταξύ των δύο πιο απομακρυσμένων μελών των ομάδων (σχ. 1.8).
3. **Κεντροειδής σύνδεση** (centroid linkage): η απόσταση μεταξύ δύο ομάδων, είναι η απόσταση μεταξύ των πολυπαραμετρικών μέσων των ομάδων (σχ. 1.8).
4. **Μέση σύνδεση** (average linkage): η απόσταση μεταξύ δύο ομάδων, είναι η απόσταση μεταξύ όλων των μελών των δύο ομάδων (σχ. 1.9).
5. **Διάμεση σύνδεση** (median linkage): η απόσταση μεταξύ δύο ομάδων, είναι η διάμεση απόσταση μεταξύ όλων των μελών των δύο ομάδων (σχ. 1.9).
6. **Μεθόδου Ward's σύνδεση** (Ward's method linkage): η απόσταση μεταξύ δύο ομάδων, είναι η μέση απόσταση μεταξύ όλων των μελών των δύο ομάδων, λαμβάνοντας υπόψη τις συνδιακυμανσεις (σχ. 1.9).



Σχήμα 1.8: Απλή και πλήρης σύνδεση ομάδων [25].



Σχήμα 1.9: Μέση, διάμεση και σύνδεση Ward's μεθόδου [25].

1.4.4. Ο δείκτης Davies-Bouldin (DB)

Ο σημαντικότερος δείκτης που χρησιμοποιείται για τη αξιολόγηση των μεθόδων συσταδοποίησης είναι ο Davies-Bouldin (DB) δείκτης. Χρησιμοποιείται και στα δίκτυα Kohonen για την εύρεση του βέλτιστου αριθμού ομάδων. Ο ορισμός του δίνεται από τις παρακάτω σχέσεις [22]:

$$s_i = \frac{1}{C_i}\sum_{x\in C_i}\|\mathbf{x}-\mathbf{z_i}\| \qquad d_{ij} = \|\mathbf{z_i}-\mathbf{z_j}\| \qquad R_{ij} = \frac{s_i + s_j}{d_{ij}}$$
$$DB_{index} = \frac{1}{K}\sum_{i=1}^{K} R_i \qquad \text{και} \qquad R_i = \max_{i=1,\dots K, i\neq j} R_{ij} \tag{1.13}$$

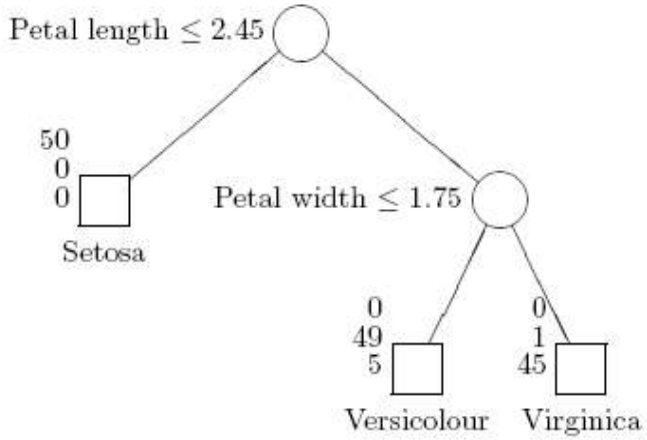
όπου **x** δείγμα της ομάδας C_i, $\mathbf{z_i}$, $\mathbf{z_j}$ κεντροειδή των ομάδων C_i και C_j αντίστοιχα, C_i ο αριθμός των δειγμάτων της ομάδας, K ο συνολικός αριθμός των ομάδων.

Από τις παραπάνω σχέσεις, είναι φανερό ότι το πηλίκο s_i εκφράζει την ομοιότητα μεταξύ των στοιχείων μιας ομάδας και του κεντροειδούς αυτής (within-cluster distance). Η διαφορά d_{ij} εκφράζει την απόσταση μεταξύ των κεντροειδών διαφορετικών ομάδων (between-clusters distance). Τέλος, ο DB δείκτης εκφράζει τη μέση ομοιότητα ανάμεσα στην κάθε ομάδα C_i και στην πιο κοντινή της. Ωστόσο είναι επιθυμητό, οι ομάδες να έχουν όσο το δυνατό τη μικρότερη ομοιότητα μεταξύ τους. Άρα, η βέλτιστη συσταδοποίηση ψάχνει για το μικρότερο DB δείκτη.

1.5. ΔΕΝΤΡΑ ΤΑΞΙΝΟΜΗΣΗΣ (CLASSIFICATION TREES, CT)

1.5.1. Παράδειγμα κατανόησης

Ως απλό παράδειγμα Δέντρων Ταξινόμησης (CT), αναφέρεται η χρήση τους στην κατάταξη των τριών (3) ποικιλιών (*Setosa, Versicolour, Virginica*) του φυτού Iris. Τα δεδομένα αποτελούν πενήντα (50) δείγματα από κάθε ποικιλία του φυτού. Οι παράμετροι που μετρήθηκαν ήταν τέσσερις: μήκος και πλάτος για τα σέπαλα, μήκος και πλάτος για τα πέταλα. Ο πρώτος διαχωρισμός στο "δέντρο", έγινε με βάση το μήκος των πετάλων και ο δεύτερος το πλάτος αυτών (σχ. 1.10). Οι δυο αυτές παράμετροι αποτελούν τις κρίσιμες για το διαχωρισμό μεταβλητές (split variables) με βάση τις κρίσιμες τιμές αυτών (split constants). Έξι (6) από τα 150 δείγματα ταξινομήθηκαν σε λάθος κατηγορία (mis-classification), δίνοντας 4% συνολικό σφάλμα [26].



Σχήμα 1.10: Δέντρο Ταξινόμησης για τα δεδομένα του λουλουδιού Iris. Η τριπλέτα κοντά σε κάθε τερματικό κόμβο, δίνει τον αριθμό των περιπτώσεων Setosa, Versicolour, Virginica, αντίστοιχα σε καθένα από αυτούς [26].

1.5.2. Δείκτης Gini

Η CART μέθοδος αξιολογεί όλες τις μεταβλητές για να καθοριστούν τελικά αυτές που δημιουργούν τους καλύτερους διαχωρισμούς, δηλαδή εκείνους με τις πιο "καθαρόαιμες" τάξεις στους κόμβους με τη βοήθεια του **δείκτη Gini** [27]. Ο δείκτης αυτός ισούται με μηδέν (0) όταν όλα τα δείγματα σε ένα κόμβο ανήκουν σε μία μόνο τάξη. Υπολογιστικά ισούται με το άθροισμα των γινομένων όλων των ζευγών των αναλογιών / κλασμάτων των τάξεων που βρίσκονται στον κόμβο [7, 28]:

$$i(t) = 1 - \sum_{j=1}^{k} (P_j\ (t))^2 \qquad (1.14)$$

όπου, i(t) είναι η καθαρότητα ενός κόμβου t, και $P_j(t)$ είναι το κλάσμα των δειγμάτων του κόμβου t που ανήκουν στην j ομάδα από τις k που είναι παρούσες σε αυτόν.

1.5.3. Εφαρμογές των Δέντρων Ταξινόμησης

Η τεχνική των CT, σπανίως έχει χρησιμοποιηθεί και μόνο πρόσφατα, για την εκτίμηση της ποιότητας του νερού [27, 29, 30, 31]. Οι Simeonova και Simeonov [29] εφάρμοσαν τις μεθόδους των PCA και CART σε μια μεγάλη βάση δεδομένων για την εκτίμηση της ποιότητας του νερού σε επεξεργασμένο και μη νερό, χρησιμοποιώντας μόνο τέσσερεις φυσικοχημικές παραμέτρους. Η Stanimirova [30] εφάρμοσε CART για να

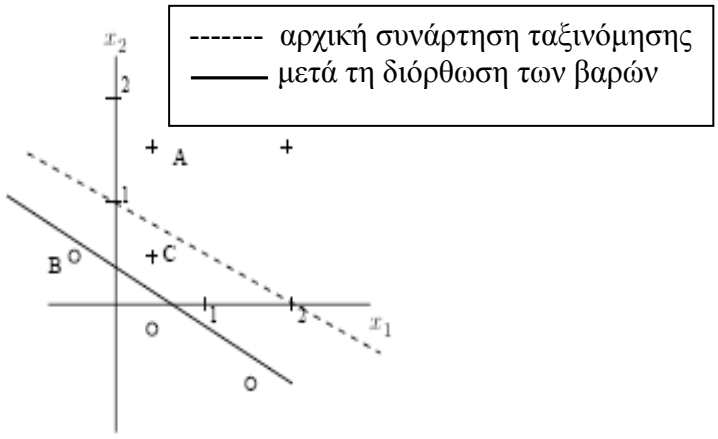
ταξινομήσει δείγματα biofilms με ακρίβεια 100 %, χρησιμοποιώντας μόνο τη συγκέντρωση του Mg, και αποδεικνύοντας τη χρησιμότητα της μεθόδου στη ανάπτυξη κανόνων ταξινόμησης. Άλλη μια εφαρμογή των CT, χρησιμοποιεί την τεχνική αυτή για την εκτίμηση της οικολογικής κατάστασης επιφανειακών νερών [31]. Πιο πρόσφατα, η τεχνική των CART χρησιμοποιήθηκε για την εύρεση των κρίσιμων μεταβλητών (μεταξύ των οποίων αγωγιμότητα, pH, υπολειμματικό Al, σκληρότητα, Cl^{-}, Ca^{2+}, Mg^{2+}) που μπορούν να διαχωρίσουν τις τέσσερις (4) μονάδες επεξεργασμένου νερού της ΕΥΔΑΠ [27] και την εκτίμηση της μόλυνσης από βαρέα μέταλλα σε χώματα [28].

ΚΕΦ. 2 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (ARTIFICIAL NEURAL NETWORKS, ANN)

2.1. ΓΕΝΙΚΗ ΘΕΩΡΙΑ ΚΑΙ MULTI-LAYER PERCEPTRON (MLP)

2.1.1. Παράδειγμα κατανόησης (perceptron)

Θεωρούμε ένα perceptron με αρχικά βάρη: $w_1 = 1$, $w_2 = 2$ (για λόγους ευκολίας χρησιμοποιούνται μικροί ακέραιοι αριθμοί) και $b = -2$ [32]. Ο κανόνας εκπαίδευσης του perceptron χρησιμοποιείται για να βρεθεί μια συνάρτηση ταξινόμησης για τα δείγματα A, B, C του σχήματος 2.1.



Σχήμα 2.1: Συνάρτηση ταξινόμησης πριν και μετά τη διόρθωση των βαρών [32]

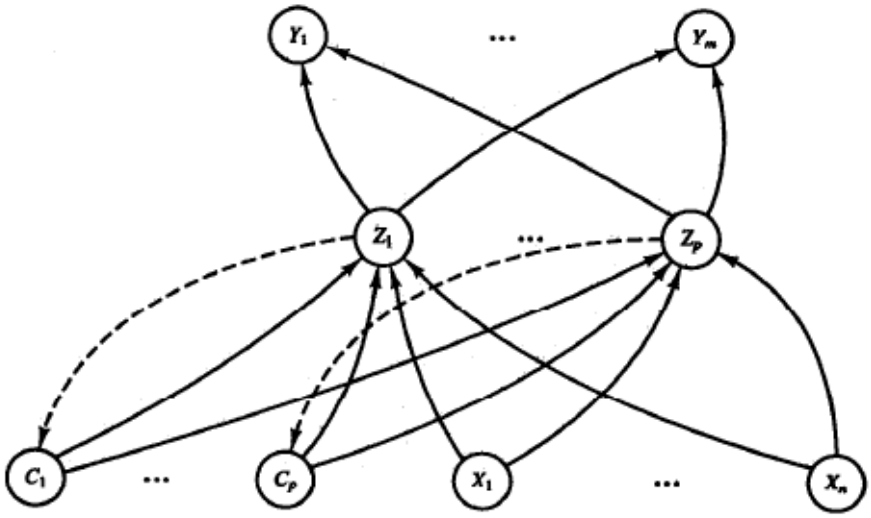
Το πρώτο δείγμα A με τιμές (0,5, 1,5) και επιθυμητή απόκριση d(x) = +1 εισάγεται στο δίκτυο. Από τις σχέσεις που ορίζουν το perceptron, μπορεί να υπολογιστεί ότι το εξερχόμενο είναι το επιθυμητό και επομένως τα βάρη δεν αλλάζουν:

$$\Sigma_A = 1\times0{,}5 + 2\times1{,}5 + (-2) = 1{,}5 > 0 \rightarrow F(\Sigma_A) = +1$$

Το ίδιο συμβαίνει για το δείγμα B με τιμές (-0,5, 0,5) και επιθυμητή απόκριση d(x) = -1. Όταν όμως εισάγεται στο δίκτυο το δείγμα C, με τιμές (0,5, 0,5) και επιθυμητή απόκριση d(x) = +1, το δίκτυο θα δώσει -1. Σύμφωνα με τον κανόνα εκπαίδευσης, οι αλλαγές των βαρών θα είναι: $\Delta w_1 = 0{,}5$, $\Delta w_2 = 0{,}5$ και $\Delta b = 1$. Τα νέα βάρη είναι: $w_1 = 1{,}5$, $w_2 = 2{,}5$ και $\Delta b = -1$ και τελικά το δείγμα C ταξινομείται σωστά.

2.1.2. Αναδρομικά δίκτυα (Recurrent)

Τα Αναδρομικά δίκτυα (Recurrent) (σχ. 2.2) σε αντίθεση με τα Εμπροσθοτροφοδοτούμενα δίκτυα (Feed-forward), περιέχουν και συνδέσεις ανατροφοδότησης (feedback). Στα δίκτυα αυτά, κάποιο ή περισσότερα από τα εισερχόμενα (σε χρόνο t) είναι τα εξερχόμενα του δικτύου (σε χρόνους t-1 ή t-2) [33]. Έτσι, το αποτέλεσμα δεν εξαρτάται μόνο από τα εισερχόμενα, αλλά και τα εξερχόμενα της προηγούμενης περιόδου [33, 34].



Σχήμα 2.2: Απεικόνιση ενός απλού αναδρομικού δικτύου. Οι μονάδες που σημειώνονται με c (c1.....cp), δέχονται επιπλέον των feed-forward και feedback πληροφορίες [34].

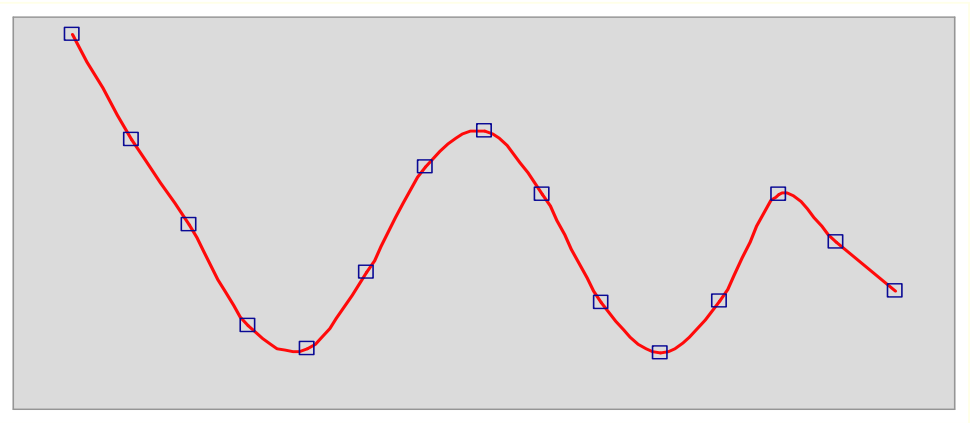
2.1.3. Παράδειγμα κατανόησης (φαινόμενο υπερ-προσαρμογής)

Μπορούμε να φανταστούμε το πρόβλημα της υπερ-προσαρμογής, χρησιμοποιώντας μια πολυωνυμική συνάρτηση. Οι συναρτήσεις αυτές περιέχουν σταθερές (παραμέτρους) και δυνάμεις των μεταβλητών, όπως για παράδειγμα:

$$y = 2x + 5 \text{ και } y = 2x^2 + 5x + 3$$

Διαφορετικά πολυώνυμα έχουν διαφορετικά σχήματα, ενώ συναρτήσεις με μεγαλύτερες δυνάμεις για τις μεταβλητές, έχουν πολυπλοκότερα σχήματα και περισσότερους όρους, σε αντιδιαστολή με τα περισσότερα βάρη των πολυπλοκότερων νευρωνικών μοντέλων. Για μια δεδομένη ομάδα δειγμάτων, μπορεί να χρειαστούμε μια πολυωνυμική καμπύλη (μοντέλο) για να ερμηνεύσουμε ή να προσομοιάσουμε τα δεδομένα. Προφανώς επίσης, υπάρχει θόρυβος στα δεδομένα και έτσι δεν περιμένουμε απαραίτητα η βέλτιστη καμπύλη να "περάσει" από όλα τα σημεία. Ένα μικρότερης δύναμης πολυώνυμο, μπορεί να μην είναι

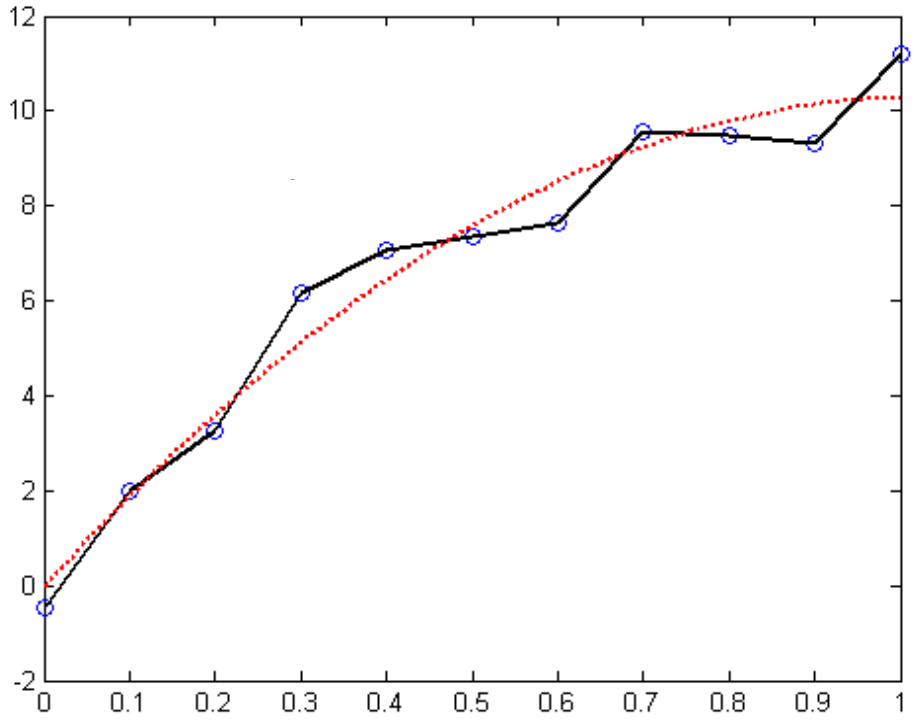
εξίσου ευέλικτο για να περάσει από όλα τα σημεία, ενώ αντίθετα, ένα μεγαλύτερης δύναμης είναι στην πραγματικότητα τόσο ευέλικτο ή προσαρμοστικό, ώστε "προσεγγίζει" ακριβώς τα σημεία, υιοθετώντας ένα τόσο παράδοξο σχήμα, το οποίο μπορεί τελικά να μη σχετίζεται με την πραγματική συνάρτηση που περιγράφει τη σχέση των δεδομένων (σχ. 2.3) [7].



Σχήμα 2.3: Πολύπλοκη συνάρτηση που "περιγράφει" ακριβώς τα δεδομένα.

2.1.4. Ομαλοποίηση (regularization)

Οι μέθοδοι που εφαρμόζονται για την αποφυγή φαινομένων υπερ-προσαρμογής στα μοντέλα ANN περιγράφονται συνολικά με τον όρο **"ομαλοποίηση"** ("regularization") (σχ. 2.4) [7, 35]. Με τον τρόπο αυτό, τα δεδομένα λαμβάνονται λιγότερο υπόψη από το μοντέλο.



Σχήμα 2.4: Ομαλοποίηση (κόκκινη στικτή γραμμή): όταν τα δεδομένα λαμβάνονται λιγότερο υπόψη από το μοντέλο [36].

Η διόρθωση των βαρών με τη χρήση ενός επιπλέον όρου (peanalty term ή factor), για παράδειγμα, συνίσταται κάτω από το γενικότερο τίτλο **"weight decay"** [35] ή **"Weigend weight regularization"** [7]. Σε αυτήν τη μέθοδο, προστίθεται ένας όρος στη συνάρτηση που μετρά το σφάλμα του μοντέλου, ώστε να "ποινικοποιούνται" τα μεγάλα βάρη, που συνήθως οδηγούν σε υπερ-προσαρμογή των μοντέλων, με τη δημιουργία εντονότερων καμπυλών.

Η κατά Bayesian ομαλοποίηση εξάλλου, ανιχνεύει και απομακρύνει κάθε "περίσσεια" παραμέτρων στο μοντέλο, έτσι ώστε να βελτιστοποιούνται βάρη και λοιπές παράμετροι, ισορροπώντας το σφάλμα του μοντέλου με την πολυπλοκότητα αυτού, ή διαφορετικά μεγιστοποιώντας συγχρόνως την ακρίβεια (bias) και τη διακύμανση/ευελιξία (variance) του μοντέλου [35]. Η ακρίβεια και η ευελιξία του μοντέλου σχετίζονται με την πολυπλοκότητα του μοντέλου με την έννοια ότι ένα πολύπλοκο μοντέλο που "περιγράφει" ακριβώς τα δεδομένα έχει μηδενικό σφάλμα. Ωστόσο, η συνάρτηση που το αντιπροσωπεύει αλλάζει δραματικά όταν αυτό πρέπει να προσαρμοστεί σε νέα δεδομένα. Αντίθετα, ένα μοντέλο που "αγνοεί" πλήρως τα δεδομένα και αυθαίρετα επιλέγει μια συνάρτηση, έχει μεγάλη ευελιξία (low variance) αλλά και υψηλό σφάλμα (high bias) [7].

Η Bayesian ομαλοποίηση μπορεί να χρησιμοποιήσει όλα τα διαθέσιμα δείγματα ως ομάδα εκπαίδευσης, χωρίς την ανάγκη δημιουργίας άλλων ομάδων (βλ. παρακάτω) [35].

Μια διαφορετική ομαλοποίησης μέθοδο αποτελεί η **"προσθήκη θορύβου"** ("noise injection") στην ομάδα εκπαίδευσης. Με τον τρόπο αυτό, ένα διάνυσμα θορύβου προστίθεται σε κάθε δείγμα εκπαίδευσης μετά το τέλος κάθε περιόδου. Αυτό προκαλεί μια "τρεμούλα" ή "θολούρα" (jitter) στα δείγματα με αποτέλεσμα τα μοντέλα ANN να δυσκολεύονται να βρουν μια λύση που να τα περιγράφει ακριβώς [37]. Έτσι βελτιώνεται η γενίκευση των μοντέλων [38, 39].

Η **"μέθοδος του πρώιμου τερματισμού"** ("early stopping method") εξάλλου, που αναφέρεται στην διατριβή, αποτελεί επίσης μια μέθοδο ομαλοποίησης.

2.1.5. Έλεγχος του φαινομένου της υπερ-προσαρμογής

Ένας κανόνας που προτείνεται για τον έλεγχο της υπερ-προσαρμογής, είναι ότι αυτή είναι πολύ πιθανό να υφίσταται όταν ο αριθμός των μεταβλητών v υπερβαίνει το (n - g)/3, όπου n είναι ο αριθμός των δειγμάτων και g ο αριθμός των ομάδων (εξερχόμενα για δίκτυα ταξινόμησης) [40]. Οι Zupan και Gasteiger [12], οι Maier και Dandy [41], ο Peres et al. [42] και η Curteanu et al. [43] και ο Palani et al. [44] αναφέρουν επίσης, ότι ο αριθμός των δειγμάτων εκπαίδευσης n_t πρέπει να είναι τουλάχιστον μεγαλύτερος από τον αριθμό των βαρών w_n, ενώ για τον Palani στα δίκτυα μίας ενδιάμεσης στιβάδας, ο αριθμός h των μονάδων αυτής, πρέπει να κυμαίνεται μεταξύ v και $2\times v + 1$. Επιπλέον, ο αριθμός h δεν θα πρέπει να είναι μικρότερος από v/3 ή τον αριθμό m των εξερχομένων. Οι Huang και Foo [45] παραλλάσσουν την παραπάνω σχέση και προτείνουν ο αριθμός h των μονάδων της ενδιάμεσης στιβάδας, να κυμαίνεται από $2\times\sqrt{v} + m$ ως $2\times v + 1$. Γενικά, ένα πλήθος συγγραφέων θεωρούν ότι ο βέλτιστος αριθμός των ενδιάμεσων μονάδων εξαρτάται από τον αριθμό των εισερχόμενων και εξερχόμενων μεταβλητών και έτσι προτείνεται ο αριθμός h των μονάδων της ενδιάμεσης στιβάδας να ισούται με v + m [46], ή (v + m)/2 [7, 47], ή (v × m)/2 [47], ή ακόμα να κυμαίνεται σε διαστήματα όπως από $2\times v + 1$ ως $m\times(v + 1)$ [48] ή από $-1/w$ μέχρι $1/w$, όπου w ο συνολικός αριθμός των βαρών προς τη στιβάδα αυτή [12]. Οι Fernandes και Lona [33] ταξινομούν και υπολογίζουν τον αριθμό των ενδιάμεσων νευρώνων και στιβάδων με βάση την αναλογία εισερχομένων και εξερχομένων (βλ. παρακάτω). Ο Zhang et al. [49] δίνει ένα εύρος για τον αριθμό των μονάδων στις ενδιά-

μεσες στιβάδες με βάση τη σχέση h = $\sqrt{i+m+1}$ + a, όπου a θετικός αριθμός που κυμαίνεται από 1 ως 10 καθορίζοντας τον ελάχιστο και μέγιστο αριθμό των νευρώνων.

Ο Darnag et al. [50] αναφέρεται στην παράμετρο ρ = n_t/ w_n και περιορίζει την τιμή της στο διάστημα: 1 (ή 0,9) < ρ < 2,2 (ή 3). Αν ρ << 1, το μοντέλο απομνημονεύει τα δείγματα εκπαίδευσης, ενώ αν ρ >> 3, το μοντέλο δεν μπορεί να γενικεύσει. Έτσι, με βάση την παραπάνω σχέση, προσδιορίζεται το εύρος των τιμών για τον αριθμό των μονάδων της ενδιάμεσης στιβάδας. Πολλαπλές παραλλαγές των παραπάνω αναφέρονται και αλλού [41].

Οι Fernandes και Lona [33] χωρίζουν τα μοντέλα ANN σε τρεις ομάδες ανάλογα με την αναλογία του αριθμού εισερχομένων και εξερχομένων και προτείνουν πρακτικούς κανόνες για τον αριθμό των ενδιάμεσων στιβάδων και νευρώνων.

Έτσι, τα MLP δίκτυα μπορούν να διαχωριστούν σε τρεις ομάδες (σχ. 2.4):

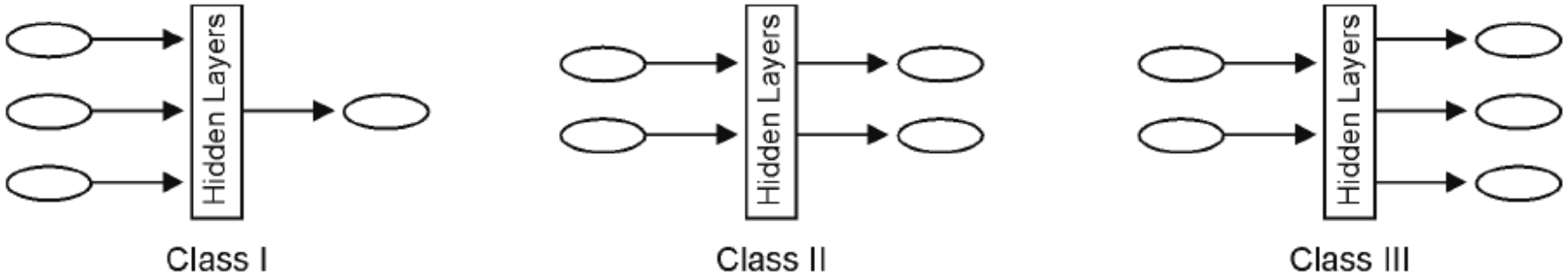
1. Ομάδα I (Class I): αναφέρεται σε δίκτυα τα οποία έχουν περισσότερες εισερχόμενες μεταβλητές από εξερχόμενες.
2. Ομάδα II (Class II): αναφέρεται σε δίκτυα τα οποία έχουν ίσο αριθμό εισερχομένων και εξερχομένων μεταβλητών.
3. Ομάδα I (Class I): αναφέρεται σε δίκτυα τα οποία έχουν λιγότερες εισερχόμενες από εξερχόμενες μεταβλητές.

Για την ομάδα I, οι συγγραφείς θεωρούν ότι στις περισσότερες περιπτώσεις, μία μόνο ενδιάμεση στιβάδα είναι αρκετή και μάλιστα αν N είναι ο αριθμός των εισερχομένων μεταβλητών και N-1 νευρώνες χρησιμοποιηθούν στην μοναδική ενδιάμεση στιβάδα, το μοντέλο μπορεί να δώσει εντυπωσιακές προβλέψεις. Αυτή η σύσταση έχει αποτέλεσμα για μικρό αριθμό εισερχομένων και η συνάρτηση μεταξύ των εισερχομένων και εξερχομένων δεν είναι πολύπλοκη. Διαφορετικά, οι συγγραφείς συνιστούν 8 – 20 νευρώνες στην ενδιάμεση στιβάδα για καλές προβλέψεις και μικρούς χρόνους εκπαίδευσης.

Όταν ο αριθμός των εξερχομένων είναι ίσος ή μεγαλύτερος από 4 και δεν είναι ανεξάρτητες μεταξύ τους (το κάθε εξερχόμενο δεν μπορεί να προβλεφθεί χωρίς τα άλλα), χρειάζεται πιθανώς μια δεύτερη ενδιάμεση στιβάδα.

Για την ομάδα II, μία μόνο ενδιάμεση στιβάδα δεν είναι πάντα αρκετή και συνίσταται ένα δίκτυο με δύο ενδιάμεσες στιβάδες για να αυξηθεί η ικανότητά του για γενίκευση. Αν χρησιμοποιείται μόνο μία ενδιάμεση στιβάδα, πρέπει αυτή να έχει 20- 40 νευρώνες. Αν χρησιμοποιούνται δύο στιβάδες, πρέπει να υπάρχουν 13 – 20 νευρώνες στην πρώτη και 18 – 25 στην δεύτερη (5 νευρώνες παραπάνω).

Για την ομάδα ΙΙΙ, χρειάζονται δύο ή τρεις ενδιάμεσες στιβάδες. Αν χρησιμοποιούνται δύο, η πρώτη πρέπει να έχει 10 – 20 νευρώνες και η δεύτερη 15 – 25. Αν προστεθεί και τρίτη ενδιάμεση στιβάδα, αυτή πρέπει να έχει τον ίδιο αριθμό νευρώνων με την δεύτερη.



Σχήμα 2.4: Ομάδες των MLP [33].

2.1.6. Επιλογή και σύνθεση των ομάδων

Με τον αλγόριθμο **Kennard-Stone (K-S)**, τα δείγματα της ομάδας εκπαίδευσης, επιλέγονται με βάση το κριτήριο της **ενδο-δειγματικής απόστασης** (inter-distance criterion). Τα δυο πρώτα δείγματα που επιλέγονται για την ομάδα εκπαίδευσης, είναι τα πιο απομακρυσμένα στο χώρο των διαστάσεων, ενώ το τρίτο είναι εκείνο με τη μεγαλύτερη ελάχιστη απόσταση, d_{min}, από τα άλλα δυο. Το κριτήριο αυτό, εφαρμόζεται διαρκώς, μέχρι να επιλεχθεί ο επιθυμητός αριθμός δειγμάτων. Στην πράξη σε κάθε K-S βήμα, υπολογίζεται η ελάχιστη απόσταση d_{min} (συνήθως η Ευκλείδεια), ανάμεσα στα ήδη επιλεγμένα δείγματα και το υποψήφιο προς επιλογή. Το αντικείμενο με τη μεγαλύτερη d_{min} προστίθεται στην ομάδα εκπαίδευσης. Ο K-S θεωρείται ότι "ανιχνεύει" την πιο αντιπροσωπευτική ομάδα εκπαίδευσης ή ότι η επιλεγμένη ομάδα καλύπτει μοναδικά (αντιπροσωπεύει) το χώρο των διαστάσεων (παραμέτρων που χαρακτηρίζουν τα δείγματα) [51, 52, 53, 54].

Ο αλγόριθμος **SPXY** (Sampling set Partioning based on joint x-y distances) αποτελεί μια βελτιωμένη εναλλακτική του K-S αλγόριθμου για την αντιπροσωπευτική επιλογή των ομάδων δειγμάτων, καθώς λαμβάνει υπόψη του τις αποστάσεις στις ανεξάρτητες **x** μεταβλητές, αλλά και την εξαρτημένη y μεταβλητή. Συνδυάζει δηλαδή, διαφορές στους άξονες X και Y για τον υπολογισμό των ενδοδειγματικών αποστάσεων. Η επιλογή της ομάδας εκπαίδευσης γίνεται ώστε σε κάθε βήμα να περιέχει τα πιο απομακρυσμένα δείγματα, μέχρι να συμπληρωθεί ο προαπαιτούμενος αριθμός αυτών [55, 56].

Ο Marini et al. [57] πρωτοπορώντας, χρησιμοποιεί δίκτυο Kohonen για την εύρεση της αντιπροσωπευτικότερης ομάδας δειγμάτων για την ομάδα εκπαίδευσης αλλά και την επιλογή των δυο άλλων ομάδων: επικύρωσης και ελέγχου.

Τέλος, η Gramatica [58] προτείνει και ο Darnag et al. [50] χρησιμοποιεί πιο προηγμένους δείκτες για την εύρεση του AD του βέλτιστου μοντέλου. Συγκεκριμένα, ελέγχεται η "ισχύς" ("leverage") ενός άγνωστου δείγματος σε σχέση με την ομάδα εκπαίδευσης. Ο δείκτης αυτός έχει σχέση με κάποιου είδους απόσταση από το κεντροειδές σημείο της ομάδας εκπαίδευσης. Υψηλή ισχύς σημαίνει ότι το δείγμα μπορεί να ενισχύσει την ομάδα εκπαίδευσης (good leverage). Αντίθετα, ένα τέτοιο δείγμα στην ομάδα ελέγχου θα μπορούσε να οδηγήσει σε αναξιόπιστα αποτελέσματα και προεκβολή του μοντέλου (bad leverage). Η κρίσιμη τιμή h* για την ισχύ υπολογίζεται από τη σχέση $3i/n_t$, όπου i ο αριθμός των εισερχόμενων μεταβλητών και n_t ο αριθμός των δειγμάτων της ομάδας εκπαί-

δευσης. Αν η ισχύς h ενός δείγματος είναι χαμηλότερη της κρίσιμης τιμής (h < h*), το δείγμα θεωρείται ότι έχει την ίδια πιθανότητα αξιόπιστης πρόβλεψης με την ομάδα εκπαίδευσης. Διαφορετικά, είναι εκτός του AD του μοντέλου. Οι παραπάνω συγκρίσεις απεικονίζονται σε διαγράμματα γνωστά ως **διαγράμματα Williamas** (Williams plots) [58].

2.1.7. Συναρτήσεις ενεργοποίησης/περιορισμοί

Εκτός από τις συνήθεις συναρτήσεις που χρησιμοποιούνται σαν ενεργοποίησης, αναφέρονται εδώ μερικές λιγότερο συνήθεις όπως:

1. Διπολική σιγμοειδής συνάρτηση (bipolar sigmoid function)

$$f(x) = \frac{1-e^{-\beta x}}{1+e^{-\beta x}} \qquad (2.1)$$

2. Υπερβολική εφαπτομενική συνάρτηση (hyperbolic tangent function)

$$f(x) = \frac{e^{\beta x}-e^{-\beta x}}{e^{\beta x}+e^{-\beta x}} \qquad (2.2)$$

Η σιγμοειδής συνάρτηση είναι η πιο συνήθης ειδικά στην περίπτωση των MLP. Εδώ θα πρέπει να τονιστεί ότι **οι υψηλές τιμές βαρών μπορούν εύκολα να μετατρέψουν τη σιγμοειδή συνάρτηση σε δυαδική βηματική** ενώ αντίθετα **μικρές τιμές αυτών, ενθαρρύνουν το δίκτυο να "αποκρίνεται" διαρκώς** [12]. Το πρόβλημα των υψηλών τιμών βαρών αλλά και της υψηλής τιμής των εισερχομένων (x) στη σιγμοειδή συνάρτηση, γίνεται αμέσως κατανοητό από τη μορφή της συνάρτησης αυτής. Αναφέρουμε βοηθητικά το παρακάτω παράδειγμα.

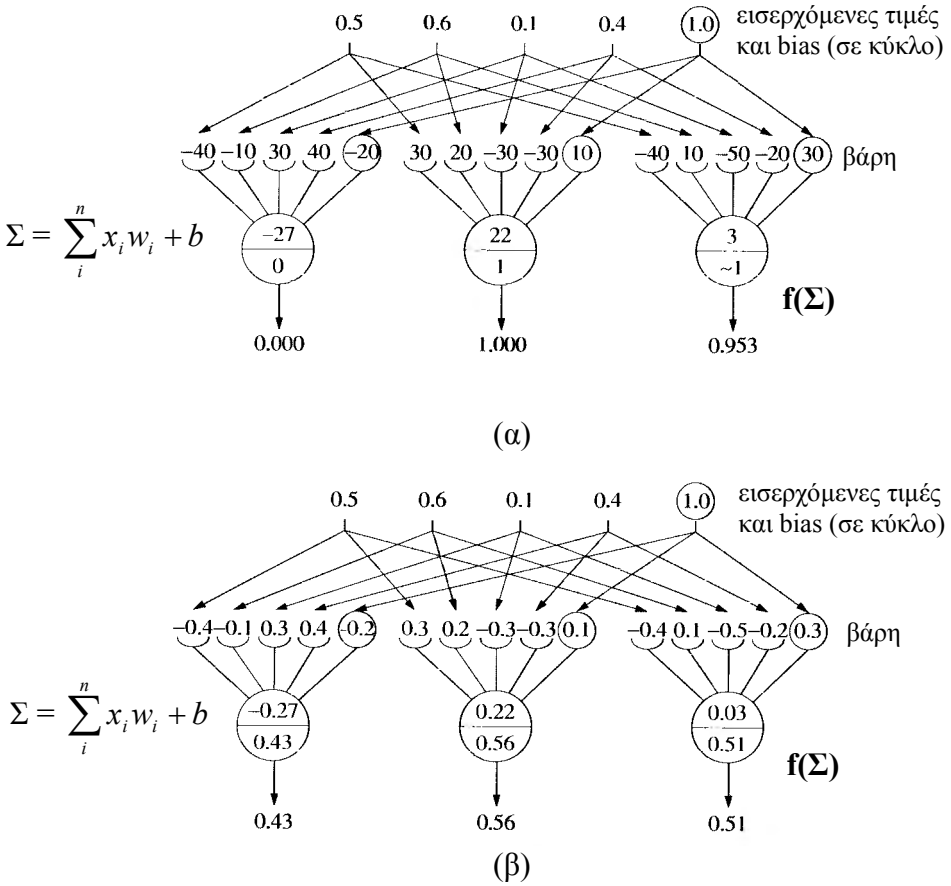
Στον πίνακα 2.1 απεικονίζονται οι τιμές της σιγμοειδούς συνάρτησης, όταν η τιμή του εισερχομένου κυμαίνεται από -10 ως 10. Τα δεδομένα αυτά επεξηγούν γιατί πρέπει να επιλέγονται μικρές τιμές για τα αρχικά βάρη: συνήθως από -1 ως +1 ή και μικρότερα. Η χρωματισμένη περιοχή βαρών δείχνει να "αποκρίνεται" στα εισερχόμενα με διαφορετικές τιμές (από 0,018 ως 0,98), ενώ αντίθετα μεγαλύτερα κατά απόλυτη τιμή εισερχόμενα "ισοπεδώνουν" το αποτέλεσμα.

Ένας κανόνας που μπορεί να χρησιμοποιηθεί είναι το άθροισμα των απολύτων τιμών των βαρών πρέπει να είναι 1.

Πίνακας 2.1: Εξερχόμενα f(x) της σιγμοειδούς συνάρτησης για διαφορετικά εισερχόμενα x

x	**f(x)**	**x**	**f(x)**
-10	0,0000	1	0,7311
-9	0,0001	2	0,8808
-8	0,0003	3	0,9526
-7	0,0009	4	0,9820
-6	0,0025	5	0,9933
-5	0,0067	6	0,9975
-4	0,0180	7	0,9991
-3	0,0474	8	0,9997
-2	0,1192	9	0,9999
-1	0,2689	10	1,0000
0	0,5000		

Επιπλέον, στο παρακάτω παράδειγμα του σχήματος 2.5 απεικονίζονται τα αποτελέσματα της ίδιας σιγμοειδούς συνάρτησης για τα ίδια εισερχόμενα διανύσματα, αλλά διαφορετικές τιμές βαρών. Είναι φανερό ότι με μεγάλα βάρη (σχ. 2.5(α)), οι εξερχόμενες τιμές είναι σχεδόν δυαδικές (οι καμπύλες πιο οξείες), ενώ αντίθετα για μικρές τιμές βαρών (σχ. 2.5(β)), τα εξερχόμενα κυμαίνονται από 0,43 ως 0,51.



Σχήμα 2.5: Παράδειγμα δικτύου με τρεις μονάδες στην ενδιάμεση στιβάδα, η καθεμιά με 5 βάρη (α) με υψηλές ή (β) χαμηλές τιμές [12].

2.1.8. Κανόνας Δέλτα (Delta-rule)

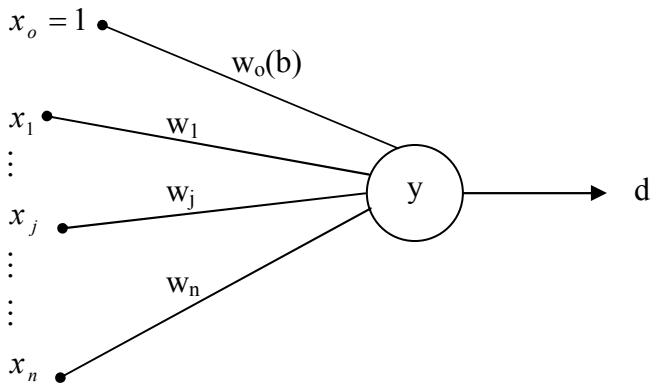
Ο κανόνας Δέλτα είναι ένας κανόνας εκπαίδευσης για Νευρωνικά Δίκτυα απλής στιβάδας (single-layer neural networks), χωρίς δηλαδή ενδιάμεση στιβάδα. Επειδή απότελεί τη βάση του πιο βασικού αλγόριθμου (back-propagation, BP) που περιγράφεται παρακάτω (§ 2.1.10), θα περιγραφεί με περισσότερες λεπτομέρειες. Γενικά, ο κανόνας δέλτα πρεσβεύει τη βελτίωση του διανύσματος **w** του βάρους με τη διόρθωση **Δw** να είναι ανάλογη μιας παραμέτρου δ (η οποία είναι ανάλογη του σφάλματος), και του εισερχόμενου διανύσματος **x** για το οποίο λήφθηκε η λανθασμένη απάντηση [12].

Για ένα δίκτυο απλής στιβάδας με γραμμική συνάρτηση ενεργοποίησης το εξερχόμενο σήμα δίνεται από τη σχέση:

$$y = \sum_{j=1}^{n} w_j x_j + b \qquad (2.3)$$

Αρχικά, πρέπει να μετρηθεί η ολική απόδοση του μοντέλου και στη συνέχεια να βελτιστοποιηθεί. Αυτό σημαίνει ότι ο αλγόριθμος θα πρέπει να αλλάζει όλα τα βάρη, ώστε

το εξερχόμενο y^q, (όπου ο εκθέτης q (=1,……N) αναφέρεται σε ολόκληρο το πλέγμα των δειγμάτων), να πλησιάσει όσο το δυνατό περισσότερο με τη θεωρητική απόκριση d^q (σχ. 2.6).



Σχήμα 2.6: Δίκτυο απλής στιβάδας με βάρη wj

Η συνολική τώρα απόδοση του μοντέλου ή η συνάρτηση σφάλματος E, για όλα τα δείγματα, με βάση τα ελάχιστα τετράγωνα (least mean square, LMS), δίνεται από τη σχέση [32, 59]:

$$E = \sum_{q=1}^{N} E^q \,, \quad \text{όπου για κάθε δείγμα} \quad E^q = \frac{1}{2}(d^q - y^q)^2 \qquad (2.4)$$

Η ιδέα είναι ότι θα πρέπει να ελαχιστοποιηθεί η E, με τις κατάλληλες διορθώσεις που γίνονται στα βάρη w_j (το w_j αναφέρεται στη σύνδεση της εισερχόμενης μεταβλητής j με την εξερχόμενη μονάδα). Θέλουμε δηλαδή να βρούμε, πως τα βάρη w_j επηρεάζουν τα σφάλμα E [36]. Η μέθοδος που θα χρησιμοποιηθεί είναι η steepest–descent minimization method [32].

Η παράγωγος $\nabla\mathbf{E}$ της συνάρτησης E ως προς **w**, είναι το διάνυσμα όλων των μερικών παραγώγων $\frac{\partial E}{\partial w_j}$. Όπως και η παράγωγος μιας συνάρτησης ως προς μια μεταβλητή, έτσι και ή $\nabla\mathbf{E}$ κατευθύνεται πάντα προς την αύξουσα διεύθυνση της E. Αντίθετα, η $-\nabla\mathbf{E}$, θα υποδεικνύει πάντα τη φθίνουσα διεύθυνση της E. Συνεπώς, για να ελαχιστοποιήσουμε την E, πρέπει να κατευθυνθούμε παράλληλα προς την $-\nabla\mathbf{E}$, διορθώνοντας κάθε φορά (όπως μετρήθηκαν για το συγκεκριμένο δείγμα) τα βάρη w_j ως εξής [60]:

$$\left.\begin{array}{l} w_j(t) \rightarrow w_j(t\text{-}1) + \Delta^q w_j \quad \text{όπου} \\ \Delta^q w_j = -n \dfrac{\partial E^q}{\partial w_j} \end{array}\right\} \quad (2.5)$$

όπου n>0 είναι **ο ρυθμός εκπαίδευσης** (learning rate). Οι όροι t και t-1, αναφέρονται στην παρούσα και την προηγούμενη περίοδο.

Η παράγωγος είναι:

$$\frac{\partial E^q}{\partial w_j} = \frac{\partial E^q}{\partial y^q} \frac{\partial y^q}{\partial w_j} \quad (2.6)$$

και επειδή η συνάρτηση είναι γραμμική (σχέση 2.3):

$$\frac{\partial y^q}{\partial w_j} = x_j \quad (2.7)$$

ενώ:

$$\frac{\partial E^q}{\partial y^q} = -(d^q - y^q) \quad (2.8)$$

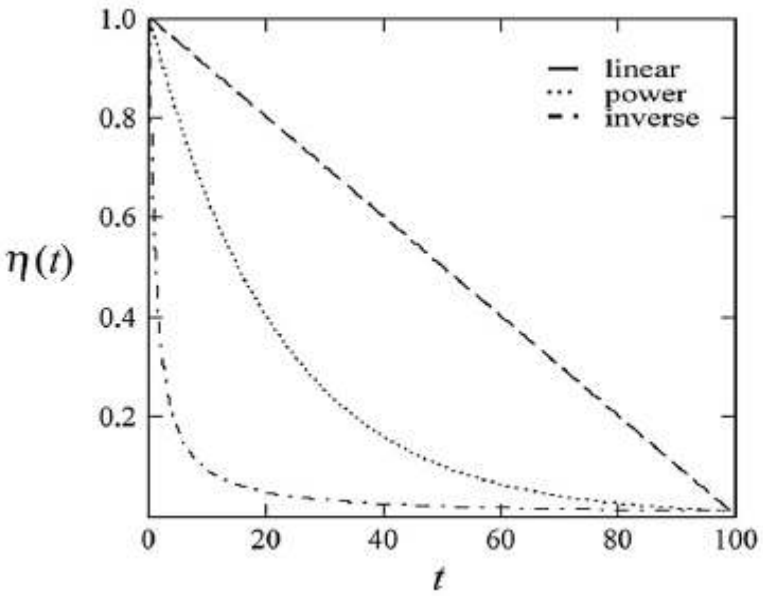
Βάση των σχέσεων (2.5), (2.6), (2.7) και (2.8) έχουμε:

$$\Delta^q w_j = n\delta^q x_j \quad (2.9)$$

αν ορίσουμε $\delta^q = d^q - y^q$.

Έτσι, ο κανόνας δέλτα διαμορφώνει τα βάρη, ελέγχοντας τα πραγματικά και θεωρητικά εξερχόμενα του δικτύου. Επιπλέον, μπορεί να εφαρμοστεί σε συνεχή ή δυαδικά εισερχόμενα και εξερχόμενα δεδομένα [32].

Ο ρυθμός εκπαίδευσης (σχέσεις 2.5, 2.6) καθορίζει σε ποιο βαθμό θα διορθωθούν τα βάρη [12, 35, 61, 62] και καθορίζει ουσιαστικά το βήμα (μαζί με μια σειρά άλλων παραμέτρων, όπως την ορμή, το χρόνο της περιόδου και τη συνάρτηση ενεργοποίησης) που γίνεται στο χώρο των βαρών [41]. Μικρές τιμές αυτού εξαναγκάζουν το μοντέλο σε αργή σύγκλιση, ενώ υπάρχει κίνδυνος παγίδευσης αυτού σε τοπικά ελάχιστα [41, 61, 62]. Αντίθετα σε περίπτωση υψηλών τιμών, το σύστημα μπορεί να είναι ασταθές και να ταλαντώνεται [35, 41, 51, 61, 63, 64]. Αρκετές φορές συνίσταται μεταβαλλόμενος ρυθμός εκπαίδευσης, με την τιμή αυτού να προσαρμόζεται κατά τη διάρκεια της εκπαίδευσης ανάλογα με το αν το σφάλμα αυξάνεται ή ελαττώνεται [41, 65]. Σπανίως δε, αναφέρονται συναρτήσεις που καθορίζουν τη μεταβολή του ρυθμού εκπαίδευσης (σχήμα 2.7) [66]. Οι τιμές του κυμαίνονται μεταξύ 0,5 και 1 για τις σιγμοειδείς συναρτήσεις και 0,001 και 0,1 για τις γραμμικές [51].



Σχήμα 2.7: Οι πιο συνήθεις συναρτήσεις για το ρυθμό εκπαίδευσης: γραμμική (linear), εκθετική (power), υπερβολή (inverse) [66].

