



Τεχνολογικό  
Πανεπιστήμιο  
Κύπρου

Σχολή Μηχανικής και  
Τεχνολογίας

**Πτυχιακή εργασία**

**Αξιολόγηση της απόδοσης και ευαισθησίας του ταξινομητή  
Random Forest για τη δημιουργία θεματικών χαρτών  
κάλυψης/χρήσης γης, με τη χρήση δεδομένων Sentinel-2 και  
Landsat-8**

**Σταύρος Πατσαλίδης**

**Λεμεσός, Απρίλιος 2018**



ΤΕΧΝΟΛΟΓΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ  
ΣΧΟΛΗ ΜΗΧΑΝΙΚΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ  
ΤΜΗΜΑ ΠΟΛΙΤΙΚΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ  
ΓΕΩΠΛΗΡΟΦΟΡΙΚΗΣ

Πτυχιακή εργασία

Αξιολόγηση της απόδοσης και ευαισθησίας του ταξινομητή  
Random Forest για τη δημιουργία θεματικών χαρτών  
κάλυψης/χρήσης γης, με τη χρήση δεδομένων Sentinel-2 και  
Landsat-8

του

Σταύρου Πατσαλίδη

Δρ. Διόφαντος Χατζημιτσής

Δρ. Άθως Αγαπίου

Λεμεσός, Απρίλιος 2018

## **Πνευματικά δικαιώματα**

Copyright © Σταύρος Πατσαλίδης, 2018

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Η έγκριση της πτυχιακής εργασίας από το Τμήμα Πολιτικών Μηχανικών και Μηχανικών Γεωπληροφορικής του Τεχνολογικού Πανεπιστημίου Κύπρου δεν υποδηλώνει απαραίτητως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

Θα ήθελα να ευχαριστήσω τους επιβλέποντες καθηγητές Δρ. Διόφαντο Χατζημιτσή και Δρ. Άθω Αγαπίου, για την εμπιστοσύνη που μου έδειξαν στην περάτωση του δύσκολου αυτού εγχειρήματος. Η καθοδήγηση και οι χρήσιμες συμβουλές τους, έπαιξαν καθοριστικό ρόλο στην συγγραφή της παρούσας μελέτης.

Ακόμα, ένα μεγάλο ευχαριστώ στην οικογένεια μου, που ήταν πάντα δίπλα μου στηρίζοντας με σε αυτό το ιδιαίτερα πιεστικό στάδιο της ζωής μου.

## ΠΕΡΙΛΗΨΗ

Η δωρεάν διαθεσιμότητα της νέας γενιάς δεδομένων Sentinel-2, προσφέρει νέες ευκαιρίες σε εφαρμογές παρακολούθησης της γης. Συνδυαστικοί ταξινομητές όπως τα Τυχαία Δάση (Random Forest), παρουσιάζουν σημαντικές δυνατότητες στην επεξεργασία των πολυφασματικών αυτών δεδομένων και την παραγωγή θεματικών χαρτών κάλυψης γης.

Ο κύριος στόχος αυτής της μελέτης, είναι η αξιολόγηση της απόδοσης και της ευαισθησίας του εν λόγω ταξινομητή στην Ανατολική Μεσόγειο, βάση των εξής κριτηρίων: τον αριθμό των δέντρων απόφασης που απαρτίζουν το δάσος, το μέγεθος του συνόλου δεδομένων που χρησιμοποιούνται για την εκπαίδευση του αλγορίθμου καθώς και το πλήθος των μεταβλητών που επιλέγονται και ελέγχονται για την εύρεση του βέλτιστου τρόπου διαχωρισμού των δεδομένων κατά την ανάπτυξη του κάθε δέντρου. Τα αποτελέσματα συγκρίνονται με ακόμα τρεις συμβατικές, παραμετρικές τεχνικές ταξινόμησης όπως η Μέγιστη Πιθανοφάνεια, η Ελάχιστη απόσταση και η απόσταση Mahalanobis, ως προς τα ίδια δεδομένα εκπαίδευσης.

Συγκεκριμένα, εφαρμόστηκαν δύο διαφορετικές στρατηγικές ταξινόμησης: η πρώτη χρησιμοποιώντας άνισο και τυχαία επιλεγμένο πλήθος δεδομένων εκπαίδευσης για καθεμία από τις 11 ομάδες κάλυψης γης και η δεύτερη με τη χρήση ίσου αριθμού δεδομένων εκπαίδευσης για όλες τις κλάσεις. Τα αποτελέσματα και στις δύο περιπτώσεις ήταν παρόμοια (< 2% διαφορά) και δείχνουν ότι ο αλγόριθμος Random Forest υπερέχει σε απόδοση συγκριτικά με τους υπόλοιπους ταξινομητές, παρουσιάζοντας υψηλά ποσοστά Ολικής Ακρίβειας (90.27 %) και δείκτη Kappa (89,11%).

Ακόμα, κρίθηκε αναγκαίο να εκτιμηθεί το κατά πόσο τα δεδομένα Sentinel-2, τα οποία θεωρούνται ως η συνέχεια της μακροβιότερης συλλογής δορυφορικών δεδομένων Landsat, μπορούν να παράγουν όμοια τελικά προϊόντα. Για αυτό το σκοπό, εφαρμόστηκαν οι ίδιες τεχνικές και δεδομένα εκπαίδευσης σε δεδομένα του αισθητήρα OLI από τον δορυφόρο Landsat 8. Τα αποτελέσματα δείχνουν ότι παρόλη τη φασματική ομοιότητα που υπάρχει μεταξύ τους, παρατηρείται μια διαφορά της τάξης του Ο.Α.: 2.5 % και 3 % για τον δείκτη Kappa, που γέρνει υπέρ των δεδομένων Landsat. Αυτή η διαφορά, οφείλεται αφενός στη μικρή διαφορά που υπάρχει μεταξύ της φασματικής ευαισθησίας των δύο και αφετέρου στη μεταβολή των ραδιομετρικών τιμών από την

επανασύσταση που εφαρμόστηκε στα δεδομένα Sentinel, για να είναι χωρικά συγκρίσιμα με αυτά του Landsat.

**Λέξεις κλειδιά:** Τυχαία Δάση, Sentinel 2, Landsat 8, Ταξινόμηση

## **ABSTRACT**

Free availability of new generation Sentinel-2 data, offers new opportunities for land monitoring applications. Ensemble classifiers such as Random Forest, have great potential in processing this multispectral data and producing land cover thematic maps.

The main objective of this study is to evaluate the efficiency and sensitivity of this classifier in the Eastern Mediterranean, based on the following criteria: the number of decision trees that make up the forest, the size of the dataset used to train the algorithm as well as the number of variables selected and tested to find the optimal way of splitting the data when growing the trees. The classification results are also compared with three conventional, parametric classification techniques such as Maximum Likelihood, Minimum Distance and Mahalanobis Distance, with respect to the same training data.

Specifically, two different classification strategies were applied: the first using an uneven and randomly selected set of training data for each one of the 11 land cover classes and the second using an equal number of training data for all classes. The results in both cases were similar (< 2% difference) and show that Random Forest algorithm outperforms the other classifiers, presenting high Overall Accuracy (90.27%) and Kappa Index (89.11%).

Furthermore, it was considered necessary to assess whether Sentinel-2 data, which were developed to be the sequel to the longest acquired collection of satellite data Landsat, can produce similar end-products. For this purpose, same training techniques and data were applied to OLI sensor data from the Landsat 8 satellite. The results show that despite the spectral similarity between them, there is a difference of about O.A.: 2.5% and 3% for the Kappa index, leaning towards Landsat data, due both to the small difference between the spectral sensitivity of the two and to the radiometric values distortion from the resampling applied to the Sentinel data in order to be spatially comparable to those of Landsat.

**Keywords:** Random Forest, Sentinel 2, Landsat 8, Classification





## ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

|   |       |
|---|-------|
| ΠΕΡΙΛΗΨΗ.....   | v     |
| ABSTRACT.....   | vii   |
| ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ .....  | ix    |
| ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ .....   | xi    |
| ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ .....  | xiii  |
| ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ .....  | xviii |
| ΑΠΟΔΟΣΗ ΟΡΩΝ .....  | xix   |
| 1 Εισαγωγή .....  | 1     |
| 1.1 Σκοπός εργασίας και σύνοψη μεθοδολογίας.....                          | 2     |
| 1.2 Η χαρτογράφηση κάλυψης γης με τη χρήση δορυφορικών δεδομένων .....    | 2     |
| 2 Βιβλιογραφική ανασκόπηση.....   | 4     |
| 2.1 Ταξινόμηση .....  | 4     |
| 2.1.1 Τα στάδια της επιβλεπόμενης ταξινόμησης.....                        | 6     |
| 2.1.2 Τεχνικές ταξινόμησης .....  | 19    |
| 2.2 Παραμετρικοί αλγόριθμοι ταξινόμησης.....                              | 21    |
| 2.2.1 Αλγόριθμος Ελάχιστης Απόστασης.....                                 | 22    |
| 2.2.2 Αλγόριθμος Απόστασης Mahalanobis .....                              | 23    |
| 2.2.3 Αλγόριθμος Μέγιστης Πιθανοφάνειας .....                             | 24    |
| 2.2.4 Δέντρα Απόφασης.....  | 26    |
| 2.3 Συνδυαστικοί ταξινομητές στην Ταξινόμηση πολυφασματικών εικόνων ..... | 28    |
| 2.3.1 Τυχαία Δάση (Random Forest).....                                    | 30    |
| 2.3.1.1 Κατασκευή του αλγόριθμου Random Forest .....                      | 30    |
| 2.3.1.2 Ευαισθησία του ταξινομητή στα δείγματα εκπαίδευσης.....           | 33    |
| 3 Περιγραφή των δεδομένων .....   | 35    |

|       |  |    |
|-------|--|----|
| 3.1   | Δεδομένα Sentinel-2 .....                                    | 35 |
| 3.2   | Δεδομένα Landsat-8 .....                                     | 38 |
| 4     | Μεθοδολογία Έρευνας.....                                     | 40 |
| 4.1   | Προ-επεξεργασία Δορυφορικών Εικόνων .....                    | 40 |
| 4.2   | Περιοχή μελέτης.....   | 44 |
| 4.3   | Στάδια ταξινόμησης δορυφορικών εικόνων.....                  | 45 |
| 4.3.1 | Ίδρυση Συστήματος ταξινόμησης.....                           | 45 |
| 4.3.2 | Συλλογή εκπαιδευτικών δειγμάτων .....                        | 53 |
| 4.4   | Κατασκευή του αλγόριθμου Random Forest .....                 | 58 |
| 4.4.1 | Μοντέλο ταξινόμησης και Αποτελέσματα .....                   | 59 |
| 4.5   | Άλλες ταξινομήσεις.....                                      | 76 |
| 4.6   | Μετά-ταξινόμηση ανάλυση .....                                | 83 |
| 5     | Συζήτηση και Σχολιασμός Αποτελεσμάτων.....                   | 89 |
| 5.1   | Χαρακτηρισμός δυνατοτήτων ταξινομητή Random Forest .....     | 89 |
| 5.2   | Σύγκριση αισθητήρων OLI και MSI.....                         | 91 |
| 6     | Μελλοντικοί ερευνητικοί στόχοι .....                         | 93 |
| 6.1   | Φασματική αποσύνθεση και ταξινόμηση υπό-εικονοστοιχείων..... | 93 |
| 6.2   | Εναρμόνιση δεδομένων Landsat και Sentinel-2 .....            | 94 |
|       | BIBΛΙΟΓΡΑΦΙΑ .....   | 95 |

## ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

|   |    |
|---|----|
| Πίνακας 1: Σύστημα Ταξινόμησης της Κάλυψης Γης (LCCS) του Gregorio Di A., (2005).....   | 8  |
| Πίνακας 2: Παράδειγμα Πίνακα Σύγκρισης τεσσάρων κλάσεων με περιεχόμενα κελιών (rij) (Olofsson et.al., 2014).....  | 17 |
| Πίνακας 3: Χαρακτηριστικά αποστολής Sentinel-2.....   | 36 |
| Πίνακας 4: Χαρακτηριστικά αισθητήρα MSI (Sentinel-2).....   | 36 |
| Πίνακας 5: Χαρακτηριστικά αποστολής Landsat-8.....  | 39 |
| Πίνακας 6: Χαρακτηριστικά του αισθητήρα OLI (Landsat-8).....  | 39 |
| Πίνακας 7: Φασματική σύγκριση δεδομένων Sentinel- 2 και Landsat-8.....  | 43 |
| Πίνακας 8: Κατηγορίες Βλάστησης βάση του δείκτη NDVI.....   | 48 |
| Πίνακας 9: Κατηγορίες εδάφους.....  | 50 |
| πίνακας 10. Κατηγορίες κάλυψης γης.....   | 52 |
| Πίνακας 11: Πλήθος εικονοστοιχείων εκπαίδευσης και ελέγχου για κάθε κατηγορία κάλυψης γης.....  | 55 |
| Πίνακας 12: Separability array (Sentinel 2).....  | 56 |
| Πίνακας 13: Separability array (Landsat 8).....   | 56 |
| Πίνακες 14. Σύγκριση φασματικών υπογραφών μεταξύ αντίστοιχων θεματικών κλάσεων για τα δεδομένα Sentinel-2 και Landsat-8.....  | 57 |
| Πίνακας 15. Μέτρα Ακρίβειας των ταξινομήσεων Random Forest σε δεδομένα Sentinel 2 και Landsat 8, για το Ισορροπημένο (Balanced) και Ανισόρροπο (Imbalanced) σύνολο δεδομένων.....                             | 74 |
| Πίνακας 16. Αριθμός σημείων ελέγχου ανά κατηγορία κάλυψη γης.....   | 79 |
| Πίνακας 17. Μέτρα Ακρίβειας των ταξινομήσεων Mamimum Likelihood, Minimum Distance και Mahalanobis Distance, σε σύγκριση με την απόδοση του αλγόριθμου Random Forest σε δεδομένα Sentinel 2 και Landsat 8..... | 80 |

|   |    |
|---|----|
| Πίνακας 18. Ποσοστιαίες διαφορές μεταξύ ταξινομήσεων Random Forest και<br>Maximum Likelihood , ανά κατηγορία κάλυψης γης..... | 87 |
|---|----|

## ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ

|   |    |
|---|----|
| Διάγραμμα 1: Τρόπος λειτουργίας Ταξινόμησης (Sabins, 1996).....   | 4  |
| Διάγραμμα 2: Η διαδικασία της Ταξινόμησης, σε διάγραμμα διαστάσεων 7*7<br>(Καρτάλης Κ. και Φείδας Χ, 2012) .....  | 5  |
| Διάγραμμα 3: Σύννεφα φασματικών σημείων εκπαίδευσης τεσσάρων θεματικών<br>κατηγοριών κάλυψης γης (Κ= καλλιέργειες, Ε=έδαφος, Δ= δάσος, Ν= νερό), στις τρεις<br>διαστάσεις (κανάλια 1, 2 και 3), πηγή: Sabins, (1996)..... | 9  |
| Διάγραμμα 4: Δεδομένα Corine 2012 (Απογραφή κάλυψης γης σε 44 κλάσεις).....   | 10 |
| Διάγραμμα 5: Οριοθέτηση σημειακών και πολυγωνικών περιοχών εκπαίδευσης .....  | 12 |
| Διάγραμμα 6: Επικάλυψη Φασματικών υπογραφών διαφορετικών τάξεων στο<br>δισδιάστατο χώρο (κανάλια 1 και 4) → Αναθεώρηση των δειγμάτων .....  | 14 |
| Διάγραμμα 7: Περιπτώσεις φασματικής διαφορετικότητας μεταξύ δειγμάτων<br>εκπαίδευσης σε δισδιάστατο φασματικό χώρο και υποψήφιων ορίων διαχωρισμού των<br>κλάσεων για κάθε περίπτωση .....                                | 14 |
| Διάγραμμα 8: Μορφές ιστογραμμάτων των δεδομένων εκπαίδευσης<br>(Unimodal (Μονοτροπική), Bimodal (Διτροπική), Multimodal (πολυτροπική) .....   | 15 |
| Διάγραμμα 9: Σχεδιάγραμμα με τους βασικότερους και πιο ευρέως χρησιμοποιούμενους<br>αλγόριθμους ανά κατηγορία ταξινόμησης.....  | 20 |
| Διάγραμμα 10: Βασικά Στάδια Επιβλεπόμενης Ταξινόμησης (Lillesand T.M. et al.,<br>2004) .....  | 21 |
| Διάγραμμα 11: Εξίσωση Φασματικής Ευκλίδειας απόστασης στο N-διάστατο<br>φασματικό χώρο (αριστερά) και παράδειγμα προσδιορισμού των ορίων διαχωρισμού<br>των θεματικών κλάσεων (δεξιά).....                                | 22 |
| Διάγραμμα 12: Ταξινόμηση με βάση τον Αλγόριθμο της ελάχιστης απόστασης .....  | 23 |
| Διάγραμμα 13: Λανθασμένη ταξινόμηση εικονοστοιχείου που εμπίπτει σε φασματικά<br>επικαλυπτόμενη περιοχή δύο υπογραφών Πηγή: Παρχαρίδης Ι. (2015), “Αρχές<br>Δορυφορικής Τηλεπισκόπησης - Θεωρία και Εφαρμογές” .....      | 25 |
| Διάγραμμα 14: Συστατικά μέρη δέντρου απόφασης .....   | 26 |

|   |    |
|---|----|
| Διάγραμμα 15: Παράδειγμα δέντρου απόφασης ιεραρχικής δομής .....  | 27 |
| Διάγραμμα 16: Φαινόμενο Overfitting – Προσαρμογή συνάρτησης στον θόρυβο που υπάρχει στα δεδομένα εκπαίδευσης → Δημιουργία πολύπλοκων ορίων διαχωρισμού, που θα προκαλέσει σφάλματα στην ταξινόμηση των άγνωστης τάξης εικονοστοιχείων ..... | 28 |
| Διάγραμμα 17: Διαδικασίας εκπαίδευσης και ελέγχου της ακρίβειας ενός συνδυαστικού αλγόριθμου με τη μέθοδο bagging .....   | 29 |
| Διάγραμμα 18: Υπολογισμός δείκτη Gini για την εύρεση των βέλτιστων ορίων διαχωρισμού.....   | 31 |
| Διάγραμμα 19: Παράδειγμα ταξινόμησης άγνωστης τάξης εικονοστοιχείου με τον αλγόριθμο Τυχαίων δασών (Criminisi et al. (2011)) .....  | 32 |
| Διάγραμμα 20: Δορυφορική σάρωση της επιφάνειας της γης, από το όργανο MSI (Sentinel-2) .....  | 35 |
| Διάγραμμα 21: Τροχιά 28 δορυφόρου Sentinel-2A και η δορυφορική εικόνα σε έγχρωμο σύνθετο (3-2-1).....   | 37 |
| Διάγραμμα 22: Δορυφορική σάρωση της επιφάνειας της γης, από το αισθητήρα OLI (Landsat 8) .....  | 38 |
| Διάγραμμα 23: Διαδικασία προ-επεξεργασίας και ετοιμασίας των προς ανάλυση εικόνων .....   | 41 |
| Διάγραμμα 24: Μοντέλο δημιουργίας πολυφασματικού προϊόντος Sentinel-2, 6 καναλιών και χωρικής ανάλυσης 30 μ.....  | 42 |
| Διάγραμμα 25: Τελικό πολυφασματικό προϊόν 6 καναλιών Sentinel-2, Ανάλυση 30 m, Έγχρωμο σύνθετο: 3-2-1 .....   | 43 |
| Διάγραμμα 26: Τελικό πολυφασματικό προϊόν 6 καναλιών Landsat-8, Ανάλυση 30 m, Έγχρωμο σύνθετο: 3-2-1 .....  | 43 |
| Διάγραμμα 27: Περιοχή μελέτης .....   | 44 |
| Διάγραμμα 28: Δεδομένα Sentinel-2, Ψευδοχρωματικό Σύνθετο 1-5-5 → εντοπισμός αστικών περιοχών .....   | 45 |

|  |    |
|--|----|
| Διάγραμμα 29: Δεδομένα Sentinel-2, Ψευδοχρωματικό Σύνθετο 4-3-4 → εντοπισμός περιοχών βλάστησης.....   | 46 |
| Διάγραμμα 30: Δεδομένα Sentinel-2, Ψευδοχρωματικό Σύνθετο 6-1-1 → Οπτικός εντοπισμός κατηγοριών εδάφους και υδάτινων επιφανειών .....                        | 46 |
| Διάγραμμα 31: Δεδομένα Sentinel-2, Ψευδοχρωματικό Σύνθετο 1-3-1 → εντοπισμός περιοχών εδάφους.....   | 46 |
| Διάγραμμα 32: Χάρτης χρήσης/κάλυψης γης CORINE (ανάλυση 100 m).....  | 47 |
| Διάγραμμα 33: Αναγνώριση της πυκνότητας βλάστησης της περιοχής.....  | 48 |
| Διάγραμμα 34: Χάρτες πυκνότητας βλάστησης των δεδομένων Sentinel 2A (1η εικόνα) και Landsat 8 (2η εικόνα) και οι διαφορές που εντοπίζονται μεταξύ τους ..... | 49 |
| Διάγραμμα 35: Χάρτες εδαφών των δεδομένων Sentinel-2 και Landsat-8.....  | 51 |
| Διάγραμμα 36: Συλλογή εκπαιδευτικών δειγμάτων .....  | 54 |
| Διάγραμμα 37: Φασματικά κανάλια Sentinel-2.....  | 60 |
| Διάγραμμα 38: Φασματικά κανάλια Landsat-8 .....  | 60 |
| Διάγραμμα 39: Εκτίμηση σφάλματος γενίκευσης OOB, ως συνάρτηση του αριθμού δέντρων που απαρτίζουν το δάσος (Landsat 8) .....                                  | 64 |
| Διάγραμμα 40: Ακρίβεια του μοντέλου συναρτήσε του αριθμού των μεταβλητών που χρησιμοποιείται σε κάθε εσωτερικό κόμβο διαχωρισμού (Landsat 8).....            | 64 |
| Διάγραμμα 41: Μέση τιμή σημαντικότητας των μεταβλητών απο όλα τα δέντρα που απαρτίζουν το δάσος (Landsat 8) .....  | 66 |
| Διάγραμμα 42: Δέντρο απόφασης, αρ.100 στο τυχαίο δάσος .....   | 67 |
| Διάγραμμα 43: Πίνακας σύγχυσης και στατιστικά μέτρα ακρίβειας (Landsat 8) .....  | 68 |
| Διάγραμμα 44: Ταξινομημένη εικόνα Sentinel-2 .....   | 70 |
| Διάγραμμα 45: Ταξινομημένη εικόνα Landsat-8 .....  | 70 |
| Διάγραμμα 46: Πίνακας Σύγχυσης και στατιστικά μέτρα ακρίβειας - Ισορροπημένο σύνολο δεδομένων (Landsat 8).....   | 72 |
| Διάγραμμα 47: Ταξινομημένη εικόνα Sentinel-2 .....   | 73 |



|   |    |
|---|----|
| Διάγραμμα 48: Ταξινομημένη εικόνα Landsat-8 .....   | 73 |
| Διάγραμμα 49: Ταξινομημένη εικόνα Sentinel-2 , Overall Acc.: 81 %, Kappa: 79.06.  | 76 |
| Διάγραμμα 50: Ταξινομημένη εικόνα Landsat 8, Overall Acc.: 83%, Kappa: 81.27 %  | 76 |
| Διάγραμμα 51: Ταξινομημένη εικόνα Sentinel-2 Overall Acc.:69%,<br>Kappa: 66 %.....  | 77 |
| Διάγραμμα 52: Ταξινομημένη εικόνα Landsat 8 Overall Acc.:<br>72%, Kappa: 67 % .....   | 77 |
| Διάγραμμα 53: Ταξινομημένη εικόνα Sentinel-2<br>Overall Acc.: 72.9%, Kappa: 69.3 % .....  | 78 |
| Διάγραμμα 54: Ταξινομημένη εικόνα Landsat 8<br>Overall Acc.: 74%, Kappa: 69.2 % .....   | 78 |
| Διάγραμμα 55: Χωρική κατανομή σημείων ελέγχου στην περιοχή μελέτης .....  | 79 |
| Διάγραμμα 56: Ιστογράμματα κατανομών των κλάσεων Low και High Intensity<br>vegetation, Bare soil και Urban areas and artificial surfaces..... | 81 |
| Διάγραμμα 57: Διαφορές Natural Waterbodies (deep) μεταξύ Random Forest και<br>Maximum Likelihood (Sentinel-2) .....                           | 83 |
| Διάγραμμα 58: Διαφορές Natural Waterbodies (shallow), μεταξύ Random Forest και<br>Maximum Likelihood (Sentinel-2) .....                       | 83 |
| Διάγραμμα 59: Διαφορές Low Intensity vegetation μεταξύ Random Forest και<br>Maximum Likelihood (Sentinel-2) .....                             | 85 |
| Διάγραμμα 60: Διαφορές High Intensity vegetation μεταξύ Random Forest και<br>Maximum Likelihood (Sentinel-2) .....                            | 85 |
| Διάγραμμα 61: Διαφορές Medium Intensity vegetation μεταξύ Random Forest και<br>Maximum Likelihood (Sentinel-2) .....                          | 85 |
| Διάγραμμα 62: Διαφορές Stony Soil, μεταξύ Random Forest και Maximum Likelihood<br>(Sentinel-2) .....  | 86 |
| Διάγραμμα 63: Διαφορές Agricultural Soil, μεταξύ Random Forest και Maximum<br>Likelihood (Sentinel-2).....                                    | 86 |

|   |    |
|---|----|
| Διάγραμμα 64: Διαφορές Bare Soil, μεταξύ Random Forest και Maximum Likelihood (Sentinel-2) .....  | 86 |
| Διάγραμμα 65: Φασματικά χαρακτηριστικά Sentinel-2, Landsat-7 και Landsat-8.....   | 91 |
| Διάγραμμα 66: Φασματική ευαισθησία (Relative Spectral Response Filters) των Sentinel-2 (μαύρες γραμμές) και Landsat-8 (κόκκινες γραμμές), οι περιοχές μη τάνυσης των RSR filters υποδεικνύονται με βελάκια<br>Πηγή: <a href="https://hls.gsfc.nasa.gov/algorithms/bandpass-adjustment/">https://hls.gsfc.nasa.gov/algorithms/bandpass-adjustment/</a> ..... | 92 |
| Διάγραμμα 67: Παράδειγμα μικτού εικονοστοιχείου και φασματικής μίξης του τελικού φασματικού διανύσματος που εντοπίζει ο αισθητήρας .....  | 94 |

## ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

|        |                                    |
|--------|------------------------------------|
| RF     | Random Forest                      |
| OA     | Overall Accuracy                   |
| PA     | Producer's Accuracy                |
| MSI    | Multispectral Instrument           |
| OLI    | Operational Land Imager            |
| RSR    | Relative Spectral Response Filters |
| TIRS   | Thermal Infrared Sensor            |
| BSI    | Bare Soil Index                    |
| NDVI   | Normalized Vegetation Index        |
| WGS 84 | World Geodetic System 1984         |
| TOA    | Top of Atmosphere                  |
| IBS    | In Bag Samples                     |
| OOB    | Out of Bag                         |
| HLS    | Harmonized Landsat_Sentinel-2      |

## ΑΠΟΔΟΣΗ ΟΡΩΝ

|            |   |
|------------|---|
| Multimodal | Μορφή κατανομής ενός συνόλου δεδομένων, όπου στο γράφημα απεικόνισης της συχνότητας των τιμών (ιστόγραμμα), παρουσιάζονται περισσότερες από μία κορυφές.  |
| Unimodal   | Μορφή κατανομής ενός συνόλου δεδομένων, όπου στη συχνότητα εμφάνισης των τιμών παρουσιάζεται ξεκάθαρα μια κορυφή, υποδεικνύοντας ότι προσαρμόζεται στην κανονική κατανομή   |
| Bias       | Κάθε είδος σφάλματος το οποίο δεν περιγράφεται από τη στατιστική (χρησιμοποιείται για να περιγράψει τα τυχαία σφάλματα)   |
| Bagging    | Διαδικασία τυχαίας επιλογής ενός υποσυνόλου από το σύνολο των δεδομένων εκπαίδευσης, για την εκπαίδευση ενός συνδυαστικού αλγόριθμου ταξινόμησης. Το τυχαία επιλεγμένο υποσύνολο, επανατοποθετείται στο σύνολο δεδομένων, έτσι οι άλλοι ταξινομητές που συνθέτουν το συνδυαστικό μοντέλο μπορούν να επιλέξουν τυχαία ξανά μέρος ή ολόκληρο το υποσύνολο των δεδομένων για σκοπούς εκπαίδευσης |

## 1 Εισαγωγή

Η κάλυψη γης αποτελεί χωρίς αμφισβήτηση μία θεμελιώδη μεταβλητή, που επηρεάζει και συνδέει άμεσα πολλά μέρη του ανθρώπινου και φυσικού περιβάλλοντος.

Η χαρτογράφηση και η συνεχής παρακολούθηση της δυναμικής αυτής μεταβλητής, έχει αναγνωριστεί ευρέως ως έναν από τους σημαντικότερους και πιο απαιτητικούς ερευνητικούς στόχους, αφού η πληροφορία που παράγεται από τις εν λόγω διαδικασίες, αποτελεί σημαντικό εργαλείο λήψης αποφάσεων σε μελέτες με περιβαλλοντικό και χωροταξικό ενδιαφέρον (Hermosilla T. et al., 2016).

Η ταξινόμηση μίας ψηφιακής εικόνας, είναι μία από τις δημοφιλέστερες μεθόδους εξαγωγής πληροφορίας από τηλεπισκοπικά δεδομένα. Σκοπός αυτής της διαδικασίας, είναι η αντικατάσταση της αναξιόπιστης φωτοερμηνείας των εικόνων, με στατιστικές / ποσοτικές τεχνικές για την αυτόματη αναγνώριση φασματικών και χωρικών προτύπων που συνεπάγονται τη παραγωγή αξιόπιστων θεματικών χαρτών κάλυψης γης (Λασπιάς Ε., 2012; Παρχαρίδης Ι., 2015)

Παρόλα αυτά, παρά το πλήθος των εφαρμογών οι οποίες στηρίζονται σε θεματικούς χάρτες χρήσης/κάλυψης γης, διατυπώνεται μια σημαντική ανησυχία η οποία αφορά την ανεπάρκεια των χαρτών αυτών σε ποιότητα και αξιοπιστία, προκειμένου να αποτελέσουν τη βάση σε αυτές τις εφαρμογές.

Αυτή η ανησυχία βασίζεται στην αξιολόγηση του παραγόμενου προϊόντος, συγκρίνοντας το με επίγεια δεδομένα αναφοράς. Οι διαφορές που εντοπίζονται μεταξύ των δυο αυτών συνόλων δεδομένων, ερμηνεύονται ως σφάλματα στο χάρτη κάλυψης γης, που προκύπτουν είτε από τα δεδομένα εισόδου είτε από τον αλγόριθμο ταξινόμησης. Η ερμηνεία αυτή, έχει οδηγήσει τη επιστημονική κοινότητα στην αναζήτηση μεθόδων μείωσης των σφαλμάτων στην ταξινόμηση δορυφορικών εικόνων.

Μεταξύ αυτών, είναι η δωρεάν διαθεσιμότητα της νέας γενιάς πολυφασματικών δεδομένων Sentinel-2, καθώς και η ανάπτυξη ισχυρών αλγορίθμων όπως είναι οι συνδυαστικοί ταξινομητές, που έχουν ωθήσει την επιστήμη της τηλεπισκόπησης σε μια νέα εποχή, προσφέροντας πρωτοφανείς ευκαιρίες σε εφαρμογές παρακολούθησης της γης.

## **1.1 Σκοπός εργασίας και σύνοψη μεθοδολογίας**

Σκοπός αυτής της εργασίας, ήταν η διερεύνηση των δυνατοτήτων χρήσης του συνδυαστικού ταξινομητή Random Forest στην τηλεπισκόπηση, δίνοντας έμφαση στην παραμετροποίηση και τις ευαισθησίες τους.

Αρχικά, ο αλγόριθμος θα αξιολογηθεί στην ταξινόμηση πολυφασματικών δεδομένων Sentinel-2, βάση των εξής κριτηρίων: τον αριθμό των δέντρων απόφασης που απαρτίζουν το “δάσος”, το μέγεθος του συνόλου δεδομένων που χρησιμοποιούνται για την εκπαίδευση του αλγορίθμου καθώς και το πλήθος των μεταβλητών που επιλέγονται και ελέγχονται για την εύρεση του βέλτιστου τρόπου διαχωρισμού των δεδομένων κατά την ανάπτυξη του κάθε δέντρου. Ακολούθως, τα αποτελέσματα που θα προκύψουν από την ταξινόμηση, θα συγκριθούν με ακόμα τρεις συμβατικές, παραμετρικές τεχνικές ταξινόμησης όπως η Μέγιστη Πιθανοφάνεια, η Ελάχιστη απόσταση και η απόσταση Mahalanobis, ως προς τα ίδια δεδομένα εκπαίδευσης.

Συγκεκριμένα, θα εφαρμοστούν δύο διαφορετικές στρατηγικές ταξινόμησης: η πρώτη χρησιμοποιώντας άνισο και τυχαία επιλεγμένο πλήθος δεδομένων εκπαίδευσης για καθεμία από τις 11 ομάδες κάλυψης γης και η δεύτερη με τη χρήση ίσου αριθμού δεδομένων εκπαίδευσης για όλες τις κλάσεις.

Ακόμα, κρίνεται σημαντικό να εκτιμηθεί το εάν και κατά πόσο τα δεδομένα Sentinel-2, τα οποία θεωρούνται ως η συνέχεια της μακροβιότερης συλλογής δεδομένων Landsat, παράγουν τελικά προϊόντα όμοια με αυτά που προέρχονται από τα δεδομένα του αισθητήρα OLI και το δορυφόρο Landsat-8. Για αυτό το σκοπό, θα εφαρμοστούν οι ίδιες τεχνικές και στρατηγικές ταξινόμησης καθώς και τα ίδια δεδομένα εκπαίδευσης, για να προκύψει η τελική σύγκριση τους.

## **1.2 Η χαρτογράφηση κάλυψης γης με τη χρήση δορυφορικών δεδομένων**

Δεδομένου ότι κύριος σκοπός της ταξινόμησης είναι η κατηγοριοποίηση ενός συνόλου εικονοστοιχείων σε τάξεις που αντιπροσωπεύουν κατηγορίες κάλυψης γης, οποιαδήποτε μέθοδος που επιδιώκει να προσδιορίσει αυτές τις κατηγορίες με βάση την αρχή της

φασματικής συσχέτισης, μπορεί να εφαρμοστεί. Ως αποτέλεσμα, διάφορες μέθοδοι ταξινόμησης εικόνων έχουν αναπτυχθεί με βάση διαφορετικές θεωρίες ή μοντέλα.

Μερικές από τις πιο κοινά χρησιμοποιούμενες μεθόδους σε εφαρμογές ταξινόμησης εικόνων, είναι η Μέγιστη Πιθανοφάνεια, η Ελάχιστη Απόσταση και η Απόσταση Mahalanobis.

Η παραμετρική μέθοδος ταξινόμησης της Μέγιστης Πιθανοφάνειας (Maximum Likelihood), βασίζεται σε υποθέσεις περί της φασματικής απόκρισης της κάθε τάξης και πιο συγκεκριμένα, στο ότι η κατανομή του νέφους των σημείων που απαρτίζουν την κατηγορία είναι κανονική. Κάτω από αυτή τη υπόθεση, ο αλγόριθμος βασίζεται σε στατιστικές παραμέτρους (μέσος όρος και τυπική απόκλιση) και χαρακτηρίζεται από ένα υποκειμενικό μοντέλο πιθανότητας, το οποίο μας παρέχει ένα μέτρο πιθανότητας για κάθε εικονοστοιχείο σε κάθε κατηγορία και όχι απλώς την τιμή της κατηγορίας στην οποία αυτό καταλήγει (Otukey J.R. and Blaschke, 2009).

Ακόμα, οι αποδόσεις άλλων παραδοσιακά χρησιμοποιούμενων ταξινομητών, όπως ο αλγόριθμος της Ελάχιστης Απόστασης και της Απόστασης Mahalanobis, εξαρτώνται από το πόσο καλά συμφωνούν τα δεδομένα με το προκαθορισμένο μοντέλο (που δημιουργείται κατά την εκπαίδευση του αλγορίθμου βάση των δεδομένων εκπαίδευσης) (Sharma R. et al, 2013). Αυτές οι μέθοδοι συνήθως αποτυγχάνουν να δώσουν τα επιθυμητά αποτελέσματα όταν τα δεδομένα είναι πολυτροπικής μορφής (multimodal), δηλαδή όταν στο ιστόγραμμα συχνότητας εμφάνισης των τιμών, εμφανίζονται περισσότερες από μία κορυφές, υποδηλώνοντας μεγάλη διακύμανση στις φασματικές τιμές.

Με την πάροδο του χρόνου, η επιστημονική κοινότητα μέσα από τις προσπάθειες για την υπέρβαση αυτών των προβλημάτων, έχει αναπτύξει πιο εξελιγμένους αλγόριθμους όπως είναι οι συνδυαστικοί ταξινομητές. Μεταξύ αυτών είναι τα Τυχαία Δάση (Breiman L., 2001), τα οποία τις τελευταίες δύο δεκαετίες χρησιμοποιούνται ολοένα και περισσότερο, λόγω της υπολογιστικής ταχύτητας και των πολύ καλών αποτελεσμάτων ταξινόμησης (Belgiu M and Dragut L., 2016). Αυτοί οι ταξινομητές είναι ουσιαστικά μη παραμετρικής φύσης, αφού δεν βασίζονται σε υποθέσεις περί της κατανομής των δεδομένων και προσαρμόζονται ευκολότερα σε κάθε είδους μορφής δεδομένων.

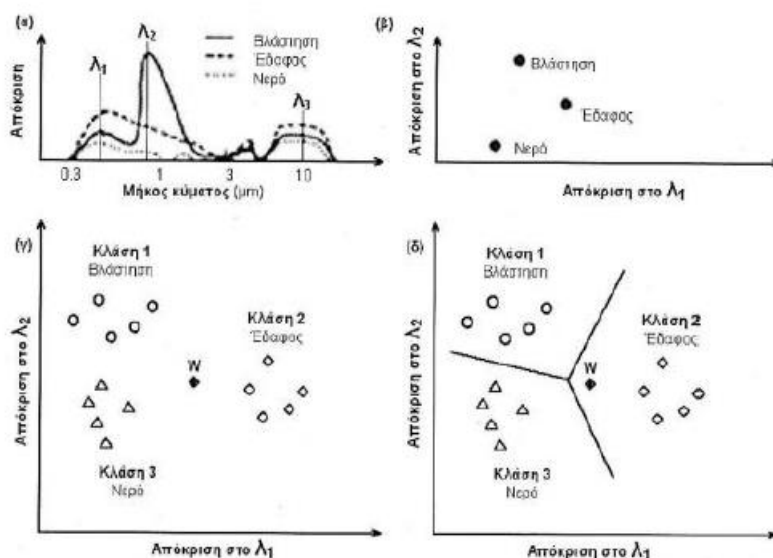
## 2 Βιβλιογραφική ανασκόπηση

### 2.1 Ταξινόμηση

Η έννοια της ταξινόμησης, αναφέρεται στην αναγνώριση, διαφοροποίηση και κατηγοριοποίηση ενός συνόλου δεδομένων και είναι ευρέως γνωστή αφού οι μέθοδοι της βρίσκουν συχνά εφαρμογή σε μελέτες και έρευνες, από διάφορα επιστημονικά πεδία.

