ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΣΧΟΛΗ ΠΕΡΙΒΑΛΛΟΝΤΟΣ
ΤΜΗΜΑ ΕΠΙΣΤΗΜΩΝ ΤΗΣ ΘΑΛΑΣΣΑΣ

Βελτιστοποίηση μαθηματικών αλγορίθμων μάθησης στη θαλάσσια οικολογία

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Ανδρονίκη Ταμβάκη

Τριμελής Συμβουλευτική Επιτροπή:

Ιωάννης Μυριτζής (Επίκουρος Καθηγητής, Πανεπιστήμιο Αιγαίου)
Γεώργιος Τσιρτσής (Αναπληρωτής Καθηγητής, Πανεπιστήμιο Αιγαίου)
Σοφία Σπαθάρη (Λέκτορας, Glasgow University)

Μυτιλήνη, Μάιος 2014

UNIVERSITY OF THE AEGEAN

SCHOOL OF THE ENVIRONMENT
DEPARTMENT OF MARINE SCIENCES

Optimization of machine learning algorithms in marine ecology

Ph.D Thesis

Androniki Tamvakis

Advisory Committee:

John Miritzis (Assistant Professor, University of the Aegean)
George Tsirtsis (Associate Professor, University of the Aegean)
Sofie Spatharis (Lecturer, Glasgow University)

Mytilene, May 2014

Τριμελής Συμβουλευτική Επιτροπή

ΙΩΑΝΝΗΣ ΜΥΡΙΤΖΗΣ, Επίκουρος Καθηγητής Πανεπιστημίου Αιγαίου (Επιβλέπων)
ΓΕΩΡΓΙΟΣ ΤΣΙΡΤΣΗΣ, Αναπληρωτής Καθηγητής Πανεπιστημίου Αιγαίου
ΣΟΦΙΑ ΣΠΑΘΑΡΗ, Λέκτορας Πανεπιστημίου Γλασκώβης

Επταμελής Επιτροπή

ΙΩΑΝΝΗΣ ΜΥΡΙΤΖΗΣ, Επίκουρος Καθηγητής Πανεπιστημίου Αιγαίου (Επιβλέπων)
ΓΕΩΡΓΙΟΣ ΤΣΙΡΤΣΗΣ, Αναπληρωτής Καθηγητής Πανεπιστημίου Αιγαίου
ΣΟΦΙΑ ΣΠΑΘΑΡΗ, Λέκτορας Πανεπιστημίου Γλασκώβης
ΓΕΩΡΓΙΟΣ ΤΣΕΚΟΥΡΑΣ, Αναπληρωτής Καθηγητής Πανεπιστημίου Αιγαίου
ΣΤΡΑΤΗΣ ΓΕΩΡΓΑΚΑΡΑΚΟΣ, Αναπληρωτής Καθηγητής Πανεπιστημίου Αιγαίου
ΓΕΩΡΓΙΟΣ ΚΟΚΚΟΡΗΣ, Επίκουρος Καθηγητής Πανεπιστημίου Αιγαίου
ΧΡΗΣΤΟΣ-ΝΙΚΟΛΑΟΣ ΑΝΑΓΝΩΣΤΟΠΟΥΛΟΣ, Επίκουρος Καθηγητής Πανεπιστημίου Αιγαίου

Advisory Committee

JOHN MIRITZIS, Assistant Professor, University of the Aegean (Supervisor)
GEORGE TSIRTSIS, Associate Professor, University of the Aegean,
SOFIE SPATHARIS, Lecturer, Glasgow University

Examining Committee

JOHN MIRITZIS, Assistant Professor, University of the Aegean (Supervisor)
GEORGE TSIRTSIS, Associate Professor, University of the Aegean
SOFIE SPATHARIS, Lecturer, Glasgow University
GEORGE TSEKOURAS, Associate Professor, University of the Aegean
STRATIS GEORGAKARAKOS, Associate Professor, University of the Aegean
GIORGOS KOKKORIS, Assistant Professor, University of the Aegean
CHRISTOS-NIKOLAOS ANAGNOSTOPOULOS, Assistant Professor, University of the Aegean

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κ. Μυριτζή για την απόφασή του να συμπορευτούμε σε μια πορεία όχι αυστηρά μαθηματική για την προσέγγιση της θαλάσσιας οικολογίας.

Επίσης, ευχαριστώ τον κ. Τσιρτσή για τον διττό του ρόλο στην εκπόνηση αυτής της διατριβής, αυτόν της καθοδήγησης πάνω σε θέματα μοντελοποίησης περιβαλλοντικών διεργασιών και αυτόν της οικολογικής επεξήγησής τους.

Η κ. Σπαθάρη είχε καίρια και πολλαπλή συμμετοχή στην ολοκλήρωση αυτής της διατριβής προσφέροντας απλόχερα επιστημονική γνώση, συντονισμό και εμψύχωση και δεν μπορώ παρά να την ευγνωμονώ για την όλη της προσφορά.

Ακόμα, θα ήθελα να ευχαριστήσω τον κάθε ένα από τα μέλη της επταμελούς επιτροπής ξεχωριστά: τον κ. Αναγνωστόπουλο για τη συνεργασία και τις ιδέες του, τον κ. Τσεκούρα για την αμέριστη στήριξή του, τον κ. Γεωργακαράκο για τις πολύτιμες συμβουλές του πάνω σε θέματα νευρωνικών δικτύων και τον κ. Κόκκορη για τις οπτικές και τις λύσεις που μου προσέφερε ειδικά σε δύσκολες στιγμές.

Τέλος, θα ήθελα να ευχαριστήσω τα μέλη της οικογένειας μου (σύζυγο, γονείς, πεθερικά και παιδιά) για την κατανόηση και τη βοήθεια που μου προσέφεραν προκειμένου να έχω χρόνο για να δουλεύω απερίσπαστα προς την ολοκλήρωση της διδακτορικής μου διατριβής.

τη

Contents

List of abbreviations

ANOVA	analysis of variance
chl α	chlorophyll α
CV	cross validation
DIN	dissolved inorganic nitrogen
DT	decision tree
EM	ensemble method
F	photoperiod
IBk	instance based learner which uses k neighbors
IBL	instance based learning
ICZM	integrated coastal zone management
IT	information technology
k-NN	nearest neighbor method with k neighbors
LM	linear (regression) model
Log	multinomial logistic regression
LOOCV	leave one out cross validation
MAE	mean absolute error
ML	machine learning
MLP	multilayer perceptron
MLR	multiple linear regression
MT	model tree
N	nitrogen
NB	naïve bayes
NN	neural network
PART	classifier that generates rules from a partial decision tree
PO_4	phosphate
R	correlation coefficient
RF	random forest
RIPPER	classifier that uses repeated incremental pruning to produce error production
RMSE	root mean squared error
S	salinity
SiO_2	silicate
SMO	sequential minimal optimization
SVM	support vector machines
T	temperature
WFD	water framework directive 2000/60/EC

ΕΚΤΕΝΗΣ ΠΕΡΙΛΗΨΗ

Ο θαλάσσιος ευτροφισμός είναι ένα σύνθετο φαινόμενο που εξαρτάται από φυσικοχημικούς παράγοντες, βιολογικές διεργασίες, χωρική ετερογένεια, εποχικές διακυμάνσεις, τοπικές ιδιαιτερότητες και χαρακτηρίζεται από στοχαστικότητα. Στα παράκτια οικοσυστήματα ο ευτροφισμός σχετίζεται με ποικίλες διεργασίες που η διερεύνησή τους αποτελεί κρίσιμο ζητούμενο της σύγχρονης θαλάσσιας οικολογίας ιδίως μετά την θέσπιση της Ευρωπαϊκής οδηγίας για τα ύδατα (European Water Framework Directive).

Στην παρούσα διατριβή η πρωτογενής παραγωγικότητα, η οικολογική κατάσταση των παράκτιων υδάτων καθώς και η βιοποικιλότητα των φυτοπλαγκτικών κοινοτήτων μοντελοποιήθηκαν χρησιμοποιώντας αποκλειστικά αβιοτικές παραμέτρους με χρήση διαφορετικών μεθόδων μηχανικής μάθησης (machine learning techniques). Συγκεκριμένα, για την πρόβλεψη της πρωτογενούς παραγωγικότητας χρησιμοποιήθηκαν δένδρα πρόβλεψης (model trees) που επέτρεψαν να περιγραφεί με επεξηγηματικό τρόπο η κατάσταση του οικοσυστήματος. Η οικολογική κατάσταση των υδάτων ταξινομήθηκε χρησιμοποιώντας τον συνδυαστικό αλγόριθμο ψηφοφορίας (voting ensemble method), ενώ ένας νέος δείκτης προτάθηκε προκειμένου να διευκολυνθεί η βελτιστοποίηση της απόδοσής του. Τέλος, τέσσερις βασικοί αλγόριθμοι μάθησης προέβλεψαν τη ποικιλότητα φυτοπλαγκτικών κοινοτήτων εκφρασμένη ως πλούτο ειδών, ισοκατανομή και επικράτηση, χρησιμοποιώντας φυσικές και προσομοιωμένες συναθροίσεις. Η παραπάνω μελέτη οδήγησε στην κατασκευή ενός ειδικού λογισμικού για την πρόβλεψη της ποικιλότητας των φυτοπλαγκτικών συναθροίσεων της Ανατολικής Μεσογείου χρησιμοποιώντας αποκλειστικά αβιοτικές μεταβλητές.

Μελέτη περίπτωσης Ι: Μετεωρολογικές επιδράσεις στο θαλάσσιο ευτροφισμό – Μοντελοποίηση με δένδρα πρόβλεψης (Model Trees)

Η πρώτη μελέτη περίπτωσης αφορά στην ποσοτική εκτίμηση της θαλάσσιας πρωτογενούς παραγωγικότητας, εκφρασμένης ως χλωροφύλλη-α (βασικότερη μεταβλητή που χαρακτηρίζει τον ευτροφισμό), αποκλειστικά από αβιοτικές φυσικοχημικές παραμέτρους. Η εκτίμηση της χλωροφύλλης επιχειρήθηκε μέσω ενός αλγορίθμου μηχανικής μάθησης, τα δένδρα πρόβλεψης (model trees). Ο αλγόριθμός αυτός διαχωρίζει τα δεδομένα σε ομογενή σύνολα (φύλλα δένδρου) και στη συνέχεια εφαρμόζει σε αυτά

γραμμικά μοντέλα πρόβλεψης. Στην συνέχεια, προκειμένου να καθοριστεί και να συγκριθεί η σχετική επίδραση της κάθε αβιοτικής παραμέτρου μέσα στα ομογενή σύνολα, εφαρμόστηκε τυποποιημένη γραμμική παλινδρόμηση έναντι της απλής. Η προβλεπτική ικανότητά του αλγόριθμου συγκρίθηκε με τα αντίστοιχα αποτελέσματα: (α) των νευρωνικών δικτύων που είναι ο ευρύτερα χρησιμοποιούμενος αλγόριθμος μηχανικής μάθησης και (β) της κλασσικής στατιστικής μεθόδου της πολλαπλής γραμμικής παλινδρόμησης. Επιπλέον, η δυνατότητα του αλγόριθμου να περιγράψει τις διεργασίες που σχετίζονται με τον ευτροφισμό διερευνήθηκε με την εφαρμογή του σε δύο διαφορετικά και εκ διαμέτρου αντίθετα έτη δειγματοληψίας: ενός ιδιαίτερα ξηρού ('04-'05) και ενός τυπικά υγρού ('09-'10) έτους για την περιοχή του κόλπου Καλλονής νήσου Λέσβου.

Τα αποτελέσματα έδειξαν ότι τα δένδρα πρόβλεψης παρέχουν αυξημένη ικανότητα πρόβλεψης της χλωροφύλλης σε σχέση με τους άλλους δύο αλγορίθμους. Αυτό το γεγονός συνδέεται με το ότι η πρόβλεψη στα δένδρα συντελείται μέσα στα διαχωρισμένα ομογενή σύνολα δεδομένων και όχι σε ένα ενιαίο σύνολο όπως γίνεται στα νευρωνικά δίκτυα και τη πολλαπλή παλινδρόμηση. Όσον αφορά στα δύο έτη δειγματοληψίας ο διαχωρισμός των δεδομένων, που ήταν ως επί το πλείστον εποχικός, προσφέρει έναν επεξηγηματικό τρόπο περιγραφής του συστήματος. Πραγματικά, οι αβιοτικές παράμετροι που χρησιμοποιήθηκαν για τον διαχωρισμό στα κλαδιά καθώς και οι συντελεστές βαρύτητάς τους στα φύλλα αποδίδουν μια χρήσιμη κλιμάκωση των παραμέτρων που επηρεάζουν τον ευτροφισμό. Επομένως, η μέθοδος της ανάπτυξης δένδρων πρόβλεψης προτείνεται ως ένα χρήσιμο εργαλείο για την εξόρυξη γνώσης που αφορά στις οικοσυστημικές διεργασίες που σχετίζονται με τον ευτροφισμό, συμβάλλοντας συγχρόνως σημαντικά στο ευρύτερο πλαίσιο της ολοκληρωμένης διαχείρισης της παράκτιας ζώνης.

Μελέτη περίπτωσης ΙΙ: Ανίχνευση της βέλτιστης ταξινόμησης που προσφέρει ο συνδυαστικός αλγόριθμος ψηφοφορίας (voting ensemble method) με χρήση ενός νέου προτεινόμενου δείκτη

Οι συνδυαστικοί αλγόριθμοί (ensemble methods) μηχανικής μάθησης είναι μια νέα κατηγορία αλγορίθμων που προσφέρουν ταξινόμηση χρησιμοποιώντας συνδυαστικά τα αποτελέσματα υφιστάμενων ταξινομητών. Έτσι, οι συνδυαστικοί αλγόριθμοι προσφέρουν ταξινόμηση που δεν βασίζεται σε μία και μόνο προσέγγιση αλλά συνδυάζουν περισσότερες διαφορετικές μεθόδους με αποτέλεσμα να παρέχουν συνήθως επιτυχέστερη πρόβλεψη. Παρόλα αυτά, η επιλογή των ταξινομητών, που θα συμμετέχουν στους αλγόριθμους αυτούς και ειδικότερα στον αλγόριθμο ψηφοφορίας (voting algorithm) που είναι και ο ευρύτερα χρησιμοποιούμενος, είναι ένα ανοιχτό και καίριο επιστημονικό πρόβλημα.

Σε αυτή την έρευνα προτείνουμε ένα νέο δείκτη (DP) που ενσωματώνει δύο σημαντικά κριτήρια για την επιτυχή επιλογή των ταξινομητών που θα συμμετέχουν στον συνδυαστικό αλγόριθμο ψηφοφορίας: την ανομοιότητα μεταξύ των αποτελεσμάτων ταξινόμησης (Dissimilarity) και την ίδια την απόδοσή τους (Performance). Προκειμένου να αναπτυχθεί ο δείκτης DP, συνδυάστηκαν σε τριάδες, δέκα απλοί ταξινομητές που αντιπροσωπεύουν όλες τις διαφορετικές κατηγορίες ταξινομητών (κανόνες, δένδρα, αλγόριθμοι υποδειγμάτων, συναρτήσεις και ταξινομητές Bayes). Επιπλέον, υπολογίσθηκε η σχέση μεταξύ υφιστάμενων δεικτών ανομοιότητας και απόδοσης του αλγορίθμου ψηφοφορίας, με τον δείκτη Jaccard να επιτυγχάνει την υψηλότερη συσχέτιση. Βάσει αυτού του αποτελέσματος ο δείκτης ανομοιότητας Jaccard συμπεριλήφθηκε μαζί με την απόδοση των ταξινομητών στο νέο δείκτη DP. Για να δοκιμαστεί η απόδοση του δείκτη DP, η εκπαίδευση των αλγορίθμων (απλών ταξινομητών και αλγόριθμου ψηφοφορίας) πραγματοποιήθηκε με χρήση δύο εντελώς διαφορετικών βάσεων δεδομένων. Η πρώτη βάση περιείχε φωνητικά δεδομένα τα οποία χρησιμοποιούνται για να ανιχνεύσουν επτά διαφορετικά συναισθήματα (θυμός, ευτυχία, άγχος/φόβος, θλίψη, ανία, αποστροφή και ουδετερότητα). Η δεύτερη βάση περιείχε περιβαλλοντικά δεδομένα που συλλέχτηκαν σε δειγματοληψίες που καλύπτουν έναν ετήσιο κύκλο στον κόλπο Καλλονής νήσου Λέσβου. Στα δεδομένα αυτά συμπεριλαμβάνονται εννέα φυσικοχημικές μεταβλητές και η ζητούμενη ταξινόμηση αφορά σε πέντε οικολογικές καταστάσεις της ποιότητας των υδάτων (υψηλή, καλή, μέτρια, φτωχή και κακή) βασισμένες στην Ευρωπαϊκή οδηγία για τα ύδατα.

Ο νέος προτεινόμενος δείκτης DP, έδειξε αφενός υψηλή συσχέτιση με την απόδοση του αλγόριθμου ψηφοφορίας και αφετέρου κατάφερε να αναγνωρίσει ποιοι είναι οι καταλληλότεροι συνδυασμοί ταξινομητών που επιτυγχάνουν τις υψηλότερες αποδόσεις όταν τροφοδοτούν τον αλγόριθμο ψηφοφορίας. Ο DP απευθύνεται στους χρήστες μεθόδων μηχανικής μάθησης που θα τον χρησιμοποιήσουν, προκειμένου να επιλέξουν τους ταξινομητές που θα τροφοδοτήσουν τον αλγόριθμο ψηφοφορίας για να επιτύχουν την βέλτιστη απόδοση ταξινόμησης. Χρησιμοποιώντας τον απλό και φιλικό δείκτη

DP, θα αποφύγουν την εξαντλητική, χρονοβόρα και υπολογιστικά απαιτητική αναζήτηση του αποδοτικότερου συνδυασμού ταξινομητών.

Μελέτη περίπτωσης ΙΙΙ: Αποτελεσματική πρόβλεψη της βιοποικιλότητας θαλάσσιων κοινοτήτων αποκλειστικά από αβιοτικές παραμέτρους

Στη παρούσα μελέτη περίπτωσης, προτείνεται μια ολοκληρωμένη μεθοδολογία για την αποτελεσματική πρόβλεψη της βιοποικιλότητας αποκλειστικά από αβιοτικές παραμέτρους. Η πρόβλεψη επιχειρείται μέσω τριών αλγορίθμων μηχανικής μάθησης: τα δένδρα πρόβλεψης (model trees), τους πολυεπίπεδους αισθητήρες (multilayer perceptron) και τον αλγόριθμο υποδειγμάτων (instance based learning). Ως αβιοτικές παράμετροι εισόδου των παραπάνω αλγορίθμων επιλέχθηκαν η θερμοκρασία, η αλατότητα, το διαλυμένο ανόργανο άζωτο και τα φωσφορικά άλατα που είναι γνωστό ότι διαμορφώνουν τη δομή των φυτοπλαγκτικών συναθροίσεων. Η βιοποικιλότητα εκφράζεται μέσω αριθμού οικολογικών δεικτών που εκφράζουν τον πλούτο ειδών, την ισοκατανομή και την επικράτηση των φυτοπλαγκτικών συναθροίσεων και οι οποίοι αποτελούν την έξοδο των αλγορίθμων. Προκειμένου να βελτιστοποιηθεί η πρόβλεψη της βιοποικιλότητας, οι οικολογικοί δείκτες υπολογίστηκαν σε ένα μεγάλο αριθμό φυσικών φυτοπλαγκτικών συναθροίσεων που συλλέχθηκαν στο πεδίο αλλά και σε προσομοιωμένες συναθροίσεις αφθονίας αντίστοιχης των συναθροίσεων πεδίου και απαλλαγμένες θορύβου. Οι προσομοιωμένες συναθροίσεις παρήχθησαν βάσει του μοντέλου της λογαριθμοκανονικής κατανομής ώστε να διατηρούνται τα αρχικά χαρακτηριστικά των φυσικών συναθροίσεων από τις οποίες προήλθαν.

Τα αποτελέσματα έδειξαν ότι η βιοποικιλότητα μπορεί να προβλεφθεί ικανοποιητικά χρησιμοποιώντας αποκλειστικά φυσικοχημικές παραμέτρους ενώ η ικανότητα πρόβλεψης διπλασιάζεται όταν χρησιμοποιούνται προσομοιωμένες συναθροίσεις. Ο αλγόριθμος υποδειγμάτων έδωσε τα βέλτιστα αποτελέσματα ειδικά για τους δείκτες: Menhinick (πλούτου), Evenness E2 (ισοκατανομής) και Berger-Parker (επικράτησης). Με βάση τον αλγόριθμο, τους δείκτες και τη μορφή των συναθροίσεων που βελτιστοποιούν την πρόβλεψη, αναπτύχθηκε ειδικό λογισμικό για την εκτίμηση της βιοποικιλότητας φυτοπλαγκτού στην περιοχή της Ανατολικής Μεσογείου.

Η προτεινόμενη προσέγγιση που βασίζεται σε δεδομένα φυτοπλαγκτικών συναθροίσεων, ενδέχεται να έχει εφαρμογή και σε άλλες ομάδες φυτικών και ζωικών οργανισμών όχι μόνο σε θαλάσσια αλλά και σε χερσαία οικοσυστήματα. Η αποτελεσματική πρόβλεψη της βιοποικιλότητας από αβιοτικές παραμέτρους παρουσιάζει πλήθος εφαρμογών όπως η ενσωμάτωση της δομής κοινοτήτων σε οικολογικά μοντέλα και η μελέτη της βιοποικιλότητας σε σενάρια παγκόσμιας αλλαγής.

ABSTRACT

The exploration of processes leading to coastal eutrophication is a major challenge in ecological research, particularly in light of important new policies such as the European Water Framework Directive. In the present study primary production, water quality status and phytoplankton diversity are modeled based on exclusively abiotic parameters using different machine learning techniques. Specifically, model trees showed increased predictive power in primary production prediction offering an explanatory description of ecosystem status. The water quality status was sufficiently classified using a voting ensemble method and a novel index was proposed in order to facilitate the optimization procedure during voting training. Finally, phytoplankton biodiversity was predicted in terms of its three components (richness, evenness and dominance) using both field and noise-free simulated assemblages. Based on the optimization of biodiversity prediction, a software package was developed for phytoplankton diversity prediction for Eastern Mediterranean waters.

The study resulted in the development of information technology tools offering useful insights into ecosystem processes affecting eutrophication in coastal ecosystems, constituting also useful components in integrated coastal zone management. Moreover, the proposed methodologies can be easily extended or adapted to any group of organisms either in marine or terrestrial ecosystems. Possible future applications include also the incorporation of community structure in ecological models and global change scenarios.

1 INTRODUCTION

Coastal areas worldwide are increasingly susceptible to eutrophication phenomena often due to anthropogenic causes such as sewage and terrestrial runoff (Beman et al., 2005). Recently, coastal eutrophication has received special attention in light of new policies e.g. the Water Framework Directive 2000/60/EC (WFD, 2000), the protocol for integrated coastal zone management (ICZM) and marine biodiversity protection (Coll et al., 2010; Karydis, 1996; Ruiz & Velasco, 2010). However, eutrophication assessment remains a complex process (Arhonditsis et al., 2003; Kitsiou & Karydis, 2011; Vollenweider, 1974) often associated with contrasting physicochemical and biological criteria, spatial heterogeneity, seasonal variability, local conditions, and stochastic processes (Spatharis et al., 2007a). Two crucial measures for understanding and predicting eutrophication phenomena are the phytoplankton biomass which is one of the most commonly used proxies (Karydis & Tsirtsis, 1996) and phytoplankton diversity which provides valuable information on changes in community structure (Collin et al., 2011). Another contemporary measurement is the quality status of coastal waters as determined for the purposes of WFD (Marin-Guirao et al., 2005), which renders the corresponding classification useful component of successful ICZM schemes. As a result, predicting phytoplankton biomass and diversity along with classifying the quality status of coastal waters through a number of biotic and abiotic parameters are current challenging issues in marine ecology (Gontier et al., 2006; Ingram & Steel, 2010).

Numerous approaches have been used for modeling phytoplankton biomass (in terms of chlorophyll *α* – chl *α*) (Kitsiou & Karydis, 2011) highlighting the importance of this undertaking. Two of the most traditional statistical approaches are linear regression models (Cho et al., 2009; Onderka, 2007) and principal component analysis (Camdevyren et al., 2005; Liu et al., 2010; Primpas et al., 2010). Bayesian statistics have also been applied for chl *α* prediction using a probabilistic, rather than a simple deterministic approach (Borsuk et al., 2004; Freeman et al., 2009; Ramin et al., 2010). More elaborate approaches include coupled models that incorporate both hydrodynamic and ecological processes (Allen et al., 2007; Lewis & Allen, 2009; Wu et al., 2009). On the other hand, few attempts have been made so far to predict phytoplankton biodiversity. Most studies are still based on classical statistical approaches such as regression analysis (Arias-Gonzalez et al., 2011; Brakstad et al., 1994; Denisenko, 2010; Thrush et al., 2001). But, estimating diversity is also essential when it comes to prioritizing sites for management purposes (Lockwood et al., 2012), for assessing the ecological status of ecosystems (WFD, 2000; (Spatharis & Tsirtsis, 2010) or for predicting effects of global change on ecosystem diversity and function (Dawson et al., 2011). Finally, the quality status of the European waters has also been studied within more theoretical and comparative frameworks e.g. by interpreting historical references (Nielsen et al., 2003; Andersen et al., 2004) or by comparing data from different areas (Borja et al., 2007). Some other studies included classical statistical approaches such as discriminant analysis (Muxika et al., 2007) or principal component analysis (Romero et al., 2007; Sondergaard et al., 2005).

In this context, alternative perspectives are called to provide a realistic prediction of phytoplankton biomass, diversity and water quality status based on a small number of abiotic parameters which are more straightforward to measure. Machine Learning (ML), an area of artificial intelligence, includes such techniques offering efficient predictive performance and interpretable results to different scientific applications. Generally, ML techniques acquire information from collected data (e.g. field samples) and yield generalization to the computational system for the effective representation of the scientific issue under consideration. This ML modeling perspective is appropriate to ecology since in such assessments there is original data availability and the oncoming generalization provides new insights on the study systems. Moreover, these techniques are effective for exploring complex ecological processes (Crisci et al., 2012; Fielding, 1999), and can handle non-linearity without relying on implicit assumptions on the relationships between parameters (Dzeroski & Drumm, 2003; Jeong et al., 2008; Junker et al., 2012; Kanevski et al., 2004). Thus, ML techniques are considered particularly useful in marine ecosystems, which are subject to stochastic and multi-dynamic phenomena often resulting in non-linearity (Olden et al., 2008).

Among the most frequently applied ML algorithms are Decision Trees (DTs), Neural Networks (NNs) including MultiLayer Perceptrons (MLPs), Support Vector Machines (SVM), Instance Based Learning (IBL) and Naïve Bayes (NB) classifiers (Kotsiantis, 2007). These algorithms represent the main ML categories (trees, functions, lazy and Bayes algorithms) that employ completely different predictive and classifying approaches (Solomatine et al., 2008). These span many applications in ecology (Dzeroski, 2001; Lek & Guegan, 1999; Recknagel, 2001) whereas in the marine environment they have been used in hydrodynamics, wave forecasting, habitat modelling, biomass prediction, and pollution assessment (e.g. Dakou et al., 2007; Etemad-Shahidi & Mahjoobi, 2009; Millie et al., 2012; Solomatine et al., 2006; Tian et al., 2011).

ML techniques have been successfully applied to phytoplankton biomass assessment focusing on the influence of different environmental conditions to chl *α* dynamics (Keiner & Yan, 1998; Zhan et al., 2003), eutrophication changes (Freeman et al., 2009; Karul et al., 2000; Kuo et al., 2007; Lamon, III et al., 2008; Scardi, 2003) and specific species abundance (Dzeroski, 2001; Dzeroski & Drumm, 2003; Kocev et al., 2010; Naumoski & Mitreski, 2010). However the applications of ML related to the classification of the quality status of coastal waters for the WFD purposes are sparse and have been accessed mainly with the training of NNs (Tison et al., 2007; Ocampo-Duque et al., 2007). Concerning biodiversity prediction in particular, application of ML techniques in both marine and terrestrial ecosystems has been based on habitat features, biotic characteristics or a combination of both with some abiotic parameters but never on abiotic variables alone (Cheng et al., 2012; Debeljak et al., 2007; Demsar et al., 2006; Dominguez-Granda et al., 2011; Dzeroski & Drumm, 2003; Knudby et al., 2010; Kocev et al., 2009; Pittman et al., 2007). These studies have also only focused on one biodiversity component (e.g. species richness or Shannon diversity) whereas so far there has been no attempt to predict different diversity components (richness, evenness, and dominance) exclusively from abiotic parameters related to the physical and chemical environment. Finally, ML techniques and specifically NNs and DTs have been used only in one occasion to classify the quality status in surface waters as required by WFD, providing impressive performance (Ocampo-Duque et al., 2007).