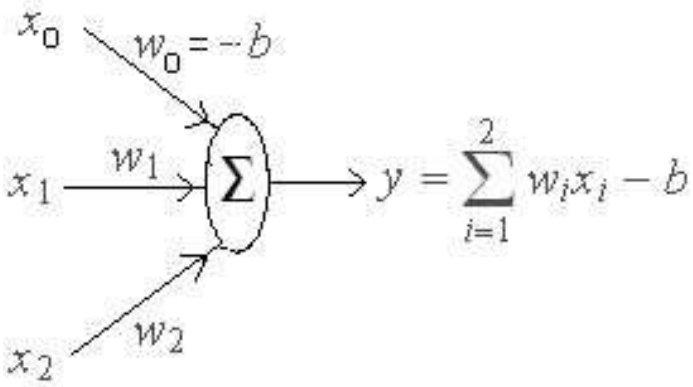
Εδώ, θα αναφέρουμε ένα παράδειγμα, ώστε να επαληθεύσουμε την ελαχιστοποίηση του σφάλματος μέσα από τον κανόνα δέλτα. Θεωρούμε την ομάδα εκπαίδευσης Τ με δυαδικά εισερχόμενα διανύσματα [60]:

$$T = \{(x^q, y^q), q = 1, 2, 3, 4\}$$

με $x^q \in \{0, 1\}$ και $y^q \in \{-1, 1\}$. Ειδικότερα:

$$\left.\begin{array}{lll} \mathbf{x^1} = (x_1^{\,1}, x_2^{\,1}) = (1,1) & \text{και} & y^1 = 1 \\ \mathbf{x^2} = (x_1^{\,2}, x_2^{\,2}) = (1,0) & \text{και} & y^2 = -1 \\ \mathbf{x^3} = (x_1^{\,3}, x_2^{\,3}) = (0,1) & \text{και} & y^3 = -1 \\ \mathbf{x^4} = (x_1^{\,4}, x_2^{\,4}) = (0,0) & \text{και} & y^4 = -1 \end{array}\right\} \quad (2.10)$$

Ένα κατάλληλο δίκτυο για το πρόβλημα αυτό, φαίνεται στο σχήμα 4.29.



Σχήμα 2.8: Αρχιτεκτονική δικτύου για το παρουσιαζόμενο παράδειγμα [60].

Η συνάρτηση ενεργοποίησης f είναι γραμμική. Έτσι, έχουμε:

$$E = \frac{1}{2} \sum_{q=1}^{4} (x_1^q w_1 + x_2^q w_2 - w_0 - y^q)^2 \qquad (2.11)$$

Μετά τις αντικαταστάσεις για τα τέσσερα εισερχόμενα (σχέσεις 2.10), η εξίσωση (2.11) γίνεται:

$$E = w_1^2 + w_2^2 + w_0^2 + 2 + w_1 w_2 - 2w_1 w_0 - 2w_2 w_0 - 2w_0 \qquad (2.12)$$

Το διάνυσμα βάρους **w*** (w_0*, w_1*, w_2*) που ελαχιστοποιεί την E, είναι η λύση του συστήματος των εξισώσεων:

$$\frac{\partial E}{\partial w_j} = 0, \quad j = 0, 1, 2 \qquad (2.13)$$

Διαδοχικά από τις (2.12) και (2.13) έχουμε:

$$\left.\begin{aligned} \frac{\partial E}{\partial w_0} &= w_1 + w_2 - 2w_0 + 1 = 0 \\ \frac{\partial E}{\partial w_1} &= 2w_1 + w_2 - 2w_0 = 0 \\ \frac{\partial E}{\partial w_2} &= w_1 + 2w_2 - 2w_0 = 0 \end{aligned}\right\} \qquad (2.14)$$

Από τις (2.14) προκύπτει: $w_0 = \frac{3}{2}$ και $w_1 = w_2 = 1$. Αυτές είναι οι τιμές των βαρών που ελαχιστοποιούν το σφάλμα E του δικτύου.

Η "μεταδοτική" λειτουργία του κανόνα δέλτα επαληθεύεται στη γενικευμένη μορφή του, με το παράδειγμα παρακάτω (§ 2.1.11) κατά την περιγραφή του BP αλγορίθμου, σε μη γραμμική συνάρτηση ενεργοποίησης και μοντέλο με ενδιάμεση στιβάδα. Ο αλγόριθμος αυτός χρησιμοποιείται ευρύτατα στα ANN και θεωρείται ως μια γενίκευση του κανόνα δέλτα (delta-rule) για μη γραμμικές συναρτήσεις ενεργοποίησης σε Νευρωνικά Δίκτυα πολλαπλών στιβάδων [32].

2.1.9. Κανόνες τερματισμού/εκτίμησης και σύγκρισης μοντέλων

Συμπληρωματικά, αναφέρονται τα παρακάτω κριτήρια τερματισμού των μοντέλων ANN:

1. Το **Μέσο Εκατοστιαίο Σχετικό Σφάλμα** (Average Relative Error, % ARE) που δίνεται από τη σχέση [67]:

$$\% \text{ ARE} = \frac{1}{K} \sum \frac{|y - d|}{d} 100 \qquad (2.15)$$

2. Το **Κανονικοποιημένο Τυπικό Σφάλμα** (Normalized Standard Error, NSE) που δίνεται από τη σχέση [51]:

$$\text{NSE} = \frac{1}{KP} \sum_{k=1}^{K} \sum_{p=1}^{P} (y_{kp} - d_{kp})^2 \qquad (2.16)$$

Το NSE δεν είναι πάντα το καλύτερο κριτήριο, καθώς μια άσχημη πρόβλεψη (μια έκτροπη τιμή δείγματος για παράδειγμα), μπορεί να επηρεάσει πολύ αρνητικά το τελικό αποτέλεσμα. Έτσι, μικρά σφάλματα (κοντά στο μηδέν, 0) για όλα τα δείγματα, αλλά και μερικά μεγάλα μπορεί να δώσουν το ίδιο χαμηλό NSE, αν ο αριθμός των δειγμάτων είναι ικανός [51]. Το ίδιο μπορεί να συμβεί και για μερικούς ακόμα από τους παραπάνω δείκτες.

3. Το **σφάλμα της διασταυρούμενης αξιολόγησης** (error of cross-validation), που χρησιμοποιείται στην εκτίμηση/σύγκριση μοντέλων και δίνεται από τη σχέση [68, 69]:

$$E = \frac{N_{wrong}}{N_o} = \frac{N_o - N_{correct}}{N_o} \qquad (2.17)$$

όπου N_o ο συνολικός αριθμός CV δειγμάτων και N_{wrong}, $N_{correct}$ ο αριθμός των δειγμάτων που αντίστοιχα ταξινομήθηκαν λάθος ή σωστά.

2.1.10. Ο back-propagation αλγόριθμος

Στις περισσότερες περιπτώσεις χρειάζονται Νευρωνικά Δίκτυα πολλαπλών στιβάδων για να προσεγγιστούν γενικότερες συσχετίσεις. Έτσι απαιτούνται αλγόριθμοι εκπαίδευσης για πιο πολύπλοκα ANN [60].

Ο **κανονικός αλγόριθμος ή αλγόριθμος οπισθοδιάδοσης** (BP, back-propagation), δημοσιεύτηκε πρώτα από τον Rumelhart [70] και αποτελεί μια γενίκευση του κανόνα δέλτα. Κάθε φορά που ένα δείγμα εισάγεται στο δίκτυο, τιμές (διαμέσου των βαρών, της προκατάληψης και της συνάρτησης ενεργοποίησης), "φτάνουν" στις εξωτερικές μονάδες. Υπολογίζονται οι τετραγωνισμένες διαφορές ανάμεσα στις ευρισκόμενες και τις θεωρητικές τιμές και αν αυτές είναι διαφορετικές από το μηδέν, η μέθοδος πρέπει να είναι **"αμείλικτη"** στη διόρθωση των βαρών [32]. Η καινοτομία ωστόσο που εισάγεται εδώ είναι ότι μπορούμε να επιφέρουμε τις κατάλληλες μεταβολές στα βάρη στις ενδιάμεσες

στιβάδες, εκεί δηλαδή που δεν υπάρχει στόχος και άρα δεν μπορεί να χρησιμοποιηθεί μια απλή τεχνική, όπως ο κανόνας Δέλτα [71].

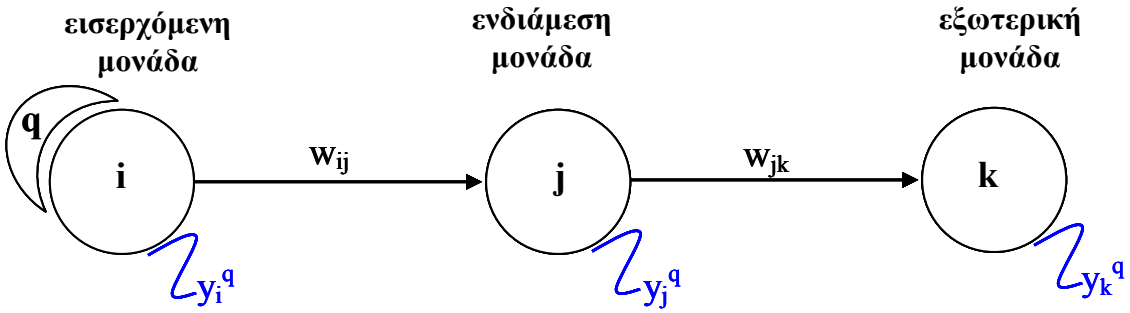
Έτσι, ο BP αλγόριθμος περιλαμβάνει δυο φάσεις:

1. Forward phase: με τη μετάδοση του σήματος από τα εισερχόμενα προς τα εξερχόμενα και
2. Backward phase: με τη μετάδοση του σφάλματος (διά της διόρθωσης των βαρών), ξεκινώντας από τα εξερχόμενα προς τα εισερχόμενα.

Η διόρθωση ενός βάρους είναι ανάλογη του **όρου σφάλματος δ** που "διαπραγματεύεται" δεδομένα της μονάδας που **λαμβάνει το σήμα** αλλά και της μονάδας από την οποία **φεύγει το σήμα.** Δηλαδή, η διόρθωση Δw_{jk} του βάρους w_{jk} που αντιστοιχεί στη σύνδεση της ενδιάμεσης μονάδας j και της εξωτερικής k (σχ. 2.9) για την παρατήρηση (δείγμα) q (μίας από το συνολικό πλέγμα των δειγμάτων), δίνεται από τη σχέση:

$$\Delta w_{jk} = n\delta_k^q y_j^q \qquad (2.18)$$

όπου n>0 είναι ο ρυθμός εκπαίδευσης (learning rate) και y_j^q το εξερχόμενο από την ενδιάμεση μονάδα j προς την εξωτερική k.



Σχήμα 2.9: Μονάδες διαδοχικών στιβάδων δικτύου με τα αντίστοιχα βάρη w και εξερχόμενα y q για το εισερχόμενο δείγμα q.

Όταν η συνάρτηση f που εφαρμόζεται στην εξωτερική στιβάδα είναι σιγμοειδής, ο όρος του σφάλματος δ_k^q για το δείγμα q, **στην εξωτερική μονάδα k** δίνεται από τη σχέση:

$$\left.\begin{aligned}\delta_k^q &= (d_k^q - y_k^q)\, f'(S_k^q) \quad \text{ή} \\ \delta_k^q &= (d_k^q - y_k^q)\, y_k^q\, (1-y_k^q)\end{aligned}\right\} \qquad (2.19)$$

όπου f΄ είναι η παράγωγος της σιγμοειδούς συνάρτησης f για το ζυγισμένο άθροισμα S_k^q όλων των εισερχομένων στην k μονάδα (ή εξερχομένων y_j^q από τις ενδιάμεσες μονάδες) και d_k^q, y_k^d η θεωρητική και υπολογιζόμενη απόκριση αντίστοιχα της εξωτερικής μονάδας k (βλ. σχέση 2.22 παρακάτω για καλύτερη κατανόηση των αντικαταστάσεων στις σχέσεις 2.19: όπου $x = S_k^q$ και όπου $y = y_k^q$). Ο όρος y_k^q $(1-y_k^q)$ είναι η τελικά υπολογιζόμενη παράγωγος της σιγμοειδούς συνάρτησης.

Ο όρος σφάλματος δ_j^q για την παρατήρηση (δείγμα) q και τη μονάδα (νευρώνα) j **της ενδιάμεσης στιβάδας,** όταν πάλι χρησιμοποιείται σιγμοειδής συνάρτηση είναι:

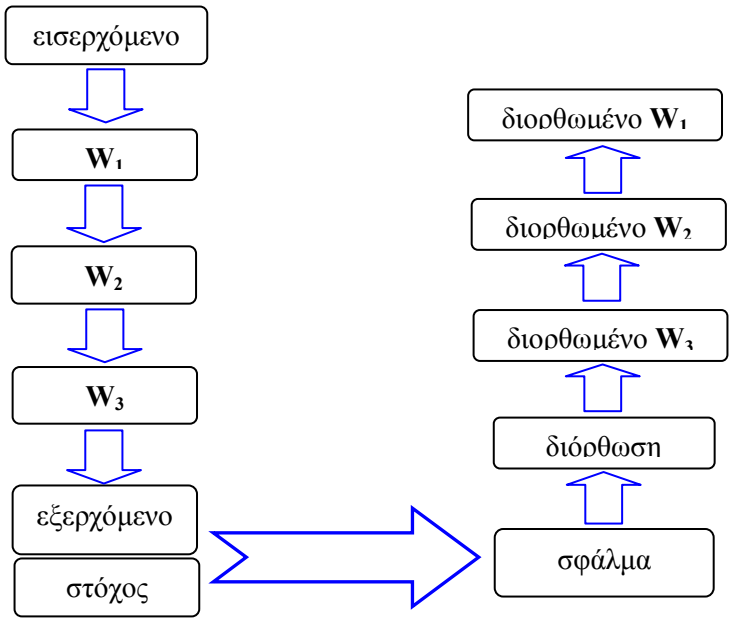
$$\delta_j^q = y_j^q (1-y_j^q) \sum_{k=1}^{K} \delta_k^q w_{jk} \qquad (2.20)$$

όπου w_{jk} το βάρος της σύνδεσης μεταξύ των μονάδων j και k, y_j η απόκριση (εξερχόμενο) της μονάδας j της ενδιάμεσης στιβάδας προς τον εξωτερικό νευρώνα k και K το σύνολο των εξωτερικών νευρώνων [32].

Έτσι το σφάλμα από την εξωτερική στιβάδα, "διαδίδεται" στην ενδιάμεση "κρυφή" μέσου του όρου δ_k^q (σχέση 2.20). Αυτό είναι το πρώτο βήμα. Με αυτόν τον τρόπο ωστόσο, δεν αλλάζουν όλα τα βάρη. Το επόμενο βήμα όμως, δίνει λύση στην "μετατόπιση" του σφάλματος και **στην πρώτη στιβάδα.** Διαδοχικά λοιπόν, το σφάλμα "μετατοπίζεται" ή διανέμεται ακόμα πιο πίσω, στην πρώτη στιβάδα, με τις κατάλληλες ρυθμίσεις που γίνονται στα βάρη ανάμεσα σε αυτές και τις ενδιάμεσες. Με τον τρόπο αυτό, δικαιολογείται το όνομα του αλγορίθμου καθώς το σφάλμα **"διαδίδεται"** (is propagated) προς τα πίσω (σχ. 2.10). Οι ρυθμίσεις που γίνονται στα βάρη ανάμεσα στην εισερχόμενη στιβάδα και τις ενδιάμεσες καθορίζονται από τη σχέση:

$$\left.\begin{aligned} \Delta w_{ij}(t) &= n\delta_j^q y_i^q \text{ ή καλύτερα} \\ \Delta w_{ij}(t) &= n\delta_j^q y_i^q + a\,\Delta w_{ij}(t-1) \end{aligned}\right\} \qquad (2.21)$$

όπου Δw_{ij} είναι η "ενημέρωση" ή διόρθωση που γίνεται στα βάρη ανάμεσα στη μονάδα j της ενδιάμεσης στιβάδας και τη μονάδα i της αρχικής στιβάδας και y_i η απόκριση (εξερχόμενο) της μονάδας i της εισερχόμενης στιβάδας προς την ενδιάμεση j. Επιπλέον, n>0 είναι ο ρυθμός εκπαίδευσης και **a η ορμή** (momentum, βλ. παρακάτω). Οι όροι t και t-1, αναφέρονται όπως και στον κανόνα δέλτα, στην παρούσα και την προηγούμενη περίοδο [72].



Σχήμα 2.10: Σχηματική αναπαράσταση της διόρθωσης των βαρών με τη βοήθεια του back-propagation αλγορίθμου [12].

Συνδυάζοντας τις σχέσεις (2.20) και (2.21), είναι φανερό ότι τιμές από **τρεις στιβάδες** επηρεάζουν τη διόρθωση των βαρών σε μια στιβάδα:

- τιμές από την "τρέχουσα" στιβάδα (εδώ ενδιάμεση μέσω του βάρους $\Delta w_{ij}(t-1)$),
- τιμές από την προηγούμενη στιβάδα (εδώ εισερχόμενη μέσω της απόκρισης y_i που προέρχεται από το νευρώνα i) και
- τιμές από την επόμενη στιβάδα (εδώ εξωτερική μέσω του όρου $\sum_{k=1}^{K} \delta_k^{\,q} w_{jk}$) [12].

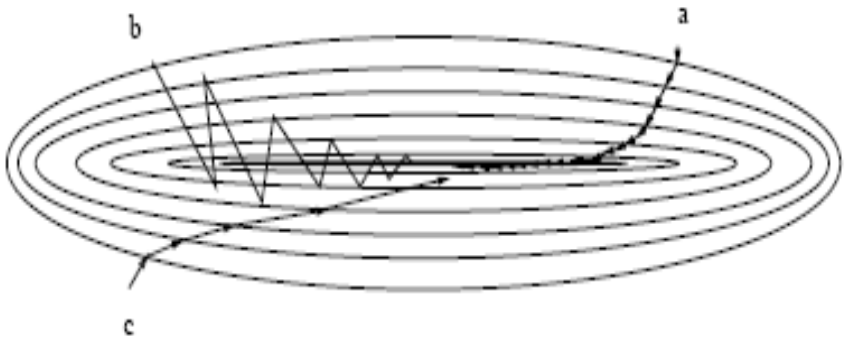
Οι σχέσεις (2.17) έως (2.20), δικαιολογούν απόλυτα την "μεταδοτική" λειτουργία του BP αλγορίθμου: αρχίζοντας από το υπολογιζόμενο σφάλμα $d_k^q - y_k^q$ στην εξωτερική στιβάδα, η διόρθωση (ρύθμιση) "περνά" στα βάρη w_{jk} της ενδιάμεσης στιβάδας και καταλήγει στην πρώτη μέσα από τα βάρη w_{ij}. Μπορούμε να φανταστούμε ένα ντόμινο το οποίο κινείται εμπρός, από τα εισερχόμενα (μονάδες της πρώτης στιβάδας) προς τα εξερχόμενα, (μονάδες της εξωτερικής στιβάδας) αλλά μετά τελείως απροσδόκητα, γυρίζει πίσω και μεταδίδει την κίνηση πάλι στις πρώτες στιβάδες [73].

Η τελική σχέση 2.20 αποτελείται από δυο προσθετέους οι οποίοι και "παρασύρουν" τη διόρθωση των βαρών προς δυο διαφορετικές κατευθύνσεις. Ο πρώτος όρος (που περιέχει το ρυθμό εκπαίδευσης) στοχεύει σε μια σύγκλιση απότομης κατάβασης, ενώ ο δεύτερος όρος (που περιέχει την ορμή) παρεμποδίζει την παγίδευση σε

τοπικά ελάχιστα (βλ. παρακάτω). Το μέγεθος των δυο αυτών σταθερών καθορίζει το βαθμό επίδρασης του κάθε όρου. [12].

Η ορμή είναι ένα συντελεστής που χρησιμοποιείται για να καθορίσει **σε ποιο βαθμό, η κάθε περίοδος εξαρτάται από την προηγούμενη** [12, 32, 35, 51, 61, 62, 63, 74]. Έτσι, αναστέλλονται ξαφνικές αλλαγές στην κατεύθυνση που γίνονται οι διορθώσεις [12, 62], το δίκτυο ανταποκρίνεται καλύτερα στις "τάσεις" της επιφάνειας σφάλματος [74], ενώ αποφεύγεται η προσκόλληση σε τοπικά ελάχιστα [54, 63, 74]. Έτσι για παράδειγμα, όταν η ορμή a είναι μεγαλύτερη του ρυθμού εκπαίδευσης n, αποφεύγονται ταλαντώσεις γύρω από το ελάχιστο του σφάλματος, αλλά μπορεί να μη γίνουν αντιληπτά στενά μονοπάτια που οδηγούν στο ολικό ελάχιστο [12]. Όταν ο ρυθμός εκπαίδευσης n είναι μεγάλος, μπορούν με την χρήση της ορμής, να αποφθεχθούν οι ταλαντώσεις γύρω από το ελάχιστο του σφάλματος (που είναι και το ζητούμενο), και το ελάχιστο να επιτευχθεί με αυτόν τον τρόπο γρηγορότερα (σχ. 2.11). Συνεπώς, η χρήση της ορμής, επιταχύνει το τελικό αποτέλεσμα [32, 61, 64, 75]. Η επίδρασή της παρόλα αυτά στη αποτελεσματικότητα του δικτύου είναι μικρότερη από του ρυθμού εκπαίδευσης και χρησιμοποιείται λιγότερο στη διαδικασία βελτιστοποίησης του δικτύου.

Γενικότερα, τόσο για το ρυθμό εκπαίδευσης όσο και για την ορμή, απλά ελέγχονται κάποιες τιμές (μέσω δοκιμών trial and error [76]) χωρίς να βελτιστοποιούνται ουσιαστικά μέσω μιας αντίστοιχης πορείας [51]. Η τιμή της ορμής πρέπει να είναι μικρότερη από 1,0 για να επιτευχθεί σύγκλιση [41]. Συνήθως χρησιμοποιούνται τιμές μεταξύ 0,6 και 0,8 [51] ή συστήνεται το άθροισμα του ρυθμού εκπαίδευσης και της ορμής να είναι κοντά στο 1 [76].



Σχήμα 2.11: Η έρευνα στο χώρο των βαρών. (a) με μικρούς ρυθμούς εκπαίδευσης n, (b) με μεγάλους ρυθμούς εκπαίδευσης: πολλές ταλαντώσεις, (c) με μεγάλους ρυθμούς εκπαίδευσης και ορμή α [32].

Η εκπαίδευση μέσω του BP αλγορίθμου, μπορεί να γίνει με δυο τρόπους:

1. **Pattern mode** (ή case-by-case ή on-line mode ή immediate correction ή incremental approach).
2. **Batch mode** (ή deferred correction).

Στην πρώτη περίπτωση, οι υπολογισμοί γίνονται μετά από κάθε δείγμα ή παρατήρηση (pattern ή case), ενώ στη δεύτερη περίπτωση, οι υπολογισμοί και συνεπώς η διόρθωση των βαρών, γίνονται μετά την παρουσίαση ολόκληρου του πλέγματος των δειγμάτων. Θεωρητικά με την χρήση του pattern mode, αποφεύγονται καλύτερα τα τοπικά ελάχιστα, καθώς τα δείγματα εισέρχονται τυχαία στο νευρωνικό μοντέλο, αλλά με χρήση της batch mode μεθόδου, υπολογίζεται ακριβέστερα το βαθμωτό διάνυσμα [41, 78]. Επιπλέον, η pattern mode θεωρείται καλύτερη επιλογή σε περιπτώσεις πολλών δειγμάτων εκπαίδευσης [60]. Εμπειρική παρατήρηση ωστόσο, έχει δείξει ότι στη δεύτερη περίπτωση, επιτυγχάνεται γρηγορότερη σύγκλιση [32]. Κατά άλλους ερευνητές ωστόσο, οι δυο προσεγγίσεις θεωρούνται ισοδύναμες [51], ενώ η batch mode προσέγγιση δεν φαίνεται να παρουσιάζει κανένα ιδιαίτερο πλεονέκτημα και έτσι δεν χρησιμοποιείται ιδιαίτερα [12].

Ιδιαίτερη μέριμνα χρειάζεται στη σειρά (order) των δειγμάτων με την οποία "τροφοδοτείται" το δίκτυο κατά την εκπαίδευση. Για παράδειγμα, όταν χρησιμοποιείται συνεχώς η ίδια σειρά δειγμάτων, το δίκτυο πιθανά να "επικεντρωθεί" στα πρώτα από αυτά. Το πρόβλημα αντιμετωπίζεται με τη συνεχή αντιμετάθεση των δειγμάτων της ομάδας εκπαίδευσης [32, 51].

Ο back-propagation έχει λιγότερες απαιτήσεις μνήμης από άλλους αλγορίθμους. Αποτελεί συνεπώς μια καλή λύση όταν υπάρχουν πολλά δεδομένα ή όταν υπάρχει περίσσεια δεδομένων (πολλές επαναλαμβανόμενες περιπτώσεις). Έτσι, δουλεύοντας σε pattern mode και διπλασιάζοντας τα δείγματα, ο χρόνος σύγκλισης των δεδομένων θα είναι μεγαλύτερος, αλλά τα αποτελέσματα θα είναι τα ίδια, χωρίς αλλοιώσεις. Αλλά ο BP είναι εξίσου καλός και για μικρό αριθμό δεδομένων: αντίθετα ένας πιο πολύπλοκος αλγόριθμος θα ήταν αμείλικτος για το μέγεθος της ομάδας εκπαίδευσης ή της ομάδας επικύρωσης. Επιπλέον, ο BP αλγόριθμος μπορεί να τροποποιηθεί με την χρήση του όρου της ορμής (momentum), ο οποίος επιταχύνει την "κατάβαση" όταν γίνονται αρκετά βήματα προς την ίδια κατεύθυνση. Το αποτέλεσμα, είναι γρηγορότερη σύγκλιση προς το ελάχιστο της επιφάνειας σφάλματος και αποφυγή των τοπικών ελαχίστων [7].

Γενικότερα, ο BP αλγόριθμος θεωρείται ότι επιτυγχάνει τα καλύτερα αποτελέσματα σε προβλήματα ταξινόμησης όταν μάλιστα συνδυάζεται με feed-forward δίκτυο [48].

Παρά τα πλεονεκτήματα που περιγράφηκαν ωστόσο, και την επιτυχία που γνωρίζει ο BP (προφανής από τις αναρίθμητες εφαρμογές), έχουν επίσης εντοπισθεί κάποια μειονεκτήματα, τα οποία και οδήγησαν σε βελτιωμένες εκδόσεις του από πολλούς ερευνητές [32]:

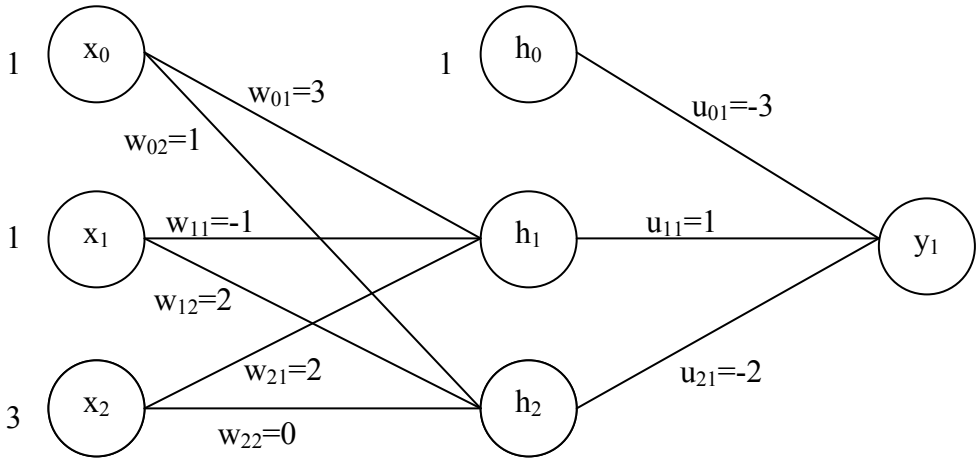
1. Το δίκτυο "παραλύει". Αυτό σημαίνει ότι κατά τη διάρκεια της εκπαίδευσης, τα βάρη μπορεί να διαμορφωθούν **σε πολύ ψηλές τιμές** και εξαιτίας της συνήθους συνάρτησης ενεργοποίησης που είναι σιγμοειδής, το αποτέλεσμα να είναι κοντά στο 0 ή το 1. Τότε, παρατηρώντας τη συνάρτηση f:

$$y = f(x) = \frac{1}{1+e^{-\beta x}} \rightarrow f'(x) = y\,(1-y) \qquad (2.22)$$

   όπου β σταθερά, είναι φανερό ότι το σύστημα μπορεί πραγματικά να παραμείνει στάσιμο [36, 41, 71, 77]. Στις περιπτώσεις αυτές, προτείνεται η αύξηση του ρυθμού εκπαίδευσης, ή η κατάλληλη διαμόρφωση (μείωση) του συντελεστή β στη σιγμοειδή συνάρτηση, ώστε η απόκριση y να πέφτει μέσα στη δυναμική περιοχή αυτής [41, 51]. Συνίσταται επίσης, η προσθήκη ενός μικρού όρου στην παράγωγο της συνάρτησης ενεργοποίησης, η χρήση τροποποιημένων συναρτήσεων ή ορμής [41], η επανεκπαίδευση του δικτύου με νέα αρχικά βάρη ή η αύξηση των μονάδων της ενδιάμεσης στιβάδας. Πράγματι, φαίνεται ότι στην περίπτωση της παράλυσης, το δίκτυο δεν έχει τη δυνατότητα να απεικονίσει επαρκώς τη συσχέτιση [51].
2. Ο αλγόριθμος μπορεί να "κολλήσει" σε τοπικά ελάχιστα, όταν υπάρχει χαμηλότερη περιοχή κοντά [41, 71, 77, 78]. Και εδώ, συνίσταται επίσης η επανεκπαίδευση του δικτύου με νέα αρχικά βάρη [41, 51], η χρήση του pattern mode (βλ. παραπάνω) ή η αύξηση των νευρώνων της ενδιάμεσης στιβάδας.
3. Μερικές εφαρμογές συγκλίνουν αργά και ο χρόνος εκπαίδευσης, είναι πραγματικά υψηλός [12, 71, 77, 78, 79]. Εδώ, σημαντικό ρόλο παίζουν και οι παράμετροι που πρέπει να βελτιστοποιηθούν ώστε να αποκτηθεί το καλύτερο BP μοντέλο [79].
4. Γενικότερα θεωρείται ότι ο BP αλγόριθμος είναι πιο επιτυχής σε προβλήματα ταξινόμησης.

2.1.11. Παράδειγμα κατανόησης (BP αλγόριθμος)

Θα αναφερθούμε τώρα σε ένα παράδειγμα για την πλήρη κατανόηση του BP αλγορίθμου. Το δίκτυο που θα εκπαιδευτεί (με μία ενδιάμεση στιβάδα), φαίνεται στο σχήμα 4.12 [60].



Σχήμα 2.12: Αρχιτεκτονική δικτύου με διπλής στιβάδας perceptron [60].

Ας υποθέσουμε ότι η επιθυμητή θεωρητική απόκριση είναι d = 0,9. Ο ρυθμός εκπαίδευσης είναι n = 0,1. Τα αρχικά βάρη φαίνονται επίσης στο σχήμα (2.12), αλλά και τις σχέσεις (2.24):

$$\left.\begin{aligned} \mathbf{w} &= \begin{bmatrix} w_{01} & w_{11} & w_{21} \\ w_{02} & w_{12} & w_{22} \end{bmatrix} = \begin{bmatrix} 3 & -1 & 2 \\ 1 & 2 & 0 \end{bmatrix} \\ \mathbf{u} &= \begin{bmatrix} u_{01} \\ u_{11} \\ u_{21} \end{bmatrix} = \begin{bmatrix} -3 \\ 1 \\ -2 \end{bmatrix} \end{aligned}\right\} \qquad (2.24)$$

Η συνάρτηση ενεργοποίησης f είναι σιγμοειδής για την ενδιάμεση και εξωτερική στιβάδα με: $f(x) = \frac{1}{1+e^{-x}}$, και το εξερχόμενο σήμα της ενδιάμεσης συνάρτησης για το εισερχόμενο διάνυσμα **x** = (1, 1, 3) είναι $h_j = f(\alpha_j) = \frac{1}{1+e^{-a_j}}$, όπου $\alpha_j = \mathbf{w_j} \times \mathbf{x}$. Έτσι:

$\alpha_1 = 3\times1 + (-1)\times1 + 2\times3 = 8$ (εισερχόμενο για την **h_1 μονάδα** της ενδιάμεσης στιβάδας),

$\alpha_2 = 1\times1 + 2\times1 + 0\times3 = 3$ (εισερχόμενο για την **h_2 μονάδα** της ενδιάμεσης στιβάδας),

$h_0 = 1$

$h_1 = \frac{1}{1+e^{-8}} = 0{,}9997$ (εξερχόμενο για την **h_1 μονάδα** της ενδιάμεσης στιβάδας),και

$h_2 = \frac{1}{1+e^{-3}} = 0{,}9526$ (εξερχόμενο για την **h_2 μονάδα** της ενδιάμεσης στιβάδας)

Το εξερχόμενο σήμα που υπολογίζεται βάση του μοντέλου είναι:

$$y_1 = f(\mathbf{u}\times\mathbf{h}) = f((-3)\times1 + 1\times0{,}9997 + (-2)\times0{,}9526 = f(-3{,}9055) = \frac{1}{1+e^{-(-3{,}9055)}} = 0{,}0197$$

Το σφάλμα στην εξερχόμενη στιβάδα που άμεσα "μεταδίδεται" στην ενδιάμεση, υπολογίζεται ως εξής:

$$\delta y_1 = (d - y_1)(y_1)(1 - y_1) = (0{,}9 - 0{,}0197)(0{,}0197)(1 - 0{,}0197) = 0{,}0170$$

$$\delta h_1 = (h_1)(1 - h_1)(\delta y_1)(u_{11}) = (0{,}9997)(1 - 0{,}9997)(0{,}0170)(1) = 5{,}098 \times 10^{-6}$$

$$\delta h_2 = (h_2)(1 - h_2)(\delta y_1)(u_{21}) = (0{,}9526)(1 - 0{,}9526)(0{,}0170)(-2) = -8{,}056 \times 10^{-4}$$

Οι διορθώσεις **Δw** και **Δu** υπολογίζονται άμεσα με βάση τα υπολογιζόμενα σφάλματα:

$$\Delta w_{01} = n(\delta h_1)(x_0) = (0{,}1)(5{,}098 \times 10^{-6})(1) = 5{,}098 \times 10^{-7}$$

$$\Delta w_{11} = n(\delta h_1)(x_1) = (0{,}1)(5{,}098 \times 10^{-6})(1) = 5{,}098 \times 10^{-7}$$

$$\Delta w_{21} = n(\delta h_1)(x_2) = (0{,}1)(5{,}098 \times 10^{-6})(3) = 1{,}5294 \times 10^{-6}$$

$$\Delta w_{02} = n(\delta h_2)(x_0) = (0{,}1)(-8{,}056 \times 10^{-4})(1) = -8{,}056 \times 10^{-5}$$

$$\Delta w_{12} = n(\delta h_2)(x_1) = (0{,}1)(-8{,}056 \times 10^{-4})(1) = -8{,}056 \times 10^{-5}$$

$$\Delta w_{22} = n(\delta h_2)(x_2) = (0{,}1)(-8{,}056 \times 10^{-4})(3) = -2{,}417 \times 10^{-4}$$

$$\Delta u_{01} = n(\delta y_1)(h_0) = (0{,}1)(0{,}0170)(1) = 0{,}0017$$

$$\Delta u_{11} = n(\delta y_1)(h_1) = (0{,}1)(0{,}0170)(0{,}9997) = 0{,}0017$$

$$\Delta u_{21} = n(\delta y_1)(h_2) = (0{,}1)(0{,}0170)(0{,}9526) = 0{,}0016$$

Με βάση τον κανόνα δέλτα, τα βάρη **w, u,** "ενημερώνονται" ως εξής:

$$\left.\begin{aligned} \mathbf{w} &= \begin{bmatrix} w_{01}+\Delta w_{01} & w_{11}+\Delta w_{11} & w_{21}+\Delta w_{21} \\ w_{02}+\Delta w_{02} & w_{12}+\Delta w_{12} & w_{22}+\Delta w_{22} \end{bmatrix} \\ \mathbf{u} &= \begin{bmatrix} u_{01}+\Delta u_{01} \\ u_{11}+\Delta u_{11} \\ u_{21}+\Delta u_{21} \end{bmatrix} \end{aligned}\right\} \quad (2.25)$$

Είναι φανερό ότι οι διορθώσεις στα βάρη είναι πολύ μικρές (σχέσεις (2.25), ενώ η απόσταση από το επιθυμητό εξερχόμενο σήμα d = 0,9 είναι πολύ μεγάλη. Θα πρέπει λοιπόν να γίνουν επαναλαμβανόμενοι υπολογισμοί και διορθώσεις, ώστε το σήμα y_1 να τείνει προς το d με όσον το δυνατό, μηδενική ανοχή.

2.1.12. Άλλοι αλγόριθμοι

Δυο βασικές τροποποιήσεις του BP αλγορίθμου: ο quick-propagation (QP) και ο Delta-bar-Delta (DBD).

Ο **Quick-propagation** είναι batch mode αλγόριθμος, που σημαίνει ότι τα βάρη διορθώνονται αφού εξεταστούν όλα τα δείγματα. Ο αλγόριθμος αρχίζει με τον ίδιο κανόνα όπως και ο BP, αλλά συνεχίζει με την παραδοχή ότι η επιφάνεια σφάλματος είναι τοπικά δευτεροβάθμια και έτσι η σύγκλιση στο ελάχιστο μπορεί να γίνει πολύ γρήγορα [80]. Το πρόβλημα δημιουργείται όταν η επιφάνεια δεν είναι κοίλη και έτσι ο αλγόριθμος οδηγείται σε λάθος κατεύθυνση. Γενικά, ο QP είναι πιο επιρρεπής σε τοπικά ελάχιστα από τον BP, ενώ επίσης θεωρείται πιο ασταθής [7].

Ο **Delta-bar-Delta** αλγόριθμος είναι επίσης επιρρεπής σε τοπικά μέγιστα, αλλά είναι πιο σταθερός αν συγκριθεί με τον QP. Είναι επίσης batch mode αλγόριθμος και ξεκινά από την παρατήρηση ότι καθώς η επιφάνεια σφάλματος μπορεί να έχει διαφορετικές κλίσεις για κάθε βάρος, καθένα από αυτά πρέπει να έχει το δικό του ρυθμό εκπαίδευσης. Συνεπώς, σε κάθε περίοδο, διορθώνονται μαζί με τα βάρη και οι ρυθμοί εκπαίδευσης, με βάση δυο βασικούς κανόνες:

1. Αν η παράγωγος έχει το ίδιο πρόσημο για αρκετές περιόδους, ο ρυθμός εκπαίδευσης αυξάνεται, αφού η κατεύθυνση φαίνεται ότι είναι προφανώς η σωστή και θεωρείται δόκιμη η διατήρηση της.
2. Αν η παράγωγος αλλάζει πρόσημο για αρκετές περιόδους, ο ρυθμός εκπαίδευσης μειώνεται.

Προκειμένου να ικανοποιηθούν τα παραπάνω, ο Delta-bar-Delta αλγόριθμος έχει ένα αρχικό ρυθμό εκπαίδευσης, που χρησιμοποιείται σε όλα τα βάρη στην πρώτη περίοδο και όταν η παράγωγος διατηρεί το πρόσημό της, εφαρμόζεται σε αυτόν ένας παράγοντας προσαύξησης, ενώ όταν το πρόσημο της παραγώγου αλλάζει, ο ρυθμός εκπαίδευσης μειώνεται με την εφαρμογή κατάλληλου συντελεστή διόρθωσης.

Μειονέκτημα του Delta-bar-Delta αλγόριθμου, αποτελεί το γεγονός ότι επηρεάζεται άμεσα, από "θορυβώδεις" επιφάνειες σφάλματος [7].

Ο **Newton** είναι ένας προηγμένος δεύτερης τάξης αλγόριθμος, δηλαδή η κατάβαση βασίζεται σε δευτεροβάθμιο μοντέλο [41]. Συνίσταται σε δίκτυα με μικρό αριθμό βαρών, επειδή έχει μεγάλες απαιτήσεις μνήμης (ανάλογες με το τετράγωνο του αριθμού των βαρών). Στους υπολογισμούς χρησιμοποιεί τον αντίστροφο πίνακα Hessian (δηλαδή ένα πίνακα με τις δεύτερης τάξης μερικές παραγώγους) που μπορεί να είναι θετικός και έτσι ο αλγόριθμος να πραγματοποιεί τελικά ανάβαση και όχι κατάβαση [41]. Επίσης, για μικρά δίκτυα οι μερικές παράγωγοι υπολογίζονται εύκολα, αλλά σε μεγαλύτερα δίκτυα δημιουργούνται προβλήματα [77].

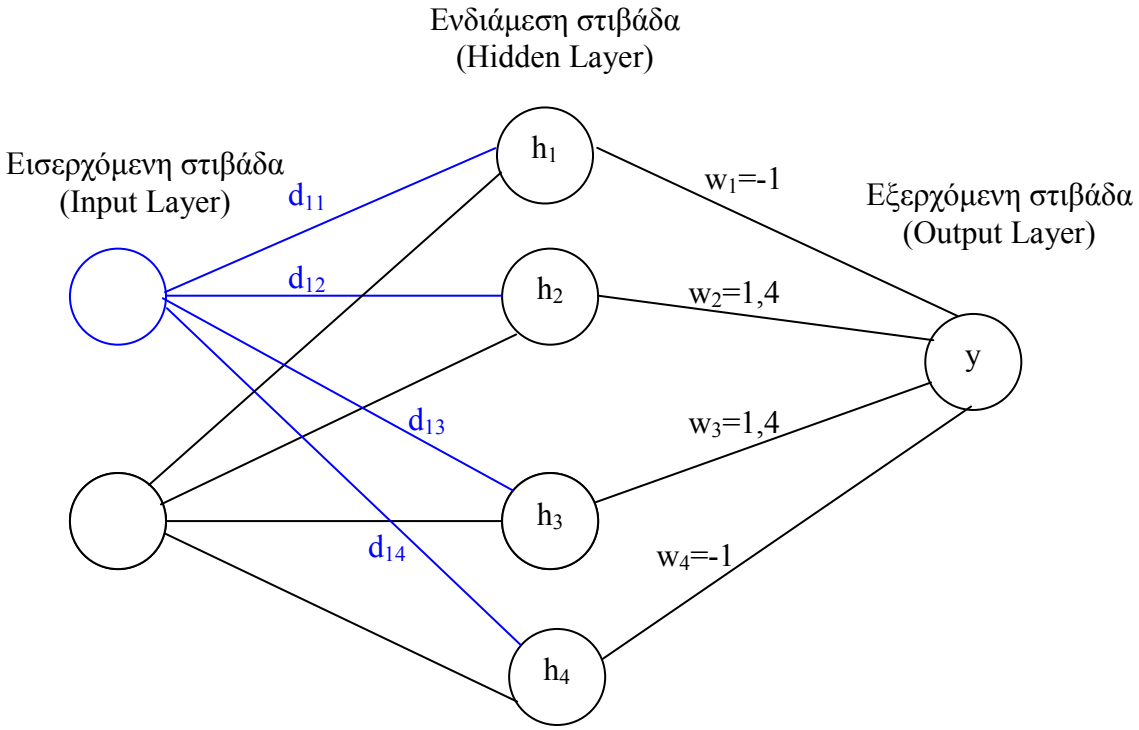
Ο **Quasi-Newton** είναι ένας επίσης προηγμένος δεύτερης τάξης αλγόριθμος που μπορεί να εφαρμοστεί όπου και ο BP, με καλύτερα γενικά αποτελέσματα. Είναι batch mode αλγόριθμος και συνίσταται σε δίκτυα με μικρό αριθμό βαρών [7]. Ο Quasi-Newton αρχίζει με την κατεύθυνση της πιο απότομης κατάβασης όπως και ο BP, αλλά συνεχίζει φτιάχνοντας μια προσέγγιση του αντιστρόφου πίνακα Hessian (αποφεύγοντας το πρόβλημα της ανάβασης) [41]. Ο Quasi-Newton αλγόριθμος έχει μεγάλες απαιτήσεις μνήμης (ανάλογες με το τετράγωνο του αριθμού των βαρών), είναι επιρρεπής σε τοπικά ελάχιστα και είναι λιγότερος σταθερός από τον Conjugate Gradient Descent. Ωστόσο, συγκρινόμενος με αυτόν, ο Quasi-Newton θεωρείται πολύ γρήγορος αλγόριθμος [7, 81].

Ο **Levenberg-Marquardt (LM)** δεύτερης τάξης αλγόριθμος αποτελεί μια τροποποίηση του Newton αλγορίθμου, ο οποίος υπερπηδά το μειονέκτημα του θετικού αντιστρόφου πίνακα Hessian. Θεωρείται υβριδικός ανάμεσα στους αλγορίθμους απότομης κατάβασης και του Newton, δεν προσκολλάται σε τοπικά ελάχιστα, είναι ταχύτατος, με υψηλές αποδόσεις [82] αλλά οι απαιτήσεις μνήμης είναι επίσης πολύ υψηλές (ανάλογες του τετραγώνου του αριθμού των παραμέτρων) [41].

2.2. ΤΑ ΑΛΛΑ ΔΙΚΤΥΑ

2.2.1. Παράδειγμα κατανόησης (RBF δίκτυα)

Ένα παράδειγμα για την κατανόηση της εκπαίδευσης ενός μοντέλου RBF, αναφέρεται παρακάτω. Το μοντέλο που θα εκπαιδευτεί αφορά το γνωστό XOR πρόβλημα με δυαδικά δείγματα (σχ. 2.13).



Σχήμα 2.13: δίκτυο RBF για την επίλυση του προβλήματος XOR.

Ορίζονται τέσσερις ενδιάμεσες μονάδες (ή κεντροειδή) h_1 – h_4 στις θέσεις: (0,0), (0,1), (1,0) και (1,1). Υπολογίζονται οι αποστάσεις κάθε δείγματος **x** (x_i, i =1-4) από την κάθε μονάδα. Για το πρώτο δείγμα x_1 (0,0) ισχύει:

$$\left.\begin{aligned} d_{11} &= (0-0)^2 + (0-0)^2 = 0 \\ d_{12} &= (0-0)^2 + (0-1)^2 = 1 \\ d_{13} &= (0-1)^2 + (0-0)^2 = 1 \\ d_{14} &= (0-1)^2 + (0-1)^2 = 2 \end{aligned}\right\} \qquad (2.26)$$

Ομοίως υπολογίζονται οι αποστάσεις των υπόλοιπων τριών σημείων από τις τέσσερις μονάδες. Ο πίνακας 2.3 συνοψίζει τα αποτελέσματα.

Πίνακας 2.3: Αποστάσεις των εισερχομένων από τις ενδιάμεσες μονάδες h1 – h4.

Εισερχόμενο x (x_i)	**Αποστάσεις από τις ενδιάμεσες μονάδες ή κεντροειδή**			
	h_1	**h_2**	**h_3**	**h_4**
(0,0)	0	1	1	2
(0,1)	1	0	2	1
(1,0)	1	2	0	1
(1,1)	2	1	1	0

Για κάθε εισερχόμενο δείγμα, υπάρχει μία και μόνο κοντινότερη μονάδα (σημειώνονται στα έντονα έγχρωμα κελιά του πίνακα 2.3). Επιπλέον, η κάθε μονάδα αντιστοιχεί σε ένα μόνο δείγμα και επομένως η νέα θέση της καθορίζεται από τις συντεταγμένες του. Έτσι, στο συγκεκριμένο παράδειγμα, οι θέσεις των μονάδων δεν αλλάζουν.

Η διασπορά σ_j (j = 1 – 4), υπολογίζεται σε κάθε μονάδα h_j από τις δυο κοντινότερες σε αυτήν μονάδες, αν θεωρήσουμε K = 2 για τον αλγόριθμο KNN που εφαρμόζεται στη συνέχεια. Έτσι αν:

$$\sigma j = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(h_j - h_k)^2} \qquad (2.27)$$

$$\sigma_1 = \sqrt{\frac{1}{2}(0-0)^2 + (0-1)^2 + (0-1)^2 + (0-0)^2} = 1$$

Ομοίως υπολογίζονται οι αποστάσεις για τις υπόλοιπες μονάδες (οι δυο γειτονικές μονάδες για κάθε μονάδα σημειώνονται στα λιγότερο έγχρωμα κελιά του πίνακα 2.3). Τελικά: $\sigma_1=\sigma_2=\sigma_3=\sigma_4=1$.

Από τη σχέση (2.28) που περιγράφει το εξερχόμενο σήμα h_j από την ενδιάμεση μονάδα j ενός δικτύου RBF:

$$h_j = e^{-(\frac{|x-Cj|}{\sigma j})^2} \qquad (2.28)$$

μπορούν να υπολογιστούν τα εξερχόμενα των τεσσάρων μονάδων για κάθε δείγμα. Για παράδειγμα, το εξερχόμενο του πρώτου δείγματος από την πρώτη μονάδα h_1 είναι:

$$h_{11} = \exp\left[-\left(\frac{\sqrt{(0-0)^2 + (0-0)^2}}{1}\right)^2\right] = 1$$

Ο πίνακας 2.4 συνοψίζει τα αποτελέσματα.

Πίνακας 2.4: Εξερχόμενα από τις ενδιάμεσες μονάδες h1 – h4 για τα τέσσερα δείγματα της ομάδας εκπαίδευσης.

Εισερχόμενο x	**Εξερχόμενα από τις ενδιάμεσες μονάδες ή κεντροειδή**			
	h_1	**h_2**	**h_3**	**h_4**
(0,0)	1	0,37	0,37	0,14
(0,1)	0,37	1	0,14	0,37
(1,0)	0,37	0,14	1	0,37
(1,1)	0,14	0,37	0,37	1

Τώρα, μπορούν να υπολογιστούν τα τελικά εξερχόμενα με τη βοήθεια κάποιου αλγορίθμου όπως τον Least Mean Squares (LMS) ή Pseudo Inverse (PI). Δεν θα αναφερθούμε σε λεπτομέρειες πάνω στο συγκεκριμένο αλγόριθμο. Με βάση ωστόσο, τα προαναφερθείσα βάρη για την εξωτερική στιβάδα, τα αποτελέσματα διαμορφώνονται όπως φαίνονται στον πίνακα 2.5.

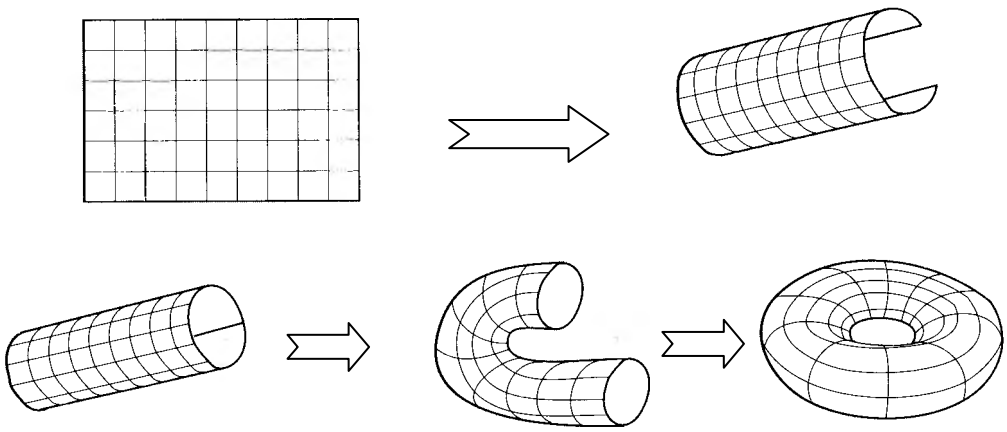
Πίνακας 2.5: Τελικά εξερχόμενα και θεωρητικές αποκρίσεις για τα τέσσερα δείγματα της ομάδας εκπαίδευσης.