Στην επιστήμη της τηλεπισκόπησης, συναντούμε μια κατά κάποιον τρόπο διαφοροποιημένη έννοια της ταξινόμησης, η οποία οφείλεται στο γεγονός ότι ο εκάστοτε αλγόριθμος βασίζεται στις πληροφορίες (ραδιομετρικές τιμές ή τιμές ανακλαστικότητας) που υπάρχουν για κάθε εικονοστοιχείο, σε κάθε φασματικό κανάλι. Με λίγα λόγια, η ταξινόμηση είναι πολύ-φασματική (Περάκης Γ. Κ., 2015) και έχει ως στόχο τη δημιουργία φασματικά ομοιογενών εσωτερικά και ανομοιογενών μεταξύ τους κλάσεων.

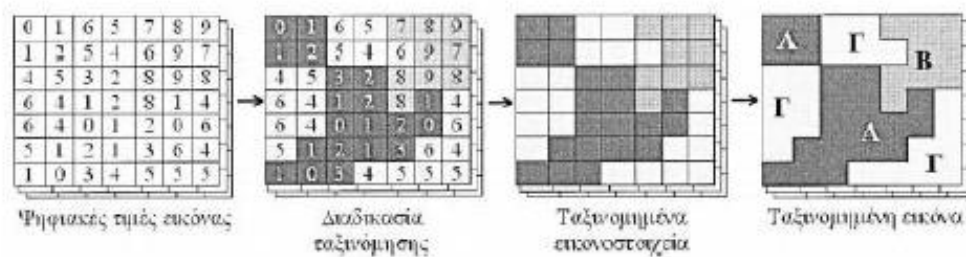
Σε παράδειγμα του τρόπου λειτουργίας της ταξινόμησης, παρατηρώντας την διάγραμμα 1 (Sabins, 1996), στο σημείο: α) απεικονίζονται οι φασματικές υπογραφές τριών τύπων επιφάνειας (Βλάστηση, Νερό και Έδαφος), οι οποίες αποτελούν τα δεδομένα εκπαίδευσης για τις θεματικές τάξεις που θα δημιουργηθούν, β) απεικονίζονται οι προαναφερόμενες κλάσεις στο δισδιάστατο φασματικό χώρο, σε δυο από τους τρεις διαθέσιμους διαύλους, γ) απεικονίζεται η διασπορά των κλάσεων στον ίδιο δισδιάστατο φασματικό χώρο και δ) απεικονίζονται τα όρια διαχωρισμού των θεματικών τάξεων, βάση των οποίων τα άγνωστα εικονοστοιχεία θα ομαδοποιηθούν.



Διάγραμμα 1: Τρόπος λειτουργίας Ταξινόμησης (Sabins, 1996)



Ο προσδιορισμός των θεματικών κλάσεων σε μια δορυφορική εικόνα, πραγματοποιείται από δυο είδη ταξινομητών. Τους ταξινομητές φάσματος, οι οποίοι αντιμετωπίζουν το κάθε εικονοστοιχείο από το οποίο απαρτίζεται η εικόνα, ως απομονωμένο αντικείμενο και τα ομαδοποιούν στις προκαθορισμένες ή όχι θεματικές κατηγορίες βάση των φασματικών χαρακτηριστικών τους. Ο δεύτερος τύπος ταξινομητών είναι οι γεωμετρικοί ταξινομητές χώρου, οι οποίοι σύμφωνα με τον Baatz et al., (2000) υποδιαιρούν την εικόνα σε περιοχές και εξετάζουν τις χωρικές σχέσεις μεταξύ των εικονοστοιχείων, εντοπίζοντας ομάδες από αυτά στις οποίες συμπεριφέρονται ως αντικείμενα (Lillesand T.M. et al., 2004).



**Διάγραμμα 2:** Η διαδικασία της Ταξινόμησης, σε διάγραμμα διαστάσεων 7\*7 (Καρτάλης Κ. και Φείδας Χ, 2012)

Σε κάθε περίπτωση, ο στόχος και τα αποτελέσματα είναι τα ίδια, δηλαδή το κάθε εικονοστοιχείο κατηγοριοποιείται σε ένα προκαθορισμένο αριθμό θεματικών κλάσεων, οδηγώντας στη δημιουργία θεματικών χαρτών.

Αν και λιγότερο ακριβείς, στην εν λόγω μελέτη θα δώσουμε έμφαση στις φασματικά προσανατολισμένες διαδικασίες ταξινόμησης, αφού την παρούσα στιγμή αποτελούν λόγο της λιγότερης πολυπλοκότητας τους, την βάση των περισσότερων πολυφασματικών διαδικασιών ταξινόμησης σε εφαρμογές χαρτογράφησης της κάλυψης γης.

### 2.1.1 Τα στάδια της επιβλεπόμενης ταξινόμησης

Η διαδικασία της επιβλεπόμενης ταξινόμησης αποτελείται από πέντε βασικά στάδια, τα οποία υποδιαιρούνται σε λεπτομερέστερες διαδικασίες και έχουν ως εξής:

1. Σχεδιασμός Ταξινόμησης
  - i. Επιλογή Συστήματος Ταξινόμησης
2. Στάδιο Εκπαίδευσης
  - ii. Πηγές Δεδομένων Αναφοράς
  - iii. Σχεδιασμός Δειγματοληψίας
    - a. Χωρική κατανομή δειγματοληπτικών περιοχών εκπαίδευσης
    - b. Διαστάσεις χωρικών οντοτήτων για τη δειγματοληψία περιοχών εκπαίδευσης
    - c. Το μέγεθος του δείγματος
    - d. Περιοχές Ελέγχου
  - iv. Αξιολόγηση και αναθεώρηση του συνόλου δεδομένων εκπαίδευσης
3. Εφαρμογή του αλγόριθμου ταξινόμησης
4. Αξιολόγηση της ακρίβειας ταξινόμησης
5. Διαμόρφωση τελικών θεματικών χαρτών

Ακολουθεί αναλυτικότερη επεξήγηση του κάθε σταδίου.

#### 1. Σχεδιασμός Ταξινόμησης

Στο στάδιο του σχεδιασμού ταξινόμησης, εμπίπτει η ίδρυση ενός συστήματος ταξινόμησης που να περιγράφει πλήρως τις θεματικές κατηγορίες κάλυψης γης στην περιοχή μελέτης και έχει σκοπό την υποστήριξη της διαδικασίας συλλογής εκπαιδευτικών δειγμάτων για την μετέπειτα ταξινόμηση.

##### i. Σύστημα Ταξινόμησης

Το σύστημα ταξινόμησης, αναφέρεται στη διαδικασία μετατροπής των υπαρχουσών φασματικών τάξεων σε μία δορυφορική εικόνα, στις κατηγορίες στις οποίες θα ταξινομηθούν τα εικονοστοιχεία της.

Τα τελευταία χρόνια, παρατηρείται μια αυξανόμενη ανάγκη για λεπτομερείς και ακριβείς πληροφορίες που αφορούν τη χρήση/κάλυψη γης σε όλες τις γεωγραφικές κλίμακες και αποτελούν τη βάση για πολλές εφαρμογές. Ο τρόπος και το επίπεδο ομαδοποίησης των χαρακτηριστικών που υπάρχουν σε μια περιοχή, ποικίλει ανάλογα με το σκοπό της χαρτογράφησης και της έρευνας.

Προσεγγίσεις για τη χωρική ταξινόμηση της κάλυψης γης, εντοπίζονται από τις δεκαετίες του 1970 και 1980, όπου ενώ οι Anderson et al. (1976) θεωρούν ότι το περιεχόμενο μιας δορυφορικής εικόνας, πρέπει να χωρίζεται σε εννέα κύριες κατηγορίες (Αστική περιοχή, Αγροτική Περιοχή, Βοσκότοπος, Δάσος, Υδάτινη Επιφάνεια, Υγροβιότοπος, Άγονο έδαφος, Τούντρα, Αιώνια χιόνια και Πάγοι), οι Καρτέρης και Τσομπανίκος (1984) προτείνουν την ομαδοποίηση των διαφορετικών τύπων κάλυψης/χρήσης γης σε έξι κύριες κατηγορίες (Αστική γη, Δασική γη, Γεωργική γη, Νερό και Άγωνα γη). Σε κάθε περίπτωση, οι κύριες αυτές κατηγορίες υποδιαίρονται ανάλογα με την κλίμακα και τη χωρική διακριτική ικανότητα των πρωτογενών δεδομένων, σε άλλες υπό-θεματικές κατηγορίες.

Παρά τις προαναφερόμενες προσεγγίσεις δημιουργίας συστημάτων ταξινόμησης, στην πράξη αποδείχτηκε δύσκολο να δημιουργηθεί μια κοινή γλώσσα ερμηνείας μεταξύ των θεματικών χαρτών που προκύπτουν από τη διαδικασία της αυτοματοποιημένης ομαδοποίησης εικονοστοιχείων. Αυτό συμβαίνει λόγω του ότι κάθε αποτύπωση ορίζει διαφορετικά παρόμοιες κατηγορίες, αφού αν για παράδειγμα σε ένα χάρτη σχεδιασμού και οικιστικής ανάπτυξης, υπάρχουν περιοχές χωρίς δέντρα, τότε μπορούν να χαρακτηριστούν ως Ζώνες Αστικού Σχεδιασμού. Αν όμως ο σκοπός δημιουργίας του χάρτη είναι ο σχεδιασμός αναδάσωσης, τότε οι ίδιες περιοχές θα χαρακτηριστούν ως Δάσος (Βάσιλας Γ., 2013).

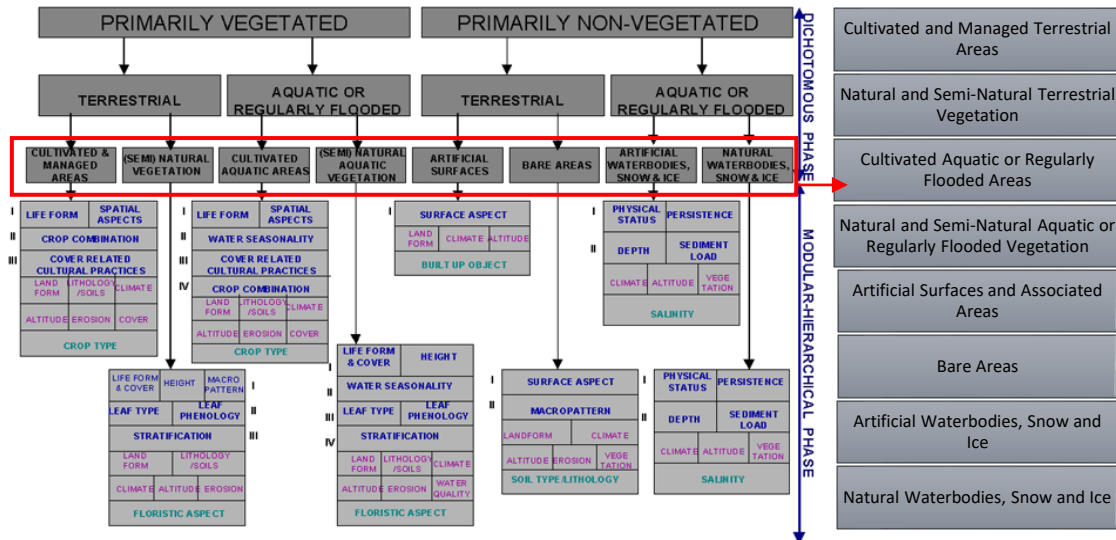
Μέσα από προσπάθειες τυποποίησης και άμβλυνσης των προβλημάτων που παρουσιάζονται στην κάθε περίπτωση, στην τρέχουσα έκδοση του FAO LCCS (Land Classification System), ο Gregorio Di A., (2005) ανέπτυξε ένα σύστημα το οποίο ανταποκρίνεται στις ειδικές απαιτήσεις των χρηστών σε παγκόσμια κλίμακα και υποστηρίζει τους ερευνητικούς στόχους ανεξαρτήτως γεωγραφικής κλίμακας.

Συγκεκριμένα, μέσα από δυο κύριες φάσεις διακρίνει το διαχωρισμό της κάλυψης γης σε οκτώ κύριες ομάδες και ακολούθως διαιρεί αυτές τις κατηγορίες επίσης ανάλογα με την

ανάλυση των πρωτογενών δεδομένων, σε υπό-κατηγορίες που ανταποκρίνονται στο επιθυμητό επίπεδο λεπτομέρειας της εφαρμογής (πίνακας 1).

**Πίνακας 1:** Σύστημα Ταξινόμησης της Κάλυψης Γης (LCCS) του Gregorio Di A., (2005)

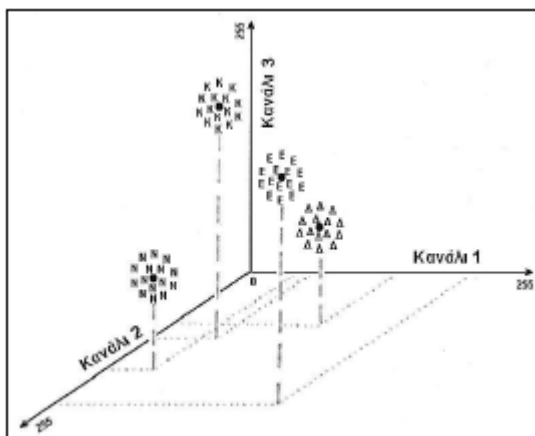
Πηγή : [http://www.fao.org/docrep/003/x0596e/x0596e01f.htm#p281\\_24459](http://www.fao.org/docrep/003/x0596e/x0596e01f.htm#p281_24459)



Σε κάθε περίπτωση, ο σημαντικότερος παράγοντας που πρέπει να λάβει υπόψη του ο αναλυτής, είναι η ελάχιστη χωρική μονάδα χαρτογράφησης (MMU), η οποία εξασφαλίζει ότι η έχει δοθεί η απαιτούμενη σημασία στην καθώς πρέπει μέτρηση της ακρίβειας κατά την τρέχουσα εφαρμογή (Olofsson et al. 2014) .

## 2. Στάδιο Εκπαίδευσης

Ο αντικειμενικός στόχος της διαδικασίας συλλογής εκπαιδευτικών δειγμάτων, είναι η συγκέντρωση ενός συνόλου στατιστικών στοιχείων τα οποία είναι ικανά να περιγράψουν επιτυχώς το πρότυπο φασματικής απόκρισης για κάθε τύπο κάλυψης γης που πρόκειται να ταξινομηθεί σε μια εικόνα. Έτσι κατά τη διάρκεια του σταδίου εκπαίδευσης, προσδιορίζεται η θέση, το μέγεθος, το σχήμα και ο προσανατολισμός των σύννεφων των φασματικών σημείων για κάθε τάξη κάλυψης εδάφους.



**Διάγραμμα 3:** Σύννεφα φασματικών σημείων εκπαίδευσης τεσσάρων θεματικών κατηγοριών κάλυψης γης (K= καλλιέργειες, E=έδαφος, Δ= δάσος, N= νερό), στις τρεις διαστάσεις (κανάλια 1, 2 και 3), πηγή: Sabins, (1996)

Για την παραγωγή ποιοτικά αποδεκτών αποτελεσμάτων ταξινόμησης, τα δεδομένα εκπαίδευσης πρέπει να είναι αντιπροσωπευτικά, αξιόπιστα και πλήρη.

Αυτό σημαίνει ότι ο αναλυτής πρέπει να συλλέξει στατιστικά στοιχεία εκπαίδευσης, για όλες τις φασματικές κατηγορίες που συμβάλλουν στη διάκριση της κάθε κλάσης πληροφορίας από τον ταξινομητή. Αν για παράδειγμα επιθυμούμε την οριοθέτηση μίας θεματικής κατηγορίας που ονομάζεται Υδάτινη Επιφάνεια, τότε πρέπει να εξετάσουμε την περιοχή ενδιαφέροντος για να εντοπίσουμε το πλήθος των υδάτινων σωμάτων (με διαφορετικά φασματικά χαρακτηριστικά), που περιέχονται στην υπό-ανάλυση εικόνα.

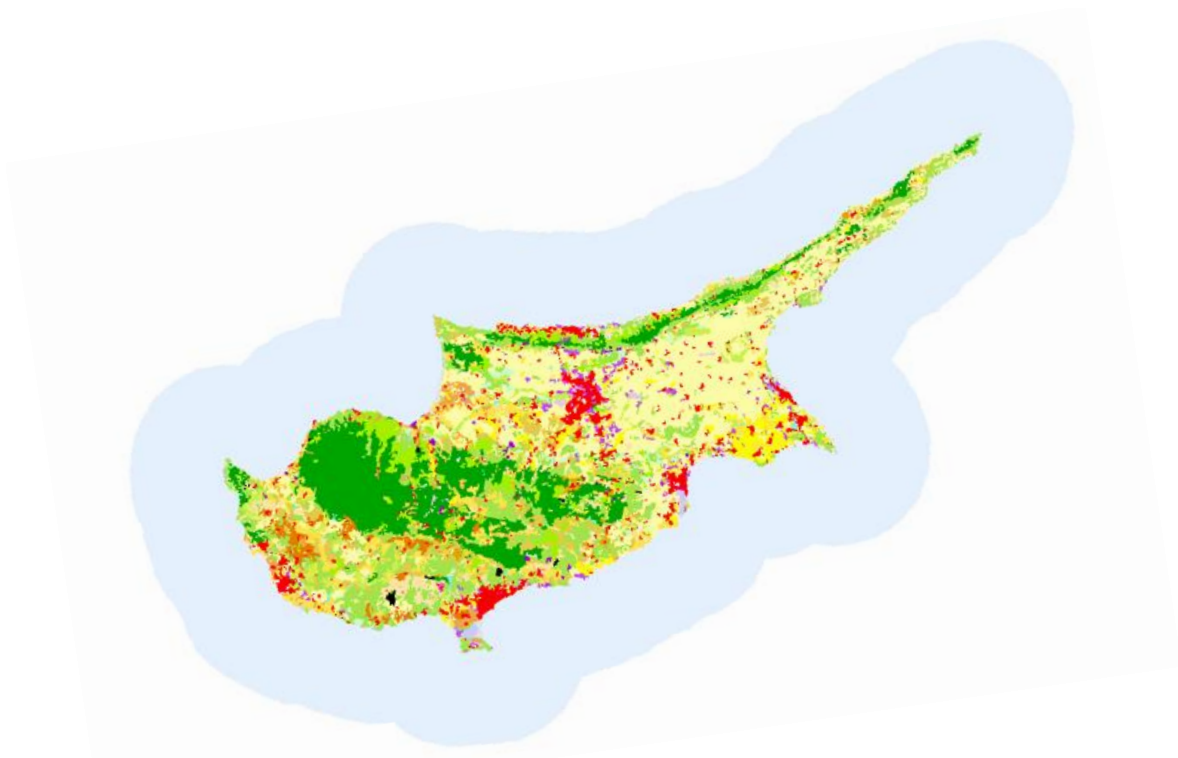
Τότε το χαρακτηριστικό Υδάτινη Επιφάνεια, θα πρέπει να παρουσιάζεται με στατιστικά στοιχεία που συλλέχθηκαν σε κάθε ξεχωριστή περιοχή ενδιαφέροντος (θαλασσινό νερό, νερό λίμνης, νερό φράγματος, διάφορα επίπεδα θολότητας του νερού κ.ο.κ.). Στο τέλος τρεις ή τέσσερις φασματικές τάξεις θα χρησιμοποιηθούν για την ταξινόμηση όλων των τύπων των υδάτινων σωμάτων και την παρουσίαση τους ως μια κλάση.

#### ii. Πηγές Δεδομένων Αναφοράς

Τα δεδομένα αναφοράς μπορούν να προσδιοριστούν από διάφορες πηγές. Αυτές κυμαίνονται από επιτόπιες επισκέψεις στις δειγματοληπτικές περιοχές, τη χρήση αεροφωτογραφιών / δορυφορικών εικόνων, παραγόμενων προϊόντων που προέρχονται από την εφαρμογή δεικτών βλάστησης και εδαφών ή ακόμη και υψηλής ποιότητας και

αξιοπιστίας ταξινομημένα προϊόντα, όπως αυτά που προέρχονται από τους χάρτες κάλυψης γης του Corine.

Για τη δημιουργία αξιόπιστων προϊόντων που αποσκοπεί στην παροχή υπηρεσιών στον τομέα παρακολούθησης των χρήσεων γης, οι υπηρεσίες διαχείρισης γης του ευρωπαϊκού προγράμματος Corine (Co-ordination of Information on the Environment) βασίζονται τόσο σε δορυφορικές εικόνες και σύνολα δεδομένων ευρωπαϊκού επιπέδου, όσο και σε δεδομένα που συλλέγονται στο πεδίο, σε εθνικό κυρίως επίπεδο. Είναι διαθέσιμα σε χωρική ανάλυση εκατό και διακοσίων πενήντα μέτρων, με τη χρονολογική σειρά εκδόσεων τους να ξεκινά το έτος 1990 και να συνεχίζεται με τρεις έως τώρα επικαιροποιημένες δημοσιεύσεις (2000, 2006 και 2012) .



**Διάγραμμα 4:** Δεδομένα Corine 2012 (Απογραφή κάλυψης γης σε 44 κλάσεις)

Πηγή: <https://land.copernicus.eu/pan-european/corine-land-cover>

### iii. Σχεδιασμός Δειγματοληψίας

Η διαδικασία του σχεδιασμού δειγματοληψίας, αφορά το μέγεθος του συνόλου των χωρικών οντοτήτων εκπαίδευσης, των διαστάσεων τους καθώς και την χωρική κατανομή τους.

Η επιλογή του τρόπου δειγματοληψίας, απαιτεί την εξέταση κυρίων κριτηρίων σχεδιασμού, όπως οι διαθέσιμοι πόροι από πλευράς χρόνου και χρήματος και οι στόχοι επίτευξης ποιότητας του παραγόμενου προϊόντος. Σύμφωνα με τον Περάκης Γ. Κ., (2015), η χωρική τοποθέτηση των δειγματοληπτικών περιοχών εκπαίδευσης μέσα στην περιοχή ενδιαφέροντος (απλή τυχαία, στρωματοποιημένη και συστηματική δειγματοληψία), επηρεάζει σε μεγαλύτερο βαθμό την ακρίβεια της ταξινόμησης απ' ό τι το πλήθος των εικονοστοιχείων εκπαίδευσης, αρκεί αυτά να περιγράφουν φασματικά πλήρως την κάθε κλάση.

#### a. Χωρική κατανομή δειγματοληπτικών περιοχών εκπαίδευσης

Όσο αφορά τη χωρική κατανομή των δειγματοληπτικών περιοχών εκπαίδευσης, είναι αποδεδειγμένο (Lillesand T.M. et al., 2004), ότι η διασπορά των δειγματοληπτικών θέσεων, αυξάνει τις πιθανότητες τα δεδομένα εκπαίδευσης να είναι αντιπροσωπευτικά ολόκληρου του εύρους φάσματος μίας θεματικής ομάδας.

Ο ορισμός του πρότυπου εκπαίδευσης μίας κλάσης, είναι προφανώς καλύτερος αναλύοντας είκοσι τοποθεσίες οι οποίες περιέχουν ανάλογα με τη χωρική ανάλυση των πρωτογενών δεδομένων, για παράδειγμα δέκα εικονοστοιχεία εκάστη, από την ανάλυση δύο ή τριών θέσεων που περιέχουν πενήντα ή εκατό χωρικές μονάδες. Με αυτό τον τρόπο μπορούμε επίσης να διατηρούμε τη φασματική ομοιογένεια στο εσωτερικό κάθε περιοχής εκπαίδευσης.

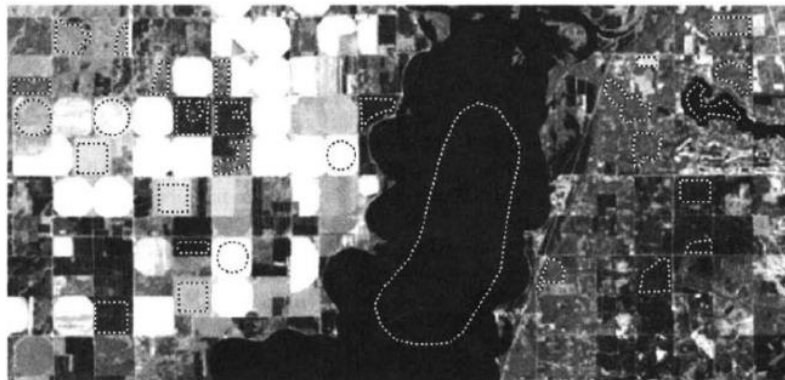
Μέχρι στιγμής, έχει γίνει σαφές ότι η διαδικασία εκπαίδευσης μπορεί να γίνει αρκετά πολύπλοκη. Ανάλογα με τη φύση των δεδομένων και την πολυπλοκότητα της γεωγραφικής περιοχής, δεν είναι καθόλου ασυνήθιστη η συλλογή δεδομένων από εκατό ή περισσότερες περιοχές εκπαίδευσης, οι οποίες θα αντιπροσωπεύσουν επαρκώς τη φασματική μεταβλητότητα που εντοπίζεται στη δορυφορική εικόνα (Μερτίκας Σ., 2006).

b. Διαστάσεις χωρικών οντοτήτων των περιοχών εκπαίδευσης

Οι χωρικές οντότητες που χρησιμοποιούνται για την συλλογή των δεδομένων εκπαίδευσης, είναι είτε σημεία είτε πολύγωνα (Stehman & Wickham, 2011).

Στη δεύτερη περίπτωση, τα όρια κάθε πολυγώνου πρέπει να ψηφιοποιούνται προσεκτικά προκειμένου να αποφευχθεί η ύπαρξη εικονοστοιχείων κατά μήκος ακμών (ορίων μεταβλητότητας). Στη συνέχεια, οι αριθμητικές τιμές των εικονοστοιχείων που απαρτίζουν την κάθε περιοχή εκπαίδευσης, χρησιμοποιούνται για το σχηματισμό των ν-στατιστικών περιγραφών τους.

Ως πιο ακριβής και αποτελεσματική, χαρακτηρίζεται η συλλογή δεδομένων εκπαίδευσης με το σημειακό τρόπο. Το εκάστοτε επιλεγόμενο εικονοστοιχείο, τείνει να είναι αντιπροσωπευτικό της περιοχής που το περιβάλλει, έτσι εικονοστοιχεία με ίδια ή παρόμοια χαρακτηριστικά που βρίσκονται στην εγγύς περιοχή συγχωνεύονται μαζί με του, για να αποτελέσουν το δείγμα εκπαίδευσης αυτής της περιοχής.



**Διάγραμμα 5:** Οριοθέτηση σημειακών και πολυγωνικών περιοχών εκπαίδευσης

a. Το μέγεθος του δείγματος

Η μεγάλη πρόκληση που έχει να αντιμετωπίσει η ανάπτυξη συνόλων δεδομένων εκπαίδευσης, είναι η συλλογή μεγέθους δείγματος, ικανοποιητικού για τον ακριβή προσδιορισμό των στατιστικών παραμέτρων που χρησιμοποιεί ο εκάστοτε ταξινομητής.

Όπως αναφέρουν οι Belgiu M. and Dragut L. (2016), το μέγεθος των δειγμάτων εκπαίδευσης επηρεάζει την απόδοση του ταξινομητή και πρέπει ταυτόχρονα να αντιπροσωπεύει τη συνολική φασματική μεταβλητότητα που υπάρχει στην περιοχή ενδιαφέροντος.



Το θεωρητικά κατώτερο όριο πλήθους εικονοστοιχείων που πρέπει να περιέχονται σε ένα σύνολο δεδομένων εκπαίδευσης όταν χρησιμοποιείται από οποιοδήποτε στατιστικό ταξινομητή (όπως ο αλγόριθμος Μέγιστης Πιθανοφάνειας), είναι σύμφωνα με τους Lillesand T.M. et al., (2004),  $v+1$  (όπου  $v$  είναι ο αριθμός των καναλιών), ενώ άλλες έρευνες αναφέρουν ότι το ποσοστό του μεγέθους των δειγμάτων πρέπει να αγγίζει περίπου το 0,25% της συνολικής περιοχής μελέτης (Colditz, 2015).

Παρά τις θεωρίες, στην πράξη εύρος πλήθους εικονοστοιχείων 10n έως 100n χρησιμοποιείται για την καθώς πρέπει αξιολόγηση της διακύμανσης και συνδιακύμανσης των φασματικών τιμών απόκρισης.

Οι Lillesand T.M. et al., (2004), αναφέρουν μεταξύ άλλων, ότι η εκτίμηση των μέσων διανυσμάτων και των πινάκων συνδιακύμανσης, βελτιώνεται καθώς ο αριθμός των εικονοστοιχείων εκπαίδευσης αυξάνεται. Όσα περισσότερα εικονοστοιχεία χρησιμοποιούνται για σκοπούς εκπαίδευσης, τόσο καλύτερη και πιο αξιόπιστη θα είναι η στατιστική αναπαράσταση της φασματικής κλάσης.

Εν ολίγοις, κανένας αναλυτής δεν επιθυμεί να παραλείψει οποιαδήποτε σημαντική για την περιγραφή της φασματικής κλάσης πληροφορία, αλλά ούτε να συμπεριλάβει περιττό αριθμό πλεοναζουσών εικονοστοιχείων στη διαδικασία ταξινόμησης (από υπολογιστική άποψη).

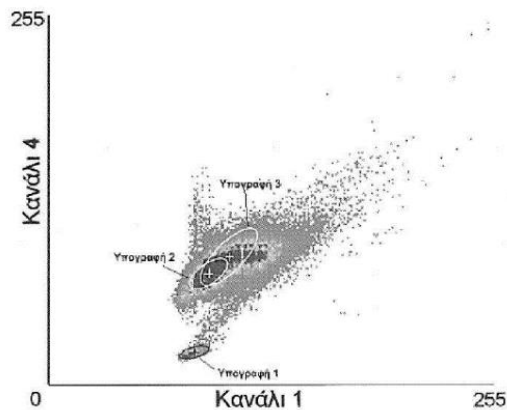
#### b. Περιοχές Ελέγχου

Οι περιοχές ελέγχου, προκύπτουν από την σκοπίμως συλλογή πλεοναζουσών περιοχών εκπαίδευσης που προορίζονται για την ανάπτυξη των στατιστικών στοιχείων της ταξινόμησης.

Ένα υποσύνολο αυτών των περιοχών, παρακρατείτε για την εκτίμηση της ακρίβειας μετά την ταξινόμηση. Οι ακρίβειες που υπολογίζονται, αποτελούν μια πρώτη προσέγγιση του επιπέδου απόδοσης του αλγόριθμου ταξινόμησης.

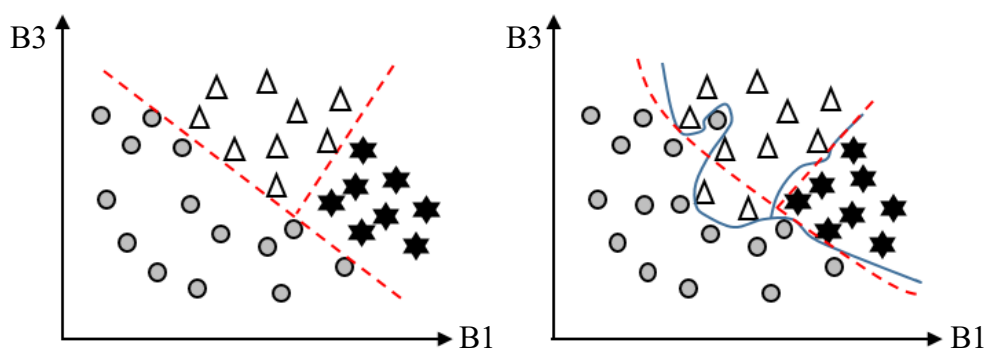
iv. Αξιολόγηση και Αναθεώρηση του συνόλου δεδομένων εκπαίδευσης

Η διαδικασία αξιολόγησης και αναθεώρησης των δειγματοληπτικών δεδομένων, είναι συνήθως μία επαναληπτική διαδικασία κατά την οποία ο αναλυτής αξιολογεί τη φασματική διαχωριστική ικανότητα μεταξύ των τάξεων και αναθεωρεί αν κρίνει αναγκαίο τις στατιστικές περιγραφές κάθε τύπου κατηγορίας, μέχρις ότου να είναι επαρκώς διαχωρίσιμες.



**Διάγραμμα 6:** Επικάλυψη Φασματικών υπογραφών διαφορετικών τάξεων στο δισδιάστατο χώρο (κανάλια 1 και 4) → Αναθεώρηση των δειγμάτων

Στο πλαίσιο αυτής της διαδικασίας, ο αναλυτής επιχειρεί να εντοπίσει πιθανά φασματικά κενά, επικαλύψεις και πέραν του φυσιολογικού μεγέθους, πλεονασμούς. Μελετάται λοιπόν η διαχωριστικότητα μεταξύ τους σε κάθε δυνατό συνδυασμό καναλιών που συνθέτουν ένα δισδιάστατο φασματικό χώρο.



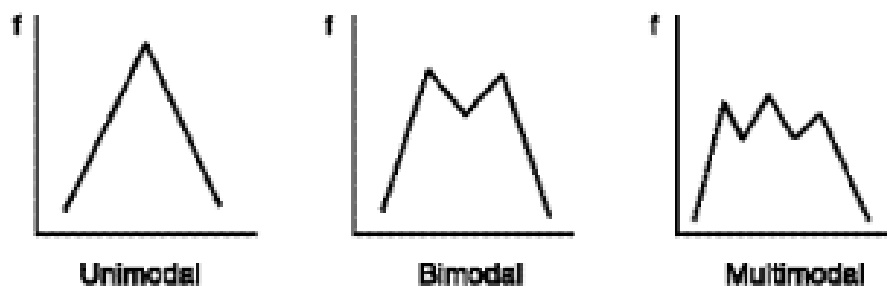
**Διάγραμμα 7:** Περιπτώσεις φασματικής διαφορετικότητας μεταξύ δειγμάτων εκπαίδευσης σε δισδιάστατο φασματικό χώρο και υποψήφιων ορίων διαχωρισμού των κλάσεων για κάθε περίπτωση

Η διαχωριστικότητα, αποτελεί ένα στατιστικό μέτρο κατά το οποίο υπολογίζεται η φασματική απόσταση μεταξύ των κέντρων κάθε δυνατού συνδυασμού κλάσεων και αξιολογείται αν αυτή είναι αρκετά σημαντική για να καταστήσει δύο κλάσεις διακριτές ή σε αντίθετη περίπτωση αν θα πρέπει αυτές να αναθεωρηθούν.

Ακόμα, ένας οπτικός τρόπος προληπτικής αξιολόγησης των δεδομένων που συλλέχθηκαν για εκπαιδευτικούς σκοπούς, είναι μέσω του ιστογράμματος των τιμών τους.

Ιδανική θεωρείται η περίπτωση όπου η συχνότητα εμφάνισης των τιμών ακολουθεί μια μονοτροπική κατανομή (unimodal μορφή), με κύριο χαρακτηριστικό την ύπαρξη μίας ξεκάθαρη κορυφής στην κατανομή. Αν και ιδανική, η συγκεκριμένη περίπτωση σπάνια συναντάται, αφού στην προσπάθεια δημιουργίας τάξεων που να περιγράφονται πλήρως φασματικά και σε σχέση πάντα με την χωρική ανάλυση των δεδομένων, η μεταβλητότητα των τιμών ανακλαστικότητα αυξάνεται επικαλύπτοντας συχνά άλλες κατηγορίες.

Στις περισσότερες περιπτώσεις, η μορφή που σχηματίζουν οι τιμές των δεδομένων εκπαίδευσης, ονομάζεται πολυτροπική (multimodal), η οποία καθιστά τη διαχωριστικότητα των κλάσεων, δύσκολο έργο.



**Διάγραμμα 8:** Μορφές ιστογραμμάτων των δεδομένων εκπαίδευσης (Unimodal (Μονοτροπική), Bimodal (Διτροπική), Multimodal (πολυτροπική))

Περιοχές που περιλαμβάνουν ακούσια περισσότερες από μια φασματικές κατηγορίες, αναγνωρίζονται και επαναπροσδιορίζονται, έτσι παραμένουν μόνο αυτές που είναι φασματικά καθαρές. Παρομοίως, μεμονωμένα ακραία εικονοστοιχεία, μπορούν να διαγραφούν από ορισμένες περιοχές εκπαίδευσης και μερικά σύνολα δεδομένων μπορούν να συγχωνευτούν ή να διαγραφούν.

### 3. Εφαρμογή του Αλγόριθμου Ταξινόμησης

Αφού δημιουργηθεί το σύνολο των δεδομένων εκπαίδευσης και αξιολογηθεί ως αξιόπιστο και πλήρες, το επόμενο βήμα είναι η επιλογή ενός κανόνα ταξινόμησης, ο οποίος θα διαχωρίσει το σύνολο των εικονοστοιχείων που απαρτίζουν την εικόνα στις προεπιλεγμένες τάξεις, σύμφωνα με τους κανόνες απόφασης του.

Ο κανόνας ταξινόμησης, αποτελεί μια μαθηματική έκφραση, η οποία είτε χρησιμοποιεί τα στατιστικά χαρακτηριστικά κάθε κλάσης (μέση τιμή, τυπική απόκλιση, πίνακα συμμεταβλητότητας) και τοποθετείται στην κατηγορία των παραμετρικών κανόνων (οι οποίοι θεωρούν ότι τα δεδομένα εκπαίδευσης για κάθε ομάδα ακολουθούν την κανονική κατανομή), είτε τα αγνοεί και ονομάζεται μη παραμετρικός. Αυτό οδηγεί στη βελτίωση της ακρίβειας της ταξινόμησης, ειδικά σε κλάσεις που παρουσιάζουν εσωτερικά μεγάλη φασματική μεταβλητότητα (κάτι που επηρεάζεται ακόμα περισσότερο αν συνδυαστεί με δεδομένα μέτριας προς χαμηλής χωρικής ανάλυσης), όπως είναι οι αστικές περιοχές.

Παραδείγματα παραμετρικών κανόνων αποτελούν οι αλγόριθμοι της Ελάχιστης Απόστασης, της Απόστασης Mahalanobis και της Μέγιστης Πιθανοφάνειας (Maximum Likelihood), ενώ στην κατηγορία των μη παραμετρικών εντάσσονται οι κανόνες του φασματικού χώρου, του παραλληλεπίπεδου και ο συνδυαστικός κανόνας των Τυχαίων Δασών, στον οποίο θα δοθεί ιδιαίτερη έμφαση στα πιο κάτω κεφάλαια.

### 4. Αξιολόγηση της ακρίβειας της ταξινόμησης

“Η ταξινόμηση δεν είναι πλήρης μέχρις ότου εκτιμηθεί η ακρίβεια της” (Lillesand T.M. et al., 2004). Σε αυτή τη φράση, περικλείεται η σημαντικότητα εκτίμησης της ποιότητας διεκπεραίωσης της αυτόματης ομαδοποίησης ενός πλήθους οντοτήτων, αφού με αυτό τον τρόπο διασφαλίζεται η ακεραιότητα των αποτελεσμάτων σε μία έρευνα.