The increased interest in ML techniques has resulted in the development of numerous classifiers (Laniak et al., 2013) differentiated in supervised or unsupervised depending on whether the training dataset is labelled *a priori* or not (Laskov et al., 2005). Despite the variety of ML approaches, there is no optimal algorithm established so far. Instead, the classification performance depends on the different characteristics of the data analyzed (e.g. selection of input variables, number of training samples) (Chaudhuri & Bhattacharya, 2000; Lu & Weng, 2007) or the method used to assess algorithm performance (Baldi et al., 2000).

Current research on ML focuses on integrating optimal prediction or classification results from the individual base classifiers using specialized techniques called ensemble methods (EMs) (Opitz & Maclin, 1999; Wozniak et al., 2014). The latter provide significantly improved performance compared to the base classifiers (e.g. Assaad et al., 2008; Chen et al., 1997). Voting is a particularly useful and comprehensible EM that collects votes (i.e. predicted values or labels of the target class) from multiple individual algorithms and predicts the value or label of the output variable by combining their single results (i.e. for prediction tasks computes uses weighted MLR to compute the output numeric value or either for classification tasks yields the label with the highest value expressed as number of votes or probability). Regarding marine ecology, voting EM has been used only recently in order to model the influence of different environmental conditions on the abundance of specific organisms (Kocev & Dzeroski, 2013; Mouton et al., 2011). Other resent studies related to marine environment have applied the voting method in order to classify marine oil spills (Xu et al., 2014; Topouzelis & Psyllos, 2012), seaports (Halabi Echeverry et al., 2012) and coral reefs (Shihavuddin et al., 2013).

2 AIM AND OBJECTIVES

ML is a very promising technique for making progress in the understanding and prediction of ecological phenomena (Olden et al., 2008). In this study different ML algorithms were used in order to assess the complex issue of coastal marine eutrophication. Special effort was put on possible coupling of ML techniques and coastal management by developing effective predictive tools for WFD and ICZM. In this context, the application and adjustment of ML algorithms were refined aiming to meet the following objectives:

a) Assessment of the main processes that determine primary production in coastal marine ecosystems affected by terrestrial inputs.

To this aim (case study I), two different ML techniques were implemented: MTs and the popular NNs in order to prioritize abiotic parameters regulating primary production in coastal ecosystems affected by terrestrial runoff.

b) Derivation of the optimal classification scheme for coastal water ecological quality using exclusively abiotic parameters.

In case study II, ten different base classifiers were implemented and their results were then integrated for improving classification performance. A new index was proposed in order to specify which base classifiers should be integrated to offer optimal classification performance.

c) Optimization of the prediction of phytoplankton community structure exclusively from abiotic parameters in coastal ecosystems.

A number of different ML algorithms were trained using both natural assemblages and noise-free simulated assemblages (case study III) in order to effectively predict the richness, evenness and dominance of phytoplankton assemblages exclusively from abiotic parameters. Based on the optimal results of ML algorithm training, a software package was developed estimating phytoplankton diversity from four abiotic parameters.

3 METHODOLOGY

3.1 STUDY AREAS

The database used was compiled using existing datasets from five coastal areas in the Aegean Sea, Eastern Mediterranean representing a wide range of productivity (Fig. 1). All stations were sampled repetitively on a monthly basis covering at least a full annual cycle. Nutrient concentrations were measured spectrophotometrically according to Parsons et al. (1984), whereas physical variables were recorded *in situ*. Moreover, available phytoplankton species-abundance data were used, analysed following the same protocol according to the inverted microscope method of Utermohl (1958).

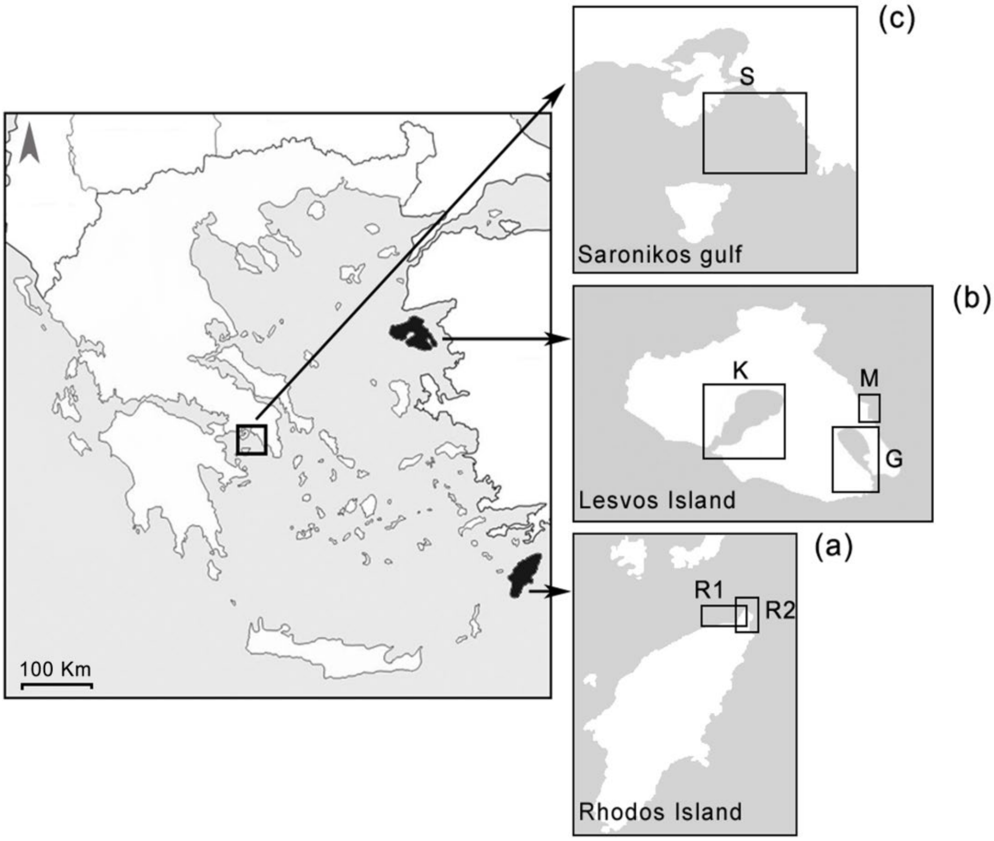


Figure 1: Maps of the five coastal areas: (a) Rhodos R1 and Rhodos R2 in the island of Rhodos, (b) Gera G and Kalloni K in the island of Lesvos, and (c) Saronikos gulf S near the metropolitan area of Athens (Spatharis et al., 2008)

From the study areas the Inner Saronikos Gulf, near Athens, and the Kalloni Gulf in Lesvos Island are characteristic of eutrophic conditions (Simboura et al., 2005). Outer Saronikos Gulf and Gera Gulf in Lesvos Island are more typical of mesotrophic conditions (Arhonditsis et al., 2000; Ignatiades et al., 1992), while offshore stations in Rhodes Island have been characterized as oligotrophic (Kitsiou et al., 2002). Detailed information about the sampling sites and data collection are provided in Spatharis et al. (2008) while an account on the eutrophication level and ecological status of these areas is provided in Spatharis & Tsirtsis (2010).

3.2 ALGORITHM DESCRIPTION

ML techniques can be used for various applications including classification and prediction (Witten & Frank, 2005). Depending on whether the output variable is categorical or numerical, ML includes algorithms that can be used exclusively for classification tasks (i.e. classifiers), others that can be used only for prediction (i.e. predictors) and a few algorithms that can be used for both tasks (Table 1). In this study different ML algorithms were used for (a) phytoplankton biomass prediction (case study I), (b) water quality status classification (case study II) and (c) phytoplankton diversity prediction (case study III).

The algorithms used in this study belong to all main ML categories such as rules, trees, lazy algorithms, functions, Bayes and meta algorithms (Table 1). Algorithms that represent each category use different approaches in order to classify or predict the value of the output variable on new unseen instances.

More specifically, rule algorithms construct rules based on disjunctions of the form“ IF … THEN …” (Frank & Witten, 1998) such as:

- IF (blood type=warm)^(eggs=yes) THEN class=bird
- IF (income<5000)^(pension=yes) THEN tax=no

The goal of rule based algorithms is to construct the smallest set of rules that is consistent with the available dataset. Thus, a large number of rules means that the rule algorithm is rather reproducing the data (i.e. overfitting), than discovering the main assumption that governs it (Kotsiantis, 2007).

Trees are conceptual schemas consisting of different paths that are followed according to comparisons on one or more input variables. Each tree path ends to a specific leaf in which the final classification or prediction of the output variable is being made (Kothari & Dong, 2001). Different tree based algorithms exists depending on (a) the tree construction method and (b) the way that the instances of each leaf are combined in order to arrive at the final classification or prediction.

Another famous ML category is the lazy learning algorithms, which postpone the induction process until classification or prediction is performed. The lazy category contains algorithms that are based on the principle that instances within a dataset generally exist in close proximity to other instances that have similar properties (Aha et al., 1991). These similar instances are properly used to provide the final prediction of the requested output label or value.

Function category, as highlights its name, contains algorithms that can be written down as simple or more complex mathematical equations in a reasonably natural way (Witten & Frank, 2005). This category includes classical statistical methods such as linear or logistic regression models. Substantially, NNs like MLPs or radial basis function networks, which are the most popular ML methods, belong to the function category.

Bayes consists of statistical algorithms that incorporate probabilities to classify the output variable. Bayes category contains algorithms that incorporate the famous Bayes rule and by assuming independence are computing the probabilities for every label of the output variable. Afterwards, these probabilities are compared to indicate the label that is the most likely to be the actual one (Aguilera et al., 2011).

Finally the meta algorithms use specialized techniques trying to improve the final performance of existing algorithms by integrating their results (Kotsiantis et al., 2006). Although meta algorithms are relatively newly proposed techniques, they are popular and span numerous applications often showing that are much more accurate than any of the single algorithms participating in them (Opitz & Maclin, 1999).

Table 1: The ML techniques used in the study

Category	Abbreviation	Description	Reference	Classifier	Predictor	Case study used
Rules	**RIPPER**	Implements the repeated incremental pruning to produce error reduction	(Cohen, 1995)	✓		II
	PART	Generates a partial decision list	(Frank & Witten, 1998)	✓		II
Trees	**J48**	Generates a pruned C4.5 decision tree	(Quinlan, 1993)	✓		II
	RF	Constructs a forest of random trees.	(Breiman, 2001)	✓		II
	MTs	Generates a tree with linear regression models at the leafs	(Quinlan, 1992)		✓	I, III
Lazy	**IBk**	Implements k-nearest neighbors method	(Aha et al., 1991)	✓	✓	II, III
	Kstar	Instance based learner with entropic distance measure	(Cleary & Trigg, 1995)	✓	✓	II
Functions	**Log**	Multinomial logistic regression	(le Cassie & van Houwelingen, 1992)	✓		II
	SMO	Implements sequential minimal optimization for training a support vector learner	(Platt, 1999)	✓		II
	MLP	Multilayer perceptron trained with back-propagation	(Pal & Mitra, 1992)	✓	✓	I, II, III
	MLR	Multiple linear regression	(Zar, 1984)		✓	I, III
Bayes	**NB**	Naïve Bayes classifier using estimator classes	(John & Langley, 1995)	✓		II
Meta	**Voting**	Ensemble method for combining learners using probability estimates	(Kittler et al., 1998)	✓	✓	II

3.2.1 RIPPER rule classifier

Repeated Incremental Pruning to Produce Error Reduction (RIPPER) is a rule learner classifier introduced by Cohen (1995) as a successor to Incremental Reduced Error Pruning (IREP) algorithm (Furnkranz, 1997). RIPPER begins the learning process by sorting (in ascending order) the training data by the output class labels beginning with the less frequent one. Thereafter RIPPER starts producing a set of rules, one at time, through two steps: growth and pruning. In the iterative growth phase, a rule is constructed to match as many instances of the minority label class (i.e. the less frequent) as possible while those instances are removed from the training set (Huhn & Hullermeier, 2009). The learner keeps producing rules in the same way until all remaining training instances belong to one single class (i.e. the last and the more frequent). Then a final default rule is added to the previous ones and the procedure ends. To prevent the produced rules from overfitting (i.e. situation where they become too specific for the training data), the pruning step eliminates conditions from the rules that do not harm the classifier's accuracy (Lorena et al., 2011). More details about the RIPPER's rule construction can be found in the Table 2.

Table 2: Logical steps of RIPPER's rules construction

RIPPER's rule classifier (for multi-labeled class problem)

1. Order instances by the label of the target class in increasing prevalence (fraction of instances that belong to a particular class label)
2. Use instances that have the less frequent label to learn the rule set and treat the rest instances as belonging to the negative class.
   For the construction of a single rule follow the steps
   a) Start from empty rule
   b) Add conjuncts as long as they improve information gain
   c) Stop when the rule no longer covers negative examples
      (accuracy achieves100%)
   d) Prune the rule using reduced error pruning
   e) Remove the instances covered by the rule
3. Repeat using instances that have the next less frequent label of the target class (treat them as positive class)

RIPPER classifier has the advantages of being (a) interpretable as it produces a set of symbolic rules, (b) flexible as new rules can be added or modified as new data are included to the database and (c) quick as it runs in linear time (Cohen & Singer, 1999). However, RIPPER has rarely been used generally in biology, having few applications related to genetics and ecology (Libralon et al., 2009; Lorena et al., 2011; Khater & Gras, 2012). In the marine environment it has been used once in order to determine the sex mechanism of a fish species in aquaculture (Palaiokostas et al., 2013).

3.2.2 PART rule classifier

PART is a rule based ML technique constructed by Frank & Witten, 1998 in order to avoid global optimization environment in which previous rule classifiers (e.g. RIPPER) used to perform, because such techniques cannot deal with problems that have many local optima (either maximal or minimal). Thus, the PART learner generates compact rule sets by combining two popular methods i.e. “separate and conquer” and “divide and conquer” (Tan et al., 2003). PART follows the same procedure as RIPPER to construct the first rule (separate and conquer method) followed by the removal of covered instances. Substantially, PART continues constructing rules recursively by generating a partial decision tree (i.e. not fully inducted) from the remaining instances of the database (divide and conquer method). The leaf of the tree with the largest coverage is converted into a rule and the tree is discarded. The analytical steps of PART classifier can be found in the Table 3.

Table 3: Logical steps of PART rule classifier

PART rule classifier (for multi-labeled class problem)
1. Build a partial decision tree on the current set of instances (for more details see Table 4)
2. Create a rule from the decision tree using the leaf with the largest coverage
3. Discard the decision tree
4. Remove the instances covered by the rule
5. Go to step one

The combined method of PART, adds flexibility and speed to the classifier while protects it from over pruning (Frank & Witten, 1998). Moreover PART maintains the essential advantage of rule classifiers offering a set of simple and comprehensible rules which contain only the crucial input variables in a scaling way. The latter can help towards the interpretation of the procedures related to the desired issue by giving new insights to it (Bibi et al., 2008). However, PART and generally the rule classifiers usually achieve medium accuracy performances and thus are considered as simple classifiers (e.g. Herrera et al., 2002; Bhasin & Raghava, 2005). PART has never been applied to marine or coastal environment.

3.2.3 Decision Tree J.48

J.48 classifier (Witten & Frank, 2005) is an open source Java re-implementation of the most popular algorithm for decision tree induction called C4.5 (Table 4) (Quinlan, 1993). A decision tree is a hierarchical structure consisting of nodes (i.e. a root, inner nodes and leaves) and branches (Fig 2). The root and the inner nodes contain tests on input variables, while leaves comprise the predicted label of the output variable. The branches connect the nodes, starting from the root or an inner node and ending in another internal node or a tree leaf (Quinlan, 1996).

Table 4: Logical steps of decision tree construction

Decision tree construction for a categorical output variable with *c* labels
1. Create a root node for the tree
2. If all instances have the same label of the output variable then return a single node tree root with that label.
3. If there are no input variables then return a single node tree root with the most common label among instances.

4. Otherwise
    a. Select the input variable A that best classifies the instances as defined by

    $$Gain\ (S, A) = Entropy(S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

    $$Entropy(S) = \sum_{i=1}^{c} -p_i log_2 p_i$$

    Where S is the collection of all training instances, S_v is the subset of S for which variable A has value v, c is the number of the labels of the output variable, p_i is the proportion of S belonging to the label i, $Values(A)$ is the set of all possible values $v_i$of variable A and $\frac{|S_v|}{|S|}$ is the fraction of examples that belong to S_v.

    b. Create tree root with the above variable A

    c. For each possible value v_i of the variable A
        - add a new tree branch bellow the tree root corresponding to the test $A = v_i$
        - Let S_{v_i} be the subset of the S that have value v_i for A
        - If S_{v_i} is empty, then below this new branch add a leaf node with the most common label of the output variable in S, else below this new branch add the subtree constructed with the same procedure and has S_{v_i} for S and possible splitting variables all the remaining variables except A.

5. Return the Root

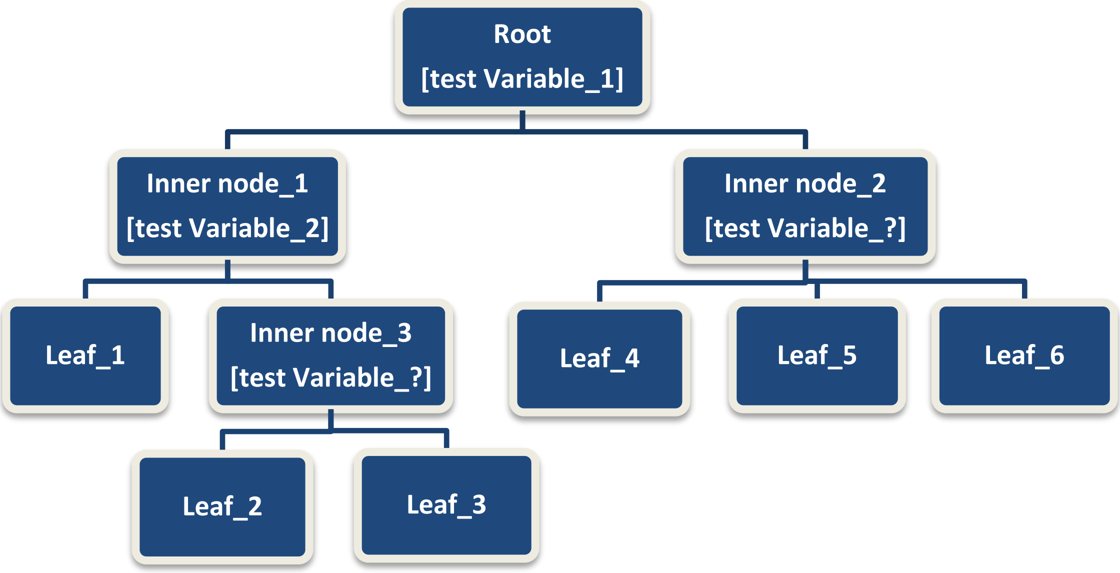


Figure 2: Diagram of a tree learner

In order to construct a new node, J.48 computes the gain of all possible splits (using a proper entropy measure) and substantially chooses the input variable with the highest gain ratio. The variables that can participate in node splitting can be either numeric or categorical. Thus, if J.48 chooses a numeric variable for this purpose then the node is divided into two branches using a proper inequality (e.g. input variable > constant value). The one of the two branches is followed when the inequality is being satisfied and the other when not. On the other hand, if the chosen splitting variable is categorical taking *n* discrete labels, the node is also divided into *n* branches i.e. one for each categorical label (Loh, 2008). The J.48 procedure is repeated until all instances are correctly classified, however it usually results to an extremely large tree (i.e. lots of nodes). The tree complexity and the resulting overfitting are being dealt through the popular tree-pruning method that reduces the tree size and enhances the classification accuracy (Quinlan, 1999).

J.48 classifier offers an interpretable extraction of hidden patterns even when dealing with long-term multivariate datasets and thus it has been used in many different classification tasks (Kothari & Dong, 2001). In the marine environment, J.48 tree induction has been applied in several studies offering sufficient classification results and new insights. Some of the most recent studies dealing with classification trees concern the identification of the factors affecting zooplankton community (Gal et al., 2013), the variation of sea water quality (Chen et al., 2010), the impact of exotic species on lakes (Everaert et al., 2011), the ciliate foraging behavior (Chang et al., 2011) and the sustainable flood management of water basins (Yang et al., 2011).

3.2.4 Model trees (MTs)

MTs are constructed using a decision tree induction algorithm (Table 4) in order to predict the value of a numeric output variable by storing a multiple linear regression equation at each leaf (Quinlan, 1992). Initially, the MT is constructed based on a criterion that determines which input variable best discriminates the input samples in distinct homogeneous subsets (nodes or leaves) (Fig. 2). For numeric prediction the criterion intends to minimize the intra-subset variation of the predicting variable down each branch (Barros et al., 2011; Witten & Frank, 2005). MT construction terminates when the variance of the predicted values in a subset is sufficiently small (Frank et al., 1998). Once the final homogeneous subsets have been defined (tree leaves) a linear regression model (LM) is constructed from data contained within each subset/leaf. This LM predicts the values of the output variable based on a number of selected input variables. Given a new sample for which the output variable's value should be predicted, the prediction procedure initiates from the tree root (the first discriminating variable). In each inner node a decision test is made to follow a particular branch based on the discriminating variable associated with that node (Quinlan, 1996). Finally, when the sample is classified into a subset/leaf, then the output value is predicted according to the corresponding linear regression model.

MTs are used to approach pattern prediction and hierarchical problems in various research fields. Applications can be found in medical science (Shao et al., 2007), quality management (Srdoc et al., 2007), agriculture (Debeljak et al., 2007; Kocev et al., 2009), water management (Bhattacharya & Solomatine, 2005) and wave forecasting (Bonakdar & Etemad-Shahidi, 2011; Etemad-Shahidi & Mahjoobi, 2009; Jain et al., 2011). Although MTs have been regularly implemented in terrestrial ecology for describing and modeling population dynamics (Demsar et al., 2006; Jurc et al., 2006; Ogris & Jurc, 2010; Stankovski et al., 1998), their applications related to modeling ecological processes in aquatic ecosystems are restricted. These studies have focused on the influence of environmental conditions on diatom assemblage abundance (Kocev et al., 2010; Naumoski & Mitreski, 2010), the effect of physical and biological factors on the spatial distribution of a sea cucumber (Dzeroski & Drumm, 2003), changes in biomass of algal species (Dzeroski, 2001), and phytoplankton dynamics of N. Adriatic Sea (Volf et al., 2011). An application for chl *α* prediction using MTs was based on a Bayesian approach to provide classification schemes of various water reservoirs characterized by different geographic, morphometric, and chemical properties (Freeman et al., 2009; Lamon, III et al., 2008).

3.2.5 Random Forest (RF)

Random Forest (RF) is an ensemble classifier developed by Breiman (2001) that combines the results of individual tree classifiers participating in the forest (Figure 3). Each of these trees is build using a bootstrap sample of the dataset while at each tree node only a small random subset of the input variables is available for the tree branch binary split. The constructed trees remain unpruned (i.e. fully grown) to ensure low-bias (i.e. flexibility in data fitting). Substantially, RF classifier yields an ensemble using majority voting over the individual tree classification results in order to predict the status of the output class (Diaz-Uriarte & Alvarez de Andres, 2006). The performance of the RF depends on the number of the trees that consist the forest, the performance of the individual trees and the correlation between their results. Applications to ecology have shown that RF can effectively model complex and non-linear relationships offering high classification accuracy and determination of the input variable importance (Cutler et al., 2007).

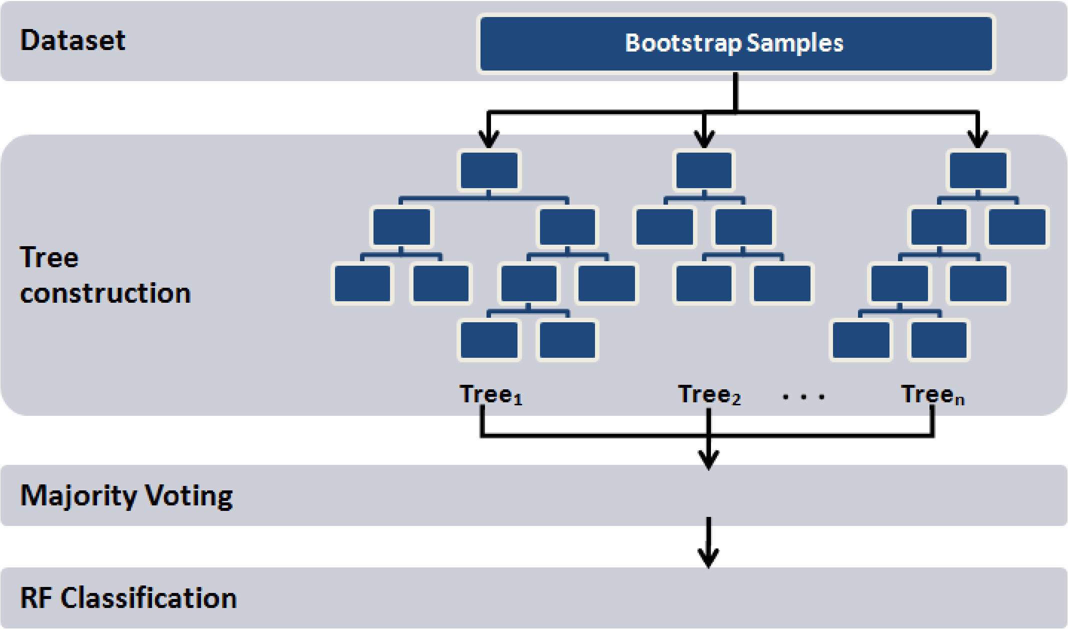


Figure 3: Conceptual diagram of the random forest classifier

Within a relative short period of time, RFs have been successfully applied to numerous classification tasks in a variety of fields, indicating their efficient performance in comparison with other ML techniques (Verikas et al., 2011). Specifically in the marine environment, RFs have been so far used to assess the mapping of fish species richness (Knudby et al., 2010), the flux of benthic light under toxic conditions (Kehoe et al., 2012), the sources of water fecal contamination (Smith et al., 2010), the discrimination of fish population (Perdiguero-Alonso et al., 2008) and the density of bacteria in water (Parkhurst et al., 2005). During the last years RF has been also sufficiently tested in different ecological tasks (Crisci et al., 2012; Cutler et al., 2007; Prasad et al., 2006) but span very few studies related to marine eutrophication mainly under a management perspective (Catherine et al., 2010; Bergstrom et al., 2013).

3.2.6 Lazy Instance Based Learner IBk

IBL algorithms are derived from the nearest neighbor pattern classifier (Cover & Hart, 1967) and are based on the idea that similar instances have similar behavior (Payne, 1995) thus the new input instances are predicted according to the stored most similar neighboring instances (Table 5) (Aha et al., 1991). The nearest neighbor classifier (*k*-NN) is one of the simplest and oldest methods to perform classification tasks (Solomatine et al., 2006). It has been used in various applications yielding excellent performances (Tsekouras 2005; Huang, 2006). IBL algorithms are also known as lazy learning algorithms since they simply store the training instances and postpone all effort until prediction time.

Table 5: Logical steps of IBL

Instance base learning (*k* neighbors)
1. For a new unseen instance, compute the distance metric between this instance and all stored training instances of the dataset
2. Define the *k* instances that have the corresponding lowest distance values (set of the nearest neighbors)
3. Compute the final prediction as the mean of the *k* values that the output variable has in the set of the defined set of the nearest neighbors (numeric output variable) or Estimate the label of the class using majority voting for the *k* labels that the output variable has in the set of the defined set of the nearest neighbors (categorical output variable)
4. Go to step one

The *k*-NN algorithm treats the input variables as dimensions of a Euclidean space and the instances as points in this space (Cover & Hart, 1967). Once a new unseen instance is given, a distance metric between this instance and all stored training instances is calculated and the *k* nearest instances are been defined. Many different distance metrics have been proposed but mostly used are:

$$\text{Euclidean:} \quad d(X,Y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$

$$\text{Manhattan:} \quad d(X,Y) = \sum_{i=1}^{m} |x_i - y_i|$$

$$\text{Chebychev:}\quad d(X,Y) = max_{i=1}^{m}|x_i - y_i|$$

where $X = (x_1, x_2, \dots, x_m)$ and $Y = (y_1, y_2, \dots, y_m)$ are two instances of a dataset that has m input variables.

