ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΑΓΡΟΝΟΜΩΝ ΤΟΠΟΓΡΑΦΩΝ ΜΗΧΑΝΙΚΩΝ – ΜΗΧΑΝΙΚΩΝ ΓΕΩΠΛΗΡΟΦΟΡΙΚΗΣ
ΤΟΜΕΑΣ ΤΟΠΟΓΡΑΦΙΑΣ – ΕΡΓΑΣΤΗΡΙΟ ΤΗΛΕΠΙΣΚΟΠΗΣΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μέθοδοι υπερδειγματοληψίας και υποδειγματοληψίας για την αύξηση της ακρίβειας ταξινομήσεων

Αναστασοπούλου Νικολέττα

Επιβλέπων Καθηγητής : Καράντζαλος Κωνσταντίνος

Αθήνα, Ιούλιος 2024

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΑΓΡΟΝΟΜΩΝ ΤΟΠΟΓΡΑΦΩΝ ΜΗΧΑΝΙΚΩΝ – ΜΗΧΑΝΙΚΩΝ ΓΕΩΠΛΗΡΟΦΟΡΙΚΗΣ
ΤΟΜΕΑΣ ΤΟΠΟΓΡΑΦΙΑΣ – ΕΡΓΑΣΤΗΡΙΟ ΤΗΛΕΠΙΣΚΟΠΗΣΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μέθοδοι υπερδειγματοληψίας και υποδειγματοληψίας
για την αύξηση της ακρίβειας ταξινομήσεων

Επιβλέπων Καθηγητής : Καράντζαλος Κωνσταντίνος

Τριμελής Εξεταστική Επιτροπή :

……..	……..	……..
Κ. Καράντζαλος	**Ι. Παπουτσής**	**Β. Ανδρώνης**

Αθήνα, Ιούλιος 2024

……

Αναστασοπούλου Νικολέττα

Διπλωματούχος Αγρονόμος Τοπογράφος Μηχανικός – Μηχανικός Γεωπληροφορικής Ε.Μ.Π.

Πρόλογος κι Ευχαριστίες

Ένα από τα σημαντικά ζητήματα που απασχολούν την Τηλεπισκόπηση αφορά την ταξινόμηση των πολυφασματικών απεικονίσεων και πως θα αυξηθεί η ακρίβεια του αποτελέσματος. Η εύρεση του καταλληλότερου αλγορίθμου ταξινόμησης και του συνόλου δεδομένων εκπαίδευσης για την ταξινόμηση απεικονίσεων αποτελεί ακόμα και σήμερα πρόκληση, δίνοντας την ευκαιρία να αναπτυχθεί πληθώρα αλγορίθμων ταξινόμησης, αλλά και μεθόδων διαχείρισης συνόλων δεδομένων, όπως υπερδειγματοληψία κι υποδειγματοληψία, επειδή τείνουν τα σύνολα δεδομένων να μη χαρακτηρίζονται ισορροπημένα, με απώτερο σκοπό το τελικό προϊόν να καταγράφει την καλύτερη δυνατή απόδοση.

Η παρούσα διπλωματική εργασία εκπονήθηκε στο Εργαστήριο Τηλεπισκόπησης του Εθνικού Μετσόβιου Πολυτεχνείου κι έχει ως στόχο τη διερεύνηση του καταλληλότερου συνόλου δεδομένων εκπαίδευσης, για την επίτευξη υψηλών αποδόσεων στις μετρικές αξιολόγησης των ταξινομήσεων στην περιοχή της νοτιοδυτικής Πελοποννήσου, εστιάζοντας στις Περιφερειακές Ενότητες Μεσσηνίας (Δήμος Καλαμάτας, Δήμος Μεσσήνης, Δήμος Οιχαλίας) κι Αρκαδίας (Δήμος Μεγαλόπολης).

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή της παρούσας διπλωματικής εργασίας κ. Καράντζαλο Κωνσταντίνο, τον κ. Παπουτσή Ιωάννη και τον κ. Ανδρώνη Βασίλειο για την εμπιστοσύνη τους να μου αναθέσουν το συγκεκριμένο θέμα, δίνοντάς μου τις κατευθυντήριες γραμμές για το επιστημονικό περιεχόμενο της εργασίας, υποδεικνύοντας παραλήψεις ,καθώς και για την τεχνική κι επιστημονική βοήθεια που προσέφεραν όποια στιγμή κι αν χρειάστηκε.

Εν συνεχεία, οφείλω να ευχαριστήσω την οικογένεια μου, η οποία δεν έλειψε στιγμή από το πλευρό μου, καθ' όλη τη διάρκεια της ακαδημαϊκής μου πορείας και με ώθησε με κάθε τρόπο να ξεπεράσω όποια δυσκολία μου παρουσιάστηκε, με σκοπό να επιτύχω τους στόχους μου.

Ακόμα, θα ήθελα να ευχαριστήσω όλους τους φίλους και τις φίλες μου που βρέθηκαν κοντά μου κατά τη διάρκεια αυτού του ταξιδιού, με τον καθένα ξεχωριστά και με το δικό του τρόπο μου έδωσαν την απαραίτητη δύναμη για να φτάσω ως το τέλος αυτής της πενταετούς πορείας.

Εν κατακλείδι, ένα μεγάλο ευχαριστώ οφείλω να αποδώσω στην οικογένεια Τσουκαλά κι ιδιαίτερα στον Φώτη και τον Ανδρέα, οι οποίοι δίνοντας μου την ευκαιρία να κάνω την πρακτική μου άσκηση στο πλάι τους, μαζί με τα υπόλοιπα μέλη της ομάδας τους, μου άνοιξαν την πόρτα για το μαγικό κόσμο των Μηχανικών, στηρίζοντας με με κάθε τρόπο από τότε, μέχρι σήμερα, βλέποντας στο πρόσωπό μου ένα μελλοντικό επιστήμονα και συνεργάτη.

Περιεχόμενα

Ευρετήριο Εικόνων

Ευρετήριο Πινάκων

Ευρετήριο Διαγραμμάτων

Ευρετήριο Σχημάτων

Περίληψη

Στην παρούσα διπλωματική εργασία αναπτύσσονται, εφαρμόζονται κι αξιολογούνται πειράματα επιβλεπόμενης ταξινόμησης σε μη ισορροπημένα σύνολα δεδομένων, με τη χρήση μεθόδων τυχαίας υπερδειγματοληψίας και τυχαίας υποδειγματοληψίας, για την αξιολόγηση της ακρίβειας που επιτυγχάνεται σε κάθε μια από τις δυο μεθόδους, για κάθε αλγόριθμο ταξινόμησης. Για τις ανάγκες των πειραματικών εφαρμογών αξιοποιείται μια Sentinel-2A πολυφασματική απεικόνιση, με τα 10 από τα 12 διαθέσιμα κανάλια, ατμοσφαιρικά διορθωμένη (προϊόν Level 2A). Εκτελούνται διαδοχικά, ένα σύνολο από πειράματα με το πρώτο να χρησιμοποιεί ισορροπημένο σύνολο δεδομένων, που συνιστά τον βασικό πυλώνα και στη συνέχεια, ακολουθούν πειράματα με μη ισορροπημένα σύνολα δεδομένων εκπαίδευσης, τα οποία αντιμετωπίζονται είτε με υπερδειγματοληψία είτε με υποδειγματοληψία, ανάλογα με το τι υποδεικνύει κάθε φορά ο δείκτης αναλογίας ισορροπίας, που δημιουργείται στα πλαίσια διεξαγωγής της εργασίας. Ο δείκτης αναλογίας ισορροπίας σχετίζεται με το πλήθος των πολυγώνων εκπαίδευσης των κατηγοριών χρήσης/κάλυψης γης. Η πρώτη εφαρμογή αναπτύσσεται με στόχο τη διερεύνηση κι αξιολόγηση της απόδοσης των αλγορίθμων επιβλεπόμενης ταξινόμησης σε ένα ισορροπημένο σύνολο δεδομένων, επιδιώκοντας υψηλή ακρίβεια. Ειδικότερα, αξιοποιώντας ένα ισορροπημένο, ως προς την έκταση των κατηγοριών χρήσης/κάλυψης γης, σύνολο δεδομένων εκπαίδευσης εκτελείται ταξινόμηση και συγκρίνονται ποιοτικά οι θεματικοί χάρτες των ταξινομήσεων με το Corine Land Cover (CLC) και τον κυρωμένο δασικό χάρτη της περιοχής μελέτης, ενώ σε δεύτερο χρόνο γίνεται η ποσοτική αξιολόγηση με τον υπολογισμό των μετρικών precision, recall και F1-score ,που προκύπτουν από τον πίνακα σύγχυσης. Σε επόμενο βήμα, υπολογίζοντας για κάθε κατηγορία το δείκτη αναλογίας ισορροπίας, εφαρμόζεται υπερδειγματοληψία ή υποδειγματοληψία, εκτελείται εκ νέου μια σειρά από εφαρμογές επιβλεπόμενης ταξινόμησης κι αξιολογείται η επίδραση που έχει σε κάθε περίπτωση η μέθοδος διαχείρισης των μη ισορροπημένων δεδομένων, στην ακρίβεια της ταξινόμησης. Η εργασία ολοκληρώνεται με τα τελικά συμπεράσματα κι ορισμένες προτάσεις για πιθανά θέματα μελλοντικής έρευνας.

Abstract

In this thesis, supervised classification experiments on unbalanced datasets are developed, implemented and evaluated using random oversampling and random undersampling methods to evaluate the accuracy achieved in each of the two methods for each classification algorithm. For the needs of the experimental applications, a Sentinel-2A multispectral imagery is utilized, with 10 of the 12 available channels, atmospherically corrected (Level 2A product). A set of experiments are performed sequentially, with the first one using a balanced dataset, which constitutes the main pillar, followed by experiments with unbalanced training datasets, which are treated either by oversampling or undersampling, depending on what is indicated each time by the balance ratio index which is made for this thesis needs. The Balance Ratio Indice (BRI) is related to the number of training polygons of the land use/land cover categories. The first application is developed to investigate and evaluate the performance of supervised classification algorithms on a balanced dataset, aiming for high accuracy. In particular, utilizing a balanced, in terms of land use/land cover categories, training dataset, classification is performed and the thematic maps of the classifications are qualitatively compared with the Corine Land Cover (CLC) and the validated forest map of the study area, while in a second step the quantitative evaluation is performed by calculating the precision, recall and F1-score, metrics derived from the confusion matrix. In the next step, by calculating for each category the Balance Ratio Indice, oversampling or undersampling is applied, a series of supervised classification applications are re-run and the effect of the method of handling unbalanced data on the accuracy of the classification is evaluated in each case. The paper concludes with the final conclusions and some suggestions for possible future research topics.

Λέξεις κλειδιά

Τηλεπισκόπηση, Δορυφόρος, Sentinel-2, Αισθητήρες, Πολυφασματική απεικόνιση, Ταξινόμηση, Μηχανική Μάθηση, Επιβλεπόμενη Ταξινόμηση, Αλγόριθμος Ταξινόμησης, Ακρίβεια Ταξινόμησης, Δεδομένα Εκπαίδευσης, Δεδομένα Ελέγχου, Corine Land Cover (CLC), Κυρωμένος Δασικός Χάρτης, Φασματική Υπογραφή, Υπερδειγματοληψία, Υποδειγματοληψία, Random Forest, Support Vector Machine, Μέθοδοι Υπερδειγματοληψίας, Μέθοδοι Υποδειγματοληψίας, Δείκτης Αναλογίας Ισορροπίας

Keywords

Remote sensing, Satellite, Sentinel-2, Sensors, Multispectral image, Classification, Machine Learning, Supervised Classification, Classification Algorithm, Classification Accuracy, Training dataset, Test dataset, Corine Land Cover (CLC), Authorized Forest Map, Spectral Signature, Oversampling, Undersampling, Random Forest, Support Vector Machine, Oversampling methods, Undersampling methods, Balance Ratio Indice (BRI)

ΚΕΦΑΛΑΙΟ 1. Ανασκόπηση Βιβλιογραφίας

Σε αυτό το κεφάλαιο γίνεται μια καταγραφή της υπάρχουσας βιβλιογραφίας που σχετίζεται με πειραματικές εφαρμογές που αφορούν την υπερδειγματοληψία και την υποδειγματοληψία δεδομένων, αλλά και μεθόδους αυτών, που αναπτύσσονται κι εφαρμόζονται στην παρούσα διπλωματική εργασία. Η βιβλιογραφία που παρατίθεται αφορά εφαρμογές που εστιάζουν στα μη ισορροπημένα σύνολα δεδομένων, πως αυτά επηρεάζουν την ακρίβεια της ταξινόμησης, με ποιες μεθόδους μπορούν να αντιμετωπιστούν και πως επηρεάζει τη διαδικασία της επαναδειγματοληψίας η μέθοδος της υπερδειγματοληψίας ή/κι υποδειγματοληψίας.

1.1.Εφαρμογές μη ισορροπημένων δεδομένων και μεθόδων υπερδειγματοληψίας κι υποδειγματοληψίας

Το πρόβλημα των μη ισορροπημένων δεδομένων ταλανίζει τους ερευνητές, λόγω του γεγονότος πως τα μη ισορροπημένα δεδομένα επηρεάζουν τα μοντέλα πρόβλεψης και την απόδοσής τους. Ένας από τους τομείς που επηρεάζουν τα μη ισορροπημένα σύνολα δεδομένων είναι ο υπολογισμός μετρικών όπως η ακρίβεια. Αναφορά γίνεται στο βιβλίο «Machine Learning for Imbalanced Data - Tackle imbalanced datasets using machine learning and deep learning techniques», Kumar Abhishek & Dr. Mounir Abdelaziz, 2023, στις εξής δυο μετρικές καμπυλών, Precision-Recall (PR) και ROC, οι οποίες χαρακτηρίζονται αμετάβλητες από την ύπαρξη κατωφλιών, που τείνουν να εκθέσουν την απόδοση του μοντέλου ταξινόμησης σε ένα ευρύ φάσμα κατωφλιών. Παρόλα αυτά, η πραγματική πρόκληση έγκειται στη δυσανάλογη επιρροή των “true negative” στοιχείων του πίνακα σύγχυσης. Μετρικές όπως precision, recall και F1-score (παράγωγο μέγεθος των δυο προηγούμενων) κρίνονται ως καταλληλότερες για την αξιολόγηση της απόδοσης του μοντέλου, αφού εστιάζουν λιγότερο στα “true negative” στοιχεία. Θα ήταν παράλειψη να μην αναφερθεί πως αυτές οι μετρικές “κρύβουν” μια υπερ-παράμετρο, το κατώφλι του αλγορίθμου ταξινόμησης, οδηγώντας με αυτόν τον τρόπο σε καλύτερη προσέγγιση των ρεαλιστικών εφαρμογών. Εκτός των άλλων, ένα σύνολο μη ισορροπημένων δεδομένων δύναται να αποτελέσει πρόκληση για τη συνάρτηση απώλειας, η οποία έχει ως σκοπό την ελαχιστοποίηση των σφαλμάτων μεταξύ των προβλεπόμενων αποτελεσμάτων και των ορθά ταξινομημένων στοιχείων, σύμφωνα με τα δεδομένα εκπαίδευσης. Τονίζεται δε, πως σε μη ισορροπημένο σύνολο δεδομένων, μια κλάση αντιπροσωπεύεται από περισσότερα δείγματα ,με αποτέλεσμα το μοντέλο εκπαίδευσης να είναι προκατειλημμένο ως προς την επικρατέστερη κλάση. Επισημαίνεται ,παράλληλα, πως η εσφαλμένη ταξινόμηση μπορεί έχει διαφορετικά αποτελέσματα για την πλειοψηφική και τη μειονοτική τάξη, δυσχεραίνοντας με αυτόν τον τρόπο τη διαχείριση ενός μη ισορροπημένου συνόλου δεδομένων. Ταυτόχρονα, τίθεται το ζήτημα της υπολογιστικής ισχύς που απαιτείται για τη διαχείριση μεγάλων συνόλων δεδομένων, όπως συνηθίζεται σε εφαρμογές που περιγράφουν τον πραγματικό κόσμο, όπου οι μέθοδοι αντιμετώπισης μη ισορροπημένων συνόλων δεδομένων αυξάνουν τις υπολογιστικές ανάγκες, όπως για παράδειγμα σε αποθηκευτικό χώρο. Φυσικά, γίνεται αντιληπτό πως σε ένα μη ισορροπημένο dataset μεγάλο πρόβλημα συνιστά η μη επαρκής ποικιλομορφία δειγμάτων της μειονοτικής κλάσης που να αντιπροσωπεύουν επαρκώς την κατανομή τους. Πιο συγκεκριμένα, το πρόβλημα δεν είναι το πλήθος των δειγμάτων, όσο το γεγονός πως τα δείγματα της μειονοτικής κλάσης πρέπει να είναι κατάλληλα κατανεμημένα στο χώρο τιμών της κλάσης. Ακόμη και σε περιπτώσεις που το σύνολο δεδομένων φαίνεται μεγάλο, η ποικιλομορφία των δειγμάτων ενδέχεται να μην είναι επαρκής, με αποτέλεσμα το μοντέλο επίβλεψης να μην εκπαιδευτεί ορθά για τα όρια της κατηγορίας ,σημειώνοντας εν τέλει χαμηλή απόδοση. Επίσης, η έρευνα του βιβλίου

αναλύει και την περίπτωση ενός μη βαθμονομημένου μοντέλου πρόβλεψης, το οποίο σε μια εφαρμογή μη ισορροπημένου συνόλου δεδομένων αναμένεται να μην αποφέρει ρεαλιστικά αποτελέσματα. Τέλος, επισημαίνεται η χαμηλή απόδοση των μοντέλων, λόγω μη προσαρμοσμένων ορίων, ενισχύοντας με αυτό το συμπέρασμα την αντίληψη πως πρέπει να χρησιμοποιούνται μοντέλα εκπαίδευσης που έχουν χρονιστεί μέσα από μη ισορροπημένα σύνολα δεδομένων, κάνοντας χρήση “έξυπνων” κατωφλιών, με την προσαρμογή κατωφλιού να κρίνεται σημαντική και σε περιπτώσεις αξιοποίησης ισορροπημένων datasets.

Στη βιβλιογραφία, το πρόβλημα των μη ισορροπημένων δεδομένων εκπαίδευσης έχει αντιμετωπιστεί με την εφαρμογή των μεθόδων υπερδειγματοληψίας κι υποδειγματοληψίας. Για το λόγο αυτό αναπτύχθηκε η μελέτη «A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining» των Tarid Wongvorachan , Surina He και Okan Bulut το 2023, όπου συγκρίνονται διάφορες τεχνικές δειγματοληψίας, τρεις στον αριθμό, για τη διαχείριση διαφορετικών αναλογιών του προβλήματος ανισορροπίας των τάξεων σε ένα σύνολο δεδομένων εκπαίδευσης από το High School Longitudinal Study of 2009, χωρίζοντας δυο κατηγορίες, τις μέτριες και τις εξαιρετικά μη ισορροπημένες ταξινομήσεις, εφαρμόζοντας τον αλγόριθμο Random Forest (RF). Αναλυτικότερα, οι τεχνικές δειγματοληψίας που εφαρμόστηκαν είναι η τυχαία υπερδειγματοληψία (Random Oversampling, ROS), τυχαία υποδειγματοληψία (Random undersampling, RUS) και ο συνδυασμός Synthetic Minority Oversampling Technique for Nominal and Continuous (SMOTE-NC) με την τυχαία υποδειγματοληψία (RUS) ως υβριδική μέθοδο δειγματοληψίας. Η μελέτη διακρίνει δυο τυπικές περιπτώσεις ταξικής ανισορροπίας στην εκπαίδευση, οι οποίες είναι η μετα-δευτεροβάθμια εγγραφή (μέτρια ανισόρροπη τάξη) και η εγκατάλειψη του σχολείου (εξαιρετικά ανισόρροπη τάξη), εστιάζοντας στη σύγκριση τεχνικών προσέγγισης σε επίπεδο δεδομένων, καθώς μια από τις κύριες εφαρμογές της Educational Data Mining (EDM) είναι η αξιοποίηση της δύναμης των υπολογιστικών αλγορίθμων για την εξαγωγή πληροφοριών από σύνολα δεδομένων εκπαίδευσης. Η προσέγγιση σε επίπεδο δεδομένων για τη μάθηση ανισορροπίας στοχεύει στη δημιουργία ενός νέου συνόλου δεδομένων με ίση αναλογία μεταξύ των κατηγοριών της μεταβλητής για τη μείωση της “προκατάληψης” του αλγορίθμου.

Όπως προκύπτει από τα αποτελέσματα της μελέτης, στην περίπτωση μέτριας ανισορροπίας, η RUS είχε χειρότερη απόδοση σε σύγκριση με την αρχική κατάσταση μη επαναδειγματοληψίας δεδομένων. Από την άλλη πλευρά, η υβριδική μέθοδος επαναδειγματοληψίας και η ROS ήταν ικανές να βελτιώσουν την απόδοση του ταξινομητή Random Forest, με τη ROS να αποδίδει καλύτερα. Αυτό το εύρημα υποδηλώνει ότι όταν η μεταβλητή/στόχος είναι μέτρια μη ισορροπημένη, η υπερδειγματοληψία θα μπορούσε να είναι πιο ωφέλιμη για την ταξινόμηση από την υποδειγματοληψία. Στην εξαιρετικά μη ισορροπημένη περίπτωση, η ROS παρουσίασε ένα πρόβλημα υπερπροσαρμογής, το οποίο πιθανόν να οφείλεται στην επαναλαμβανόμενη αναπαραγωγή της μειονοτικής τάξης. Συνολικά, η προσέγγιση της υβριδικής επαναδειγματοληψίας φαίνεται να λειτουργεί καλά σε μια εξαιρετικά ανισόρροπη κατάσταση, ενώ η ROS δε συνιστάται λόγω της ευπάθειας της στην υπερπροσαρμογή. Σε περιπτώσεις μέτριας και εξαιρετικής ανισορροπίας, η RUS δεν συνιστάται ως η πρώτη επιλογή, καθώς οδηγεί στην απώλεια δυνητικά χρήσιμων δεδομένων για τον ταξινομητή.

Το πρόβλημα της ανισορροπίας των δεδομένων και την εφαρμογή των τεχνικών της υπερδειγματοληψίας και της υποδειγματοληψίας έχουν μελετήσει και οι Julio

Hernandez, Jesus Ariel Carrasco-Ochoa και Jose Francisco Martınez-Trinidad, το 2013, στην έρευνα «An Empirical Study of Oversampling and Undersampling for Instance Selection Methods on Imbalance Datasets». Σε αυτή την έρευνα παρουσιάζεται μια εμπειρική μελέτη σχετικά με τη χρήση μεθόδων υπερδειγματοληψίας κι υποδειγματοληψίας για βελτίωση της ακρίβειας των μεθόδων επιλογής δειγμάτων σε μη ισορροπημένα σύνολα δεδομένων εκπαίδευσης. Για τη διεξαγωγή των πειραμάτων αξιοποιούνται 18 databases από το αποθετήριο KEEL. Τα σύνολα δεδομένων ταξινομήθηκαν σε αύξουσα σειρά σύμφωνα με το λόγο ανισορροπίας τους (Imbalance Ratio, IR) που υπολογίζεται ως ο λόγος μεταξύ του μεγέθους των τάξεων της πλειοψηφίας και της μειοψηφίας. Ο λόγος ανισορροπίας είναι πολύ διαφορετικός για κάθε σύνολο δεδομένων, για αυτόν τον λόγο ομαδοποιούνται τα δεδομένα ως εξής: IR:1-3, IR:3-9 και IR>9. Για κάθε σετ δεδομένων, πραγματοποιήθηκε σε 10 δοκιμές ο υπολογισμός της διασταυρούμενης επικύρωσης (cross validation) λαμβάνοντας υπόψη το μέσο όρο της ακρίβειας της ταξινόμησης για τις κατηγορίες μειοψηφίας και πλειοψηφίας ,ξεχωριστά, καθώς και τη συνολική ακρίβεια, συμπεριλαμβάνοντας το μέγεθος F-Measure.

Αναφορά για το λόγο ανισορροπίας εντοπίζεται και στο βιβλίο «Dealing with Imbalanced and Weakly Labelled Data in Machine Learning using Fuzzy and Rough Set Methods, Sarah Vluymans ,2019» περιγράφοντας τη λειτουργία του και τη συμπεριφορά του στο πεδίο τιμών. Ως δείκτης ανισορροπίας (imbalance ratio, IR) ορίζεται ο λόγος των μεγεθών της συνολικής πλειοψηφικής τάξης με τη μειοψηφική τάξη, αναμένοντας το αποτέλεσμα του λόγου να είναι μια τιμή μεγαλύτερη ή ίση της μονάδας. Φυσικά, η περίπτωση που ο δείκτης λαμβάνει την τιμή 1, ισοδυναμεί με την περίπτωση ενός απόλυτα ισορροπημένου συνόλου δεδομένων, που κατά συνέπεια δεν επιδέχεται επεξεργασία. Όσο μεγαλύτερες τιμές δέχεται ο λόγος μεταξύ πλειοψηφικής και μειονοτικής τάξης, τόσο μεγαλύτερη και η διαφορά στα μεγέθη των τάξεων. Κάθε σύνολο δεδομένων με λόγο ανισορροπίας άνω της τιμής 1.5 θεωρείται μη ισορροπημένο, ενώ σε μελέτες η τιμή IR=9 σηματοδοτεί το άνω όριο, που χαρακτηρίζει τα σύνολα δεδομένων εξαιρετικά μη ισορροπημένα. Αποδεικνύεται επίσης, πως ο δείκτης IR δεν είναι το μόνο χαρακτηριστικό ενός μη ισορροπημένου συνόλου δεδομένων που θέτει προκλήσεις στους αλγορίθμους ταξινόμησης. Αρκετοί συγγραφείς περιγράφουν πως τα γενικά ζητήματα που σχετίζονται με τα δεδομένα , όπως το μικρό μέγεθος δείγματος, η επικάλυψη των κατηγοριών, ο θόρυβος που εντοπίζεται μεταξύ των δεδομένων και το πρόβλημα μετατόπισης συνόλου δεδομένων δύνανται να έχουν πιο έντονη επίδραση όταν πρόκειται για ένα μη ισορροπημένο σύνολο δεδομένων. Ο δείκτης ανισορροπίας είναι ανεπαρκής ως ένα μόνο μέγεθος να προβλέψει την απόδοση μιας μεθόδου ταξινόμησης σε ένα σύνολο μη ισορροπημένων δεδομένων, καθώς μέρος η ικανότητα αναγνώρισης των κατηγοριών επηρεάζεται σημαντικά από την πολυπλοκότητα του συνόλου δεδομένων που περιγράφουν τις κατηγορίες.

Ο δείκτης ανισορροπίας IR αποτελεί πηγή έμπνευσης και το βήμα για την ανάπτυξη αντίστοιχου δείκτη που αρμόζει στα δεδομένα της περιοχής μελέτης της παρούσας διπλωματικής εργασίας, με τον δείκτη να παρουσιάζεται στην αντίστοιχη ενότητα της μεθοδολογίας.

Η κύρια συμβολή αυτής της εργασίας είναι μια εμπειρική μελέτη συνδυασμού τεχνικών υπερδειγματοληψίας και υποδειγματοληψίας με ορισμένες μεθόδους επιλογής περιπτώσεων με βάση τον κανόνα του πλησιέστερου γείτονα (Nearest Neighbor, NN), τους εξελικτικούς αλγόριθμους και τους αλγορίθμους κατάταξης. Τα αποτελέσματα δείχνουν ότι αυτός ο συνδυασμός βελτιώνει την ακρίβεια της κατηγορίας μειοψηφίας

σε σχέση με το αρχικό σύνολο δεδομένων. Για μη ισορροπημένες σύνολα δεδομένων με IR στο διάστημα 1-9, η καλύτερη επιλογή είναι να χρησιμοποιείται SMOTE & IRB, ενώ για σετ δεδομένων με IR μεγαλύτερο από 9 υπάρχουν δύο κύριοι συνδυασμοί: Resample & IRB, αποκτώντας υψηλή συνολική ακρίβεια, και Spread Subsample & IRB, που αποκτά υψηλή ακρίβεια για τη μειονοτική τάξη.

Το άρθρο «Evolutionary Undersampling for Classification with Imbalanced Datasets: Proposals and Taxonomy» των Salvador Garcia και Francisco Herrera αναπτύχθηκε το 2009, μελετώντας την τεχνική της υποδειγματοληψίας. Σε αυτή τη μελέτη παρατίθεται ένα σύνολο από μεθόδους εξελικτικής υπερδειγματοληψίας, λαμβάνοντας υπόψη τη φύση του προβλήματος ανισορροπίας δεδομένων, χρησιμοποιώντας διαφορετικές προσεγγίσεις για την επίτευξη ενός “συμβιβασμού” μεταξύ της ισορροπημένης κατανομής των δεδομένων και της απόδοσης της μεθόδου. Η μελέτη περιλαμβάνει μια ταξινόμηση των προσεγγίσεων και μια συνολική σύγκριση μεταξύ των εξεταζόμενων μοντέλων και των σύγχρονων μεθόδων υποδειγματοληψίας. Από την αλληλουχία πειραμάτων και συγκρίσεων μεταξύ μεθόδων και τα αποτελέσματα να έχουν αντιπαραβληθεί με τη χρήση μη παραμετρικών στατιστικών διαδικασιών, τα συμπεράσματα που προκύπτουν από αυτή τη μελέτη είναι πως :

1.Οι αλγόριθμοι επιλογής πρωτοτύπων δεν πρέπει να χρησιμοποιούνται για το χειρισμό μη ισορροπημένων δεδομένων, καθώς είναι επιρρεπείς στο να σημειώσουν υψηλή συνολική ακρίβεια, εξαλείφοντας δείγματα της μειονοτικής τάξης, θεωρώντας τα ως θόρυβο.

2.Κατά τη διάρκεια της εξελικτικής διαδικασίας υποδειγματοληψίας, η χρήση του μηχανισμού επιλογής δειγμάτων από την πλειοψηφική κατηγορία βοηθά στην απόκτηση ακριβέστερων υποσυνόλων.

3.Σύνολα δεδομένων με χαμηλό λόγο ανισορροπίας ενδέχεται να αντιμετωπίσουν τα μοντέλα evolutionary undersampling guided by classification measures (EUSCM), ιδίως με τη χρήση του μοντέλου με καθολικό μηχανισμό επιλογής κι αξιολόγησης μέσω του μέτρου geometric mean (GM).

4.Αν και πάνω από σύνολα δεδομένων με υψηλό λόγο ανισορροπίας, όλα τα μοντέλα EUS έχουν καλά αποτελέσματα, δίνεται έμφαση στα μοντέλα evolutionary balancing undersampling (EBUS) με ιδιαίτερο ενδιαφέρον σε αυτό που εκτελεί πλειοψηφική επιλογή χρησιμοποιώντας το μέτρο GM. Η υπεροχή αυτού του μοντέλου σε σχέση με τους αλγορίθμους υποδειγματοληψίας τελευταίας τεχνολογίας έχει αποδειχθεί εμπειρικά.

Συνεχίζεται η μελέτη της μεθόδου υποδειγματοληψίας με το άρθρο «The impact of Under-Sampling Techniques on Classification Accuracy in multi-class Imbalance Data», των Suwanto Sanjaya, Rahmad Abdillah και Iis Afrianty, που εκδόθηκε το 2022. Στη συγκεκριμένη μελέτη γίνεται χρήση του αλγορίθμου LVQ-3, διαθέτοντας το χαρακτηριστικό να σταθεροποιείται μόνος του, οδηγώντας σε καλύτερη απόδοση, ενώ ταυτόχρονα χρησιμοποιείται και η k-fold διασταυρούμενη επικύρωση, ώστε να καθοριστούν τα δεδομένα εκπαίδευσης κι ελέγχου.

Η μέθοδος της υποδειγματοληψίας εξαλείφει δεδομένα στην πλειοψηφική τάξη μέχρι να εξισορροπήσει τον όγκο των δεδομένων της μειονοτικής τάξης. Ως αποτέλεσμα, τα δεδομένα που δεν χρησιμοποιούνται στην τεχνική γίνονται “άχρηστα”, παρόλο που τα δεδομένα δεν είναι δυνατό να χαρακτηριστούν με αυτόν τον όρο, ειδικά όσον αφορά την ακρίβεια. Η συγκεκριμένη έρευνα θέτει αυτό το ζήτημα αποδεικνύοντας ότι τα δεδομένα που δεν χρησιμοποιούνται έχουν την ικανότητα να επηρεάσουν την ποιότητα

της ακρίβειας. Αναπτύσσεται με αυτόν τον τρόπο η μεθοδολογία εκμάθησης συνόλου κι επιλογή χαρακτηριστικών, λόγω βιβλιογραφίας που υποδεικνύει ότι το σύνολο μάθησης, οι τεχνικές και η επιλογή χαρακτηριστικών θα μπορούσαν να αποδώσουν ακόμα καλύτερη ακρίβεια, ειδικά στην κατηγορία που μελετάται.

Έχει ήδη αναφερθεί το πρόβλημα πως οι αλγόριθμοι ταξινόμησης έχουν σχεδιαστεί για ισορροπημένα σύνολα δεδομένων εκπαίδευσης, ενώ ένα ακόμη πρόβλημα που παρατηρείται στην περίπτωση των μη ισορροπημένων δεδομένων εκπαίδευσης είναι αυτό της αλληλοεπικάλυψης των κλάσεων, η οποία φαίνεται να έχει μεγάλο αντίκτυπο στην ακρίβεια της ταξινόμησης, ακόμα μεγαλύτερο κι από την ανισορροπία των τάξεων. Σε αυτό το πρόβλημα εστίασαν οι Pattaramon Vuttipittayamongkol, Eyad Elyan, Andrei Petrovski και Chrisina Jayne, το 2018, αναπτύσσοντάς το στο άρθρο «Overlap-Based Undersampling for Improving Imbalanced Data Classification». Πιο συγκεκριμένα, σε αυτή τη μελέτη προτείνεται μια νέα μέθοδος υποδειγματοληψίας που εξαλείφει τις αρνητικές περιπτώσεις (δεδομένα πλειοψηφικής κλάσης) από την επικαλυπτόμενη περιοχή κι ως εκ τούτου βελτιώνει την ορατότητα των μειονοτικών περιπτώσεων.

Με τη μέθοδο Overlap-Based Undersampling εισάγεται ένας αλγόριθμος ομαδοποίησης για τον προσδιορισμό των επικαλυπτόμενων περιπτώσεων. Σε επόμενο βήμα, με τη χρήση της μεθόδου, οι αρνητικά επικαλυπτόμενες περιοχές αφαιρούνται. Καθ' αυτόν τον τρόπο, πλέον βελτιώνεται η ορατότητα μεταξύ της μειονοτικής κατηγορίας και του αλγορίθμου εκπαίδευσης, προσδίδοντας καλύτερα αποτελέσματα στη διαδικασία της ταξινόμησης, χωρίς να απαιτείται η εκ νέου εξισορρόπηση των δεδομένων.

Στη συγκεκριμένη μελέτη εφαρμόζονται 3 διαδοχικά πειράματα με χρήση 36 συνόλων δεδομένων από τα αποθετήρια των UCI και KEEL, για την αξιολόγηση της προτεινόμενης μεθόδου. Αρχικά, τα σύνολα δεδομένων ταξινομήθηκαν μετά την εφαρμογή της μεθόδου OBU. Σε δεύτερο χρόνο, συγκρίθηκαν τα αποτελέσματα των μελετητών με τη baseline που απλώς ταξινομούσε τα σύνολα δεδομένων χρησιμοποιώντας τον αλγόριθμο Random Forest χωρίς υποδειγματοληψία. Στο τρίτο πείραμα, αναπαρήχθη μια από της τελευταίας τεχνολογίας μέθοδος, Clustering-based undersampling in class-imbalanced data, και συγκρίθηκε με την προτεινόμενη από τους ερευνητές τεχνική.

Με την αφαίρεση των αρνητικών περιπτώσεων από την επικαλυπτόμενη περιοχή, σημειώνεται μια σημαντική βελτίωση στην ακρίβειας της μειοψηφικής κατηγορίας με σχετικά μικρό συμβιβασμό του true negative rate (TNR), βελτιώνοντας με αυτόν τον τρόπο και την ευαισθησία της μεθόδου. Αυτή η τεχνική έχει αποδειχθεί ότι ενισχύει την ταξινόμηση γνωστών συνόλων δεδομένων και ξεπέρασε τις επιδόσεις της τελευταίας τεχνολογίας στις περισσότερες από τις αποδεδειγμένες περιπτώσεις. Αυτά τα αποτελέσματα μπορούν να αποδοθούν σε πολλά πλεονεκτήματα της προτεινόμενης μεθόδου σε σχέση με άλλες κοινές τεχνικές υποδειγματοληψίας. Αυτά περιλαμβάνουν: πρώτον, το ποσοστό της υποδειγματοληψίας που προκύπτει από τη μέθοδο OBU, είναι ανάλογο του βαθμού επικάλυψης και δεύτερον, η μέθοδος OBU είναι απίθανο να εξαλείψει περιπτώσεις εκτός της επικαλυπτόμενης περιοχής, γεγονός που ελαχιστοποιεί την απώλεια πληροφοριών.

ΚΕΦΑΛΑΙΟ 2. Μη Ισορροπημένα Δεδομένα

Ένα σύνηθες φαινόμενο κατά την επεξεργασία πολυφασματικών απεικονίσεων είναι η δημιουργία συνόλων δεδομένων, στα οποία δεν υπάρχει κανονική κατανομή μεταξύ των μελετώμενων κατηγοριών, δηλαδή υπάρχουν παρατηρήσεις κατηγοριών οι οποίες σημειώνουν μεγαλύτερη συχνότητα από άλλες. Συνέπεια του φαινομένου αυτού είναι το πρόβλημα της ανισορροπίας μεταξύ των κλάσεων.

Ένας από τους ερευνητές που έχουν μελετήσει το πρόβλημα ανισορροπίας των κλάσεων, ο Weiss (2004) παρουσίασε μια μελέτη επί του πεδίου της μάθησης από μη ισορροπημένα σύνολα δεδομένων ,ορίζοντας δυο τύπους σπανιότητας. Ο ένας τύπος σπανιότητας αφορά τις σπάνιες κατηγορίες, που αναφέρονται στη μεταβλητή κλάση ή αλλιώς στη μεταβλητή απόκριση, ενώ από την άλλη, ο δεύτερος τύπος σχετίζεται με τις σπάνιες περιπτώσεις που εστιάζουν στο μέγεθος του δείγματος. Να σημειωθεί πως μια σπάνια κατηγορία περιέχει σχετικά μικρότερο αριθμό παρατηρήσεων από τις άλλες κατηγορίες ,εν αντιθέσει με τη σπάνια περίπτωση που καταδεικνύει ένα μικρό υποσύνολο του χώρου των δεδομένων. Συνοψίζοντας και με πιο απλά λόγια, παρατηρούνται δυο περιπτώσεις στις οποίες σημειώνεται το πρόβλημα της ανισορροπίας των κλάσεων, είτε επειδή υπάρχει φυσική ανισορροπία μεταξύ των κατηγοριών, είτε λόγω σπανιότητας των περιπτώσεων/δειγμάτων.

2.1.Ορισμός προβλήματος ανισορροπίας δεδομένων (Class Imbalance Problem)

Το πρόβλημα ανισορροπίας στα σύνολα δεδομένων παρουσιάζεται στο στάδιο της ταξινόμησης, με τον αριθμό των παρατηρήσεων που ανήκουν σε μια κατηγορία να είναι κατά πολύ μικρότερος ή μεγαλύτερος από τον αριθμό των άλλων κατηγοριών. Το μεγαλύτερο ζήτημα που τίθεται στο πρόβλημα της ανισορροπίας μεταξύ των κατηγοριών είναι ότι η κατηγορία που αποτελεί τη μειοψηφία είναι και η κατηγορία ενδιαφέροντος. Παρόλα αυτά, η κλασική μοντελοποίηση των αλγορίθμων ταξινόμησης υποθέτει ότι τα δεδομένα ακολουθούν κανονική κατανομή, με αποτέλεσμα οι ταξινομητές να είναι μεροληπτικοί απέναντι στην κατηγορία πλειοψηφίας, αγνοώντας την κλάση που μειοψηφεί. Συμπεραίνεται λοιπόν, πως οι αλγόριθμοι ταξινόμησης έχουν την τάση να εστιάζουν στην κλάση πλειοψηφίας και να αγνοούν τη μειοψηφική κλάση, η οποία ως επί των πλείστων αντιπροσωπεύει την κλάση ενδιαφέροντος.

Διάφορες προσεγγίσεις έχουν προταθεί για την αντιμετώπιση του συγκεκριμένου προβλήματος, οι οποίες δύναται να κατηγοριοποιηθούν σε δυο ομάδες :

1. Εσωτερικές προσεγγίσεις (internal approaches)

Δημιουργία καινοτόμων αλγορίθμων ή αλλαγή υπαρχόντων, ούτως ώστε να ληφθεί υπόψιν το πρόβλημα ανισορροπίας μεταξύ των κατηγοριών.

2. Εξωτερικές προσεγγίσεις (external approaches)

Προ-επεξεργασία των δεδομένων, ώστε να μειωθεί η επίδραση που προκαλείται από την ανισορροπία μεταξύ των κλάσεων.

Το μειονέκτημα των εσωτερικών προσεγγίσεων είναι ότι αποτελούν μια αλγοριθμική προσέγγιση, ενώ οι εξωτερικές προσεγγίσεις είναι ανεξάρτητες από τον αλγόριθμο ταξινόμησης.

2.2.Χειρισμός Προβλήματος Μη Ισορροπημένων Δεδομένων

Σημειώνονται τρεις κατηγορίες ομαδοποίησης των προσεγγίσεων που χρησιμοποιούνται για την αντιμετώπιση του προβλήματος της ανισορροπίας μεταξύ των κλάσεων, οι οποίες είναι :

- Αλλαγή της κατανομής των κλάσεων, τροποποιώντας τα ίδια τα δεδομένα ,ώστε να εξισορροπηθεί η ασυμμετρία των συνόλων δεδομένων
- Προσαρμογή των ταξινομητών, προσαρμόζοντας βασικούς αλγορίθμους ταξινόμησης σε μη ισορροπημένα σύνολα δεδομένων αποδίδοντας ένα βάρος στις λανθασμένα ταξινομημένες περιπτώσεις
- Σύνολο μεθόδων μάθησης, χρησιμοποιώντας ένα συνδυασμό πολλαπλών ταξινομητών με πολλαπλά σύνολα δεδομένων

Το σφάλμα ταξινόμησης στην κλάση πλειοψηφίας κυριαρχεί του σφάλματος ταξινόμησης στην κλάση μειοψηφίας, με αποτέλεσμα αλγόριθμοι μηχανικής μάθησης να αδυνατούν να διαχειριστούν προβλήματα με μη ισορροπημένα δεδομένα ,χωρίς να ταξινομείται με ιδιαίτερη επιτυχία το σύνολο των δεδομένων.

Οι τεχνικές για τη διαχείριση του προβλήματος ανισορροπίας των δεδομένων μπορούν να κατηγοριοποιηθούν ως τεχνικές :

1)Στο επίπεδο των δεδομένων
2)Σε αλγοριθμικό επίπεδο
3)Στο επίπεδο του ευαίσθητου κόστους
4)Σε επίπεδο επιλογής χαρακτηριστικών ,και
5)Στο επίπεδο ενός συνόλου μεθόδων μάθησης

Η παρούσα εργασία εστιάζει στην προσέγγιση του επιπέδου δεδομένων, με την προσέγγιση αυτή να λειτουργεί σε ένα στάδιο προ-επεξεργασίας των δεδομένων, προσπαθώντας να εξισορροπήσει τις κατανομές μεταξύ των κλάσεων.

Η προσέγγιση αλλαγής της κατανομής των κλάσεων στο επίπεδο των δεδομένων, προκειμένου να τροποποιηθεί η κατανομή στα σύνολα δεδομένων εκπαίδευσης, πρέπει να λαμβάνει υπόψιν ότι υπάρχουν πολλές περισσότερες περιπτώσεις που ανήκουν στην κλάση πλειοψηφίας από την κλάση μειοψηφίας. Δεδομένου αυτού, η κατανομή των κλάσεων είναι δυνατό να εξισορροπηθεί χρησιμοποιώντας μεθόδους υποδειγματοληψίας (undersampling) της κλάσης πλειοψηφίας, υπερδειγματοληψίας (oversampling) της κλάσης μειοψηφίας ,αλλά και συνδυασμό αυτών των δυο.

Μέσω πληθώρας μελετών αποδεικνύεται πως ένα ισορροπημένο σύνολο δεδομένων παρέχει βελτιωμένη απόδοση ταξινόμησης συγκριτικά με ένα μη ισορροπημένο σύνολο δεδομένων. Χαρακτηριστική είναι η συμβολή των Laurikkala (2001) και των Estabrooks et al. (2004) με μελέτες που αφορούν την αλλαγή της κατανομής των κλάσεων, με τον Weiss (2003) να διερευνά την επίδραση της κατανομής των κλάσεων στην περίπτωση της ταξινόμησης με δέντρα αποφάσεων (decision trees), αλλάζοντας την κατανομή των κλάσεων για την επίτευξη διαφορετικών ποσοστών μεταξύ της κλάσης μειοψηφίας και της κλάσης πλειοψηφίας και μετρώντας την απόδοση, χρησιμοποιώντας την ακρίβεια και την περιοχή κάτω από την ROC καμπύλη*.

Ο Japkowicz (2000) σύγκρινε πολλαπλές μεθόδους εξισορρόπησης και κατέληξε στο συμπέρασμα ότι τόσο οι τεχνικές υποδειγματοληψίας όσο και οι τεχνικές υπερδειγματοληψίας είναι πολύ αποτελεσματικές για την αντιμετώπιση του προβλήματος ανισορροπίας μεταξύ των κλάσεων.

*Καμπύλες ROC : καμπύλες λειτουργικού χαρακτηριστικού δέκτη ,μια τυποποιημένη μέθοδος που συνοψίζει την απόδοση ενός ταξινομητή σε σχέση με μια σειρά από «ανταλλαγές» μεταξύ αληθώς θετικών (TP) και ψευδώς θετικών (FP) ποσοστών σφάλματος.

2.3.Υπερδειγματοληψία (Oversampling)

Ο μηχανισμός της μεθόδου υπερδειγματοληψίας στηρίζεται στην προσθήκη ενός είτε τυχαία επιλεγμένου, είτε κατευθυνόμενου δείγματος, επιπλέον περιπτώσεων από την κλάση μειοψηφίας. Με τον τρόπο αυτό, ο αριθμός των συνολικών περιπτώσεων μειοψηφίας αυξάνεται, με αποτέλεσμα η κατανομή των κλάσεων να είναι πιο ισορροπημένη.

2.3.1.Μέθοδοι Υπερδειγματοληψίας

1.Τυχαία υπερδειγματοληψία (Random oversampling ,ROS)

Η τυχαία υπερδειγματοληψία περιλαμβάνει την τυχαία αντιγραφή παραδειγμάτων από την τάξη μειοψηφίας και την προσθήκη τους στο σύνολο δεδομένων εκπαίδευσης. Παραδείγματα από το σύνολο δεδομένων εκπαίδευσης επιλέγονται τυχαία με αντικατάσταση. Η διαδικασία της τυχαίας επιλογής συνεπάγεται ότι παραδείγματα από την κατηγορία μειοψηφίας μπορούν να επιλεγούν και να προστεθούν στο νέο, ισορροπημένο σύνολο δεδομένων εκπαίδευσης πολλές φορές. Πιο συγκεκριμένα, η τυχαία υπερδειγματοληψία γίνεται με την επιλογή από το αρχικό σύνολο δεδομένων εκπαίδευσης δειγμάτων από την κατηγορία μειοψηφίας ,προστίθενται στο νέο σύνολο δεδομένων εκπαίδευσης και στη συνέχεια επιστρέφονται ή «αντικαθίστανται» στο αρχικό σύνολο δεδομένων, δίνοντας τη δυνατότητα να επιλεγούν ξανά.

Η τεχνική αυτή κρίνεται αποτελεσματική για αλγορίθμους μηχανικής μάθησης που επηρεάζονται από μια ανισοκατανεμημένη κατανομή δεδομένων ,όπου η προσαρμογή του μοντέλου επηρεάζεται από πολλά διπλά παραδείγματα για μια δεδομένη τάξη. Τέτοιους αλγορίθμους αποτελούν κι αυτοί που μαθαίνουν επαναληπτικά τους συντελεστές ,όπως τα τεχνητά νευρωνικά δίκτυα, που χρησιμοποιούν στοχαστική κλίση καθόδου, όπως και μοντέλα που αναζητούν καλούς διαχωρισμούς των δεδομένων, όπως οι Μηχανές Υποστήριξης Διανυσμάτων (SVM) και τα Δέντρα Αποφάσεων (Decision Trees).

Παρόλα αυτά ,ενέχεται ο κίνδυνος της υπερβολικής προσαρμογής των επηρεαζόμενων αλγορίθμων στην κατηγορία μειοψηφίας, κατά την αναζήτηση μιας ισορροπημένης κατανομής για ένα σύνολο δεδομένων με εμφανή ανισορροπία, οδηγώντας σε αυξημένο σφάλμα γενίκευσης (overfitting). Το σφάλμα γενίκευσης έχει ως αποτέλεσμα το μοντέλο αντί να μάθει να ξεχωρίζει τα δείγματα που επαναλαμβάνονται πολλές φορές, απλά τα απομνημονεύει. Αποτέλεσμα αυτής της διαδικασίας είναι καλύτερη απόδοση στο σύνολο δεδομένων εκπαίδευσης, αλλά χειρότερη απόδοση στο συγκρατημένο ή δοκιμαστικό σύνολο δεδομένων.

2.Τεχνική υπερδειγματοληψία συνθετικής μικρότερης κλάσης (Synthetic Minority Oversampling Technique, SMOTE)

Κάνοντας λόγο για την τεχνική υπερδειγματοληψία συνθετικής μικρότερης κλάσης, πρόκειται για την πιο διαδεδομένη τεχνική αύξησης του συνόλου δεδομένων. Συγκεκριμένα, η SMOTE επιλέγει σε πρώτο βήμα τυχαία ένα δείγμα από τη μικρότερη κλάση και βρίσκει τους k πλησιέστερους γείτονές της, που επίσης ανήκουν στην κλάση μειοψηφίας. Σε δεύτερο χρόνο, επιλέγει πάλι τυχαία έναν από τους πλησιέστερους γείτονες και στην ευθεία που δημιουργούν τα δυο δείγματα στο χώρο των χαρακτηριστικών επιλέγεται ένα σημείο που είναι το καινούργιο συνθετικό δείγμα που δημιουργείται. Με την παρούσα μέθοδο δημιουργούνται όσα συνθετικά δείγματα της μικρότερης κλάσης χρειάζονται.

Το μειονέκτημα που σημειώνει η μέθοδος SMOTE είναι πως δε λαμβάνει υπόψιν την επικρατούσα κλάση, με συνέπεια, σε περίπτωση που οι κλάσεις επικαλύπτονται, τα συνθετικά δείγματα να είναι πιθανό να συμπίπτουν ή να είναι υπερβολικά κοντά με ορισμένα από τα δείγματα της κλάσης πλειοψηφίας.

3.SMOTE οριογραμμής (Borderline SMOTE)
Η μέθοδος SMOTE οριογραμμής αποτελεί τη βελτιωμένη εκδοχή της μεθόδου SMOTE και περιλαμβάνει την επιλογή των δειγμάτων εκείνων της μικρότερης κλάσης, που έχουν κατηγοριοποιηθεί λανθασμένα. Τα δείγματα που ταξινομήθηκαν λάθος είναι πιο πιθανό να ανήκουν στα δείγματα της οριογραμμής ή κοντά σε αυτήν, αφού αυτές είναι περιοχές όπου προσεγγίζονται ή επικαλύπτονται οι δυο κλάσεις. Καθ' αυτόν τον τρόπο, τα δείγματα που έχουν καταχωρηθεί λάθος είναι πιο σημαντικά και η μέθοδος εστιάζει σε αυτά, ώστε να δημιουργηθεί ένα πιο αποτελεσματικό μοντέλο ταξινόμησης.

4.Προσαρμοστική συνθετική δειγματοληψία (Adaptive Synthetic Sampling, ADASYN)
Όσον αφορά τη μέθοδο ADASYN ,δημιουργεί συνθετικά δείγματα έχοντας σε κάθε σημείο του χώρου ως κριτήριο το πλήθος των συνθετικών δειγμάτων να είναι αντιστρόφως ανάλογο με την πυκνότητα των δειγμάτων της κλάσης μειοψηφίας. Επί της ουσίας, τα δείγματα της μειοψηφικής κλάσης σταθμίζονται ανάλογα με την πυκνότητά τους κι όπου εντοπίζεται μικρή συγκέντρωση πυκνότητας, απαιτείται αυξημένη δημιουργία συνθετικών δειγμάτων.

Πρόβλημα συναντάται στις περιπτώσεις των ακραίων τιμών, καθώς η μέθοδος ADASYN τις θεωρεί σημεία με χαμηλή πυκνότητα και λανθασμένα δημιουργεί νέα δείγματα σε αυτά τα σημεία, έχοντας ως αποτέλεσμα την αρνητική επιρροή στην αποτελεσματικότητα του αλγορίθμου.

5.Ενισχυμένο SMOTE (SMOTEboost)
Η χρήση του boosting βελτιώνει τη συνολική ακρίβεια του συνόλου δεδομένων, εστιάζοντας στις αμφισβητούμενες περιπτώσεις της μικρότερης κλάσης. Στόχος είναι να μειωθεί η μεροληψία προς την κλάση πλειοψηφίας, που οφείλεται στην ανισορροπία του συνόλου δεδομένων. Οπότε ενώ η διαδικασία ενίσχυσης δίνει ίσα βάρη σε όλα τα λανθασμένα δείγματα, με τη μέθοδο SMOTEboost αυξάνονται τα βάρη στα δείγματα της μειοψηφικής κλάσης.

6.Ενίσχυση της βαθμολογημένης υπερδειγματοληψίας της μικρότερης κλάσης
(Ranked Minority Oversampling in Boosting, RAMOBoost)
Η μείωση της μεροληψίας που προκαλείται λόγω της ανισορροπίας των κλάσεων είναι ο στόχος της μεθόδου RAMOBoost, καθώς και η προσαρμογή της εκπαίδευσης του μοντέλου με βάση την κατανομή. Ο συγκεκριμένος στόχος επιτυγχάνεται με δυο σκέλη. Το πρώτο σκέλος είναι μια διαδικασία ρύθμισης ενός προσαρμοστικού βάρους που υπάρχει στη μέθοδο RAMOBoost και μετατοπίζει την οριογραμμή ανάμεσα στις δυο κλάσεις προς δείγματα που είναι δύσκολα στην εκμάθηση κι ανήκουν όχι μόνο στη μικρότερη, αλλά και στην επικρατούσα κλάση. Από την άλλη, το δεύτερο σκέλος αποτελεί μια κατανομή πιθανότητας βαθμολογημένης δειγματοληψίας που χρησιμοποιείται ,ώστε να δημιουργεί συνθετικά δείγματα της μικρότερης κλάσης προκειμένου να εξισορροπήσει μια ασύμμετρη κατανομή.

7.Υπερδειγματοληψία βασισμένη στην απόσταση Mahalanobis (Mahalanobis Distance-based Oversampling, MDO)

Η συγκεκριμένη μέθοδος υπερδειγματοληψίας βασίζεται στην απόσταση Mahalanobis, η οποία αφορά την απόσταση μεταξύ ενός σημείου και μιας κατανομής. Πρόκειται για μια επέκταση της Ευκλείδειας απόστασης. Ουσιαστικά, η μέθοδος MDO δημιουργεί συνθετικά δείγματα που έχουν την ίδια απόσταση Mahalanobis από τη θεωρούμενη ως μέση τιμή της μειοψηφικής κλάσης, λαμβάνοντας υπόψη όλα τα δείγματά της.

8.Μέθοδος υπερδειγματοληψίας με βάρη στη μικρότερη κλάση βάσει την επικρατούσα κλάση (Majority Weighted Minority Oversampling Technique, MWMOTE)

Για τον έλεγχο και τη στάθμιση με βάρη των σημαντικών δειγμάτων της κλάσης μειοψηφίας, χρησιμοποιείται η μέθοδος MWMOTE, βάσει των πλησιέστερων δειγμάτων της κλάσης πλειοψηφίας και στη συνέχεια, αυτά χρησιμοποιούνται για την εφαρμογή της υπερδειγματοληψίας, δημιουργώντας συνθετικά δείγματα. Επισημαίνεται σε αυτό το σημείο, πως η μέθοδος MWMOTE χρησιμοποιεί πληροφορίες κι από τις δυο κλάσεις του συνόλου δεδομένων.

Η μεθοδολογία, σε πρώτο στάδιο ταυτοποιεί τα δείγματα της μικρότερης κλάσης που δυσχεραίνουν την εκμάθηση και τους αποδίδει βάρη ανάλογα με τη σημαντικότητά τους, σύμφωνα με πληροφορίες που προκύπτουν για την απόστασή τους από το πλησιέστερο δείγμα της επικρατούσας κλάσης. Στη συνέχεια, προσδιορίζονται οι συστάδες της μειοψηφικής κλάσης και χρησιμοποιούνται σταθμισμένα δείγματα αυτής, για τη δημιουργία συνθετικών δειγμάτων εντός των συστάδων. Η διαδικασία αυτή συμβαίνει, ώστε να διασφαλιστεί ότι τα παραγόμενα δείγματα βρίσκονται πάντα μέσα σε κάποια συστάδα της κλάσης μειοψηφίας και δεν επικαλύπτονται με τις περιοχές της πλειοψηφικής κλάσης.

2.4.Υποδειγματοληψία (Undersampling)

Η μέθοδος υποδειγματοληψίας λειτουργεί με τον αντίθετο τρόπο από τη μέθοδο υπερδειγματοληψίας, καθώς εν προκειμένω αφαιρούνται περιπτώσεις από την τάξη πλειοψηφία, διατηρώντας παράλληλα όλες τις περιπτώσεις της κλάσης μειοψηφίας λόγω της σπάνιας εμφάνισής τους ,με απόρροια αυτής τη λιγοστή πληροφορία που παρέχουν. Στη βιβλιογραφία συναντάται πληθώρα μεθόδων υποδειγματοληψίας, με την πιο απλή μέθοδο να είναι αυτή της τυχαίας υποδειγματοληψίας, οι σύνδεσμοι Tomek, η μέθοδος μονόπλευρης επιλογής κι αρκετές ακόμη ,που παρουσιάζονται παρακάτω.

2.4.1.Μέθοδοι Υποδειγματοληψίας

1.Τυχαία υποδειγματοληψία (Random undersampling ,RUS)

Η τυχαία υποδειγματοληψία περιλαμβάνει την τυχαία επιλογή παραδειγμάτων από την πλειοψηφική τάξη προς διαγραφή από το σύνολο δεδομένων εκπαίδευσης. Αποτέλεσμα αυτής της διαδικασίας είναι η μείωση του αριθμού παραδειγμάτων στην πλειοψηφική τάξη στη μετασχηματισμένη έκδοση του συνόλου δεδομένων εκπαίδευσης. Υπάρχει η δυνατότητα επανάληψης της διαδικασίας έως ότου να επιτευχθεί η επιθυμητή κατανομή κλάσης, όπως για παράδειγμα ίσος αριθμός δειγμάτων για κάθε κατηγορία.

Ο περιορισμός που θέτει η μέθοδος υποδειγματοληψίας σχετίζεται με τη διαγραφή δειγμάτων από την κλάση πλειοψηφίας, τα οποία μπορεί να είναι χρήσιμα, σημαντικά ή ακόμα και κρίσιμα για την τοποθέτηση ενός ισχυρού ορίου απόφασης κι αυτό διότι η διαγραφή των δειγμάτων γίνεται με τυχαίο τρόπο. Φυσικά, η τυχαιότητα στη διαγραφή των δειγμάτων από την πλειοψηφική κατηγορία και η πιθανότητα απώλειας πληροφορία από σημαντικά δεδομένα έχει αντίκτυπο στην απόδοση, ακρίβεια του μοντέλου.

2.Near Miss υποδειγματοληψία

Η Near Miss υποδειγματοληψία στηρίζεται στη μέθοδο των k-πλησιέστερων γειτόνων, αποτελώντας μια συλλογή τριών παραπλήσιων τεχνικών που επιλέγουν δείγματα βάσει της απόστασης των δειγμάτων της επικρατούσας κλάσης από τα δείγματα μικρότερης κλάσης.

Πιο συγκεκριμένα, οι τρεις αυτές τεχνικές είναι :

1.Η Near Miss-1 ,που επιλέγει να διατηρήσει τα δείγματα της επικρατούσας κλάσης χρησιμοποιώντας ως λογική τη μικρότερη μέση απόσταση από τα τρία πλησιέστερα δείγματα της μειοψηφικής κλάσης.

Σε αυτή την τεχνική αναμένεται οι συστάδες των δειγμάτων της κλάσης μειοψηφίας που θα παραμείνουν να είναι αυτές που βρίσκονται γύρω από τα δείγματα της μειοψηφικής κλάσης, δηλαδή στην περιοχή επικάλυψης.

2.Αναφερόμενοι στην τεχνική Near Miss-2, γίνεται λόγος για τη διατήρηση δειγμάτων της επικρατούσας κλάσης, χρησιμοποιώντας ως λογική τη μικρότερη μέση απόσταση από τα τρία πιο απομακρυσμένα δείγματα της μειοψηφικής κλάσης. Αναμένεται λοιπόν, τα επιλεγμένα δείγματα της επικρατούσας κλάσης να βρίσκονται στο κεντρικότερο σημείο της επικάλυψης των δυο κατηγοριών.

3.Στην τρίτη και τελευταία τεχνική, την Near Miss-3, διατηρείται ένας συγκεκριμένος αριθμός δειγμάτων της πλειοψηφικής κλάσης για κάθε δείγμα της μειοψηφικής κατηγορίας που είναι πιο κοντά. Η λογική αυτής της τεχνικής επιδιώκει στο αποτέλεσμα

τα δείγματα της μειοψηφικής κατηγορίας που βρίσκονται στην περιοχή της επικάλυψης των δυο κατηγοριών να έχουν σχετικά κοντά τους μέχρι ν γείτονες από την επικρατούσα κλάση.

3.Μέθοδος συμπυκνωμένων πλησιέστερων γειτόνων (Condensed Nearest Neighbors, CNN)

Πρόκειται για μια τεχνική υποδειγματοληψίας που αναζητά ένα υποσύνολο δειγμάτων που να επιτυγχάνει 100% επιτυχία του μοντέλου, με το υποσύνολο να ονομάζεται ελάχιστο συνεπές σύνολο. Η συγκεκριμένη μέθοδος αποτελεί μια εναλλακτική της k-Nearest Neighbors μεθόδου, KNN, με τη μέθοδο CNN να διαγράφει τα δείγματα της επικρατούσας κλάσης που βρίσκονται μακριά από την οριογραμμή και δεν προσφέρουν κάτι στην κατηγοριοποίηση των νέων δειγμάτων.

Σε πρώτο στάδιο απαριθμούνται τα δείγματα στο σύνολο δεδομένων και μετέπειτα αποθηκεύονται μόνο στην περίπτωση που δεν μπορούν να κατηγοριοποιηθούν σωστά από τα ήδη αποθηκευμένα δείγματα. Στην περίπτωση μη ισορροπημένου συνόλου δεδομένων, τα αποθηκευμένα δείγματα αποτελούνται από όλα τα δείγματα της μειοψηφικής κλάσης και στη συνέχεια αποθηκεύονται δείγματα της επικρατούσας κλάσης που κατηγοριοποιούνται εσφαλμένα. Στο τελικό στάδιο, ενδέχεται η αναλογία των δυο κλάσεων να μην είναι ένα προς ένα, αλλά σίγουρα θα είναι μικρότερη από την αρχική αναλογία του συνόλου δεδομένων.

4.Σύνδεσμοι Tomek (Tomek links)

Εφαρμόζοντας τη μέθοδο των συνδέσμων Tomek αναζητούνται δείγματα που είναι πολύ κοντινά μεταξύ τους, αλλά προέρχονται από διαφορετικές κλάσεις. Η κοντινή απόσταση μπορεί να αξιολογηθεί με κάποια μετρική, όπως η Ευκλείδεια απόσταση.

Οι προϋποθέσεις που πρέπει να πληρούνται ώστε να θεωρηθεί ένα ζευγάρι δειγμάτων ως σύνδεσμος Tomek είναι οι εξής :

i. Το δείγμα Α να έχει πιο κοντινό δείγμα το Β

ii. Το δείγμα Β να έχει πιο κοντινό δείγμα το Α

iii. Τα δείγματα Α και Β να ανήκουν σε διαφορετικές κατηγορίες.

Ο τρόπος χειρισμού στη μέθοδο των συνδέσμων Tomek είναι:
i) είτε να σβηστούν και τα δυο δείγματα, προκειμένου να αυξηθεί η απόσταση μεταξύ των δυο κλάσεων, διευκολύνοντας τη διαδικασία της κατηγοριοποίησης.
ii) είτε να σβηστεί μόνο το δείγμα που ανήκει στην επικρατούσα κλάση, θεωρώντας πως αποτελεί θόρυβο.

5.Μέθοδος επεξεργασμένων πλησιέστερων γειτόνων (Edited Nearest Neighbors, ENN)

Αρχικά, η μέθοδος εφαρμόζει μια διαδικασία για τους 3 πλησιέστερους γείτονες των δειγμάτων που η κατηγοριοποίησή τους μέσω του μοντέλου δεν είναι σωστή και λειτουργεί ως μια πρώτη μείωση του πλήθους των δεδομένων, αφού αφαιρούνται δείγματα από την επικρατούσα κλάση.

Η πρώτη φάση με τους 3 πλησιέστερους γείτονες λειτουργεί ως εξής :

Υπολογίζονται οι 3 πλησιέστεροι γείτονες για κάθε δείγμα. Αν το δείγμα είναι της επικρατούσας κλάσης και δεν κατηγοριοποιούνται σωστά, τότε διαγράφεται. Αν όμως το δείγμα ανήκει στη μικρότερη κλάση και δεν κατηγοριοποιείται σωστά, τότε διαγράφονται οι γείτονες που ανήκουν στην πλειοψηφική κλάση.

Η δεύτερη φάση είναι μια κατηγοριοποίηση με ένα μόνο πλησιέστερο γείτονα, από όπου προκύπτει η τελική εκτίμηση.

6.Μέθοδος μονόπλευρης επιλογής (One-sided Selection, OSS)
Η μέθοδος της μονόπλευρης επιλογής αποτελεί ένα συνδυασμό των συνδέσμων Tomek και της μεθόδου CNN. Οι σύνδεσμοι Tomek απομακρύνουν τα δείγματα της επικρατούσας κλάσης που βρίσκονται στην οριογραμμή μεταξύ δυο κατηγοριών, αλλά και τα δείγματα που αποτελούν θόρυβο. Από την άλλη, η μέθοδος CNN διαγράφει όσα δείγματα της επικρατούσας κλάσης δεν προσφέρουν στην κατηγοριοποίηση των νέων δειγμάτων.

7.Μέθοδος εκκαθάρισης γειτόνων (Neighborhood Cleaning Rule, NCR)
Η παρούσα μέθοδος αποτελεί συνδυασμό της μεθόδου συμπυκνωμένων πλησιέστερων γειτόνων, CNN και της μεθόδου επεξεργασμένων πλησιέστερων γειτόνων, ENN. Η διαφορά με τη μέθοδο μονόπλευρης επιλογής είναι πως διαγράφονται λιγότερα από τα περιττά δείγματα που βρίσκονται μακριά από την οριογραμμή.

Η μέθοδος NCR εστιάζει στον “καθαρισμό” των δειγμάτων που διατηρούνται, επιδιώκοντας το σύνολο που θα παραμείνει να είναι πιο ποιοτικό και κατά συνέπεια πιο αποδοτικό. Εν κατακλείδι ,προκύπτει ένα λιγότερο εξισορροπημένο σύνολο δεδομένων μεταξύ των δυο κατηγοριών, αλλά τα δείγματα που παραμένουν είναι πιο συγκεκριμένα ως προς τα όρια της κάθε κατηγορίας, ενώ έχει διαγραφεί και ο θόρυβος κι επομένως η κατηγοριοποίηση είναι αποτελεσματικότερη.

8.Μέθοδος με χρήση συστάδων (Clustering)
Μια άλλη μέθοδος, που περιέχει μεγαλύτερο βαθμό τυχαιότητας, είναι η δημιουργία συστάδων, δηλαδή η ομαδοποίηση των δειγμάτων της πλειοψηφικής κατηγορίας. Αφαιρούνται κάποια δείγματα από κάθε συστάδα, μειώνοντας τον αριθμό, αλλά παράλληλα επιδιώκοντας το δείγμα που θα παραμείνει να είναι αντιπροσωπευτικό της κατηγορίας.

Ο τρόπος επιλογής των τελικών δειγμάτων της επικρατούσας κλάσης είναι :

1.με τη μέθοδο των κεντρικών δειγμάτων των συστάδων, όπου διατηρείται μόνο το κεντρικό σημείο κάθε συστάδας, δημιουργώντας το καινούργιο σύνολο δεδομένων της πλειοψηφικής κλάσης.

2.με την αντικατάσταση του κεντρικού σημείου κάθε συστάδας με το πλησιέστερο γειτονικό δείγμα, που ήδη υπάρχει, διατηρώντας μόνο αυτά τα δείγματα, ώστε να υπάρχει ένα υποσύνολο του αρχικού συνόλου δεδομένων.

9.Μέθοδος βελτιστοποίησης του υποσυνόλου δειγμάτων (Sample Subset Optimization, SSO)

Η μέθοδος SSO βασίζεται στην επιλογή ενός συγκεκριμένου υποσυνόλου από όλα τα διαθέσιμα δείγματα, χρησιμοποιώντας ως κριτήριο την ελαχιστοποίηση του αναμενόμενου σφάλματος, το οποίο προκύπτει με μια διαδικασία διασταυρούμενης επικύρωσης (cross validation) στα δεδομένα εκπαίδευσης.

ΚΕΦΑΛΑΙΟ 3. Ταξινόμηση

3.1.Μηχανική Μάθηση (Machine Learning)

Η Μηχανική Μάθηση αποτελεί έναν κλάδο της Τεχνητής Νοημοσύνης, ο οποίος αφορά τη μελέτη κι εφαρμογή αλγορίθμων και μοντέλων της στατιστικής για την εκτέλεση ενός έργου χωρίς προκαθορισμένες οδηγίες, αλλά με την αναγνώριση κι αξιοποίηση μοτίβων και τεκμηρίων.

Ο ορισμός που δόθηκε από τον Άρθουρ Σάμουελ (1959) για τη μηχανική μάθηση, τη χαρακτηρίζει ως το πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν χωρίς να έχουν προγραμματιστεί ρητά για το αποτέλεσμα που θα δώσουν. Πιο συγκεκριμένα, η μηχανική μάθηση αποτελεί υποπεδίο της επιστήμης των υπολογιστών που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη.

Οι αλγόριθμοι που χρησιμοποιούνται κι εκπληρώνουν τους σκοπούς της μηχανικής μάθησης, βελτιώνουν την απόδοσή τους, ευστοχία πρόβλεψης & περιγραφής αντικειμένου, στην εργασία που τους έχει ανατεθεί χρησιμοποιώντας και χτίζοντας πάνω στην προηγούμενη εμπειρία που έχουν αποκτήσει.

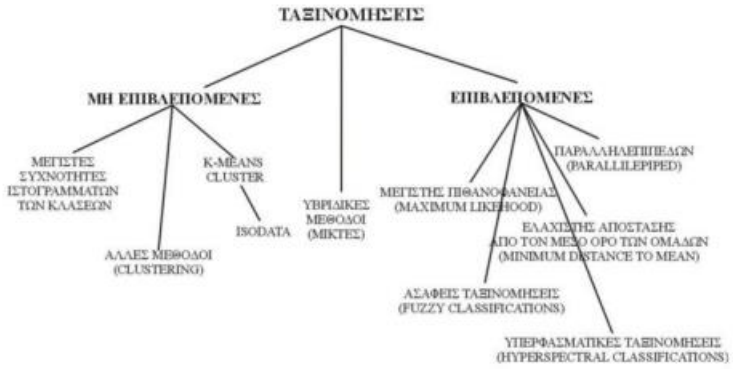
3.2.Ταξινόμηση Πολυφασματικών Απεικονίσεων

Ως ταξινόμηση ψηφιακής πολυφασματικής εικόνας ορίζεται η διαδικασία κατηγοριοποίησης των εικονοστοιχείων σε ομάδες ή τάξεις, με κοινά χαρακτηριστικά από άποψη φασματικής απόκρισης ή/κι υφής.