Πιο κάτω περιγράφονται μερικές από τις αρχές και πρακτικές που χρησιμοποιούνται για την αξιολόγηση της ακρίβειας της ταξινόμησης, σύμφωνα με τους Congalton and Green (1999).

Ένα από τα πιο κοινά μέσα έκφρασης της ακρίβειας της ταξινόμησης, είναι ο πίνακας σφαλμάτων της ταξινόμησης ή αλλιώς πίνακας σύγχυσης (πίν. 2).

**Πίνακας 2:** Παράδειγμα Πίνακα Σύγκυσης τεσσάρων κλάσεων με περιεχόμενα κελιών ( $p_{ij}$ ) (Olofsson et.al., 2014)

|     |         | Reference     |               |               |               |              |
|-----|---------|---------------|---------------|---------------|---------------|--------------|
|     |         | Class 1       | Class 2       | Class 3       | Class 4       | Total        |
| Map | Class 1 | $p_{11}$      | $p_{12}$      | $p_{13}$      | $p_{14}$      | $p_{1\cdot}$ |
|     | Class 2 | $p_{21}$      | $p_{22}$      | $p_{23}$      | $p_{24}$      | $p_{2\cdot}$ |
|     | Class 3 | $p_{31}$      | $p_{32}$      | $p_{32}$      | $p_{34}$      | $p_{3\cdot}$ |
|     | Class 4 | $p_{41}$      | $p_{42}$      | $p_{43}$      | $p_{44}$      | $p_{4\cdot}$ |
|     | Total   | $p_{\cdot 1}$ | $p_{\cdot 2}$ | $p_{\cdot 3}$ | $p_{\cdot 4}$ | 1            |

Οι πίνακες σφαλμάτων συγκρίνουν κατηγορία ανά κατηγορία τη σχέση μεταξύ των δεδομένων αναφοράς (επίγεια αληθή δεδομένα) και τα αντίστοιχα αποτελέσματα που προκύπτουν από τη διαδικασία της ταξινόμησης. Είναι τετραγωνικοί, με τον αριθμό των σειρών και στηλών να ισούται με το πλήθος των κατηγοριών, των οποίων η ακρίβεια ταξινόμησης αξιολογείται.

Διάφορα περιγραφικά μέτρα σχετικά με την απόδοση της ταξινόμησης, μπορούν να εκφραστούν μέσα από τον πίνακα σύγκυσης. Μεταξύ αυτών, η Ολική ακρίβεια (Overall Accuracy), η οποία εκφράζει το συνολικό ποσοστό των ορθά ταξινομημένων εικονοστοιχείων.

**Εξίσωση 1:** 
$$O = \sum_{j=1}^q p_{ij}$$

Η ακρίβεια του χρήστη, είναι ένα ακόμα μέτρο που προκύπτει από τον πίνακα σύγκυσης και εκφράζει το ποσοστό των εικονοστοιχείων που τοποθετήθηκαν σωστά στην κλάση  $i$  ως προς τον συνολικό αριθμό εικονοστοιχείων που τοποθετήθηκαν σε αυτή.

**Εξίσωση 2:** 
$$U_i = p_{ii}/p_{i\cdot}$$

Αυτό που αναφέρεται στη βιβλιογραφία ως ακρίβεια του παραγωγού (Producer's Accuracy), υπολογίζεται ως το πλήθος των ορθά ταξινομημένων εικονοστοιχείων κάθε κατηγορίας, ως προς το σύνολο των εικονοστοιχείων αναφοράς που ανήκαν σε κάθε κλάση.

**Εξίσωση 3:** 
$$P_j = p_{ij}/p_{\cdot j}$$

Τέλος, ο δείκτης kappa αποτελεί μια πολύ-μεταβλητή ανάλυση, η οποία λαμβάνει υπόψη τις πιθανότητες τυχαίας κατηγοριοποίησης των ορθά ταξινομημένων εικονοστοιχείων στις κλάσεις τους και αφού τις συγκρίνει με την πραγματική ταξινόμηση εκφράζει το ποσοστό των σφαλμάτων που απέφυγε η πραγματική ταξινόμηση σε σχέση με την τυχαία. Αυτό το μέτρο, μπορεί επίσης να εκφραστεί για κάθε κατηγορία ξεχωριστά.

**Εξίσωση 4:**

$$K = \frac{N * \sum_{i=1}^r X_{ii} - \sum_{i=1}^r (X_{i+} * X_{+i})}{N^2 - \sum_{i=1}^r (X_{i+} * X_{+i})}$$

Όπου :

N → Πλήθος των εικονοστοιχείων

i → Πλήθος κλάσεων ταξινόμησης

r → Αριθμός γραμμών

X<sub>ii</sub> → Τα στοιχεία που βρίσκονται στη διαγώνιο του πίνακα σύγχυσης

X<sub>+i</sub>, X<sub>i+</sub> → Αθροίσματα κατά γραμμή και κατά στήλη αντίστοιχα

Στην ιδανική περίπτωση, τα δείγματα που χρησιμοποιούνται για την εκπαίδευση και έλεγχο ενός ταξινομητή, αντιπροσωπεύουν την πραγματικότητα (επίγεια αληθή δεδομένα). Βασισμένα σε ένα σύνολο αναφοράς, τα δεδομένα αυτά υπόκεινται σε ένα μέτρο αβεβαιότητας, επομένως θεωρείται αναγκαία η αξιολόγηση αυτού του μέτρου.

Δύο από τις πιθανές πηγές της αβεβαιότητας στο σύνολο δεδομένων εκπαίδευσης/ελέγχου, σχετίζονται αφενός με τη γεωγραφική κατανομή (Pontius, 2010), και αφετέρου συνδέονται με την ερμηνεία των δεδομένων αναφοράς (Pontius & Lippitt, 2006).

Η αβεβαιότητα της φωτοερμηνείας, διακρίνεται σε δυο κατηγορίες: Το σφάλμα του ερμηνευτή (σφάλμα αντιστοίχισης της σωστής τάξης αναφοράς στη χωρική οντότητα) και το σφάλμα μεταβλητότητας του ερμηνευτή, δηλαδή της διαφοράς μεταξύ της τάξης

αναφοράς που ανατίθεται στην ίδια χωρική οντότητα από διαφορετικούς ερμηνευτές (Olofsson et al., 2014).

## 5. Διαμόρφωση τελικών θεματικών χαρτών

### 2.1.2 Τεχνικές ταξινόμησης

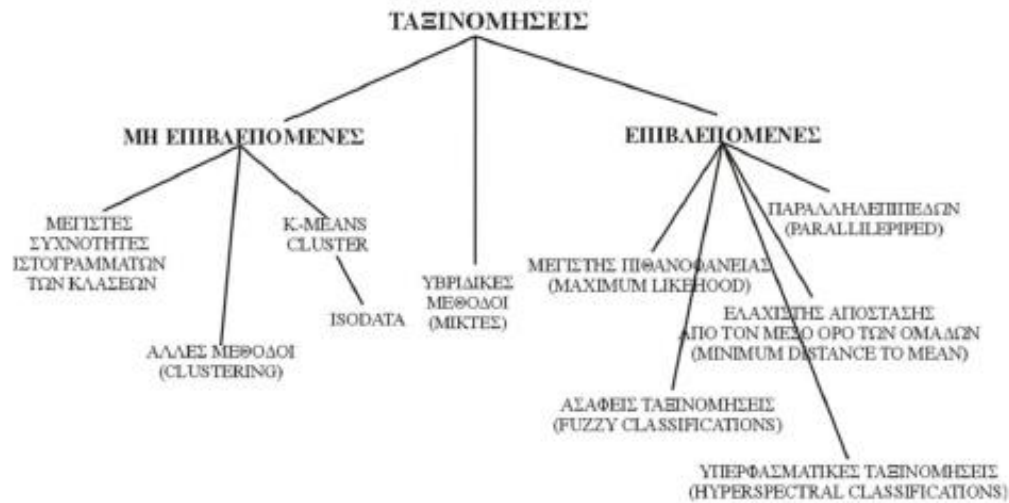
Η πρώτη τεχνική ταξινόμησης στην οποία θα αναφερθούμε, είναι γνωστή ως Επιβλεπόμενη (Supervised). Σε αυτό τον τύπο ταξινόμησης, η διαδικασία αυτόματης κατηγοριοποίησης βασίζεται σε αριθμητικές περιγραφές των διαφόρων τύπων κάλυψης, που προσδιορίζονται από τον αναλυτή στον αλγόριθμο του ταξινομητή.

Πιο συγκεκριμένα, αντιπροσωπευτικά δείγματα τοποθεσιών γνωστών τύπων κάλυψης, χρησιμοποιούνται για την κατάρτιση ενός αριθμητικού κλειδιού, το οποίο περιγράφει τα φασματικά χαρακτηριστικά για κάθε αντικείμενο ενδιαφέροντος. Στη συνέχεια, κάθε εικονοστοιχείο (ν-διάστατο διάνυσμα στον φασματικό χώρο) στο σύνολο δεδομένων συγκρίνεται αριθμητικά με την κάθε προκαθορισμένη κατηγορία στο αριθμητικό κλειδί και επισημαίνεται με το όνομα της κατηγορίας με την οποία “μοιάζει” περισσότερο (Lillesand T.M. et al., 2004).

Στον αντίποδα των αρχών της επιβλεπόμενης ταξινόμησης, βρίσκεται η δεύτερη βασική κατηγορία που ονομάζεται Μη επιβλεπόμενη. Οι αλγόριθμοι που εμπίπτουν σε αυτή την κατηγορία ταξινόμησης, αποβλέπουν στην εξαγωγή των κυρίων φασματικών κλάσεων που υπάρχουν σε μία δορυφορική εικόνα, βάση των φυσικών τους ορίων (Μερτίκας Σ.,2006).

Διαδικαστικά διαφέρουν από τις επιβλεπόμενες ταξινομήσεις, ως προς το γεγονός πως δεν απαιτούν δείγματα εικονοστοιχείων που θα εκπαιδεύσουν τον εκάστοτε αλγόριθμο ταξινόμησης και κατά συνέπεια δεν προκαθορίζεται ο τύπος των ομάδων γης που θα δημιουργηθούν, παρά μόνο ο αριθμός τους. Έτσι ο αναλυτής ορίζει το τι αντιπροσωπεύει η κάθε κλάση, μετά το πέρας της διαδικασίας ταξινόμησης (Περάκης Γ. Κ., 2015).

Αμφότεροι οι Μερτίκας Π.Σ. (2006) και οι Lillesand και Kiefer (1994), αναφέρουν άλλες μεθόδους ταξινόμησης που έχουν αναπτυχθεί και περιλαμβάνουν πτυχές και των δύο βασικών κατηγοριών, στοχεύοντας στη βελτίωση της ακρίβειας ή της αποτελεσματικότητας (ή και των δύο), των αποτελεσμάτων της ταξινόμησης. Αυτές οι μέθοδοι ανήκουν στην κατηγορία των Υβριδικών Ταξινομήσεων.



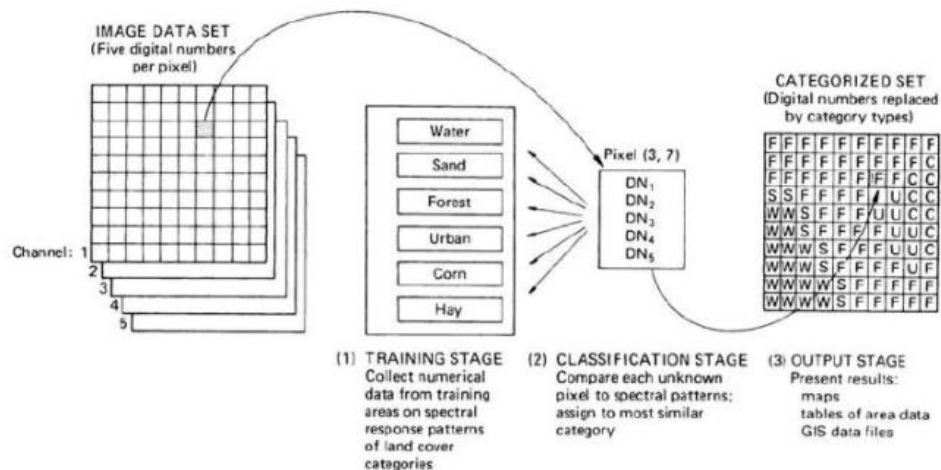
**Διάγραμμα 9:** Σχεδιάγραμμα με τους βασικότερους και πιο ευρέως χρησιμοποιούμενους αλγόριθμους ανά κατηγορία ταξινόμησης

Με την πάροδο των χρόνων, υπήρξαν πολλές συγκρίσεις μεταξύ των δυο κύριων τεχνικών ταξινόμησης (Επιβλεπόμενη και Μη Επιβλεπόμενη) στη βιβλιογραφία, που δείχνουν ότι οι επιβλεπόμενες μέθοδοι όπως η Μέγιστη Πιθανοφάνεια (Maximum Likelihood), ο αλγόριθμος Ελάχιστης Απόστασης, τα Support Vector Machines (SVM) και τα Δυαδικά δέντρα αποφάσεων (DT), ξεπερνούν σε απόδοση τις μη επιβλεπόμενες μεθόδους (Szuster et al., 2011; Khatami et al., 2016)



## 2.2 Παραμετρικοί αλγόριθμοι ταξινόμησης

Στην κατηγορία της επιβλεπόμενης ταξινόμησης, όπως προαναφέρθηκε οι αλγόριθμοι χρησιμοποιούν εικονοστοιχεία γνωστής ταυτότητας ως δείγματα εκπαίδευσης, των οποίων τα φασματικά χαρακτηριστικά σε κάθε κανάλι θα οδηγήσουν στην ομαδοποίηση κάθε εικονοστοιχείου άγνωστης ταυτότητας.



**Διάγραμμα 10:** Βασικά Στάδια Επιβλεπόμενης Ταξινόμησης (Lillesand T.M. et al., 2004)

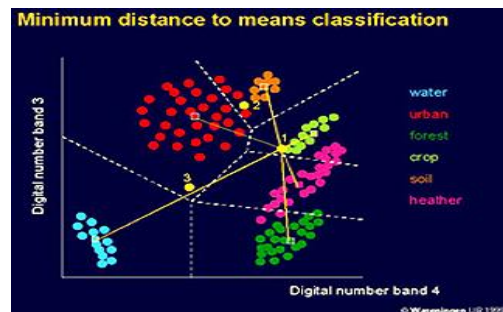
Ο τρόπος κατηγοριοποίησης των εικονοστοιχείων που απαρτίζουν μία δορυφορική εικόνα, καθορίζεται από τις αρχές του χρησιμοποιούμενου κάθε φορά, αλγόριθμου ταξινόμησης. Υπάρχουν πολλά είδη ταξινομητών, παρόλα αυτά θα αναφερθούμε σε αυτούς που παραδοσιακά χρησιμοποιούνται στις περισσότερες εφαρμογές δημιουργίας θεματικών χαρτών. Αυτοί είναι ο αλγόριθμος της ελάχιστης απόστασης, ο αλγόριθμος της απόστασης Mahalanobis και ο αλγόριθμος της Μέγιστης πιθανοφάνειας (Maximum Likelihood).

### 2.2.1 Αλγόριθμος Ελάχιστης Απόστασης

Σε αυτή τη μέθοδο, η κατηγοριοποίηση των εικονοστοιχείων βασίζεται στην ευκλείδεια απόσταση του εκάστοτε εικονοστοιχείου, από τη μέση φασματική τιμή κάθε ομάδας στην ταξινόμηση και ειδικότερα στην ομάδα από την οποία απέχει τη μικρότερη φασματική απόσταση.

Συγκεκριμένα, η μέση τιμή κάθε κατηγορίας κάλυψης/χρήσης γης, εκτιμάται από τις ραδιομετρικές τιμές των δειγμάτων που ο αναλυτής επιλέγει για την εκπαίδευση του αλγόριθμου (Περάκης Γ. Κ., 2015) και αντιστοιχούν στο κέντρο κάθε περιοχής εκπαίδευσης. Επομένως, σε περίπτωση πεπερασμένου πλήθους ομάδων, θα προσδιοριστούν αντίστοιχοι πλήθους μέσες φασματικές τιμές (κέντρα), από τις οποίες θα υπολογιστεί η ευκλείδεια απόσταση κάθε εικονοστοιχείου και θα αποφασιστεί η κλάση στην οποία θα τοποθετηθεί.

$$\begin{aligned} 1\text{-dim} &: \sqrt{(p - q)^2} \\ 2\text{-dim} &: \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2} \\ n\text{-dim} &: \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \end{aligned}$$

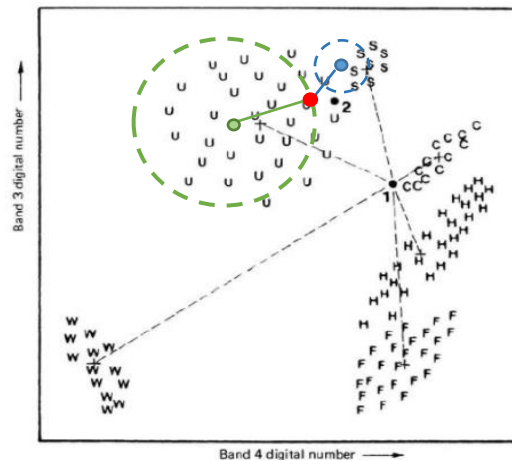


**Διάγραμμα 11:** Εξίσωση Φασματικής Ευκλείδειας απόστασης στο N-διάστατο φασματικό χώρο (αριστερά) και παράδειγμα προσδιορισμού των ορίων διαχωρισμού των θεματικών κλάσεων (δεξιά)

Σύμφωνα με τους Περάκη Γ.Κ. (2015) και Lillesand T.M. et al., (2004), η στρατηγική ταξινόμησης της ελάχιστης απόστασης από τους μέσους κάθε κλάσης, είναι μαθηματικά απλή και υπολογιστικά αποδοτική αλλά έχει συγκεκριμένους περιορισμούς. Από τα πιο σημαντικά μειονεκτήματα του είναι το γεγονός ότι δεν είναι ευαίσθητος στο στατιστικό μέτρο της διακύμανσης των δεδομένων φασματικής απόκρισης. Αυτό είναι ένα πρόβλημα που μπορεί να προκαλέσει αύξηση του σφάλματος ταξινόμησης.

Στο παράδειγμα του διαγράμματος 12, το κόκκινο εικονοστοιχείο θα τοποθετηθεί στην κλάση S, παρά το γεγονός ότι η μεταβλητότητα που υπάρχει στην κατηγορία U υποδηλώνει ότι θα ήταν σοφότερη η επισήμανση του εικονοστοιχείου σε αυτή την κατηγορία.

Εξαιτίας τέτοιων προβλημάτων, αυτός ο ταξινομητής χρησιμοποιείται όλο και λιγότερο σε εφαρμογές όπου η διασπορά των εικονοστοιχείων που απαρτίζουν μία θεματική κατηγορία, είναι μεγάλη.



**Διάγραμμα 12:** Ταξινόμηση με βάση τον Αλγόριθμο της ελάχιστης απόστασης

### 2.2.2 Αλγόριθμος Απόστασης Mahalanobis

Όπως αναφέρθηκε στην προηγούμενη παράγραφο, οι φασματικές αποστάσεις που υπολογίζονται από τον αλγόριθμο της ελάχιστης απόστασης για την ταξινόμηση των αγνώστων εικονοστοιχείων δεν λαμβάνουν υπόψη την διακύμανση της κλάσης, παρά μόνο την φασματική απόσταση από το κέντρο της.

Παραλλαγή του εν λόγω αλγόριθμου αποτελεί η απόσταση Mahalanobis, όπου κατά τον υπολογισμό της απόστασης, λαμβάνει υπόψη ότι η τιμή της διακύμανσης σε κάθε κατεύθυνση είναι διαφορετική και υπολογίζει επίσης την τιμή της συνδιακύμανσης μεταξύ των μεταβλητών, μειώνοντας με αυτόν τον τρόπο τη γνωστή ευκλείδεια απόσταση για μη συσχετισμένες μεταβλητές.

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}$$

**Εξίσωση 5.** Απόσταση Mahalanobis

### 2.2.3 Αλγόριθμος Μέγιστης Πιθανοφάνειας

Ο αλγόριθμος μέγιστης πιθανοφάνειας (Maximum Likelihood), χρησιμοποιεί τα εκπαιδευτικά δεδομένα για την εύρεση των φασματικών μέσων τιμών (κέντρων) και τον υπολογισμό της συνδιασποράς τους για κάθε κλάση και ακολούθως τα χρησιμοποιεί για την εκτίμηση της πιθανότητας κάθε εικονοστοιχείου να εμπίπτει σε καθεμιά από τις προκαθορισμένες τάξεις (Παρχαρίδης Ι. (2015)).

Η βασική εξίσωση της μέγιστης πιθανοφάνειας, έχει ως βάση δύο υποθέσεις . Αρχικά ότι κάθε εικονοστοιχείο παρουσιάζει ίσες πιθανότητες να ανήκει σε κάθε τάξη και ότι τα ιστογράμματα των εκπαιδευτικών τιμών των τάξεων σε κάθε κανάλι, είναι μονοτροπικής μορφής και περιγράφονται από κανονικές κατανομές. Αν η υπόθεση αυτή δεν ισχύει και τα δεδομένα παρουσιάζουν για μια ή περισσότερες θεματικές κατηγορίες περισσότερες από μία κορυφές στο ιστόγραμμα, τότε η απόδοση του επηρεάζεται αρνητικά. Συνεπώς, όσο πιο ακριβής και αξιόπιστη είναι η συλλογή των δεδομένων εκπαίδευσης, τόσο πιο ισχυρή και αποτελεσματική γίνεται η εφαρμογή της εν λόγω μεθόδου.

Σε περιγραφή του τρόπου λειτουργίας του, αφού ο κανόνας ταξινόμησης βρει το βέλτιστο τρόπο προσαρμογής μιας κανονικής κατανομής στο σύνολο δεδομένων κάθε κλάσης, ούτως ώστε να μεγιστοποιηθούν οι συναρτήσεις πιθανότητας τους, εφαρμόζεται η πιο κάτω εξίσωση :

**Εξίσωση 6:** 
$$p(x_k | i) = \frac{1}{\sqrt{2\pi} \sqrt{|M_i|}} \exp\left(-\frac{1}{2} D^2\right)$$

Όπου :

$p(X_k | i)$  : η πιθανότητα ένα εικονοστοιχείο  $X_k$  να ανήκει στην τάξη  $i$

$|\sqrt{M_i}|$  : ο πίνακας μεταβλητότητας της κατηγορίας  $i$  (προκύπτει από τον υπολογισμό της τυπικής απόκλισης)

$D^2 = \left(\frac{X - \mu_i}{\sigma_i}\right)^2$  : η απόσταση mahalanobis μεταξύ του εικονοστοιχείου  $k$  και του φασματικού κέντρου κάθε τάξης

Ακολούθως, αφού υπολογιστούν οι πιθανότητες του εικονοστοιχείου για την κάθε κλάση, η παρακάτω εξίσωση καθορίζει το που τελικά θα καταλήξει το εικονοστοιχείο μέσω του υπολογισμού της a-posteriori πιθανότητας και εύρεσης της μέγιστης αυτής τιμής.

**Εξίσωση 7:**

$$L(i | x_k) = \frac{P_i p(x_{k|i})}{\sum_{j=1}^c P_j p(x_k | j)}$$

Όπου :

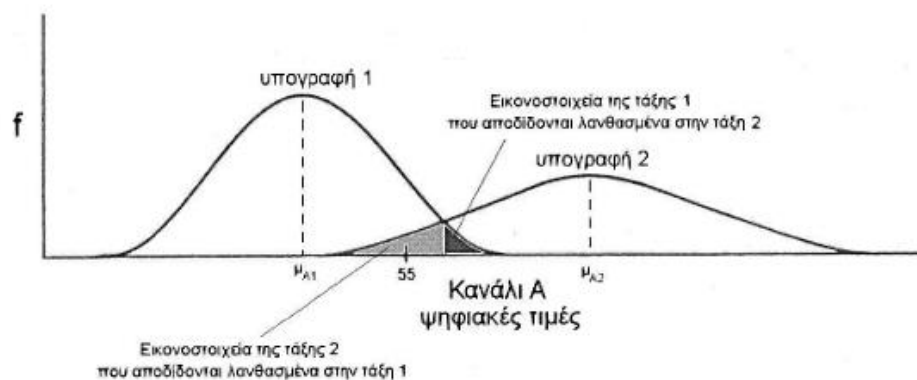
$L(i | X_k)$  : η a posteriori πιθανότητα ένα εικονοστοιχείο  $X_k$  να ανήκει στην τάξη  $i$

$P_i$  : η πιθανότητα κάθε κλάσης

$p(X_k | i)$  : η πιθανότητα ένα εικονοστοιχείο  $X_k$  να ανήκει στην τάξη  $i$

$c$  : ο συνολικός αριθμός των κλάσεων

Πλεονέκτημα του αλγόριθμου μέγιστης πιθανοφάνειας, αποτελεί το γεγονός ότι βασίζεται στα στατιστικά στοιχεία που υπολογίζει (μέση τιμή και πίνακα συμμεταβλητότητας) για να παρέχει μία εκτίμηση των φασματικά επικαλυπτόμενων περιοχών που προκύπτουν. Με αυτό τον τρόπο, αποτρέπει εικονοστοιχεία που εμπίπτουν στα επικαλυπτόμενα τμήματα δύο κλάσεων να ταξινομηθούν σε λάθος τάξη (διάγραμμα 13), αφού λαμβάνει υπόψη τη μεταβλητότ



**Διάγραμμα 13:** Λανθασμένη ταξινόμηση εικονοστοιχείου που εμπίπτει σε φασματικά επικαλυπτόμενη περιοχή δύο υπογραφών Πηγή: Παρχαρίδης Ι. (2015), “Αρχές Δορυφορικής Τηλεπισκόπησης - Θεωρία και Εφαρμογές”

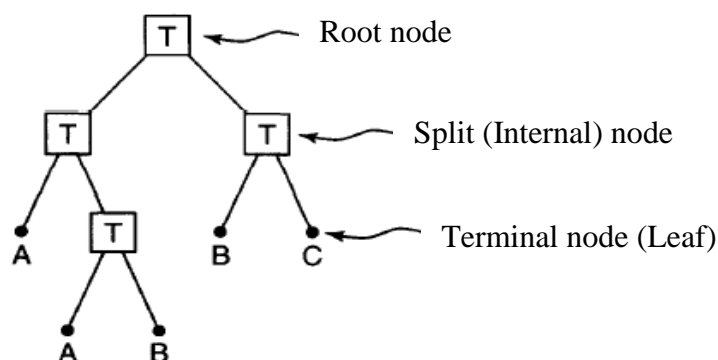
## 2.2.4 Δέντρα Απόφασης

Η κατηγοριοποίηση των δεδομένων με τη χρήση ενός μηχανισμού ιεραρχικής διάσπασης, έχει ως σκοπό την ολοκληρωμένη κατανόηση των σχέσεων μεταξύ των οντοτήτων σε διαφορετικά επίπεδα λεπτομέρειας (Hastie T. et al, 2009). Η απλούστερη αναπαράσταση τέτοιου είδους αλγόριθμων, παίρνει τη μορφή ενός ανεστραμμένου δέντρου, στο οποίο τα διάφορα επίπεδα ιεραρχίας αντιπροσωπεύουν τα διαφορετικά επίπεδα ταξινόμησης.

Ο σχεδιασμός ενός δέντρου αποφάσεων όταν εφαρμόζεται σε πολύ-φασματικά δεδομένα, βασίζεται στη γνώση των φασματικών ιδιοτήτων της κάθε κλάσης αλλά και των σχέσεων που υπάρχουν μεταξύ των κλάσεων. Σύμφωνα με τους Tso B. and Mather P.M. (2009), το κύριο πλεονέκτημα της χρήσης ενός δέντρου απόφασης για την πραγματοποίηση μίας ταξινόμησης, είναι ότι η δομή του καθιστά ευκολότερη την κατανόηση και την ερμηνεία της σχέσης μεταξύ των δεδομένων εισόδου και των πληροφοριών που εξάγονται από την ανάλυση που διεξάγεται στο μεταξύ.

### *i. Βασικά Χαρακτηριστικά*

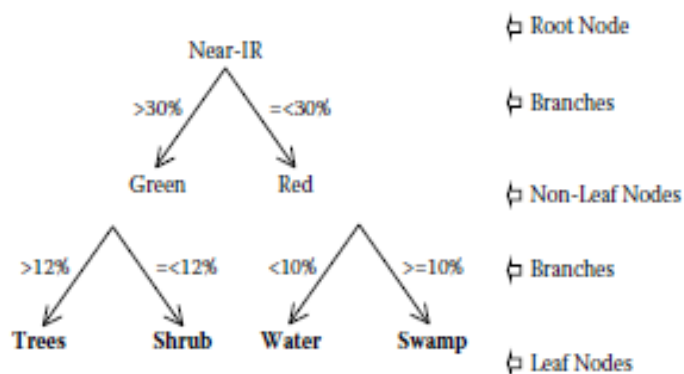
Ένα δέντρο αποφάσεων, αποτελείται από το ριζικό κόμβο (root node), ένα σύνολο από εσωτερικούς κόμβους διαχωρισμού (split nodes) και τους τερματικούς κόμβους ή αλλιώς φύλλα (terminal nodes – leaves). Η ρίζα του δέντρου (root node) και οι εσωτερικοί κόμβοι οι οποίοι αναφέρονται στη βιβλιογραφία και ως μη τερματικοί κόμβοι, συνδέονται με τα στάδια λήψης των αποφάσεων, ενώ τα φύλλα (terminal node ή leaf) αντιπροσωπεύουν την κλάση στην οποία κατατάσσεται το κάθε εικονοστοιχείο.



**Διάγραμμα 14:** Συστατικά μέρη δέντρου απόφασης

Η διαδικασία της ταξινόμησης, υλοποιείται από ένα σύνολο κανόνων οι οποίοι καθορίζουν τη διαδρομή που θα ακολουθήσει το κάθε εικονοστοιχείο. Συγκεκριμένα, σε κάθε εσωτερικό κόμβο πρέπει να ληφθεί μια απόφαση βάσει των φασματικών κριτηρίων που θέτει ο αναλυτής και έτσι τα εικονοστοιχεία συνεχίζουν την πορεία τους προς τα φύλλα και την ομάδα που φασματικά ανήκουν.

Στο διάγραμμα 16, απεικονίζεται ένας απλός ταξινομητής δέντρου αποφάσεων, όπου βλέπουμε πως χρησιμοποιείται η τιμή ανακλαστικότητα των εικονοστοιχείων ως δεδομένα εισόδου, προκειμένου να αποφασιστεί σε ποια ομάδα (Δέντρο, Θάμνος, Βάλτος, Νερό) ανήκει. Σε κάθε εσωτερικό κόμβο διαχωρισμού, επιλέγεται ένας διάυλος στον οποίο θα εξεταστούν τα φασματικά κριτήρια που θέτει ο αναλυτής.

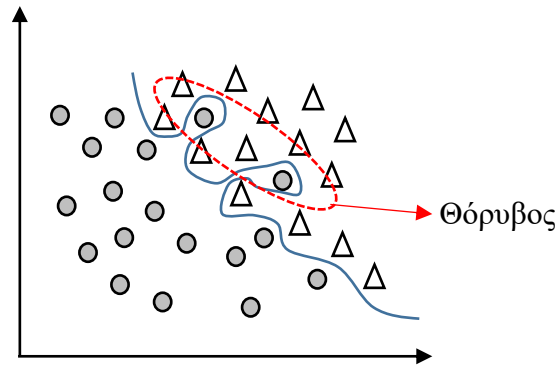


**Διάγραμμα 15:** Παράδειγμα δέντρου απόφασης ιεραρχικής δομής

Αν τα φασματικά κριτήρια εξετάζονται ως προς ένα χαρακτηριστικό των δεδομένων εισόδου (π.χ. το κανάλι Near –IR) σε κάθε εσωτερικό κόμβο διαχωρισμού, τότε το δέντρο αποφάσεων ονομάζεται μονό-μεταβλητό. Αν όμως χρησιμοποιηθούν περισσότερα από ένα χαρακτηριστικά για την επιλογή της πορείας του κάθε εικονοστοιχείου, το πολύ-μεταβλητό δέντρο αποφάσεων που κατασκευάζεται είναι πιο συμπαγές και οδηγεί συνήθως στην παραγωγή ακριβέστερων αποτελεσμάτων (Friedl M.A. and Brodley C.E., 1997 ; Swain and Hauska, 1969 ; Brodley and Utgoff, 1995).

Παρόλο που οι εν λόγω αλγόριθμοι παρουσιάζουν υψηλές αποδόσεις, το βασικό πρόβλημα που αντιμετωπίζεται συνήθως κατά την εφαρμογή τους, είναι η ευαισθησία τους στο φαινόμενο overfitting (διάγραμμα 16), το οποίο αναφέρεται στο πόσο πολύπλοκη και ευέλικτη εκπαιδεύεται να γίνει μια συνάρτηση. Η επιλογή ενός τέτοιου αλγόριθμου, συνεπάγει ότι θα εντοπίσει και θα διαχωρίσει τέλεια τις δύο κλάσεις αλλά

θα προσαρμοστεί ακόμα και στο θόρυβο που βρίσκεται μέσα στα δεδομένα εκπαίδευσης και αποτελεί τιμές που πιθανώς να μην ξανά εμφανιστούν.



**Διάγραμμα 16:** Φαινόμενο Overfitting – Προσαρμογή συνάρτησης στον θόρυβο που υπάρχει στα δεδομένα εκπαίδευσης → Δημιουργία πολύπλοκων ορίων διαχωρισμού, που θα προκαλέσει σφάλματα στην ταξινόμηση των άγνωστης τάξης εικονοστοιχείων

Ακόμα ένα βασικό μειονέκτημα του συγκεκριμένου κανόνα ταξινόμησης αποτελεί η ύπαρξη μεγάλου αριθμού biases (τυχαία σφάλματα – δεν περιγράφονται από νόμους της στατιστικής).

Τέλος, ο ιεραρχικός χαρακτήρας αυτής της διαδικασίας, έχει ως αποτέλεσμα την μετάδοση ενός σφάλματος που ξεκίνησε από την κορυφή (ρίζα) του δέντρου και προκάλεσε τη δημιουργία λανθασμένων αποφάσεων και εκτιμήσεων σε όλους τους κόμβους διαχωρισμού κάτω από αυτή, μέχρι το εκάστοτε φύλλο.

### 2.3 Συνδυαστικοί ταξινομητές στην Ταξινόμηση πολυφασματικών εικόνων

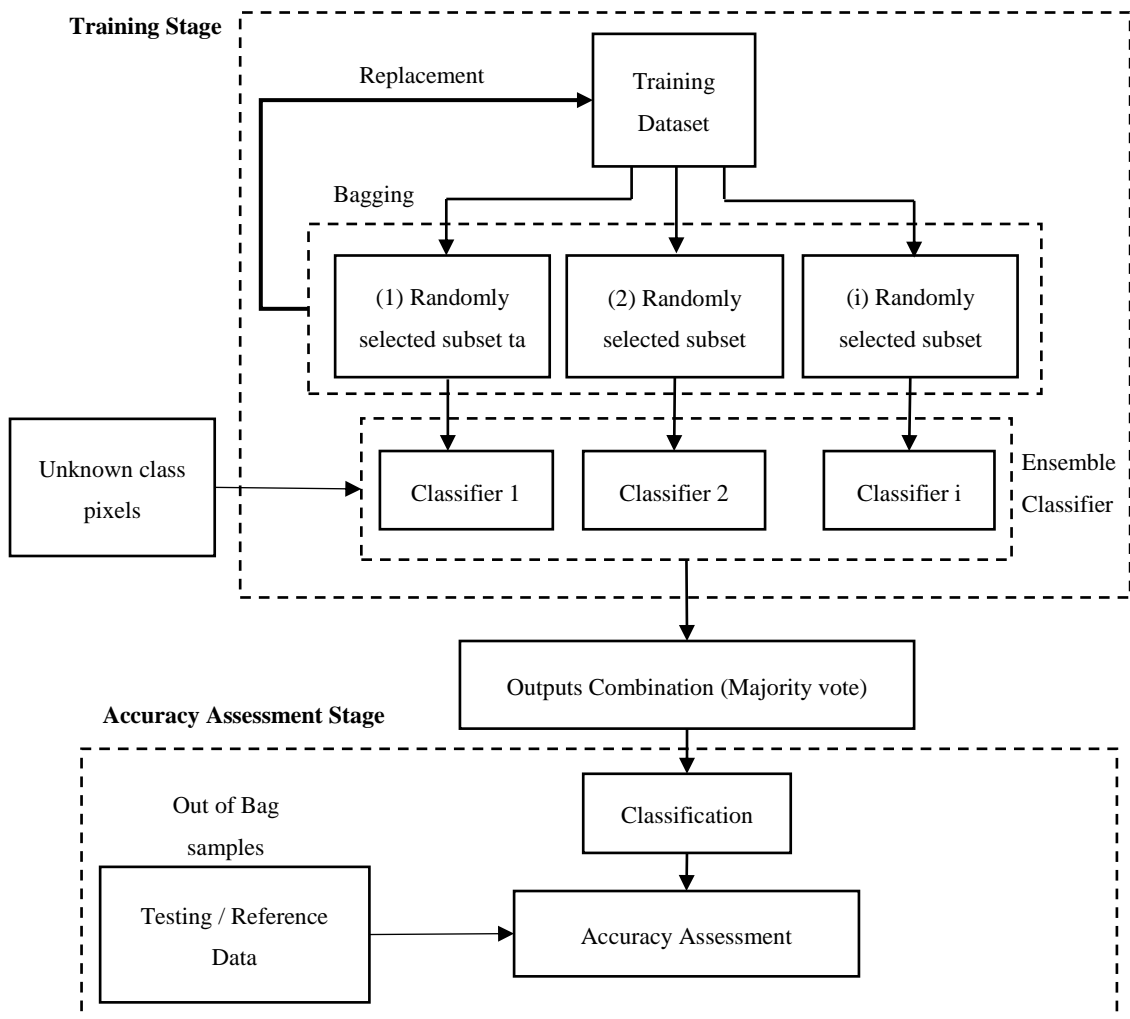
Τα τελευταία χρόνια, η προσοχή της ερευνητικής κοινότητας έχει στραφεί σε ένα νέο είδος αλγόριθμων οι οποίοι ονομάζονται συνδυαστικοί. Η ιδέα πίσω από αυτή την έννοια, είναι η δημιουργία ενός εκτιμητικού μοντέλου, μέσω της συνένωσης πολλαπλών όμοιων ή ανόμοιων μεταξύ τους μοντέλων. Το συνδυαστικό μοντέλο που προκύπτει, εκπαιδεύεται είτε μέσω της διαδικασίας Bagging (ή αλλιώς Bootstrap Aggregation) κατά την οποία ο κάθε ταξινομητής από τον οποίο απαρτίζεται το συνδυαστικό μοντέλο, εκπαιδεύεται με βάση ένα τυχαία επιλεγμένο υποσύνολο του συνόλου των δεδομένων



εκπαίδευσης (Rokach L. (2009)), είτε μέσω της διαδικασίας Boosting, όπου το μοντέλο εκπαιδεύεται επανειλημμένα χρησιμοποιώντας ολόκληρο το σύνολο των δεδομένων εκπαίδευσης.