Then, the prediction of the output variable is estimated as the mean of k values that the output variable has in the set of the defined nearest instances. The k-NN algorithm can be improved by weighing each of the k nearest neighbors (X_i) (Wettschereck et al., 1997) according to their distance $d(X_q, X_i)$ from the new query point (X_q) based on the following two functions:

$$f(X_q) = \frac{\sum_{i=1}^{k} w_i f(X_i)}{\sum_{i=1}^{k} w_i}$$

where w_i is a function of the distance $d(X_q, X_i)$ with the following two weight functions being commonly used:

$$w_i = 1 - d(X_q, X_i) \quad (Linear)$$

$$w_i = \frac{1}{d(X_q, X_i)} \qquad (Inverse)$$

IBk is a popular ML technique already applied either as predictor or classifier in few studies related to the marine environment in order to assess hydrologic and wave modeling, sea water quality or marine species habitat preference (e.g. Dzeroski & Drumm, 2003, Hatzikos et al., 2008; Solomatine et al., 2008; Zamani et al., 2008).

3.2.7 Lazy KStar

KStar is an instance based algorithm proposed by Cleary & Trigg (1995), operating either as classifier or predictor and able to handle both numerical and categorical input variables. The difference of KStar in relation to the classic IBk algorithm is that the former uses a different approach to calculate the distance between instances, based to an entropy measure (Morrison et al., 2007). This entropy measure has been inspired from information theory and can be defined as the complexity of transforming one instance into another. More specifically KStar defines a finite set of transformations in order to map instances to instances. Substantially, finite sequences of transformations starting from an instance and terminating to another are defined covering all instance combinations. Finally, the entropy measure is estimated as the length of the shortest sequence connecting two instances. This entropy measure assessment makes KStar algorithm much more general and greedy than the classic IBk, especially when dealing with missing values (Yucel & Ozel, 2012). Thus, when using KStar it is considered that each instance exerts a “sphere of influence” with soft boundaries rather than the hard edged cutoff implied by the k-NN rule in which any particular instance of the dataset either participates or not to the final prediction (Witten & Frank, 2005).

Although KStar is not so popular compared to IBk, it has been applied in various studies with good results (e.g. Rocha et al., 2007; Grabar & Krivine, 2007; Uygun et al., 2010). In the coastal environment it has been used once to assess biomass of mangroves i.e. type of trees that grow in saline coastal sediment habitats (Jachowski et al., 2013).

3.2.8 Multinomial Logistic Regression (MLR)

Logistic Regression (Log) is a statistical method used in classification to predict the outcome of a categorical variable (i.e. target class) based on input variables that can be either numerical or categorical. According to the total number of categories (i.e. labels) that the target class owes, the logistic regression is called binary if the number of labels is two (e.g. “male=1” vs “female=0”) or multinomial if this number is larger.

During binary logistic regression, coefficients (as long as its standard errors and significance levels) are generated in order to predict a logit transformation of the probability of the occurrence of a situation (recorded with label “1” vs the other label “0”).

$$logit(p) = \ln\frac{p}{1-p} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

where p is the probability of the occurrence of a situation (usually the presence of a characteristic of interest in biological studies), b_i are the linear regression coefficients estimated using maximum likelihood (McCullagh & Nelder, 1989) and x_i are the k independent input variables. Thus, the general multiple logistic regression model in terms of p is:

$$p = \frac{e^{b_0+b_1x_1+b_2x_2+\cdots+b_kx_k}}{1+e^{b_0+b_1x_1+b_2x_2+\cdots+b_kx_k}}$$

Multinomial logistic regression is a simple extension of the binary one.

Logistic regression is commonly used in several environmental tasks (e.g. Pearce & Ferrier, 2000; Keating & Cherry, 2004). More specifically, logistic regression has been used to assess marine eutrophication tasks such as toxic diatom blooms estimation (Lane et al., 2009), eutrophic classification of hypoxic waters (Lowery, 1998), species presence-absence along with different environmental factors (Bini & Thomaz, 2005) or sea grass pattern modeling (Fonseca et al., 2002).

3.2.9 Sequential Minimal Optimization (SMO)

Sequential minimal optimization (SMO) implements a method proposed by Platt (1999) that trains a support vector machine (SVM) classifier using polynomial kernels. A normal SVM tries to solve a quadratic programming problem that is expressed in the dual form as follows:

$$max_a W(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j K(x_i, x_j) a_i a_j$$

subject to:

$$0 \leq a_i \leq c \forall i = 1,2,\ldots,n$$

$$\sum_{i=1}^{n} y_i a_i = 0$$

where a_i are Lagrange multipliers, n is the number of training instances (i.e. examples), x_i is the input variables vector, y_i is the output label of the target binary variable with $y_i \in \{-1,+1\}$, $K(x_i, x_j)$ is the kernel function selected by the user, and c is an appropriate parameter.

Essentially, training a SVM involves large matrix operations that solve the above n^{th} (n is equal to the number of training instances) dimensional quadratic problem. However, if the training set is too large, the SVM requires a lot of computational effort (memory and time) making the algorithm very slow and impractical (Keerthi et al., 2001). The SMO solves the quadratic problem by decomposing it into smaller problems, each one being a reduced problem of the quadratic one that can be described as follows:

$$0 \leq a_1, a_2 \leq \mathrm{c}$$

$$\mathrm{y}_1 a_1 + \mathrm{y}_2 a_2 = \mathrm{k}$$

During training process and for each iteration, SMO proceeds as follows: (a) picks a pair of Lagrange multipliers to optimize the solution of a smaller quadratic programming problem and (b) repeats the same process until it converges on a solution. The advantage of SMO is that the solution for two Lagrange multipliers can be done analytically and thus an entire inner iteration is avoided. Even though more sub-problems are solved during SMO training, each solution is so fast that the overall optimization is achieved rather quickly. Additionally, SMO requires small data storages as it stores only the required 2x2 matrix for each iteration (Platt, 1999). However, the SMO classifier is binary and in case of a multi-class problem (i.e. target class with more than two labels), it must be reduced to a set of multiple binary classification problems (Crammer & Singer, 2002).

SMO classifier is easy to implement and has already yielded excellent generalization performance on a wide range of problems (Keerthi et al., 2001). In the marine environment it has been used to predict water quality (Hatzikos et al., 2008), to monitor seagrass population (Musavi et al., 2007), to estimate an aquatic fern species distribution (Sadeghi et al., 2012), and to retrieve chlorophyll concentration from remote sensing (Haigang et al., 2003).

3.2.10 Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) is a statistical approach to model the relationship between a numeric dependent output variable Y and more than one explanatory input variables X_i. Given a dataset $\{X_{1j}, X_{2j}, \ldots, X_{mj}, Y_j\}$ containing m input variables and n instances, the linear model takes the form:

$$Y_j = a + b_1 X_{1j} + b_2 X_{2j} + \cdots + b_m X_{mj} + e_j, \qquad j = 1, \ldots, n$$

The parameters $b_1, b_2, \ldots, b_m$ are called partial regression coefficients and express how much Y would change for a unit change of each input variable.

The intercept a, is the value of Y when all input variables X_i are zero. The residual or error e_j is the amount by which y_j differs from what is predicted by $\hat{Y}_j = a + b_1 X_{1j} + b_2 X_{2j} + \cdots + b_m X_{mj}$. Note that the sum of all e's is zero (Zar, 1984).

The criterion for defining the best fit (i.e. optimal $a, b_1, b_2, \ldots, b_m$) of the MLR equation is the minimum residual sum of squares i.e. the minimum value of $\sum_{j=1}^{n}(Y_j - \hat{Y}_j)^2$ (Flury & Riedwyl, 1988).

3.2.11 Multilayer Perceptron (MLP)

MLP is an artificial neural network that maps input instances onto values or labels of the output variable. A MLP architecture consists of one or more layers of nodes (neurons) between the input and output layers in a directed graph (feedforward), while each layer is fully connected with weighted connections to the next one (Fig.4) (Lek & Park, 2008). The input layer typically contains as many neurons as the number of the input variables; the hidden layer has a number of neurons which can be selected arbitrarily or determined empirically, while the output layer has usually one neuron referring to the output variable.

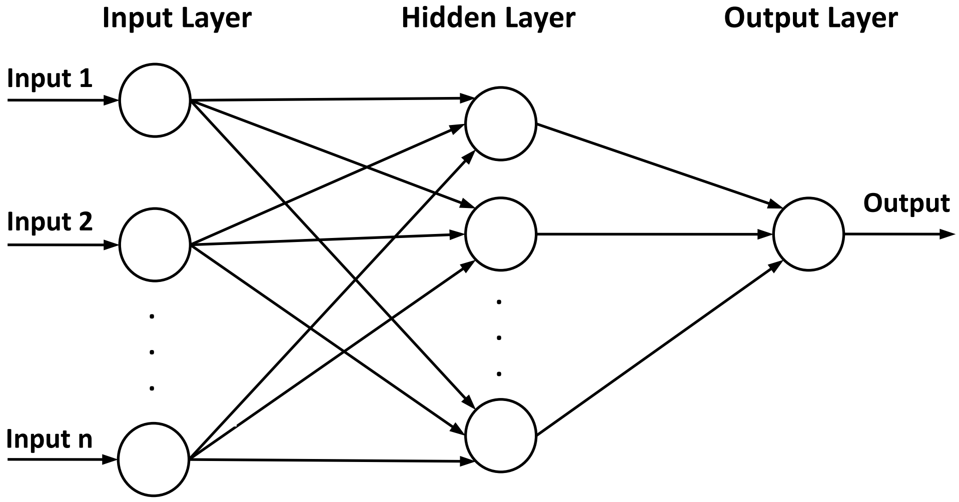


Figure 4: The classical MLP architecture consisting by three layers of neurons

Technically, each neuron receives weighted input signals which are used as a sum to feed an activation function for producing an output signal that substantially activates the neurons of the next layer (Table 6) (Lek & Guegan, 1999). During the training phase a set of instances (having values for both the input and output variables) is presented to the MLP again and again. The MLP is being trained by an update procedure based to the simple concept: if the network gives an insufficient response, the connection weights are corrected so that the error is reduced and future responses of the network are more likely to be closer to the real wishing outputs (Olden et al., 2008).Thus, the information hidden in the input data flows within the network from the input to output layer in order to improve the MLP's predictive performance. More details about the MLP training technique can be found in the Table 6.

Table 6: Logical steps of MLP (Lek and Guegan, 1999)

Feed-forward MLP training by back-propagation algorithm with the use of sigmoid activation function

1. Initialize the number of hidden nodes
2. Initialize the maximum number of iterations and the learning rate (η). Set all connection weights $W_{ij}^{h}$and thresholds to small random numbers. Thresholds are weights with corresponding inputs always equal to 1.
3. For each training instance (input X_p=(x_1, x_2, …, x_n), output Y) repeat steps 4-7.
4. Present the input X_p to the input nodes and the output Y to the output node;
5. Calculate the input to the hidden nodes: $a_j^h = \sum_{l=1}^{n} W_{ij}^h x_l - \theta_j$

   Calculate the output from the hidden nodes: $x_j^h = f\left(a_j^h\right) = \frac{1}{1+e^{-a_j^h}}$

   Calculate the inputs to the output nodes: $a_k = \sum_{j=1}^{L} W_{jk} x_j^h - \theta_k$

   Calculate the output from the output nodes: $\hat{Y}_k = f(a_k) = \frac{1}{1+e^{-a_k}}$

   If the network has a single output and one hidden layer then: k = 1, $\hat{Y}_k = \hat{Y}$

   L is the number of nodes of the hidden layer, θ_j, θ_k are the thresholds
6. Calculate the error term for the output node: $\delta_k = (Y - \hat{Y}_k) \cdot f^{'}(a_k)$

   Calculate the error for the hidden nodes: $\delta_j^h = f^{'}(a_j^h) \cdot \sum_k \delta_\kappa W_{jk}$

f' is the derivative of the sigmoid function

7. Update weights on the output layer: $W_{jk}(t+1) = W_{jk}(t) + \eta \delta_\kappa x_j^h$
   and on the hidden layer: $W_{ij}(t+1) = W_{ij}(t) + \eta \delta_j^h x_i$
   as long as the network errors are larger than a predefined threshold or the number of iterations is smaller than the maximum number of iterations envisaged, repeat steps 4-7.

All MLPs used in this study belong to the classic group of feed-forward neural networks with one hidden layer in which sigmoid activation function is used to all neurons while it is trained by the backpropagation algorithm (Rumelhart et al., 1986).

Among ML algorithms, NNs including MLPs are the most commonly used and span numerous and various applications (Bhattacharya & Solomatine, 2005; Tsekouras & Tsimikas, 2013). In the marine environment MLPs has been used in eutrophication modeling (Karul et al., 2000; Kuo et al., 2007), wave forecasting (Altunkaynak, 2013; Etemad-Shahidi & Mahjoobi, 2009), biomass prediction (Musavi et al., 2007; Scardi, 1996) and pollution assessment (Tian et al., 2011; Topouzelis et al., 2008).

3.2.12 Naïve Bayes (NB)

A Naïve Bayes (NB) classifier is a probabilistic method based on the Bayes rule in combination with the independence assumption (Naïve) of the input variables (Lewis, 1998). The NB classifier assigns every new instance $E = (x_1, x_2, \dots, x_n)$ into a class label c of the output target variable C. According to Bayes rules the probability of an instance E to belong to class c is:

$$p(c|E) = \frac{p(E|c) \cdot p(c)}{p(E)}$$

By assuming that all input variables (categorical or numeric) are independent given the label of the output class, the conditional probability $p(E|c)$ can be calculated as:

$$p(E|c) = p(x_1, x_2, \dots, x_n|c) = \prod_{i=1}^{n} p(x_i|c)$$

Combining the above two notions, the Naïve Bayes classifier picks the label of the class that is the most probable to happen (maximum *a posteriori* decision rule). As a result the NB function can be written:

$$f_{NB}(E) = f_{NB}(x_1, x_2, \dots, x_n) = argmax_c p(c) \prod_{i=1}^{n} p(x_i|c)$$

Note: $p(E)$ is a constant for every category (Peng et al., 2004) and $argmax_c$ returns the label of the output class with the maximum probability.

If an input variable x_i is numerical then the method uses the variable's mean μ_c and variance σ_c^2 for each class label of the output variable. Then the probability density of a value v given a class label c can be computed as follows:

$$p(x_i = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

Although NB classifier is a popular machine learning technique (Lewis, 1998), it has been rarely used to classification assessments in environmental modelling (Aguilera et al., 2011). Only recently, the NB classifier has been applied in the marine environment in order to estimate the phytoplankton structure and composition, to map the seafloor using image data and to predict fish recruitment in fisheries management (Fernandes et al., 2010; Ludtke et al., 2012; Zarauz et al., 2009).

3.2.13 Voting

Current research on ML focuses on integrating optimal classification results from the individual base classifiers using specialized techniques called ensemble methods (EMs) (Opitz & Maclin, 1999; Wozniak et al., 2014). The latter provide significantly improved classification performance compared to the base classifiers (Assaad et al., 2008; Chen et al., 1997).

Voting is a particularly useful and comprehensible EM that collects votes (i.e. predicted labels of the target class) from multiple individual classifiers and predicts the label of the target class yielding the highest value expressed as number of votes (simple majority voting). One refinement on simple majority voting, weights the participating classifiers by using probability estimates than just a simple classification decision. Using the "average probability" method during voting, for every new instance $E = (x_1, x_2, \ldots, x_n)$, the corresponding class-probability estimate is calculated as follows:

$$p(f(E) = c) = \frac{1}{L}\sum_{l=1}^{L} p(f(E) = c|h_l)$$

where n is the number of input variables,c states for every label of the output variable, L is the number of base classifiers $h_1 \ldots h_L$ participating in the voting schema, and finally $p(f(E) = c|h_l)$ is the probability that the true (i.e. correct) label is c. Note that the predicted label of Eresulted from the above equation for all labels, is the one with the highest computed probability.

Voting is the most widely applicable EM method, as other EMs (including bagging and boosting) employ voting approaches in order to provide their own final outcome (Bauer & Kohavi, 1999; Dietterich, 2000a). Voting is also the simplest and easiest way to combine classifiers (Tan & Gilbert, 2003), demanding no extra training except when applying the voting scheme (Dzeroski & Zenko, 2004). For these reasons, voting spans many applications ranging from simple classification tasks (Saha & Ekbal, 2013; Srinivas et al., 2009) to more complex implementations such as clustering (Dimitriadou et al., 2001), pairwise comparison (Loza Mencia et al., 2010) and fuzzy systems (Ishibuchi et al., 1999; Kaburlasos & Pachidis, 2014).

3.3 ALGORITHM EVALUATION

3.3.1 Cross Validation

Cross Validation (CV) is a popular technique for estimating the error of algorithm predictions. CV is efficient for datasets containing neither few (few tens) nor too many (tens of thousands) records (Stone, 1978) providing a nearly unbiased estimate using exclusively original data (Efron, 1983). The main advantage of this method is that it protects the system from overlearning (i.e. overfitting) and for this reason it is more commonly used in data analysis (Witten and Frank, 2005).

In K-fold CV the dataset is randomly partitioned into K subsamples, K minus 1 of which are used as training data while the remaining subsample is retained for testing the algorithm. This process is repeated K times (the folds) and results are averaged to produce the performance estimation. Leave-One-Out (LOOCV) is a specific category of CV, in which the parameter K is equal to the number of instances of the dataset. During LOOCV, a single instance is used for the validation of the algorithm and the remaining instances are used for training. Thus, the same procedure is repeated as many times as the number of instances of the dataset and then the results produce the overall algorithm performance (Cawley & Talbot, 2003).

3.3.2 Measures of performance

Three measures of performance were considered in order to evaluate the numeric prediction (i.e. prediction of chl a or phytoplankton diversity) of the algorithms (Table 7). These measures are: a) the correlation coefficient (R) which measures the statistical correlation between predicted and observed values, b) the mean absolute error (MAE) which averages the magnitude of the differences between predicted and observed values ignoring their sign and c) the root mean squared error (RMSE) which represents the standard deviation of the above differences (Witten & Frank, 2005).

Table 7: Measures of performance used to evaluate the algorithm's numeric prediction

Measures of performance		
Name	Abbreviation	Formula
Correlation coefficient	R	$\frac{S_{PA}}{\sqrt{S_P S_A}}$ $where\ S_{PA} = \frac{\sum_{i=1}^{n}(p_i - \bar{p})(a_i - \bar{a})}{n-1},$ $S_P = \frac{\sum_{i=1}^{n}(p_i - \bar{p})^2}{n-1},\ S_A = \frac{\sum_{i=1}^{n}(a_i - \bar{a})^2}{n-1}$ $\bar{p} = \frac{\sum_{i=1}^{n} p_i}{n}, \bar{a} = \frac{\sum_{i=1}^{n} a_i}{n}$
Mean Absolute Error	MAE	$\frac{\sum_{i=1}^{n} \lvert p_i - a_i \rvert}{n}$
Root Mean Squared Error	RMSE	$\sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}$

where p_i are the predicted values of the algorithm, a_i are the actual values according to the dataset and n is the number of instances in the dataset.

Furthermore, in classification tasks the principal measure of a classifier's performance is the percentage of the correctly classified instances over the total number of instances in the dataset (CCI). Another measure is the Cohen's kappa statistic (*κ*) (Cohen, 1960) which is calculated as the proportion of all possible cases of the presence or absence that are predicted correctly by a classifier after accounting for chance predictions (Everaert et al., 2011). Classifiers with CCI higher than 70% and *κ* higher than 0.4 can be considered reliable (Dakou et al., 2007). The classification performance of a classifier can be also determined using a scaling system for *κ* proposed by Landis & Koch (1977), that is: ≤ 0 (poor), 0-0.2 (slight), 0.2-0.4 (fair), 0.4-0.6 (moderate), 0.6-0.8 (substantial), and 0.8-1 (almost perfect).

4 CASE STUDY I: EFFECTS OF METEOROLOGICAL FORCING ON COASTAL EUTROPHICATION: MODELLING WITH MODEL TREES

4.1 SUMMARY

In the present case study primary production (in terms of chlorophyll *α* – chl *a*) is modeled based on a number of abiotic parameters using MTs, a ML approach whereby linear regressions are induced within homogeneous subsets of samples (tree leaves). Standardized regression was applied to determine the relative weight of abiotic parameters in the MT tree leaves whereas the efficiency of the MT method in chl *α* prediction was tested against NNs which is the most frequently used ML approach, and the classical MLR. To assess the efficiency of models to describe eutrophication-related responses under different environmental conditions, the methods were applied on a coastal ecosystem affected by terrestrial runoff for two meteorologically contrasting annual cycles: a typical dry ('04-'05) and a typical wet ('09-'10). MTs showed increased predictive power in chl *α* prediction attributed to the discrimination of input data space into tree leaves, instead of using a uniform space as in NNs and MLR. By grouping samples of each tested annual cycle (wet and dry) on a seasonal basis into discrete groups/leaves, MTs offer a much more explanatory description of ecosystem status than NNs and MLR. The discriminating variables forming tree leaves and the weighing coefficients of Linear Models (LMs) in each leaf provided a useful scaling of abiotic parameters driving chl *α* dynamics. The MT method is thus proposed as an efficient tool for obtaining insights into ecosystem processes leading to eutrophication events in coastal ecosystems and a useful component in integrated coastal zone management.

4.2 INTRODUCTION

ML algorithms, including MTs and MLPs, are considered as appropriate in ecological studies because of their efficiency when dealing with non-linearity (Huang & Foo, 2002; Ornella & Tapia, 2010). This advantage of MTs could be particularly useful in marine ecosystems, which are subject to highly complex and multi-dynamic phenomena (Olden et al., 2008) often resulting in non-linearity. In this chapter, an MT approach was applied in order to evaluate the efficiency of this methodology to model chl *α* dynamics in coastal waters but also to verify whether the method can be used to prioritize factors regulating primary production in coastal ecosystems. The two datasets used were collected from an enclosed coastal ecosystem influenced by terrestrial runoff, under two contrasting meteorological regimes, namely a dry and a wet annual cycle. The main objectives of the study were: (a) to assess the efficiency of MTs in modeling chl *α* compared to two alternative techniques: the most widely used ML method of MLPs and secondly the classical statistical approach of MLR, (b) to evaluate the relative weight of environmental factors regulating chl *α* variability in the study area, (c) to compare the results of the two contrasting meteorological regimes and discuss whether the approach may assist in the understanding of eutrophication-related processes in coastal ecosystems affected by terrestrial runoff.

4.3 METHODOLOGY

4.3.1 Datasets

Kalloni gulf is a semi-enclosed shallow water body located in the southwestern part of Lesvos Island, Greece in E. Mediterranean. The surrounding watershed of 413 km^2 is used for horticulture and agriculture, mainly of olive trees (Spatharis et al., 2007b). These cultivations involve the application of fertilizers during winter, coinciding with the period of high precipitation that usually occurs in February (Spatharis et al., 2007a; Spyropoulou et al., 2013)

The compiled database included information from two annual cycles corresponding to contrasting meteorological conditions. In the dry annual cycle (August ‘04 to July ’05) the total amount of rainfall was low (291 mm) and so was the corresponding amount of terrestrial runoff into the gulf (1.4x10^6 m^3 month^{-1}). On the other hand, in the typical wet cycle (August ‘09 to July ’10) rainfall was high (755 mm), resulting to an increase of one order of magnitude in runoff (14x10^6 m^3 month^{-1}) (Spyropoulou et al., 2013). Previous studies (Spatharis et al., 2007a; Spatharis et al., 2007b) have demonstrated that the interior part of the gulf is characterized by high nutrient and chl *α* concentrations compared to the E. Mediterranean typical levels. This is due to nutrient enrichment from intermittent rivers flowing from November to April, mainly in the northern part of the gulf.

For August ’04 to July ’05 the dataset was compiled from 140 samples collected on a monthly basis from the water column (1 and 5 m depth) from six stations (K3-K8) located in the inner part of the gulf. For August ’09 to July ’10 information on a monthly basis was available for 120 samples from a similar network of stations in the interior of the gulf (KA3-KA7). The two sampling networks have been described in detail in previous studies (Spatharis et al., 2007a; Spyropoulou et al., 2013). Each dataset included information on physical, chemical, and biotic variables. More specifically, physico-chemical parameters selected as input variables for the model were temperature (T), salinity (S), photoperiod (F), nitrogen (N), phosphate (PO_4), and silicate (SiO_2). Chl *α* was the output variable that is predicted by the model, which was compared with observed chl *α* values from the field samples. Summary statistics of the parameters for both annual cycles used as inputs in the LMs during the MT development are provided in Table 8. Irradiance, being highly correlated with photoperiod, was excluded from the input variables in tree construction.

Table 8: Mean, standard deviation (in parenthesis), and number of samples (n) in each of the predicted LMs, for the parameters used in MT method.

		Dry annual cycle '04-'05				Wet annual cycle '09-'10		
Variables	Units	LM1 (n=20)	LM2 (n=23)	LM3 (n=57)	LM4 (n=40)	LM1 (n=61)	LM2 (n=25)	LM3 (n=34)
Temperature - T	°C	10.3	15.5	15.7	24.0	19.5	21.0	15.5
		(0.7)	(3.7)	(4.6)	(2.9)	(5.4)	(2.7)	(5.0)
Salinity - S	psu	36.3	36.9	39.7	38.9	38.4	40.4	38.6
		(1.2)	(0.8)	(0.7)	(0.6)	(1.1)	(0.3)	(2.0)
Photoperiod - F	hrs	9.9	11.8	10.1	13.6	11.2	10.8	12.5
		(0.1)	(1.0)	(1.2)	(0.3)	(1.7)	(1.1)	(1.7)
Nitrogen - N	μM	12.50	1.76	1.84	2.11	0.51	0.49	0.86
		(13.5)	(0.8)	(1.1)	(0.8)	(0.5)	(0.4)	(0.7)
Phosphate - PO_4	μM	0.385	0.036	0.059	0.062	0.036	0.042	0.147
		(0.58)	(0.03)	(0.05)	(0.04)	(0.03)	(0.03)	(0.04)
Silicate - SiO_2	μM	34.9	13.2	13.3	8.1	17.0	7.3	18.5
		(31.7)	(7.6)	(8.0)	(3.2)	(10.3)	(2.6)	(15.6)
Chl *α*	μg/L	3.16	1.01	0.66	1.06	0.76	1.34	1.75
		(0.51)	(0.60)	(0.26)	(0.46)	(0.63)	(0.77)	(0.71)

The size of the two datasets (n=140 for ’04-’05 and n=120 for ’09-’10) is considered sufficient for the application of the MT method since even a small number of training samples (50-100) is sufficient to design a reliable tree decision rule when the number of tree rules is not too large (<10) (Raudys & Jain, 1991) as in the present case study. Moreover, in order to ensure that samples do not violate the condition of independence, a multifactor ANOVA analysis was performed to test chl *α* and nutrient variability within each annual cycle (’04-’05 and ’09-’10). The effect of time is stronger than space (higher F values) suggesting a higher temporal than spatial system turnover. However, since both time and space have a significant effect on the variables (ANOVA, P<0.01), the system seems to present sufficient heterogeneity in space and time.

4.3.2 Details of MTs construction

A number of algorithms exists for inducing MTs from samples, such as CART (Wu et al., 2009), and M5P (Wang & Witten, 1997) which is the most frequently used for MT induction. The package WEKA was used for the analysis (Hall et al., 2009). The parameters of M5 were set to their default values and the important mechanism of tree pruning (Quinlan, 1999) was applied on model construction. Smoothing was not applied, since it has the undesirable property of altering the weight of the original regression coefficients of input variables.

In linear regression, useful indications concerning the ecosystem functioning may be drawn by evaluating the relative importance of independent/input variables in the chl *α* prediction process. This cannot be done with the original regression coefficients because of the different measurement units and variances of the variables (Zar, 1984). In order to render the variables directly comparable to each other, we performed a standardization of the ordinary regression coefficients contained in the equations of each MT leaf. The standardization of LMs was not provided by the WEKA package and was thus carried out using the SPSS statistical package version 16.

4.3.3 Details of MLPs construction

The MLP system that is used in the present case study belongs to the feed-forward group, and it is being trained by the back-propagation algorithm with the use of the sigmoid activation function. The used MLP contains three layers: the input layer comprising by as many neurons as the number of the input parameters (i.e. six), the hidden layer of neurons whose number was set to the default value that is provided by the WEKA package (i.e. three) and the output layer which has a single neuron referring to the output variable (i.e. chl *a*).

4.3.4 Comparison of MTs vs the MLPs and MLR approaches

In order to compare the efficiency of MTs against the MLPs and MLR approaches a 10-fold cross validation technique was performed (Stone, 1974) to assess the model performance on unseen input data (paragraph 3.3.1). Three measures of performance were considered in order to compare the results of MTs with MLPs and MLR in modeling chl *α*: R, MAE and RMSE (paragraph 3.3.2). This procedure was carried out for both annual cycles ('04-'05 and '09-'10) in order to compare the performance of the three approaches using two independent and contrasting datasets.