Εισερχόμενο x	**Εξερχόμενο y**	**Θεωρητική απόκριση**
(0,0)	-0,10	0
(0,1)	0,86	1
(1,0)	0,86	1
(1,1)	-0,10	0

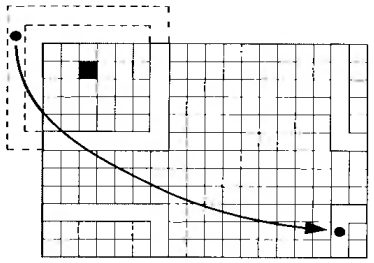
2.2.2. Δίκτυα Kohonen

Ένα επίπεδο Kohonen μπορεί να "διπλωθεί" σε μια δακτυλιοειδή καμπύλη (torus ή toroid) ώστε να δημιουργηθούν "ισοδύναμοι" (ισάριθμοι) γείτονες για κάθε νευρώνα, χωρίς βέβαια να παραποιείται το δίκτυο Στο σχήμα 2.6 φαίνεται πως ένα επίπεδο μπορεί να "διπλωθεί" σε μια δακτυλιοειδή καμπύλη (torus ή toroid). Έτσι, δημιουργούνται

"ισοδύναμοι" (ισάριθμοι) γείτονες για κάθε νευρώνα, χωρίς βέβαια να παραποιείται το δίκτυο (σχ. 2.7).

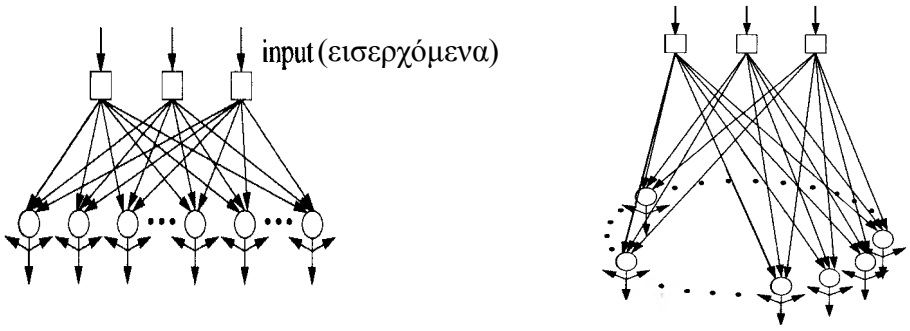


Σχήμα 2.6: "Διπλώνοντας" τη δισδιάστατη δομή σε toroid [12].



Σχήμα 2.7: Ο τρίτος νευρώνας της δεύτερης γραμμής αποκτά 4ης τάξης γείτονες με την κατάλληλη "δίπλωση" [12].

Στη μονοδιάστατη δομή είναι ακόμα ευκολότερη η "δίπλωση" του δικτύου σε toroid: οι νευρώνες απλά διευθετούνται κυκλικά (σχ. 2.8). Το σημαντικό εδώ είναι ότι τελικά όλοι οι νευρώνες δέχονται το ίδιο πολυδιάστατο εισερχόμενο και ο καθένας από αυτούς συνδέεται μόνο **με αυτούς (ισάριθμους για κάθε νευρώνα)** που τοπολογικά βρίσκονται κοντινότερα (local feedback). Έτσι, οι νευρώνες που είναι κοντά μεταξύ τους συμπεριφέρονται παρόμοια στα εισερχόμενα σήματα [12].



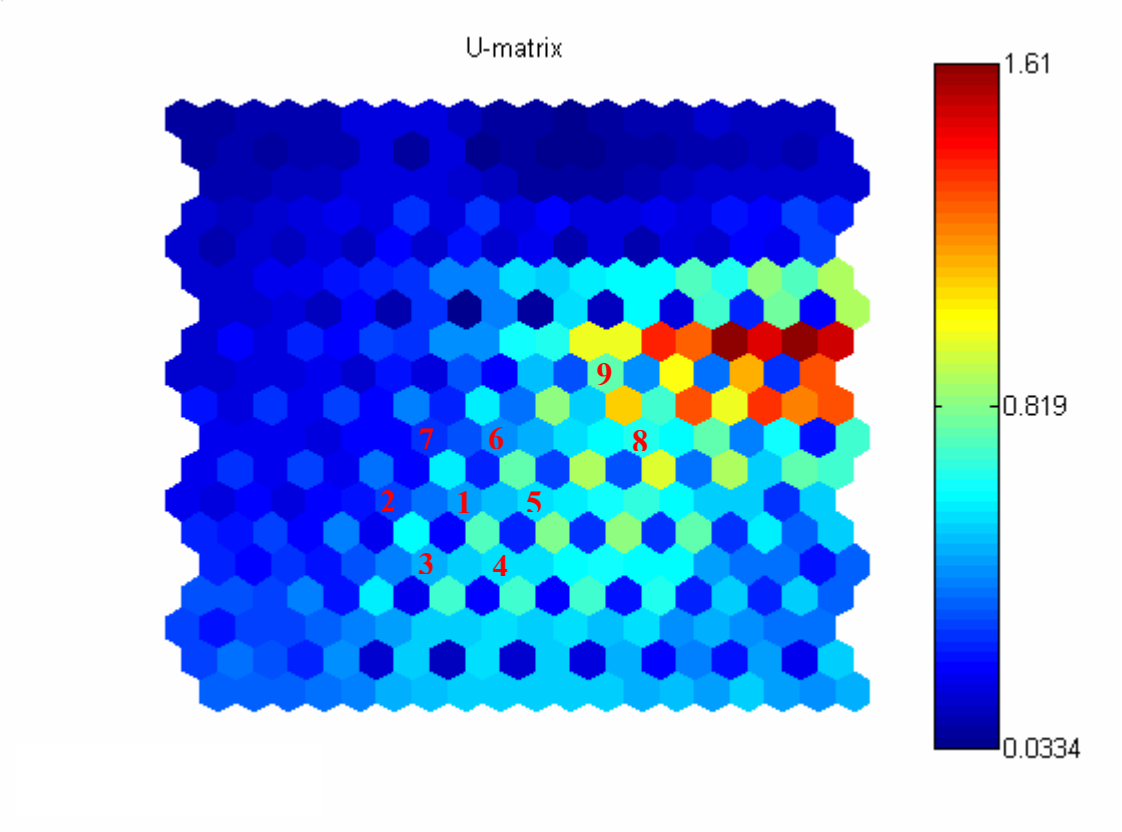
(α) (β)

Σχήμα 2.8: Δίκτυο Kohonen με γραμμική δομή νευρώνων (α). Η toroidal διάταξη διευθετεί το δίκτυο σε κύκλο (β) [12].

Για την οπτικοποίηση των αποστάσεων μεταξύ των γειτονικών νευρώνων ιδιαίτερα δημοφιλής είναι ο **ενιαίος πίνακας αποστάσεων** (U-matrix, unified distance matrix, σχ. 2.9) [22, 83, 84, 85]. Οι υψηλές τιμές στο U-matrix (κόκκινο και κίτρινο χρώμα στο σχήμα), δείχνουν όριο ομάδων. Δεδομένα στην ίδια ομάδα επισημαίνονται από τις ομοιόμορφες περιοχές των χαμηλών τιμών (μπλε χρώμα) [22, 84].

Πιο συγκεκριμένα, το U-matrix απεικονίζει την Ευκλείδεια απόσταση μεταξύ δυο γειτονικών νευρώνων, αλλά και το μέσο όρο των αποστάσεων ενός νευρώνα από όλους τους γειτονικούς σε αυτόν. Έτσι στο παρακάτω σχήμα 2.9, το χρώμα του νευρώνα 1, δείχνει τη μέση Ευκλείδεια απόσταση από τους γειτονικούς του νευρώνες: 2, 3, 4, 5, 6 και 7. Η απόσταση φαίνεται να είναι πραγματικά μικρή. Αντίθετα, το κίτρινο εξάγωνο μεταξύ των νευρώνων 8 και 9 απεικονίζει την πραγματικά μεγάλη απόσταση μεταξύ τους. Άρα οι νευρώνες 8 και 9 δεν ανήκουν στην ίδια ομάδα [83].

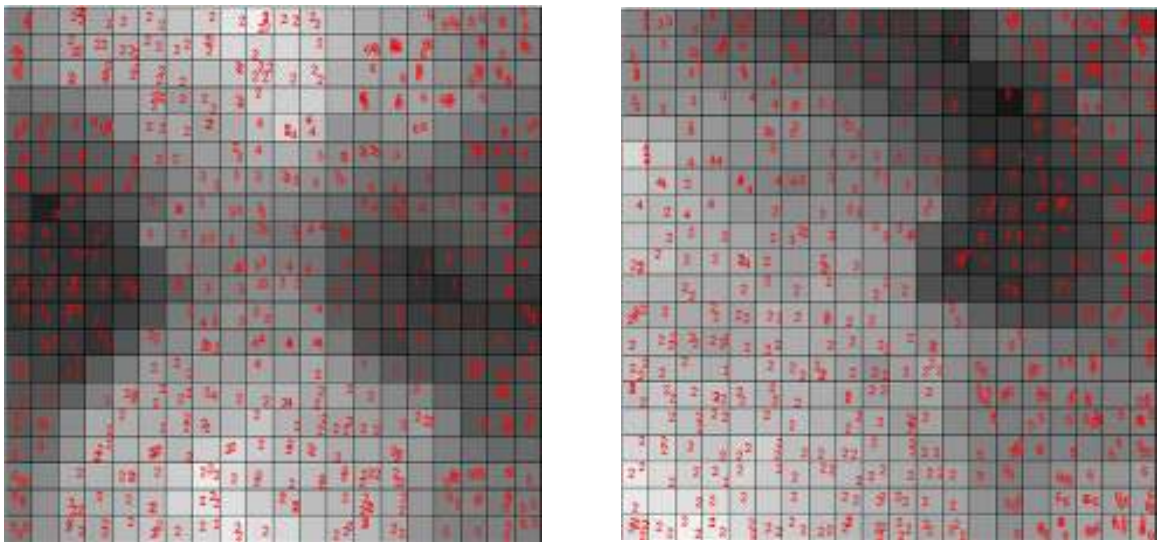
Η ανάγκη δημιουργίας του U-matrix, παρουσιάστηκε εξαιτίας των προβλημάτων που υπήρξαν κατά την ερμηνεία χαρτών Kohonen. Συγκεκριμένα, σε ένα χάρτη διαστάσεων ίσο με το αρχικό δίκτυο Kohonen, οι ενδιαμέσου χρώματος περιοχές, θα μπορούσαν να ερμηνευτούν είτε ως νευρώνες που απέχουν μετρίου μεγέθους αποστάσεις μεταξύ τους, είτε ως νευρώνες με πολύ μεγάλη απόσταση από κάποιους γείτονες από τη μια πλευρά και πολύ μικρή από την άλλη [86]. Ο U-matrix διπλασιάζει το αρχικό Kohonen δίκτυο, ώστε κάποια κελιά να απεικονίζουν ενεργούς νευρώνες και κάποια τις αποστάσεις μεταξύ γειτονικών νευρώνων [83, 86].



Σχήμα 2.9: U-matrix για ένα δίκτυο Kohonen 10x10 [84].Ο πίνακας U-matrix για ένα 10x10 δίκτυο είναι 19x19, καθώς οι ενεργοί νευρώνες διαχωρίζονται μεταξύ τους από ανενεργά εξάγωνα κελιά (ή σκιές, "shades")[83].

Ένα ενδιαφέρον γνώρισμα του Kohonen δικτύου, είναι ότι κάθε νευρώνας έχει τον ίδιο αριθμό βαρών, ενώ σε κάθε επίπεδο βαρών, "διαχειρίζονται" δεδομένα μόνο από μια συγκεκριμένη μεταβλητή. Σε κάθε βάρος, σε μια καθορισμένη και σταθερή θέση στο νευρώνα, θα "περνά" πάντα η ίδια μεταβλητή. Στο τέλος λοιπόν της εκπαίδευσης, σε κάθε επίπεδο του χάρτη, απεικονίζεται η κατανομή μίας και μόνο μεταβλητής. Ενώ δηλαδή συγκεντρωτικά, με την αξιοποίηση όλων των μεταβλητών, παίρνουμε τον τοπολογικό χάρτη που απεικονίζει την ομαδοποίηση/ταξινόμηση όλων των δειγμάτων, σε κάθε επιμέρους επίπεδο, μπορούμε να δούμε την ομαδοποίηση/ταξινόμηση που επιτυγχάνει ή όχι η κάθε μεταβλητή, της οποίας "ανήκει" το επίπεδο. Τα επίπεδα αυτά ονομάζονται **χάρτες βαρών.**

Στο σχήμα 2.10 που ακολουθεί απεικονίζονται επίσης χάρτες βαρών για μια μεταβλητή (Oleic acid για το διαχωρισμό λαδιών), με δυο ωστόσο διαφορετικές τεχνικές: toroidal και κανονική (βλ. παραπάνω).



(α) (β)

Σχήμα 2.10: Kohonen χάρτες βαρών για το oleic acid. Κάθε νευρώνας "περιέχει" δείγματα από τις ομάδες 1, 2, ….7 και χρωματίζεται ανάλογα με τη συγκέντρωση του oleic acid σε αυτά: (α) toroidal τεχνική γειτνίασης νευρώνων (β) κανονική [87].

Η αξιολόγηση των δικτύων Kohonen μπορεί να γίνει με δυο δείκτες που εκφράζουν ουσιαστικά την ελαχιστοποίηση των σφαλμάτων [22, 85, 88, 89, 90, 91]:

1. **Τοπογραφικό** (topographic error): το ήδη εκπαιδευμένο Kohonen δίκτυο θα πρέπει να διατηρεί την τοπολογία των εισερχομένων δειγμάτων. Έτσι, όμοια δείγματα πρέπει να αντιστοιχίζονται σε γειτονικούς νευρώνες. Αν αυτό δεν συμβαίνει, θεωρείται τοπολογικό σφάλμα και μπορεί να υπολογιστεί με βάση το σύνολο των δειγμάτων για τα οποία οι δυο πιο κοντινοί νευρώνες δεν είναι γειτονικοί:

$$e_t = \frac{1}{N} \sum_{k=1}^{N} u(x_k) \qquad (2.29)$$

όπου: N είναι ο αριθμός των δειγμάτων, x_k είναι το k τάξης δείγμα και $u(x_k)$ είναι μια συνάρτηση η οποία είναι ίση με τη μονάδα (1) όταν οι δυο πιο κοντινοί νευρώνες ενός δείγματος δεν είναι γειτονικοί, ενώ διαφορετικά είναι ίση μηδέν (0). Το τοπογραφικό σφάλμα μπορεί να χρησιμοποιηθεί ως ένα μέτρο σύγκρισης μόνο ανάμεσα σε δίκτυα Kohonen των οποίων οι νευρώνες έχουν ίδιο αριθμό βαρών [91].

2. **Κβαντικό** (resolution ή quantization error): το όνομά του προέρχεται από το κύριο στόχο των δικτύων Kohonen, που είναι η "κβαντοποίηση" του χώρου των εισερχομένων σε ένα πεπερασμένο αριθμό νευρώνων. Για να υπολογιστεί το

σφάλμα αυτό, χρησιμοποιείται το σύνολο των δειγμάτων και ο κοντινότερος σε κάθε δείγμα νευρώνας (νικητής):

$$e_q = \frac{1}{N} \sum_{k=1}^{N} \|\mathbf{x_i} - \mathbf{w}\| \qquad (2.30)$$

όπου: x_i είναι το i τάξης δείγμα και **w** ο νικητής νευρώνας.

Ο αριθμός των νευρώνων καθορίζει ουσιαστικά την ακρίβεια αλλά και τη δυνατότητα γενίκευσης ενός δικτύου Kohonen. Όσο μεγαλύτερος είναι ο χάρτης, τόσο χαμηλότερο είναι το κβαντικό σφάλμα e_q, αλλά τόσο μεγαλύτερο το τοπογραφικό e_t [22]. Αυτό οφείλεται στο γεγονός, ότι στους μεγάλους χάρτες, υπάρχει καλύτερη αντιπροσώπευση του κάθε εισερχομένου δείγματος, αλλά το δείγμα μπορεί να μεταπηδά από τη μια περιοχή στην άλλη [92]. Στην πραγματικότητα, γίνεται ένας συμβιβασμός για μια βέλτιστη επιλογή του αριθμού των νευρώνων στον τελικό χάρτη [22, 90].

2.2.3. Παράδειγμα κατανόησης 1 (δίκτυο Kohonen)

Θα αναφερθούμε τώρα σε ένα παράδειγμα για την πλήρη κατανόηση των Kohonen δικτύων. Τα δείγματα (διανύσματα 4 διαστάσεων) που θα ταξινομηθούν είναι τέσσερα [34]:

$$(1, 1, 0, 0), (0, 0, 0, 1), (1, 0, 0, 0) \text{ και } (0, 0, 1, 1) \qquad (2.31)$$

Υποθέτουμε ότι ο αριθμός των ομάδων (και των νευρώνων) είναι m = 2 και ο ρυθμός εκπαίδευσης n = 0,6 με:

$$n(0) = 0,6 \qquad \text{και} \qquad n(t+1) = 0,5\, n(t) \qquad (2.32)$$

Εφόσον υπάρχουν μόνο δυο ομάδες, η ακτίνα γειτνίασης R τίθεται στο μηδέν (μόνο τα βάρη της νικήτριας ομάδας θα διορθώνονται σε κάθε βήμα).

Τα αρχικά βάρη για κάθε ομάδα είναι:

$$\begin{bmatrix} 0,2 & 0,8 \\ 0,6 & 0,4 \\ 0,5 & 0,7 \\ 0,9 & 0,3 \end{bmatrix} \qquad (2.33)$$

Για το πρώτο δείγμα υπολογίζουμε τις αποστάσεις από τους δυο νευρώνες:

$$\left.\begin{aligned} D(1) &= (0,2-1)^2 + (0,6-1)^2 + (0,5-0)^2 + (0,9-0)^2 = 1,86 \\ D(2) &= (0,8-1)^2 + (0,4-1)^2 + (0,7-0)^2 + (0,3-0)^2 = 0,98 \end{aligned}\right\} \qquad (2.34)$$

Ο νευρώνας 2 είναι "νικητής" για το πρώτο δείγμα. Έτσι τα βάρη του νευρώνα διορθώνονται ως εξής:

$$w_{i2}(new) = w_{i2}(old) + 0,6[xi - w_{i2}(old)] \qquad \text{και}$$

$$\begin{bmatrix} 0,2 & 0,8 \\ 0,6 & 0,4 \\ 0,5 & 0,7 \\ 0,9 & 0,3 \end{bmatrix} \rightarrow \begin{bmatrix} 0,2 & 0,92 \\ 0,6 & 0,76 \\ 0,5 & 0,28 \\ 0,9 & 0,12 \end{bmatrix} \tag{2.35}$$

Για το δεύτερο δείγμα ισχύει:

$$\left.\begin{aligned} D(1) &= (0,2-0)^2 + (0,6-0)^2 + (0,5-0)^2 + (0,9-1)^2 = 0,66 \\ D(2) &= (0,92-0)^2 + (0,76-0)^2 + (0,28-0)^2 + (0,12-0)^2 = 2,28 \end{aligned}\right\} \tag{2.36}$$

Ο νευρώνας 1 είναι " νικητής " για το δεύτερο δείγμα. Έτσι τα βάρη του νευρώνα διορθώνονται ως εξής:

$$w_{i1}(new) = w_{i1}(old) + 0,6[xi - w_{i1}(old)] \qquad \text{και}$$

$$\begin{bmatrix} 0,2 & 0,92 \\ 0,6 & 0,76 \\ 0,5 & 0,28 \\ 0,9 & 0,12 \end{bmatrix} \rightarrow \begin{bmatrix} 0,08 & 0,92 \\ 0,24 & 0,76 \\ 0,20 & 0,28 \\ 0,96 & 0,12 \end{bmatrix} \tag{2.37}$$

Ομοίως για το τρίτο δείγμα νικητής είναι ο νευρώνας 2 και διορθώνονται τα βάρη αυτού:

$$\begin{bmatrix} 0,08 & 0,92 \\ 0,24 & 0,76 \\ 0,20 & 0,28 \\ 0,96 & 0,12 \end{bmatrix} \rightarrow \begin{bmatrix} 0,08 & 0,97 \\ 0,24 & 0,30 \\ 0,20 & 0,11 \\ 0,96 & 0,05 \end{bmatrix} \tag{2.38}$$

Το τέταρτο δείγμα ανήκει στο νευρώνα 1 και τα βάρη αυτού διορθώνονται ως εξής:

$$\begin{bmatrix} 0,08 & 0,97 \\ 0,24 & 0,30 \\ 0,20 & 0,11 \\ 0,96 & 0,05 \end{bmatrix} \rightarrow \begin{bmatrix} 0,03 & 0,97 \\ 0,10 & 0,30 \\ 0,68 & 0,11 \\ 0,98 & 0,05 \end{bmatrix} \tag{2.39}$$

Εδώ τελειώνει η πρώτη περίοδος και ο ρυθμός εκπαίδευσης βάση της σχέσης (2.32) μειώνεται:

$$n(1) = 0,5\ n(0) = 0,3 \tag{2.40}$$

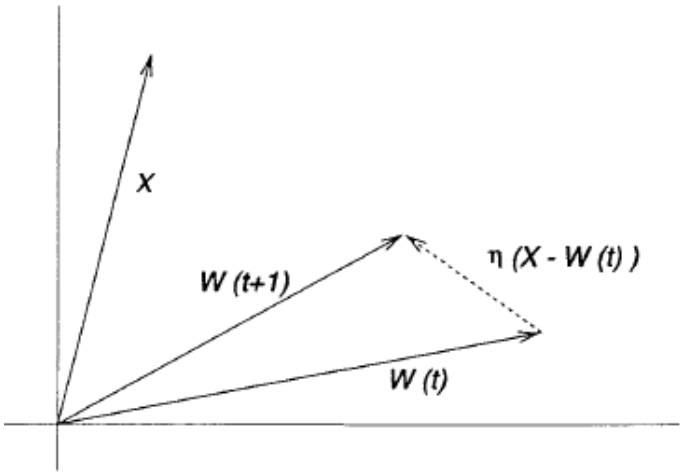
Η ίδια διαδικασία συνεχίζεται για 100 περιόδους με τα βάρη να διαμορφώνονται ως εξής:

$$\begin{bmatrix} 6,7 \times 10^{-7} & 1,0000 \\ 2,0 \times 10^{-16} & 0,49 \\ 0,5 & 2,3 \times 10^{-16} \\ 1,0000 & 1,0 \times 10^{-16} \end{bmatrix} \tag{2.41}$$

Συνολικά τα βάρη τείνουν να συγκλίνουν προς τον πίνακα:

$$\begin{bmatrix} 0,0 & 1,0 \\ 0,0 & 0,5 \\ 0,5 & 0,0 \\ 1,0 & 0,0 \end{bmatrix} \quad (2.42)$$

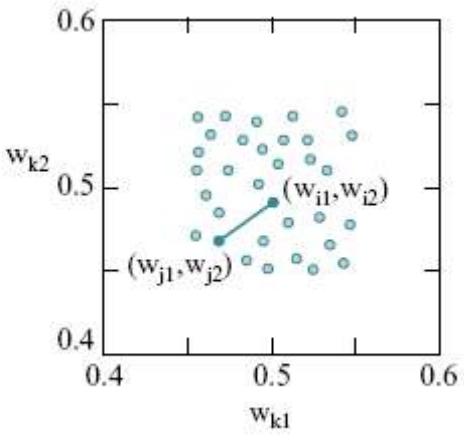
Είναι φανερό, ότι η πρώτη στήλη του τελευταίου πίνακα (2.42), αποτελεί το μέσο όρο των τιμών που αντιπροσωπεύουν τα δείγματα 2 και 4 (σχέση 2.31), ενώ παράλληλα η δεύτερη στήλη, αποτελεί το μέσο όρο των τιμών που αντιπροσωπεύουν τα δείγματα 1 και 3. Αποδεικνύεται έτσι ότι η ταξινόμηση γίνεται με βάση τις παραμέτρους (συνιστώσες του διανύσματος) που χαρακτηρίζουν το κάθε δείγμα. Τα τελικά βάρη αναπαριστούν την κατανομή των εισερχομένων (σχ. 2.14).



Σχήμα 2.14: Παράδειγμα διόρθωσης βαρών του νικητήριου νευρώνα σε ένα Kohonen δίκτυο. Το βάρος W "προσεγγίζει" διαρκώς το εισερχόμενο X [51].

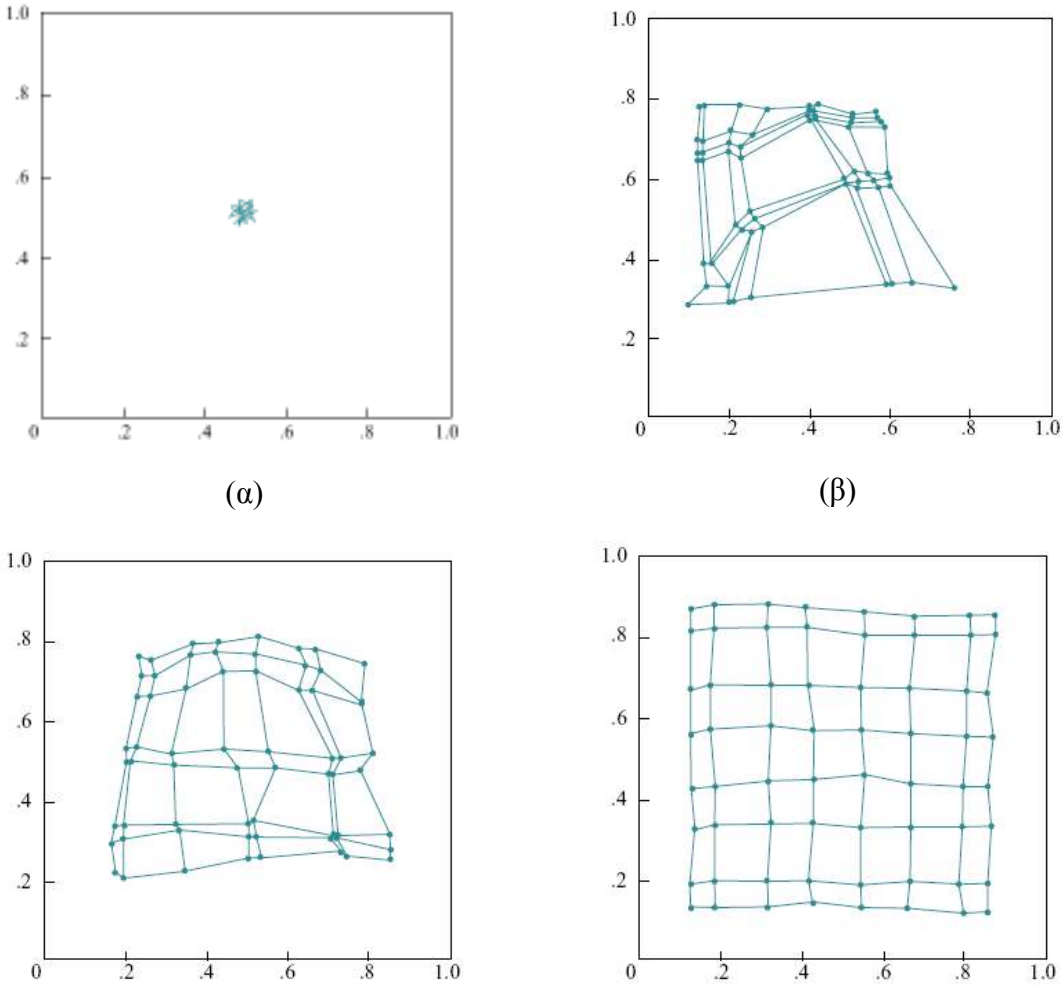
2.2.4. Παράδειγμα κατανόησης 2 (δίκτυο Kohonen)

Έστω ότι έχουμε ένα δίκτυο Kohonen με 2 εισερχόμενα και ένα 8x8 χάρτη. Κάθε εισερχόμενο παίρνει τιμές στο διάστημα (0, 1) που επιλέγονται τυχαία, από μια κατανομή τυχαίων αριθμών. Οι αρχικές τιμές για τα βάρη w επιλέγονται επίσης τυχαία ως εξής: Ξεκινάμε από μια τιμή w_{ij} = 0,5, και προσθέτουμε έναν μικρό τυχαίο αριθμό από το διάστημα (–0,05, 0,05). Έτσι οι αρχικές τιμές των βαρών θα είναι στο διάστημα (0,45, 0,55). Τα αρχικά βάρη φαίνονται στο σχήμα 2. 15 [71].



Σχήμα 2.15: Οι τιμές των αρχικών βαρών wij που επιλέγονται στο διάστημα (0,45, 0,55) για ένα δίκτυο με δυο εισερχόμενες μεταβλητές και 8x8 χάρτη [71].

Στο σχήμα 2.16 (α) έχουμε τη γραφική παράσταση των αρχικών βαρών, όμοια με το σχήμα 2.15, αλλά τώρα οι άξονες έχουν τιμή από 0 έως 1 και έτσι τα σημεία συσσωρεύονται στη μέση του τετραγώνου. Το δίκτυο Kohonen σταδιακά οργανώνεται ξεκινώντας σε χρόνο t = 0 με τη δομή του σχήματος 2.16 (α). Καθόσον το δίκτυο εκπαιδεύεται, μετά από 1000 περιόδους, η δομή του δικτύου φαίνεται στο σχήμα 2.16 (β), όπου αρχίζει σιγά–σιγά να φαίνεται η φυσική τοποθέτηση των νευρώνων. Οι συνδέσεις μεταξύ των νευρώνων είναι παραποιημένες και διαστρεβλωμένες. Παρόλα αυτά, αρχίζει ήδη να διαφαίνεται ότι οι νευρώνες διασυνδέονται μεταξύ τους σε μια δομή πλέγματος. Μετά από t = 6000 περιόδους έχουμε το σχήμα 2.16 (γ). Καθόσον η εκπαίδευση προχωρεί τα βάρη των μονάδων έχουν "απλωθεί" περισσότερο. Μετά από t = 20000 περιόδους έχουμε το σχήμα 2.16 (δ), όπου εδώ οι μονάδες έχουν πάρει πλέον τη φυσική τους θέση και καλύπτουν το τετράγωνο ομοιόμορφα. Οι άξονες x και y έχουν τιμές από 0 ως 1, καθώς αυτές ήταν οι αρχικές τιμές εισόδου στην ομάδα εκπαίδευσης που χρησιμοποιήσαμε.



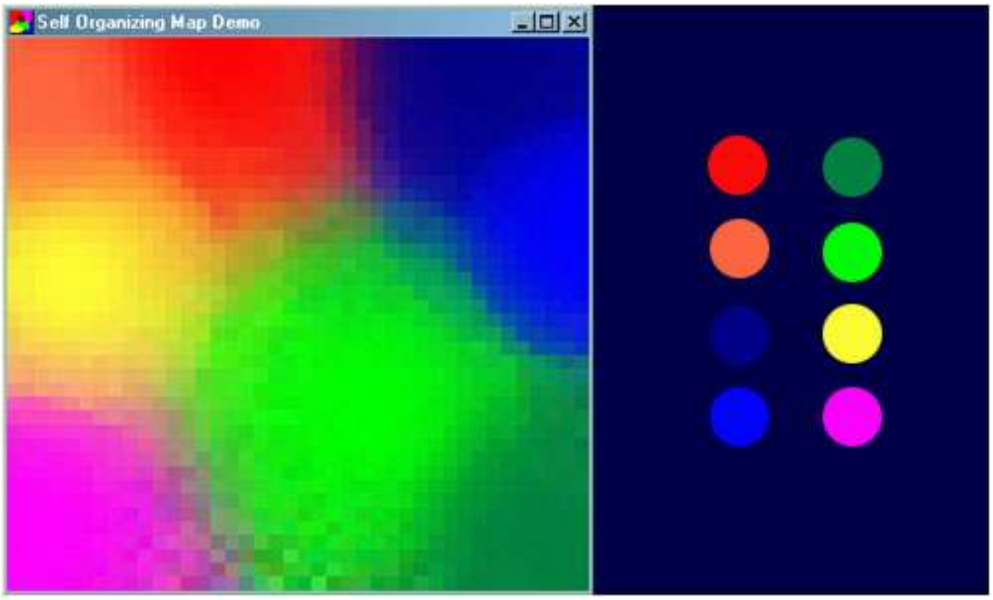
Σχήμα 2.16: (α) Η εκπαίδευση του δικτύου δεν έχει αρχίσει ακόμα, t = 0. (β) Το δίκτυο μετά από εκπαίδευση 1000 περιόδων, t = 1000. (γ) Το δίκτυο μετά από εκπαίδευση 6000 περιόδων, t = 6000. (δ) Το δίκτυο μετά από εκπαίδευση 20000 περιόδων, t = 20000. Στο σημείο αυτό θεωρούμε ότι η εκπαίδευση του μοντέλου έχει τελειώσει [71].

Θεωρούμε ότι το δίκτυο αυτό είναι ένα παράδειγμα αυτοοργάνωσης, καθώς το δίκτυο Kohonen μπορεί από μόνο του να εκπαιδευθεί και αρχίζοντας από τυχαίες τιμές των βαρών να καταλήξει σε μια οργανωμένη δομή, όπως είδαμε στα προηγούμενα σχήματα. Η όλη διαδικασία γίνεται με το να απλωθούν οι νευρώνες του χάρτη, έτσι ώστε κάθε νευρώνας να ομοιάζει σε κάποιο αριθμό δειγμάτων εκπαίδευσης [71].

2.2.5. Παράδειγμα κατανόησης 3 (δίκτυο Kohonen)

Το παρακάτω παράδειγμα χρησιμοποιεί το χάρτη Kohonen για την ταξινόμηση των χρωμάτων [93]. Τα τρία εισερχόμενα αποτελούν τις RGB (RedGreenBlue) διαστάσεις ενός χρώματος. Έτσι, ένας 2x4 χάρτης (σχ. 2.17) θα αντιπροσώπευε 8 διαφορετικές αποχρώσεις. Στο χάρτη αυτό, τα βασικά χρώματα κόκκινο (Red), πράσινο (Green) και μπλε (Blue) θα καταλάμβαναν τις 3 γωνίες. Αν αντίθετα τα εισερχόμενα αφορούσαν μόνο διαστάσεις του κόκκινου, ο χάρτης θα περιείχε μόνο νευρώνες με τις αντίστοιχες αποχρώσεις.

Με το παραπάνω παράδειγμα επιτεύχθηκε η μείωση των τριών διαστάσεων σε δυο. Επίσης, εύκολα γίνεται αντιληπτό ότι εκτός της ταξινόμησης των χρωμάτων, παρατηρείται η τοποθέτηση παρόμοιων αποχρώσεων σε γειτονικούς νευρώνες



Σχήμα 2.17: Ταξινόμηση των 3D χρωμάτων (αριστερά) σε δυσδιάστατο επίπεδο (δεξιά) [93].

2.2.6. Υβριδικά Δίκτυα (Ensembles Networks) και επαναδειγματοληψία (resampling)

Με την επαναδειγματοληψία τελευταία, για κάθε νέο δίκτυο (από τα υβριδικά), γίνεται νέος διαχωρισμός της βάσης δεδομένων (σε ομάδες εκπαίδευσης, επικύρωσης και ελέγχου), ώστε τελικά όλα τα δείγματα να χρησιμοποιούνται στην ομάδα εκπαίδευσης. Έτσι, η διακύμανση του τελικού υβριδικού μοντέλου μπορεί να ελαττωθεί αποτελεσματικά, εφόσον κάθε φορά χρησιμοποιείται διαφορετική ομάδα εκπαίδευσης. Το ίδιο

αποτέλεσμα θα έχουμε όταν επιλέγουμε διαφορετικά αρχικά δίκτυα (βλ. παρακάτω) [94]. Επιπλέον, η απόδοση του υβριδικού δικτύου μπορεί να βελτιωθεί αποτελεσματικά. Η διαδικασία εύρεσης του μέσου όρου (ή της πλειοψηφίας) των μοντέλων, μειώνει την variance χωρίς να αυξάνει το bias. Από την άλλη μεριά, μπορούμε να "ανεχθούμε" πολυπλοκότερα μοντέλα (με ανάλογο bias), με τη λογική ότι το υβριδικό μοντέλο θα "απαλύνει" την τελική variance [7].

Με βάση τα παραπάνω, η ικανότητα γενίκευσης των υβριδικών μοντέλων μπορεί να είναι καλύτερη από την αντίστοιχη για το βέλτιστο των μεμονωμένων δικτύων, παρά το γεγονός ότι αυτό εξαρτάται από το πόσο καλά είναι αυτά τα δίκτυα [7, 95]. Ωστόσο, υπάρχουν δυο σημεία στα οποία πρέπει να δοθεί έμφαση κατά την "κατασκευή" των υβριδικών δικτύων:

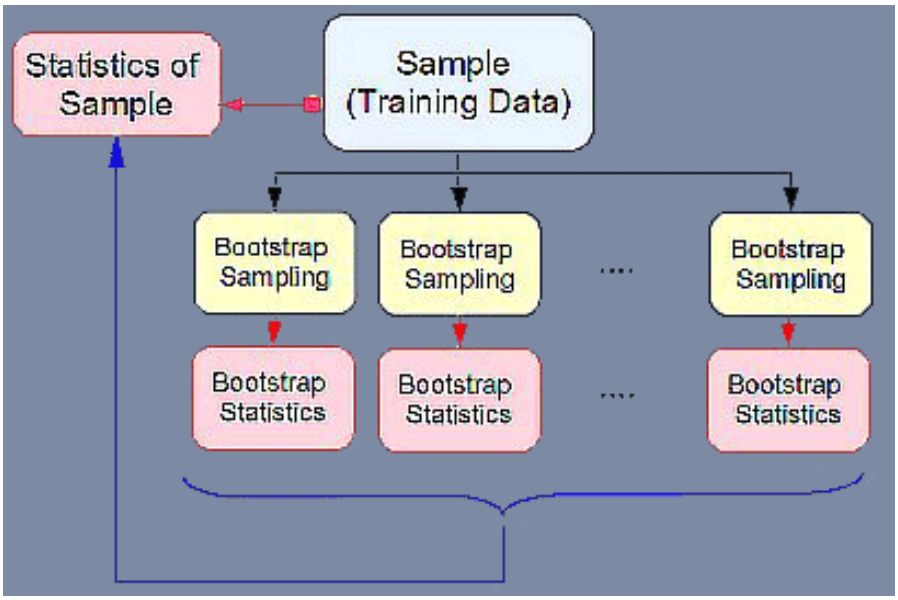
1. Όταν τα μεμονωμένα δίκτυα είναι μη συσχετιζόμενα, το υβριδικό δίκτυο μειώνει το σφάλμα κατά ένα παράγοντα N, όπου N ο αριθμός των αρχικών μεμονωμένων δικτύων. Αυτό σημαίνει ότι όσο μη συσχετιζόμενα είναι τα αρχικά δίκτυα, τόσο πιο αποτελεσματικό είναι το τελικό υβριδικό δίκτυο [7]. Ένας συνδυασμός δικτύων είναι χρήσιμος μόνο όταν τα αρχικά δίκτυα "διαφωνούν" για κάποια αποτελέσματα. "Είναι σαφές, ότι καμιά παραπάνω πληροφορία δεν θα μπορέσουμε να εξάγουμε από ακόμα και ένα εκατομμύριο ολόιδια δίκτυα, από ότι μόνο από ένα από αυτά" [94]. Τα διαφορετικά δίκτυα κάνουν συστηματικά σφάλματα σε διαφορετικά σημεία του χώρου των εισερχομένων. Ο μέσος όρος αυτών ωστόσο, μπορεί να "ισοπεδώσει" κάποιο από αυτό το bias [7].
2. Το αναμενόμενο σφάλμα του υβριδικού δικτύου είναι τουλάχιστον τόσο όσο ο μέσος όρος του σφάλματος των μεμονωμένων δικτύων και συνήθως καλύτερο από αυτό. Κάτι τέτοιο βέβαια, προϋποθέτει κάποιο κόστος στην ταχύτητα της διεργασίας, αλλά αυτή για τις περισσότερες εφαρμογές είναι ικανοποιητική.
3. Η απόδοση του τελικού δικτύου εξαρτάται από τα αρχικά επιλεγόμενα δίκτυα. Είναι εύλογο λοιπόν, ότι το υβριδικό δίκτυο δεν θα έχει καλά αποτελέσματα αν επιλεγούν "κακά" αρχικά δίκτυα [33].

Η επαναδειγματοληψία των δεδομένων μπορεί να γίνει με τρεις τρόπους:

1. Η πιο απλή εκδοχή είναι η **τυχαία** (Monte-Carlo), όπου οι ομάδες εκπαίδευσης, επικύρωσης και ελέγχου επιλέγονται τυχαία από τη βάση των δεδομένων, κρατώντας απλά σταθερό τον αριθμό των δειγμάτων [7].
2. Η δεύτερη επιλογή είναι η **διασταυρούμενη αξιολόγηση** (cross-validation). Οι Krogh και Velesby εδώ προτείνουν ότι αν υπάρχει ένα εισερχόμενο δείγμα, για το

οποίο τα δίκτυα "διαφωνούν" έντονα, είναι καλύτερα να συμπεριληφθεί στην ομάδα εκπαίδευσης [94].

3. Η τρίτη εκδοχή είναι το **bootstrapping.** Η ονομασία της τεχνικής προκύπτει από τη φράση "pull up by your own bootstraps", με την έννοια ότι βασίζεται στις δικές της δυνατότητες [96]. Η τεχνική δεν έχει ανάγκη την χρήση μιας ανεξάρτητης ομάδας ελέγχου [37], καθώς μια νέα ομάδα εκπαίδευσης δημιουργείται με **δειγματοληψία αντικατάστασης ή επανατοποθέτηση** ("sampling with replacement") από τη διαθέσιμη ομάδα. Με αυτό τον τρόπο, τυχαία δείγματα επιλέγονται από τη βάση δεδομένων, οι συνδυασμοί δειγμάτων (στις ομάδες εκπαίδευσης-ελέγχου) αυξάνονται πολλαπλά, ενώ η ίδια πιθανότητα αντιστοιχεί σε καθένα από αυτά. Κάποια από τα δεδομένα μπορεί να μην επιλεγούν, οπότε και χρησιμοποιούνται για την ομάδα ελέγχου, ενώ άλλα μπορεί να "διπλοεγγραφούν" [7, 46, 97, 98]. Η πορεία του bootstrapping δημιουργεί κατά μια έννοια πολλαπλά δεδομένα από την αρχική βάση. Το αποτέλεσμα είναι η μείωση της προκατάληψης (ακριβέστερα αποτελέσματα) και η εξισορρόπηση (average out) της διακύμανσης [7, 58]. Επιπλέον, η θεωρία της τεχνικής λέει ότι η εκτιμήτρια της δειγματοληψίας αντικατάστασης (δηλαδή της bootstrapping δειγματοληψίας) είναι ίση με την εκτιμήτρια του δείγματος. Έτσι με τον τρόπο αυτό, αποκτά κανείς καλύτερη πρόσβαση στα στατιστικά δεδομένα του δείγματος: οι τιμές από τις N bootstrapping δειγματοληψίες του δείγματος (σχ. 2.18) μπορούν να χρησιμοποιηθούν για την αξιολόγηση των στατιστικών του αρχικού δείγματος [96].



Σχήμα 2.18: Παράδειγμα bootstrapping δειγματοληψίας για την εκπαίδευση ενός μοντέλου [96].

2.2.7. Δίκτυα Προσαρμόσιμου Συντονισμού

Τα **Δίκτυα Προσαρμόσιμου Συντονισμού** (Adaptive Resonance theory Networks, ART), σχεδιάστηκαν επίσης ως μια μη-επιβλεπόμενη τεχνική αναγνώρισης προτύπων. Επικράτησαν για να καλύψουν ένα μειονέκτημα των τεχνικών ταξινόμησης (πχ των MLP), που αφορά το γεγονός ότι μετά το τέλος της "εκπαίδευσης", τα βάρη είναι πια αμετάβλητα και μη προσαρμόσιμα. Το κάθε δίκτυο ή μοντέλο ταξινόμησης, σχεδιάζεται για ένα συγκεκριμένο καθήκον: "εκπαιδεύεται" από μια αντιπροσωπευτική ομάδα δειγμάτων και πειραματίζεται σε νέα δείγματα που επίσης είναι ή πρέπει να είναι σταθερής και αναμενόμενης "σύστασης". Δυστυχώς όμως, σε πραγματικά προβλήματα, δεν συμβαίνει πάντα αυτό. Μπορεί λοιπόν να συμβεί η εμφάνιση νέων δειγμάτων, που δεν αντιπροσωπεύουν την ομάδα εκπαίδευσης. Η μοναδική λύση στην περίπτωση αυτή, είναι η επανεκπαίδευση του μοντέλου με την χρήση δειγμάτων που ανήκουν στη νέα ομάδα. Μπορεί επίσης δείγματα που ανήκουν σε παλιές ομάδες να αλλάζουν σύσταση με το χρόνο. Και εδώ η λύση είναι η επανεκπαίδευση του δικτύου με δείγματα που αντιπροσωπεύουν τώρα τη νέα σύσταση της ομάδας. Το δίλημμα που προκύπτει λοιπόν είναι: πως μπορεί ένα σύστημα να είναι προσαρμόσιμο σε νέα δεδομένα, ενώ συγχρόνως να παραμένει σταθερό σε "άσχετες" αλλαγές αυτών;

Σε απάντηση, αναπτύχθηκαν τα ART δίκτυα: όταν ένα "οικείο" δείγμα εισέρχεται στο δίκτυο (δηλαδή δείγμα που ικανοποιεί τις συνθήκες των προηγούμενων δειγμάτων), το

δίκτυο το αναγνωρίζει και ενσωματώνει τη νέα, περιορισμένη πληροφορία που αυτό μεταφέρει για την προσαρμογή των βαρών του. Όταν όμως, ένα "ξένο" δείγμα παρουσιάζεται που δεν ικανοποιεί τις συνθήκες των προηγούμενων δειγμάτων, η δομή του δικτύου αναπροσαρμόζεται και το "ξένο" δείγμα αναγνωρίζεται ως ο πρώτος αντιπρόσωπος της νέας τάξης [51]. Με αυτόν τον τρόπο, τα ART δίκτυα δίνουν λύση στο δίλημμα της σταθερότητας-ευελιξίας ("stability-plasticity dilemma"[99]): είναι σταθερά και μπορούν να διατηρήσουν τους κανόνες της παλιάς εκπαίδευσης, αλλά συγχρόνως παραμένουν προσαρμόσιμα, ώστε να "απορροφούν" κάθε νέα πληροφορία, εφόσον αυτό χρειαστεί [51, 99].

2.2.8. Μέθοδοι επικύρωσης (validation of the models)

Εδώ αναφέρονται μερικές λιγότερο συνηθισμένες μέθοδοι επικύρωσης των ANN μοντέλων.

Έτσι, ο Rao et al. [100] εξάλλου, χρησιμοποιεί την **y-scrambling ή randomization test μέθοδο** για να επικυρώσει τα μοντέλα του. Η μέθοδος ελέγχει παράλληλα για την ύπαρξη τυχαίας συσχέτισης (chance correlation), δηλαδή συσχέτισης που επιτυγχάνεται τυχαία από το μοντέλο, χωρίς το ίδιο να διαθέτει ικανότητα πρόβλεψης. Σύμφωνα με τη μέθοδο αυτή, οι χαρακτηρισμοί των ομάδων (ή οι αποκρίσεις για ένα μοντέλο συσχέτισης), μετατίθενται τυχαία πολλές φορές και παράγονται πολλές νέες βάσεις δεδομένων. Στη συνέχεια, ακολουθείται η συνήθης διαδικασία βελτιστοποίησης του μοντέλου και τα υπολογιζόμενα ποσοστά πρόβλεψης των νέων μοντέλων πρέπει να είναι μικρότερα από της αρχικής αληθινής βάσης δεδομένων. Αναφορά ή / και εφαρμογή της μεθόδου αυτής, βλέπουμε και αλλού [50, 58].

Ο **συντελεστής k** ("Cohen's kappa coefficient") παρέχει επίσης μια εναλλακτική μέθοδο για τον έλεγχο της τυχαίας συσχέτισης στα μοντέλα ταξινόμησης. Λεπτομέρειες για τον υπολογισμό του συντελεστή k αναφέρονται αναλυτικά στη βιβλιογραφία [101, 102].

2.2.9. Αλλαγή κλίμακας (scaling) των δεδομένων

Μερικές λιγότερο συνηθισμένες μέθοδοι αλλαγής κλίμακας (scaling) είναι:

1. **Constant row sum:** κάθε μεταβλητή διαιρείται με το άθροισμα όλων των μεταβλητών για ένα δείγμα [40].

2. **Normalization variable:** οι μεταβλητές κανονικοποιούνται (normalized), σε σχέση με μια μεταβλητή [40, 103].
3. **Range transformation:** η ελάχιστη τιμή μιας μεταβλητής τίθεται στο μηδέν (0), η μέγιστη στο ένα (1), και όλες οι ενδιάμεσες τιμές κυμαίνονται μεταξύ αυτών [40, 104].
4. **Log transformation:** ελαττώνει την ασυμμετρία (skewness) μιας κατανομής και μπορεί να μειώσει τη επίδραση ακραίων τιμών [105].