Σύμφωνα με έρευνες των Briem et al., (2002) και Miao et al., (2012), τα συνδυαστικά μοντέλα που προκύπτουν από τις εν λόγω διαδικασίες εκπαίδευσης, ξεπερνούν σε απόδοση κάθε ένα από τα μοντέλα από τα οποία απαρτίζονται και είναι επίσης πιο σταθερά και αποτελεσματικά έναντι στο θόρυβο και στη φασματική διακύμανση που παρουσιάζουν στις περισσότερες περιπτώσεις τα δεδομένα εκπαίδευσης.

Στο διάγραμμα 16, παρουσιάζεται λεπτομερώς η διαδικασία εκπαίδευσης, εφαρμογής και τέλος της εκτίμησης της ακρίβειας της απόδοσης ενός συνδυαστικού μοντέλου ταξινόμησης, που ακολούθησε τη μέθοδο bagging.



**Διάγραμμα 17:** Διαδικασίας εκπαίδευσης και ελέγχου της ακρίβειας ενός συνδυαστικού αλγόριθμου με τη μέθοδο bagging

### 2.3.1 Τυχαία Δάση (Random Forest)

Ο ταξινομητής Random Forest (RF) , είναι ένας από τους πιο γνωστούς ομοιογενής, συνδυαστικούς ταξινομητές, ο οποίος χρησιμοποιεί πολλά, ασυσχέτιστα μεταξύ τους δέντρα αποφάσεων για να πραγματοποιήσει μία εκτίμηση (Belgiu M. and Dragut L. (2016)). Η βασική ιδέα πίσω από το μοντέλο Random Forest, είναι η μείωση της συσχέτισης μεταξύ των ταξινομητών που το απαρτίζουν και του φαινομένου overfitting που προκαλεί την προσαρμογή του μοντέλου στον θόρυβο που υπάρχει στα δεδομένα.

#### 2.3.1.1 Κατασκευή του αλγόριθμου Random Forest

Η ανεξάρτητη ανάπτυξη των δέντρων που απαρτίζουν το συνδυαστικό μοντέλο, βασίζεται στην επιλογή τυχαίων υποσυνόλων δεδομένων εκπαίδευσης (Bagging), με αντικατάσταση. Αυτό σημαίνει ότι μέρος του συνόλου των δειγμάτων που χρησιμοποιούνται για την εκπαίδευση ενός δέντρου, μπορεί να επιλεγεί περισσότερες από μία φορές για την εκπαίδευση και άλλων δέντρων μέσα στο σύμπλεγμα, ενώ άλλα δείγματα ενδέχεται να μην επιλεγούν καθόλου. Σε κάθε περίπτωση, τα υποσύνολα εκπαίδευσης που δημιουργούνται είναι διαφορετικά μεταξύ τους και κατά συνέπεια στατιστικά ανεξάρτητα (Rokach L. (2010))

Η δεύτερη πηγή τυχειότητας στο μοντέλο Random Forest, αφορά την τυχαία επιλογή των μεταβλητών (καναλιών), βάση των οποίων θα τεθούν τα φασματικά όρια διαχωρισμού σε κάθε εσωτερικό κόμβο (internal splitting node). Το πλήθος των καναλιών που χρησιμοποιούνται σε κάθε δέντρο απόφασης, καθορίζεται από το σύνολο των μεταβλητών ( $V$ ) που χρησιμοποιούνται στην ταξινόμηση και συνηθίζεται να είναι  $mtry = \sqrt{V}$ .

Η επιλογή ενός υποσυνόλου εκπαίδευσης, επιδρά θετικά στη μείωση του υπολογιστικού χρόνου ενώ η επιλογή μερικώς ή εξ' ολοκλήρου διαφορετικού υποσυνόλου δεδομένων εκπαίδευσης για κάθε δέντρο, βοηθά στη δημιουργία διαφορετικών εκτιμητών, ικανών να προσαρμοστούν σε διαφορετικά είδη δεδομένων, για να βελτιωθεί έτσι η απόδοση του αλγορίθμου (Criminisi et al., 2011). Ακόμα, σύμφωνα με τους Hastie T. et al, (2009), η επιλογή κάθε φορά  $mtry$  τυχαίων μεταβλητών σε κάθε εσωτερικό κόμβο διαχωρισμού (internal splitting node), βοηθά επίσης στη μείωση της συσχέτισης ( $\rho$ ) μεταξύ κάθε πιθανού ζεύγους δέντρων και κατ' επέκταση μείωση της μεγάλης αστάθειας (ύπαρξη

τυχαίων σφαλμάτων) που παρουσιάζουν τα δενδροειδές μοντέλα από τα οποία απαρτίζεται το δάσος.

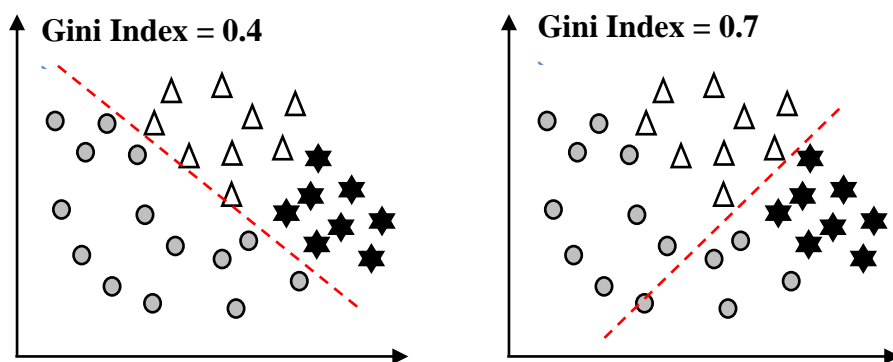
$$\overline{\sigma^2} = \rho\sigma^2 + \left(\frac{1-\rho}{B}\right)\sigma^2$$

**Εξίσωση 8.** Μέση διακύμανση του αλγόριθμου τυχαίου δάσους

Στην εξίσωση 5, βλέπουμε τη μέση διακύμανση ( $\overline{\sigma^2}$ ) που υπάρχει στο δάσος και μπορούμε πολύ εύκολα να συμπεράνουμε ότι αν διατηρήσουμε όλους τους όρους σταθερούς και αυξήσουμε το πλήθος των δέντρων ( $B$ ) που απαρτίζουν το δάσος, τότε ο δεύτερος όρος τείνει στο 0 και μας μένει ο 1ος όρος. Η μείωση της συσχέτισης ( $\rho$ ) μεταξύ κάθε πιθανού ζεύγους δέντρων, επιτυγχάνεται μέσω της επιλογής τυχαίου πλήθους  $m$  μεταβλητών για κάθε εσωτερικό κόμβο διαχωρισμού

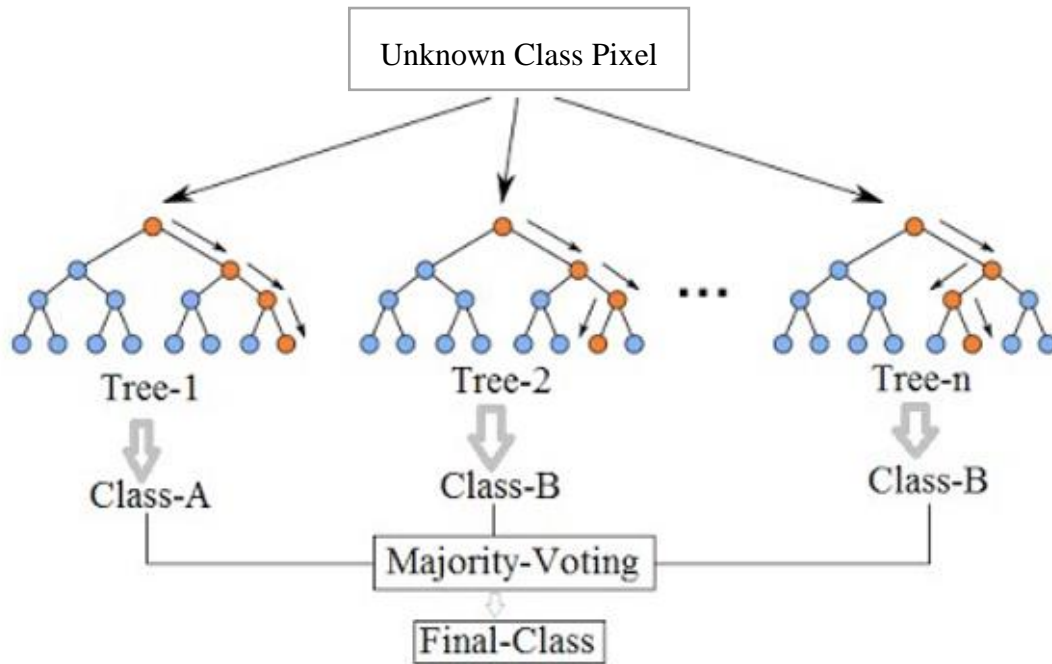
Αφού επιλεγούν τα κανάλια και τα υποσύνολα των δεδομένων εκπαίδευσης που θα χρησιμοποιηθούν για την ανάπτυξη του κάθε δέντρου, σε κάθε εσωτερικό κόμβο επιλέγεται το σημείο διαχωρισμού μεταξύ των υποψήφιων φασματικών κλάσεων.

Τα κριτήρια που θα καθορίσουν τη διαδρομή αυτού του εικονοστοιχείου μέχρι έναν τερματικό κόμβο (φύλλο), καθορίζονται από τα όρια διαχωρισμού. Αυτά προσδιορίζονται σε κάθε εσωτερικό κόμβο μέσα από μια επαναληπτική διαδικασία, κατά την οποία διάφορα όρια εξετάζονται και αξιολογούνται μέσω του υπολογισμού του δείκτη gini (διάγραμμα 18). Επιλέγεται το όριο με την χαμηλότερη τιμή του δείκτη, αφού με αυτό τον τρόπο επιτυγχάνεται η υπολογιστική ταχύτητα του αλγόριθμου, χωρίς να σπαταλιέται χρόνος που δε θα οδηγήσει το συντομότερο σε μία απόφαση.



**Διάγραμμα 18:** Υπολογισμός δείκτη Gini για την εύρεση των βέλτιστων ορίων διαχωρισμού

Η τελική απόφαση κατηγοριοποίησης του κάθε εικονοστοιχείου, λαμβάνεται βάση της ψήφου πλειοψηφίας στο σύνολο των αποφάσεων από τα δέντρα (διάγραμμα 19).



**Διάγραμμα 19:** Παράδειγμα ταξινόμησης άγνωστης τάξης εικονοστοιχείου με τον αλγόριθμο Τυχαίων δασών (Criminisi et al. (2011))

Απόφαση του κάθε δέντρου απόφασης  $\rightarrow \{T_b\}_1^B$

Τελική απόφαση κατηγοριοποίησης του εικονοστοιχείου από το δάσος βάσει του ψήφου

πλειοψηφίας από τις αποφάσεις των επιμέρους ταξινομητών  $\rightarrow \hat{C}_{if}^B(x) = \text{majority vote } \{\hat{C}_b(x)\}_1^B$

Εκτός από τα δεδομένα εκπαίδευσης και τον αριθμό των μεταβλητών που επιλέγεται τυχαία για την εκπαίδευση του κάθε δέντρου, ακόμα ένας παράγοντας πρέπει να καθοριστεί για την κατασκευή του αλγορίθμου και αυτός είναι ο αριθμός των δέντρων απόφασης που απαρτίζουν το δάσος.

Θεωρητικά, όσο το πλήθος των δέντρων αυξάνεται, τόσο πιο ομαλά είναι τα όρια διαχωρισμού και κατά συνέπεια 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) έχουν δείξει ότι η ακρίβεια της ταξινόμησης όσο αφορά την παράμετρο του πλήθους των δέντρων, δεν είναι τόσο ευαίσθητη όσο στην παράμετρο

του πλήθους των μεταβλητών που θα καθορίσουν τα όρια διαχωρισμού σε κάθε εσωτερικό κόμβο. Στην πλειοψηφία των περιπτώσεων, η παράμετρος τίθεται στα 200, αφού τα σφάλματα παρατηρείται ότι σταθεροποιούνται λίγο πριν φτάσουμε σε αυτήν την τιμή.

### ***2.3.1.2 Ευαισθησία του ταξινομητή στα δείγματα εκπαίδευσης***

Οι δειγματοληπτικές περιοχές που χρησιμοποιούνται για την εκπαίδευση των επιβλεπόμενων ταξινομητών, πρέπει να ικανοποιούν συγκεκριμένες απαιτήσεις. Μεταξύ αυτών όπως αναφέρθηκαν στην ενότητα 3.1.1, πρέπει να είναι αντιπροσωπευτικά του φαινομένου που απεικονίζουν, να είναι φασματικά πλήρη και αξιόπιστα. Επιπλέον, πρέπει να συντελούν στη δημιουργία φασματικά ομοιογενών κλάσεων, οι οποίες μεταξύ τους να είναι ανομοιογενής.

Εντούτοις, κάτι που παραλείφθηκε, είναι η κατάσταση ισορροπίας που υπάρχει μεταξύ των κλάσεων που απαρτίζουν το σύνολο δεδομένων.

Αυστηρά μιλώντας, σχεδόν όλα τα σύνολα δεδομένων που υπάρχουν στον πραγματικό κόσμο είναι ανισόρροπα. Σε συνδυασμό με το γεγονός ότι η συλλογή τους είναι μια επίπονη, χρονοβόρα και δαπανηρή διαδικασία, οδηγούμαστε πολλές φορές στη δημιουργία ενός συνόλου δεδομένων εκπαίδευσης το οποίο αποτελείται από ομάδες που δεν αντιπροσωπεύονται εξίσου στο σύνολο των δεδομένων εκπαίδευσης που τις απαρτίζουν.

Η πρόκληση προέρχεται από το γεγονός ότι τα σπανιότερα εμφανιζόμενα δείγματα που αφορούν μια τυχαία κλάση A, συνήθως “επικαλύπτονται” από τα δείγματα της τυχαίας κλάσης B (κλάση πλειοψηφίας) και είναι κατά συνέπεια πολύ πιο δύσκολο να ταυτοποιηθούν κατά τη διαδικασία της ταξινόμησης.

Οι αλγόριθμοι ταξινόμησης, συνήθως στοχεύουν στην επίτευξη υψηλής ολικής ακρίβειας, γεγονός που δημιουργεί μια εγγενή μεροληψία υπέρ των κλάσεων πλειοψηφίας, αφού οι κλάσεις μειονότητας επηρεάζουν λιγότερο την ακρίβεια της ταξινόμησης (Yuan B. And Liu W. (2012)). Έτσι οι κλάσεις με τα μικρότερα πλήθη εικονοστοιχείων, καταλήγουν με υψηλότερα ποσοστά σφάλματος.

Επομένως η ανάλυση της ευαισθησίας του ταξινομητή ως προς την κατανομή των δειγμάτων εκπαίδευσης είναι απαραίτητη.

Στη βιβλιογραφία προτείνονται διάφορες τεχνικές για την τροποποίηση των κατανομών των κλάσεων. Μεταξύ αυτών η μείωση του πλήθους των παρατηρήσεων για τις κλάσεις που υπερέχουν αριθμητικά τις κλάσεις μειονότητας (Under-sampling method), η οποία είναι και η βέλτιστη στην περίπτωση που το προς ταξινόμηση σύνολο δεδομένων είναι τεράστιο (π.χ. δορυφορικές εικόνες) και η μείωση του πλήθους των δεδομένων εκπαίδευσης θα βοηθήσει στην υπολογιστική απόδοση του αλγόριθμου και αποφυγή προβλημάτων αποθήκευσης των δεδομένων. Παρ' όλα αυτά, η τυχαία απόρριψη πληροφορίας από τις κλάσεις, επηρεάζει την φασματική πληρότητα στην περιγραφή κάθε θεματικής κατηγορίας (Yap B.W. et al (2013)).

Ακόμα, η μέθοδος oversampling, κατά την οποία προστίθενται επιπλέον δεδομένα εκπαίδευσης για όλες τις κλάσεις μέχρις ότου να έχουν ίσο πλήθος παρατηρήσεων με την κλάση πλειοψηφίας. Το πλεονέκτημα αυτής της μεθόδου, είναι η αποφυγή απώλειας σημαντικής πληροφορίας από τις κλάσεις, ενώ στα μειονεκτήματα το γεγονός ότι η επιπλέον πληροφορία θα προστεθεί στις κλάσεις είτε μέσω δημιουργίας επαναλήψεων των ήδη υπάρχοντων δεδομένων, κάτι που δεν θα ωφελήσει ιδιαίτερα την εκπαίδευση του αλγόριθμου, είτε μέσω συλλογής περισσότερων δεδομένων που μπορεί να μην υπάρχουν (Yap B.W. et al (2013)).

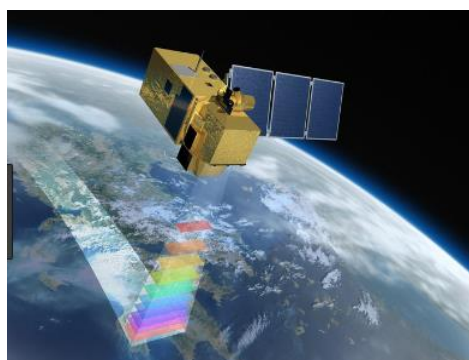
### 3 Περιγραφή των δεδομένων

Τα προϊόντα που μας παρέχουν οι αποστολές των δορυφόρων Sentinel-2 και Landsat-8, αντιπροσωπεύουν τα πιο ευρέως προσβάσιμα, πολυφασματικά, μεσαίας προς υψηλής ανάλυσης δεδομένα. Στη συγκεκριμένη μελέτη, χρησιμοποιήθηκαν δεδομένα Sentinel-2 με ημερομηνία λήψης 11/09/2017 ( ID : S2A\_MSIL1C\_T36SVD \_A011601 \_ 20170911T083628) και Landsat-8 με ημερομηνία λήψης 15/09/2017 (ID : LC08\_L1TP\_ 176036\_20170915\_20170928\_01\_T1.tar), περισσότερες λεπτομέρειες για τα οποία αναφέρονται πιο κάτω.

#### 3.1 Δεδομένα Sentinel-2

Χάρη στις υπερσύγχρονες προδιαγραφές της, η αποστολή Sentinel-2 έχει σχεδιαστεί για να συνεχίσει τις πολυφασματικές απεικονίσεις που παρείχαν οι αποστολές Spot και Landsat, συνεισφέροντας σε μια πληθώρα εφαρμογών που αφορούν την παρακολούθηση της γήινης επιφάνειας και των παράκτιων περιοχών (Pesaresi M. et al(2016) ; Immitzer M. et al (2016)).

Πέραν από την ιδιαίτερα καλή χωρική ανάλυση που φθάνει μέχρι και τα 10 μέτρα, προσφέρει πληροφορία σε 13 φασματικά κανάλια που καλύπτουν το φάσμα από το μπλε (BLUE) μήκος κύματος μέχρι το κοντινό υπέρυθρο μήκος κύματος (SWIR), συμπεριλαμβανομένων και των καναλιών Red Edge τα οποία έχουν ήδη αποδειχθεί ότι είναι χρήσιμα σε εφαρμογές χαρτογράφησης χρήσης/κάλυψης γης (Schuster C. et al., (2012)). Ακόμα, το πολυφασματικό όργανο (MSI) με το οποίο είναι εξοπλισμένος ο δορυφόρος, παρέχει λωρίδα κάλυψης πλάτους 290 χιλιομέτρων, η οποία είναι σημαντικά μεγαλύτερη από αυτή του οργάνου OLI στον δορυφόρο Landsat 8.



**Διάγραμμα 20:** Δορυφορική σάρωση της επιφάνειας της γης, από το όργανο MSI (Sentinel-2)

**Πίνακας 4:** Χαρακτηριστικά αισθητήρα MSI (Sentinel-2)

Πηγή: <https://earth.esa.int/web/sentinel/user->

| Κανάλι | Κέντρο κανάλιού (nm) | Εύρος κανάλιού (nm) | Φάσμα               | Ανάλυση (m) | Αισθητήρας |
|--------|----------------------|---------------------|---------------------|-------------|------------|
| 1      | 0.443                | 20                  | Coastal aerosol     | 60          | MSI        |
| 2      | 0.490                | 65                  | Blue                | 10          |            |
| 3      | 0.560                | 35                  | Green               | 10          |            |
| 4      | 0.665                | 30                  | Red                 | 10          |            |
| 5      | 0.705                | 15                  | Vegetation Red Edge | 20          |            |
| 6      | 0.740                | 15                  | Vegetation Red Edge | 20          |            |
| 7      | 0.783                | 20                  | Vegetation Red Edge | 20          |            |
| 8      | 0.842                | 115                 | NIR                 | 10          |            |
| 8A     | 0.865                | 20                  | Narrow NIR          | 20          |            |
| 9      | 0.945                | 20                  | Water vapour        | 60          |            |
| 10     | 1.375                | 20                  | SWIR – Cirrus       | 60          |            |
| 11     | 1.610                | 90                  | SWIR                | 20          |            |
| 12     | 2,190                | 180                 | SWIR                | 20          |            |

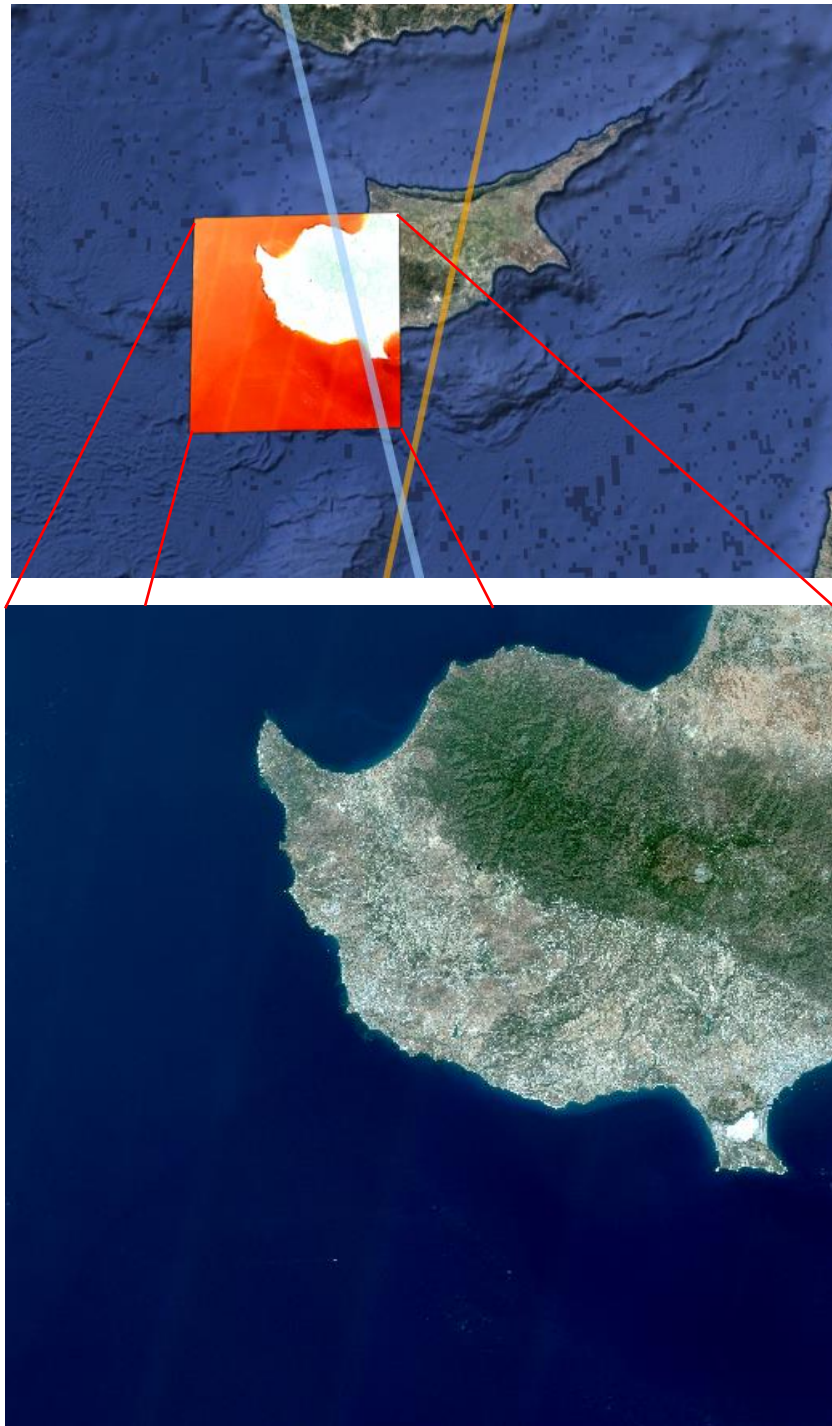
**Πίνακας 3:** Χαρακτηριστικά αποστολής Sentinel-2

Πηγή: <https://earth.esa.int/web/sentinel/user->

|                               |   |
|-------------------------------|---|
| Δορυφορικά Οχήματα σε τροχιά  | 2A & 2B (σε διαφορά φάσης 180°)   |
| Ημερομηνία Εκκίνησης          | Sentinel 2A ( 23 Ιανουάριου 2015 )<br>Sentinel 2B ( 7 Μαρτίου 2017)                     |
| Τροχιά                        | Ηλιο-σύγχρονη (διατήρηση σταθερής γωνίας ήλιου)   |
| Διάκριση Ζώνης κάθε δορυφόρου | 7,25 – 12 χρόνια  |
| Βάρος                         | 1,2 τόνοι έκαστος   |
| Τροχιακό υψόμετρο             | 786 km  |
| Κλίση Τροχιάς                 | 98.62°  |
| Λειτουργία                    | Παθητική  |
| Παγκόσμια κάλυψη              | 143 τροχιές   |
| Πλάτος λωρίδας κάλυψης        | 290 km  |
| Όργανο - Αισθητήρας           | Multispectral Instrument (MSI)  |
| Είδος σάρωσης                 | Push-broom  |
| Ραδιομετρική ανάλυση          | 12 bits (4096 διαβαθμίσεις του γκρι)  |
| Χρονική ανάλυση Αποστολής     | 5 ημέρες (10 ημέρες το κάθε δορυφορικό όχημα)   |
| Προϊόντα                      | Level-1C (Top of Atmosphere reflectance)<br>Level-2A (Bottom of Atmosphere reflectance) |
|                               | WGS 84  |



Από το σύνολο των 143 σχετικών τροχιών του δορυφόρου, οι τροχιές 28 και 21 του δορυφορικού οχήματος sentinel 2A καλύπτουν την περιοχή ενδιαφέροντος. Ωστόσο τα δεδομένα που θα επεξεργαστούμε πάρθηκαν από την τροχιά 28, η πορεία της οποίας φαίνεται στο πιο κάτω διάγραμμα μαζί με το αποτύπωμα της δορυφορικής εικόνας.

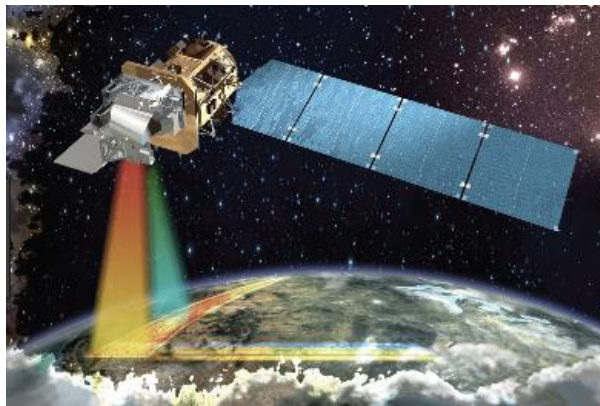


**Διάγραμμα 21:** Τροχιά 28 δορυφόρου Sentinel-2A και η δορυφορική εικόνα σε έγχρωμο σύνθετο (3-2-1)

### 3.2 Δεδομένα Landsat-8

Η οικογένεια των δορυφόρων Landsat, αποτελεί τη μακροβιότερη πηγή τηλεπισκοπικών δεδομένων εδάφους, μέτριας ανάλυσης. Τα πολύ-φασματικά χαρακτηριστικά αλλά και η συχνή, χωρίς κόστος διαθεσιμότητά τους, τον καθιστούν κατάλληλο για μια πληθώρα περιβαλλοντικών εφαρμογών σε παγκόσμια κλίμακα. Ο δορυφόρος Landsat 8, εκτοξεύτηκε τον Φεβρουάριο του 2013 και από τότε αποτελεί τον έβδομο δορυφόρο της σειράς που τίθεται σε τροχιά.

Αν εξαιρεθεί η πληροφορία που παρέχεται στο θερμικό φάσμα από τον αισθητήρα TIRS, τα φασματικά κανάλια του Landsat 8 (OLI), είναι παρόμοια με αυτά του αισθητήρα MSI από τον δορυφόρο Sentinel 2, ενώ παρόλη την ομοιότητα του και με τον δορυφόρο Landsat 7 ETM+, παρουσιάζει φασματική βελτίωση με την προσθήκη δύο επιπλέον φασματικών καναλιών στο οπτικό φάσμα (Band 1 – Deep blue και Band 9 – Infrared\_Cirrus clouds).



**Διάγραμμα 22:** Δορυφορική σάρωση της επιφάνειας της γης, από το αισθητήρα OLI (Landsat 8)

Στους πίνακες 5 και 6, παρουσιάζονται γενικά χαρακτηριστικά που αφορούν την αποστολή του δορυφόρου Landsat 8 και τα φασματικά και γεωμετρικά χαρακτηριστικά του αισθητήρα του αντίστοιχα.

**Πίνακας 5:** Χαρακτηριστικά αποστολής Landsat-8

Πηγή: <https://landsat.gsfc.nasa.gov/landsat-8/landsat-8-bands/>

|                           |   |        |
|---------------------------|---|--------|
| Ημερομηνία Εκτόξευσης     | 11 Φεβρουαρίου 2013                                       |        |
| Τροχιά                    | Ήλιο-σύγχρονη (διατήρηση σταθερής γωνίας ήλιου)           |        |
| Διάρκεια Ζωής             | 5,25 -10 χρόνια   |        |
| Βάρος                     | 1,5 τόνοι   |        |
| Τροχιακό υψόμετρο         | 705 km  |        |
| Κλίση Τροχιάς             | 98.22°  |        |
| Λειτουργία                | Παθητική  |        |
| Πλάτος λωρίδας κάλυψης    | 185 km  |        |
| Όργανα - Αισθητήρες       | Operational Land Imager (OLI)                             |        |
|                           | Thermal Infrared Sensor (TIRS)                            |        |
| Είδος σάρωσης             | Push-broom  |        |
| Ραδιομετρική ανάλυση      | 12 bits rescaled to 16 bits (55000 διαβαθμίσεις του γκρι) |        |
| Χρονική ανάλυση Αποστολής | 16 ημέρες   |        |
| Προϊόντα                  | Level-1T  | WGS 84 |
|                           | Level-1GT   |        |
|                           | Level-1G  |        |

**Πίνακας 6:** Χαρακτηριστικά του αισθητήρα OLI (Landsat-8)

Πηγή: <https://landsat.gsfc.nasa.gov/landsat-8/landsat-8-bands/>

| Κανάλι | Εύρος Μήκους<br>Κόματος (nm) | Φάσμα           | Αισθητήρας | Ανάλυση<br>(m) |
|--------|------------------------------|-----------------|------------|----------------|
| 1      | 433-453                      | Coastal-Aerosol | OLI        | 30             |
| 2      | 450-515                      | Blue            | OLI        | 30             |
| 3      | 525-600                      | Green           | OLI        | 30             |
| 4      | 630-680                      | Red             | OLI        | 30             |
| 5      | 845-885                      | NIR             | OLI        | 30             |
| 6      | 1560-1660                    | SWIR 1          | OLI        | 30             |
| 7      | 2100-2300                    | SWIR 2          | OLI        | 30             |
| 8      | 500-680                      | Panchromatic    | OLI        | 15             |
| 9      | 1360-1390                    | Cirrus          | OLI        | 30             |
| 10     | 10600-11200                  | TIR 1           | TIRS       | 100            |
| 11     | 11500-12500                  | TIR 2           | TIRS       | 100            |

## **4 Μεθοδολογία Έρευνας**

Σε αυτό το κεφάλαιο, παρατίθεται η μεθοδολογία που αναπτύχθηκε κατά τη διάρκεια της περιόδου έρευνας και περιγράφεται το σύνολο των τεχνικών και μεθόδων που εφαρμόστηκαν.

### **4.1 Προ-επεξεργασία Δορυφορικών Εικόνων**

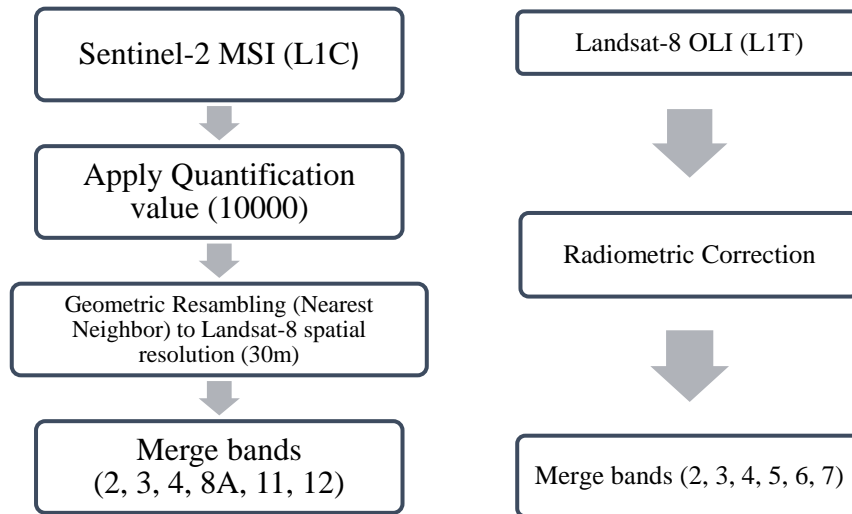
Το στάδιο της προ-επεξεργασίας, περιλαμβάνει τις λειτουργίες που απαιτούνται πριν από την κύρια ανάλυση των δεδομένων και εξαγωγή των πληροφοριών που χρειάζονται σε κάθε έρευνα και διακρίνονται στις ραδιομετρικές, γεωμετρικές και ατμοσφαιρικές διορθώσεις των εικόνων.

Όπως αναφέρθηκε πιο πάνω, στην παρούσα μελέτη χρησιμοποιήθηκε μια εικόνα Sentinel 2A \_ Level 1C, μεγέθους 100 km x 100 km, η οποία είχε ως σύστημα αναφοράς το WGS 84. Οι τιμές στο κάθε εικονοστοιχείο της εικόνας, αντιπροσωπεύουν τιμές ανακλαστικότητας στο πάνω τμήμα της ατμόσφαιρας (Top of Atmosphere), οι οποίες εμφανίζονται για σκοπούς εξοικονόμησης χωρητικότητας ως υπό κλίμακα ακέραιες τιμές.

Επομένως, αφού οι τιμές δεν επηρεάστηκαν από τα διάφορα συστατικά της ατμόσφαιρας όπως τα αερολύματα, τους υδρατμούς και άλλα σωματίδια (που μπορούν να προκαλέσουν φαινόμενα όπως η απορρόφηση και η σκέδαση) και δεδομένου ότι και τα δεδομένα που προέρχονται από το δορυφόρο Landsat 8 θα μεταφερθούν στο ίδιο επίπεδο (TOA reflectance), δεν χρειάζεται να διορθώσουμε ατμοσφαιρικά τα δεδομένα, παρά μόνο να εφαρμόσουμε τον παράγοντα κλίμακας για να επαναφέρουμε τις τιμές στο εύρος της ποσοστιαίας ανακλαστικότητας.

Τα δεδομένα που προήλθαν από το δορυφόρο Landsat 8 L1T (Level 1), είναι επίσης γεωμετρικά διορθωμένα (χρησιμοποιώντας GCPs και DEMs), χωρίς παραμορφώσεις που σχετίζονται με τον αισθητήρα (π.χ. γωνίες λήψης), με τον δορυφόρο (αποκλίσεις της συμπεριφοράς του που σχετίζεται είτε με την τροχιά είτε με το υψόμετρο) και τη γη (περιστροφή, καμπυλότητα και άλλους γεωφυσικούς παράγοντες).

Στο διάγραμμα 23, παρουσιάζεται η μεθοδολογία προ-επεξεργασίας και προετοιμασίας των προς ανάλυση δεδομένων, ενώ στο διάγραμμα 24 παρουσιάζεται το μοντέλο που συντέθηκε στο λογισμικό Snar, για την επανασύσταση και συγχώνευση των δεδομένων Sentinel-2, ούτως ώστε να μπορούν συγκριθούν με τα δεδομένα Landsat-8.



**Διάγραμμα 23:** Διαδικασία προ-επεξεργασίας και ετοιμασίας των προς ανάλυση εικόνων

Το προϊόν (Level 1T Landsat-8), παρουσιάζεται σε βαθμονομημένη κλίμακα ψηφιακών τιμών (DN), οι οποίες μετατρέπονται σε τιμές ανακλαστικότητας στο πάνω μέρος της ατμόσφαιρας (TOA reflectance) μέσω των πιο κάτω εξισώσεων :

$$\rho_{\lambda} = \frac{\rho_{\lambda'}}{\cos(\theta_{SZ})} = \frac{\rho_{\lambda'}}{\sin(\theta_{SE})} \quad , \text{όπου} \quad \rho_{\lambda'} = M_p Q_{cal} + A_p$$

**Εξίσωση 9.** Μετατροπή Ψηφιακών τιμών προϊόντος Landsat 8 (Level 1T), σε Top of Atmosphere Reflectance

Όπου:

$\rho_{\lambda}$  → Top Of Atmosphere Reflectance

$\rho_{\lambda'}$  → Top Of Atmosphere Reflectance (χωρίς τη διόρθωση για τη γωνία ύψους του ήλιου)

$\theta_{SZ}$  → Γωνία ύψους του ήλιου

$\theta_{SE}$  → Ζενίθια Γωνία

$M_p \rightarrow$  Rescaling factor

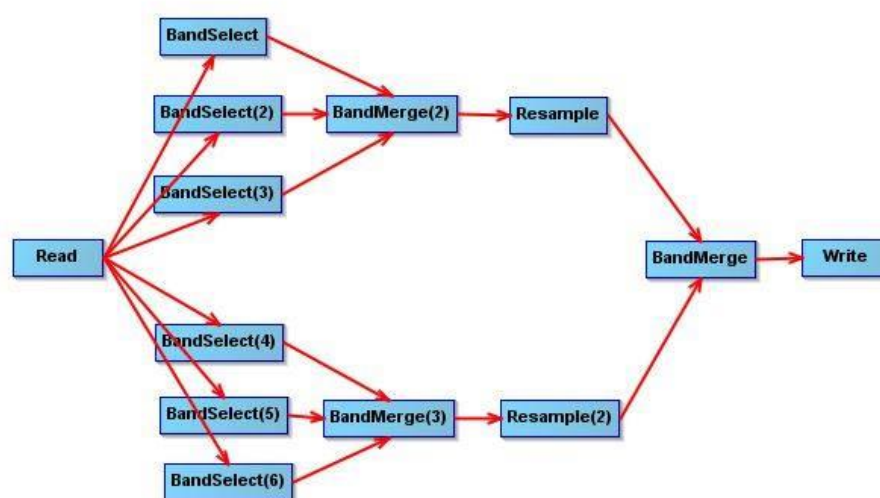
$Q_{cal} \rightarrow$  DNs

$A_p \rightarrow$  Reflectance Rescaling factor

Ακολούθως, δεδομένου ότι τόσο ο Sentinel 2 όσο και ο Landsat 8 καλύπτουν φασματικό εύρος μεταξύ 440 nm και 2300 nm, θα συγχωνεύσουμε διαύλους που παρουσιάζουν παρόμοια φασματικά χαρακτηριστικά, προκειμένου να προκύψει μια σύγκριση της απόδοσης τους κατά την μετέπειτα ανάλυση. Η συγχώνευση αυτών των καναλιών για τον δορυφόρο Landsat είναι πολύ απλή, αφού οι δίαυλοι ενδιαφέροντος βρίσκονται στην ίδια ανάλυση (30 μ.), κάτι που δεν ισχύει για τους δίαυλους του δορυφόρου Sentinel 2.