Είναι διαθέσιμη πληθώρα αυτοματοποιημένων μεθόδων ταξινόμησης, στις οποίες αξιοποιούνται διάφορα κεφάλαια των μαθηματικών και της πληροφορικής, όπως για παράδειγμα η θεωρία αλγορίθμων, οι διανυσματικοί χώροι, τα νευρωνικά δίκτυα, ο προγραμματισμός, με τη χρήση του οποίου αναπτύσσεται το κατάλληλο λογισμικό, με σκοπό να υλοποιηθεί σε σύντομο χρόνο και με ποσοτικά κριτήρια η διαδικασία ταυτοποίησης του κάθε εικονοστοιχείου σε μια κατηγορία.

Στην περίπτωση που η ταξινόμηση πραγματοποιείται με βάση τη φασματική υπογραφή του κάθε εικονοστοιχείου ξεχωριστά, τότε πρόκειται για πολυφασματική ταξινόμηση (multispectral classification). Θεμελιώδη σημασία σε αυτό το είδος ταξινόμησης έχει η έννοια της περιοχής ομαδοποίησης (cluster). Κάνοντας λόγο για περιοχή ομαδοποίησης, πρόκειται για μια περιοχή στο φασματικό χώρο, στην οποία ανήκουν οι φασματικές υπογραφές πολλών εικονοστοιχείων. Η διαδικασία πολυφασματικής ταξινόμησης είναι μια διαδικασία ένταξης του κάθε εικονοστοιχείου σε μια περιοχή ομαδοποίησης, με βάση κάποιο ποσοτικό κριτήριο. Εικονοστοιχεία που δεν είναι δυνατόν να ενταχθούν σε καμία περιοχή ομαδοποίησης παραμένουν αταξινόμητα, δηλαδή μη αναγνωρισμένα.

Από τη μελέτη των ταξινομήσεων έχουν προκύψει δυο βασικές κατηγορίες, οι «Επιβλεπόμενες Ταξινομήσεις» και οι «Μη Επιβλεπόμενες Ταξινομήσεις». Η πρώτη περίπτωση αλγορίθμου ταξινόμησης βασίζεται στην εκ των προτέρων γνώση (a priori) της κατηγορίας κάλυψης γης και στη δυνατότητα πρόσβασης σε υπάρχουσες κατηγορίες, με στόχο την κατηγοριοποίηση των περιοχών που έχουν δεχθεί δειγματοληψία, ενώ από την άλλη πλευρά, στη μη επιβλεπόμενη ταξινόμηση, οι κλάσεις που προκύπτουν με την ταξινόμηση δεν είναι γνωστές αρχικά (a posteriori καταγραφή κατηγορίας). Άρα, οι δυο βασικές κατηγορίες ταξινόμησης διαφέρουν ως προς το διαδικαστικό μέρος.



Σχήμα 1. Σχεδιάγραμμα ειδών Ταξινομήσεων κι αλγορίθμων αυτών

3.2.1.Επιβλεπόμενη Ταξινόμηση (Supervised Classification)

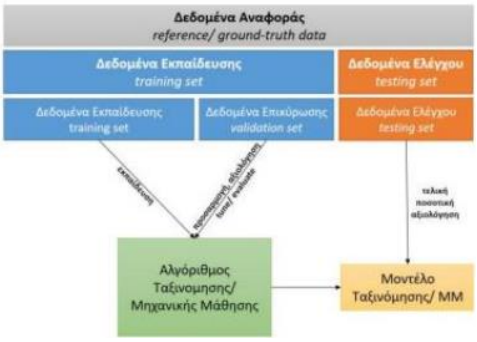
Ο αλγόριθμος που εφαρμόζεται κατά την επιβλεπόμενη ταξινόμηση εκπαιδεύεται στον εντοπισμό των ραδιομετρικών τιμών κάθε κατηγορίας κάλυψης γης σύμφωνα με τις ομογενείς ή όχι δειγματοληπτικές περιοχές, οι οποίες περιγράφουν εξαρχής την κάθε κατηγορία γης (π.χ. σκληροφυλλική βλάστηση). Οι αλγόριθμοι ταξινόμησης είναι πολυάριθμοι και μπορεί να βασίζονται σε μεθόδους στατιστικής ανάλυσης είτε σε τεχνικές μάθησης (Machine learning). Αποτέλεσμα αυτής της διαδικασίας είναι η δημιουργία φασματικών υπογραφών από τις περιοχές δειγματοληψίας για κάθε κατηγορία, από τις οποίες αξιολογούνται τα χαρακτηριστικά κάθε εικονοστοιχείου κι επιλέγεται σε ποια από τις υφιστάμενες καλύψεις γης ανήκει το εικονοστοιχείο ή αν τοποθετείται στην κατηγορία αταξινόμητα, επειδή οι ραδιομετρικές τιμές που του αντιστοιχούν δεν ανήκουν σε κάποια από τις κατηγορίες που έχουν οριστεί.

Με τις επιβλεπόμενες ταξινομήσεις στόχος είναι η δημιουργία θεματικών χαρτών με αυτόματο ή ημι-αυτόματο τρόπο, για τον οποίο έχουν εντοπιστεί οι κατηγορίες χρήσης γης της εικόνας κι έχουν κατηγοριοποιηθεί σε επίπεδο εικονοστοιχείου. Σε πρώτο στάδιο δημιουργούνται δεδομένα εκπαίδευσης με τα οποία προσδιορίζονται οι περιοχές της εικόνας που ανήκουν σε συγκεκριμένες κατηγορίες. Η δημιουργία δεδομένων εκπαίδευσης κρίνεται σκόπιμη, ώστε να εκπαιδευτεί ένας αλγόριθμος για να αναγνωρίζει και να ξεχωρίζει τις κατηγορίες στα δεδομένα εισόδου. Παράλληλα, για την αξιολόγηση των αποτελεσμάτων της ταξινόμησης δημιουργούνται δεδομένα επικύρωσης κι ελέγχου.

Πιο συγκεκριμένα, εφόσον έχει μελετηθεί η εικόνα κι είναι γνωστές οι χρήσεις γης που εντοπίζονται σε αυτή, εντοπίζονται και προσδιορίζονται μέσω δειγματοληψίας τμήματα της περιοχής μελέτης που ανήκουν στις κατηγορίες αυτές, εξάγεται η φασματική υπογραφή αυτών και σε τελευταία στάδιο με βάση έναν αλγόριθμο, κάθε εικονοστοιχείο της εικόνας ταξινομείται στην κατηγορία χρήσης γης στην οποία ανήκει, αν βέβαια, αυτή έχει οριστεί στον αλγόριθμο.

Ο σκοπός της επιλογής δεδομένων εκπαίδευσης που αντιστοιχούν σε συγκεκριμένες περιοχές/κατηγορίες χρήσης γης είναι ο υπολογισμός στατιστικών μεγεθών που αφορούν την κάθε θεματική κατηγορίας ξεχωριστά. Τα δεδομένα εκπαίδευσης δημιουργούνται με επισκέψεις στο πεδίο, για να εντοπιστούν και να καταγραφούν τα τμήματα της περιοχής στα οποία εντοπίζονται οι θεματικές κατηγορίες, ή αν δεν είναι εφικτή η επίσκεψη στο πεδίο, υπάρχει η δυνατότητα ο προσδιορισμός των περιοχών εκπαίδευσης να γίνει με ψηφιοποίηση δειγμάτων των θεματικών τάξεων της εικόνας, εφόσον έχει μελετηθεί η εικόνα προσεκτικά από άποψη φωτοερμηνείας κι έχει γίνει χρήση συμβουλευτικών πηγών.

Με τον ίδιο τρόπο δημιουργούνται και τα δεδομένα επικύρωσης κι ελέγχου της ταξινόμησης, τα οποία χρησιμοποιούνται μόνο για την ποσοτική αξιολόγηση και τον έλεγχο της ακρίβειας του θεματικού χάρτη που προκύπτει από τα αποτελέσματα της ταξινόμησης.



Σχήμα 2. Σχεδιάγραμμα αλγορίθμου για την εκπόνηση Επιβλεπόμενης Ταξινόμησης

3.2.2.Μη Επιβλεπόμενη Ταξινόμηση (Unsupervised Classification)

Στην περίπτωση μη επιβλεπόμενης ταξινόμησης δε χρησιμοποιούνται ζεύγη τιμών εισόδου-εξόδου. Ως στόχος ορίζεται η αναγνώριση των μοτίβων στα δεδομένα εισόδου χωρίς την ανατροφοδότηση από τα δεδομένα εξόδου, με το μοντέλο να καλείται να ανακαλύψει μοτίβα και να καταλάβει από μόνο του τη δομή των χαρακτηριστικών των δεδομένων εισόδου. Η ικανότητά του να ανακαλύπτει ομοιότητες και διαφορές στις πληροφορίες που του δίνονται ,το καθιστά ιδανική λύση για διερευνητικές αναλύσεις δεδομένων, στρατηγικές διασταυρωμένων πωλήσεων κι αναγνώριση εικόνας. Τα μοντέλα μάθησης χωρίς επίβλεψη χρησιμοποιούνται για τρεις κύριες εργασίες : ομαδοποίηση, συσχέτιση και μείωση διαστάσεων.

- Ομαδοποίηση (clustering) : Πρόκειται για μια τεχνική, η οποία ομαδοποιεί δεδομένα χωρίς ετικέτα με βάση τις ομοιότητες και τις διαφορές τους. Οι αλγόριθμοι ομαδοποίησης αξιοποιούνται σε εφαρμογές επεξεργασίας ακατέργαστων, μη ταξινομημένων δεδομένων σε ομάδες που αντιπροσωπεύονται από δομές ή μοτίβα στις πληροφορίες.
- Συσχέτιση (association) : Ένας κανόνας συσχέτισης είναι μια μέθοδος βασισμένη σε κανόνες για τον εντοπισμό σχέσεων μεταξύ μεταβλητών σε ένα δοσμένο σύνολο δεδομένων. Τέτοιοι είδους μέθοδοι χρησιμοποιούνται σε περιπτώσεις ανάλυσης του καλαθιού στις αγορές, επιτρέποντας στις εταιρίες να κατανοήσουν καλύτερα τις σχέσεις μεταξύ διαφορετικών προϊόντων.
- Μείωση διαστάσεων (dimensionality reduction) : Η απόδοση ενός αλγορίθμου επηρεάζεται από τη χρήση περισσότερων δεδομένων, παρόλο που αποδίδουν πιο ακριβή αποτελέσματα (overfitting) κι ενδέχεται να δυσχεραίνεται η απεικόνιση συνόλων δεδομένων. Η μείωση διαστάσεων είναι μια τεχνική που χρησιμοποιείται όταν ο αριθμός των χαρακτηριστικών ή των διαστάσεων σε ένα σύνολο δεδομένων είναι πολύ υψηλός. Ουσιαστικά, μειώνει τον αριθμό των δεδομένων που εισάγονται στο μοντέλο, σε ένα διαχειρίσιμο μέγεθος, διατηρώντας παράλληλα την ακεραιότητα του συνόλου δεδομένων όσο το δυνατόν περισσότερο.

3.3.Δεδομένα αλγορίθμων ταξινόμησης

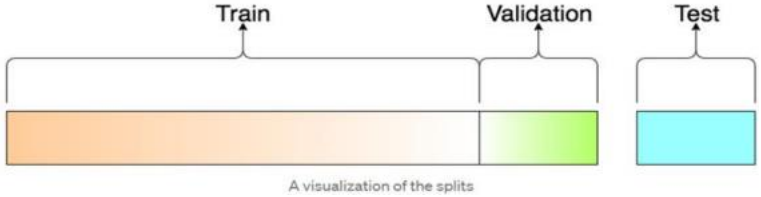
Τα δεδομένα που χρησιμοποιούνται κατά την εκπόνηση των αλγορίθμων μηχανικής μάθησης κατατάσσονται σε τρεις κατηγορίες προκειμένου να γίνουν ακριβείς ταξινομήσεις. Οι τρεις αυτές κατηγορίες είναι τα δεδομένα εκπαίδευσης, ελέγχου κι επικύρωσης (train, test, validation data αντίστοιχα).

- Training data. Τα δεδομένα αυτά δίνονται στον αλγόριθμο σαν είσοδος, εκπαιδεύοντας τον με την πληροφορία για κάθε κατηγορία κάλυψης/χρήσης γης. Το μοντέλο αξιολογεί επανειλημμένα τα δεδομένα για να μάθει περισσότερα για τη συμπεριφορά τους και στη συνέχεια προσαρμόζεται αναλόγως.
- Validation data. Κατά τη διάρκεια της εκπαίδευσης του μοντέλου, τα δεδομένα επικύρωσης είναι νέα δεδομένα που εισάγονται στο μοντέλο και δεν έχουν αξιολογηθεί σε προηγούμενο στάδιο, χρονίζοντας έτσι τον αλγόριθμο ταξινόμησης, δίνοντας τις πρώτες υπερπαραμέτρους. Ουσιαστικά, αυτά τα δεδομένα αποτελούν την πρώτη δοκιμή σε νέα δεδομένα και με αυτόν τον τρόπο επιτρέπουν στους μελετητές/παρατηρητές να αξιολογήσουν πόσο καλά το μοντέλο κάνει προβλέψεις με βάση τα νέα δεδομένα.

  Δε χρησιμοποιούνται πάντοτε δεδομένα επικύρωσης από τους παρατηρητές κατά την εφαρμογή αλγορίθμων ταξινόμησης, παρόλο που αυτά είναι πιθανό να παρέχουν χρήσιμες πληροφορίες για τη βελτιστοποίηση των υπερπαραμέτρων.

- Testing data. Ολοκληρώνοντας την κατασκευή του μοντέλου, τα δεδομένα ελέγχου επιβεβαιώνουν για ακόμη μια φορά ότι δύνανται να γίνουν ακριβείς προβλέψεις. Επί της ουσίας, τα δεδομένα ελέγχου παρέχουν έναν τελικό, πραγματικό κόσμο ενός αόρατου συνόλου δεδομένων για να επιβεβαιώσουν ότι ο αλγόριθμος μηχανικής μάθησης εκπαιδεύτηκε αποτελεσματικά.

Συνοψίζοντας, οι αλγόριθμοι ταξινόμησης απαιτούν δεδομένα εκπαίδευσης, τα οποία ο αλγόριθμος θα αναλύσει, θα ταξινομήσει τις εισόδους και τις εξόδους και στη συνέχεια θα τα αναλύσει ξανά. Με αυτόν τον τρόπο, ο αλγόριθμος απομνημονεύει όλα τα χαρακτηριστικά των δεδομένων εκπαίδευσης, με τη διαδικασία αυτή να ενέχει τον κίνδυνο να δημιουργούνται προβλήματα όταν χρειάζεται να ληφθούν υπόψη δεδομένα από άλλες πηγές. Τη λύση δίνουν τα δεδομένα επικύρωσης, όπου παρέχουν έναν αρχικό έλεγχο ότι το μοντέλο είναι ικανό να πραγματοποιήσει χρήσιμες προβλέψεις. Ο αλγόριθμος τότε μπορεί να αξιολογήσει δεδομένα εκπαίδευσης και δεδομένα επικύρωσης ταυτόχρονα.



Σχήμα 3. Διαχωρισμός δεδομένων για αλγορίθμους ταξινόμησης
https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

3.4.Ακρίβεια Ταξινόμησης

Κατά τη διαδικασία ταξινόμησης πολυφασματικών απεικονίσεων, όποιος αλγόριθμος κι αν χρησιμοποιείται, σημαντική είναι η ακρίβεια των αποτελεσμάτων που αποδίδει κάθε ταξινομητής.

Ο απλούστερος τρόπος αξιολόγησης της απόδοσης ενός ταξινομητή είναι η οπτική αξιολόγηση του χάρτη ταξινόμησης, συγκρίνοντάς τον με το Corine Land Cover και την ίδια την πολυφασματική απεικόνιση που παράγει τα δεδομένα για την εκπαίδευση και τον έλεγχο του αλγορίθμου. Όπως είναι κατανοητό, μια τέτοια μέθοδος αξιολόγησης δεν είναι ιδιαίτερα ακριβής ,καθώς κρίνεται από υποκειμενικότητα και τη διακριτική ικανότητα κι ευχέρεια του παρατηρητή.

Ένας ακόμη τρόπος για την αξιολόγηση και τον έλεγχο των αποτελεσμάτων της ταξινόμησης είναι η δημιουργία ενός πίνακα σύγχυσης, ο οποίος περιγράφει την καταλληλότητα μεταξύ των κλάσεων και των δεδομένων αναφοράς, χρησιμοποιώντας διάφορους συντελεστές για να εκφράσει την επιτυχία του αλγορίθμου, όπως είναι η συνολική ακρίβεια, η ακρίβεια παραγωγού, η ακρίβεια χρήστη.

Για την παρούσα εργασία, ο πίνακας σύγχυσης θα αποτελέσει το βασικό παράγοντα αξιολόγησης των αποτελεσμάτων των πειραμάτων κι ιδιαίτερα μέσα από τη χρήση των ακριβειών που παράγονται από αυτόν (ακρίβεια παραγωγού κι ακρίβεια χρήστη, όχι η συνολική), που αναπτύσσονται σε ακόλουθη ενότητα (Ενότητα 3.4.2.).

3.4.1.Πίνακας Σύγχυσης

Ο πίνακας σύγχυσης (ή πίνακας σφαλμάτων ή πίνακας σύμπτωσης) χρησιμοποιείται κυρίως σε εφαρμογές μηχανικής μάθησης κι ειδικότερα σε προβλήματα που αφορούν στατιστική ταξινόμηση. Οι πίνακες σφαλμάτων συγκρίνουν τη σχέση μεταξύ γνωστών επίγειων δεδομένων αναφοράς (αληθών δεδομένων) και των αντίστοιχων αποτελεσμάτων μιας αυτόματης διαδικασίας ταξινόμησης κατηγορία προς κατηγορία. Τέτοιοι πίνακες είναι τετραγωνικοί (MxM), με αριθμό γραμμών (M) και στηλών (M) ίσο με τον αριθμό των κατηγοριών (M), των οποίων εκτιμάται η ακρίβεια. (Φωτοερμηνεία-Τηλεπισκόπηση, Αργιαλάς, Δ.,1999)

Σε έναν πίνακα σύγχυσης ,κάθε ένα στοιχείο έχει τη δική του ερμηνεία σχετικά με την ακρίβεια της ταξινόμησης. Αναλύοντας αυτό, ισχύει ότι οι σειρές του πίνακα αντιστοιχούν στις κλάσεις που αντιστοιχούν σε αληθή δεδομένα, οι στήλες αντιστοιχούν στις κλάσεις που ταξινομούνται από τον αλγόριθμο και με τη σειρά τους τα διαγώνια στοιχεία του πίνακα αντιπροσωπεύουν τον αριθμό των σωστά ταξινομημένων εικονοστοιχείων κάθε κατηγορίας, δηλαδή τον αριθμό των εικονοστοιχείων των δεδομένων ελέγχου, με ένα ορισμένο όνομα κατηγορίας, που το ίδιο έχει προκύψει κι από την ταξινόμηση τω εικονοστοιχείων. Επίσης, για τα εκτός διαγωνίου στοιχεία ισχύει ότι αντιστοιχούν σε λανθασμένα ταξινομημένα εικονοστοιχεία ή σφάλματα κατάταξης. Δηλαδή, εκτός διαγωνίου τοποθετούνται τα εικονοστοιχεία όπου ως δεδομένα ελέγχου ανήκουν σε άλλη κατηγορία από αυτή που περιγράφουν τα εικονοστοιχεία που ταξινομήθηκαν από τον αλγόριθμο. Σε αυτή την περίπτωση ισχύουν δυο περιπτώσεις :

1.Τα εκτός διαγωνίου στοιχεία στις <u>σειρές</u> του πίνακα αντιπροσωπεύουν εικονοστοιχεία των δεδομένων ελέγχου, μιας συγκεκριμένης κατηγορίας τα οποία εξαιρέθηκαν από την κατηγορία αυτή κατά την ταξινόμηση. Αυτά τα σφάλματα είναι επίσης γνωστά ως <u>σφάλματα παράλειψης ή αποκλεισμού</u>.

2.Τα εκτός διαγωνίου στοιχεία των στηλών του πίνακα αντιστοιχούν σε εικονοστοιχεία των δεδομένων ελέγχου, άλλων κατηγοριών που συμπεριλήφθηκαν σε μια συγκεκριμένη κατηγορία κατά την ταξινόμηση. Τέτοια σφάλματα είναι γνωστά κι ως σφάλματα συμπερίληψης ή προμήθειας.

3.4.2.Μεγέθη Αξιολόγησης Ακρίβειας

Ένα χαρακτηριστικό μέγεθος που γίνεται γνωστό μέσα από τον πίνακα σφαλμάτων είναι η συνολική ακρίβεια της ταξινόμησης που υπολογίζεται διαιρώντας το συνολικό αριθμό των ορθά ταξινομημένων εικονοστοιχείων, με το συνολικό αριθμό εικονοστοιχείων αναφοράς.

Το μέγεθος που περιγράφει την ακρίβεια της ταξινόμησης για την κάθε κλάση ονομάζεται ακρίβεια του παραγωγού. Προκύπτει από το λόγο των σωστά ταξινομημένων εικονοστοιχείων σε σχέση με όλα τα εικονοστοιχεία της εν λόγω κλάσης των δεδομένων ελέγχου. Για κάθε κλάση εικονοστοιχείων των δεδομένων ελέγχου (σειρά), ο αριθμός των σωστά ταξινομημένων εικονοστοιχείων διαιρείται με το συνολικό αριθμό των εικονοστοιχείων των δεδομένων ελέγχου αυτής της κλάσης.

Το μέγεθος που αντιπροσωπεύει την αξιοπιστία της κάθε κλάσης στην ταξινομημένη εικόνα ορίζεται ως ακρίβεια χρήστη. Πρόκειται για το κλάσμα των σωστά ταξινομημένων εικονοστοιχείων σε σχέση με όλα τα εικονοστοιχεία που ταξινομούνται σε αυτή την κατηγορία στην ταξινομημένη εικόνα. Για κάθε κλάση της ταξινομημένης εικόνας (στήλη), ο αριθμός των σωστά ταξινομημένων εικονοστοιχείων διαιρείται με το συνολικό αριθμό των εικονοστοιχείων που ταξινομήθηκαν σε αυτή την κατηγορία.

Στα στοιχεία που αντλούνται από τον πίνακα σφαλμάτων συμπεριλαμβάνεται η στατιστική τιμή k-hat. Η τιμή αυτή είναι ένα μέτρο της διαφοράς μεταξύ της πραγματικής συμφωνίας μεταξύ των επίγειων δεδομένων αναφοράς κι ενός αυτόματου αλγόριθμου ταξινόμησης και της τυχαίας συμφωνίας μεταξύ των επίγειων δεδομένων αναφοράς κι ενός τυχαίου αλγορίθμου ταξινόμησης. Η στατιστική παράμετρος k ορίζεται ως

k = (παρατηρηθείσα ακρίβεια – τυχαία συμφωνία) / (1 – τυχαία συμφωνία)

Αυτή η στατιστική παράμετρος είναι μια ένδειξη του βαθμού, κατά τον οποίο οι ποσοστιαίες ορθές τιμές ενός πίνακα σφαλμάτων οφείλονται σε αληθή συμφωνία έναντι τυχαίας συμφωνίας, λαμβάνοντας συνήθως τιμές από 0 έως 1. Στις περιπτώσεις που η τιμή k ισούται με 0 ισχύει ότι μια δεδομένη ταξινόμηση δεν είναι καλύτερη από μια τυχαία καταχώρηση των εικονοστοιχείων. Συναντώνται και περιπτώσεις που η τυχαία συμφωνία είναι αρκετά υψηλή, με αποτέλεσμα η στατιστική παράμετρος να λαμβάνει αρνητικές τιμές, που συνεπάγεται μια κακή ταξινόμηση. Για τον ορισμό της στατιστικής παραμέτρου k-hat χρησιμοποιείται η μαθηματική σχέση :

$$\hat{k} = \frac{N \cdot \Sigma x_{ii} - \sum(x_{i+} \cdot x_{+i})}{N^2 - \Sigma(x_{i+} \cdot x_{+i})} \quad \textbf{(Εξίσωση 1)}$$

Όπου

x_{ii} : ο αριθμός των παρατηρήσεων στη γραμμή i και τη στήλη i (στην κύρια διαγώνιο)

x_{i+} : το σύνολο των παρατηρήσεων στη γραμμή i ,παρουσιάζεται ως περιθωριακό σύνολο στο δεξί μέρος του πίνακα

x_{+i} : το σύνολο των παρατηρήσεων στη στήλη i ,παρουσιάζεται ως περιθωριακό σύνολο στο κάτω μέρος του πίνακα

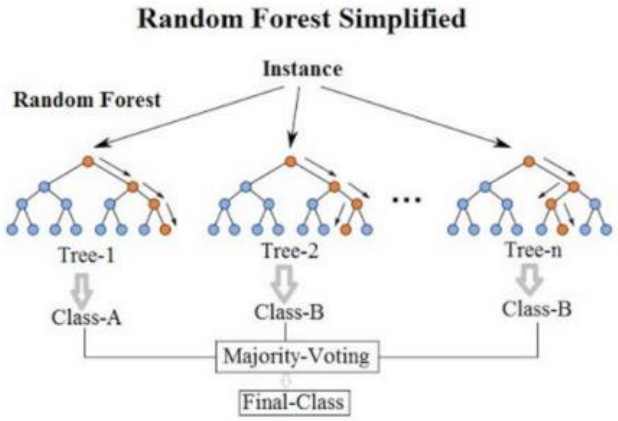
N : ο συνολικός αριθμός των παρατηρήσεων που περιλαμβάνονται στον πίνακα.

3.5.Αλγόριθμοι Επιβλεπόμενης Ταξινόμησης

3.5.1.Αλγόριθμος Random Forest (Random Forest)

Ένας από τους πιο γνωστούς συνδυαστικούς αλγορίθμους ταξινόμησης είναι ο Random Forest (RF), ο οποίος χρησιμοποιεί πολλά, ασυσχέτιστα μεταξύ τους δέντρα αποφάσεων για να πραγματοποιήσει μια εκτίμηση (Belgiu M. and Dragut L. (2016)). Ο αλγόριθμος Random Forests ή Random Decision Forests είναι μια συνολική μέθοδος μάθησης για την ταξινόμηση, την ανάλυση παλινδρόμησης κι άλλες εργασίες, που εκτελούνται βασιζόμενες στη δημιουργία ενός πλήθους δέντρων απόφασης κατά τη διάρκεια της εκπαίδευσης. Η βασική ιδέα πίσω από τον αλγόριθμο αυτό είναι η μείωση της συσχέτισης μεταξύ των ταξινομητών που τον απαρτίζουν και του φαινομένου overfitting που προκαλεί την προσαρμογή του μοντέλου στο θόρυβο που υπάρχει στα δεδομένα.

Από τη διαδικασία της ταξινόμησης, το αποτέλεσμα που προκύπτει με τον αλγόριθμο των Random Forests είναι η κλάση που επιλέγεται από τα περισσότερα δέντρα. Κατά την παλινδρόμηση, επιστρέφεται ο μέσος όρος ή η μέση τιμή της πρόβλεψης των μεμονωμένων δέντρων. Υπάρχει και η μέθοδος των Decision Trees, αλλά φυσικά, τα Random Forests έχουν καλύτερη απόδοση, ενώ ταυτόχρονα σημειώνουν χαμηλότερη ακρίβεια από τα ανώτερης βαθμίδας Random Trees. Ωστόσο, τα χαρακτηριστικά των δεδομένων εισόδου είναι δυνατό να επηρεάσουν το αποτέλεσμα, την απόδοση του αλγορίθμου απόφασης.



Σχήμα 4. Σχεδιάγραμμα υλοποίησης αλγορίθμου Random Forest
https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d

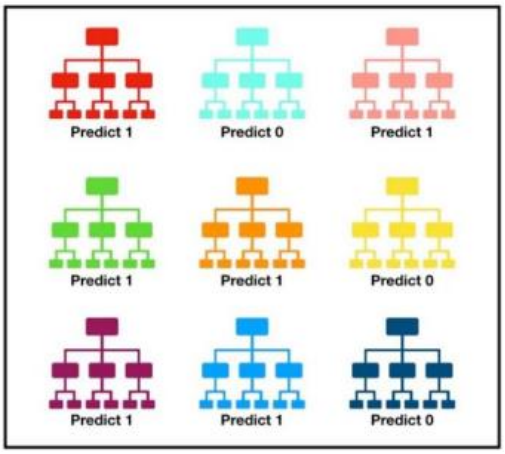
Η εφαρμογή του αλγορίθμου Random Forest στηρίζεται στην ανεξάρτητη ανάπτυξη δέντρων που απαρτίζουν ένα συνδυαστικό μοντέλο και βασίζεται στην επιλογή τυχαίων υποσυνόλων δεδομένων εκπαίδευσης (bagging), με αντικατάσταση. Κατά συνέπεια, μέρος του συνόλου των δειγμάτων που χρησιμοποιούνται ,ώστε να εκπαιδευτεί ένα δέντρο ,υπάρχει η πιθανότητα να επιλεγεί περισσότερες από μια φορές και να εκπαιδεύσει παράλληλα κι άλλα δέντρα μέσα στο σύμπλεγμα κι άλλα δείγματα να μην επιλεχθούν καθόλου. Εν πάση περιπτώσει, τα υποσύνολα δεδομένων εκπαίδευσης που δημιουργούνται είναι διαφορετικά μεταξύ τους και συνεπώς στατιστικά ανεξάρτητα (Rokach, 2010).

Η μέθοδος των Random Forests είναι ένας τρόπος υπολογισμού του μέσου όρου πολλών δέντρων τυχαίας απόφασης, που εκπαιδεύονται σε διαφορετικά μέρη του ίδιου συνόλου εκπαίδευσης, με στόχο τη μείωση της διακύμανσης. Το γεγονός ότι προκύπτει η τελική απόφαση μέσω ενός μέσου όρου δέντρων έχει ως αποτέλεσμα την απώλεια

ερμηνείας, μολονότι γενικά ενισχύεται σημαντικά το τελικό αποτέλεσμα, αφού μειώνεται η διακύμανση των τιμών.

Τα δάση που δημιουργούνται είναι εφικτό να θεωρηθούν ως η συνένωση των επιμέρους δέντρων αποφάσεων που δημιουργούνται κατά τον αντίστοιχο αλγόριθμο. Μάλιστα, ακριβώς επειδή πρόκειται για ένα συνολικό μοντέλο δεδομένων, βελτιώνεται και η απόδοση των επιμέρους δέντρων που αναπτύσσονται μέσα από τα δεδομένα εκπαίδευσης του αλγορίθμου.

Τα τυχαία δάση χρησιμοποιούν έναν τροποποιημένο αλγόριθμο εκμάθησης δέντρων που επιλέγει, σε κάθε υποψήφια διαίρεση στη διαδικασία εκμάθησης, ένα τυχαίο υποσύνολο χαρακτηριστικών. Ο λόγος για να γίνει αυτό είναι ο συσχετισμός των δέντρων σε ένα συνηθισμένο δείγμα εκκίνησης ,όπου εάν ένα ή μερικά χαρακτηριστικά είναι πολύ ισχυροί προγνωστικοί παράγοντες για τη μεταβλητή απόκρισης (στόχος έξοδος), αυτά τα χαρακτηριστικά θα επιλεγούν σε πολλά από τα δέντρα Β, προκαλώντας τα να γίνει συσχετισμένος.



Σχήμα 5. Σχεδιάγραμμα σχηματισμού συνόλου προβλέψεων Τυχαίων Δασών
*https://miro.medium.com/max/526/1*VHDtVaDPNepRglIAv72BFg.jpeg*

Υπάρχει η περίπτωση προσθήκης Extra Trees, τα οποία παρόλο που είναι παρόμοια με τα συνηθισμένα τυχαία δάση, υπάρχουν δύο κύριες διαφορές μιας κι αποτελούν ένα σύνολο μεμονωμένων δέντρων :
(1) κάθε δέντρο εκπαιδεύεται χρησιμοποιώντας ολόκληρο το δείγμα εκμάθησης (κι όχι ένα δείγμα εκκίνησης) και
(2)ο διαχωρισμός από πάνω προς τα κάτω στην εκπαίδευση του δέντρου τυχαιοποιείται.

Αντί να υπολογιστεί το τοπικά βέλτιστο σημείο αποκοπής για κάθε υπό εξέταση χαρακτηριστικό ,επιλέγεται ένα τυχαίο σημείο αποκοπής. Αυτή η τιμή επιλέγεται από μια ομοιόμορφη κατανομή εντός του εμπειρικού εύρους του χαρακτηριστικού (στο σετ εκπαίδευσης του δέντρου). Στη συνέχεια, από όλα τα τυχαία δημιουργούμενα splits, η διάσπαση που αποδίδει την υψηλότερη βαθμολογία επιλέγεται για να χωρίσει τον κόμβο. Παρόμοια με τα συνηθισμένα τυχαία δάση, μπορεί να καθοριστεί ο αριθμός των τυχαία επιλεγμένων χαρακτηριστικών που θα ληφθούν υπόψη σε κάθε κόμβο.

Το θέμα είναι ότι η συσχέτιση μεταξύ των μοντέλων είναι χαμηλή. Αυτό το αποτέλεσμα οφείλεται στο γεγονός ότι τα δέντρα προστατεύουν το ένα το άλλο από τα αντίστοιχα λάθη τους ,αρκεί να μην σφάλλουν πάντα προς την ίδια κατεύθυνση. Ενώ ορισμένα δέντρα μπορεί να είναι λάθος, πολλά άλλα ταξινομούν ορθά, κάνοντας το δέντρο στο

σύνολό του να κινείται προς τη σωστή κατεύθυνση. Επομένως, οι απαιτήσεις για καλή τυχαία απόδοση δασών είναι οι εξής :

Α. Οι συναρτήσεις πρέπει να περιλαμβάνουν πραγματικά σήματα, έτσι ώστε τα μοντέλα που κατασκευάζονται με αυτές τις συναρτήσεις να μπορούν να ξεπεράσουν τις τυχαίες εκτιμήσεις.

Β. Οι προβλέψεις κι επομένως τα λάθη των μεμονωμένων δέντρων πρέπει να παρουσιάζουν χαμηλή συσχέτιση μεταξύ τους.

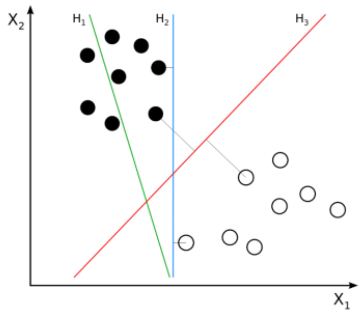
Τα δέντρα αποφάσεων είναι πολύ ευαίσθητα στα δεδομένα στα οποία εκπαιδεύονται, με συνέπεια μικρές αλλαγές στο σετ εκπαίδευσης να έχουν ως πιθανό αποτέλεσμα σημαντικά διαφορετικές δομές δέντρων. Ο αλγόριθμος Random Forest εκμεταλλεύεται αυτή τη συνθήκη, επιτρέποντας σε κάθε μεμονωμένο δέντρο να δειγματοληπτεί τυχαία από το σύνολο δεδομένων με αντικατάσταση, με αποτέλεσμα να δημιουργούνται διαφορετικά δέντρα. Αυτή η διαδικασία είναι γνωστή ως bagging.

Παρόλο που τα τυχαία δάση γενικά επιτυγχάνουν υψηλότερη ακρίβεια από τα μεμονωμένα δέντρα απόφασης, “θυσιάζουν” την εγγενή ερμηνεία των δέντρων αποφάσεων. Τα δέντρα αποφάσεων ανήκουν σε μια σχετικά μικρή οικογένεια μοντέλων μηχανικής μάθησης που είναι εύκολο να ερμηνευτούν μαζί με γραμμικά, βασισμένα σε κανόνες μοντέλα που βασίζονται στην προσοχή. Αυτή η δυνατότητα ερμηνείας είναι μια από τις πιο επιθυμητές ιδιότητες των δέντρων αποφάσεων. Αυτό επιτρέπει στους προγραμματιστές να επιβεβαιώσουν ότι το μοντέλο έχει μάθει αληθείς πληροφορίες από τα δεδομένα και οι τελικοί χρήστες μπορούν να έχουν εμπιστοσύνη στις αποφάσεις που παίρνει το μοντέλο.

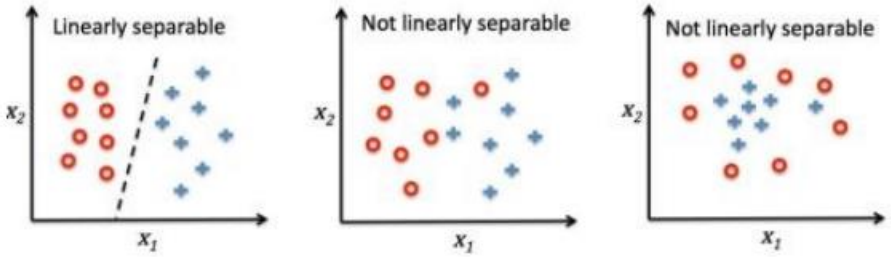
Θεωρητικά, όσο το πλήθος των δέντρων αυξάνεται, τόσο πιο ομαλά είναι τα όρια διαχωρισμού και κατά συνέπεια γίνεται καλύτερη η απόδοση του ταξινομητή. Εμπειρικά, όπως αναφέρουν οι Belgiu M. and Dragut L. (2016) σε έρευνες των Ghosh et al., (2014) και Kulkarni and Sinha, (2012), η ακρίβεια της ταξινόμησης όσον αφορά την παράμετρο του πλήθους των δέντρων, δεν είναι τόσο ευαίσθητη όσο η παράμετρος του πλήθους των μεταβλητών που θα καθορίσουν τα όρια διαχωρισμού σε κάθε εσωτερικό κόμβο.

3.5.2.Αλγόριθμος Support Vector Machine (Support Vector Machine)

Μια πολύ δημοφιλής τεχνική που ανήκει στις μεθόδους μηχανικής μάθησης είναι ο αλγόριθμος Support Vector Machine (SVM). Σαν ιδέα, οι Μηχανές Διανυσματικής Υποστήριξης αναπτύχτηκαν από τους Cortes και Vapnik (1995), με τη μέθοδο αυτή να βασίζεται στη στατιστική θεωρία της μηχανικής μάθησης και να χρησιμοποιείται για την πρόβλεψη μελλοντικών δεδομένων. Εκπαιδεύεται από την επίλυση ενός περιορισμένου προβλήματος ταξινόμησης κι υλοποιεί τη χαρτογράφηση των συντελεστών παραγωγής σε ένα υψηλό τρισδιάστατο χώρο αξιοποιώντας ένα σύνολο γραμμικών και μη γραμμικών βασικών συναρτήσεων.



Σχήμα 6. Γραμμικός διαχωρισμός κατηγοριών σε δισδιάστατο χώρο με χρήση 3 υπερεπιπέδων



Σχήμα 7. Παράδειγμα Γραμμικά και Μη Γραμμικά Διαχωρισμένων κλάσεων

Ο αλγόριθμος SVM προσφέρει τη δυνατότητα να χρησιμοποιηθεί για ποικίλες αναπαραστάσεις ,όπως τα νευρωνικά δίκτυα, τα splines, τους πολυγωνικούς εκτιμητές, αλλά υπάρχει μια μοναδική βέλτιστη λύση για κάθε επιλογή των SVM παραμέτρων, το οποίο είναι διαφορετικό σε άλλες μηχανές μάθησης, όπως τα τυποποιημένα Νευρωνικά Δίκτυα που χρησιμοποιούν την προς τα πίσω διάδοση. Εν ολίγοις, η ανάπτυξη της μεθόδου SVM διαφέρει σε σημαντικό βαθμό από τους συνήθεις αλγορίθμους που χρησιμοποιούνται στη μηχανική μάθηση.

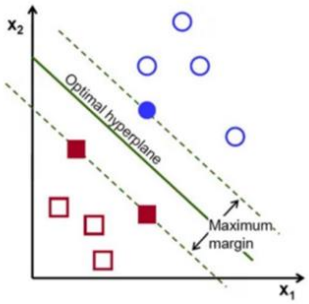
Οι Μηχανές Διανυσματικής Υποστήριξης είναι μια μέθοδος μάθησης με πλήρη επίβλεψη. Ένα βήμα για τη μέθοδο SVM είναι η αναγνώριση, κάτι το οποίο είναι άρρηκτα συνδεδεμένο με τις γνωστές κατηγορίες. Με αυτόν τον τρόπο ο συγκεκριμένος αλγόριθμος είναι σε θέση να χρησιμοποιηθεί για να προσδιορίσει τα βασικά σύνολα που εμπλέκονται στις διεργασίες για διάκριση μεταξύ των κλάσεων.

Ο αλγόριθμος ταξινόμησης SVM στηρίζεται στην αναπαράσταση των δεδομένων εισόδου στο χώρο, δημιουργώντας ένα κατάλληλο υπερεπίπεδο ,με τον αλγόριθμο να διαχωρίζει τα δεδομένα στις κλάσεις εξόδου. Το υπερεπίπεδο πρέπει να διαχωρίζει τις κλάσεις με τη μεγαλύτερη δυνατή ακρίβεια, ενώ ταυτόχρονα μεγιστοποιεί την απόσταση των δεδομένων εισόδου κάθε κλάσης ως προς αυτό. Οι κατηγορίες εξόδου μπορεί να είναι γραμμικά ή μη γραμμικά διαχωρισμένες ,αναλόγως τα δεδομένα εισόδου. Στην περίπτωση που τα δεδομένα δεν είναι γραμμικά διαχωρισμένα,

εφαρμόζονται κατάλληλοι μετασχηματισμοί ,με τη χρήση συγκεκριμένων συναρτήσεων πυρήνα (kernel functions), με στόχο τη μεταφορά των δεδομένων σε ένα νέο χώρο, άλλων διαστάσεων, με την επιλογή της συνάρτησης πυρήνα να είναι αυτή που καθορίζει άμεσα το αποτέλεσμα ενός SVM.

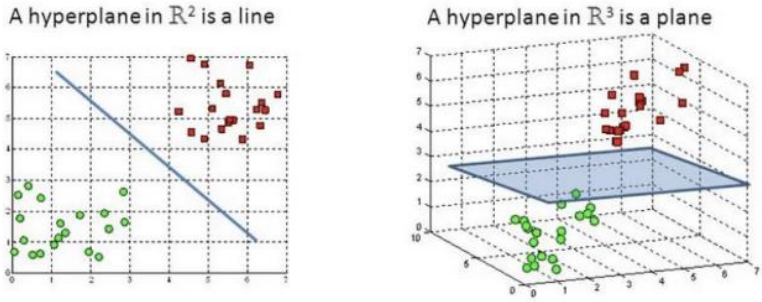
Όπως έχει ήδη αναφερθεί, η βασική λειτουργία της μεθόδου είναι ο διαχωρισμός των δεδομένων εισόδου από κάποιο βέλτιστο υπερεπίπεδο, με τα κοντινότερα σημεία σε αυτό το υπερεπίπεδο να ονομάζονται Διανύσματα Υποστήριξης (Support Vectors), τα οποία φέρουν τη μεγαλύτερη πρόκληση ως προς την ταξινόμησή τους. Έχοντας ως κριτήριο τη μεγιστοποίηση της απόστασής τους από το υπερεπίπεδο ,επηρεάζουν άμεσα τη βέλτιστη τοποθεσία του υπερεπιπέδου.

Αξίζει να σημειωθεί πως ένα σημαντικό χαρακτηριστικό της μεθόδου SVM είναι πως δεν υπάρχει πρόβλημα στην κατηγοριοποίηση δεδομένων πολλών διαστάσεων. Ουσιαστικά, ο στόχος του αλγορίθμου Support Vector Machine (SVM) είναι να βρει ένα υπερεπίπεδο σε ένα χώρο N-διαστάσεων που θα ταξινομεί ευδιάκριτα τα σημεία δεδομένων. Για να διαχωριστούν όμως, οι δυο κατηγορίες σημείων δεδομένων, υπάρχουν πολλά πιθανά υπερεπίπεδα που θα μπορούσαν να επιλεγούν. Αυτό που επιδιώκεται είναι να βρεθεί ένα επίπεδο που έχει το μέγιστο περιθώριο, δηλαδή τη μέγιστη απόσταση μεταξύ των σημείων δεδομένων και των δυο κατηγοριών. Η μεγιστοποίηση της απόστασης περιθωρίου παρέχει κάποια ενίσχυση, έτσι ώστε τα μελλοντικά σημεία δεδομένων να μπορούν να ταξινομηθούν με μεγαλύτερη σιγουριά.



Σχήμα 8. Γραφική απεικόνιση λειτουργίας μεγιστοποίησης απόστασης για το διαχωρισμό κλάσεων
https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Τα υπερεπίπεδα είναι όρια απόφασης που βοηθούν στην ταξινόμηση των σημείων δεδομένων. Τα σημεία δεδομένων που εμπίπτουν σε κάθε πλευρά του υπερεπιπέδου μπορούν να αποδοθούν σε διαφορετικές κατηγορίες. Επίσης, η διάσταση του υπερεπιπέδου εξαρτάται από τον αριθμό των χαρακτηριστικών. Εάν ο αριθμός των χαρακτηριστικών εισόδου είναι 2, τότε το υπερεπίπεδο είναι απλώς μια γραμμή. Εάν ο αριθμός των χαρακτηριστικών εισόδου είναι 3, τότε το υπερεπίπεδο γίνεται δισδιάστατο επίπεδο.

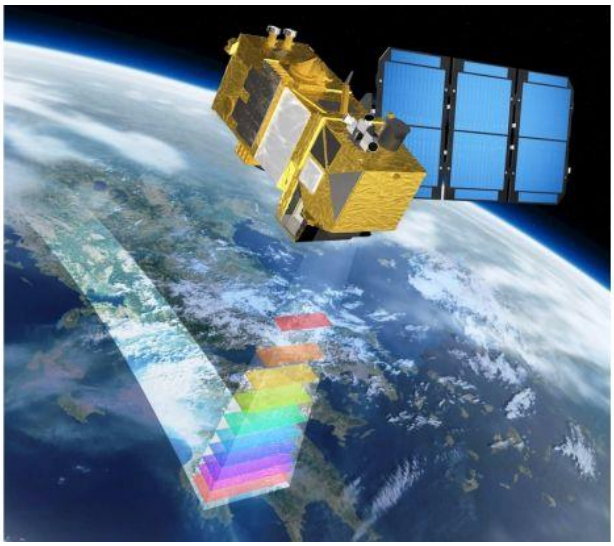


Σχήμα 9. Γραφική απεικόνιση των υπερεπιπέδων σε χώρους διαφορετικών διατάσεων
https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

ΚΕΦΑΛΑΙΟ 4. Περιοχή Μελέτης και Δεδομένα

4.1.Ο δορυφόρος Sentinel-2

Η αποστολή Copernicus SENTINEL-2 περιλαμβάνει έναν αστερισμό δύο δορυφόρων, τους δίδυμους δορυφόρους Sentinel-2A και Sentinel-2B ,δύο πανομοιότυποι δορυφόροι, που αναπτύχθηκαν από τον Ευρωπαϊκό Οργανισμό Διαστήματος (ESA), για την εκτέλεση επίγειων παρατηρήσεων, με σκοπό τη χρήση τους σε εφαρμογές, όπως η ανίχνευση αλλαγών χρήσεων γης και η διαχείριση φυσικών καταστροφών, και κινούνται στην ίδια τροχιά, αλλά με διαφορά φάσης 180^{o}. Επειδή το σύστημα των δυο δορυφόρων είναι παθητικό κι απαιτεί την ακτινοβολία του ήλιου για να λειτουργήσει, θεωρείται ηλιοσύγχρονο. Ο δορυφόρος με σκοπό την καταγραφή ολόκληρης της γης χρειάζεται να διαγράψει 143 τροχιές, όπου η περίοδος της μιας τροχιάς είναι 100.6 λεπτά, με το δορυφόρο να βρίσκεται σε μέσο υψόμετρο 786 χιλιόμετρα και με τη χρονική διακριτική ικανότητα να είναι 5 ημέρες με τη λειτουργία και των 2 δορυφόρων. Αξίζει να σημειωθεί ότι η γωνία μεταξύ του τροχιακού επιπέδου και του Ισημερινού είναι 98.62^{o}.



Εικόνα 1. Αποστολή Copernicus Sentinel-2
https://labo.obs-mip.fr/multitemp/directional-effects-what-field-of-view-for-the-next-generation-of-sentinel-2/

Οι δορυφόροι Sentinel 2 διαθέτουν 13 κανάλια με τους δύο δορυφόρους, Sentinel-2A και Sentinel-2B να εμφανίζουν διαφορετικά κεντρικά μήκη κύματος, εύρη τιμών και χωρική ανάλυση. Τα κανάλια των Sentinel 2 είναι κανάλι 1-Coastal aerosol, κανάλι 2-Blue, κανάλι 3-πράσινο, κανάλι 4-κόκκινο, τα κανάλια 5,6 και 7-Vegetation red edge που αποδίδουν κόκκινη τη βλάστηση, κανάλι 8-NIR (υπέρυθρο), κανάλι 8A-Narrow NIR (εγγύς υπέρυθρο), κανάλι 9-Water vapour ,εντοπισμός υδάτινων στοιχείων, κανάλι 10-SWIR-Cirrus και τα κανάλια 11,12- SWIR. Τα κανάλια αυτά εμφανίζουν διαφορετικές τιμές ως προς τη χωρική ανάλυσή τους, όπως φαίνεται και στον παρακάτω πίνακα (Πίνακας 1) ,με τα κανάλια 1,9 και 10 να έχουν χωρική ανάλυση των 60 μέτρων, τα κανάλια 2,3,4 και 8 διαθέτουν χωρική ανάλυση 10 μέτρα και τα κανάλια 5,6,7,8-A,11 και 12 20 μέτρα. Ο λόγος για τον οποίο χρησιμοποιήθηκαν 13 φασματικά κανάλια είναι πως ο σχεδιασμός του πολυφασματικού οργάνου του Sentinel 2 βασίστηκε στην ανάγκη για μεγάλο εύρος πληροφορίας, υψηλής γεωμετρικής και φασματικής ακρίβειας.

	Sentinel-2A		Sentinel-2B		
Bands	Κεντρικό μήκος κύματος (nm)	Εύρος μήκους κύματος	Κεντρικό μήκος κύματος (nm)	Εύρος μήκους κύματος	Χωρική ανάλυση (m)
B1-Coastal aerosol	443.9	27	442.3	45	60
B2-Blue	496.6	98	492.1	98	10
B3-Green	560.0	45	559.0	46	10
B4-Red	664.5	38	665.0	39	10
B5-Vegetation Red Edge	703.9	19	703.8	20	20
B6-Vegetation Red Edge	740.2	18	739.1	18	20
B7-Vegetation Red Edge	782.5	28	779.7	28	20
B8-NIR	835.1	145	833.0	133	10
B8A-Narrow NIR	864.8	33	864.0	32	20
B9-Water vapour	945.0	26	943.2	27	60
B10-SWIR Cirrus	1373.5	75	1376.9	76	60
B11-SWIR	1613.7	143	1610.4	141	20
B12-SWIR	2202.4	242	2185.7	238	20

Πίνακας 1. Κανάλια (bands) των Sentinel-2 δορυφόρων και τα χαρακτηριστικά τους

Όλα τα παραπάνω χαρακτηριστικά γνωρίσματα των δίδυμων δορυφόρων Sentinel-2 κι ιδιαίτερα η χωρική ανάλυση που διαθέτουν τα κανάλια τους, οδηγούν στην επιλογή πολυφασματικής απεικόνισης που προέρχεται από αυτούς.

4.2.Επιλογή ατμοσφαιρικά διορθωμένης πολυφασματικής απεικόνισης

Το αντικείμενο της παρούσας διπλωματικής εργασίας στηρίζεται σε μεγάλο βαθμό στην ποιότητα της πολυφασματικής απεικόνισης, που αποτελεί το βασικό δεδομένο για την εφαρμογή των συνολικών πειραμάτων. Οι πολυφασματικές απεικονίσεις στο ορατό και στο υπέρυθρο μήκος κύματος του ηλεκτρομαγνητικού φάσματος επηρεάζονται από τα σωματίδια και τα αέρια της ατμόσφαιρας μέσω των διεργασιών της σκέδασης και της απορρόφησης. Ακολουθεί λοιπόν η εφαρμογή ενός μοντέλου διόρθωσης, που διορθώνει τις επιπτώσεις στην εικόνα όπως το θόλωμα και οι αλλαγές στις ραδιομετρικές τιμές. Αυτός είναι και ο λόγος για τον οποίο επιδιώκεται η λήψη πολυφασματικής απεικόνισης ,η οποία είναι ατμοσφαιρικά διορθωμένη, καθώς σε περιπτώσεις που κατά την προεπεξεργασία της απεικόνισης απαιτείται η εφαρμογή ατμοσφαιρικών διορθώσεων, αυτές πρέπει να γίνονται προσεκτικά, ανάλογα με το αντικείμενο και το στόχο της εφαρμογής.

Η χρήση ατμοσφαιρικά διορθωμένων απεικονίσεων σε εργασίες που σχετίζονται με την ανάλυση πολυφασματικών δεδομένων είναι απαραίτητη για αρκετούς λόγους:

- Οι αλγόριθμοι ταξινόμησης και ανάλυσης εικόνας βασίζονται σε μεγάλο βαθμό στις φασματικές υπογραφές των αντικειμένων. Η ατμοσφαιρική διόρθωση εξασφαλίζει ότι οι φασματικές υπογραφές είναι αντιπροσωπευτικές των πραγματικών χαρακτηριστικών των αντικειμένων στην επιφάνεια της γης και

δεν επηρεάζονται από ατμοσφαιρικά φαινόμενα, βελτιώνοντας έτσι την αξιοπιστία των ταξινομήσεων.

- Η αφαίρεση των ατμοσφαιρικών επιδράσεων μπορεί να ενισχύσει την φασματική διακριτική ικανότητα, επιτρέποντας καλύτερη διάκριση μεταξύ διαφορετικών κατηγοριών κάλυψης γης. Αυτό είναι ιδιαίτερα σημαντικό για την ανίχνευση ακόμη και μικρών φασματικών διαφορών μεταξύ αντικειμένων.
- Οι ατμοσφαιρικές επιδράσεις εισάγουν σφάλμα και αβεβαιότητα στις μετρήσεις. Με την ατμοσφαιρική διόρθωση, μειώνεται το σφάλμα και αυξάνεται η αξιοπιστία των δεδομένων, γεγονός που οδηγεί σε πιο αξιόπιστα αποτελέσματα και συμπεράσματα.
- Οι ατμοσφαιρικά διορθωμένες απεικονίσεις βελτιώνουν την ακρίβεια και την ποιότητα των δεδομένων, απομακρύνοντας τις επιδράσεις της ατμόσφαιρας όπως η σκέδαση και η απορρόφηση. Αυτό εξασφαλίζει ότι οι παρατηρήσεις αντανακλούν πιο πιστά τις πραγματικές ιδιότητες των αντικειμένων στην επιφάνεια της Γης.

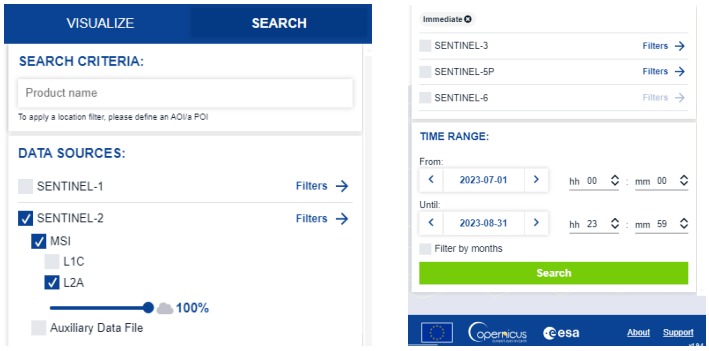
Υπάρχουν βέβαια κι άλλοι λόγοι για την ανάγκη ατμοσφαιρικών διορθώσεων, αλλά αυτοί είναι οι κυριότεροι που επηρεάζουν την πορεία αυτής της εργασίας. Χρησιμοποιώντας ατμοσφαιρικά διορθωμένες απεικονίσεις Sentinel-2 L2A, εξασφαλίζεται ότι τα δεδομένα που χρησιμοποιούνται είναι όσο το δυνατόν πιο ακριβή κι αξιόπιστα, επιτρέποντας την πραγματοποίηση πιο έγκυρων και αξιόπιστων αναλύσεων και συμπερασμάτων.

Αξίζει να αναφερθεί ότι οι εικόνες από τον δορυφόρο Sentinel-2 διατίθενται σε διάφορα επίπεδα επεξεργασίας, με τα δύο πιο συνηθισμένα να είναι τα επίπεδα L1C και L2A. Πιο συγκεκριμένα, οι εικόνες L1C είναι γεωμετρικά διορθωμένες. Περιλαμβάνουν διόρθωση για παραμορφώσεις λόγω της θέσης του δορυφόρου και του αισθητήρα, αλλά όχι για ατμοσφαιρικές επιδράσεις. Παρέχουν δεδομένα σε 13 φασματικά κανάλια (bands), που καλύπτουν από το ορατό έως το βραχύ υπέρυθρο (SWIR) φάσμα. Διαθέτουν χωρική ανάλυση 10m, 20m και 60m, ανάλογα με το φασματικό κανάλι και προσφέρουν παγκόσμια κάλυψη κάθε 5 ημέρες (στην περιοχή του ισημερινού) από τους δύο δορυφόρους Sentinel-2A και Sentinel-2B. Από την άλλη, οι εικόνες L2A είναι επεξεργασμένες για ατμοσφαιρικές διορθώσεις (δηλαδή διορθώσεις για ατμοσφαιρικές επιδράσεις όπως σκέδασης και απορρόφησης) για να παρέχουν τιμές ανακλαστικότητας επιφάνειας (bottom-of-atmosphere reflectance). Αυτό βελτιώνει την ακρίβεια και την ποιότητα των φασματικών υπογραφών των επιφανειακών χαρακτηριστικών. Διαθέτουν δεδομένα σε 12 φασματικά κανάλια (λείπει το κανάλι B9 - Water vapour (Ατμοσφαιρική υγρασία) στα 945 nm), με τις ίδιες χωρικές αναλύσεις όπως οι εικόνες L1C.

4.3.Αναζήτηση & λήψη πολυφασματικής απεικόνισης

Η διαδικασία αναζήτησης πολυφασματικής απεικόνισης πραγματοποιείται στο περιβάλλον της σελίδας scihub ,που λειτουργεί ως browser της Copernicus, στο περιβάλλον του οποίου ορίζονται οι απαραίτητες παράμετροι ώστε να γίνει η τελική επιλογή και λήψη των πολυφασματικών Sentinel-2 απεικονίσεων.

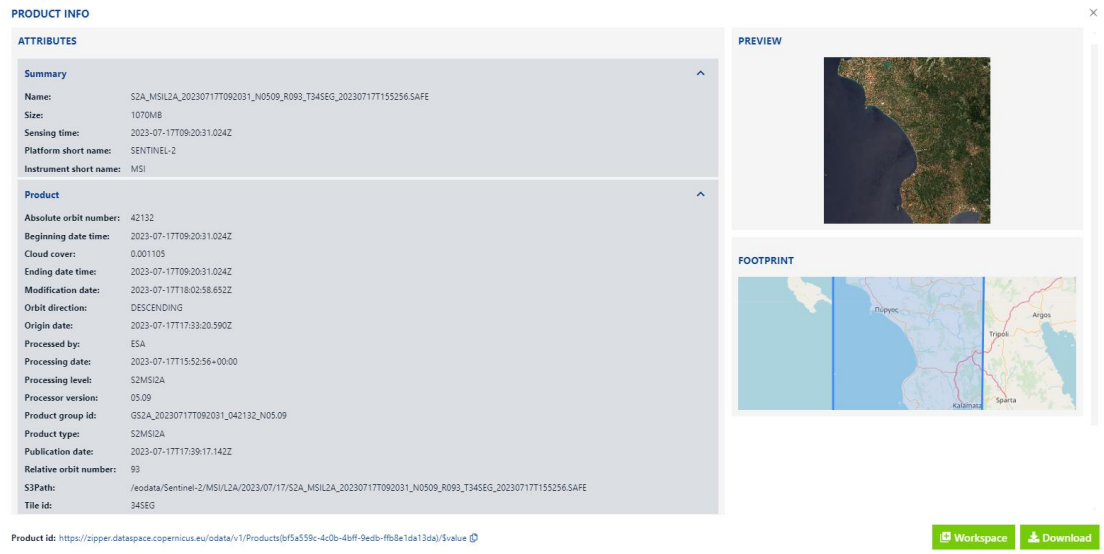
Αρχικά, επιλέγεται η περίοδος στην οποία πρέπει να έχει πραγματοποιηθεί η λήψη της απεικόνισης ,όπου λόγω του γεγονότος ότι είναι επιθυμητό να υπάρχει σταθερότητα στις κλιματικές συνθήκες που επικρατούν, αλλά και μικρό ποσοστό νεφοκάλυψης, το χρονικό διάστημα είναι επιθυμητό να βρίσκεται μεταξύ των μηνών Ιούλιο με Σεπτέμβριο. Αναφορικά με το ποσοστό νεφοκάλυψης στην απεικόνιση, αποδεκτό ορίζεται εάν είναι της τάξης 0-3%. Επίσης, ζητούμενο είναι να είναι ατμοσφαιρικά διορθωμένη η απεικόνιση, οπότε επιλέγεται ο τύπος του προϊόντος S2MSIL2A.



Εικόνα 2. Κριτήρια αναζήτησης πολυφασματικής απεικόνισης Sentinel-2

Εκτός αυτών σχηματίζεται ένα πολύγωνο, εντός του οποίου είναι επιθυμητό να υπάρχει πολυφασματική απεικόνιση, με σκοπό σε επόμενο βήμα να σχηματιστεί και η περιοχή μελέτης. Εκτελώντας την αναζήτηση, αποτυπώνονται επί του χάρτη πλήθος πολυγώνων που αποδίδουν τις απεικονίσεις και τις τροχιές αυτών, εντός των οποίων υπάρχει τμήμα του πολυγώνου που έχει σχηματιστεί από το χρήστη, δίνοντάς του τη δυνατότητα να επιλέξει την τροχιά που θεωρεί πως έχει τις πιο αντιπροσωπευτικές απεικονίσεις και θα αποτυπώνουν πλήρως την περιοχή μελέτης.

Λαμβάνοντας όλα τα παραπάνω υπόψιν, η πολυφασματική Sentinel-2 απεικόνιση που επιλέγεται, αποδίδει την παρακάτω περιοχή κι έχει τα εξής χαρακτηριστικά :



Εικόνα 3. Περιοχή κάλυψης και χαρακτηριστικά πολυφασματικής απεικόνισης

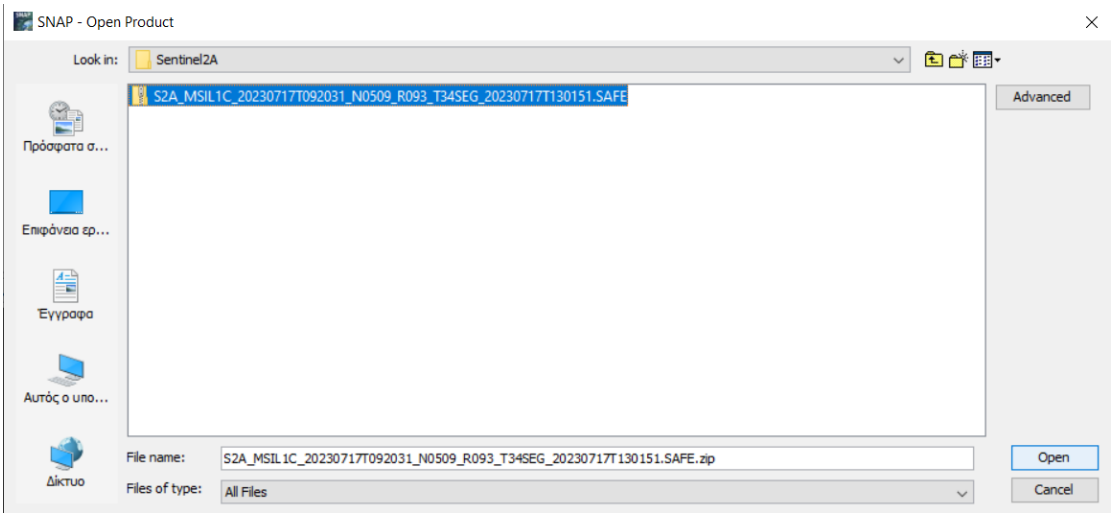
Πιο συγκεκριμένα, όπως προκύπτει τόσο από το όνομα της απεικόνισης :
S2A_MSIL2A_20230717T092031_N0509_R093_T34SEG_20230717T155256.SAFE
αλλά κι από την Εικόνα 3 ,η πολυφασματική απεικόνιση έχει τα εξής χαρακτηριστικά

Ανάλυση χαρακτηριστικών πολυφασματικής απεικόνισης :

Satellite name : Sentinel-2
Satellite number : A
Instrument abbreviation : MSI
Processing level : S2MSI2A
Sensing start: 2023-07-17 T09:20:31
Sensing stop: 2023-07-17 T09:20:31
Processing baseline : 05:09
Relative orbit (start) : 093
Tile Identifier: 34SEG
Processing date : 2023-07-17 T15:52:56

4.4.Προ-επεξεργασία απεικόνισης στο SNAP

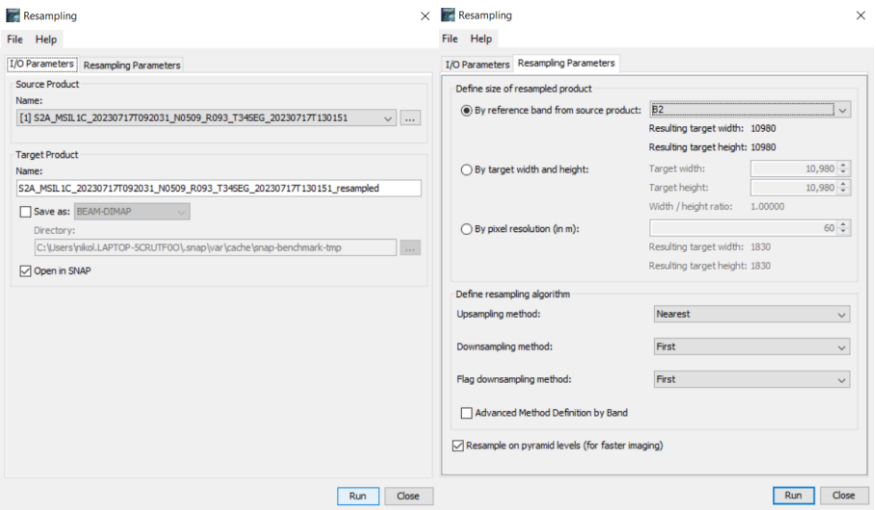
Η απεικόνιση εισάγεται ως αρχείο zip στο περιβάλλον του λογισμικού Sentinel Application Platform (SNAP), ώστε να γίνει η περικοπή της και να επιλεγεί η τελική περιοχή ενδιαφέροντος και παράλληλα να επιλεγούν μόνο τα απαραίτητα για τις εφαρμογές ταξινόμησης κανάλια.



Σχήμα 10. Εισαγωγή απεικόνισης στο λογισμικό SNAP

Αφού εισαχθεί η απεικόνιση στο λογισμικό, ακολουθεί η διαδικασία της αναδόμησης κατά την οποία όλα τα κανάλια της εικόνας φτάνουν στη βέλτιστη χωρική ανάλυση, αρκεί να γίνει χρήση ενός δεκάμετρου καναλιού, όπως είναι τα bands 2,3,4,8, τα οποία διαθέτουν καλύτερη γεωμετρική ανάλυση. Σε αυτή την περίπτωση επιλέγεται το δεύτερο κανάλι της απεικόνισης, B2 μπλε, το οποίο διαθέτει χωρική ανάλυση 10 μέτρων, ενώ ως upsampling μέθοδος επιλέγεται η μέθοδος του πλησιέστερου γείτονα που προτιμάται γενικότερα στις εφαρμογές της τηλεπισκόπησης, με στόχο τον ευκολότερο εντοπισμό των περιγραμμάτων διαφόρων θεματικών κατηγοριών και των μεταβολών που αυτές παρουσιάζουν.

Για τη διαδικασία της αναδόμησης επιλέγεται Raster>Geometric>Resampling, ορίζοντας κάθε φορά ως reference band το B2 και upsampling method αυτή του πλησιέστερου γείτονα, Nearest Neighbor.



Σχήμα 11. Παράμετροι για την υλοποίηση της διαδικασίας αναδόμησης της Sentinel-2A απεικόνισης

Ουσιαστικά, με την αναδόμηση ένα προϊόν πολλών διαστάσεων, με τις διαστάσεις σε αυτή την περίπτωση να αναφέρονται στο πλήθος των bands, με διαφορετικά μεγέθη ή/κι ανάλυση, να μετατρέπεται σε μονοδιάστατο προϊόν. Αυτό γίνεται κυρίως για το πως διαχειρίζονται διάφορες άλλες διαδικασίες επεξεργασίας τις απεικονίσεις και για να οριστεί το band από το οποίο θα αντλείται η πληροφορία.

Ολοκληρώνοντας την αναδόμηση, υπάρχει η δυνατότητα η πολυφασματική απεικόνιση να προβληθεί σε RGB:432, φυσικό έγχρωμο σύνθετο, για να μελετηθεί εκτενέστερα και να εντοπιστεί η περιοχή ενδιαφέροντος, προτού υλοποιηθεί η περικοπή της.

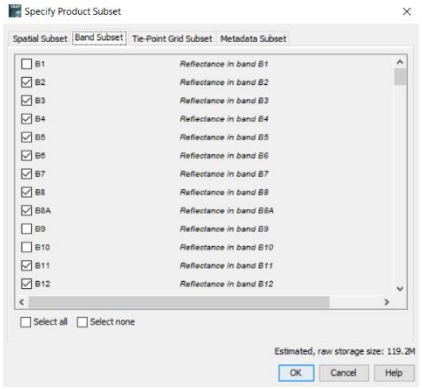
Σχήμα 12. Έγχρωμο σύνθετο RGB:432 ,Sentinel-2A πολυφασματικής απεικόνισης, με ημερομηνία λήψης 17/07/2023, δυτικό τμήμα Πελοποννήσου

4.4.1.Ορισμός περιοχής ενδιαφέροντος και καναλιών (bands) απεικόνισης

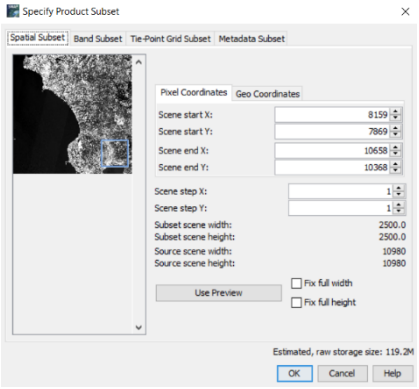
Για την περικοπή των απεικονίσεων επιλέγεται Raster>Subset κι ακολουθεί ο σχηματισμός του παραλληλογράμμου που περιέχει την περιοχή που είναι επιθυμητό να μελετηθεί, χωρίς περιορισμό στην έκταση, ορίζοντας παράλληλα και τα bands που θα περιλαμβάνει η τελική απεικόνιση.

Αξίζει σε αυτό το σημείο να γίνει αναφορά στα κανάλια που θα περιλαμβάνει η τελική απεικόνιση, τα οποία ορίζονται στο στάδιο προσδιορισμού της περιοχής μελέτης (εργαλείο Subset-Σχήμα 14). Εστιάζοντας στα πολλαπλά φασματικά δεδομένα που προσφέρει ο δορυφόρος Sentinel-2 ,διαθέτοντας 13 κανάλια -12 κανάλια στην περίπτωση Level-2A προϊόντος-, η χωρική ανάλυση των καναλιών, καθώς και οι ανάγκες ως προς τη φασματική πληροφορία για τη διεξαγωγή των πειραμάτων που ακολουθούν, είναι οι παράγοντες που επηρεάζουν την επιλογή καναλιών που θα διαθέτει η τελική απεικόνιση. Λόγω του γεγονότος πως τα πειράματα βασίζονται κυρίως στη συλλογή των δεδομένων εκπαίδευσης κι ελέγχου, για να αξιολογηθεί η ακρίβεια που επιτυγχάνεται, απαιτείται υψηλή χωρική ανάλυση, που διευκολύνει τη συλλογή των δεδομένων, με συνέπεια να επιλέγονται όλα τα κανάλια που διαθέτουν την υψηλότερη

χωρική ανάλυση των 10 μέτρων, με την ταυτόχρονη χρήση των καναλιών με την αμέσως χαμηλότερη χωρική ανάλυση ,20 μέτρα, καθώς είναι αναγκαία και η πληροφορία που αφορά το μικροκυματικό τμήμα του φάσματος. Όπως φαίνεται από το Σχήμα 13 κι υπενθυμίζεται ανατρέχοντας στον Πίνακας 1, έχει αποφευχθεί η χρήση των καναλιών με χωρική ανάλυση 60 μέτρα, ώστε να διατηρηθεί η ευκρίνεια της απεικόνισης.