Υπάρχουν ακόμα λιγότερο συνήθεις μέθοδοι, που χρησιμοποιούνται σε ειδικές περιπτώσεις.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Marini, F., Magrì, A.L., Marini, D., Balestrieri, F., Characterization of the lipid fraction of Niger seeds (Guizotia abyssinica cass.) from different regions of Ethiopia and India and chemometric authentication of their geographical origin, Eur. J. Lipid Sci. Technol., 2003, **105,** 697-704.
2. Marini, F., Balestrieri, F., Bucci, R., Magrì, A.L., Marini, D., Supervised pattern recognition to discriminate the geographical origin of rice bran oils: a first study, Microchem. J., 2003, **74,** 239-248.
3. Alvarez-Guerra, M., Ballabio, D., Amigo, J.M., Viguri, J.R., Bro, R., A chemometric approach to the environmental problem of predicting toxicity in contaminated sediments, J. Chemometr., 2009, **24,** 379-386.
4. Ντζούφρας, Ι., Σημειώσεις για το μάθημα Ανάλυση Δεδομένων Ι, Στοιχεία Πολυμεταβλητής Ανάλυσης Δεδομένων, Τμήμα Διοίκησης Επιχειρήσεων, Πανεπιστήμιο Αιγαίου, 2001.
5. Moreno, I.M., González-Weller, D., Gutierrez, V., Marino, M., Cameán, A.M., González, A.G., Hardisson, A., Differentiation of two Canary DO red wines according to their metal content from inductively coupled plasma optical emission spectrometry and graphite furnace atomic absorption spectrometry by using Probabilistic Neural Networks, Talanta, 2007, **72,** 263-268.
6. http://www.scribd.com/doc/25044857/discriminant-Function-Analysis (τελευταία επίσκεψη 12/12/2010).
7. STATISTICA 7th edition, software, StatSoft, Inc., 2004.
8. Friel, C.M., Notes on Discriminat Analysis, Criminal Justice Center, Sam Houston State University.
9. http://127.0.0.1:49312/help/index.jsp?topic=/com.ibm.spss.statistics.cs/discriminant_table.htm (τελευταία επίσκεψη 15/12/2010).
10. Miller, J.N., and Miller, J.C., Statistics and Chemometrics for Analytical Chemistry, 5th ed., Pearson, England, 2005.
11. Παπαγρηγορίου Ν., Πτυχιακή εργασία, Τμήμα Τεχνολογίας Τροφίμων, Σχολής Τεχνολογίας Τροφίμων και Διατροφής, ΤΕΙ Θεσσαλονίκης, 2001.
12. Zupan, J., Gasteiger, J., Neural Networks for Chemists; An Introduction. VCH Verlagsgescllschaft, Weinhcim, Germany and VCH Publishers, New York, USA, 1993.

13. Massart, D.L., Vandeginste, B.G.M., Deming, S.N., Michotte, Y. and Kaufman, L, Chemometrics: a textbook, Elsevier, 1988.
14. Smith, L.I., A tutorial on Principal Components analysis, February 26, 2002, http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf (τελευταία επίσκεψη 5/2/2010).
15. http://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors (τελευταία επίσκεψη 1/4/2011).
16. Φαρμάκη, Ε., Ερευνητική εργασία διπλώματος ειδίκευσης, Τμήμα Αναλυτικής Χημείας, ΕΚΠΑ, 2007
17. Wold, S., Esbensen, K., Geladi, P., Principal Component Analysis, Chemometr. Intell. Lab., 1987, **2**, 37-52.
18. Spanos, T., Simeonov, V., Simeonova, P., Apostolidou, E., Stratis, J., Environmetrics to evaluate marine environment quality, Environ. Monit. Assess., 2008, **143,** 215-225.
19. http://onlinecourses.science.psu.edu/stat505/node/80 (τελευταία επίσκεψη Σεπτέμβριος 2011)
20. IBM SPSS Statistics 19th edition, software, SPSS, Inc., 2010.
21. http://people.revoledu.com/kardi/tutorial/kMean/WhatIs.htm (τελευταία επίσκεψη Απρίλιος 2011).
22. Garcia, H.L., González, I.M., Self-organizing map and clustering for wastewater treatment monitoring, Artif. Intell., 2004, **17,** 215-225.
23. Chan, Y.H., Biostatistics 304. Cluster analysis, Singapore Med. J., 2005, **46(4),** 153-160.
24. Vesanto, J., Alhoniemi, E., Clustering of the Self-Organizing Map, IEEE Transactions on Neural Networks, 2000, **11,** 586-600.
25. L10, Bio 8100S Applied Multivariate Biostatistics, University of Ottawa, 2001.
26. Loh, W-Y., Shih, Y-S., Split Selection Methods for Classification Trees, Stat. Sinica, 1997, **7**, 815-840.
27. Smeti, E.M., Thanasoulias, N.C., Lytras, E.S., Tzoumerkas, P.C., Golfinopoulos, S.K., Treated water quality assurance and description of distribution networks by multivariate chemometrics, Water Res., 2009, **43,** 4676-4684.
28. Cheng, W., Zhang X., Wang, K., Dai, X., Integrating classification and regression tree (CART) with GIS for assessment of heavy metals pollution, Environ. Monit. Assess., 2009, **158,** 419-431.

29. Simeonova, P., Simeonov, V., Chemometrics to evaluate the quality of water sources for human consumption, Microchim. Acta, 2007, **156**, 315-320.
30. Stanimirova, I., Kubik, A., Walczak, B., Einax, J.W., Discrimination of biofilm samples using pattern recognition techniques, Anal. Bioanal. Chem., 2008, **390**, 1273-1282.
31. Ocampo-Duque, W., Schuhmacher, M., Domingo, J.L., A neural-fuzzy approach to classify the ecological status in surface waters, Environ. Pollut., 2007, **148**, 634-641.
32. Kröse, B., Van der Smagt, P., An introduction to Neural Networks, The University of Amsterdam, Amsterdam 1996.
33. Fernandes, F.A.N., Lona, L.M.F., Neural network application in polymerization processes, Braz. J. Chem. Eng. 2005, **22**, 401–418.
34. Fausett, L., Fundamentals of Neural Networks–Architectures, Algorithms and Applications, Prentice-Hall, Inc. Upper Saddle River, NJ, USA 1994.
35. Lancashire, L.J., Lemetre, C., Ball, G.R., An introduction to artificial neural networks in bioinformatics application to complex microarray and mass spectrometry datasets in cancer studies, Briefings in Bioinformatics, 2009, **10,** 315-329.
36. http://www.sussex.ac.uk/Users/andrewop/Courses/NN/NNs5_6_MLP.ppt (τελευταία επίσκεψη Σεπτέμβριος 2010).
37. Zur, R.M., Jiang, Y., Pesce, L.L.,Drukker, K., Noise injection for training artificial neural networks: A comparison with weight decay and early stopping, Med. Phys., 2009, **36,** 4810-4818.
38. LeBouf, R.F., Schuckers, S.A., Rossner, A., Preliminary assessment of a model to predict mold contamination based on microbial volatile organic compound profiles, Sci. Total Environ., 2010, **408,** 3648-3653.
39. Sietsma, J., Dow, R.J.F., Creating artificial neural networks that generalize, Neural Networks, 1991, **4,** 67-79.
40. Berrueta, L.A., Alonso-Salces, R.M., Héberger, K., Supervised pattern recognition in food analysis, J. Chromatogr. A, 2007, **1158,** 196-214.
41. Maier, H.R., Dandy, G.C. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications, Environ. Modell. Softw., 2000, **15,** 101-124.
42. Peres, A.M., Baptista, P., Malheiro, R., Dias, L.G., Bento, A., Pereira, J.A., Chemometric classification of several olive cultivars from Trás-os-Montes region

(northeast of Portugal) using artificial neural networks, Chemometr. Intell. Lab., 2011, **105,** 65-73.

43. Curteanu, S., Cartwright, H., Neural networks applied in chemistry. I. Determination of the optimal topology of multilayer perceptron neural networks, J. Chemometr., 2011, **25,** 527-549.
44. Palani, S., Liong, S-Y., Tkalich, P., An ANN application for water quality forecasting, Mar. Pollut. Bull., 2008, **56,** 1586-1597.
45. Huang, W., Foo, S., Neural network modeling of salinity variation in Apalachicola River, Water Res., 2002, **36,** 356-362.
46. Das, R., Sengur, A., Evaluation of ensemble methods for diagnosing of valvular heart disease, Expert Syst., Appl., 2010, **37,** 5110-5115.
47. Lee, S.C., Lin, H.T., Yang, T.Y., Artificial neural network analysis for reliability prediction of regional runoff utilization, Environ. Monit. Assess., 2010, **161,** 315-326.
48. Hanafizadeh, P., Ravasan, A.Z., Khaki, H.R., An expert system for perfume selection using artificial neural network, Expert Syst. Appl., 2010, **37,** 8879-8887.
49. Zhang, H., Zhou, Y., Cheng, P., Deng, S., Cui, X., Wang, H., Multi-objective simultaneous prediction of waterborne coating properties J. Math. Chem., 2009, **46,** 1050-1059.
50. Darnag, R., Schmitzer, A., Belmiloud, Y., Villemin, D., Jarid, A., Chat, A., Mazouz, E., Cherqaoui, D., Quantitative structure-activity relationship studies of TIBO derivatives using support vector machines, SAR QSAR Environ. Res., 2010, **21,** 231-246.
51. Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., De Jong, S., Lewi, P.J., Smeyers-Verbeke, J., Handbook of Chemometrics and Qualimetrics: Part B., Elsevier, Amsterdam, 1998.
52. Carlucci, G., D' Archivio, A.A., Maggi, M. A., Mazzeo, P., Ruggieri, F., Investigation of retention behaviour of non-steroidal anti-inflammatory drugs in high-erformance liquid chromatography by using quantitative structure–retention relationships, Anal. Chim. Acta, 2007, **601,** 68-76.
53. Marini, F. Balestieri F., Bucci, R., Magrì, A.D., Magrì, A.L., Marini, D., Supervised pattern recognition to authenticate Italian extra virgin olive oil varieties, Chemometr. Intell. Lab., 2004, **73,** 85-93.

54. Marini, F., Bucci, R., Magrì, A. L., Magrì, A. D., Acquistucci, R., Francisci R., Classification of 6 durum wheat cultivars from Sicily (Italy) using artificial neural networks, Chemometr. Intell. Lab., 2008, **90,** 1-7.
55. Galvão, R.K.H., Araujo, M.C.U., José, G.E., Pontes, M.J.C., Silva, E.C., Saldanha, T.C.B., A method for calibration and validation subset partitioning, Talanta, 2005, **67,** 736-740.
56. Fatemi, M.H., Ghorbanzad'e, M., Classification of drugs according to their milk/plasma concentration ratio, Eur. J. Med. Chem., 2010, **45,** 5051-5055.
57. Marini, F. Magrì, A.L., Bucci, R., Magrì, A.D., Use of different artificial neural networks to resolve binary blends of monocultivar Italian olive oils, Anal. Chim. Acta, 2007, **599,** 232-240.
58. Gramatica, P., Principles of QSAR models validation: internal and external, QSAR Comb. Sci., 2007, **26,** 694-701.
59. Kutlu, Y., Kuntalp, M., Kuntalp, D., Optimizing the performance of an MLP classifier for the automatic detection of epileptic spikes, Expert Syst., Appl., 2009, **36,** 7567-7575.
60. Nguyen, H.T., Prasad, N.R., Walker, C.L., Walker, E.A., A First Course in Fuzzy and Neural Control, Chapman & Hall/CRC, Boca Raton, Florida, USA 2003.
61. Zhang, G., Patuwo, B.E., Hu, M.Y., Forecasting with artificial neural networks: The state of the art, Int. J. Forecasting, 1998, **14,** 35-62.
62. Wesolowski, M., Suchacz, B., Halkiewicz, J., The analysis of seasonal air pollution pattern with application of neural networks, Anal. Bioanal. Chem.., 2006, **384,** 458-467.
63. Yesilnacar, M.I., Sahinkaya, E., Naz, M., Ozkaya, B., Neural network prediction of nitrate in groundwater of Harran Plain, Turkey, Environ. Geol., 2008, **56,** 19-25.
64. Tagluk, M.E., Akin, M., Sezgin, N., Classıfication of sleep apnea by using wavelet transform and artificial neural networks, Expert Syst. Appl., 2010, **37,** 1600-1607.
65. Kuo, J-T., Hsieh, M-H., Lung, W-S., She, N., Using artificial neural network for reservoir eutrophication prediction, Ecol. Model., 2007, **200,** 171-177.
66. Kuzmanovski, I., Novič, M., Counter-propagation neural networks in Matlab, Chemometr. Intell. Lab., 2008, **90,** 84-91.
67. Rene, E. R., Saidutta, M. B., Prediction of water quality indices by regression analysis and artificial neural networks, Int. J. Environ. Res., 2008, **2(2),** 183-188.

68. Balabin, R.M., Safieva, R.Z., Biodiesel classification by base stock type (vegetable oil) using near infrared spectroscopy data, Anal. Chim. Acta, 2011, **689,** 190-197.
69. Balabin, R.M., Safieva, R.Z., Lomakina, E.I., Near-infrared (NIR) spectroscopy for motor oil classification: From discriminant analysis to support vector machines, Microchem. J., 2011, **98,** 121-128.
70. Rumelhart, D.E., Hinton, G.E., Williams, R. J., Nature, Learning representations by back-propagating errors, 1986, **323,** 533-536.
71. Αργυράκης Π., Νευρωνικά Δίκτυα και Εφαρμογές, Ελληνικό Ανοιχτό Πανεπιστήμιο, Πάτρα 2001.
72. Kompany-Zareh, M., Massoumi, A., Pezeshk-Zadeh, Sh., Simultaneous spectrophotometric determination of Fe and Ni with xylenol orange using principal component analysis and artificial neural networks in some industrial samples, Talanta, 1999, **48,** 283-292.
73. Farmaki, E.G., Thomaidis, N.S., Efstathiou, C.E., Artificial Neural Networks in water analysis: Theory and applications, Int. J. Environ. An. Ch., 2010, **90,** 85-105.
74. Ramadan, Z., Hopke, P.K., Johnson M.J., Scow K.M., Application of PLS and Back-Propagation Neural Networks for the estimation of soil properties, Chemometr., Intell. Lab., 2005, **75,** 23-30.
75. Kim, Y.S., Performance evaluation for classification methods: A comparative simulation study, Expert Syst. Appl., 2010, **37,** 2292-2306.
76. Alfassi, Z.B., Boger, Z., Ronen, Y., Statistical Treatment of Analytical Data, Blackwell Science Ltd, Oxford, UK 2005.
77. Svozil, D., Kvasnička, V., Pospíchal, J., Introduction to multi-layer feed-forward neural networks, Chemometr. Intell. Lab., 1997, **39,** 43-62.
78. Luengo, J., García, S., Herrera, F., A study on the use of statistical tests for experimentation with neural networks: Analysis of parametric test conditions and non-parametric tests, Expert Syst. Appl., 2009, **36,** 7798-7808.
79. Hernández-Caraballo, E.A., Ávila de Hernández, R.M., Rivas-Echeverría, F., Capote-Luna, T., Discrimination of Venezuelan spirituous beverages by a trace element-radial basis neural network approach, Talanta, 2008, **74,** 871-878.
80. Spanilá, M., Pazourek, J., Farková, M., Havel, J., Optimization of solid-phase extraction using artificial neural networks in combination with experimental design for determination of resveratrol by capillary zone electrophoresis in wines, J. Chromatogr. A., 2005, **1084,** 180-185.

81. Kruzlicova, D., Mocak, J., Balla, B., Petka J., Farkova, M., Havel, J., Classification of Slovak white wines using artificial neural networks and discriminant techniques, Food Chem., 2009, **112,** 1046-1052.
82. Gulbag, A., Temurtas, F., Tasaltin, C., Öztürk, Z.Z., A study on radial basis function neural network size reduction for quantitative identification of individual gas concentrations in their gas mixtures, Sensor. Actuat. B-Chem., 2007, **124,** 383-392.
83. Stavrou, E.T., Charalambous, C., Spiliotis, S., Human resource management and performance: A neural network analysis, Eur. J. Oper. Res., 2007, **181,** 453-467.
84. http://www.ivie.es/downloads/ws/bf/2003/05/08/ponencia05.pdf (τελευταία επίσκεψη 17/11/2010).
85. Bieroza, M., Baker, A., Bridgeman, J., Classification and calibration of organic matter fluorescence data with multiway analysis methods and artificial neural networks: an operational tool for improved drinking water treatment, Environmetrics, 2011, **22,** 256-270.
86. Samsonova, E., Kok, J. N., Ijzerman, A.P., TreeSOM: Cluster analysis in the self-organizing map, Neural Networks, 2006, **19,** 935-949.
87. Ballabio, D., Consonni, V., Todeschini, R., The Kohonen and CP-ANN toolbox: A collection of MATLAB modules for Self Organizing Maps and Counterpropagation Artificial Neural Networks, Chemometr. Intell. Lab., 2009, **98,** 115-122.
88. Alvarez-Guerra, M., González-Piñuela, C., Andrés, A., Galán, B., Viguri, J.R., Assessment of Self-Organizing Map artificial neural networks for the classification of sediment quality, Environ. Int., 2008, **34,** 782-790.
89. Alvarez-Guerra, E., Molina, A., Viguri, J., Alvarez-Guerra, M., A SOM-Based Methodology for Classifying Air Quality Monitoring Stations, Environ. Prog. Sust. Energy, 2011, **30(3),** 424-438.
90. Lamrini, B., Lakhal, El-K., Le Lann, M-V., Data validation and missing data reconstruction using self-organizing map for water treatment, Neural Comput. Appl., 2011, **20,** 575-588.
91. Xiao, Y.-D., Clauset, A., Harris, R., Bayram, E., Santago, P., Schmitt, J.D., Supervised Self-Organizing Maps in Drug Discovery. 1. Robust Behavior with Overdetermined Data Sets, J. Chem. Inf. Model., 2005, **45,** 1749-1758.
92. Bodt, E., Cottrell, M., Verleysen, M., Statistical tools to assess the reliability of self-organizing maps, Neural Networks, 2002, **15,** 967-978.
93. http://www.ai-junkie.com/ann/som/som1

(τελευταία επίσκεψη Νοέμβριος 2011).

94. Krogh, A., Vedelsby, J., Neural Network Ensembles, Cross Validation, and Active Learning, in: Tesauro, G., Touretzky, D.S., Leen, T.K., (Eds.), Advances in Neural Information Processing Systems 7, p. 231-238, MIT Press, Cambridge MA, 1995.
95. Zhou, Z-H., Wu, J., Tang, W., Ensembling neural networks: Many could be better than all, Artif. Intell., 2002, **137,** 239–263.
96. http:\\people.revoledu.com\kardi\tutorial\bootstrap
(τελευταία επίσκεψη Μάρτιος 2011).
97. Otto, M., Chemometrics, Statistics and Computer Application in Analytical Chemistry, Wiley-VCH, Weinheim, Germany 1999.
98. Brereton, R.G., Lloyd, G.R., Support Vector Machines for classification and regression, Analyst, 2010, **135,** 230-267.
99. Tan, S.C., Lim, C.P., Integration of supervised ART-based neural networks with a hybrid genetic algorithm, Soft Comput., 2011, **15,** 205-219.
100. Rao, H., Yang, G., Tan, N., Li, P., Li, Z., Li, X., Prediction of HIV-1 Protease Inhibitors Using Machine Learning Approaches, QSAR Comb. Sci., 2009, **28,** 1346-1357.
101. Everaert, G., Boets, P., Lock K., Džeroski, S., Goethals, P.L.M., Using classification trees to analyze the impact of exotic species on the ecological assessment of polder lakes in Flanders, Belgium, Ecol. Model., 2011, **222,** 2202-2212.
102. Landis, J.R., Koch, G.G., The Measurement of Observer Agreement for Categorical Data, Biometrics, 1977, **33,** 159-174.
103. Cajka, T., Riddellova, K., Klimankova, E., Cerna, M., Pudil F., Hajslova, Traceability of olive oil based on volatiles pattern and multivariate analysis, Food Chem., 2010, **121,** 282-289.
104. Cajka, T., Riddellova, K., Tomaniova, M., Hajslova J., Recognition of beer brand based on multivariate analysis of volatile fingerprint, J. Chromatogr. A, 2010, **1217,** 4195-4203.
105. Jin, Y.-H., Kawamura, A., Park, S.-C., Nakagawa, N., Amaguchi, H., Olsson, J., Spatiotemporal classification of environmental monitoring data in the Yeongsan River basin, Korea, using self-organizing maps, J. Environ. Monit., 2011, **13,** 2886-2894.

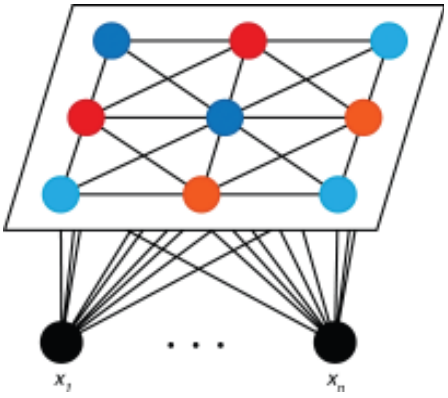
ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΧΗΜΕΙΑΣ

ΠΕΙΡΑΜΑΤΙΚΟ ΜΕΡΟΣ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ

ΕΦΑΡΜΟΓΕΣ ΠΟΛΥΠΑΡΑΜΕΤΡΙΚΩΝ ΣΤΑΤΙΣΤΙΚΩΝ ΤΕΧΝΙΚΩΝ ΣΤΗ ΧΗΜΙΚΗ ΑΝΑΛΥΣΗ



ΕΛΕΝΗ ΦΑΡΜΑΚΗ

ΧΗΜΙΚΟΣ

ΑΘΗΝΑ

ΙΟΥΝΙΟΣ 2012

ΠΑΡΑΡΤΗΜΑ (ΠΕΙΡΑΜΑΤΙΚΟ ΜΕΡΟΣ)

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΕΦ. 1 ΤΑΜΙΕΥΤΗΡΕΣ ΠΟΣΙΜΟΥ ΥΔΑΤΟΣ ΑΤΤΙΚΗΣ

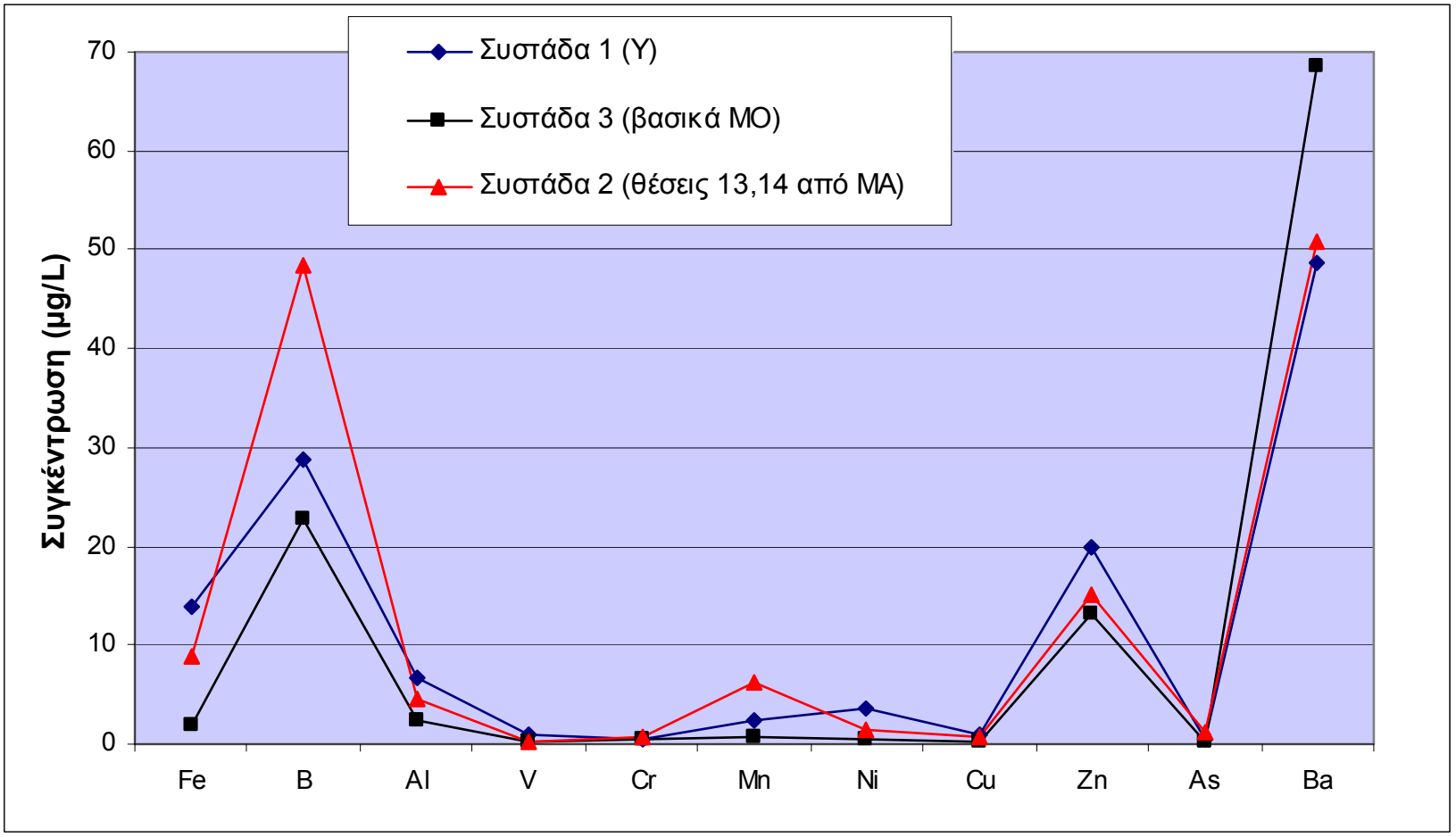
1.1. ΣΥΜΠΛΗΡΩΜΑ ΣΤΟ ΠΕΙΡΑΜΑΤΙΚΟ ΜΕΡΟΣ (ΚΛΑΣΙΚΕΣ ΤΕΧΝΙΚΕΣ)

Πίνακας 1.1: Συσχετίσεις Pearson για τα "ενοποιημένα" συνολικά δεδομένα

	Fe	**B**	**Al**	**V**	**Cr**	**Mn**	**Ni**	**Cu**	**Zn**	**As**	**Ba**
Fe	1,00	0,17	0,43	0,28	0,10	0,05	0,30	0,52	0,21	0,12	-0,12
B	0,17	1,00	-0,01	0,47	0,38	-0,02	0,42	0,40	0,19	**0,65**	0,21
Al	0,43	-0,01	1,00	0,19	0,28	-0,07	-0,02	0,10	0,12	-0,10	-0,17
V	0,28	0,47	0,19	1,00	0,54	-0,19	0,33	0,46	-0,01	0,21	-0,12
Cr	0,10	0,38	0,28	0,54	1,00	-0,24	-0,05	0,12	-0,02	0,15	-0,04
Mn	0,05	-0,02	-0,07	-0,19	-0,24	1,00	0,43	0,01	0,21	0,44	0,15
Ni	0,30	0,42	-0,02	0,33	-0,05	0,43	1,00	0,45	0,33	0,56	-0,03
Cu	0,52	0,40	0,10	0,46	0,12	0,01	0,45	1,00	0,01	0,32	-0,10
Zn	0,21	0,19	0,12	-0,01	-0,02	0,21	0,33	0,01	1,00	0,21	0,55
As	0,12	0,65	-0,10	0,21	0,15	0,44	0,56	0,32	0,21	1,00	0,17
Ba	-0,12	0,21	-0,17	-0,12	-0,04	0,15	-0,03	-0,10	0,55	0,17	1,00

Πίνακας 1.2: Συσχετίσεις Spearman για τα “ενοποιημένα” συνολικά δεδομένα

	Fe	**B**	**Al**	**V**	**Cr**	**Mn**	**Ni**	**Cu**	**Zn**	**As**	**Ba**
Fe	1,00	0,53	0,18	0,29	0,15	0,56	0,58	0,04	0,35	0,53	-0,09
B	0,53	1,00	-0,21	0,14	0,18	0,43	0,41	0,00	0,66	**0,63**	0,34
Al	0,18	-0,21	1,00	0,20	0,18	0,05	-0,04	0,08	-0,21	-0,14	-0,26
V	0,29	0,14	0,20	1,00	0,35	0,18	0,43	0,39	0,01	0,27	-0,16
Cr	0,15	0,18	0,18	0,35	1,00	0,19	0,06	0,07	-0,03	0,10	-0,12
Mn	0,56	0,43	0,05	0,18	0,19	1,00	0,50	0,11	0,27	0,54	-0,02
Ni	0,58	0,41	-0,04	0,43	0,06	0,50	1,00	0,40	0,23	**0,72**	-0,16
Cu	0,04	0,00	0,08	0,39	0,07	0,11	0,40	1,00	-0,29	0,30	-0,19
Zn	0,35	0,66	-0,21	0,01	-0,03	0,27	0,23	-0,29	1,00	0,35	0,52
As	0,53	**0,63**	-0,14	0,27	0,10	0,54	**0,72**	0,30	0,35	1,00	0,06
Ba	-0,09	0,34	-0,26	-0,16	-0,12	-0,02	-0,16	-0,19	0,52	0,06	1,00



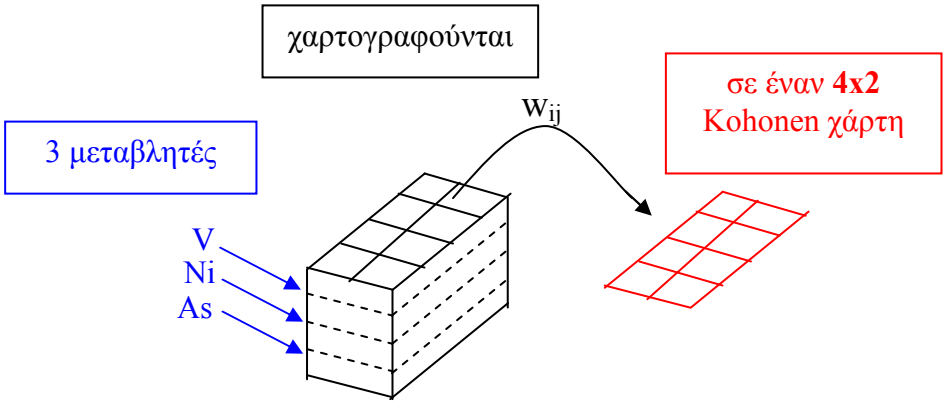
Σχήμα 1.1: Μεταβλητές που είναι υπεύθυνες για τον διαχωρισμό των ομάδων και υπο-όμάδων στην CA.

1.2. ΣΥΜΠΛΗΡΩΜΑ ΣΤΟ ΠΕΙΡΑΜΑΤΙΚΟ ΜΕΡΟΣ (ΔΙΚΤΥΟ KOHONEN)

1.2.1. Πίνακες και σχήματα

Πίνακας 1.3: Παράμετροι εκπαίδευσης δικτύων Kohonen

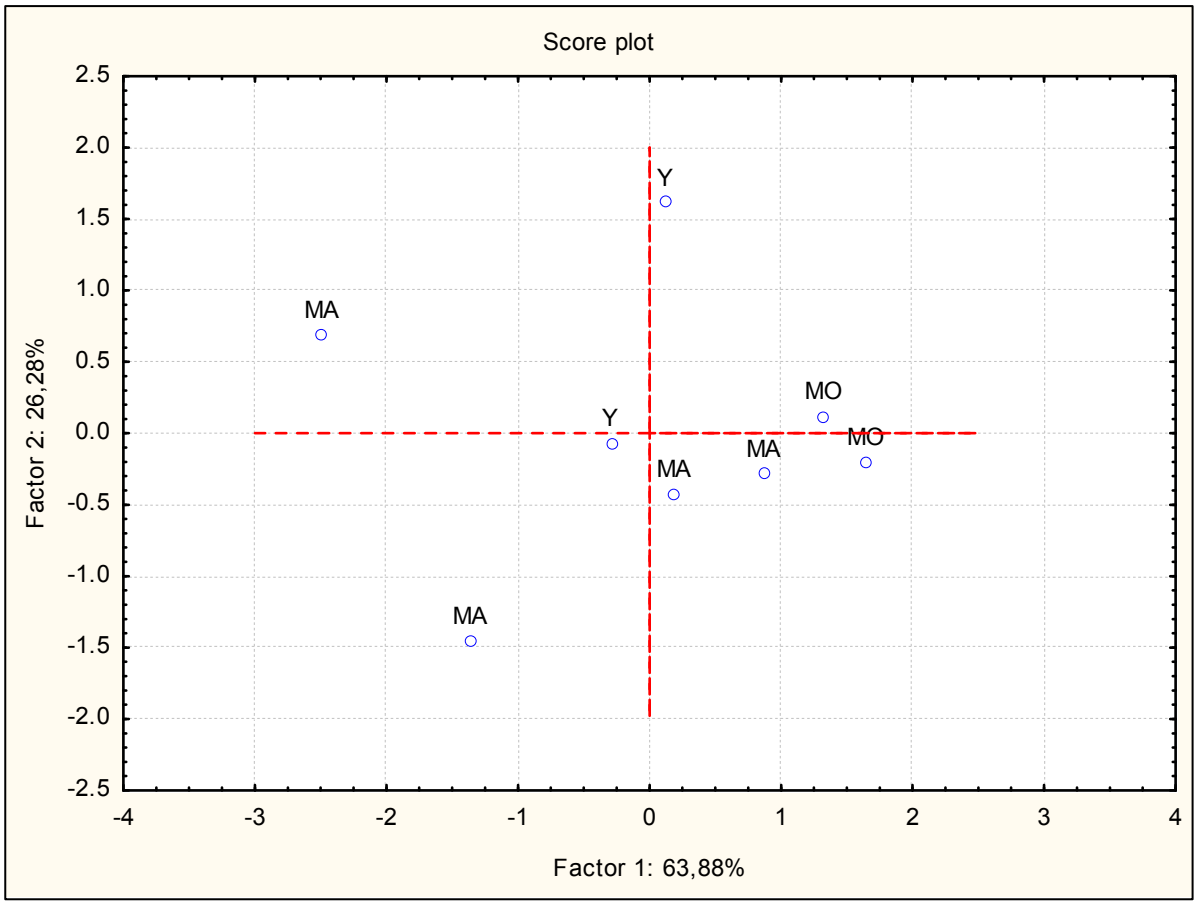
	Α Φάση	**Β Φάση**
Περίοδοι (Epochs)	20000	20000
Ρυθμός εκπαίδευσης (Learning rate)	0,1 – 0,02	0,1 – 0,01
Ακτίνα γειτνίασης R	3 - 1	0 - 0
Training / Selection / Test	45 / 22 / 22	45 / 22 / 22



Σχήμα 1.2: Σχηματική παράσταση της πορείας Kohonen (για την 4x2 δομή).

1.2.2. Αξιολόγηση βαρών του βελτιστοποιημένου μοντέλου Kohonen

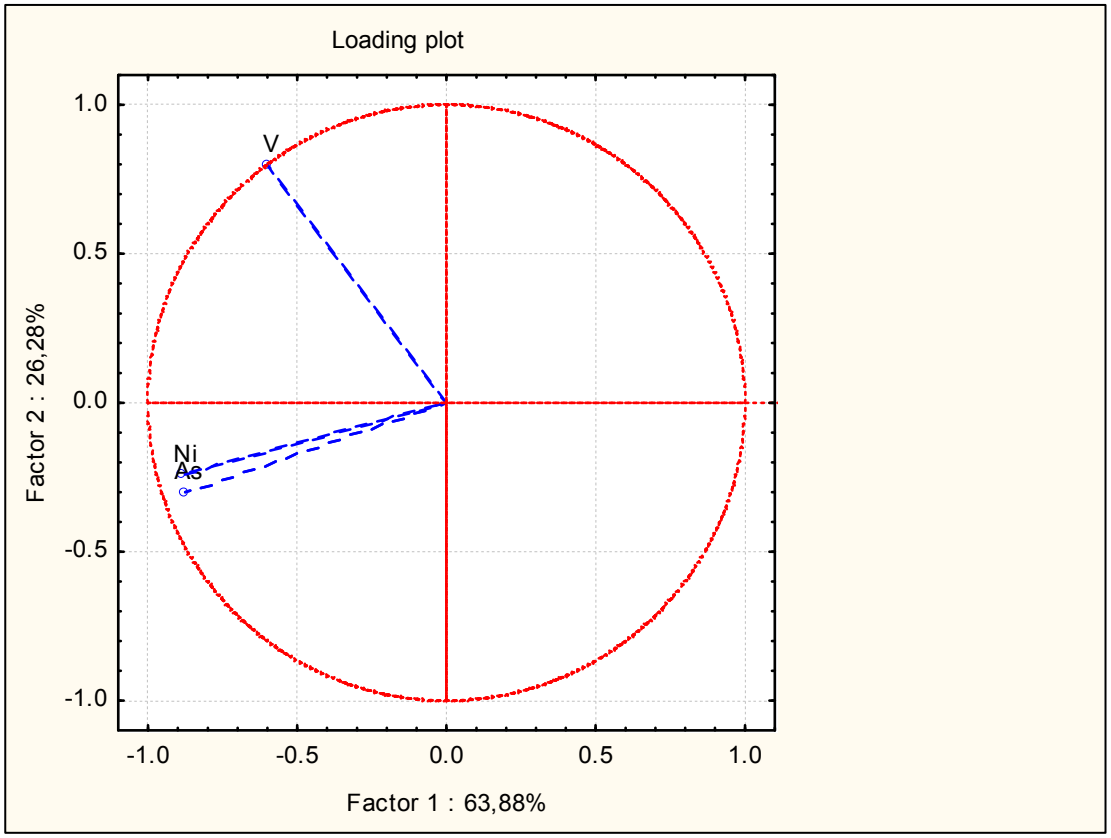
Η αξιολόγηση των βαρών του βελτιστοποιημένου Kohonen μοντέλου 3:3-8:1, μπορεί να γίνει με την βοήθεια των αναλύσεων PCA και DA των βαρών των 8 νευρώνων.



Σχήμα 1.3: Διάγραμμα σκορ της PCA στα βάρη του βελτιστοποιημένου μοντέλου Kohonen 3:3-8:1

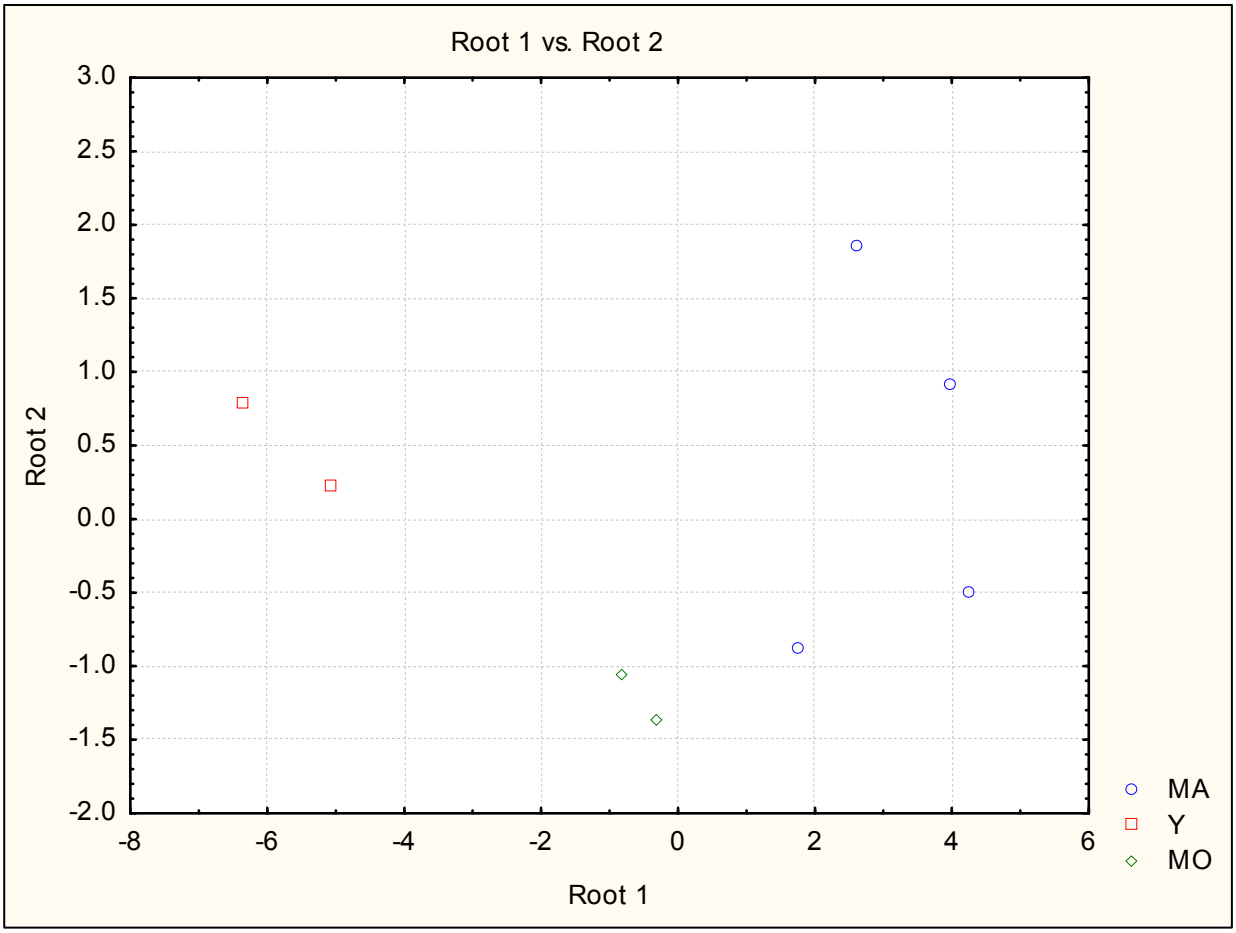
Έτσι με σκοπό την παραπέρα εξερεύνηση της σχέσης ανάμεσα στα βάρη των νευρώνων και τις ομάδες των δειγμάτων (Y, MO, MA), εφαρμόζεται αρχικά PCA σε αυτά. Στο διάγραμμα σκορ (σχ. 1.3) αναπαρίσταται οι 8 νευρώνες του μοντέλου Kohonen: καθώς οι νευρώνες χαρακτηρίζονται από τον αριθμό των δειγμάτων που έχουν "πέσει" σε αυτούς, το ίδιο μπορεί να γίνει και στα σημεία του σκορ διαγράμματος. Εδώ, οι νευρώνες των διαφορετικών ομάδων περιγράφονται πολύ καλά από τις δύο συνιστώσες PC1 και PC2 οποίες ερμηνεύουν συνολικό ποσοστό ίσο με 90, 2 %. Καθώς οι συνιστώσες αυτές, μπορούν να διαχωρίσουν τις τρεις ομάδες, είναι δυνατό να βρεθούν από το αντίστοιχο διάγραμμα φορτίσεων (σχ. 1.4) οι μεταβλητές που περιγράφουν αυτές τις ομάδες. Έτσι

είναι φανερό, ότι κάποιες από τις παραμέτρους δεν παίζουν σημαντικό ρόλο σε κάποιο διαχωρισμό: το V δεν μπορεί να διαχωρίσει τις ομάδες των ΜΟ και ΜΑ, αλλά αντίθετα το είναι σημαντικό για την διαφοροποίηση της Υλίκης (Υ). Τα As, Ni συμβάλλουν στην διαφοροποίηση ΜΟ και ΜΑ.



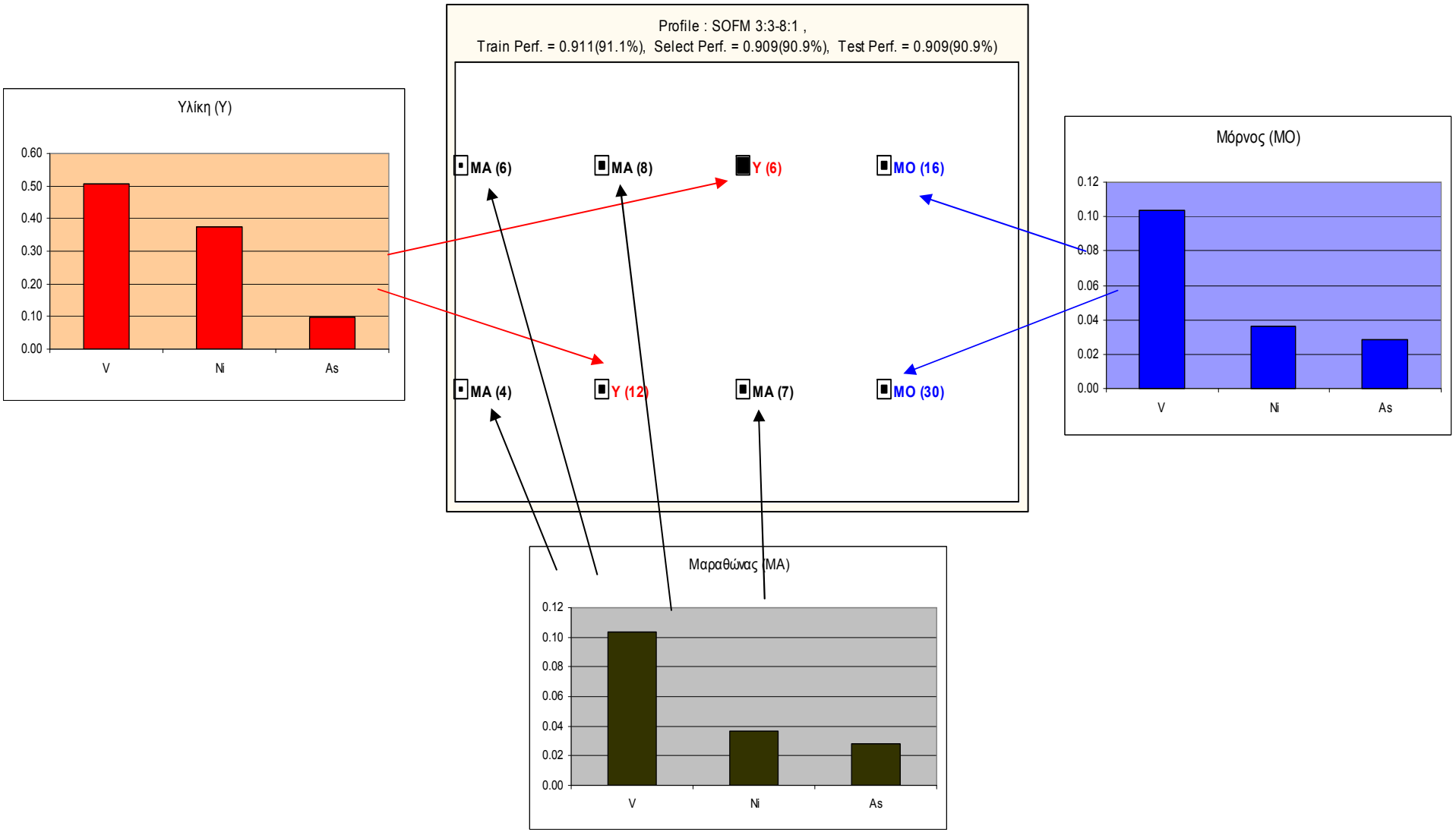
Σχήμα 1.4: Διάγραμμα φορτίσεων PCA στα βάρη του βελτιστοποιημένου μοντέλου Kohonen 3:3-8:1

Παρότι η PCA των βαρών μοντέλου Kohonen, έχει γίνει στο παρελθόν [1], η DA αυτών δεν έχει επαναληφθεί. Εφαρμόζεται λοιπόν DA στα βάρη των 8 νευρώνων και τα αποτελέσματα φαίνονται στο canonical plot (σχ. 1.5). Τα σημεία του διαγράμματος αναπαριστούν τους 8 νευρώνες. Ο διαχωρισμός τους είναι πολύ καλός με το As να κρίνεται ως σημαντικότερη μεταβλητή: As>Ni>V. Είναι φανερό ότι μέσω του μοντέλου Kohonen, ο διαχωρισμός των τριών λιμνών ανάγεται σε ένα γραμμικό πρόβλημα.



Σχήμα 1.5: Canonical plot από την κλασική DA στα βάρη του βελτιστοποιημένου μοντέλου Kohonen 3:3-8:1

Στο σχήμα 1.6 ανασκοπούνται όλα τα παραπάνω συμπεράσματα.



Σχήμα 1.6: Συνολική αξιολόγηση των βαρών του βελτιστοποιημένου μοντέλου Kohonen 3:3-8:1

ΚΕΦ. 2 ΠΕΡΙΒΑΛΛΟΝΤΙΚΗ ΕΠΙΔΡΑΣΗ ΤΗΣ ΕΝΤΑΤΙΚΗΣ ΙΧΘΥΟΚΑΛΛΙΕΡΓΕΙΑΣ (ΜΕΛΕΤΗ ΤΗΣ ΣΥΓΚΕΝΤΡΩΣΗΣ ΜΕΤΑΛΛΩΝ ΚΑΙ ΘΡΕΠΤΙΚΩΝ ΣΥΣΤΑΤΙΚΩΝ ΣΕ ΘΑΛΑΣΣΙΑ ΙΖΗΜΑΤΑ ΣΤΗΝ ΕΛΛΑΔΑ)

2.1. ΕΙΣΑΓΩΓΗ

Στη διάρκεια των τελευταίων χρόνων, εκτεταμένες ιχθυοκαλλιέργειες έχουν αναπτυχθεί κατά μήκος των ηπειρωτικών και νησιωτικών ελληνικών ακτών. Αυτή η ανάπτυξη απαιτεί περαιτέρω μελέτη όχι τόσο στην υδάτινη στήλη, αλλά στα θαλάσσια ιζήματα γύρω από τους κλωβούς των ψαριών. Πραγματικά, οι ιχθυοκαλλιέργειες φαίνεται να επηρεάζουν το προφίλ των συγκεντρώσεων της οργανικής ύλης, των θρεπτικών συστατικών [2, 3] αλλά και των μετάλλων [4, 5, 6]. Εργασίες συντήρησης, ιχθυοτροφές, μεταβολικά προϊόντα και περιττώματα συνεισφέρουν επίσης στην επιμόλυνση της περιοχής.

Η αλλαγή της σύστασης των θαλάσσιων ιζημάτων που συνορεύουν με ιχθυοκαλλιέργειες έχει μελετηθεί πολύ στην διάρκεια των δύο τελευταίων δεκαετιών. Ο Chou et al. [7] διαπίστωσε σημαντική μείωση στην επίδραση των μονάδων ιχθυοκαλλιέργειας στα ιζήματα απόστασης 50 m από τους κλωβούς των ψαριών στην New Brunswick (Καναδάς). Βρέθηκαν επίσης υψηλές συγκεντρώσεις Cu, Zn, C και σωματιδίων <63 μm. PCA και CA χρησιμοποιήθηκαν για την αξιολόγηση των αποτελεσμάτων. Ο Mendiguchía et al. [8] επιβεβαίωσε την συγκέντρωση μετάλλων όπως Cu, Zn και Pb και οργανικής ύλης σε θαλάσσια ιζήματα σαν συνέπεια της εντατικής ιχθυοκαλλιέργειας. Επιπλέον, η συγκέντρωση του P βρέθηκε να μειώνεται δραστικά με την αύξηση της απόστασης από ιχθυοκαλλιέργειες στο Σούνιο (Ελλάδα) στην Alicante (Ισπανία) και την Sicily (Ιταλία) [9]. Ο Dalman et al. [10] έχει επίσης προσδιορίσει υψηλές συγκεντρώσεις Zn, Cd και Cu εξαιτίας παρακείμενων μονάδων ιχθυοκαλλιέργειας στην Gulluk Bay (Τουρκία). Ο Tovar et al. [4] αξιολόγησε την ποιότητα του νερού σε ένα ποτάμι στην Cadiz Bay (Ισπανία), όπου λειτουργούν εκτεταμένες μονάδες ιχθυοκαλλιέργειας. Πολλές παράμετροι προσδιορίστηκαν (pH, θερμοκρασία, αλατότητα, διαλυμένο οξυγόνο, θρεπτικά συστατικά) κατά μήκος των ακτών του ποταμού στην διάρκεια διαφορετικών εποχών του έτους. Το αμμώνιο και τα αιωρούμενα στερεά ήταν οι πιο σημαντικοί ρυπαντές. Ο Barasan et al. [11] επιβεβαίωσε την επίδραση των ιχθυοκαλλιεργειών στα θαλάσσια ιζήματα (η οποία ήταν πολύ μικρότερη στην υδάτινη στήλη). Βρέθηκαν σημαντικές διαφορές ανάμεσα στα “τυφλά”

σημεία ελέγχου και τα παρακείμενα των κλωβών σημεία για παραμέτρους όπως η οργανική ύλη και βαρέα μέταλλα (Fe και Zn). Οι συγγραφείς χρησιμοποίησαν τις δοκιμές Kruskal–Wallis και U Mann–Whitney για την ανίχνευση διαφορών ανάμεσα στις θέσεις δειγματοληψίας και τις εποχές. Ο Neofitou et al. [12] μελέτησε επίσης τις χωρικές και εποχιακές διακυμάνσεις στην επίδραση των μονάδων ιχθυοκαλλιέργειας στα θαλάσσια ιζήματα και τους οργανισμούς σε δύο μονάδες στον Παγασητικό κόλπο. Από τους συγγραφείς χρησιμοποιήθηκε ANOVA για την ταυτοποίηση των χωρικών διακυμάνσεων, ενώ δεν ανιχνεύτηκαν εποχιακές διαφορές. Τέλος, οι Mente et al. [3], Cao et al. [13] και Sapkota et al. [14] επιβεβαίωσαν την επίδραση των μονάδων ιχθυοκαλλιέργειας σε πρόσφατες εργασίες ανασκόπησης.