Αυτό το πρόβλημα, θα αντιμετωπιστεί μέσω της διαδικασίας της επανασύστασης των εικονοστοιχείων που βρίσκονται σε αναλύσεις 10 μ. και 20 μ., εφαρμόζοντας τη μέθοδο του Εγγύτερου Γείτονα (Nearest Neighbor). Η συγκεκριμένη μέθοδος συστήνεται σε εφαρμογές ταξινόμησης, λόγω του ότι δεν μεταβάλλει τις τιμές των εικονοστοιχείων, κάτι που δεν συμβαίνει στις μεθόδους της Διγραμμικής (Bilinear) και Κυβικής (Cubic) παρεμβολής, αφού η νέες τιμές των εικονοστοιχείων προκύπτουν από τον μέσο όρο των τιμών τεσσάρων και δεκαέξι εικονοστοιχείων αντίστοιχα.

Οι εικόνες που προκύπτουν από τις πιο πάνω διαδικασίες φαίνονται στα διαγράμματα 25 και 26



**Διάγραμμα 24:** Μοντέλο δημιουργίας πολυφασματικού προϊόντος Sentinel-2, 6 καναλιών και χωρικής ανάλυσης 30 μ.

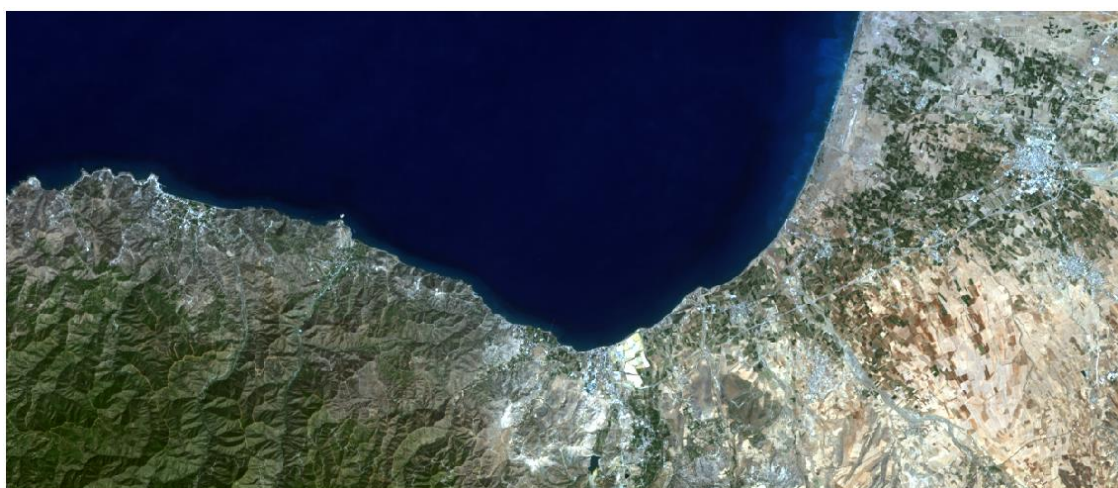


**Πίνακας 7:** Φασματική σύγκριση δεδομένων Sentinel- 2 και Landsat-8

| Sentinel 2A |                     |        | Landsat 8 |                     |            | Ανάλυση<br>(m) |
|-------------|---------------------|--------|-----------|---------------------|------------|----------------|
| Κανάλι      | Εύρος φάσματος (nm) | Φάσμα  | Κανάλι    | Εύρος φάσματος (nm) | Φάσμα      |                |
| 2           | 450-515             | Blue   | 2         | 0.460-0.520         | Blue       | 30             |
| 3           | 525-600             | Green  | 3         | 0.542-0.578         | Green      | 30             |
| 4           | 630-680             | Red    | 4         | 0.650-0.680         | Red        | 30             |
| 5           | 845-885             | NIR    | 8A        | 0.855-0.875         | Narrow NIR | 30             |
| 6           | 1560-1660           | SWIR 1 | 11        | 1.565-1.655         | SWIR       | 30             |
| 7           | 2100-2300           | SWIR 2 | 12        | 2,100-2.280         | SWIR       | 30             |



**Διάγραμμα 25:** Τελικό πολυφασματικό προϊόν 6 καναλιών Sentinel-2, Ανάλυση 30 m, Έγχρωμο σύνθετο: 3-2-1



**Διάγραμμα 26:** Τελικό πολυφασματικό προϊόν 6 καναλιών Landsat-8, Ανάλυση 30 m, Έγχρωμο σύνθετο: 3-2-1

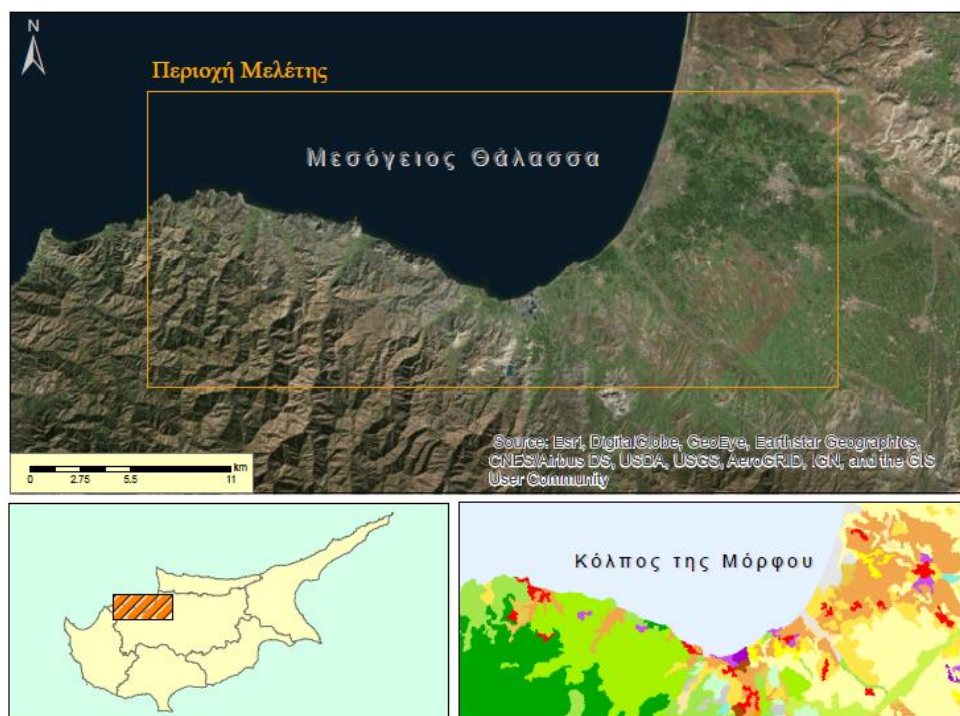
## 4.2 Περιοχή μελέτης

Η προς ταξινόμηση περιοχή, βρίσκεται στο βόρειο τμήμα του νησιού της Κύπρου σε απόσταση 40-60 χλμ. βόρεια της Λεμεσού και της Πάφου και 50 χλμ. ανατολικά της Λευκωσίας, καλύπτοντας συνολική έκταση περίπου 607.025 km<sup>2</sup> (674472 pixels).

Εντός των ορίων της περιοχής μελέτης (37.68 km x 16.11 km), παρουσιάζεται πληθώρα τύπων κάλυψης γης. Μεταξύ αυτών τα διάφορα επίπεδα έντασης της βλάστησης, συμπεριλαμβανομένων του δάσους της Πάφου και της οροσειράς του Τροόδου στα δυτικά και φυσικών περιοχών που εμπίπτουν στο ευρωπαϊκό πρόγραμμα προστασίας NATURA 2000, νότια της περιοχής μας.

Ακόμα, εντοπίζονται πολλές τεχνητές επιφάνειες (πόλεις και χωριά) και πολλά επίπεδα καλλιεργήσιμων εκτάσεων στα ανατολικά της περιοχής μελέτης, ενώ το βόρειο τμήμα καλύπτεται πλήρως από υδάτινες επιφάνειες.

Η ποικιλία που υπάρχει στις εδαφοκαλύψεις, και κατά συνέπεια, η μεγάλη διακύμανση στις τιμές ανακλαστικότητας που θα κληθούν να αναγνωρίσουν και εντέλει να διαχωρίσουν / ομαδοποιήσουν οι προς εξέταση αλγόριθμοι, θα βοηθήσει στην αυστηρότερη σύγκριση της απόδοσης τους.



Διάγραμμα 27: Περιοχή μελέτης



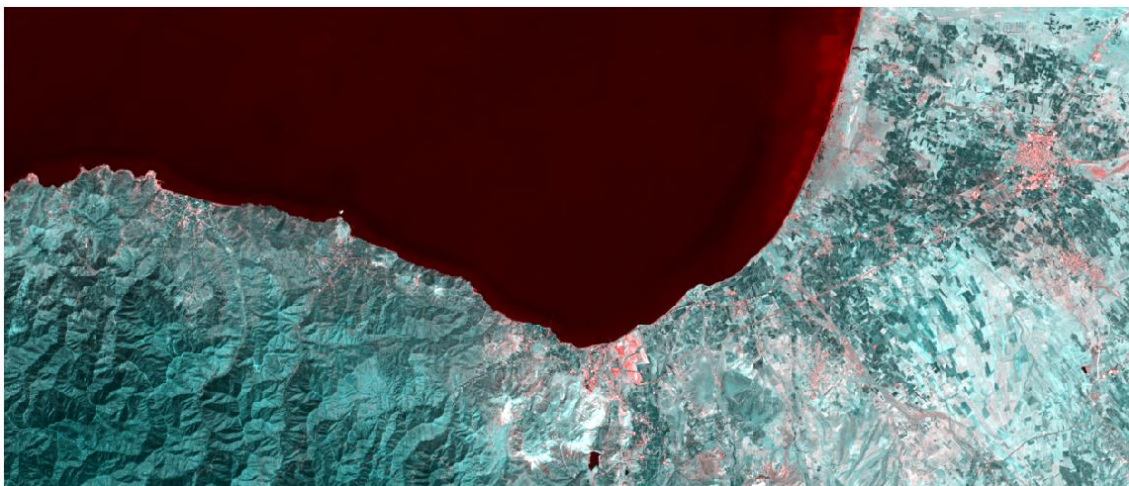
## 4.3 Στάδια ταξινόμησης δορυφορικών εικόνων

### 4.3.1 Ίδρυση Συστήματος ταξινόμησης

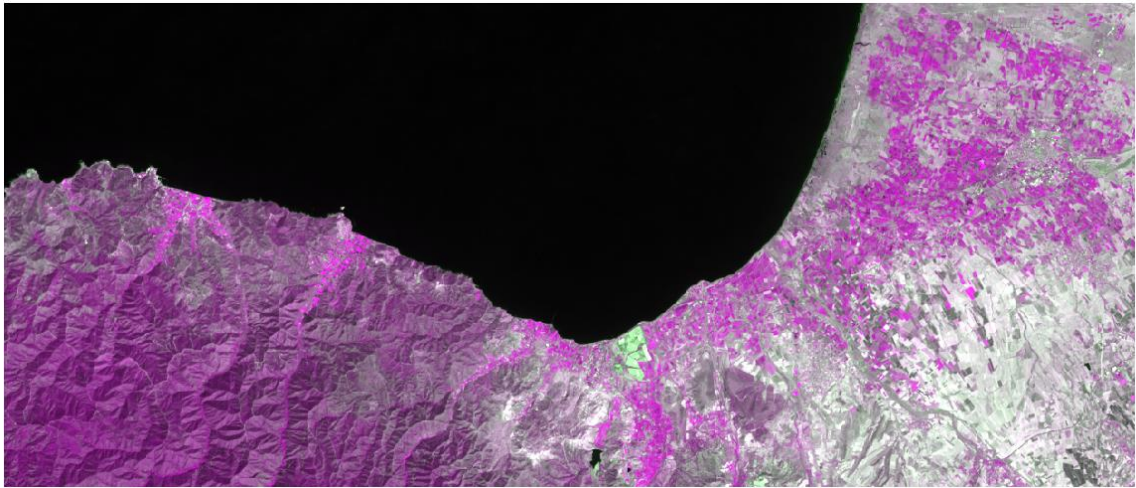
Εφαρμογές με υψηλό επίπεδο σημαντικότητας όπως η χαρτογράφηση χρήσης/κάλυψης γης, απαιτούν συστήματα ταξινόμησης που να είναι όσο το δυνατό πληρέστερα. Για αυτό το λόγο, είναι απαραίτητο να περιλαμβάνουν όλες τις χρήσεις και καλύψεις γης που υπάρχουν στην περιοχή μελέτης, ορίζοντας τις με σαφήνεια και ομαδοποιώντας τις όπου θεωρηθεί αναγκαίο.

Για τη δημιουργία του δικού μας συστήματος ταξινόμησης, βασιστήκαμε στο σύστημα ταξινόμησης που αναπτύχθηκε από τον Gregorio Di A. το 2005, ( FAO LCCS (Land Classification System)), αναπροσαρμόζοντας το όπου κριθεί απαραίτητο, προκειμένου να ανταποκρίνεται στις απαιτήσεις της περιοχής μελέτης.

Παρόλο που η γνώση για την επί τόπου κατάσταση της περιοχής μελέτης (γεωλογία, είδη βλάστησης και κλιματικές συνθήκες) δεν είναι πολύ καλή, η φωτοερμηνεία των εικόνων βασίστηκε σε ψευδοχρωματικά σύνθετα (διαγράμματα 27 -30), σε θεματικούς χάρτες κάλυψης γης που προέρχονται από τις υπηρεσίες διαχείρισης γης του ευρωπαϊκού προγράμματος Corine (διάγραμμα 31) και τέλος σε θεματικούς χάρτες που προέκυψαν από την εφαρμογή δεικτών εδάφους στα προς ταξινόμηση δεδομένα.



**Διάγραμμα 28:** Δεδομένα Sentinel-2, Ψευδοχρωματικό Σύνθετο 1-5-5 → εντοπισμός αστικών περιοχών



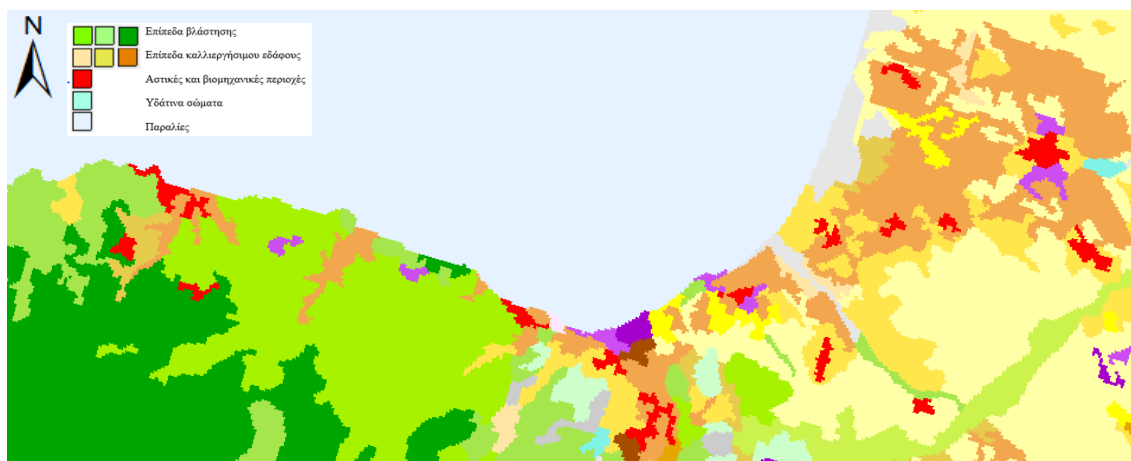
**Διάγραμμα 29:** Δεδομένα Sentinel-2, Ψευδοχρωματικό Σύνθετο 4-3-4 → εντοπισμός περιοχών βλάστησης



**Διάγραμμα 30:** Δεδομένα Sentinel-2, Ψευδοχρωματικό Σύνθετο 6-1-1 → Οπτικός εντοπισμός κατηγοριών εδάφους και υδάτινων επιφανειών



**Διάγραμμα 31:** Δεδομένα Sentinel-2, Ψευδοχρωματικό Σύνθετο 1-3-1 → εντοπισμός περιοχών εδάφους



**Διάγραμμα 32:** Χάρτης χρήσης/κάλυψης γης CORINE (ανάλυση 100 m)

Ακόμα, εφαρμόστηκε ο κανονικοποιημένος δείκτης βλάστησης (NDVI), ο οποίος ορίζεται ως ο λόγος της διαφοράς και της πρόσθετης των τιμών ανακλαστικότητας στο εγγύς υπέρυθρο και οπτικό φάσμα και χρησιμοποιείται αφενός για τον εντοπισμό των περιοχών που καλύπτονται με βλάστηση και αφετέρου για τον υπολογισμό της ζωτικότητας της βλάστησης σε αυτές τις περιοχές (pixels).

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

**Εξίσωση 10.** Κανονικοποιημένος δείκτης βλάστησης (NDVI)

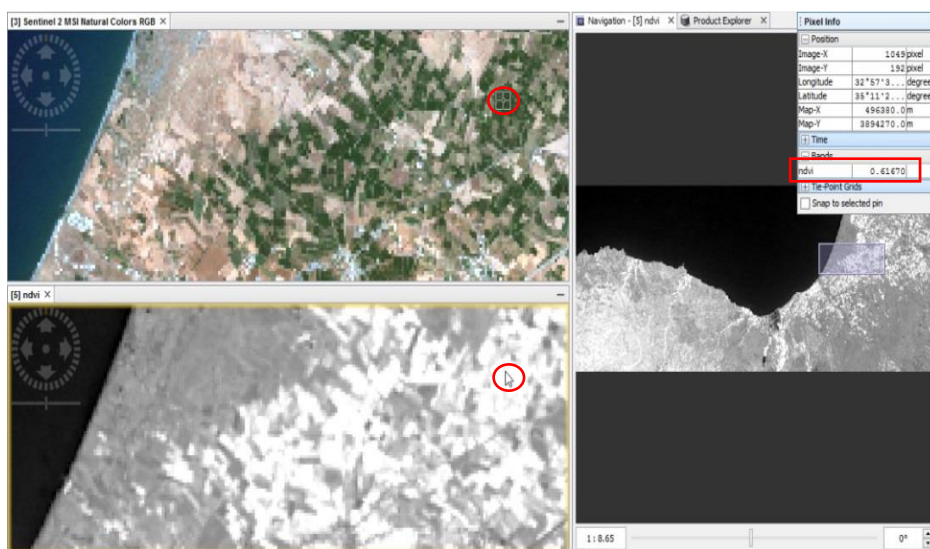
Όπου :

NIR : υπέρυθρη ακτινοβολία στο εγγύς υπέρυθρο φάσμα

RED: ορατή ακτινοβολία στο κόκκινο φάσμα

Η δημιουργία των χαρτών πυκνότητας της βλάστησης (διαγράμματα 34) έγινε στο λογισμικό SNAP, όπου αρχικά έγινε αναγνώριση της περιοχής (διάγραμμα 33) προκειμένου να εντοπισθούν οι κατηγορίες πυκνότητας της βλάστησης που θα δημιουργηθούν (πίνακας 5).

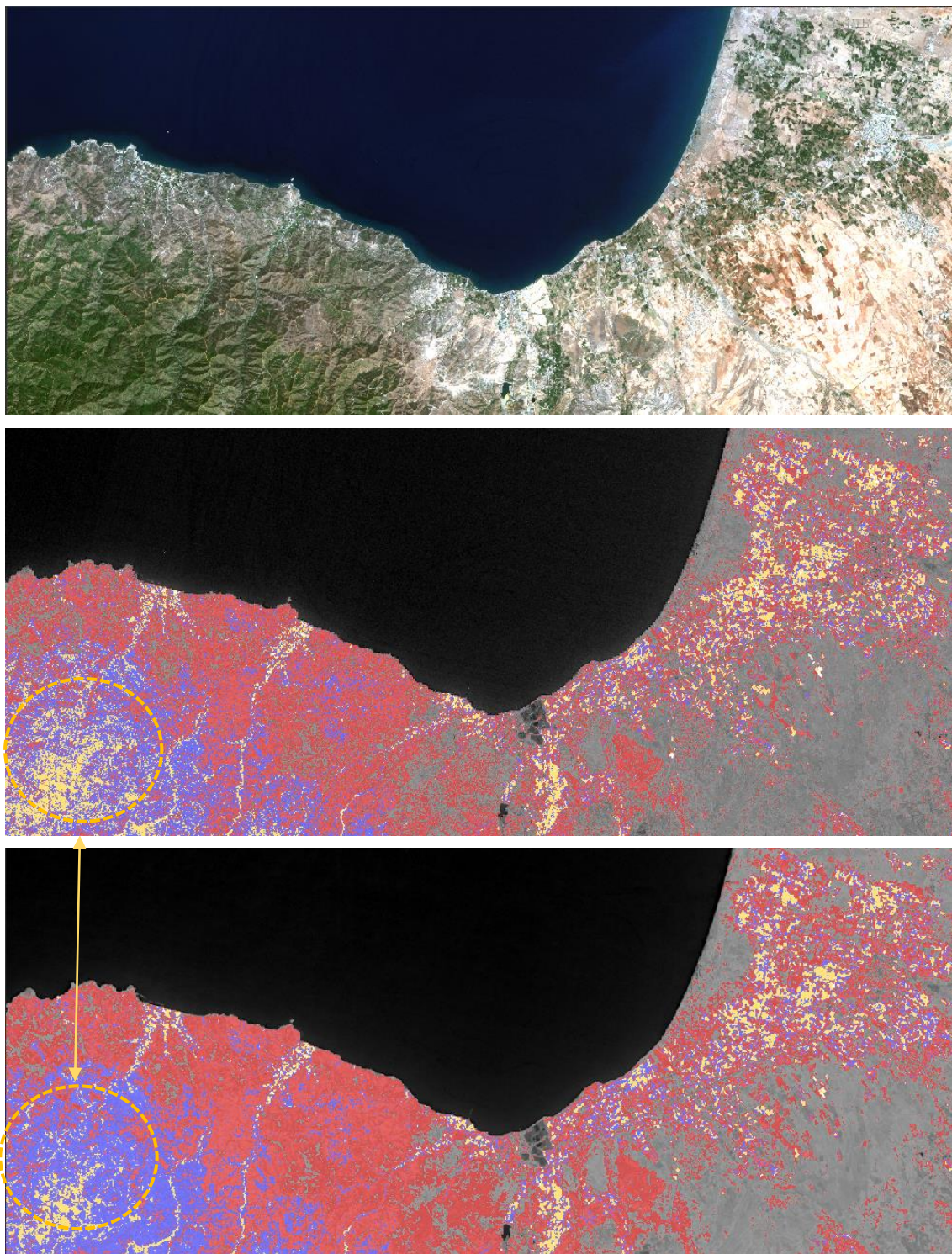




**Διάγραμμα 33:** Αναγνώριση της πυκνότητας βλάστησης της περιοχής

**Πίνακας 8:** Κατηγορίες Βλάστησης βάση του δείκτη NDVI

| <b>Πυκνότητα Βλάστησης</b> | <b>NDVI</b>                     |
|----------------------------|---------------------------------|
| Υψηλή                      | $ndvi > 0.45$ and $ndvi < 0.65$ |
| Μέτρια                     | $ndvi > 0.35$ and $ndvi < 0.45$ |
| Χαμηλή                     | $ndvi > 0.2$ and $ndvi < 0.35$  |



**Διάγραμμα 34:** Χάρτες πυκνότητας βλάστησης των δεδομένων Sentinel 2A (1η εικόνα) και Landsat 8 (2η εικόνα) και οι διαφορές που εντοπίζονται μεταξύ τους

Υψηλή πυκνότητα-Πορτοκαλί, Μεσαία πυκνότητα-Μπλε, Χαμηλή πυκνότητα-κόκκινο

Ακολουθώς, στα δύο σύνολα δεδομένων εφαρμόστηκε ο δείκτης Bare Soil, ο οποίος βασίζεται στις τιμές ανακλαστικότητας του μπλε, κόκκινου, εγγύς και κοντινού υπέρυθρου φάσματος των δεδομένων. Πιο κάτω φαίνονται οι εξισώσεις του δείκτη για τα δεδομένα Sentinel-2 και Landsat-8 αντίστοιχα.

$$BSI_{Sentinel2} = \frac{(Band11 + Band4) - (Band8A + Band2)}{(Band11 + Band4) + (Band8A + Band2)}$$

$$BSI_{Landsat8} = \frac{(Band6 + Band4) - (Band5 + Band2)}{(Band6 + Band4) + (Band5 + Band2)}$$

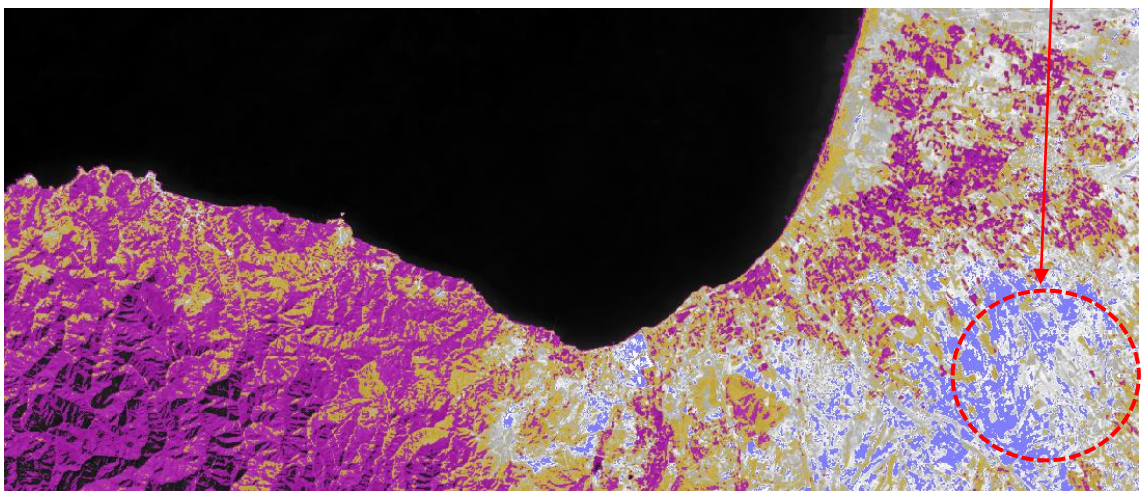
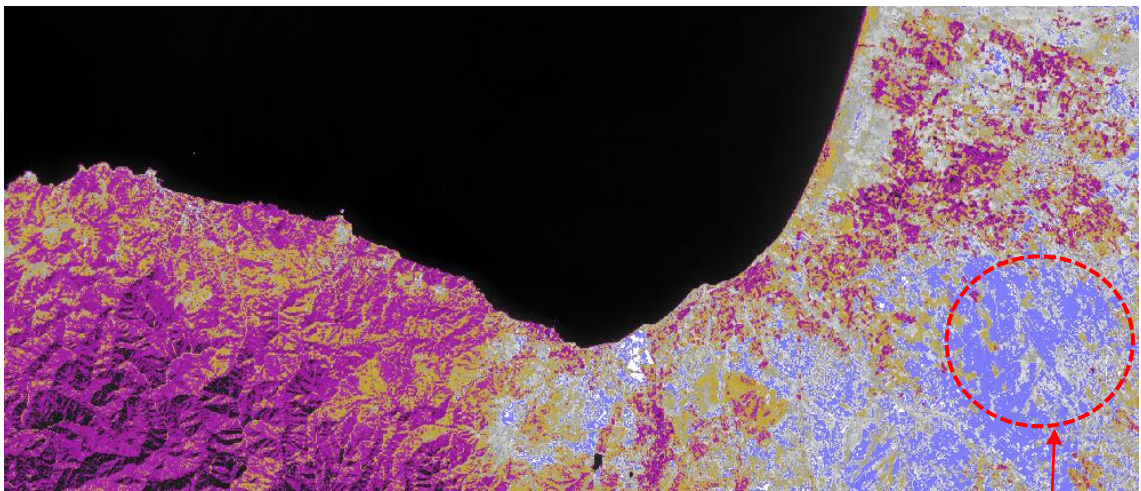
#### Εξίσωση 11. Δείκτης εδαφών για τα δεδομένα Sentinel-2 και Landsat-8

Η κατηγοριοποίηση των τύπων εδάφους που υπάρχουν στην περιοχή μελέτης, έγινε με τρόπο παρόμοιο με αυτό της βλάστησης. Συγκεκριμένα, εντοπίστηκαν οι εξής κατηγορίες εδαφών (με αυξανόμενη την τιμή του δείκτη BSI), το Γυμνό έδαφος (bare), το γεωργικό έδαφος (agricultural) και το πετρώδες έδαφος (stony). Ακόμα, υπάρχει και η κατηγορία του γκρι εδάφους, το οποίο βρέθηκε οπτικά, χωρίς να απαιτείται η σύγκριση του με τους υπόλοιπους τύπους εδαφών.

#### Πίνακας 9: Κατηγορίες εδάφους

| Τύποι εδάφους     | BSI                 |
|-------------------|---------------------|
| Bare soil         | bi>0.2 and bi<0.25  |
| Agricultural soil | bi>0.12 and bi<0.15 |
| Stony soil        | bi>0.08 and bi<0.12 |










**Διάγραμμα 35:** Χάρτες εδαφών των δεδομένων Sentinel-2 και Landsat-8

Γεωργικό έδαφος-Πορτοκαλί, Γυμνό έδαφος-Μπλε, Πετρώδες έδαφος-ροζ


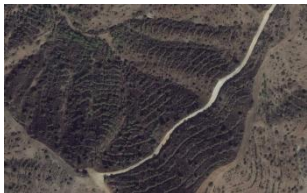
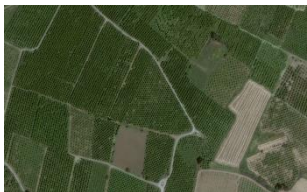
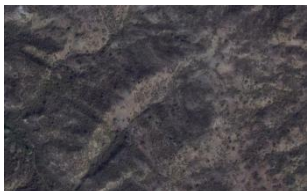


Μεταξύ των δύο εικόνων στις δύο περιπτώσεις, εντοπίστηκαν κάποιες διαφορές οι οποίες οφείλονται σε μικρές μεταβολές των ραδιομετρικών τιμών που υπέστηκαν τα δεδομένα Sentinel-2 κατά τη διαδικασία της επανασύστασης τους, προκειμένου να έρθουν στη χωρική ανάλυση που βρίσκονται τα δεδομένα Landsat-8.

Αφού λοιπόν έγιναν οι απαραίτητες μελέτες για την ίδρυση ενός πλήρους συστήματος ταξινόμησης, προέκυψαν οι εξής 11 θεματικές κατηγορίες.

**πίνακας 10.** Κατηγορίες κάλυψης γης

| A/A | Αρχικές κλάσεις                     | Παράδειγμα  | Περιγραφή   |
|-----|-------------------------------------|---|---|
| 1   | Natural Waterbodies (deep)          |   | Φυσικές υδάτινες επιφάνειες (Βαθιά νερά)  |
| 2   | Natural Waterbodies (shallow)       |  | Φυσικές υδάτινες επιφάνειες (Ρηχά νερά κοντά στην ακτή)                                   |
| 3   | Artificial Waterbodies              |  | Τεχνητές υδάτινες επιφάνειες (Φράγματα και δεξαμενές)                                     |
| 4   | Urban areas and artificial surfaces |  | Εκτάσεις καλυμμένες από κτίρια (αστικός ιστός, χωριά) ή άλλες τεχνητές επιφάνειες, δρόμοι |
| 5   | Grey soil                           |  | Γκρι έδαφος<br>Μεταλλεία, Ορυχεία, αποθέσεις ποταμών                                      |



|    |                             |   |   |
|----|-----------------------------|---|---|
| 6  | High Intensity vegetation   |    | Βλάστηση υψηλής πυκνότητας<br>(Πυκνά δάση από κωνοφόρα και πλατύφυλλα δέντρα)   |
| 7  | Medium Intensity vegetation |    | Βλάστηση μεσαίας πυκνότητας<br>(Μεσαίας πυκνότητας δάση και “νεαρά” δάση )  |
| 8  | Low Intensity vegetation    |    | Βλάστηση χαμηλής πυκνότητας<br>(Λιβάδια, Εποχιακά καλλιεργούμενες εκτάσεις, καλλιέργειες σε αρχικό στάδιο, υποβρύχιοι και χαμηλοί θάμνοι) |
| 9  | Stony soil                  |   | Έδαφος που καλύπτεται από πολλές πέτρες   |
| 10 | Agricultural soil           |  | Γεωργικό έδαφος, Καλλιεργημένες εκτάσεις γης  |
| 11 | Bare soil                   |  | Άγονο έδαφος, Έδαφος που δεν καλύπτεται από κανένα είδος βλάστησης  |

#### 4.3.2 Συλλογή εκπαιδευτικών δειγμάτων

Αν και η ταξινόμηση πολυφασματικών δεδομένων είναι μια ιδιαίτερα αυτοματοποιημένη διαδικασία, όπως αναφέραμε στο κεφάλαιο 2.1.1 στην πραγματικότητα η συλλογή των δεδομένων εκπαίδευσης που απαιτούνται από τον αλγόριθμο, είναι οτιδήποτε άλλο εκτός από αυτόματη.

Απαιτεί στενή αλληλεπίδραση μεταξύ του αναλυτή και της πραγματικότητας, αφού η ποιότητα των δεδομένων εκπαίδευσης (επίγεια αληθή δεδομένα) και γενικότερα της εκπαιδευτικής διαδικασίας, καθορίζει το επίπεδο ακρίβειας και αξιοπιστίας της πληροφορίας που προκύπτει από την όλη προσπάθεια ταξινόμησης.

Η συλλογή των δεδομένων εκπαίδευσης για τις φασματικές κατηγορίες, πραγματοποιήθηκε στο λογισμικό arcmap 10.2.2 (διάγραμμα 36), χρησιμοποιώντας κυρίως σημειακές αλλά και πολυγωνικές οντότητες.



**Διάγραμμα 36:** Συλλογή εκπαιδευτικών δειγμάτων

Στον πίνακα 8, παρουσιάζεται το πλήθος των εικονοστοιχείων που συλλέχθηκε για κάθε κλάση, ενώ όπως αναφέραμε στην ενότητα, μόνο τα δύο τρίτα των δειγματοληπτικών περιοχών χρησιμοποιήθηκαν για εκπαιδευτικό σκοπό και αναφέρονται ως In Bag Samples (IBS). Το εναπομένον μερίδιο του ενός τρίτου (Out of Bag Samples - OOB), χρησιμοποιήθηκε για την εκτίμηση της απόδοσης του αλγόριθμου και το σφάλμα που προκύπτει ονομάζεται Out of Bag error (OOB error).

Το ίδιο σύνολο δεδομένων θα χρησιμοποιηθεί για να εκπαιδεύσει όλους τους αλγόριθμους επιβλεπόμενης ταξινόμησης.

**Πίνακας 11:** Πλήθος εικονοστοιχείων εκπαίδευσης και ελέγχου για κάθε κατηγορία κάλυψης γης

| A/A | Αρχικές κλάσεις                     | Πλήθος<br>εικονοστοιχείων | In Bag Data<br>(δεδομένα<br>εκπαίδευσης) | Out of Bag Data<br>(δεδομένα<br>ελέγχου) |
|-----|-------------------------------------|---------------------------|--|--|
| 1   | Natural Waterbodies (deep)          | 90                        | 69                                       | 21                                       |
| 2   | Natural Waterbodies (shallow)       | 89                        | 61                                       | 28                                       |
| 3   | Artificial Waterbodies              | 110                       | 77                                       | 33                                       |
| 4   | Urban areas and artificial surfaces | 175                       | 118                                      | 57                                       |
| 6   | Grey soil                           | 84                        | 57                                       | 27                                       |
| 7   | High Intensity vegetation           | 161                       | 118                                      | 43                                       |
| 8   | Medium Intensity vegetation         | 215                       | 147                                      | 68                                       |
| 9   | Low Intensity vegetation            | 137                       | 92                                       | 45                                       |
| 10  | Stony soil                          | 143                       | 96                                       | 47                                       |
| 11  | Agricultural soil                   | 202                       | 151                                      | 51                                       |
| 12  | Bare soil                           | 291                       | 204                                      | 87                                       |

Στη συνέχεια, ακολουθεί η επαναληπτική διαδικασία αξιολόγησης των δειγμάτων εκπαίδευσης που συλλέχθηκαν, βάση της φασματικής διαχωριστικότητας που υπάρχει μεταξύ κάθε πιθανού συνδυασμού κλάσεων.

Η εν λόγω αξιολόγηση, επιτυγχάνεται μέσω της δημιουργίας του τριγωνικού πίνακα διαχωριστικότητας (separability array), στον οποίο η διαχωριστική ικανότητα υπολογίζεται μέσω του αλγόριθμου transformed divergence (εξ.9) και εκφράζεται σε εύρος τιμών 0 – 2000 (0-1000 → χαμηλή διαχωριστικότητα, 1000-1900 → μέτρια διαχωριστικότητα και 1900-2000→ Πολύ καλή διαχωριστικότητα).

$$TD_{ij} = 2000 \left( 1 - \exp \left( \frac{-D_{ij}}{8} \right) \right)$$

$$D_{ij} = \frac{1}{2} \text{tr}((C_i - C_j)(C_i^{-1} - C_j^{-1})) + \frac{1}{2} \text{tr}((C_i^{-1} - C_j^{-1})(\mu_i - \mu_j)(\mu_i - \mu_j)^T)$$

**Εξίσωση 12.** Αλγόριθμος Transformed Divergence (Erdas Imagine 2014)

Στους πιο πάνω πίνακες διαχωριστικότητας, παρατηρούμε ότι τα δείγματα που εμπίπτουν στις περισσότερες κλάσεις, διαχωρίζονται σύμφωνα με τον αλγόριθμο Transformed Divergence αρκετά καλά. Οι μόνες ανησυχητικές περιπτώσεις, παρουσιάζονται μεταξύ των κλάσεων Low Intensity vegetation - Medium Intensity vegetation και Bare Soil – Urban area and Artificial surfaces, οι οποίες σύμφωνα με τις τιμές του αλγόριθμου, εμπίπτουν στην κατηγορία μέτριας διαχωριστικής ικανότητας.