Σχήμα 13. Επιλογή bands τελικής απεικόνισης



Σχήμα 14. Ορισμός περιοχής ενδιαφέροντος

Η τελική απεικόνιση είναι

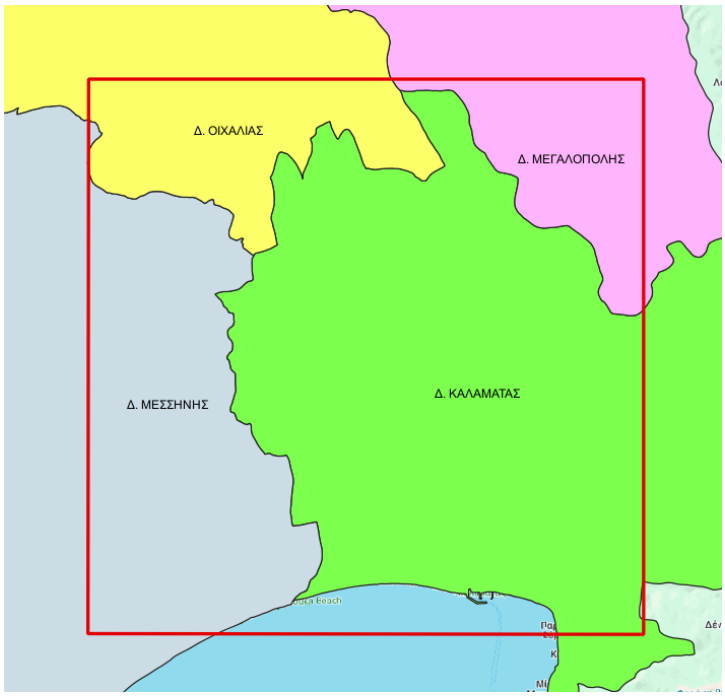
Σχήμα 15. Έγχρωμο σύνθετο RGB:432, Sentinel-2A απεικόνισης, με ημερομηνία λήψης 17/07/2023

Έχοντας διαμορφώσει την τελική απεικόνιση της περιοχής ενδιαφέροντος όπου θα εφαρμοσθούν πειράματα υπερδειγματοληψίας κι υποδειγματοληψίας σε μεθόδους ταξινόμησης με απώτερο σκοπό την αύξηση ακρίβειας του τελικού προϊόντος, η επεξεργασία των εικόνων στο λογισμικό SNAP ολοκληρώνεται με την αποθήκευσή της πολυφασματικής εικόνας σε μορφή GeoTIFF, ώστε να είναι δυνατή η εισαγωγή της στο περιβάλλον QGIS, όπου θα υλοποιηθούν όλα τα απαραίτητα στάδια για τη διεξαγωγή των πειραμάτων.

4.5.Γεωγραφική θέση περιοχής μελέτης

Η περιοχή μελέτης που έχει επιλεγεί για την εκπόνηση της παρούσας διπλωματικής εργασίας εντοπίζεται στην Περιφέρεια Πελοποννήσου και πιο συγκεκριμένα στην Περιφερειακή Ενότητα Μεσσηνίας κι ένα μικρό τμήμα της Περιφερειακής Ενότητας Αρκαδίας, περιλαμβάνοντας τμήματα των Δήμων Καλαμάτας, Μεσσήνης κι Οιχαλίας από την πρώτη Π.Ε. και του Δήμου Μεγαλόπολης από τη δεύτερη.

Σχήμα 16. Θέση περιοχής μελέτης στο Google Maps



Σχήμα 17. Δήμοι περιοχής μελέτης

4.6.Διοικητική διαίρεση

Η περιοχή ενδιαφέροντος περιλαμβάνει στη μεγαλύτερή της έκταση το Δήμο Καλαμάτας, με έδρα του την πόλη της Καλαμάτας και συνολική έκταση 440.3 km^2 ,με την πόλη αυτή μάλιστα να συνιστά την πρωτεύουσα της Περιφερειακής Ενότητας Μεσσηνίας, στη συνέχεια το Δήμο Μεσσήνης, που σημειώνει συνολική έκταση 563.7 km^2 κι έδρα του την κωμόπολη της Μεσσήνης, το Δήμο Οιχαλίας με έδρα του την κωμόπολη του Μελιγαλά, με συνολική έκταση 411.4 km^2 και τέλος, ο Δήμος Μεγαλόπολης, συνολικής έκτασης 727,33 km^2 με έδρα την κωμόπολη Μεγαλόπολη.

Σχήμα 18. Όρια δήμων επί της πολυφασματικής απεικόνισης

Η Μεσσηνία αποτελείται περιμετρικά από ομαλά και χαμηλά βουνά, με το κέντρο της να περιγράφεται από την εύφορη πεδιάδα του Παμίσου ποταμού. Στην ανατολική πτέρυγα της Π.Ε. Μεσσηνίας βρίσκεται ο ορεινός όγκος του Ταϋγέτου, με υψηλότερες κορυφές τον Προφήτη Ηλία (2407 μ.), τη Νεραϊδοβούνα (2025 μ.) και την Ξεροβούνα (1852 μ.) στα όρια με τη Λακωνία και χαμηλότερες στα σύνορα με την Αρκαδία, όπως το Ξεροβούνι (1521 μ.) και ο Προφήτης Ηλίας (1389 μ.). Το βορειοανατολικό τμήμα της Μεσσηνίας έχει ως φυσικό σύνορο με την Αρκαδία το Λύκαιο όρος (1421 μ.).

Ο ποταμός Πάμισος ,μήκους 43 χιλιομέτρων, διαρρέει τη μεσσηνιακή πεδιάδα, ο οποίος πηγάζει από τον Άγιο Φλώρο στις υπώρειες του Ταϋγέτου, συλλέγει τα νερά των χειμάρρων από την Ιθώμη και το βόρειο τμήμα του Ταϋγέτου κι εκβάλλει κοντά στη Μεσσήνη.

Από την άλλη, στο Δήμο Μεγαλόπολης αναπτύσσεται το υδρογραφικό δίκτυο του ποταμού Νέδωντα που αποστραγγίζει τον κύριο όγκο των κατακρημνισμάτων που δέχεται η περιοχή. Ο Νέδων πηγάζει από τα όρη της Αλαγονίας στον Ταΰγετο, διατρέχει την πόλη της Καλαμάτας κι εκβάλλει στο Μεσσηνιακό κόλπο.

Η Μεσσηνία βρέχεται από το Ιόνιο Πέλαγος, διαμορφώνοντας κόλπους, το Μεσσηνιακό και τον Κυπαρισσιακό, ενώ εντοπίζονται και πολλές παραλίες, οργανωμένες και μη. Μεγάλη πληθώρα παραλιών σημειώνεται στη Μεσσηνιακή Μάνη και στην περιοχή της

Πύλου, με το Δήμο Καλαμάτας να σημειώνει μικρό αριθμό, και το Δήμο Οιχαλίας να αντιπροσωπεύει την ορεινή Μεσσηνία, χωρίς να σημειώνει ύπαρξη παραλιών, παρά μόνο ποταμών και καταρρακτών.

Αξιοσημείωτες και γνωστές στο ευρύ κοινό είναι οι παραλίες :

Δήμος Καλαμάτας Δυτική Καλαμάτα, Μικρή Μαντίνεια, Ανατολική Καλαμάτα – Βέργα
Δήμος Μεσσήνης Μπούκα, Ανάληψη, Πεταλίδι, Επισκοπή

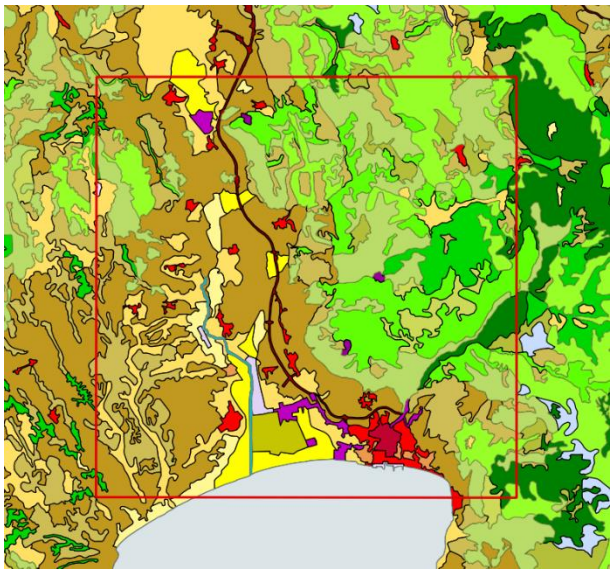
4.7.Δημογραφικά χαρακτηριστικά

Σύμφωνα με την τελευταία απογραφή πληθυσμού, από την ΕΛΣΤΑΤ το 2021, ο Δήμος Καλαμάτας έχει 71.894 κατοίκους, αποτελώντας το μόνο δήμο της Π.Ε. Μεσσηνίας με αύξηση πληθυσμού από την προηγούμενη απογραφή το 2011 (69.849), ο Δήμος Μεσσήνης 19.200 κατοίκους, ο Δήμος Οιχαλίας 8.508 κατοίκους και τέλος, ο Δήμος Μεγαλόπολης 8.791 κατοίκους, με αυτούς τους τρεις δήμους να σημειώνουν μείωση πληθυσμού της τάξης των 2000-3000 κατοίκων συγκριτικά με την απογραφή του 2011.

4.8.Κατηγορίες κάλυψης γης του CLC στην περιοχή μελέτης

Ο όρος χρήσεις γης αναφέρεται μόνο στις δραστηριότητες του ανθρώπου που σχετίζονται άμεσα με τη γη. Οι κοινωνικοοικονομικές δραστηριότητες χωρίζονται σε αγροτικές κι αστικές, με τις πολυφασματικές απεικονίσεις να καταγράφουν τη φασματική πληροφορία των αντικειμένων της επιφάνειας της γης ,αποτυπώνοντας τα φυσικά υλικά του εδάφους. Με τη χρήση της αυτόματης ταξινόμησης των χαρακτηριστικών της γης, με τους υπάρχοντες/διαθέσιμους αλγορίθμους των τεχνικών ταξινόμησης διακρίνονται τελικά οι μορφές κάλυψης γης κι όχι οι χρήσεις.

Λόγω του γεγονότος ότι η εργασία βασίζεται στη μελέτη κατηγοριών χρήσεων/καλύψεων γης, γίνεται χρήση του Corine Land Cover (CLC), που είναι ένα ευρωπαϊκό πρόγραμμα, το οποίο δημιουργήθηκε για την καταγραφή των καλύψεων/χρήσεων γης των 27 μελών της Ευρωπαϊκής Ένωσης, ώστε σε πρώτο βαθμό να γίνουν γνωστές οι κατηγορίες που εντοπίζονται στην περιοχή μελέτης και στην πορεία να αποτελέσουν οδηγό, με την ταυτόχρονη φωτοερμηνεία της απεικόνισης, για τη δημιουργία των δεδομένων εκπαίδευσης/ελέγχου των αλγορίθμων ταξινόμησης και την αξιολόγηση των αποτελεσμάτων.



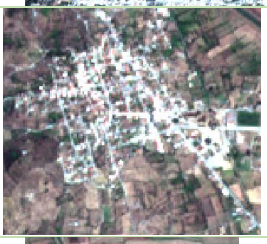
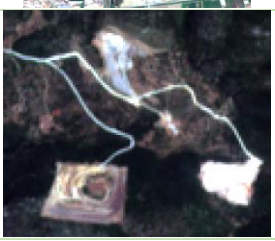
Σχήμα 19. Κατηγορίες κάλυψης/χρήσεις γης του CLC στην περιοχή ενδιαφέροντος

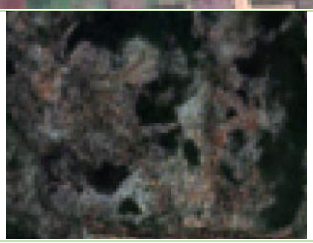
Αξίζει βέβαια να αναφερθεί πως το Corine Land Cover αποτελεί οδηγό για την εκπόνηση της εργασίας, μιας και η πρώτη έρευνα για τον εντοπισμό περιοχής ενδιαφέροντος γίνεται μέσω αυτού, ώστε να δοθεί στον παρατηρητή μια συνολική εικόνα για το πως κατανέμονται οι κατηγορίες κάλυψης/χρήσεις γης στο χώρο, ποια είναι η συγκέντρωσή τους και σε ποιες περιοχές εμφανίζεται πληθώρα κατηγοριών, για να επιλεγεί μια περιοχή πλούσια σε κατηγορίες ,με διαφορετικές εκτάσεις, ούτως ώστε τα πειράματα που θα ακολουθήσουν να ανταποκρίνονται σε ρεαλιστικές συνθήκες.

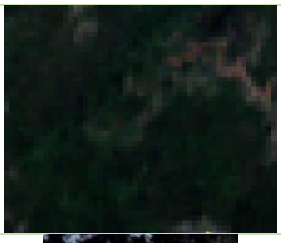
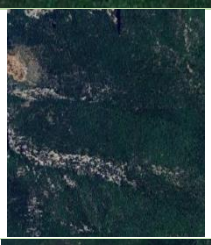
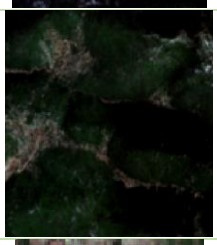
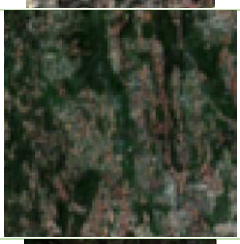
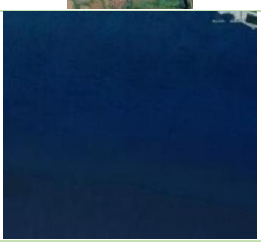
Αναλυτικότερα, οι κατηγορίες που απαντώνται στην περιοχή μελέτης, όπως προκύπτει από το Σχήμα 19, αλλά και το υπόμνημα του CLC είναι

1.Συνεχής αστικός ιστός (1.1.1.), με την κατηγορία αυτή να περιλαμβάνει εκτάσεις γης που καλύπτονται από διάφορες κατασκευές και το δίκτυο μεταφορών.
2.Ασυνεχής αστικός ιστός (1.1.2.), όπου το μεγαλύτερο μέρος γης καλύπτεται από κτίσματα και γενικότερα ζώνες τεχνητών επιφανειών, σε συνδυασμό με ζώνες βλάστησης και γυμνού εδάφους, που καλύπτουν ασυνεχείς, αλλά εκτενείς επιφάνειες.
3.Βιομηχανικές ή εμπορικές ζώνες (1.2.1.), που πρόκειται για ζώνες τεχνητών ζωνών, χωρίς βλάστηση, καλύπτοντας το μεγαλύτερο μέρος της έκτασης ,η οποία επίσης περιλαμβάνει κτίρια ή/και ζώνες βλάστησης.
4.Οδικά ,σιδηροδρομικά δίκτυα και γειτνιάζουσα γη (1.2.2.), όπου σε αυτή την κατηγορία συμπεριλαμβάνονται αυτοκινητόδρομοι, σιδηρόδρομοι, με ελάχιστο πλάτος για ένταξη στην κατηγορία τα 100 μέτρα.
5.Λιμενικές ζώνες (1.2.3.), δηλαδή υποδομές λιμενικών ζωνών, συμπεριλαμβανομένου ναυπηγείων, μαρίνων κι αποβάθρων.
6.Αεροδρόμια (1.2.4.), περιγράφοντας όλες τις εγκαταστάσεις των αεροδρομίων
7.Χώροι εξορύξεως ορυκτών (1.3.1.), δηλαδή περιοχές υπαίθριας εξόρυξης βιομηχανικών ορυκτών ή άλλων ορυκτών.
8.Μη αρόσιμη γη (2.1.1.), που περιλαμβάνει ανθοκομικές καλλιέργειες, δενδροκαλλιέργειες, καθώς κι οπωροκηπευτικά, είτε σε ανοικτό χωράφι, είτε κάτω από πλαστικό ή γυαλί.
9.Μόνιμα αρδευόμενη γη (2.1.2.), στην οποία εντάσσονται καλλιέργειες που ποντίζονται μόνιμα ή περιοδικά, χρησιμοποιώντας μόνιμη υποδομή.
10.Ορυζώνες (2.1.3.), δηλαδή γη διαμορφωμένη για καλλιέργεια ρυζιού, με επίπεδες επιφάνειες με αρδευτικά κανάλια, περιοδικά πλημμυρισμένες.
11.Οπωρώνες (2.2.2.), αποτελούμενοι από αγροτεμάχια που φυτεύονται με οπωροφόρα δέντρα, θάμνους, συνήθως συνδεόμενοι με ποώδη βλάστηση.
12.Ελαιώνες (2.2.3.), δηλαδή περιοχές φυτεμένες με ελαιόδεντρα, συμπεριλαμβανομένων αυτών με μίξη ελαιόδεντρων κι αμπελιών στο ίδιο αγροτεμάχιο.
13.Λιβάδια (2.3.1.), περιοχές με πυκνή κάλυψη από ποώδη βλάστηση, στην οποία κυριαρχούν τα αγρωστώδη φυτά.
14.Σύνθετα συστήματα καλλιέργειας (2.4.2.), που πρόκειται για σύνθεση μικρών αγροτεμαχίων με διάφορες ετήσιες καλλιέργειες, λιβάδια ή/και μόνιμες καλλιέργειες.
15.Γη που καλύπτεται κυρίως από γεωργία με σημαντικές εκτάσεις φυσικής βλάστησης (2.4.3.), δηλαδή περιοχές που καλύπτονται κυρίως από τη γεωργία με διάσπαρτες περιοχές με φυσική βλάστηση.
16.Δάσος πλατύφυλλων (3.1.1.) ,βλάστηση που αποτελείται κυρίως από δέντρα συμπεριλαμβανομένων υπορόφων με θάμνους κι άλλη χαμηλή βλάστηση, όπου κυριαρχούν τα πλατύφυλλα είδη.
17.Δάσος κωνοφόρων (3.1.2.), που πρόκειται για βλάστηση, η οποία αποτελείται κυρίως από δέντρα, συμπεριλαμβανομένων υπορόφων με θάμβους κι άλλη χαμηλή βλάστηση, όπου κυριαρχούν τα κωνοφόρα είδη.

18.Μικτό δάσος (3.1.3.), βλάστηση που αποτελείται από δέντρα, συμπεριλαμβανομένων υπορόφων με θάμνους κι άλλη χαμηλή βλάστηση, όπου δεν κυριαρχούν ούτε τα πλατύφυλλα είδη, ούτε τα κωνοφόρα.
19.Φυσικοί βοσκότοποι (3.2.1.), όπου αντιστοιχεί σε περιοχές με ανώμαλο, ανισόπεδο έδαφος ,χαμηλής παραγωγικότητας.
20.Σκληροφυλλική βλάστηση (3.2.3.), όπου περιλαμβάνει θαμνώδη, σκληροφυλλική βλάστηση, καθώς και τα φρύγανα και τη μακκία.
21.Μεταβατικές δασώδεις – θαμνώδεις εκτάσεις (3.2.4.), μπορεί να αντιπροσωπεύει είτε υποβαθμισμένο δασικό οικοσύστημα, είτε δασική αναγέννηση, αποτυπώνοντας θαμνώδη ή ποώδη βλάστηση με διεσπαρμένα δέντρα.
22.Ροές υδάτων (5.1.1.), που αφορούν για φυσικά ή τεχνητά ρεύματα που λειτουργούν ως αποστραγγιστικά κανάλια.
23.Θάλασσα (5.2.3.), ζώνη προς τη θάλασσα του χαμηλότερου ορίου της παλίρροιας.

Κωδικός κλάσης	Όνομα κλάσης	Στην απεικόνιση (RGB:432)	Στο Google Satellite
1.1.1.	Συνεχής αστικός ιστός		
1.1.2.	Ασυνεχής αστικός ιστός		
1.2.1.	Βιομηχανικές ή εμπορικές ζώνες		
1.2.2.	Οδικά ,σιδ/μικά δίκτυα		
1.2.3.	Λιμάνια		
1.2.4	Αεροδρόμια		
1.3.1.	Ορυχεία		

2.1.1.	Μη αρόσιμη γη		
2.1.2.	Μόνιμα αρδευόμενη γη		
2.1.3.	Ορυζώνες		
2.2.2.	Οπωρώνες		
2.2.3.	Ελαιώνες		
2.3.1.	Λιβάδια		
2.4.2.	Σύνθετες καλλιέργειες		
2.4.3.	Εκτάσεις φυσ. βλάστησης/γεωργία		

3.1.1.	Δάσος πλατύφυλλων		
3.1.2.	Δάσος κωνοφόρων		
3.1.3.	Μεικτό δάσος		
3.2.1.	Φυσικοί βοσκότοποι		
3.2.3.	Σκληροφυλλική βλάστηση		
3.2.4.	Δασώδεις-θαμνώδεις εκτάσεις		
5.1.1.	Ροές υδάτων		
5.2.3.	Θάλασσα		

Πίνακας 2. Εικονιστικά παραδείγματα κατηγοριών CLC

ΚΕΦΑΛΑΙΟ 5. Ερμηνεία Πολυφασματικής Απεικόνισης

5.1.Κανάλια (bands) απεικόνισης

Σε πρώτο επίπεδο είναι δυνατή η ερμηνεία της πολυφασματικής απεικόνισης με προβολή και μελέτη κάθε band ξεχωριστά, μελέτη των ιστογραμμάτων τους, αλλά και γενικότερα εφαρμογή τεχνικών που διευκολύνουν τον παρατηρητή να αντιληφθεί το περιεχόμενο της επιλεγμένης πολυφασματικής εικόνας, να αναγνωρίσει/ξεχωρίσει κατηγορίες κάλυψης/χρήσεων γης.

Θα ήταν παράλειψη να μην αναφερθεί, πριν την προβολή των bands, πως κατά την επιλογή των καναλιών της απεικόνισης στο λογισμικό SNAP (όπως φαίνεται και στο Σχήμα 13), έχουν επιλεγεί τα bands B2, B3, B4, B5, B6, B7, B8, B8-A, B11 & B12, τα οποία στο λογισμικό QGIS αντιστοιχίζονται ως εξής :

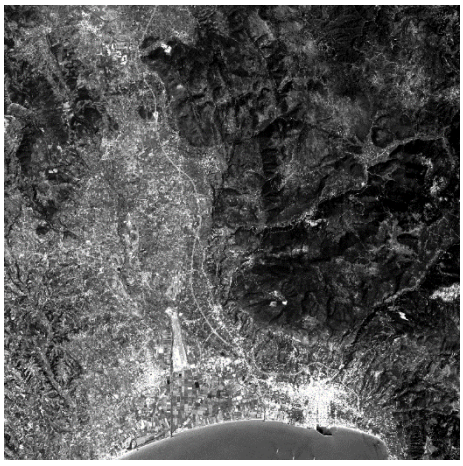
Bands	SNAP	QGIS
Blue	B2	B1
Green	B3	B2
Red	B4	B3
Vegetation red edge	B5	B4
Vegetation red edge	B6	B5
Vegetation red edge	B7	B6
NIR	B8	B7
Narrow NIR	B8-A	B8
SWIR11	B11	B9
SWIR12	B12	B10

Πίνακας 3. Αρίθμηση καναλιών (bands) Sentinel-2A απεικόνισης

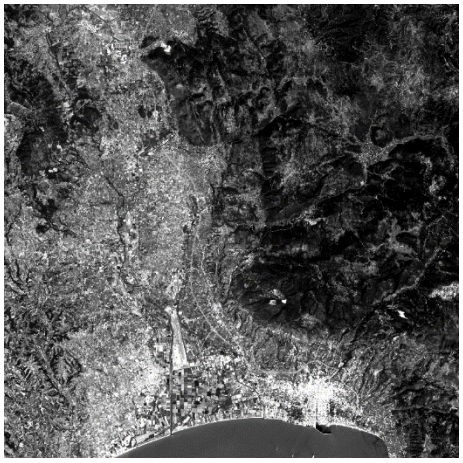
με την αρίθμηση που θα ακολουθείται στην παρούσα εργασία να είναι αυτή του SNAP.

Σε αυτό το σημείο κρίνεται σκόπιμο να αναφερθεί πως τα παρακάτω σχήματα που παρουσιάζονται τα Bands της απεικόνισης έχουν δεχθεί ενίσχυση με αποκοπή στα άκρα 2-98%, ώστε να είναι πιο εύκολη η ερμηνεία τους.

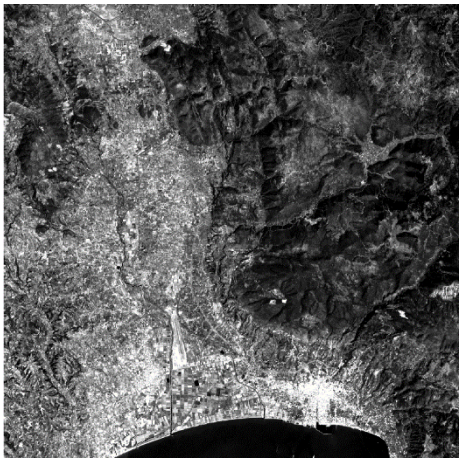
Σχήμα 20. Blue (B2) band, Sentinel-2A πολυφασματικής απεικόνισης, με ημερομηνία λήψης 17/07/2023



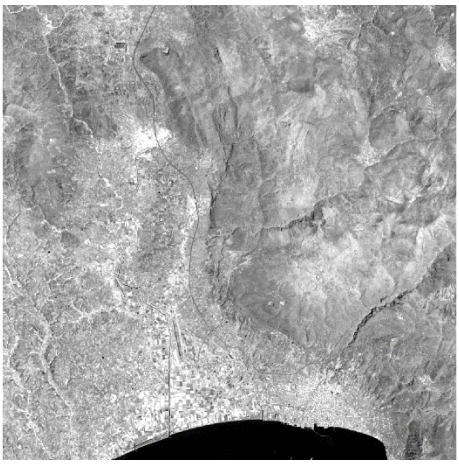
Σχήμα 21. Green (B3) band, Sentinel-2A πολυφασματικής απεικόνισης, με ημερομηνία λήψης 17/07/2023



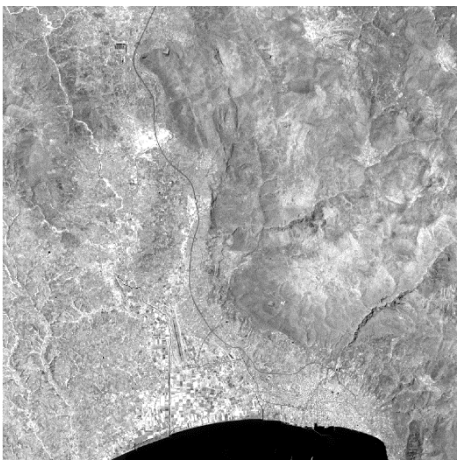
Σχήμα 22. Red (B4) band, Sentinel-2A πολυφασματικής απεικόνισης, με ημερομηνία λήψης 17/07/2023



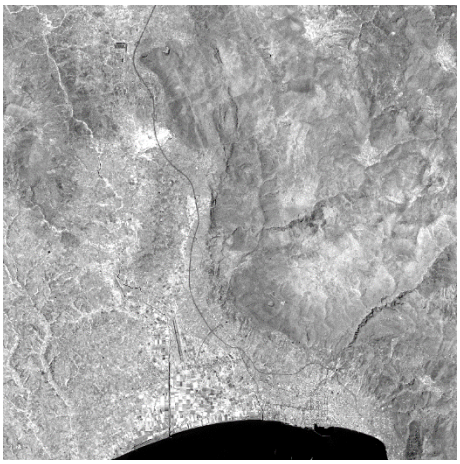
Σχήμα 23. Vegetation Red Edge (B5) band, Sentinel-2A πολυφασματικής απεικόνισης, με ημερομηνία λήψης 17/07/2023



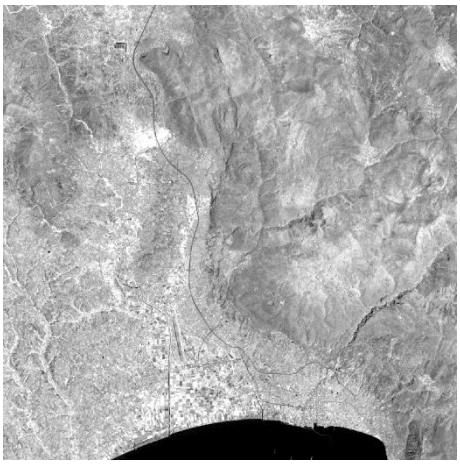
Σχήμα 24. Vegetation Red Edge (B6) band, Sentinel-2A πολυφασματικής απεικόνισης, με ημερομηνία λήψης 17/07/2023



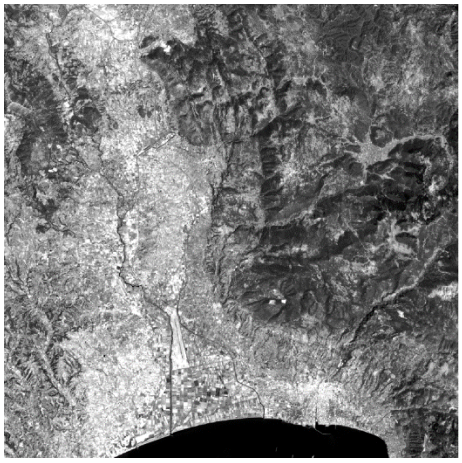
Σχήμα 25. Vegetation Red Edge (B7) band, Sentinel-2A πολυφασματικής απεικόνισης, με ημερομηνία λήψης 17/07/2023



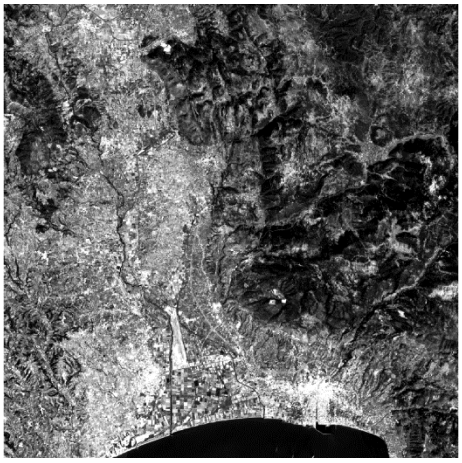
Σχήμα 26. NIR (B8) band, Sentinel-2A πολυφασματικής απεικόνισης, με ημερομηνία λήψης 17/07/2023



Σχήμα 27. Narrow NIR (B8A) band, Sentinel-2A πολυφασματικής απεικόνισης, με ημερομηνία λήψης 17/07/2023



Σχήμα 28. SWIR (B11) band, Sentinel-2A πολυφασματικής απεικόνισης, με ημερομηνία λήψης 17/07/2023



Σχήμα 29. SWIR (B12) band, Sentinel-2A πολυφασματικής απεικόνισης, με ημερομηνία λήψης 17/07/2023

Γίνεται αντιληπτό τόσο από το φυσικό έγχρωμο σύνθετο RGB:432 της απεικόνισης, αλλά κι από την προβολή των καναλιών μεμονωμένα, πως τα τεχνητά χαρακτηριστικά, επιφάνειες της περιοχής μελέτης αναγνωρίζονται σε κάθε περίπτωση, απεικονιζόμενα με λευκούς τόνους λόγω της υψηλής ανακλαστικότητας που σημειώνουν και παρόλο που εντοπίζεται σημαντικός αριθμός των σχετικών κατηγοριών του CLC , επειδή έχουν χαρακτηριστική μορφή κι υφή είναι εύκολα αντιληπτές στον παρατηρητή, με αποτέλεσμα να μην προκαλείται σύγχυση στην αναγνώρισή τους. Αντιθέτως, σημειώνοντας μία προς μία τις κατηγορίες που παρατηρούνται επί της απεικόνισης εντοπίζεται σημαντικός αριθμός κατηγοριών κάλυψης γης, που σχετίζονται με τη βλάστηση. Επομένως το ζητούμενο στο παρόν στάδιο είναι να βρεθούν χαρακτηριστικά όπως σχήματα, υφές κι εντάσεις, με απώτερο σκοπό οι κατηγορίες βλάστησης να μπορούν να διαχωρίζονται, με το CLC να συνιστά βοηθητικό εργαλείο.

Όσον αφορά τις κατηγορίες βλάστησης, ανάλογα την περιεκτικότητα των φυτών σε νερό, αλλά και την υγεία τους, λαμβάνοντας παράλληλα υπόψιν το γεγονός ότι η εποχή λήψης είναι καλοκαίρι, έχει ως αποτέλεσμα να διαφέρει ο τόνος με τον οποίο απεικονίζονται στην πολυφασματική απεικόνιση συγκριτικά με το αναμενόμενο σύμφωνα με τις ρεαλιστικές προσεγγίσεις αποτέλεσμα. Σημαντικό ρόλο παίζει και η ποσότητα εδάφους που περιλαμβάνουν τα εικονοστοιχεία που περιγράφουν τις κατηγορίες βλάστησης, καθώς και ο τύπος του εδάφους συμβάλει στη διαμόρφωση διαφορετικών τόνων του γκρι.

Ιδιαίτερα βοηθητική στην αναγνώριση των κατηγοριών βλάστησης είναι η περιεκτικότητα του εδάφους σε νερό, όπως αναφέρεται και παραπάνω, υποδεικνύοντας ποιες εκτάσεις της απεικόνισης ανήκουν στην κατηγορία της μόνιμα αρδευόμενης γης και ποιες στη μη αρόσιμη γη. Φυσικά, το συμπέρασμα ότι πρόκειται για αυτές τις δυο κατηγορίες προκύπτει από το γνώμονα του σχήματος, αφού τα γεωτεμάχια παρουσιάζουν παραλληλόγραμμο σχηματισμό και σημείωση των ορίων μεταξύ τους. Πιο συγκεκριμένα, διαπιστώνεται πως οι μόνιμα αρδευόμενες εκτάσεις απεικονίζονται με σκουρότερες τόνους από τις μη αρόσιμες.

Από την άλλη, όσον αφορά τις δασικές εκτάσεις σημαντικό ρόλο παίζει η συγκέντρωση βλάστησης και το σχήμα των φύλλων των δένδρων που εντοπίζονται εντός αυτών, από την άποψη πως ένα δένδρο που ανήκει στην κατηγορία των κωνοφόρων ανακλά διαφορετικά από ένα που ανήκει στα πλατύφυλλα. Από τα παραπάνω σχήματα παρατηρείται πως τα κωνοφόρα δάση αποτυπώνονται με τους σκουρότερους τόνους,

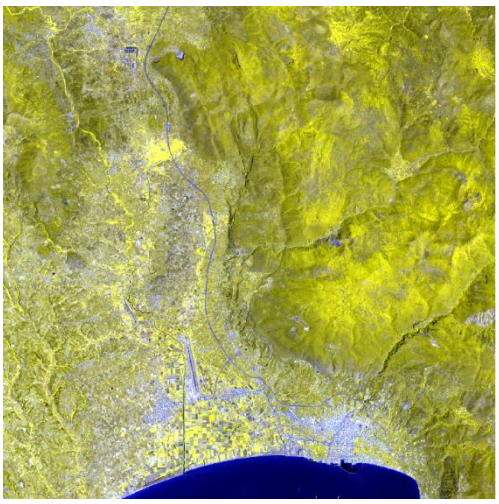
ενώ η κατηγορία των πλατύφυλλων με ανοιχτότερους τόνους και τα μικτά δάση, όπως είναι αναμενόμενο.

Παρόλα αυτά, γίνεται αντιληπτό πως η απεικόνιση κάθε καναλιού ξεχωριστά δε διευκολύνει σημαντικά τη διαδικασία αναγνώρισης και διαχωρισμού των κατηγοριών κάλυψης/χρήσης γης, γι' αυτό ακολουθεί η δημιουργία ψευδέγχρωμων σύνθετων RGB για περεταίρω ερμηνεία.

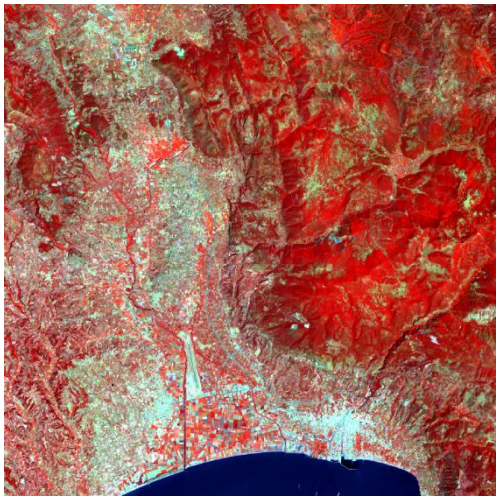
5.2.Έγχρωμα σύνθετα RGB

Κρίνεται σκόπιμη η δημιουργία έγχρωμων σύνθετων, με στόχο τη σύγκριση των ένθετων για τις θεματικές κατηγορίες που απεικονίζει με τον καλύτερο φωτοερμηνευτικό τρόπο το καθένα. Σε αυτό το σημείο αξίζει να σημειωθεί πως ενδείκνυται η επιλογή καναλιών με διακριτική ανάλυση των 10 μέτρων, ώστε να μην αλλοιώνεται η ποιοτική απόδοση των έγχρωμων σύνθετων. Επίσης, λόγω της μεγάλης έκτασης πρασίνου στην περιοχή μελέτης, είτε από άποψη καλλιεργειών, είτε με τη μορφή δασών, παρουσιάζεται ιδιαίτερο ενδιαφέρον για τη μελέτη της βλάστησης στα διάφορα έγχρωμα σύνθετα που κατασκευάζονται, για αυτό μάλιστα επιδιώκεται η χρήση ενός από τα Vegetation Red Edge bands.

Σχήμα 31. Ψευδέγχρωμο σύνθετο RGB:843 απεικόνισης Sentinel-2A, με ημερομηνία λήψης 17/07/2023



Σχήμα 30. Ψευδέγχρωμο σύνθετο RGB:873 απεικόνισης Sentinel-2A, με ημερομηνία λήψης 17/07/2023



Σχήμα 32. Ψευδέγχρωμο σύνθετο RGB:8124 απεικόνισης Sentinel-2A, με ημερομηνία λήψης 17/07/2023

Στο έγχρωμο σύνθετο RGB:873 οι εκτάσεις πρασίνου αποδίδονται με κίτρινες αποχρώσεις λόγω του συνδυασμού του NIR(B8), Vegetation red edge(B7) και Green(B3). Ωστόσο, επειδή οι τεχνητές κατασκευές ανακλούν σημαντικά στο φάσμα του NIR ο αστικός ιστός, συνεχής κι ασυνεχής, εντοπίζονται με ανάγλυφο τρόπο, ενώ είναι ευδιάκριτες και οι βιομηχανικές-εμπορικές ζώνες, με τη λογική αυτή να ακολουθείται σε όλες τις κατηγορίες που σχετίζονται με τεχνητές κατασκευές. Παράλληλα, στο σύνθετο αυτό υπάρχει η δυνατότητα να παρατηρηθούν στις εκτάσεις της βλάστησης ποια τμήματα παρουσιάζουν πυκνή βλάστηση και σε ποια τμήματα αυτή είναι αραιή. Μια πρώτη διάκριση μεταξύ των δασικών εκτάσεων σημειώνεται με την έντονη κίτρινη απόχρωση των δασών πλατύφυλλων, ενώ οι υπόλοιπες δασικές εκτάσεις περιγράφονται με μια απόχρωση που περιέχει κίτρινους, καφέ και γκρι τόνους. Σχετικά με τις εκτάσεις αρόσιμης και μη γης διαπιστώνεται πως κατά κύριο λόγο οι μη αρδεύσιμες εκτάσεις αποδίδονται με μπλε αποχρώσεις, ενώ οι μόνιμα αρδευόμενες με κίτρινους τόνους, ιδιότητα που απορρέει από την ύπαρξη νερού, τις ιδιότητες των φυτών και την αναλογία φυτοκάλυψης/εδάφους στα εξεταζόμενα εικονοστοιχεία.

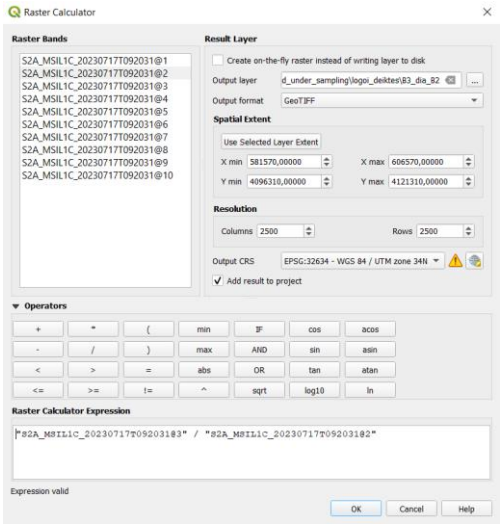
Για το έγχρωμο σύνθετο RGB:843, η βλάστηση αποδίδεται με κόκκινους τόνους λόγω της τοποθέτησης του RED(B4) στο πράσινο. Επειδή και σε αυτό το σύνθετο συμπεριλαμβάνεται το κανάλι NIR(B8) οι τεχνητές επιφάνειες είναι ευδιάκριτες, ενώ ταυτόχρονα γίνονται γνωστές και οι επιφάνειες με αραιή βλάστηση, λόγω της ανακλαστικότητας του εδάφους. Ειδικά στις εκτάσεις των καλλιεργειών, τμήματα που εμφανίζονται με πράσινες αποχρώσεις αντιστοιχούν σε τμήματα που είτε είναι ξερά, είτε έχουν οργωθεί, ενώ τα τμήματα που αποδίδονται με έντονο κόκκινο ακόμα και στις περιπτώσεις των δασών, όχι μόνο στις καλλιέργειες αντιστοιχούν σε πυκνή βλάστηση, αυτός είναι ο βασικός λόγος για τον οποίο διακρίνονται τα δάση πλατύφυλλων από τα υπόλοιπα- λόγω μεγαλύτερης ανακλαστικής επιφάνειας φύλλων και λιγότερης περιεκτικότητας σε εδαφικό υλικό. Σε αυτό το έγχρωμο σύνθετο εντοπίζεται και πλήθος εκτάσεων όπου με σκούρους κόκκινους τόνους απεικονίζονται τα δάση κωνοφόρων, ενώ με αντίστοιχη απόχρωση και ταυτόχρονη έντονη παρεμβολή ,εντός της βλάστησης, γκρίζων-μπλε τόνων απεικονίζεται η σκληροφυλλική βλάστηση.

Τέλος, όσον αφορά το έγχρωμο σύνθετο RGB:8124 περιλαμβάνει τα κανάλια NIR(B8), SWIR(B12) και RED(B4), η εφαρμογή των οποίων έχει ως αποτέλεσμα την απόδοση των καλλιεργήσιμων εκτάσεων με διαφορετικές αποχρώσεις ,ανάλογα το είδος τους. Το συγκεκριμένο σύνθετο παρουσιάζει περισσότερο ενδιαφέρον για μελέτη κατά διαχωρισμό των χρήσεων γης συγκριτικά με τα άλλα δυο, επειδή οι εκτάσεις πρασίνου δεν αποδίδονται με διαφορετικούς τόνους μιας απόχρωσης, αλλά με διαφορετικά χρώματα, γεγονός που οφείλεται στις ανακλαστικές ιδιότητες που παρουσιάζει κάθε καλλιέργεια ή δάσος, όπου για παράδειγμα οι εκτάσεις σκληροφυλλικής βλάστησης αποδίδονται με αποχρώσεις που τείνουν στο πράσινο, ενώ τα δάση πλατύφυλλων με έντονους τόνους του πορτοκαλί.

5.3.Αριθμητικές πράξεις καναλιών πολυφασματικής απεικόνισης

Σε δύο ή περισσότερες εικόνες που έχουν το ίδιο σύστημα αναφοράς κι απεικονίζουν την ίδια γεωγραφική περιοχή οι αριθμητικές πράξεις που είναι δυνατό να εφαρμοστούν είναι η πρόσθεση ,η αφαίρεση, ο πολλαπλασιασμός και η διαίρεση. Οι πράξεις αυτές εφαρμόζονται μεταξύ των ψηφιακών τιμών των εικονοστοιχείων των φασματικών καναλιών, χωρίς να εμπλέκονται γειτονικά εικονοστοιχεία. Οι αριθμητικές πράξεις γίνονται είτε στα φασματικά κανάλια της ίδιας δορυφορικής εικόνας, είτε στα φασματικά κανάλια εικόνων που προέρχονται από διαφορετικές ημερομηνίες. Σε αυτό το στάδιο, με λόγους και δείκτες φασματικών καναλιών επιδιώκεται κυρίως να αναδειχθούν κατηγορίες που δεν είναι ευδιάκριτες μεταξύ τους και δεν μπόρεσαν να διακριθούν με χαρακτηριστικό τρόπο στην προηγούμενη ενότητα, με τη δημιουργία έγχρωμων σύνθετων.

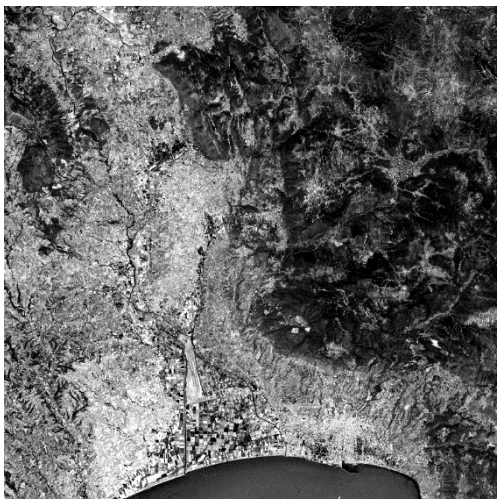
Για τη δημιουργία λόγου και δεικτών στο περιβάλλον του λογισμικού QGIS αξιοποιείται το εργαλείο Raster>Raster Calculator ,όπου εμπεριέχονται ένα προς ένα τα φασματικά κανάλια της εικόνας κι υπάρχει η δυνατότητα εκτέλεσης οποιασδήποτε πράξεις με τη χρήση των φασματικών καναλιών, ακόμα και η απομόνωση και παρουσίαση μόνο ενός καναλιού της εικόνας.



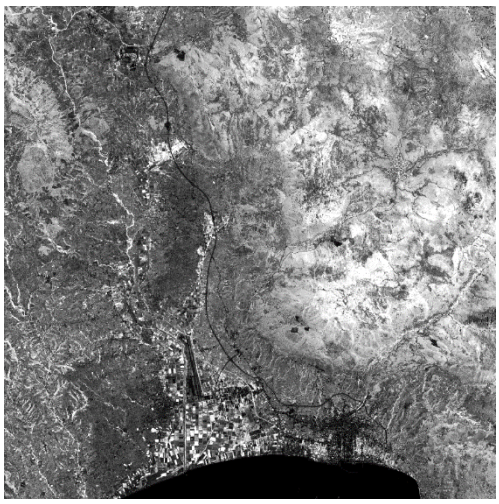
Σχήμα 33. Εκτέλεση εργαλείου Raster Calculator για τη δημιουργία λόγων και δεικτών

Εφόσον σημαντικό πρόβλημα εντοπίζεται στο διαχωρισμό κατηγοριών που αφορούν τη βλάστηση, ένας από τους δείκτες που κρίνεται σκόπιμο κι αναγκαίο να χρησιμοποιηθεί στην πολυφασματική εικόνα ,είναι αυτός του NDVI. Υπενθυμίζεται πως ο δείκτης NDVI είναι ένας από τους δείκτες βλάστησης, όπου αυτοί βασίζονται στην αλληλεπίδραση της ηλεκτρομαγνητικής ακτινοβολίας με τα φύλλα των φυτών. Τα φύλλα περιέχουν ειδικές χρωστικές ουσίες όπως οι χλωροφύλλες, οι οποίες χρησιμεύουν στην απορρόφηση ενέργειας φωτός για την επιτέλεση της λειτουργίας της φωτοσύνθεσης. Να σημειωθεί σε αυτό το σημείο πως τα φυτά χρησιμοποιούν ενέργεια από συγκεκριμένα μήκη κύματος της ηλεκτρομαγνητικής ακτινοβολίας για να παράγουν τα απαραίτητα στοιχεία για τη θρέψη τους. Επομένως, γίνεται αντιληπτό πως οι βιοφυσικές ιδιότητες της βλάστησης επηρεάζουν την απορρόφηση, ανάκλαση και μετάδοση της ηλεκτρομαγνητικής ακτινοβολίας στα διαφορετικά μήκη κύματος του ηλεκτρομαγνητικού φάσματος. Εστιάζοντας στον κανονικοποιημένο δείκτη βλάστησης διαφοράς (NDVI) ,στα πλεονεκτήματά του είναι η ελαχιστοποίηση των τοπογραφικών επιδράσεων, είναι σχεδόν αναλλοίωτος από τις διάφορες συνθήκες εξαιτίας των

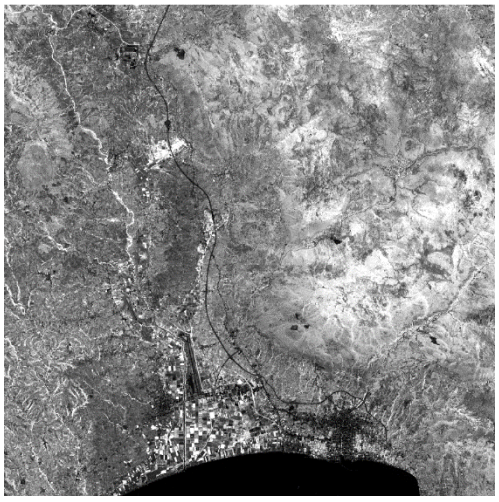
κανονικοποιημένων τιμών, το εύρος τιμών του είναι από -1 έως και +1, με το 0 να εκφράζει απουσία βλάστησης, ενώ οι αρνητικές τιμές περιγράφουν καλύψεις γης όπως νερό κι ανθρωπογενείς κατασκευές. Ωστόσο, στα μειονεκτήματα του NDVI εντοπίζονται δείγματα κορεσμού σε πολύ υψηλές συγκεντρώσεις βλάστησης, καθώς κι υπερεκτίμηση σε χαμηλές συγκεντρώσεις βλάστησης εξαιτίας της ανακλαστικότητας του εδάφους.



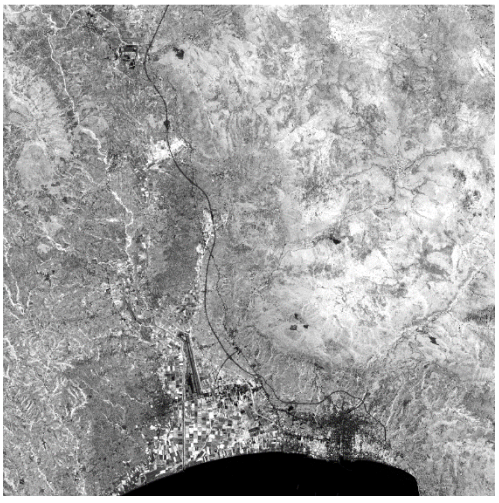
Σχήμα 34. Λόγος Red(B4)/Green(B3) φασματικών καναλιών πολυφασματικής απεικόνισης Sentinel-2A, με ημερομηνία λήψης 17/07/2023



Σχήμα 35. Λόγος Vegetation Red Edge(B6)/Red(B4) φασματικών καναλιών πολυφασματικής απεικόνισης Sentinel-2A, με ημερομηνία λήψης 17/07/2023



Σχήμα 36. Λόγος NIR(B8)/Vegetation red edge(B5) φασματικών καναλιών πολυφασματικής απεικόνισης Sentinel-2A, με ημερομηνία λήψης 17/07/2023



Σχήμα 37. Δείκτης NDVI φασματικών καναλιών πολυφασματικής απεικόνισης Sentinel-2A, με ημερομηνία λήψης 17/07/2023

Ο λόγος Red(B4)/Green (B3) έχει την ιδιότητα να περιγράφει με έντονο τρόπο την ανακλαστικότητα του εδάφους, καθώς εκτάσεις γυμνού εδάφους, ή τεχνητές κατασκευές, όπως ο αστικός ιστός, το οδικό δίκτυο απεικονίζονται με ανοιχτόχρωμους τόνους, όπου μια λεία ανακλαστική επιφάνεια, στην οποία δεν εντοπίζεται κάποια θεματική κατηγορία, όπως αυτή της βλάστησης, αποτυπώνεται στην απεικόνιση με λευκούς τόνους. Αντίθετα, οι εκτάσεις πρασίνου, ανάλογα με την ποσότητα της βλάστησης απεικονίζονται και με τον αντίστοιχο σκούρο τόνο του γκρι. Πιο

συγκεκριμένα, εκτάσεις όπως αυτές των δασών πλατύφυλλων δένδρων απεικονίζονται με τους πιο σκούρους τόνους, λόγω του γεγονότος ότι τα εικονοστοιχεία στις περιοχές αυτές εμφανίζουν ελάχιστη ποσότητα εδάφους, με τα δάση κωνοφόρων να αποτυπώνονται με μέσες τιμές του γκρίζου, καθώς στα εικονοστοιχεία τους εμφανίζεται μια μέση περιεκτικότητα σε έδαφος και βλάστηση.

Ο λόγος Vegetation red edge (B6)/Red(B4) προσφέρει στον παρατηρητή τη δυνατότητα μιας στοιχειώδους εκτίμησης της φυτοκάλυψης των περιοχών μελέτης ανάλογα με τους τόνους του γκρι. Μάλιστα, όσο πιο ανοιχτόχρωμοι τόνοι περιγράφουν τα εικονοστοιχεία που αποτελούνται μόνο από βλάστηση, διευκολύνοντας την αναγνώριση των περιοχών που καλύπτονται από πλατύφυλλα δάση, ενώ και στην περίπτωση της αρόσιμης γης ξεχωρίζουν τα γεωτεμάχια με την πιο πυκνή φυτοκάλυψη, από τα υπόλοιπα που έχουν λιγότερη. Παράλληλα, αυτή η ποσοτικοποίηση της πυκνότητας της βλάστησης εξυπηρετεί στην αναγνώριση των ομάδων εικονοστοιχείων που απεικονίζουν δάση κωνοφόρων μιας και το πλάτος του φύλλου είναι μικρό, άρα η ακτινοβολία φτάνει έως το έδαφος, με αποτέλεσμα σε ένα εικονοστοιχείο αυτής της θεματικής κατηγορίας να καταγράφεται κυρίως έδαφος, παρά βλάστηση.

Στη συνέχεια ανάλυσης των λόγων φασματικών καναλιών, ακολουθεί η ερμηνεία του λόγου NIR(B8)/Vegetation red edge(B5), ο οποίος θα μπορούσε να χαρακτηριστεί ως αντίστοιχος του προηγούμενου, Vegetation red edge (B6)/Red(B4), έχοντας ως στόχο να διακρίνει τις κατηγορίες βλάστησης ανάλογα με την ποσότητα πρασίνου, όπου όσο μεγαλύτερη είναι, τόσο πιο ανοιχτόχρωμα τα εικονοστοιχεία που τις απεικονίζουν.

Εφόσον, οι κατηγορίες που επιδιώκεται να διαχωρισθούν ανήκουν στο πεδίο της βλάστησης ,θα ήταν παράλειψη να μην εφαρμοστεί ο παγκόσμιος κανονικοποιημένος δείκτης βλάστησης NDVI, ο οποίος υπολογίζεται ως NIR(B8)-Red(B4)/NIR(B8)+Red(B4). Επί της ουσίας, με την εφαρμογή του δείκτη αυτού ποσοτικοποιείται η βλάστηση στις απεικονίσεις και παράλληλα διακρίνονται τεχνικά χαρακτηριστικά, όπως δρόμοι, που παρεμβάλλονται μεταξύ των εκτάσεων πρασίνου.

Να σημειωθεί πως η ανάλυση τόσο των σχημάτων των λόγων και δεικτών σε αυτή την ενότητα, όσο και αυτών των έγχρωμων σύνθετων στην προηγούμενη, εστιάζει σε θεματικές κατηγορίες με ακανόνιστο σχήμα, μη ορισμένη γεωμετρία, όπως αυτές των δασών και γενικότερα όσων σχετίζονται με τη βλάστηση, καθώς είναι αυτές που εντοπίζονται πιο δύσκολα στις απεικονίσεις, χωρίς τη χρήση του Corine Land Cover. Θεματικές ενότητες όπως αυτή του αστικού ιστού ή των υδάτινων σωμάτων, μπορεί να μην έχουν ορισμένη γεωμετρία, όμως η υφή τους, η έκτασή τους σε ορισμένες περιπτώσεις κι άλλα χαρακτηριστικά, είναι αυτά που τις κάνουν απευθείας αντιληπτές από τον παρατηρητή.

5.4.Ενοποίηση κατηγοριών CLC

Από τις Ενότητες 5.1.Κανάλια (bands) απεικόνισης, 5.2.Έγχρωμα σύνθετα RGB και 5.3.Αριθμητικές πράξεις καναλιών πολυφασματικής απεικόνισης, όπου γίνεται η παράθεση των απομονωμένων καναλιών της πολυφασματικής απεικόνισης που αναπαριστά την περιοχή ενδιαφέροντος και με το συνδυασμό των καναλιών, είτε με τη μορφή των ψευδέγχρωμων σύνθετων ,είτε τη χρήση αριθμητικών πράξεων μεταξύ των καναλιών (λόγοι και δείκτης NDVI), γίνεται αντιληπτό πως το πλήθος των κατηγοριών και δη η μελέτη συγγενών κατηγοριών (βλ. τρεις κατηγορίες δασών), τόσο κατηγοριών που σχετίζονται με τη βλάστηση, όσο και των κατηγοριών που αφορούν τεχνητές επιφάνειες και τεχνικά έργα, δυσχεραίνει τη διαδικασία της ταξινόμησης κι ιδιαίτερα την ακρίβεια του παραγόμενου προϊόντος. Αυτός είναι ο βασικός λόγος για τον οποίο θα ακολουθήσει η συγχώνευση συγγενών κατηγοριών, για την καλύτερη διαχείρισή τους κατά τη διεξαγωγή των πειραμάτων.

Ένας παράγοντας που θα υποδείξει σε πρώτο στάδιο ποιο δύναται να είναι το πλήθος των κατηγοριών είναι η εφαρμογή μη επιβλεπόμενης ταξινόμησης. Με τις πειραματικές δοκιμές μη επιβλεπόμενης ταξινόμησης δίνεται μια πρώτη εκτίμηση για το πόσες κατηγορίες μπορούν να αναγνωριστούν στην περιοχή μελέτης, χωρίς εκπαίδευση, με βάση κοινά χαρακτηριστικά των εικονοστοιχείων της πολυφασματικής απεικόνισης, όπως είναι οι τιμές φωτεινότητας.

Από τη βιβλιογραφία παράλληλα, είναι γνωστό πως σε πολλές περιπτώσεις απαιτείται συγχώνευση ορισμένων κλάσεων για να σχηματίσουν έναν τύπο κάλυψης γης ή διαχωρισμός μίας γενικής κατηγορίας σε περισσότερους από έναν τύπους κάλυψης γης που παρουσιάζουν ενδιαφέρον, για την εκάστοτε εφαρμογή (Zhu H., 2008). Άρα, ο παρατηρητής μπορεί να προβεί σε διαδικασία σύμπτυξης κατηγοριών που σχετίζονται μεταξύ τους, δηλαδή κατηγοριών με παρόμοια χαρακτηριστικά ή/και συμπεριφορά στο φάσμα, φαινόμενο που αποδεικνύεται από τις φασματικές υπογραφές τους.

Επομένως, η μεθοδολογία της εργασίας συνεχίζει με τη μελέτη των πειραμάτων μη επιβλεπόμενης ταξινόμησης και σε επόμενο βήμα των φασματικών υπογραφών των κατηγοριών/χρήσεων γης της περιοχής μελέτης, ώστε να διαμορφωθούν οι τελικές προς μελέτη κατηγορίες, έπειτα από ενοποίηση των αρχικών ,23 σε αριθμό, κατηγοριών.

5.4.1.Εφαρμογή Μη Επιβλεπόμενης Ταξινόμησης

Στη μη επιβλεπόμενη ταξινόμηση, οι περιοχές ομαδοποίησης προσδιορίζονται από τον ίδιο τον αλγόριθμο ταξινόμησης, χωρίς να αξιοποιούνται δεδομένα από περιοχές εκπαίδευσης. Στο τέλος της αυτοματοποιημένης αλγοριθμικής διαδικασίας, κάθε εικονοστοιχείο έχει ταυτοποιηθεί σε μια τάξη, όμως δεν είναι γνωστή η φυσική οντότητα που εκφράζει η τάξη αυτή. Έγκειται ,λοιπόν, στο χρήστη να αξιοποιήσει πληροφορίες για την περιοχή μελέτης, ώστε να ερμηνεύσει το φυσικό περιεχόμενο της κάθε τάξης. Εν προκειμένω, η διαδικασία εφαρμογής μη επιβλεπόμενης ταξινόμησης επιλέγεται να εφαρμοστεί με αυτοματοποιημένο τρόπο, με τη χρήση του προγράμματος SNAP, αξιοποιώντας τον αλγόριθμο k-means.

Σχετικά με τον αλγόριθμο K-Means, πρόκειται για έναν αλγόριθμο που ταξινομεί τα δεδομένα με βάση τα χαρακτηριστικά των K διαφορετικών κατηγοριών/τάξεων. Ο αλγόριθμος υποθέτει ότι τα χαρακτηριστικά των τάξεων δημιουργούν ένα χώρο διανυσμάτων κι ουσιαστικά σκοπός του είναι να ελαχιστοποιήσει τη συνολική

διακύμανση κάθε τάξης/φασματικής κατηγορίας ή τη συνάρτηση τετραγωνικού σφάλματος.

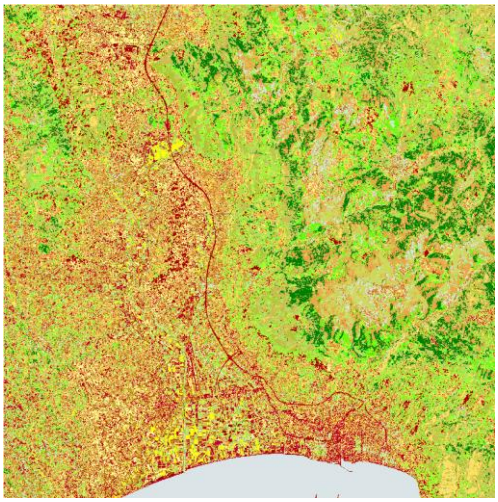
Τα βασικά βήματα του αλγορίθμου είναι τα εξής :

1. Επιλογή του αριθμού των k τάξεων,
2. Αρχικοποίηση με τυχαίο ορισμό των κεντροειδών των k τάξεων,
3. Σύνδεση του κάθε εικονοστοιχείου με το κεντροειδές της κοντινότερης, με βάση κάποιο κριτήριο, τάξης
4. Υπολογισμός των νέων κεντροειδών κάθε τάξης κι
5. Επανάληψη των βημάτων 3 και 4, μέχρι να συγκλίνει ο αλγόριθμος.

Ο αλγόριθμος συγκλίνει όταν ικανοποιηθεί κάποιο από τα κριτήρια σύγκλισης, μερικά εκ των οποίων μπορούν να είναι η πραγματοποίηση του μέγιστου αριθμού επαναλήψεων, που καθορίζεται από το χρήστη αρχικά, ή όταν τα κέντρα των τάξεων δε μεταβάλλονται σημαντικά μεταξύ δυο διαδοχικών επαναλήψεων.

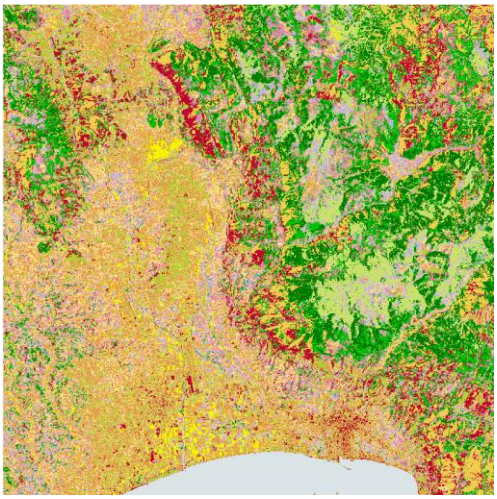
Τονίζεται πως ο συγκεκριμένος αλγόριθμος συγκλίνει σχετικά γρήγορα κι είναι ο λόγος για τον οποίο θεωρείται δημοφιλής, ενώ θα ήταν παράλειψη να μην αναφερθεί πως ως μειονέκτημά του σημειώνεται ότι ο αριθμός των τάξεων πρέπει να οριστεί εξ' αρχής από το χρήστη. Σε τηλεπισκοπικές απεικονίσεις που καλύπτουν μεγάλες εκτάσεις δεν είναι εύκολο ή εφικτό να είναι γνωστός από την αρχή και πριν τα πειράματα ο αριθμός των τάξεων που μπορεί ένας αλγόριθμος να αναγνωρίσει. Για το λόγο αυτό, είναι σκόπιμο να εφαρμοστούν πληθώρα πειραμάτων με διαφορετικούς αριθμούς τάξεων.

Σε αυτό το σημείο, η μεθοδολογία κινείται με τον ορισμό 23 τάξεων στο πρώτο πείραμα, όσες δηλαδή είναι και οι κατηγορίες του CLC που σημειώνονται στην περιοχή μελέτης και στη συνέχεια, διαδοχική μείωση του αριθμού των τάξεων στα πειράματα έως ότου να βρεθεί ένα πλήθος τάξεων που θα θεωρείται αντιπροσωπευτικό.

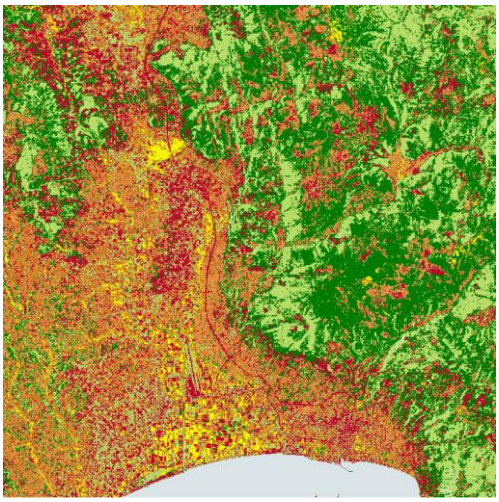


Σχήμα 38. Μη Επιβλεπόμενη Ταξινόμηση, Αλγόριθμος K-Means, με 23 τάξεις

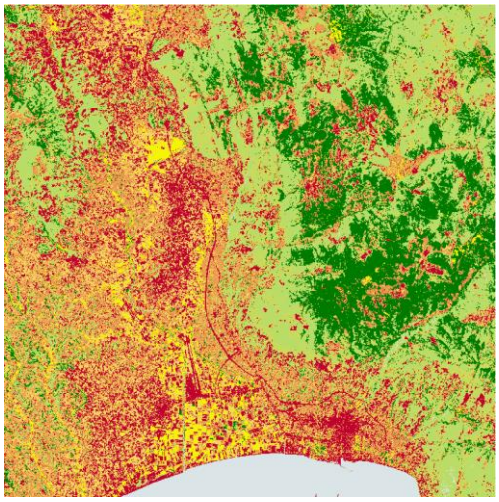
Σχήμα 39. Μη Επιβλεπόμενη Ταξινόμηση, Αλγόριθμος K-Means, με 20 τάξεις



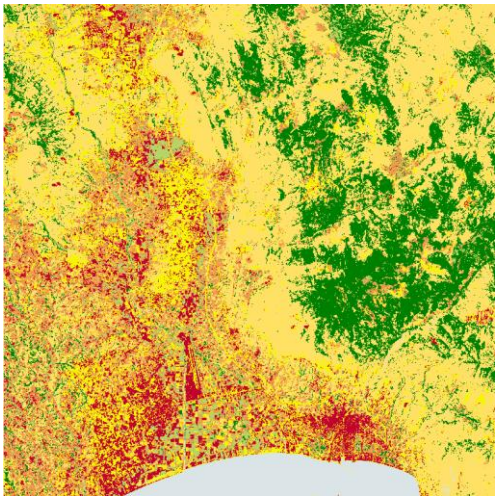
Σχήμα 40. Μη Επιβλεπόμενη Ταξινόμηση, Αλγόριθμος K-Means, με 15 τάξεις



Σχήμα 41. Μη Επιβλεπόμενη Ταξινόμηση, Αλγόριθμος K-Means, με 10 τάξεις



Σχήμα 42. Μη Επιβλεπόμενη Ταξινόμηση, Αλγόριθμος K-Means, με 9 τάξεις



Σχήμα 43. Μη Επιβλεπόμενη Ταξινόμηση, Αλγόριθμος K-Means, με 8 τάξεις

Σε αυτό το στάδιο γίνεται αντιπαραβολή των θεματικών χαρτών που παράγονται με την εφαρμογή μη επιβλεπόμενης ταξινόμησης με το χάρτη χρήσεων/καλύψεων γης CLC, επιδιώκοντας τη σύγκριση και τον πιθανό εντοπισμό/αναγνώριση κατηγοριών χρήσεων γης, για αυτό γίνεται και χρήση αντιπροσωπευτικών αποχρώσεων αντίστοιχων αυτών που χρησιμοποιούνται από το CLC.

Η ερμηνεία και σύγκριση των θεματικών χαρτών υποδεικνύει πως ενώ και οι 23 τάξεις αναγνωρίζονται με έναν ικανοποιητικό βαθμό, αναλογιζόμενοι πως πρόκειται για ταξινόμηση χωρίς εκπαίδευση, η άμεση συσχέτιση και σύγχυση μεταξύ των κατηγοριών, οδηγούν στην ανάγκη για μείωση του πλήθους των τάξεων ανά πείραμα. Με αυτόν τον τρόπο φαίνεται πως τα πειράματα με τις 8,9 και 10 τάξεις ανταποκρίνονται ικανοποιητικά στις πραγματικές συνθήκες που επικρατούν στην περιοχή μελέτης, αναγνωρίζοντας κι αντιστοιχίζοντας σε μια κατηγορία τεχνητές επιφάνειες, ενοποιώντας τις δασικές εκτάσεις, αποτυπώνοντας με σημαντική ακρίβεια εκτάσεις σκληροφυλλικής βλάστησης, με το πείραμα των 9 τάξεων σε περιοχές, πλησίον των δασικών, όπου σύμφωνα με την ερμηνεία της πολυφασματικής

απεικόνισης χαρακτηρίζονται από αραιή ή/και καθόλου βλάστηση και κατά τόπους πυκνή, φαίνεται να σημειώνονται σαν δυο διαφορετικές κατηγορίες, αλλά πρόκειται για τις συγγενείς κατηγορίες των φυσικών βοσκοτόπων και της σκληροφυλλικής βλάστησης. Εν ολίγοις, σε πρώτο στάδιο οι πειραματικές εφαρμογές μη επιβλεπόμενης ταξινόμησης υποδεικνύουν πως η ενοποίηση κλάσεων που παρουσιάζονται στην περιοχή, φτάνοντας σε πλήθος 8-10 κατηγορίες, είναι ικανές να περιγράψουν με σημαντική ακρίβεια την περιοχή, χωρίς να υποβιβάζονται κατηγορίες.

Στην επόμενη ενότητα (5.4.2.Υπόδειξη κατηγοριών από το CLC) μελετώνται οι κατηγορίες κάλυψης γης σύμφωνα με την ομαδοποίηση που ορίζει το CLC και πως αυτές μπορούν να συγχωνευθούν σε ένα πλαίσιο 8-10 κατηγοριών, που θα αποτελέσουν τις τελικές, προς μελέτη κατηγορίες.