Έτσι είναι φανερό, ότι είναι σημαντική η εκτίμηση της τύχης των αποβλήτων των μονάδων ιχθυοκαλλιέργειας για την βιοσιμώτητα της βιομηχανίας αυτής. Συγκεκριμένα, ο σκοπός του κεφαλαίου αυτού, είναι η μελέτη της συσσώρευσης οργανικών C, N και P και οκτώ μετάλλων (Cu, Cd, Pb, Hg, Fe, Mn, Ni και Zn) και As σε θαλάσσια ιζήματα σε μονάδες ιχθυοκαλλιέργειας κατά μήκος τριών παράκτιων περιοχών της Ελλάδας: Χίος-Οινούσσες στο Βόρειο Αιγαίο και Ναύπακτος, Αστακός στην κεντρική Ελλάδα. Παράλληλα μελετάται η ιδιαιτερότητα των μονάδων αυτών και τυχόν εποχιακές διακυμάνσεις. Έτσι, συλλέγονται θαλάσσια ιζήματα κοντά (μηδενική απόσταση) ή μακρύτερα (50 και 100 m) από τους κλωβούς των ψαριών με πέντε βασικούς στόχους:

1. να εκτιμηθούν οι συγκεντρώσεις των μετάλλων/μεταλλοειδών και θρεπτικών συστατικών σε θαλάσσια ιζήματα παρακείμενα σε ιχθυοκαλλιέργειες,
2. να συγκριθούν τα αποτελέσματα αυτά με τα αντίστοιχα ιζημάτων μακρύτερα των κλωβών ιχθυοκαλλιέργειας, ώστε να παρουσιαστούν οι επιπτώσεις των εγκαταστάσεων αυτών στην ποιότητα των παρακειμένων θαλασσίων ιζημάτων,
3. να μελετηθούν διαφορές μεταξύ των μονάδων που σχετίζονται με την εποχή δειγματοληψίας,
4. να εφαρμοστούν βασικές πολυπαραμετρικές τεχνικές για την αξιολόγηση των αποτελεσμάτων,
5. να ανιχνευτούν οι πιο σημαντικοί από τους παράγοντες επιμόλυνσης στο σύνολο των μετάλλων και θρεπτικών συστατικών που προσδιορίζονται.

Μη επιβλεπόμενες και επιβλεπόμενες τεχνικές όπως PCA/FA και DA αντίστοιχα χρησιμοποιήθηκαν για την εκτίμηση των αποτελεσμάτων. Οι PCA/FA εφαρμόστηκαν για την ταυτοποίηση των συσχετίσεων μεταξύ των μεταβλητών και την εύρεση των παραγόντων που επηρεάζουν την ρύπανση στην περιοχή ή καθορίζουν την διαφοροποίηση

των τριών μονάδων. Η DA επιβεβαίωσε την διαφοροποίηση των σημείων δειγματοληψίας εξαιτίας της απόστασής τους από τους κλωβούς ιχθυοκαλλιέργειας, τα ιδιαίτερα χαρακτηριστικά της κάθε μονάδας ή το διαφορετικό γεωλογικό υπόβαθρο. Είναι η πρώτη φορά που αποτελέσματα από τόσες χημειομετρικές τεχνικές χρησιμοποιούνται σαν μια ισχυρή ένδειξη της συσσώρευσης μετάλλων και θρεπτικών συστατικών στα παρακείμενα των κλωβών ιζήματα.

2.2. ΠΕΙΡΑΜΑΤΙΚΟ ΜΕΡΟΣ

2.2.1. Δειγματοληψία

Επιλέχθησαν τρεις μονάδες: Χίος-Οινούσσες (CH), Ναύπακτος (NA) και Αστακός (AS) (σχ. 2.1, πίνακας 2.1) οι οποίες χρησιμοποιούνται για την εκτροφή τσιπούρας και λαυρακιού. Όλα τα βασικά χαρακτηριστικά των μονάδων περιγράφονται στον πίνακα 2.1. Παράγοντες όπως τα χρόνια λειτουργίας, το βάθος, η κατεύθυνση των ανέμων, ο τύπος του ιζήματος και η δυναμικότητα της μονάδας ενδέχεται να έχουν επίδραση στην εκτίμηση των αποτελεσμάτων.

Οι θέσεις δειγματοληψίας επιλέχθησαν με βάση σχέδια προτεινόμενα από άλλες μελέτες για την περιβαλλοντική επίδραση των ιχθυοκαλλιεργειών στα ιζήματα [15, 16, 17, 18, 19], όπου ένα δείγμα συλλέγεται κάτω από τον κλωβό, ένα δεύτερο σε απόσταση από αυτόν ώστε αν αξιολογηθεί η έκταση της ρύπανσης και ένα τρίτο σε σημείο τελείως ανεπηρέαστο από την δραστηριότητα. Το τελευταίο αναφέρεται σαν δείγμα ελέγχου ή αναφοράς (“control” ή “reference”).

Ενενήντα δείγματα (90) συλλέχθησαν συνολικά (56 CH, 16 NA και 18 AS), κάτω από τον κλωβό των ψαριών (μηδενική απόσταση) και 50 ή 100 m μακριά από αυτά, κατά μήκος των κυριότερων θαλάσσιων ρευμάτων. Δεκαέξι (16) δείγματα από αυτά ήταν “δείγματα-μάρτυρες” ή ελέγχου και βρίσκονταν μακριά από τις ιχθυοκαλλιέργειες (πίνακες 2.2-2.4). Η δειγματοληψία έγινε σε τέσσερις περιόδους: τον Δεκέμβριο του 2005, τον Ιούνιο και Δεκέμβριο του 2006 και τον Δεκέμβριο του 2007. Τα βάθη δειγματοληψίας κυμαίνονταν από 15 ως 42 m. Όλα τα δείγματα συλλέχθησαν από δύτες σε πλαστικά δοχεία 1 L και φυλλάχθησαν στους -20°C μέχρι την ανάλυση.

Σχήμα 2.1: Μονάδες ιχθυοκαλλιέργειας που χρησιμοποιήθηκαν στον προσδιορισμό των μετάλλων/μεταλλοειδών και θρεπτικών συστατικών σε θαλάσσια ιζήματα (1: ΝΑ, Ναύπακτος, 2: CH, Χίος-Οινούσσες, 3: AS, Αστακός

Πίνακας 2.1: Περιγραφή των θέσεων δειγματοληψίας των θαλασσίων ιζημάτων

Κωδικός δειγματοληψίας*	**Μονάδα ιχθυοκαλλιέργειας**	**Φάρμα**	**Χαρακτηριστικά φάρμας**				
			Έτη λειτουργίας	**Βάθος (m)**	**Κατεύθυνση ανέμων (μέγιστη ταχύτητα)**	**Τύπος ιζήματος**	**Δυναμικότητα (t/έτος)**
108.1.S	N-E των Οινουσσών	Γαβάθι	9	26-38	S	αμμώδες	230
108.2.S	Χίος	Αιγέας	12	12-40	S-E	αμμώδες	230
108.3.S	N-E of Inousses	Ιχθύς	11	17-28	E-W	αμμώδες	230
108.5.S	Χίος	Πρωτέας	14	30-50	S-E	αμμώδες	290
108.6.S	N-E της Χίου	Ροδόσσι	9	20-50	N & S-E	αμμώδες-βραχώδες	190
108.7.S	N-E των Οινουσσών	Οινούσσες	12	16-27	---	αμμώδες	230
CH.6.M	Χίος	Δείγμα ελέγχου	---	15-20	N-SE	αμμώδες	---
CH.IN.M	Οινούσσες	Δείγμα ελέγχου	---	15-20	N-SE	αμμώδες	---
107.1.S	Ναύπακτος	Εκάλ	19	19-38	S-SW	αμμώδες	190
107.2.S	Ναύπακτος	Υδροκάλ	12	20-70	E-SE	αμμώδες-χαλίκι	280
107.3.S	Ναύπακτος	Αγ. Νικόλαος	16	19-35	S-SW	χαλίκι	160
NA.M1	Ναύπακτος	Δείγμα ελέγχου Εκάλ-Υδροκάλ	---	8	E	βραχώδες-χαλίκι	---
NA.M3	Ναύπακτος	Δείγμα ελέγχου Αγ. Νικολάου	---	6	W	βραχώδες-χαλίκι	---
109.1.S	Αστακός	Προβάτι	6	42-49	N-NE & SE (10.4)	λασπώδες	700

Κωδικός δειγματοληψίας*	Μονάδα ιχθυοκαλλιέργειας	Φάρμα	Χαρακτηριστικά φάρμας				
			Έτη λειτουργίας	Βάθος (m)	Κατεύθυνση ανέμων (μέγιστη ταχύτητα)	Τύπος ιζήματος	Δυναμικότητα (t/έτος)
109.2.S	Αστακός	Ποντικός	13	33-42	N (6.2)	λασπώδες	700
109.3.S	Αστακός	Παλαιά Δραγονέρα	11	25-47	NW	λασπώδες	150
109.4.S	Αστακός	Νέα Δραγονέρα	3	45-62	N-NE & SW	λασπώδες-αμμώδες	390
AS.M1	Αστακός	Δείγμα ελέγχου Δραγονέρας	---	20-25	NW	αμμώδες	---
AS.M2	Αστακός	Δείγμα ελέγχου Προβάτι-Ποντικός	---	20-25	N	αμμώδες	---

* Με την προσθήκη της κατάλληλης τιμής: 0, 50,100 σε αυτόν τον κωδικό, καθορίζεται η απόσταση από τον κλωβό (βλ. παρακάτω πίνακες 2.2, 2.3, 2.4).

Πίνακας 2.2: Αποτελέσματα Cu, Cd, Pb, Hg, As, Fe, Mn, Zn, Ni, C, N & P για την μονάδα Χίου-Οινουσσών (CH)

Δείγμα / απόσταση από τον κλωβό (m)	**Κωδικός**	**Cu**	**Cd**	**Pb**	**Hg**	**As**	**Fe**	**Mn**	**Zn**	**Ni**	**C**	**N**	**P**
1η δειγματοληψία: 12/2005													
108.1.S 100	DI	13,4	90,0	0,52	0,031	6,30	3177	67,7	31,5	17,1	15,0	0,66	0,45
108.1.S 0	ZE	47,9	253	0,73	0,047	6,74	6617	74,8	136	18,4	29,8	3,64	6,41
108.7.S 100	DI	17,9	128	0,75	0,044	7,46	7418	97,8	45,4	35,1	**29,6**	**2,59**	0,66
108.7.S 0	ZE	83,1	311	0,77	0,054	7,82	8020	101	256	29,3	**36,2**	**3,87**	19,5
108.3.S 100	DI	10,1	101	0,79	0,042	7,61	6722	110	30,2	26,3	16,2	1,33	0,32
108.3.S 0	ZE	20,8	218	0,69	0,056	14,2	7917	117	141	42,4	36,0	4,18	6,26
108.6.S 100	DI	22,0	204	0,95	0,079	16,1	8096	117	47,8	28,3	22,6	2,24	0,94
108.6.S 0	ZE	72,5	619	0,74	0,339	9,52	8577	140	448	27,8	65,6	5,11	17,9
108.2.S 100	DI	3,14	109	0,23	0,029	5,77	117	29,4	11,6	1,83	7,38	0,82	0,45
108.2.S 0	ZE	68,1	305	0,34	0,038	7,39	558	45,5	114	10,9	15,3	2,47	9,69
108.6.S 0	ZE	20,9	243	0,45	0,131	5,16	2678	96,2	77,5	8,33	16,2	1,45	6,65
108.6.S 100	DI	14,5	373	0,40	0,089	5,50	2902	133	104	7,77	11,7	1,04	1,05
CH.IN.M	Δείγμα ελέγχου	8,59	140	0,71	0,053	8,00	6341	93,0	50,3	20,8	21,2	1,87	0,33
CH.6.M	Δείγμα ελέγχου	11,4	117	0,29	0,065	7,97	3956	112	25,2	13,9	18,5	1,52	0,43
2η δειγματοληψία: 6/2006													
108.2.S 0	ZE	23,6	245	0,27	2,44	12,8	4167	89,3	95,7	13,6	17,2	1,60	3,85
108.2.S 100	DI	6,69	109	0,15	0,247	7,97	872	42,8	11,2	5,21	3,93	0,360	0,17

Δείγμα / απόσταση από τον κλωβό (m)	Κωδικός	Cu	Cd	Pb	Hg	As	Fe	Mn	Zn	Ni	C	N	P
108.5.S 0	ZE	33,6	168	0,24	**66,6**	14,4	7977	109	116	108	**20,4**	**1,62**	2,85
108.5.S 100	DI	27,3	174	1,20	3,67	21,2	8895	131	65,9	88,9	**22,4**	**1,72**	0,83
108.6.S 0	ZE	39,9	223	0,36	4,52	7,68	4976	129	74,5	73,5	19,9	1,63	3,47
108.6.S 100	DI	16,6	103	0,71	3,78	6,54	5090	94,2	18,2	65,3	21,6	1,87	0,53
108.1.S 0	ZE	106	480	0,29	0,041	7,10	3540	60,0	270	9,64	20,4	2,55	8,28
108.1.S 100	DI	25,9	89,9	0,53	0,029	6,60	3820	81,1	49,9	6,61	7,18	0,75	0,77
108.3.S 0	ZE	113	772	0,70	0,092	7,30	7048	101	702	16,8	36,5	4,79	13,9
108.3.S 100	DI	19,2	86,8	1,02	0,050	9,04	6585	118	27,2	24,9	13,5	0,99	0,33
108.7.S 0	ZE	119	365	0,64	0,236	10,9	7591	95,2	237	52,9	**26,8**	**3,65**	10,8
108.7.S 100	DI	76,2	343	2,57	0,184	9,40	8127	102	154	59,6	**30,8**	**1,07**	3,73
CH.6.M	Δείγμα ελέγχου	9,77	88,4	0,19	1,10	9,37	1758	59,7	5,10	39,5	7,33	0,57	0,22
CH.IN.M	Δείγμα ελέγχου	7,90	48,2	0,85	0,099	9,01	10076	76,3	19,2	11,5	3,61	0,39	0,13
3η δειγματοληψία: 12/2006													
108.2.S 0	ZE	21,3	309	1,37	0,067	12,8	4167	89,3	92,1	13,6	31,4	2,06	6,21
108.2.S 100	DI	28,4	176	0,013	0,069	7,97	872	42,8	55,7	5,21	4,60	1,28	0,91
108.5.S 0	ZE	98,7	634	1,17	0,090	14,4	7977	109	532	108	**28,8**	**3,68**	5,02
108.5.S 100	DI	69,4	97	4,44	0,074	21,2	8895	131	61,4	88,9	**4,60**	**1,91**	1,52
108.6.S 0	ZE	24,6	313	0,59	0,044	7,68	4976	129	276	73,5	16,9	2,21	3,72
108.6.S 100	DI	19,6	152	0,03	0,022	6,54	5090	94,2	30,5	65,3	2,00	1,04	0,91
108.1.S 0	ZE	64,5	769	0,33	0,022	30,1	6090	299	211	38,0	20,3	3,03	7,22

Δείγμα / απόσταση από τον κλωβό (m)	Κωδικός	Cu	Cd	Pb	Hg	As	Fe	Mn	Zn	Ni	C	N	P
108.1.S 100	DI	18,3	113	0,53	0,020	3,68	20505	374	35,6	110	9,20	1,08	1,06
108.3.S 0	ZE	269	864	2,83	0,067	21,0	8139	285	1007	50,5	35,2	5,58	5,26
108.3 S 100	DI	21,5	60	1,67	0,035	11,2	2350	185	27,7	22,7	13,8	1,02	0,32
108.7.S 0	ZE	266	**1296**	1,82	0,052	6,72	1786	181	434	17,8	**26,8**	**6,61**	13,7
108.7.S 100	DI	39,2	154	2,20	0,044	9,27	2814	225	59,1	24,3	**31,4**	**2,21**	1,03
CH.6.M	Δείγμα ελέγχου	23,4	581	0,10	0,047	9,37	1758	59,7	78,2	39,5	9,80	1,57	4,91
CH.IN.M	Δείγμα ελέγχου	11,4	170	3,06	0,089	7,06	2730	206	32,1	40,3	3,90	0,51	0,22
4η δειγματοληψία: 12/2007													
108.1.S 0	ZE	800	286	0,10	0,040	4,60	2679	129	161	19,0	14,6	1,56	5,57
108.1.S 50	DI	44,0	219	0,27	0,020	10,1	7454	202	172	24,0	10,7	1,09	2,72
108.3.S 0	ZE	92,0	457	2,91	0,060	5,42	4087	199	755	33,3	22,0	1,70	4,58
108.3.S 50	DI	19,0	38,2	0,56	0,070	4,99	6092	198	22,3	31,4	13,4	0,45	0,48
108.7.S 50	DI	26,0	64,9	1,92	0,10	5,48	8375	239	36,9	50,1	**23,0**	**1,02**	0,86
108.7.S 0	ZE	766	494	1,13	0,79	6,93	6789	195	2790	38,8	**34,5**	**4,80**	6,31
108.2.S 0	ZE	53,2	74,2	0,41	0,10	8,25	4886	256	54,2	37,0	15,1	0,67	1,73
108.2.S 50	DI	111	23,8	0,040	0,050	3,95	829	60,7	3,74	10,8	25,2	0,60	0,48
108.6.S 0	ZE	38,4	135	0,23	0,31	6,51	3967	269	86,3	21,8	15,7	0,31	3,22
108.6.S 50	DI	14,5	48,8	0,10	0,080	6,93	5288	245	11,1	17,7	13,1	0,51	0,37
108.5.S 0	ZE	39,8	57,8	0,67	0,060	7,17	3782	139	48,4	21,9	**16,6**	**0,47**	1,72
108.5.S 50	DI	10,3	32,5	0,27	0,14	5,81	2704	145	9,35	15,91	**8,29**	**0,30**	0,25

Δείγμα / απόσταση από τον κλωβό (m)	**Κωδικός**	**Cu**	**Cd**	**Pb**	**Hg**	**As**	**Fe**	**Mn**	**Zn**	**Ni**	**C**	**N**	**P**
CH.IN.M	Δείγμα ελέγχου	27,5	6,23	1,93	0,090	10,9	14216	273	26,1	26,3	5,41	0,51	0,12
CH.6.M	Δείγμα ελέγχου	6,03	35,7	0,35	0,10	5,32	4847	356	9,27	13,4	12,0	0,62	0,24

Τα αποτελέσματα στα μέταλλα/μεταλλοειδή αναφέρονται σε mg/kg, εκτός του Cd που αναγράφεται σε μg/kg. Τα θρεπτικά συστατικά C, N, P δίνονται σε mg/g.

Πίνακας 2.3: Αποτελέσματα Cu, Cd, Pb, Hg, As, Fe, Mn, Zn, Ni, C, N & P για την μονάδα Ναυπάκτου (NA)

Δείγμα / απόσταση από τον κλωβό (m)	**Κωδικός**	**Cu**	**Cd**	**Pb**	**Hg**	**As**	**Fe**	**Mn**	**Zn**	**Ni**	**C**	**N**	**P**
2η δειγματοληψία: 6/2006													
107.1.S 0	ZE	46,7	153	1,91	0,066	13,0	21879	316	66,7	97,0	**12,1**	**0,53**	**0,54**
107.1.S 100	DI	49,7	159	6,45	0,049	12,9	30524	281	72,4	119	**22,5**	**0,59**	**0,33**
107.2.S 0	ZE	33,9	291	0,287	0,034	7,76	3504	186	86,7	25,9	13,4	1,18	3,33
107.2 S 100	DI	19,3	183	0,205	0,031	9,65	3356	215	25,7	22,4	6,85	0,59	0,61
107.3.S 0	ZE	250,4	217	0,263	0,036	10,9	3992	272	26,3	241	7,78	1,13	1,65
107.3.S 100	DI	27,3	203	0,740	0,017	10,0	10260	238	38,4	86,5	8,78	0,53	0,41
NA.M1	Δείγμα ελέγχου	18,0	166	0,479	0,024	4,58	3909	254	21,1	42,9	6,71	0,68	0,24
NA.M3	Δείγμα ελέγχου	27,1	173	0,436	0,020	5,44	8306	306	31,8	60,9	5,86	0,57	0,28
4η δειγματοληψία: 12/2007													
107.1.S 0	ZE	54,9	50,1	13,0	0,087	10,2	41515	916	74,9	182	**9,24**	**0,34**	**0,37**
107.1.S 50	DI	49,2	145	9,55	0,085	6,99	31787	1417	105	173	**10,2**	**0,28**	**1,48**
107.2.S 0	ZE	15,6	186	0,04	0,038	8,26	2794	496	53,0	35,6	6,55	0,89	0,99
107.2.S 50	DI	16,1	147	0,16	0,049	8,74	3784	415	14,9	38,9	7,44	0,76	0,99
107.3.S 0	ZE	104,8	382	0,04	0,013	8,65	5138	483	136	68,3	14,4	2,20	6,20
107.3.S5 0	DI	20,5	209	0,83	0,031	10,8	10661	643	43,0	90,0	7,46	0,14	0,74
NA.M1	Δείγμα ελέγχου	12,5	51,6	0,10	0,060	6,48	3284	351	3,69	41,8	4,01	0,66	0,12
NA.M3	Δείγμα	11,2	84,8	0,04	0,014	8,35	2534	502	12,6	28,9	2,87	0,68	0,12

Δείγμα / απόσταση από τον κλωβό (m)	Κωδικός	Cu	Cd	Pb	Hg	As	Fe	Mn	Zn	Ni	C	N	P
	ελέγχου												

Τα αποτελέσματα στα μέταλλα/μεταλλοειδή αναφέρονται σε mg/kg, εκτός του Cd που αναγράφεται σε μg/kg. Τα θρεπτικά συστατικά C, N, P δίνονται σε mg/g.

Πίνακας 2.4: Αποτελέσματα Cu, Cd, Pb, Hg, As, Fe, Mn, Zn, Ni, C, N & P για την μονάδα Αστακού (AS)

Δείγμα / απόσταση από τον κλωβό (m)	Κωδικός	Cu	Cd	Pb	Hg	As	Fe	Mn	Zn	Ni	C	N	P
2η δειγματοληψία: 6/2006													
109.1.S 0	ZE	44,2	16533	51,8	0,100	8,00	25386	445	130	129	11,6	1,05	0,940
109.1 S 100	DI	16,8	129	0,497	0,058	30,1	6090	299	53,5	38,0	4,67	0,430	0,320
109.2.S 0	ZE	113	6873	24,5	0,081	3,68	20505	374	596	110	50,1	5,87	12,9
109.2.S 100	DI	24,1	125	0,576	0,051	21,1	8139	285	27,1	50,5	5,17	0,460	0,370
109.3.S 0	ZE	145	1096	0,362	0,026	11,2	2350	185	955	22,7	**18,6**	3,13	20,0
109.3.S 100	DI	27,8	164	0,141	0,013	6,72	1786	181	37,7	17,8	**7,34**	0,790	1,56
109.4.S 100	DI	16,3	140	0,347	0,024	9,27	2814	225	27,8	24,3	5,34	0,560	0,310
AS.M1	Δείγμα ελέγχου	15,0	131	0,319	0,094	7,06	2730	206	16,6	40,3	8,58	0,820	0,270
AS.M2	Δείγμα ελέγχου	19,3	153	0,788	0,101	18,7	6678	318	39,1	57,7	8,10	0,660	0,230
4η δειγματοληψία: 12/2007													
109.1.S 0	ZE	39,4	95,8	9,69	0,056	9,85	33093	977	63,8	149	14,4	0,835	0,370

Δείγμα / απόσταση από τον κλωβό (m)	**Κωδικός**	**Cu**	**Cd**	**Pb**	**Hg**	**As**	**Fe**	**Mn**	**Zn**	**Ni**	**C**	**N**	**P**
109.1 S 50	DI	13,2	41,6	0,13	0,017	25,1	9607	805	15,0	56,6	3,99	0,516	1,483
109.2 S 0	ZE	187	1068	9,87	0,080	5,81	26299	1077	677	155	45,2	4,978	5,205
109.2 S 50	DI	46,8	121	9,13	0,060	9,51	35884	1434	80,8	187	7,62	0,548	0,492
109.3 S 0	ZE	1081	546	0,15	0,067	6,56	1762	302	709	25,4	**18,9**	1,28	8,99
109.3 S 50	DI	47,6	109	0,040	0,006	4,20	2484	320	16,4	27,3	**12,3**	0,372	0,367
109.4 S 50	DI	24,2	50,0	0,040	0,027	6,00	14202	358	11,2	17,3	4,53	0,185	0,367
AS.M1	Δείγμα ελέγχου	10,0	27,4	0,040	0,017	5,37	974	288	9,25	12,3	2,85	0,441	0,122
AS.M2	Δείγμα ελέγχου	22,3	84,5	2,20	0,086	7,82	11845	851	35,3	84,6	11,2	0,546	0,490

Τα αποτελέσματα στα μέταλλα/μεταλλοειδή αναφέρονται σε mg/kg, εκτός του Cd που αναγράφεται σε μg/kg. Τα θρεπτικά συστατικά C, N, P δίνονται σε mg/g.

2.2.2. Μέθοδοι και οργανολογία

Μετά τη συλλογή τους, τα δείγματα των θαλασσίων ιζημάτων καταψύχτηκαν. Πριν την ανάλυση κοσκινίστηκαν και το κλάσμα των 0,63 mm χρησιμοποιήθηκε για τον προσδιορισμό των 12 μετάλλων/μεταλλοειδών και ανόργανων στοιχείων. Ειδικότερα, ο οργανικός C (TOC) προσδιορίστηκε με την ογκομετρική μέθοδο Walkey και Black [20], ενώ τα N, P προσδιορίστηκαν συγχρόνως με αυτόματη peroxodisulfate μέθοδο οξείδωσης. Για τον προσδιορισμό των μετάλλων και του As, έγινε χώνευση των δειγμάτων με HF και βασιλικό νερό. Για τα Fe και Zn χρησιμοποιήθηκε φασματόμετρο Ατομικής Απορρόφησης Perkin Elmer 2380 (τεχνική φλόγας), ενώ τα Cu, Cd, Pb, Mn, As, και Ni προσδιορίστηκαν με την τεχνική φούρνου γραφίτη Zeeman THGA σε φασματόμετρο Ατομικής Απορρόφησης Perkin Elmer SIMAA 6000. Ο Hg προσδιορίστηκε με την cold-vapor τεχνική (CVAAS).

Οι αναλύσεις πραγματοποιήθηκαν από τους Ν. Θωμαΐδη, Ι. Πασιά, Κ. Μπαρκονίκο, Α. Καστρίτη και Π. Νησιανάκη στα Εργαστήρια Αναλυτικής Χημείας και Χημείας Περιβάλλοντος του Ε.Κ.Π.Α. Ο ποιοτικός έλεγχος των αποτελεσμάτων διασφαλίστηκε μέσω τυφλών δειγμάτων, διπλών προσδιορισμών και διεργαστηριακών δειγμάτων, (river sediment/Quality Consult, Italy, πίνακας 2.5). Όλες οι μετρήσεις πραγματοποιήθηκαν εις τριπλούν, ενώ χρησιμοποιήθηκε ο μέσος όρων για την εκτίμηση των αποτελεσμάτων.

Πίνακας 2.5: Ποιοτικός έλεγχος (υλικό: river sediment)

Στοιχείο	**Συγκέντρωση(mg/Kg)**	**z-score**
Cr	96,0	+1,10
As	21,9	+2,40
Cd	2,68	+2,50
Pb	79,7	+0,40
Ni	39,7	+1,70

2.2.3. Ανάλυση των δεδομένων και στατιστικές μέθοδοι

Τα 90 δείγματα χαρακτηρίζονταν από 12 μεταβλητές. Στα αποτελέσματα αυτά εφαρμόστηκαν βασικές μη παραμετρικές και παραμετρικές δοκιμές και διαφορετικές πολυπαραμετρικές τεχνικές. Για την πρώτη ανάλυση (διαφοροποίηση μεταξύ των μονάδων ιχθυοκαλλιέργειας), καθορίστηκαν οι τρεις ομάδες, δηλαδή οι αντίστοιχες ανεξάρτητες μονάδες: CH (Χίος-Οινούσσες), NA (Ναύπακτος) και AS (Αστακός). Επιπλέον, ερευνήθηκαν εποχιακές διακυμάνσεις για κάθε είδος. Μόνο τα απομακρυσμένα (DISTANT, DI) δείγματα, δηλαδή εκείνα που συλλέχθηκαν 50 ή 100 m μακριά από τους κλωβούς ή τα παρακείμενα (ZERO, ZE) δείγματα χρησιμοποιήθηκαν αντίστοιχα για την κάθε ανάλυση. Οι τελευταίες δύο ομάδες δειγμάτων (DI και ZE) χρησιμοποιήθηκαν για την δεύτερη ανάλυση (την μελέτη της επίδρασης των ιχθυοκαλλιεργειών στην σύσταση των θαλασσίων ιζημάτων).

Έτσι, αρχικά οι μη παραμετρικές δοκιμές Kruskal-Wallis και U Mann-Whitney χρησιμοποιήθηκαν, ώστε να διερευνηθεί η σημαντικότητα των διαφορών στις μεταβλητές. Επιπλέον, διαγράμματα Whiskers χρησιμοποιήθηκαν για την οπτικοποίηση των αποτελεσμάτων.

Η PCA εφαρμόστηκε στα DI δείγματα, ώστε να αποκαλυφθούν τυχόν "γηγενείς" διαφορές μεταξύ των μονάδων ιχθυοκαλλιέργειας, αλλά και στο σύνολο των δειγμάτων, ώστε να ανιχνευτούν διαφορές μεταξύ των DI και ZE δειγμάτων. Οι συσχετίσεις μεταξύ των μεταβλητών μελετήθηκαν με την βοήθεια της FA. Έτσι, ταυτοποιήθηκαν επίσης οι παράγοντες που καθορίζουν την διαφοροποίηση των μονάδων (στην πρώτη ανάλυση) ή την μόλυνση στην περιοχή (στην δεύτερη ανάλυση).

Η DA στο κεφάλαιο αυτό, χρησιμοποιήθηκε για την αποκάλυψη των διαφορών/ομοιοτήτων μεταξύ των θέσεων δειγματοληψίας, Εφαρμόστηκε στα αρχικά δεδομένα με την βοήθεια της κλασικής προσέγγισης. Τριάντα-οκτώ (38) DI δείγματα χρησιμοποιήθηκαν για την διαφοροποίηση των μονάδων ιχθυοκαλλιέργειας, ενώ το σύνολο αυτών (48 CH, 12 NA και 14 AS) για την διαφοροποίηση των ZE και DI δειγμάτων. Στην τελευταία αυτή ανάλυση, τα DI δείγματα διαχωρίστηκαν ανάλογα με την απόστασή τους από τους κλωβούς (50 ή 100 m), καθώς η DA απαιτεί τουλάχιστον τρεις ομάδες για την κατασκευή των LDF. Επικύρωση των DA μοντέλων δεν έγινε, καθώς ήταν εκτός του σκοπού του κεφαλαίου αυτού.

2.3. ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΣΥΖΗΤΗΣΗ

2.3.1. Πρώτες παρατηρήσεις

Από τους πίνακες 2.2-2.4 των αποτελεσμάτων, μπορούν να γίνουν οι πρώτες βασικές παρατηρήσεις.

Έτσι, η μονάδα των Χίου-Οινουσσών (CH) έδωσε υψηλές τιμές σε θρεπτικά συστατικά, Hg και Zn συγκρινόμενη με την μονάδα της Ναυπάκτου (NA). Επιπλέον, υπάρχει συσσώρευση σε Cu, Cd, Hg, As, Zn, C, N και P για τα δείγματα που βρίσκονται κάτω από τους κλωβούς. Σε πολλές περιπτώσεις, η συγκέντρωση των C και N σε απόσταση 100 m από τους κλωβούς ήταν αμελητέα και συγκρίσιμη με των δειγμάτων ελέγχου. Εξαίρεση αποτελούν οι φάρμες των Οινουσσών (108.7.S) και του Πρωτέα (108.5.S), οι οποίες είναι πολύ καλά προστατευμένες από τα θαλάσσια ρεύματα (σχ. 2.2). Η συσσώρευση του C κάτω από τους κλωβούς, κυμαίνεται από 83 ως 350 % σε σχέση με τα δείγματα ελέγχου. Αυτές οι τιμές είναι χαμηλότερες από τις αντίστοιχες που καταγράφηκαν από παρόμοια μελέτη [8], όπου αναφέρθηκαν τιμές αύξησης 445 % (μέση τιμή) και 641 % (μέγιστη). Σημαντική συσσώρευση παρατηρείται επίσης για τον P με τις συγκεντρώσεις κάτω από τους κλωβούς να καταγράφονται ως 100 φορές μεγαλύτερες των δειγμάτων ελέγχου. Σε απόσταση 100 m από τους κλωβούς οι συγκεντρώσεις ήταν αμελητέες.

Σχήμα 2.2 : Σημεία δειγματοληψίας για την μονάδα των Χίου-Οινουσσών (CH).

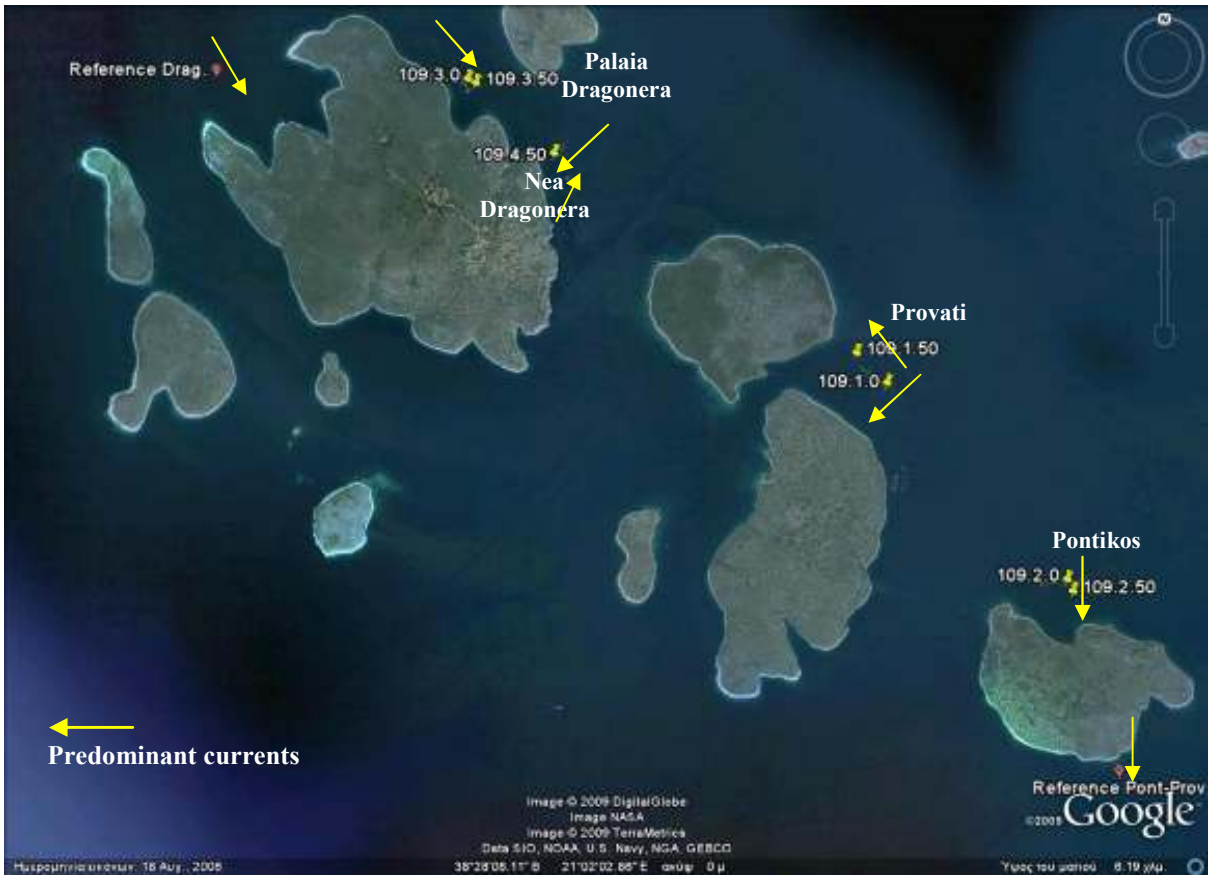
Ο Cu έδωσε ιδιαίτερα αυξημένες τιμές στα απομακρυσμένα δείγματα (μεγαλύτερες των δειγμάτων ελέγχου). Έτσι, στις περισσότερες των περιπτώσεων, υψηλές τιμές Cu (αν και χαμηλότερες των ΖΕ) καταγράφηκαν και για τα DI δείγματα. Τα μέταλλα Cd και Zn έδωσαν πολύ υψηλές τιμές για τα DI δείγματα. Για τα Pb, As, Fe, Mn και Ni δεν μπορούν να εξαχθούν ασφαλή αποτελέσματα που να αφορούν τις διαφορές μεταξύ DI και ΖΕ δειγμάτων. Ο Hg έδωσε μια πολύ υψηλή τιμή στην θέση Πρωτέας (ΖΕ σημείο). Τα Fe, Mn και Ni ήταν εντυπωσιακά χαμηλότερα στα δείγματα ΖΕ σε σχέση με τα αντίστοιχα δείγματα των άλλων δύο μονάδων. Για το Cd, οι Οινούσσες έδωσαν μια πολύ υψηλή τιμή. Αυτές οι μεμονωμένα υψηλές τιμές για τα Hg και Cd στον Πρωτέα και τις Οινούσσες, φαίνεται να αντανακλούν σημειακές τιμές εκπομπών καθώς πρόκειται για πολύ καλά προστατευμένες περιοχές.

Οι μονάδες των Ναυπάκτου (NA) και Αστακού (AS) βρέθηκαν φορτισμένες με Pb, Fe, Mn και Ni συγκρινόμενες με την CH. Οι ρύποι αυτοί δεν φαίνεται να προέρχονται από την δραστηριότητα της ιχθυοκαλλιέργειας, όπως περιγράφεται και παρακάτω.

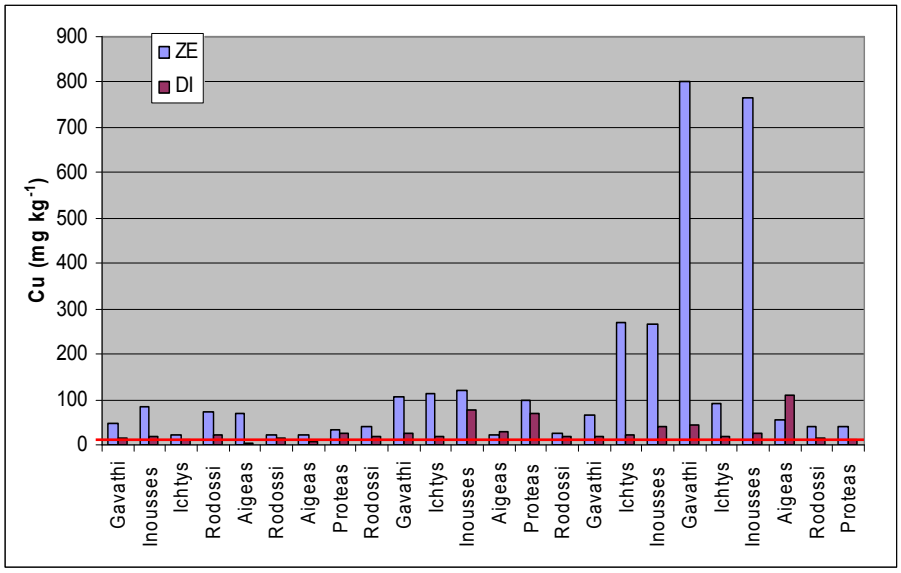
Η Ναύπακτος εμφάνισε συσσώρευση θρεπτικών συστατικών, όχι μόνο κάτω από τους κλωβούς αλλά και σε μεγαλύτερη απόσταση. Στην Εκάλ (107.1.S), το φαινόμενο ήταν ακόμα πιο έντονο εξαιτίας των ισχυρών ανέμων (σχ. 2.3, πίνακας 2.3). Η μακρά περίοδος λειτουργίας και η αμμώδης υφή του ιζήματος στην περιοχή αυτή (πίνακας 2.1), πιθανώς συνεισέφεραν στην προσρόφηση των ρυπαντών. Οι τιμές των C και P στα ΖΕ και DI δείγματα ήταν υψηλότερα των δειγμάτων ελέγχου. Αντίθετα, οι τιμές του N σε όλα σχεδόν τα DI σημεία ήταν χαμηλότερα των δειγμάτων ελέγχου. Ωστόσο, εξαιτίας της μικρότερης δυναμικότητας των μονάδων, (μόνο η Υδροκάλ (107.2.S) έχει αντίστοιχη δυναμικότητα με την CH), τα ΖΕ δείγματα βρέθηκαν λιγότερο φορτισμένα. Αυτή η τελευταία παρατήρηση, αφορά θρεπτικά συστατικά και μέταλλα-μεταλλοειδή όπως Hg, As, Zn. Τα Fe και Ni έδωσαν υψηλές τιμές σε ΖΕ και DI δείγματα. Γενικότερα, οι τιμές των συγκεντρώσεων ήταν πιο ομοιόμορφες σε αυτήν την μονάδα σε σχέση με τις άλλες.

Σχήμα 2.3 : Σημεία δειγματοληψίας για την μονάδα της Ναυπάκτου (ΝΑ).

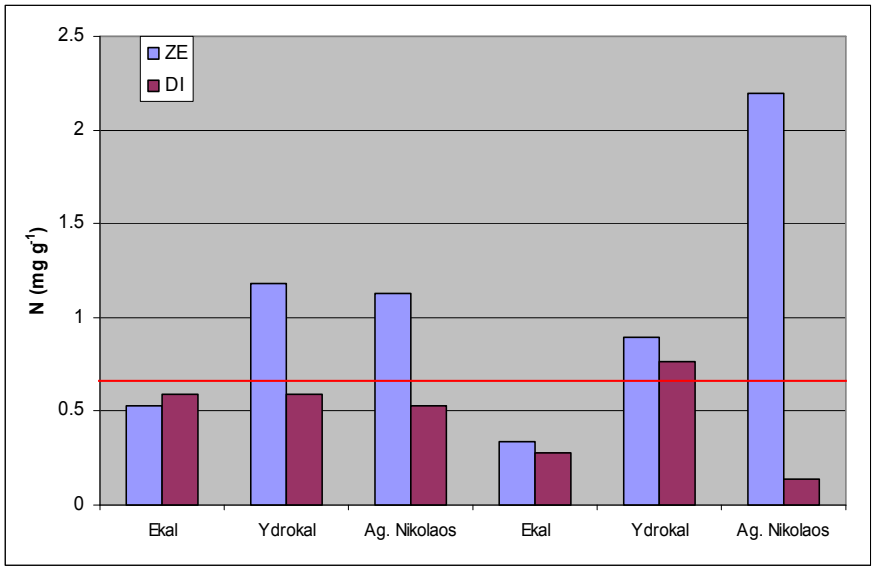
Στον Αστακό, οι μονάδες ιχθυοκαλλιέργειες δείχνουν καθαρά ότι είναι υπεύθυνες για την συσσώρευση θρεπτικών συστατικών κάτω από τους κλωβούς. Ο Ποντικός (109.2.S) ήταν η πιο προβληματική περιοχή, ενώ το Προβάτι (109.1.S) η καθαρότερη (σχ. 2.4, πίνακας 2.4) εξαιτίας ασθενέστερων ανέμων (πίνακας 2.1) που δεν μπορούν να διαμοιράσουν τους ρυπαντές σε όλη την περιοχή. Επιπλέον, ο Ποντικός λειτουργεί για μεγάλο χρονικό διάστημα. Το Προβάτι δέχεται δυνατούς βόρειους ανέμους που "ξεπλένουν" την περιοχή. Όπως και στην περιοχή των Οινουσσών, η επίδραση της ιχθυοκαλλιέργειας (ειδικότερα για τον C) εκτείνεται σε μεγαλύτερη έκταση για την Παλαιά Δραγονέρα (109.3.S), αν και οι απόλυτες τιμές δεν είναι πολύ υψηλές, προφανώς λόγω της μικρής δυναμικότητας της φάρμας. Σχετικά με τα μέταλλα-μεταλλοειδή, τα Cu, Cd, Pb και Zn έδειξαν πολύ υψηλές τιμές στα ΖΕ σημεία, ενώ αντίθετα αποτελέσματα έδωσε το As. Συγκρινόμενη με τις άλλες μονάδες, ο Αστακός είναι η μόνη που χαρακτηρίζεται από υψηλές τιμές Pb σε ΖΕ δείγματα, προφανώς λόγω των μηχανισμών μεταφοράς (φερτά υλικά), καθώς ολόκληρη η μονάδα βρίσκεται σε ένα ιδιαίτερα κλειστό κόλπο (σχ. 2.1).



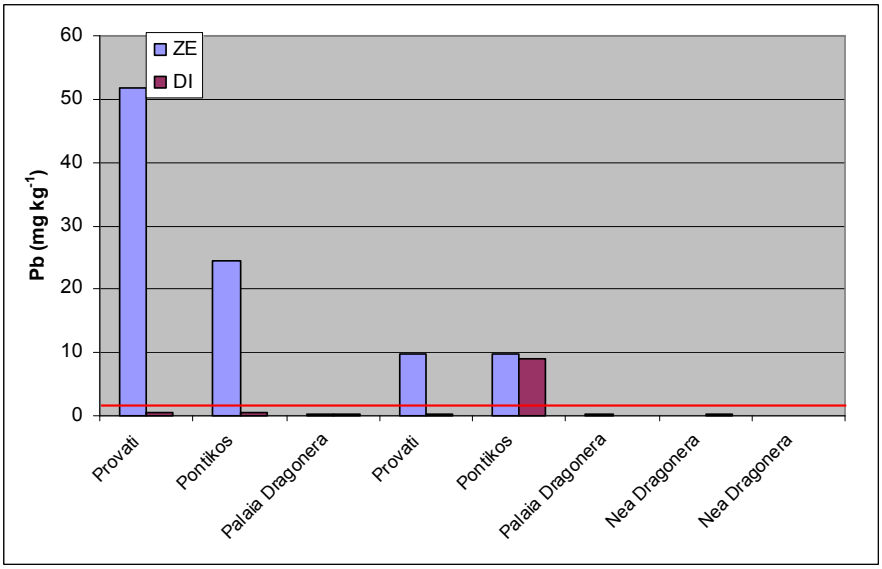
Σχήμα 2.4 : Σημεία δειγματοληψίας για την μονάδα του Αστακού (AS).



(α)



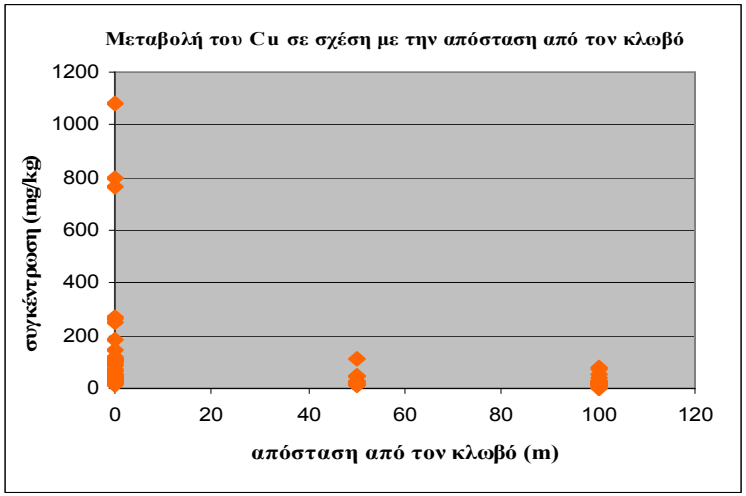
(β)



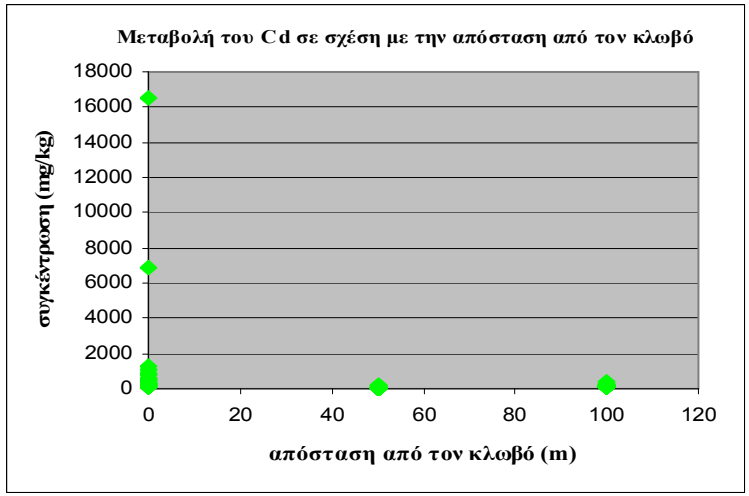
(γ)

Σχήμα 2.5 : Συγκεντρώσεις χαρακτηριστικών μεταβλητών για κάθε φάρμα (ξεχωριστά κάθε μονάδα): (α) Cu στην CH, (β) N στην NA και (γ) Pb στον AS. Η κόκκινη γραμμή δείχνει τα δείγματα ελέγχου.