Οι χαμηλές τιμές διαχωρισμού τους, μας οδήγησαν σε επαναληπτικές διαδικασίες αναθεώρησης των δειγμάτων που περιέχονται σε κάθε μια από τις προβληματικές κλάσεις, χωρίς ωστόσο να επιλυθεί το πρόβλημα. Έτσι, γνωρίζοντας ότι σύμφωνα με την χωρική ανάλυση των δεδομένων (30 μ.), η μεταβλητότητα των (μέσων) τιμών ανακλαστικότητας σε κάθε εικονοστοιχείο επηρεάζεται και αυξάνεται επικαλύπτοντας συχνά άλλες κατηγορίες, αποδεχτήκαμε τις κλάσεις ως έχουν και προχωρήσαμε στο στάδιο της ταξινόμησης.

**Πίνακας 12:** Separability array (Sentinel 2)

Distance Measure: Transformed Divergence  
Using Layers: 1 2 3 4 5 6  
Taken 6 at a time  
Best Average Separability: 1986.05  
Combination: 1 2 3 4 5 6

| Signature Name                     | 1  | 2    | 3    | 4    | 5       | 6       | 7       | 8       | 9       | 10      | 11      |         |
|------------------------------------|----|------|------|------|---------|---------|---------|---------|---------|---------|---------|---------|
| Natural Waterbodies (deep)         | 1  | 0    | 2000 | 2000 | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    |
| Natural Waterbodies (shallow)      | 2  | 2000 | 0    | 2000 | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    |
| Artificial Waterbodies             | 3  | 2000 | 2000 | 0    | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    |
| Urban Area and artificial surfaces | 4  | 2000 | 2000 | 2000 | 0       | 1999.81 | 2000    | 2000    | 1937.67 | 2000    | 2000    | 1757.79 |
| Grey soil                          | 5  | 2000 | 2000 | 2000 | 1999.81 | 0       | 2000    | 2000    | 1999.97 | 1999.94 | 1999.94 | 1925.75 |
| High Intensity vegetation          | 6  | 2000 | 2000 | 2000 | 2000    | 2000    | 0       | 1995.38 | 2000    | 1993.03 | 2000    | 2000    |
| Medium Intensity vegetation        | 7  | 2000 | 2000 | 2000 | 2000    | 2000    | 1995.38 | 0       | 1657.28 | 1999.51 | 1999.87 | 1999.9  |
| Low Intensity vegetation           | 8  | 2000 | 2000 | 2000 | 1937.67 | 2000    | 2000    | 1657.28 | 0       | 1999.93 | 1988.34 | 1980.2  |
| Stony soil                         | 9  | 2000 | 2000 | 2000 | 2000    | 1999.97 | 1993.03 | 1999.51 | 1999.93 | 0       | 1999.33 | 1999.97 |
| Agricultural soil                  | 10 | 2000 | 2000 | 2000 | 2000    | 1999.94 | 2000    | 1999.87 | 1988.34 | 1999.33 | 0       | 1998.89 |
| Bare soil                          | 11 | 2000 | 2000 | 2000 | 1757.79 | 1925.75 | 2000    | 1999.9  | 1980.2  | 1999.97 | 1999.94 | 0       |

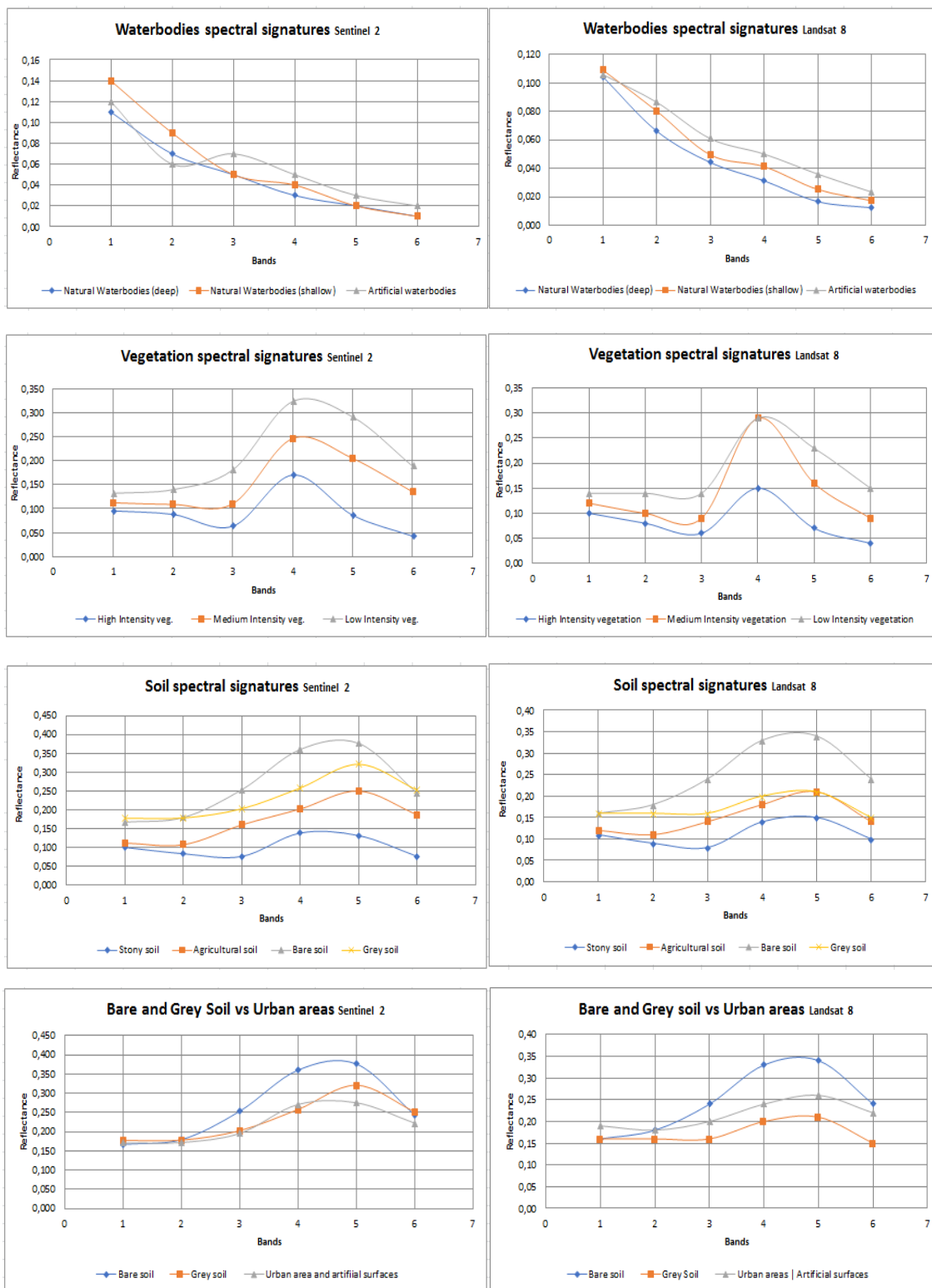
**Πίνακας 13:** Separability array (Landsat 8)

Distance Measure: Transformed Divergence  
Using Layers: 1 2 3 4 5 6  
Taken 6 at a time  
Best Average Separability: 1990.32  
Combination: 1 2 3 4 5 6

| Signature Name                     | 1  | 2    | 3    | 4    | 5       | 6       | 7       | 8       | 9       | 10      | 11      |         |
|------------------------------------|----|------|------|------|---------|---------|---------|---------|---------|---------|---------|---------|
| Natural Waterbodies (deep)         | 1  | 0    | 2000 | 2000 | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    |
| Natural Waterbodies (shallow)      | 2  | 2000 | 0    | 2000 | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    |
| Artificial waterbodies             | 3  | 2000 | 2000 | 0    | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    | 2000    |
| Urban area and artificial surfaces | 4  | 2000 | 2000 | 2000 | 0       | 1999.99 | 2000    | 2000    | 1882.44 | 2000    | 2000    | 1785.31 |
| Grey soil                          | 5  | 2000 | 2000 | 2000 | 1999.99 | 0       | 2000    | 2000    | 1999.99 | 2000    | 2000    | 1982.49 |
| High Intensity vegetation          | 6  | 2000 | 2000 | 2000 | 2000    | 2000    | 0       | 1996.38 | 2000    | 1997.24 | 2000    | 2000    |
| Medium Intensity vegetation        | 7  | 2000 | 2000 | 2000 | 2000    | 2000    | 1996.38 | 0       | 1846.67 | 1999.87 | 1999.99 | 2000    |
| Low Intensity vegetation           | 8  | 2000 | 2000 | 2000 | 1882.44 | 2000    | 2000    | 1846.67 | 0       | 1999.99 | 1990.31 | 1989.49 |
| Stony soil                         | 9  | 2000 | 2000 | 2000 | 2000    | 1999.99 | 1997.24 | 1999.87 | 1999.99 | 0       | 1999.5  | 1999.79 |
| Agricultural soil                  | 10 | 2000 | 2000 | 2000 | 2000    | 2000    | 2000    | 1999.99 | 1990.31 | 1999.5  | 0       | 1998.04 |
| Bare soil                          | 11 | 2000 | 2000 | 2000 | 1785.31 | 1982.49 | 2000    | 2000    | 1989.49 | 1999.79 | 1998.04 | 0       |

Στην επόμενη σελίδα, παρουσιάζεται το μέσο φασματικό προφίλ της κάθε κατηγορίας για τα δεδομένα Sentinel και Landsat αντίστοιχα, όπου μπορούμε να δούμε ποιές κατηγορίες μπορεί να αντιμετωπίσουν προβλήματα φασματικής συσχέτισης και να προκαλέσουν σφάλματα στην ταξινόμηση.

**Πίνακες 14.** Σύγκριση φασματικών υπογραφών μεταξύ αντίστοιχων θεματικών κλάσεων για τα δεδομένα Sentinel-2 και Landsat-8



#### 4.4 Κατασκευή του αλγόριθμου Random Forest

Όπως αναφέρθηκε και στην εισαγωγή, στα πλαίσια της παρούσας διπλωματικής πραγματοποιήθηκε η επεξεργασία και ανάλυση δεδομένων που προέρχονται από τους δορυφόρους Sentinel-2 και Landsat-8 με τη μέθοδο των Τυχαίων Δασών (Random Forest). Αυτή η μέθοδος θα αξιολογηθεί ως προς την απόδοση και την ευαισθησία των παραμέτρων της και θα συγκριθεί στη συνέχεια με άλλες μεθόδους επιβλεπόμενης ταξινόμησης.

Για το σκοπό αυτό, δημιουργήθηκε ένας μοντέλο ταξινόμησης γραμμένο στην ανοικτή γλώσσα προγραμματισμού R, δίνοντας μας τη δυνατότητα να επεξεργαστούμε τα μεγάλου όγκου δεδομένα που έχουμε στη διάθεση μας, να πραγματοποιήσουμε υπολογιστική στατιστική και να δημιουργήσουμε διάφορα γραφήματα σχετικά με τις απαιτήσεις της έρευνας μας.

Συγκεκριμένα, εφαρμόστηκαν δύο στρατηγικές ταξινόμησης, η πρώτη βασίστηκε σε ένα ανισόρροπο σύνολο δεδομένων, χρησιμοποιώντας άνισο πλήθος εικονοστοιχείων εκπαίδευσης για κάθε μια από τις 11 ομάδες κάλυψης γης και η δεύτερη εφαρμόζοντας τη μέθοδο under-sampling για την εξισορρόπηση του πλήθους των δεδομένων εκπαίδευσης για όλες τις κλάσεις, δημιουργώντας ένα ισορροπημένο σύνολο δεδομένων.

Στην υπό-ενότητα Μοντέλο Ταξινόμησης, παρουσιάζεται επεξηγημένος ο κώδικας με αναλυτικές περιγραφές των όσων προκύπτουν από αυτόν. Για σκοπούς εξοικονόμησης χώρου, μερικά από τα γραφήματα και τα αποτελέσματα που συνοδεύουν τον κώδικα, παρουσιάζονται μόνο για τα δεδομένα Landsat 8, ενώ στο τέλος της υπό-ενότητας, παρουσιάζονται συνοπτικά όλα τα σημαντικά αποτελέσματα που προέκυψαν από την ανάλυση και για τις δύο εικόνες (Sentinel-2 και Landsat-8).

#### 4.4.1 Μοντέλο ταξινόμησης και Αποτελέσματα

##### Imbalanced Classification Model

```
# Φόρτωση των απαραίτητων υπολογιστικών πακέτων

library(sp)

library(raster)

library(caret)

library(rgdal)

library(randomForest)

library(e1071)

# Φόρτωση των δεδομένων (Δορυφορικές εικόνες και δεδομένα εκπαίδευσης)

Limg <- brick("F:/BSc_Thesis/Data/L_30m/L8_30m_f.tif")

Simg <- brick("F:/BSc_Thesis/Data/S_30m/s2_30m_f.tif")

names(Simg) <- c(paste0("B", 2:4, coll=""), "B8A", "B11", "B12")

names(Limg) <- c(paste0("B", 2:7))

plot(Simg)

plot(Limg)

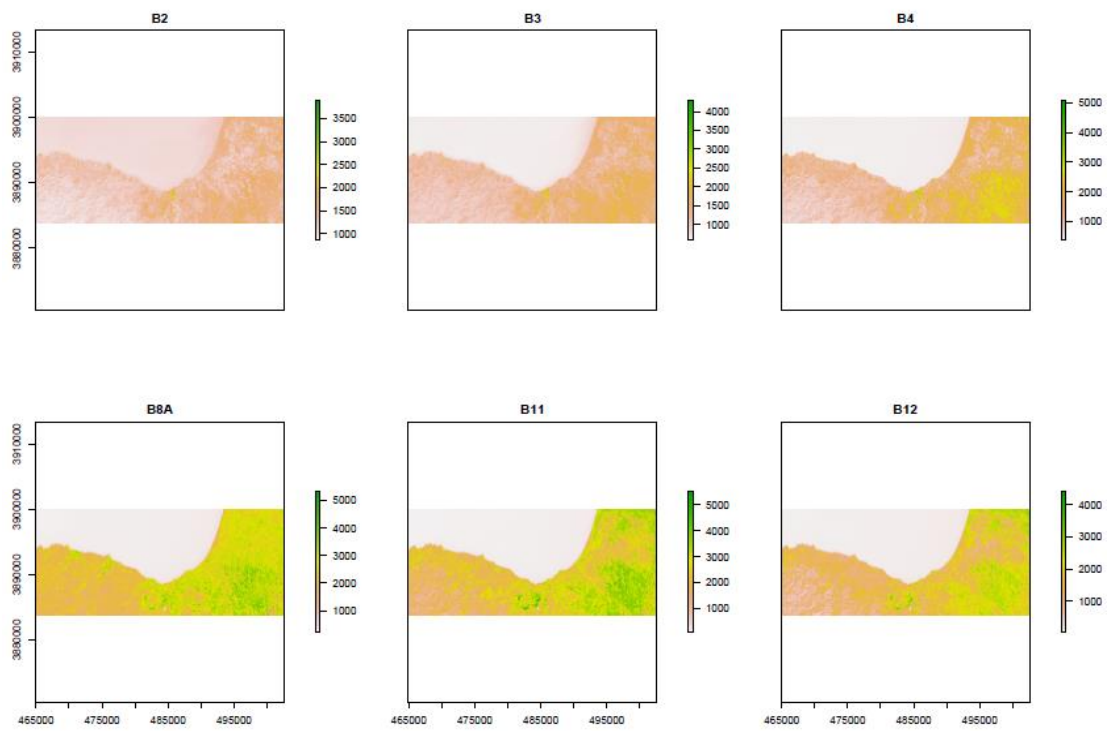
trainData <- shapefile("F:/BSc_Thesis/Data/Training Data/train_Data.shp")

responseCol <- "Classvalue"
```

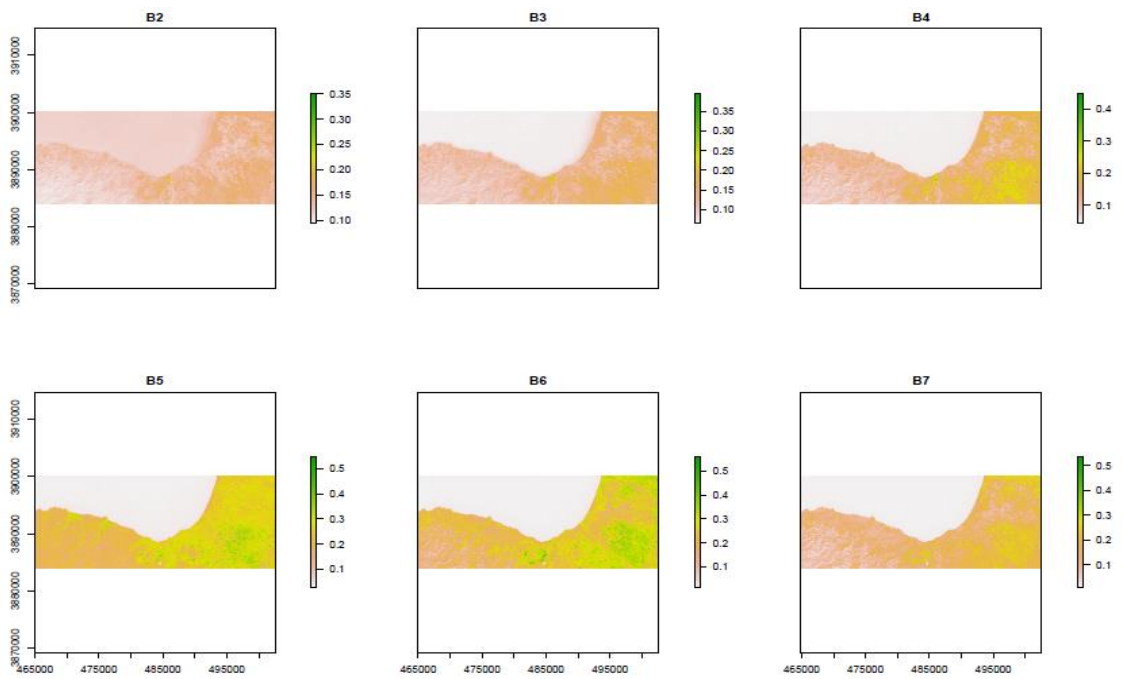
|    | Classname                          | Classvalue | RED | GREEN | BLUE | Count |
|----|------------------------------------|------------|-----|-------|------|-------|
| 0  | Natural Waterbodies (deep)         | 1          | 0   | 197   | 255  | 79    |
| 1  | Natural Waterbodies (shallow)      | 2          | 115 | 255   | 223  | 87    |
| 2  | Artificial Waterbodies             | 3          | 105 | 47    | 82   | 111   |
| 3  | Urban area and artificial surfaces | 4          | 255 | 0     | 0    | 172   |
| 4  | Grey soil                          | 5          | 0   | 0     | 0    | 148   |
| 5  | High Intensity vegetation          | 6          | 76  | 115   | 0    | 163   |
| 6  | Medium Intensity vegetation        | 7          | 56  | 168   | 0    | 219   |
| 7  | Low Intensity vegetation           | 8          | 163 | 255   | 115  | 134   |
| 8  | Stony Soil                         | 9          | 115 | 76    | 0    | 91    |
| 9  | Agricultural Soil                  | 10         | 230 | 152   | 0    | 210   |
| 10 | Bare soil                          | 11         | 137 | 112   | 68   | 288   |

Συνολικός αριθμός  
δειγμάτων εκπαίδευσης  
**1702**





**Διάγραμμα 37:** Φασματικά κανάλια Sentinel-2



**Διάγραμμα 38:** Φασματικά κανάλια Landsat-8



```
# Εξαγωγή των τιμών ανακλαστικότητας του συνόλου των δεδομένων εκπαίδευσης  
από κάθε κανάλι σε ένα dataframe (dfAll)
```

```
dfAll = data.frame(matrix(vector(), nrow = 0, ncol = length(names(Simg)) + 1))
```

```
for
```

```
(i in 1:length(unique(trainData[[responseCol]]))){
```

```
  category <- unique(trainData[[responseCol]][i]
```

```
  categorymap <- trainData[trainData[[responseCol]] == category,]
```

```
  dataSet <- extract(Simg, categorymap)
```

```
  dataSet <- dataSet[!unlist(lapply(dataSet, is.null))]
```

```
  dataSet <- lapply(dataSet, function(x){ cbind(x, class = as.numeric(rep(category,  
nrow(x))))})
```

```
  df <- do.call("rbind", dataSet)
```

```
  dfAll <- rbind(dfAll, df) }
```

|    | B2        | B3         | B4         | B8A        | B11        | B12        | class |
|----|-----------|------------|------------|------------|------------|------------|-------|
| 1  | 0.1166610 | 0.06801482 | 0.04712366 | 0.03124938 | 0.01624868 | 0.01170604 | 1     |
| 2  | 0.1164364 | 0.06793994 | 0.04717358 | 0.03087498 | 0.01542501 | 0.01135660 | 1     |
| 3  | 0.1158873 | 0.06759050 | 0.04682415 | 0.03067531 | 0.01567461 | 0.01103213 | 1     |
| 4  | 0.1156127 | 0.06734091 | 0.04689903 | 0.03040075 | 0.01540005 | 0.01095725 | 1     |
| 5  | 0.1156127 | 0.06716619 | 0.04664943 | 0.03042571 | 0.01532517 | 0.01065774 | 1     |
| 6  | 0.1153382 | 0.06721611 | 0.04659951 | 0.03114954 | 0.01577445 | 0.01090733 | 1     |
| 7  | 0.1158374 | 0.06759050 | 0.04724846 | 0.03085002 | 0.01557477 | 0.01105709 | 1     |
| 8  | 0.1154879 | 0.06766538 | 0.04672431 | 0.03122442 | 0.01574949 | 0.01085741 | 1     |
| 9  | 0.1155129 | 0.06791499 | 0.04702383 | 0.03127434 | 0.01604900 | 0.01098221 | 1     |
| 10 | 0.1156876 | 0.06766538 | 0.04699887 | 0.03087498 | 0.01557477 | 0.01123181 | 1     |
| 11 | 0.1155878 | 0.06739083 | 0.04664943 | 0.03035083 | 0.01537509 | 0.01065774 | 1     |
| 12 | 0.1159621 | 0.06784011 | 0.04694895 | 0.03080011 | 0.01547493 | 0.01108205 | 1     |
| 13 | 0.1160869 | 0.06816458 | 0.04714862 | 0.03157385 | 0.01637348 | 0.01160620 | 1     |
| 14 | 0.1154380 | 0.06789003 | 0.04677423 | 0.03062539 | 0.01542501 | 0.01100717 | 1     |

Showing 1 to 14 of 1,698 entries

Συνολικός αριθμός εικονοστοιχείων εκπαίδευσης

```
# Διαχωρισμός του συνόλου δεδομένων εκπαίδευσης στο υποσύνολο που θα  
χρησιμοποιηθεί για εκπαιδευτικούς σκοπούς (70% - τυχαία επιλεγμένο) και το  
υποσύνολο που θα χρησιμοποιηθεί για σκοπούς αξιολόγησης της ακρίβειας (το  
εναπομένον 30%)
```

```
inBuild <- createDataPartition(y = dfAll$class, p = 0.7, list = FALSE)
```

```
training <- dfAll[inBuild,]
```

```
testing <- dfAll[-inBuild,]
```

```
table(training$class)
```

```
table(testing$class)
```

```
> table(training$class)
 1  2  3  4  5  6  7  8  9 10 11
69 61 77 118 57 118 147 92 96 151 204
> table(testing$class)
 1  2  3  4  5  6  7  8  9 10 11
21 28 33 57 27 43 68 45 47 52 87
```



*Πλήθος εικονοστοιχείων που προορίζεται για εκπαιδευτικούς (training) και σκοπούς εκτίμησης της ακρίβειας (testing)*

*Η κλάση μειονότητας – Natural Waterbodies deep (69 εικονοστοιχεία εκπαίδευσης) είναι μέχρι και 3 φορές μικρότερη από την κλάση πλειοψηφίας Bare soil (204 εικονοστοιχεία εκπαίδευσης).*

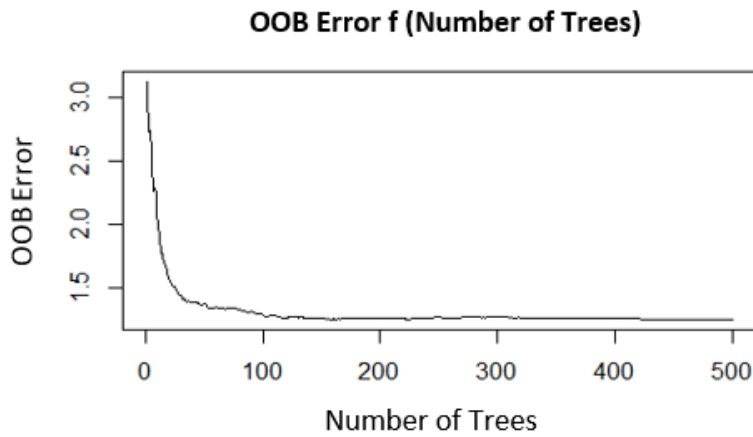
```

# Δημιουργία ενός Ανισόρροπου συνόλου δεδομένων
training_imb <- training[sample(1:nrow(training)), ]
table(training_imb$class)

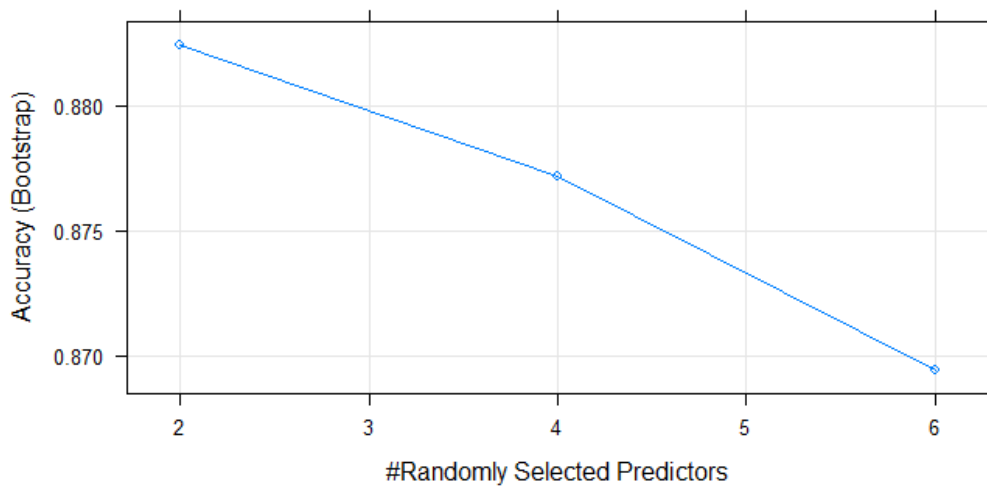
# Εκπαίδευση του μοντέλου Random Forest, με το τυχαία επιλεγμένο υποσύνολο
δεδομένων εκπαίδευσης (training_imb)

# Δημιουργία ενός γραφήματος που να υποδεικνύει την εκτίμηση του σφάλματος
γενίκευσης OOB (Out Of Bag Error – testing dataset) ως συνάρτηση του αριθμού
των δέντρων που απαρτίζουν το δάσος και εφαρμογή του μοντέλου για εύρεση της
τιμής μεταβλητών – καναλιών (mtry) που βελτιστοποιεί την απόδοση του
αλγορίθμου
x<-0
repeat {
  modFit_rf_imb <- train(as.factor(class) ~ B2 + B3 + B4 + B8A + B11 + B12,
  method = "rf", data = training_imb, ntree = x)
  x=x+10
  print(modFit_rf_imb)
  if(x==110)
  repeat {
    modFit_rf_imb <- train(as.factor(class) ~ B2 + B3 + B4 + B8A + B11 + B12,
    method = "rf", data = training_imb, ntree = x, doBest=TRUE)
    x=x+50
    print(modFit_rf_imb)
    if(x==550) { break } }
  plot (modFit_rf_imb) { break } }
  plot (modFit_rf_imb)

```



**Διάγραμμα 39:** Εκτίμηση σφάλματος γενίκευσης OOB, ως συνάρτηση του αριθμού δέντρων που απαρτίζουν το δάσος (Landsat 8)



| mtry | Accuracy  | Kappa     |
|------|-----------|-----------|
| 2    | 0.8824468 | 0.8703767 |
| 4    | 0.8771530 | 0.8645456 |
| 6    | 0.8694147 | 0.8560116 |

Accuracy was used to select the optimal model using the largest value. The final value used for the model was mtry = 2.

**Διάγραμμα 40:** Ακρίβεια του μοντέλου συναρτήσει του αριθμού των μεταβλητών που χρησιμοποιείται σε κάθε εσωτερικό κόμβο διαχωρισμού (Landsat 8)

Τα μέτρα ακρίβειας που φαίνονται πιο πάνω, αναφέρονται στην ικανότητα του αλγόριθμου να προβλέπει τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του, αφού υπολογίζονται ως προς το ποσοστό των δειγμάτων εκπαίδευσης που ταιριάζουν στο μοντέλο που δημιουργήθηκε από τον αλγόριθμο και όχι ως προς την ταξινόμηση που έγινε στα δεδομένα Sentinel-2 και Landsat-8 και τα δεδομένα ελέγχου (πραγματικότητα)

### **Ερμηνεία και Σχολιασμός Αποτελεσμάτων:**

Το διάγραμμα 40, υποδεικνύει τη σημαντικότητα της παραμέτρου *πλήθος των μεταβλητών (καναλιών)* που θα χρησιμοποιηθούν για το διαχωρισμό των δεδομένων σε κάθε εσωτερικό κόμβο, ως προς την απόδοση του αλγόριθμου.

Συγκεκριμένα, η μείωση του αριθμού μεταβλητών οδηγεί στη δημιουργία ακριβέστερων μοντέλων, καταλήγοντας στην εύρεση του βέλτιστου πλήθους μεταβλητών  $mtry=2$ , το οποίο θα χρησιμοποιηθεί και στην συνέχεια της ταξινόμησης.

Επιπλέον, για την εκτίμηση της επίδρασης της παραμέτρου *πλήθος των δέντρων απόφασης* σε κάθε δάσος, παρατηρώντας το διάγραμμα 39, συμπεραίνουμε ότι παρόλο που θεωρητικά όσο αυξάνεται το πλήθος των δέντρων στο δάσος, επηρεάζει αναλογικά με θετικό τρόπο την ακρίβεια της ταξινόμησης, παρατηρούμε πως μετά από κάποιο σημείο (περίπου 100 δέντρα), η ακρίβεια (και το σφάλμα γενίκευσης αντίστοιχα) δεν επηρεάζεται ουσιαστικά (“ασήμαντες” τιμές βελτίωσης.), αφού το γράφημα συγκλίνει.

Συμπεραίνουμε λοιπόν, ότι ο αλγόριθμος δεν είναι τόσο ευαίσθητος στη συγκεκριμένη παράμετρο όσο είναι στην παράμετρο του πλήθους των μεταβλητών που θα καθορίσουν τα όρια διαχωρισμού σε κάθε εσωτερικό κόμβο.

Ακόμα, μέσω των συγκεκριμένων διαγραμμάτων μπορούμε να εντοπίσουμε το βέλτιστο αριθμό δέντρων απόφασης, τα οποία θα δημιουργήσουν ένα δάσος που αφενός θα αποδώσει τα μεγαλύτερα ποσοστά ακρίβειας και αφετέρου θα είναι όσο το δυνατό πιο “ελαφρύ” από άποψη υπολογιστικού χρόνου, έχοντας υπόψη πως διαχειριζόμαστε μεγάλο όγκο δεδομένα.

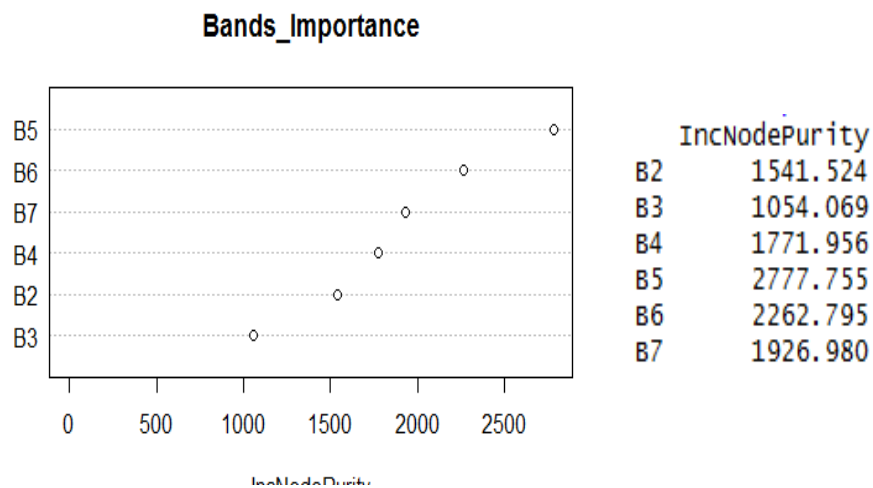
```
# Υπολογισμός της σημαντικότητας κάθε καναλιού στο διαχωρισμό των κλάσεων στην ταξινόμηση
```

```
Bands_Importance<- randomForest(class~.,data=training_imb)
```

```
Bands_Importance
```

```
importance(Bands_Importance, type=2)
```

```
varImpPlot(Bands_Importance)
```



**Διάγραμμα 41:** Μέση τιμή σημαντικότητας των μεταβλητών απο όλα τα δέντρα που απαρτίζουν το δάσος (Landsat 8)

Το μέτρο **IncNodePurity**, αντιπροσωπεύει το πόσο καλά διαχωρίζονται τα δεδομένα στα δέντρα απόφαση, ως συνάρτηση του κάθε καναλιού. Υπάρχουν διάφορα μέτρα έκφρασης αυτής της τιμής, ένας εκ των οποίων είναι ο δείκτης **GINI** που φαίνεται πιο πάνω.

$$\text{Gini}(t) = \sum_{i=1}^L p_{\omega_i}(1 - p_{\omega_i})$$

Καθώς αυτό το μέτρο υπολογίζεται για κάθε μεταβλητή (κανάλι), όσο πιο υψηλή είναι η τιμή του, τόσο πιο χρήσιμη είναι η εν λόγω μεταβλητή στο διαχωρισμό των δεδομένων σε κάθε εσωτερικό κόμβο διαχωρισμού.