5.4.2.Υπόδειξη κατηγοριών από το CLC

Το Corine Land Cover (CLC) σημειώνει 5 βασικές κατηγορίες, “ομπρέλες”, στις οποίες κατατάσσονται χαρακτηριστικές υποκατηγορίες αυτών, όπως υποδεικνύονται και στην Εικόνα 4, αποτελώντας βασικό οδηγό για την ενοποίηση των κατηγοριών στην περιοχή μελέτης, σε συνδυασμό με τις φασματικές υπογραφές που παρουσιάζονται παρακάτω.

Εικόνα 4. Υπόμνημα Corine Land Cover (CLC)
https://www.michanikos.gr/uploads/monthly_05_2016/post-94283-0-30282500-1462953499.jpg

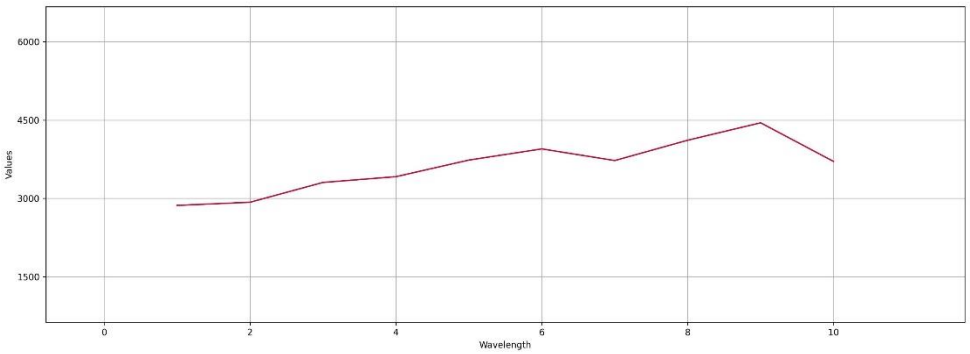
Σύμφωνα με το υπόμνημα του CLC, μια πρώτη εκτίμηση για τη συγχώνευση των κατηγοριών και τη διαμόρφωση των τελικών κατηγοριών προς μελέτη στηρίζεται στην ενοποίηση όλων των τεχνητών επιφανειών, όπως αυτές υπάγονται στην κύρια κατηγορία, ενώ στην περίπτωση των κατηγοριών που σχετίζονται με τη βλάστηση και στα ύδατα, αξιοποιούνται οι υποκατηγορίες αρόσιμη γη, μόνιμες καλλιέργειες, λιβάδια, ετερογενείς γεωργικές περιοχές, δάση, συνδυασμοί θαμνώδους και ποώδους βλάστησης, και για τα ύδατα αντίστοιχα, χερσαία ύδατα και θαλάσσια ύδατα.

5.4.3.Φασματικές υπογραφές κατηγοριών CLC

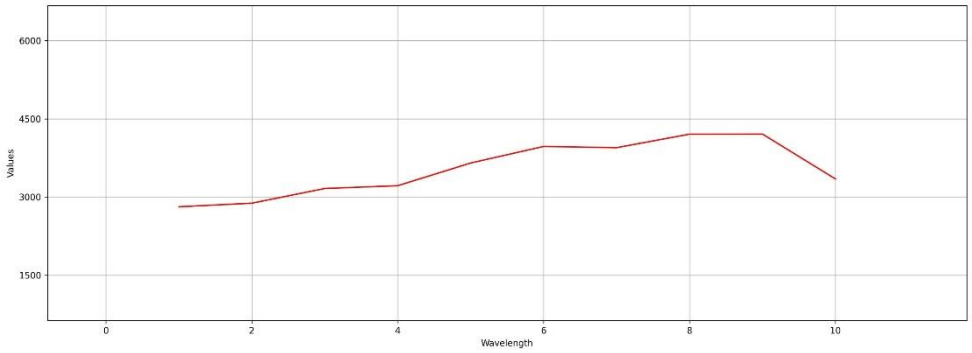
Με τη συλλογή περιοχών εκπαίδευσης καταγράφονται αντιπροσωπευτικές ραδιομετρικές τιμές για την κάθε κατηγορία κάλυψης, με τις μέσες τιμές που συγκεντρώνονται για κάθε κανάλι να δημιουργούν ένα διάνυσμα, το οποίο είναι χαρακτηριστικό για κάθε κατηγορία. Με αυτόν τον τρόπο δημιουργούνται τα διαγράμματα ανακλαστικότητας των κατηγοριών κάλυψης/χρήσεων γης, γνωστά κι ως φασματικές υπογραφές.

Εφόσον το βασικό εργαλείο για την εκπόνηση της παρούσας εργασίας είναι το λογισμικό περιβάλλον QGIS και πιο συγκεκριμένα το Semi-Automatic Classification Plugin, που παρέχει τη δυνατότητα μελέτης των φασματικών υπογραφών, η παράθεσή τους γίνεται με χρήση αυτού. Αναλυτικότερα, η διαδικασία ξεκινά με το σχεδιασμό πολυγώνων επί της απεικόνισης μέσω του περιβάλλοντος του SCP, στα πλαίσια του Training input, με κάθε πολύγωνο να περιγράφεται από το όνομα της κλάσης που αντιπροσωπεύει και το Class ID, που λαμβάνει αύξοντα αριθμό κατά το σχεδιασμό πολυγώνων, ενώ το Macroclass ID να παραμένει σταθερό για όλες τις κατηγορίες. Δημιουργώντας τα πολύγωνα, δίνεται η δυνατότητα σχεδιασμού των φασματικών υπογραφών και η παράθεσή τους σε αντίστοιχη καρτέλα, όπου υπάρχει η επιλογή να παρουσιασθούν όλες συγκεντρωμένες ή/και μεμονωμένες.

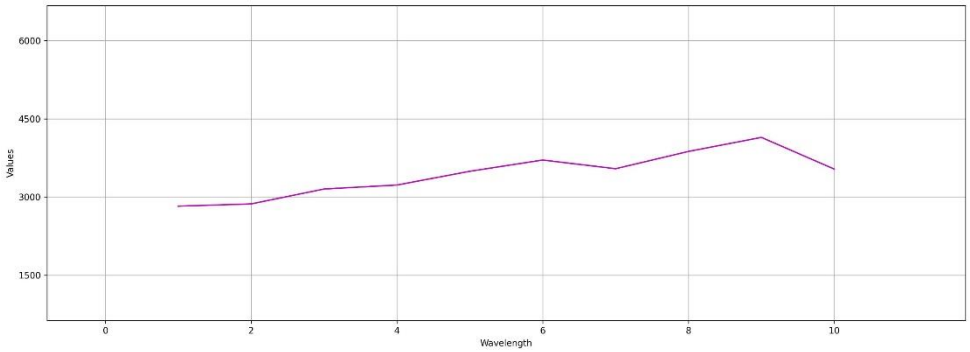
Ήδη έχει γίνει υπόδειξη των κατηγοριών μελέτης, αρκεί τώρα να επισφραγισθεί αυτή με τη μελέτη των φασματικών υπογραφών που παρουσιάζονται στα παρακάτω σχήματα.



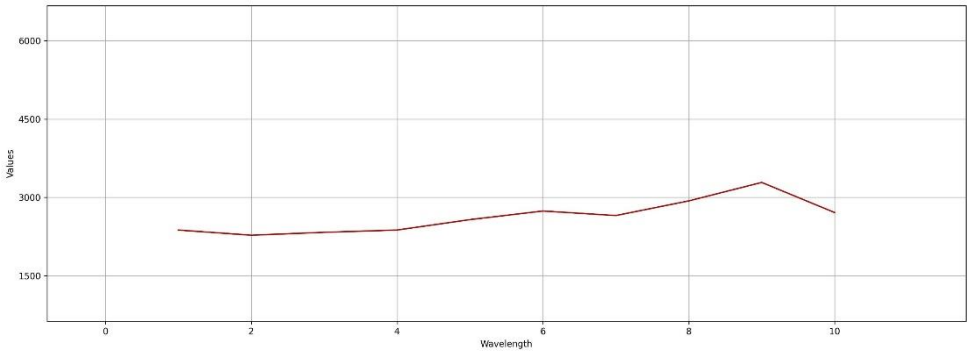
Σχήμα 44. Φασματική υπογραφή συνεχή αστικού ιστού της περιοχής μελέτης



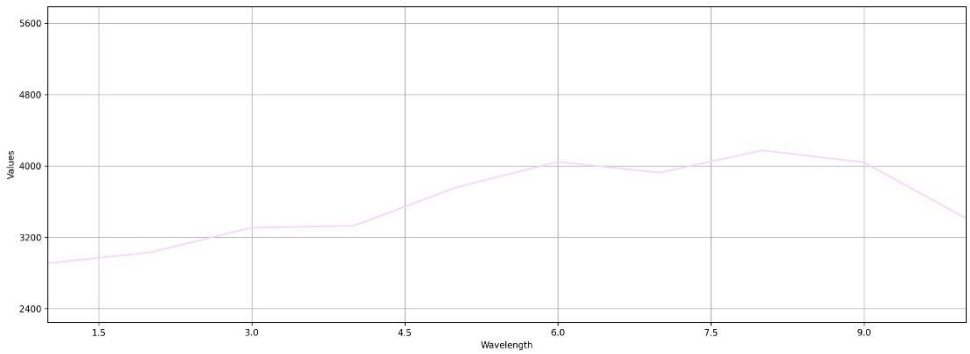
Σχήμα 45. Φασματική υπογραφή ασυνεχούς αστικού ιστού της περιοχής μελέτης



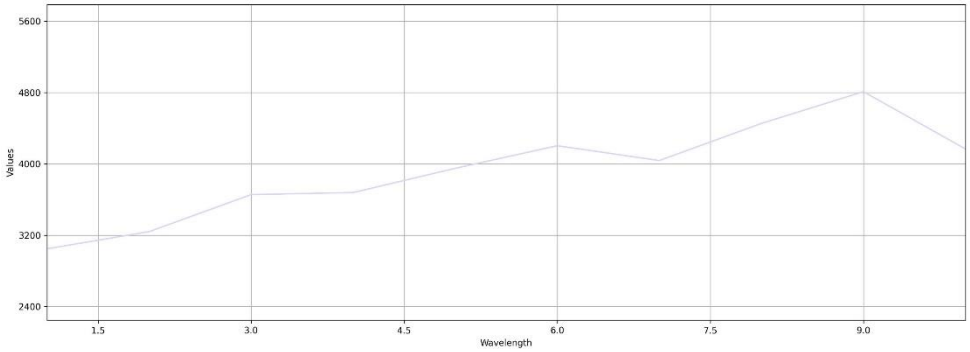
Σχήμα 46. Φασματική υπογραφή βιομηχανικής ζώνης της περιοχής μελέτης



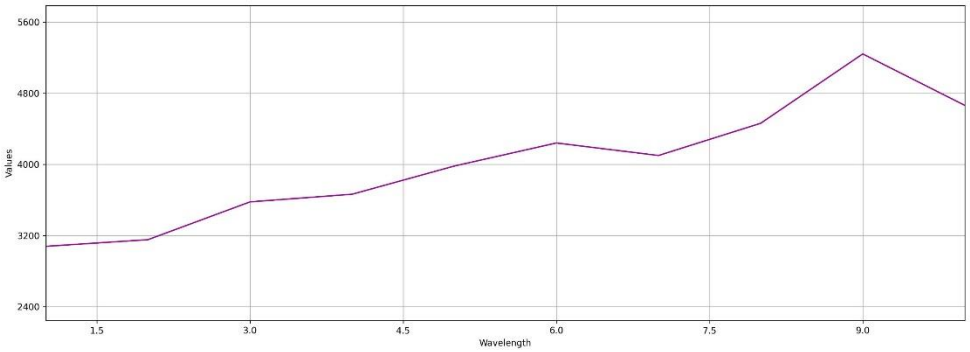
Σχήμα 47. Φασματική υπογραφή οδικού δικτύου της περιοχής μελέτης



Σχήμα 48. Φασματική υπογραφή λιμανιού της περιοχής μελέτης

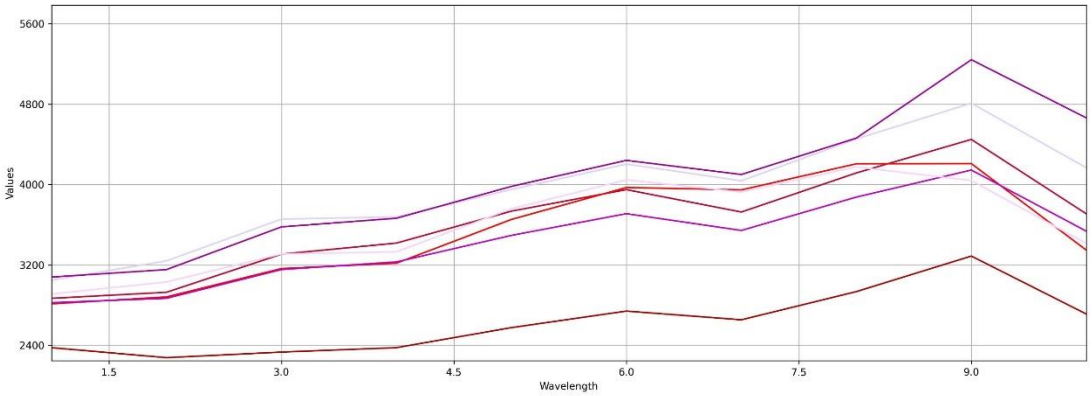


Σχήμα 49. Φασματική υπογραφή αεροδρομίου της περιοχής μελέτης



Σχήμα 50. Φασματική υπογραφή χώρου εξορύξεως ορυκτών της περιοχής μελέτης

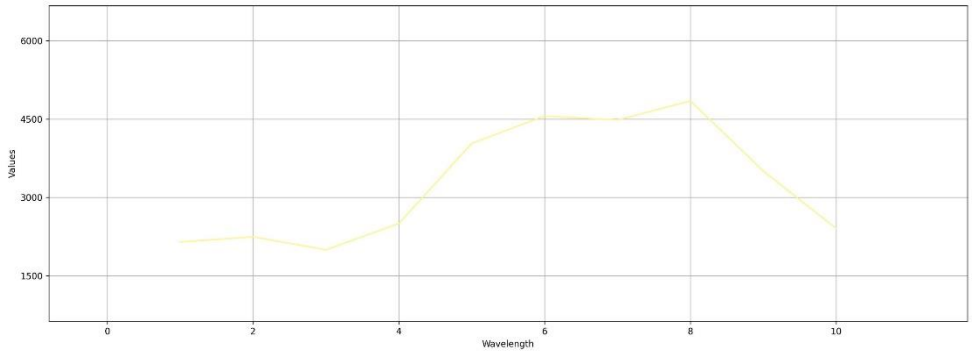
Οι κατηγορίες συνεχής κι ασυνεχής αστικός ιστός, η βιομηχανική ζώνη, το οδικό δίκτυο, το λιμάνι ,το αεροδρόμιο και ο χώρος εξορύξεως ορυκτών αποτελούν λείες ανακλαστικές επιφάνειες κατασκευασμένες από ίδια ή/και παρεμφερή υλικά, με αποτέλεσμα να σημειώνουν στο σύνολό τους παρόμοια συμπεριφορά στο Η/Μ φάσμα, με τον παράγοντα που τις επηρεάζει να είναι η τραχύτητα που εμφανίζουν και η περιεκτικότητά τους σε άλλα υλικά, όπως για παράδειγμα στην περίπτωση του ασυνεχή αστικού ιστού που υπάρχουν μεγαλύτερες αποστάσεις μεταξύ των κτηρίων και σημειώνεται μεσολάβηση μικρών εκτάσεων πρασίνου.



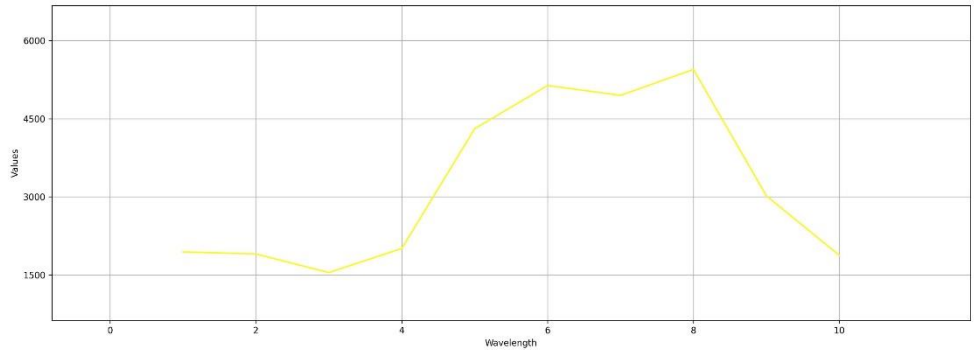
Σχήμα 51. Σύνολο φασματικών υπογραφών κατηγοριών CLC που υπάγονται στην ομάδα των τεχνητών επιφανειών

Σχετικά με τη φασματική υπογραφή των τεχνητών επιφανειών στο σύνολό τους πρόκειται για μια τεθλασμένη γραμμή που παρουσιάζει κατά κύριο λόγο τοπικά μέγιστα ανά φασματικό κανάλι, με τα κανάλια NIR(B8) και SWIR(B12) να εμφανίζουν τοπικά ελάχιστα. Το μικροκυματικό υπέρυθρο SWIR(B11,B12) ενδείκνυται για τη χαρτογράφηση διαφόρων τύπων εδάφους κι είναι το τμήμα όπου η φασματική υπογραφή, σε όλες τις περιπτώσεις των κατηγοριών, παρουσιάζει μεγάλη τυπική απόκλιση μεταξύ των τιμών που χαρακτηρίζουν τα δύο αυτά κανάλια. Στο Σχήμα 51 φαίνεται πως οι φασματικές υπογραφές των μελετώμενων κατηγοριών παρουσιάζουν μικρές αποκλίσεις μεταξύ τους, με πολύ κοντινές μεταξύ τους τιμές, πέραν της υπογραφής του ασυνεχούς αστικού ιστού, που σημειώνει μια μετατόπιση της τάξης των 500 περίπου ραδιομετρικών τιμών. Στο σύνολο, οι αποκλίσεις που εμφανίζονται μεταξύ των φασματικών υπογραφών των κατηγοριών είναι μικρές, επιβεβαιώνοντας ως αυτές οι κατηγορίες δύνανται να ενοποιηθούν και να αποτελέσουν μια κατηγορία, τις τεχνητές επιφάνειες.

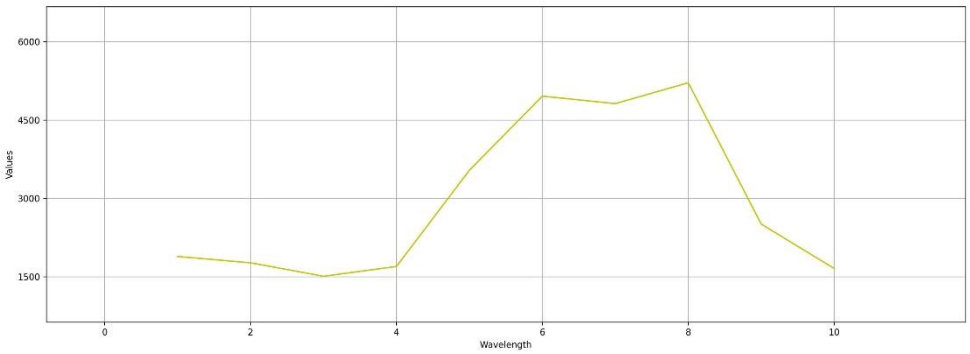
Με τη σειρά τους, οι κατηγορίες αρόσιμη γη, μη αρόσιμη γη κι ορυζώνες, σύμφωνα με την κατηγοριοποίηση του CLC υπάγονται στην υποκατηγορία των γεωργικών περιοχών, αρόσιμη γη. Όπως φαίνεται από τις φασματικές υπογραφές των τριών αυτών κατηγοριών πρόκειται για σχεδόν πανομοιότυπες υπογραφές με μικρές αποκλίσεις που αφορούν ένα επιπλέον “σπάσιμο” σε τοπικό μέγιστο ή ελάχιστο, είτε γίνεται λόγος για μετατόπιση της υπογραφής μιας κατηγορίας σχετικά με τις υπόλοιπες για μικρή διαφορά ραδιομετρικών τιμών.



Σχήμα 52. Φασματική υπογραφή μη αρόσιμης γης της περιοχής μελέτης



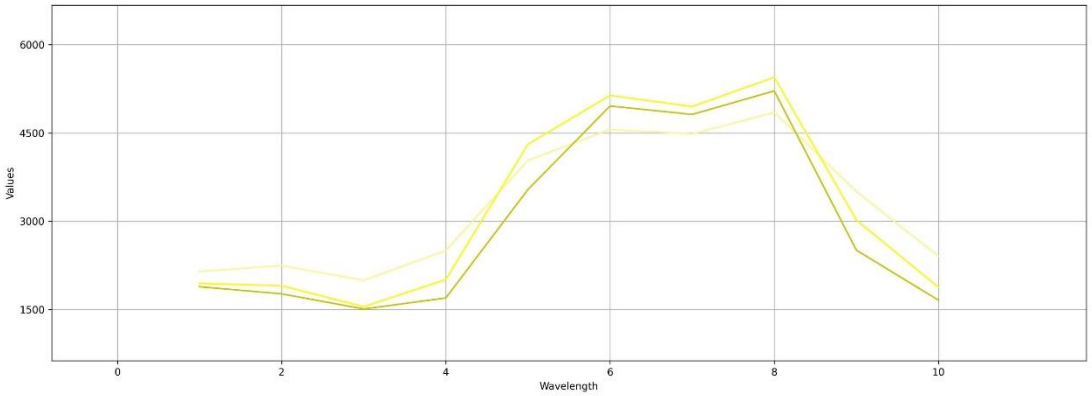
Σχήμα 53. Φασματική υπογραφή αρόσιμης γης της περιοχής μελέτης



Σχήμα 54. Φασματική υπογραφή ορυζώνων της περιοχής μελέτης

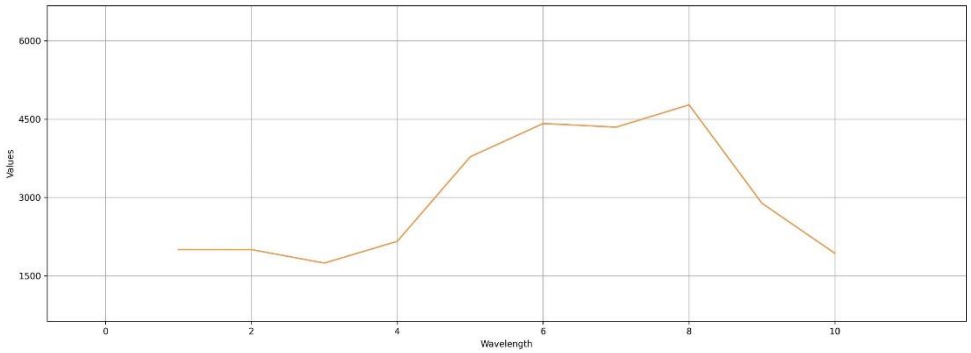
Φαίνεται, λοιπόν, από τη συγκεντρωτική απεικόνιση των φασματικών υπογραφών που σχετίζονται με την αρόσιμη γη πως είναι εφικτή η ενοποίηση των τριών κατηγοριών, δημιουργώντας με αυτόν τον τρόπο αντιπροσωπευτικά σύνολα δεδομένων εκπαίδευσης, χωρίς να εισάγουν τιμές με μεγάλη απόκλιση μεταξύ τους, για την

κατηγορία αρόσιμη γη. Βέβαια, η κατάσταση αυτή οφείλεται και στην εποχή λήψης της πολυφασματικής απεικόνισης.

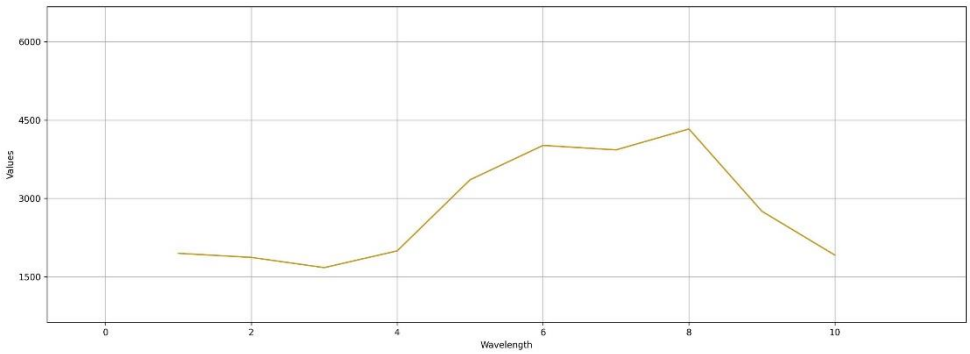


Σχήμα 55. Σύνολο φασματικών υπογραφών κατηγοριών CLC που υπάγονται στην υποκατηγορία της αρόσιμης γης

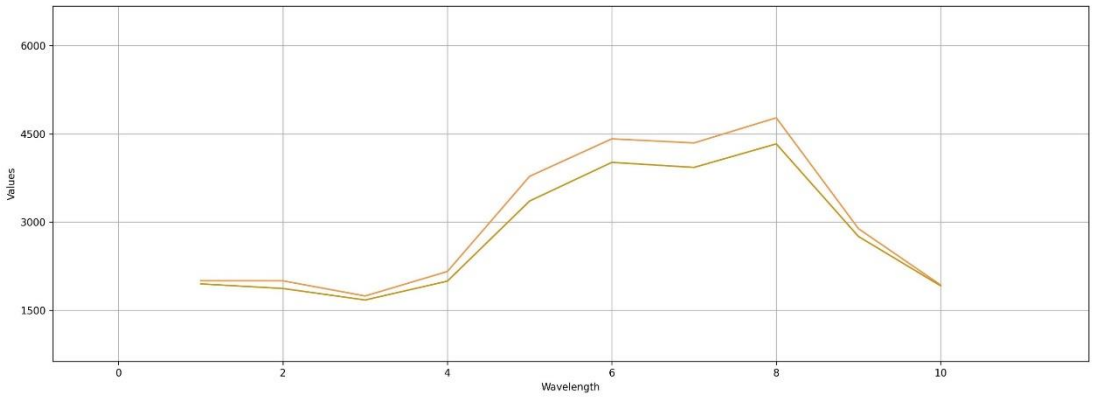
Σύμφωνα με την κατηγοριοποίηση του CLC, η κατηγορία των οπωρώνων και των ελαιώνων υπάγονται στην υποκατηγορία μόνιμες καλλιέργειες των γεωργικών περιοχών, με τις φασματικές τους υπογραφές να επιβεβαιώνουν ότι όχι μόνο λόγω φυσικών ιδιοτήτων, αλλά και λόγω της ανακλαστικότητας τους και των ραδιομετρικών τιμών που λαμβάνουν στο φάσμα, οι δυο κατηγορίες είναι δυνατόν να αποτελούν μια κατηγορία, τις μόνιμες καλλιέργειες.



Σχήμα 56. Φασματική υπογραφή οπωρώνων της περιοχής μελέτης

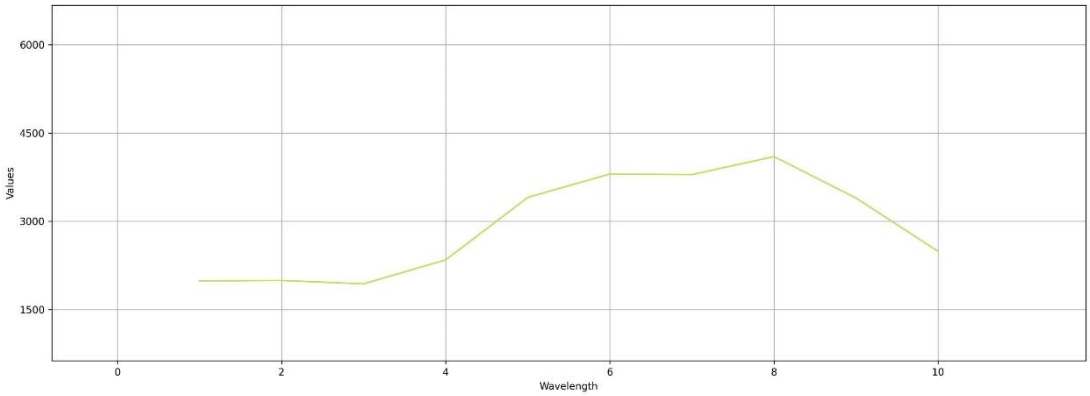


Σχήμα 57. Φασματική υπογραφή ελαιώνων της περιοχής μελέτης



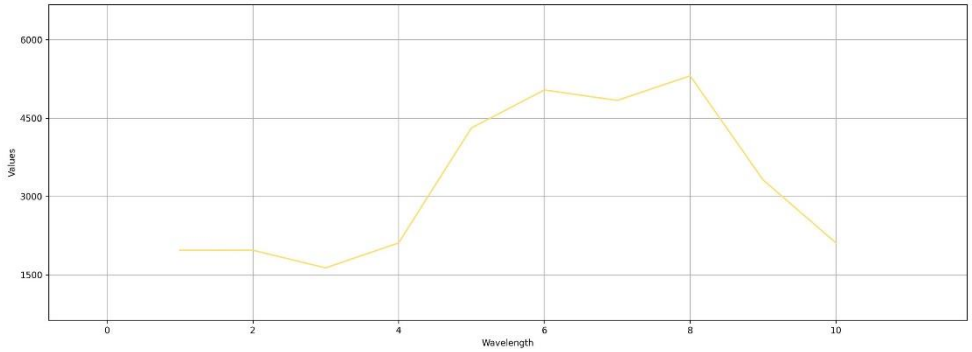
Σχήμα 58. Σύνολο φασματικών υπογραφών κατηγοριών CLC που υπάγονται στην υποκατηγορία μόνιμες καλλιέργειες

Στην περίπτωση της κατηγορίας των λιβαδιών συνεχίζει να ακολουθείται η κατηγοριοποίηση του Corine, με αποτέλεσμα να μην υπάρχει περεταίρω συγχώνευση κλάσεων, αλλά να διατηρείται η κατηγορία λιβάδια ως έχει.

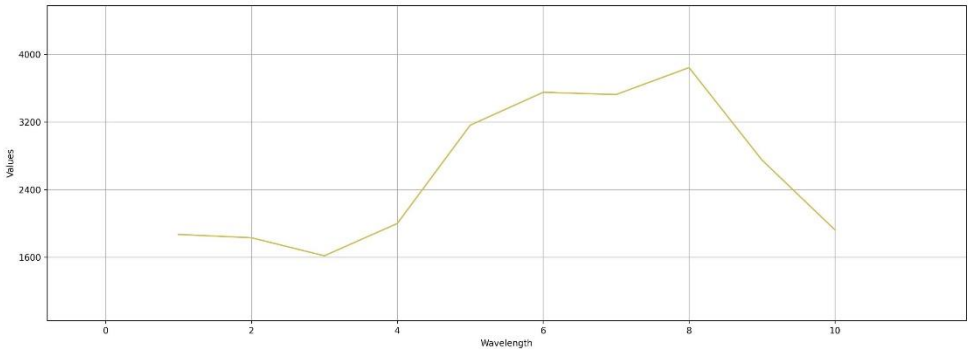


Σχήμα 59. Φασματική υπογραφή λιβαδιών της περιοχής μελέτης

Στη συνέχεια μελέτης των κατηγοριών κάλυψης γης της περιοχής μελέτης συναντώνται κατηγορίες βλάστησης, οι οποίες όμως εμφανίζουν διάφορες ετήσιες καλλιέργειες, σχηματίζοντας αγροτεμάχια με ανομοιομορφία μεταξύ τους. Οι κατηγορίες που υπάγονται στην υποκατηγορία ετερογενείς γεωργικές περιοχές του CLC κι εντοπίζονται στην περιοχή μελέτης είναι οι σύνθετες καλλιέργειες και η γη που καλύπτεται κυρίως από γεωργία με σημαντικές εκτάσεις φυσικής βλάστησης.

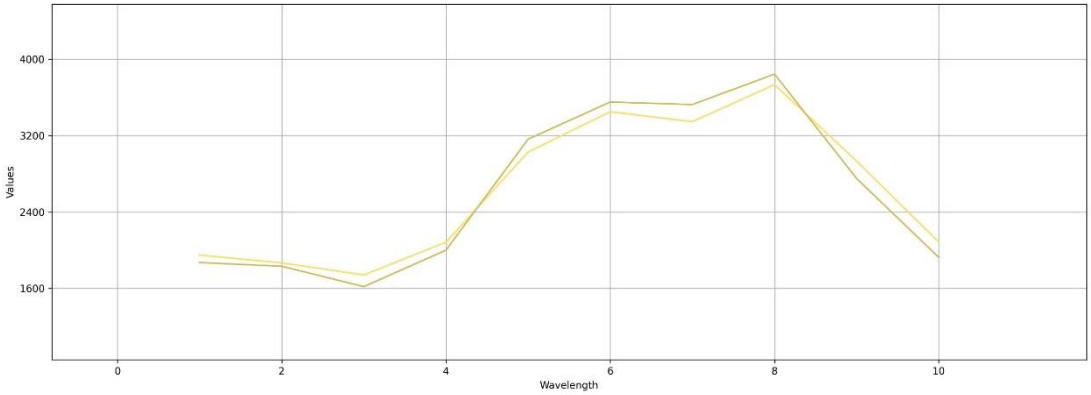


Σχήμα 60.Φασματική υπογραφή σύνθετων καλλιεργειών της περιοχής μελέτης

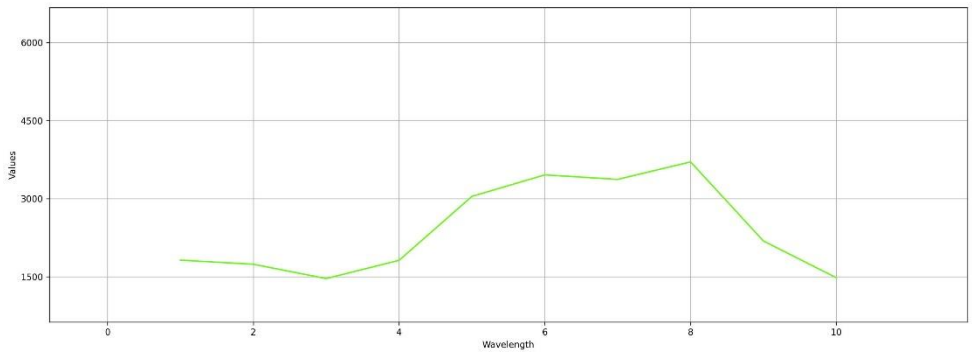


Σχήμα 61. Φασματική υπογραφή γης που καλύπτεται κυρίως από γεωργία με σημαντικές εκτάσεις φυσικής βλάστησης της περιοχής μελέτης

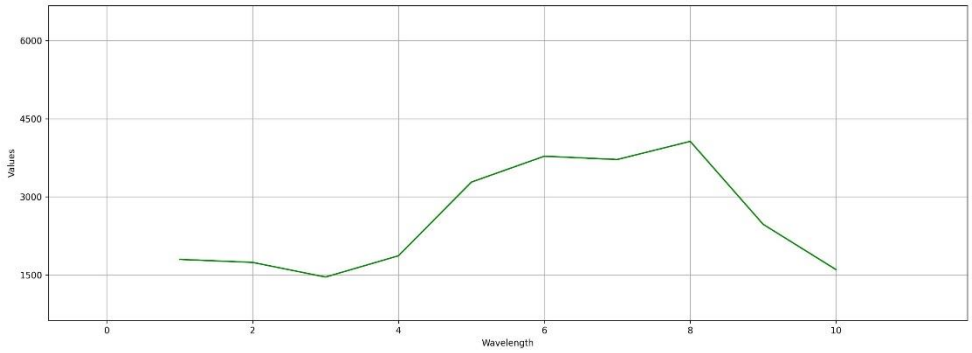
Σε πρώτο επίπεδο ανάλυσης των φασματικών υπογραφών των δυο κατηγοριών διακρίνεται η ομοιότητα μεταξύ τους, με μικρές αποκλίσεις στις ραδιομετρικές τιμές κι όχι σε διαφορετικά τοπικά μέγιστα-ελάχιστα που έχει παρατηρηθεί σε άλλες περιπτώσεις κατηγοριών. Αυτό γίνεται αντιληπτό κι από την επίθεση των φασματικών υπογραφών στο Σχήμα 62, ένα γεγονός που διευκολύνει την ενοποίηση των δυο κατηγοριών, δημιουργώντας την τελική κατηγορία προς μελέτη, την κατηγορία ετερογενείς γεωργικές περιοχές.



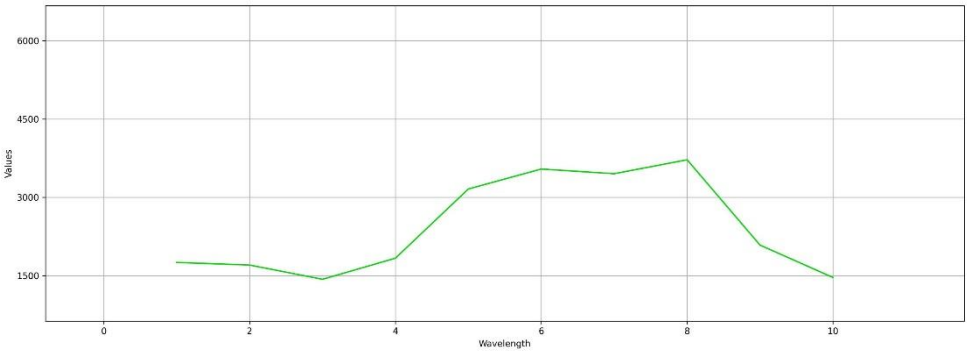
Σχήμα 62. Σύνολο φασματικών υπογραφών κατηγοριών CLC που υπάγονται στην υποκατηγορία ετερογενείς γεωργικές περιοχές



Σχήμα 63. Φασματική υπογραφή δάσους πλατύφυλλων της περιοχής μελέτης

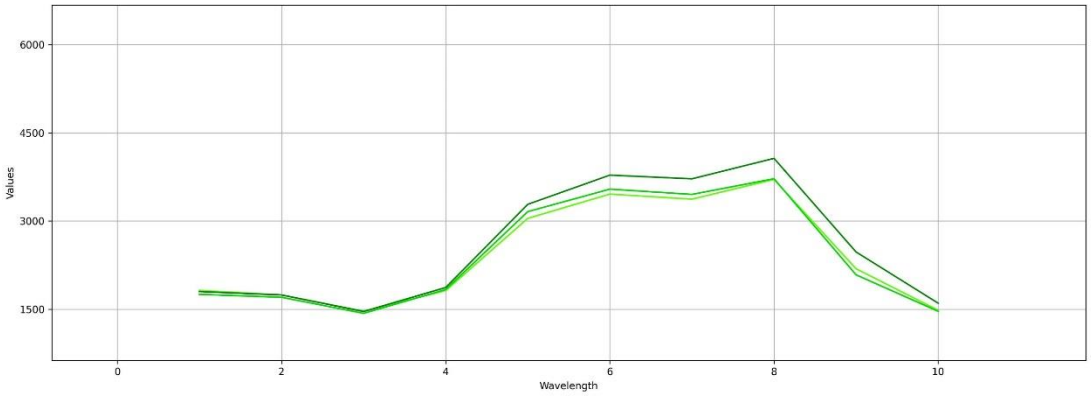


Σχήμα 64. Φασματική υπογραφή δάσους κωνοφόρων της περιοχής μελέτης



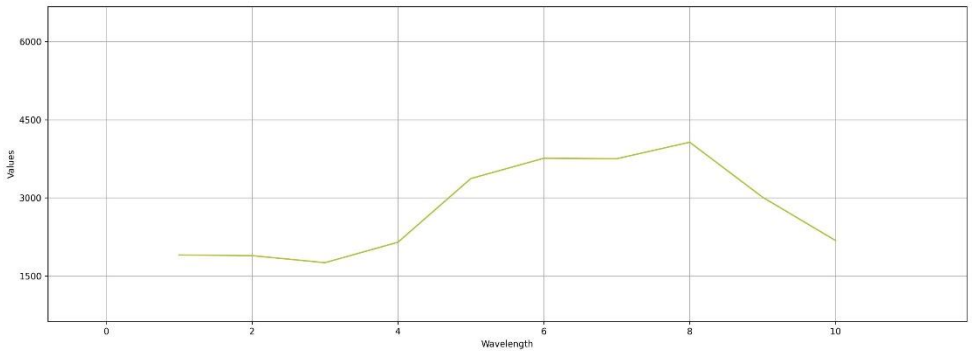
Σχήμα 65. Φασματική υπογραφή μικτού δάσους της περιοχής μελέτης

Όπως προκύπτει από την κατηγοριοποίηση των κλάσεων από το CLC κι επιβεβαιώνει η φασματική υπογραφή των τριών κατηγοριών ,σχετικών με δασικές εκτάσεις, τα κωνοφόρα ,τα πλατύφυλλα και τα μικτά δάση, για τις ανάγκες των πειραμάτων στη συνέχεια της εργασίας, θα αποτελούν μια κλάση, τα δάση.

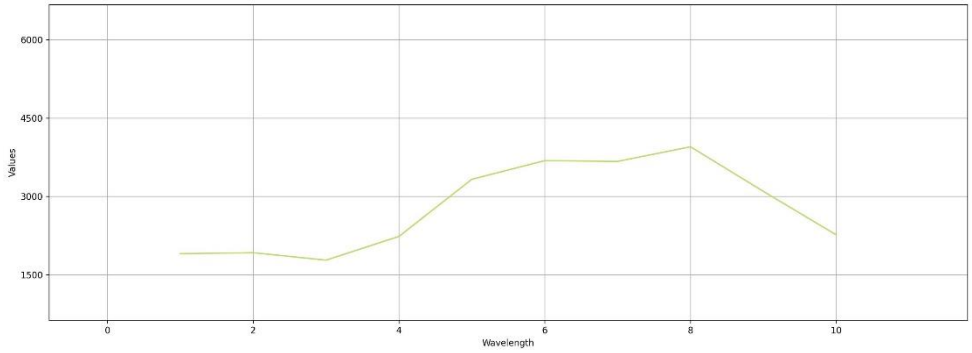


Σχήμα 66. Σύνολο φασματικών υπογραφών κατηγοριών CLC που υπάγονται στην υποκατηγορία δάση

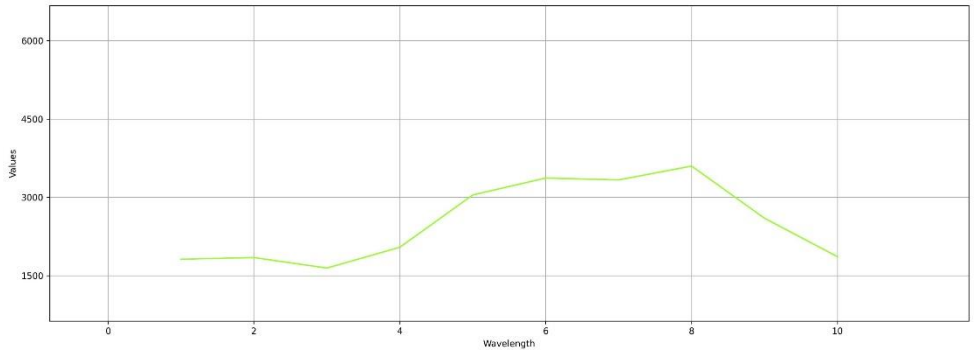
Επόμενες κατηγορίες στην περιοχή μελέτης είναι οι φυσικοί βοσκότοποι, η σκληροφυλλική βλάστηση και δασώδεις και θαμνώδεις εκτάσεις, οι οποίες ανήκουν στην υποκατηγορία συνδυασμοί θαμνώδους και/ή ποώδους βλάστησης του CLC.



Σχήμα 67. Φασματική υπογραφή φυσικών βοσκότοπων της περιοχής μελέτης

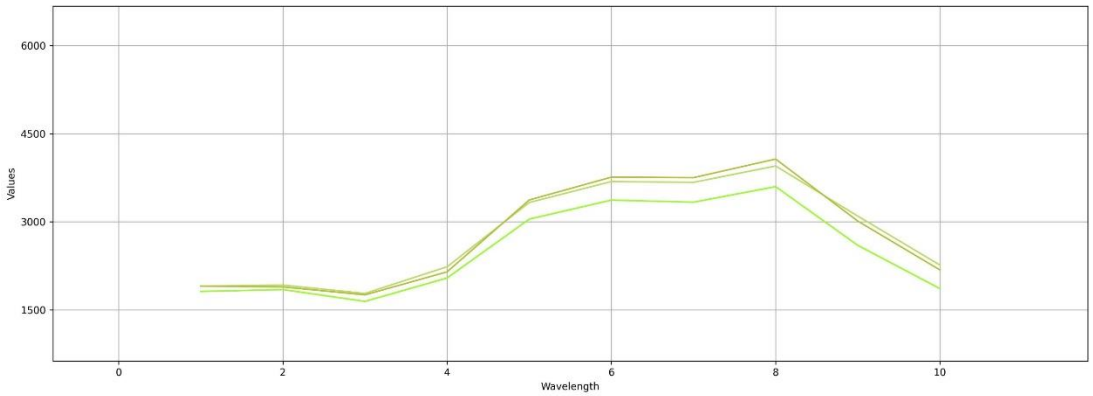


Σχήμα 68. Φασματική υπογραφή σκληροφυλλικής βλάστησης της περιοχής μελέτης



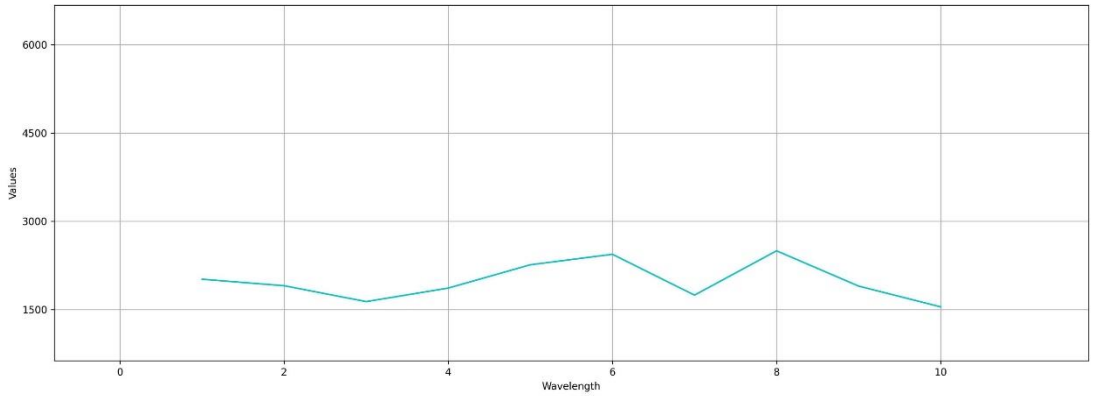
Σχήμα 69. Φασματική υπογραφή δασώδων - θαμνώδων εκτάσεων της περιοχής μελέτης

Παρατηρώντας μια προς μια τις φασματικές υπογραφές (Σχήμα 6978-Σχήμα 6880), αλλά και με την παράθεσή τους σε ένα διάγραμμα (Σχήμα 70), σημειώνεται πως δεν παρατηρούνται σημαντικές διαφορές ή αποκλίσεις, οπότε αυτές οι τρεις κατηγορίες δύνανται να αποτελούν μια, χωρίς να εισάγονται ακραίες τιμές κατά τη συλλογή δεδομένων εκπαίδευσης κι ελέγχου από αυτές τις εκτάσεις ,για την υλοποίηση πειραμάτων ταξινόμησης κι εκτίμησης της ακρίβειας του αποτελέσματος.

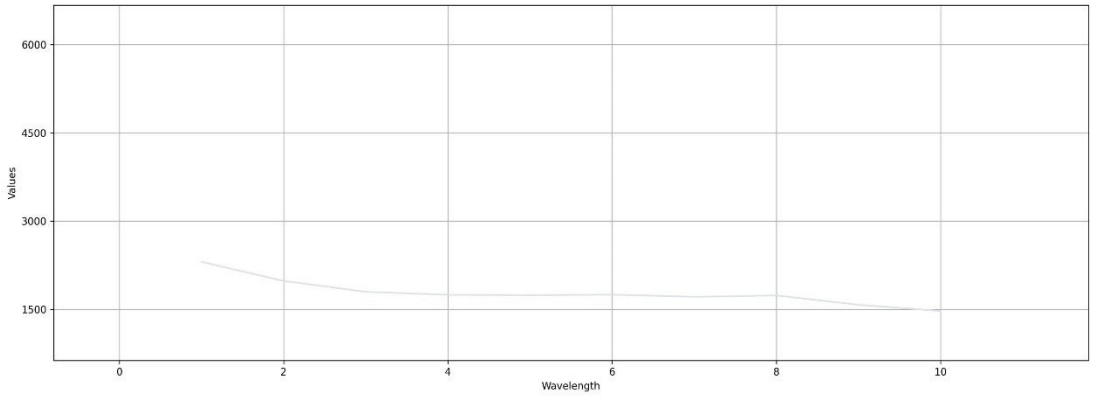


Σχήμα 70. Σύνολο φασματικών υπογραφών κατηγοριών CLC που υπάγονται στην υποκατηγορία συνδυασμοί θαμνώδους και/ή ποώδους βλάστησης

Οι δυο επόμενες κατηγορίες, χερσαία ύδατα και θαλάσσια ύδατα, περιγράφουν τις υδάτινες επιφάνειες της περιοχής μελέτης, με βασικό παράγοντα που τις ξεχωρίζει να είναι το βάθος τους και τι υλικά μπορούν να σημειωθούν εντός αυτού και να επηρεάσουν την ανακλαστικότητα κάθε επιφάνειας. Επομένως, γίνεται αντιληπτό κι από τις φασματικές υπογραφές τους, πως σε αυτή την περίπτωση κατηγοριών δε γίνεται συνένωση.



Σχήμα 71. Φασματική υπογραφή χερσαίων υδάτων της περιοχής μελέτης



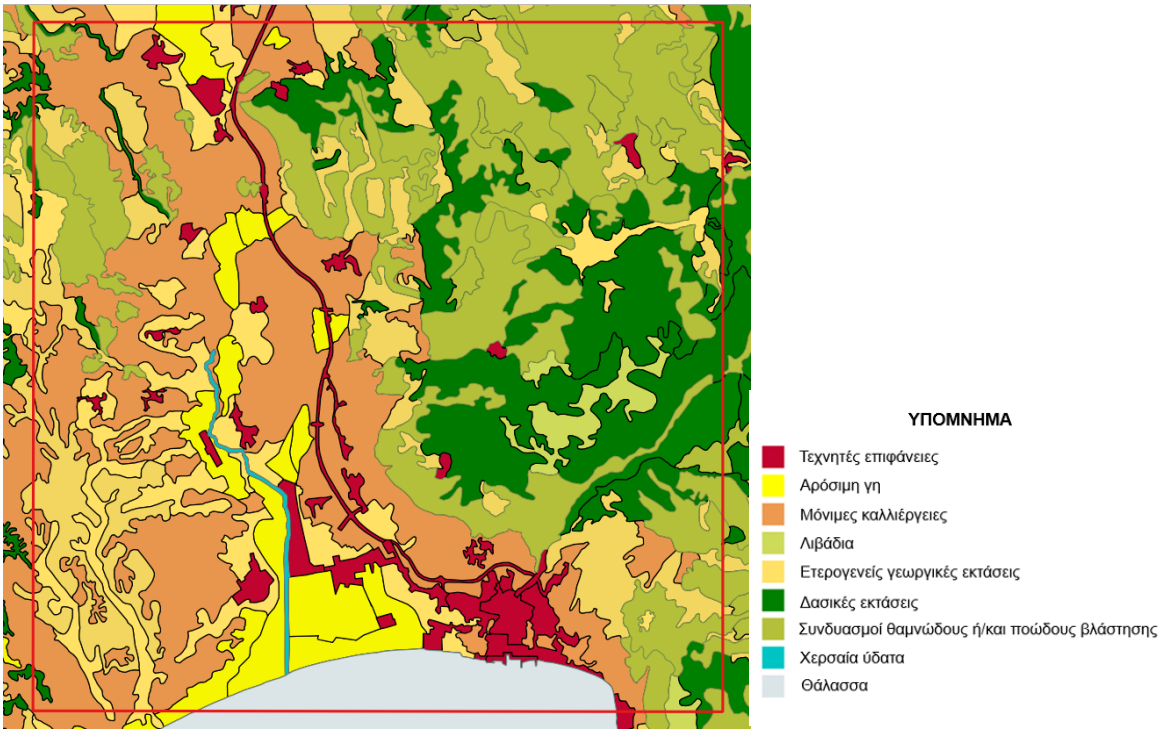
Σχήμα 72. Φασματική υπογραφή θαλάσσιων υδάτων της περιοχής μελέτης

Συνοψίζοντας, οι τελικές κατηγορίες προς μελέτη θα είναι

	Κατηγορία	Υποκατηγορίες που περιλαμβάνει
1	Τεχνητές επιφάνειες	Συνεχής αστικός ιστός, Ασυνεχής αστικός ιστός, Βιομηχανική ζώνη, Οδικό δίκτυο, Λιμάνι, Αεροδρόμιο, Ορυχείο
2	Αρόσιμη γη	Μη αρόσιμη γη, Αρόσιμη γη, Ορυζώνες
3	Μόνιμες καλλιέργειες	Οπωρώνες, Ελαιώνες
4	Λιβάδια	Λιβάδια
5	Ετερογενείς γεωργικές εκτάσεις	Σύνθετες καλλιέργειες, Γεωργικές εκτάσεις & φυσική βλάστηση
6	Δάση	Δάσος πλατύφυλλων, Δάσος κωνοφόρων, Μικτό δάσος
7	Συνδυασμοί θαμνώδους ή/και ποώδους βλάστησης	Φυσικοί βοσκότοποι, Σκληροφυλλική βλάστηση, Δασώδεις-θαμνώδεις εκτάσεις
8	Χερσαία ύδατα	Ροές υδάτων
9	Θαλάσσια ύδατα	Θάλασσα

Πίνακας 4. Κατηγορίες κάλυψης γης CLC προς μελέτη

Ακολουθώντας την κατηγοριοποίηση των χρήσεων γης, όπως παρουσιάζεται στον Πίνακας 4 και το σχετικό συμβολισμό από το CLC, ο τελικός χάρτης χρήσεων/καλύψεων γης που περιγράφει την περιοχή ενδιαφέροντος είναι της μορφής:



Σχήμα 73. Τελική μορφή του θεματικού χάρτη CLC μετά την ενοποίηση των κατηγοριών της περιοχής ενδιαφέροντος

ΚΕΦΑΛΑΙΟ 6. Μεθοδολογία

Σε αυτό το κεφάλαιο παρουσιάζεται βήμα προς βήμα η μεθοδολογία που ακολουθείται για την υλοποίηση πειραματικών εφαρμογών επιβλεπόμενης ταξινόμησης με μη ισορροπημένα δεδομένα, ώστε να αξιολογηθεί η ακρίβεια των αποτελεσμάτων και πως αυτή επηρεάζεται ανάλογα με το εάν έχει εφαρμοσθεί υπερδειγματοληψία ή υποδειγματοληψία στο σύνολο των δεδομένων εκπαίδευσης, για την αντιμετώπιση του προβλήματος των μη ισορροπημένων δεδομένων.

6.1.Λεπτομέρειες υλοποίησης

Έχοντας ολοκληρώσει την ερμηνεία της απεικόνισης κι έχοντας καταλήξει στις τελικές προς μελέτη κατηγορίες καλύψεων/χρήσεων γης, ακολουθεί ένα από τα σημαντικότερα και βασικότερα βήματα της μεθοδολογίας, η συλλογή των δεδομένων εκπαίδευσης των αλγορίθμων ταξινόμησης. Η δημιουργία των πολυγώνων μέσω του Semi-Automatic Classification Plugin γίνεται με ιδιαίτερη προσοχή, ώστε τα εικονοστοιχεία που συμπεριλαμβάνει κάθε πολύγωνο, να είναι ομοιόμορφα μεταξύ τους και να θεωρούνται αντιπροσωπευτικά για την κατηγορία που σχεδιάζονται. Αντίστοιχη λογική ακολουθείται και για το σχηματισμό πολυγώνων για τον έλεγχο των ταξινομήσεων (test set). Έχοντας σχεδιάσει ένα ισορροπημένο σύνολο δεδομένων, με τα πολύγωνα να καλύπτουν ίδια έκταση σε όλες τις κατηγορίες χρήσεων γης, ακολουθεί η εφαρμογή επιβλεπόμενης ταξινόμησης με τον αλγόριθμο Random Forest (RF). Ο θεματικός χάρτης που παράγεται από τα πειράματα ταξινόμησης συγκρίνεται με το CLC και τον κυρωμένο δασικό χάρτη της περιοχής μελέτης, αλλά το μεγαλύτερο ζητούμενο σε αυτή την εργασία είναι η αξιολόγηση της ποσοτικής ακρίβειας που επιτυγχάνεται. Για το λόγο αυτό ακολουθεί μια σειρά από υπολογισμούς για τον προσδιορισμό των μετρικών precision, recall ,F1-score, όπου σύμφωνα με τους ορισμούς που δίνονται από την Τηλεπισκόπηση, η μετρική precision είναι αντίστοιχη της ακρίβειας του παραγωγού και η recall είναι ίδια με την ακρίβεια του χρήστη, με το μέγεθος F1-score να προκύπτει ως ένα μέγεθος που συμπεριλαμβάνει τις δυο προηγούμενες μετρικές. Η συνολική ακρίβεια που σημειώνει η ταξινόμηση, όπως προκύπτει από τον πίνακα σύγχυσης δε λαμβάνεται υπόψιν, καθώς πρόκειται για ένα μέγεθος που επηρεάζεται σημαντικά από μεμονωμένα υψηλές ακρίβειες, όπως για παράδειγμα την υψηλή ακρίβεια που συνηθίζεται να σημειώνεται σε κατηγορίες όπως αυτή του νερού (συνήθως $\approx$1).

Έχοντας ολοκληρώσει το στάδιο αξιολόγησης της επιβλεπόμενης ταξινόμησης με ισορροπημένο σύνολο δεδομένων, επόμενο βήμα είναι η εφαρμογή του δείκτη αναλογίας ισορροπίας, που έχει σχεδιαστεί αποκλειστικά για τις ανάγκες της παρούσας εργασίας, βασιζόμενος στο δείκτη ανισορροπίας, imbalance ratio (IR), που περιγράφεται σε πληθώρα μελετών, όπως σημειώνεται και στο αντίστοιχο κεφάλαιο των βιβλιογραφικών αναφορών. Σκοπός λοιπόν, του δείκτη αναλογίας είναι μέσα από διαδοχικούς υπολογισμούς, που αφορούν το πλήθος των πολυγώνων εκπαίδευσης, να υποδείξει σε ποιες περιπτώσεις το πλήθος των πολυγώνων κάθε κατηγορίας, μεμονωμένα, οδηγεί σε μέθοδο υποδειγματοληψίας ή σε μέθοδο υπερδειγματοληψίας ή ακόμα κι αν υπάρχει ενδεχόμενο συνδυασμού των δυο μεθόδων, για να προκύψουν συμπεράσματα για το πως και κατά πόσο τελικά επηρεάζεται η ακρίβεια της ταξινόμησης. Ανάλογα με το τι προκύπτει από το δείκτη αναλογίας αφαιρούνται ή προστίθενται δείγματα στο σύνολο των δεδομένων εκπαίδευσης κι ακολουθεί μια σειρά από εφαρμογές επιβλεπόμενης ταξινόμησης και η ποσοτική αξιολόγηση του αποτελέσματος με τον υπολογισμό των μετρικών που αναφέρθηκαν προηγουμένως.

6.1.1.Δημιουργία πολυγώνων εκπαίδευσης & ελέγχου

Όπως επισημαίνεται καθ' όλη την έκταση της παρούσας εργασίας, βασικός οδηγός της διαδικασίας των πειραμάτων είναι τα πολύγωνα εκπαίδευσης που χρησιμοποιούνται, ώστε να εκπαιδευτεί ο αλγόριθμος στα χαρακτηριστικά των κατηγοριών κάλυψης/χρήσεων γης, που εντοπίζονται στην περιοχή μελέτης. Τα πειράματα ξεκινούν με ένα ισορροπημένο, ως προς την έκταση των πολυγώνων, σύνολο δεδομένων εκπαίδευσης και στη συνέχεια με την εφαρμογή του δείκτη αναλογίας ισορροπίας, που αφορά το πλήθος των πολυγώνων, αποφαίνεται αν τα επακόλουθα πειράματα θα γίνονται με τη μέθοδο υπερδειγματοληψίας ή τη μέθοδο υποδειγματοληψίας. Αυτό που αξίζει να αναφερθεί είναι πως τα δεδομένα που χρησιμοποιούνται και η αντιμετώπισή τους ,τόσο κατά το στάδιο της προ-επεξεργασίας, όσο και κατά την εφαρμογή των πειραμάτων, σκόπιμα προσεγγίζουν ρεαλιστικές συνθήκες, καθώς η ανισορροπία των δεδομένων είναι ένα σύνηθες πρόβλημα, ιδίως σε περιπτώσεις που έστω μια εκ των παρατηρούμενων κατηγοριών σημειώνει μικρή έκταση, συγκριτικά με τις υπόλοιπες.

Το πρόβλημα περιορισμένης έκτασης συναντάται και στην περιοχή μελέτης, ιδιαίτερα στην περίπτωση που γίνεται λόγος για τα χερσαία ύδατα, γεγονός που γίνεται αντιληπτό και μόνο με την παρατήρηση του Corine Land Cover, όπως αυτό έχει διαμορφωθεί μετά την ενοποίηση των κατηγοριών (Σχήμα 84). Λόγω της περιορισμένης έκτασης που σημειώνει η κατηγορία των χερσαίων υδάτων, η διαδικασία δημιουργίας πολυγώνων εκπαίδευσης ξεκινά από αυτή, με απώτερο σκοπό να εκτιμηθεί η έκταση που είναι δυνατό να καλυφθεί μέσω των πολυγώνων και να αποτελέσει τη βάση για τη δημιουργία των υπολοίπων, χωρίς να αποκλίνουν από αυτή την έκταση. Βέβαια, πρόκειται για μια κατηγορία, που παρά το γεγονός ότι ανήκει στην κατηγορία των νερών, δεν παρουσιάζει παρόμοια συμπεριφορά στο φάσμα με την κατηγορία της θάλασσας (Σχήμα 71 & Σχήμα 72) κι αυτό διότι το βάθος που σημειώνεται σε κάθε περίπτωση δείγματος επιδρά σε σημαντικό βαθμό στο πως θα διαμορφωθεί η φασματική υπογραφή της κατηγορίας. Εστιάζοντας στην κατηγορία των χερσαίων υδάτων, τόσο το βάθος, όσο και τα φερτά υλικά που μπορεί να εμφανίζονται εντός τους, αλλά και η περιεκτικότητά τους σε είδη πρασίνου, έχουν ως αποτέλεσμα όχι μόνο την απόκλιση των φασματικών υπογραφών μεταξύ των δυο υδάτινων κατηγοριών, αλλά και την απόκλιση των φασματικών υπογραφών ανάμεσα στα πολύγωνα εκπαίδευσης (κατ' επέκταση κι ελέγχου) της ίδια κατηγορίας, των χερσαίων υδάτων, κατά μήκος της ροής τους. Γίνεται ιδιαίτερη αναφορά στην κατηγορία των χερσαίων υδάτων, καθώς είναι η μόνη κλάση που εμφανίζει αυτή την "ιδιομορφία", ενώ για τις υπόλοιπες τα πολύγωνα εκπαίδευσης κι ελέγχου καταγράφουν σχεδόν ταυτόσημες φασματικές υπογραφές. Η συνθήκη αυτή προϊδεάζει τον παρατηρητή για τυχόν αστοχίες που είναι πιθανό να σημειωθούν στο στάδιο αξιολόγησης του αποτελέσματος της ταξινόμησης.

Βασικό εργαλείο για την εκτίμηση κι αξιολόγηση της ομοιομορφίας που σχηματίζεται μεταξύ των πολυγώνων εκπαίδευσης μιας κατηγορίας κάλυψης γης είναι η φασματική υπογραφή. Όπως είναι γνωστό από τη βιβλιογραφία, οι φασματικές υπογραφές δίνουν τη δυνατότητα να αξιολογούνται οι πληροφορίες που προκύπτουν για τις φυσικές ή χημικές ιδιότητες ενός αντικειμένου/μιας κατηγορίας και να διαχωρίζονται οι κατηγορίες/τα αντικείμενα μεταξύ τους. Αυτός είναι ο βασικός λόγος για τον οποίο αναπτύχθηκε το αντίστοιχο κεφάλαιο, ενώ ταυτόχρονα, λόγω παρόμοιων χαρακτηριστικών προσφέρει τη δυνατότητα προ-εκτίμησης σε ποιες περιπτώσεις κατηγοριών υπάρχει πιθανότητα σύγχυσης και λανθασμένης ταξινόμησης εικονοστοιχείων.

Ο λόγος, λοιπόν, για τον οποίο σχολιάζεται η ανάγκη για καταγραφή πολυγώνων εκπαίδευσης με ίδιες φασματικές υπογραφές είναι γιατί η ομοιομορφία αποτελεί ένα από τα βασικά χαρακτηριστικά που ορίζουν ένα σύνολο δεδομένων εκπαίδευσης ποιοτικό. Είναι σαφές πως το σύνολο των δεδομένων εκπαίδευσης πρέπει να απαρτίζεται από ομοιόμορφα κι αντιπροσωπευτικά πολύγωνα, ώστε ο αλγόριθμος ταξινόμησης να καταγράψει υψηλή απόδοση.

Αντίστοιχη λογική ακολουθείται και κατά το σχεδιασμό των πολυγώνων ελέγχου, τα οποία καλούνται να ελέγξουν ποια εικονοστοιχεία ταξινομήθηκαν ορθώς σε μια κατηγορία και ποια όχι. Η σημαντική διαφορά που σημειώνεται είναι πως υπό κανονικές συνθήκες τα δεδομένα ελέγχου συλλέγονται in situ, δηλαδή με επίσκεψη στο πεδίο, ωστόσο, για τις ανάγκες της εργασίας αυτό δεν είναι εφικτό και η διαδικασία συλλογής τους γίνεται αντίστοιχα με αυτή των πολυγώνων εκπαίδευσης, δηλαδή με ερμηνεία της πολυφασματικής απεικόνισης, με υποδείξεις του Corine Land Cover και σε ορισμένες περιπτώσεις με τη συμβολή του Google Satellite.

6.1.2.Δείκτης Αναλογίας Ισορροπίας (Balance Ratio Indice, BRI)

Πριν την εκκίνηση των πειραμάτων κρίνεται σκόπιμο να γίνει μια σύντομη περιγραφή του δείκτη αναλογίας ισορροπίας με ένα παράδειγμα, στο οποίο έστω ότι γίνεται λόγος για 3 κατηγορίες κάλυψης γης, έναντι των 9 κατηγοριών της περιοχής μελέτης της εργασίας. Γίνεται υπόθεση πως η κατηγορία Α απαρτίζεται από 100 πολύγωνα στο σύνολο δεδομένων εκπαίδευσης, η κατηγορία Β από 300 πολύγωνα και τέλος, την κατηγορία Γ την αντιπροσωπεύουν 50 πολύγωνα στο σετ, συνεπώς το σετ δεδομένων εκπαίδευσης αποτελείται από 450 πολύγωνα. Σε δεύτερο χρόνο υπολογίζεται η συχνότητα εμφάνισης κάθε κατηγορίας στο σύνολο των δεδομένων, άρα

- f(A) = 100 / 450 = 0.222
- f(B) = 300 / 450 = 0.667
- f(Γ) = 50 / 450 = 0.111

Ο δείκτης αναλογίας για κάθε κατηγορία μπορεί να υπολογιστεί συγκρίνοντας την αναλογία κάθε κατηγορίας, με την ιδανική αναλογία, εάν τα δεδομένα ήταν τέλεια ισορροπημένα. Η ιδανική αναλογία είναι: 1 / συνολικό αριθμό των κατηγοριών (στο παράδειγμα 3, στην περιοχή μελέτης 9). Οπότε, με τρεις κατηγορίες, η ιδανική αναλογία για κάθε κατηγορία είναι 1 / 3 = 0.333.

- Δείκτης αναλογίας για την κατηγορία Α: 0.222 / 0.333 = 0.667
- Δείκτης αναλογίας για την κατηγορία Β: 0.667 / 0.333 = 2.003
- Δείκτης αναλογίας για την κατηγορία Γ: 0.111 / 0.333 = 0.333

Ο δείκτης αναλογίας BRI δείχνει με ποιον τρόπο συγκρίνεται κάθε κατηγορία με ένα απόλυτα ισορροπημένο σύνολο:

1. Μια αναλογία κοντά στο 1 δείχνει μια ισορροπημένη κατάσταση.
2. Μια αναλογία σημαντικά μικρότερη από 1 υποδηλώνει υποεκπροσώπηση της κατηγορίας (και πιθανή ανάγκη για υπερδειγματοληψία).
3. Μια αναλογία σημαντικά μεγαλύτερη από 1 υποδηλώνει υπερεκπροσώπηση (και πιθανή ανάγκη για υποδειγματοληψία).

Μπορεί με βάση το παράδειγμα να αποφασιστεί να ακολουθήσει υπερδειγματοληψία στην Κατηγορία Α και την Κατηγορία Γ, για να αυξηθεί η εκπροσώπησή τους. Από την

άλλη, υπάρχει το ενδεχόμενο υποδειγματοληψίας της Κατηγορίας Β ή απλά να μείνει ως έχει. Εξαρτάται σε κάθε περίπτωση τι επιδιώκει κι επιθυμεί ο παρατηρητής.

6.2.Τρόποι Αξιολόγησης των Αποτελεσμάτων

Είναι απαραίτητο τα μοντέλα να αξιολογούνται κατά την εφαρμογή τους στο σετ ελέγχου μέσω διάφορων μετρικών που προκύπτουν από τον πίνακα σύγχυσης, αλλά κι από τον τελικό θεματικό χάρτη που παράγουν, δηλαδή την απεικόνιση των αποτελεσμάτων της ταξινόμησης των κατηγοριών χρήσεων/καλύψεων γης, με το ενδιαφέρον να επικεντρώνεται στην ποσοτική αξιολόγηση του αποτελέσματος.

6.2.1.Πίνακας σύγχυσης – Μετρικές για τα δεδομένα ελέγχου

Για την αξιολόγηση της απόδοσης των αλγορίθμων είναι σημαντικό να οριστούν ποσοτικά μέτρα αξιολόγησης (Daudt, 2020). Για την δυαδική περίπτωση ταξινόμησης, έστω ότι y_n ~ Y είναι η n-ιοστή ετικέτα στο σετ απεικονίσεων είτε στο σύνολο εκπαίδευσης είτε στο σύνολο ελέγχου και η $\hat{y}_n$ ~ $\hat{Y}$ η n-ιοστή ετικέτα που προβλέπει ο αλγόριθμος υπό εξέταση, όπου y_n, $\hat{y}_n \in \{0,1\}$. Οι τέσσερις πιθανοί συνδυασμοί δίνονται στον πίνακα που ακολουθεί, που συνιστά τον πίνακα σύγχυσης (confusion matrix).

	y_n=0	y_n=1
$\hat{y}_n$=0	Αληθώς αρνητικό (True Negative-TN)	Ψευδώς Αρνητικό (False Negative-FN)
$\hat{y}_n$=1	Ψευδώς θετικό (False Positive-FP)	Αληθώς θετικό (True Positive-TP)

Πίνακας 5. Πίνακας σύγχυσης για 2 κατηγορίες

Ως αληθώς θετικό ορίζεται το εικονοστοιχείο που απεικονίζει αλλαγή και έχει προβλεφθεί ως αλλαγή, ενώ ψευδώς θετικό θεωρείται όταν έχει προβλεφθεί ως αλλαγή ενώ στην πραγματικότητα δεν είναι. Ως αληθώς αρνητικό ορίζεται ένα εικονοστοιχείο που δεν αντιστοιχεί σε αλλαγή και ούτε έχει προβλεφθεί από το μοντέλο ως αλλαγή. Τέλος, ψευδώς αρνητικό θεωρείται όταν έχει προβλεφθεί από το μοντέλο ως μη αλλαγή, ενώ στην πραγματικότητα είναι αλλαγή.