4.4 RESULTS

4.4.1 Efficiency of the MT over the MLPs and MLR approaches

The MLR equations for predicting the dependent variable chl *α* for the two studied annual cycles are given below, the numbers in parentheses showing the standardized coefficients:

For '04-'05:

$$chl\ a = -\ 0.285(-\ 0.475) * S - \ 1.362(-\ 0.364) * PO4 \ +\ 0.097(+\ 0.656) * N \\ +\ 11.963(0.0) \qquad (1)$$

For '09-'10:

$$chl\ a\ =\ +\ 5.893(+\ 0.430) * PO4 +\ 0.757(0.0) \qquad (2)$$

For the '04-'05 cycle (Equation 1), a significant influence of nitrogen and a weaker negative effect of phosphate and salinity on chl *α* concentration was observed. For the '09-'10 cycle (Equation 2) the only statistically significant variable in the MLR having a positive effect on chl *α* was phosphate.

MT provided a more realistic estimation of chl *α* concentrations in Kalloni gulf than MLPs and MLR based on all three performance criteria (Table 8) for both annual cycles ('04-'05 and '09-'10). More particularly for '04-'05, MTs had higher correlation coefficient (R) and lower estimation errors (MAE and RMSE) than MLPs and MLR (Table 9), whereas for '09-'10, MTs performed slightly better than MLPs(same R but lower errors) and MLR (higher R with lower errors).

Table 9: Validation of the Model Tree (MT), Neural Network (NN) and Multiple Linear Regression (MLR) methods for chl *α* prediction using three validation criteria: multiple correlation coefficient (R), mean absolute error (MAE) and root mean squared error (RMSE).

Annual cycle	Method	R	MAE	RMSE
Dry '04-'05	MT	0.849	0.342	0.491
	NN	0.768	0.430	0.614
	MLR	0.676	0.519	0.680
Wet '09-'10	MT	0.376	0.565	0.732
	NN	0.377	0.586	0.756
	MLR	0.344	0.587	0.755

As MTs offer the better predictions than the other two approaches, a comparison of measured chl *α* values with those predicted by MTs is made in Figure 5.

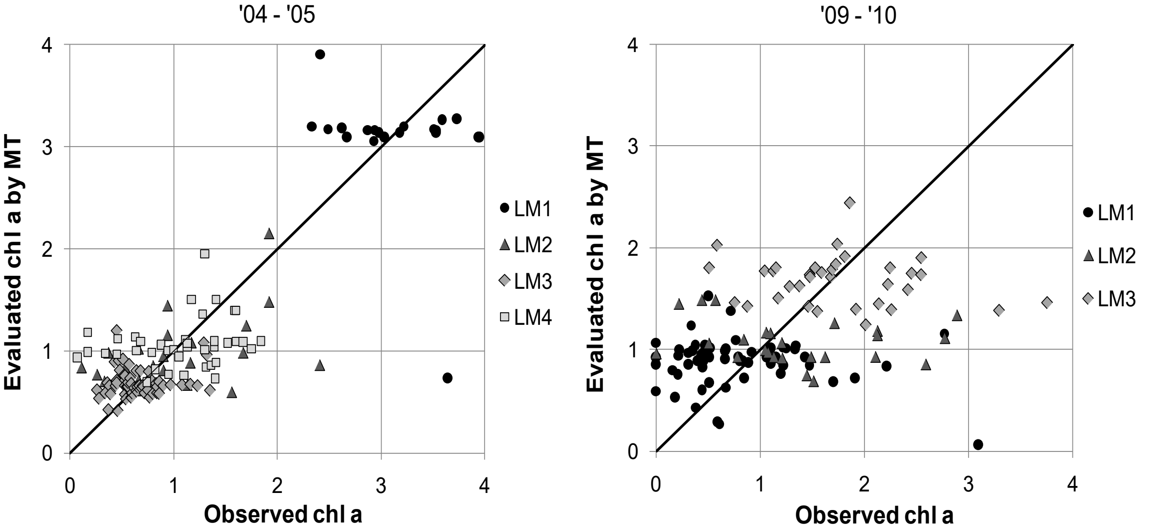


Figure 5: Comparison of observed and predicted chl *α* values by each Linear Model (LM) based on the MT method for the '04-'05 (left) and '09-'10 (right) annual cycles. The black line corresponds to the dichotomous ($y = x$) line.

The points near the dichotomous ($y = x$) line are better approximations of observed chl α values compared to more distant points representing larger prediction errors. Thus, chl α modeling based on the available parameters is much more accurate for ’04-’05 than ’09-’10, which is in agreement with the performance criteria (R, MAE and RMSE) described above. Considering some LMs the predictive capability of the model seems weak. In particular, for LM1 of ’04-’05 and LM2, LM3 of the ’09-’10 annual cycle, the model gives a very narrow range of predicted chl α values (y) for a wide range of observed chl α values (x). In these LMs, either the number of samples was relatively small (LM1 of ’04-’05) or the corresponding equation was independent of the input variables (LM2, LM3 of ’09-’10 annual cycle) (Fig. 6).

4.4.2 Resulting LMs

MTs and the resulting LMs are shown in Figure 6 for the two annual cycles. For ’04-’05, the 140 samples fall within four well defined subsets corresponding to distinct and continuous time periods (Fig. 6a). The grouping of samples into each of the four subsets was based on three discriminating variables, namely salinity, temperature, and photoperiod. For the ’09-’10 annual cycle, the resulting MT is much different comprising of three subsets/leaves, and the 120 samples were grouped within leaves based on phosphate and salinity (Fig. 6b). As in the ’04-’05 cycle, subsets are comprised of samples corresponding to different months, however, in ’09-’10, months are not always continuous within a subset, therefore not always representing continuous time periods of the year.

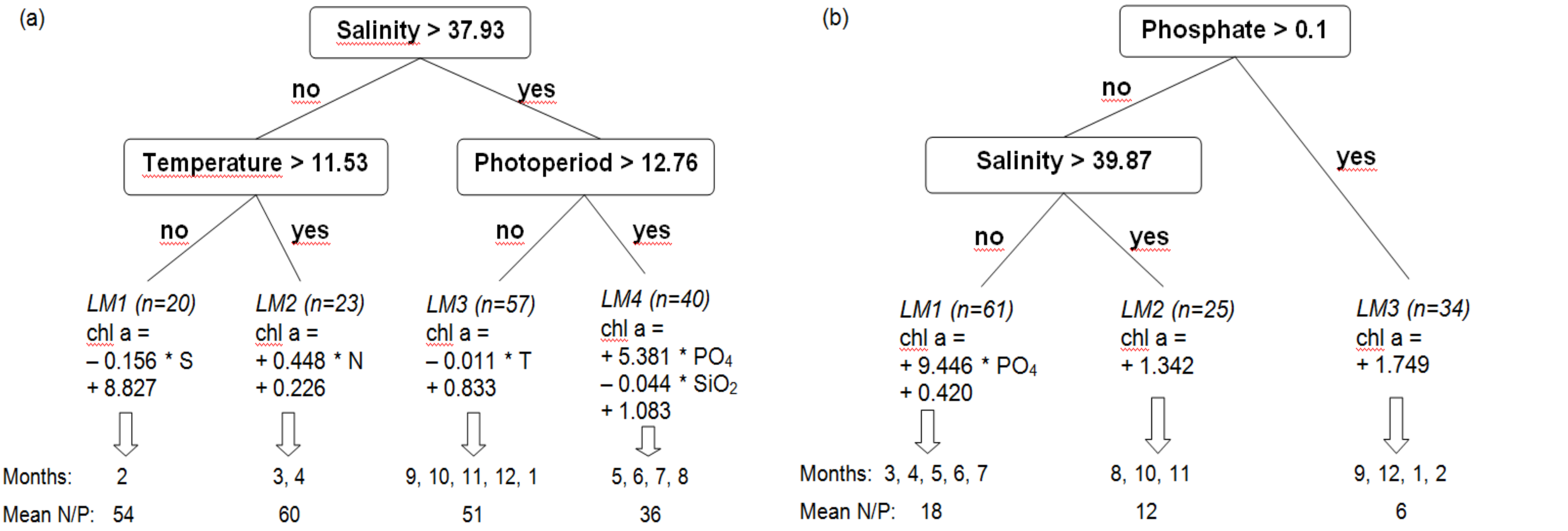


Figure 6: Model Trees (MTs) showing the grouping of input samples based on discriminating variables into Linear Model (LM) subsets for (a) the ‘04-’05 (dry) and (b) the ’09-’10 (wet) annual cycles. Each LM provides a regression equation of the output variable (chl *α*) on the significant input variables, as well as the number of samples (n) grouped within each subset. The number at the bottom shows the months falling inside each LM.

Trying to further explore the subsets (LMs) formed by the MTs for each annual cycle, samples were plotted on a two dimensional graph based on two discriminating variables (Fig. 7). For ’04-’05, the 140 input samples were plotted on a temperature vs salinity plain superimposing chl *α* concentrations and indicating the four subsets of samples corresponding to each LM. The LM1 subset was induced using 20 input samples collected during February, reflecting peak chl *α* concentrations and the lowest salinity and temperature of the year. For LM2, a total of 23 input samples was used, corresponding to March and April characterized by medium to high chl *α* concentrations, low salinity, and medium temperature. LM3 was developed using 57 samples collected from September to January characterized by low chl *α* concentrations, high salinity, and a wide temperature range. Finally, LM4 subset comprised of 40 samples corresponding to summer conditions from May to August characterized by medium chl *α* concentrations and high temperature and salinity (see also Table 8).

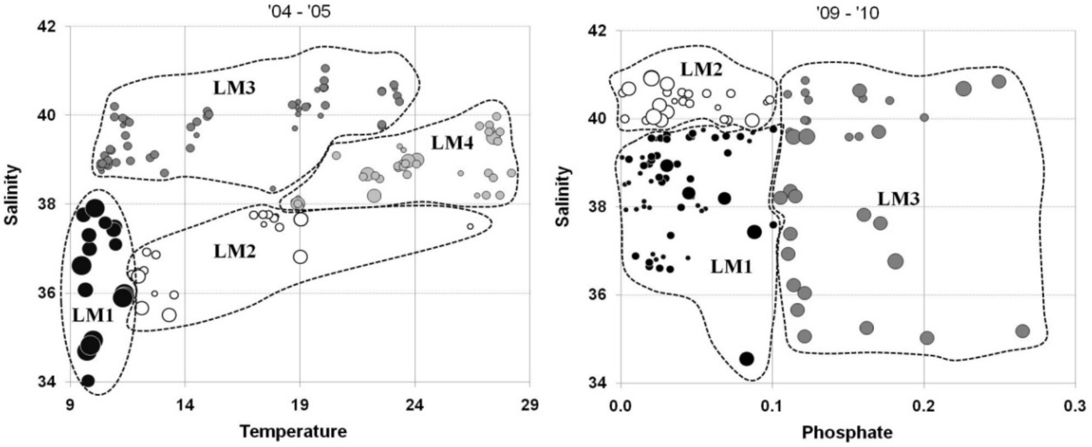


Figure 7: Two-dimensional plots based on discriminating variables from the MT method. For the ‘04-‘05 cycle (n=140 samples) aggregation was based on temperature and salinity whereas for the ‘09-‘10 cycle (n=120 samples) on phosphate and salinity. Each point has a diameter proportional to the measured chl *α* concentration for the specific sample. Also shown are the groupings of samples based on the Linear Model (LM) subsets.

For the ’09-’10 annual cycle the 120 samples were aggregated in a phosphate vs salinity plain (Fig. 7) based on the three LMs formed by MT. LM1 was constructed of 61 samples collected from March to July presenting the lowest chl *α* concentrations, medium salinity, and very low phosphate concentrations, whereas other nutrients had a medium to high concentration compared to background annual means (see Table 7). LM2 represents warmer conditions since it was constructed by 25 samples collected in August, October, and November and describes fairly high chl *α* concentrations in combination with higher temperature, salinity, and phosphate than the previous time period (LM1). Finally, LM3 contains 34 samples from all winter months plus September and describes cold winter conditions with the highest nutrient concentrations and chl *α* values. Based on this analysis, it is evident that for both annual cycles the aggregation of samples in subsets corresponding to different LMs is achieved on a seasonal basis with temperature, photoperiod, salinity, and phosphate being the most important discriminating variables. Sample aggregation was entirely unaffected by the location of sampling stations in the gulf since no such classification was observed in the LM formation.

During the ’04-’05 annual cycle, the N:P ratio was rather high; in particular, the subsets of data in LM1, LM2, and LM3 have mean N:P values close or above the threshold for P limitation (Fig. 8) as defined in Guildford and Hecky (2000). An exception to this was LM4 which had an N:P ratio closer to N limitation due to higher PO_4 concentrations during the warm period. On the other hand all LMs of the wet annual cycle (’09-’10) are characterized by mean N:P values below the threshold of N limitation (Guildford and Hecky 2000).

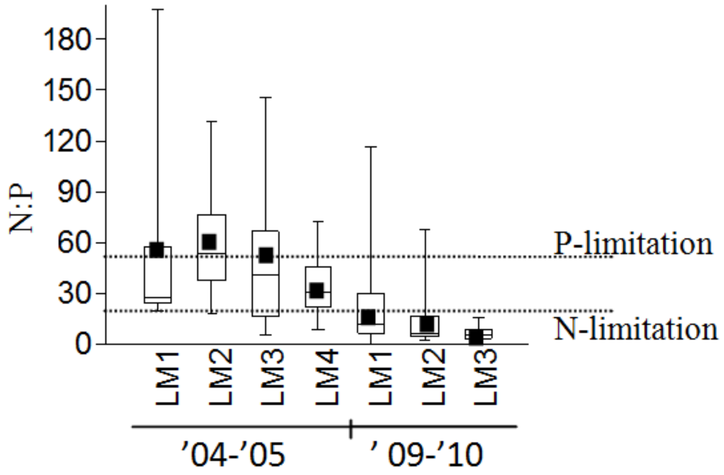


Figure 8: Box-and-whisker plot of the N:P ratio for the LMs of Model trees constructed for both annual cycles compared to thresholds for N deficient (N:P<20) and P deficient (N:P>50) phytoplankton growth according to Guildford and Hecky (2000). Boxes show lower and upper quartiles with median (line) and mean (square) inside the box.

4.4.3 Weighing the LM variables for chl α prediction

In order to detect the relative importance of input variables for chl *α* prediction, standardized regression coefficients were computed for the LMs of each MT (Table 10) corresponding to the two annual cycles. For '04-'05, different variables seem to be important for modeling chl *α* throughout the year. Salinity seemed to be the most important variable during peak chl *α* conditions of February (LM1) although the effect of this variable was not statistically significant (Table 10). This variable was probably selected during the LM construction since it presented a relatively higher correlation coefficient with chl *α* (Pearson R= –0.365, p=0.114, n=20) compared to other variables. Among the variables that played a statistically significant role, nitrogen affected the period following the peak chl *α* concentrations of February (LM2), phosphate affected the medium chl *α* values from May to August (LM4), whereas temperature was correlated with the low winter chl *α* from September to January (LM3).

Table 10: Results for the Linear Models (LMs) resulting from MT application for '04-'05 and '09-'10 annual cycles. B are the unstandardized and Beta the standardized regression coefficients with the corresponding t-test results.

Annual cycle	Linear Model	Parameter	B	Beta	t
Dry '04-'05	LM1	Constant	8.827		2.586*
		S	–0.156	–0.365	–1.663
	LM2	Constant	0.226		0.884
		N	0.448	0.593	3.371**
	LM3	Constant	0.833		6.812**
		T	–0.011	–0.191	–2.121*
	LM4	Constant	1.083		5.907**
		PO_4	5.381	0.483	3.380**
		SiO_2	–0.044	–0.312	–2.183*
Wet '09-'10	LM1	Constant	0.420		3.194**
		PO_4	9.446	0.383	3.182**
	LM2	Constant	1.342		8.731**
	LM3	Constant	1.749		14.410**

* Statistically significant relation at the 0.05 level

** Statistically significant relation at the 0.01 level

For the '09-'10 annual cycle, phosphate seems to be the most important variable for modeling chl *α* variability since it is the main tree separation variable and the only prediction variable for LM1 (Fig. 6b) corresponding to spring and summer conditions. Instead of a linear regression equation, subsets LM2 and LM3 predict chl *α* concentration as a constant value, resulting from the mean of samples contained within each LM. This is because within each of these two LMs, none of the input variables had a statistically significant importance or correlation with chl *α*.

4.5 DISCUSSION

According to the results of the current study, when modeling phytoplankton biomass in an enclosed coastal area, the MT method seems to have increased predictive power on unseen cases (as estimated with 10-fold cross validation) compared to MLPs and MLR statistical approach. This is consistent with previous studies showing a better performance of MTs over the MLPs (Ajmera and Goyal, 2012; Bhattacharya and Solomatine, 2005; Solomatine and Siek, 2006). Apart from the higher predictive power, MTs offer more insight into the generated model (Singh et al., 2010). Indeed, MTs provide the opportunity to easily interpret the effects of input variables to the output variable (e.g. chl *α*). This is not the case in MLPs where special treatment for the weighing of input variables is required to evaluate their contribution (Gevrey et al., 2003; Ruck et al., 1990; Tirelli and Pessani, 2011). Considering the MLR approach, MTs have also shown a better performance which was also confirmed by previous works (e.g. Dzeroski and Drumm, 2003; Jurc et al., 2006). The main advantage of MTs, is that they subdivide the initial dataset into homogeneous subsets/leaves with distinct characteristics based on a number of discriminating variables, instead of the use of a uniform space as in MLPs and MLR.

The data discrimination process in MT induction is based on selected input variables that may reflect characteristics of ecosystem functioning (salinity, temperature, photoperiod for ’04-’05 and phosphate, salinity for ’09-’10). In the resulting subsets the method focuses on the most important variables (if any), incorporating them in the LMs constructed by the MT. Consequently each final subset, expressing an ecosystem state, is described by a linear regression equation with its own input variables affecting chl *α* concentration in contrast to the MLR approach where a single equation originating from the whole dataset aims to predict the output variable (e.g. chl *α*). However, predictions must be made with caution since the predictive power of the method is occasionally low within LMs. Possible reasons for this low predictive capability may be the small number of samples in the LM, or the possibility that chl *α* variability cannot be described by any of the input variables.

Using MTs, inferences about processes regulating ecosystem functioning can be made considering the discriminating variables, the subsets formed (LMs), and the regression coefficients in each LM (e.g. Dzeroski and Drumm, 2003; Lamon III et al., 2008; Kocev et al., 2010). These coefficients may be used for ranking the importance of independent variables, provided that they are standardized prior to analysis. This standardization is essential since many variables of different orders of magnitude are involved. However the explanation of the physical meaning of the weighing coefficients must be carried out with caution, since the number of samples within each subset is small and consequently the statistical power of the linear regression analysis is low. For example in the present application of MTs, some LM subsets contained only 20 samples or weak predictive parameter, indicating that more data may be needed to improve the models. This fact may act as a limitation to the MT application compared to the classic MLR approach which develops linear models using the whole sample dataset.

Two independent datasets were used in the current application of MTs, characteristic of two contrasting meteorological regimes, a typical wet and dry annual cycle (Spyropoulou et al., 2011). The aim was to assess the efficiency of the method to reveal factors regulating primary production. In agreement with the two other applications of MTs on marine ecosystems (Pereira et al. 2009; Volf et al. 2011), salinity seems to play a crucial role on ecosystem functioning, since it was selected as discriminating variable for both the dry and wet annual cycle in the coastal area under consideration. The effect of salinity is probably indirect and is related to the important role of freshwater inputs from the surrounding watershed. These inputs affect both the hydrodynamic regime and the nutrient content of the receiving water body (Tsirtsis et al., 2008). Previous attempts to explain phytoplankton structure (but not chl *α*) for the dry annual cycle ('04-'05) have also shown that salinity and temperature were the two most important parameters explaining assemblage variability (Spatharis et al. 2007a). High freshwater inputs seem to develop a well-formed pycnocline and also decrease residence time in the gulf (Spyropoulou et al., 2011), whereas nutrient-rich freshwater inputs were identified in the past as the driving factor for development of winter algal blooms (Spatharis et al., 2007a; 2009).

The role of freshwater inputs and the seasonal pattern are further stressed due to the fact that temperature and photoperiod were also identified as discriminating variables during the dry annual cycle. Moreover, during winter and particularly February when chl *α* is generally high, salinity and temperature are low, underlining the already observed trend that winter blooms are driven by the cold, nutrient-rich freshwater from the watershed (Spatharis et al., 2007b). A strong seasonal pattern is also revealed when considering the subsets (LMs) formed. For the wet annual cycle three periods were identified characterized by high, medium precipitation, and dryness. For the dry annual cycle however, four periods were formed (summer, autumn/early winter, February and early spring) with considerable fluctuations in chl *α* values possibly related to strong seasonal variability in the physical setting of the system (residence time and stratification). Depth which was found as the main discriminating variable when studying eutrophication in lakes (Lamon III et al., 2008), does not play a significant role in Kalloni gulf as it was also observed in previous studies (Spatharis et al., 2007b), possibly due to the shallowness of the system.

Considering nutrients, the ’09-’10 annual cycle (wet) seems to be driven mostly by phosphate, although a higher number of samples would probably be needed to improve the model predictive power. Phosphate was identified both as the discriminating variable in LM construction and it was also included as a significant variable in the subset corresponding to the warm period (LM1). It was also included in the significant variables affecting chl *α* during the warm period (LM4) of ‘04-‘05 (dry annual cycle). In both cases (the ‘09-’10 tree and LM4 of ‘04-‘05) the N:P ratio was low, close to the threshold for N limitation. These results seem contradictory because although chl *α* variability during these periods should be depending on nitrogen, it is better explained by PO_4according to the MT results. The reverse trend was observed for March and April of ‘05 (LM2), where nitrogen affected the post peak chl *α* concentrations although the N:P ratio in this subset suggested P-limitation. Previous studies (Spatharis et al., 2007a) have attributed this phenomenon to the presence of nitrophilous species such as the diatom *Pseudo-nitzschia calliantha*. These trends are in agreement with Carstensen et al. (2011) who found that TP was a better predictor of chl *α* in regions having TN:TP ratios consistent with nitrogen limitation and vice versa. It seems therefore that nitrogen is the driving factor for the growth of phytoplankton biomass (in terms of chl *α*) during periods of high freshwater input and low renewal rate, whereas phosphate plays a key-role when nutrients are generally low and renewal rate is high. A possible explanation may be related to the tendency of phosphate ions to be adsorbed on particles and consequently be removed from the water column (Krom et al., 2010). During periods of low renewal rate (e.g. February), phosphate is removed from the water column and nitrogen plays a key-role since nitrophilous phytoplankton species form the winter bloom. However, during periods of high renewal rate (e.g. winter of wet annual cycle or summer), phosphate plays a major role driving primary production.

5 CASE STUDY II: OPTIMIZING CLASSIFICATION TASKS WITH A NEW INDEX FOR COMBINING MACHINE LEARNING ALGORITHMS

5.1 SUMMARY

Voting is a commonly used ensemble method that combines base classifier results in order to improve classification in the output variable. However, the selection of proper classifiers to participate in the voting algorithm is currently an open issue. In this study we developed a novel Dissimilarity-Performance (DP) index which incorporates two important criteria for the selection of the base algorithms: their different response in classification (dissimilarity) when combined in triads and their individual performance. To develop this index we firstly evaluated the relationship between voting results and different measures of dissimilarity among classifiers covering heterogeneous algorithm groups (rules, trees, lazy classifiers, functions and Bayes) and using two substantially different datasets (i.e. emotion recognition based on speaker data and ecological state prediction based on physicochemical data). The Jaccard dissimilarity index computed among the classifier triads has shown the strongest relationship (R>0.60) with the corresponding voting results for both datasets and was thus selected as the most appropriate index to represent dissimilarity in our newly proposed DP index. The DP index is highly correlated with the voting performance and can efficiently identify the best and worst performing classifier triads.

5.2 INTRODUCTION

ML techniques can be applied for classification tasks whereby an output variable (target class) with discrete and unordered values (labels) is predicted from a set of collected samples (instances) that consist the training set (Kotsiantis et al., 2006). This approach spans cutting edge applications over a wide variety of scientific fields such as bioinformatics (Cline & Karchin, 2011; Pinero et al., 2004), computing (Huang et al., 2010; Nigam et al., 2000), astronomy (Brescia et al., 2012) and the environment (Cutler et al., 2007; Kubat et al., 1998). The above classification approach could be extremely useful to the evaluation of the ecological quality status of coastal waters for the purposes of WFD (2000); however it has not so far been studied using different ML techniques except NNs (Tison et al., 2007; Ocampo-Duque et al., 2007).

Voting is the simplest and easiest ensemble method for combining classifiers (Tan & Gilbert, 2003) in order to achieve better classifying performance. The challenging step when employing a voting algorithm is selecting the base classifiers to be combined. When the number of potential classifier combinations and the size of the dataset are rather small, then the optimal classifier combination can be found exhaustively. Otherwise, such sequential search is impossible due to the exponential increase of the system's complexity and the amount of time required (Ruta & Gabrys, 2005). To simplify this process, appropriate criteria must be applied for the selection of optimal classifiers to participate in the voting algorithm. For instance when classifiers in a voting scheme are highly dissimilar or independent (as assessed with dissimilarity indices), the classification performance may be significantly improved (Banfield et al., 2005; Kuncheva et al., 2003; Shipp & Kuncheva, 2002). However, previous attempts to incorporate dissimilarity indices within the voting procedure resulted in highly complex and user-unfriendly techniques (Li et al., 2012; Opitz & Shavlik, 1996; Ruta & Gabrys, 2005). Selection of the best combination of base algorithms should thus be based on simple and flexible criteria that will jointly consider the dissimilarity or independency of classifiers along with their individual performance in classification tasks.

In this study we aim to develop a user-friendly index capable of identifying the optimal combination of base classifiers maximizing the classification performance of the voting algorithm. To this aim, the specific objectives are: (a) to assess the efficiency of individual classifiers in two substantially different classification tasks (i.e. classify emotion recognition based on speaker data and eutrophication state based on physicochemical data), (b) to identify combinations of individual classifiers that have markedly different behavior (i.e. high dissimilarity), (c) to test whether these combinations also have a corresponding high performance in voting classification, and (d) to develop a new user-friendly index that joints the two criteria of classifier dissimilarity and individual classifier performance. We expect that this dissimilarity-performance (DP) index will be more efficient in quantifying the classifying performance of different combinations of base classifiers than the traditionally applied dissimilarity indices. DP will be confronted with two essentially different classification tasks to subjectively evaluate the index efficiency.

5.3 METHODOLOGY

5.3.1 Outline

The development of the DP index, identifying the optimal combination of base classifiers to perform classification tasks, was carried out as follows (Figure 9). Initially 10 base classifiers were trained in order to assess their individual performance in two substantially different classification tasks (i.e. emotion recognition based on speaker data and prediction of eutrophication state based on physicochemical data). This information is useful, as it enables direct comparison of classifiers and secondly, it clarifies whether sufficient classifiers participate afterwards in the devolvement of the new index. The next step involves the training of the voting algorithm with all possible combinations of the 10 base classifiers in triads using both datasets. Thereafter, Binary Dissimilarity Indices (BDIs) were computed for all possible classifier triads, to assess possible differences in the outcome of base classifiers. Additionally, this classifier dissimilarity within each triad measured by BDIs and the corresponding voting performance were related using the correlation coefficient. The last step is the development of the new DP index that will take into account the classification performance of each base classifier and the dissimilarity of classifiers in the triads during the voting procedure. In order to assess the efficiency of DP index we tested whether its performance values were correlated with the corresponding voting performance of triads. To further test the efficiency of our new index, we checked whether the triads identified by DP as having the best or worst index value are the same as those giving the best or worst classification performance based on voting. The latter will show whether DP (which is considerably less computationally intensive than the exhaustive search) can identify the optimal classifier combinations.

All base classifiers were trained with the WEKA machine learning package (Hall et al., 2009). The same package was also used for the training of the voting algorithm with all possible classifier triads. The purpose of DP index is to identify the optimal combination of classifiers maximizing the classification performance during voting and is thus not concerned whether base classifiers participate with their highest potential performance. For this reason, each classifier was trained using the default parameter values of the WEKA package.