# Σχεδιασμός της δομής ενός συγκεκριμένου δέντρου (αρ.δέντρου ->100) στο δάσος

```
> getTree(modFit_rf_imb$finalModel, k = 100, labelVar = FALSE)
  left daughter right daughter split var split point status prediction
1      2          3           3 0.12272619      1      0
2      4          5           1 0.10629032      1      0
3      6          7           2 0.13308442      1      0
4      8          9           5 0.09764183      1      0
5     10         11           3 0.07027366      1      0
6     12         13           2 0.12061711      1      0
7     14         15           4 0.25884320      1      0
8      0          0           0 0.00000000     -1      6
9     16         17           1 0.10566632      1      0
10    18         19           3 0.06130070      1      0
11    20         21           2 0.09297439      1      0
12    22         23           5 0.24170845      1      0
13    24         25           5 0.27527906      1      0
14    26         27           2 0.15937932      1      0
15    28         29           5 0.28965579      1      0
16     0          0           0 0.00000000     -1      6
17     0          0           0 0.00000000     -1      7
18    30         31           4 0.03463140      1      0
19    32         33           1 0.10703910      1      0
20    34         35           4 0.16375972      1      0
21    36         37           2 0.12059215      1      0
22    38         39           4 0.21632457      1      0
23    40         41           2 0.11950640      1      0
24    42         43           4 0.21929476      1      0
25     0          0           0 0.00000000     -1      8
26    44         45           2 0.14148331      1      0
27    46         47           5 0.29799227      1      0
28    48         49           5 0.25136781      1      0
29    50         51           2 0.13746481      1      0
```

Τερματικοί  
κόμβοι Κλάσεις

Διάγραμμα 42: Δέντρο απόφασης, αρ.100 στο τυχαίο δάσος

# Ερμηνεία δέντρου απόφασης

**Left Daughter and Right Daughter** : Πλήθος εσωτερικών κόμβων διαχωρισμού ( ~ 249 στο σύνολο)

**Split Var** : Η μεταβλητή (κανάλι) που χρησιμοποιήθηκε για το διαχωρισμό του εικονοστοιχείου σε κάθε κόμβο. Η τιμή 0 αναφέρεται σε μη τερματικό κόμβο

**Split Point** : Η τιμή ανακλαστικότητας στην οποία ορίστηκε το βέλτιστο όριο διαχωρισμού (με βάση το μέτρο information gain)

**Status** : Αν ο κόμβος είναι τερματικός τότε παίρνει την τιμή -1 αλλιώς αν το εικονοστοιχείο συνεχίζει την πορεία του παίρνει την τιμή 1

**Prediction** : Είναι η πρόβλεψη, η εκτίμηση της κλάσης που ορίστηκε για ένα εικονοστοιχείο (εύρος τιμών: 1-11, ενώ παίρνει την τιμή 0 αν ο κόμβος δεν είναι τερματικός)

```

# Υπολογισμός του πίνακα σύγχυσης

# Αξιολόγηση της ακρίβειας ( Ολική ακρίβεια, στατιστικό kappa και ακρίβεια
του αναλυτή (producer's accuracy (για κάθε κλάση)) για το ανισόρροπο
σύνολο δεδομένων, βάση των δεδομένων ελέγχου (testing dataset)

confusionMatrix(modFit_rf_imb,"none", positive=NULL, dnn = c("Prediction",
"Reference"))

OA_imb <- predict(modFit_rf_imb, testing)

y=1

repeat {

print(confusionMatrix(OA_imb, testing$class)$overall[y])

y=y+1

if(y==3) {break}}

confusionMatrix(OA_imb, testing$class)$byClass[, 1]

```

|            | Reference |     |     |     |     |     |     |     |     |     |     |
|------------|-----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Prediction | 1         | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  |
| 1          | 518       | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   |
| 2          | 0         | 504 | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   |
| 3          | 0         | 0   | 488 | 0   | 0   | 5   | 0   | 0   | 0   | 0   | 0   |
| 4          | 0         | 0   | 0   | 361 | 54  | 0   | 0   | 22  | 0   | 0   | 22  |
| 5          | 0         | 0   | 0   | 37  | 417 | 0   | 0   | 19  | 0   | 1   | 31  |
| 6          | 0         | 0   | 5   | 0   | 0   | 516 | 2   | 0   | 8   | 0   | 0   |
| 7          | 0         | 0   | 0   | 22  | 0   | 3   | 430 | 48  | 6   | 0   | 11  |
| 8          | 0         | 0   | 0   | 48  | 0   | 0   | 31  | 370 | 0   | 22  | 62  |
| 9          | 0         | 0   | 0   | 0   | 0   | 9   | 14  | 0   | 473 | 6   | 0   |
| 10         | 0         | 0   | 0   | 0   | 1   | 0   | 8   | 5   | 10  | 473 | 21  |
| 11         | 0         | 0   | 0   | 34  | 43  | 0   | 3   | 33  | 0   | 10  | 369 |

Accuracy (average) : 0.8823

```

> confusionMatrix(OA_bal, testing$class)$overall[1]
Accuracy
0.9036609
> confusionMatrix(OA_bal, testing$class)$overall[2]
Kappa
0.8926093
> confusionMatrix(OA_bal, testing$class)$byClass[, 1]
Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
Class: 9 Class: 10 Class: 11
0.9615385 1.0000000 1.0000000 0.6400000 0.9117647 0.9607843 0.9491525 0.8387097
0.9791667 0.9710145 0.8275862

```

**Διάγραμμα 43:** Πίνακας σύγχυσης και στατιστικά μέτρα ακρίβειας (Landsat 8)



---

```
# Διαδικασία Parallel Processing για την ταξινόμηση των εικόνων
```

```
beginCluster()
```

```
system.time(preds_rf <- clusterR(Simg, raster::predict, args = list(model =  
modFit_rf_imb),df=T))
```

```
endCluster()
```

```
user  system elapsed  
9.73   8.08  118.70
```

```
RF_Imb=preds_rf
```

```
RF_Bal=preds_rf
```

---

```
# Δημιουργία Θεματικού χάρτη
```

```
library(RColorBrewer)
```

```
colours <- c("deepskyblue2", "deepskyblue", "navy", "red", "dimgrey", "seagreen4",  
"seagreen3", "palegreen", "saddlebrown", "chocolate2", "burlywood3" )
```

```
# Δημιουργία υπομνήματος
```

```
par(xpd = FALSE)
```

```
plot(RF_Bal, col=colours,main="Balanced Classification Results", legend = F)
```

```
par(xpd = TRUE)
```

```
legend(par()$usr[2], 3900000,
```

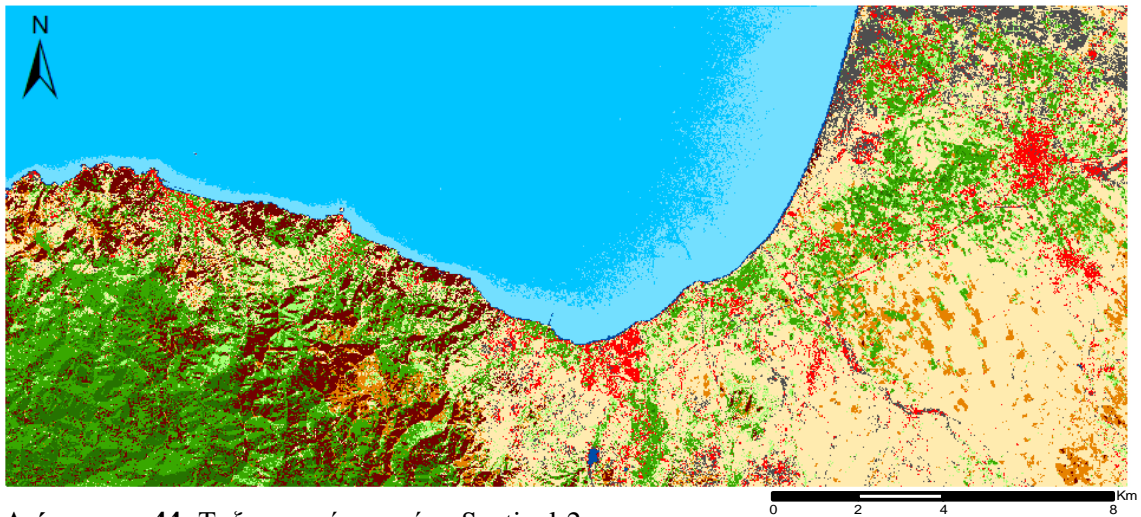
```
legend = c("Natural Waterbodies (deep)", "Natural Waterbodies  
(shallow)", "Artificial waterbodies", "Urban Area and artificial surfaces", "Grey  
soil", "High Intensity vegetation", "MediumIntensity vegetation", "Low Intensity  
vegetation", "Stony soil", "Agricultural soil", "Bare soil"), fill = colours)
```

```
# Εξαγωγή προϊόντος σε μορφή Geotiff για περαιτέρω ανάλυση σε λογισμικά GIS
```

```
RF_Imb <-
```

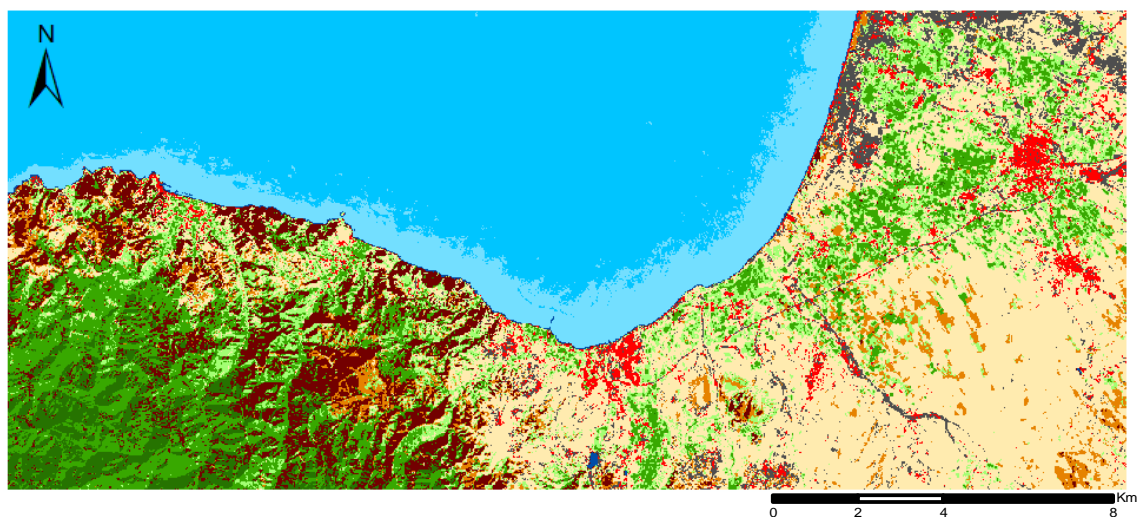
```
writeRaster(RF_Imb,filename="F:/RF_Classified_imgs/S2_RF_Imb.tif",  
format="GTiff", overwrite=TRUE)
```

## Προϊόντα Ανισόρροπης Ταξινόμησης Random Forest (Imbalanced Classification)



**Διάγραμμα 44:** Ταξινομημένη εικόνα Sentinel-2

Overall Acc.: 88.35%, Kappa: 87.26 %



**Διάγραμμα 45:** Ταξινομημένη εικόνα Landsat-8

Overall Acc.: 90.65%, Kappa: 89.69 %

## Balanced Classification Model

```
# Δημιουργία ενός Ισορροπημένου συνόλου δεδομένων, μέσω της εφαρμογής της
# μεθόδου under-sampling, η οποία θα εξισορροπήσει το πλήθος των εικονοστοιχείων
# που θα χρησιμοποιηθούν για εκπαιδευτικούς σκοπούς στην κάθε κλάση.

# Αριθμός εικονοστοιχείων σε κάθε κλάση = 55
# Συνολικός αριθμός εικονοστοιχείων εκπαίδευσης = 605

undersample <- function(x, classCol, nsamples_class){
  for (i in 1:length(unique(x[, classCol]))){
    class.i <- unique(x[, classCol])[i]
    if((sum(x[, classCol] == class.i) - nsamples_class) != 0){
      x <- x[-sample(which(x[, classCol] == class.i),
                     sum(x[, classCol] == class.i) - nsamples_class), ] } }
  return(x) }

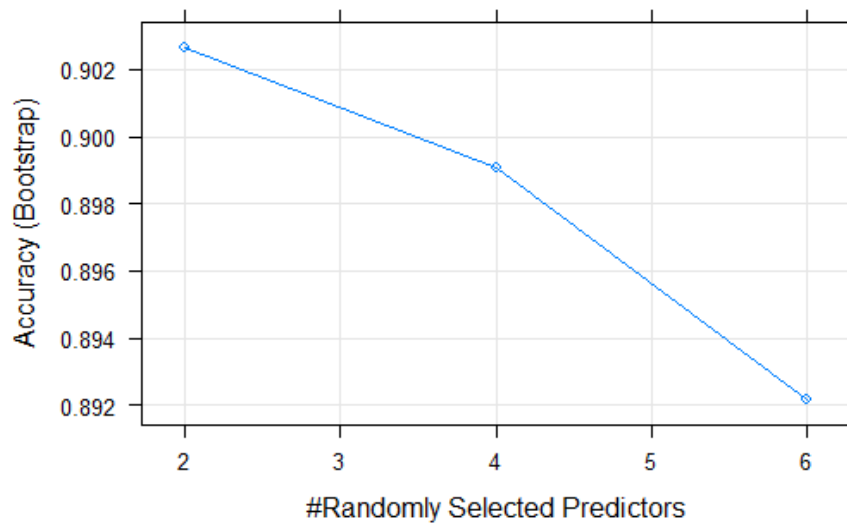
nsamples_class <- 55

training_bal <- undersample(training, "class", nsamples_class)

table(training_bal$class)

 1  2  3  4  5  6  7  8  9 10 11
55 55 55 55 55 55 55 55 55 55 55
```

Αφού δημιουργήθηκε ένα σύνολο δεδομένων με ίσο πλήθος εικονοστοιχείων εκπαίδευσης για κάθε κλάση, το μοντέλο Random Forest εκπαιδεύτηκε βάση αυτών και ακολουθώντας την ίδια σειρά εργασιών υπολογίστηκαν όλα τα γραφήματα και τα μέτρα ακρίβειας που φαίνονται στις επόμενες σελίδες.



```

Random Forest
1216 samples
  6 predictor
 11 classes: '1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 1216, 1216, 1216, 1216, 1216, 1216, ...
Resampling results across tuning parameters:

```

| mtry | Accuracy  | Kappa     |
|------|-----------|-----------|
| 2    | 0.9026641 | 0.8910683 |
| 4    | 0.8990637 | 0.8870431 |
| 6    | 0.8921569 | 0.8793352 |

Accuracy was used to select the optimal model using the largest value.  
The final value used for the model was mtry = 2.

| Prediction | Reference |     |     |     |     |      |      |     |     |      |      |
|------------|-----------|-----|-----|-----|-----|------|------|-----|-----|------|------|
|            | 1         | 2   | 3   | 4   | 5   | 6    | 7    | 8   | 9   | 10   | 11   |
| 1          | 549       | 1   | 0   | 0   | 0   | 0    | 0    | 0   | 0   | 0    | 0    |
| 2          | 11        | 659 | 0   | 0   | 0   | 0    | 0    | 0   | 0   | 0    | 0    |
| 3          | 0         | 0   | 711 | 0   | 0   | 3    | 0    | 0   | 0   | 0    | 0    |
| 4          | 0         | 0   | 0   | 689 | 35  | 0    | 0    | 23  | 0   | 0    | 81   |
| 5          | 0         | 0   | 0   | 68  | 391 | 0    | 0    | 16  | 0   | 0    | 38   |
| 6          | 0         | 0   | 12  | 0   | 0   | 1072 | 37   | 0   | 2   | 0    | 0    |
| 7          | 0         | 0   | 0   | 13  | 0   | 16   | 1352 | 99  | 20  | 0    | 14   |
| 8          | 0         | 0   | 0   | 87  | 0   | 0    | 44   | 579 | 0   | 0    | 48   |
| 9          | 0         | 0   | 0   | 0   | 0   | 32   | 32   | 0   | 899 | 11   | 0    |
| 10         | 0         | 0   | 0   | 4   | 0   | 0    | 15   | 1   | 8   | 1434 | 50   |
| 11         | 0         | 0   | 0   | 147 | 60  | 0    | 2    | 78  | 0   | 8    | 1684 |

Accuracy (average) : 0.8998

```

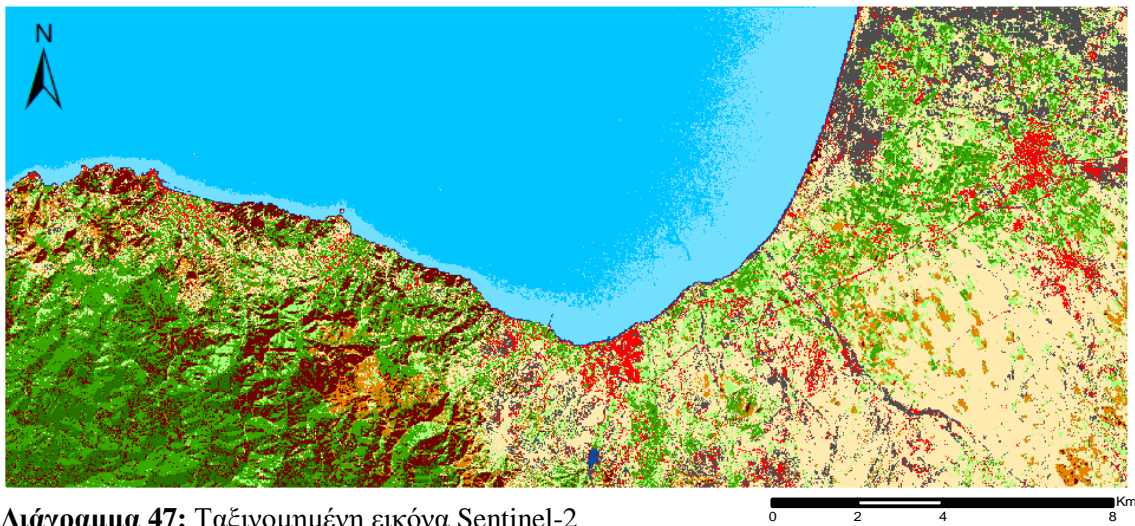
Accuracy
0.9132948
kappa
0.9031367
> confusionMatrix(OA_imb, testing$class)$byClass[, 1]
Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8 Class: 9
0.9583333 1.0000000 1.0000000 0.6875000 0.8648649 0.9791667 1.0000000 0.8181818 1.0000000
Class: 10 Class: 11
0.9811321 0.8965517

```

**Διάγραμμα 46:** Πίνακας Σύγκρισης και στατιστικά μέτρα ακρίβειας - Ισοροπημένο σύνολο δεδομένων (Landsat 8)

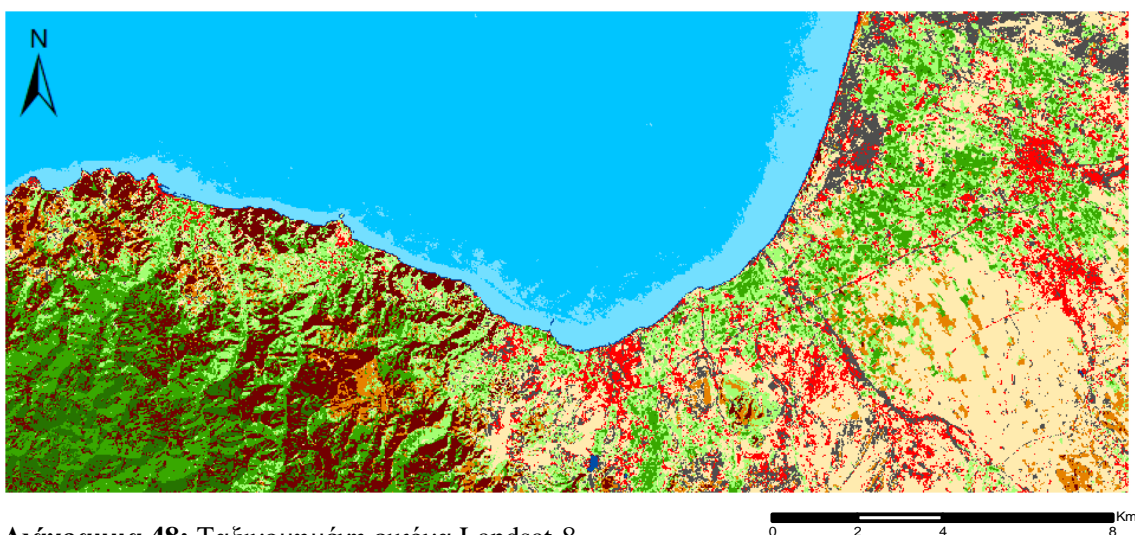


## Προϊόντα Ισορροπημένης Ταξινόμησης Random Forest (Balanced Classification)



**Διάγραμμα 47:** Ταξινομημένη εικόνα Sentinel-2

Overall Acc.: 90.27 %, Καρρα: 89.11 %



**Διάγραμμα 48:** Ταξινομημένη εικόνα Landsat-8

Overall Acc.: 92.88 %, Καρρα : 92.03 %

**Πίνακας 15.** Μέτρα Ακρίβειας των ταξινομήσεων Random Forest σε δεδομένα Sentinel 2 και Landsat 8, για το Ισορροπημένο (Balanced) και Ανισόρροπο (Imbalanced) σύνολο δεδομένων

|                 |                                    | Sentinel-2    |            | Landsat-8     |            |
|-----------------|------------------------------------|---------------|------------|---------------|------------|
|                 |                                    | Random Forest |            | Random Forest |            |
|                 |                                    | Balanced      | Imbalanced | Balanced      | Imbalanced |
|                 | Overall Acc.                       | 90.27%        | 88.35%     | 92.03%        | 89.69%     |
|                 | Kappa St.                          | 89.11%        | 87.26%     | 92.88%        | 90.65%     |
| Producer's Acc. | Natural Waterbodies (deep)         | 100%          | 100%       | 100%          | 100%       |
|                 | Natural Waterbodies (Shallow)      | 100%          | 100%       | 100%          | 100%       |
|                 | Artificial Waterbodies             | 100%          | 100%       | 96.67%        | 96.67%     |
|                 | Urban Area and artificial surfaces | 65.96%        | 68.09%     | 86.79%        | 84.91%     |
|                 | Grey soil                          | 92.59%        | 74.00%     | 87.50%        | 97.50%     |
|                 | High Intensity vegetation          | 93.88%        | 95.92%     | 98.00%        | 96.00%     |
|                 | Medium Intensity vegetation        | 94.12%        | 86.76%     | 96.88%        | 89.01%     |
|                 | Low Intensity vegetation           | 79.31%        | 65.51%     | 73.91%        | 71.74%     |
|                 | Stony soil                         | 92.31%        | 80.77%     | 100.00%       | 100.00%    |
|                 | Agricultural soil                  | 89.51%        | 95.52%     | 98.41%        | 98.40%     |
| Bare soil       | 81.61%                             | 72.41%        | 88.51%     | 82.76%        |            |
| User's Acc.     | Natural Waterbodies (deep)         | 97.70%        | 98.30%     | 98.70%        | 99.80%     |
|                 | Natural Waterbodies (Shallow)      | 99.50%        | 99.00%     | 97.80%        | 98.40%     |
|                 | Artificial Waterbodies             | 99.70%        | 99.80%     | 100.00%       | 99.60%     |
|                 | Urban Area and artificial surfaces | 84.40%        | 86.60%     | 90.10%        | 83.20%     |
|                 | Grey soil                          | 81.70%        | 78.20%     | 92.10%        | 76.20%     |
|                 | High Intensity vegetation          | 95.90%        | 94.10%     | 97.10%        | 95.50%     |
|                 | Medium Intensity vegetation        | 89.50%        | 87.90%     | 88.80%        | 89.30%     |
|                 | Low Intensity vegetation           | 73.90%        | 74.60%     | 82.50%        | 76.40%     |
|                 | Stony soil                         | 96.40%        | 95.60%     | 96.10%        | 92.30%     |
|                 | Agricultural soil                  | 95.60%        | 96.50%     | 95.50%        | 94.80%     |
| Bare soil       | 85.80%                             | 76.50%        | 89.20%     | 97.80%        |            |

### Ερμηνεία και Σχολιασμός Αποτελεσμάτων:

Όπως βλέπουμε από τον πιο πάνω πίνακα, τα ποσοστά ακρίβειας που προέκυψαν από τη διαδικασία ελέγχου είναι ιδιαίτερα υψηλά και στις δύο περιπτώσεις ταξινομήσεων (Ισορροπημένο και Ανισόρροπο σύνολο δεδομένων εκπαίδευσης).

Ωστόσο, μπορούμε να παρατηρήσουμε ότι η ολική ακρίβεια και ο δείκτης kappa, υπερέρχουν κατά 1,92 % και 1,85 % αντίστοιχα στα δεδομένα Sentinel -2 και κατά 2,23 % και 2,34 % αντίστοιχα στα δεδομένα Landsat 8, στην περίπτωση της ισορροπημένης ταξινόμησης (Balanced classification). Τα συγκεκριμένα ποσοστά διαφοράς μπορεί να μην θεωρηθούν αρκετά σημαντικά, αλλά αυτό μπορεί να αποδοθεί στο γεγονός ότι η αναλογία μεταξύ της κλάσης πλειοψηφίας (Bare soil → 288 δείγματα εκπαίδευσης) και της κλάσης μειονότητας (Stony soil → 91 δείγματα εκπαίδευσης) είναι μόλις 1 προς 3, έτσι το σύνολο δεδομένων μπορεί να μην είναι και τόσο ανισόρροπο τελικά.

Επιπλέον, επιβεβαιώνονται τα όσα προαναφέρθηκαν στην ενότητα 2.3.1.1 (Ευαισθησία του ταξινομητή στα δείγματα εκπαίδευσης), αφού παρατηρούμε ότι η ακρίβεια του παραγωγού στην κλάση μειονότητας Stony soil, είναι χαμηλότερη κατά 11,54 %, στην περίπτωση του ανισόρροπου συνόλου δεδομένων από ότι στην περίπτωση του ισορροπημένου συνόλου δεδομένων.

Όσο αφορά τη σύγκριση μεταξύ των αισθητήρων MSI και OLI, παρατηρείται μια μικρή υπεροχή της τάξης του 2,61% για τα δεδομένα Landsat-8 στην ολική ακρίβεια. Ακόμα, το στατιστικό kappa το οποίο εκφράζει το ποσοστό των σφαλμάτων που απέφυγε η πραγματοποιημένη ταξινόμηση σε σχέση με τα σφάλματα που θα είχε μια τυχαία ταξινόμηση, είναι 89,26 % για τα δεδομένα Sentinel-2 και 92.03 % για τα δεδομένα Landsat-8.

Στα αποτελέσματα της ακρίβειας του παραγωγού, τα ποσοστά κυμαίνονται από 84 % έως 100 % στις περισσότερες κλάσεις, υποδεικνύοντας την υψηλή απόδοση του αλγόριθμου. Η κλάση “Urban areas and artificial surfaces” και “Low Intensity vegetation”, παρουσιάζουν τα χαμηλότερα ποσοστά ακρίβειας, επηρεάζοντας αρνητικά την απόδοση του ταξινομητή.

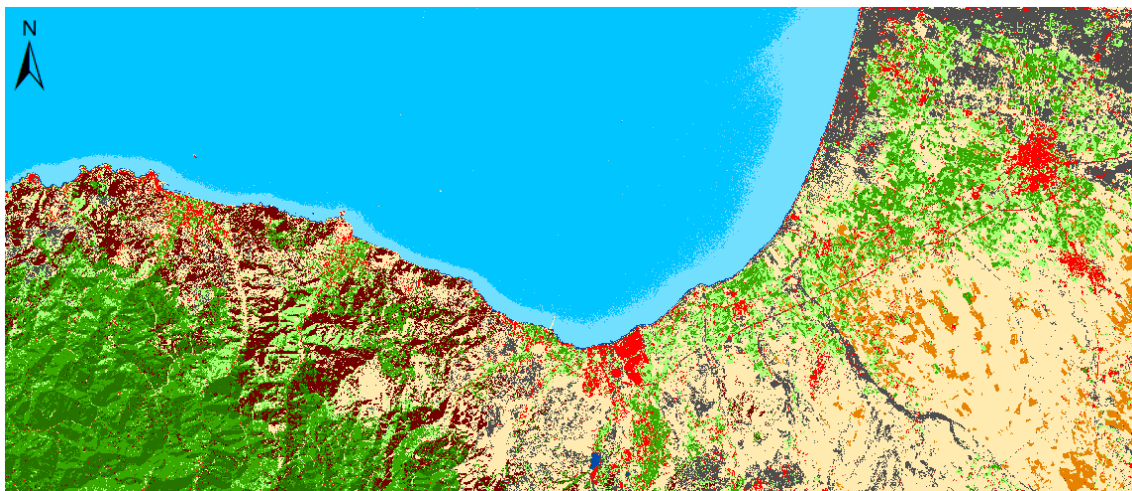
Παρατηρούμε επίσης, ότι υπάρχει διακύμανση στα ποσοστά ακρίβειας της κλάσης “Urban areas and artificial surfaces” → 65,96 % για τον αισθητήρα MSI και 86,79 % για τον αισθητήρα OLI. Αυτό το αποτέλεσμα ήταν “αναμενόμενο”, αφού όπως είδαμε στην ενότητα 4.3.2 η φασματική διαχωριστικότητα της συγκεκριμένης κλάσης χαρακτηρίστηκε ως μέτρια, λόγω της ομοιότητας της με την κλάση Bare soil. Εντούτοις, η ακρίβεια του χρήστη στην εν λόγω κατηγορία, ήταν υψηλή και για τις δύο εικόνες, παρέχοντας μας υψηλό επίπεδο αξιοπιστίας του τελικού προϊόντος.

Σχετικά με τη “σημαντικότητα” του κάθε καναλιού ως προς τον διαχωρισμό των δεδομένων στους κόμβους των δέντρων απόφασης, όπως βλέπουμε από το γράφημα στο διάγραμμα 41, τα κανάλια στο ορατό φάσμα (B2, B3, B4) είναι λιγότερο σημαντικά από τα κανάλια του υπέρυθρου φάσματος (B5, B6, B7) και στις δύο περιπτώσεις ταξινόμησης. Συγκεκριμένα, το σημαντικότερο κανάλι φαίνεται να είναι το Narrow NIR στην περίπτωση της ανισόρροπης ταξινόμησης και το Shortwave IR 1 στην περίπτωση της ισορροπημένης.

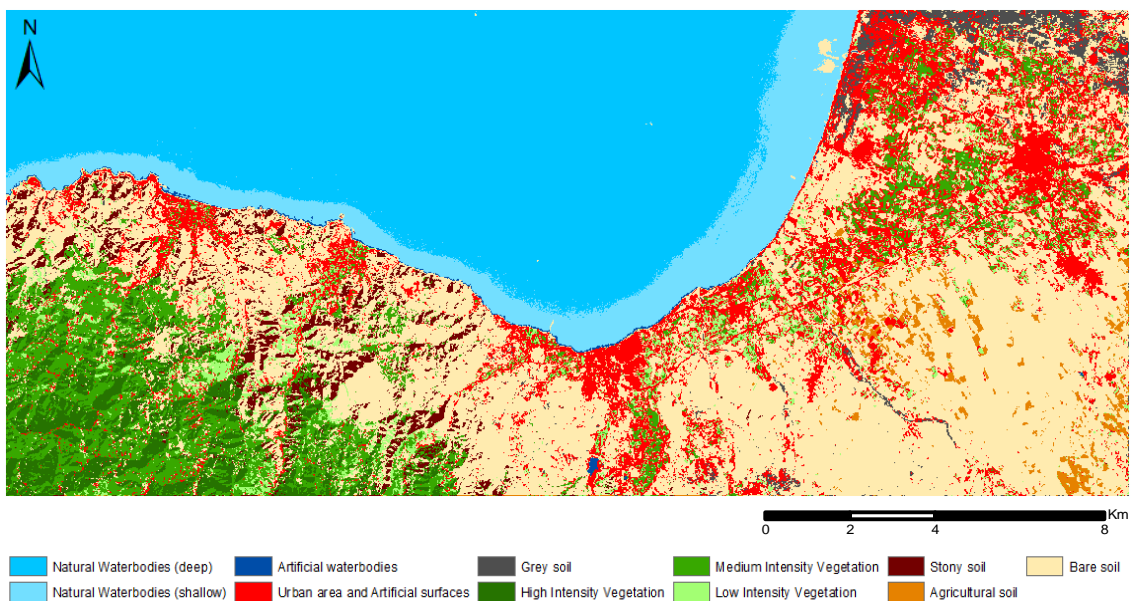
## 4.5 Άλλες ταξινομήσεις

Παρά τα υψηλά επίπεδα ακρίβειας και αξιοπιστίας που προέκυψαν από την εφαρμογή του αλγόριθμου Τυχαίων Δασών, θεωρήθηκε επίσης αναγκαία η σύγκριση του με άλλους γνωστούς ταξινομητές, ούτως ώστε να δημιουργήσουμε μια ολοκληρωμένη εικόνα των δυνατοτήτων του. Έτσι, ακολούθησε η εφαρμογή των παραμετρικών επιβλεπόμενων ταξινομήσεων της Ελάχιστης Απόστασης, της Απόστασης Mahalanobis και της Μέγιστης Πιθανοφάνειας, ως προς τα ίδια δείγματα εκπαίδευσης. Τα τελικά προϊόντα παρουσιάζονται στις επόμενες σελίδες.

### Προϊόντα Ταξινόμησης Μέγιστης Πιθανοφάνειας (Maximum Likelihood)



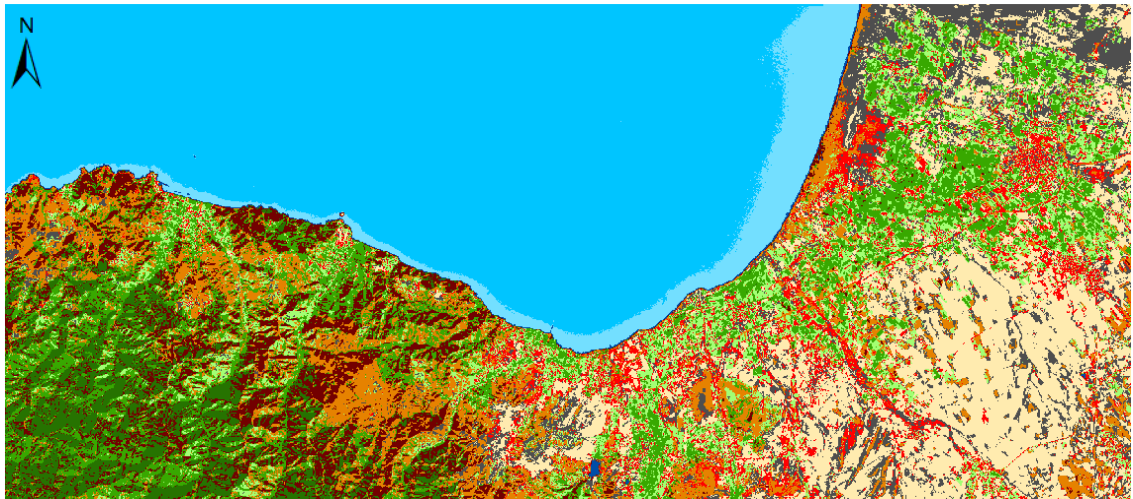
**Διάγραμμα 49:** Ταξινομημένη εικόνα Sentinel-2 , Overall Acc.: 81 % , Kappa: 79.06



**Διάγραμμα 50:** Ταξινομημένη εικόνα Landsat 8, Overall Acc.: 83%, Kappa: 81.27 %



## Προϊόντα Ταξινόμησης Ελάχιστης Απόστασης (Minimum Distance)



**Διάγραμμα 51:** Ταξινομημένη εικόνα Sentinel-2

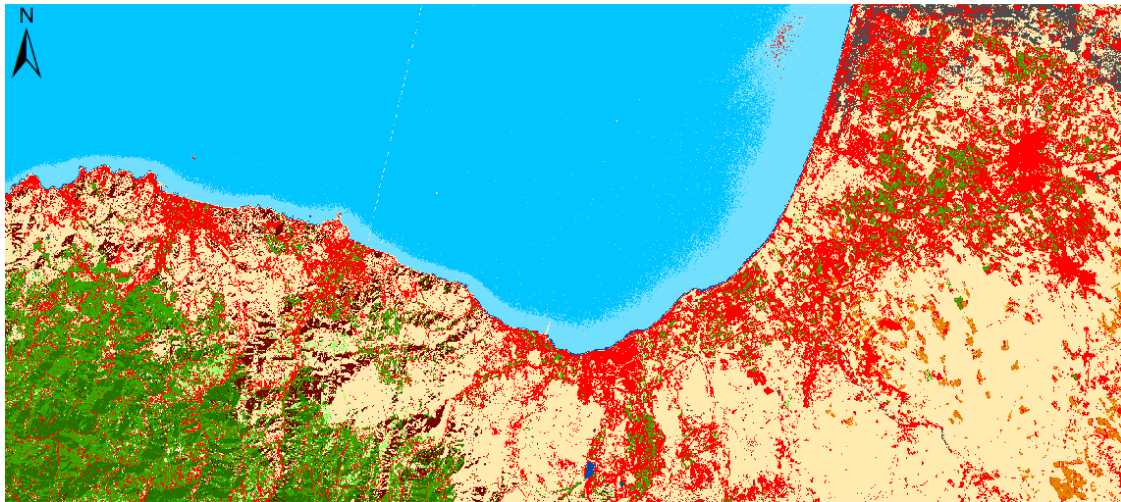
Overall Acc.:69%, Kappa: 66 %



**Διάγραμμα 52:** Ταξινομημένη εικόνα Landsat 8

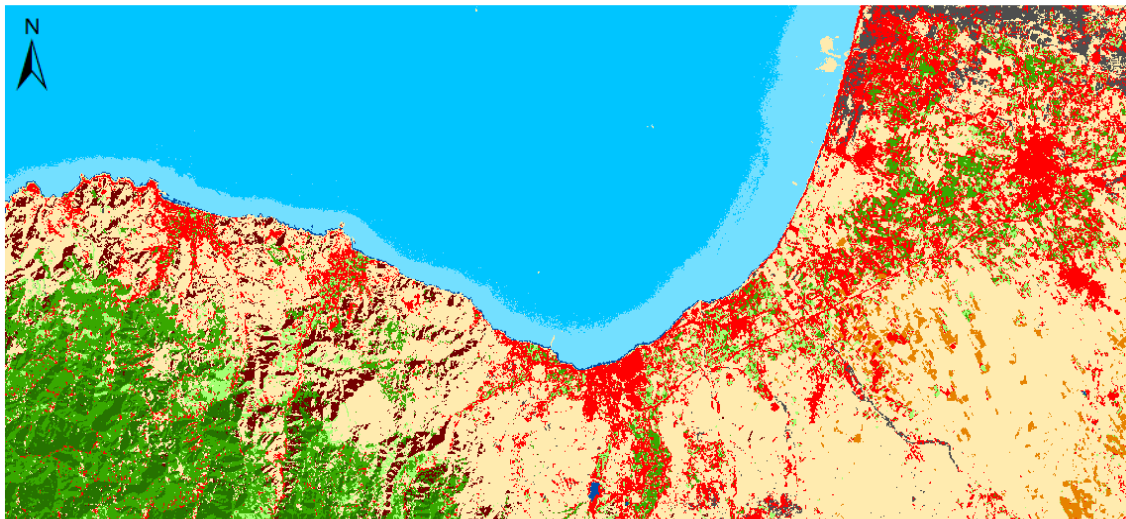
Overall Acc.: 72%, Kappa: 67 %

### Προϊόντα Ταξινόμησης Απόστασης Mahalanobis (Mahalanobis Distance)



**Διάγραμμα 53:** Ταξινομημένη εικόνα Sentinel-2

Overall Acc.: 72.9%, Kappa: 69.3 %



**Διάγραμμα 54:** Ταξινομημένη εικόνα Landsat 8

Overall Acc.: 74%, Kappa: 69.2 %

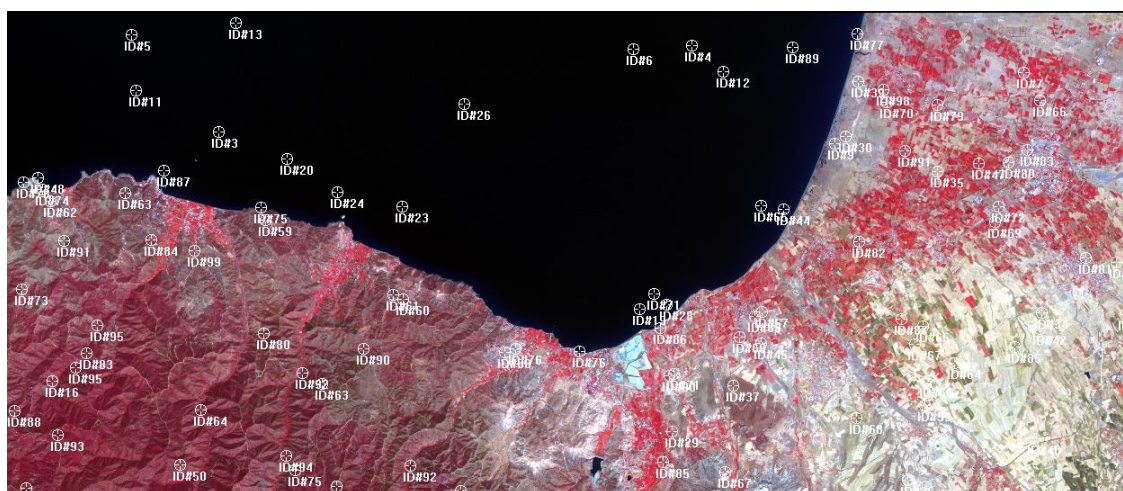


Η αξιολόγηση της κάθε μίας από τις ταξινομήσεις, πραγματοποιήθηκε στο λογισμικό erdas imagine 2014 με τη χρήση εκατό σημείων ελέγχου, κατανεμημένων όσο το δυνατό ισάριθμα σε όλες τις κατηγορίες.

Ο αριθμός των σημείων ελέγχου σε κάθε κατηγορία φαίνεται στον πίνακα 13, ενώ η χωρική κατανομή τους απεικονίζεται στην διάγραμμα 53. Τα αποτελέσματα που προέκυψαν από την κάθε ταξινόμηση, απεικονίζονται συγκριτικά με τα αποτελέσματα απόδοσης του αλγόριθμου Random Forest στον πίνακα 14.