Χρησιμοποιώντας τα στοιχεία του πίνακα σύγχυσης προκύπτουν διάφορες ενδιαφέρουσες μετρικές. Η ακρίβεια (accuracy) ή συνολική ακρίβεια (overall accuracy) αφορά το ποσοστό των σωστά ταξινομημένων εικονοψηφίων στο σύνολο των εικονοψηφίων. Μπορεί να πάρει τιμές από 0 έως και 1, με το 1 να είναι η καλύτερη πιθανή τιμή.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Παρόλα αυτά, όπως έχει ήδη αναφερθεί, η συνολική ακρίβεια της ταξινόμησης δε θα είναι από τις βασικές μετρικές που θα απασχολήσουν στην αξιολόγηση των πειραμάτων.

Από την άλλη, μια από τις μετρικές που κρίνονται σημαντικές είναι το precision ,το οποίο καλείται να προσδιορίσει τα ορθά κατανεμημένα εικονοστοιχεία. Δηλαδή πόσα

εικονοστοιχεία στην ταξινομημένη απεικόνιση ανήκουν όντως στην κατηγορία που ταξινομήθηκαν βάσει τα δεδομένα ελέγχου. Η μετρική αυτή επίσης ανήκει στο πεδίο τιμών από 0 έως 1, με τη μονάδα να συνιστά την καλύτερη πιθανή τιμή.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Η παρούσα εργασία στέκεται μεταξύ άλλων και στο μέγεθος του recall, που μετρά πόσα εικονοστοιχεία ανιχνεύθηκαν σωστά από τον αλγόριθμο. Ωστόσο, να σημειωθεί πως δεν παρέχει πληροφορία για το πλήθος των ψευδώς θετικά κατανεμημένων εικονοστοιχείων. Όπως τα προηγούμενα μεγέθη αξιολόγησης, έτσι και το recall λαμβάνει τιμές από το 0 έως το 1, με την περίπτωση να είναι ίσο με 1 να είναι η καλύτερη πιθανή τιμή.

$$\text{Recall} = \frac{TP}{TP+FN}$$

Τέλος, άλλη μια ποσότητα που θα εφαρμοστεί για την αξιολόγηση των πειραμάτων είναι το F1-score, που ως ο αρμονικός μέσος των δυο μετρικών precision και recall εξάγει ένα ισορροπημένο αποτέλεσμα μεταξύ αυτών των δυο. Σε περίπτωση που μια από τις δυο μετρικές έχει χαμηλή τιμή, θα μειωθεί το F1-score, χαρακτηριστικό που την καθιστά μια εύρωστη μετρική για την ποσοτικοποίηση των αλγορίθμων ταξινόμησης.

Ένας σημαντικός λόγος, που καθιστά το F1-score προτιμότερο μέγεθος για την αξιολόγηση των αποτελεσμάτων, είναι επειδή χαρακτηρίζεται καλύτερο από τη συνολική ακρίβεια όταν υπάρχει ανισορροπία μεταξύ των παραδειγμάτων κάθε κατηγορίας, μιας και η συνολική ακρίβεια είναι ισχυρά προκατειλημμένη προς την κατηγορία με τα περισσότερα παραδείγματα. Παίρνει τιμές από το 0 έως το 1, με την περίπτωση να είναι ίσο με 1 να είναι η καλύτερη πιθανή τιμή.

$$\text{F1-score} = \frac{2TP}{2TP+FP+FN} \quad \text{ή} \quad \text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

6.2.2.Ποιοτική αξιολόγηση απεικονίσεων

Αν και δεν αποτελεί το κύριο αντικείμενο της εργασίας, εκτός από την ποσοτική αξιολόγηση των αποτελεσμάτων, πραγματοποιείται και ποιοτική αξιολόγηση των ταξινομημένων απεικονίσεων με οπτική σύγκριση αυτών με διαθέσιμα ground truth δεδομένα.

Στην προκειμένη περίπτωση, ως ground truth δεδομένα χρησιμοποιούνται το Corine Land Cover, που είναι διαθέσιμο στο ευρύ κοινό μέσω της υπηρεσίας Copernicus και τον Κυρωμένο Δασικό Χάρτη, όπως προκύπτει από τα διαγράμματα που είναι διαθέσιμα από το Ελληνικό Κτηματολόγιο. Βέβαια, αναμένεται να σημειώνονται διαφοροποιήσεις μεταξύ των δυο αυτών χαρτών, για δυο βασικούς λόγους. Ο πρώτος είναι το χρονικό πλαίσιο που έχουν σχεδιαστεί, με πιο πρόσφατο το δασικό χάρτη (2022), έναντι του CLC (2018), με το δεύτερο λόγο να έγκειται στον τρόπο αντιμετώπισης και κατηγοριοποίησης των περιοχών, που διαφέρει κατά κράτος στους δύο χάρτες.

Από το Ελληνικό Κτηματολόγιο ακολουθείται η λογική της σύγκρισης πρόσφατων αεροφωτογραφιών, με παλαιότερες λήψεις του 1945, με ταυτόχρονη αξιολόγηση του ιδιοκτησιακού καθεστώτος ,ώστε να αποφανθεί αν μια έκταση χαρακτηρίζεται δασική στο παρόν ή σε παρελθοντικό χρόνο. Αυτός είναι και ο λόγος που σημειώνεται

σημαντικός αριθμός κωδικών ,όπως ΑΑ ή ΑΔ ή ΔΔ, στο υπόμνημά του, που επεξηγούνται παρακάτω.

ΔΔ	ΔΑΣΗ ΚΑΙ ΔΑΣΙΚΕΣ ΕΚΤΑΣΕΙΣ ΣΤΙΣ Α/Φ ΠΑΛΑΙΟΤΕΡΗΣ ΛΗΨΗΣ Ή ΠΡΟΫΦΙΣΤΑΜΕΝΑ ΣΤΟΙΧΕΙΑ ΔΑΣΗ ΚΑΙ ΔΑΣΙΚΕΣ ΕΚΤΑΣΕΙΣ ΣΤΙΣ Α/Φ ΠΡΟΣΦΑΤΗΣ ΛΗΨΗΣ & ΣΤΙΣ ΑΥΤΟΨΙΕΣ*
ΔΑ	ΔΑΣΗ ΚΑΙ ΔΑΣΙΚΕΣ ΕΚΤΑΣΕΙΣ ΣΤΙΣ Α/Φ ΠΑΛΑΙΟΤΕΡΗΣ ΛΗΨΗΣ Ή ΠΡΟΫΦΙΣΤΑΜΕΝΑ ΣΤΟΙΧΕΙΑ ΑΛΛΗΣ ΜΟΡΦΗΣ / ΚΑΛΥΨΗΣ ΕΚΤΑΣΕΙΣ ΣΤΙΣ Α/Φ ΠΡΟΣΦΑΤΗΣ ΛΗΨΗΣ & ΣΤΙΣ ΑΥΤΟΨΙΕΣ*
ΑΔ	ΑΛΛΗΣ ΜΟΡΦΗΣ / ΚΑΛΥΨΗΣ ΕΚΤΑΣΕΙΣ ΣΤΙΣ Α/Φ ΠΑΛΑΙΟΤΕΡΗΣ ΛΗΨΗΣ ΔΑΣΗ ΚΑΙ ΔΑΣΙΚΕΣ ΕΚΤΑΣΕΙΣ ΣΤΙΣ Α/Φ ΠΡΟΣΦΑΤΗΣ ΛΗΨΗΣ & ΣΤΙΣ ΑΥΤΟΨΙΕΣ*
ΑΑ	ΑΛΛΗΣ ΜΟΡΦΗΣ / ΚΑΛΥΨΗΣ ΕΚΤΑΣΕΙΣ ΣΤΙΣ Α/Φ ΠΑΛΑΙΟΤΕΡΗΣ ΛΗΨΗΣ ΑΛΛΗΣ ΜΟΡΦΗΣ / ΚΑΛΥΨΗΣ ΕΚΤΑΣΕΙΣ ΣΤΙΣ Α/Φ ΠΡΟΣΦΑΤΗΣ ΛΗΨΗΣ & ΣΤΙΣ ΑΥΤΟΨΙΕΣ*
ΠΔ	ΤΕΛΕΣΙΔΙΚΕΣ ΠΡΑΞΕΙΣ & ΑΠΟΦΑΣΕΙΣ ΧΑΡΑΚΤΗΡΙΣΜΟΥ - ΔΑΣΙΚΕΣ
ΠΑ	ΤΕΛΕΣΙΔΙΚΕΣ ΠΡΑΞΕΙΣ & ΑΠΟΦΑΣΕΙΣ ΧΑΡΑΚΤΗΡΙΣΜΟΥ - ΜΗ ΔΑΣΙΚΕΣ
ΠΧ	ΤΕΛΕΣΙΔΙΚΕΣ ΠΡΑΞΕΙΣ & ΑΠΟΦΑΣΕΙΣ ΧΑΡΑΚΤΗΡΙΣΜΟΥ - ΧΟΡΤΟΛΙΒΑΔΙΚΕΣ
ΑΝ	ΑΝΑΔΑΣΩΤΕΕΣ Ή ΔΑΣΩΤΕΕΣ ΕΚΤΑΣΕΙΣ
ΧΧ	ΧΟΡΤΟΛΙΒΑΔΙΚΕΣ ΕΚΤΑΣΕΙΣ ΣΤΙΣ Α/Φ ΠΑΛΑΙΟΤΕΡΗΣ ΛΗΨΗΣ ΧΟΡΤΟΛΙΒΑΔΙΚΕΣ ΕΚΤΑΣΕΙΣ ΣΤΙΣ Α/Φ ΠΡΟΣΦΑΤΗΣ ΛΗΨΗΣ & ΣΤΙΣ ΑΥΤΟΨΙΕΣ*
ΧΑ	ΧΟΡΤΟΛΙΒΑΔΙΚΕΣ ΕΚΤΑΣΕΙΣ ΣΤΙΣ Α/Φ ΠΑΛΑΙΟΤΕΡΗΣ ΛΗΨΗΣ ΑΛΛΗΣ ΜΟΡΦΗΣ / ΚΑΛΥΨΗΣ ΕΚΤΑΣΕΙΣ ΣΤΙΣ Α/Φ ΠΡΟΣΦΑΤΗΣ ΛΗΨΗΣ & ΣΤΙΣ ΑΥΤΟΨΙΕΣ*
ΑΧ	ΑΛΛΗΣ ΜΟΡΦΗΣ / ΚΑΛΥΨΗΣ ΕΚΤΑΣΕΙΣ ΣΤΙΣ Α/Φ ΠΑΛΑΙΟΤΕΡΗΣ ΛΗΨΗΣ ΧΟΡΤΟΛΙΒΑΔΙΚΕΣ ΕΚΤΑΣΕΙΣ ΣΤΙΣ Α/Φ ΠΡΟΣΦΑΤΗΣ ΛΗΨΗΣ & ΣΤΙΣ ΑΥΤΟΨΙΕΣ*

Εικόνα 5. Υπόμνημα Κυρωμένου Δασικού Χάρτη
https://gis.ktimanet.gr/gis/forestfinal/Samples/LegendKyrosi.html

Έχοντας συλλέξει τα διανυσματικά αρχεία που προσφέρει το Κτηματολόγιο, ο κυρωμένος δασικός χάρτης της περιοχής μελέτης έχει τη μορφή :

Σχήμα 74. Κυρωμένος Δασικός Χάρτης περιοχής ενδιαφέροντος

Στον Κυρωμένο Δασικό Χάρτη που παρουσιάζει το Ελληνικό Κτηματολόγιο αποτυπώνονται με διαφορετικό συμβολισμό μεταξύ τους οι δυο περιπτώσεις περιοχών, δασικές και μη δασικές. Ωστόσο, για την αποφυγή σύγχυσης, κατά την

ποιοτική αξιολόγηση των θεματικών χαρτών που παράγονται κατά την εκπόνηση των πειραμάτων, διατηρούνται στο χάρτη μόνο οι δασικές εκτάσεις. Παρόλα αυτά, για να υπάρχει ακριβής απόδοση του κυρωμένου δασικού χάρτη του Κτηματολογίου, παρακάτω παρατίθεται απόσπασμα αυτού που περιέχει το νότιο τμήμα της περιοχής μελέτης.

Σχήμα 75. Απόσπασμα Κυρωμένου Δασικού Χάρτη του Ελληνικού Κτηματολογίου
https://gis.ktimanet.gr/gis/forestfinal

ΚΕΦΑΛΑΙΟ 7. Πειραματικές εφαρμογές & Αξιολόγηση

Αρχικός στόχος της παρούσας διπλωματικής εργασίας είναι η υλοποίηση επιβλεπόμενης ταξινόμησης με τη χρήση δυο αλγορίθμων εκπαίδευσης, τον Random Forest (RF) και τον Support Vector Machine (SVM). Όμως, οι δοκιμές εκτέλεσης του δεύτερου δεν αποδίδουν σημειώνοντας κιόλας τεράστιες χρονικές καθυστερήσεις. Ο κυριότερος λόγος που ο αλγόριθμος SVM αργεί πολύ, οφείλεται στο ότι η πολυπλοκότητα της εκπαίδευσης ενός SVM είναι γενικά της τάξης από O(n2) έως O(n3) όπου n είναι ο αριθμός των δειγμάτων. Επομένως, εάν το σύνολο των δεδομένων εκπαίδευσης είναι μεγάλο, εν προκειμένω σύνολο εκπαίδευσης με 400-500 πολύγωνα, ο χρόνος που απαιτείται για την εκπαίδευση ενός SVM αυξάνεται εκθετικά.

Γενικά, η διαφορά στην απόδοση και τον χρόνο επεξεργασίας μεταξύ των αλγορίθμων Random Forest (RF) και Support Vector Machine (SVM) κατά την ταξινόμηση πολυφασματικών εικόνων σε R (ή οποιοδήποτε άλλο περιβάλλον) μπορεί να είναι αρκετά σημαντική. Αυτή η διαφορά οφείλεται κυρίως στα εγγενή χαρακτηριστικά και την υπολογιστική πολυπλοκότητα αυτών των αλγορίθμων.

- Random Forest: Το RF είναι εγγενώς «παραλληλιζόμενο», πράγμα που σημαίνει ότι μπορεί εύκολα να χωρίσει την εργασία σε πολλούς πυρήνες. Κάθε δέντρο στο δάσος είναι χτισμένο ανεξάρτητα από τα άλλα, καθιστώντας το εξαιρετικά αποδοτικό και επεκτάσιμο, ιδιαίτερα όταν έχουμε να κάνουμε με μεγάλο αριθμό χαρακτηριστικών εισόδου και δεδομένων εκπαίδευσης.
- Support Vector Machine: Το SVM περιλαμβάνει την επίλυση ενός πολύπλοκου προβλήματος βελτιστοποίησης που απαιτεί τον υπολογισμό των αποστάσεων μεταξύ όλων των ζευγών σημείων στον χώρο χαρακτηριστικών (των καναλιών), ο οποίος γίνεται υπολογιστικά "ακριβός", καθώς αυξάνεται ο αριθμός των δειγμάτων και των χαρακτηριστικών. Ο υπολογισμός περιλαμβάνει αντιστροφή πινάκων, η οποία είναι υπολογιστικά εντατική. Επίσης ο SVM είναι ευαίσθητος στον αριθμό των χαρακτηριστικών στα δεδομένα (χώρος υψηλών διαστάσεων). Εάν τα δεδομένα της εικόνας έχουν πολλά κανάλια (χαρακτηριστικά), το SVM θα αργεί περισσότερο σε σύγκριση με το RF. Ακόμη, ο αλγόριθμος SVM συνήθως απαιτεί τον υπολογισμό ενός πίνακα πυρήνα, ο οποίος περιλαμβάνει υπολογισμούς για κάθε ζεύγος σημείων των δεδομένων. Αυτό δεν είναι μόνο υπολογιστικά δαπανηρό, αλλά απαιτεί και εντατική χρήση της μνήμης RAM, καθώς μπορεί να χρειαστεί να αποθηκευτούν μεγάλοι πίνακες στη μνήμη.

Λόγω των παραπάνω, και την μεγάλης χρονικής καθυστέρησης στην εκτέλεση του SVM με τα δεδομένα εκπαίδευσης που έχουν σχεδιαστεί για τα πειράματα, αποφαίνεται η μη χρήση του αλγορίθμου SVM.

7.1.Εφαρμογή Ταξινόμησης με Ισορροπημένο Σύνολο Δεδομένων

Η διαδικασία υλοποίησης πειραματικών εφαρμογών ξεκινά με ένα ισορροπημένο σύνολο δεδομένων εκπαίδευσης, με την ισορροπία να έγκειται στην έκταση που καλύπτουν τα πολύγωνα εκπαίδευσης κάθε κατηγορίας. Το αποτέλεσμα ταξινόμησης με ένα ισορροπημένο σύνολο δεδομένων θα έχει καθοριστικό ρόλο τόσο για την εφαρμογή του Δείκτη Αναλογίας Ισορροπίας με την αξιοποίηση του πλήθους των πολυγώνων εκπαίδευσης ,όσο και στη σύγκριση των αποτελεσμάτων, ιδίως των μετρικών που είναι σχετικές με την ακρίβεια, τα οποία προκύπτουν από μη ισορροπημένα, ως προς την έκταση, σύνολα δεδομένων εκπαίδευσης.

Το ισορροπημένο σύνολο δεδομένων εκπαίδευσης που χρησιμοποιείται για την υλοποίηση του πρώτου πειράματος αποτελεί από :

Κατηγορία CLC	Πλήθος πολυγώνων
Τεχνητές επιφάνειες	63
Αρόσιμη γη	57
Μόνιμες καλλιέργειες	50
Λιβάδια	47
Ετερογενείς γεωργικές εκτάσεις	57
Δασικές εκτάσεις	38
Συνδυασμοί βλάστησης	56
Χερσαία ύδατα	45
Θάλασσα	34
Συνολικός αριθμός πολυγώνων	**447**

Πίνακας 6. Περιγραφή ισορροπημένου συνόλου δεδομένων εκπαίδευσης

Πέραν όμως του συνόλου εκπαίδευσης, σημαντικό ρόλο παίζει η επιλογή παραμέτρων για την εκτέλεση αλγορίθμων, όπως ο Random Forest, για τον οποίο προσδιορίζεται το πλήθος των δέντρων (Number of trees), οι διαχωρισμοί (splits) που θα σημειώνονται μεταξύ των δέντρων, με τελική επιλογή αυτών να είναι 100 δέντρα και 2 διαχωρισμοί μεταξύ των δέντρων.

Αξίζει να γίνει πιο ολοκληρωμένη και σαφής αιτιολόγηση στην επιλογή της παραμέτρου του πλήθους των δέντρων. Η παράμετρος "Number of Trees" σε έναν αλγόριθμο ταξινόμησης Random Forest αναφέρεται στον συνολικό αριθμό των δέντρων απόφασης που δημιουργούνται και συνδυάζονται στο μοντέλο. Αυτή η παράμετρος είναι σημαντική γιατί επηρεάζει θεωρητικά τόσο την απόδοση όσο και την ακρίβεια του μοντέλου.

Ο αλγόριθμος Random Forest λειτουργεί με τη λήψη μέσου όρου πολλαπλών δέντρων αποφάσεων, τα οποία μεμονωμένα μπορεί να έχουν υψηλή διακύμανση και θα μπορούσαν να προσαρμόζονται υπερβολικά στα δεδομένα (overfitting). Συνδυάζοντας πολλά δέντρα, η συνολική διακύμανση του μοντέλου μειώνεται. Αυτή είναι μια τεχνική, που ιστορικά ήρθε από τις τεχνικές εκμάθησης συνόλου (ensemble learning) όπου πολλά μοντέλα (δέντρα) ψηφίζουν για την τελική έξοδο, μειώνοντας με αυτόν τον τρόπο το σφάλμα που εισάγεται από οποιοδήποτε δέντρο.

Γενικά, όσο περισσότερα δέντρα στο δάσος, τόσο μεγαλύτερη είναι η ακρίβεια των προβλέψεων, μέχρι ένα σημείο. Κάθε δέντρο εκπαιδεύεται σε ένα τυχαίο υποσύνολο δεδομένων (bootstrapping) και λαμβάνει αποφάσεις με βάση ένα τυχαίο υποσύνολο χαρακτηριστικών. Περισσότερα δέντρα σημαίνουν καλύτερη προσέγγιση των ορίων απόφασης λόγω του μέσου όρου πολλαπλών διαφορετικών διαδρομών αποφάσεων.

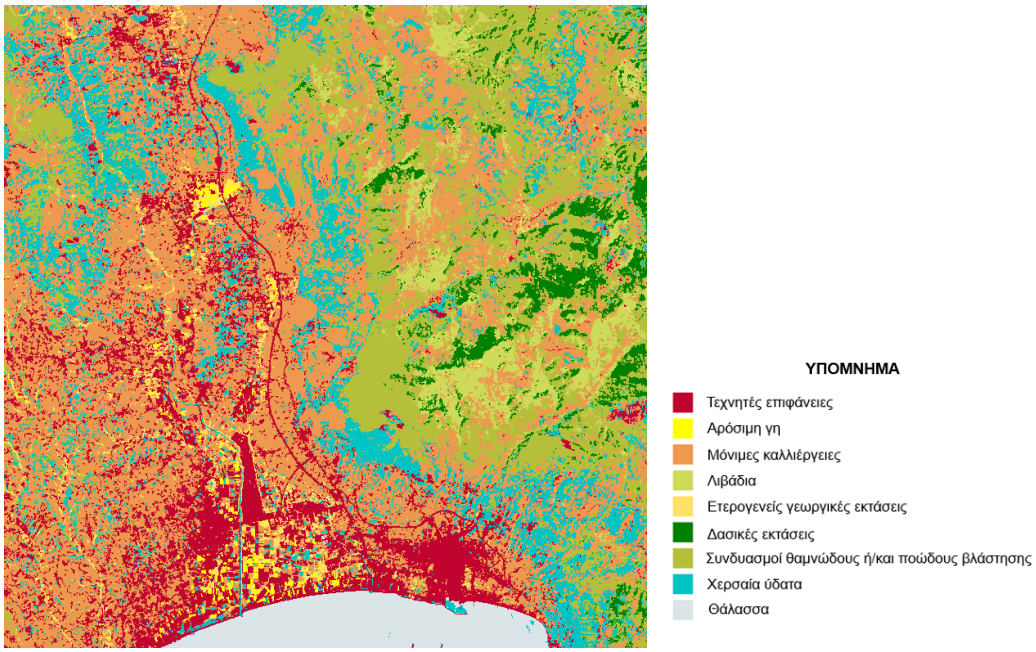
Ενώ η αύξηση του αριθμού των δέντρων μπορεί να βελτιώσει την ακρίβεια του μοντέλου, πέρα από έναν ορισμένο αριθμό, συνήθως υπάρχει αμελητέα βελτίωση στην απόδοση του μοντέλου. Αυτό συμβαίνει γιατί οι προβλέψεις αρχίζουν να σταθεροποιούνται καθώς ο πλειοψηφικός μηχανισμός ψηφοφορίας των δέντρων καταλήγει σε κάποιας μορφής συναίνεση. Θεωρητικά, πολλές δοκιμές ξεκινούν με "Number of Trees" όσες και οι κατηγορίες ταξινόμησης και σταδιακά αυξάνουν τον αριθμό, έως ότου επιτύχουν τα βέλτιστα αποτελέσματα.

Γενικά θα πρέπει να βρεθεί μια ισορροπία μεταξύ της υπολογιστικής απόδοσης και της ακρίβειας του μοντέλου. Έρευνες και πρακτικές εφαρμογές έχουν δείξει ότι η ύπαρξη περίπου 100 δέντρων στις περισσότερες περιπτώσεις βελτιώνει την απόδοση του μοντέλου και σταθεροποιεί τη διακύμανση χωρίς υπερβολικό κόστος υπολογισμού (ιδιαίτερα εάν χρησιμοποιείται ένα laptop με διαμόρφωση καθημερινής χρήσεως). Εξάλλου η κατασκευή και η αξιολόγηση μεγάλου αριθμού δέντρων μπορεί να είναι υπολογιστικά δαπανηρή, τόσο από πλευράς μνήμης όσο και χρόνου επεξεργασίας.

Ο βέλτιστος αριθμός δέντρων γενικά θα εξαρτηθεί από τα ειδικά χαρακτηριστικά των δεδομένων, αυτός είναι και ο λόγος πολλών σχετικών πειραμάτων σε αυτήν την εργασία.

7.1.1.Ποιοτική αξιολόγηση

Για τη μελέτη της απόδοσης της μεθόδου ταξινόμησης ακολουθεί μια διαδικασία οπτικής σύγκρισης των θεματικών χαρτών της ταξινόμησης με το χάρτη καλύψεως γης CLC, αλλά και τον κυρωμένο δασικό χάρτη, όπως προκύπτει από το Ελληνικό Κτηματολόγιο.

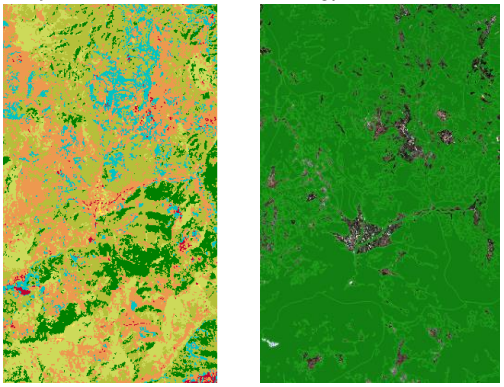


Σχήμα 76. Αποτέλεσμα ταξινόμησης Random Forest, με 100 δέντρα, πολυφασματικής απεικόνισης Sentinel-2 ημερομηνίας λήψης 17/07/2023

Το πρώτο οφθαλμοφανές σφάλμα της ταξινόμησης αφορά πολλές περιπτώσεις εικονοστοιχείων που ανήκουν σε κατηγορία βλάστησης να ταξινομούνται τελικά ως χερσαία ύδατα. Ωστόσο, ένα τέτοιο φαινόμενο μπορεί μόνο να εξηγηθεί με την παρουσία νερού στις εκτάσεις βλάστησης, ειδικά όταν γίνεται λόγος σε καλλιεργήσιμες περιοχές, όπως σε αυτές που συναντώνται ελαιώνες κι όχι δασικές. Σε κάθε άλλη περίπτωση, πρόκειται για λάθος κατά την ταξινόμηση. Πέραν όμως αυτού, είναι γνωστό από τα αντίστοιχα σχήματα πως η περιοχή μελέτης χαρακτηρίζεται από δάση, ιδίως στο ανατολικό της τμήμα. Παρόλα αυτά στο Σχήμα 76 φαίνεται μόνο ένα μικρό τμήμα να έχει ταξινομηθεί όντως στην κατηγορία των δασικών εκτάσεων. Το αποτέλεσμα αυτό πιθανώς οφείλεται στα συγγενή χαρακτηριστικά μεταξύ των δασικών

εκτάσεων και της σκληροφυλλικής βλάστησης, η οποία είναι η κυρίαρχη υποκατηγορία στην κατηγορία των συνδυασμών βλάστησης. Συγκρίνοντας δε το αποτέλεσμα της ταξινόμησης με τον κυρωμένο δασικό χάρτη προκύπτει πως όσα εικονοστοιχεία έχουν αντιστοιχηθεί στις δασικές εκτάσεις, αναγνωρίζονται σε αυτές κι από το Ελληνικό Κτηματολόγιο.

Βέβαια, όπως είναι αναμενόμενο το παραγόμενο από την ταξινόμηση αποτέλεσμα δε δύναται να ταυτίζεται σε απόλυτο βαθμό με το θεματικό χάρτη CLC, καθώς το πρώτο εστιάζει σε επίπεδο εικονοστοιχείου, ενώ το δεύτερο παρουσιάζει επιφανειακά μια γενικευμένη κατηγορία χρήσης γης που παρατηρείται σε μια έκταση. Για το λόγο αυτό αναμένεται να υπάρχουν αποκλίσεις και σφάλματα, αφού σημειώνονται παραδείγματα όπου τμήματα γυμνού εδάφους/εκτάσεις βλάστησης σε αγρανάπαυση κατατάσσονται στην κατηγορία των τεχνητών επιφανειών (συμπεριλαμβάνουν την υποκατηγορία των ορυχείων που προσομοιάζει με το γυμνό έδαφος).



Σχήμα 77. Σύγκριση Κυρωμένου Δασικού Χάρτη με το αποτέλεσμα της ταξινόμησης

7.1.2.Ποσοτική αξιολόγηση

Έχει επισημανθεί πολλάκις πως βασικός στόχος της εργασίας είναι η ποσοτική αξιολόγηση του αποτελέσματος της ταξινόμησης και πως η ακρίβεια του παραγόμενου προϊόντος επηρεάζεται ανάλογα με το εάν έχει εφαρμοσθεί υπερδειγματοληψία ή υποδειγματοληψία, όπως θα ακολουθήσει στα επόμενα πειράματα. Βασικός οδηγός για τη διαδικασία αυτή είναι ο υπολογισμός της ακρίβειας στο αποτέλεσμα της ταξινόμησης που έχει χρησιμοποιηθεί το αρχικά ισορροπημένο σύνολο δεδομένων εκπαίδευσης. Οι μετρικές που χρησιμοποιούνται για την αξιολόγηση του αποτελέσματος είναι το precision, αντίστοιχο μέγεθος της ακρίβειας παραγωγού, το recall που αντιστοιχεί στη μετρική της ακρίβειας του χρήστη όπως ονομάζεται στην Τηλεπισκόπηση, το F1-score, το οποίο στην περίπτωση του ισορροπημένου δείγματος δεν είναι ιδιαίτερα χρήσιμο, αποτελεί κυρίως μετρική αξιολόγησης για τα μη ισορροπημένα σύνολα ,ενώ ταυτόχρονα προσδιορίζεται και ο δείκτης συμφωνίας k-hat. Όλα τα παραπάνω μεγέθη αξιολόγησης προκύπτουν μέσα από τον πίνακα σύγχυσης, που για το πείραμα που αναλύεται στην ενότητα του ισορροπημένου πειράματος, έχει τη μορφή

V_Classified	Τεχνητές	Αρόσιμη	Μόνιμες	Λιβάδια	Ετερογενείς	Δάση	Συνδυασμοί	Χερσαία	Θάλασσα	Total
Τεχνητές	**719**	0	0	0	0	0	0	0	0	**719**
Αρόσιμη	0	**436**	0	0	123	0	0	0	0	**559**
Μόνιμες	0	12	**661**	61	2	0	11	0	0	**747**
Λιβάδια	0	3	16	**267**	21	78	240	0	0	**625**
Ετερογενείς	0	244	0	0	**552**	0	0	0	0	**796**
Δάση	0	0	0	38	0	**550**	51	0	0	**639**
Συνδυασμοί	0	2	3	324	0	76	**384**	0	0	**789**
Χερσαία	0	6	11	0	0	0	4	**699**	0	**720**
Θάλασσα	0	0	0	0	0	0	0	0	**692**	**692**
Total	**719**	**703**	**691**	**690**	**698**	**704**	**690**	**699**	**692**	**6286**

Πίνακας 7. Πίνακας σύγχυσης ταξινόμησης πολυφασματικής απεικόνισης με ημερομηνία λήψης 17/07/2023 με χρήση του αλγορίθμου Random Forest, με ισορροπημένο σύνολο εκπαίδευσης

Από τη βιβλιογραφία είναι γνωστό πως στον πίνακα σύγχυσης τα στοιχεία της κύριας διαγωνίου είναι αυτά που έχουν ταξινομηθεί ορθά, ενώ τα εκτός διαγωνίου στοιχεία στις σειρές του πίνακα αντιπροσωπεύουν εικονοστοιχεία των δεδομένων ελέγχου, μιας συγκεκριμένης κατηγορίας τα οποία εξαιρέθηκαν από την κατηγορία αυτή κατά την ταξινόμηση κι αντίστοιχα, τα εκτός διαγωνίου στοιχεία των στηλών του πίνακα αντιστοιχούν σε εικονοστοιχεία των δεδομένων ελέγχου, άλλων κατηγοριών που συμπεριλήφθηκαν σε μια συγκεκριμένη κατηγορία κατά την ταξινόμηση. Στην πρώτη περίπτωση γίνεται λόγος για σφάλματα παράλειψης ή αποκλεισμού και στη δεύτερη για σφάλματα συμπερίληψης ή προμήθειας.

Συνδέοντας την ποσοτική με την ποιοτική αξιολόγηση, ο πίνακας σύγχυσης καταγράφοντας τα 11 εσφαλμένα ταξινομημένα εικονοστοιχεία της κατηγορίας των μόνιμων καλλιεργειών ως χερσαία ύδατα καθώς και τα 16 εικονοστοιχεία των ετερογενών καλλιεργειών, επιβεβαιώνει το σφάλμα της ταξινόμησης όπως περιεγράφηκε κατά την ποιοτική αξιολόγηση. Παρόλα αυτά δεν είναι το μόνο σφάλμα που σημειώνεται, με σημαντικά πλήθη λανθασμένα ταξινομημένων εικονοστοιχείων να καταγράφονται σε κατηγορίες σχετικές με τη βλάστηση. Αναλυτικότερα, ένα από τα σημαντικότερα σφάλμα, από ποσοτική άποψη, είναι η ταξινόμηση εικονοστοιχείων των ετερογενών καλλιεργειών στην κατηγορία της αρόσιμης γης. Ωστόσο, το γεγονός πως πρόκειται για δυο “συγγενείς” κατηγορίες, με την κατηγορία των ετερογενών να περιλαμβάνει τις υποκατηγορίες 2.4.2. Σύνθετα συστήματα καλλιέργειας και 2.4.3. Γη που καλύπτεται κυρίως από γεωργία με σημαντικές εκτάσεις φυσικής βλάστησης, οι οποίες σύμφωνα με τη βιβλιογραφία ως επί το πλείστον περιγράφουν μικρά αγροτεμάχια με διάφορες ετήσιες καλλιέργειες, μόνιμες καλλιέργειες και γενικότερα γεωργικές περιοχές, έχει συνέπεια να δημιουργείται αυτή η σύγχυση κατά την ταξινόμηση. Ένα ακόμη σφάλμα της διαδικασίας ταξινόμησης που υποδεικνύει ο πίνακας σύγχυσης είναι τα εικονοστοιχεία των λιβαδιών που έχουν ταξινομηθεί ως αρόσιμη γη, μόνιμες κι ετερογενείς καλλιέργειες, δάση και συνδυασμοί βλάστησης. Το συγκεκριμένο σφάλμα πιθανότατα οφείλεται σε περιπτώσεις όπου αντίστοιχα σε αυτές τις κατηγορίες η βλάστηση είναι χαμηλή και κατά τόπους αραιή, γεγονός που προκύπτει κι από την παράθεση της πολυφασματικής απεικόνισης με το CLC της περιοχής μελέτης και την ταξινομημένη απεικόνιση. Από την άλλη, η ταξινόμηση των δασικών εκτάσεων, όπου κατά την ποιοτική αξιολόγηση δε φαίνεται να επιτυγχάνει σημαντική ακρίβεια, σύμφωνα με τον πίνακα σύγχυσης καταγράφει σημαντικό μέρος ορθά ταξινομημένων εικονοστοιχείων, καταγράφοντας όμως κι εικονοστοιχεία των δασών ως λιβάδια και συνδυασμούς βλάστησης, ενώ συμβαίνει και το αντίστροφο, εικονοστοιχεία αυτών των δυο να ταξινομούνται ως δασικές εκτάσεις (σφάλματα παράλειψης και συμπερίληψης, αντίστοιχα). Ήδη βέβαια, τόσο από την ερμηνεία της πολυφασματικής απεικόνισης, όσο κι από την ερμηνεία των φασματικών υπογραφών ο παρατηρητής προϊδεάζεται για το ενδεχόμενο τέτοιου είδους σφαλμάτων, καθώς είναι εμφανή τα κοινά χαρακτηριστικά των κατηγοριών, ιδίως στην περίπτωση της υποκατηγορίας της σκληροφυλλικής βλάστησης, που ανήκει στους συνδυασμούς βλάστησης (Πίνακας 4). Σχετικά με τη μερίδα εικονοστοιχείων των δασικών εκτάσεων που φαίνεται να έχουν να ταξινομηθεί ως λιβάδια, η εξήγηση αυτού του αποτελέσματος βασίζεται σε ενδεχόμενο τμήματος πιο πυκνής βλάστησης στην κατηγορία των λιβαδιών, όπως υποδεικνύεται κι από την παράθεση των Σχημάτων 73 και 76. Σημαντικά σφάλματα σημειώνονται και κατά την ταξινόμηση της κατηγορίας των συνδυασμών βλάστησης, όπου το μεγαλύτερο πλήθος εικονοστοιχείων έχει ταξινομηθεί στην κατηγορία των λιβαδιών και μια μικρότερη μερίδα στα δάση. Ολοκληρώνοντας σε αυτό το σημείο την ανάλυση των αποτελεσμάτων του πίνακα σύγχυσης που αφορούν τις κατηγορίες βλάστησης προκύπτει πως οι κατηγορίες που

εμπλέκονται σε μείζονα κλίμακα μεταξύ τους είναι αυτές των μόνιμων καλλιεργειών, των ετερογενών καλλιεργειών, των δασικών εκτάσεων και των συνδυασμών βλάστησης και σε μικρότερο βαθμό η αρόσιμη γη. Βασικός παράγοντας της σύγχυσης του αλγορίθμου κατά την ταξινόμηση αυτών είναι τα χαρακτηριστικά της βλάστησης, που σημειώνει χαρακτηριστική φασματική υπογραφή, γνωστή κι αναλυμένη στο σύνολο της βιβλιογραφίας της Τηλεπισκόπησης, από την οποία δεν αποκλίνουν σε μεγάλο βαθμό οι φασματικές υπογραφές των επιμέρους κατηγοριών που μελετώνται στην περιοχή ενδιαφέροντος. Ο παράγοντας δε, που οδηγεί την κατηγορία της αρόσιμης γης να αποκλίνει από τις υπόλοιπες και σύμφωνα με τον πίνακα σύγχυσης να σημειώνει την κατηγορία βλάστησης με τα λιγότερα λανθασμένα ταξινομημένα εικονοστοιχεία, είναι η παρουσία του νερού, λόγω της άρδευσης αυτών των εκτάσεων. Από την άλλη, το γεγονός πως εκτάσεις βλάστησης ταξινομούνται ως χερσαία ύδατα προκαλεί μεγάλη σύγχυση. Όμως, η μια κατηγορία που ταξινομείται ως χερσαία ύδατα είναι αυτή των μόνιμων καλλιεργειών, με τους ελαιώνες να είναι αυτοί που καλύπτουν μεγάλη έκταση της περιοχής μελέτης, σημειώνοντας αραιή βλάστηση, τόσο λόγω του μικρού φυλλώματος, όσο και λόγω της απόστασης μεταξύ των δέντρων, με την πολυφασματική απεικόνιση να υποδεικνύει το έδαφος κάτω από τα ελαιόδεντρα να χαρακτηρίζεται από χαμηλή βλάστηση ή/και μηδαμινή (γυμνό έδαφος). Ταυτόχρονα, λαμβάνοντας υπόψιν την εποχή λήψης της πολυφασματικής απεικόνισης, 17/07/2023, μια εποχή με υψηλές θερμοκρασίες, χωρίς βροχόπτωση, εξαιτίας των οποίων οι καλλιεργητές οδηγούνται στο να αρδεύουν τις καλλιέργειές τους, συμπεραίνεται πως το σφάλμα του αλγορίθμου έγκειται σε αυτό το λόγο. Η δεύτερη κατηγορία κάλυψης γης που αποτυπώνεται στην παραγόμενη απεικόνιση ως χερσαία ύδατα, είναι αυτή της αρόσιμης γης. Όπως υποδεικνύει ο θεματικός χάρτης CLC της περιοχής μελέτης, Σχήμα 73, η μια κατηγορία που περιβάλει τα χερσαία ύδατα είναι αρόσιμη γη, επομένως η παρεμβολή βλάστησης αυτής της κατηγορίας επί των χερσαίων υδάτων, πιθανώς αποτελεί παράγοντα σύγχυσης κατά την εκπαίδευση του αλγορίθμου, λαμβάνοντας υπόψιν ταυτόχρονα ότι γίνεται λόγος για αρδεύσιμες εκτάσεις.

Χωρίς ιδιαίτερη σημαντικότητα στην αξιολόγηση του αποτελέσματος, προσδιορίζεται και η συνολική ακρίβεια του αποτελέσματος της ταξινόμησης, ως το άθροισμα των στοιχείων της κύριας διαγωνίου του πίνακα σύγχυσης (ορθά ταξινομημένα εικονοστοιχεία) προς το συνολικό πλήθος των εικονοστοιχείων κι ισούται με 78,91%.

Περνώντας πλέον στις μετρικές αξιολόγησης του αποτελέσματος που αφορούν επιμέρους τις κατηγορίες χρήσης γης, προσδιορίζονται το μέγεθος precision, αντίστοιχο της ακρίβειας παραγωγού, περιγράφοντας για κάθε κλάση εικονοστοιχείων των δεδομένων ελέγχου (σειρά), τον αριθμό των σωστά ταξινομημένων εικονοστοιχείων που διαιρείται με το συνολικό αριθμό των εικονοστοιχείων των δεδομένων ελέγχου αυτής της κλάσης και το μέγεθος recall, αντίστοιχο της ακρίβειας χρήστη, που πρόκειται για το κλάσμα των σωστά ταξινομημένων εικονοστοιχείων σε σχέση με όλα τα εικονοστοιχεία που ταξινομούνται σε αυτή την κατηγορία στην ταξινομημένη εικόνα, όπου για κάθε κλάση της ταξινομημένης εικόνας (στήλη), ο αριθμός των σωστά ταξινομημένων εικονοστοιχείων διαιρείται με το συνολικό αριθμό των εικονοστοιχείων που ταξινομήθηκαν σε αυτή την κατηγορία.

	Precision	Recall	F1-score	Συνολική ακρίβεια (%)	K-hat (%)
Τεχνητές	1	1	1	78,91	76,27
Αρόσιμη	0,7800	0,6202	0,6910		
Μόνιμες	0,8849	0,9566	0,9193		
Λιβάδια	0,4272	0,3870	0,4061		
Ετερογενείς	0,6935	0,7908	0,7390		
Δάση	0,8607	0,7813	0,8191		
Συνδυασμοί	0,4867	0,5565	0,5193		
Χερσαία	0,9708	1	0,9852		
Θάλασσα	1	1	1		

Πίνακας 8. Μετρικές αξιολόγησης της ταξινόμησης της πολυφασματικής απεικόνισης με τον αλγόριθμο Random Forest, με ισορροπημένο σύνολο δεδομένων

Τις χαμηλότερες ακρίβειες σημειώνουν οι κατηγορίες που σχετίζονται με τη βλάστηση, με τη “χειρότερη” περίπτωση να αποτελεί η κατηγορία των λιβαδιών που καταγράφουν ακρίβειες κάτω της τάξης του 45%. Οι κατηγορίες που σημειώνουν σημαντικά υψηλές ακρίβειες είναι οι τεχνητές επιφάνειες, τα χερσαία ύδατα και η θάλασσα.

Μια ακόμη μετρική που χρησιμοποιείται για την αξιολόγηση της ταξινόμησης είναι το F1-score, το οποίο δεν είναι ιδιαίτερα χρηστικό για ένα ισορροπημένο σύνολο δεδομένων, αλλά είναι υψίστης σημασίας για την αξιολόγηση σε πείραμα με μη ισορροπημένα σύνολα δεδομένων. Το F1-score ως αρμονικός μέσος των μεγεθών precision και recall, με τιμές πλησίον της μονάδας να περιγράφουν την καλύτερη απόδοση των δυο επιμέρους μετρικών, σε αντίθεση με τις χαμηλές τιμές του F1-score, που τείνουν στο 0 να υποδεικνύουν χαμηλή απόδοση του μοντέλου. Μέτρια θα μπορούσε να χαρακτηριστεί η απόδοση των λιβαδιών με τιμή πλησίον του 0.4 και στη συνέχεια οι αποδόσεις των συνδυασμών βλάστησης, με τιμές περί το 0.5. Οι κατηγορίες με τις καλύτερες αποδόσεις είναι τα χερσαία ύδατα, η θάλασσα και οι τεχνητές επιφάνειες με τιμές που τείνουν ή/κι ίσες με τη μονάδα, με εξίσου σημαντικό μέγεθος F1-score να σημειώνει η αρόσιμη γη με τιμή πλησίον του 0.92.

Τέλος, προσδιορίζεται ο δείκτης συμφωνίας k-hat ***(Εξίσωση 1)*** και για την ταξινόμηση με τη χρήση του ισορροπημένου συνόλου δεδομένων προκύπτει ίσος με 76,27%. Να σημειωθεί πως η τιμή k-hat = 0.76 είναι λίγο μικρότερη από τη συνολική ακρίβεια ≈ 0.79 που υπολογίστηκε νωρίτερα. Η συνολική ακρίβεια περιλαμβάνει μόνο τα δεδομένα κατά μήκος της κύριας διαγωνίου κι αποκλείει τα σφάλματα παράλειψης και συμπερίληψης. Από την άλλη πλευρά, η μετρική k-hat εμπεριέχει τα μη διαγώνια στοιχεία του πίνακα σύγχυσης, ως το γινόμενο των περιθωριακών γραμμών και στηλών. Ένα από τα κυριότερα πλεονεκτήματα του υπολογισμού του k-hat είναι η δυνατότητα χρήσης αυτής της τιμής ως βάση για τον προσδιορισμό της στατιστικής σημασίας ενός συγκεκριμένου πίνακα σύγχυσης ή των διαφορών μεταξύ πινάκων σύγχυσης, όπου εν προκειμένω θα διευκολύνει στη σύγκριση μεταξύ των πινάκων σύγχυσης που θα προκύψουν από το σύνολο των πειραματικών εφαρμογών.

7.2.Προσδιορισμός Δείκτη Αναλογίας Ισορροπίας (Balance Ratio Indice, BRI)
Ο σκοπός για τον οποίο αναπτύχθηκε ο Δείκτης Αναλογίας Ισορροπίας (Balance Ratio Indice, BRI) είναι η εκτίμηση του βαθμού και της κατεύθυνσης της ανισορροπίας των δεδομένων. Εν ολίγοις, με τη χρήση του δείκτη αποφαίνεται ποια τεχνική δειγματοληψίας θα χρησιμοποιηθεί για τη διαχείριση της ανισορροπίας των κατηγοριών, όπως αυτή προκύπτει από το πλήθος των πολυγώνων εκπαίδευσης.

Σε πρώτο στάδιο προσδιορίζεται ο αριθμός των πολυγώνων εκπαίδευσης σε κάθε κατηγορία και ο συνολικός τους αριθμός στο dataset :

Κατηγορία CLC	Πλήθος πολυγώνων
Τεχνητές επιφάνειες	63
Αρόσιμη γη	57
Μόνιμες καλλιέργειες	50
Λιβάδια	47
Ετερογενείς γεωργικές εκτάσεις	57
Δασικές εκτάσεις	38
Συνδυασμοί βλάστησης	56
Χερσαία ύδατα	45
Θάλασσα	34
Συνολικός αριθμός πολυγώνων	**447**

Πίνακας 9. Αριθμός πολυγώνων εκπαίδευσης ανά κατηγορία

Το δεύτερο βήμα αφορά τη συχνότητα εμφάνισης, δηλαδή το λόγο αναλογίας κάθε κατηγορίας στο σύνολο των δεδομένων εκπαίδευσης :

Κατηγορία CLC	Λόγος αναλογίας
Τεχνητές επιφάνειες	0,141
Αρόσιμη γη	0,128
Μόνιμες καλλιέργειες	0,112
Λιβάδια	0,105
Ετερογενείς γεωργικές εκτάσεις	0,128
Δασικές εκτάσεις	0,085
Συνδυασμοί βλάστησης	0,125
Χερσαία ύδατα	0,101
Θάλασσα	0,076

Πίνακας 10. Συχνότητες εμφάνισης ανά κατηγορία

Ο δείκτης αναλογίας για κάθε κατηγορία μπορεί να υπολογιστεί συγκρίνοντας την αναλογία κάθε κατηγορίας, με την ιδανική αναλογία, εάν τα δεδομένα ήταν τέλεια ισορροπημένα. Η ιδανική αναλογία ,εφόσον γίνεται λόγος για 9 κατηγορίας κάλυψης γης στην περιοχή μελέτης, είναι 1/9 = 0,111.

Κατηγορία CLC	Συχνότητας Εμφάνισης προς Ιδανική Αναλογία	Δείκτης BRI
Τεχνητές επιφάνειες	0,141/0,111	1,268
Αρόσιμη γη	0,128/0,111	1,148
Μόνιμες καλλιέργειες	0,112/0,111	1,007
Λιβάδια	0,105/0,111	0,946
Ετερογενείς γεωργικές εκτάσεις	0,128/0,111	1,148
Δασικές εκτάσεις	0,085/0,111	0,765
Συνδυασμοί βλάστησης	0,125/0,111	1,128
Χερσαία ύδατα	0,101/0,111	0,906
Θάλασσα	0,076/0,111	0,685

Πίνακας 11. Δείκτης Αναλογίας Ισορροπίας για κάθε κατηγορία

Δείκτης αναλογίας για κάθε κατηγορία		Σύγκριση		Πιθανή αντιμετώπιση
Τεχνητές επιφάνειες	1,268	>	1	υποδειγματοληψία
Αρόσιμη γη	1,148	>	1	υποδειγματοληψία
Μόνιμες καλλιέργειες	1,007	≈	1	συνδυασμός μεθόδων
Λιβάδια	0,946	<	1	υπερδειγματοληψία
Ετερογενείς γεωργικές εκτάσεις	1,148	>	1	υποδειγματοληψία
Δασικές εκτάσεις	0,765	<	1	υπερδειγματοληψία
Συνδυασμοί βλάστησης	1,128	>	1	υποδειγματοληψία
Χερσαία ύδατα	0,906	<	1	υπερδειγματοληψία
Θάλασσα	0,685	<	1	υπερδειγματοληψία

Πίνακας 12. Ερμηνεία Δείκτη Αναλογίας Ισορροπίας

Ο δείκτης αναλογίας BRI δείχνει με ποιον τρόπο συγκρίνεται κάθε κατηγορία με ένα απόλυτα ισορροπημένο σύνολο:

1. Μια αναλογία κοντά στο 1 δείχνει μια ισορροπημένη κατάσταση.
2. Μια αναλογία σημαντικά μικρότερη από 1 υποδηλώνει υποεκπροσώπηση της κατηγορίας (και πιθανή ανάγκη για υπερδειγματοληψία).
3. Μια αναλογία σημαντικά μεγαλύτερη από 1 υποδηλώνει υπερεκπροσώπηση (και πιθανή ανάγκη για υποδειγματοληψία).

Όπως επισημαίνεται στην αντίστοιχη Ενότητα 6.1.2., ο δείκτης υποδεικνύει την πιθανή αντιμετώπιση της ανισορροπίας των δεδομένων, χωρίς όμως αυτή να είναι η μόνη λύση. Η τελική απόφαση για τον τρόπο αντιμετώπισης της ανισορροπίας των δεδομένων έγκειται στις επιδιώξεις του παρατηρητή.

Η επιλογή της μεθόδου αντιμετώπισης του προβλήματος της ανισορροπίας των δεδομένων εκπαίδευσης στηρίζεται αρχικά στην υπόδειξη του Δείκτη Αναλογίας Ισορροπίας, όπως παρουσιάζεται στον Πίνακα 12, ενώ ταυτόχρονα αξιολογείται αν στην περιοχή μελέτης είναι δυνατή η εφαρμογή της μεθόδου που υποδεικνύεται, με χαρακτηριστικό παράδειγμα την κατηγορία των χερσαίων υδάτων, όπου ο δείκτης υποδεικνύει υπερδειγματοληψία αυτών, αλλά η έκτασή τους και τα υπάρχοντα δεδομένα εκπαίδευσης δεν προσφέρουν τη δυνατότητα αυτή, με αποτέλεσμα να ακολουθήσει μια διαδικασία υποδειγματοληψίας των υπολοίπων. Αντίστοιχη λογική ακολουθείται για όλες τις κατηγορίες κάλυψης γης, με την πιο ιδιαίτερη περίπτωση να είναι αυτή των μόνιμων καλλιεργειών όπου η τιμή της είναι σχεδόν ίση με τη μονάδα, οδηγώντας με αυτόν τον τρόπο στο συνδυασμό υπερδειγματοληψίας κι υποδειγματοληψίας των κατηγοριών, ώστε ο δείκτης για τις μόνιμες καλλιέργειες να λάβει την τιμή 1. Για τους λόγους αυτούς ο τρόπος αντιμετώπισης της ανισορροπίας για την κάθε κατηγορία αναφέρεται ξεχωριστά στις αντίστοιχες ενότητες των πειραμάτων.

7.3.Εφαρμογές Μεθόδων Τυχαίας Υποδειγματοληψίας & Υπερδειγματοληψίας

Για την προηγούμενη εφαρμογή επιβλεπόμενης ταξινόμησης δημιουργήθηκε ένα σύνολο δεδομένων εκπαίδευσης, το οποίο θεωρείται ισορροπημένο ως προς την έκταση που καλύπτουν τα πολύγωνα εκπαίδευσης του αλγορίθμου, με διαφορετικά ωστόσο πλήθη πολυγώνων εκπαίδευσης. Για την ανισορροπία μεταξύ των πολυγώνων χρησιμοποιείται ο Δείκτης Αναλογίας Ισορροπίας, ο οποίος συγκρίνεται με τη μονάδα κι αποφαίνεται αν θα εφαρμοστεί υπερδειγματοληψία ή υποδειγματοληψία για να λάβει την τιμή της μονάδας ο δείκτης για την εξεταζόμενη κατηγορία.

Έτσι λοιπόν, στις πειραματικές εφαρμογές επιβλεπόμενης ταξινόμησης που ακολουθούν για κάθε μια από τις 9 κατηγορίες κάλυψης γης της περιοχής ενδιαφέροντος, λαμβάνοντας υπόψιν τα αποτελέσματα που παρουσιάζονται στην Ενότητα 7.2.Προσδιορισμός Δείκτη Αναλογίας Ισορροπίας (Balance Ratio Indice, BRI), υλοποιείται μια σειρά από δοκιμές με αύξηση ή μείωση του πλήθους των πολυγώνων, με στόχο η τιμή του δείκτη για την κατηγορία να λάβει την τιμή 1. Ο στόχος της ενότητας αυτής είναι να βρεθεί το ιδανικότερο σενάριο, ώστε να αυξάνεται η ακρίβεια στα αποτελέσματα της ταξινόμησης.

7.3.1.Τεχνητές επιφάνειες

Η κατηγορία των τεχνητών επιφανειών, όπως ορίζει ο Πίνακας 12, έχει δείκτη αναλογίας ισορροπίας ίσο με 1.268 ,τιμή μεγαλύτερη της μονάδας, η οποία σύμφωνα με την ερμηνεία του δείκτη υποδεικνύει τη μέθοδο της υποδειγματοληψίας για την αντιμετώπιση της ανισορροπίας, οδηγώντας με αυτόν τον τρόπο σε μια σειρά από δοκιμές, με μείωση του πλήθους των πολυγώνων εκπαίδευσης της κατηγορίας. Οι τεχνητές επιφάνειες είναι μια κατηγορία που προσφέρει τη δυνατότητα υποδειγματοληψίας, καθώς ο αρχικός αριθμός των πολυγώνων της φτάνει τα 63.

Ολοκληρώνοντας τις δοκιμές, το σύνολο δεδομένων εκπαίδευσης παίρνει τη μορφή που περιγράφεται παρακάτω :

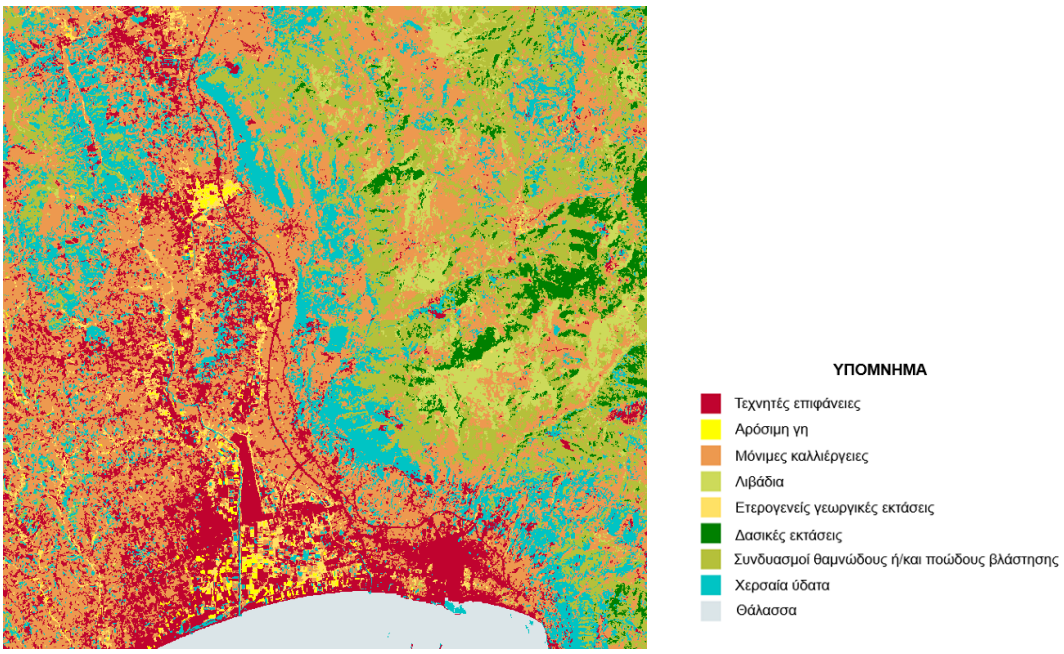
Κατηγορίες CLC	Πλήθος πολυγώνων	Συχνότητα	BRI
Τεχνητές επιφάνειες	48	0,111	1,000
Αρόσιμη γη	57	0,132	1,188
Μόνιμες καλλιέργειες	50	0,116	1,042
Λιβάδια	47	0,109	0,979
Ετερογενείς γεωργικές εκτάσεις	57	0,132	1,188
Δασικές εκτάσεις	38	0,088	0,792
Συνδυασμοί βλάστησης	56	0,130	1,167
Χερσαία ύδατα	45	0,104	0,938
Θάλασσα	34	0,079	0,708

Πίνακας 13. Μορφή συνόλου δεδομένων εκπαίδευσης για την ισορροπία των τεχνητών επιφανειών

Σύμφωνα με τα αποτελέσματα του παραπάνω πίνακα στο υπάρχον dataset εκπαίδευσης του αλγορίθμου μειώνονται με τυχαίο τρόπο τα πολύγωνα των τεχνητών επιφανειών, εφαρμόζεται δηλαδή τυχαία υποδειγματοληψία, χωρίς να υπάρχει παρέμβαση στις υπόλοιπες κατηγορίες.

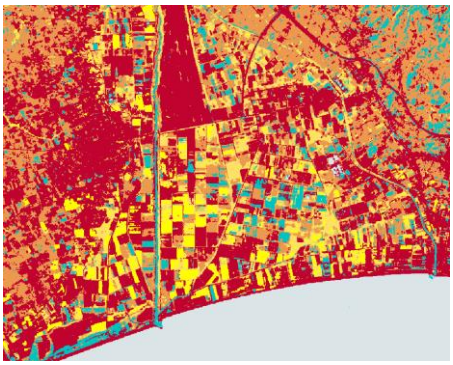
7.3.1.1.Ποιοτική Αξιολόγηση

Πρώτο στάδιο έπειτα από την εφαρμογή της ταξινόμησης είναι η ποιοτική αξιολόγηση του αποτελέσματος, μια διαδικασία που διευκολύνει και τη μετέπειτα ερμηνεία του πίνακα σύγχυσης κατά την ποσοτική αξιολόγηση.

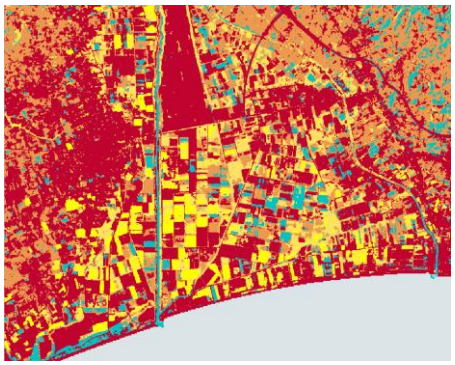


Σχήμα 78. Αποτέλεσμα ταξινόμησης Random Forest, με 100 δέντρα, πολυφασματικής απεικόνισης Sentinel-2 ημερομηνίας λήψης 17/07/2023, με τη μέθοδο υποδειγματοληψίας για τις τεχνητές επιφάνειες

Η βασική σύγκριση που γίνεται σε αυτό το στάδιο αφορά το παραγόμενο μέσω υποδειγματοληψίας αποτέλεσμα, με την αρχική απεικόνιση της ταξινόμησης, που έχει παραχθεί με ισορροπημένο ως προς την έκταση dataset. Παραθέτοντας το Σχήμα 78 έναντι του Σχήματος 76 παρατηρείται πως παρά το γεγονός ότι έχει εφαρμοσθεί υποδειγματοληψία στα πολύγωνα της κατηγορίας, οι τεχνητές επιφάνειες αναγνωρίζονται με καλύτερη ακρίβεια στο 2ο πείραμα, πυκνώνοντας τα εικονοστοιχεία που την περιγράφουν και διορθώνοντας σφάλματα που σχετίζονται με την κατηγορία της αρόσιμης γης στο νότιο τμήμα της περιοχής ενδιαφέροντος.



Σχήμα 79. Απόσπασμα από το 1ο πείραμα ταξινόμησης



Σχήμα 80. Απόσπασμα από το 2ο πείραμα -υποδειγματοληψία τεχνητών επιφανειών

Δεν εξαλείφεται απόλυτα η εσφαλμένη ταξινόμηση εκτάσεων σχετικών με τη βλάστηση ως τεχνητές επιφάνειες. Επισημαίνεται για ακόμη μια φορά πως είναι εκτάσεις αρόσιμης γης, οι οποίες σύμφωνα με την ερμηνεία της πολυφασματικής απεικόνισης στην εποχή λήψης δεν είναι καλλιεργημένες, με αποτέλεσμα το γυμνό έδαφος, που παρουσιάζει κοινά χαρακτηριστικά με τις τεχνητές επιφάνειες, να υπάγεται σε αυτές

κατά την ταξινόμηση. Πέραν της ενισχυμένης αποτύπωσης των τεχνητών επιφανειών δεν καταγράφονται διαφοροποιήσεις στις απεικονίσεις των πειραμάτων ταξινόμησης, με τη μόνη αισθητή να είναι αυτή που αφορά την αρόσιμη γη.

7.3.1.2.Ποσοτική Αξιολόγηση

Οι μετρικές που αφορούν την ποσοτική αξιολόγηση προκύπτουν από τον πίνακα σύγχυσης, ο οποίος για το πείραμα υποδειγματοληψίας των τεχνητών επιφανειών έχει τη μορφή :

V_Classified	Τεχνητές	Αρόσιμη	Μόνιμες	Λιβάδια	Ετερογενείς	Δάση	Συνδυασμοί	Χερσαία	Θάλασσα	Total
Τεχνητές	**719**	0	0	0	0	0	0	0	0	**719**
Αρόσιμη	0	**434**	0	0	129	0	0	0	0	**563**
Μόνιμες	0	13	**663**	62	3	0	10	0	0	**751**
Λιβάδια	0	0	16	**274**	22	74	250	0	0	**636**
Ετερογενείς	0	249	0	0	**544**	0	0	0	0	**793**
Δάση	0	0	0	37	0	**555**	45	0	0	**637**
Συνδυασμοί	0	1	4	317	0	75	**379**	0	0	**776**
Χερσαία	0	6	8	0	0	0	6	**699**	0	**719**
Θάλασσα	0	0	0	0	0	0	0	0	**692**	**692**
Total	**719**	**703**	**691**	**690**	**698**	**704**	**690**	**699**	**692**	**6286**

Πίνακας 14. Πίνακας σύγχυσης ταξινόμησης πολυφασματικής απεικόνισης με ημερομηνία λήψης 17/07/2023 με χρήση του αλγορίθμου Random Forest, με τη μέθοδο υποδειγματοληψίας για τις τεχνητές επιφάνειες

Την απόλυτη ακρίβεια φαίνεται να επιτυγχάνει η κατηγορία των τεχνητών επιφανειών, όπου και τα 719 εικονοστοιχεία των πολυγώνων ελέγχου καταγράφονται σε αυτή την κατηγορία, χωρίς απόκλιση. Ταυτόχρονα, βελτιωμένη φαίνεται και η ακρίβεια που έχει επιτευχθεί και για τις υπόλοιπες κατηγορίες, με την αύξηση έστω και για μερικά εικονοστοιχεία, των ορθά ταξινομημένων εικονοστοιχείων. Αντίστοιχα, διαφορά μερικών εικονοστοιχείων καταγράφεται και για τα στοιχεία που δεν είναι ορθά ταξινομημένα στις κατηγορίες που τα ενέταξε ο αλγόριθμος ταξινόμησης. Αυτή η διαπίστωση από τον πίνακα σύγχυσης έρχεται να επιβεβαιώσει το αντίστοιχο συμπέρασμα που προέκυψε κατά την ποιοτική αξιολόγηση.

Μια γενική αναφορά σε μικρές διαφορές εικονοστοιχείων δεν προσφέρει σημαντικά στην αξιολόγηση του αποτελέσματος, με τα μεγέθη των ακριβειών να είναι αυτά που όντως κρίνουν το αποτέλεσμα. Το πρώτο μέγεθος ποσοτικής αξιολόγησης που υπολογίζεται είναι η συνολική ακρίβεια που καταγράφει η εφαρμογή υποδειγματοληψίας στις τεχνητές επιφάνειες ,η οποία ανέρχεται στο 78,89%, καταγράφοντας μικρή αύξηση συγκριτικά με το αρχικό πείραμα. Οι επιμέρους ακρίβειες παραγωγού και χρήστη αποτελούν πιο περιγραφικά μεγέθη, αποδίδοντας με γλαφυρό τρόπο το αποτέλεσμα του πίνακα σύγχυσης και παρουσιάζονται στον παρακάτω πίνακα.

	Precision	**Recall**	**F1-score**	**Συνολική ακρίβεια (%)**	**K-hat (%)**
Τεχνητές	1	1	1	78,89	76,25
Αρόσιμη	0,7709	0,6174	0,6856		
Μόνιμες	0,8828	0,9595	0,9196		
Λιβάδια	0,4308	0,3971	0,4133		
Ετερογενείς	0,6860	0,7794	0,7297		
Δάση	0,8713	0,7884	0,8277		
Συνδυασμοί	0,4884	0,5493	0,5171		
Χερσαία	0,9722	1	0,9859		
Θάλασσα	1	1	1		

Πίνακας 15. Μετρικές αξιολόγησης της ταξινόμησης της πολυφασματικής απεικόνισης με τον αλγόριθμο Random Forest, με υποδειγματοληψία των τεχνητών επιφανειών

Τις χαμηλότερες ακρίβειες σημειώνουν οι κατηγορίες που σχετίζονται με τη βλάστηση, με τη χειρότερη περίπτωση να αποτελεί η κατηγορία των λιβαδιών που καταγράφουν ακρίβειες κάτω της τάξης του 50%. Οι κατηγορίες που σημειώνουν σημαντικά υψηλές ακρίβειες είναι οι τεχνητές επιφάνειες, τα χερσαία ύδατα και η θάλασσα, με τις μόνιμες καλλιέργειες να επιτυγχάνουν υψηλότερη ακρίβεια από κάθε άλλη κατηγορία βλάστησης.

Έχοντας πλέον ένα μη ισορροπημένο σύνολο δεδομένων εκπαίδευσης (μη ισορροπημένο ως προς την έκταση), η μετρική που χαρακτηρίζεται ενδεικτική για την αξιολόγηση της απόδοσης του αλγορίθμου είναι αυτή του F1-score, δηλαδή τον αρμονικό μέσο των μεγεθών precision και recall. Αναμενόμενη δε η χαμηλή απόδοση που σημειώνει η κατηγορία των λιβαδιών. Χαμηλή επίσης είναι η τιμή του F1-score για την κατηγορία των συνδυασμών βλάστησης, μιας και παρά το γεγονός πως έχει ταξινομηθεί ορθά σημαντικό πλήθος εικονοστοιχείων, είναι μεγάλη και η μερίδα των εικονοστοιχείων που έχουν ταξινομηθεί λανθασμένα σε άλλες κατηγορίες βλάστησης. Οι αποδόσεις των δασικών εκτάσεων, της αρόσιμης γης και των μόνιμων καλλιεργειών σημειώνουν μια αύξηση, με τιμές περί το 0.68 με 0.92. Οι κατηγορίες με τις καλύτερες αποδόσεις είναι τα χερσαία ύδατα, η θάλασσα και οι τεχνητές επιφάνειες με τιμές που τείνουν ή/κι ίσες με τη μονάδα. Εν ολίγοις, γίνεται λόγος για τιμές πολύ κοντινές με αυτές που υπολογίστηκαν κατά την ταξινόμηση με ισορροπημένο σύνολο, καταγράφοντας μικρές αυξήσεις ή μειώσεις.

Ο δείκτης συμφωνίας k-hat για την ταξινόμηση με εφαρμογή τυχαίας υποδειγματοληψίας στην κατηγορία των τεχνητών επιφανειών προκύπτει ίσος με 76,25%. Να σημειωθεί πως η τιμή k-hat = 0.76 είναι λίγο μικρότερη από τη συνολική ακρίβεια ≈ 0.79 που υπολογίστηκε νωρίτερα.

7.3.2.Αρόσιμη γη

Επόμενη κατηγορία κάλυψης γης της περιοχής μελέτης για την οποία κρίνεται αν κατά την πειραματική εφαρμογή επιβλεπόμενης ταξινόμησης θα ακολουθήσει τυχαία υπερδειγματοληψία ή υποδειγματοληψία είναι η αρόσιμη γη, της οποίας ο δείκτης αναλογίας ισορροπίας είναι ίσος με 1,148(>1), υποδεικνύοντας με αυτόν τον τρόπο τη μέθοδο της υποδειγματοληψίας. Έπειτα από διαδοχικές δοκιμές, η κατηγορία έρχεται σε ισορροπίας (δείκτης αναλογίας ισορροπίας = 1), μόνο με την υποδειγματοληψία των δειγμάτων της και χωρίς επιπλέον τροποποιήσεις σε πολύγωνα άλλων κατηγοριών.

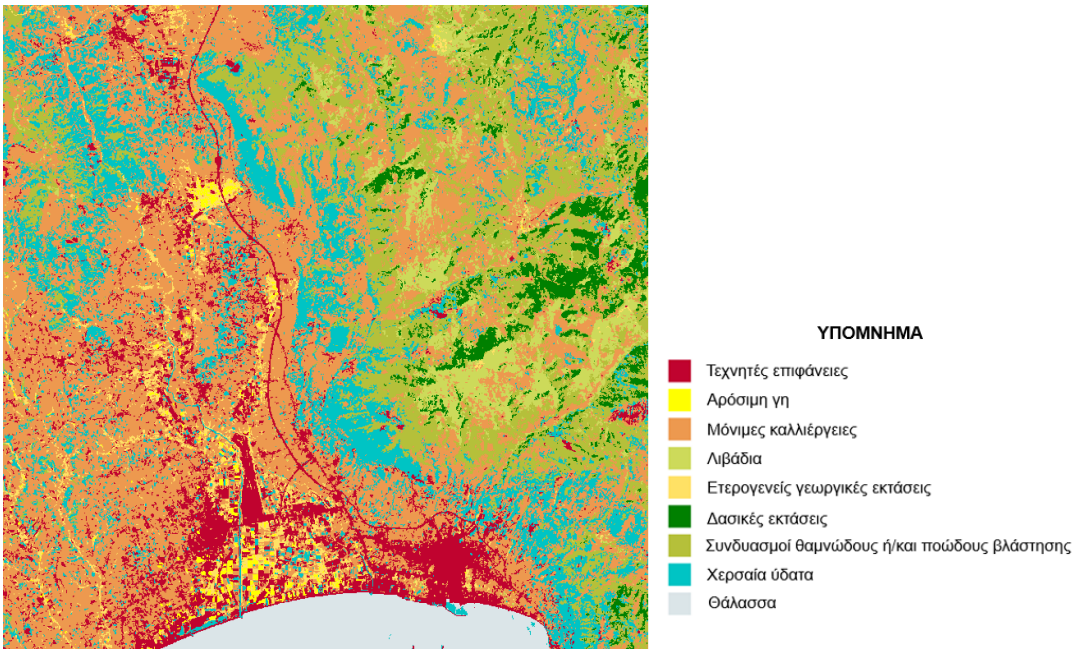
Κατηγορίες CLC	Πλήθος πολυγώνων	Συχνότητα	BRI
Τεχνητές επιφάνειες	63	0,143	1,286
Αρόσιμη γη	49	0,111	1,000
Μόνιμες καλλιέργειες	50	0,118	1,061
Λιβάδια	47	0,107	0,959
Ετερογενείς γεωργικές εκτάσεις	57	0,129	1,163
Δασικές εκτάσεις	38	0,086	0,776
Συνδυασμοί βλάστησης	56	0,127	1,143
Χερσαία ύδατα	45	0,102	0,918
Θάλασσα	34	0,077	0,694

Πίνακας 16. Μορφή συνόλου δεδομένων εκπαίδευσης για την ισορροπία της αρόσιμης γης

Συνοψίζοντας, τα αποτελέσματα που παρατίθενται παρακάτω προκύπτουν με την εφαρμογή τυχαίας υποδειγματοληψίας στην κατηγορία αρόσιμης γης, που στο αρχικά ισορροπημένο σύνολο περιγράφεται με 57 πολύγωνα εκπαίδευσης και καταλήγει με 8 στο νέο dataset.