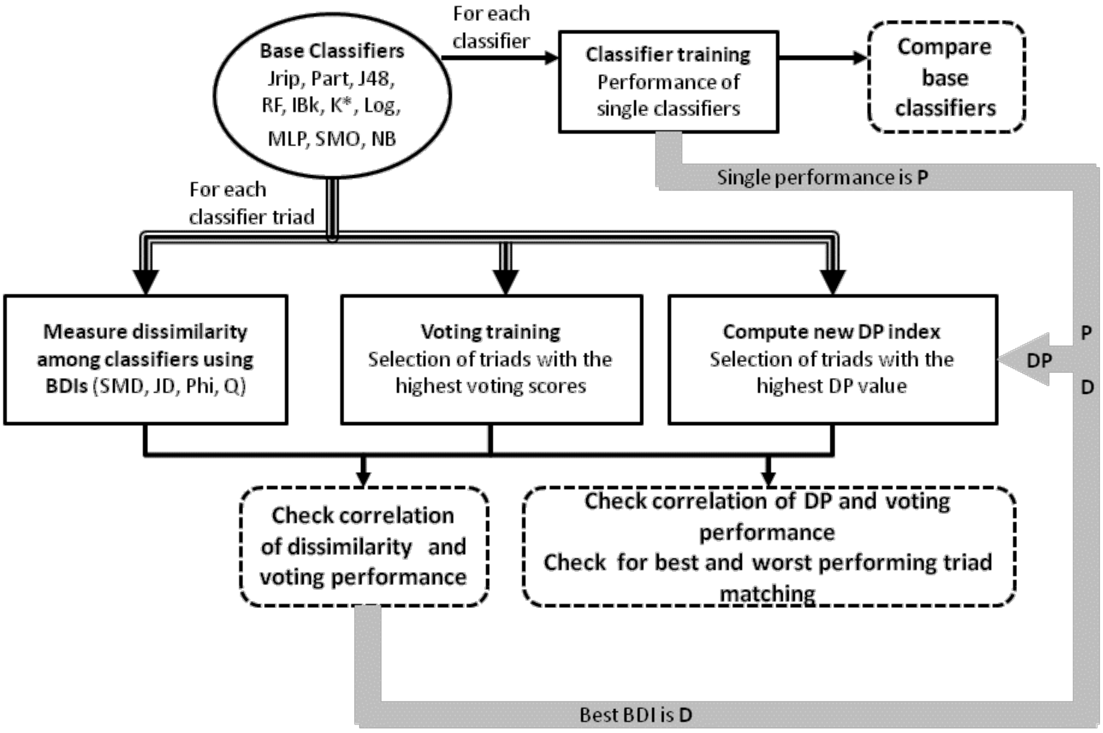

Base Classifiers
Jrip, Part, J48,
RF, IBk, K*, Log,
MLP, SMO, NB
For each classifier
Classifier training
Performance of single classifiers
Compare base classifiers
Single performance is P
For each classifier triad
Measure dissimilarity among classifiers using BDIs (SMD, JD, Phi, Q)
Voting training
Selection of triads with the highest voting scores
Compute new DP index
Selection of triads with the highest DP value
DP
P
D
Check correlation of dissimilarity and voting performance
Check correlation of DP and voting performance
Check for best and worst performing triad matching
Best BDI is D

Figure 9: Schematic diagram of the methodological steps followed for the development and testing of the proposed Dissimilarity-Performance (DP) index. This index takes into account both the individual performance of base classifiers (D) and the dissimilarity of classifiers results -measured with Binary Dissimilarity Indices (BDIs) - when these are combined in triads.

5.3.2 Datasets

Two substantially different datasets were used for the training of base classifiers and voting EM. The first dataset includes voice speaker data used to recognize 7 emotion states and the second dataset includes physicochemical parameters used to classify 5 ecological status levels of seawater. These datasets are different in the number of input variables (133 vs 9), samples (525 vs 188), and predicted labels of the target class (7 vs 5). It is therefore expected, that these two datasets will show different classification efficiency due to the aforementioned data characteristics. Additionally, datasets differ both in structure and functionality as the first deals with human emotional states having unclear boundaries (due to differences among humans) (Anagnostopoulos et al., 2012), whereas the second is subject to high stochasticity and noise, inherent in ecological data (Kitsiou & Karydis, 2011).

More specifically, the first dataset was based on the Berlin Emotional database (EMO-DB) (Burkhardt et al., 2005), which contains 535 utterances of 10 actors (5 male, 5 female) simulating 7 emotional states (anger, happiness, anxiety/fear, sadness, boredom, disgust and neutral). After processing with PRAAT software (Boersma & Weenink, 2005) each utterance was converted to a 133-dimensional prosodic feature vector based on well-established speech features, such as Pitch, Mel Frequency Cepstral Coefficients (MFCCs), energy and formant frequencies (Anagnostopoulos & Iliou, 2010). Thus, the dataset consists of 535 samples with 133 prosodic inputs to be categorized in 7 class labels.

The second dataset comprises of 188 seawater samples collected on monthly campaigns during one annual cycle (August ’04-July ’05) in Kalloni Gulf, Lesvos Island, Greece (Spatharis et al., 2007a). The dataset includes 9 physico-chemical input parameters (e.g. temperature and nutrients) and one target class including 5 ecological status levels (high, good, moderate, poor and bad) (Table 11).The latter is based on chlorophyll *α* limits set by Simboura et al. (2005) for the evaluation of ecological quality of coastal waters for the purposes of the WFD.

Table 11: Classification schemes developed for chl *α* and the corresponding water quality status (Simboura et al., 2005).

	Water quality status				
Index	High	Good	Moderate	Low	Bad
Chl *α*	< 0.10	0.10 - 0.40	0.40 - 0.60	0.60 - 2.21	> 2.21

5.3.3 Training of base classifiers and voting EM

The 10 base classifiers were selected in order to represent all different categories of classification such as rules, trees, lazy classifiers, functions, and Bayes (Table 12). The voting EM combines the results of the base classifiers in triads to offer its own classification for all samples (Kuncheva, 2004). In this work, an exhaustive training of the voting algorithm was achieved by combing the 10 base classifiers at all possible triads (i.e. 120 different classifier triads). We used classifier triads because during voting the combination of an odd number of classifiers avoids the risk of ties (Ruta & Gabrys, 2005). Additionally, three is the minimum odd number that can be used in voting and thus combining classifiers in triads simplifies the whole procedure with respect to complexity and time.

Table 12: Predictive performance in terms of CCI and κ (number in parenthesis) of the 10 base classifiers for both datasets.

Category	Abbreviation	Dataset	
		Emotion recognition	Ecological state prediction
Rules	JRip	58.7 (0.51)	60.1 (0.38)
	Part	64.4 (0.57)	50.5 (0.29)
Trees	J48	62.4 (0.55)	55.9 (0.37)
	RF	73.1 (0.65)	62.2 (0.45)
Lazy	IBk	80.6 (0.77)	63.3 (0.47)
	KStar	78.1 (0.74)	55.3 (0.36)
Functions	Log	67.5 (0.62)	52.1 (0.29)
	MLP	81.7 (0.79)	59.0 (0.42)
	SMO	78.7 (0.75)	45.7 (0.07)
Bayes	NB	51.6 (0.43)	46.3 (0.25)
Meta	Vote best triad	86.8 (0.84)	70.0 (0.59)

The efficiency of the 10 base classifiers and the voting algorithm to perform emotion recognition and ecological state prediction was evaluated using the 10-fold cross validation procedure (Stone, 1978). The voting EM was trained based on the averaged probability estimates of the base classifiers (Witten & Frank, 2005). The classification performance was assessed on the basis of two criteria i.e. the percentage of CCI and the Cohen's κ statistic (Cohen, 1960).

5.3.4 Binary diversity indices (BDIs)

BDIs quantify the dissimilarity or independency of results among base classifiers combined in triads. This is later used to determine whether combinations of dissimilar or independent classifiers also have a corresponding high performance during voting. Dissimilarity of classifiers is an essential measure because combinations of classifiers that are markedly different (i.e. commit classification mistakes on different instances) are expected to improve classification results during voting (Kuncheva &

Whitaker, 2003). BDIs expressing dissimilarity, measure the differences in classification results among classifiers (Wonda, 1981), whereas BDIs expressing independency are used to assess correlation between classifiers. Both BDI categories have been extensively used in various disciplines (e.g. psychology, engineering or economics) for assessing the relation between situations consisting of potential occurrences of a specific event (Seifoddini & Djassemi, 1991; Taylor et al., 2012; Yin & Yasuda, 2005).

In the present case study, the correct classification of an instance by a classifier was assigned a “1” score, whereas misclassification was assigned a “0” score. Using this binary assessment for all 10 classifiers, four well-known BDIs (Table 13) were computed. The first three BDIs can be estimated by combining the classification results of two classifiers. Thus, to express dissimilarity (simple matching distance SMD, Jaccard distance (JD) or independency (Phi) in triads, an average of the paired combinations was calculated. The last index (Q), being also a measure of independency (positive or negative) between classifiers, is estimated by using three classifiers as described in Kuncheva et al. (2003).

Table 13: Definition and ranges of four binary similarity (or dissimilarity) indices

Coefficient	No of alg. required	Range	Formula (S_{ij} or S_{ijk})	Reference
Simple Matching Distance (SMD)	2	[0,1]	$\frac{N^{10}+N^{01}}{N^{11}+N^{10}+N^{01}+N^{00}}$	(Sokal & Sneath, 1963)
Jaccard Distance (JD)	2	[0,1]	$\frac{N^{10}+N^{01}}{N^{11}+N^{10}+N^{01}}$	(Jaccard, 1908)
Phi	2	[-1,1]	$\frac{N^{11}N^{00}-N^{10}N^{01}}{\sqrt{(N^{11}+N^{10})(N^{11}+N^{01})(N^{10}+N^{00})(N^{01}+N^{00})}}$	(Yule, 1912)
Q	3	[-1,1]	$\frac{N^{111}N^{001}N^{010}N^{100}-N^{011}N^{101}N^{110}N^{000}}{N^{111}N^{001}N^{010}N^{100}+N^{011}N^{101}N^{110}N^{000}}$	(Yule, 1900)

N^{11} Number of instances that have been correctly classified by both classifiers

N^{10}, N^{01} Number of instances that have been correctly classified by the 1st classifier but not by the 2nd and likewise respectively

N^{00} Number of instances that have been correctly classified by neither classifier

N^{111} Number of instances that have been correctly classified by all three classifiers

$N^{011}, N^{101}, N^{110}$ Number of instances that have been correctly classified by the 2nd and 3rd classifiers but not by 1st and likewise respectively

$N^{001}, N^{010}, N^{100}$ Number of instances that have been correctly classified by 3rd classifiers but not by 1st nor 2nd and likewise respectively

N^{000} Number of instances that have been correctly classified by neither classifier

5.3.5 Dissimilarity-Performance index (DP)

Apart from the traditionally used BDIs, the individual performance of classifiers was also considered in the present case study to provide further information on the classifiers to be combined in order to achieve improved classification (Sharkey & Sharkey, 1997). To identify the best performing classifier triad it is thus essential to jointly consider the criterion of dissimilarity among classifiers along with their individual performance in the development of an integrated Dissimilarity-Performance (DP) index. The formula proposed for this DP index is the following:

$$DP = \frac{\sum_{i<j}^{n} J_{i,j} + \sum_{i=1}^{n} p_i}{6} \qquad i,j = 1,2,3$$

where $J_{i,j}$ is the JD index calculated from the binary classification results of the$i$-th and j-th classifiers and p_i is the ratio of the correctly classified samples by the i-th classifier to the total number of instances. The first addend in the numerator represents the sum of the JD for all classifier pairs, while the second is the sum of the single performance of each classifier used in voting. The denominator is used to standardize results on 0 to 1 scale. JD index was selected as a dissimilarity measure in the new DP index as it showed the best correlation with voting performance compared to other BDIs. The characteristics of diversity and performance of the classifiers existing in each triad have equal contributions in DP index.

The efficiency of BDIs and DP based on the performance criteria (i.e. CCI and κ) for the 120 different classifier triads was assessed with Spearman's rank correlation coefficient. DP was further tested for monotonicity (consistent increase or decrease along the CCI spectrum) as this is an important prerequisite for an index (Spatharis & Tsirtsis, 2010).

5.4 RESULTS

The classification performance of the ten base classifiers for both datasets is presented in Table 13. Overall, the results of the emotional recognition are significantly better for all classifiers than the corresponding results of ecological state classification. The best classifier is IBk, as it is the most efficient in classifying ecological state and the second more efficient for emotional recognition. RF and MLP can also be considered as satisfactory classifiers for both classification tasks. On the other hand, classification results of Jrip and SMO algorithms were contradicting. Although Jrip showed a satisfactory classification of ecological states compared to other classifiers, it failed to give statistically significant results for emotional recognition. The opposite was found for the SMO classifier. Finally, NB had low performance for both datasets while the remaining classifiers (i.e. Part, J48, K* and Log) were characterized by moderate predictive performance.

Voting EM, combining the aforementioned base classifiers in triads, has shown higher classification performance compared to the performance of individual classifiers (Table 13). The best classifier triad for each of the two datasets achieved an increase in performance higher than 5%, based on the CCI and κ performance criteria*,* compared to the corresponding results of the best base classifier (i.e. MLP for emotional recognition and IBk for eutrophication). Thus, the best classifier triad for emotion recognition (i.e. IBk, MLP, SMO) classified correctly 86.8% of the samples whereas the κ performance criterion indicates that the classification performance is almost “perfect”. Considering ecological state classification, the best triad was JRip, RF and MLP classifiers, which correctly classified 70% of samples with performance that can be characterized as almost “substantial” ($\kappa = 0.59$).

The voting performance of various classifier combinations in terms of the CCI and κ performance criteria has shown statistically significant correlation ($p<0.01$) with JD and DP indices (Table 14) for both datasets. Other BDIs such as SMD were more weakly but significantly correlated with CCI and κ for both datasets, whereas Q was significantly correlated with CCI and κ only for emotional recognition. The positive correlation between SMD or JD and the performance measures shows that when combining highly dissimilar classifiers, the resulting classification performance is also high. On the other hand, the negative correlation with the Q index is observed due to the ambivalent relationship with voting classification performance (Kuncheva et al., 2000). Finally, Phi index was not correlated with the performance measures for both classifications tasks. Therefore, assuming that voting performance can be expressed with CCI and κ, the most efficient among the indices considered is DP showing high correlation coefficient values for both datasets ($R>0.80$).

Table 14: Results of correlation analysis (Spearman) between the performance (based on CCI and *κ*) of the voting algorithm and the BDIs trained on the 120 classifier combination triads.

	Emotion recognition		Ecological state prediction	
Index	CCI	*κ*	CCI	*κ*
Phi	0.005	0.003	0.163	0.177
SMD	0.529**	0.520**	0.218*	0.234*
JD	0.618**	0.610**	0.429**	0.451**
Q	-0.385**	-0.378**	-0.178	-0.146
DP	0.811**	0.813**	0.824**	0.845**

** Correlation is significant at the 0.01 level (2-tailed)

The efficiency of the new DP index to identify classifier triads with significantly high or low performance in voting procedure is shown in Table 15. DP has determined in the best decade, 9 out of 10 classifier combinations having the higher voting performance in terms of both CCI and *κ* for emotional recognition. The corresponding values were 8 and 9 respectively, for eutrophication state classification. On the other hand, in the worst decade of classifier triads DP managed to identify 7 out of 10 with the worse CCI after voting for both datasets. The DP performance was slightly improved (i.e. 8 out of 10 identifications), when in the worst decade the classifier triads with the lower *κ* value were only considered. Additionally, the combination triad identified as best by DP for each dataset, was the triad that finally presented the best performance during voting.

Table 15: Number of classifier combination triads that both one of the performance measures (CCI or *κ)* and DP placed in the worst or best tens for each datasets.

Classifier	Emotion recognition		Ecological state prediction	
combinations	CCI	*κ*	CCI	*κ*
10 worse	7	8	7	8
10 better	9	9	8	9

The monotonic behavior of DP was checked by plotting its variability in specific percentiles of CCI performing measure (Fig. 10) for comparative reasons. To this aim, six percentiles of CCI were selected for both datasets: the minimum and maximum values, and the 20th, 40th, 60th and 80th percentiles. DP has shown consistent increase along the CCI spectrum for both datasets and thus its behavior is considered as monotonic.

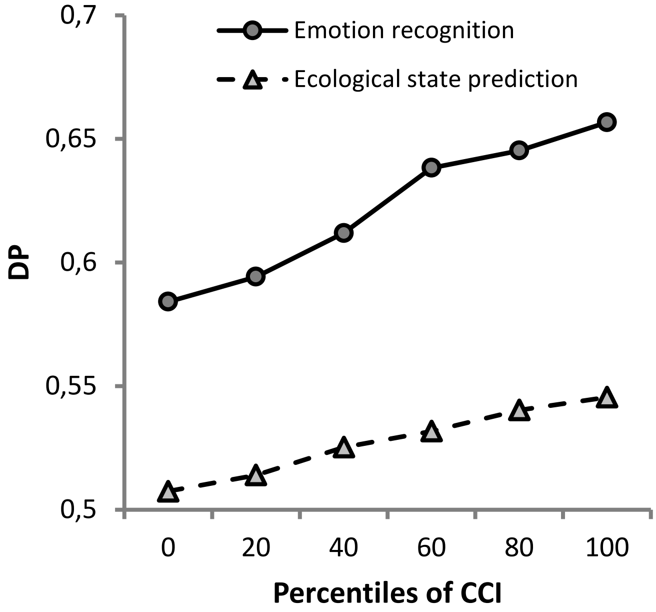


Figure 10: Monotonic behavior of DP along CCI gradient for both datasets.

5.5 DISCUSSION

In the present case study, 10 base classifiers corresponding to various ML categories, were trained using two substantially different datasets (i.e. recognition of emotion and eutrophication state classification) in order to access their classifying efficiency. Best performance in recognition of emotion was achieved by MLP, although IBk and SMO were also efficient. This is in agreement with previous applications where these three classifiers accurately recognized emotions from data offering significantly better performance (Fragopanagos & Taylor, 2005; Iliou & Anagnostopoulos, 2010; Morrison et al., 2007; Rani et al., 2006; Shami & Verhelst, 2007). On the other hand, classification performance of ecological state using base classifiers was moderate, an observation also holding for previous studies on this topic (Tamvakis et al., 2014). Higher performance was observed for IBk, MLP and tree classifiers in agreement with previous studies on eutrophication analysis by ML techniques (Recknagel, 2001; Tamvakis et al., 2012; Volf et al., 2011).

The voting algorithm was trained with all possible classifier triads resulting from combinations of the 10 base classifiers in order to give its own combined classification. The best triad has shown improved performance in agreement with the general principle that ensembles of classifiers are often substantially more accurate than their individual base classifiers (Dietterich, 1997; Pal & Mather, 2003; Saha & Ekbal, 2013; Tsai, 2014; Wozniak et al., 2014). For both datasets the % increase of CCI was over 5%, which is considered as remarkable improvement in classification performance (Pal & Mather, 2003). Moreover, according to κ performance criterion, voting increased the classification performance from “substantial” to “almost perfect” for the recognition of emotion, whereas the performance for seawater ecological state increased significantly to the lower limit of “substantial” classification.

Each base classifier employs a different learning strategy to give its own classification results which are fed into the voting algorithm for the final classification outcome. When the individual results are similar then the voting outcome will be based more or less on the same information (errors and corrects) (Dietterich, 2000b). Thus, combining classifiers with similar results does not offer any additive value in voting, increasing however the system complexity (Ruta & Gabrys, 2005). On the other hand, EMs consisting of classifiers offering different results have the potential to achieve significantly better performance compared to those of individual base classifiers (Ruta & Gabrys, 2005; Tan & Gilbert, 2003; Tsymbal et al., 2003). This finding was confirmed in the present case study, with JD dissimilarity measure showing the highest statistically significant correlation with the two measures of voting performance for both datasets. The positive correlation indicates that classifier triads with highly differentiated results (as expressed by JD) tend to be more accurate during voting. These results are in agreement with Kuncheva & Hadjitodorov, 2004 who employed JD in cluster ensembles. However the two measures of independency (i.e. Q and Phi) being considered to offer improvement in voting accuracy (Kuncheva et al., 2000; 2003), showed low relationship with voting performance also in agreement with previous studies (Banfield et al., 2005; Ruta & Gabrys, 2005; Shipp & Kuncheva, 2002).

Although dissimilarity among individual classifiers combined to develop EMs may be the key towards the improvement of classification efficiency (e.g. (Canuto et al., 2007; Mao et al., 2011), dissimilar but powerless classifiers are unlikely to bring any benefits in EMs performance (Ruta & Gabrys, 2005). These two crucial characteristics, classifier dissimilarity and efficiency in individual performance, have been coupled in the current study to propose a new index highlighting the optimum classifier combinations to train voting algorithms. The DP index integrates dissimilarity using the JD measure, which is considered as an efficient and stable indicator (Yin & Yasuda, 2005), while it is sensitive on following the voting performance variability for both datasets. In addition, DP index integrates the performance characteristic using the individual performance of the classifiers, as it is reasonable to assume that optimal combinations should include classifiers with high individual performances (Sharkey & Sharkey, 1997).

Considering (a) the high and significant correlation between DP and the voting performance, (b) the fact that DP achieved to determine the best performing classifier triads and (c) the consistent monotonic behavior of DP for both datasets, this newly proposed index is very efficient to identify base classifiers that should be combined in order to optimize the classification performance during voting. DP is recommended to individual ML users (rather than EMs designers) seeking to optimize their classification performance by selecting appropriate base classifiers. Indeed, DP can be easily calculated in three steps. First, the user trains any set of base classifiers, being composed by representatives of any ML category, trained through any learning platform or even composed by a single classifier trained by different training sets. Then, using the obtained classification results for every instance of the database (i.e. correctly or falsely classified), the user calculates the JD index for every triad of base classifiers. Finally, using the derived information (performances from the first step and dissimilarities from the second) DP values are computed for every triad and subsequently the triad that possibly offers the best voting performance (i.e. the triad with the grater DP value) is identified.

Apart from the easiness in application, DP has a number of additional advantages: (a) simplicity as it uses only three combined classifiers, (b) efficiency in the selection of classifiers to participate in voting schema as proven in the present case study for two different datasets, (c) flexibility as any base classifier can be included in the voting scheme and (d) innovation in the joint consideration of the dissimilarity among classifiers as well as their individual performance. On the other hand, DP cannot be compared with complex EM schemes that perform thorough search towards inducing all possible kinds of classification errors which however need highly qualified designers to apply them to new tasks. DP mainly aims at individual users aiming to achieve a combined and more accurate classification using their own familiar and tested ML algorithms.

6 CASE STUDY III: OPTIMIZING BIODIVERSITY PREDICTION FROM ABIOTIC PARAMETERS USING MACHINE LEARNING TECHNIQUES

6.1 SUMMARY

An integrated methodology is proposed for the effective prediction of biodiversity exclusively from abiotic parameters. Prediction is based on three machine learning techniques: MTs, MLP and IBk algorithms. Abiotic parameters (input parameters) include temperature, salinity, dissolved inorganic nitrogen and phosphates that are known to affect phytoplankton assemblage structure. Biodiversity is expressed as a number of indices (output variables) representing richness, evenness and dominance. To optimize diversity prediction, indices were calculated on a large number of phytoplankton field assemblages, but also on corresponding noise-free simulated assemblages that retain the structure of field ones. Results indicate that biodiversity can be accurately predicted using exclusively abiotic parameters and the efficiency is doubled with simulated assemblages. The Instance Based learning algorithm was the most effective and achieved the best prediction for Menhinick richness (R = 0.80), Evenness E2 (R = 0.81) and Berger Parker dominance (R = 0.80) indices. Based on the optimal algorithm, indices, and dataset, a software package was developed for phytoplankton diversity prediction typical for Eastern Mediterranean waters.

6.2 INTRODUCTION

Diversity can be expressed through a number of indices which quantify community structure and the changes it undergoes due to natural or anthropogenic stress (Magurran, 2004). However, field communities are also driven by multiple stochastic factors such as seasonality and spatial heterogeneity which impose a degree of uncertainty and distortion on data (Straten, 1992). This ‘environmental noise’ inherent in field communities is also reflected on the subsequent calculation of indices (Vounatsou & Karydis, 1991). This problem can be overcome with the use of simulated communities *via* a species abundance distribution (e.g. the log-series, lognormal) however retaining the structure of field ones (Blackwood et al., 2007; Lyashevska & Farnsworth, 2012; Schloss & Handelsman, 2006; Spatharis & Tsirtsis, 2010). Calculations on noise-free simulated communities seem appropriate when trying to establish cause-and-effect relationships, e.g. between diversity and abiotic parameters, due to the removal of noise or distortion that more easily supports the revealing of possible signals.

In this paper we propose an integrated methodology for the optimization of diversity prediction exclusively from abiotic parameters (Fig. 11). The diversity is expressed by diversity, evenness, and dominance indices calculated on both field and simulated phytoplankton assemblages covering a wide productivity range typical of Eastern Mediterranean waters. Predictions were carried out based on three ML algorithms. The objectives of the study were thus: (a) to distinguish the ML technique offering the most accurate prediction, (b) to select the indices representative of all three diversity components (richness, evenness, and dominance) (c) to optimize prediction by calibrating the methodology with indices calculated on simulated assemblages, and (d) to develop a software tool for biodiversity prediction based on the proposed methodology.

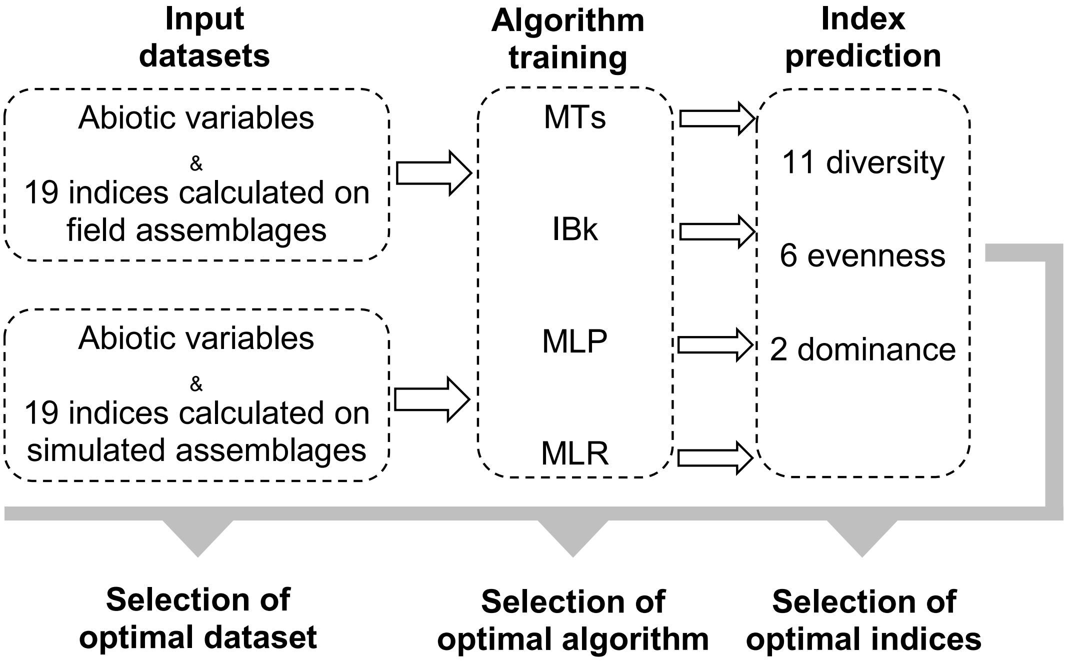

Input datasets
Algorithm training
Index prediction
Abiotic variables & 19 indices calculated on field assemblages
Abiotic variables & 19 indices calculated on simulated assemblages
MTs
IBk
MLP
MLR
11 diversity
6 evenness
2 dominance
Selection of optimal dataset
Selection of optimal algorithm
Selection of optimal indices

Figure 11: Conceptual diagram of the methodological procedure followed in order to optimize diversity prediction from abiotic parameters.

6.3 METHODOLOGY

6.3.1 Datasets

The first dataset employed in the study includes 658 field samples and was compiled using existing data from coastal areas of the Aegean Sea, E. Mediterranean representing a wide range of productivity. Among the various abiotic parameters available in the dataset, a subset was selected for the aims of the present case study, including: (a) concentrations of limiting nutrients, Dissolved Inorganic Nitrogen (DIN) and Phosphates (PO_4), that directly influence the growth and composition of phytoplankton in the areas under consideration (Spatharis et al., 2008) and (b) Salinity (S) and Temperature (T), which may also indirectly affect phytoplankton synthesis through stratification in coastal waters (Spyropoulou et al., 2013). Dataset information and summary statistics of the above parameters in each of the four areas are provided in Table 16. The dataset covers a wide range of phytoplankton abundance (10^3-9×10^6 cells/L) and species richness (4-39 species). There were no missing values in the dataset and no special treatment was performed for outlying values. It was considered that the latter often correspond to extreme events such as algal blooms due to episodic terrestrial inputs (Spatharis et al., 2007b) or to the photoperiod increase during spring, that have to be included in the models to be developed.The variables' positive skeweness (Table 16), that is almost always observed for environmental data, was taken into account in the application of the ML algorithms. According to the requirements of each algorithm standardization or normalization procedures were applied, described in detail below.

Table 16: Dataset information (mean annual values, range in parenthesis and skeweness) of abiotic (input) and phytoplankton parameters for the coastal areas in Aegean Sea.