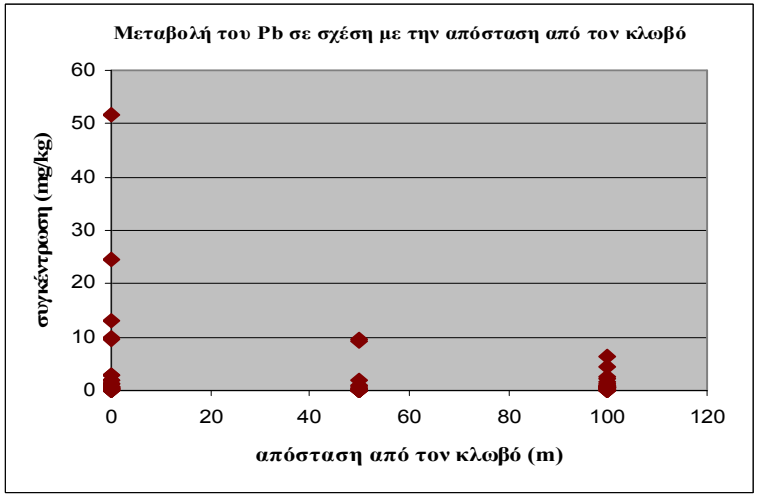
Οι παραπάνω παρατηρήσεις επαληθεύονται στο σχήμα 2.5 ξεχωριστά για κάθε μονάδα και το σχήμα 2.6, για όλα τα δείγματα των τριών μονάδων (74 συνολικά, εκτός των δειγμάτων ελέγχου). Τα C, N, P, Cu, Zn και λιγότερο το Cd φαίνεται ότι αποτελούν τις κρισιμότερες μεταβλητές για τον διαχωρισμό των σημείων που γειτονεύουν με τους κλωβούς και τα πιο απομακρυσμένα από αυτά. Η πολυπαραμετρική ανάλυση που έγινε παρακάτω, επιβεβαίωσε αυτές τις παρατηρήσεις.


Μεταβολή του Cu σε σχέση με την απόσταση από τον κλωβό
συγκέντρωση (mg/kg)
απόσταση από τον κλωβό (m)
0
200
400
600
800
1000
1200
0
20
40
60
80
100
120

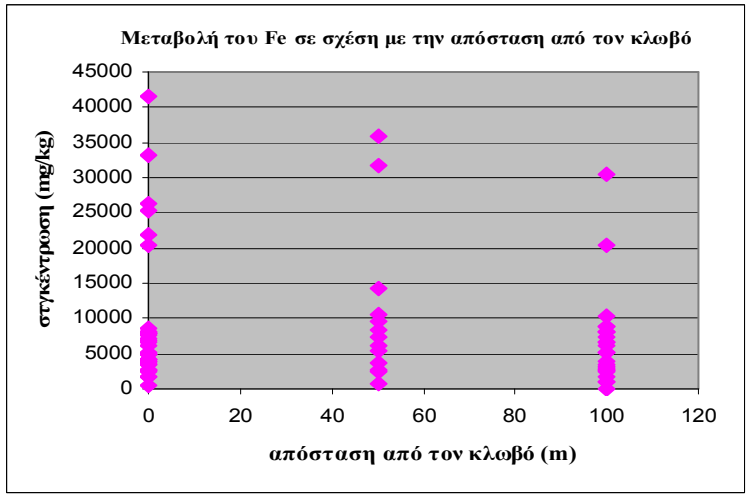
(α)


Μεταβολή του Cd σε σχέση με την απόσταση από τον κλωβό
συγκέντρωση (mg/kg)
απόσταση από τον κλωβό (m)
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
0
20
40
60
80
100
120

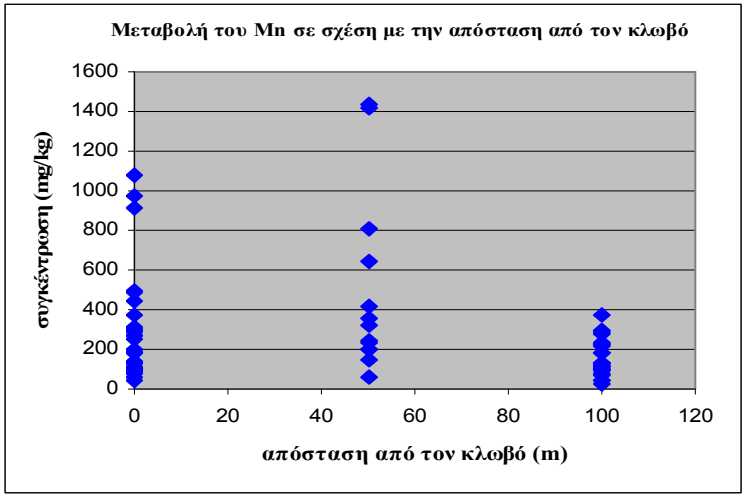
(β)


Μεταβολή του Pb σε σχέση με την απόσταση από τον κλωβό
συγκέντρωση (mg/kg)
απόσταση από τον κλωβό (m)
0
10
20
30
40
50
60
0
20
40
60
80
100
120

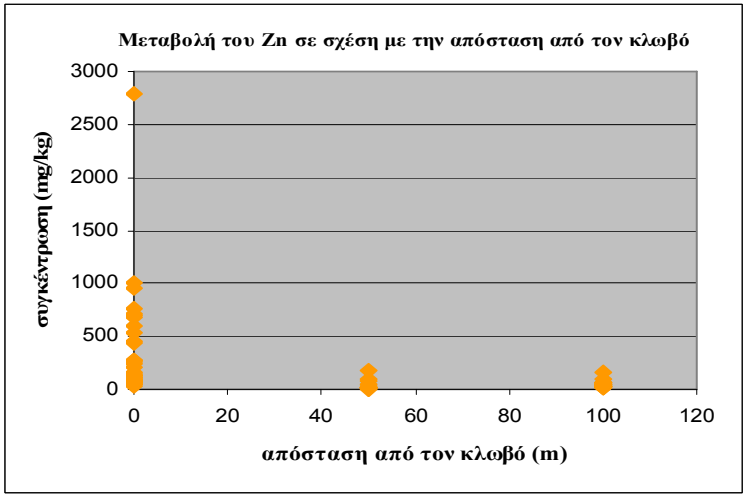
(γ)


Μεταβολή του Fe σε σχέση με την απόσταση από τον κλωβό
στγκέντρωση (mg/kg)
απόσταση από τον κλωβό (m)
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
0
20
40
60
80
100
120

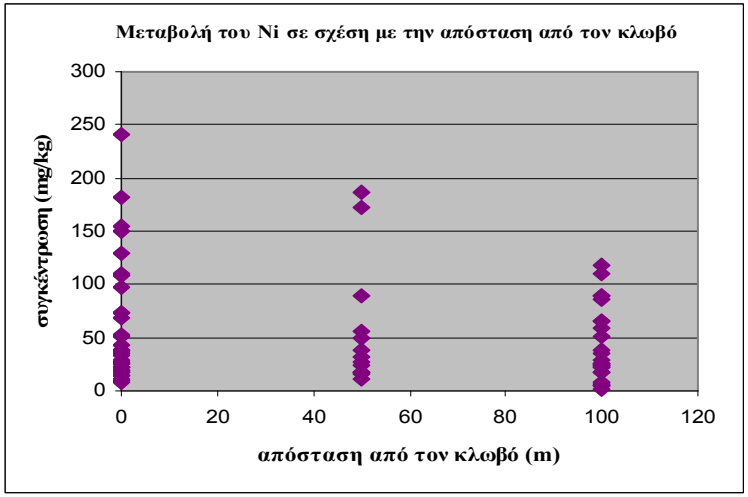
(δ)


Μεταβολή του Mn σε σχέση με την απόσταση από τον κλωβό
συγκέντρωση (mg/kg)
απόσταση από τον κλωβό (m)
0
200
400
600
800
1000
1200
1400
1600
0
20
40
60
80
100
120

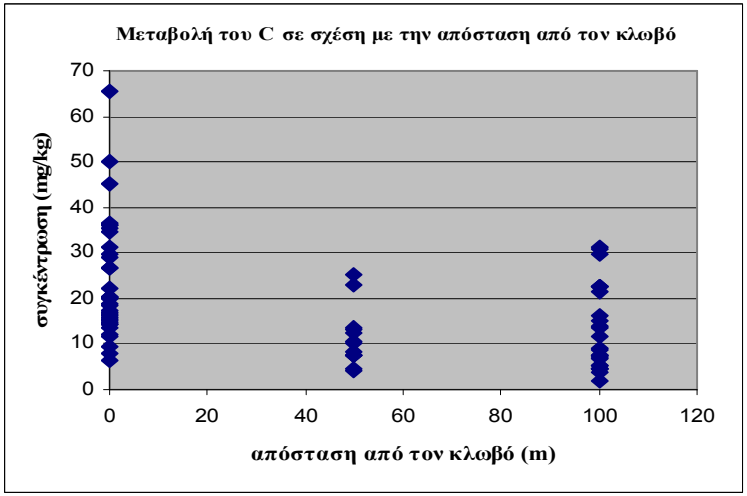
(ε)


Μεταβολή του Zn σε σχέση με την απόσταση από τον κλωβό
συγκέντρωση (mg/kg)
απόσταση από τον κλωβό (m)
0
500
1000
1500
2000
2500
3000
0
20
40
60
80
100
120

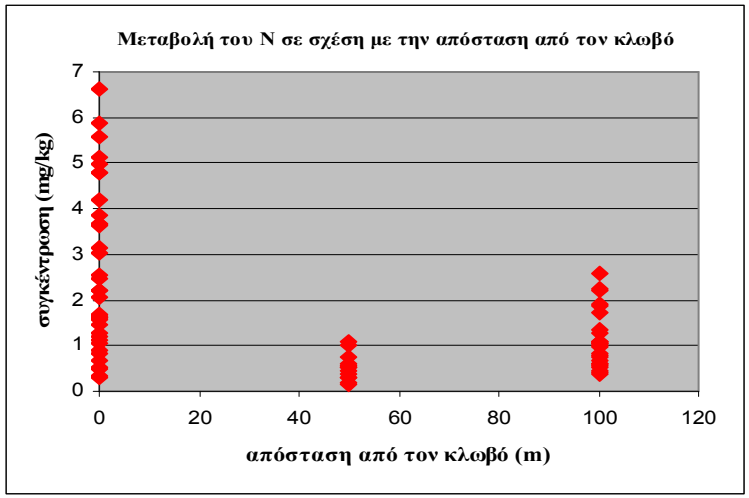
(στ)


Μεταβολή του Ni σε σχέση με την απόσταση από τον κλωβό
συγκέντρωση (mg/kg)
απόσταση από τον κλωβό (m)
0
50
100
150
200
250
300
0
20
40
60
80
100
120

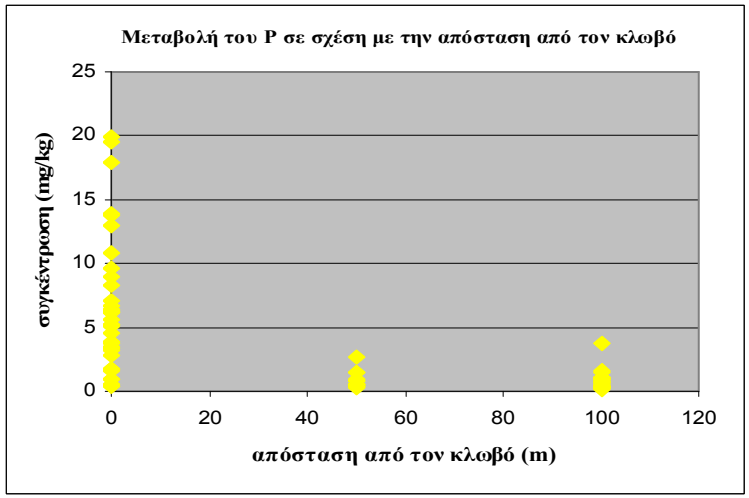
(ζ)


Μεταβολή του C σε σχέση με την απόσταση από τον κλωβό
συγκέντρωση (mg/kg)
απόσταση από τον κλωβό (m)
0
10
20
30
40
50
60
70
0
20
40
60
80
100
120

(η)



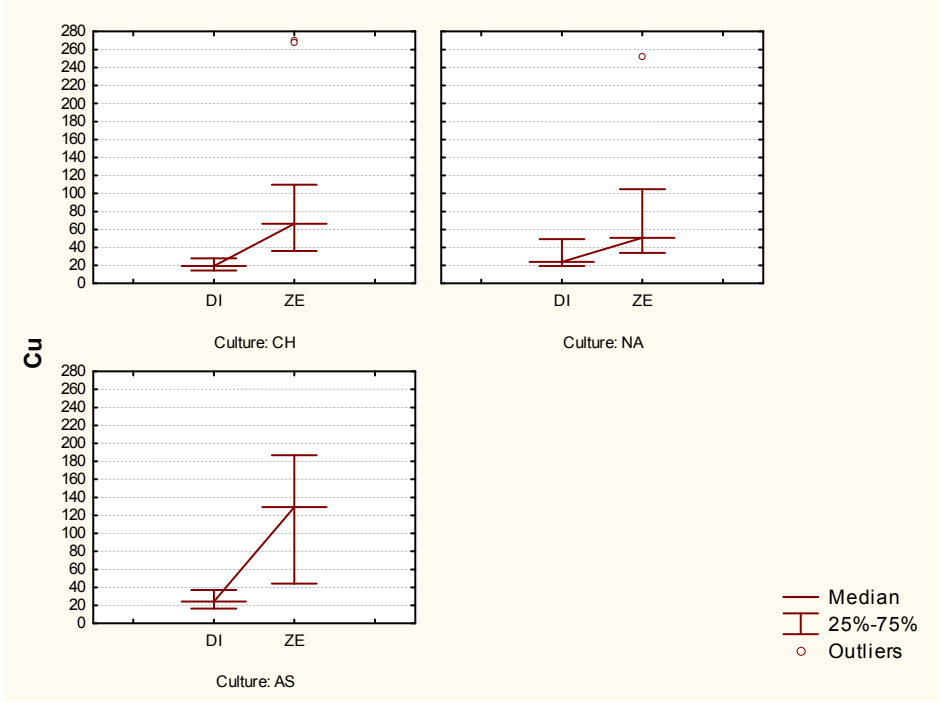
(θ)



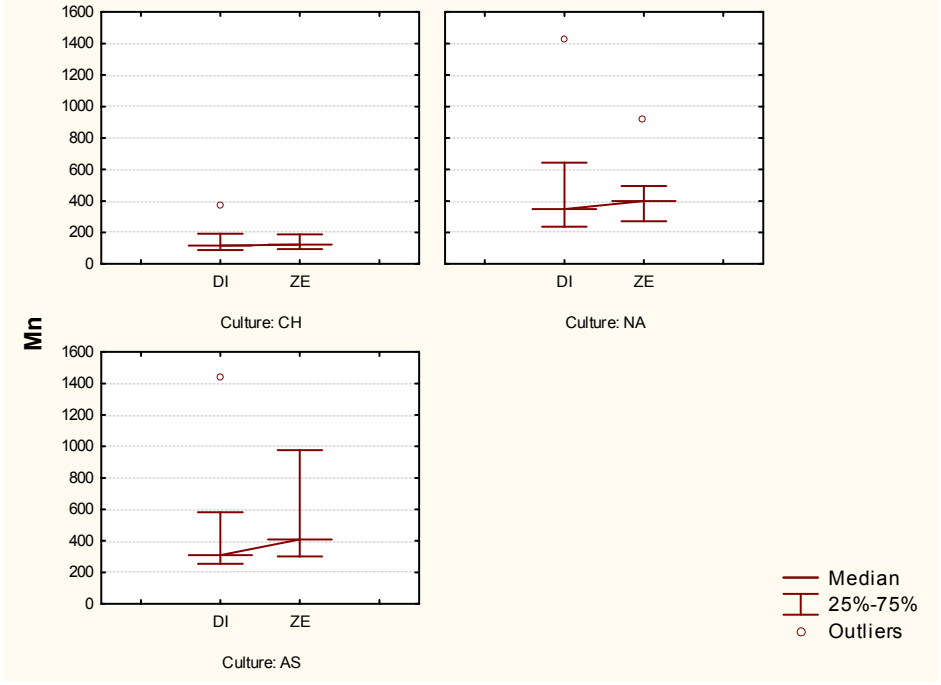
(ι)

Σχήμα 2.6: Διαγράμματα συγκεντρώσεων για τις πιο χαρακτηριστικές μεταβλητές, σε σχέση με την απόσταση των σημείων από τους κλωβούς.

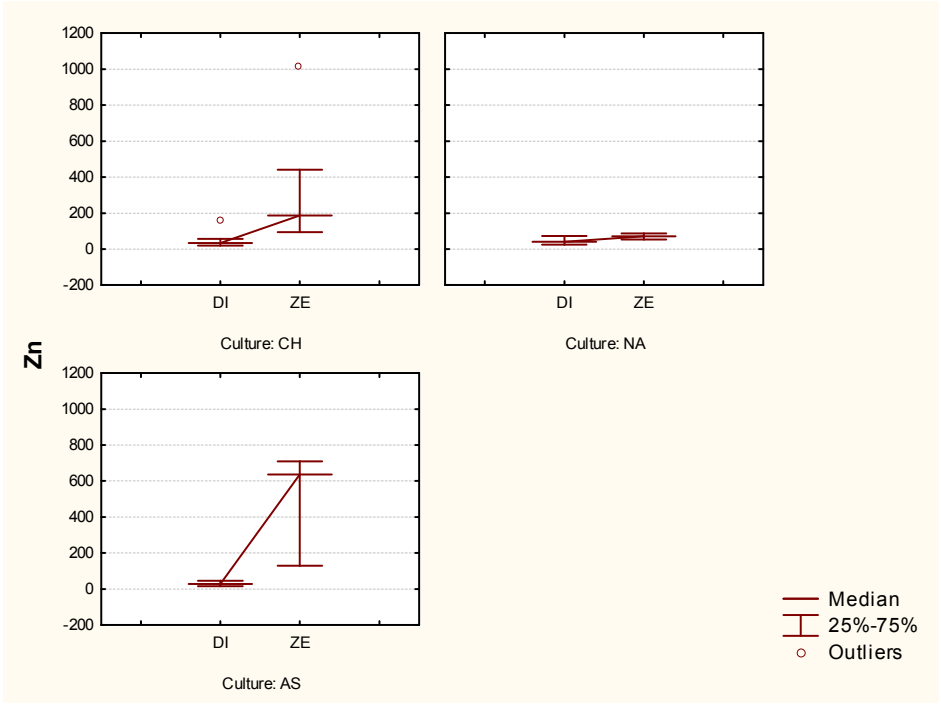
Επιπλέον στο σχήμα 2.7, δίνονται τα διαγράμματα whiskers των πιο χαρακτηριστικών μεταβλητών σε σχέση με την απόσταση από τους κλωβούς, ξεχωριστά για κάθε μονάδα ιχθυοκαλλιέργειας. Είναι φανερό, ότι στοιχεία σαν τα Cu, Zn, C, N, P και Mn, Ni τα οποία διαφοροποιούνται ή όχι αντίστοιχα για τα ZE και DI σημεία, δείχνουν παρόμοια συμπεριφορά για όλες τις μονάδες.



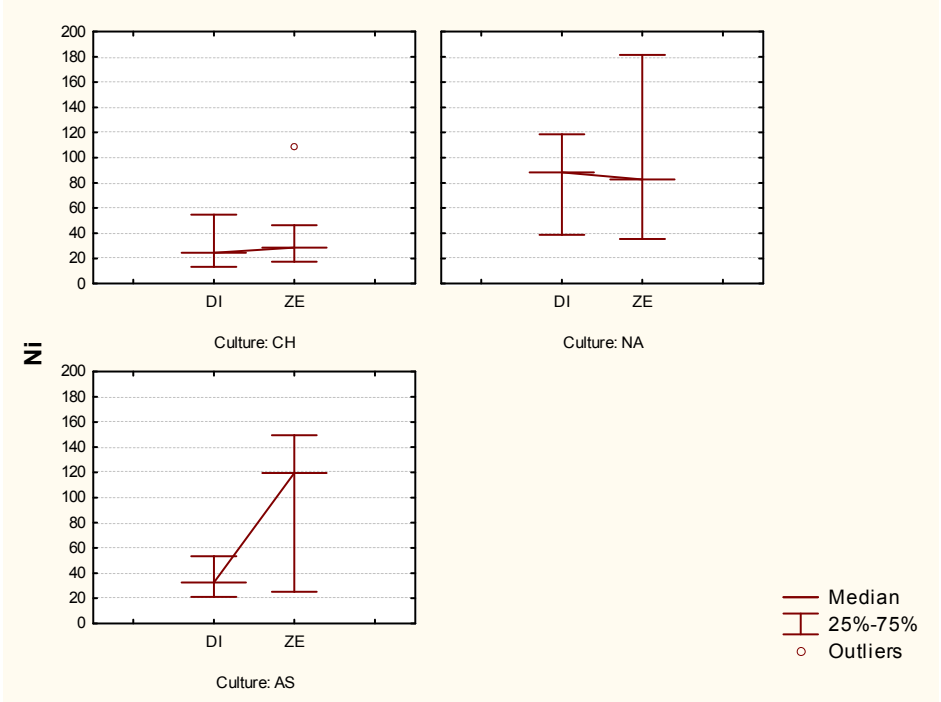
(α)



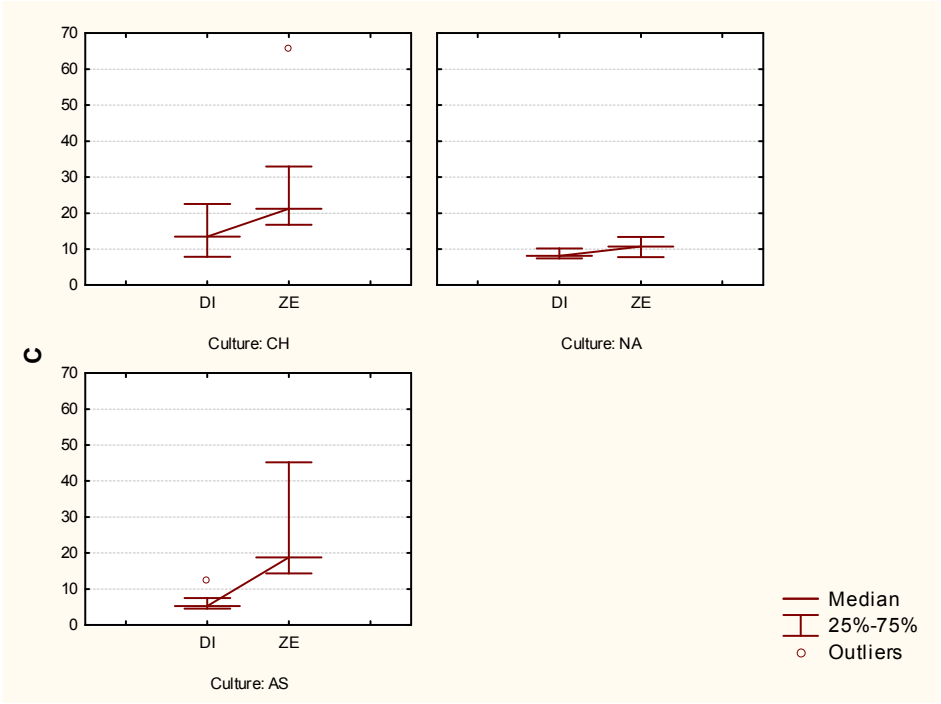
(β)


Zn
1200
1000
800
600
400
200
0
-200
DI
ZE
Culture: CH
Culture: NA
Culture: AS
Median
25%-75%
Outliers

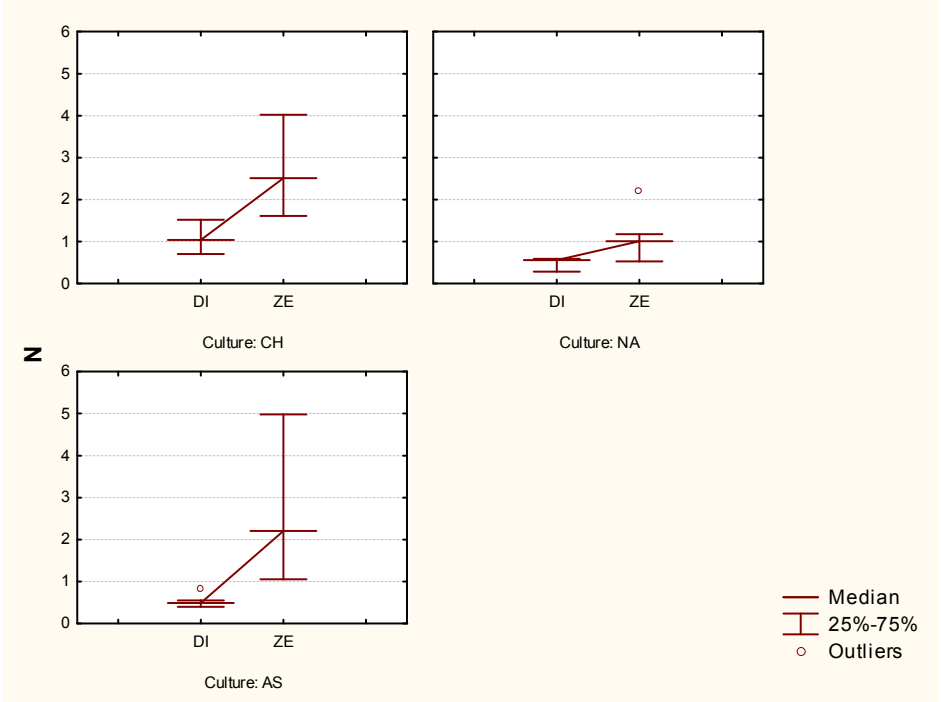
(γ)


Ni
200
180
160
140
120
100
80
60
40
20
0
DI
ZE
Culture: CH
Culture: NA
Culture: AS
Median
25%-75%
Outliers

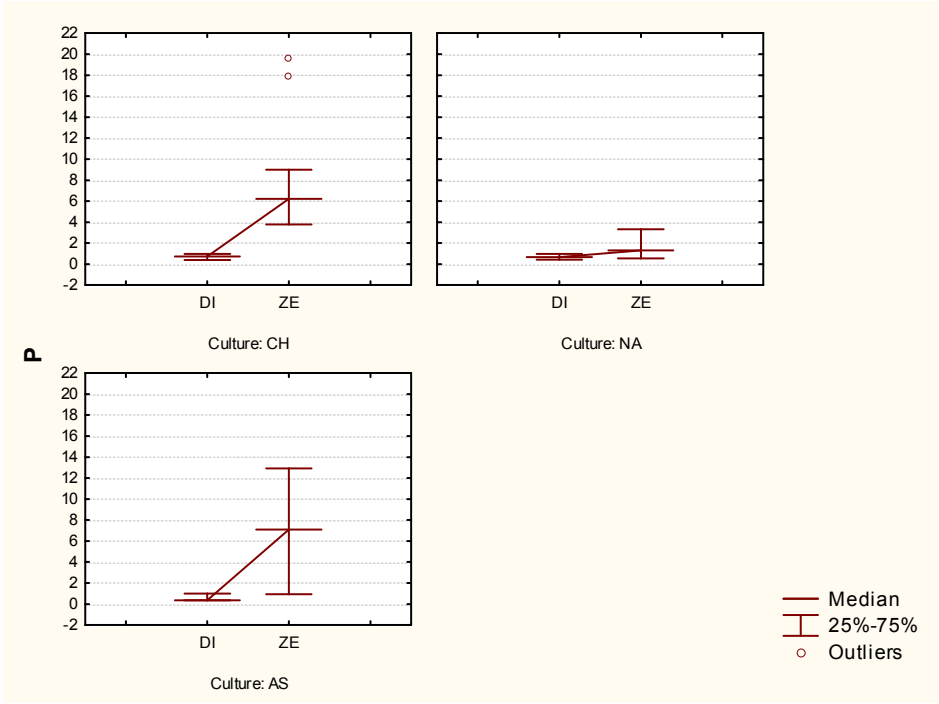
(δ)


C
70
60
50
40
30
20
10
0
DI
ZE
Culture: CH
Culture: NA
Culture: AS
Median
25%-75%
Outliers

(ε)


N
6
5
4
3
2
1
0
DI
ZE
Culture: CH
Culture: NA
Culture: AS
Median
25%-75%
Outliers

(στ)



(ζ)

Σχήμα 2.7: Διαγράμματα whiskers για τις πιο χαρακτηριστικές μεταβλητές σε σχέση με την απόσταση από τους κλωβούς για κάθε μονάδα ξεχωριστά (καταγράφονται η μέση τιμή και τα 25, 75 % τεταρτημόρια).

Οι εποχιακές διακυμάνσεις (σχ. 2.8) μεταξύ χειμώνα και καλοκαιριού, επιβεβαίωσαν την επίδραση των μονάδων ιχθυοκαλλιέργειας στα ιζήματα, καθώς οι διαφορές μεταξύ των DI δειγμάτων φαίνεται να μηδενίζονται. Αντίθετα, τα ΖΕ δείγματα παρουσιάζουν μεγαλύτερες διακυμάνσεις, εξαιτίας της ρύπανσης, αφού τουλάχιστον θεωρητικά, κατά την διάρκεια του καλοκαιριού, τα ψάρια καταναλώνουν περισσότερη τροφή και ο μεταβολισμός είναι ταχύτερος. Ωστόσο, δεν μπορούν να εξαχθούν ασφαλή συμπεράσματα ακόμα και γι΄αυτά τα δείγματα: χαμηλές διαφορές καταγράφηκαν στις μέσες τιμές, ενώ τα εύρη διακύμανσης ποικίλουν. Μη εποχιακές διακυμάνσεις έχουν καταγραφεί και σε σχετικές πρόσφατες δημοσιεύσεις [11, 12].

Σφάλμα! Τα αντικείμενα δεν μπορούν να δημιουργηθούν από την επεξεργασία κωδικών πεδίων.

(α)

Σφάλμα! Τα αντικείμενα δεν μπορούν να δημιουργηθούν από την επεξεργασία κωδικών πεδίων.

(β)

Σφάλμα! Τα αντικείμενα δεν μπορούν να δημιουργηθούν από την επεξεργασία κωδικών πεδίων.

(γ)

Σφάλμα! Τα αντικείμενα δεν μπορούν να δημιουργηθούν από την επεξεργασία κωδικών πεδίων.

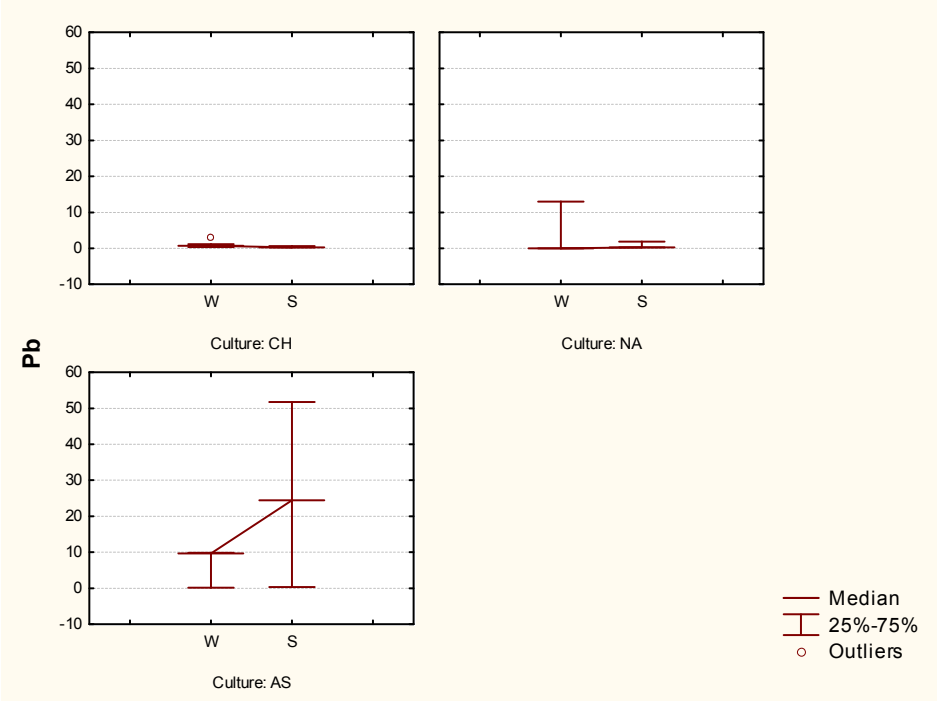
(δ)

Σφάλμα! Τα αντικείμενα δεν μπορούν να δημιουργηθούν από την επεξεργασία κωδικών πεδίων.

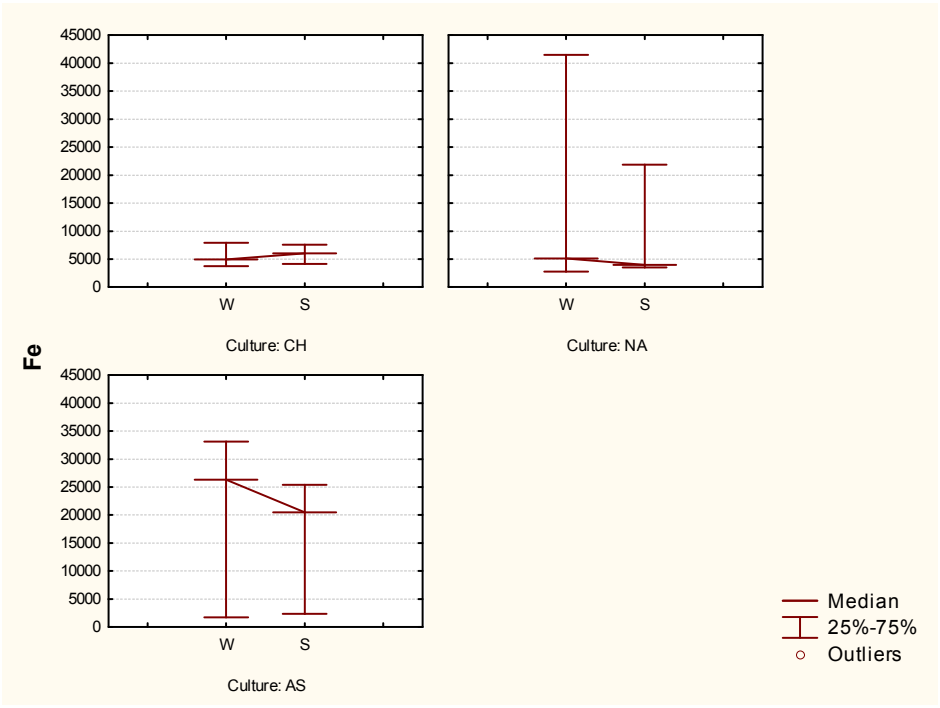
(ε)

*Σχήμα 2.8: Διαγράμματα whiskers για τις πιο χαρακτηριστικές μεταβλητές σε σχέση με την απόσταση από τους κλωβούς για τις δύο εποχές δειγματοληψίας (**W**inter και **S**ummer) ξεχωριστά (καταγράφονται η μέση τιμή και τα 25, 75 % τεταρτημόρια).*

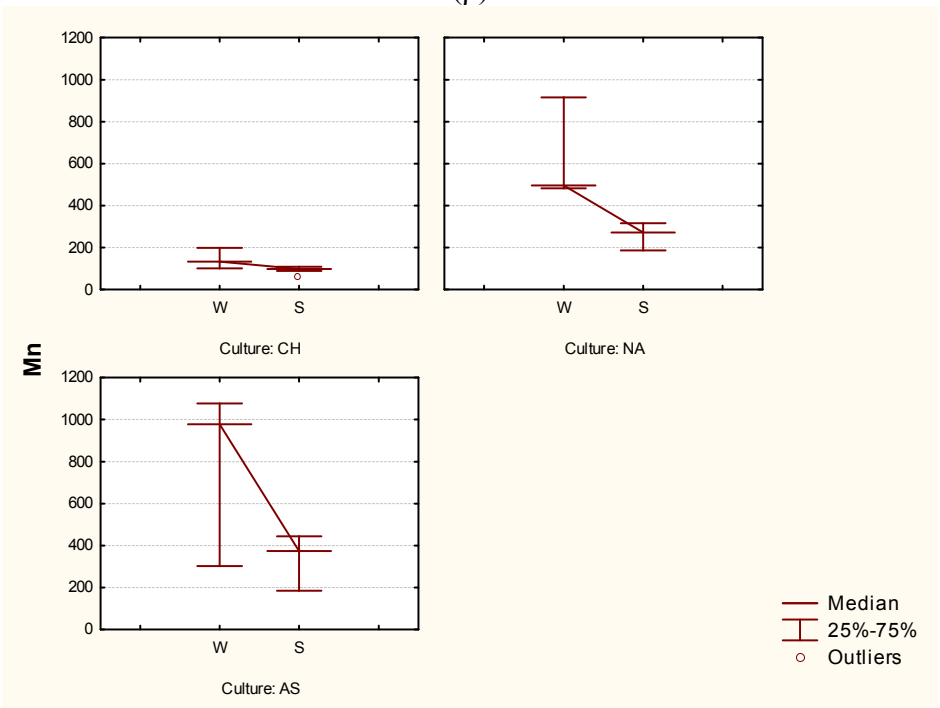
Τα ίδια αποτελέσματα επιβεβαιώθηκαν και στο σχήμα 2.9, όπου μόνο τα ZE δείγματα χρησιμοποιήθηκαν. Μέταλλα όπως τα Pb, Fe, Mn έδειξαν διαφορές ανάμεσα στις δύο εποχές για τις μονάδες NA και AS προφανώς εξαιτίας μηχανισμών μεταφοράς (θαλάσσια ρεύματα). Γενικά. ωστόσο, οι εποχιακές διαφορές δεν ήταν ιδιαίτερα σημαντικές όπως έδειξε και η δοκιμή U Mann-Whitney (πίνακας 2.6).



(α)



(β)



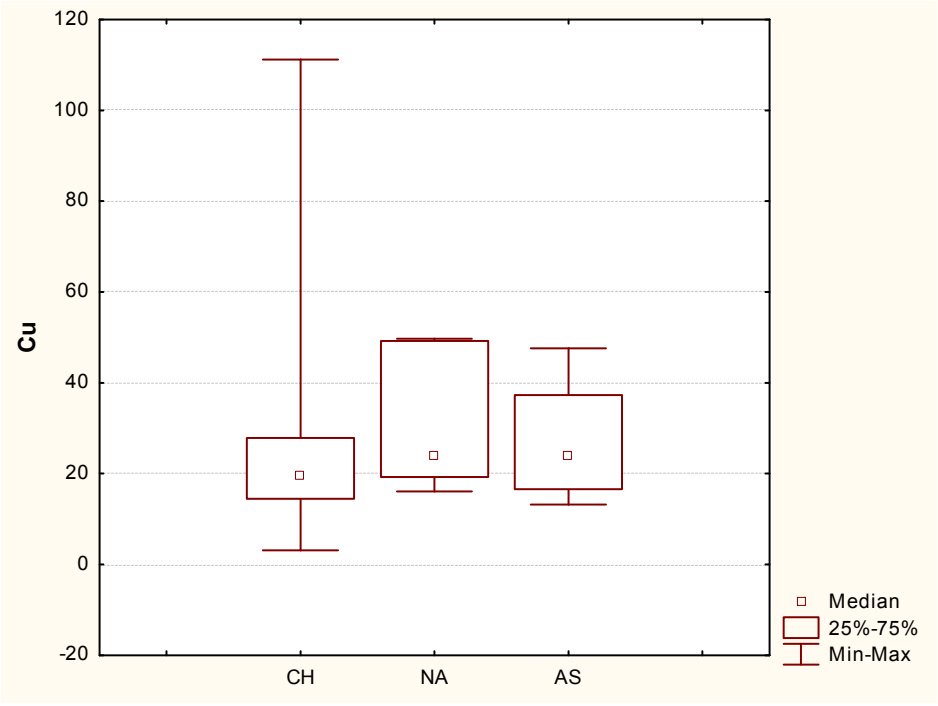
(γ)

*Σχήμα 2.9: Διαγράμματα whiskers για τις πιο χαρακτηριστικές μεταβλητές σε σχέση με την απόσταση από τους κλωβούς για τις δύο εποχές δειγματοληψίας (**W**inter και **S**ummer ξεχωριστά (καταγράφονται η μέση τιμή και τα 25, 75 % τεταρτημόρια). Χρησιμοποιήθηκαν μόνο ZE δείγματα.*

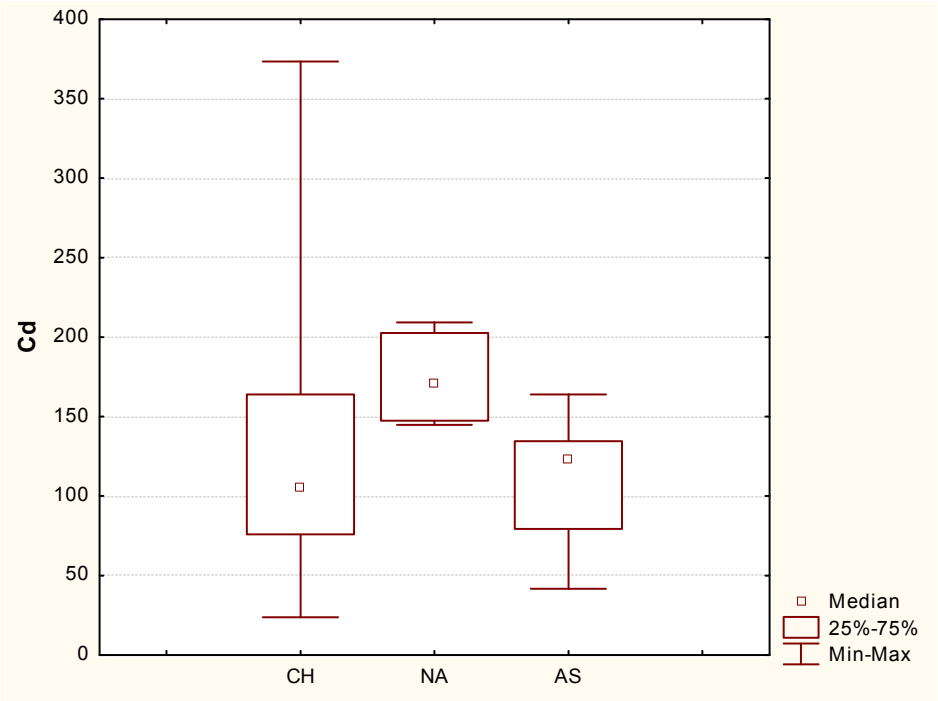
Πίνακας 2.6: Στατιστική εκτίμηση των διαφορών ανάμεσα στις δυο εποχές για τα ZE δείγματα (δοκιμή U Mann-Whitney). Για τον χειμώνα N=24 και το καλοκαίρι N=12.

Στοιχείο	p-level	Στοιχείο	p-level
Cu	0,92	**Mn**	0,43
Cd	0,52	**Zn**	0,66
Pb	0,59	**Ni**	0,43
Hg	0,35	**C**	0,56
As	0,32	**N**	0,76
Fe	0,67	**P**	0,84

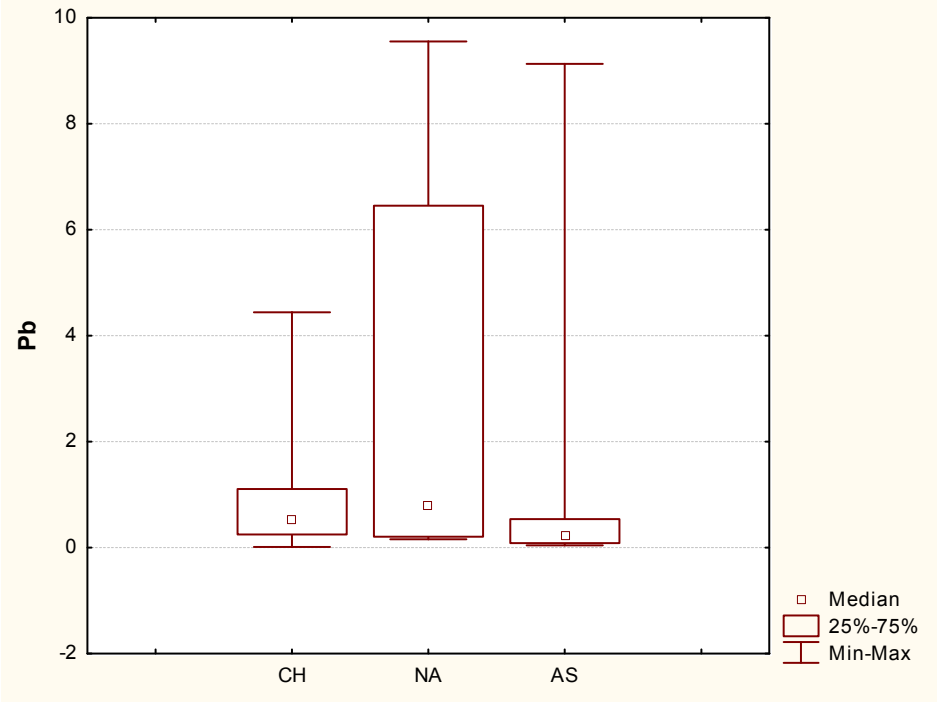
Στο σχήμα 2.10, αναλύονται μόνο τα DI δείγματα, ώστε να μελετηθούν οι γηγενείς διαφορές (λόγω θαλασσίων ρευμάτων και γεωλογικού υποβάθρου) μεταξύ των μονάδων. Τα Cd, Mn, Ni, C, N και λιγότερο τα Pb και Fe φαίνεται να είναι οι κρισιμότερες μεταβλητές, όπως έχει ήδη αναφερθεί.



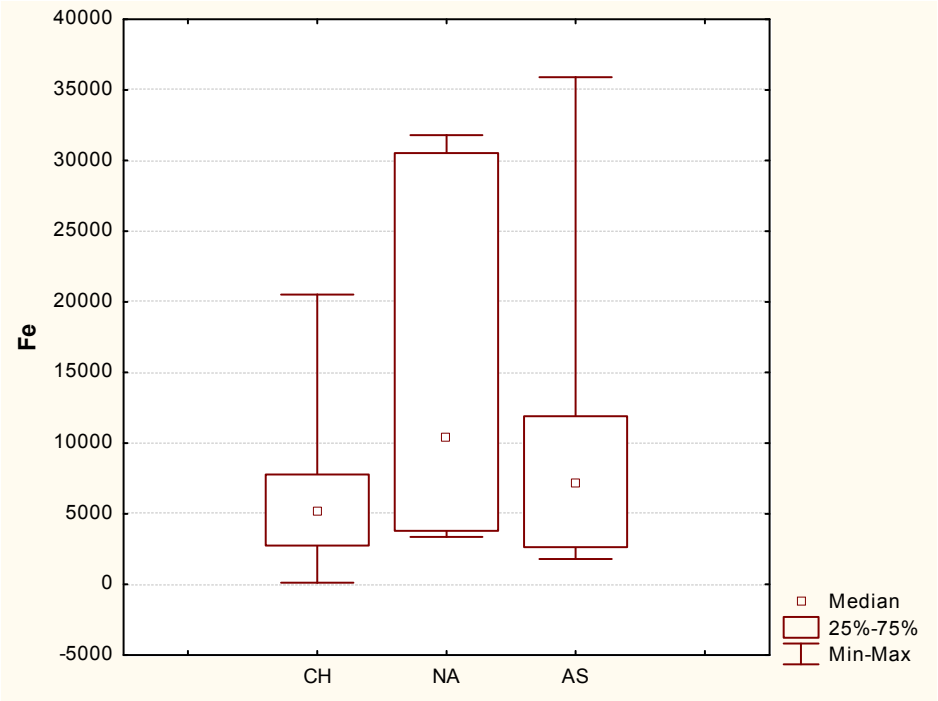
(α)


Cd
400
350
300
250
200
150
100
50
0
CH
NA
AS
Median
25%-75%
Min-Max

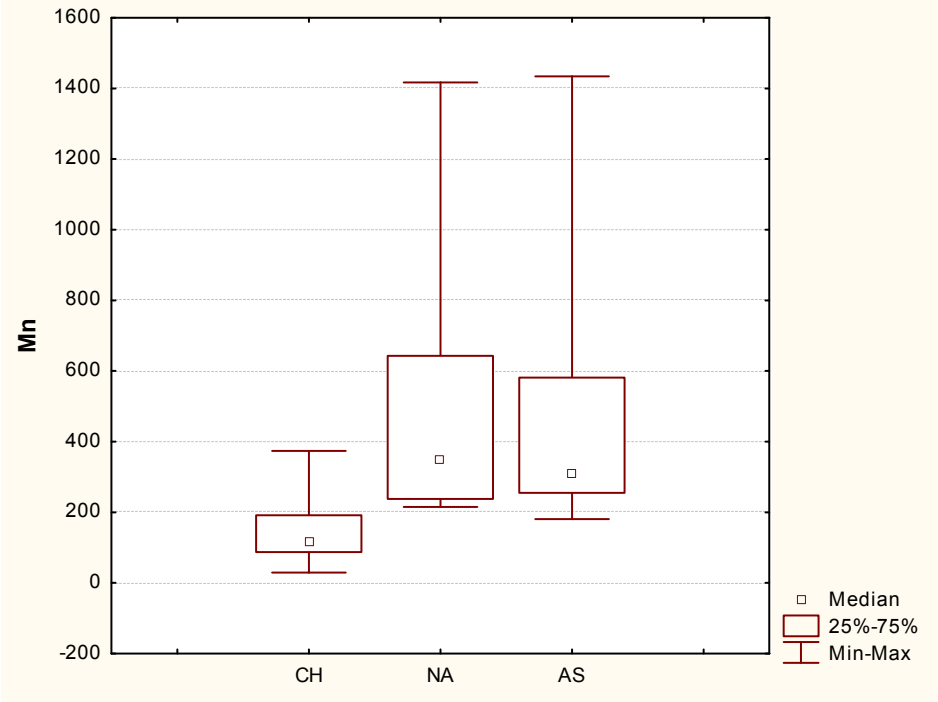
(β)


Pb
10
8
6
4
2
0
-2
CH
NA
AS
Median
25%-75%
Min-Max

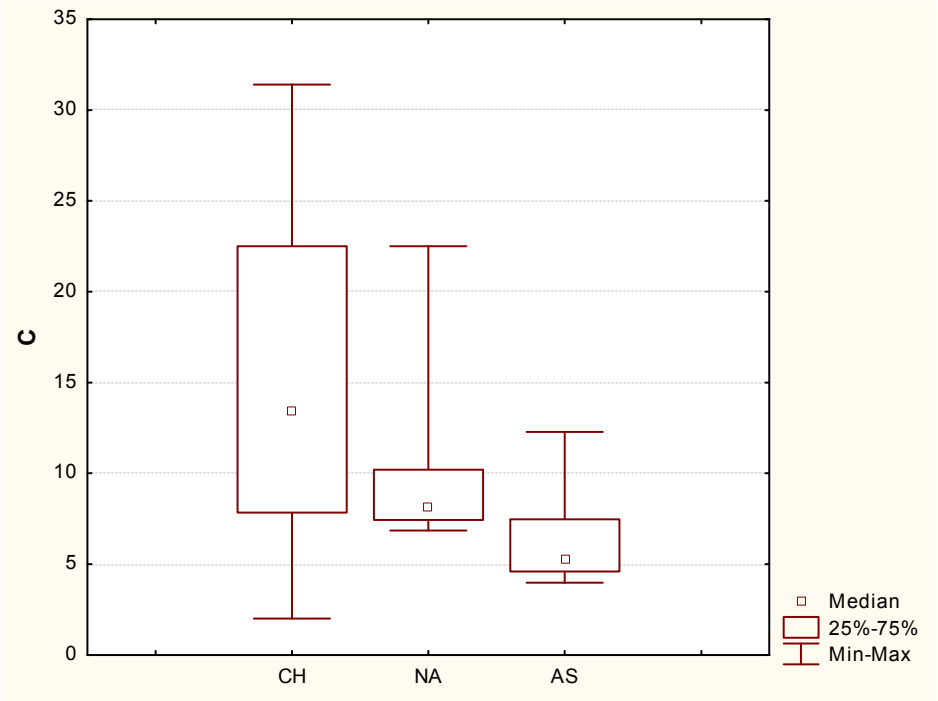
(γ)


Fe
40000
35000
30000
25000
20000
15000
10000
5000
0
-5000
CH
NA
AS
Median
25%-75%
Min-Max

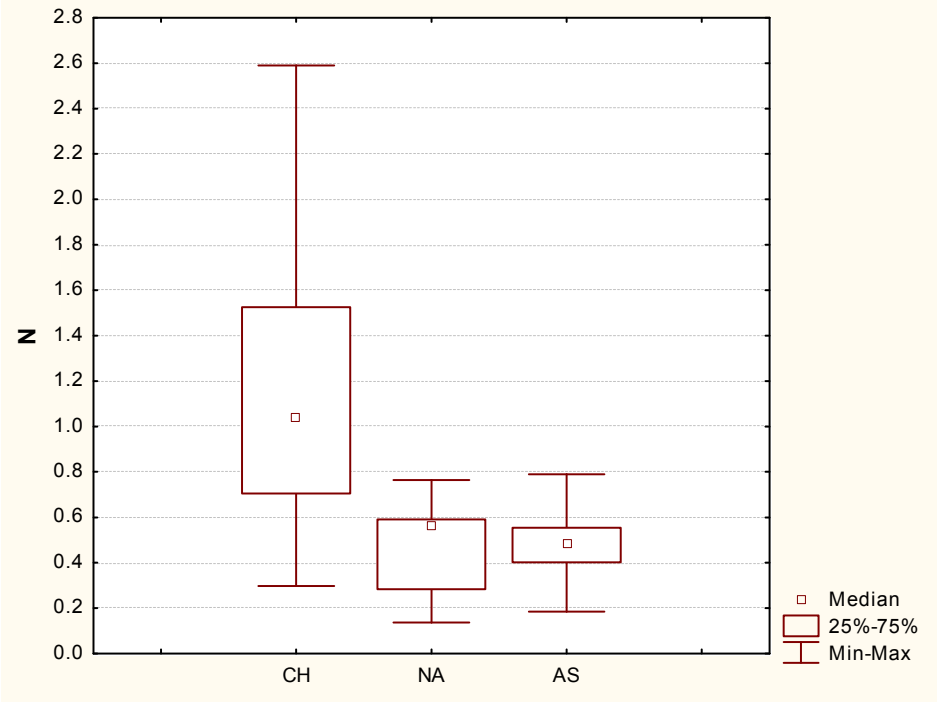
(δ)


Mn
1600
1400
1200
1000
800
600
400
200
0
-200
CH
NA
AS
Median
25%-75%
Min-Max

(ε)



(στ)



(ζ)

Σχήμα 2.10: Διαγράμματα box & whiskers για τις πιο χαρακτηριστικές μεταβλητές. Χρησιμοποιήθηκαν μόνο DI δείγματα.