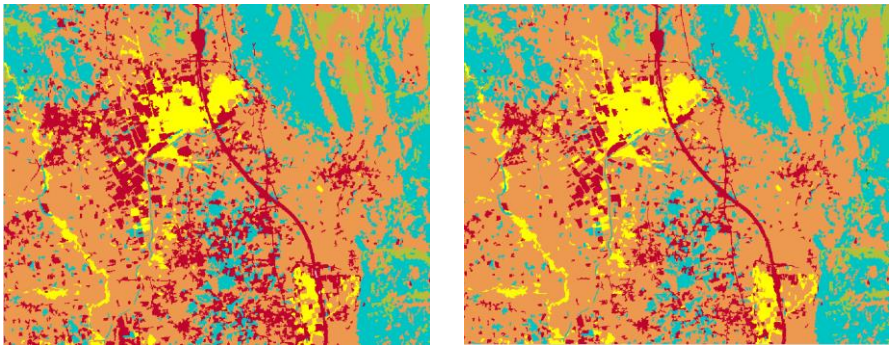
7.3.2.1.Ποιοτική Αξιολόγηση

Έπειτα από εφαρμογή τυχαίας υποδειγματοληψίας για την κατηγορία της αρόσιμης γης, η ταξινομημένη απεικόνιση που προκύπτει είναι

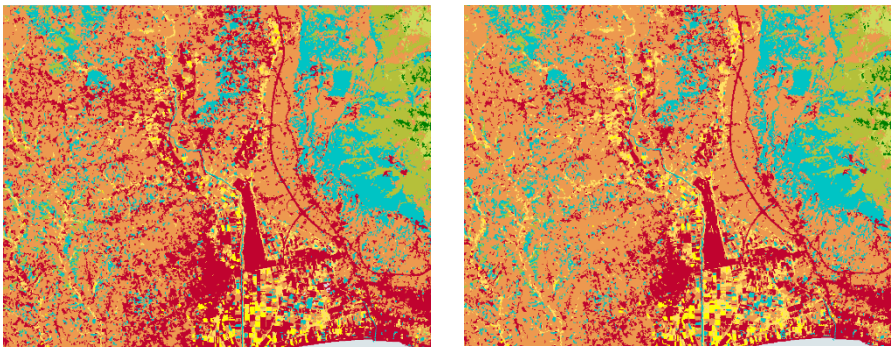


Σχήμα 81. Αποτέλεσμα ταξινόμησης Random Forest, με 100 δέντρα, πολυφασματικής απεικόνισης Sentinel-2 ημερομηνίας λήψης 17/07/2023, με τη μέθοδο τυχαίας υποδειγματοληψίας για την αρόσιμη γη

Στο πείραμα ταξινόμησης της πολυφασματικής απεικόνισης Sentinel-2A με εφαρμογή υποδειγματοληψίας στο σύνολο δεδομένων εκπαίδευσης στην κατηγορία της αρόσιμης γης, τρεις φαίνεται να είναι οι κατηγορίες που επηρεάζονται, οι τεχνητές επιφάνειες, η αρόσιμη γη και οι μόνιμες καλλιέργειες. Ουσιαστικά, από την ποιοτική ερμηνεία του Σχήματος 81 προκύπτει η ενίσχυση της αρόσιμης γης, “διώχνοντας” αρκετά εικονοστοιχεία εντός των εκτάσεών της που είχαν κατηγοριοποιηθεί στις τεχνητές επιφάνειες, με το ίδιο φαινόμενο να παρουσιάζεται και για τις μόνιμες καλλιέργειες. Παρακάτω παρουσιάζονται ενδεικτικά αποσπάσματα αυτών των παρατηρήσεων.



Σχήμα 82. Ταξινόμηση με ισορροπημένο dataset εκπαίδευσης (αριστερά) - Ταξινόμηση με τυχαία υποδειγματοληψία αρόσιμης γης (δεξιά)



Σχήμα 83. Ταξινόμηση με ισορροπημένο dataset εκπαίδευσης (αριστερά) - Ταξινόμηση με τυχαία υποδειγματοληψία αρόσιμης γης (δεξιά)

7.3.2.2.Ποσοτική Αξιολόγηση

Οι μετρικές που αφορούν την ποσοτική αξιολόγηση προκύπτουν από τον πίνακα σύγχυσης, ο οποίος για το πείραμα υποδειγματοληψίας της αρόσιμης γης έχει τη μορφή :

V_Classified	Τεχνητές	Αρόσιμη	Μόνιμες	Λιβάδια	Ετερογενείς	Δάση	Συνδυασμοί	Χερσαία	Θάλασσα	Total
Τεχνητές	**719**	0	0	0	0	0	0	0	0	**719**
Αρόσιμη	0	**440**	0	0	118	0	0	0	0	**558**
Μόνιμες	0	13	**659**	61	2	0	11	0	0	**746**
Λιβάδια	0	1	16	**272**	13	76	246	0	0	**624**
Ετερογενείς	0	242	0	0	**565**	16	0	0	0	**823**
Δάση	0	0	0	38	0	**532**	44	0	0	**614**
Συνδυασμοί	0	1	4	319	0	80	**383**	0	0	**787**
Χερσαία	0	6	12	0	0	0	6	**699**	0	**723**
Θάλασσα	0	0	0	0	0	0	0	0	**692**	**692**
Total	**719**	**703**	**691**	**690**	**698**	**704**	**690**	**699**	**692**	**6286**

Πίνακας 17. Πίνακας σύγχυσης ταξινόμησης πολυφασματικής απεικόνισης με ημερομηνία λήψης 17/07/2023 με χρήση του αλγορίθμου Random Forest, με τη μέθοδο υποδειγματοληψίας για την αρόσιμη γη

Από τον πίνακα σύγχυσης προκύπτει πως στην κατηγορία της αρόσιμης γης, σύμφωνα με τα δεδομένα ελέγχου 118 εικονοστοιχεία ανήκουν στην κατηγορία, αλλά δεν ταξινομήθηκαν σε αυτή, ενώ τα εσφαλμένα ταξινομημένα στην κατηγορία εικονοστοιχεία φτάνουν τα 263, με τα 242 να ανήκουν στις ετερογενείς καλλιέργειες, γεγονός που επηρεάζει την ακρίβεια χρήστη. Βέβαια, όπως έχει αναφερθεί και κατά την ερμηνεία του πίνακα σύγχυσης του ισορροπημένου συνόλου, τέτοιου είδους σφάλμα συμπερίληψης είναι αποτέλεσμα των εγγενών χαρακτηριστικών των κατηγοριών. Στις υπόλοιπες καλύψεις γης της περιοχής μελέτης παρατηρούνται αντίστοιχα σφάλματα συμπερίληψης και παράλειψης με τα προηγούμενα πειράματα.

Οι παρατηρήσεις που απορρέουν από τον πίνακα σύγχυσης γίνονται καλύτερα κατανοητές μέσα από τα μεγέθη που περιγράφουν την ακρίβεια και παρουσιάζονται στον παρακάτω πίνακα.

	Precision	**Recall**	**F1-score**	**Συνολική ακρίβεια (%)**	**K-hat (%)**
Τεχνητές	1	1	1	78,92	76,29
Αρόσιμη	0,7885	0,6259	0,6979		
Μόνιμες	0,8834	0,9537	0,9172		
Λιβάδια	0,4359	0,3942	0,4140		
Ετερογενείς	0,6865	0,8095	0,7429		
Δάση	0,8664	0,7557	0,8073		
Συνδυασμοί	0,4867	0,5551	0,5186		
Χερσαία	0,9668	1	0,9831		
Θάλασσα	1	1	1		

Πίνακας 18. Μετρικές αξιολόγησης της ταξινόμησης της πολυφασματικής απεικόνισης με τον αλγόριθμο Random Forest, με υποδειγματοληψία της αρόσιμης γης

Στις κατηγορίες αρόσιμη γη, μόνιμες καλλιέργειες, λιβάδια κι ετερογενείς καλλιέργειες τείνει να έχει βελτιωμένη απόδοση το precision συγκριτικά με τα προηγούμενα πειράματα, σε αντίθεση με το recall που καταγράφει μειωμένη απόδοση στο συγκεκριμένο πείραμα για τις μόνιμες καλλιέργειες, τις δασικές εκτάσεις και τους συνδυασμούς βλάστησης. Ιδιαίτερο ενδιαφέρον παρουσιάζει η περίπτωση των μόνιμων καλλιεργειών που καταγράφουν ταυτόχρονα αύξηση και μείωση στις επιμέρους ακρίβειες, με την αυξομείωση αυτή να καταγράφεται στον αρμονικό τους μέσο ,F1-score, με αποτέλεσμα στο παρόν πείραμα το μέγεθος του F1-score να καταγράφει τη μικρότερη απόδοση συγκριτικά με τα προηγούμενα πειράματα. Βέβαια, αξίζει να σημειωθεί η απόλυτη ακρίβεια (=1) που επιτυγχάνει σταθερά η κατηγορία των τεχνητών επιφανειών και της θάλασσας σε όλα τα πειράματα, όπως επίσης και η σταθερότητα στην απόδοση των χερσαίων υδάτων.

Κρίνοντας την απόδοση του μοντέλου από το μέγεθος του F1-score, καταγράφονται 8 από τις 9 κατηγορίες με τιμή μεγαλύτερη του 0,5, με τη μονάδα να χαρακτηρίζει την καλύτερη απόδοση που θα μπορούσε να καταγράψει το μοντέλο, ενώ σταθερά χαμηλή είναι η απόδοση των λιβαδιών, καθώς σε όλα τα πειράματα μέχρι στιγμής ταξινομούνται η ακρίβειά τους κυμαίνεται στο 0,40.

Από την άλλη, σταθερά κοντά στο 79% η συνολική ακρίβεια της ταξινόμησης, με λίγο μικρότερη ακρίβεια να καταγράφει ο δείκτης συμφωνίας K-hat κοντά στο 76%.

7.3.3.Μόνιμες καλλιέργειες

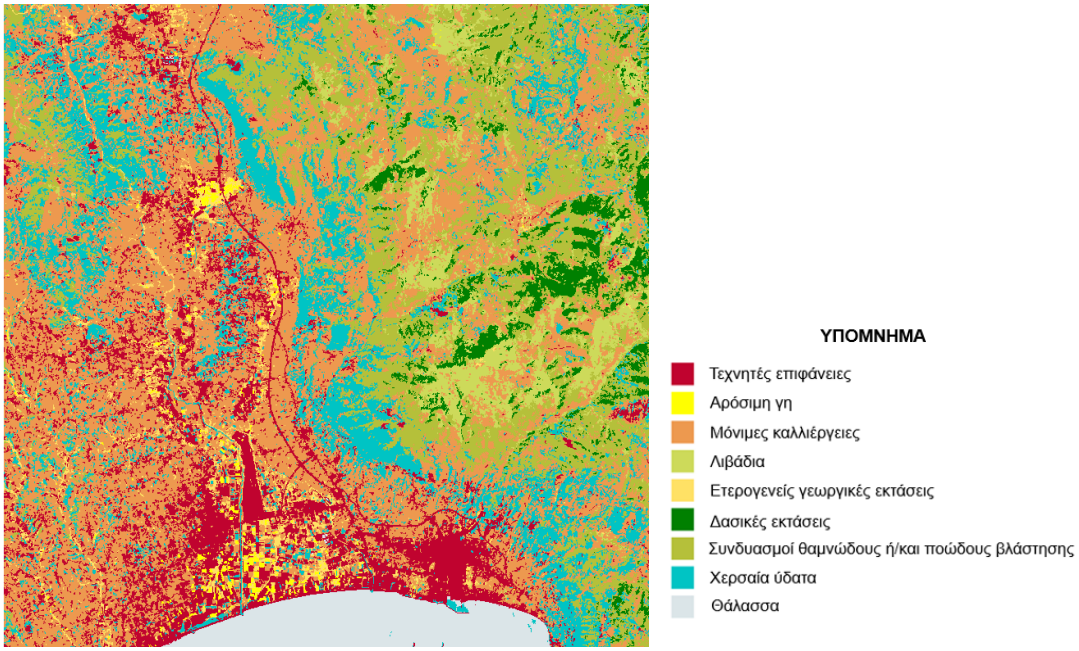
Ως πρόκληση αντιμετωπίζεται η κατηγορία των μόνιμων καλλιεργειών, που έχουν τιμή δείκτη αναλογίας ίσο με 1.007, μια τιμή που τείνει στη μονάδα κι υπό άλλες συνθήκες δε θα δεχόταν καμία επεξεργασία. Ωστόσο, στην προκειμένη περίπτωση θεωρείται ευκαιρία για το συνδυασμό των δυο μεθόδων και να εφαρμοσθεί τυχαία υπερδειγματοληψία και τυχαία υποδειγματοληψία, με παράλληλο στόχο την αξιολόγηση του πόσο δύναται να επηρεαστεί το αποτέλεσμα της ταξινόμησης, με την προσθήκη ή την αφαίρεση ενός ή δυο πολυγώνων σε κάθε κατηγορία. Οι προσαυξήσεις αυτές οδηγούν στην παρακάτω μορφή του dataset.

Κατηγορίες CLC	Πλήθος πολυγώνων	Συχνότητα	BRI
Τεχνητές επιφάνειες	63	0,140	1,260
Αρόσιμη γη	58	0,129	1,160
Μόνιμες καλλιέργειες	50	0,111	1,000
Λιβάδια	46	0,102	0,920
Ετερογενείς γεωργικές εκτάσεις	58	0,129	1,160
Δασικές εκτάσεις	40	0,089	0,800
Συνδυασμοί βλάστησης	56	0,124	1,120
Χερσαία ύδατα	44	0,098	0,880
Θάλασσα	35	0,078	0,700

Πίνακας 19. Μορφή συνόλου δεδομένων εκπαίδευσης για την ισορροπία των μόνιμων καλλιεργειών

7.3.3.1.Ποιοτική Αξιολόγηση

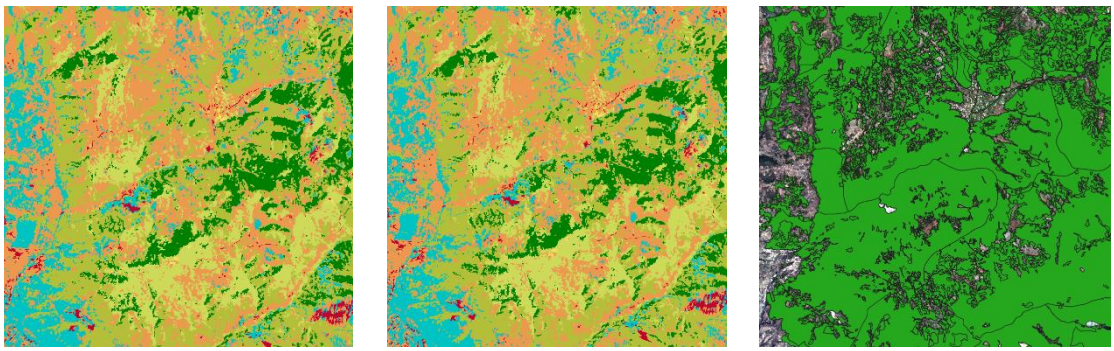
Ο συνδυασμός των δυο μεθόδων με μικρές αλλαγές στο πλήθος των πολυγώνων κάθε κατηγορίας έχει ως αποτέλεσμα η μορφή της ταξινομημένης απεικόνισης να είναι :



Σχήμα 84. Αποτέλεσμα ταξινόμησης Random Forest, με 100 δέντρα, πολυφασματικής απεικόνισης Sentinel-2 ημερομηνίας λήψης 17/07/2023, με τη μέθοδο τυχαίας υποδειγματοληψίας & τυχαίας υπερδειγματοληψίας για την ισορροπίας των μόνιμων καλλιεργειών

Οι μικρές τροποποιήσεις που έγιναν στο αρχικό, ισορροπημένο dataset, ώστε να έρθει στην απόλυτη ισορροπία ο δείκτης αναλογίας των μόνιμων καλλιεργειών, είχαν ως

αποτέλεσμα να σημειώνεται ποιοτικά μια αύξηση της έκτασης των δασικών εκτάσεων, ενώ για την κατηγορία ενδιαφέροντος, μόνιμες καλλιέργειες, δε σημειώνεται διαφορά.



Σχήμα 85. Ταξινόμηση με ισορροπημένο dataset εκπαίδευσης (αριστερά) - Ταξινόμηση με συνδυασμό τυχαίας υπερδειγματοληψίας & τυχαίας υποδειγματοληψίας (μέση) – Κυρωμένος Δασικός Χάρτης (δεξιά)

Η παράθεση των τριών αποσπασμάτων γίνεται για την ενίσχυση του ισχυρισμού ότι αυξάνεται η έκταση των δασικών εκτάσεων στο συγκεκριμένο πείραμα κι επαληθεύεται ότι πρόκειται για δασικές εκτάσεις από τον Κυρωμένο Δασικό χάρτη της περιοχής ενδιαφέροντος.

7.3.3.2.Ποσοτική Αξιολόγηση

Οι μετρικές που αφορούν την ποσοτική αξιολόγηση προκύπτουν από τον πίνακα σύγχυσης, ο οποίος για το συγκεκριμένο πείραμα που υλοποιείται συνδυασμός μεθόδων υποδειγματοληψίας & υπερδειγματοληψίας για την απόλυτη ισορροπία του δείκτη των μόνιμων καλλιεργειών έχει τη μορφή :

V_Classified	Τεχνητές	Αρόσιμη	Μόνιμες	Λιβάδια	Ετερογενείς	Δάση	Συνδυασμοί	Χερσαία	Θάλασσα	Total
Τεχνητές	**719**	0	0	0	0	0	0	0	0	**719**
Αρόσιμη	0	**437**	0	0	120	0	0	0	0	**557**
Μόνιμες	0	12	**661**	58	2	0	12	0	0	**745**
Λιβάδια	0	0	15	**275**	21	79	243	0	0	**633**
Ετερογενείς	0	246	0	0	**555**	0	0	0	0	**801**
Δάση	0	0	0	36	0	**553**	51	0	0	**640**
Συνδυασμοί	0	2	4	321	0	72	**379**	0	0	**778**
Χερσαία	0	6	11	0	0	0	5	**699**	0	**721**
Θάλασσα	0	0	0	0	0	0	0	0	**692**	**692**
Total	**719**	**703**	**691**	**690**	**698**	**704**	**690**	**699**	**692**	**6286**

Πίνακας 20. Πίνακας σύγχυσης ταξινόμησης πολυφασματικής απεικόνισης με ημερομηνία λήψης 17/07/2023 με χρήση του αλγορίθμου Random Forest, με το συνδυασμό υποδειγματοληψίας κι υπερδειγματοληψίας για τις μόνιμες καλλιέργειες

Αναλύοντας τον πίνακα σύγχυσης έπειτα από μικρές διορθώσεις του αρχικού συνόλου δεδομένων εκπαίδευσης οι μόνιμες καλλιέργειες που είναι η βασική κατηγορία του πειράματος σημειώνει 84 εικονοστοιχεία που δε συμπεριλήφθησαν σε αυτή κατά την ταξινόμηση, ενώ ταξινομήθηκαν 30 εικονοστοιχεία ως μόνιμες καλλιέργειες, τα οποία σύμφωνα με τα δεδομένα ελέγχου ανήκουν στα λιβάδια, στους συνδυασμούς βλάστησης και τα χερσαία ύδατα. Όπως έχει αναφερθεί και σε προηγούμενες πειραματικές εφαρμογές, σφάλματα μεταξύ των κατηγοριών βλάστησης είναι αναμενόμενα, αλλά το σφάλμα με τα χερσαία ύδατα να ταξινομούνται ως βλάστηση είναι σαφώς ένα πρόβλημα, το οποίο εξηγείται με ενδεχόμενη άρδευση των καλλιεργειών.

Αναλυτικότερα, για μια ολοκληρωμένη ποσοτική αξιολόγηση του αποτελέσματος της ταξινόμησης της πολυφασματικής απεικόνισης υπολογίζονται οι μετρικές precision, recall, F1-score, συνολική ακρίβεια και δείκτης συμφωνίας k-hat.

	Precision	Recall	F1-score	Συνολική ακρίβεια (%)	K-hat (%)
Τεχνητές	1	1	1	79,06	76,45
Αρόσιμη	0,7846	0,6216	0,6937		
Μόνιμες	0,8872	0,9566	0,9206		
Λιβάδια	0,4344	0,3986	0,4157		
Ετερογενείς	0,6929	0,7951	0,7405		
Δάση	0,8641	0,7855	0,8229		
Συνδυασμοί	0,4871	0,5493	0,5163		
Χερσαία	0,9695	1	0,9845		
Θάλασσα	1	1	1		

Πίνακας 21. Μετρικές αξιολόγησης της ταξινόμησης της πολυφασματικής απεικόνισης με τον αλγόριθμο Random Forest, με συνδυασμό υποδειγματοληψία & υπερδειγματοληψία για τις μόνιμες καλλιέργειες

Όπως προκύπτει από τις μετρικές αξιολόγησης, η επιβλεπόμενη ταξινόμηση με συνδυασμό μεθόδων τυχαίας υποδειγματοληψίας κι υπερδειγματοληψίας για την ισορροπία των μόνιμων καλλιεργειών, το σύνολο των κατηγοριών κάλυψης γης καταγράφει υψηλή ακρίβεια precision, με τις μόνες κάτω του 0,5 να είναι τα λιβάδια και οι συνδυασμοί βλάστησης, καθώς σημαντική μερίδα εικονοστοιχείων των λιβαδιών έχουν ταξινομηθεί συνδυασμοί βλάστησης και το αντίστροφο, εικονοστοιχεία των συνδυασμών βλάστησης, στην ταξινομημένη απεικόνιση κατατάσσονται στα λιβάδια. Οι ετερογενείς καλλιέργειες επίσης, σημειώνουν λίγο χαμηλότερη ακρίβεια precision συγκριτικά με τις υπόλοιπες, λόγω του γεγονότος ότι 246 εικονοστοιχεία τους, έχουν ταξινομηθεί ως αρόσιμη γη, που κατ' επέκταση επηρεάζει το recall της αρόσιμης.

Κρίνοντας την απόδοση του μοντέλου από το μέγεθος του F1-score, καταγράφονται 8 από τις 9 κατηγορίες με τιμή μεγαλύτερη του 0,5, με τη μονάδα να χαρακτηρίζει την καλύτερη απόδοση που θα μπορούσε να καταγράψει το μοντέλο, με τη μέτρια απόδοση των συνδυασμών βλάστησης κοντά στο 0,5 και τη χαμηλότερη επίδοση των λιβαδιών με τιμή πλησίον του 0,42.

Από την άλλη, η συνολική ακρίβεια της ταξινόμησης ίση με 79,06%, με το συγκεκριμένο πείραμα να σημειώνει τη μικρότερη συνολική ακρίβεια συγκριτικά με όλα τα πειράματα που έχουν υλοποιηθεί έως αυτό το σημείο, με λίγο μικρότερη ακρίβεια να καταγράφει ο δείκτης συμφωνίας K-hat κοντά στο 76%.

7.3.4.Λιβάδια

Στη σειρά των πειραματικών εφαρμογών για την αξιολόγηση της ακρίβειας που επιτυγχάνεται, ο δείκτης αναλογίας ισορροπίας για την κατηγορία των λιβαδιών υποδεικνύει την ανάγκη για υπερδειγματοληψία, αφού έχει τιμή ίση με 0,976 (<1). Εξετάζοντας την περιοχή ενδιαφέροντος προκύπτει ότι είναι δυνατή η δειγματοληψία επιπλέον πολυγώνων εκπαίδευσης για τα λιβάδια, επομένως, το σύνολο εκπαίδευσης του νέου πειράματος διαμορφώνεται ως εξής

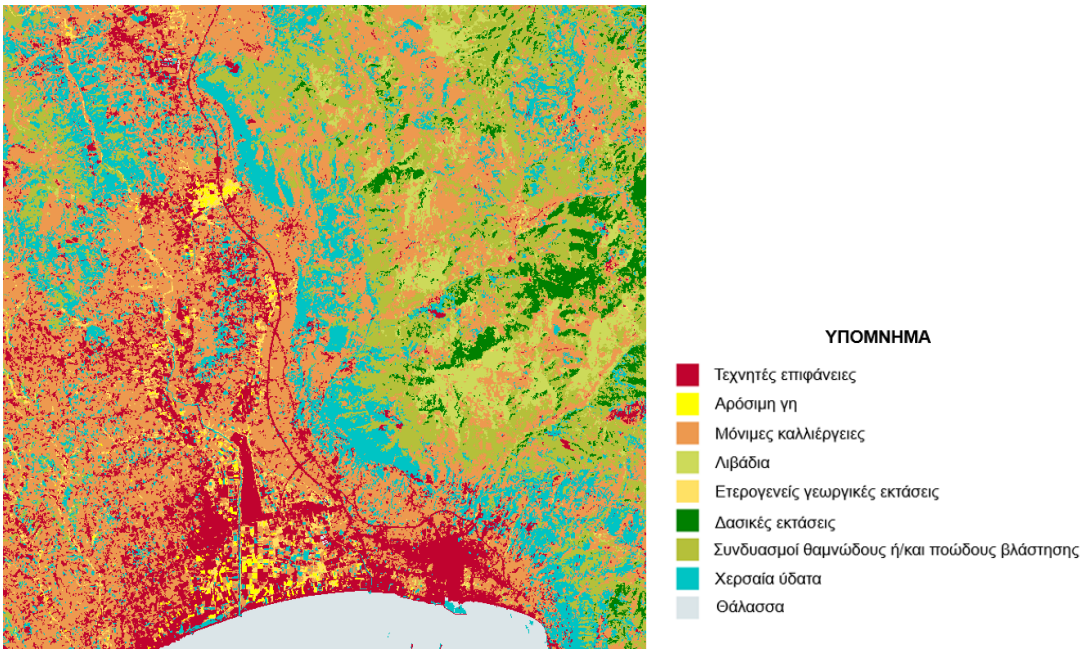
Κατηγορίες CLC	Πλήθος πολυγώνων	Συχνότητα	BRI
Τεχνητές επιφάνειες	63	0,140	1,260
Αρόσιμη γη	57	0,127	1,140
Μόνιμες καλλιέργειες	50	0,111	1,000
Λιβάδια	50	0,111	1,000
Ετερογενείς γεωργικές εκτάσεις	57	0,127	1,140
Δασικές εκτάσεις	38	0,084	0,760
Συνδυασμοί βλάστησης	56	0,124	1,120
Χερσαία ύδατα	45	0,100	0,900
Θάλασσα	34	0,076	0,680

Πίνακας 22. Μορφή συνόλου δεδομένων εκπαίδευσης για την ισορροπία των λιβαδιών

Αξίζει δε, να σημειωθεί σε αυτό το σημείο πως με το παραπάνω σύνολο δεδομένων έρχεται σε ισορροπία και η κατηγορία των μόνιμων καλλιεργειών, διαμορφώνοντας με αυτόν τον τρόπο ένα ενδεικτικό σενάριο για την κατηγορία που εξετάστηκε προηγουμένως με άλλη μεθοδολογία.

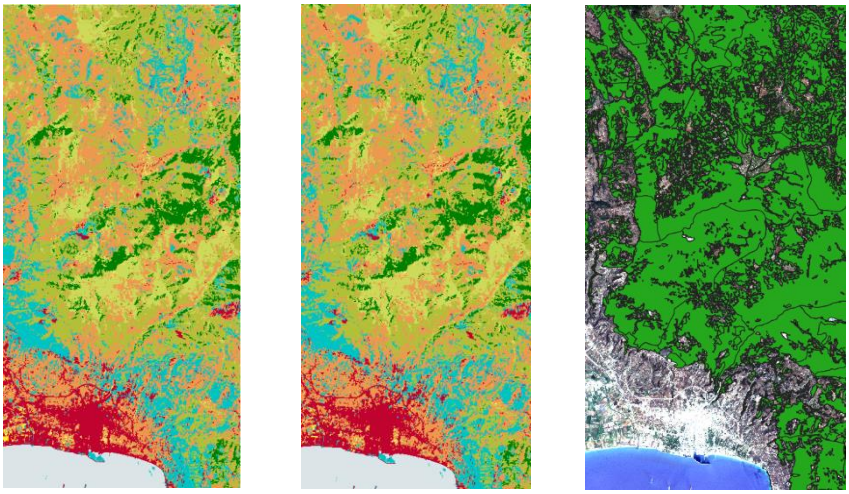
7.3.4.1.Ποιοτική Αξιολόγηση

Ακολουθώντας τη μέθοδο τυχαίας υπερδειγματοληψίας με την προσθήκη τριών πολυγώνων εκπαίδευσης για την κατηγορία των λιβαδιών, η ταξινόμηση της πολυφασματικής απεικόνισης Sentinel-2A έχει ως αποτέλεσμα



Σχήμα 86. Αποτέλεσμα ταξινόμησης Random Forest, με 100 δέντρα, πολυφασματικής απεικόνισης Sentinel-2 ημερομηνίας λήψης 17/07/2023, με τη μέθοδο τυχαίας υπερδειγματοληψίας για τα λιβάδια

Η υπερδειγματοληψία στην κατηγορία των λιβαδιών βοηθά ποιοτικά στην πύκνωση των εικονοστοιχείων των δασικών εκτάσεων έναντι αυτών που αρχικά, στην εφαρμογή του ισορροπημένου συνόλου εκπαίδευσης είχαν κατηγοριοποιηθεί ως λιβάδια και συνδυασμοί βλάστησης, τρεις κατηγορίες οι οποίες παρουσιάζουν παρόμοια συμπεριφορά στο φάσμα, γεγονός αντιληπτό και κατά την ερμηνεία του έγχρωμου σύνθετου της πολυφασματικής απεικόνισης.



Σχήμα 87. Ταξινόμηση με ισορροπημένο dataset εκπαίδευσης (αριστερά) - Ταξινόμηση με τυχαία υπερδειγματοληψία (μέση) – Κυρωμένος Δασικός Χάρτης (δεξιά)

Η παράθεση των τριών αποσπασμάτων γίνεται για την ενίσχυση του ισχυρισμού ότι αυξάνεται η έκταση των δασικών εκτάσεων στο συγκεκριμένο πείραμα κι επαληθεύεται ότι πρόκειται για δασικές εκτάσεις από τον Κυρωμένο Δασικό χάρτη της περιοχής ενδιαφέροντος.

7.3.4.2.Ποσοτική Αξιολόγηση

Οι μετρικές που αφορούν την ποσοτική αξιολόγηση προκύπτουν από τον πίνακα σύγχυσης, ο οποίος για το πείραμα υπερδειγματοληψίας των λιβαδιών σχηματίζεται ως εξής :

V_Classified	Τεχνητές	Αρόσιμη	Μόνιμες	Λιβάδια	Ετερογενείς	Δάση	Συνδυασμοί	Χερσαία	Θάλασσα	Total
Τεχνητές	**719**	0	0	0	0	0	0	0	0	**719**
Αρόσιμη	0	**431**	0	0	129	0	0	0	0	**560**
Μόνιμες	0	11	**660**	57	2	0	11	0	0	**741**
Λιβάδια	0	4	16	**271**	19	82	240	0	0	**632**
Ετερογενείς	0	249	0	0	**548**	0	0	0	0	**797**
Δάση	0	0	0	37	0	**547**	50	0	0	**634**
Συνδυασμοί	0	2	4	325	0	75	**385**	0	0	**791**
Χερσαία	0	6	11	0	0	0	4	**699**	0	**720**
Θάλασσα	0	0	0	0	0	0	0	0	**692**	**692**
Total	**719**	**703**	**691**	**690**	**698**	**704**	**690**	**699**	**692**	**6286**

Πίνακας 23. Πίνακας σύγχυσης ταξινόμησης πολυφασματικής απεικόνισης με ημερομηνία λήψης 17/07/2023 με χρήση του αλγορίθμου Random Forest, με υπερδειγματοληψία για τα λιβάδια

Σταθερά αποδίδουν οι τεχνητές επιφάνειες και η θάλασσα, με όλα τα εικονοστοιχεία των δεδομένων ελέγχου να αναγνωρίζουν μόνο ορθά ταξινομημένα εικονοστοιχεία στο θεματικό χάρτη που παράγεται από την επιβλεπόμενη ταξινόμηση της πολυφασματικής απεικόνισης. Οι κατηγορίες βλάστησης δε, συνεχίζουν να καταγράφουν σφάλματα συμπερίληψης και παράλειψης, με εικονοστοιχεία της αρόσιμης γης να ταξινομούνται ως ετερογενείς καλλιέργειες, με ταυτόχρονη

ταξινόμηση εικονοστοιχείων των ετερογενών ως αρόσιμη γη, φαινόμενο που οδηγεί σε αντίστοιχες τιμές στο precision και το recall των δυο αυτώ κατηγοριών, όπως προκύπτει κι από τον παρακάτω πίνακα. Το μεγαλύτερο σφάλμα παράλειψης σημειώνεται για την κατηγορία των συνδυασμών βλάστησης, με τα ορθά ταξινομημένα εικονοστοιχεία να αγγίζουν τα 385 ,με 325 εικονοστοιχεία της κατηγορίας να έχουν ταξινομηθεί ως λιβάδια. Αντίστοιχα, παρόμοιας διακύμανσης σφάλμα παράλειψης εμφανίζεται στην περίπτωση των λιβαδιών με 271 ορθά ταξινομημένα εικονοστοιχεία και 240 εικονοστοιχεία λιβαδιών να εντάσσονται στους συνδυασμούς βλάστησης. Όλα τα συμπεράσματα που απορρέουν από τον πίνακα σύγχυσης, καταγράφονται και ποσοτικοποιούνται στον Πίνακα 24, με τη μετρική F1-score να περιγράφει ολοκληρωμένα πως έχουν επηρεάσει τα σφάλματα παράλειψης και συμπερίληψης τη συνολική απόδοση κάθε κατηγορίας χρήσεων γης.

	Precision	**Recall**	**F1-score**	**Συνολική ακρίβεια (%)**	**K-hat (%)**
Τεχνητές	1	1	1	78,78	76,13
Αρόσιμη	0,7696	0,6131	0,6825		
Μόνιμες	0,8907	0,9551	0,9218		
Λιβάδια	0,4288	0,3928	0,4100		
Ετερογενείς	0,6876	0,7851	0,7331		
Δάση	0,8628	0,7770	0,8176		
Συνδυασμοί	0,4867	0,5580	0,5199		
Χερσαία	0,9708	1	0,9852		
Θάλασσα	1	1	1		

Πίνακας 24. Μετρικές αξιολόγησης της ταξινόμησης της πολυφασματικής απεικόνισης με τον αλγόριθμο Random Forest, με υπερδειγματοληψία για τα λιβάδια

Τα λανθασμένα ταξινομημένα εικονοστοιχεία (129) της αρόσιμης γης έχουν ως αποτέλεσμα τη μειωμένη απόδοση του recall της κατηγορίας. Από την άλλη, λόγω της σύγχυσης του αλγορίθμου να ταξινομεί μεγάλη μερίδα εικονοστοιχείων των λιβαδιών ως συνδυασμούς βλάστησης κι εικονοστοιχείων των συνδυασμών ως λιβάδια οδηγεί στη χαμηλή απόδοση των δυο κατηγοριών, τόσο στο precision ,όσο και στο recall, με μέτρια προς χαμηλή απόδοση της μετρικής F1-score. Οι υπόλοιπες κατηγορίες καταγράφουν μερικά μόνο εικονοστοιχεία που είναι εσφαλμένα ταξινομημένα σε άλλες κατηγορίες, συγκριτικά με τα ορθά ταξινομημένα, με συνέπεια η απόδοση των χρήσεων γης να είναι υψηλή, με τη χαμηλότερη τιμή να σημειώνει το F1-score της αρόσιμης γης.

7.3.5.Ετερογενείς γεωργικές εκτάσεις

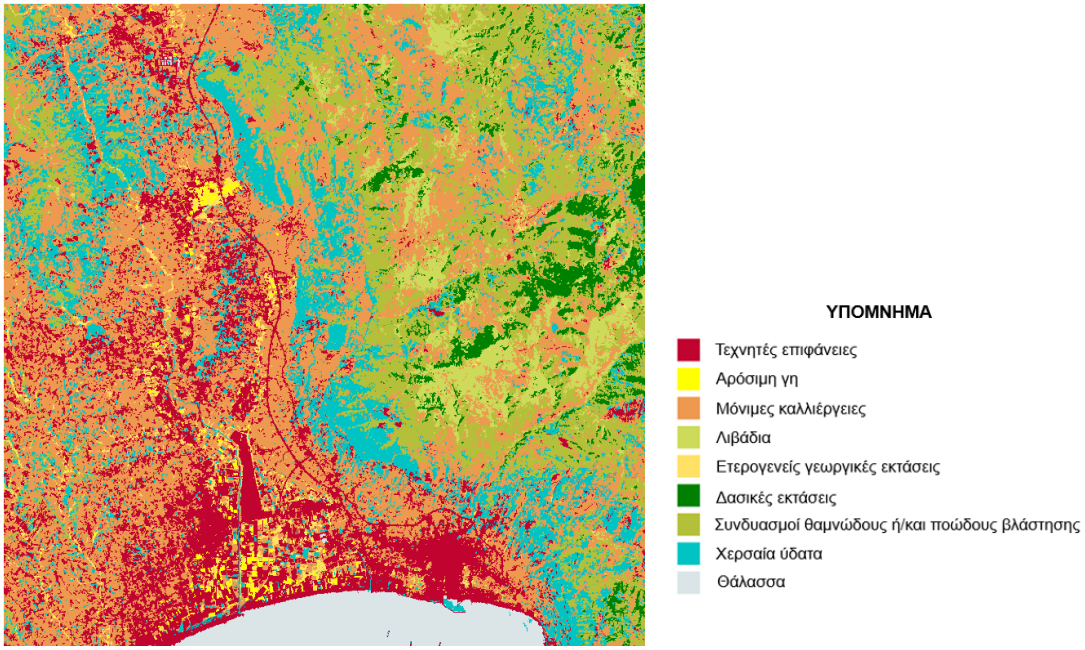
Κατά τον προσδιορισμό του Δείκτη Αναλογίας Ισορροπίας (BRI), Ενότητα 7.2.,όπου συγκρίνοντας την τιμή του δείκτη για την κάθε κατηγορία με τη μονάδα, δίνεται μια πρώτη πιθανή αντιμετώπιση του προβλήματος ανισορροπίας των πολυγώνων εκπαίδευσης της, η οποία με διαδοχικές δοκιμές υποδεικνύει όχι μόνο τον τρόπο αντιμετώπισης, αλλά και κατά πόσο θα αυξηθεί ή θα μειωθεί το πλήθος των πολυγώνων εκπαίδευσης, ώστε να ισορροπήσει ο δείκτης για την εξεταζόμενη κατηγορία. Με αυτόν τον τρόπο, για την κατηγορία των ετερογενών καλλιεργειών εξετάζεται το ενδεχόμενο εφαρμογής υποδειγματοληψίας, το οποίο είναι εφικτό, καταλήγοντας στη μείωση των πολυγώνων της κατηγορίας και το σύνολο πολυγώνων εκπαίδευσης σχηματίζεται ως :

Κατηγορίες CLC	Πλήθος πολυγώνων	Συχνότητα	BRI
Τεχνητές επιφάνειες	63	0,144	1,292
Αρόσιμη γη	57	0,130	1,169
Μόνιμες καλλιέργειες	50	0,114	1,025
Λιβάδια	47	0,107	0,964
Ετερογενείς γεωργικές εκτάσεις	49	0,111	1,000
Δασικές εκτάσεις	38	0,087	0,779
Συνδυασμοί βλάστησης	56	0,128	1,148
Χερσαία ύδατα	45	0,103	0,923
Θάλασσα	34	0,077	0,697

Πίνακας 25. Μορφή συνόλου δεδομένων εκπαίδευσης για την ισορροπία των ετερογενών γεωργικών εκτάσεων

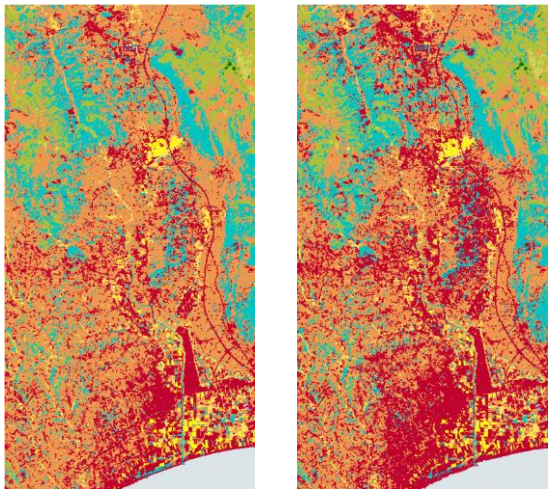
7.3.5.1.Ποιοτική Αξιολόγηση

Η εφαρμογή επιβλεπόμενης ταξινόμησης με σύνολο δεδομένων εκπαίδευσης στο οποίο έχει εφαρμοσθεί υποδειγματοληψία της κατηγορίας των ετερογενών γεωργικών εκτάσεων έχει ως αποτέλεσμα την παρακάτω απεικόνιση



Σχήμα 88. Αποτέλεσμα ταξινόμησης Random Forest, με 100 δέντρα, πολυφασματικής απεικόνισης Sentinel-2 ημερομηνίας λήψης 17/07/2023, με τη μέθοδο τυχαίας υποδειγματοληψίας για τις ετερογενείς καλλιέργειες

Η υποδειγματοληψία των ετερογενών καλλιεργειών έχει ως αποτέλεσμα την ενίσχυση των τεχνητών επιφανειών, καθώς σε όλο το δυτικό τομέα της περιοχής ενδιαφέροντος μεγάλο μέρος εικονοστοιχείων κατηγοριών βλάστησης σημειώνονται ως τεχνητές επιφάνειες. Βέβαια, όπως έχει αναφερθεί και σε άλλο πείραμα, ένα τέτοιο σφάλμα ενδέχεται να οφείλεται σε εκτάσεις γυμνού εδάφους, όπως είναι οι περιπτώσεις καλλιεργήσιμων εκτάσεων σε περίοδο αγρανάπαυσης, σε αγροτεμάχια των οποίων το έδαφος έχει υποστεί κατεργασία ή ακόμα και βραχώδεις/πετρώδεις επιφάνειες. Ταυτόχρονα στον ανατολικό τομέα της περιοχής ενισχύονται με μερικά εικονοστοιχεία οι δασικές εκτάσεις.



Σχήμα 89. Ταξινόμηση με ισορροπημένο dataset εκπαίδευσης (αριστερά) - Ταξινόμηση με τυχαία υποδειγματοληψία (δεξιά)

7.3.5.2.Ποσοτική Αξιολόγηση

Οι μετρικές που αφορούν την ποσοτική αξιολόγηση προκύπτουν από τον πίνακα σύγχυσης, ο οποίος για το συγκεκριμένο πείραμα έχει τη μορφή :

V_Classified	Τεχνητές	Αρόσιμη	Μόνιμες	Λιβάδια	Ετερογενείς	Δάση	Συνδυασμοί	Χερσαία	Θάλασσα	Total
Τεχνητές	**719**	0	0	0	1	0	0	0	0	**720**
Αρόσιμη	0	**449**	0	0	130	0	0	0	0	**579**
Μόνιμες	0	12	**660**	60	2	0	12	0	0	**746**
Λιβάδια	0	4	15	**272**	18	78	237	0	0	**624**
Ετερογενείς	0	230	0	0	**547**	0	0	0	0	**777**
Δάση	0	0	0	38	0	**554**	54	0	0	**646**
Συνδυασμοί	0	2	5	320	0	72	**383**	0	0	**782**
Χερσαία	0	6	11	0	0	0	4	**699**	0	**720**
Θάλασσα	0	0	0	0	0	0	0	0	**692**	**692**
Total	**719**	**703**	**691**	**690**	**698**	**704**	**690**	**699**	**692**	**6286**

Πίνακας 26. Πίνακας σύγχυσης ταξινόμησης πολυφασματικής απεικόνισης με ημερομηνία λήψης 17/07/2023 με χρήση του αλγορίθμου Random Forest, με υποδειγματοληψία για τις ετερογενείς καλλιέργειες

Απόρροια της υποδειγματοληψίας των ετερογενών καλλιεργειών είναι το σφάλμα παράλειψης των εικονοστοιχείων της αρόσιμης γης που ταξινομήθηκαν ως ετερογενείς γεωργικές εκτάσεις και παράλληλα το σφάλμα συμπερίληψης εικονοστοιχείων των μόνιμων καλλιεργειών, το σημαντικότερο των ετερογενών και το χειρότερο από άποψη συνάφειας, των χερσαίων υδάτων. Βέβαια, συγκριτικά με τα προηγούμενα πειράματα είναι το μόνο που καταγράφεται ελαττωμένη σύγχυση μεταξύ καλλιεργειών βλάστησης. Ένα ακόμη σφάλμα που καταγράφεται είναι αυτό της ταξινόμησης εικονοστοιχείων των λιβαδιών ως συνδυασμοί βλάστησης και το αντίστροφο, το οποίο όμως συμβαίνει σε

όλα τα πειράματα. Υψηλό είναι το σφάλμα συμπερίληψης 130 εικονοστοιχείων της αρόσιμης γης ως ετερογενείς γεωργικές εκτάσεις. Συνοψίζοντας, η αξιολόγηση των πειραμάτων μέσω του πίνακα σύγχυσης δε σημειώνει σημαντικές διαφοροποιήσεις από πείραμα σε πείραμα, με μικρές αυξομειώσεις στα πλήθη των εικονοστοιχείων.

	Precision	**Recall**	**F1-score**	**Συνολική ακρίβεια (%)**	**K-hat (%)**
Τεχνητές	0,9986	1	0,9993	79,14	76,54
Αρόσιμη	0,7755	0,6387	0,7005		
Μόνιμες	0,8847	0,9551	0,9186		
Λιβάδια	0,4359	0,3942	0,4140		
Ετερογενείς	0,7040	0,7837	0,7417		
Δάση	0,8576	0,7869	0,8207		
Συνδυασμοί	0,4898	0,5551	0,5204		
Χερσαία	0,9708	1	0,9852		
Θάλασσα	1	1	1		

Πίνακας 27. Μετρικές αξιολόγησης της ταξινόμησης της πολυφασματικής απεικόνισης με τον αλγόριθμο Random Forest, με υποδειγματοληψία για τις ετερογενείς καλλιέργειες

Είναι πλέον αντιληπτό ότι στις περιπτώσεις κατηγοριών στις οποίες καταγράφεται σημαντικό πλήθος εικονοστοιχείων που έχουν παραληφθεί από την κατηγορία που εξετάζεται ή έχουν συμπεριληφθεί σε αυτή, χωρίς να ανήκουν σε αυτή, τότε επηρεάζεται σε ανάλογο μέγεθος το precision και το recall της, αντίστοιχα, φαινόμενο που συνοψίζεται με τον αρμονικό μέσο F1-score. Λόγω της υψηλής σύγχυσης μεταξύ των λιβαδιών και των συνδυασμών βλάστησης η απόδοσή τους θεωρείται μέτρια, με τιμές ≤ 0,52, με τις υπόλοιπες κατηγορίες να εμφανίζουν ακρίβειες πάνω από 0,70.

Η συνολική ακρίβεια της ταξινόμησης αγγίζει το 79,14%, όμως είναι ένα μέγεθος που δεν μπορεί να ληφθεί υπόψιν στην αξιολόγηση του αποτελέσματος, καθώς επιδρά σημαντικά σε αυτό η αυξημένη ακρίβεια των τεχνητών επιφανειών, των χερσαίων υδάτων και της θάλασσας. Τέλος, ο δείκτης συμφωνίας k-hat σταθερά έχει μια απόκλιση της τάξης του 3% από τη συνολική ακρίβεια της ταξινόμησης.

7.3.6.Δασικές Εκτάσεις

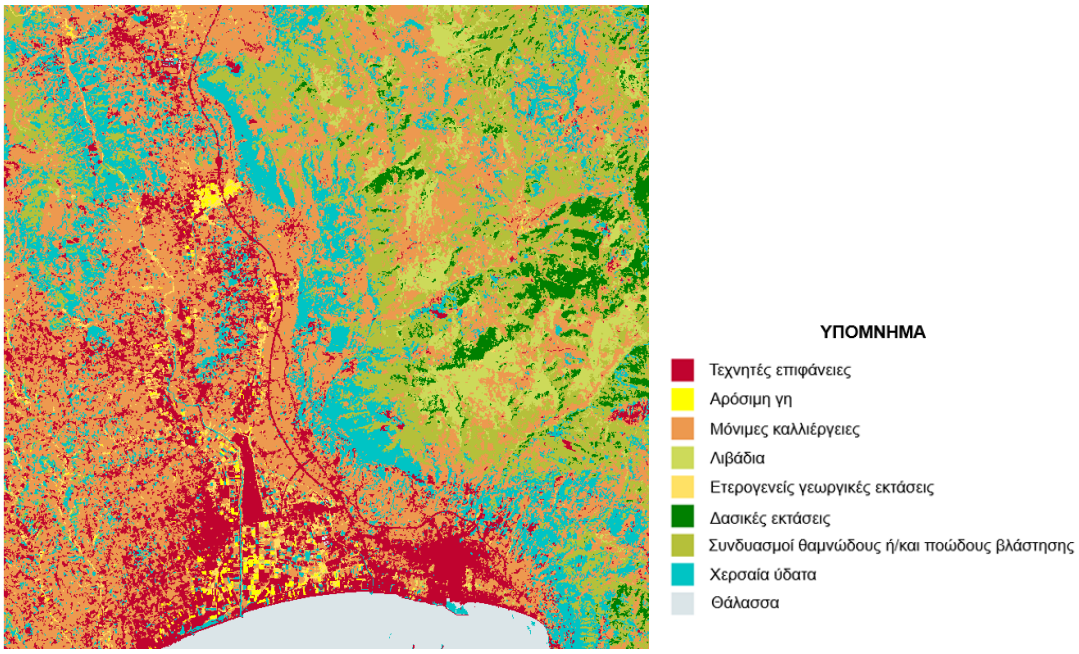
Στην περίπτωση των δασικών εκτάσεων, ο δείκτης αναλογίας ισορροπίας λαμβάνει τιμή ίση με 0.765 (<1), οδηγώντας στην περίπτωση εφαρμογής τυχαίας υπερδειγματοληψίας πολυγώνων εκπαίδευσης, με σκοπό η κατηγορία να εμφανίζει την ιδανική αναλογία. Λόγω του γεγονότος ότι τα δάση καλύπτουν σημαντική έκταση της περιοχής ενδιαφέροντος, είναι δυνατή η εφαρμογή υπερδειγματοληψίας, με την προσθήκη δεδομένων εκπαίδευσης για την κατηγορία των δασών. Έπειτα από δοκιμές, το σύνολο δεδομένων εκπαίδευσης που θα χρησιμοποιηθεί για την υλοποίηση του πειράματος διαμορφώνεται ως εξής :

Κατηγορίες CLC	Πλήθος πολυγώνων	Συχνότητα	BRI
Τεχνητές επιφάνειες	63	0,137	1,233
Αρόσιμη γη	57	0,124	1,115
Μόνιμες καλλιέργειες	50	0,109	0,978
Λιβάδια	47	0,102	0,920
Ετερογενείς γεωργικές εκτάσεις	57	0,124	1,115
Δασικές εκτάσεις	51	0,111	1,000
Συνδυασμοί βλάστησης	56	0,122	1,096
Χερσαία ύδατα	45	0,098	0,880
Θάλασσα	34	0,074	0,665

Πίνακας 28. Μορφή συνόλου δεδομένων εκπαίδευσης για την ισορροπία των δασικών εκτάσεων

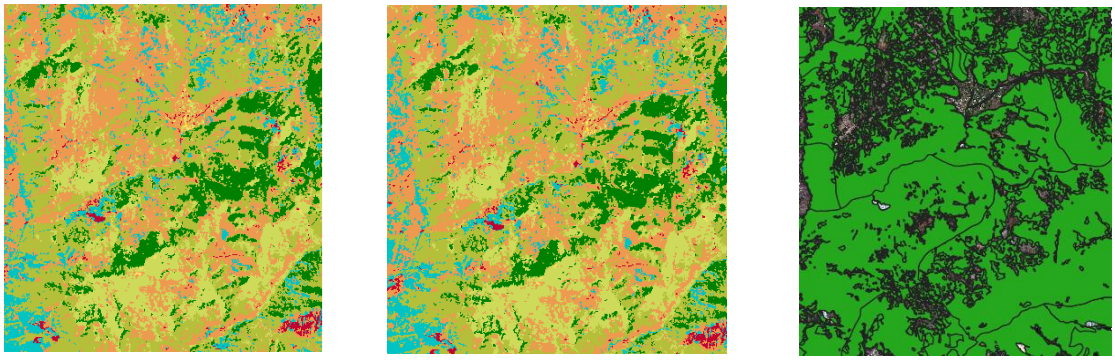
7.3.6.1.Ποιοτική Αξιολόγηση

Η εφαρμογή τυχαίας υπερδειγματοληψίας στα πολύγωνα εκπαίδευσης των δασικών εκτάσεων έχει ως αποτέλεσμα η επιβλεπόμενη ταξινόμηση της πολυφασματικής απεικόνισης να δημιουργεί τον παρακάτω θεματικό χάρτη.



Σχήμα 90. Αποτέλεσμα ταξινόμησης Random Forest, με 100 δέντρα, πολυφασματικής απεικόνισης Sentinel-2 ημερομηνίας λήψης 17/07/2023, με τη μέθοδο τυχαίας υπερδειγματοληψίας για τις δασικές εκτάσεις

Παρά το γεγονός πως πραγματοποιείται υπερδειγματοληψία των πολυγώνων εκπαίδευσης των δασικών εκτάσεων, στο αποτέλεσμα της ταξινόμησης δεν εμφανίζεται σημαντική βελτίωση στην απόδοση της κατηγορίας. Πιο συγκεκριμένα, συγκρίνοντας την απεικόνιση του αποτελέσματος της ταξινόμησης με το ισορροπημένο σύνολο δεδομένων ,Σχήμα 76 με την ταξινόμηση με υπερδειγματοληψία των δασικών εκτάσεων είναι μόνο μερικά επιπλέον εικονοστοιχεία που έχουν αναγνωριστεί ορθά ως δασικές εκτάσεις, αντικαθιστώντας την αρχική, λανθασμένη ταξινόμησή τους ως λιβάδια και συνδυασμούς βλάστησης.



Σχήμα 91. Ταξινόμηση με ισορροπημένο dataset εκπαίδευσης (αριστερά) - Ταξινόμηση με τυχαία υπερδειγματοληψία (μέση) – Κυρωμένος Δασικός Χάρτης (δεξιά)

7.3.6.2.Ποσοτική Αξιολόγηση

Ο βασικός οδηγός ποσοτικής αξιολόγησης του αποτελέσματος της ταξινόμησης είναι ο πίνακας σύγχυσης, που για το συγκεκριμένο πείραμα έχει σχηματιστεί ως εξής :

V_Classified	Τεχνητές	Αρόσιμη	Μόνιμες	Λιβάδια	Ετερογενείς	Δάση	Συνδυασμοί	Χερσαία	Θάλασσα	Total
Τεχνητές	**719**	0	0	0	0	0	0	0	0	**719**
Αρόσιμη	0	**442**	0	0	118	0	0	0	0	**560**
Μόνιμες	0	12	**661**	67	3	0	11	0	0	**754**
Λιβάδια	0	7	19	**266**	18	73	246	0	0	**629**
Ετερογενείς	0	235	1	0	**559**	0	0	0	0	**795**
Δάση	0	0	0	38	0	**558**	48	0	0	**644**
Συνδυασμοί	0	1	3	319	0	73	**379**	0	0	**775**
Χερσαία	0	6	7	0	0	0	6	**699**	0	**718**
Θάλασσα	0	0	0	0	0	0	0	0	**692**	**692**
Total	**719**	**703**	**691**	**690**	**698**	**704**	**690**	**699**	**692**	**6286**

Πίνακας 29. Πίνακας σύγχυσης ταξινόμησης πολυφασματικής απεικόνισης με ημερομηνία λήψης 17/07/2023 με χρήση του αλγορίθμου Random Forest, με υπερδειγματοληψία για τις δασικές εκτάσεις

Απόρροια της υπερδειγματοληψίας των δασικών εκτάσεων είναι το σφάλμα παράλειψης των εικονοστοιχείων της αρόσιμης γης που ταξινομήθηκαν ως ετερογενείς γεωργικές εκτάσεις και παράλληλα το σφάλμα συμπερίληψης εικονοστοιχείων των μόνιμων καλλιεργειών, των λιβαδιών, το σημαντικότερο των ετερογενών, των συνδυασμών βλάστησης και το χειρότερο από άποψη συνάφειας, των χερσαίων υδάτων. Ένα ακόμη σφάλμα που καταγράφεται είναι αυτό της ταξινόμησης εικονοστοιχείων των λιβαδιών ως συνδυασμοί βλάστησης και το αντίστροφο, το οποίο όμως συμβαίνει σε όλα τα πειράματα. Υψηλό είναι το σφάλμα συμπερίληψης 118 εικονοστοιχείων της αρόσιμης γης ως ετερογενείς γεωργικές εκτάσεις. Συνοψίζοντας, η αξιολόγηση των πειραμάτων μέσω του πίνακα σύγχυσης δε σημειώνει σημαντικές διαφοροποιήσεις από πείραμα σε πείραμα, με μικρές αυξομειώσεις στα πλήθη των εικονοστοιχείων.

Πιο συγκεκριμένα, εστιάζοντας στην απόδοση κάθε κατηγορίας, υπολογίζονται οι μετρικές precision, recall, F1-score, συνολική ακρίβεια και δείκτης συμφωνίας k-hat.

	Precision	Recall	F1-score	Συνολική ακρίβεια (%)	K-hat (%)
Τεχνητές	1	1	1	79,14	76,54
Αρόσιμη	0,7893	0,6287	0,6999		
Μόνιμες	0,8767	0,9566	0,9149		
Λιβάδια	0,4229	0,3855	0,4033		
Ετερογενείς	0,7031	0,8009	0,7488		
Δάση	0,8665	0,7926	0,8279		
Συνδυασμοί	0,4890	0,5493	0,5174		
Χερσαία	0,9735	1	0,9866		
Θάλασσα	1	1	1		

Πίνακας 30. Μετρικές αξιολόγησης της ταξινόμησης της πολυφασματικής απεικόνισης με τον αλγόριθμο Random Forest, με υπερδειγματοληψία για τις δασικές εκτάσεις

Τα εικονοστοιχεία που ταξινομούνται σε λάθος κατηγορίες, όπως προκύπτει από τα πολύγωνα ελέγχου, επιδρούν στην απόδοση της ακρίβειας της κατηγορίας στην οποία αναφέρονται. Είναι πλέον αντιληπτό ότι στις περιπτώσεις κατηγοριών στις οποίες καταγράφεται σημαντικό πλήθος εικονοστοιχείων που έχουν παραληφθεί από την κατηγορία που εξετάζεται ή έχουν συμπεριληφθεί σε αυτή, χωρίς να ανήκουν σε αυτή, τότε επηρεάζεται σε ανάλογο μέγεθος το precision και το recall της, αντίστοιχα, φαινόμενο που συνοψίζεται με τον αρμονικό μέσο F1-score. Λόγω της υψηλής σύγχυσης μεταξύ των λιβαδιών και των συνδυασμών βλάστησης η απόδοσή τους θεωρείται μέτρια, με τιμές ≤ 0,5, με τις υπόλοιπες κατηγορίες να εμφανίζουν ακρίβειες πάνω από 0,70.

Η συνολική ακρίβεια της ταξινόμησης αγγίζει το 79,14%, όμως είναι ένα μέγεθος που δεν μπορεί να ληφθεί υπόψιν στην αξιολόγηση του αποτελέσματος, καθώς επιδρά σημαντικά σε αυτό η αυξημένη ακρίβεια των τεχνητών επιφανειών, των χερσαίων υδάτων και της θάλασσας. Τέλος, ο δείκτης συμφωνίας k-hat σταθερά έχει μια απόκλιση της τάξης του 3% από τη συνολική ακρίβεια της ταξινόμησης.

7.3.7.Συνδυασμοί θαμνώδους ή/και ποώδους βλάστησης

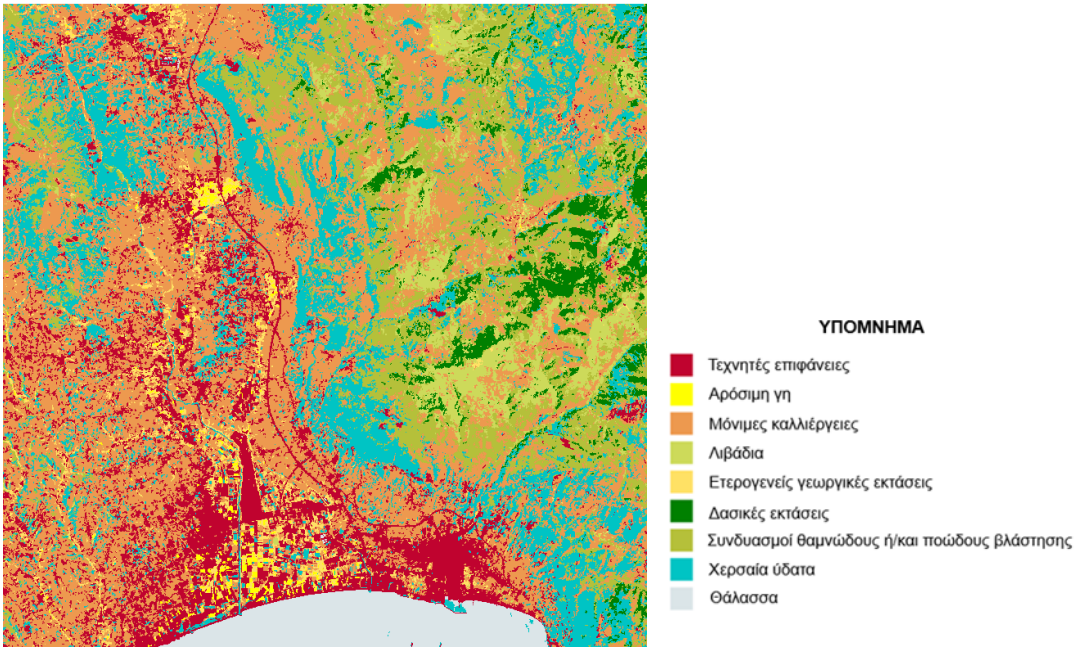
Για την κατηγορία των συνδυασμών βλάστησης ορίζεται η μέθοδος της υποδειγματοληψίας, BRI = 1.128>1, για την αντιμετώπιση της ανισορροπίας του πλήθους των πολυγώνων εκπαίδευσης. Το dataset λειτουργεί με αυτή τη μέθοδο, ορίζοντας εν τέλει 49 πολύγωνα εκπαίδευσης για τη συγκεκριμένη κατηγορία (αρχικά 57), σχηματίζοντας το παρακάτω σύνολο δεδομένων εκπαίδευσης

Κατηγορίες CLC	Πλήθος πολυγώνων	Συχνότητα	BRI
Τεχνητές επιφάνειες	63	0,143	1,289
Αρόσιμη γη	57	0,130	1,166
Μόνιμες καλλιέργειες	50	0,114	1,023
Λιβάδια	47	0,107	0,961
Ετερογενείς γεωργικές εκτάσεις	57	0,130	1,166
Δασικές εκτάσεις	38	0,086	0,777
Συνδυασμοί βλάστησης	49	0,111	1,002
Χερσαία ύδατα	45	0,102	0,920
Θάλασσα	34	0,077	0,695

Πίνακας 31. Μορφή συνόλου δεδομένων εκπαίδευσης για την ισορροπία των συνδυασμών βλάστησης

7.3.7.1.Ποιοτική Αξιολόγηση

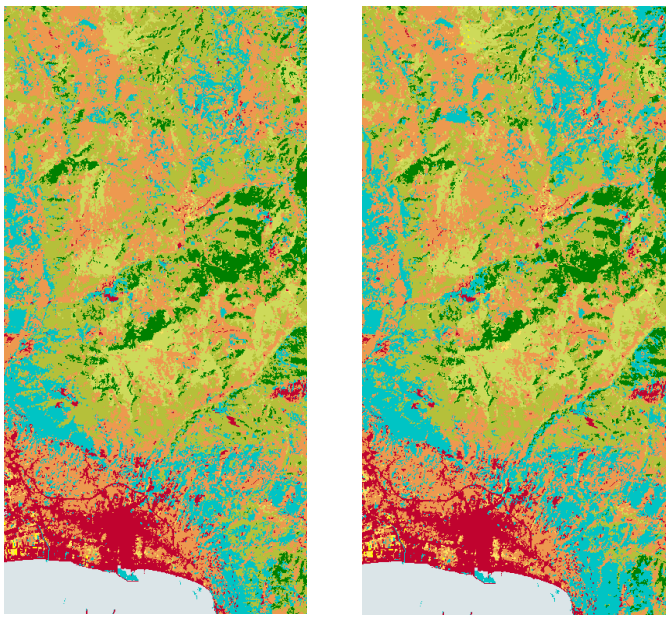
Η υποδειγματοληψία του συνδυασμού βλάστησης έχει ως αποτέλεσμα την παρακάτω απεικόνιση.



Σχήμα 92. Αποτέλεσμα ταξινόμησης Random Forest, με 100 δέντρα, πολυφασματικής απεικόνισης Sentinel-2 ημερομηνίας λήψης 17/07/2023, με τη μέθοδο τυχαίας υποδειγματοληψίας για τους συνδυασμούς βλάστησης

Στην περίπτωση ταξινόμησης με υποδειγματοληψία στους συνδυασμούς βλάστησης οξύνεται ένα από τα κυριότερα σφάλματα των πειραμάτων, η ταξινόμηση εκτάσεων βλάστησης ως χερσαία ύδατα. Το μεγαλύτερο ποσοστό εικονοστοιχείων που αρχικά είχαν ταξινομηθεί ως χερσαία ύδατα και στο παρόν πείραμα πολλαπλασιάζονται,

παρατηρούνται στο ανατολικό τμήμα της περιοχής μελέτης κι ένα δεύτερο μέρος στο βορειοδυτικό τομέα.



Σχήμα 93. Ταξινόμηση με ισορροπημένο dataset εκπαίδευσης (αριστερά) - Ταξινόμηση με τυχαία υποδειγματοληψία (δεξιά)

Γίνεται αντιληπτό πως η επέκταση του σφάλματος ταξινόμησης κατηγοριών βλάστησης ως χερσαία ύδατα θα επηρεάζει αρνητικά την απόδοση/ακρίβεια του μοντέλου, γεγονός που γίνεται αντιληπτό μέσα από τον πίνακα σύγχυσης και τις μετρικές που υπολογίζονται μέσω αυτού, όπως παρουσιάζονται στην ακόλουθη ενότητα.

7.3.7.2.Ποσοτική Αξιολόγηση

Ο πιο ενδεικτικός τρόπος για την αξιολόγηση του αποτελέσματος μιας ταξινόμησης είναι οι μετρικές ακρίβειας μέσα από τον πίνακα σύγχυσης. Ο πίνακας σύγχυσης του πειράματος επιβλεπόμενης ταξινόμησης με την εφαρμογή μεθόδου υποδειγματοληψίας στην κατηγορίας των συνδυασμών βλάστησης παρουσιάζεται παρακάτω.

V_Classified	Τεχνητές	Αρόσιμη	Μόνιμες	Λιβάδια	Ετερογενείς	Δάση	Συνδυασμοί	Χερσαία	Θάλασσα	Total
Τεχνητές	**719**	0	0	0	0	0	0	0	0	**719**
Αρόσιμη	0	**437**	0	0	128	0	0	0	0	**565**
Μόνιμες	0	14	**661**	54	2	0	11	0	0	**742**
Λιβάδια	0	0	14	**274**	31	76	245	0	0	**640**
Ετερογενείς	0	246	0	0	**537**	1	0	0	0	**784**
Δάση	0	0	0	36	0	**553**	49	0	0	**638**
Συνδυασμοί	0	0	5	326	0	72	**379**	0	0	**782**
Χερσαία	0	6	11	0	0	2	6	**699**	0	**724**
Θάλασσα	0	0	0	0	0	0	0	0	**692**	**692**
Total	**719**	**703**	**691**	**690**	**698**	**704**	**690**	**699**	**692**	**6286**

Πίνακας 32. Πίνακας σύγχυσης ταξινόμησης πολυφασματικής απεικόνισης με ημερομηνία λήψης 17/07/2023 με χρήση του αλγορίθμου Random Forest, με υποδειγματοληψία για τους συνδυασμούς βλάστησης

Απόρροια της υποδειγματοληψίας των συνδυασμών βλάστησης είναι το σφάλμα παράλειψης των εικονοστοιχείων της αρόσιμης γης που ταξινομήθηκαν ως ετερογενείς γεωργικές εκτάσεις και παράλληλα το σφάλμα συμπερίληψης εικονοστοιχείων των μόνιμων καλλιεργειών, το σημαντικότερο των ετερογενών και το χειρότερο από άποψη

συνάφειας, των χερσαίων υδάτων. Βέβαια, συγκριτικά με τα προηγούμενα πειράματα είναι το μόνο που καταγράφεται ελαττωμένη σύγχυση μεταξύ καλλιεργειών βλάστησης. Ένα ακόμη σφάλμα που καταγράφεται είναι αυτό της ταξινόμησης εικονοστοιχείων των λιβαδιών ως συνδυασμοί βλάστησης και το αντίστροφο, το οποίο όμως συμβαίνει σε όλα τα πειράματα. Υψηλό είναι το σφάλμα συμπερίληψης 128 εικονοστοιχείων της αρόσιμης γης ως ετερογενείς γεωργικές εκτάσεις. Εστιάζοντας στους συνδυασμούς βλάστησης που αποτελούν κατηγορία ενδιαφέροντος για το παρόν πείραμα, συνεχίζει η αυξημένη απόδοση εικονοστοιχείων της κατηγορίας στην αρόσιμη γη (326 εικονοστοιχεία), αλλά και μεγάλο μέρος εικονοστοιχείων της αρόσιμης ως συνδυασμοί (245 εικονοστοιχεία). Συνοψίζοντας, η αξιολόγηση των πειραμάτων μέσω του πίνακα σύγχυσης δε σημειώνει σημαντικές διαφοροποιήσεις από πείραμα σε πείραμα, με μικρές αυξομειώσεις στα πλήθη των εικονοστοιχείων.

	Precision	**Recall**	**F1-score**	**Συνολική ακρίβεια (%)**	**K-hat (%)**
Τεχνητές	1	1	1	78,76	76,11
Αρόσιμη	0,7735	0,6216	0,6893		
Μόνιμες	0,8908	0,9566	0,9225		
Λιβάδια	0,4281	0,3971	0,4120		
Ετερογενείς	0,6849	0,7693	0,7247		
Δάση	0,8668	0,7855	0,8241		
Συνδυασμοί	0,4847	0,5493	0,5149		
Χερσαία	0,9655	1	0,9824		
Θάλασσα	1	1	1		

Πίνακας 33. Μετρικές αξιολόγησης της ταξινόμησης της πολυφασματικής απεικόνισης με τον αλγόριθμο Random Forest, με υποδειγματοληψία για τους συνδυασμούς βλάστησης

Είναι πλέον αντιληπτό ότι στις περιπτώσεις κατηγοριών στις οποίες καταγράφεται σημαντικό πλήθος εικονοστοιχείων που έχουν παραληφθεί από την κατηγορία που εξετάζεται ή έχουν συμπεριληφθεί σε αυτή, χωρίς να ανήκουν σε αυτή, τότε επηρεάζεται σε ανάλογο μέγεθος το precision και το recall της, αντίστοιχα, φαινόμενο που συνοψίζεται με τον αρμονικό μέσο F1-score. Λόγω της υψηλής σύγχυσης μεταξύ των λιβαδιών και των συνδυασμών βλάστησης η απόδοσή τους θεωρείται μέτρια, με τιμές ≤ 0,5, με τις υπόλοιπες κατηγορίες να εμφανίζουν ακρίβειες πάνω από 0,69.