		Rhodes offshore n=143	Gera Gulf n=114	Kalloni Gulf n=186	Saronikos Gulf n=215
Abiotic parameters	T (°C)	19.67	19.06	17.73	19.21
		(15.86-26.39)	(9.90-26.70)	(9.43-28.20)	(13.10-27.60)
		0.66	0.32	0.11	0.33
	S (pcu)	39.16	38.92	38.58	38.30
		(38.92-39.39)	(36.39-40.28)	(34.02-41.06)	(37.20-39.70)
		11.63	0.23	0.91	8.50
	DIN (μM)	0.91	1.48	3.94	2.70
		(0.21-12.45)	(0.40-5.82)	(0.47-45.20)	(0.36-37.95)
		9.08	2.36	4.66	5.18
	PO_4(μM)	0.0700	0.194	0.088	0.236
		(0.010-4.090)	(0.050-0.850)	(0.00-1.577)	(0.010-6.00)
		11.60	2.11	5.88	7.54
Biotic parameters	Cell No.	6,291	47,237	592,441	283,201
		(10^3-6×10^4)	(2×10^3-4×10^5)	(3×10^3-9×10^6)	(10^3-6×10^6)
		4.37	3.04	5.03	6.30
	Species No.	12	16	23	19
		(5-23)	(4-37)	(4-39)	(5-39)
		0.18	0.95	0.40	0.35

The second dataset includes 658 simulated phytoplankton assemblages with abundances corresponding exactly to the abundances of the 658 field samples. The simulation was based on the log-series statistical distribution which assumes that most species in an assemblage are rare (Fisher et al., 1943). The log-series distribution is shaped by parameters *x* and *a*, that can be calculated knowing the ratio of species richness to total abundance (S/N) in an assemblage. The S/N ratio was estimated via a simple linear regression equation between S and N using the 658 field samples as described in Spatharis & Tsirtsis (2010). Regression analysis was also used to identify the relation of the abundance of the most dominant species N1 with the total phytoplankton abundance N in the 658 field samples. When parameters *x* and *a* were estimated, the expected number of species S was allocated for each abundance (total cells N). By feeding the previous two relationships which characterize field phytoplankton assemblages onto the log-series distribution, simulated assemblages are generated that retain the structure of the initial field ones (Fig. 12). This approach has been described in detail in previous studies (Spatharis & Tsirtsis, 2010; Tsirtsis et al., 2008) resulting in a wide range of assemblage diversity closely matching reality (Spatharis et al., 2011).

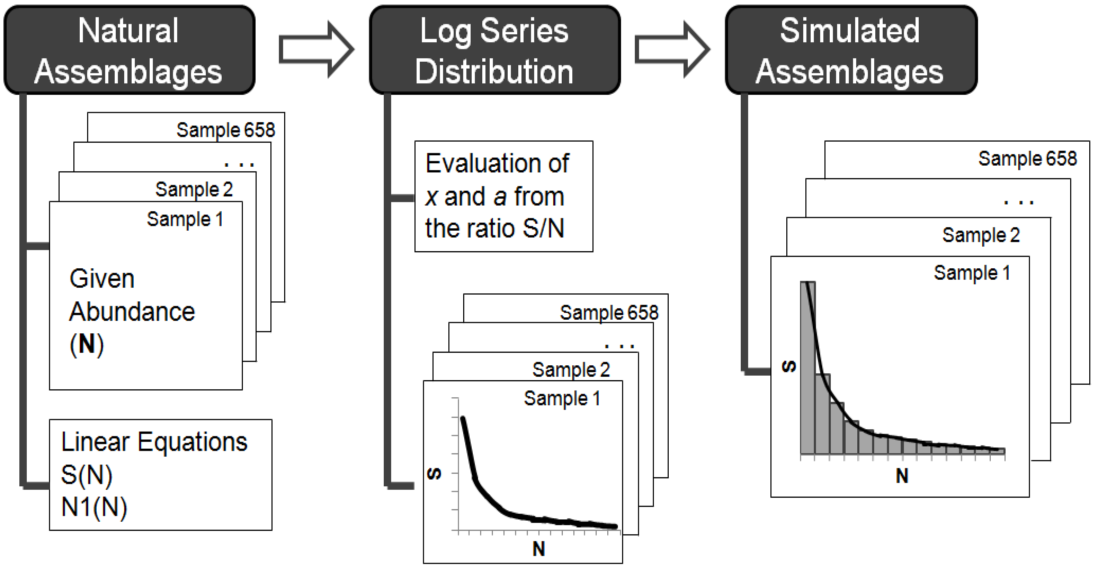


Figure 12: Schematic presentation of the procedure followed for the generation of 658 simulated phytoplankton assemblages corresponding to the 658 field assemblages.

6.3.2 Indices expressing diversity components

Indices can express different aspects of biological diversity such as richness, evenness, and dominance. Thus, diversity indices weigh more on the richness component of assemblages, evenness indices account more for the distribution of individuals to species, and dominance indices consider only the proportion of most abundant species in an assemblage (Karydis & Tsirtsis, 1996). In the current study, the most commonly used diversity, evenness and dominance indices (Krebs, 1999; Magurran, 2004) were used in order to express all aspects of phytoplankton diversity (Table 17). These indices were considered as output parameters for the ML algorithms described below.

6.3.3 Details of the ML algorithms

MTs

The M5 algorithm is one of the most well-known MT induction methods. The M5P algorithm in Java implementation which is part of WEKA machine learning package (Hall et al., 2009) was used for the MT induction. An optimization of the method was attempted based on the minimum number of instances reaching a leaf that is crucial since it controls the tree pruning (Quinlan, 1999). To this aim, different values were used in order to optimize results, that is 4 (default), 8, 16, 32 and 64 instances. Prior to analysis, abiotic parameters were standardized using the z-score procedure to ensure equal weights during tree induction.

IBk

The IBk algorithm was applied with the use of *k* nearest training instances (*k*-NN) in order to predict the value of the output variable in new unseen instances. The Manhattan (city-block) distance was used as distance metric, as it was found more powerful compared to the classic Euclidean distance. In the software package WEKA (Hall et al., 2009) the initial setting of parameter *k* may significantly affect the prediction power. To optimise results, different values of this parameter were tested, i.e. 2, 4, 8, 12 and 20. Prior to data analysis, abiotic parameters were standardized using the z-scores, as performed in the application of MTs.

MLP

The MLPs used in this study belong to the classic group of feed-forward NNs with one hidden layer in which sigmoid activation function is used to all nodes while it is being trained by the backpropagation algorithm. To select the network's topology that maximizes the algorithm effectiveness, five numbers of neurons were tested (4, 8, 10, 15, and 20).

The performance of the three algorithms was also compared to the classic multiple linear regression (MLR) technique. Prior to data analysis, abiotic parameters were log transformed to approach normality, as it is common for natural data to follow positively skewed distributions. Although NNs do not require any assumption regarding input data, it has been shown that their performance is often improved through data transformation using mathematical functions (Shi, 2000).

6.3.4 Assessment of optimal diversity prediction

To estimate the prediction accuracy of different algorithms on unseen data, the K-fold CV approach was employed (paragraph 3.3.1) (Stone, 1974). To optimize algorithm results, biodiversity prediction was based on three numbers of CV folds: 10, 20 and 658 i.e. LOOCV. The basic measure of performance for assessing the predictive power of the three algorithms is the R coefficient between the calculated values of indices (based on field or simulated assemblages) and those predicted by the algorithm while the RMSE is also presented. The variability of R coefficient within the K-folds of each algorithm was estimated by the coefficient of variation which quantifies the variability (or stability) of the results. A two-factor ANOVA was used to determine the relative effect of testing different CV folds and different values of algorithm parameters (i.e. number of instances reaching an MT leaf, number of neighbours for IBk, or number of neurons for MLP).Percent errors between calculated and predicted values were used as an additional measure of performance. Instead of using solely instances (e.g. the 658 samples) to assess the performance of predicted indices, we also assessed the behavior of predictions using average monthly values of biodiversity for each of the four areas.

6.4 RESULTS

6.4.1 Selection of optimal parameters for algorithm training

The relative effect of three numbers of CV folds (10, 20, 658) on algorithm performance based on R was tested using both field and simulated data. This factor was not significant for MTs and IBk (ANOVA, p>0.05), but was statistically significant for MLP (ANOVA, p<0.001). For the latter, the best performance was achieved with the 10-fold CV. The minimum number of instances reaching an MT leaf (4, 8, 16, 32, 64) was not statistically different for field dataset (ANOVA, p>0.05) but was strongly significant when using indices calculated on simulated data (ANOVA, p<0.005). The optimal number of instances selected for MT parameterization was 8. Significant differences in R were observed among the different numbers of neighbors for IBk (2, 4, 8, 12, 20) and neurons for MLP (4, 8, 10, 15, 20) (ANOVA, p<0.05). The optimal number of neighbors was 8 and the optimal number of neurons was 10.

6.4.2 Selection of optimal dataset, algorithm and indices

The performance of the three ML algorithms (using the optimal parameters described above) with the respective performance of MLR in terms of R for all indices is presented in Table 17. Additionally, the performance of the algorithms in terms of RMSE is presented in Table 18. Overall, the use of indices calculated on simulated instead of field assemblages resulted in significantly improved predictive power. This was observed for all tested algorithms and almost all indices as indicated by the higher and in some cases doubled correlation coefficients. The most efficient algorithm for diversity prediction was IBk and the least efficient was MLR. Based on IBk, the most effective indices calculated on simulated data were Species Richness, Menhinick, Evenness E1, Evenness E2, Evenness E3 and Berger-Parker ($R \geq 0.80$). On the other hand, Shannon and Hill N1 indices had lower predictive power (R<0.72). According to the coefficient of variation, variability of R among the 10 folds of CV was low (<25%) for the majority of indices tested, whereas it was minimised for IBk on simulated data (e.g. 5.1% for Species Richness, 5.9% for Menhinick, 5.3% for Evenness E2, 5.1% for Berger-Parker). Therefore, Species Richness, Menhinick, Evenness E2 and Berger-Parker were selected as representative of the three components of assemblage diversity i.e. richness, evenness and dominance.

Table 17 Predictive performance in terms of R of MTs, IBk, MLP and MLR for all indices using data from field (F.A.) and simulated assemblages (S.A.) evaluated by 10-CV.

		Reference	MTs		IBk		MLP		MLR	
			F.A.	S.A.	F.A.	S.A.	F.A.	S.A.	F.A.	S.A.
	Abundance		0.76	0.77	0.73	0.73	0.39	0.39	0.26	0.26
	Sp. Richness	(Ludwig & Reynolds, 1988)	0.33	0.73	0.69	0.81	0.47	0.54	0.28	0.31
	Margalef	(Margalef, 1958)	0.39	0.69	0.60	0.79	0.33	0.58	0.19	0.29
	Gleason	(Ludwig & Reynolds, 1988)	0.37	0.69	0.59	0.79	0.29	0.58	0.18	0.28
	Menhinick	(Menhinick, 1964)	0.62	0.71	0.77	0.80	0.55	0.59	0.28	0.29
Diversity indices	Odum	(Odum et al., 1960)	0.35	0.65	0.75	0.74	0.58	0.59	0.26	0.23
	Simpson	(Ludwig & Reynolds, 1988)	0.60	0.67	0.66	0.77	0.35	0.51	0.25	0.35
	H2-Shannon	(Shannon & Weaver, 1949)	0.58	0.57	0.67	0.71	0.40	0.44	0.21	0.32
	Hill N1	(Ludwig & Reynolds, 1988)	0.49	0.53	0.63	0.70	0.32	0.43	0.10	0.30
	Hill N2	(Ludwig & Reynolds, 1988)	0.34	0.68	0.60	0.79	0.30	0.53	0.12	0.32
	Hurlbert	(Hulbert, 1971)	0.60	0.63	0.66	0.77	0.33	0.51	0.25	0.35
	McIntosh	(McIntosh, 1967)	0.55	0.61	0.66	0.78	0.32	0.52	0.22	0.35
	Evenness E1	(Pielou, 1975)	0.59	0.70	0.68	0.80	0.33	0.56	0.27	0.33
	Evenness E2	(Sheldon, 1969)	0.53	0.72	0.67	0.81	0.35	0.58	0.24	0.30
Evenness indices	Evenness E3	(Ludwig & Reynolds, 1988)	0.43	0.71	0.67	0.80	0.35	0.57	0.24	0.30
	Evenness E4	(Ludwig & Reynolds, 1988)	0.23	0.71	0.45	0.78	0.20	0.54	0.03	0.27
	Evenness E5	(Ludwig & Reynolds, 1988)	0.35	0.70	0.54	0.79	0.24	0.56	0.23	0.31
	Redundancy	(Pattern, 1962)	0.59	0.69	0.64	0.78	0.35	0.55	0.27	0.33
Dominance indices	Berger-Parker	(Berger & Parker, 1970)	0.51	0.70	0.64	0.80	0.34	0.54	0.23	0.34
	McNaughton	(McNaughton, 1967)	0.51	0.64	0.64	0.74	0.32	0.47	0.18	0.32

Table 18: Predictive performance in terms of RMSE of MTs, IBk, MLP and MLR for all indices using data from field (F.A.) and simulated assemblages (S.A.) evaluated by 10-CV.

		MTs		IBk		MLP		MLR	
		F.A.	S.A.	F.A.	S.A.	F.A.	S.A.	F.A.	S.A.
Diversity indices	Abundance	4.2E+04	4.2E+04	4.9E+04	4.9E+04	7.0E+04	7.0E+04	9.0E+04	9.0E+04
	Sp. Richness	5.57	2.63	3.93	2.19	5.30	3.23	7.48	4.99
	Margalef	0.36	0.13	0.31	0.11	0.43	0.16	0.54	0.24
	Gleason	0.36	0.12	0.31	0.10	0.44	0.15	0.53	0.23
	Menhinick	0.06	0.02	0.03	0.02	0.04	0.03	0.06	0.05
	Odum	1.28	0.71	0.67	0.55	0.86	0.78	1.40	1.14
	Simpson	0.14	0.04	0.13	0.03	0.17	0.05	0.23	0.07
	H2-Shannon	0.60	0.10	0.54	0.09	0.78	0.12	1.00	0.17
	Hill N1	2.33	0.60	2.05	0.50	2.95	0.66	3.71	0.96
	Hill N2	1.90	0.61	1.60	0.49	2.20	0.69	2.82	1.07
	Hurlbert	0.14	0.04	0.13	0.03	0.17	0.05	0.23	0.07
	McIntosh	0.12	0.04	0.10	0.03	0.14	0.05	0.19	0.07
Evenness indices	Evenness E1	0.17	0.07	0.12	0.06	0.17	0.09	0.23	0.14
	Evenness E2	0.15	0.11	0.12	0.09	0.16	0.13	0.23	0.21
	Evenness E3	0.19	0.12	0.13	0.09	0.17	0.14	0.24	0.22
	Evenness E4	0.08	0.05	0.07	0.04	0.08	0.06	0.11	0.09
	Evenness E5	0.11	0.06	0.10	0.05	0.12	0.07	0.16	0.11
Dominance indices	Redundancy	0.17	0.07	0.12	0.06	0.16	0.09	0.23	0.14
	Berger-Parker	0.15	0.04	0.13	0.04	0.17	0.06	0.23	0.10
	McNaughton	0.13	0.05	0.11	0.04	0.15	0.05	0.20	0.08

6.4.3 Prediction performance of optimal algorithm

The distribution of the percent error of prediction for the above indices is depicted in Fig. 13. Half of the produced errors fall within a ±10% range for all four indices. Moreover, almost all errors do not exceed a ±30% limit. Menhinick and Berger-Parker seem to be overestimated by IBk giving positive error values. On the other hand, the median is close to zero for Species Richness and Evenness E2, while the skewness is similar to a normal distribution indicating that IBk does not unilaterally overestimate or underestimate these indices.

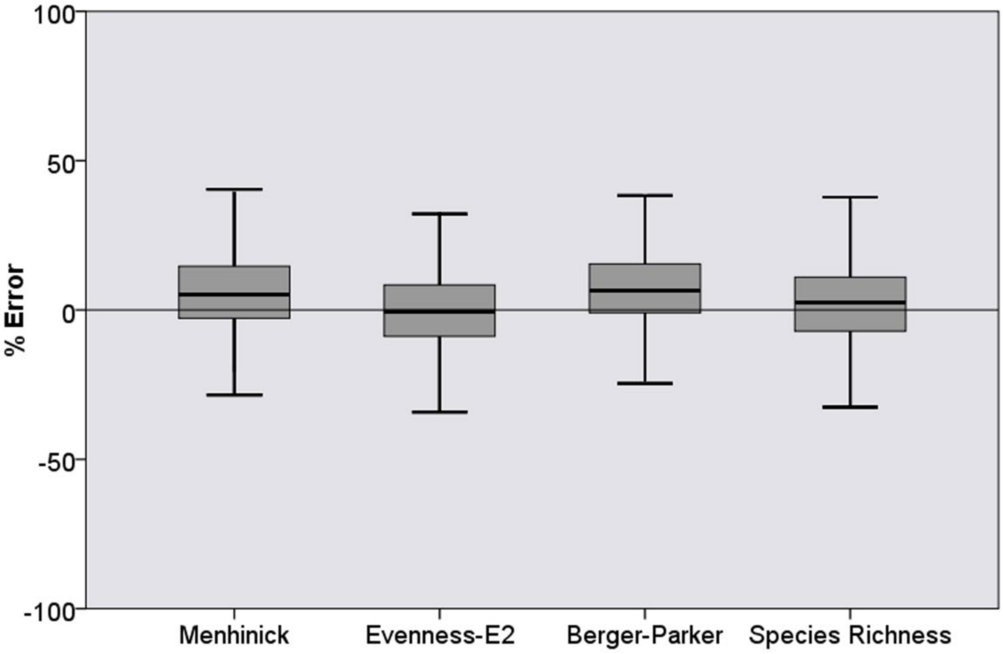


Figure 13: Box-plots of the percent errors for predicted values by the 4 best performing diversity indices that were calculated on simulated assemblages. Prediction was based on IBk algorithm.

For each of the four sampling areas, monthly data of Evenness E2 index calculated on simulated and field assemblages were compared with the corresponding predicted values by the IBk algorithm (Fig. 14). The deviation between predicted and field or simulated values was expressed quantitatively by calculating the Mean Absolute Error (MAE). Monthly predictions of Evenness E2 using the simulated assemblages shown in Fig. 14a were more accurate (MAE=0.048) than the values calculated on field data in Fig. 14b

(MAE=0.053). However, the latter can be also considered as satisfactory indicating that IBk performs with high precision using mean monthly values not only for noise-free simulated data but also for field data. IBk predictions were least accurate for both simulated and field data in the case of Gera Gulf. This area is characterized by mesotrophic conditions, and for this reason the response of phytoplankton diversity to physico-chemical parameters is likely to be more unpredictable.

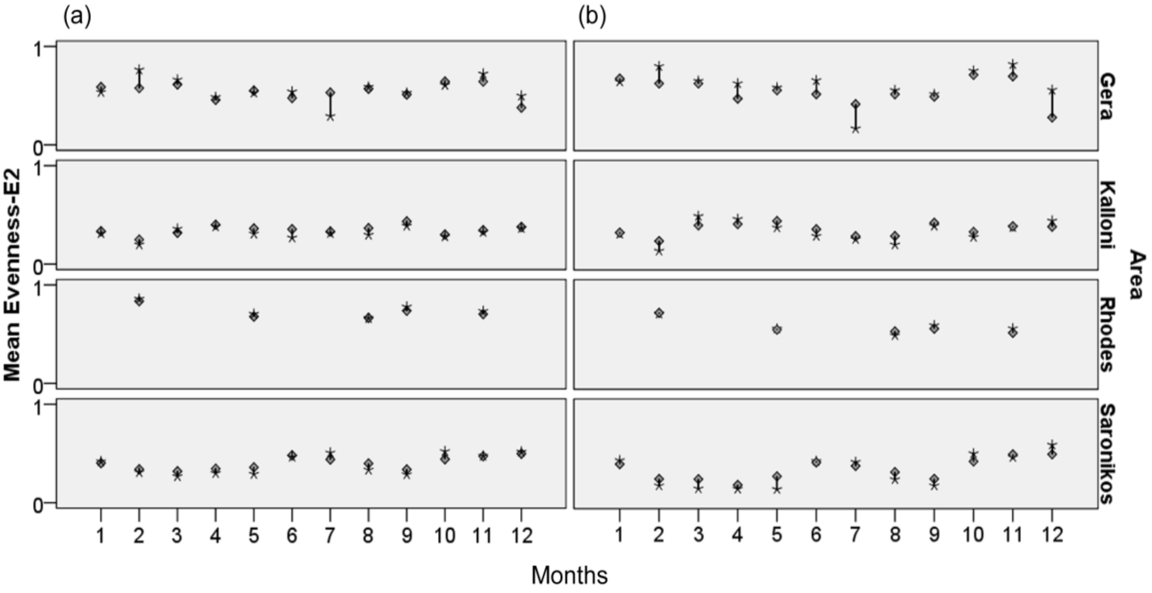


Figure 14: Monthly IBk predictions of Evenness E2 (shown with rhombus) for each sampling area in comparison with the corresponding (a) simulated and (b) field data (shown with stars).

6.5 DISCUSSION

Three novel ML techniques and 19 indices were tested in order to achieve the best biodiversity prediction using exclusively abiotic parameters. Algorithm training was based on an extensive dataset containing biotic (phytoplankton species abundances) and physico-chemical information representative of a wide productivity range of E. Mediterranean Sea. Biodiversity prediction, particularly in the marine environment, is a complex task as multiple factors and stochastic processes are acting upon community structure (Adjou et al., 2012; Gontier et al., 2006). This problem was overcome by using diversity indices calculated on simulated assemblages, free of environmental noise.

The use of powerful modelling tools such as MLs and the further optimization of the methodology with simulated assemblages provided an integrated framework for biodiversity prediction with high predictive power (R>0.80 for all selected indices between predicted and simulated values).

The simulated phytoplankton assemblages used in this study maintained the structural characteristics of the corresponding field assemblages across a wide productivity range (Tsirtsis et al., 2008), but were also free of noise related to stochastic extrinsic factors such as patchiness, grazing, and seasonality (Karydis, 1996). This property improved the relationship of diversity indices with abiotic parameters, given that noise renders algorithms sensitive to misleading (McCune, 1997; Van Straten, 1992). It also increased or even doubled the predictive power of algorithms while maintaining the realism of the natural system. Simulated communities originating from field ones have been successfully used in the past to investigate the behavior of diversity indices in microbes (Blackwood et al., 2007; Schloss and Handelsman, 2006), benthos (Lyashevska and Farnsworth, 2012), and phytoplankton (Tsirtsis and Spatharis, 2011).

Our results indicate that ML techniques can greatly increase the predictive power of models; however, the three algorithms presented significant differences in their predictive performance. IBk was the most efficient and reliable in biodiversity prediction in agreement with other marine applications of this algorithm (Dzeroski and Drumm, 2003; Hatzikos et al., 2008) or other scientific disciplines such as hydrology, weather forecasting, bioinformatics, banking and forensics (Bannayan and Hoogenboom, 2008; Bhasin et al., 2005; Buchholz et al., 2009; Diplaris et al., 2005; Hinwood et al., 2006; Solomatine et al., 2006, 2008). The observed increased efficiency of IBk in our study can be explained considering the heterogeneous structure of our dataset compiled from four different coastal areas, each one showing variability on a monthly basis. In this algorithm, every single input instance can be dynamically used with equal weight during prediction (Aha et al., 1991). Therefore, when indices are associated to the abiotic information, IBk maintains the localized information of the data in the heterogeneous dataset (Solomatine et al., 2008). This also makes IBk sensitive to instances that deviate from the main trends giving a more accurate prediction.

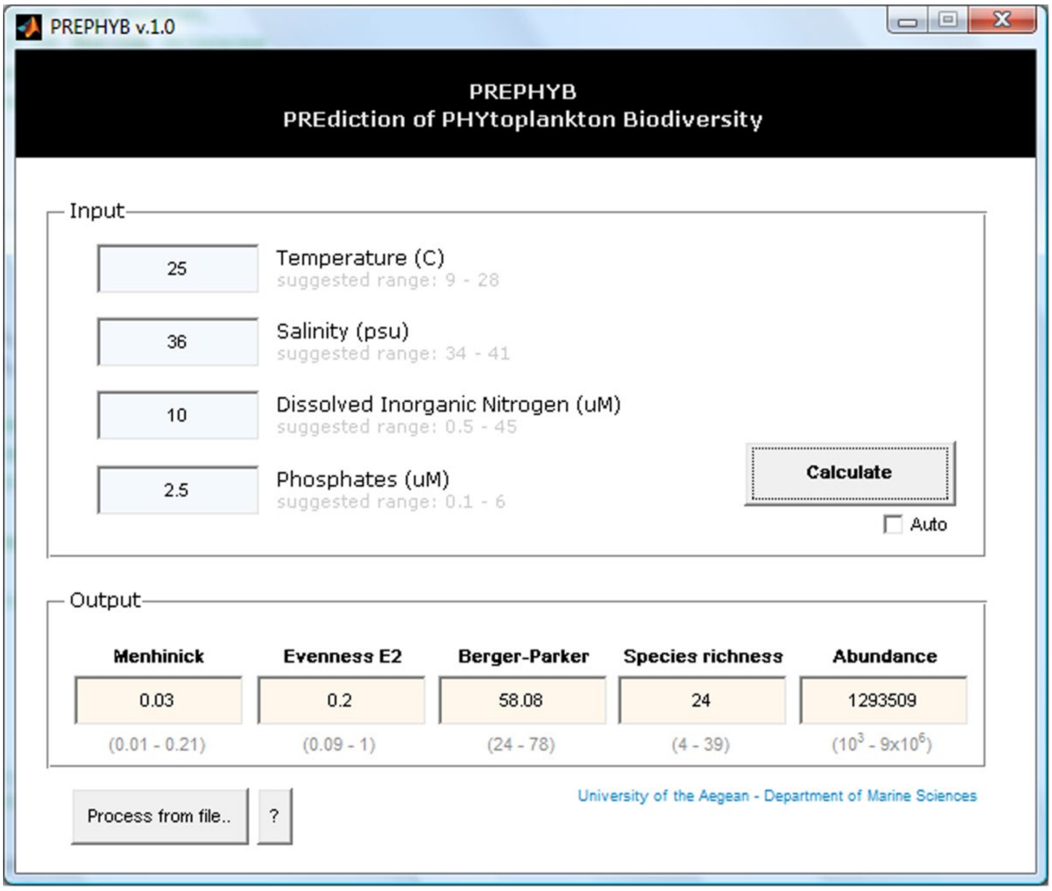
Instances that deviate from main trends (that characterize our heterogeneous dataset) are missed by algorithms such as MTs and MLP, resulting in reduced sensitivity. Contrary to IBk, these algorithms attempt to derive general relationships between diversity indices and abiotic parameters, described by linear models in MTs or weighted neurons in MLP. The latter has been proposed as a reliable model of ecological processes (Basheer and Hajmeer, 2000; Lek et al., 1996) however, its efficiency depends upon choosing the correct topology (i.e. number of layers and neurons) and applying elaborate adjustments such as pruning, constructive algorithms or recurrence (Rocha et al., 2007; Wang et al., 1994). Although these adjustments may improve predictive performance, they dramatically increase algorithm complexity and thus application runtime. In the present case study, MLP was applied in its simplest form and its efficiency was inferior compared to both MT and IBk algorithms. MLP has also shown weakness to give accurate predictions compared to other ML techniques in several other studies (e.g. Etemad-Shahidi and Mahjoobi, 2009; Nisanci et al., 2011; Solomatine and Siek, 2006; Soysal and Schmidt, 2010).

Almost all indices calculated on simulated assemblages were sufficiently predicted by the IBk algorithm. However, Menhinick's, Evenness E2 and Berger-Parker which scored higher based on their R values, are proposed for predicting the three diversity components namely richness, evenness and dominance. Although calculations on simulated data increase the predictive power of algorithms, satisfactory predictions can be also made with field data. We tested the predictive power using 658 discrete samples but also by pooling together data from different stations within a sampling campaign at a given study site. The latter predictions were much more accurate since the use of averaged data smoothed the effect of time, space (local dimensionality), and outlying values in agreement with previous studies (Kumar, 2000; More and Deo, 2003).

Presently we propose an optimization procedure for biodiversity prediction based on few abiotic parameters. Although optimization was based on phytoplankton data, this methodology can be easily adapted for any group of organisms, provided that there are sufficient samples covering a wide range of environmental conditions so that biodiversity can be fully represented. The proposed models are based on a black-box approach and do not offer mechanistic explanations for the observed relations between abiotic variables and diversity; however the performance of a sensitivity analysis in a future work could reveal the underlying processes and shed light on theoretical aspects (Refsgaard et al., 2007). The high predictive power (expressed with R correlation coefficient) in diversity prediction that the proposed methodology provides, enables its integration in various crucial ecological implementations.