Για τον ίδιο σκοπό χρησιμοποιήθηκαν οι μη παραμετρικές δοκιμές Kruskal-Wallis και U Mann-Whitney για τα δείγματα DI (πίνακας 2.7). Εξήχθησαν τα ίδια αποτελέσματα.

Πίνακας 2.7: Στατιστική εκτίμηση των διαφορών μεταξύ των μεταβλητών για τα DI δείγματα. Χρησιμοποιήθηκαν οι δοκιμές Kruskal-Wallis και U Mann-Whitney για τις τρεις μονάδες: CH (N=48), NA (N=12) and AS (N=14). Όταν p<0,001 καταγράφεται ως "++", ενώ όταν 0,001<p<0,05 καταγράφεται "+" [21].

Στοιχείο	**Kruskal-Wallis**	**U Mann-Whitney**			**Σημαντικότερες μεταβλητές**
		CH-NA	**CH-AS**	**NA-AS**	
Cu	0,57	0,32	0,571	0,60	
Cd	0,056	+	0,90	+	√
Pb	0,22	0,41	0,21	+	
Hg	+	0,19	+	0,37	
As	0,13	0,069	0,18	0,80	
Fe	0,098	+	0,33	0,30	
Mn	++	++	++	1,0	√
Zn	0,63	0,57	0,51	0,37	
Ni	+	+	0,31	0,16	√
C	+	0,23	+	0,053	√

Στοιχείο	Kruskal-Wallis	U Mann-Whitney			Σημαντικότερες μεταβλητές
		CH-NA	CH-AS	NA-AS	
N	++	+	+	0,70	√
P	0,58	0,98	0,34	0,37	

Επιπλέον, για την εκτίμηση των διαφορών μεταξύ των δειγμάτων DI και ZE, οι ίδιες δοκιμές χρησιμοποιήθηκαν για το σύνολο των δειγμάτων. Τα αποτελέσματα φαίνονται στον πίνακα 2.8. Cu, Cd, Zn, C, N και P ήταν οι κρισιμότερες μεταβλητές. Η δοκιμή U Mann-Whitney εφαρμόστηκε για τον παραπέρα διαχωρισμό των DI δειγμάτων (50 και 100 m απόσταση από τους κλωβούς).

Πίνακας 2.8 : Στατιστική εκτίμηση των διαφορών μεταξύ των μεταβλητών για όλα τα δείγματα (δοκιμή Kruskal-Wallis για δύο ομάδες: ZE (N=36) και DI (N=38) και δομική U Mann-Whitney για τρεις ομάδες: 0 (N=36), 50 (N=13) και 100 (N=25). Όταν $p<0,001$ καταγράφεται ως "++", ενώ όταν $0,001<p<0,05$ καταγράφεται "+" [21].

Στοιχείο	Kruskal-Wallis	U Mann-Whitney			Σημαντικότερες μεταβλητές
		0-50	50-100	0-100	
Cu	++	+	0,49	++	√
Cd	++	++	+	++	√
Pb	0,32	0,094	0,26	0,80	
Hg	+	0,14	0,96	+	
As	0,60	0,12	0,14	0,76	
Fe	0,99	0,40	0,21	0,57	
Mn	0,90	+	+	0,11	

Στοιχείο	Kruskal-Wallis	U Mann-Whitney			Σημαντικότερες μεταβλητές
		0-50	50-100	0-100	
Zn	++	++	0,087	++	√
Ni	0,35	0,82	0,59	0,26	
C	++	++	0,94	++	√
N	++	++	+	++	√
P	++	++	0,77	++	√

Παρόμοια αποτελέσματα καταγράφηκαν και από το t-test για το σύνολο των δειγμάτων (διαφοροποίηση μεταξύ DI και ZE σημείων). Τα αποτελέσματα φαίνονται στον πίνακα 2.9. Η μηδενική υπόθεση Ho που αφορά την ισότητα των μέσων τιμών για τα Cu, Zn, N, C και P απορρίφθηκε.

Πίνακας 2.9: t-test για τις δύο ομάδες DI και ZE (df = 72)

Μεταβλητή	Μέση τιμή		SD		t-value	p-value(1)	Αποδοχή της Ho(2)
	DI	ZE	DI	ZE			
Cu	28,2	150	21,2	237	-3,15	+	X
Cd	132	1038	75,1	2882	-1,94	0,0565	✓
Pb	1,33	3,66	2,30	9,55	-1,46	0,150	✓
Hg	0,250	2,14	0,83	11,1	-1,05	0,298	✓
As	9,94	9,63	6,12	4,95	0,239	0,812	✓
Fe	8012	8811	8386	9477	-0,385	0,702	✓
Mn	273	263	317	252	0,154	0,878	✓

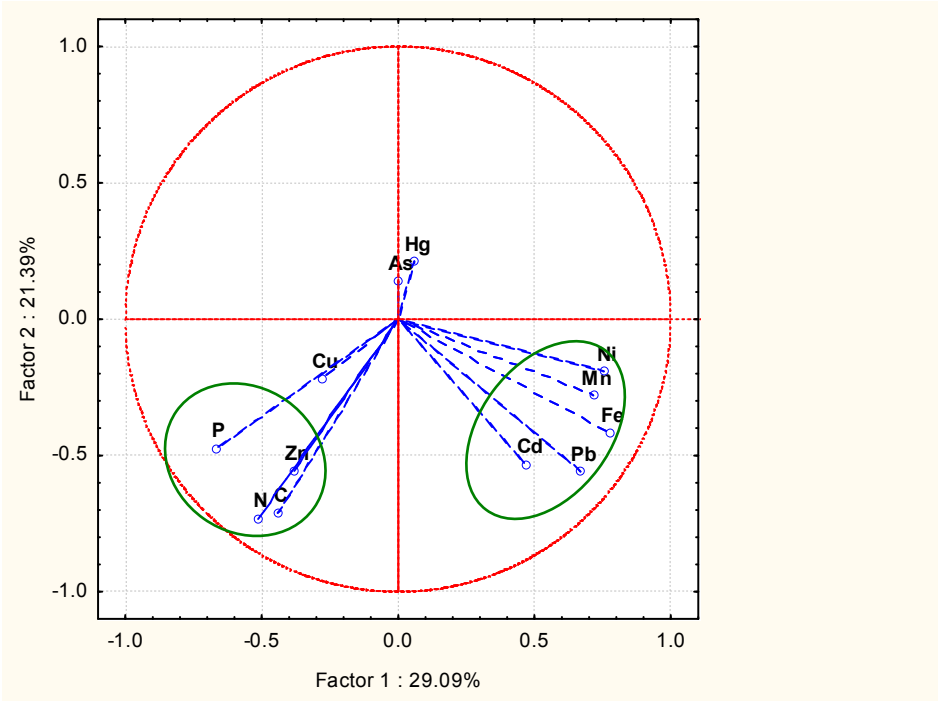
Μεταβλητή	Μέση τιμή		SD		t-value	p-value[(1)]	Αποδοχή της Ho[(2)]
	DI	ZE	DI	ZE			
Zn	44,5	352	37,5	501	-3,78	++	X
Ni	46,9	59,0	44,0	55,8	-1,05	0,299	✓
C	12,5	23,3	8,24	12,7	-4,37	++	X
N	0,913	2,57	0,61	1,77	-5,47	++	X
P	0,815	6,54	0,70	5,28	-6,62	++	X

(1) + για 0,001 < p-value < 0,05 και ++ για p-value < 0,001
(2) η Ho αφορά την ισότητα των μέσων τιμών

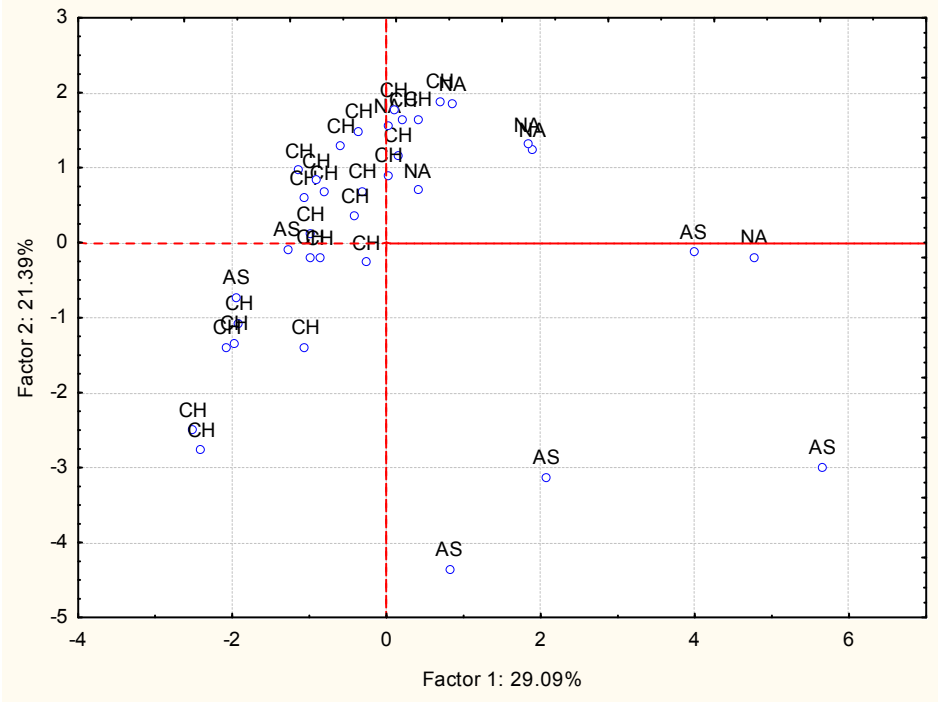
2.3.2. Εφαρμογή της PCA

Η PCA εφαρμόστηκε στα ZE δείγματα (36), ώστε να αποκαλυφθούν οι παράγοντες που διαμορφώνουν την ρύπανση στην κάθε περιοχή ξεχωριστά, αλλά και στο σύνολο των δειγμάτων ώστε να ανιχνευτούν δυνητικές διαφορές ανάμεσα στα DI και ZE δείγματα και συσχετίσεις μεταξύ των μεταβλητών.

Τα διαγράμματα φορτίσεων και σκορ, φαίνονται στο σχήμα 2.11. Οι περισσότερες μεταβλητές Zn, C, N, P, Cd, Pb, Fe, Mn και Ni φαίνονται να συνεισφέρουν στην διαφοροποίηση των τριών μονάδων (σχ. 2.11(α)). Οι παρατηρήσεις αυτές καταγράφηκαν και παραπάνω (§ 2.3.1). Η CH μονάδα διαφοροποιείται από την AS, ενώ δείγματα των CH και NA συγχέονται μεταξύ τους (σχ. 2.11(β)).



(α)

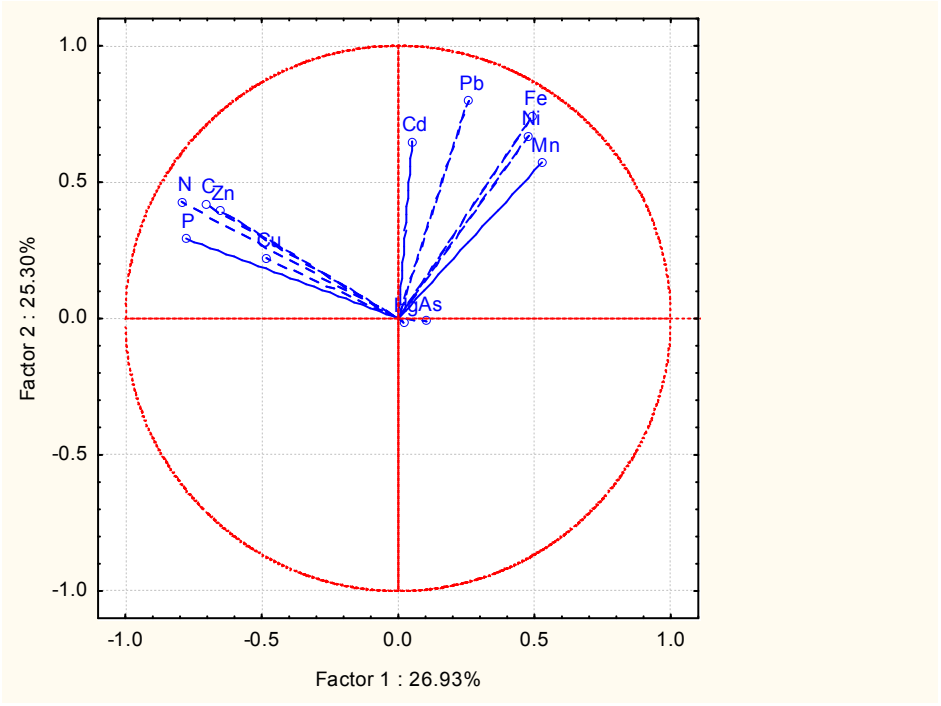


(β)

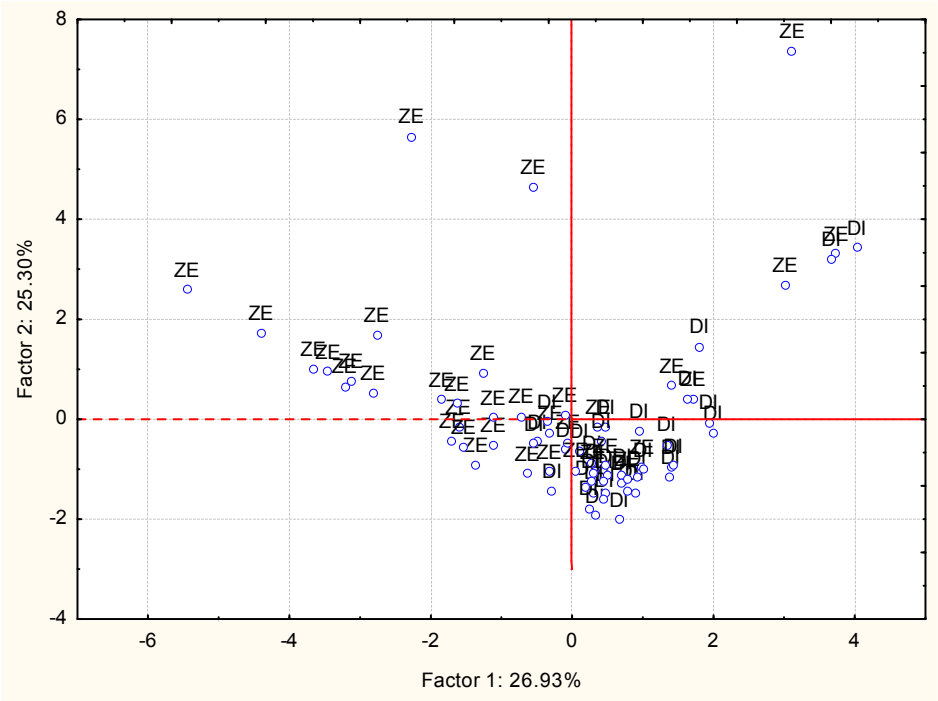
Σχήμα 2.11: Διαγράμματα φορτίσεων (α) και σκορ (β) της PCA για τα ΖΕ μόνο δείγματα.

Για την συνολική PCA, τρεις ομάδες μεταβλητών ταυτοποιήθηκαν (σχ. 2.12 (α)). Τα P, C, N, Zn και Cu αποτελούν την πρώτη ομάδα, υπεύθυνες για την πλειοψηφία των ΖΕ δειγμάτων (βλ. το αριστερό τμήμα στο σκορ διάγραμμα, σχ. 2.12(β)). Αυτά είναι τα ίδια στοιχεία που απέρριψαν την μηδενική υπόθεση Ho που αφορούσε την ισότητα των μέσων τιμών των δύο ομάδων (πίνακας 2.9). Η δεύτερη ομάδα μεταβλητών περιέχει τα

Cd, Pb, Fe, Ni και Mn υπεύθυνα για τον λιγότερο επιτυχή διαχωρισμό των DI και ZE δειγμάτων στον άξονα Y. Τα Hg και As δεν φαίνονται να συνεισφέρουν στην διαφοροποίηση των σημείων. Όλα τα παραπάνω σχόλια έχουν ήδη αναφερθεί και παραπάνω σαν πρώτες παρατηρήσεις (§ 2.3.1).



(α)



(β)

Σχήμα 2.12: Διαγράμματα φορτίσεων (α) και σκορ (β) της συνολικής PCA.

2.3.3. Ταυτοποίηση των πηγών ρύπανσης

Αρχικά, μόνο τα δείγματα ΖΕ χρησιμοποιήθηκαν για την εφαρμογή της FA. Οι μεταβλητές προς ανάλυση ήταν τα προσδιοριζόμενα στοιχεία, ενώ η εξαρτημένη μεταβλητή ήταν οι μονάδες ιχθυοκαλλιέργειας (CH, NA, AS). Η FA (Varimax normalized rotation) επέτυχε την μείωση των αρχικών μεταβλητών σε τέσσερις παράγοντες με ερμηνεία ποσοστού ίσου με το 74 % της αρχικής διακύμανσης. Οι αντίστοιχες φορτίσεις φαίνονται στον πίνακα 2.10.

Πίνακας 2.10: Φορτίσεις για τους τέσσερις επιλεγμένους παράγοντες (μόνο ΖΕ δείγματα χρησιμοποιούνται).

Μεταβλητή	**VF1**	**VF2**	**VF3**	**VF4**
Cu	0,048	0,116	***0,808***	-0,272
Cd	0,174	0,062	0,065	***0,910***
Pb	0,424	0,011	0,050	***0,877***
Hg	0,078	-0,034	-0,423	-0,204
As	0,156	0,150	-0,578	-0,322
Fe	***0,861***	0,009	-0,101	0,318
Mn	***0,876***	-0,107	0,080	0,084
Zn	0,123	***0,564***	***0,583***	-0,203
Ni	***0,835***	-0,140	-0,141	0,117
C	-0,015	***0,906***	-0,030	0,052
N	-0,052	***0,940***	0,045	0,012
P	-0,420	***0,735***	0,084	0,056
Ιδιοτιμές	3,49	2,57	1,40	1,36
% Ποσοστό διακύμανσης	29,1	50,5	62,2	73,5

Ο πρώτος παράγοντας VF1 (Varimax Factor) ερμήνευσε 29,1 % από την αρχική διακύμανση και περιλαμβάνει τα μέταλλα Fe, Mn, Ni που διαχωρίζουν την μονάδα CH από τις υπόλοιπες λόγω των χαμηλότερων τιμών τους. Ο VF2 ερμηνεύοντας ποσοστό 21,4 % της αρχικής διακύμανσης αντιπροσωπεύει τα θρεπτικά συστατικά C, N και P υπεύθυνα για την διαφοροποίηση της NA (οι τιμές τους ήταν χαμηλότερες σε σχέση με των CH και AS). Ο VF3 αντιπροσωπεύει τα μέταλλα Cu και Zn και ερμήνευσε 11,7 % της αρχικής διακύμανσης. Τελικά, ο VF4 περιλαμβάνει τα Cd και Pb. Οι δύο παράγοντες VF3 και VF4 χαρακτηρίζουν την AS μονάδα.

Πίνακας 2.11: Φορτίσεις για τους πέντε επιλεγμένους παράγοντες (χρησιμοποιούνται όλα τα δείγματα)

Variable	**VF1**	**VF2**	**VF3**	**VF4**	**VF5**
Cu	0,190	0,008	-0,050	***0,894***	0,008
Cd	0,123	0,084	***0,960***	0,004	0,007
Pb	0,056	0,398	***0,902***	-0,024	-0,030
Hg	-0,054	-0,046	0,042	0,050	***0,934***
As	0,152	0,214	-0,273	-0,387	0,414
Fe	0,000	***0,902***	0,242	-0,071	-0,029
Mn	-0,133	***0,899***	0,014	0,045	-0,115
Zn	0,550	0,061	-0,008	***0,674***	0,032
Ni	-0,050	***0,865***	0,169	0,006	0,190
C	***0,879***	-0,012	0,086	0,062	0,012
N	***0,937***	-0,037	0,062	0,134	0,033
P	***0,833***	-0,147	0,025	0,161	-0,050
Ιδιοτιμές	3,22	3,05	1,40	1,19	0,96
% Ποσοστό διακύμανσης	26,9	52,3	64,0	73,9	81,9

Στην συνολική FA, η εξαρτημένη μεταβλητή ήταν οι ομάδες DI και ZE. Η FA ερμήνευσε το 82 % της συνολικής διακύμανσης σε πέντε παράγοντες. Οι αντίστοιχες φορτίσεις φαίνονται στον πίνακα 2.11.

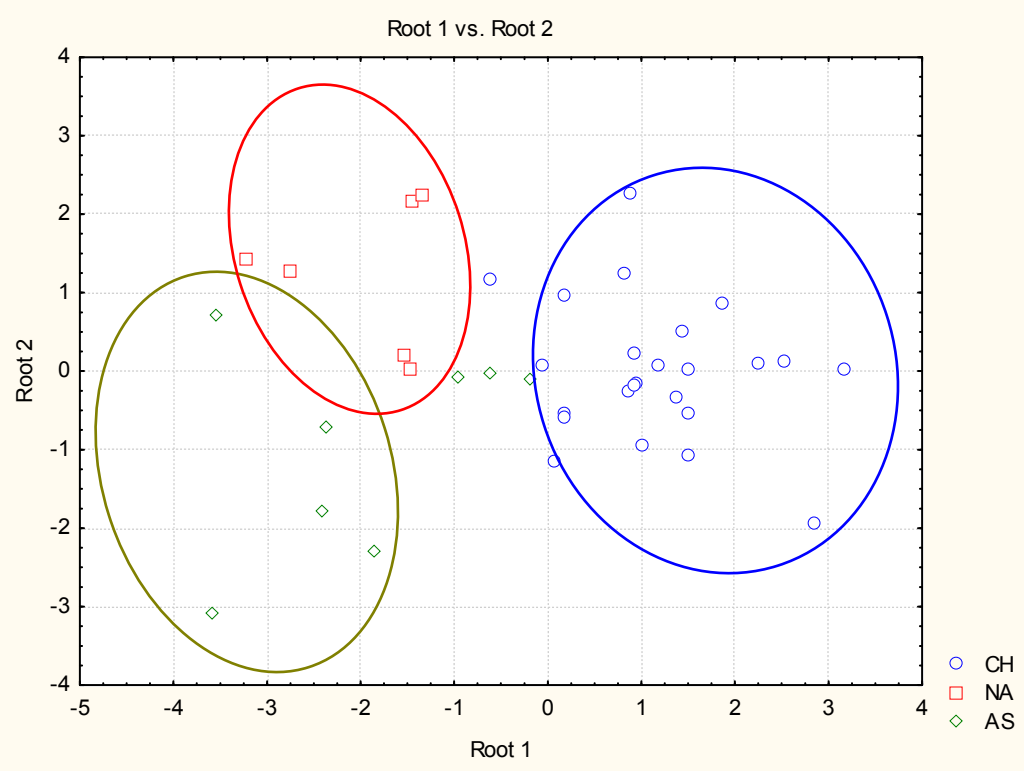
Οι φορτίσεις για τον VF1 δείχνουν ότι ο παράγοντας αυτός ερμηνεύοντας ποσοστό ίσο με 26,9 % της αρχικής διακύμανσης περιλαμβάνει τα C, N και P που προέρχονται από την δραστηριότητα της ιχθυοκαλλιέργειας (ιχθυοτροφές). Οι τιμές τους είναι πάντα υψηλότερες εξαιτίας των παρακείμενων ιχθυοκαλλιεργειών [3, 8, 9]. Ο VF2 ερμηνεύοντας ποσοστό 25,4 %, αντιπροσωπεύει τα μέταλλα Fe, Mn, Ni με φυσική προέλευση, εφόσον οι τιμές τους δεν παρουσιάζουν διαφορές στα δείγματα κοντά και μακρύτερα των κλωβών. Ο VF3 περιέχει τα Cd και Pb που οφείλουν την προέλευσή τους σε ανθρωπογενείς πηγές [22, 23], όπως φαίνεται από τις υψηλές τιμές τους στα ZE δείγματα (σχ. 2.5 (γ), 2.6 (β) και 2.6(γ)). Τα μέταλλα Cu και Zn συνεισέφεραν στο παράγοντα VF4 ερμηνεύοντας ποσοστό 9,9 % της διακύμανσης τα οποία προέρχονται από ανθρωπογενείς πηγές [8, 10, 11]. Συγκεκριμένα τα Cu και Zn χρησιμοποιούνται σαν αντισκουριακά στους κλωβούς [7, 14, 22, 23, 24, 25]. Βρίσκονται ακόμα στις ιχθυοτροφές σαν προσθετικά ή συντηρητικά [22] και στα περιττώματα των ψαριών [24, 25]. Τέλος, ο πέμπτος παράγοντας VF5 περιλαμβάνει μόνο τον Hg (ο οποίος βρέθηκε σε περίσσεια στην μονάδα CH) και λιγότερο το As και έχουν την προέλευσή τους σε φυσικές και ανθρωπογενείς πηγές. Οι τιμές και των δύο στοιχείων ήταν γενικά πολύ χαμηλές.

2.3.4. Εφαρμογή της Διαχωριστικής Ανάλυσης (DA)

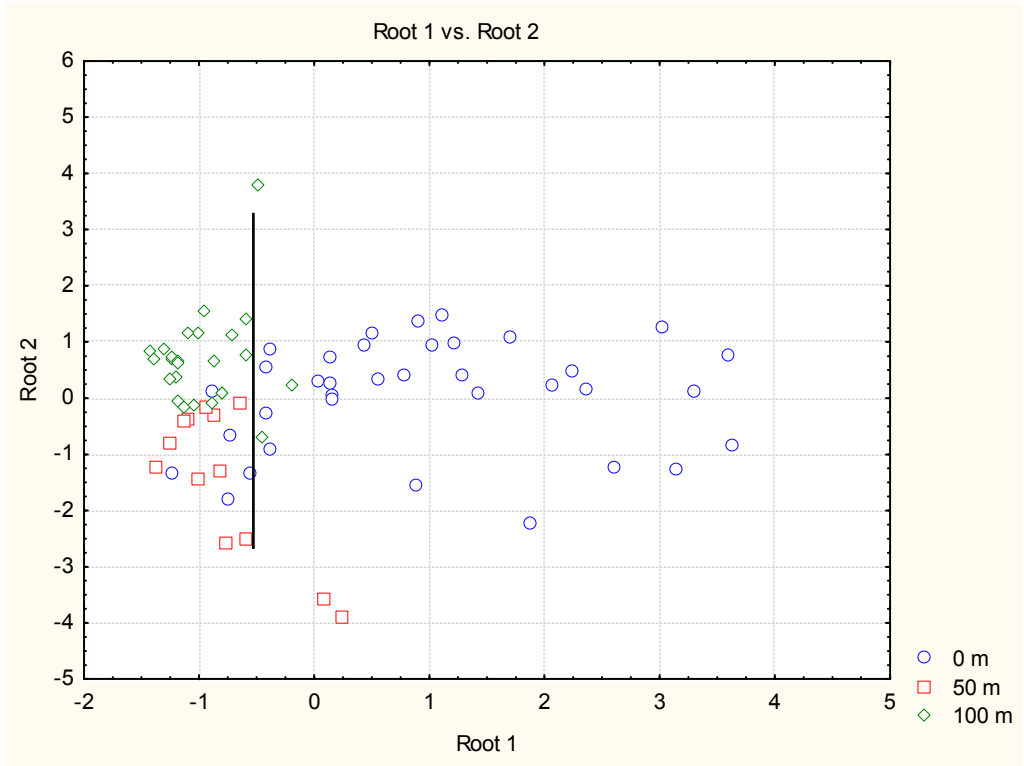
Η DA εφαρμόστηκε στα δεδομένα με την κλασική της προσέγγιση. Χρησιμοποιήθηκε στα DI δείγματα (38) αλλά και στο σύνολο αυτών (74).

Για την πρώτη ανάλυση, ερευνήθηκαν οι γηγενείς διαφορές των τριών μονάδων. Το canonical plot της ανάλυσης φαίνεται στο σχήμα 2.13. Ο διαχωρισμός είναι πολύ καλός εξαιτίας μερικών κρίσιμων μεταβλητών όπως οι Cd, As, Mn.

Το canonical plot της συνολικής ανάλυσης φαίνεται στο σχήμα 2.14. Τρεις ομάδες δειγμάτων ταυτοποιούνται ανάλογα με την απόστασή τους από τους κλωβούς (0, 50 και 100 m). Όλα τα απομακρυσμένα δείγματα (τετράγωνα και ρόμβοι) διαφοροποιήθηκαν εμφανώς.



Σχήμα 2.13: Canonical plot της DA μόνο για τα DI δείγματα.



Σχήμα 2.14: Canonical plot της συνολικής DA.

2.4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στο κεφάλαιο αυτό, μελετήθηκαν θαλάσσια ιζήματα από τρεις μονάδες ιχθυοκαλλιέργειας της χώρας. Τα αποτελέσματα έδειξαν ότι υπάρχει συσσώρευση των θρεπτικών συστατικών P, N, C στα δείγματα κοντά στους κλωβούς των ψαριών. Αυτό έρχεται σε συμφωνία με το γεγονός ότι τα απόβλητα από τις ιχθυοκαλλιέργειες αποτελούνται κυρίως από σωματίδια που προέρχονται από υπολείμματα τροφών και περιττώματα περιέχοντα οργανικές ενώσεις, άζωτο και φώσφορο.

Ο προσδιορισμός των μετάλλων και As στα ιζήματα έδειξε μια συσσώρευση Cu και Zn και λιγότερο των Cd και Pb. Οι κύριες πηγές των μετάλλων αυτές είναι οι ιχθυοτροφές και οι εργασίες για την αντισκουριακή προστασία των κλωβών. Η συσσώρευση των Fe, Ni, As, Hg και Mn στα δείγματα κάτω από τους κλωβούς ήταν αμελητέα, καθώς ήταν παρόμοιες οι συγκεντρώσεις τους σε όλα τα σημεία (μακριά ή κοντά στους κλωβούς).

Επιπλέον, μελετήθηκαν τα ιδιαίτερα χαρακτηριστικά της κάθε μονάδας. Τα C, N, Cd, Mn και Ni ήταν οι πιο κρίσιμες μεταβλητές για την διαφοροποίησή τους.

Οι πολυπαραμετρικές τεχνικές που χρησιμοποιήθηκαν συνήνεσαν στα παραπάνω συμπεράσματα. Έτσι, οι PCA και DA επικύρωσαν τις χωρικές διαφοροποιήσεις μεταξύ των σημείων κοντά και μακριά των κλωβών. Στην αξιολόγηση των μεταβλητών, η FA υπέδειξε τα P, C, N, Cu και Zn σαν τις πιο κρίσιμες μεταβλητές για την διαμόρφωση των ομάδων και υπεύθυνων για την ρύπανση των παρακείμενων σημείων. Είναι φανερό, ότι τεχνητά παρασκευασμένες τροφές, συμπληρώματα, εργασίες συντήρησης και επισκευών είναι υπεύθυνες για την επιμόλυνση των θαλάσσιων ιζημάτων.

Οι ίδιες τεχνικές εφαρμόστηκαν και για την διαφοροποίηση των μονάδων μεταξύ τους. Μόνο ΖΕ ή DI δείγματα χρησιμοποιήθηκαν. Στην πρώτη περίπτωση τα Cd, Pb, Fe, Mn, Ni, C, N, P ήταν οι κρίσιμες μεταβλητές, ενώ τα Cd, Mn, Ni, C, N διαφοροποίησαν τις μονάδες στη δεύτερη περίπτωση. Θαλάσσια ρεύματα και το γεωλογικό υπόβαθρο είναι οι παράγοντες που συνεισφέρουν σε αυτόν το διαχωρισμό.

ΚΕΦ. 3 ΠΟΛΥΠΑΡΑΜΕΤΡΙΚΕΣ ΤΕΧΝΙΚΕΣ ΣΤΗΝ ΜΕΛΕΤΗ ΤΗΣ ΕΠΙΔΡΑΣΗΣ ΤΩΝ ΙΧΘΥΟΚΑΛΛΙΕΡΓΕΙΩΝ ΣΤΑ ΘΑΛΑΣΣΙΑ ΙΖΗΜΑΤΑ

3.1. ΑΠΟΣΠΑΣΜΑΤΑ ΘΕΩΡΙΑΣ

3.1.1. ROC (Receiving Operating Characteristic) καμπύλες

Οι καμπύλες αυτές αποτελούν ένα **μη-παραμετρικό δείκτη σύγκρισης μοντέλων** [26] και διευκολύνουν στην επιλογή των καλύτερων από αυτά [27] και την απόρριψη των λιγότερων καλών, παρέχοντας άμεσα τα αποτελέσματα της πρόβλεψης σε μια δυαδική ταξινόμηση. Οι ROC καμπύλες πρωτοεμφανίστηκαν κατά τη διάρκεια του Β΄ Παγκοσμίου Πολέμου για την ανάλυση των σημάτων radar που χρησιμοποιούνταν στην ανίχνευση των εχθρικών αεροπλάνων [28]. Εδώ αντικατοπτρίζεται ίσως, η χρήση του πρώτου συνθετικού (receiving για τη λήψη σημάτων), στο ακρώνυμο της λέξης ROC. Οι καμπύλες εξάλλου, είναι επίσης γνωστές ως Relative Operating Characteristic curves, επειδή συγκρίνουν βασικά λειτουργικά χαρακτηριστικά (TP vs FN, βλ. παρακάτω).

Οι ROC καμπύλες συνοψίζουν την απόδοση ενός δικτύου ταξινόμησης δυο ομάδων, κατά την αλλαγή στις τιμές των κατωφλιών ταξινόμησης (classification thresholds). Ο άξονας Y απεικονίζει την ευαισθησία (sensitivity) του δικτύου, δηλαδή την αναλογία **των αληθώς θετικά ταξινομημένων σημείων (true positive, TP)** της δεύτερης ομάδας (σχ. 3.1). Ο άξονας X απεικονίζει συνολικά τη σχέση: 1-εξειδίκευση (1-specificity), δηλαδή την αναλογία των ψευδώς αρνητικά ταξινομημένων (false negative, FN) της πρώτης ομάδας ή **των ψευδώς θετικά ταξινομημένων σημείων (false positive, FP)** της δεύτερης ομάδας [29, 30, 31].

Με αλλαγή των κατωφλιών ταξινόμησης, κατασκευάζονται δίκτυα με εναλλαγές στην αναλογία των ψευδώς θετικά ταξινομημένων και των ψευδώς αρνητικά ταξινομημένων. Με ένα κατώφλι ταξινόμησης ίσο με το 0, όλα τα σημεία θα κατατάσσονταν στη δεύτερη ομάδα και έτσι θα αποδίδονταν τα περισσότερα εσφαλμένα θετικά (για τη δεύτερη ομάδα) και καθόλου εσφαλμένα αρνητικά (για την πρώτη ομάδα). Το αντίθετο θα συνέβαινε με κατώφλι ταξινόμησης ίσο με το 1. Καθώς το κατώφλι ταξινόμησης αυξάνεται, η καμπύλη "μετακινείται" από αριστερά προς τα δεξιά.

Όσο μεγαλύτερη είναι η περιοχή κάτω από την καμπύλη (AUC, Area Under Curve), τόσο μεγαλύτερη είναι η ικανότητα ταξινόμησης του μοντέλου που αντιπροσωπεύει [32], καθώς η AUC ερμηνεύεται ως την πιθανότητα της σωστής ταξινόμησης [33]. Μια ROC καμπύλη ιδανικού δικτύου προσεγγίζει την πάνω αριστερή καμπύλη του

διαγράμματος και η AUC που καλύπτει είναι 1,0. Το σημείο (0,1) (σχ. 7.15) ονομάζεται **το τέλειο σημείο ταξινόμησης** (perfect classification, [28]), και αντιπροσωπεύει 100 % ευαισθησία (μηδενικά ψευδώς αρνητικά) και 100 % εξειδίκευση (μηδενικά ψευδώς θετικά). Αυτό σημαίνει ότι το μοντέλο θα έχει 100 % επιτυχία στη εύρεση της σωστής ομάδας [33]. Η μαθηματική σχέση που περιγράφει τα παραπάνω (δηλαδή τα ποιοτικά χαρακτηριστικά ενός μοντέλου), δίνεται από το γεωμετρικό μέσο όρο G των ποσοστών της ευαισθησίας και εξειδίκευσης [32]:

$$G = \sqrt{\%\ \text{ευαισθησία} \times \%\ \text{εξειδίκευση}}$$

Ο Rao et al. [34] αντί αυτού ωστόσο, προτιμά τους δείκτες συνολικής ακρίβειας (overall prediction accuracy, Q) και το συντελεστή Mattheews (Mattheews coefficient, C) που δίνονται αντίστοιχα από τις σχέσεις:

$$Q = \frac{TP + TN}{TP + FN + TN + FP}$$

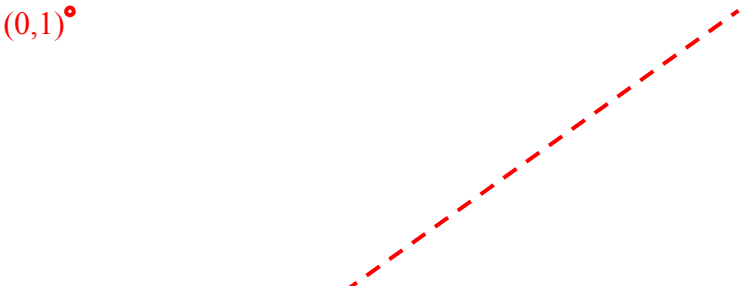
$$C = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}$$

Η ιδανική ROC καμπύλη επιτυγχάνεται μόνο για τις πολύ καλά διαφοροποιημένες ομάδες, μερικές φορές, απλά θεωρείται σημαντικότερο να αποφθεχθούν λάθη στη μια ομάδα, από όσο στη δεύτερη. Αυτό συμβαίνει για παράδειγμα, σε ιατρικές εφαρμογές.

Ένα μοντέλο με AUC 0,7 για την ROC καμπύλη θεωρείται ότι επιτυγχάνει ικανοποιητικό διαχωρισμό των ομάδων, με AUC 0,8 καλό διαχωρισμό και με AUC 0,9 πολύ καλό [35]. Ένα πείραμα τυχαίας πρόβλεψης (random classifier [29]) δίνει AUC περίπου 0,5. Ένα τέτοιο σημείο (random guess [28], που θα λαμβάναμε για παράδειγμα με το γνωστό παιγνίδι του κορώνα-γράμματα σε ένα νόμισμα), θα βρισκόταν πάνω στη διαγώνια κόκκινη γραμμή (line of no-discrimination, σχ. 3.1) που ενώνει την κάτω αριστερή γωνία του διαγράμματος με την πάνω δεξιά. Η διαγώνια αυτή χωρίζει το χώρο της καμπύλης ROC, σε περιοχές καλής και κακής ταξινόμησης. Σημεία πάνω από τη γραμμή δηλώνουν ορθά αποτελέσματα, ενώ σημεία κάτω από αυτή, εσφαλμένα. Εδώ επισημαίνεται, ότι με την "αντιστροφή" της μεθόδου πρόβλεψης (δηλαδή αντιστροφή των αποφάσεων της μεθόδου, όλα τα εσφαλμένα αποτελέσματα μετατρέπονται σε ορθά).

Σφάλμα! Τα αντικείμενα δεν μπορούν να δημιουργηθούν από την επεξεργασία κωδικών πεδίων.

Σχήμα 3.1: ROC καμπύλη με AUC = 0,95.



3.2. ΣΥΜΠΛΗΡΩΜΑ ΣΤΟ ΠΕΙΡΑΜΑΤΙΚΟ ΜΕΡΟΣ

Πίνακας 3.1: Στοιχεία της κατασκευής του δέντρου (μέθοδος CART).

Κόμβοι	**Αριστ. κλάδος**	**Δεξιός κλάδος**	**DI**	**ZE**	**Πρόβλεψη**	**Σταθερά διαχ/μου**	**Μεταβλητή διαχ/μου**
1	2	3	31	29	DI	-1,64	**P**
2	4	5	29	4	DI	-11,2	**Pb**
3*			2	25	ZE		
4	6	7	29	2	DI	-307,5	**Mn**
5*			0	2	ZE		
6*			23	0	DI		
7	8	9	6	2	DI	-148,8	**Cd**
8*			5	0	DI		
9*			1	2	ZE		

* Τερματικοί κόμβοι

ΚΕΦ. 4 ΤΑΞΙΝΟΜΗΣΗ ΕΛΑΙΟΛΑΔΩΝ ΜΕ ΒΑΣΗ ΤΗΝ ΓΕΩΓΡΑΦΙΚΗ ΤΟΥΣ ΠΡΟΕΛΕΥΣΗ

4.1. ΣΥΜΠΛΗΡΩΜΑ ΣΤΟ ΠΕΙΡΑΜΑΤΙΚΟ ΜΕΡΟΣ

Πίνακας 4.1: Συγκεντρωτικά αποτελέσματα για την περιεκτικότητα ελαιολάδων από έντεκα (11) περιοχές της Ελλάδας σε 14 σπάνιες γαίες.