Η συνολική ακρίβεια της ταξινόμησης αγγίζει το 78,76%, η χαμηλότερη μέχρι στιγμής συνολική ακρίβεια πειράματος, όμως είναι ένα μέγεθος που δεν μπορεί να ληφθεί υπόψιν στην αξιολόγηση του αποτελέσματος, καθώς επιδρά σημαντικά σε αυτό η αυξημένη ακρίβεια των τεχνητών επιφανειών, των χερσαίων υδάτων και της θάλασσας. Τέλος, ο δείκτης συμφωνίας k-hat σταθερά έχει μια απόκλιση της τάξης του 3% από τη συνολική ακρίβεια της ταξινόμησης.

7.3.8.Χερσαία ύδατα

Ιδιαίτερη χαρακτηρίζεται η περίπτωση των χερσαίων υδάτων, που είναι η κατηγορία με βάση την οποία έχει διαμορφωθεί το αρχικό, ισορροπημένο σύνολο δεδομένων εκπαίδευσης, έχοντας με αυτόν τον τρόπο καθοριστικό ρόλο στη διαμόρφωση όλων των πειραμάτων. Στην Ενότητα 7.2. όπου προσδιορίζεται ο Δείκτης Αναλογίας Ισορροπίας (BRI) , τα χερσαία ύδατα λαμβάνουν τιμή δείκτη ίση με 0.906, γεγονός που υποδεικνύει την ανάγκη υπερδειγματοληψίας, ώστε η κατηγορία να σημειώσει την ιδανική αναλογία. Ωστόσο, το πλήθος των πολυγώνων εκπαίδευσης που περιγράφουν την κατηγορία στο αρχικό dataset είναι οριακό λόγω της έκτασης που καλύπτει, με συνέπεια να μην είναι εφικτή η εφαρμογή της μεθόδου υπερδειγματοληψίας. Αποφαίνεται λοιπόν, η εφαρμογή της μεθόδου υποδειγματοληψίας όλων των υπόλοιπων κατηγοριών, για να λάβουν την ιδανική τιμή τα χερσαία ύδατα. Η επιβλεπόμενη ταξινόμηση στην πολυφασματική απεικόνιση εφαρμόζεται με χρήση του παρακάτω dataset

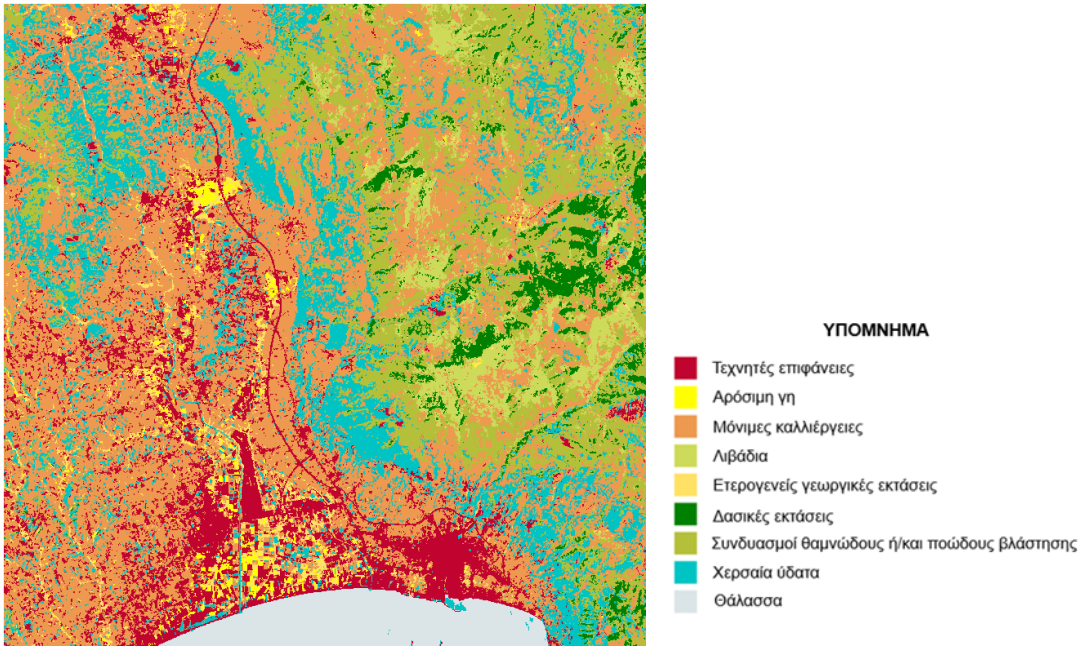
Κατηγορίες CLC	Πλήθος πολυγώνων	Συχνότητα	BRI
Τεχνητές επιφάνειες	50	0,123	1,111
Αρόσιμη γη	50	0,123	1,111
Μόνιμες καλλιέργειες	45	0,111	1,000
Λιβάδια	45	0,111	1,000
Ετερογενείς γεωργικές εκτάσεις	55	0,136	1,222
Δασικές εκτάσεις	35	0,086	0,778
Συνδυασμοί βλάστησης	50	0,123	1,111
Χερσαία ύδατα	45	0,111	1,000
Θάλασσα	30	0,074	0,667

Πίνακας 34. Μορφή συνόλου δεδομένων εκπαίδευσης για την ισορροπία των συνδυασμών βλάστησης

Ενδιαφέρον προκαλεί αν η ταξινόμηση με αυτό το σύνολο δεδομένων εκπαίδευσης θα καταφέρει να αμβλύνει ή να οξύνει το σφάλμα ταξινόμησης εκτάσεων βλάστησης ως χερσαία ύδατα. Αυτό μελετάται στην επόμενη ενότητα αξιολόγησης της ταξινόμησης.

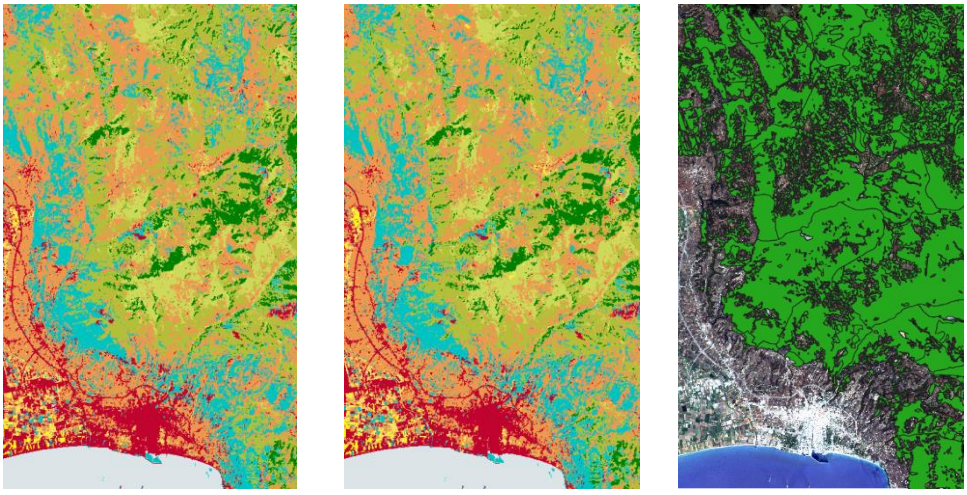
7.3.8.1.Ποιοτική Αξιολόγηση

Η επιβλεπόμενη ταξινόμηση της πολυφασματικής απεικόνισης με σύνολο εκπαίδευσης στο οποίο έχει εφαρμοστεί υποδειγματοληψία στις υπόλοιπες κατηγορίες πλην των χερσαίων έχει ως αποτέλεσμα

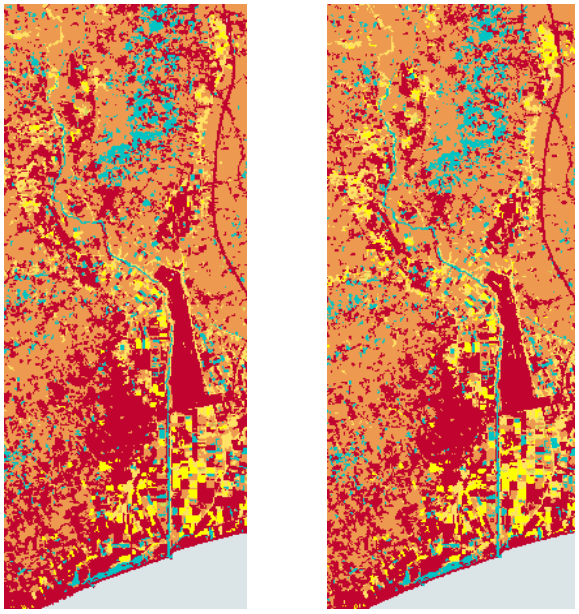


Σχήμα 94. Αποτέλεσμα ταξινόμησης Random Forest, με 100 δέντρα, πολυφασματικής απεικόνισης Sentinel-2 ημερομηνίας λήψης 17/07/2023, με τη μέθοδο τυχαίας υποδειγματοληψίας όλων των κατηγοριών ,πλην των χερσαίων υδάτων

Το κυριότερο που επιτυγχάνεται με την υποδειγματοληψία των πολυγώνων των κατηγοριών είναι η ενίσχυση των δασικών εκτάσεων σε όλο το ανατολικό τμήμα της περιοχής ενδιαφέροντος, με όλες αυτές να επιβεβαιώνονται ως δάση από τον κυρωμένο δασικό χάρτη. Παράλληλα, η κατηγορία που επηρεάζεται από την υποδειγματοληψία είναι οι τεχνητές επιφάνειες, περιορίζοντας το πλήθος τους σε όλη την έκταση της περιοχής ,με απόρροια την καλύτερη απόδοση της αρόσιμης γης και των μόνιμων καλλιεργειών, καθώς είναι οι δυο κατηγορίες στις οποίες παρεμβάλλονται αρκετά εικονοστοιχεία των τεχνητών επιφανειών.



Σχήμα 95. Ταξινόμηση με ισορροπημένο dataset εκπαίδευσης (αριστερά) - Ταξινόμηση με τυχαία υποδειγματοληψία (μέση) – Κυρωμένος Δασικός Χάρτης (δεξιά)



Σχήμα 96. Ταξινόμηση με ισορροπημένο dataset εκπαίδευσης (αριστερά) - Ταξινόμηση με τυχαία υποδειγματοληψία (δεξιά)

7.3.8.2.Ποσοτική Αξιολόγηση

Ο πίνακας σύγχυσης που προκύπτει για το πείραμα υποδειγματοληψίας όλων των κατηγοριών εκτός των χερσαίων υδάτων σχηματίζεται ως εξής

V_Classified	Τεχνητές	Αρόσιμη	Μόνιμες	Λιβάδια	Ετερογενείς	Δάση	Συνδυασμοί	Χερσαία	Θάλασσα	Total
Τεχνητές	**719**	0	0	0	0	0	0	0	0	**719**
Αρόσιμη	0	**419**	0	0	108	0	0	0	0	**527**
Μόνιμες	0	12	**661**	67	4	0	11	0	0	**755**
Λιβάδια	0	3	14	**267**	20	76	243	0	0	**623**
Ετερογενείς	0	261	0	0	**566**	0	0	0	0	**827**
Δάση	0	0	0	23	0	**542**	33	0	0	**598**
Συνδυασμοί	0	2	4	333	0	86	**397**	0	0	**822**
Χερσαία	0	6	12	0	0	0	6	**699**	0	**723**
Θάλασσα	0	0	0	0	0	0	0	0	**692**	**692**
Total	**719**	**703**	**691**	**690**	**698**	**704**	**690**	**699**	**692**	**6286**

Πίνακας 35. Πίνακας σύγχυσης ταξινόμησης πολυφασματικής απεικόνισης με ημερομηνία λήψης 17/07/2023 με χρήση του αλγορίθμου Random Forest, με υποδειγματοληψία όλων των κατηγοριών, πλην των χερσαίων υδάτων

Σταθερά αποδίδουν οι τεχνητές επιφάνειες και η θάλασσα, με όλα τα εικονοστοιχεία των δεδομένων ελέγχου να αναγνωρίζουν μόνο ορθά ταξινομημένα εικονοστοιχεία στο θεματικό χάρτη που παράγεται από την επιβλεπόμενη ταξινόμηση της πολυφασματικής απεικόνισης. Οι κατηγορίες βλάστησης δε, συνεχίζουν να καταγράφουν σφάλματα συμπερίληψης και παράλειψης, με εικονοστοιχεία της αρόσιμης γης να ταξινομούνται ως ετερογενείς καλλιέργειες, με ταυτόχρονη ταξινόμηση εικονοστοιχείων των ετερογενών ως αρόσιμη γη, φαινόμενο που οδηγεί σε αντίστοιχες τιμές στο precision και το recall των δυο αυτώ κατηγοριών, όπως προκύπτει κι από τον παρακάτω πίνακα. Το μεγαλύτερο σφάλμα παράλειψης σημειώνεται για την κατηγορία των συνδυασμών βλάστησης, με τα ορθά ταξινομημένα εικονοστοιχεία να αγγίζουν τα 397 ,με 333 εικονοστοιχεία της κατηγορίας να έχουν ταξινομηθεί ως λιβάδια. Αντίστοιχα, παρόμοιας διακύμανσης σφάλμα παράλειψης εμφανίζεται στην περίπτωση των λιβαδιών με 267 ορθά ταξινομημένα εικονοστοιχεία και 243 εικονοστοιχεία λιβαδιών να εντάσσονται στους συνδυασμούς βλάστησης. Όλα

τα συμπεράσματα που απορρέουν από τον πίνακα σύγχυσης, καταγράφονται και ποσοτικοποιούνται στον Πίνακα 36, με τη μετρική F1-score να περιγράφει ολοκληρωμένα πως έχουν επηρεάσει τα σφάλματα παράλειψης και συμπερίληψης τη συνολική απόδοση κάθε κατηγορίας χρήσεων γης.

	Precision	Recall	F1-score	Συνολική ακρίβεια (%)	K-hat (%)
Τεχνητές	1	1	1	78,94	76,31
Αρόσιμη	0,7951	0,5960	0,6813		
Μόνιμες	0,8755	0,9566	0,9142		
Λιβάδια	0,4286	0,3870	0,4067		
Ετερογενείς	0,6844	0,8109	0,7423		
Δάση	0,9064	0,7699	0,8326		
Συνδυασμοί	0,4830	0,5754	0,5251		
Χερσαία	0,9668	1	0,9831		
Θάλασσα	1	1	1		

Πίνακας 36. Μετρικές αξιολόγησης της ταξινόμησης της πολυφασματικής απεικόνισης με τον αλγόριθμο Random Forest, με υποδειγματοληψία όλων των κατηγοριών, πλην των χερσαίων υδάτων

Τα εικονοστοιχεία που ταξινομούνται σε λάθος κατηγορίες, όπως προκύπτει από τα πολύγωνα ελέγχου, επιδρούν στην απόδοση της ακρίβειας της κατηγορίας στην οποία αναφέρονται. Είναι πλέον αντιληπτό ότι στις περιπτώσεις κατηγοριών στις οποίες καταγράφεται σημαντικό πλήθος εικονοστοιχείων που έχουν παραληφθεί από την κατηγορία που εξετάζεται ή έχουν συμπεριληφθεί σε αυτή, χωρίς να ανήκουν σε αυτή, τότε επηρεάζεται σε ανάλογο μέγεθος το precision και το recall της, αντίστοιχα, φαινόμενο που συνοψίζεται με τον αρμονικό μέσο F1-score. Λόγω της υψηλής σύγχυσης μεταξύ των λιβαδιών και των συνδυασμών βλάστησης η απόδοσή τους θεωρείται μέτρια, με τιμές ≤ 0,52, με τις υπόλοιπες κατηγορίες να εμφανίζουν ακρίβειες πάνω από 0,68.

Η συνολική ακρίβεια της ταξινόμησης αγγίζει το 78,94%, όμως είναι ένα μέγεθος που δεν μπορεί να ληφθεί υπόψιν στην αξιολόγηση του αποτελέσματος, καθώς επιδρά σημαντικά σε αυτό η αυξημένη ακρίβεια των τεχνητών επιφανειών, των χερσαίων υδάτων και της θάλασσας. Τέλος, ο δείκτης συμφωνίας k-hat έχει μια απόκλιση από τη συνολική ακρίβεια της ταξινόμησης κι είναι ίσος με 76,31%.

7.3.9.Θάλασσα

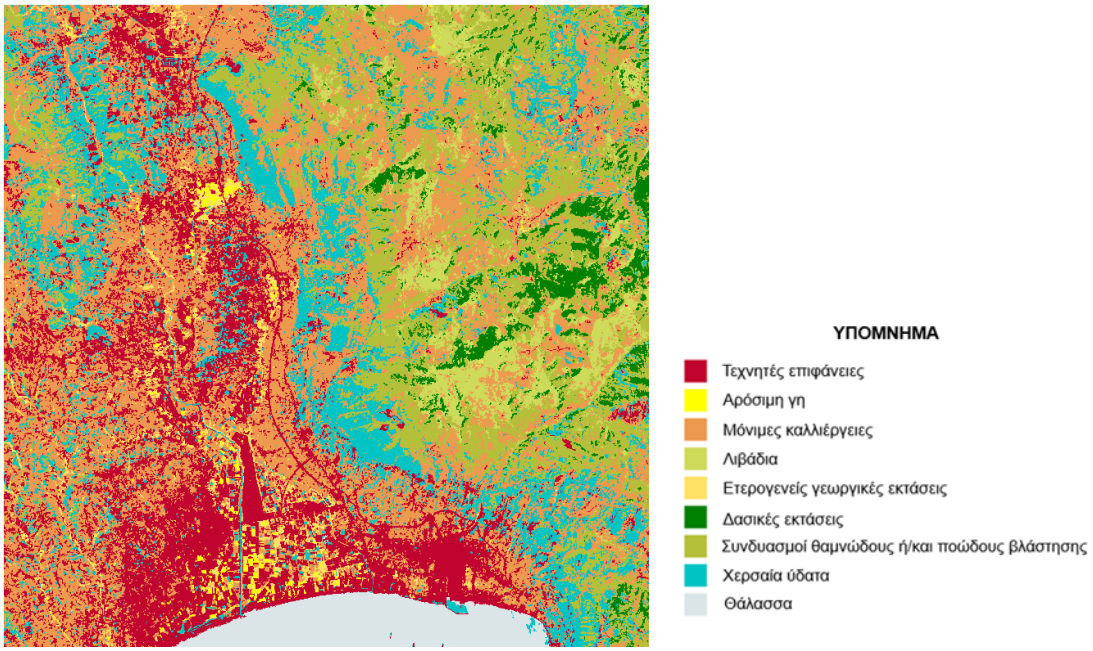
Η κατηγορία της θάλασσας, ως η κατηγορία με τα λιγότερα πολύγωνα εκπαίδευσης είναι αναμενόμενο πως χρίζει υπερδειγματοληψίας, γεγονός που είναι εφικτό λόγω της έκτασης που καταλαμβάνει στην περιοχή μελέτης. Οι δοκιμές αύξησης του πλήθους των πολυγώνων της σχηματίζουν το εξής σύνολο δεδομένων :

Κατηγορίες CLC	Πλήθος πολυγώνων	Συχνότητα	BRI
Τεχνητές επιφάνειες	64	0,137	1,233
Αρόσιμη γη	58	0,124	1,118
Μόνιμες καλλιέργειες	50	0,107	0,964
Λιβάδια	47	0,101	0,906
Ετερογενείς γεωργικές εκτάσεις	57	0,122	1,099
Δασικές εκτάσεις	38	0,081	0,732
Συνδυασμοί βλάστησης	56	0,120	1,079
Χερσαία ύδατα	45	0,096	0,867
Θάλασσα	52	0,111	1,000

Πίνακας 37. Μορφή συνόλου δεδομένων εκπαίδευσης για την ισορροπία της θάλασσας

7.3.9.1.Ποιοτική Αξιολόγηση

Η εφαρμογή υπερδειγματοληψίας στη θάλασσα έχει ως αποτέλεσμα τη δημιουργία της παρακάτω απεικόνισης.

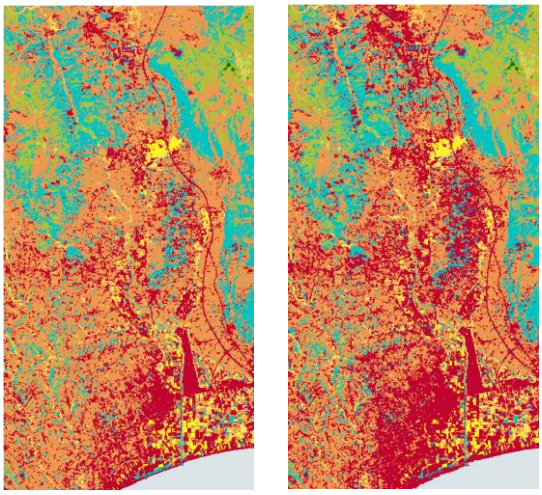


Σχήμα 97. Αποτέλεσμα ταξινόμησης Random Forest, με 100 δέντρα, πολυφασματικής απεικόνισης Sentinel-2 ημερομηνίας λήψης 17/07/2023, με τη μέθοδο τυχαίας υπερδειγματοληψίας της θάλασσας

Το πείραμα υπερδειγματοληψίας της θάλασσας έχει ως αποτέλεσμα την ενίσχυση των τεχνητών επιφανειών, καθώς σε όλο το δυτικό τομέα της περιοχής ενδιαφέροντος μεγάλο μέρος εικονοστοιχείων κατηγοριών βλάστησης σημειώνονται ως τεχνητές επιφάνειες. Βέβαια, όπως έχει αναφερθεί και σε άλλο πείραμα, ένα τέτοιο σφάλμα ενδέχεται να οφείλεται σε εκτάσεις γυμνού εδάφους, όπως είναι οι περιπτώσεις καλλιεργήσιμων εκτάσεων σε περίοδο αγρανάπαυσης, σε αγροτεμάχια των οποίων το έδαφος έχει υποστεί κατεργασία ή ακόμα και βραχώδεις/πετρώδεις επιφάνειες.

Μια ακόμη παρατήρηση που είναι εμφανής σε όλα τα πειράματα, αλλά δεν έχει σχολιαστεί μέχρι στιγμής, είναι η εμφάνιση εικονοστοιχείων θάλασσας ως χερσαία, φαινόμενο που οφείλεται στο βάθος αυτών των σημείων ,το οποίο είναι μικρό κι ιδιαίτερα στην περίπτωση των υδάτων πλησίον του λιμανιού της Καλαμάτας, όπου χαρακτηριστικό ρόλο διαδραματίζει και η παρουσία των τεχνητών επιφανειών να απαρτίζουν το λιμάνι. Αυτός είναι και ο βασικός λόγος για τον οποίο δεν έχει δοθεί έκταση σε αυτό το σφάλμα.

Σχήμα 98. Ενδεικτικό απόσπασμα αποτύπωσης του σφάλματος ταξινόμησης θαλάσσιων υδάτων ως χερσαία



Σχήμα 99. Ταξινόμηση με ισορροπημένο dataset εκπαίδευσης (αριστερά) - Ταξινόμηση με τυχαία υπερδειγματοληψία (δεξιά)

7.3.9.2.Ποσοτική Αξιολόγηση

Ο πίνακας σύγχυσης αυτού του πειράματος έχει τη μορφή :

V_Classified	Τεχνητές	Αρόσιμη	Μόνιμες	Λιβάδια	Ετερογενείς	Δάση	Συνδυασμοί	Χερσαία	Θάλασσα	Total
Τεχνητές	**719**	0	0	0	0	0	0	0	0	**719**
Αρόσιμη	0	**434**	0	0	119	0	0	0	0	**553**
Μόνιμες	0	12	**660**	61	2	0	12	0	0	**747**
Λιβάδια	0	0	16	**269**	24	75	242	0	0	**626**
Ετερογενείς	0	249	0	0	**553**	0	0	0	0	**802**
Δάση	0	0	0	37	0	**550**	47	0	0	**634**
Συνδυασμοί	0	2	4	323	0	79	**383**	0	0	**791**
Χερσαία	0	6	11	0	0	0	6	**699**	0	**722**
Θάλασσα	0	0	0	0	0	0	0	0	**692**	**692**
Total	**719**	**703**	**691**	**690**	**698**	**704**	**690**	**699**	**692**	**6286**

Πίνακας 38. Πίνακας σύγχυσης ταξινόμησης πολυφασματικής απεικόνισης με ημερομηνία λήψης 17/07/2023 με χρήση του αλγορίθμου Random Forest, με υπερδειγματοληψία της θάλασσας

Σταθερά αποδίδουν οι τεχνητές επιφάνειες και η θάλασσα, με όλα τα εικονοστοιχεία των δεδομένων ελέγχου να αναγνωρίζουν μόνο ορθά ταξινομημένα εικονοστοιχεία στο θεματικό χάρτη που παράγεται από την επιβλεπόμενη ταξινόμηση της πολυφασματικής απεικόνισης. Οι κατηγορίες βλάστησης δε, συνεχίζουν να καταγράφουν σφάλματα συμπερίληψης και παράλειψης, με εικονοστοιχεία της αρόσιμης γης να ταξινομούνται ως ετερογενείς καλλιέργειες, με ταυτόχρονη ταξινόμηση εικονοστοιχείων των ετερογενών ως αρόσιμη γη, φαινόμενο που οδηγεί σε αντίστοιχες τιμές στο precision και το recall των δυο αυτώ κατηγοριών, όπως προκύπτει κι από τον παρακάτω πίνακα. Όλα τα συμπεράσματα που απορρέουν από τον πίνακα σύγχυσης, καταγράφονται στον παρακάτω πίνακα, με τη μετρική F1-score να περιγράφει ολοκληρωμένα πως έχουν επηρεάσει τα σφάλματα παράλειψης και συμπερίληψης τη συνολική απόδοση κάθε κατηγορίας χρήσεων γης.

	Precision	**Recall**	**F1-score**	**Συνολική ακρίβεια (%)**	**K-hat (%)**
Τεχνητές	1	1	1	78,89	76,25
Αρόσιμη	0,7848	0,6174	0,6911		
Μόνιμες	0,8835	0,9551	0,9179		
Λιβάδια	0,4297	0,3899	0,4088		
Ετερογενείς	0,6895	0,7923	0,7373		
Δάση	0,8675	0,7813	0,8221		
Συνδυασμοί	0,4842	0,5551	0,5172		
Χερσαία	0,9681	1	0,9838		
Θάλασσα	1	1	1		

Πίνακας 39. Μετρικές αξιολόγησης της ταξινόμησης της πολυφασματικής απεικόνισης με τον αλγόριθμο Random Forest, με υπερδειγματοληψία της θάλασσας

Όπως προκύπτει από τις μετρικές αξιολόγησης, η επιβλεπόμενη ταξινόμηση με συνδυασμό μεθόδων τυχαίας υπερδειγματοληψίας για την ισορροπία της θάλασσας, το σύνολο των κατηγοριών κάλυψης γης καταγράφει υψηλή ακρίβεια precision, με τις μόνες κάτω του 0,5 να είναι τα λιβάδια και οι συνδυασμοί βλάστησης, καθώς σημαντική μερίδα εικονοστοιχείων των λιβαδιών έχουν ταξινομηθεί συνδυασμοί βλάστησης και το αντίστροφο, εικονοστοιχεία των συνδυασμών βλάστησης, στην ταξινομημένη απεικόνιση κατατάσσονται στα λιβάδια. Οι ετερογενείς καλλιέργειες επίσης, σημειώνουν λίγο χαμηλότερη ακρίβεια precision συγκριτικά με τις υπόλοιπες, λόγω του γεγονότος ότι 249 εικονοστοιχεία τους, έχουν ταξινομηθεί ως αρόσιμη γη, που κατ' επέκταση επηρεάζει το recall της αρόσιμης.

Κρίνοντας την απόδοση του μοντέλου από το μέγεθος του F1-score, καταγράφονται 8 από τις 9 κατηγορίες με τιμή μεγαλύτερη του 0,5, με τη μονάδα να χαρακτηρίζει την καλύτερη απόδοση που θα μπορούσε να καταγράψει το μοντέλο, με τη μέτρια απόδοση των συνδυασμών βλάστησης κοντά στο 0,5 και τη χαμηλότερη επίδοση των λιβαδιών με τιμή πλησίον του 0,41.

Από την άλλη, η συνολική ακρίβεια της ταξινόμησης ίση με 78,89%, με το συγκεκριμένο πείραμα να σημειώνει την ίδια συνολική ακρίβεια συγκριτικά με το πείραμα υποδειγματοληψίας των τεχνητών επιφανειών, με λίγο μικρότερη ακρίβεια να καταγράφει ο δείκτης συμφωνίας K-hat κοντά στο 76%.

7.4.Προβλήματα κατά την υλοποίηση πειραμάτων

Καθ' όλη την πορεία της Διπλωματικής Εργασίας δεν παρέλειψαν να υπάρχουν προβλήματα, ιδίως στο στάδιο υλοποίησης των πειραμάτων κι όχι τόσο στα στάδια προ-επεξεργασίας της πολυφασματικής απεικόνισης και της δημιουργίας των δεδομένων εκπαίδευσης κι ελέγχου.

Τα προβλήματα που παρουσιάστηκαν αφορούσαν είτε άμεσα είτε έμμεσα την υπολογιστική ισχύ του υπολογιστή, ιδίως σε θέματα μνήμης -σκληρός δίσκος- και δευτερεύοντος σε στοιχεία όπως η μνήμη RAM και η κάρτα γραφικών. Το σημαντικότερο πρόβλημα λοιπόν, εμφανίστηκε όσο το λογισμικό QGIS "έτρεχε" τον αλγόριθμο επιβλεπόμενης ταξινόμησης, όπου για κάθε πείραμα κατανάλωνε περίπου 100 GB από το σκληρό δίσκο. Μάλιστα, υπήρχαν φορές που η απελευθέρωση του χώρου που καταναλώθηκε από τα πειράματα, ήταν χρονοβόρα, χρειάστηκε αρκετές προσπάθειες, δημιουργώντας με αυτόν τον τρόπο καθυστερήσεις στην κύλιση όλων των απαραίτητων εργασιών. Παράλληλα, λόγω του γεγονότος πως ο υπολογιστής που χρησιμοποιείτο για την εκπόνηση της εργασίας είναι διαμορφωμένος για οικιακή, απλή χρήση, άρα πρόκειται για χαμηλότερων ικανοτήτων υπολογιστή, υπήρχαν κολλήματα και κυρίως καθυστερήσεις σε περιπτώσεις που πέραν της διαδικασίας της ταξινόμησης ,εκτελούνταν κι άλλη διεργασία, όπως η συγγραφή της τεχνικής έκθεσης. Ενδεικτικά, οι χρόνοι υλοποίησης των πειραμάτων ήταν :

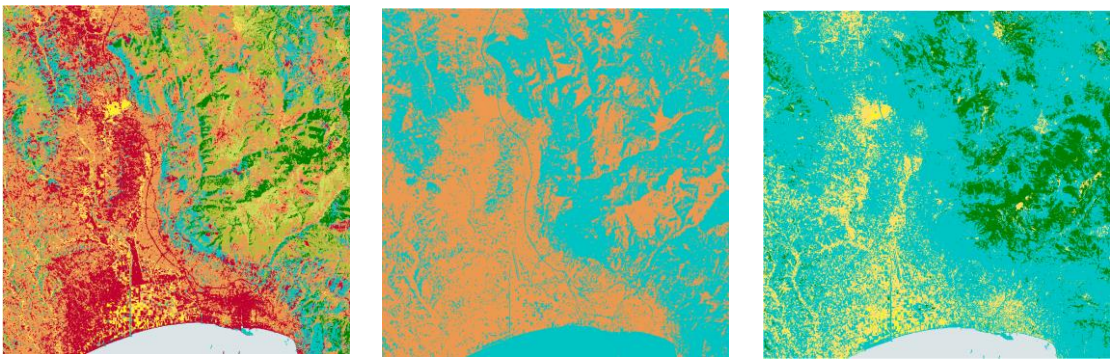
Πείραμα	Αντικείμενο	Χρόνος εκτέλεσης	Πλήθος πολυγώνων dataset
Πείραμα 1	ισορροπημένο σύνολο	47,8 min	447
Πείραμα 2	υποδειγματοληψία τεχνητών	45,6 min	432
Πείραμα 3	υποδειγματοληψία αρόσιμης	46,1 min	441
Πείραμα 4	συνδυασμός μεθόδων	51,2 min	450
Πείραμα 5	υπερδειγματοληψία λιβαδιών	50,3 min	450
Πείραμα 6	υποδειγματοληψία ετερογενών	48,9 min	439
Πείραμα 7	υπερδειγματοληψία δασών	49,6 min	460
Πείραμα 8	υποδειγματοληψία συνδυασμών	54,1 min	440
Πείραμα 9	υποδειγματοληψία όλων των κατηγοριών ,εκτός των χερσαίων υδάτων	51,4 min	405
Πείραμα 10	υπερδειγματοληψία θάλασσας	49,6 min	465

Πίνακας 40. Λεπτομέρειες υλοποίησης πειραμάτων

Ωστόσο, το σημαντικότερο όλων πρόβλημα αφορούσε την απώλεια της δυνατότητας υλοποίησης δεύτερου αλγορίθμου, τον SVM, λόγω της πολυπλοκότητάς τους και της ταυτόχρονης πολυπλοκότητας του μεγάλου dataset εκπαίδευσης του αλγορίθμου, με κατά μέσο όρο 450 πολύγωνα τη φορά. Η μη υλοποίηση του αλγορίθμου SVM είχε ως αποτέλεσμα να μην εκπληρωθεί ένας από τους στόχους της εργασίας, τη σύγκριση της απόδοσης όχι μόνο ανάλογα τη μέθοδο τυχαίας δειγματοληψίας που εφαρμόστηκε, αλλά και με βάση τον αλγόριθμο ταξινόμησης. Το πρόβλημα αυτό οδήγησε στην εναλλακτική λύση εφαρμογής ενός απλοϊκότερου αλγορίθμου, όπως οι αλγόριθμοι Μεγίστης Πιθανοφάνειας κι Ελάχιστης Απόστασης. Ακολούθησαν διαδοχικές δοκιμές ώστε να επιλεγεί ένας από τους δυο αλγορίθμους ,με κριτήριο την επίτευξη καλής ακρίβειας, όμως με τα αποτελέσματα που προέκυψαν από τους αλγορίθμους δεν ευοδώθηκε η εναλλακτική αυτή λύση.

Ο αλγόριθμος Maximum Likelihood λόγω της πολυπλοκότητας του συνόλου δεδομένων εκπαίδευσης, δεν ήταν ικανός να αποδώσει κάποιο αποτέλεσμα, παράγοντας μια "ταξινομημένη" απεικόνιση, στην οποία υπάρχει μόνο η κατηγορία των τεχνητών επιφανειών.

Από την άλλη, ο αλγόριθμος Minimum Distance αρχικά παρήγαγε ταξινομημένες απεικονίσεις, οι οποίες ποιοτικά απέδιδαν με υψηλή ακρίβεια, αντιμετωπίζοντας κιόλας σε ένα ποσοστό το σφάλμα απεικόνισης περιορισμένης έκτασης των δασών, αλλά στην πορεία των πειραμάτων, με την εφαρμογή τυχαίας δειγματοληψίας αναγνωρίζονταν μόνο μερικές κατηγορίες ,με συνέπεια την παύση των δοκιμών.



Σχήμα 100. Δοκιμές εφαρμογής αλγορίθμου Minimum Distance
Πείραμα με ισορροπημένο σύνολο δεδομένων (αριστερά) – Πείραμα υποδειγματοληψίας τεχνητών επιφανειών (μέση) – Πείραμα υποδειγματοληψίας αρόσιμης γης (δεξιά)

ΚΕΦΑΛΑΙΟ 8. Συμπεράσματα & Προοπτικές

Από την υλοποίηση των πειραματικών εφαρμογών επιβλεπόμενης ταξινόμησης, ως προς την ποιοτική αξιολόγηση των αποτελεσμάτων δεν προκύπτει σημαντικό συμπέρασμα για την απόδοση των μεθόδων υποδειγματοληψίας κι υπερδειγματοληψίας. Στα περισσότερα πειράματα ,ποιοτικά, καταγράφονται διαφοροποιήσεις των μερικών εικονοστοιχείων από την ταξινόμηση των κατηγοριών χρήσης γης, με τα δύο σημαντικότερα σφάλματα να παραμένουν σε όλα τα πειράματα. Τα δυο σφάλματα που αναφέρονται, αφορούν τη μη ορθή ταξινόμηση περιοχών βλάστησης ως χερσαία ύδατα και το δεύτερο σχετίζεται με την έκταση που καλύπτουν οι δασικές εκτάσεις, που είναι πολύ μικρή συγκριτικά με αυτή που ορίζουν τα ground trough δεδομένα, CLC και Κυρωμένος Δασικός Χάρτης. Τα δυο αυτά σφάλματα αιτιολογούνται με την άρδευση και χαμηλή περιεκτικότητα σε βλάστηση/αραιή βλάστηση των καλλιεργήσιμων εκτάσεων και στη δεύτερη περίπτωση, αφού οι περισσότερες δασικές εκτάσεις ταξινομούνται ως λιβάδια και συνδυασμούς βλάστησης, φαινόμενο που στηρίζεται στα εγγενή χαρακτηριστικά των κατηγοριών. Από την άλλη, από τις κατηγορίες που παρουσιάζει αλλαγές από πείραμα σε πείραμα είναι οι τεχνητές επιφάνειες, τμήματα των οποίων είναι γεγονός πως εμφανίζονται εντός καλλιεργήσιμων εκτάσεων, το οποίο πιθανότατα οφείλεται σε γυμνά ή πετρώδη εδάφη, τα οποία είναι γνωστό ότι λόγω των φυσικών ,αλλά και των φασματικών χαρακτηριστικών τους, εμφανίζουν παρόμοια συμπεριφορά στο φάσμα με τις κατηγορίες που συγκαταλέγονται στις τεχνητές επιφάνειες.

Πέραν όμως της ποιοτικής αξιολόγησης, ολοκληρώνοντας τις πειραματικές εφαρμογές, διαδοχικά, επιμέρους για κάθε πείραμα, παρατέθηκαν στις αντίστοιχες ενότητες ο πίνακας σύγχυσης και οι μετρικές ακρίβειας που απορρέουν από αυτόν, ώστε να ποσοτικοποιηθεί το ποιοτικό αποτέλεσμα της ταξινόμησης. Για το λόγο αυτό, στην παρακάτω ενότητα, παρατίθεται ένα σύνολο από διαγράμματα που παρουσιάζουν τις διακυμάνσεις των ακριβειών και πως αυτές επηρεάζονται από τη μέθοδο δειγματοληψίας που έχει εφαρμοσθεί σε κάθε πείραμα.

Στον παρακάτω πίνακα υπενθυμίζεται το αντικείμενο κάθε πειράματος.

Πείραμα 1	ισορροπημένο σύνολο
Πείραμα 2	υποδειγματοληψία τεχνητών
Πείραμα 3	υποδειγματοληψία αρόσιμης
Πείραμα 4	συνδυασμός μεθόδων
Πείραμα 5	υπερδειγματοληψία λιβαδιών
Πείραμα 6	υποδειγματοληψία ετερογενών
Πείραμα 7	υπερδειγματοληψία δασών
Πείραμα 8	υποδειγματοληψία συνδυασμών
Πείραμα 9	υποδειγματοληψία όλων των κατηγοριών ,εκτός των χερσαίων υδάτων
Πείραμα 10	υπερδειγματοληψία θάλασσας

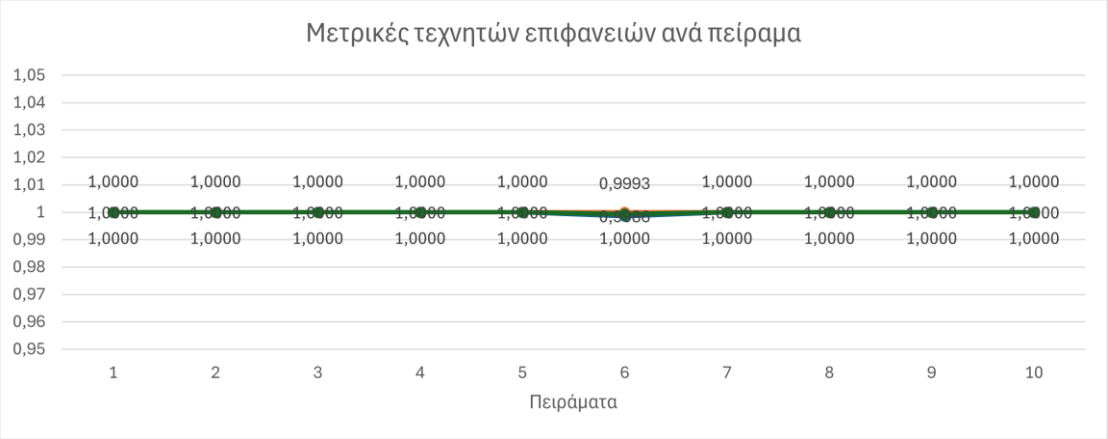
Πίνακας 41. Αντικείμενο πειραματικών εφαρμογών

8.1.Ποσοτικά συμπεράσματα για τα πειράματα ανά κατηγορία

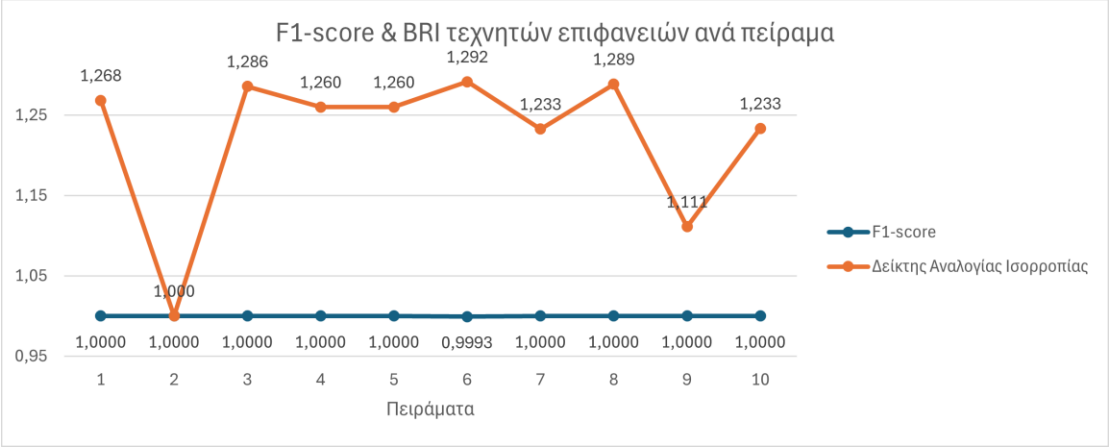
Ο πρώτος και κύριος στόχος της παρούσας Διπλωματικής Εργασίας είναι η ποσοτική αξιολόγηση πειραματικών εφαρμογών επιβλεπόμενης ταξινόμησης, με σύνολα δεδομένων εκπαίδευσης, όπου έχει εφαρμοσθεί η μέθοδος τυχαίας υπερδειγματοληψίας ή τυχαίας υποδειγματοληψίας ή ακόμα και ο συνδυασμός των δυο μεθόδων στα πολύγωνα εκπαίδευσης των επιμέρους κατηγοριών κάλυψης/χρήσης γης. Τα βασικά μεγέθη που υπολογίζονται μέσω του πίνακα σύγχυσης είναι το precision, το recall και το F1-score, που έχουν παρουσιαστεί σε συγκεντρωτικούς πίνακες κατά την υλοποίηση των πειραμάτων, για το κάθε πείραμα ξεχωριστά. Στο παρόν στάδιο επιλέγεται η προβολή της μεταβολής των τριών μετρικών ακρίβειας ανά πείραμα, για κάθε κατηγορία του CLC ξεχωριστά. Ο λόγος που τα διαγράμματα παρουσιάζουν τη διακύμανση των ακριβειών με βάση την κατηγορία ανά πείραμα, είναι επειδή ήδη τα πειράματα εστιάζουν κάθε φορά σε μια κατηγορία. Μάλιστα, επειδή οι μετρικές είναι συνυφασμένες, αφού το μέγεθος F1-score αποτελεί τον αρμονικό μέσο των precision και recall, επιλέγεται η κοινή αποτύπωσή τους σε διάγραμμα, ώστε να τονισθεί και πως επηρεάζουν τα δυο επιμέρους μεγέθη τον αρμονικό τους μέσο.

8.1.1.Τεχνητές Επιφάνειες

Οι τεχνητές επιφάνειες ως μια περίπτωση κατηγορίας, που δύσκολα συσχετίζεται με κάποια άλλη στην περιοχή ενδιαφέροντος, επιτυγχάνει μεγάλη ακρίβεια και στα τρία επιμέρους μεγέθη ακρίβειας, με τη μόνη περίπτωση πειράματος στο οποίο “χάνει” ελάχιστα η ακρίβεια του precision, να είναι αυτό της υποδειγματοληψίας των ετερογενών καλλιεργειών, με ένα εικονοστοιχείο των τεχνητών επιφανειών να έχει ταξινομηθεί ως ετερογενείς καλλιέργειες. Πρόκειται λοιπόν, για μια κατηγορία που δεν την επηρεάζει το πλήθος των πολυγώνων εκπαίδευσης, τόσο αυτής, όσο και των υπολοίπων κατηγοριών, κυρίως όμως λόγω των μοναδικών χαρακτηριστικών που εμφανίζει, αφού η περιοχή ενδιαφέροντος στη μεγαλύτερή της επιφάνεια καλύπτεται από βλάστηση.



Διάγραμμα 1. Μετρικές ακρίβειας τεχνητών επιφανειών ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

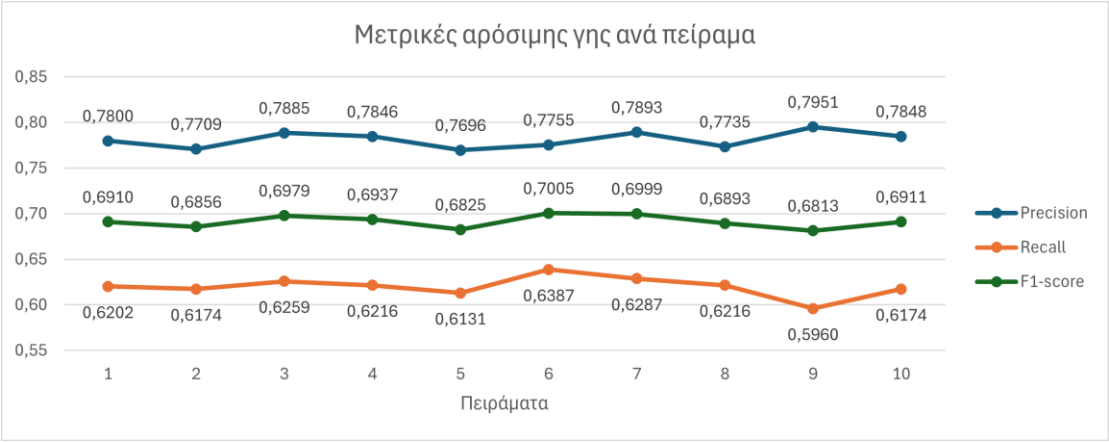


Διάγραμμα 2. Διακύμανση μετρικής F1-score Δείκτη Αναλογίας Ισορροπίας (BRI) τεχνητών επιφανειών ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

Το συμπέρασμα πως οι τεχνητές επιφάνειες δεν επηρεάζονται από τη μέθοδο δειγματοληψίας που εφαρμόζεται σε κάθε πείραμα αποτυπώνεται και στο παραπάνω διάγραμμα, όπου η ακρίβεια του F1-score παραμένει σταθερή, παρά τις όποιες διακυμάνσεις παίρνει η τιμή του Δείκτη Αναλογίας Ισορροπίας (BRI), ο οποίος εκφράζει αν η κατηγορίας βρίσκεται σε ισορροπία στο εκάστοτε πείραμα ή όχι.

8.1.2.Αρόσιμη γη

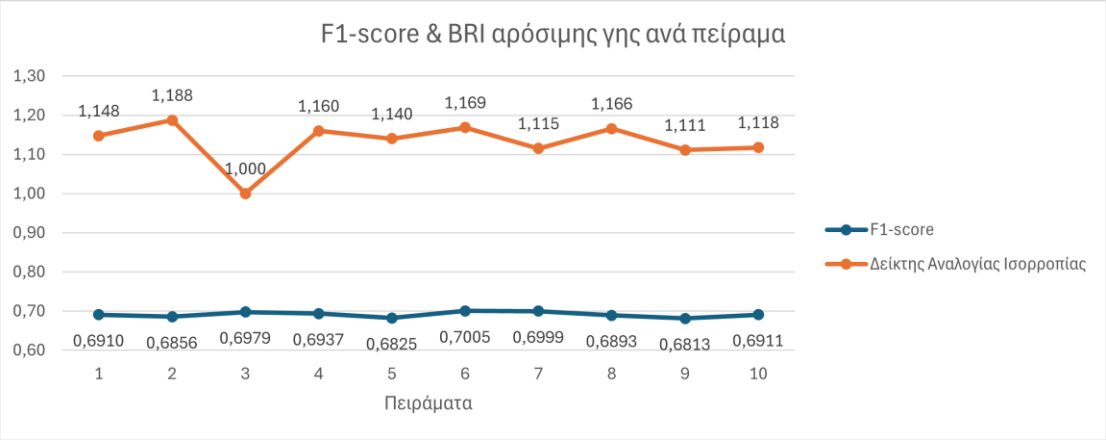
Υπενθυμίζεται πως η μετρική του precision είναι αντίστοιχη της ακρίβειας παραγωγού, δηλαδή το μέγεθος που περιγράφει την ακρίβεια της ταξινόμησης για την κάθε κλάση. Στην περίπτωση της αρόσιμης γης, το precision δεν παρουσιάζει σημαντικές μεταβολές από πείραμα σε πείραμα, με τη μικρότερη τιμή που καταγράφει να είναι ίση με 0,7696, ενώ η μεγαλύτερη 0,7951, άρα όλες οι τιμές του κυμαίνονται σε αυτό το πλαίσιο διαφοράς 0,0254. Το recall, αντίστοιχο της ακρίβειας χρήστη, είναι το μέγεθος που αντιπροσωπεύει την αξιοπιστία της κάθε κλάσης στην ταξινομημένη εικόνα και για την αρόσιμη γη παρουσιάζει λίγο μεγαλύτερη διαφορά μεταξύ της μικρότερης (0,5960) και της μεγαλύτερης (0,6387) τιμής του



Διάγραμμα 3. Μετρικές ακρίβειας μόνιμων καλλιεργειών ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

Παρατηρώντας και το Διάγραμμα 3, διαπιστώνεται πως στην πλειοψηφία των πειραμάτων υπάρχει αύξηση ή μείωση και για τα δυο μεγέθη και κατά συνέπεια το F1-

score. Στην περίπτωση του πειράματος που είναι σε ισορροπία είναι η αρόσιμη γη, έπειτα από υποδειγματοληψία, σημειώνεται η 3η καλύτερη ακρίβεια, για όλες τις μετρικές συγκριτικά με το σύνολο των πειραμάτων, ενώ, η συνολικά καλύτερη απόδοση της κατηγορίας επιτυγχάνεται στο πείραμα υποδειγματοληψίας των ετερογενών καλλιεργειών. Βέβαια, το συμπέρασμα αυτό είναι αναμενόμενο, αναλογιζόμενοι την άμεση συσχέτιση μεταξύ των δυο κατηγοριών και τα σφάλματα παράλειψης και συμπερίληψης που παρουσιάζονται εξαιτίας αυτής.

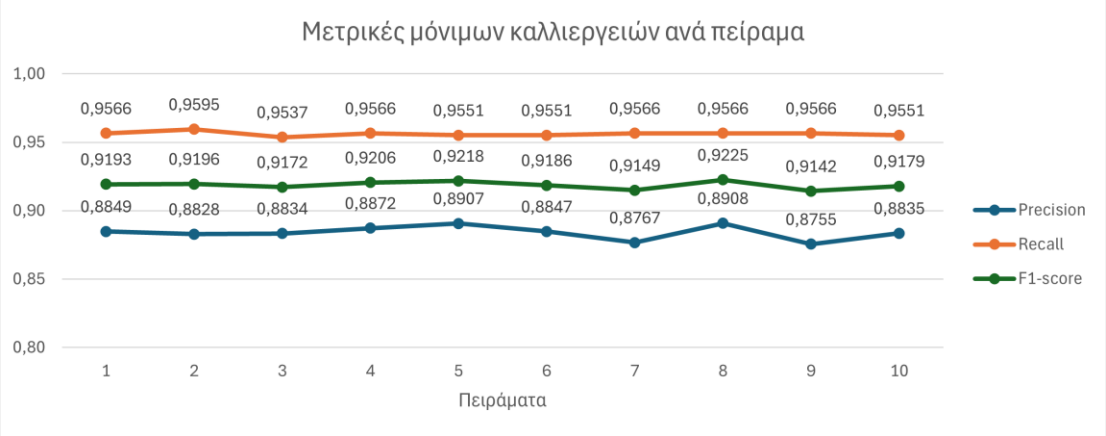


Διάγραμμα 4. Διακύμανση μετρικής F1-score Δείκτη Αναλογίας Ισορροπίας (BRI) αρόσιμης γης ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

Παράλληλα, με τη σχηματική απόδοση των ακριβειών φαίνεται πόσο επηρεάζει την αρόσιμη γη, το πείραμα όπου εφαρμόζεται σε όλες τις κατηγορίες υποδειγματοληψία, εκτός από τα χερσαία ύδατα, ώστε αυτά να αποκτήσουν την ιδανική αναλογία. Μάλιστα, πρόκειται για το πείραμα όπου έχει τη 2η κοντινότερη στη μονάδα τιμή ο δείκτης αναλογίας. Παρόλα αυτά, στο συγκεκριμένο πείραμα η αρόσιμη γη “διχάζεται”, σημειώνοντας το υψηλότερο precision και το χαμηλότερο recall, επηρεάζοντας τον αρμονικό μέσο, καταγράφοντας εν τέλει τη χειρότερη απόδοση για την κατηγορία.

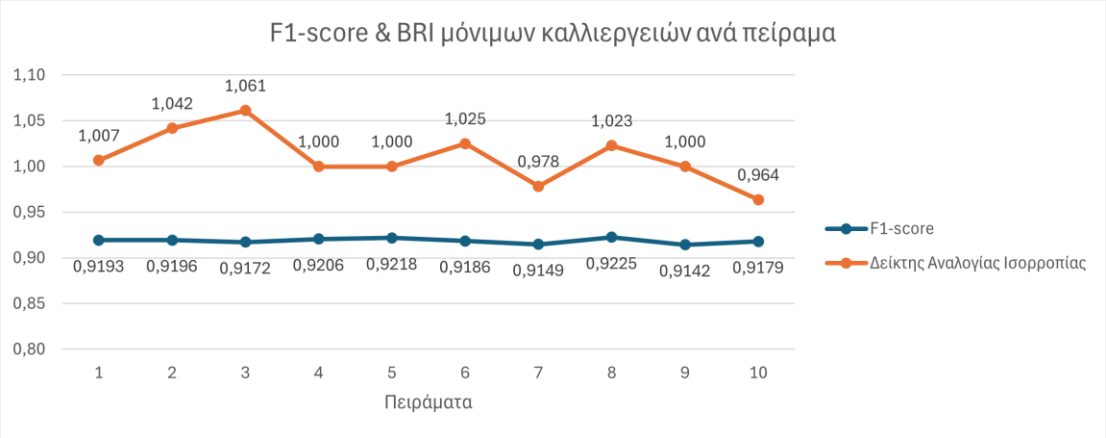
8.1.3.Μόνιμες καλλιέργειες

Οι μόνιμες καλλιέργειες παρά το γεγονός ότι κατά την ποιοτική αξιολόγηση είναι μια αμφιλεγόμενη κατηγορία, αφού είναι μια από τις περιπτώσεις που επηρεάζονται από την κατηγοριοποίηση των χερσαίων υδάτων, ποσοτικά κατατάσσεται στις κατηγορίες που εμφανίζουν υψηλές ακρίβειες.



Διάγραμμα 5. Μετρικές ακρίβειας μόνιμων καλλιεργειών ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

Μάλιστα, πρόκειται για μια κατηγορία με σχεδόν σταθερές τιμές ακρίβειας, ιδιαίτερα στο precision, που η διαφοροποίηση καταγράφεται στο τρίτο δεκαδικό ψηφίο, με αντίστοιχη πορεία να ακολουθείται και στο recall, με το μόνο “σπάσιμο” να καταγράφεται στο πείραμα υποδειγματοληψίας των συνδυασμών βλάστησης, με τις μόνιμες καλλιέργειες να αυξάνουν ακρίβεια. Αν έπρεπε να σημειωθεί ένα πείραμα το οποίο επηρεάζει αρνητικά την απόδοση των μόνιμων καλλιεργειών, παραμένοντας όμως μεγαλύτερη του 0,90, είναι αυτό όπου όλες οι κατηγορίες επιδέχονται υποδειγματοληψία για την ισορροπία των χερσαίων υδάτων. Η μείωση λοιπόν των πολυγώνων εκπαίδευσης των μόνιμων καλλιεργειών προκαλεί μικρή μείωση στην απόδοσή τους.

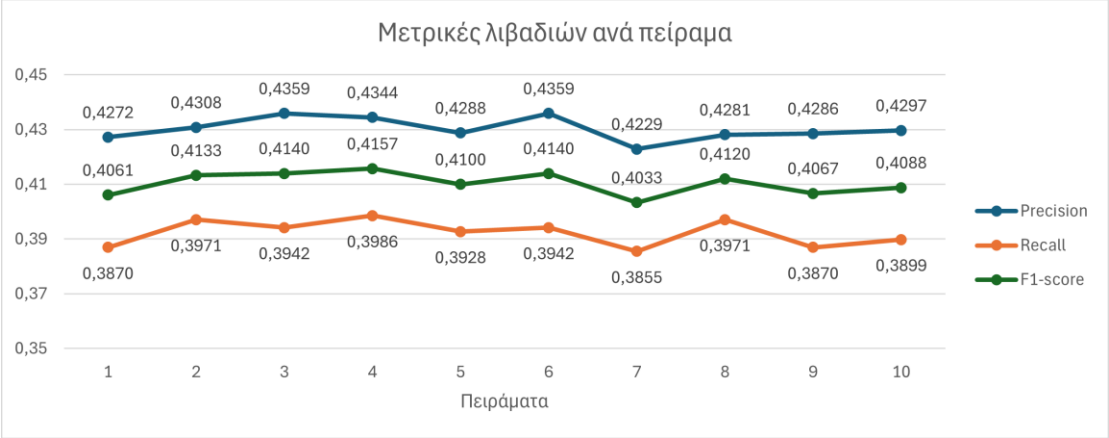


Διάγραμμα 6. Διακύμανση μετρικής F1-score Δείκτη Αναλογίας Ισορροπίας (BRI) αρόσιμης γης ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

Λαμβάνοντας δε υπόψιν ότι πρόκειται για την κατηγορία όπου ήδη, στο αρχικό, ισορροπημένο σύνολο δεδομένων εκπαίδευσης τείνει να θεωρείται ισορροπημένη, ενώ ακόμα και στο πείραμα που διερευνάται η απόλυτη ισορροπία της, μέσα από συνδυασμό τυχαίας υπερδειγματοληψίας και τυχαίας υποδειγματοληψίας των υπόλοιπων κατηγοριών, διατηρεί σταθερά τα πολύγωνα εκπαίδευσής της, με τιμές δείκτη αναλογίας μικρής απόκλισης από τη μονάδα. Ταυτόχρονα, λαμβάνεται υπόψιν και το γεγονός πως στο πείραμα που γίνεται υποδειγματοληψία των πολυγώνων, για την ισορροπίας των χερσαίων υδάτων, οι μόνιμες καλλιέργειες είναι επίσης ισορροπημένες, αλλά η μείωση των πολυγώνων τους, είχε αποτέλεσμα τη μικρή μείωση της απόδοσής τους, λόγω της πτώσης της ακρίβειας του recall.

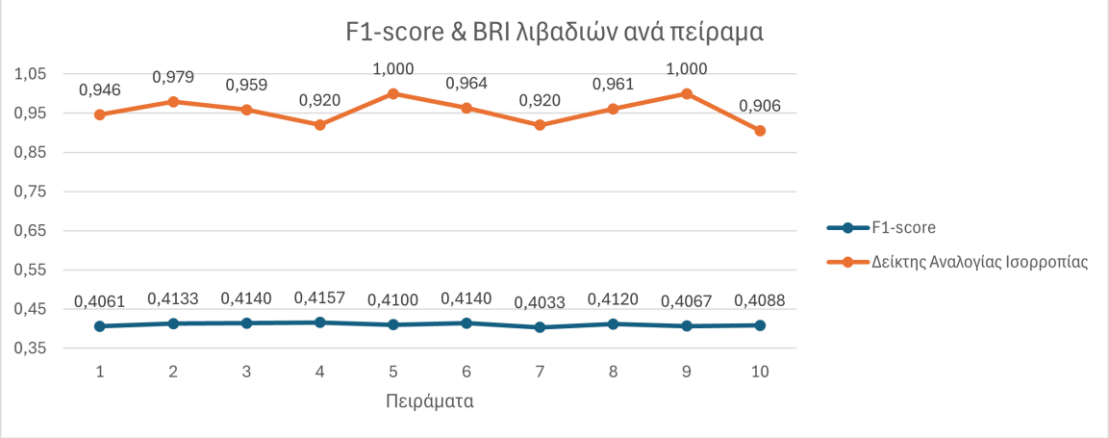
8.1.4.Λιβάδια

Καθ’ όλη την έκταση των πειραματικών εφαρμογών και του σταδίου αξιολόγησής τους τονίζεται η μέτρια προς κακή απόδοση των λιβαδιών, λόγω του γεγονότος ότι είναι άρρηκτα συνδεδεμένη κατηγορία με τους συνδυασμούς βλάστησης σε πρώτο βαθμό και σε δεύτερο με τις δασικές εκτάσεις. Παρά την προσεκτική ερμηνεία της κατηγορίας επί της πολυφασματικής απεικόνισης και κατ’ επέκταση τον επιφυλακτικό σχεδιασμό πολυγώνων εκπαίδευσης κι ελέγχου για την κατηγορία, οι ακρίβειες που καταγράφει κυμαίνονται για το precision μεταξύ 0,42 και 0,43, ενώ για το recall, ακόμα χειρότερα αποτελέσματα, με τιμές 0,38-0,40, με αποτέλεσμα ο αρμονικός μέσος να παίρνει τιμές περί το 0,41.



Διάγραμμα 7. Μετρικές ακρίβειας λιβαδιών ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

Σύμφωνα με το διάγραμμα, εκτός των άλλων, γίνεται αντιληπτό πως σε τρία πειράματα οι ακρίβειες έρχονται σε “ρήξη”, καθώς το precision τείνει να αυξάνεται, σε αντίθεση με το recall που έχει καθοδική πορεία, η οποία είναι σε τέτοιο βαθμό που επιφέρει μείωση και στο F1-score.

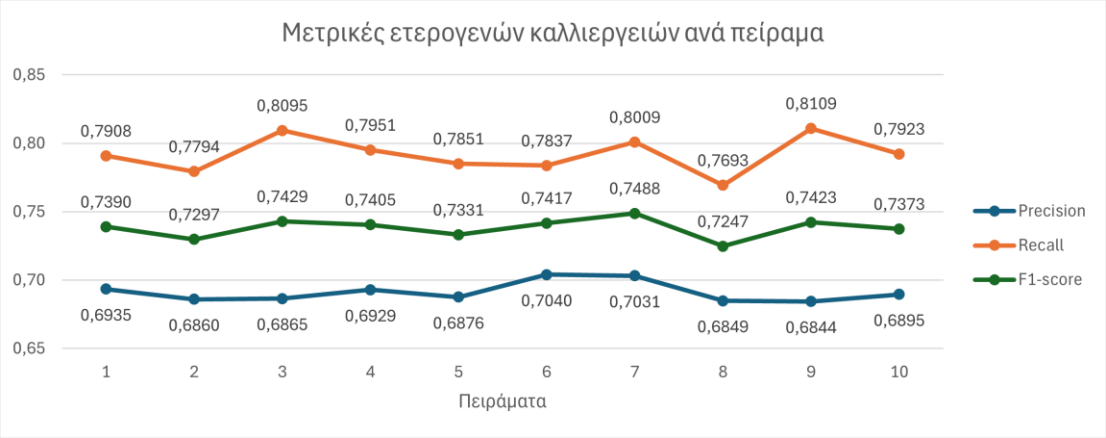


Διάγραμμα 8. Διακύμανση μετρικής F1-score Δείκτη Αναλογίας Ισορροπίας (BRI) λιβαδιών ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

Εντύπωση προκαλεί η παρατήρηση πως το πείραμα στο οποίο έχει δεχθεί υποδειγματοληψία η κατηγορία των λιβαδιών -Πείραμα 4 , είναι και το πείραμα στο οποίο αποδίδει τις καλύτερες ακρίβειες. Θα μπορούσε να χαρακτηριστεί μια από τις κατηγορίες που επηρεάζονται από την ευρύτερη συνθήκη που επικρατεί στο σύνολο δεδομένων εκπαίδευσης, μιας και στο Πείραμα 5, που αφορά την υπερδειγματοληψία της κατηγορίας ,σημειώνει μια μέτρια απόδοση, ενώ στο Πείραμα 9, που δέχεται υποδειγματοληψία, όπως όλες οι κατηγορίες ,εκτός των χερσαίων υδάτων, σημειώνει μια μέτρια πως χαμηλή ακρίβεια, παρά το γεγονός πως ο δείκτης αναλογίας ισορροπίας είναι ίσος με τη μονάδα.

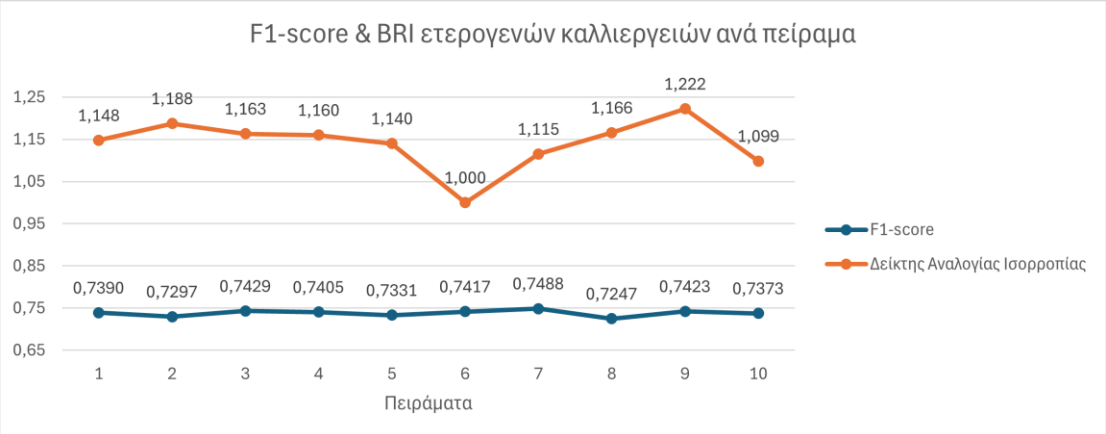
8.1.5.Ετερογενείς γεωργικές εκτάσεις

Οι ετερογενείς καλλιέργειες είναι μια αμφιλεγόμενη κατηγορία, διότι εμφανίζει παρόμοια χαρακτηριστικά με την αρόσιμη γη, με αποτέλεσμα να σημειώνονται σφάλματα ανάμεσα στις δυο και να μειώνεται η συνολική τους απόδοση.



Διάγραμμα 10. Μετρικές ακρίβειας ετερογενών καλλιεργειών ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

Πιο συγκεκριμένα, από το Διάγραμμα 9 προκύπτει πως υπάρχουν διακυμάνσεις στις ακρίβειες των ετερογενών καλλιεργειών από πείραμα σε πείραμα, με τις μεγαλύτερες να εμφανίζονται στο precision. Το ένα πείραμα στο οποίο καταγράφεται σημαντική αύξηση του precision των ετερογενών καλλιεργειών είναι αυτό της υποδειγματοληψίας της αρόσιμης γης, ενώ σημαντική μείωση προκύπτει στο πείραμα υποδειγματοληψίας των συνδυασμών βλάστησης. Η μεγαλύτερη τιμή που καταγράφει το precision είναι ίση με 0,8109 στο πείραμα υποδειγματοληψίας όλων των κατηγοριών ,εκτός των χερσαίων υδάτων, με τη χαμηλότερη να σημειώνεται στο αμέσως προηγούμενο πείραμα, υποδειγματοληψίας των συνδυασμών βλάστησης. Στον αντίποδα, το recall εμφανίζει μέγιστη τιμή στο πείραμα υποδειγματοληψίας των ετερογενών καλλιεργειών και τη μικρότερη στην εφαρμογή υποδειγματοληψίας για την ισορροπία των χερσαίων υδάτων. Όλα αυτά έχουν ως αποτέλεσμα “σκαμπανεβάσματα” στο διάγραμμα για τον αρμονικό μέσο F1-score, στην προσπάθειά του να ακολουθήσει τις διακυμάνσεις των δυο προηγούμενων μετρικών.

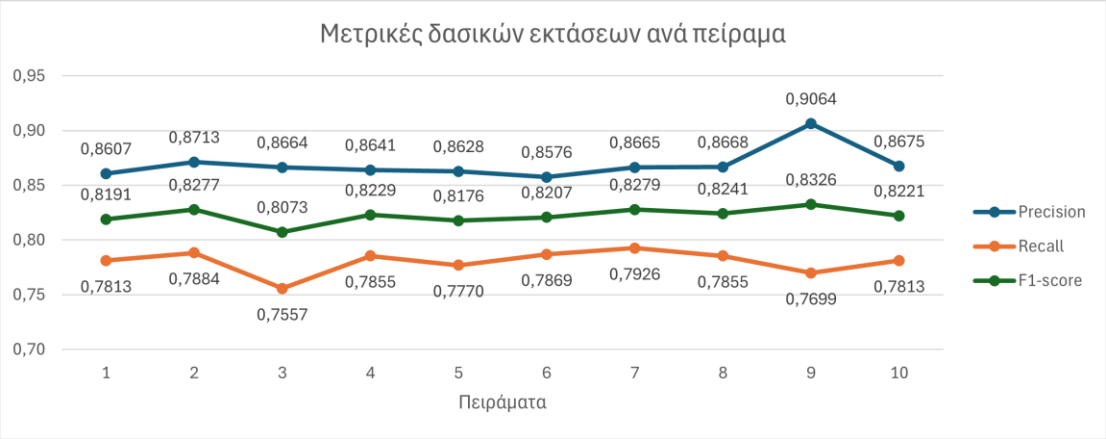


Διάγραμμα 9. Διακύμανση μετρικής F1-score Δείκτη Αναλογίας Ισορροπίας (BRI) ετερογενών καλλιεργειών ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

Αξιολογώντας την πορεία των μετρικών ακρίβειας των ετερογενών καλλιεργειών και πως αυτές επηρεάζονται από το πόσο κοντά ή μακριά από τη μονάδα είναι ο δείκτης αναλογίας ισορροπίας -Διαγράμματα 9 &10-, δεν προκύπτει σαφές συμπέρασμα για το εάν επιδρούν θετικά ή αρνητικά στην απόδοση της κατηγορίας οι δυο μέθοδοι δειγματοληψίας, αφού υπάρχουν περιπτώσεις που η ίδια μέθοδος στη μια περίπτωση αυξάνει την απόδοση των ετερογενών καλλιεργειών και στο δεύτερο πείραμα, η ίδια μέθοδος αποφέρει μείωση της απόδοσης της κατηγορίας. Επομένως, κρίνοντας κι από τις τιμές που λαμβάνει ο δείκτης αναλογίας ισορροπίας και πως κυμαίνεται στα αντίστοιχα πειράματα ο αρμονικός μέσος, F1-score γίνεται αντιληπτό πως το σύνολο των πολυγώνων εκπαίδευσης των ετερογενών καλλιεργειών επηρεάζεται από τη διαμόρφωση του ευρύτερου συνόλου δεδομένων εκπαίδευσης, που αφορά όλες τις κατηγορίες.