6.6 SOFTWARE FEATURES

PREdiction of PHYtoplankton Biodiversity (PREPHYB) is a MATLAB-based software with a user-friendly interface (Fig. 15) that is freely downloadable at http://www.mar.aegean.gr/biodiv/Prephyb. The software provides the optimal phytoplankton diversity prediction with high predictive power, implementing the IBk algorithm and methodological scheme described in Fig. 9. PREPHYB incorporates an extensive dataset of 658 samples, and the built-in IBk is trained through the relationship between physico-chemical parameters and indices that are calculated on noise-free simulated phytoplankton assemblages. User input is limited to four abiotic variables i.e. temperature, salinity, DIN and PO_4; these can be either entered manually, or automatically processed in batch mode through a standard comma-separated ASCII file. The output consists of the predicted diversity, which corresponds to a wide productivity range typical of coastal and offshore waters of the Eastern Mediterranean Sea (10^3-9×10^6 cells/L), expressed by indices representing all three diversity components (richness, evenness, dominance), as well as additional descriptors of phytoplankton assemblage structure such as species richness and cell number. It must be noted that the dataset used for model training (abiotic variables and corresponding phytoplankton assemblages) is characteristic of Eastern Mediterranean waters as mentioned above. Therefore the use of PREPHYB as it is for biodiversity prediction with the already stated accuracy is limited for waters of similar characteristics.



PREPHYB v.1.0

PREPHYB
PREdiction of PHYtoplankton Biodiversity

Input

Temperature (C): 25
suggested range: 9 - 28

Salinity (psu): 36
suggested range: 34 - 41

Dissolved Inorganic Nitrogen (uM): 10
suggested range: 0.5 - 45

Phosphates (uM): 2.5
suggested range: 0.1 - 6

□ Auto

Output

Menhinick	Evenness E2	Berger-Parker	Species richness	Abundance
0.03	0.2	58.08	24	1293509
(0.01 - 0.21)	(0.09 - 1)	(24 - 78)	(4 - 39)	(10^3 - $9x10^6$)

University of the Aegean - Department of Marine Sciences

Figure 15: Graphical user interface of the PREPHYB software developed in MATLAB. Prediction of four indices and abundance of phytoplankton assemblages is based on abiotic variables that are either manually entered by the user, or batch processed from a comma-separated ASCII file.

7 CONCLUSIONS - CONTRIBUTION TO ICZM

In the present work a wide range of supervised ML techniques was used in order to investigate various aspects of coastal eutrophication i.e. chl *α* prediction, water quality classification and phytoplankton diversity prediction. All these tasks were performed for Aegean waters in Eastern Mediterranean by exclusively using abiotic parameters as cause variables.

MTs, being less popular as an ML method, were assessed for their efficiency to predict chl *α*under high environmental variability usually encountered in coastal ecosystems affected by terrestrial runoff. Compared to MLPs and MLR, MT method showed (a) increased predictive power, (b) higher sensitivity to discriminate different abiotic conditions driving chl *α* variability, (c) ability to scale parameters affecting chl *α* variability, and (d) easiness of application. For these reasons, MTs are recommended for the investigation of eutrophication-related ecosystem processes offering new knowledge on chl *a* dynamics from existing datasets. Based on the MT results, within each annual cycle (wet and dry) chl *α* variability occurred on a seasonal basis (and not spatial) and important differences were detected between the two meteorological regimes since chl *α* seasonality was affected by quite different abiotic factors. The efficiency of MTs to identify variables driving chl *α*, and thus eutrophication, can be invaluable in ICZM, since most of these variables are strongly linked to terrestrial processes. By reducing nutrient inputs (e.g. phosphate), or altering freshwater inflow that affects salinity, effects on chl *α* can be estimated using MTs. Therefore useful cause-and-effect relationships can be established between terrestrial processes and the response of the marine ecosystem (Tsirtsis et al., 2008), a prerequisite of modern approaches in ICZM. It must be stressed however, that a sufficient number of samples must be available in each tree leaf and variables need to be standardized in order to scale their importance in describing chl *α* variability.

Chl *α*variability was also studied in a different context that is water quality status classification for the needs of the European WFD. The ML algorithm training for this aim (a) highlighted the base classifiers with the higher accuracy and (b) showed that EMs such as voting algorithm can offer better performance than single base classifiers. Moreover, the proposed DP index can effectively show the way that voting EM could succeed higher performance during ecological quality classification of coastal water bodies, by identifying the best feeding combination triad of classifiers. Therefore, DP in combination with the newly proposed EMs can be incorporated as an information technology (IT) tool to assist one of the main aims of WFD i.e. the continuous monitoring for the protection of coastal waters. Moreover, the proposed methodology can link water quality status with basic abiotic parameters, and therefore offer new insights towards the prevention of water bodies' deterioration due to nutrient enrichment and support the achievement of the demand for *good quality status* by 2015.

The successful prediction of phytoplankton diversity from abiotic parameters offers important new insights on 'phytoplankton', often represented in ecological models in terms of biomass of one or few components characteristic of different size classes or main groups (Arhonditsis et al., 2006). Based on the proposed methodology, a link is being established between the most important abiotic variables and diversity, therefore the whole diversity spectrum and its dynamics can be incorporated into an ecological model (Laniak et al., 2013). This approach supports both the testing of ecological questions regarding diversity, as well as environmental quality assessment and protection, since changes in diversity are a focal point in recent environmental protection measures, as in the WFD. In this context, diversity prediction can be incorporated in models testing the effect of different scenarios of climate change, habitat loss, or ecosystem management. For phytoplankton in particular, diversity across a wide productivity range was predicted from temperature, salinity, DIN and PO_4. Therefore important changes in phytoplankton structure can be foreseen based on temperature projections related to climate change scenarios, or in connection to nutrient loading originating from potential changes in land use and management practices.

8 FUTURE WORK

Future perspectives may either fall into the more ‘technical’ part aiming to the optimization/elaboration of existing approaches by exploring ML capabilities and sensitivity, or in the more ‘ecological’ part by raising and answering ‘new’ ecological questions, or both of the above.

For the first approach, application of promising ML ensembles such as stacking, bagging or boosting, which have not been yet applied to the marine environment, may succeed in better predictive performance, or interpret the eutrophication phenomenon in a more comprehensive way. Other studies may include the identification of the most appropriate input variables or training instances to participate in the ML techniques, as well as the better calibration of the various algorithms. Furthermore, recent studies introduce methods such as genetic algorithms or fuzzy logic in combination with NNs to achieve improved accuracy. Application of such combinations in our ecological questions with the existing databases, may improve the less efficient performance of NNs compared to other ML techniques, observed in the present study.

Towards the second approach of ecological perspective, the ML techniques and the IT tools developed in the present work can be further used to quantify the relative impact of each input abiotic parameter to primary production, phytoplankton biodiversity and water quality status. To this aim an extensive sensitivity analysis can be performed using the available predictive tools (i.e. MTs, DP or PREPHYB) contributing to the identification of the main mechanisms involved in eutrophication processes. Moreover, model sets of instances representative of extreme conditions, climate changes or stressed ecosystems can be developed and feed the IT predictive tools aiming to assess the resulting changes in phytoplankton biomass or biodiversity. This information is valuable for managerial purposes either in local/regional, or in global scales. In this context, possible coupling of the ML methods with other powerful tools such as GIS, hydrodynamic or watershed models would be desirable.

9 REFERENCES

Aguilera, P. A., Fernandez, A., Fernandez, R., Rumi, R., and Salmeron, A. (2011). Bayesian networks in environmental modelling. Environmental Modelling & Software 26, 1376-1388.

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. Machine Learning 6, 37-66.

Allen, J. I., Holt, J. T., Blackford, J., and Proctor, R. (2007). Error quantification of a high-resolution coupled hydrodynamic-ecosystem coastal-ocean model: Part 2. Chlorophyll-a, nutrients and SPM. Journal of Marine Systems 68, 381-404.

Altunkaynak, A. (2013). Prediction of significant wave heigh using geno-multilayer perceptron. Ocean Engineering 58, 144-153.

Anagnostopoulos, C. N. and Iliou, T. (2010). Towards Emotion Recognition from Speech: Definition, Problems and the Materials of Research. In M. Wallace, I. Anagnostopoulos, P. Mylonas, and M. Bielikova (Eds.), Semantics in Adaptive and Personalized Services, Springer Berlin Heidelberg, pp. 127-143.

Anagnostopoulos, C. N., Iliou, T., and Giannoukos, I. (2012). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artificial Intelligence Review, 1-23.

Andersen, J. H., Conley, D. J., and Hedal, S. (2004). Palaeoecology, reference conditions and classification of ecological status: the EU Water Framework Directive in practice. Marine Pollution Bulletin 49, 283-290.

Arhonditsis, G., Tsirtsis, G., Angelidis, M. O., and Karydis, M. (2000). Quantification of the effects of nonpoint nutrient sources to coastal marine eutrophication: applications to a semi-enclosed gulf in the Mediterranean Sea. Ecological Modelling 129, 209-227.

Arhonditsis, G., Karydis, M., and Tsirtsis, G. (2003). Analysis of phytoplankton community structure using similarity indices: A new methodology for discriminating among eutrophication levels in coastal marine ecosystems. Environmental Management 31, 619-632.

Arias-Gonzalez, J. E., Acosta-Gonzalez, G., Membrillo, N., Garza-Perez, J. R., and Castro-Perez, J. M. (2012). Predicting spatially explicit coral reef fish abundance, richness and Shannon-Weaver index from habitat characteristics. Biodiversity and Conservation 21, 115-130.

Assaad, M., Bone, R., and Cardot, H. (2008). A new boosting algorithm for improved time-series forecasting with recurrent neural networks. Information Fusion 9, 41-55.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412-424.

Banfield, R. E., Hall, L. O., Bowyer, K. W., and Kegelmeyer, W. P. (2005). Ensemble diversity measures and their application to thinning. Information Fusion 6, 49-62.

Barros, R. C., Ruiz, D. D., and Basgalupp, M. P. (2011). Evolutionary model trees for handling continuous classes in machine learning. Information Sciences 181, 954-971.

Bauer, E. and Kohavi, R. (1999). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning 36, 105-139.

Beman, J. M., Arrigo, K. R., and Matson, P. A. (2005). Agricultural runoff fuels larg phytoplankton blooms in vulnerable areas of the ocean. Nature 434, 211-214.

Berger, W. H. and Parker, F. L. (1970). Diversity of planktonic Foraminifera in deep sediments. Science 168, 1345-1347.

Bergstrom, U., Sundblad, G., Downie, A. L., Snickars, M., Bostrom, C., and Lindegarth, M. (2013). Evaluating eutrophication management scenarios in the Baltic Sea using species distribution modelling. Journal of Applied Ecology 50, 680-690.

Bhasin, M. and Raghava, G. P. S. (2005). Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. Nucleic Acids Research 33, 202-207.

Bhattacharya, B. and Solomatine, D. P. (2005). Neural networks and M5 model trees in modelling water level discharge relationship. Neurocomputing 63, 381-396.

Bibi, S., Tsoumakas, G., Stamelos, I., and Vlahavas, I. (2008). Regression via Classification applied on software defect estimation. Expert Systems with Applications 34, 2091-2101.

Bini, L. M. and Thomaz, S. M. (2005). Prediction of Egeria najas and Egeria densa occurrence in a large subtropical reservoir (Itaipu Reservoir, Brazil-Paraguay). Aquatic Botany 83, 227-238.

Blackwood, C. B., Hudleston, D., Zak, D. R., and Buyer, Jeffrey S. B. (2007). Interpreting ecological diversity indices applied to terminal restriction fragment length polymorphism data: insights from simulated microbial communities. Applied & Environmental Microbiology 73, 5276-5283.

Boersma, P. and Weenink, D. (2005). Praat: doing phonetics by computer (version 4.6.09). http://www.praat.org.

Bonakdar, L. and Etemad-Shahidi, A. (2011). Predicting wave run-up on rubble-mound structures using M5 model tree. Ocean Engineering 38, 111-118.

Borja, A., Josefson, A. B., Miles, A., Muxika, I., Olsgard, F., Phillips, G., Rodriguez, J. G., and Rygg, B. (2007). An approach to the intercalibration of benthic ecological status assessment in the North Atlantic ecoregion, according to the European Water Framework Directive. Marine Pollution Bulletin 55, 42-52.

Borsuk, M. E., Stow, C. A., and Reckhow, K. H. (2004). A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. Ecological Modelling 173, 219-239.

Brakstad, F., Kvalheim, O. M., Ugland, K. I., Tjessem, K., and Bryne, K. (1994). Prediction of the Shannon Wiener diversity index from trace element profiles in sediments around the Statfjord platforms. Chemosphere 29, 1441-1465.

Breiman, L. (2001). Random Forests. Machine Learning 45, 5-32.

Brescia, M., Cavuoti, S., Paolillo, M., Longo, G., and Puzia, T. (2012). The detection of globular clusters in galaxies as a data mining problem. Monthly Notices of the Royal Astronomical Society 421, 1155-1165.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A database of German emotional speech. 9th International Conference on Spoken Language Processing (INTERSPEECH-ICSLP). ISCA, 1517-1520.

Camdevyren, H., Demyr, N., Kanik, A., and Keskyn, S. (2005). Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservois. Ecological Modelling 181, 581-589.

Canuto, A. M. P., Abreu, M. C. C., de Melo Oliveira, L., Xavier, J., and Santos, A. D. (2007). Investigating the influence of the choice of the ensemble members in accuracy and diversity of selection-based and fusion-based methods for ensembles. Pattern Recognition Letters 28, 472-486.

Carstensen, J., Sanchez-Camacho, M., Duarte, C. M., Krause-Jensen, D., and Marba, N. (2011). Connecting the dots: responses of coastal ecosystems to changing nutrient concentrations. Environmental Science & Technology 45, 9122-9132.

Catherine, A., Mouillot, D., Escoffier, N., Bernard, C., and Troussellier, M. (2010). Cost effective prediction of the eutrophication status of lakes and reservoirs. Freshwater Biology 55, 2425-2435.

Cawley, G. C. and Talbot, N. L. C. (2003). Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. Pattern Recognition 36, 2585-2592.

Chang, Y. C., Yan, J. C., Hwang, J. S., Wu, C. H., and Lee, M. T. (2011). Data-oriented analyses of ciliate foraging behaviors. Hydrobiologia 666, 223-237.

Chaudhuri, B. B. and Bhattacharya, U. (2000). Efficient training and improved performance of multilayer perceptron in pattern classification. Neurocomputing 34, 11-27.

Chen, K., Wang, L., and Chi, H. (1997). Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification. International Journal of Pattern Recognition & Artificial Intelligence 11, 417-445.

Chen, C. Y., Shyue, S. W., and Chang, C. J. (2010). Association rule mining for evaluation of regional environments: case study of Dapeng Bay, Taiwan. International Journal of Innovative Computing, Information & Control 6, 3425-3436.

Cheng, L., Lek, S., Lek-Ang, S., and Li, Z. (2012). Predicting fish assemblages and diversity in shallow lakes in the Yangtze River basin. Limnologica - Ecology & Management of Inland Waters 42, 127-136.

Cho, K. H., Kang, J. H., Ki, S. J., Park, Y., Cha, S. M., and Kim, J. H. (2009). Determination of the optimal parameters in regression models for the prediction of chlorophyll-a: a case study of the Yeongsan Reservoir, Korea. Science of the Total Environment 407, 2536-2545.

Cleary, J. G. and Trigg, L. E. (1995). K*: An instance-based learner using an entropic distance measure. (pp. 108-114). Morgan Kaufmann.

Cline, M. S. and Karchin, R. (2011). Using bioinformatics to predict the functional impact of SNVs. Bioinformatics 27, 441-448.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational & Psychological Measurement 20, 37-46.

Cohen, W. W. (1995). Fast effective rule induction. (pp. 108-114). Morgan Kaufmann.

Cohen, W. W. and Singer, Y. (1999). Context-sensitive learning methods for text categorization. ACM Transactions on Infromation Systems, 17, 141-173.

Coll, M., Piroddi, C., Steenbeek, J., Kaschner, K., Lasram, F. B., Aguzzi, J., Ballesteros, E., Bianchi, C. N., Corbera, J., Dailianis, T., Danovaro, R., Estrada, M., Froglia, C., Galil, B. S., Gasol, J. M., Gertwagen, R., Gil, J., Guilhaumon, F., Kesner-Reyes, K., Kitsos, M. S., Koukouras, A., Lampadariou, N., Laxamana, E., de la Cuadra, C. M. L. F., Lotze, H. K., Martin, D., Mouillot, D., Oro, D., Raicevich, S., Rius-Barile, J., Saiz-

Salinas, J. I., San Vicente, C., Somot, S., Templado, J., Turon, X., Vafidis, D., Villanueva, R., and Voultsiadou, E. (2010). The Biodiversity of the Mediterranean Sea: estimates, patterns, and threats. Plos One 5.

Collin, A., Archambault, P., and Long, B. (2011). Predicting Species Diversity of Benthic Communities within Turbid Nearshore Using Full-Waveform Bathymetric LiDAR and Machine Learners. Plos One 6, 1-16.

Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory IT-13, 21-27.

Crammer, K. and Singer, Y. (2002). On the learnability and design of output codes for multiclass problems. Machine Learning 47, 201-233.

Crisci, C., Ghattas, B., and Perera, G. (2012). A review of supervised machine learning algorithms and their applications to ecological data. Ecological Modelling, 240, 113-122.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J. (2007). Randomforests for classification in ecology. Ecology 88, 2783-2792.

Dakou, E., D'heygere, T., Dedecker, A. P., Goethals, P., Lazaridou-Dimitriadou, M., and De Pauw, N. (2007). Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece). Aquatic Ecology 41, 399-411.

Dawson, T. P., Jackson, S. T., House, J. I., Prentice, I. C., and Mace, G. M. (2011). Beyond predictions:biodiversity conservation in a changing climate. Science 332, 53-58.

Debeljak, M., Cortet, J., Demaar, D., Krogh, P. H., and Dzeroski, S. (2007). Hierarchical classification of environmental factors and agricultural practices affecting soil fauna under cropping systems using Bt maize. Pedobiologia 51, 229-238.

Demsar, D., Dzeroski, S., Larsen, T., Struyf, J., Axelsen, J., Pedersen, M. B., and Krogh, P. H. (2006). Using multi-objective classification to model communities of soil microarthropods. Ecological Modelling 191, 131-143.

Denisenko, N. V. (2010). The description and prediction of benthic biodiversity in high arctic and freshwater-dominated marine areas: The southern Onega Bay (the White Sea). Marine Pollution Bulletin 61, 224-233.

Diaz-Uriarte, R. and Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3.

Dietterich, T. (1997). Machine learning research: Four current directions. AI Magazine 18, 97-136.

Dietterich, T. (2000a). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning 40, 139-157.

Dietterich, T. (2000b). Ensemble Methods in Machine Learning. Multiple Classifier Systems (pp. 1-15). Springer Berlin Heidelberg.

Dimitriadou, E., Weingessel, A., and Hornik, K. (2001). Voting-Merging: An Ensemble Method for Clustering. In G. Dorffner, H. Bischof, and K. Hornik (Eds.), Artificial Neural Networks - ICANN 2001 (pp. 217-224). Springer Berlin Heidelberg.

Dominguez-Granda, L., Lock, K., and Goethals, P. L. M. (2011). Using multi-target clustering trees as a tool to predict biological water quality indices based on benthic macroinvertebrates and environmental parameters in the Chaguana watershed (Ecuador). Ecological Informatics 6, 303-308.

Dzeroski, S. (2001). Applications of symbolic machine learning to ecological modelling. Ecological Modelling 146, 263-273.

Dzeroski, S. and Drumm, D. (2003). Using regression trees to identify the habitat preference of the sea cucumber (Holothuria leucospilota) on Rarotonga, Cook Islands. Ecological Modelling 170, 219-226.

Dzeroski, S. and Zenko, B. (2004). Is Combining Classifiers with Stacking Better than Selecting the Best One? Machine Learning, 54 255-273.

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. Journal of the American Statistical Association 78, 316-331.

Etemad-Shahidi, A. and Mahjoobi, J. (2009). Comparison between M5' model tree and neural networks for prediction of significant wave height in Lake Superior. Ocean Engineering 36, 1175-1181.

Everaert, G., Boets, P., Lock, K., Dzeroski, S., and Goethals, P. L. M. (2011). Using classification trees to analyze the impact of exotic species on the ecological assessment of polder lakes in Flanders, Belgium. Ecological Modelling 222, 2202-2212.

Fernandes, J. A., Irigoien, X., Goikoetxea, N., Lozano, J. A., Inza, I., Perez, A., and Bode, A. (2010). Fish recruitment prediction, using robust supervised classification methods. Ecological Modelling 221, 338-352.

Fielding, A. H. (1999). Machine learning methods for ecological applications. Boston: Kluwer Academic Publishers.

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individurals in a random sample of an animal population. Journal of Animal Ecology 12, 42-58.

Flury, B. and Riedwyl, H. (1988). Multiple linear regression. Multivariate Statistics (pp. 54-74). Springer Netherlands.

Fonseca, M., Whitfield, P. E., Kelly, N. M., and Bell, S. S. (2002). Modeling seagrass landscape pattern and associated ecological attributes. Ecological Applications 12, 218-237.

Fragopanagos, N. and Taylor, J. G. (2005). Emotion recognition in human-computer interaction. Neural Networks 18, 389-405.

Frank, E., Wang, Y., Inglis, S., Holmes, G., and Witten, I. H. (1998). Technical note: Using model trees for classification. Machine Learning 32, 63-76.

Frank, E. and Witten, I. H. (1998). Generating accurate rule sets without global optimization. (pp. 144-151). Morgan Kaufmann, San Francisco, CA.

Freeman, A. M., Lamon III, E. C., and Stow, C. A. (2009). Nutrient criteria for lakes, ponds, and reservoirs: A Bayesian TREED model approach. Ecological Modelling 220, 630.

Furnkranz, J. (1997). Pruning Algorithms for Rule Learning. Machine Learning 27, 139-172.

Gal, G., Skerjanec, M., and Atanasova, N. (2013). Fluctuations in water level and the dynamics of zooplankton: a data-driven modelling approach. Freshwater Biology 58, 800-816.

Gontier, M., Balfors, B., and Mortberg, U. (2006). Biodiversity in environmental assessment-current practice and tools for prediction. Environmental Impact Assessment Review 26, 268-286.

Grabar, N. and Krivine, S. (2007). Application of cross-language criteria for the automatic distinction of expert and non expert online health documents. In R. Bellazzi, A. bu-Hanna, and J. Hunter (Eds.), Artificial Intelligence in Medicine (pp. 252-256). Springer Berlin Heidelberg.

Haigang, Z., Ping, S., and Chen, C. (2003). Retrieval of oceanic chlorophyll concentration using support vector machines. IEEE Transactions on Geoscience & Remote Sensing 41, 2947-2951.

Halabi Echeverry, A. X., Richards, D., and Bilgin, A. (2012). Identifying characteristics of seaports for environmental benchmarks based on meta-learning. In D. Richards and B. Kang (Eds.), Knowledge Management and Acquisition for Intelligent Systems (pp. 350-363). Springer Berlin Heidelberg.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. SIGKDD Exlporations 11, 10-18.

Hatzikos, E. V., Tsoumakas, G., Tzanis, G., Bassiliades, N., and Vlahavas, I. (2008). An empirical study on sea water quality prediction. Knowledge-Based Systems 21, 471-478.

Herrera, P., Yeterian, A., and Gouyon, F. (2002). Automatic Classification of Drum Sounds: A Comparison of Feature Selection Methods and Classification Techniques. In C. Anagnostopoulou, M. Ferrand, and A. Smaill (Eds.), Music and Artificial Intelligence (pp. 69-80). Springer Berlin Heidelberg.

Huang, C. C. (2006). A novel gray-based reduced NN classification method. Pattern Recognition 39, 1979-1986.

Huang, G. B., Ding, X., and Zhou, H. (2010). Optimization method based extreme learning machine for classification. Neurocomputing 74, 155-163.

Huang, W. and Foo, S. (2002). Neural network modeling of salinity variation in Apalachicola River. Water Research 36, 356-362.

Huhn, J. and Hullermeier, E. (2009). FURIA: an algorithm for unordered fuzzy rule induction. Data Mining & Knowledge Discovery 19, 293-319.

Hulbert, S. H. (1971). The nonconcept of species diversity: a critique and alternative parameters. Ecology 59, 67-77.

Ignatiades, L., Karydis, M., and Vounatsou, P. (1992). A possible method for evaluating oligotrophy and eutrophication based on nutrient concentration scales. Marine Pollution Bulletin 24, 238-243.

Iliou, T. and Anagnostopoulos, C. N. (2010). Classification on speech emotion recognition - A comparative study. International Journal on Advances in Life Sciences 2, 18-28.

Ingram, T. and Steel, M. (2010). Modelling the unpredictability of future biodiversity in ecological networks. Journal of theoretical biology 264, 1047-1056.

Ishibuchi, H., Nakashima, T., and Morisawa, T. (1999). Voting in fuzzy rule-based systems for pattern classification problems. Fuzzy Sets & Systems 103, 223-238.

Jaccard, P. (1908). Nouvelles recherctus sur la distribution floral. Bulletin de la Societe Vaudense des Sciences Naturelles 44, 223-270.

Jachowski, N. R. A., Quak, M. S. Y., Friess, D. A., Duangnamon, D., Webb, E. L., and Ziegler, A. D. (2013). Mangrove biomass estimation in Southwest Thailand using machine learning. Applied Geography 45, 311-321.

Jain, P., Deo, M. C., Latha, G., and Rajendran, V. (2011). Real time wave forecasting using wind time history and numerical model. Ocean Modelling 36, 26-39.

Jeong, K. S., Kim, D. K., Jung, J. M., Kim, M. C., and Joo, G. J. (2008). Non-linear autoregressive modelling by temporal recurrent neural networks for the prediction of freshwater phytoplankton dynamics. Ecological Modelling 211, 292-300.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, San Mateo. (pp. 338-345). Morgan Kaufmann Publishers.

Junker, K., Sovilj, D., Kroncke, I., and Dippner, J. W. (2012). Climate induced changes in benthic macrofauna - a non-linear model approach. Journal of Marine Systems 96-97, 90-94.

Jurc, M., Perko, M., Dzeroski, S., Demsar, D., and Hrasovec, B. (2006). Spruce bark beetles (Ips typographus, Pityogenes chalcographus, Col.: Scolytidae) in the Dinaric mountain forests of Slovenia: Monitoring and modeling. Ecological Modelling 194, 219-226.

Kaburlasos, V. G. and Pachidis, T. (2014). A Lattice-Computing ensemble for reasoning based on formal fusion of disparate data types, and an industrial dispensing application. Information Fusion 16, 68-83.

Kanevski, M., Parkin, R., Pozdnukhov, A., Timonin, V., Maignan, M., Demyanov, V., and Canu, S. (2004). Environmental data mining and modeling based on machine learning algorithms and geostatistics. Environmental Modelling & Software 9, 845-855.

Karul, C., Soyupak, S., Cilesiz, A. F., Akbay, N., and Germen, E. (2000). Case studies on the use of neural networks in eutrophication modeling. Ecological Modelling 134, 145-152.

Karydis, M. (1996). Quantitative assessment of eutrophication: A scoring system for characterising water quality in coastal marine ecosystems. Environmental Monitoring & Assessment 41, 233-246.

Karydis, M. and Tsirtsis, G. (1996). Ecological indices: a biometric approach for assessing eutrophication levels in the marine environment. Science of The Total Environment 186, 209-219.

Keating, K. A. and Cherry, S. (2004). Use and interpretation of logistic regression in habitat selection studies. Journal of Wildlife Management 68, 774-789.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. K. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation 13, 637-649.

Kehoe, M., O' Brien, K., Grinham, A., Rissik, D., Ahern, K. S., and Maxwell, P. (2012). Random forest algorithm yields accurate quantitative prediction models of benthic light at intertidal sites affected by toxic Lyngbya majuscula blooms. Harmful Algae 19, 46-52.

Keiner, L. E. and Yan, X. H. (1998). A Neural Network Model for Estimating Sea Surface Chlorophyll and Sediments from Thematic Mapper Imagery. Remote Sensing of Environment 66, 153-165.

Khater, M. and Gras, R. (2012). Adaptation and Genomic Evolution in EcoSim. In T. Ziemke, C. Balkenius, and J. Hallam (Eds.), From Animals to Animats 12 (pp. 219-229). Springer Berlin Heidelberg.

Kitsiou, D., Coccossis, H., and Karydis, M. (2002). Multi-dimensional evaluation and ranking of coastal areas using GIS and multiple criteria choice methods. Science of тhe Total Environment 284, 1-17.

Kitsiou, D. and Karydis, M. (2011). Coastal marine eutrophication assessment: A review on data analysis. Environment International 37, 778-801.

Kittler, J., Hatef, M., Duin, R. W. D., and Matas, J. (1998). On combining classifiers. IEEE Transactions on Pattern Analysis & Machine Intelligence 20, 226-239.

Knudby, A., LeDrew, E., and Brenning, A. (2010). Predictive mapping of reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques. Remote Sensing of Environment 114, 1230-1241.

Kocev, D. and Dzeroski, S. (2013). Habitat modeling with single and multi target trees and ensembles. Ecological Informatics 18, 79-92.