Προέλευση	**Κωδικο ποίηση**	**Y**	**La**	**Ce**	**Pr**	**Nd**	**Sm**	**Gd**	**Tb**	**Dy**	**Ho**	**Er**	**Tm**	**Yb**	**Th**
Αρκαδία	---	318	<475	**<849**	135	**232**	211	103	**<29,6**	124	116	118	94,1	159	**132**
Εύβοια	---	297	<475	**<849**	126	**242**	197	117	**<29,6**	129	117	123	96,6	126	**155**
Ζάκυνθος	ΖΑ	344	797	**<849**	140	**309**	189	122	**<29,6**	122	110	112	89,0	129	**148**
Ζάκυνθος	ΖΑ	220	<475	**<849**	109	**<200**	201	108	**2490**	112	103	110	87,8	118	**122**
Ζάκυνθος	ΖΑ	323	834	**1243**	197	**469**	230	138	**<29,6**	129	104	110	86,9	127	**213**
Ζάκυνθος	ΖΑ	238	<475	**<849**	124	**235**	177	115	**<29,6**	113	106	114	93,7	125	**142**
Ζάκυνθος	ΖΑ	356	<475	**<849**	132	**265**	179	130	**<29,6**	140	118	124	95,4	136	**143**
Ζάκυνθος	ΖΑ	198	<475	**<849**	114	**233**	153	102	**<29,6**	104	100	96,6	88,0	118	**127**

Προέλευση	Κωδικο ποίηση	Y	La	Ce	Pr	Nd	Sm	Gd	Tb	Dy	Ho	Er	Tm	Yb	Th
Ζάκυνθος	ZA	580	706	**895**	206	**510**	305	160	**73,3**	276	142	97,7	72,5	134	**<111**
Ζάκυνθος	ZA	376	<475	**<849**	159	**<200**	198	76,2	**53,6**	152	108	103	47,6	70,8	**<111**
Ζάκυνθος	ZA	374	615	**915**	193	**420**	230	87,9	**47,0**	160	95,5	134	42,4	59,7	**119**
Ζάκυνθος	ZA	364	609	**<849**	228	**426**	184	114	**49,9**	179	96,5	129	42,6	55,8	**124**
Ζάκυνθος	ZA	260	<475	**<849**	135	**<200**	200	<64,5	**46,1**	152	78,7	98,8	38,9	57,6	**<111**
Ζάκυνθος	ZA	325	<475	**<849**	<84,2	**<200**	185	103	**47,8**	172	89,9	82,2	45,9	88,8	**<111**
Ζάκυνθος	ZA	385	544	**<849**	177	**<200**	253	101	**51,4**	145	103	108	47,9	79,2	**121**
Ζάκυνθος	ZA	322	<475	**<849**	146	**<200**	181	73,3	**36,8**	163	105	109	46,9	58,7	**<111**
Ζάκυνθος	ZA	352	525	**<849**	171	**<200**	257	118	**56,8**	167	99,8	110	44,1	59,8	**<111**
Ζάκυνθος	ZA	219	<475	**<849**	120	**<200**	175	64,7	**33,4**	123	98,4	81,8	36,0	64,2	**<111**
Ζάκυνθος	ZA	341	<475	**<849**	144	**<200**	172	82,3	**50,4**	176	97,9	127	42,8	49,9	**<111**
Ζάκυνθος	ZA	236	<475	**<849**	<84,2	**<200**	179	<64,5	**29,6**	155	91,1	77,8	50,1	49,4	**<111**
Ζάκυνθος	ZA	394	654	**923**	193	**468**	224	94,6	**50,8**	149	90,9	97,1	46,9	76,2	**142**

Προέλευση	Κωδικο ποίηση	Y	La	Ce	Pr	Nd	Sm	Gd	Tb	Dy	Ho	Er	Tm	Yb	Th
Ζάκυνθος	ZA	340	541	**<849**	152	**<200**	200	88,3	**30,6**	162	68,2	110	28,9	49,6	**<111**
Ζάκυνθος	ZA	377	570	**<849**	181	**<200**	216	94,8	**44,5**	169	89,4	125	48,8	71,3	**128**
Ζάκυνθος	ZA	420	681	**1103**	209	**491**	257	138	**64,6**	193	93,9	125	39,5	86,3	**151**
Ζάκυνθος	ZA	281	<475	**<849**	157	**<200**	181	69,3	**43,6**	128	59,6	102	47,7	36,0	**130**
Ζάκυνθος	ZA	385	2579	**5934**	343	**<200**	228	162	**52,0**	180	106	124	45,0	82,6	**149**
Ηράκλειο	I	274	<475	**<849**	125	**243**	201	117	**339**	129	109	122	91,0	127	**129**
Ηράκλειο	I	277	<475	**<849**	131	**229**	186	110	**<29,6**	116	111	113	96,8	126	**138**
Ηράκλειο	I	310	<475	**<849**	130	**218**	196	108	**<29,6**	116	107	120	92,0	131	**140**
Ηράκλειο	I	281	<475	**<849**	170	**272**	237	117	**<29,6**	152	132	110,	97,3	137	**130**
Ηράκλειο	I	579	886	**1335**	221	**676**	307	170	**52,3**	207	85,1	120	54,3	101	**212**
Ηράκλειο	I	307	<475	**<849**	158	**<200**	210	90,2	**41,4**	146	82,8	84,7	37,0	61,9	**117**
Ηράκλειο	I	579	669	**1075**	192	**578**	294	150	**63,6**	216	105	134,	53,8	86,7	**201**
Ηράκλειο	I	554	805	**1202**	235	**651**	234	139	**45,9**	236	83,6	154	40,2	108	**204**

Προέλευση	**Κωδικοποίηση**	**Y**	**La**	**Ce**	**Pr**	**Nd**	**Sm**	**Gd**	**Tb**	**Dy**	**Ho**	**Er**	**Tm**	**Yb**	**Th**
Ηράκλειο	I	410	517	**<849**	172	**<200**	212	111	**56,9**	176	106	137	54,0	74,1	**128**
Ηράκλειο	I	464	752	**1252**	199	**461**	240	135	**51,9**	182	97,0	134	37,4	82,8	**145**
Ηράκλειο	I	483	760	**<849**	227	**569**	268	165	**61,0**	225	90,1	134	59,1	85,0	**243**
Ηράκλειο	I	261	<475	**<849**	145	**<200**	190	83,7	**37,0**	144	103	110	43,4	57,1	**<111**
Ηράκλειο	I	349	486	**<849**	135	**<200**	189	82,6	**48,0**	154	105	95,8	45,2	52,2	**<111**
Ηράκλειο	I	348	1108	**1161**	213	**470**	191	101	**37,4**	151	82,6	95,0	39,1	60,6	**<111**
Ηράκλειο	I	388	530	**<849**	174	**<200**	209	92,4	**43,2**	165	92,7	109	43,7	71,4	**<111**
Ηράκλειο	I	389	550	**<849**	183	**<200**	254	123	**50,8**	168	93,8	131	42,0	84,9	**130**
Ηράκλειο	I	355	528	**<849**	158	**<200**	204	97,1	**56,9**	159	102	104	48,5	79,5	**<111**
Ηράκλειο	I	379	531	**<849**	150	**<200**	215	98,4	**54,4**	164	89,3	118	47,0	64,6	**117**
Ηράκλειο	I	430	830	**1265**	235	**465**	272	128	**47,1**	214	75,0	132	37,5	49,2	**146**
Ηράκλειο	I	285	<475	**<849**	142	**<200**	207	86,2	**56,8**	142	85,9	121	50,2	69,5	**<111**
Ηράκλειο	I	215	507	**<849**	142	**<200**	188	100	**<29,6**	153	86,6	92,0	31,3	49,3	**<111**

Προέλευση	Κωδικοποίηση	Y	La	Ce	Pr	Nd	Sm	Gd	Tb	Dy	Ho	Er	Tm	Yb	Th
Ηράκλειο	I	310	<475	**<849**	141	**<200**	183	72,6	**34,9**	148	79,9	80,7	35,8	66,1	**<111**
Λακωνία	La	462	1718	**3018**	401	**1224**	348	271	**<29,6**	157	104	127	90,0	141	**264**
Λακωνία	LA	967	2399	**3888**	523	**1671**	436	346	**<29,6**	244	122	176	97,3	176	**857**
Λακωνία	LA	248	<475	**<849**	134	**238**	220	130	**<29,6**	126	116	122	100	141	**<111**
Λακωνία	LA	287	<475	**<849**	<84,2	**<200**	254	90,7	**102**	125	101	81,2	36,7	47,0	**<111**
Λακωνία	LA	201	<475	**<849**	135	**<200**	201	81,4	**111**	140	88,1	101	38,1	61,1	**<111**
Λακωνία	LA	481	632	**925**	234	**481**	310	170	**107**	214	88,9	131	41,4	70,8	**161**
Λακωνία	LA	235	<475	**<849**	<84,2	**<200**	167	108	**78,4**	132	111	83,0	30,2	47,0	**<111**
Λακωνία	LA	272	<475	**<849**	<84,2	**<200**	190	110	**91,5**	123	86,5	104	46,1	43,1	**<111**
Λακωνία	LA	373	<475	**<849**	163	**<200**	262	134	**119**	174	112	110	40,3	81,1	**117**
Λακωνία	LA	316	<475	**<849**	<84,2	**<200**	220	113	**98,0**	162	93,8	122	28,0	75,9	**<111**
Λακωνία	LA	314	<475	**<849**	<84,2	**<200**	252	99,7	**106**	153	111	112	40,3	70,4	**<111**
Λακωνία	LA	406	475	**<849**	179	**465**	255	164	**98,9**	163	80,5	118	35,5	70,3	**206**

Προέλευση	Κωδικοποίηση	Y	La	Ce	Pr	Nd	Sm	Gd	Tb	Dy	Ho	Er	Tm	Yb	Th
Λακωνία	LA	420	570	**907**	204	**522**	293	152	**102**	161	108	119	38,5	91,7	**165**
Λακωνία	LA	513	609	**1025**	229	**580**	231	183	**106**	191	109	118	37,0	94,8	**205**
Λακωνία	LA	391	<475	**<849**	173	**<200**	245	162	**105**	145	77,3	120	43,3	66,8	**168**
Λακωνία	LA	444	518	**<849**	172	**441**	241	140	**95,2**	191	93,4	133	25,5	78,7	**163**
Λακωνία	LA	428	<475	**<849**	216	**462**	287	162	**99,6**	173	118	143	40,2	121	**279**
Λέσβος	---	358	661	**<849**	159	**247**	208	123	**249**	138	128	121	93,3	129	**147**
Λέσβος	---	288	<475	**<849**	127	**265**	154	115	**<29,6**	105	95,0	103	87,5	120	**143**
Λέσβος	---	780	2067	**6455**	338	**848**	284	272	**<29,6**	167	112	126	89,9	138	**581**
Μεσσηνία	ΜΕ	282	648	**<849**	127	**225**	182	124	**144**	111	98,4	123	96,9	128	**129**
Μεσσηνία	ΜΕ	312	683	**<849**	159	**346**	199	137	**<29,6**	107	104	121	89,3	122	**157**
Μεσσηνία	ΜΕ	660	2433	**2518**	398	**945**	313	246	**<29,6**	186	112	141	96,3	155	**176**
Μεσσηνία	ΜΕ	760	835	**<849**	177	**341**	196	140	**<29,6**	147	115	136	95,8	131	**240**

Προέλευση	Κωδικοποίηση	Y	La	Ce	Pr	Nd	Sm	Gd	Tb	Dy	Ho	Er	Tm	Yb	Th
Μεσσηνία	ΜΕ	350	930	**1995**	193	**467**	218	143	**<29,6**	130	102	112	83,7	118	**294**
Μεσσηνία	ΜΕ	345	719	**<849**	167	**297**	188	129	**<29,6**	125	112	131	93,4	134	**147**
Μεσσηνία	ΜΕ	275	5311	**1811**	241	**492**	190	124	**<29,6**	112	103	108	90,7	130	**142**
Μεσσηνία	ΜΕ	1489	2632	**3501**	463	**1248**	450	320	**<29,6**	261	187	244	118	245	**400**
Μεσσηνία	ΜΕ	356	<475	**<849**	164	**<200**	188	110	**50,6**	155	82,1	109	42,0	42,1	**<111**
Μεσσηνία	ΜΕ	355	549	**<849**	153	**<200**	194	124	**40,3**	165	106	93,1	33,3	56,7	**114**
Μεσσηνία	ΜΕ	320	479	**<849**	154	**<200**	184	106	**36,1**	142	93,3	97,7	31,2	54,3	**<111**
Μεσσηνία	ΜΕ	367	<475	**<849**	148	**<200**	173	136	**50,0**	171	101	104	39,7	59,1	**<111**
Μεσσηνία	ΜΕ	338	676	**<849**	164	**<200**	194	142	**38,8**	155	82,7	102	32,7	52,6	**142**
Μεσσηνία	ΜΕ	367	670	**1015**	179	**<200**	185	121	**41,4**	150	88,0	132	45,2	73,6	**131**
Μεσσηνία	ΜΕ	290	<475	**<849**	<84,2	**<200**	171	96,6	**47,5**	126	94,2	96,9	37,3	69,1	**<111**
Μεσσηνία	ΜΕ	406	754	**924**	187	**440**	231	130	**54,2**	166	102	102	40,6	63,4	**140**
Μεσσηνία	ΜΕ	348	<475	**<849**	146	**<200**	187	121	**53,9**	150	94,9	94,1	40,0	55,9	**<111**

Προέλευση	Κωδικός ποίησης	Y	La	Ce	Pr	Nd	Sm	Gd	Tb	Dy	Ho	Er	Tm	Yb	Th
Μεσσηνία	ΜΕ	292	<475	**<849**	142	**<200**	159	108	**44,2**	155	98,9	109	46,1	69,8	**<111**
Μεσσηνία	ΜΕ	308	560	**<849**	171	**<200**	185	83,3	**37,4**	128	82,3	106	44,8	42,5	**<111**
Μεσσηνία	ΜΕ	303	<475	**<849**	135	**<200**	173	99,0	**38,4**	153	82,0	98,9	38,7	72,2	**<111**
Μεσσηνία	ΜΕ	294	<475	**<849**	147	**<200**	175	99,2	**51,4**	137	101	95,9	40,4	45,3	**<111**
Μεσσηνία	ΜΕ	257	<475	**<849**	<84,2	**<200**	154	77,6	**34,4**	116	69,5	97,9	34,8	46,1	**<111**
Μεσσηνία	ΜΕ	334	632	**1018**	185	**487**	190	136	**<29,6**	130	91,9	98,8	36,2	<35,5	**<111**
Μεσσηνία	ΜΕ	358	711	**1117**	171	**<200**	178	114	**30,7**	167	73,5	77,0	29,5	61,6	**<111**
Μεσσηνία	ΜΕ	380	558	**<849**	191	**<200**	198	108	**<29,6**	163	68,2	84,4	33,9	47,2	**139**
Μεσσηνία	ΜΕ	368	498	**<849**	143	**<200**	159	93,8	**<29,6**	114	79,7	90,6	34,4	36,7	**114**
Μεσσηνία	ΜΕ	302	547	**<849**	187	**<200**	172	75,5	**31,5**	129	54,8	77,9	28,5	39,6	**<111**
Μεσσηνία	ΜΕ	480	723	**1237**	220	**501**	224	151	**40,9**	167	75,3	83,2	31,3	45,4	**<111**
Μεσσηνία	ΜΕ	364	1188	**1764**	250	**665**	212	149	**30,2**	161	69,1	73,5	31,9	44,9	**133**
Μεσσηνία	ΜΕ	390	762	**1043**	194	**416**	198	146	**34,8**	154	62,9	92,5	27,7	52,1	**136**

Προέλευση	Κωδικός ποίησης	Y	La	Ce	Pr	Nd	Sm	Gd	Tb	Dy	Ho	Er	Tm	Yb	Th
Μεσσηνία	ΜΕ	306	552	**<849**	153	**<200**	200	82,1	**40,7**	143	90,2	73,4	38,7	<35,5	**117**
Μεσσηνία	ΜΕ	440	2986	**3552**	405	**1074**	247	196	**<29,6**	157	82,8	102	33,8	52,9	**133**
Μεσσηνία	ΜΕ	265	<475	**<849**	<84,2	**<200**	135	66,3	**<29,6**	104	89,4	63,9	22,0	<35,5	**<111**
Μεσσηνία	ΜΕ	479	1081	**1361**	247	**620**	201	128	**36,9**	144	62,7	84,6	26,7	61,2	**166**
Πιερία	---	421	911	**1461**	214	**434**	258	178	**<29,6**	157	112	139	91,9	142	**240**
Πιερία	---	398	1587	**3640**	300	**856**	257	228	**<29,6**	156	110	127	92,4	144	**539**
Ρέθυμνο	---	326	<475	**940**	180	**485**	282	151	**98,3**	163	99,1	80,3	23,9	48,8	**<111**
Ρέθυμνο	---	623	643	**<849**	216	**563**	362	203	**135**	300	165	112	76,9	143	**123**
Ρέθυμνο	---	374	507	**<849**	187	**435**	264	117	**104**	159	106	145	40,0	59,8	**129,8**
Ρέθυμνο	---	256	<475	**<849**	<84,2	**<200**	231	82,9	**102**	150	87,9	106	36,1	57,3	**<111**
Ρέθυμνο	---	326	<475	**<849**	<84,2	**<200**	218	133	**105**	173	101	89,2	44,0	89,2	**<111**
Ρέθυμνο	---	388	<475	**<849**	172	**<200**	289	131	**109**	144	115	117	46,2	79,6	**132**

Προέλευση	Κωδικοποίηση	Y	La	Ce	Pr	Nd	Sm	Gd	Tb	Dy	Ho	Er	Tm	Yb	Th
Ρέθυμνο	---	353	<475	**<849**	165	**<200**	294	150	**114**	168	111	119	42,1	60,1	**119**
Ρέθυμνο	---	218	<475	**<849**	<84,2	**<200**	208	93,3	**89,5**	122	110	89,3	33,7	64,9	**<111**
Ρέθυμνο	---	342	<475	**<849**	154	**417**	288	168	**102**	170	119	137	46,9	64,4	**117**
Χαλκιδική	---	317	4751	**<849**	140	**287**	182	125	**<29,6**	110	124	117	91,8	137	**144**
Χαλκιδική	---	248	<475	**<849**	127	**231**	168	100	**<29,6**	111	112	110	89,8	125	**117**
Χαλκιδική	---	359	611	**<849**	136	**248**	202	117	**<29,6**	134	115	109	90,0	137	**145**
Χαλκιδική	---	571	876	**<849**	205	**333**	252	147	**<29,6**	154	147	152	102	154	**172**
Χανιά	---	265	598	**<849**	150	**303**	190	140	**<29,6**	119	124	141	97,1	138	**128**
Χανιά	---	259	<475	**<849**	135	**254**	157	109	**<29,6**	113	98,5	102	86,1	116	**132**
Χανιά	---	313	624	**<849**	150	**305**	180	110	**<29,6**	124	110	115	96,3	129	**168**
Χανιά	---	265	<475	**<849**	126	**219**	197	109	**<29,6**	116	102	110	87,1	125	**128**
Χανιά	---	252	<475	**<849**	121	**<200**	159	107	**<29,6**	129	116	116	90,3	121	**<111**

Προέλευση	**Κωδικο ποίηση**	**Y**	**La**	**Ce**	**Pr**	**Nd**	**Sm**	**Gd**	**Tb**	**Dy**	**Ho**	**Er**	**Tm**	**Yb**	**Th**
Χανιά	---	257	<475	**<849**	119	**224**	165	113	**<29,6**	118	103	112	93,3	123	**129**
Χανιά	---	342	<475	**<849**	135	**245**	215	102	**<29,6**	137	125	113	98,7	131	**137**
Χανιά	---	712	698	**<849**	161	**308**	179	95,8	**<29,6**	136	114	124	95,7	126	**138**
Μέσος όρος		374	663	**822**	171	**345**	218	125	**70,3**	153	99,9	112	57,4	87,5	**132**
Διάμεση τιμή		348	518	**436**	159	**229**	201	117	**41,4**	152	101	110	45,2	74,1	**128**
Εύρος		198-1489	238-5310	**424-6455**	63.0-523	**100-1671**	135-450	32,3-346	**14,8-2490**	104-300	54,8-187	63,9-244	22,0-118	17,8-245	**55,7-857**

Τα αποτελέσματα δίνονται σε ng/kg.

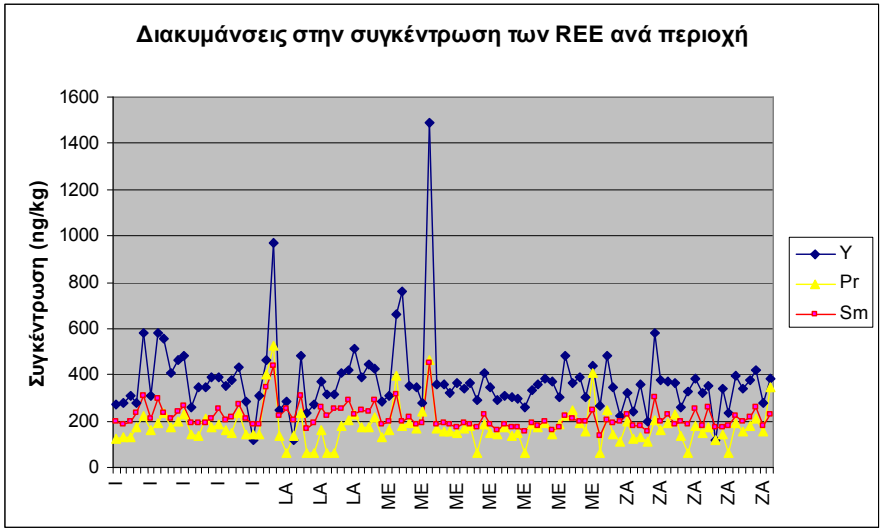
Πίνακας 4.2: Παράμετροι λειτουργίας του ICP-MS για τον προσδιορισμό REE σε δείγματα ελαιολάδου.

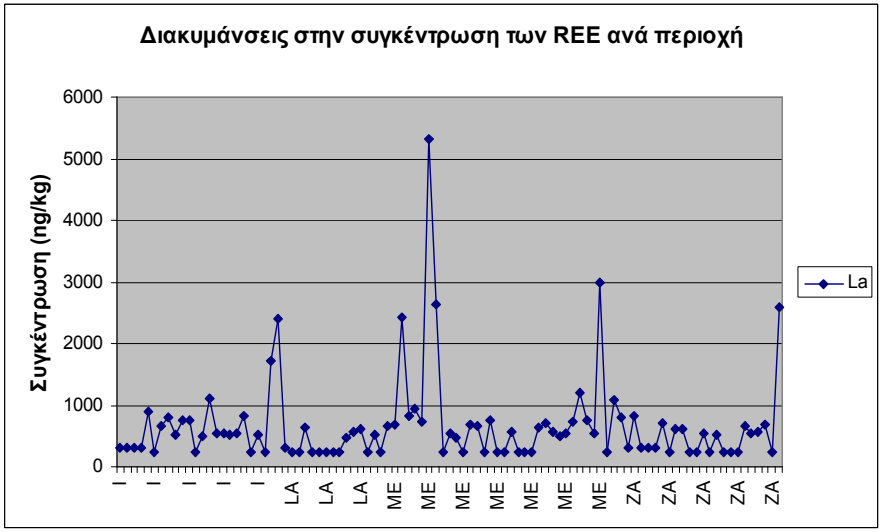
RF Power (W)	950
Nebulizer (carrier gas) flow rate (L min^{-1})	0,78
Lens Voltage (V)	5,75
Analog stage voltage (V)	-1900
Pulse stage voltage (V)	950
Discrimination threshold (V)	17
AC Rod Offset (V)	-2
Resolution (amu)	0,7
Detector	Dual
Speed of peristaltic pump (rpm)	24
Sweeps/ reading	3
Dwell time (ns)	60

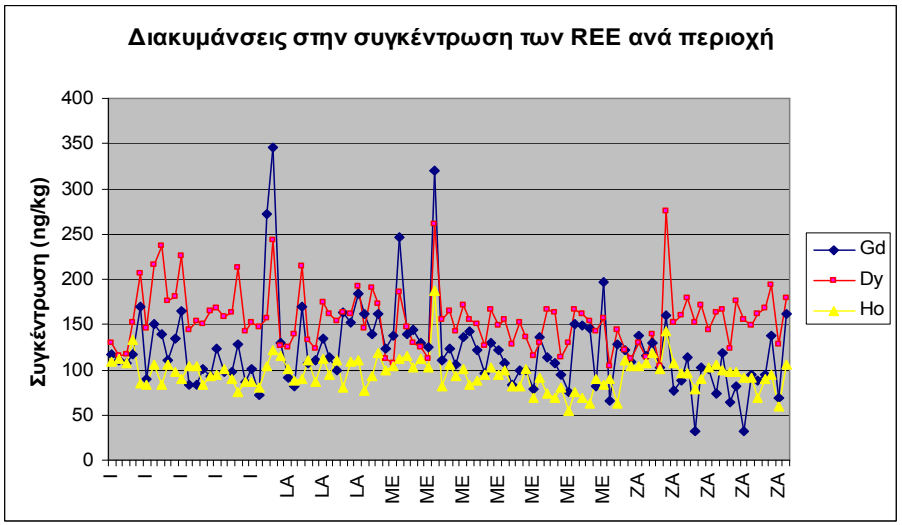
Πίνακας 4.3: Δεδομένα ποιοτικού ελέγχου και επικύρωσης για τον προσδιορισμό REE σε δείγματα ελαιολάδου.

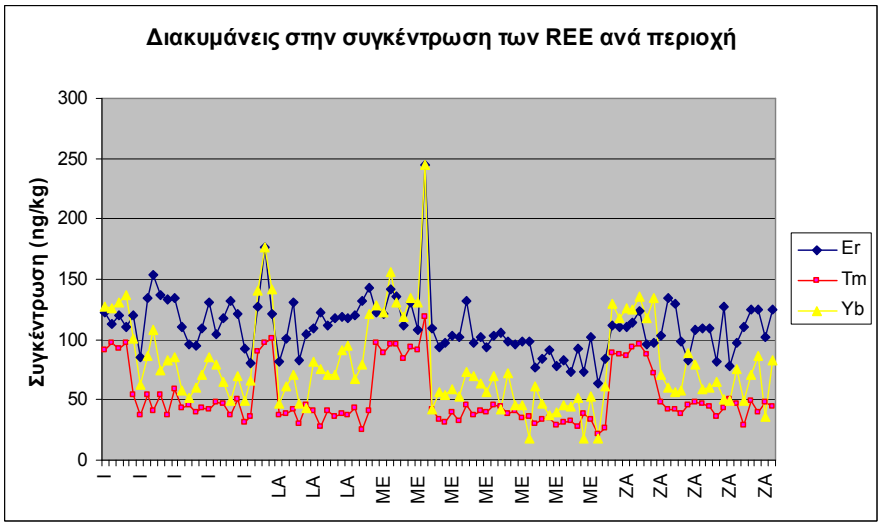
	m/z	**LOD (ng/kg)**	**Ανακτησιμότητα QC* (%) (250 ng/kg) (% RSD, n=3)**	**Ανακτησιμότητα QC* (%) (500 ng/kg) (% RSD, n=3)**
Y	89	228,8	88 (5,2)	126 (2,4)
La	139	475,2	97 (6,4)	125 (11)
Ce	140	872,2	99 (14)	113 (10)
Pr	141	125,9	93 (10)	100 (15)
Nd	142	415,1	98 (9,3)	108 (5,6)
Sm	152	75,4	92 (5,4)	102 (9,3)
Gd	158	64,6	108 (9,5)	106 (9,6)
Tb	159	29,6	87 (5,5)	97 (7,8)
Dy	164	81,8	96 (8,9)	106 (7,6)
Ho	165	10,5	83 (4,1)	96 (5,9)
Er	166	31,3	123 (3,7)	117 (8,5)
Tm	169	14,6	92 (6,4)	100 (7,3)
Yb	174	35,5	101 (12)	102 (13)
Th	232	111	95 (10)	112 (5,5)

*QC: Δείγμα ποιοτικού ελέγχου









Σχήμα 4.1: Διαγράμματα των μεταβλητών (λανθανίδες) ανά περιοχή (Ι: Ηράκλειο, LA: Λακωνία, ΜΕ: Μεσσηνία, ΖΑ: Ζάκυνθος)

Πίνακας 4.4: Roots 1, 2, 3 όπως προκύπτουν μετά από DA κλασική ανάλυση

Προέλευση	**Κωδικοποίηση**	**Root 1**	**Root 2**	**Root 3**
Ηράκλειο	I	-0,026	-1,224	0,421
Ηράκλειο	I	-1,000	-0,665	0,995
Ηράκλειο	I	-0,598	-0,869	0,643
Ηράκλειο	I	-0,585	-2,959	2,217
Ηράκλειο	I	1,288	-0,735	-1,782
Ηράκλειο	I	-0,340	-0,487	0,653
Ηράκλειο	I	0,830	-1,345	-1,500
Ηράκλειο	I	-1,404	-2,059	-3,131
Ηράκλειο	I	-0,794	-1,064	-0,722
Ηράκλειο	I	0,813	-0,387	-0,824
Ηράκλειο	I	-0,083	-1,572	-2,471
Ηράκλειο	I	-0,379	-0,969	0,851
Ηράκλειο	I	-1,417	-0,012	0,795
Ηράκλειο	I	-1,799	0,409	0,236
Ηράκλειο	I	-1,285	-0,899	-0,052
Ηράκλειο	I	1,559	-1,458	-0,516
Ηράκλειο	I	-0,843	-0,660	0,569
Ηράκλειο	I	-0,467	-0,779	-0,849
Ηράκλειο	I	-0,114	-2,317	-2,441
Ηράκλειο	I	-0,049	-1,511	-0,467
Ηράκλειο	I	1,127	-0,969	0,456
Ηράκλειο	I	-1,567	-0,448	0,469

Προέλευση	Κωδικοποίηση	Root 1	Root 2	Root 3
Λακωνία	LA	3,605	0,613	0,881
Λακωνία	LA	2,570	1,138	-1,344
Λακωνία	LA	0,774	-1,682	0,943
Λακωνία	LA	3,346	0,064	1,529
Λακωνία	LA	1,126	-2,183	0,722
Λακωνία	LA	2,327	-1,566	-1,695
Λακωνία	LA	1,962	2,119	1,517
Λακωνία	LA	2,024	1,630	-0,693
Λακωνία	LA	2,371	-1,021	0,968
Λακωνία	LA	3,463	-0,239	-0,913
Λακωνία	LA	3,385	-1,109	0,588
Λακωνία	LA	3,082	0,926	-1,192
Λακωνία	LA	3,539	-0,688	1,068
Λακωνία	LA	1,201	1,667	0,089
Λακωνία	LA	2,901	1,269	-1,289
Λακωνία	LA	1,998	-0,328	-1,196
Λακωνία	LA	3,823	-1,646	0,995
Μεσσηνία	ΜΕ	-0,371	0,079	-0,202
Μεσσηνία	ΜΕ	0,344	0,622	0,592
Μεσσηνία	ΜΕ	-0,191	1,070	-0,161
Μεσσηνία	ΜΕ	-3,630	2,805	-0,867
Μεσσηνία	ΜΕ	-0,129	0,291	0,580
Μεσσηνία	ΜΕ	-0,910	-0,118	0,397

Προέλευση	Κωδικοποίηση	Root 1	Root 2	Root 3
Μεσσηνία	ΜΕ	-2,126	1,837	0,810
Μεσσηνία	ΜΕ	0,519	1,721	-0,977
Μεσσηνία	ΜΕ	-0,858	0,660	-0,960
Μεσσηνία	ΜΕ	0,303	1,262	0,795
Μεσσηνία	ΜΕ	0,137	1,092	0,613
Μεσσηνία	ΜΕ	-0,163	1,686	-0,240
Μεσσηνία	ΜΕ	1,087	1,952	-0,717
Μεσσηνία	ΜΕ	-0,240	0,535	-1,063
Μεσσηνία	ΜΕ	1,346	1,493	0,332
Μεσσηνία	ΜΕ	0,470	0,523	0,539
Μεσσηνία	ΜΕ	0,064	1,461	0,310
Μεσσηνία	ΜΕ	-1,000	0,533	0,208
Μεσσηνία	ΜΕ	-1,362	0,208	0,009
Μεσσηνία	ΜΕ	-0,609	0,174	-0,398
Μεσσηνία	ΜΕ	-0,498	0,933	1,005
Μεσσηνία	ΜΕ	0,340	1,451	-0,985
Μεσσηνία	ΜΕ	0,511	2,901	0,329
Μεσσηνία	ΜΕ	-1,283	1,372	-0,383
Μεσσηνία	ΜΕ	-1,548	0,735	-0,742
Μεσσηνία	ΜΕ	-1,076	2,795	0,168
Μεσσηνία	ΜΕ	-2,189	0,724	-0,339
Μεσσηνία	ΜΕ	-0,246	2,502	-0,510
Μεσσηνία	ΜΕ	-0,594	2,142	-0,127
Μεσσηνία	ΜΕ	0,428	2,530	-1,308

Προέλευση	Κωδικοποίηση	Root 1	Root 2	Root 3
Μεσσηνία	ΜΕ	-1,300	0,592	0,962
Μεσσηνία	ΜΕ	-0,415	3,086	0,623
Μεσσηνία	ΜΕ	-0,348	3,166	1,475
Μεσσηνία	ΜΕ	-1,452	2,576	-0,336
Ζάκυνθος	ΖΑ	-0,770	0,379	0,809
Ζάκυνθος	ΖΑ	0,526	-0,970	0,810
Ζάκυνθος	ΖΑ	0,013	-0,589	0,968
Ζάκυνθος	ΖΑ	-0,512	-0,442	0,681
Ζάκυνθος	ΖΑ	-1,122	-0,134	0,365
Ζάκυνθος	ΖΑ	-1,155	0,137	1,084
Ζάκυνθος	ΖΑ	-1,089	-3,260	0,712
Ζάκυνθος	ΖΑ	-1,756	-0,943	1,241
Ζάκυνθος	ΖΑ	-0,484	-1,878	-0,308
Ζάκυνθος	ΖΑ	-2,021	-0,583	-0,659
Ζάκυνθος	ΖΑ	-2,142	-3,146	0,082
Ζάκυνθος	ΖΑ	0,115	0,159	-0,472
Ζάκυνθος	ΖΑ	0,659	-1,153	1,210
Ζάκυνθος	ΖΑ	-2,044	-1,259	0,492
Ζάκυνθος	ΖΑ	1,349	-1,132	0,275
Ζάκυνθος	ΖΑ	0,244	-1,265	2,232
Ζάκυνθος	ΖΑ	-1,934	-1,061	-0,950
Ζάκυνθος	ΖΑ	-2,318	-2,025	0,377
Ζάκυνθος	ΖΑ	-1,141	-0,597	0,745

Προέλευση	Κωδικοποίηση	Root 1	Root 2	Root 3
Ζάκυνθος	ZA	-0,488	-0,500	-1,511
Ζάκυνθος	ZA	-1,147	-1,565	-0,809
Ζάκυνθος	ZA	0,997	-1,397	-0,750
Ζάκυνθος	ZA	-2,052	-0,479	-1,281
Ζάκυνθος	ZA	-1,097	0,323	0,697

Πίνακας 4.5: Πίνακας ταξινόμησης με τη βάση την ομάδα εκπαίδευσης (3 τεχνικές DA: 10 μεταβλητές και 97 δείγματα).

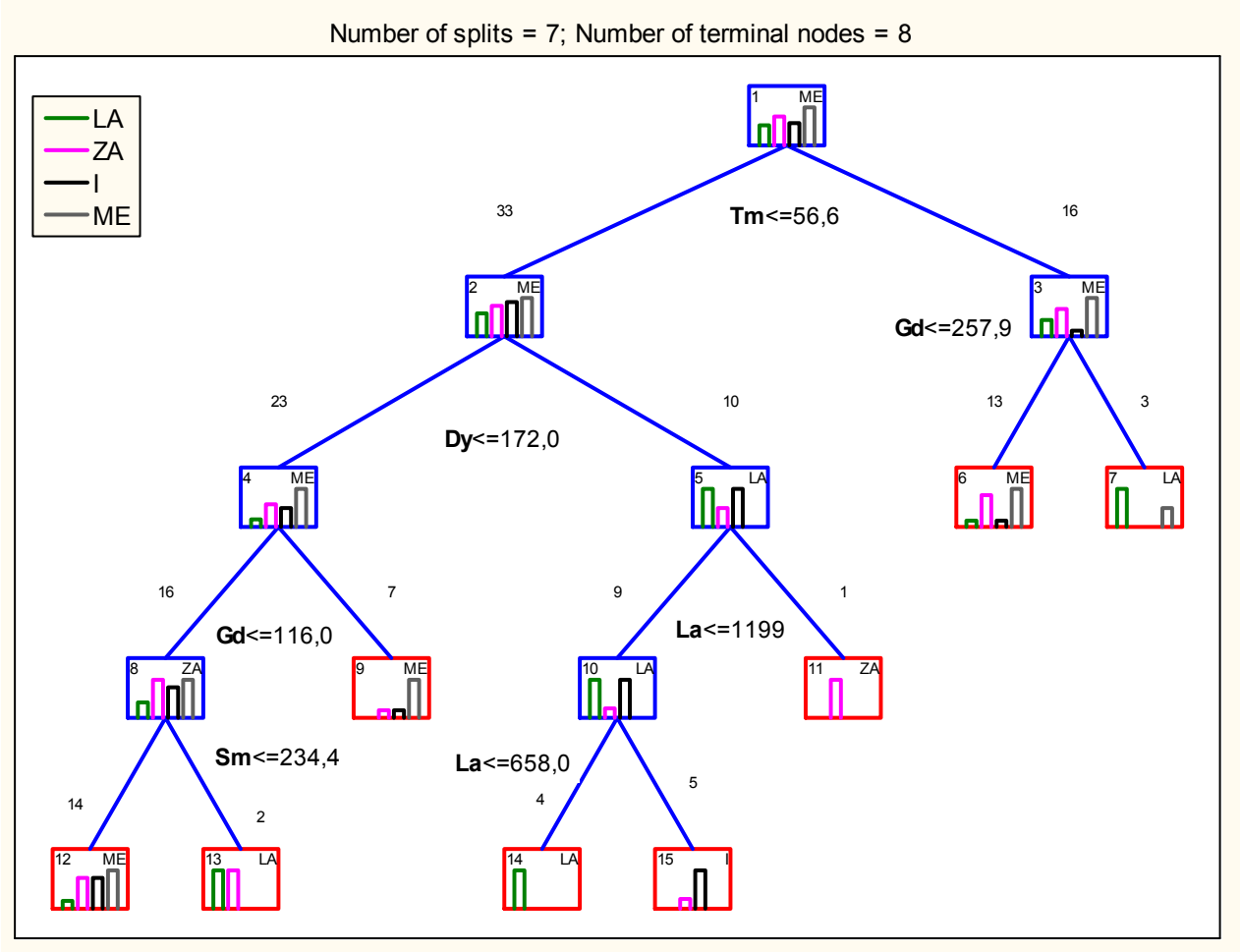
Περιοχή	**Συνολικός αρ. δειγμάτων**	**Κλασική**				**FW**				**BW**			
		I	**LA**	**ME**	**ZA**	**I**	**LA**	**ME**	**ZA**	**I**	**LA**	**ME**	**ZA**
I	22	11	0	1	**10**	8	1	2	**11**	7	1	2	**12**
LA	17	2	14	1	0	1	14	1	1	1	14	1	1
ME	34	0	0	33	1	0	0	32	2	0	2	28	4
ZA	24	**5**	0	3	16	**6**	0	3	15	**6**	0	6	12

Πίνακας 4.6: Κατασκευή δέντρου, (tree structure), κόμβοι (nodes) που δημιουργούνται, παρατηρούμενες (στήλες) έναντι προβλεπόμενων (σειρές) θέσεων, σταθερές και μεταβλητές διαχωρισμού. Με αστερίσκο (*) σημειώνονται οι τερματικοί κόμβοι (μέθοδος LCM)

Κόμβοι	**Αριστ. κλάδος**	**Δεξιός κλάδος**	**LA**	**ZA**	**I**	**ME**	**Πρόβλεψη**	**Σταθερά διαχ/μου**	**Y**	**La**	**Pr**	**Sm**	**Gd**	**Dy**	**Ho**	**Er**	**Tm**	**Yb**
1	2	3	9	13	10	17	ME	0,4436	0,0009	0,0001	-0,0002	-0,0027	0,0006	-0,0013	0,0006	-0,0010	0,0026	-0,0025
2	4	5	9	10	8	3	ZA	0,8067	0,0010	0,0002	0,0004	-0,0010	-0,0053	0,0005	-0,0012	-0,0048	0,0043	-0,0002
3*			0	3	2	14	ME											
4*			6	0	0	0	LA											
5	6	7	3	10	8	3	ZA	0,7442	0,0027	-0,0005	0,0040	-0,0090	0,0060	-0,0023	0,0043	-0,0040	0,0099	-0,0123
6	8	9	3	10	7	1	ZA	2,5414	0,0042	-0,0005	0,0056	-0,0104	0,0027	-0,0045	0,0048	-0,0132	0,0090	-0,0112
7*			0	0	1	2	ME											
8*			3	0	0	0	LA											
9	10	11	0	10	7	1	ZA	2,5811	0,0047	-0,0004	0,0054	-0,0116	0,0022	-0,0060	0,0050	-0,0166	0,0127	-0,0116
10*	12	13	0	8	7	0	ZA	-0,8476	-0,0020	-0,0009	0,0078	-0,0010	0,0075	0,0014	-0,0091	0,0045	0,0007	0,0020
11*			0	2	0	1	ZA											
12*			0	7	1	0	ZA											
13*			0	1	6	0	I											

Πίνακας 4.7: Πίνακας ταξινόμησης για την ομάδα εκπαίδευσης (μέθοδος Classic CT): Παρατηρούμενες θέσεις (στήλες) έναντι προβλεπόμενων (σειρές)

Παρατηρήσεις / Προβλέψεις	**LA**	**ZA**	**I**	**ME**	**Συνολικά**
LA	7	1	0	1	
ZA	0	1	0	0	
I	0	1	4	0	
ME	2	10	6	16	
Συνολικός αρ. δειγμάτων	7/9	1/13	4/10	16/17	28/49
% Ποσοστά επιτυχίας	77,8	7,69	40,0	94,1	57,1



Σχήμα 4.2: Δέντρο ταξινόμησης για τα δείγματα (ελαιόλαδα) των τεσσάρων περιοχών. Μέθοδος: Discriminant-based univariate method, Classic CT.

Πίνακας 4.8: Κατασκευή δέντρου (μέθοδος Classic CT)

Κόμβοι	**Αριστ. κλάδος**	**Δεξιός κλάδος**	**LA**	**ZA**	**I**	**ME**	**Πρόβλεψη**	**Σταθερά διαχ/μου**	**Μεταβλητή διαχ/μου**
1	2	3	9	13	10	17	ME	-56,64	Tm
2	4	5	6	8	9	10	ME	-171,95	Dy
3	6	7	3	5	1	7	ME	-257,94	Gd
4	8	9	2	6	5	10	ME	-115,98	Gd
5	10	11	4	2	4	0	LA	-1199,0	La
6*			1	5	1	6	ME		
7*			2	0	0	1	LA		
8	12	13	2	5	4	5	ZA	-234,45	Sm
9*			0	1	1	5	ME		
10	14	15	4	1	4	0	LA	-658,01	La
11*			0	1	0	0	ZA		
12*			1	4	4	5	ME		
13*			1	1	0	0	LA		
14*			4	0	0	0	LA		
15*			0	1	4	0	I		

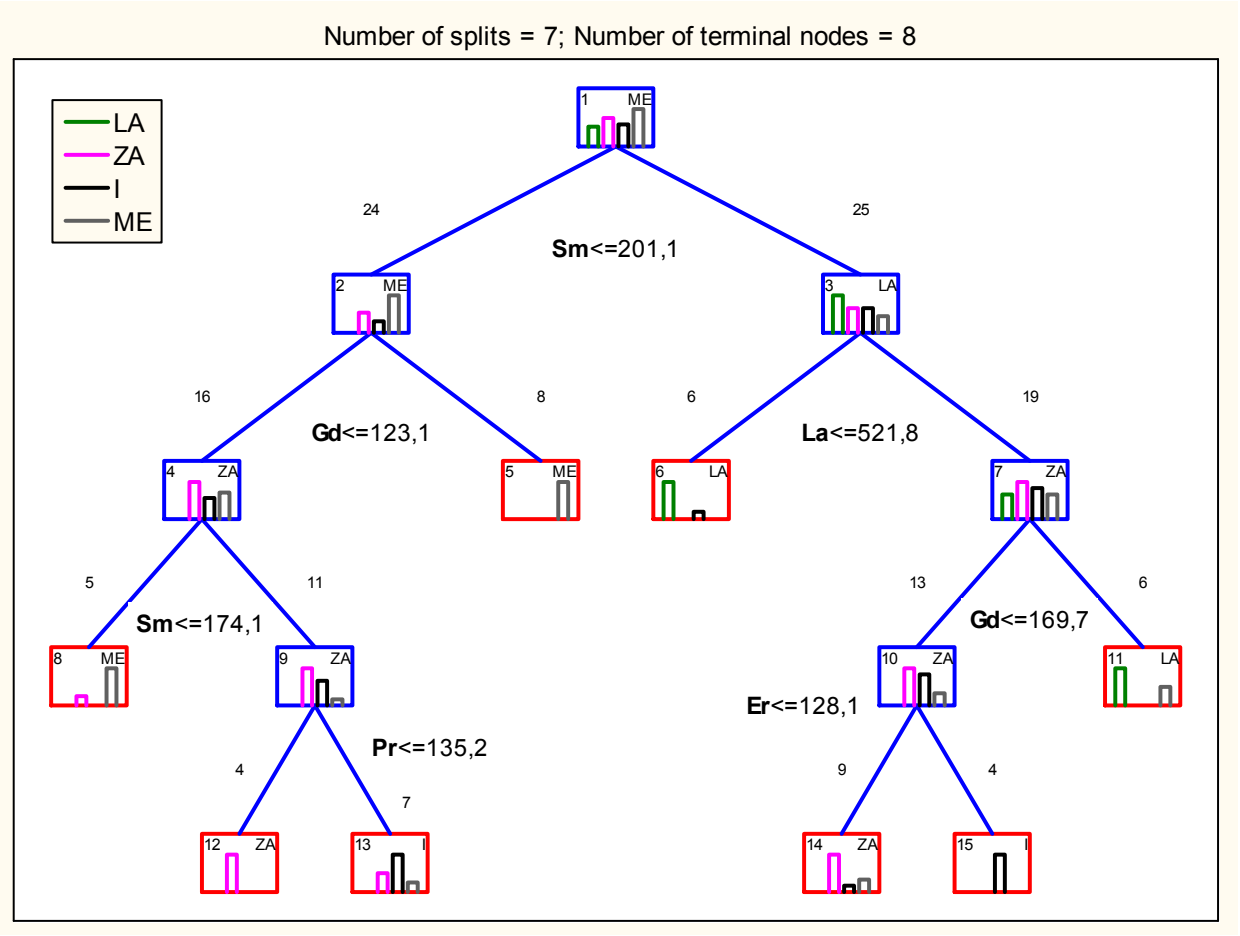
* Τερματικοί κόμβοι

Πίνακας 4.9: Πίνακας ταξινόμησης για την ομάδα ελέγχου (μέθοδος Classic CT): Παρατηρούμενες θέσεις (στήλες) έναντι προβλεπόμενων (σειρές)

Παρατηρήσεις / Προβλέψεις	**LA**	**ZA**	**I**	**ME**	**Συνολικά**
LA	2	2	1	0	
ZA	0	0	0	0	
I	0	0	1	0	
ME	**6**	**9**	**10**	17	
Συνολικός αρ. δειγμάτων	2/8	0/11	1/12	17/17	20/48
% Ποσοστά επιτυχίας	25,0	0,0	8,33	100,0	41,7

Πίνακας 4.10: Πίνακας ταξινόμησης για την ομάδα εκπαίδευσης (μέθοδος CART): Παρατηρούμενες θέσεις (στήλες) έναντι προβλεπόμενων (σειρές)

Παρατηρήσεις / Προβλέψεις	**LA**	**ZA**	**I**	**ME**	**Συνολικά**
LA	9	0	1	2	
ZA	0	10	1	2	
I	0	2	8	1	
ME	0	1	0	12	
Συνολικός αρ. δειγμάτων	9/9	10/13	8/10	12/17	39/49
% Ποσοστά επιτυχίας	100,0	76,9	80,0	94,1	79,6



Σχήμα 4.3: Δέντρο ταξινόμησης για τα δείγματα (ελαιόλαδα) των τεσσάρων περιοχών. Μέθοδος: CART.

Πίνακας 4.11: Κατασκευή δέντρου (μέθοδος CART)

Κόμβοι	**Αριστ. κλάδος**	**Δεξιός κλάδος**	**LA**	**ZA**	**I**	**ME**	**Πρόβλεψη**	**Σταθερά διαχ/μου**	**Μεταβλητή διαχ/μου**
1	2	3	9	13	10	17	ME	-201,1	Sm
2	4	5	0	7	4	13	ME	-123,1	Gd
3	6	7	9	6	6	4	LA	-521,8	La
4	8	9	0	7	4	5	ZA	-174,1	Sm
5*			0	0	0	8	ME		
6*			5	0	1	0	LA		
7	10	11	4	6	5	4	ZA	-169,7	Gd
8*			0	1	0	4	ME		
9	12	13	0	6	4	1	ZA	-135,2	Pr
10	14	15	0	6	5	2	ZA	-128,1	Er
11*			4	0	0	2	LA		
12*			0	4	0	0	ZA		
13*			0	2	4	1	I		
14*			0	6	1	2	ZA		
15*			0	0	4	0	I		

* Τερματικοί κόμβοι

Πίνακας 4.12: Πίνακας ταξινόμησης για την ομάδα ελέγχου (μέθοδος CART): Παρατηρούμενες θέσεις (στήλες) έναντι προβλεπόμενων (σειρές)

Παρατηρήσεις / Προβλέψεις	**LA**	**ZA**	**I**	**ME**	**Συνολικά**
LA	5	0	3	1	
ZA	2	4	**6**	2	
I	0	**5**	3	**8**	
ME	1	2	0	6	
Συνολικός αρ. δειγμάτων	5/8	4/11	3/12	6/17	18/48
% Ποσοστά επιτυχίας	62,5	36,4	25,0	35,3	37,5

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Cosio, M.S., Ballabio, D., Benedetti, S., Gigliotti, C., Geographical origin and authentication of extra virgin olive oils by an electronic nose in combination with artificial neural networks. Anal. Chim. Acta, 2006, **567,** 202-210.
2. Banas, D., Masson, G., Leglize, L., Usseglio-Polatera, P., Boyd, C.E., Assessment of sediment concentration and nutrient loads in effluents drained from extensively managed fishponds in France. Environ. Pollut., 2008, **152,** 679-685.
3. Mente, E., Graham, J.P., Santos, M.B., Neofitou, C., Effect of feed and feeding in the culture of salmonids on the marine aquatic environment: a synthesis for European aquaculture. Aquacult. Int., 2006, **14,** 499-522.
4. Tovar, A., Moreno, C., Mánuel-Vez, M.P., García-Vargas, M., Environmental impacts of intensive aquaculture in marine waters. Water Res., 2000, **34(1),** 334-342.
5. Li, Q., Wu, Z., Chu, B., Zhang, N., Cai, S., Fang, J., Heavy metals in coastal wetland sediments of the Pearl River Estuary, China, Environ. Pollut., 2007, **149,** 158-164.
6. Nghia, N.D., Lunestad, B.T., Trung, T.S., Son, N.T., Maage, A., Heavy Metals in the Farming Environment and in some Selected Aquaculture Species in the Van Phong Bay and Nha Trang Bay of the Khanh Hoa Province in Vietnam, Bull. Environ. Contam. Toxicol., 2009, **82,** 75–79.
7. Chou, C.L., Haya, K., Paon, L.A., Burridge, L., Moffatt J.D., Aquaculture-related trace metals in sediments and lobsters and relevance to environmental monitoring program ratings for near-field effects, Mar. Pollut. Bull., 2002, 44, 1259-1268.
8. Mendiguchía, C., Moreno, C., Mánuel-Vez, M.P., García-Vargas, M. Preliminary investigation on the enrichment of heavy metals in marine sediments originated from intensive aquaculture effluents, Aquaculture, 2006, **254,** 317–325.
9. Apostolaki, E.T., Tsagaraki, T., Tsapakis, M., Karakassis, I., Fish farming impact on sediments and macrofauna associated with seagrass meadows in the Mediterranean, Estuar. Coast. Shelf S., 2007, **75,** 408-416.
10. Dalman, O., Demiral, A. Balci, A., Determination of heavy metals (Cd, Pb) and trace elements (Cu, Zn) in sediments and fish of the Southeastern Aegean, Food chem., 2006, **95,** 157-162.
11. Barasan, A.K., Aksu, M., Egemen, O., Impacts of the fish farms on the water column nutrient concentrations and accumulation of heavy metals in the sediments in the eastern Aegean Sea (Turkey), Environ. Monit. Assess., 2010, **162,** 439–451.

12. Neofitou, N., Vafidis, D., Klaoudatos, S., Spatial and temporal effects of fish farming on benthic community structure in a semi-enclosed gulf of the Eastern Mediterranean, Aquacult. Environ. Interact., 2010, **1,** 95-105.
13. Cao, L., Wang, W., Yang, Y., Yang, C., Yuan, Z., Xiong, S., Diana J., Environmental Impact of Aquaculture and Countermeasures to Aquaculture Pollution in China, Env. Sci. Pollut. Res., 2007, **14,** 452–462.
14. Sapkota A., Sapkota, A. R., Kucharski, M., Burke, J., McKenzie, S., Walker, P., Lawrence, R., Aquaculture practices and potential human health risks: Current knowledge and future priorities, Environ. Int., 2008, **34,** 1215-1226.
15. Schendel, E.K., Nordström, S.E.. Lavkulich, L.M., Floc and sediment properties and their environmental distribution from a marine fish farm, Aquacult. Res., 2004, **35,** 483-493.
16. Mazzola, A., Mirto, S., Danovaro, R., Initial fish-farm impact on meiofaunal assemblages in coastal sediments of the Western Mediterranean. Mar. Pollut. Bull., 1999, **38,** 1126-1133.
17. Mazzola, A., Mirto, S., La Rosa, T., Fabiano, M., Danovaro, R., Fish-farming effects on benthic community structure in coastal sediments: analysis of meiofaunal recovery. J. Mar. Sci., 2000, **57,** 1454-1461.
18. Crawford, C.M., Mitchell, M.I., Macleod, C.K.A., Video assessment of environmental impacts of salmon farms. J. Mar. Sci., 2001, **58,** 445-452.
19. Karakassis, I., Tsapakis, M., Hatziyanni, E., Papadopoulou, K.N., Plaiti, W., Impact of cage farming of fish on the seabed in three Mediterranean coastal areas. J. Mar. Sci., 2000, **57,** 1462-1471.
20. Walkey, A., A critical examination of a rapid method for determining organic carbon in soil. Soil Sci., 1947, **63,** 251–263.
21. Astel, A., Tsakovski, S., Barbieri, P., Simeonov, V., Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. Water Res., 2007, **41,** 4566-4578.
22. Dean, R.J., Shimmield, T.M., Black, K.D., Copper, zinc and cadmium in marine cage fish farm sediments: An extensive survey. Environ. Pollut., 2007, **145,** 84-95.
23. Jaysankar, D., Fukami, K., Iwasaki, K., Okamura, K., Occurrence of heavy metals in the sediments of Uranouchi Inlet, Kochi prefecture, Japan. Fish Sci., 2009, **75,** 413–423.

24. Chou, C.L., Haya, K., Paon, L.A., Moffatt J.D., A regression model using sediment chemistry for the evaluation of marine environmental impacts associated with salmon aquaculture cage wastes, Mar. Pollut. Bull., 2004, **49,** 465-472.
25. Sutherland, T.F., Petersen, S.A., Levings, C.D., Martin, A.J., Distinguishing between natural and aquaculture-derived sediment concentrations of heavy metals in the Broughton Archipelago, British Columbia, Mar. Pollut. Bull., 2007, **54**, 1451–1460.
26. Zur, R.M., Jiang, Y., Pesce, L.L.,Drukker, K., Noise injection for training artificial neural networks: A comparison with weight decay and early stopping, Med. Phys., 2009, **36,** 4810-4818.
27. Paz Suárez Araujo, C., García Báez, P., Sánchez Rodríguez, Á., Juan Santana Rodríguez, J., HUMANN-based system to identify benzimidazole fungicides using multi-synchronous fluorescence spectra: An ensemble approach, Anal. Bioanal. Chem., 2009, **394,** 1059-1072.
28. http://en.wikipedia.org/wiki/ROC_curve (τελευταία επίσκεψη 9/1/2009).
29. STATISTICA 7th edition, software, StatSoft, Inc., 2004.
30. Alfassi, Z.B., Boger, Z., Ronen, Y., Statistical Treatment of Analytical Data, Blackwell Science Ltd, Oxford, UK 2005.
31. Alvarez-Guerra, M., Ballabio, D., Amigo, J.N., Bro, R., Viguri, J.R., Development of models for predicting toxicity from sediment chemistry by partial least squares-discriminant analysis and counter-propagation artificial neural networks, Environ. Pollut., 2010, **158,** 607-614.
32. Tortajada, S., García-Gómez, J.M., Vicente, J., Sanjuán, J., de frutos, R., Martín-Santos, R., García-Esteve, L., Gornemann, I., Gutiérrez-Zotes, A., Canellas, F., Carracedo, Á., Gratacos, M., Guillamat, R., Baca-García, E., Robles, M., Prediction of Postpartum Depression Using Multilayer Perceptrons and Pruning, Methods Inf., Med., 2009, **48,** 291-298.
33. Román, R.C., Hernández, O.G., Urtubia, U.A., Prediction of problematic wine fermentations using artificial neural networks, Bioproc. Biosyst. Eng., 2011, **34,** 1057-1065.
34. Rao, H., Yang, G., Tan, N., Li, P., Li, Z., Li, X., Prediction of HIV-1 Protease Inhibitors Using Machine Learning Approaches, QSAR Comb. Sci., 2009, **28,** 1346-1357.

35. Tirelli, T., Pessani, D., Use of Decision tree and Artificial neural network approaches to model presence/absence of Teleste Muticellus in Piedmont (north-western Italy), River Res. Appl., 2009, **25,** 1001-1012.