8.1.6.Δασικές εκτάσεις

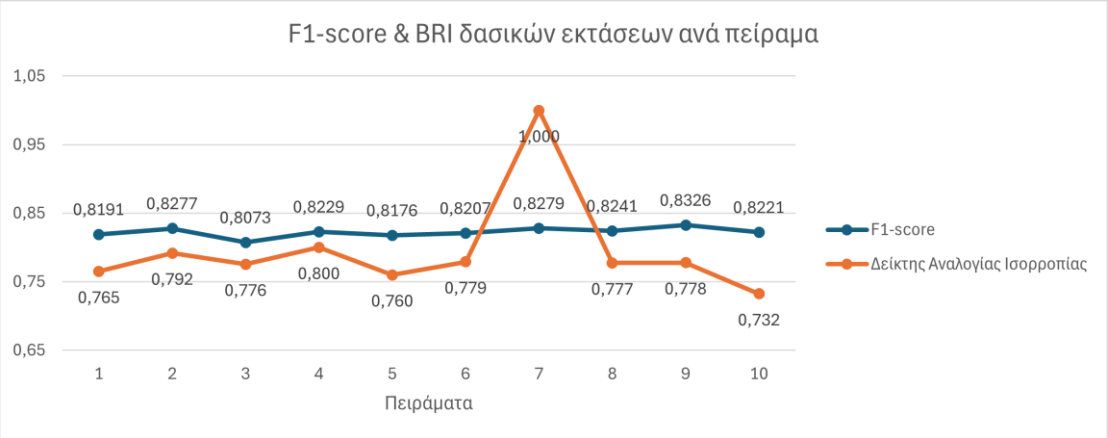
Οι δασικές εκτάσεις ,παρά την άμεση συσχέτισή τους με την κατηγορία των συνδυασμών βλάστησης πρωτίστως και δευτερεύοντος με τα λιβάδια, επιτυγχάνουν σημαντικά ποσοστά ακρίβειας σε όλες τις πειραματικές εφαρμογές, ανεξαρτήτως της μεθόδου δειγματοληψίας που έχει εφαρμοστεί στο εκάστοτε σύνολο δεδομένων εκπαίδευσης.



Διάγραμμα 11. Μετρικές ακρίβειας δασικών εκτάσεων ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

Δυο είναι τα πειράματα στα οποία μεταβάλλεται αρκετά η ακρίβεια των δασικών εκτάσεων συγκριτικά με τις τιμές που λαμβάνει στο σύνολο των πειραμάτων. Το ένα πείραμα αφορά την υποδειγματοληψία των πολυγώνων αρόσιμης γης και το άλλο την υποδειγματοληψία όλων των κατηγοριών, ώστε να ισορροπήσει ο δείκτης αναλογίας ισορροπίας για τα χερσαία ύδατα. Στην πρώτη περίπτωση καταγράφεται σημαντική μείωση του recall, σημειώνοντας και τη μικρότερη τιμή του, σε αντίθεση με το δεύτερο πείραμα που αυξάνει κατακόρυφα η ακρίβεια precision, σημειώνοντας τη μέγιστη τιμή του precision. Παρά τις όποιες διακυμάνσεις των precision και recall, ο αρμονικός τους μέσος ακολουθεί πιο ομαλές μεταβολές μεταξύ των τιμών του από πείραμα σε πείραμα. Η μικρότερη απόδοση της κατηγορίας των δασών είναι στο πείραμα υποδειγματοληψίας της αρόσιμης γης, που σημειώνεται και η ελάχιστη τιμή του recall, ενώ αντίστοιχα, η μεγαλύτερη τιμή εμφανίζεται στο πείραμα που καταγράφει την υψηλότερη τιμή του το precision, με την υποδειγματοληψία όλων εκτός των χερσαίων υδάτων. Διαπιστώνεται λοιπόν, για ακόμα μια φορά η θετική επίδραση της

υποδειγματοληψίας των κατηγοριών για την ισορροπία των χερσαίων υδάτων με την αύξηση της απόδοσης κάθε κατηγορίας.

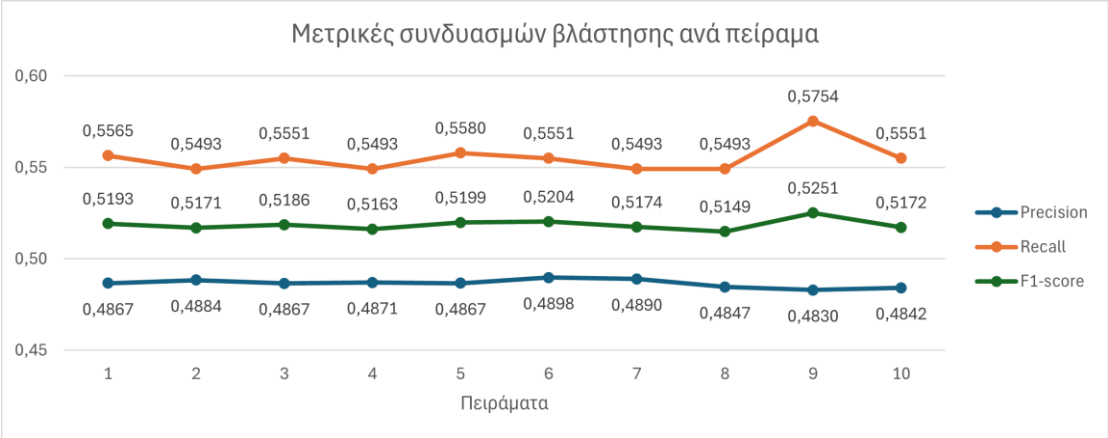


Διάγραμμα 12. Διακύμανση μετρικής F1-score Δείκτη Αναλογίας Ισορροπίας (BRI) ετερογενών καλλιεργειών ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

Στο σύνολο των πειραμάτων δε φαίνεται να επηρεάζεται σημαντικά η κατηγορία των δασών από τις τροποποιήσεις στο ευρύτερο σύνολο δεδομένων εκπαίδευσης, καθώς αρχικά ο ίδιος ο δείκτης αναλογίας ισορροπίας δε μεταβάλλεται ιδιαίτερα από τις μεθόδους δειγματοληψίας που έχουν εφαρμοστεί στα dataset των πειραμάτων, παρά μόνο στην περίπτωση τυχαίας υπερδειγματοληψίας πολυγώνων των δασών “εκτινάσσεται”, για να φτάσει τη μονάδα, με τον αρμονικό μέσο να “ακολουθεί” σχεδόν αντίστοιχη πορεία με αυτή του δείκτη αναλογίας ισορροπίας.

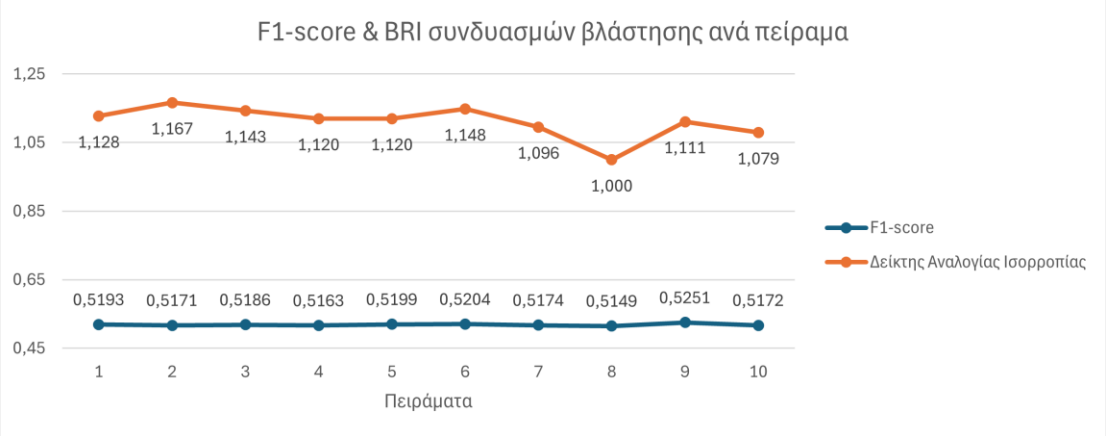
8.1.7. Συνδυασμοί βλάστησης

Οι συνδυασμοί βλάστησης είναι η μια από τις δυο κατηγορίες που παρουσιάζουν μέτριες ακρίβειες συγκριτικά με τις υψηλές επιδόσεις των υπολοίπων, που σημειώνουν κατά μέσο όρο τιμές μεγαλύτερες του 0,7. Σε όλη την έκταση των πειραματικών εφαρμογών αντικείμενο σχολιασμού αποτελεί η εσφαλμένη ταξινόμηση μεγάλης μερίδας εικονοστοιχείων των συνδυασμών βλάστησης ως λιβάδια και των λιβαδιών ως συνδυασμούς βλάστησης, καθώς πρόκειται για ένα φαινόμενο που αποδεδειγμένα επηρεάζει σημαντικά την ακρίβεια των κατηγοριών, αφού οι μετρικές ποσοτικής αξιολόγησης σημειώνουν τιμές πλησίον του 0,5.



Διάγραμμα 13. Μετρικές ακρίβειας συνδυασμών βλάστησης ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

Στο σύνολο των πειραμάτων ,οι συνδυασμοί βλάστησης σημειώνουν μικρές διακυμάνσεις των ακριβειών που επιτυγχάνουν από πείραμα σε πείραμα. Εμφανή παρόλα αυτά είναι τα τοπικά ελάχιστα στο διάγραμμα για το precision, στα οποία καταγράφεται και η μικρότερη ακρίβεια του precision σε όλα τα πειράματα κι αφορούν τα πειράματα υποδειγματοληψίας των τεχνητών επιφανειών, στο συνδυασμό μεθόδων για την απόλυτη ισορροπία των μόνιμων καλλιεργειών, στην υπερδειγματοληψία των δασικών εκτάσεων και τέλος, στην υποδειγματοληψία των πολυγώνων των συνδυασμών βλάστησης. Υψίστης σημασίας είναι το πρόβλημα πως ενώ στόχος της εφαρμογής υποδειγματοληψίας ή υπερδειγματοληψία στην εκάστοτε κατηγορία είναι η επίτευξη της ακρίβειας της κατηγορίας και σε δεύτερο χρόνο της συνολικής ακρίβειας της ταξινόμησης. Παρόλα αυτά, στην προκειμένη περίπτωση ,η τυχαία υποδειγματοληψία των πολυγώνων των συνδυασμών βλάστησης αντί να επιτύχει αύξηση της ακρίβειας της κατηγορίας, κατάφερε να καταγράψει τη μικρότερη ακρίβεια από όλα τα πειράματα. Ταυτόχρονα, για ακόμη μια κατηγορία, η εφαρμογή τυχαίας υποδειγματοληψίας ώστε να έχει τιμή ίση με τη μονάδα ο δείκτης αναλογίας ισορροπίας των χερσαίων υδάτων, προσφέρει στη βελτίωση της απόδοσης της κατηγορίας, επιτυγχάνοντας μάλιστα τη μέγιστη τιμή, όπως προκύπτει από τη σύγκριση των τιμών όλων των πειραμάτων.

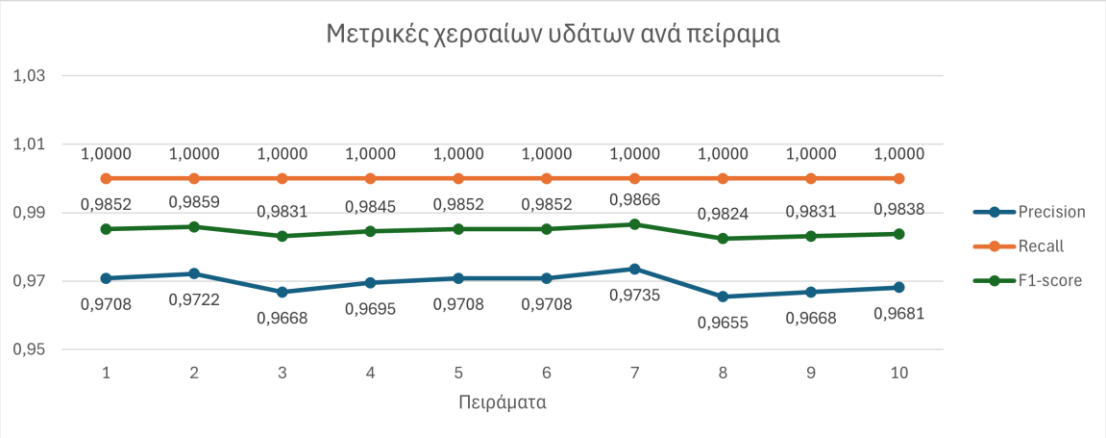


Διάγραμμα 14. Διακύμανση μετρικής F1-score Δείκτη Αναλογίας Ισορροπίας (BRI) συνδυασμών βλάστησης ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

Τα μόνα πειράματα στα οποία αλλάζει το πλήθος των πολυγώνων των συνδυασμών βλάστησης, είναι το 8ο και το 9ο ,στα οποία τα πολύγωνα επιδέχονται μείωση και παρά το γεγονός ότι εφαρμόζεται η ίδια μέθοδος δειγματοληψίας και στα δυο πειράματα, σημειώνονται οι δυο περιπτώσεις με τη χείριστη και την καλύτερη ακρίβεια της κατηγορίας, αντίστοιχα. Στα υπόλοιπα πειράματα, ο δείκτης αναλογίας ισορροπίας επηρεάζεται από τις μεταβολές που σημειώνονται για τα πολύγωνα εκπαίδευσης των υπολοίπων κατηγοριών, με τις μεταβολές αυτές να μην είναι σημαντικές, ενώ στην προσπάθεια παραλληλισμού της διακύμανσης των τιμών του δείκτη, με τις τιμές του αρμονικού μέσου, δεν προκύπτει συγκεκριμένο συμπέρασμα, αφού σε περιπτώσεις αύξησης του δείκτη υπάρχει μείωση της τιμής του F1-score, αλλά και το αντίστροφο.

8.1.8.Χερσαία ύδατα

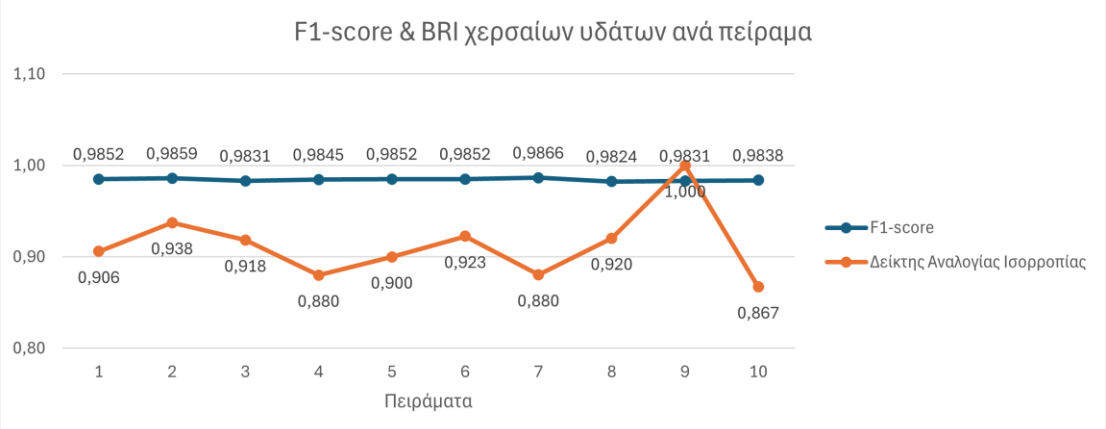
Η πιο “ιδιαίτερη” κατηγορία, τα χερσαία ύδατα, μιας κι αποτέλεσαν τη βάση για τη δημιουργία του αρχικού, ισορροπημένου dataset, φαίνεται μέχρι στιγμής, να είναι και η κατηγορία, όπου το πείραμα για την ιδανική της αναλογία, αποδίδει για όλες τις κατηγορίες, σημειώνοντας σε αυτό την υψηλότερή τους ακρίβεια.



Διάγραμμα 15. Μετρικές ακρίβειας χερσαίων υδάτων ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

Έρχεται λοιπόν, το Διάγραμμα 15 να καταρρίψει το συμπέρασμα που ίσχυε μέχρι στιγμής, καθώς το πείραμα υποδειγματοληψίας όλων των κατηγοριών εκτός των χερσαίων υδάτων, ενώ παρουσιάζει για όλες τις κατηγορίες τη μέγιστη ακρίβεια που καταγράφουν σε πείραμα, για τα χερσαία ύδατα δεν αποδίδει με τον ίδιο τρόπο.

Ο λόγος που “πέφτει” ο αρμονικός μέσος για τα χερσαία ύδατα, είναι οι διακυμάνσεις του recall, το μέγεθος που αντιπροσωπεύει την αξιοπιστία της κλάσης στην ταξινομημένη εικόνα κι αυτό γιατί καταγράφεται το σφάλμα ταξινόμησης περιοχών βλάστησης ως χερσαία ύδατα, καλύπτοντας μάλιστα σημαντικές εκτάσεις ,όπως προκύπτει από την ποιοτική αξιολόγηση των πειραμάτων. Οι περιπτώσεις που επηρεάζεται η μετρική recall για τα χερσαία ύδατα είναι το πείραμα υποδειγματοληψίας της αρόσιμης γης, κατηγορία, η οποία λόγω των αρδεύσιμων εκτάσεων συσχετίζεται άμεσα με τα χαρακτηριστικά που έχουν αναφερθεί για τα χερσαία ύδατα. Από την άλλη, Η μέγιστη τιμή του recall καταγράφεται στο πείραμα υπερδειγματοληψίας των δασικών εκτάσεων ,αφού στο συγκεκριμένο πείραμα ενισχύθηκε η εξεταζόμενη κατηγορία, αποβάλλοντας από τα εικονοστοιχεία της άλλες κατηγορίες, μια εκ των οποίων ήταν και τα χερσαία ύδατα, με συνέπεια να αυξάνεται η αξιοπιστία των εικονοστοιχείων που περιγράφουν τα χερσαία ύδατα στην ταξινομημένη απεικόνιση.



Διάγραμμα 16. Διακύμανση μετρικής F1-score Δείκτη Αναλογίας Ισορροπίας (BRI) χερσαίων υδάτων ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

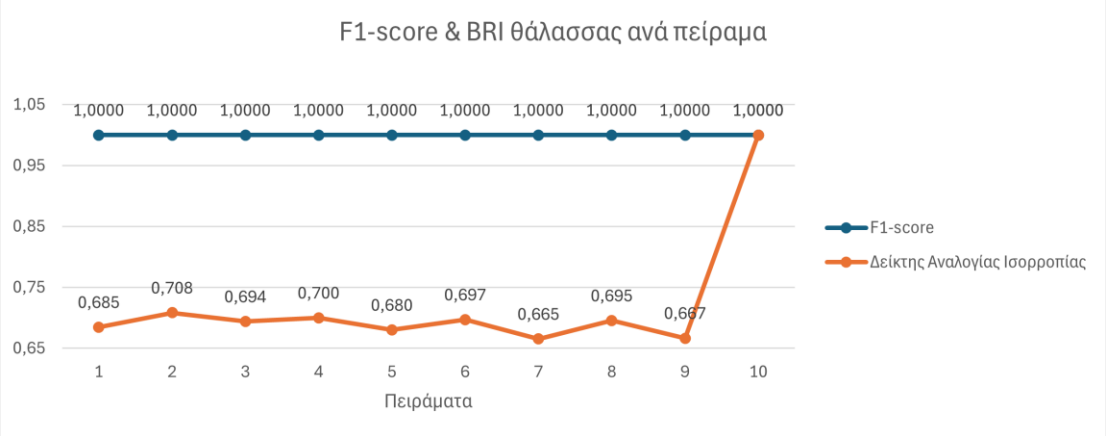
Παρά το γεγονός πως τα χερσαία ύδατα, ιδιαίτερα μέσω της ποιοτικής αξιολόγησης, φάνηκε να αποδίδουν εσφαλμένα, σύμφωνα με τον πίνακα σύγχυσης και τις μετρικές που υπολογίζονται από αυτόν καταγράφει μεγάλα ποσοστά ακρίβειας, τα οποία παρατηρώντας το Διάγραμμα 16 αυξάνονταν στις περιπτώσεις που ο δείκτης αναλογίας απομακρυνόταν από τη μονάδα -τιμές ≤0,88.

8.1.9.Θάλασσα

Τέλος, η κατηγορία της θάλασσας, η μοναδικότητα της οποίας δεν επιτρέπει σφάλματα, καθ' όλη την έκταση των πειραμάτων αποδίδει τα μέγιστα, χωρίς σφάλματα, με αποτέλεσμα στο διάγραμμα που απεικονίζονται οι μετρικές ποσοτικής αξιολόγησης, οι αναπαραστάσεις του precision, του recall και του F1-score να συμπίπτουν.

Διάγραμμα 17. Μετρικές ακρίβειας θάλασσας ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

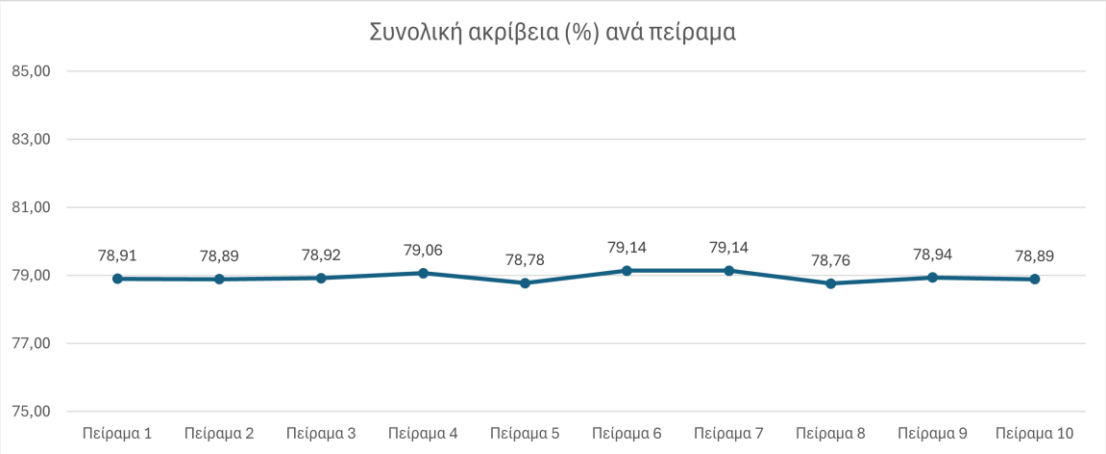
Παρά το γεγονός πως πρόκειται για την κατηγορία με τη λιγότερη εκπροσώπηση από άποψη πολυγώνων εκπαίδευσης, λόγω του ότι δεν υπήρχε άλλη κατηγορία με παρόμοια συμπεριφορά στο φάσμα, αφού τα χερσαία ύδατα επηρεάζονται σημαντικά από το βάθος και τα λοιπά υλικά που εντοπίζονται εντός αυτών, καταφέρνει την απόλυτη ακρίβεια, παρά τις όποιες τροποποιήσεις που δέχεται το dataset και την όποια μέθοδο δειγματοληψίας έχει εφαρμοσθεί, με το δείκτη αναλογίας ισορροπίας της να απέχει σημαντικά από τη μονάδα, που σηματοδοτεί την αρμονία του συνόλου εκπαίδευσης.



Διάγραμμα 18. Διακύμανση μετρικής F1-score Δείκτη Αναλογίας Ισορροπίας (BRI) θάλασσας ανά πείραμα επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

8.2.Ποσοτικά συμπεράσματα για τα πειράματα συνολικά

Στα πλαίσια αξιολόγησης των πειραμάτων ,μέσω του πίνακα σύγχυσης, για κάθε πείραμα υπολογίζεται και η συνολική ακρίβεια που επιτυγχάνεται σε κάθε εφαρμογή και ο δείκτης συμφωνίας k-hat. Έχει τονισθεί πολλάκις πως η συνολική ακρίβεια, για τέτοιου είδους πειραματικές εφαρμογές δεν αποτελεί ένα απαραίτητα αντιπροσωπευτικό μέγεθος ,καθώς αυξάνεται εκθετικά, επηρεαζόμενη από τις υψηλές ακρίβειες που καταγράφουν οι κατηγορίες ξεχωριστά. Ωστόσο, αφού στο σύνολο του πλήθους των κατηγοριών κάλυψης/χρήσης γης καταγράφονται αρκετά καλά μεγέθη ακρίβειας, δημιουργείται ένα διάγραμμα για την απεικόνιση της διακύμανσης της συνολικής ακρίβειας ανά πείραμα.



Διάγραμμα 19. Συνολική ακρίβεια (%) πειραματικών εφαρμογών επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

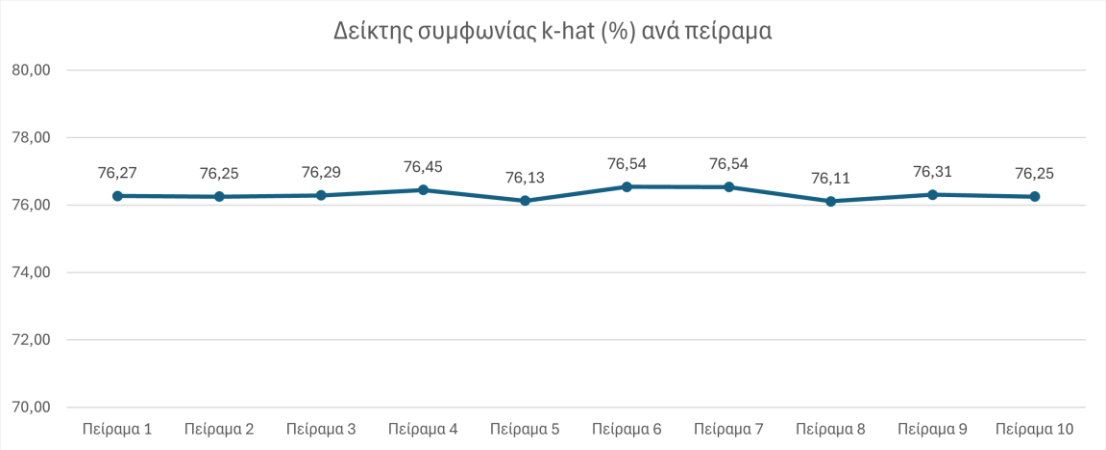
Βασικός στόχος ήταν η αύξηση της ακρίβειας με τη χρήση των μεθόδων τυχαίας υπερδειγματοληψίας κι υποδειγματοληψίας, όμως αυτό δε συμβαίνει σε όλα τα πειράματα. Στις ταξινομημένες απεικονίσεις, η συνολική ακρίβεια που επιτυγχάνεται κυμαίνεται πλησίον του 79%, με μικρές αποκλίσεις από πείραμα σε πείραμα, αλλά και με βάση το 1ο πείραμα που αποτελεί τον οδηγό.

Ένας από τους λόγους που οι πειραματικές εφαρμογές ξεκίνησαν με το ισορροπημένο ως προς την έκταση των δεδομένων εκπαίδευσης είναι για τη δημιουργία ενός μέτρου σύγκρισης των λοιπών αποτελεσμάτων και για την αξιολόγηση της επίδρασης της μεθόδου δειγματοληψίας σε κάθε πείραμα. Δεν είναι όλα τα πειράματα ικανά για την αύξηση της ακρίβειας συγκριτικά με το 1ο ,γεγονός που φάνηκε κι από την ανάλυση των αποτελεσμάτων ανά κατηγορία, με το άξιο αναφοράς, τις περιπτώσεις που στο πείραμα που η εξεταζόμενη κατηγορία κάλυψης γης είναι σε ισορροπία, με δείκτη αναλογίας ισορροπίας ίσο με τη μονάδα, να αποδίδει τελικά τη χειρότερή της ακρίβεια. Χαρακτηριστικό δε, είναι το πείραμα όπου όλες οι χρήσεις γης υποδειγματοληπτούνται, με στόχο την ισορροπία των χερσαίων υδάτων. Στο συγκεκριμένο πείραμα, όλες οι κατηγορίες πέτυχαν τη μέγιστη τιμή ακρίβειας στον αρμονικό μέσο, F1-score, συγκρίνοντας τις επιδόσεις τους με τα υπόλοιπα πειράματα, ενώ η κατηγορία των χερσαίων υδάτων, δεν απέδωσε το ίδιο με τις υπόλοιπες. Έχοντας υπόψιν τη μέγιστη ακρίβεια που επιτυγχάνεται για τις 8 από τις 9 κατηγορίες χρήσης γης, στο προαναφερθέν πείραμα, ίσως άμεση συνέπεια να θεωρείτο και η μέγιστη συνολική ακρίβεια του πειράματος να είναι η μέγιστη όλων των πειραμάτων, γεγονός που δεν επαληθεύεται από το παραπάνω Διάγραμμα.

Αναμενόμενα “κακή” η ακρίβεια των πειραμάτων υπερδειγματοληψίας των λιβαδιών κι υποδειγματοληψίας των συνδυασμών βλάστησης, καθώς τα εγγενή χαρακτηριστικά τους, τα οποία αναγνωρίζονται τόσο ποσοτικά μέσω της φασματικής τους υπογραφής, που κατ’ επέκταση υποδεικνύει την παρόμοια συμπεριφορά τους στο φάσμα, όσο και ποιοτικά από την ερμηνεία της πολυφασματικής απεικόνισης, αποτελούν τροχοπέδη για την επίτευξη καλής ακρίβειας για τις δυο κατηγορίες μεμονωμένα σε κάθε πείραμα και ταυτόχρονα στα πειράματα που είναι οι εξεταζόμενες κατηγορίες -Πειράματα 5 & 8, γι’ αυτό κι εμφανίζουν σε όλα τα πειράματα μέτριες και κάτω του μετρίου ακρίβειες. Ένα ακόμη πείραμα στο οποίο δεν ανεμένετο λιγότερο καλή ακρίβεια από το αρχικό, είναι το πείραμα υποδειγματοληψίας των δεδομένων των τεχνητών επιφανειών, το οποίο ποιοτικά φαίνεται να ενισχύσει τις εκτάσεις των τεχνητών επιφανειών, χωρίς να σφάλει σε εκτάσεις πρασίνου, ενώ από τα διαγράμματα που παρουσιάζουν τη διακύμανση των μετρικών αξιολόγησης των ακριβειών των κατηγοριών ανά πείραμα, προκύπτει ότι έχει ως αποτέλεσμα τη μέτρια απόδοση όλων των κατηγοριών κάλυψης γης της περιοχής ενδιαφέροντος (εκτός της θάλασσας).

Σε κάθε άλλη περίπτωση, συγκριτικά με τη συνολική ακρίβεια που επιτυγχάνεται από το ισορροπημένο ως προς την έκταση των δεδομένων dataset σημειώνεται αύξηση της συνολικής ακρίβειας των πειραμάτων όποια μέθοδος δειγματοληψίας κι αν έχει χρησιμοποιηθεί. Μάλιστα, τα πειράματα που καταγράφουν την υψηλότερη συνολική ακρίβεια είναι 2 κι αφορούν την τυχαία υποδειγματοληψία των ετερογενών γεωργικών εκτάσεων και την τυχαία υπερδειγματοληψία των δασικών εκτάσεων. Παρόλα αυτά, επισημαίνεται για ακόμη μια φορά πως η συνολική ακρίβεια δεν είναι το κατάλληλο μέγεθος, για να κριθεί το αποτέλεσμα μιας πειραματικής εφαρμογής.

Σε συνδυασμό με τη συνολική ακρίβεια, έχει υπολογιστεί για κάθε πείραμα και ο δείκτης συμφωνίας k-hat, που παρουσιάζεται στο παρακάτω διάγραμμα.



Διάγραμμα 20. Δείκτης συμφωνίας k-hat (%) πειραματικών εφαρμογών επιβλεπόμενης ταξινόμησης με τον αλγόριθμο RF

Ένα από τα πλεονεκτήματα του υπολογισμού του k-hat είναι η δυνατότητα χρήσης αυτής της τιμής ως βάση για τον προσδιορισμό της στατιστικής σημασίας των διαφορών μεταξύ πινάκων σύγχυσης. Πιο συγκεκριμένα, σε περίπτωση σύγκρισης πινάκων σύγχυσης ,οι οποίοι εν προκειμένω έχουν δημιουργηθεί από διαφορετικές μεθόδους δειγματοληψίας σε πειράματα επιβλεπόμενης ταξινόμησης, με τέτοιου είδους τεστ να βασίζονται στον υπολογισμό εκτίμησης της μεταβλητότητας του $\hat{k}$, με σκοπό να ελεγχθεί εάν ένας συγκεκριμένος πίνακας διαφέρει σημαντικά από ένα τυχαίο αποτέλεσμα κι αν οι τιμές του $\hat{k}$ από δυο ξεχωριστούς πίνακες διαφέρουν

σημαντικά μεταξύ τους. Υπό κανονικές συνθήκες, η αξιολόγηση των αποτελεσμάτων ενός πίνακα σύγχυσης μέσω του δείκτη συμφωνίας απαιτεί και την παρεμβολή του σε μια κατανομή Z. Όμως, επειδή στην παρούσα εργασία λειτουργεί συμπληρωματικά με τη συνολική ακρίβεια, αποτελώντας μεγέθη που δεν έχουν καθοριστικό ρόλο στην αξιολόγηση των πειραμάτων, για αυτό δε γίνεται κι εκτενής σχολιασμός τους. Κατά τον υπολογισμό των δυο αυτών μεγεθών διαπιστώνεται μια απόκλιση της τάξης του 2,6%, ενώ το γεγονός πως οι διακυμάνσεις του δείκτη συμφωνίας είναι ανάλογες της συνολικής ακρίβειας, μπορεί να οδηγήσει στο συμπέρασμα πως το αποτέλεσμα των ταξινομήσεων δεν επηρεάζεται από την τυχαιότητα της μεθόδου δειγματοληψίας που εφαρμόζεται κάθε φορά.

8.3.Σύνοψη

Η επιλογή συγκεκριμένων τεχνικών δειγματοληψίας σε διαφορετικά πειράματα βασίστηκε σε διάφορους παράγοντες, όπως η φύση των δεδομένων, οι στόχοι της ανάλυσης και η απόδοση των μεθόδων σε προηγούμενες βιβλιογραφικές μελέτες που εξετάστηκαν (Ενότητα 1.1.).

Ο θεματικός χαρακτήρας των δεδομένων διαδραματίζει αποφασιστικό ρόλο στην επιλογή της κατάλληλης τεχνικής δειγματοληψίας. Για παράδειγμα:

•Τυχαία Υπερδειγματοληψία (Random Oversampling): Εφαρμόστηκε σε περιπτώσεις όπου οι μειονοτικές τάξεις είναι μικρές και υπάρχει ανάγκη αύξησης του αριθμού των δειγμάτων τους για να αποφευχθεί η προκατάληψη του ταξινομητή προς τις πλειοψηφικές τάξεις.

•Τυχαία Υποδειγματοληψία (Random Undersampling): Χρησιμοποιήθηκε όταν οι πλειοψηφικές τάξεις ήταν (έγιναν) πολύ μεγάλες, ώστε να μειωθεί ο αριθμός των δειγμάτων τους και να εξισορροπηθούν οι τάξεις χωρίς να επηρεαστεί η απόδοση του ταξινομητή.

Οι στόχοι της ανάλυσης καθόρισαν την προτεραιότητα στη μεθοδολογία δειγματοληψίας:

1. Βελτίωση της Ακρίβειας: Σε πειράματα όπου ο στόχος ήταν η βελτίωση της ακρίβειας της ταξινόμησης, επιλέχθηκαν μέθοδοι που έχουν αποδειχθεί αποδοτικές στην αντιμετώπιση της ανισορροπίας, όπως η τυχαία υπερδειγματοληψία και υποδειγματοληψία.
2. Μείωση της Μεροληψίας (προκατάληψης): Εάν ο κύριος στόχος είναι να μειωθεί η μεροληψία του ταξινομητή, μπορούν να χρησιμοποιηθούν υβριδικές μέθοδοι όπως το SMOTE σε συνδυασμό με Tomek Links, οι οποίες όχι μόνο αυξάνουν τα δείγματα των μειονοτικών τάξεων αλλά και αφαιρούν περιττά δείγματα από τις πλειοψηφικές τάξεις.

Η επιλογή συγκεκριμένων μεθόδων αυτής της εργασίας, βασίστηκε επίσης σε αποδείξεις από προηγούμενες μελέτες και βιβλιογραφία που μελετήθηκε:

I. Ευκολία Εφαρμογής: Οι τυχαίες μέθοδοι υπερδειγματοληψίας και υποδειγματοληψίας είναι απλές στην εφαρμογή και παρέχουν γρήγορα αποτελέσματα, κάτι που είναι σημαντικό σε πειραματικές εφαρμογές με περιορισμένο χρόνο και πόρους.
II. Αποτελεσματικότητα στη μεροληψία: Μέθοδοι όπως το SMOTE έχουν αποδειχθεί ιδιαίτερα αποτελεσματικές στη βελτίωση της μεροληψίας αλλά και της ακρίβειας στη γενίκευση, σε πολλές εφαρμογές μηχανικής μάθησης.

Η διαθεσιμότητα υπολογιστικών πόρων επηρέασε την επιλογή της τεχνικής που χρησιμοποιήθηκε:

- Εάν οι υπολογιστικοί πόροι είναι περιορισμένοι (όπως σε αυτήν την εργασία), οι απλούστερες μέθοδοι πρέπει να προτιμηθούν.
- Μέθοδοι όπως το SMOTE είναι πιο απαιτητικές υπολογιστικά σε σύγκριση με τις τυχαίες μεθόδους.

8.4.Προτάσεις

Στο πλαίσιο της παρούσας διπλωματικής εργασίας χρησιμοποιήθηκαν οι μέθοδοι αντιμετώπισης μη ισορροπημένων δεδομένων, τυχαία υπερδειγματοληψία και τυχαία υποδειγματοληψία. Για μελλοντική έρευνα, θα είχε ιδιαίτερο ενδιαφέρον να γίνει μια περαιτέρω διερεύνηση του πως επηρεάζεται η ακρίβεια της ταξινόμησης με μια διαφορετική μέθοδο δειγματοληψίας, πέραν της τυχαίας, όπως για παράδειγμα η μέθοδος SMOTE ή η υβριδική μέθοδος Tomek links. Αυτές οι μέθοδοι μπορεί να προσφέρουν βελτιωμένα αποτελέσματα ιδιαίτερα στη μεροληψία, αντιμετωπίζοντας καλύτερα την ανισορροπία των δεδομένων.

Επιπλέον, θα μπορούσε να δοκιμαστεί η τμηματοποίηση της περιοχής ενδιαφέροντος, ώστε να διαμορφώνεται διαφορετικά σε κάθε πείραμα και η έκταση κάθε κατηγορίας, μεταβάλλοντας με αυτόν τον τρόπο τις ισορροπίες μεταξύ πλειοψηφικής και μειονοτικής τάξης. Αυτό θα επιτρέψει μια ενδιαφέρουσα αξιολόγηση της επίδρασης της θεματικής κατανομής των δεδομένων στην ακρίβεια της ταξινόμησης.

Ακόμη, αναφορικά με τον αλγόριθμο επιβλεπόμενης ταξινόμησης που χρησιμοποιήθηκε, θα μπορούσαν να διεξαχθούν περισσότερα πειράματα που να εξετάζουν την επίδραση της παραμετροποίησης του αλγορίθμου, όπως οι παράμετροι «Minimum number to split» και «Max features». Η ανάλυση των αποτελεσμάτων με διαφορετικές τιμές αυτών των παραμέτρων είναι δυνατόν να προσφέρει σημαντικές πληροφορίες για τη βέλτιστη διαμόρφωση του αλγορίθμου.

Επιπροσθέτως, από τη βιβλιογραφία γίνεται γνωστή η χρήση διαφόρων μετρικών για την αξιολόγηση των αποτελεσμάτων ταξινόμησης, όπως οι καμπύλες ROC (Receiver Operating Characteristic). Υπολογίζονται και οπτικοποιούνται εύκολα σε ένα περιβάλλον προγραμματισμού (π.χ σε Python ή R) αλλά δυσκολότερα στο QGIS που δομήθηκε η παρούσα εργασία. Επομένως, θα ήταν σκόπιμο να εξεταστούν τέτοιες μετρικές για να ενισχυθεί διαγραμματικά το στάδιο της ποσοτικής αξιολόγησης και να παρασχεθούν πιο ολοκληρωμένες και αξιόπιστες εκτιμήσεις της απόδοσης των αλγορίθμων.

Για τις ανάγκες της εργασίας δημιουργήθηκε ο Δείκτης Αναλογίας Ισορροπίας (Balance Ratio Indice ,BRI), για την υπόδειξη πιθανής λύσης για την αντιμετώπιση της ανισορροπίας στο σύνολο δεδομένων εκπαίδευσης και τη διαδοχική επιλογή του τελικού αριθμού πολυγώνων εκπαίδευσης. Ο δείκτης αυτός βασίστηκε στο δείκτη ανισορροπίας IR (Imbalance Ratio). Λαμβάνοντας υπόψιν τους δυο αυτούς δείκτες, δύναται να γίνει επιπλέον διερεύνηση για την εύρεση τρίτου, ο οποίος θα βασίζεται στην έκταση που καλύπτουν οι κατηγορίες κάλυψης γης της περιοχής μελέτης ή στην έκταση που καλύπτουν τα πολύγωνα εκπαίδευσης των κατηγοριών κι όχι στο πλήθος τους, όπως συνέβη στην παρούσα εργασία (αν και τα πολύγωνα ήταν κατά το δυνατόν ισομεγέθη, οπότε υπήρχε άμεση αναλογία του πλήθους τους, με την έκταση που καλύπτουν).

Τέλος, αποτελεί πρόκληση η εφαρμογή των μοντέλων που έχουν δημιουργηθεί για μια συγκεκριμένη περιοχή ενδιαφέροντος, με εφαρμογή μεθόδων υπερδειγματοληψίας ή υποδειγματοληψίας, σε μια παρόμοια και σε μια αρκετά διαφορετική περιοχή. Αυτή η διαδικασία θα επιτρέψει την αξιολόγηση του πώς αυτές οι μέθοδοι επηρεάζουν τη γενίκευση της ταξινόμησης, παρέχοντας σημαντικά ευρήματα για την αποτελεσματικότητα και τη σταθερότητά τους σε διαφορετικά γεωγραφικά και θεματικά περιβάλλοντα. Με αυτόν τον τρόπο, θα εξεταστεί η ικανότητα των μεθόδων να αποδίδουν αξιόπιστα αποτελέσματα και σε άλλες περιοχές, πέραν της αρχικής περιοχής μελέτης, εξασφαλίζοντας την ευρωστία και την προσαρμοστικότητά τους.

Βιβλιογραφία

Vluymans S., 2019. Dealing with Imbalanced and Weakly Labelled Data in Machine Learning using Fuzzy and Rough Set Methods. ISBN 978-3-030-04663-7 (eBook), https://doi.org/10.1007/978-3-030-04663-7 ,pages 82–91.

Fernández A., García S., Galar M., Prati C. R., Krawczyk B., Herrera F., 2018. Learning from Imbalanced Data Sets. ISBN 978-3-319-98074-4 (eBook), https://doi.org/10.1007/978-3-319-98074-4 , pages 82-117.

Abhishek K. , Dr. Abdelaziz M., 2023. Machine Learning for Imbalanced Data – Tackle imbalanced datasets using machine learning and deep learning techniques. ISBN 978-1-80107-083-6, pages 20-23.

Dal Pozzolo A., Caelen O., Johnson A. R., Bontempi G., 2015. Calibrating Probability with Undersampling for Unbalanced Classification. IEEE Symposium Series on Computational Intelligence

Wongvorachan T., He S., Bulut O., 2023. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. Information 2023,15,54 https://doi.org/10.3390/info14010054

Hernandez J., Carrasco-Ochoa J.A. ,Martínez-Trinidad J.F., 2013. An Empirical Study of Oversampling and Undersampling for Instance Selection Methods on Imbalance Datasets.

García S., Herrera F., 2009. Evolutionary Undersampling for Classification with Imbalanced Datasets: Proposals and Taxonomy.

Vuttipittayamongkol P., Elyan E., Petrovski A. Jayne Ch., 2018. Overlap-Based Undersampling for Improving Imbalanced Data Classification.

Sanjaya S., Abdillah R., Afrianty I., 2022. The impact of Under-Sampling Techniques on Classification Accuracy in multi-class Imbalance Data. ISBN 978-1-6654-5434-6.

Φιλανδριανού Χ., 2022. Εμπλουτισμός δεδομένων μέσω δημιουργικών νευρωνικών δικτύων για την ανίχνευση κακόβουλων ηλεκτρονικών συναλλαγών. Διπλωματική Εργασία, Εθνικό Μετσόβιο Πολυτεχνείο, σελ. 36-42.

Δρόσου Π. Κ., 2015. Μέθοδοι για την Ταξινόμηση μη Ισορροπημένων Δεδομένων με Μηχανές Διανυσματική Υποστήριξης, Μεταπτυχιακή Διπλωματική Εργασία, Εθνικό Μετσόβιο Πολυτεχνείο, σελ. 25-112.

Σταματάκης Αργ., 2022. Τεχνικές διαχείρισης μη ισορροπημένων συνόλων δεδομένων δυαδικής κατηγοριοποίησης στη μηχανική μάθηση. Μεταπτυχιακή Διπλωματική Εργασία, Εθνικό Μετσόβιο Πολυτεχνείο, σελ. 52-91.

Νάι Μ., 2019. Επιβλεπόμενη Μηχανική Μάθηση και το Πρόβλημα της Ταξινόμησης. Διπλωματική Εργασία ,Εθνικό Μετσόβιο Πολυτεχνείο.

Αλεξάνδρου Κ., 2021. Πρόβλεψη Καθυστέρησης Πτήσεων με Τεχνικές Μηχανικής Μάθησης. Διπλωματική Εργασία, Εθνικό Μετσόβιο Πολυτεχνείο, σελ. 30-34 & 42-50.

towardsdatascience.com. (χ.χ.). 7 Over Sampling techniques to handle Imbalanced Data - Deep dive analysis of various oversampling techniques. Ανάκτηση από

towardsdatascience.com: https://towardsdatascience.com/7-over-sampling-techniques-to-handle-imbalanced-data-ec51c8db349f

machinelearningmastery.com. (χ.χ.). Undersampling Algorithms for Imbalanced Classification. Ανάκτηση από machinelearningmastery.com: https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/

sciencedirect.com. (χ.χ.). Class Imbalance Problem - Multi-class imbalance problems involve imbalances between multiple classes, with some classes having fewer instances. Ανάκτηση από sciencedirect.com: https://www.sciencedirect.com/topics/computer-science/class-imbalance-problem

machinelearningmastery.com. (χ.χ.). A Gentle Introduction to Imbalanced Classification. Ανάκτηση από machinelearningmastery.com: https://machinelearningmastery.com/what-is-imbalanced-classification/

simplilearn.com. (χ.χ.). An Overview on Multilayer Perceptron (MLP). Ανάκτηση από simplilearn.com: https://www.simplilearn.com/tutorials/deep-learning-tutorial/multilayer-perceptron

towardsdatascience.com. (χ.χ.). Perceptron Learning Algorithm: A Graphical Explanation Of Why It Works. Ανάκτηση από towardsdatascience.com: https://towardsdatascience.com/perceptron-learning-algorithm-d5db0deab975

edu.ellak.gr. (χ.χ.). Νευρωνικά δίκτυα και μηχανική μάθηση. Ανάκτηση από edu.ellak.gr: https://edu.ellak.gr/2023/04/11/nevronika-diktia-ke-michaniki-mathisi/

ΠΑΡΑΡΤΗΜΑ Α

AREA BASED ERROR MATRIX											
	> Reference										
V_Classified	1	2	3	4	5	6	7	8	9	Area	Wi
1	0.1144	0	0	0	0	0	0	0	0	71900	0.1144
2	0	0.0690	0	0	0.0205	0	0	0	0	56300	0.0896
3	0	0.0021	0.1055	0.0099	0.0005	0	0.0016	0	0	75100	0.1195
4	0	0	0.0025	0.0436	0.0035	0.0118	0.0398	0	0	63600	0.1012
5	0	0.0396	0	0	0.0865	0	0	0	0	79300	0.1262
6	0	0	0	0.0059	0	0.0883	0.0072	0	0	63700	0.1013
7	0	0.0002	0.0006	0.0504	0	0.0119	0.0603	0	0	77600	0.1234
8	0	0.0010	0.0013	0	0	0	0.0010	0.1112	0	71900	0.1144
9	0	0	0	0	0	0	0	0	0.1101	69200	0.1101
Total	0.1144	0.1118	0.1099	0.1098	0.1110	0.1120	0.1098	0.1112	0.1101	628600	1
Estimated area	71900	70300	69100	69000	69800	70400	69000	69900	69200	628600	
SE	0	0.0027	0.0016	0.0033	0.0027	0.0023	0.0032	0.0007	0		
SE area	0	1704.11	1026.26	2087.61	1717.31	1431.39	2009.58	441.26	0		
95% CI area	0	3340.06	2011.47	4091.71	3365.92	2805.53	3938.77	864.86	0		
PA [%]	1.000.000	617.354	959.479	397.101	779.370	788.352	549.275	1.000.000	1.000.000		
UA [%]	1.000.000	770.870	882.823	430.818	686.003	871.272	488.402	972.184	1.000.000		

Πίνακας 42. Πίνακας σύγχυσης βάσει εμβαδού εφαρμογής επιβλεπόμενης ταξινόμησης με ισορροπημένο dataset

AREA BASED ERROR MATRIX											
	> Reference										
V_Classified	1	2	3	4	5	6	7	8	9	Area	Wi
1	0.1144	0	0	0	0	0	0	0	0	71900	0.1144
2	0	0.0694	0	0	0.0196	0	0	0	0	55900	0.0889
3	0	0.0019	0.1052	0.0097	0.0003	0	0.0017	0	0	74700	0.1188
4	0	0.0005	0.0025	0.0425	0.0033	0.0124	0.0382	0	0	62500	0.0994
5	0	0.0388	0	0	0.0878	0	0	0	0	79600	0.1266
6	0	0	0	0.0060	0	0.0875	0.0081	0	0	63900	0.1017
7	0	0.0003	0.0005	0.0515	0	0.0121	0.0611	0	0	78900	0.1255
8	0	0.0010	0.0017	0	0	0	0.0006	0.1112	0	72000	0.1145
9	0	0	0	0	0	0	0	0	0.1101	69200	0.1101
Total	0.1144	0.1118	0.1099	0.1098	0.1110	0.1120	0.1098	0.1112	0.1101	628600	1
Estimated area	71900	70300	69100	69000	69800	70400	69000	69900	69200	628600	
SE	0	0.0027	0.0016	0.0033	0.0027	0.0023	0.0032	0.0007	0		
SE area	0	1697.93	1027.89	2088.70	1696.62	1462.41	2018.11	451.84	0		
95% CI area	0	3327.94	2014.67	4093.85	3325.37	2866.33	3955.49	885.60	0		
PA [%]	1.000.000	620.199	956.585	386.957	790.831	781.250	556.522	1.000.000	1.000.000		
UA [%]	1.000.000	779.964	884.873	427.200	693.467	860.720	486.692	970.833	1.000.000		

Πίνακας 43. Πίνακας σύγχυσης βάσει εμβαδού εφαρμογής επιβλεπόμενης ταξινόμησης με υποδειγματοληψία των τεχνητών επιφανειών

AREA BASED ERROR MATRIX											
	> Reference										
V_Classified	1	2	3	4	5	6	7	8	9	Area	Wi
1	0.1144	0	0	0	0	0	0	0	0	71900	0.1144
2	0	0.0700	0	0	0.0188	0	0	0	0	55800	0.0888
3	0	0.0021	0.1048	0.0097	0.0003	0	0.0017	0	0	74600	0.1187
4	0	0.0002	0.0025	0.0433	0.0021	0.0121	0.0391	0	0	62400	0.0993
5	0	0.0385	0	0	0.0899	0.0025	0	0	0	82300	0.1309
6	0	0	0	0.0060	0	0.0846	0.0070	0	0	61400	0.0977
7	0	0.0002	0.0006	0.0507	0	0.0127	0.0609	0	0	78700	0.1252
8	0	0.0010	0.0019	0	0	0	0.0010	0.1112	0	72300	0.1150
9	0	0	0	0	0	0	0	0	0.1101	69200	0.1101
Total	0.1144	0.1118	0.1099	0.1098	0.1110	0.1120	0.1098	0.1112	0.1101	628600	1
Estimated area	71900	70300	69100	69000	69800	70400	69000	69900	69200	628600	
SE	0	0.0027	0.0017	0.0033	0.0027	0.0024	0.0032	0.0008	0		
SE area	0	1688.23	1041.03	2086.66	1689.07	1502.27	2009.58	482.03	0		
95% CI area	0	3308.93	2040.41	4089.85	3310.58	2944.46	3938.78	944.78	0		
PA [%]	1.000.000	625.889	953.690	394.203	809.456	755.682	555.072	1.000.000	1.000.000		
UA [%]	1.000.000	788.530	883.378	435.897	686.513	866.450	486.658	966.805	1.000.000		

Πίνακας 44. Πίνακας σύγχυσης βάσει εμβαδού εφαρμογής επιβλεπόμενης ταξινόμησης με υποδειγματοληψία της αρόσιμης γης

AREA BASED ERROR MATRIX											
	> Reference										
V_Classified	1	2	3	4	5	6	7	8	9	Area	Wi
1	0.1144	0	0	0	0	0	0	0	0	71900	0.1144
2	0	0.0695	0	0	0.0191	0	0	0	0	55700	0.0886
3	0	0.0019	0.1052	0.0092	0.0003	0	0.0019	0	0	74500	0.1185
4	0	0	0.0024	0.0437	0.0033	0.0126	0.0387	0	0	63300	0.1007
5	0	0.0391	0	0	0.0883	0	0	0	0	80100	0.1274
6	0	0	0	0.0057	0	0.0880	0.0081	0	0	64000	0.1018
7	0	0.0003	0.0006	0.0511	0	0.0115	0.0603	0	0	77800	0.1238
8	0	0.0010	0.0017	0	0	0	0.0008	0.1112	0	72100	0.1147
9	0	0	0	0	0	0	0	0	0.1101	69200	0.1101
Total	0.1144	0.1118	0.1099	0.1098	0.1110	0.1120	0.1098	0.1112	0.1101	628600	1
Estimated area	71900	70300	69100	69000	69800	70400	69000	69900	69200	628600	
SE	0	0.0027	0.0016	0.0033	0.0027	0.0023	0.0032	0.0007	0		
SE area	0	1687.48	1020.43	2078.84	1695.03	1449.01	2020.84	462.15	0		
95% CI area	0	3307.45	2000.05	4074.53	3322.25	2840.05	3960.84	905.82	0		
PA [%]	1.000.000	621.622	956.585	398.551	795.129	785.511	549.275	1.000.000	1.000.000		
UA [%]	1.000.000	784.560	887.248	434.439	692.884	864.062	487.147	969.487	1.000.000		

Πίνακας 45. Πίνακας σύγχυσης βάσει εμβαδού εφαρμογής επιβλεπόμενης ταξινόμησης με συνδυασμό υποδειγματοληψίας & υπερδειγματοληψίας για τις μόνιμες καλλιέργειες

AREA BASED ERROR MATRIX											
	> Reference										
V_Classified	1	2	3	4	5	6	7	8	9	Area	Wi
1	0.1144	0	0	0	0	0	0	0	0	71900	0.1144
2	0	0.0686	0	0	0.0205	0	0	0	0	56000	0.0891
3	0	0.0017	0.1050	0.0091	0.0003	0	0.0017	0	0	74100	0.1179
4	0	0.0006	0.0025	0.0431	0.0030	0.0130	0.0382	0	0	63200	0.1005
5	0	0.0396	0	0	0.0872	0	0	0	0	79700	0.1268
6	0	0	0	0.0059	0	0.0870	0.0080	0	0	63400	0.1009
7	0	0.0003	0.0006	0.0517	0	0.0119	0.0612	0	0	79100	0.1258
8	0	0.0010	0.0017	0	0	0	0.0006	0.1112	0	72000	0.1145
9	0	0	0	0	0	0	0	0	0.1101	69200	0.1101
Total	0.1144	0.1118	0.1099	0.1098	0.1110	0.1120	0.1098	0.1112	0.1101	628600	1
Estimated area	71900	70300	69100	69000	69800	70400	69000	69900	69200	628600	
SE	0	0.0027	0.0016	0.0033	0.0027	0.0023	0.0032	0.0007	0		
SE area	0	1713.69	1013.38	2084.06	1706.87	1465.02	2019.71	451.84	0		
95% CI area	0	3358.82	1986.23	4084.77	3345.46	2871.44	3958.63	885.60	0		
PA [%]	1.000.000	613.087	955.137	392.754	785.100	776.989	557.971	1.000.000	1.000.000		
UA [%]	1.000.000	769.643	890.688	428.797	687.578	862.776	486.726	970.833	1.000.000		

Πίνακας 46. Πίνακας σύγχυσης βάσει εμβαδού εφαρμογής επιβλεπόμενης ταξινόμησης υπερδειγματοληψίας των λιβαδιών

AREA BASED ERROR MATRIX											
	> Reference										
V_Classified	1	2	3	4	5	6	7	8	9	Area	Wi
1	0.1144	0	0	0	0	0	0	0	0	71900	0.1144
2	0	0.0710	0	0	0.0196	0	0	0	0	56900	0.0905
3	0	0.0019	0.1052	0.0092	0.0003	0	0.0017	0	0	74400	0.1184
4	0	0	0.0024	0.0428	0.0049	0.0115	0.0383	0	0	62800	0.0999
5	0	0.0377	0	0	0.0862	0	0	0	0	77900	0.1239
6	0	0	0	0.0057	0	0.0869	0.0076	0	0	63000	0.1002
7	0	0.0003	0.0006	0.0520	0	0.0130	0.0611	0	0	79900	0.1271
8	0	0.0010	0.0017	0	0	0.0006	0.0010	0.1112	0	72600	0.1155
9	0	0	0	0	0	0	0	0	0.1101	69200	0.1101
Total	0.1144	0.1118	0.1099	0.1098	0.1110	0.1120	0.1098	0.1112	0.1101	628600	1
Estimated area	71900	70300	69100	69000	69800	70400	69000	69900	69200	628600	
SE	0	0.0027	0.0016	0.0033	0.0027	0.0023	0.0032	0.0008	0		
SE area	0	1677.70	1016.55	2085.62	1712.32	1464.32	2024.04	510.21	0		
95% CI area	0	3288.29	1992.45	4087.82	3356.14	2870.06	3967.12	1000.02	0		
PA [%]	1.000.000	634.424	956.585	389.855	776.504	775.568	556.522	1.000.000	1.000.000		
UA [%]	1.000.000	783.831	888.441	428.344	695.764	866.667	480.601	962.810	1.000.000		

Πίνακας 47. Πίνακας σύγχυσης βάσει εμβαδού εφαρμογής επιβλεπόμενης ταξινόμησης με συνδυασμό υποδειγματοληψίας & υπερδειγματοληψίας για τις ετερογενείς καλλιέργειες

AREA BASED ERROR MATRIX											
> Reference											
V_Classified	1	2	3	4	5	6	7	8	9	Area	Wi
1	0.1144	0	0	0	0	0	0	0	0	71900	0.1144
2	0	0.0703	0	0	0.0188	0	0	0	0	56000	0.0891
3	0	0.0019	0.1052	0.0107	0.0005	0	0.0017	0	0	75400	0.1199
4	0	0.0011	0.0030	0.0423	0.0029	0.0116	0.0391	0	0	62900	0.1001
5	0	0.0374	0.0002	0	0.0889	0	0	0	0	79500	0.1265
6	0	0	0	0.0060	0	0.0888	0.0076	0	0	64400	0.1024
7	0	0.0002	0.0005	0.0507	0	0.0116	0.0603	0	0	77500	0.1233
8	0	0.0010	0.0011	0	0	0	0.0010	0.1112	0	71800	0.1142
9	0	0	0	0	0	0	0	0	0.1101	69200	0.1101
Total	0.1144	0.1118	0.1099	0.1098	0.1110	0.1120	0.1098	0.1112	0.1101	628600	1
Estimated area	71900	70300	69100	69000	69800	70400	69000	69900	69200	628600	
SE	0	0.0027	0.0017	0.0033	0.0027	0.0023	0.0032	0.0007	0		
SE area	0	1687.49	1053.69	2094.36	1673.20	1433.43	2013.06	430.38	0		
95% CI area	0	3307.48	2065.23	4104.94	3279.46	2809.52	3945.59	843.55	0		
PA [%]	1.000.000	628.734	956.585	385.507	800.860	792.614	549.275	1.000.000	1.000.000		
UA [%]	1.000.000	789.286	876.658	422.893	703.145	866.460	489.032	973.538	1.000.000		

Πίνακας 48. Πίνακας σύγχυσης βάσει εμβαδού εφαρμογής επιβλεπόμενης ταξινόμησης υπερδειγματοληψίας των δασικών εκτάσεων

AREA BASED ERROR MATRIX											
> Reference											
V_Classified	1	2	3	4	5	6	7	8	9	Area	Wi
1	0.1144	0	0	0	0	0	0	0	0	71900	0.1144
2	0	0.0695	0	0	0.0204	0	0	0	0	56500	0.0899
3	0	0.0022	0.1052	0.0086	0.0003	0	0.0017	0	0	74200	0.1180
4	0	0	0.0022	0.0436	0.0049	0.0121	0.0390	0	0	64000	0.1018
5	0	0.0391	0	0	0.0854	0.0002	0	0	0	78400	0.1247
6	0	0	0	0.0057	0	0.0880	0.0078	0	0	63800	0.1015
7	0	0	0.0008	0.0519	0	0.0115	0.0603	0	0	78200	0.1244
8	0	0.0010	0.0017	0	0	0.0003	0.0010	0.1112	0	72400	0.1152
9	0	0	0	0	0	0	0	0	0.1101	69200	0.1101
Total	0.1144	0.1118	0.1099	0.1098	0.1110	0.1120	0.1098	0.1112	0.1101	628600	1
Estimated area	71900	70300	69100	69000	69800	70400	69000	69900	69200	628600	
SE	0	0.0027	0.0016	0.0033	0.0028	0.0023	0.0032	0.0008	0		
SE area	0	1696.81	1008.92	2077.11	1732.39	1446.80	2022.70	491.63	0		
95% CI area	0	3325.75	1977.49	4071.14	3395.49	2835.73	3964.49	963.60	0		
PA [%]	1.000.000	621.622	956.585	397.101	769.341	785.511	549.275	1.000.000	1.000.000		
UA [%]	1.000.000	773.451	890.836	428.125	684.949	866.771	484.655	965.470	1.000.000		

Πίνακας 49. Πίνακας σύγχυσης βάσει εμβαδού εφαρμογής επιβλεπόμενης ταξινόμησης υποδειγματοληψίας των συνδυασμών βλάστησης

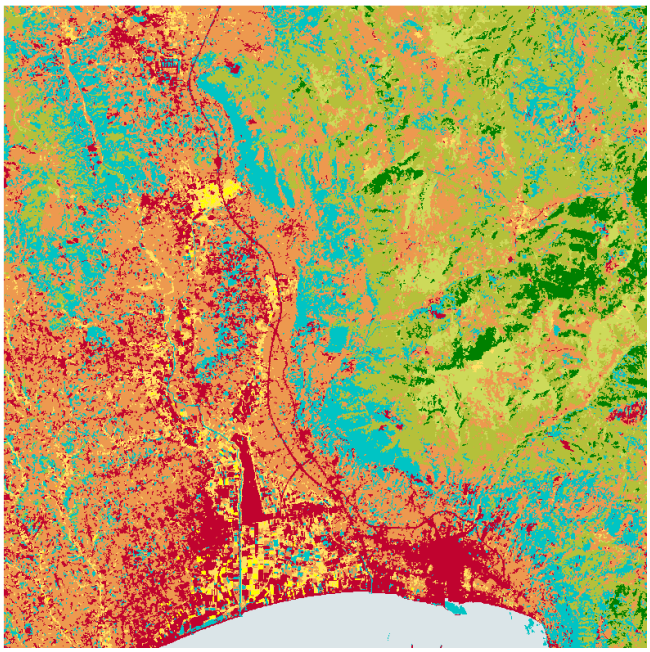
AREA BASED ERROR MATRIX											
> Reference											
V_Classified	1	2	3	4	5	6	7	8	9	Area	Wi
1	0.1144	0	0	0	0	0	0	0	0	71900	0.1144
2	0	0.0667	0	0	0.0172	0	0	0	0	52700	0.0838
3	0	0.0019	0.1052	0.0107	0.0006	0	0.0017	0	0	75500	0.1201
4	0	0.0005	0.0022	0.0425	0.0032	0.0121	0.0387	0	0	62300	0.0991
5	0	0.0415	0	0	0.0900	0	0	0	0	82700	0.1316
6	0	0	0	0.0037	0	0.0862	0.0052	0	0	59800	0.0951
7	0	0.0003	0.0006	0.0530	0	0.0137	0.0632	0	0	82200	0.1308
8	0	0.0010	0.0019	0	0	0	0.0010	0.1112	0	72300	0.1150
9	0	0	0	0	0	0	0	0	0.1101	69200	0.1101
Total	0.1144	0.1118	0.1099	0.1098	0.1110	0.1120	0.1098	0.1112	0.1101	628600	1
Estimated area	71900	70300	69100	69000	69800	70400	69000	69900	69200	628600	
SE	0	0.0027	0.0017	0.0033	0.0027	0.0022	0.0032	0.0008	0		
SE area	0	1696.02	1057.90	2084.33	1697.79	1395.61	2005.02	482.03	0		
95% CI area	0	3324.19	2073.49	4085.29	3327.67	2735.40	3929.85	944.78	0		
PA [%]	1.000.000	596.017	956.585	386.957	810.888	769.886	575.362	1.000.000	1.000.000		
UA [%]	1.000.000	795.066	875.497	428.571	684.401	906.355	482.968	966.805	1.000.000		

Πίνακας 50. Πίνακας σύγχυσης βάσει εμβαδού εφαρμογής επιβλεπόμενης ταξινόμησης υποδειγματοληψίας όλων των κατηγοριών, πλην των χερσαίων υδάτων

AREA BASED ERROR MATRIX											
	> Reference										
V_Classified	1	2	3	4	5	6	7	8	9	Area	Wi
1	0.1144	0	0	0	0	0	0	0	0	71900	0.1144
2	0	0.0690	0	0	0.0189	0	0	0	0	55300	0.0880
3	0	0.0019	0.1050	0.0097	0.0003	0	0.0019	0	0	74700	0.1188
4	0	0	0.0025	0.0428	0.0038	0.0119	0.0385	0	0	62600	0.0996
5	0	0.0396	0	0	0.0880	0	0	0	0	80200	0.1276
6	0	0	0	0.0059	0	0.0875	0.0075	0	0	63400	0.1009
7	0	0.0003	0.0006	0.0514	0	0.0126	0.0609	0	0	79100	0.1258
8	0	0.0010	0.0017	0	0	0	0.0010	0.1112	0	72200	0.1149
9	0	0	0	0	0	0	0	0	0.1101	69200	0.1101
Total	0.1144	0.1118	0.1099	0.1098	0.1110	0.1120	0.1098	0.1112	0.1101	628600	1
Estimated area	71900	70300	69100	69000	69800	70400	69000	69900	69200	628600	
SE	0	0.0027	0.0016	0.0033	0.0027	0.0023	0.0032	0.0008	0		
SE area	0	1688.93	1036.43	2088.04	1704.65	1450.19	2019.54	472.21	0		
95% CI area	0	3310.29	2031.40	4092.55	3341.11	2842.38	3958.29	925.53	0		
PA [%]	1.000.000	617.354	955.137	389.855	792.264	781.250	555.072	1.000.000	1.000.000		
UA [%]	1.000.000	784.810	883.534	429.712	689.526	867.508	484.197	968.144	1.000.000		

Πίνακας 51. Πίνακας σύγχυσης βάσει εμβαδού εφαρμογής επιβλεπόμενης ταξινόμησης υπερδειγματοληψίας της θάλασσας

ΠΑΡΑΡΤΗΜΑ Β



Σχήμα 101. Αποτέλεσμα ταξινόμησης Random Forest, με 100 δέντρα, πολυφασματικής απεικόνισης Sentinel-2 ημερομηνίας λήψης 17/07/2023, με ισορροπημένο σύνολο εκπαίδευσης ως προς το πλήθος των πολυγώνων (45 πολύγωνα/κατηγορία)

V_Classified	Τεχνητές	Αρόσιμη	Μόνιμες	Λιβάδια	Ετερογενείς	Δάση	Συνδυασμοί	Χερσαία	Θάλασσα	Total
Τεχνητές	**719**	0	0	0	0	0	0	0	0	**719**
Αρόσιμη	0	**440**	0	0	103	0	0	0	0	**543**
Μόνιμες	0	13	**654**	34	4	0	7	0	0	**712**
Λιβάδια	0	1	14	**262**	22	70	240	0	0	**609**
Ετερογενείς	0	243	0	0	**569**	0	0	0	0	**812**
Δάση	0	0	0	34	0	**565**	50	0	0	**649**
Συνδυασμοί	0	0	14	360	0	69	**387**	0	0	**830**
Χερσαία	0	6	9	0	0	0	6	**699**	0	**720**
Θάλασσα	0	0	0	0	0	0	0	0	**692**	**692**
Total	**719**	**703**	**691**	**690**	**698**	**704**	**690**	**699**	**692**	**6286**

Πίνακας 52. Πίνακας σύγχυσης ταξινόμησης πολυφασματικής απεικόνισης με ημερομηνία λήψης 17/07/2023 με χρήση του αλγορίθμου Random Forest, με ισορροπημένο σύνολο εκπαίδευσης ως προς το πλήθος των πολυγώνων (45 πολύγωνα/κατηγορία)

	Precision	Recall	F1-score	Συνολική ακρίβεια (%)	K-hat (%)
Τεχνητές	1	1	1	79,34	76,75
Αρόσιμη	1	0,6259	0,7063		
Μόνιμες	0,9185	0,9465	0,9323		
Λιβάδια	0,4302	0,3797	0,4034		
Ετερογενείς	0,7007	0,8152	0,7536		
Δάση	0,8706	0,8026	0,8352		
Συνδυασμοί	0,4663	0,5609	0,5092		
Χερσαία	0,9708	1	0,9852		
Θάλασσα	1	1	1		

Πίνακας 53. Μετρικές αξιολόγησης της ταξινόμησης της πολυφασματικής απεικόνισης με τον αλγόριθμο Random Forest, με ισορροπημένο σύνολο εκπαίδευσης ως προς το πλήθος των πολυγώνων (45 πολύγωνα/κατηγορία)

AREA BASED ERROR MATRIX											
	> Reference										
V_Classified	1	2	3	4	5	6	7	8	9	Area	Wi
1	0.1144	0	0	0	0	0	0	0	0	71900	0.1144
2	0	0.0700	0	0	0.0164	0	0	0	0	54300	0.0864
3	0	0.0021	0.1040	0.0054	0.0006	0	0.0011	0	0	71200	0.1133
4	0	0.0002	0.0022	0.0417	0.0035	0.0111	0.0382	0	0	60900	0.0969
5	0	0.0387	0	0	0.0905	0	0	0	0	81200	0.1292
6	0	0	0	0.0054	0	0.0899	0.0080	0	0	64900	0.1032
7	0	0	0.0022	0.0573	0	0.0110	0.0616	0	0	83000	0.1320
8	0	0.0010	0.0014	0	0	0	0.0010	0.1112	0	72000	0.1145
9	0	0	0	0	0	0	0	0	0.1101	69200	0.1101
Total	0.1144	0.1118	0.1099	0.1098	0.1110	0.1120	0.1098	0.1112	0.1101	628600	1
Estimated area	71900	70300	69100	69000	69800	70400	69000	69900	69200	628600	
SE	0	0.0026	0.0015	0.0033	0.0027	0.0022	0.0032	0.0007	0		
SE area	0	1654.83	947.27	2045.29	1671.31	1409.39	2028.75	451.84	0		
95% CI area	0	3243.46	1856.64	4008.77	3275.78	2762.40	3976.34	885.60	0		
PA [%]	1.000.000	625.889	946.454	379.710	815.186	802.557	560.870	1.000.000	1.000.000		
UA [%]	1.000.000	810.313	918.539	430.213	700.739	870.570	466.265	970.833	1.000.000		

Πίνακας 54. Πίνακας σύγχυσης βάσει εμβαδού εφαρμογής επιβλεπόμενης ταξινόμησης με ισορροπημένο σύνολο εκπαίδευσης ως προς το πλήθος των πολυγώνων (45 πολύγωνα/κατηγορία)