Kocev, D., Dzeroski, S., White, M. D., Newell, G. R., and Griffioen, P. (2009). Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition. Ecological Modelling 220, 1159-1168.

Kocev, D., Naumoski, A., Mitreski, K., Krsti, S., and Dzeroski, S. (2010). Learning habitat models for the diatom community in Lake Prespa. Ecological Modelling 221, 330-337.

Kothari, R. and Dong, M. (2001). Decision trees for classification: a review and some new results. Pattern Recognition (pp. 169-184). World Scientific.

Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review 26, 159-190.

Kotsiantis, S. B. (2007). Supervised learning: A review of classification techniques. Informatica 31, 249-268.

Krebs, C. J. (1999). Ecological methodology. Monlo Park, California: Addison Wesley Longman.

Kubat, M., Holte, R., and Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. Machine Learning 30, 195-215.

Kuncheva, L. I., Whitaker, C. J., Shipp, C. A., and Duin, R. P. W. (2000). Is independence good for combining classifiers? Proceedings of the 15th International Conference on Pattern Recognition, 168-171.

Kuncheva, L. and Whitaker, C. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning 51, 181-207.

Kuncheva, L. I., Whitaker, C. J., Shipp, C. A., and Duin, R. P. W. (2003). Limits on the majority vote accuracy in classifier fusion. Pattern Analysis & Applications 6, 22-31.

Kuncheva, L. I. (2004). Combining pattern classifiers: methods and algorithms. Hoboken, New Jersey: John Wiley & Sons, Inc.

Kuncheva, L. I. and Hadjitodorov, S. T. (2004). Using diversity in cluster ensembles. IEEE International Conference on Systems, Man & Cybernetics 2, 1214-1219.

Kuo, J. T., Hsieh, M. H., Lung, W. S., and She, N. (2007). Using artificial neural network for reservoir eutrophication prediction. Ecological Modelling 200, 171-177.

Lamon, E. C., III, Malve, O., and Pietilainen, O. P. (2008). Lake classification to enhance prediction of eutrophication endpoints in Finnish lakes. Environmental Modelling & Software 23, 938-947.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics 33, 159-174.

Lane, J. Q., Raimondi, P. T., and Kudela, R. M. (2009). Development of a logistic regression model for the prediction of toxigenic Pseudo-nitzschia blooms in Monterey Bay, California. Marine Ecology Progress Series 383, 37-51.

Laniak, G. F., Olchin, G., Goodall, J., Voinov, A., Hill, M., Glynn, P., Whelan, G., Geller, G., Quinn, N., Blind, M., Peckham, S., Reaney, S., Gaber, N., Kennedy, R., and Hughes, A. (2013). Integrated environmental modeling: a vision and roadmap for the future. Environmental Modelling & Software 39, 3-23.

Laskov, P., D+-ssel, P., Sch+vfer, C., and Rieck, K. (2005). Learning Intrusion Detection: Supervised or Unsupervised? In F. Roli and S. Vitulano (Eds.), Image Analysis and Processing ICIAP 2005 (pp. 50-57). Springer Berlin Heidelberg.

le Cassie, S. and van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. Applied Statistics 41, 191-201.

Lek, S. and Guegan, J. F. (1999). Artificial neural networks as a tool in ecological modeling - an introduction. Ecological Modelling 120, 65-73.

Lek, S. and Park, Y. S. (2008). Multilayer Perceptron. In Editors-in-Chief:TıTıSven Erik Jorgensen and F. Brian (Eds.), Encyclopedia of Ecology (pp. 2455-2462). Oxford: Academic Press.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In C. Nedellec and C. Rouveirol (Eds.), Machine Learning: ECML-98 (pp. 4-15). Springer Berlin Heidelberg.

Lewis, K. and Allen, J. I. (2009). Validation of a hydrodynamic-ecosystem model simulation with time-series data collected in the western English Channel. Journal of Marine Systems 77, 296-311.

Li, K., Liu, Z., and Han, Y. (2012). Study of Selective Ensemble Learning Methods Based on Support Vector Machine. Physics Procedia 33, 1518-1525.

Libralon, G. L., Carvalho, A. C. P. de L. F. de, and Lorena, A. C. (2009). Pre-processing for noise detection in gene expression classification data. Journal of the Brazilian Computer Society 15, 3-11.

Liu, Y., Guo, H., and Yang, P. (2010). Exploring the influence of lake water chemistry on chlorophyll a: A multivariate statistical model analysis. Ecological Modelling 221, 681-688.

Lockwood, M., Davidson, J., Hockings, M., Haward, M., and Kriwoken, L. (2012). Marine biodiversity conservation governance and management: Regime requirements for global environmental change. Ocean & Coastal Management 69, 160-172.

Loh, W. Y. (2008). Classification and Regression Tree Methods. Encyclopedia of Statistics in Quality and Reliability John Wiley & Sons, Ltd.

Lorena, A. C., Jacintho, L. F. O., Siqueira, M. F., Giovanni, R. D., Lohmann, L. G., de Carvalho, A. C. P. L., and Yamamoto, M. (2011). Comparing machine learning classifiers in potential distribution modelling. Expert Systems with Applications 38, 5268-5275.

Lowery, T. A. (1998). Modelling estuarine eutrophication in the context of hypoxia, nitrogen loadings, stratification and nutrient ratios. Journal of Environmental Management 52, 289-305.

Loza Mencia, E., Park, S. H., and Furnkranz, J. (2010). Efficient voting prediction for pairwise multilabel classification. Neurocomputing 73, 1164-1176.

Lu, D. and Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing 28, 823-870.

Ludtke, A., Jerosch, K., Herzog, O., and Schluter, M. (2012). Development of a machine learning technique for automatic analysis of seafloor image data: Case example, Pogonophora coverage at mud volcanoes. Computers & Geosciences 39, 120-128.

Ludwig, A. J. and Reynolds, J. F. (1988). Statistical Ecology: A primer on methods and computing. New York: Wiley Press.

Lyashevska, O. and Farnsworth, K. D. (2012). How many dimensions of biodiversity do we need? Ecological Indicators 18, 485-492.

Magurran, A. E. (2004). Measuring Biological Ecology. (2^{nd} edition). Oxford: Blackwell Science.

Mao, S., Jiao, L. C., Xiong, L., and Gou, S. (2011). Greedy optimization classifiers ensemble based on diversity. Pattern Recognition 44, 1245-1261.

Margalef, R. (1958). Information theory in ecology. General Systems 3, 36-71.

Marin-Guirao, L., Cesar, A., Marin, A., Lioret, J., and Vita, R. (2005). Establishing the ecological quality status of soft-bottom mining-impacted coastal water bodies in the scope of Water Framework Directive. Marine Pollution Bulletin 50, 374-387.

McCullagh, P. and Nelder, J. A. (1989). Generalized linear models. (2^{nd} edition). London: Chapman and Hall.

McIntosh, R. P. (1967). An index of diversity and the relation of certain concepts to diversity. Ecology 48, 392-404.

McNaughton, J. (1967). Relationship among functional properties of California grassland. Nature 216, 168-169.

Menhinick, E. P. (1964). A comparison of some species-individuals diversity indices applied to samples of field insects. Ecology 45, 859-861.

Millie, D. F., Weckman, G. R., Young II, W. A., Ivey, J. E., Carrick, H. J., and Fahnenstiel, G. L. (2012). Modeling microalgal abundance with artificial neural networks: Demonstration of a heuristic 'Grey-Box' to deconvolve and quantify environmental influences. Environmental Modelling & Software 38, 27-39.

Morrison, D., Wang, R., and De Silva, L. C. (2007). Ensemble methods for spoken emotion recognition in call-centres. Speech Communication 49, 98-112.

Mouton, A. M., Alcaraz-Hernandez, J. D., De Baets, B., Goethals, P. L. M., and Martinez-Capel, F. (2011). Data driven fuzzy habitat suitability models for brown trout in Spanish Mediterranean rivers. Environmental Modelling & Software 26, 615-622.

Musavi, M. T., Ressom, H., Srirangam, S., Natarajan, P., Virnstein, R. W., Morris, L. J., and Tweedale, W. (2007). Neural network-based light attenuation model for monitoring seagrass population in the Indian river lagoon. Journal of Intelligent Information Systems 29, 63-77.

Muxika, I., Borja, A., and Bald, J. (2007). Using historical data, expert judgement and multivariate analysis in assessing reference conditions and benthic ecological status, according to the European Water Framework Directive. Marine Pollution Bulletin 55, 16-29.

Naumoski, A. and Mitreski, K. (2010). Classifying diatoms into trophic state index classes with novel classification algorithm. Procedia Environmental Sciences 2, 1124-1138.

Nielsen, K., Somod, B., Ellegaard, C., and Krause-Jensen, D. (2003). Assessing reference conditions according to the European Water Framework Directive using modelling and analysis of historical data: an example from Randers Fjord, Denmark. AMBIO A Journal of the Human Environment 32, 287-294.

Nigam, K., Mccallum, A., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. Machine Learning 39, 103-134.

Ocampo-Duque, W., Schuhmacher, M., and Domingo, J. L. (2007). A neural-fuzzy approach to classify the ecological status in surface waters. Environmental Pollution 148, 634-641.

Odum, H. T., Cantlon, J. E., and Kornicker, L. S. (1960). An organizational hierarchy postulate for the interpretation of species-individuals distribution, species entropy and ecosystem evolution and the meaning of a species variety index. Ecology 41, 395-399.

Ogris, N. and Jurc, M. (2010). Sanitary felling of Norway spruce due to spruce bark beetles in Slovenia: A model and projections for various climate change scenarios. Ecological Modelling 221, 290.

Olden, J. D., Lawler J.Joshua, and Poff LeRoy N. (2008). Machine learning methods without tears: A primer for ecologists. The Quarterly Review of Biology 83, 171-193.

Onderka, M. (2007). Correlations between several environmental factors affecting the bloom events of cyanobacteria in Liptovska Mara reservoir (Slovakia)-A simple regression model. Ecological Modelling 209, 412-416.

Opitz, D. W. and Shavlik, J. W. (1996). Actively searching for an effective neural network ensemble. Connection Science 8, 337-354.

Opitz, D. and Maclin, R. (1999). Popular ensemble methods: an empirical study. Journal of Artificial Intelligence Recearch 11, 169-198.

Ornella, L. and Tapia, E. (2010). Supervised machine learning and heterotic classification of maize (Zea mays L.) using molecular marker data. Computers & Electronics in Agriculture 74, 250-257.

Pal, S. K. and Mitra, S. (1992). Multilayer perceptron, fuzzy sets, and classification. IEEE transactions on neural networks, a publication of the IEEE Neural Networks Council 3, 683-697.

Pal, M. and Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment 86, 554-565.

Palaiokostas, C., Bekaert, M., Davie, A., Cowan, M. E., Oral, M., Taggart, J. B., Gharbi, K., McAndrew, B. J., Penman, D. J., and Migaud, H. (2013).

Mapping the sex determination locus in the Atlantic halibut (Hippoglossus hippoglossus) using RAD sequencing. BMC Genomics 14, 566.

Parkhurst, D. F., Brenner, K. P., Dufour, A. P., and Wymer, L. J. (2005). Indicator bacteria at five swimming beaches analysis using random forests. Water Research 39, 1354-1360.

Parsons, T. R., Maita, Y., and Lalli, C. M. (1984). A Manual of Chemical and Biological Methods for Seawater Analysis. Oxford: Pergamon Press.

Pattern, B. C. (1962). Species diversity in net plankton of Raritan Bay. Journal of Marine Research 20, 57-75.

Payne, T. R. (1995). Instance-Based Prototypical Learning of Set Valued Attributes. Morgan Kaufmann Publishers.

Pearce, J. and Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133, 225-245.

Peng, F., Schuurmans, D., and Wang, S. (2004). Augmenting Naive Bayes classifiers with statistical language models. Information Retrieval 7, 317-345.

Perdiguero-Alonso, D., Montero, F. E., Kostadinova, A., Raga, J. A., and Barrett, J. (2008). Random forests, a novel approach for discrimination of fish populations using parasites as biological tags. International Journal for Parasitology, 38, 1425-1434.

Pielou, E. C. (1975). Ecological Diversity. New York: Wiley InterScience.

Pinero, P., Garcia, P., Arco, L., Alvarez, A., Garcia, M. M., and Bonal, R. (2004). Sleep stage classification using fuzzy sets and machine learning techniques. Neurocomputing 58-60, 1137-1143.

Pittman, S. J., Christensen, J. D., Caldow, C., Menza, C., and Monaco, M. E. (2007). Predictive mapping of fish species richness across shallow-water seascapes in the Caribbean. Ecological Modelling 204, 9-21.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In B. Scholkopf, C. Burges, and A. Smola (Eds.), Advances in Kernel Methods-Support Vector Learning (pp. 185-208). MIT Press.

Prasad, A. M., Iverson, L. R., and Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and Random Forests for ecological prediction. Ecosystems 9, 181-199.

Primpas, I., Tsirtsis, G., Karydis, M., and Kokkoris, G. D. (2010). Principal component analysis: Development of a multivariate index for assessing eutrophication according to the European water framework directive. Ecological Indicators 10, 178-183.

Quinlan, J. R. (1992). Learning with continuous classes. (pp. 343-348). World Scientific.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufman.

Quinlan, J. R. (1996). Learning decision tree classifiers. Acm Computing Surveys 28, 71-72.

Quinlan, J. R. (1999). Simplifying decision trees. International Journal of Human-Computer Studies 51, 497-510.

Ramin, M., Stremilov, S., Labencki, T., Gudimov, A., Boyd, D., and Arhonditsis, G. B. (2011). Integration of numerical modeling and Bayesian analysis for setting water quality criteria in Hamilton Harbour, Ontario, Canada. Environmental Modelling & Software 26, 337-353.

Rani, P., Liu, C., Sarkar, N., and Vanman, E. (2006). An empirical study of machine learning techniques for affect recognition in human-robot interaction. Pattern Analysis & Applications 9, 58-69.

Raudys, S. I. and Jain, A. K. (1991). Small sample effects in statistical pattern recognition: Recommendations for practitioners. IEEE Transactions on Pattern Analysis & Machine Intelligence 13, 252-264.

Recknagel, F. (2001). Applications of machine learning to ecological modelling. Ecological Modelling 146, 303-310.

Rocha, M., Cortez, P., and Neves, J. (2007). Evolution of neural networks for classification and regression. Neurocomputing 70, 2809-2816.

Romero, J., Martinez-Creco, B., Alcoverro, T., and Perez, M. (2007). A multivariate index based on the seagrass Posidonia oceanica (POMI) to assess ecological status of coastal waters under the water framework directive (WFD). Marine Pollution Bulletin 55, 196-204.

Ruiz, M. and Velasco, J. (2010). Nutrient Bioaccumulation in Phragmites australis: Management Tool for reduction of pollution in the Mar Menor. Water Air and Soil Pollution 205, 173-185.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. Nature 323, 533-536.

Ruta, D. and Gabrys, B. (2005). Classifier selection for majority voting. Information Fusion 6, 63-81.

Sadeghi, R., Zarkami, R., Sabetraftar, K., and Van Damme, P. (2012). Use of support vector machines (SVMs) to predict distribution of an invasive water fern Azolla filiculoides (Lam.) in Anzali wetland, southern Caspian Sea, Iran. Ecological Modelling 244, 117-126.

Saha, S. and Ekbal, A. (2013). Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. Data & Knowledge Engineering 85, 15-39.

Scardi, M. (1996). Artificial neural networks as empirical models for estimating phytoplankton production. Marine Ecology Progress Series 139, 289-299.

Scardi, M. (2003). Chapter 19 Neural network applications in coastal ecological modeling. In V. C. Lakhan (Ed.), Elsevier Oceanography Series, Advances in Coastal Modeling (pp. 505-532). Elsevier.

Schloss, P. D. and Handelsman, J. (2006).Toward a census of bacteria in soil. PLoS computational biology 2 e92, 786-793

Seifoddini, H. and Djassemi, M. (1991). The production data-based similarity coefficient versus Jaccard's similarity coefficient. Computers & Industrial Engineering 21, 263-266.

Shami, M. and Verhelst, W. (2007). An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Communication, 49, 201-212.

Shannon, C. E. and Weaver, W. (1949). The Mathematical Theory of Communication. Urbana: University of Illinois Press.

Shao, Q., Rowe, R. C., and York, P. (2007). Investigation of an artificial intelligence technology - Model trees. Novel applications for an immediate release tablet formulation database. European Journal of Pharmaceutical Sciences 31, 137.

Sharkey, A. J. C. and Sharkey, N. E. (1997). Combining diverse neural nets. The Knowledge Engineering Review 12, 231-247.

Sheldon, A. L. (1969). Equitability indices: dependence on species count. Ecology 50, 466-467.

Shi, J. J. (2000). Reducing prediction error by transforming input data for neural networks. Journal of Computing in Civil Engineering 14, 109-116.

Shihavuddin, A. S. M., Gracias, N., Garcia, R., Gleason, A. C. R., and Gintert, B. (2013). Image-based coral reef classification and thematic mapping. Remote Sensing 5, 1809-1841.

Shipp, C. A. and Kuncheva, L. I. (2002). Relationships between combination methods and measures of diversity in combining classifiers. Information Fusion 3, 135-148.

Simboura, N., Panayotidis, P., and Papathanassiou, E. (2005). A synthesis of the biological quality elements for the implementation of the European Water Framework Directive in the Mediterranean ecoregion: The case of Saronikos Gulf. Ecological Indicators 5, 253-266.

Smith, A., Sterba-Boatwright, B., and Mott, J. (2010). Novel application of a statistical technique, Random Forests, in a bacterial source tracking study. Water Research 44, 4067-4076.

Sokal, R. R. and Sneath, P. H. A. (1963). Principles of numerical taxonomy. San Francisco: WH Freeman.

Solomatine, D. P., Maskey M., and Durga L.S. (2006). Eager and Lazy Learning Methods in the Context of Hydrologic Forecasting. International Joint Conference on Neural Networks, Vancouver, Canada, 4847-4853.

Solomatine, D. P., Maskey, M., and Shrestha, D. L. (2008). Instance-based learning compared to other data-driven methods in hydrological forecasting. Hydrological Processes 22, 275-287.

Sondergaard, M., Jeppesen, E., Jensen, J. P., and Amsinck, S. L. (2005). Water Framework Directive: ecological classification of Danish lakes. Journal of Applied Ecology 42, 616-629.

Spatharis, S., Tsirtsis, G., Danielidis, D. B., Thang, D. C., and Mouillot, D. (2007a). Effects of pulsed nutrient inputs on phytoplankton assemblage structure and blooms in an enclosed coastal area. Estuarine, Coastal & Shelf Science, 73, 807-815.

Spatharis, S., Danielidis, D. B., and Tsirtsis, G. (2007b). Recurrent Pseudo-nitzschia calliantha (Bacillariophyceae) and Alexandrium insuetum (Dinophyceae) winter blooms induced by agricultural runoff. Harmful Algae 6, 811.

Spatharis, S., Mouillot, D., Danielidis, D. B., Karydis, M., Chi, T. D., and Tsirtsis, G. (2008). Influence of terrestrial runoff on phytoplankton species richness-biomass relationships: A double stress hypothesis. Journal of Experimental Marine Biology & Ecology 362, 55-62.

Spatharis, S., Roelke, D. L., Dimitrakopoulos, P. G., and Kokkoris, G. D. (2011). Analyzing the (mis) behavior of Shannon index in eutrophication studies using field and simulated phytoplankton assemblages. Ecological Indicators 11, 697-703.

Spatharis, S. and Tsirtsis, G. (2010). Ecological quality scales based on phytoplankton for the implementation of Water Framework Directive in the Eastern Mediterranean. Ecological Indicators 10, 840-847.

Spyropoulou, A., Spatharis, S., Papantoniou, G., and Tsirtsis, G. (2013). Potential response of a semi-arid ecosystem to climate change. Hydrobiologia, 705, 87-99.

Srdoc , A., Bratko, I., and Sluga, A. (2007). Machine learning applied to quality management-A study in ship repair domain. Computers in Industry 58, 464-473.

Srinivas, M., Supreethi, K. P., and Prasad, D. E. V. (2009). Combining the classifiers and LSI method for efficient and accurate text classification. International Journal of Information Technology & Knowledge Management 2, 263-267.

Stankovski, V., Debeljak, M., Bratko, I., and Adamic, M. (1998). Modelling the population dynamics of red deer (Cervus elaphus L.) with regard to forest development. Ecological Modelling 108, 145-153.

Stone, M. (1978). Cross-validation:a review 2. Series Statistics 9, 127-139.

Straten, G. v. (1992). The predicting power of models for eutrophication. In D. W. Sutcliffe and J. G. Jones (Eds.), Eutrophication: research and application to water supply (pp. 44-48). Ampleside, Cumbria, UK: Freshwater Biological Association.

Tamvakis, A., Miritzis, J., Tsirtsis, G., Spyropoulou, A., and Spatharis, S. (2012). Effects of meteorological forcing on coastal eutrophication: Modeling with model trees. Estuarine, Coastal & Shelf Science 115, 210-217.

Tamvakis, A., Trygonis, V., Miritzis, J., Tsirtsis, G., and Spatharis, S. (2014). Optimizing biodiversity prediction from abiotic parameters. Environmental Modelling & Software 53, 112-120.

Tan, A. C. and Gilbert, D. (2003) Ensemble machine learning on gene expression data for cancer classification. Applied bioinformatics 2, S75-83.

Tan, A. C., Gilbert, D., and Deville, Y. (2003) Multi-class protein fold classification using a new ensemble machine learning approach. Genome informatics.International Conference on Genome Informatics 14, 206-217.

Taylor, P. J., Donald, I. J., Jacques, K., and Conchie, S. M. (2012). Jaccard's heel: Radex models of criminal behaviour are rarely falsifiable when derived using Jaccard coefficient. Legal & Criminological Psychology 17, 41-58.

Thrush, S. F., Hewitt, J. E., Funnell, G. A., Cummings, V. J., Ellis, J., Schultz, D., Talley, D., and Norkko, A. (2001). Fishing disturbance and marine biodiversity: the role of habitat structure in simple soft-sediment systems. Marine Ecology Progress Series 233, 277-286.

Tian, X., Ju, M., Shao, C., and Fang, Z. (2011). Developing a new grey dynamic modeling system for evaluation of biology and pollution indicators of the marine environment in coastal areas. Ocean & Coastal Management 54, 750-759.

Tirelli, T., and Pessani, D. (2011). Importance of feature selection in decision-tree and artificial-neural-network ecological applications. Alburnus alburnus alborella: A practical example. Ecological Informatics 6, 309-315.

Tison, J., Park, Y. S., Coste, M., Wasson, J. G., Rimet, F., Ector, L., and Delmas, F. (2007). Predicting diatom reference communities at the French hydrosystem scale: A first step towards the definition of the good ecological status. Ecological Modelling 203, 99-108.

Topouzelis, K., Karathanassi, V., Pavlakis, P., and Rokos, D. (2008). Dark formation detection using neural networks. International Journal of Remote Sensing 29, 4705-4720.

Topouzelis, K. and Psyllos, A. (2012). Oil spill feature selection and classification using decision tree forest on SAR image data. ISPRS Journal of Photogrammetry & Remote Sensing 68, 135-143.

Tsai, C. F. (2014). Combining cluster analysis with classifier ensembles to predict financial distress. Information Fusion 16, 46-58.

Tsekouras G. E. (2005). Fuzzy modeling based on ordinary fuzzy partitions and nearest neighbor clustering. Journal of Intelligent & Robotic Systems 43, 255-282.

Tsekouras G. E. and Tsimikas J. (2013). On training RBF neural networks using input-output fuzzy clustering and particle swarm optimization. Fuzzy Sets & Systems 221, 65-89.

Tsirtsis, G., Spatharis, S., and Karydis, M. (2008). Application of the lognormal equation to assess phytoplankton community structural changes induced by marine eutrophication. Hydrobiologia 605, 89-98.

Tsymbal, A., Puuronen, S., and Patterson, D. W. (2003). Ensemble feature selection with the simple Bayesian classification. Information Fusion 4, 87-100.

Utermohl, H. (1958). Zur Vervollkommnung der quantitativen Phytoplankton-Methodik. Mitt Internationale Ver Theoretische und Angewandte Limnologie 9, 1-38.

Uygun, K., Tolboom, H., Izamis, M. L., Uygun, B., Sharma, N., Yagi, H., Soto-Gutierrez, A., Hertl, M., Berthiaume, F., and Yarmush, M. L. (2010). Diluted blood reperfusion as a model for transplantation of ischemic rat livers: alanine aminotransferase is a direct indicator of viability. Transplantation Proceedings 42, 2463-2467.

Verikas, A., Gelzinis, A., and Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. Pattern Recognition 44, 330-349.

Volf, G., Atanasova, N., Kompare, B., Precali, R., and Ozanic, N. (2011). Descriptive and prediction models of phytoplankton in the northern Adriatic. Ecological Modelling 222, 2502-2511.

Vollenweider, R. A. (1974). A manual on methods for measuring primary production in aquatic environments. Oxford, UK : Blackwell Scientific Publishers.

Vounatsou, P. and Karydis, M. (1991). Environmental characteristics in oligotrophic waters: Data evaluation and statistical limitations in water quality studies. Environmental Monitoring & Assessment 18, 211-220.

Wang, Y. and Witten, I. H. (1997). Induction of model trees for predicting continuous classes. Proceedings of the poster papers of the European Conference on Machine Learning. University of Economics, Faculty of Informatics and Statistics, Prague.

Wettschereck, D., Aha, D. W., and Mohri, T. (1997). A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artificial Intelligence Review 273-314.

WFD, 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 – establishing framework for community action in the field of water policy. Official Journal of the European Communities L327, 1-71.

Witten, I. H. and Frank, E. (2005). Data Mining, Practical Machine Learning Tools and Techniques. (2nd edition). Morgan Kaufmann.

Wonda, H. (1981). Similarity indices, sample size and diversity. Oecologia 50, 296-302.

Wozniak, M., Grana, M., and Corchado, E. (2014). A survey of multiple classifier systems as hybrid systems. Information Fusion 16, 3-17.

Wu, T., Luo, L., Qin, B., Cui, G., Yu, Z., and Yao, Z. (2009). A vertically integrated eutrophication model and its application to a river-style reservoir-Fuchunjiang, China. Journal of Environmental Sciences 21, 319-327.

Xu, L., Li, J., and Brenning, A. (2014). A comparative study of different classification techniques for marine oil spill identification using RADARSAT-1 imagery. Remote Sensing of Environment 141, 14-23.

Yang, Q., Shao, J., Scholz, M., and Plant, C. (2011). Feature selection methods for characterizing and classifying adaptive Sustainable Flood Retention Basins. Water Research 45, 993-1004.

Yin, Y. and Yasuda, K. (2005). Similarity coefficient methods applied to the cell formation problem: a comparative investigation. Computers & Industrial Engineering 48, 471-489.

Yucel, K. T. and Ozel, C. (2012). Modeling of mechanical properties and bond relationship using data mining process. Advances in Engineering Software 45, 54-60.

Yule, G. U. (1900). On the association of attributes in statistics. Philosophy of Transactions A 194, 257-319.

Yule, G. U. (1912). On the methods of measuring the association between two variables. Journal of the Royal Statistical Society 75, 579-642.

Zamani, A., Solomatine, D., Azimian, A., and Heemink, A. (2008). Learning from data for wind-wave forecasting. Ocean Engineering 35, 953-962.

Zar, J. H. (1984). Biostatistical Analysis. London: Prentice-Hall International.

Zarauz, L., Irigoien, X., and Fernandes, J. A. (2009). Changes in plankton size structure and composition, during the generation of a phytoplankton bloom, in the central Cantabrian Sea. Journal of Plankton Research 31, 193-207.

Zhan, H., Shi, P., and Chen, C. (2003). Retrieval of oceanic chloropyll concentation using support vector machines. IEEE Transactions on Geoscience & Remote Sensing 41, 2947-2951.

10 PUBLISHED WORK

Tamvakis, A., Miritzis, J., Tsirtsis, G., Spyropoulou, A., and Spatharis, S. (2012). Effects of meteorological forcing on coastal eutrophication: Modeling with model trees. Estuarine, Coastal & Shelf Science 115, 210-217.

Tamvakis, A., Trygonis, V., Miritzis, J., Tsirtsis, G., and Spatharis, S. (2014). Optimizing biodiversity prediction from abiotic parameters. Environmental Modelling & Software 53, 112-120.

Tamvakis, A., Anagnostopoulos, C. N., Tsirtsis, G., and Spatharis, S. (2014). Optimizing classification tasks with a new index for combining machine learning algorithms. Submitted.

Ταμβάκη, Α., Σπαθάρη, Σ., Μυριτζής, Ι., και Τσιρτσής, Γ. (2012). Είναι δυνατή η αποτελεσματική πρόβλεψη της βιοποικιλότητας από αβιοτικές παραμέτρους; Πρακτικά 6ου Πανελλήνιου Συνεδρίου Οικολογίας, 152.