**Πίνακας 16.** Αριθμός σημείων ελέγχου ανά κατηγορία κάλυψη γης

| Κατηγορίες κάλυψης γης             | Αριθμος σημειων ελεγχου |
|------------------------------------|-------------------------|
| Natural Waterbodies (deep)         | 8                       |
| Natural Waterbodies (Shallow)      | 11                      |
| Artificial Waterbodies             | 5                       |
| Urban Area and artificial surfaces | 11                      |
| Grey soil                          | 10                      |
| High Intensity vegetation          | 10                      |
| Medium Intensity vegetation        | 9                       |
| Low Intensity vegetation           | 7                       |
| Stony soil                         | 12                      |
| Agricultural soil                  | 9                       |
| Bare soil                          | 8                       |
| <b>Σύνολο</b>                      | <b>100</b>              |



**Διάγραμμα 55:** Χωρική κατανομή σημείων ελέγχου στην περιοχή μελέτης

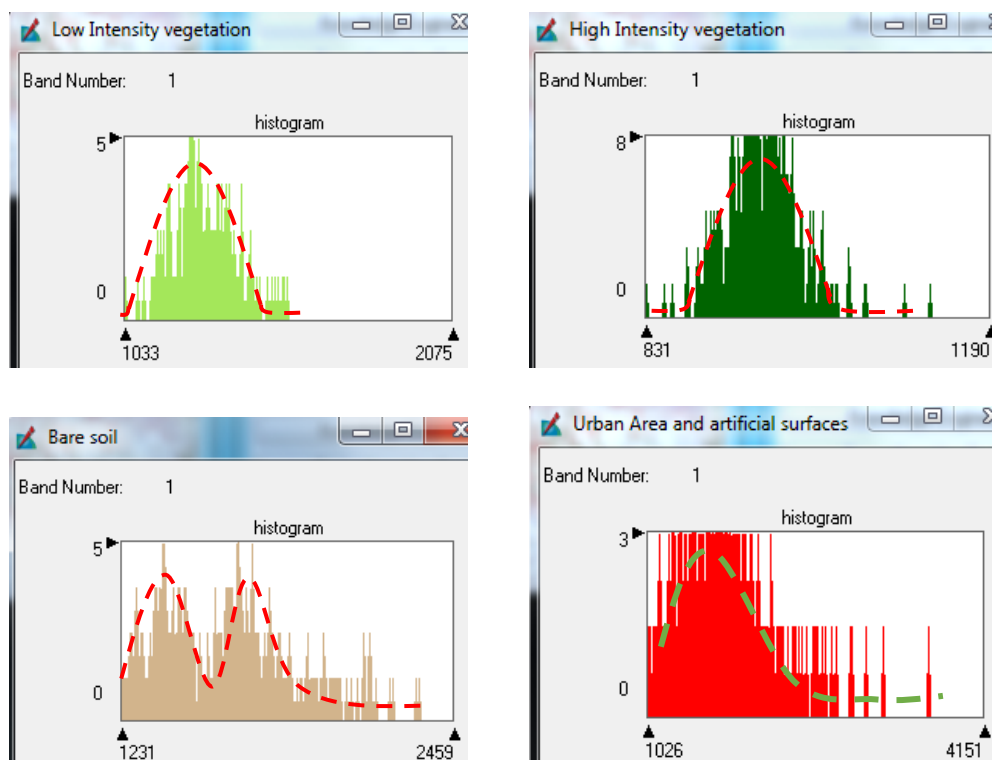
**Πίνακας 17.** Μέτρα Ακρίβειας των ταξινόμησεων Mamimum Likelihood, Minimum Distance και Mahalanobis Distance, σε σύγκριση με την απόδοση του αλγόριθμου Random Forest σε δεδομένα Sentinel 2 και Landsat 8

|                                    | Overall Acc.<br>Kappa St. | Sentinel 2    |                 |            |          |          | Landsat 8     |                 |            |          |          |
|------------------------------------|---------------------------|---------------|-----------------|------------|----------|----------|---------------|-----------------|------------|----------|----------|
|                                    |                           | RF (Balanced) | RF (Imbalanced) | Max.Likel. | Mah.Dist | Min.Dist | RF (Balanced) | RF (Imbalanced) | Max.Likel. | Mah.Dist | Min.Dist |
|                                    |                           | 90,27%        | 88,35%          | 81%        | 72,98%   | 69%      | 92,88%        | 90,65%          | 83,00%     | 74%      | 72%      |
|                                    | Number of Training pixels | 89,11%        | 87,25%          | 79,06%     | 69,30%   | 66%      | 92,03%        | 89,69%          | 81,27%     | 69,20%   | 67%      |
| Natural Waterbodies (deep)         | 55                        | 100%          | 100%            | 90%        | 100%     | 100%     | 100%          | 100%            | 87,50%     | 100%     | 100%     |
| Natural Waterbodies (Shallow)      | 61                        | 100%          | 100%            | 88,89%     | 100%     | 100%     | 100%          | 100%            | 72,73%     | 100%     | 100%     |
| Artificial Waterbodies             | 78                        | 100%          | 100%            | 100%       | 100%     | 100%     | 96,67%        | 100%            | 100%       | 95%      | 100%     |
| Urban Area and artificial surfaces | 120                       | 65,96%        | 68,09%          | 70%        | 32,50%   | 26,67%   | 86,79%        | 84,91%          | 81,82%     | 36%      | 62%      |
| Grey soil                          | 104                       | 74%           | 92,59%          | 58,33%     | 80%      | 75%      | 87,50%        | 87,50%          | 70%        | 70%      | 64%      |
| High Intensity vegetation          | 114                       | 93,88%        | 95,92%          | 90,00%     | 55,00%   | 45,45%   | 98%           | 96%             | 90%        | 53%      | 38%      |
| Medium Intensity vegetation        | 153                       | 94,12%        | 86,76%          | 88,89%     | 54,55%   | 80%      | 96,88%        | 89,01%          | 77,78%     | 54%      | 60%      |
| Low Intensity vegetation           | 94                        | 65,51%        | 79,31%          | 100%       | 40%      | 45,45%   | 73,91%        | 71,74%          | 85,71%     | 40%      | 73%      |
| Stony soil                         | 64                        | 92,31%        | 80,77%          | 71,43%     | 85%      | 92%      | 100%          | 100%            | 83,33%     | 88%      | 90%      |
| Agricultural soil                  | 147                       | 89,51%        | 95,52%          | 100%       | 86%      | 95%      | 98,41%        | 98,40%          | 100%       | 86%      | 98%      |
| Bare soil                          | 202                       | 81,61%        | 72,41%          | 62,50%     | 92%      | 80%      | 88,51%        | 82,76%          | 75%        | 93%      | 83%      |
| Natural Waterbodies (deep)         | 55                        | 97,70%        | 98,30%          | 100%       | 96%      | 100%     | 98,70%        | 99,80%          | 77,78%     | 100%     | 100%     |
| Natural Waterbodies (Shallow)      | 61                        | 99,50%        | 99%             | 88,89%     | 76%      | 72,50%   | 97,80%        | 98,40%          | 88,89%     | 100%     | 84%      |
| Artificial Waterbodies             | 78                        | 99,70%        | 99,80%          | 66,67%     | 100%     | 100%     | 100%          | 99,60%          | 83,33%     | 100%     | 100%     |
| Urban Area and artificial surfaces | 120                       | 84,40%        | 86,60%          | 70%        | 66,67%   | 67,50%   | 90,10%        | 83,20%          | 90%        | 76%      | 62,50%   |
| Grey soil                          | 104                       | 81,70%        | 78,20%          | 77,78%     | 66,67%   | 45%      | 92,10%        | 76,20%          | 77,78%     | 62%      | 60%      |
| High Intensity vegetation          | 114                       | 95,90%        | 94,10%          | 90%        | 66%      | 83,33%   | 97,10%        | 95,50%          | 90%        | 63%      | 44,44%   |
| Medium Intensity vegetation        | 153                       | 89,50%        | 87,90%          | 88,89%     | 55,60%   | 46%      | 88,80%        | 89,30%          | 77,78%     | 33%      | 50%      |
| Low Intensity vegetation           | 94                        | 73,90%        | 74,60%          | 70%        | 45%      | 28,57%   | 96,10%        | 76,40%          | 60%        | 55%      | 46%      |
| Stony soil                         | 64                        | 96,40%        | 95,60%          | 100%       | 72,40%   | 84%      | 96,10%        | 92,30%          | 100%       | 100%     | 50%      |
| Agricultural soil                  | 147                       | 95,60%        | 96,50%          | 70%        | 100%     | 84%      | 95,50%        | 94,80%          | 90%        | 100%     | 57%      |
| Bare soil                          | 202                       | 85,80%        | 76,50%          | 62,50%     | 81,88%   | 73,08%   | 89,20%        | 87,80%          | 75%        | 75,23%   | 81,43%   |

### Ερμηνεία και Σχολιασμός Αποτελεσμάτων:

Τα αποτελέσματα που προκύπτουν από την εφαρμογή του κανόνα της Μέγιστης Πιθανοφάνειας, είναι τα αμέσως καλύτερα μετά από αυτά των Τυχαίων Δασών (OA: 81%, kappa: 79,06 % για τα δεδομένα Sentinel-2 και OA: 83%, kappa:81,27 % για τα δεδομένα Landsat-8). Αυτό μας βοηθά να κατανοήσουμε ότι ο αλγόριθμος αυτός αποτελεί μια πολύ ακριβή μέθοδο ταξινόμησης, με την προϋπόθεση ότι τα δεδομένα εισόδου έχουν οριστεί σωστά. Αυτό σημαίνει ότι είναι ταυτόχρονα αντιπροσωπευτικά της κάθε θεματικής κατηγορίας αλλά και ομοιογενή, καταλήγοντας στη δημιουργία τάξεων των οποίων τα ιστογράμματα προσαρμόζονται όσο το δυνατό καλύτερα στην κανονική κατανομή.

Αντιπροσωπευτικά παραδείγματα του περιορισμού περί της μορφής της κατανομής που έχουν τα δεδομένα, αποτελούν (για τα δεδομένα Sentinel 2) αφενός οι κλάσεις Low Intensity vegetation (PA → 90%), High Intensity vegetation (PA → 100%) και Stony soil (PA → 100%) %, και αφετέρου οι προβληματικές κλάσεις Urban areas and artificial surfaces (PA → 70%) και Bare soil (PA → 62.5%), που επηρέασαν αρνητικά την συνολική ακρίβεια της ταξινόμησης (OA→ 79.06%).



**Διάγραμμα 56:** Ιστογράμματα κατανομών των κλάσεων Low και High Intensity vegetation, Bare soil και Urban areas and artificial surfaces

Στα ιστογράμματα Bare soil και Urban areas and artificial surfaces, προφανώς η υπόθεση κανονικότητας δεν ισχύει, αφού στις κατανομές υπάρχουν περισσότερες από μία κορυφές και μεγάλη μεταβλητότητα. Για αυτό το λόγο, πριν από κάθε ταξινόμηση, απαιτείται ο έλεγχος των ιστογραμμάτων των δεδομένων που έχουν συλλεχθεί και στην περίπτωση που εντοπισθούν μεγάλες αποκλίσεις από την κανονική κατανομή, η χρήση του αλγόριθμου δεν ενδείκνυται.

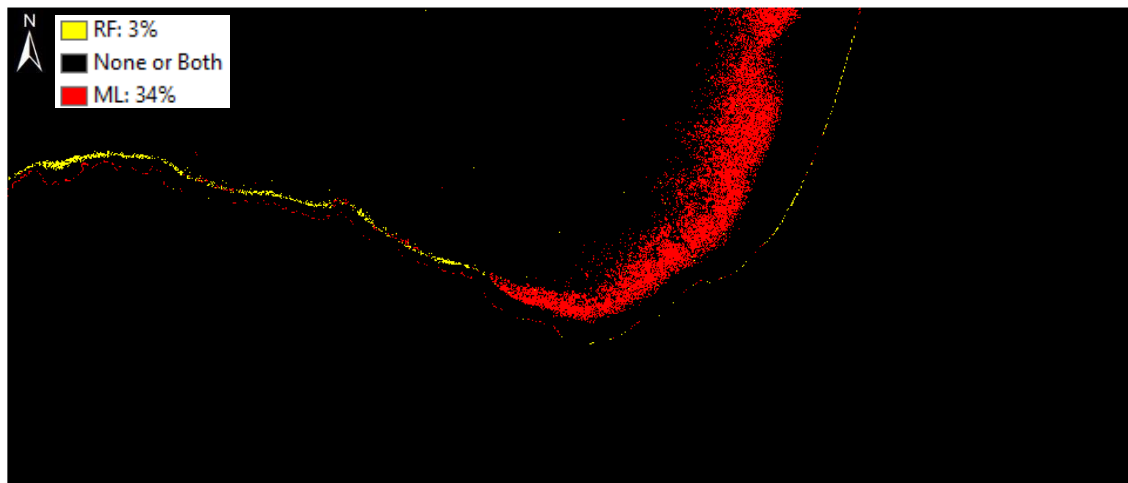
Ακολούθως, ο αλγόριθμος της Απόστασης Mahalanobis ταξινόμησε σωστά το 74 % του συνόλου των εικονοστοιχείων, ενώ το ποσοστό των σφαλμάτων που απέφυγε η πραγματοποιημένη ταξινόμηση σε σχέση με τα σφάλματα που θα είχε μια τυχαία ταξινόμηση είναι 69,2 % για τα δεδομένα Landsat που παρουσίασαν γενικότερα ποιοτικότερα αποτελέσματα σε σχέση με τα δεδομένα Sentinel 2.

Τις χαμηλότερες ακρίβειες ταξινόμησης παρουσιάζει ο αλγόριθμος της Ελάχιστης Απόστασης (O.A. →72% , Kappa→67%), ο οποίος κατά την ταξινόμηση αγνοεί την διασπορά των δεδομένων και δεν μειώνει με αυτόν τον τρόπο τη ευκλείδεια απόσταση για μη συσχετισμένες μεταβλητές.

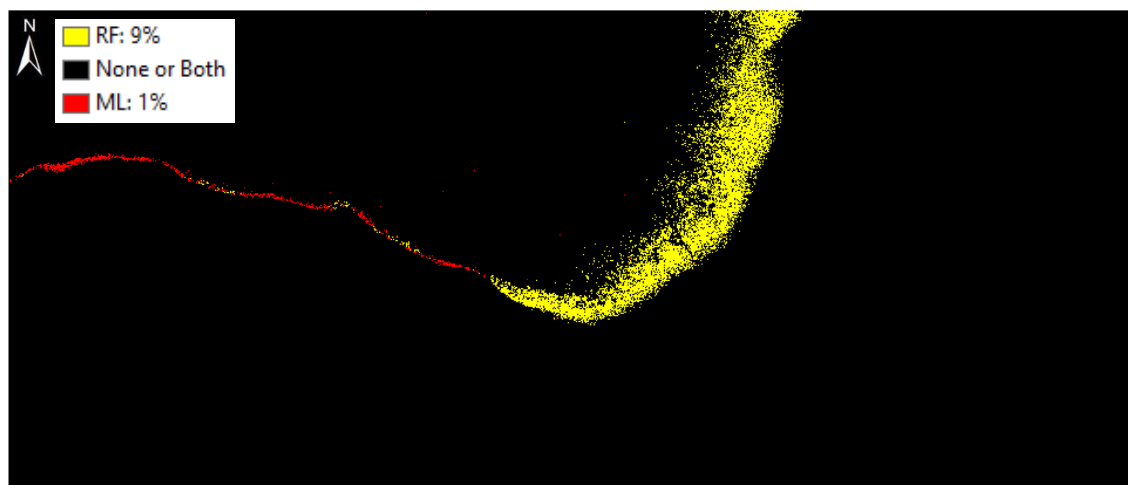
#### 4.6 Μετά-ταξινόμηση ανάλυση

Τα αποτελέσματα που προέκυψαν από την εφαρμογή των διαφόρων μεθόδων ταξινόμησης, μας καθόρισαν την ακριβέστερη τεχνική που ήταν ο αλγόριθμος των Τυχαίων Δασών και την αμέσως καλύτερη που ήταν ο αλγόριθμος της Μέγιστης Πιθανοφάνειας.

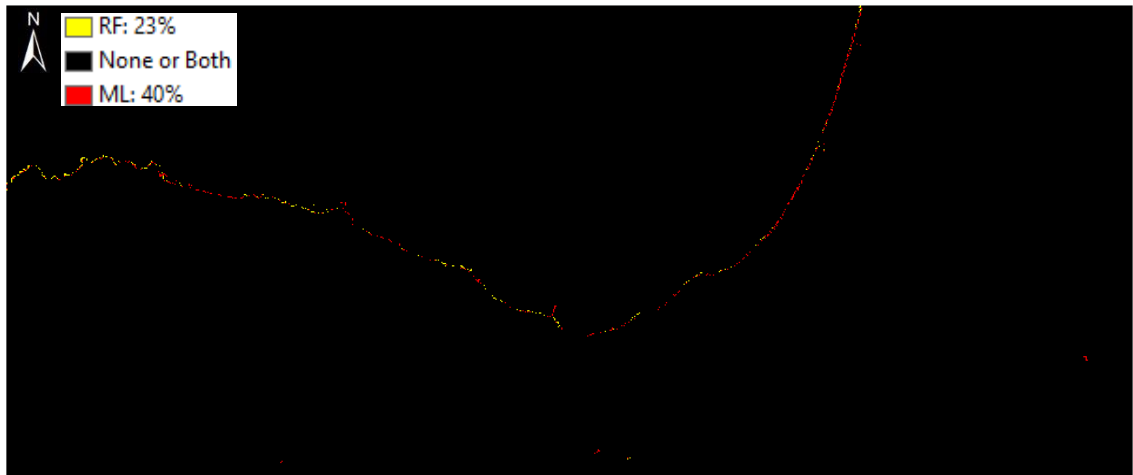
Στα τελικά προϊόντα αυτών των δύο ταξινομητών για τα δεδομένα Sentinel-2, εντοπίστηκαν επίσης οι χωρικές διαφορές των εικονοστοιχείων μεταξύ κάθε τάξης.



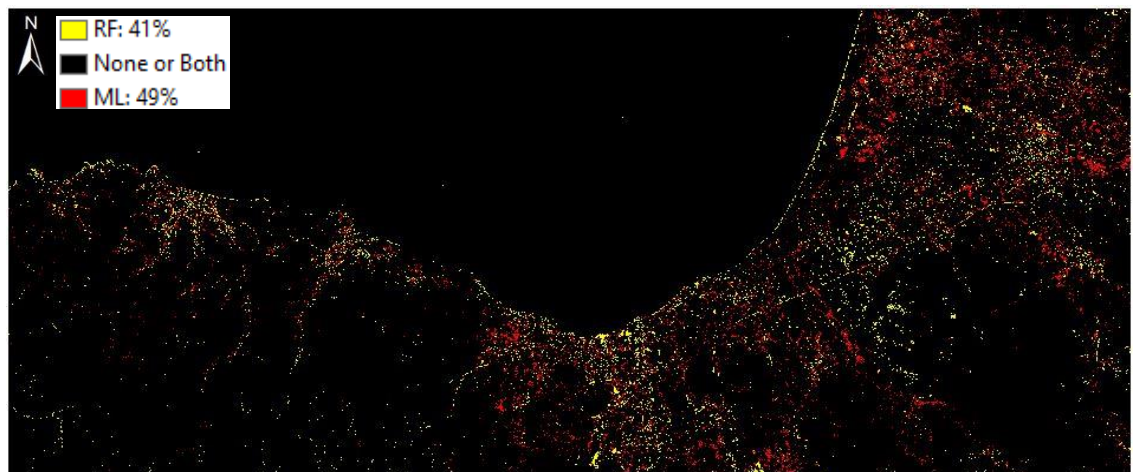
**Διάγραμμα 57:** Διαφορές Natural Waterbodies (deep) μεταξύ Random Forest και Maximum Likelihood (Sentinel-2)



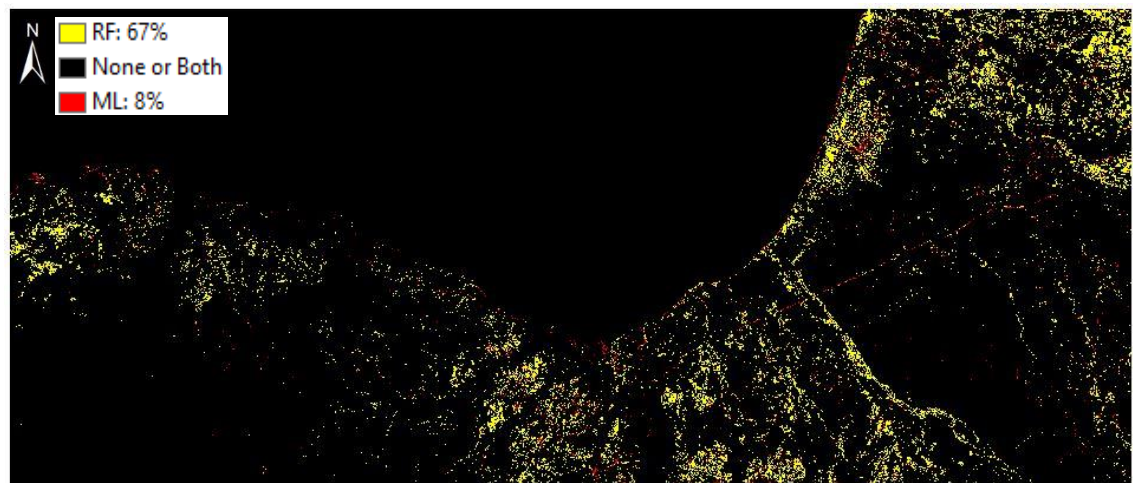
**Διάγραμμα 58:** Διαφορές Natural Waterbodies (shallow), μεταξύ Random Forest και Maximum Likelihood (Sentinel-2)



**Διάγραμμα 57.** Διαφορές Artificial Waterbodies, μεταξύ Random Forest και Maximum Likelihood (Sentinel-2)



**Διάγραμμα 58.** Διαφορές Urban areas and Artificial surfaces, μεταξύ Random Forest και Maximum Likelihood (Sentinel-2)

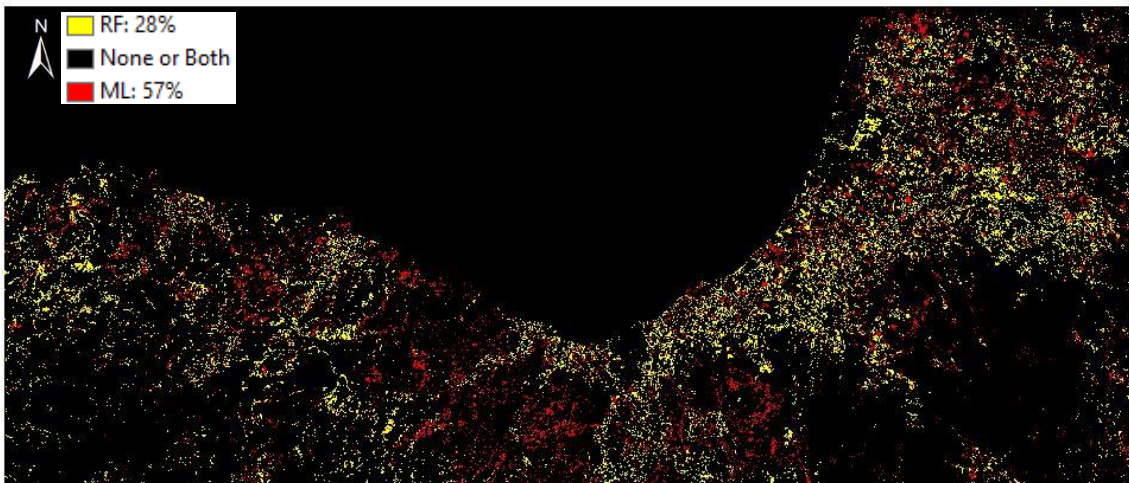


**Διάγραμμα 59.** Διαφορές Grey soil, μεταξύ Random Forest και Maximum Likelihood (Sentinel-2)

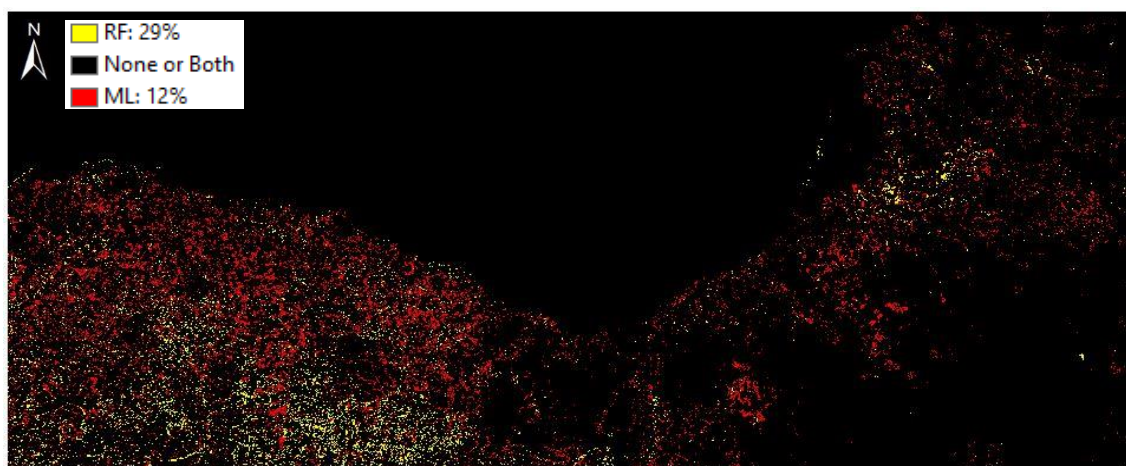




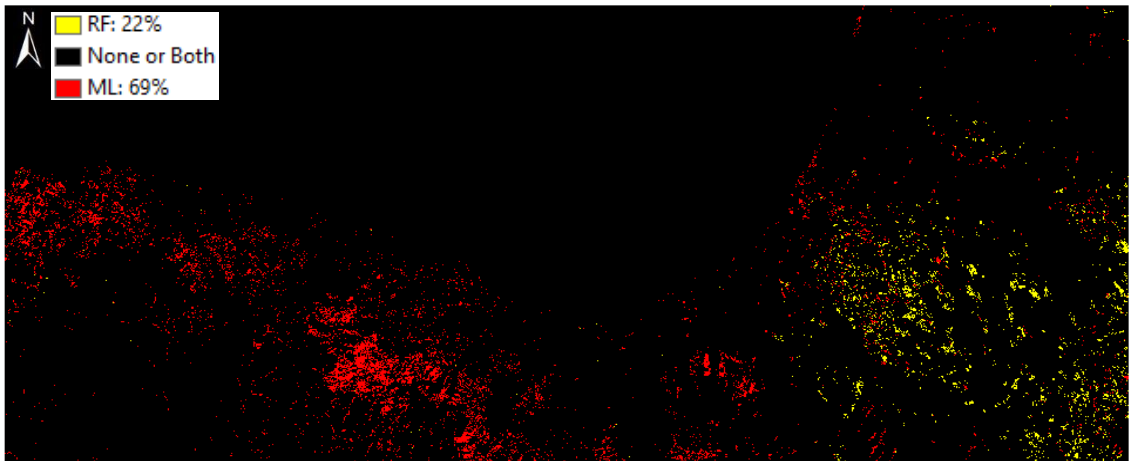
**Διάγραμμα 60:** Διαφορές High Intensity vegetation μεταξύ Random Forest και Maximum Likelihood (Sentinel-2)



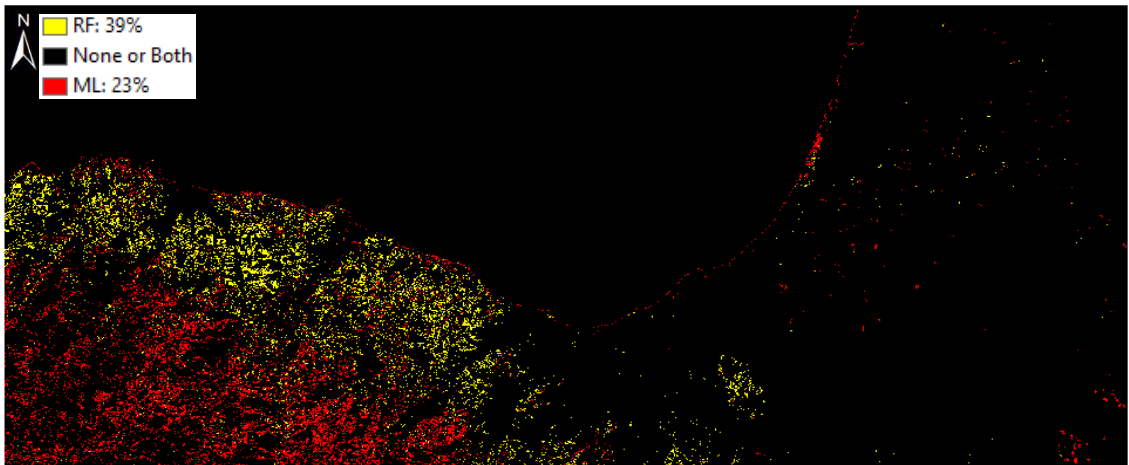
**Διάγραμμα 59:** Διαφορές Low Intensity vegetation μεταξύ Random Forest και Maximum Likelihood (Sentinel-2)



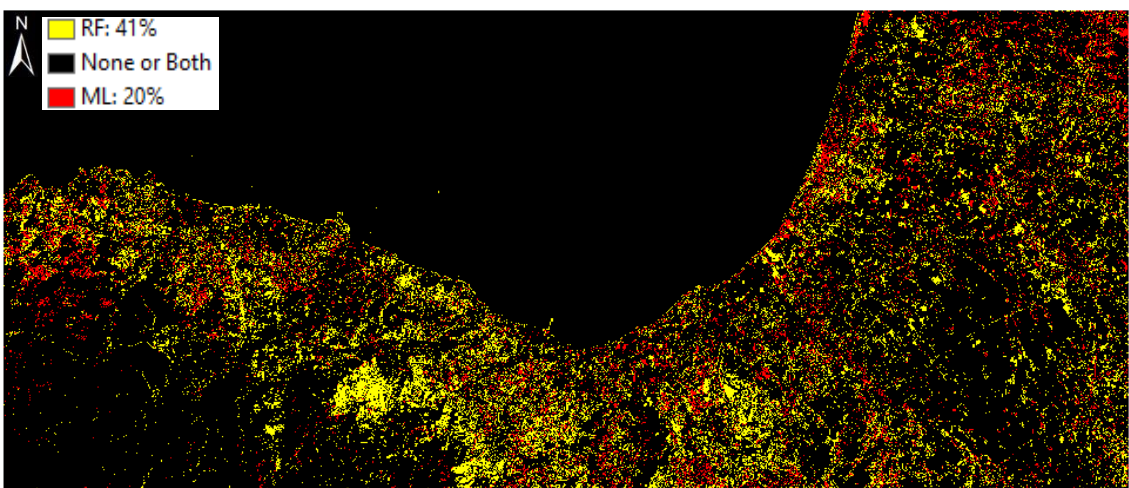
**Διάγραμμα 61:** Διαφορές Medium Intensity vegetation μεταξύ Random Forest και Maximum Likelihood (Sentinel-2)



**Διάγραμμα 63:** Διαφορές Agricultural Soil, μεταξύ Random Forest και Maximum Likelihood (Sentinel-2)



**Διάγραμμα 62:** Διαφορές Stony Soil, μεταξύ Random Forest και Maximum Likelihood (Sentinel-2)



**Διάγραμμα 64:** Διαφορές Bare Soil, μεταξύ Random Forest και Maximum Likelihood (Sentinel-2)

**Πίνακας 18.** Ποσοστιαίες διαφορές μεταξύ ταξινομήσεων Random Forest και Maximum Likelihood , ανά κατηγορία κάλυψης γης

| Class number | Class Name                         | Random Forest |        |     | Maximum Likelihood |        |     |
|--------------|------------------------------------|---------------|--------|-----|--------------------|--------|-----|
|              |                                    | RF            | Total  | %   | Max. Lik.          | Total  | %   |
|              |                                    | pixels        |        | %   | pixels             |        | %   |
| 1            | Natural waterbodies (deep)         | 17441         | 191223 | 9%  | 1318               | 207346 | 1%  |
| 2            | Natural waterbodies (shallow)      | 1482          | 51859  | 3%  | 17854              | 35487  | 50% |
| 3            | Artificial Waterbodies             | 373           | 1589   | 23% | 640                | 1322   | 48% |
| 4            | Urban area and artificial surfaces | 10407         | 25414  | 41% | 12364              | 22315  | 55% |
| 5            | Grey soil                          | 26702         | 39712  | 67% | 3021               | 43296  | 7%  |
| 6            | High Intensity vegetation          | 2114          | 21515  | 10% | 9348               | 29376  | 32% |
| 7            | Medium Intensity vegetation        | 26158         | 91444  | 29% | 10955              | 75648  | 14% |
| 8            | Low Intensity vegetation           | 15414         | 55095  | 28% | 31507              | 49929  | 63% |
| 9            | Stony soil                         | 13544         | 41709  | 32% | 9748               | 36734  | 27% |
| 10           | Agricultural soil                  | 4721          | 21145  | 22% | 2677               | 11322  | 24% |
| 11           | bare soil                          | 54332         | 133767 | 41% | 26402              | 161697 | 16% |

### Ερμηνεία και Σχολιασμός Αποτελεσμάτων:

Όπως βλέπουμε από τον πιο πάνω πίνακα, οι μεγαλύτερες διαφορές μεταξύ των δύο ταξινομητών, εντοπίζονται στην κλάση Grey soil, στην κλάση Bare soil και τέλος στην κλάση Urban areas and Artificial Surfaces. Random Forest.

Συγκεκριμένα, το 67%, το 41% και το 41% του συνολικού πλήθους εικονοστοιχείων που ταξινομήθηκαν σε κάθε μία από τις εν λόγω κλάσεις, δεν εντοπίστηκαν από τον αλγόριθμο της Μέγιστης Πιθανοφάνειας. Τα αντίστοιχα ποσοστά σε αυτές τις κλάσεις που εντοπίστηκαν μόνο από τον αλγόριθμο της Μέγιστης Πιθανοφάνειας, είναι 7% για την κλάση Grey soil, 16% για την κλάση Bare soil και 55% για την κλάση Urban areas and Artificial Surfaces.

Εντούτοις, αν παρατηρήσουμε τις ακρίβειες του παραγωγού στην Ισορροπημένη (Balanced) ταξινόμηση RF και στην ταξινόμηση Maximum Likelihood αντίστοιχα από τον πίνακα 12, βλέπουμε ότι στις κλάσεις Grey και Bare Soil, η ακρίβεια RF είναι πολύ υψηλή (PA: 92.59% και PA: 81.61%), ενώ αντιθέτως στην ταξινόμηση Max. Lik., αυτές είναι οι δύο χαμηλότερες ακρίβειες παραγωγού από όλες τις κλάσεις (PA: 58.33% και PA: 62.50%). Αυτό “αναιρεί” τις όποιες υποψίες αδυναμίας εντοπισμού των εν λόγω κλάσεων από τον αλγόριθμο RF, λόγω των υψηλών ποσοστιαίων διαφορών από τον αλγόριθμο Maximum Likelihood.

Στην κλάση Urban areas and Artificial Surfaces, αφενός παρατηρούμε την χαμηλότερη ακρίβεια από όλες τις κλάσεις για την ταξινόμηση RF (PA: 65.96%), ωστόσο η ακρίβεια του παραγωγού για την ταξινόμηση Max. Lik. αν και υψηλότερη (PA: 70%),

εξακολουθεί να είναι πολύ χαμηλή και μια από τις τρεις χαμηλότερες από το σύνολο των κλάσεων. Αυτές οι παρατηρήσεις, επιβεβαιώνουν την ιδιαιτερότητα της συγκεκριμένης κλάσης, αφού η φασματική πολυπλοκότητα της σε συνδυασμό με τη μέτρια χωρική ανάλυση των 30 μ. στην οποία πραγματοποιείται η ανάλυση, την καθιστούν δύσκολη στον εντοπισμό.

Ακόμα μια “προβληματική” κλάση, φαίνεται να είναι η κλάση Low Intensity Vegetation, αφού το 63% των εικονοστοιχείων που εντοπίστηκαν από τον αλγόριθμο ML, δεν εντοπίστηκαν από τον αλγόριθμο RF. Ανατρέχοντας και πάλι πίσω στις ακρίβειες του παραγωγού για τις δύο ταξινομήσεις, παρατηρούμε ότι ο ML σκόραρε το εκπληκτικό 100%, ενώ ο RF μόλις 79.31%. Η συγκεκριμένη διαφορά ήταν και πάλι αναμενόμενη, αφού σύμφωνα με τον πίνακα φασματικής διαχωριστικότητας (πίνακας 12), η κλάση εμπίπτει στην κατηγορία μέτριας διαχωριστικότητας λόγω της φασματικής ομοιότητας της με την κλάση Medium Intensity vegetation.

Άλλα αποτελέσματα που ξεχωρίζουν από τον πίνακα 18, είναι τα ποσοστά εικονοστοιχείων που εντόπισε μόνο ο αλγόριθμος της μέγιστης πιθανοφάνειας, στις κλάσεις Natural Waterbodies (shallow) και Artificial Waterbodies, χωρίς να εμπνέουν ανησυχία λόγω και πάλι των υψηλών ποσοστών ακρίβεια του παραγωγού στην ταξινόμηση RF (PA: 100%).

## 5 Συζήτηση και Σχολιασμός Αποτελεσμάτων

Σε αυτή την μελέτη, εξετάστηκαν και συγκρίθηκαν διαφορετικοί παραμετρικοί και μη αλγόριθμοι ταξινόμησης, σε εφαρμογές δημιουργίας θεματικών χαρτών κάλυψης γης στην ανατολική Μεσόγειο. Ο κύριος στόχος, ήταν η μελέτη της συμπεριφοράς του αλγόριθμου ταξινόμησης Random Forest, ως συνάρτηση: του αριθμού των δέντρων απόφασης που δημιουργούν το σύμπλεγμα, το μέγεθος του συνόλου δεδομένων που χρησιμοποιούνται για την εκπαίδευση του καθώς και το πλήθος των μεταβλητών που επιλέγονται και ελέγχονται για την εύρεση του βέλτιστου τρόπου διαχωρισμού των δεδομένων κατά την ανάπτυξη του κάθε δέντρου.

Για την διεκπεραίωση της έρευνας, χρησιμοποιήθηκαν πολυφασματικά δεδομένα Sentinel-2 και Landsat-8.

### 5.1 Χαρακτηρισμός δυνατοτήτων ταξινομητή Random Forest

Συγκεκριμένα, ο συνδυαστικός μη παραμετρικός αλγόριθμος Random Forest (OA: 92,03%, kappa: 92.88%), ξεπέρασε σε απόδοση τους παραμετρικούς ταξινομητές της Μέγιστης Πιθανοφάνειας (OA: 83%, kappa: 81.27%), Ελάχιστης Απόστασης (OA: 72%, kappa: 67%) και Απόστασης Mahalanobis (OA: 74%, kappa: 69.2%).

Ακριβώς όπως οι άνθρωποι τείνουν να αναζητούν πολλές απόψεις προτού πάρουν οποιαδήποτε σημαντική απόφαση, με τον ίδιο τρόπο ο συνδυαστικός ταξινομητής RF, ζυγίζει τις αποφάσεις των αλγόριθμων που τον απαρτίζουν (δέντρα απόφασης) και τις συνδυάζει για να φτάσει στην τελική απόφαση. Με αυτό τον τρόπο, αντιμετωπίζονται τα βασικά προβλήματα που χαρακτηρίζουν τα συστατικά του μέρη, όπως είναι η ύπαρξη μεγάλου αριθμού τυχαίων σφαλμάτων (biases) και η δημιουργία ενός κανόνα τόσο ευέλικτου, ο οποίος θα προσαρμοστεί ακόμα και στο θόρυβο που βρίσκεται μέσα στα δεδομένα εκπαίδευσης και αποτελεί τιμές που πιθανώς να μην ξανά εμφανιστούν στα δεδομένα. Ακόμα, βασικό πλεονέκτημα τους αποτελεί το γεγονός πως δεν βασίζονται σε υποθέσεις περί της κατανομής των δεδομένων και προσαρμόζονται ευκολότερα με αυτόν τον τρόπο σε κάθε είδους μορφής δεδομένων.

Ο αριθμός των δέντρων, είναι ευθέως ανάλογος προς την ακρίβεια του ταξινομητή, μέχρι τα 100 περίπου δέντρα απόφασης, όπου το σφάλμα γενίκευσης συγκλίνει, με “ασήμαντες” τιμές βελτίωσης.

Ο αριθμός των μεταβλητών (καναλιών) που χρησιμοποιούνται για το διαχωρισμό των δεδομένων σε κάθε κόμβο, είναι ακόμη μια σημαντική παράμετρος η οποία επηρεάζει με αντιστρόφως ανάλογο τρόπο την απόδοση του ταξινομητή. Η βέλτιστη τιμή βρέθηκε να είναι  $mtry=2$ , με μικρά ωστόσο ποσοστά διαφοράς ακρίβειας από τις υπόλοιπες τιμές ελέγχου (4 μεταβλητές → ΔΟΑ: 0,53 % , Δkappa: 0,6 % και 6 μεταβλητές → ΔΟΑ: 0,77 % και Δkappa: 0,85 %).

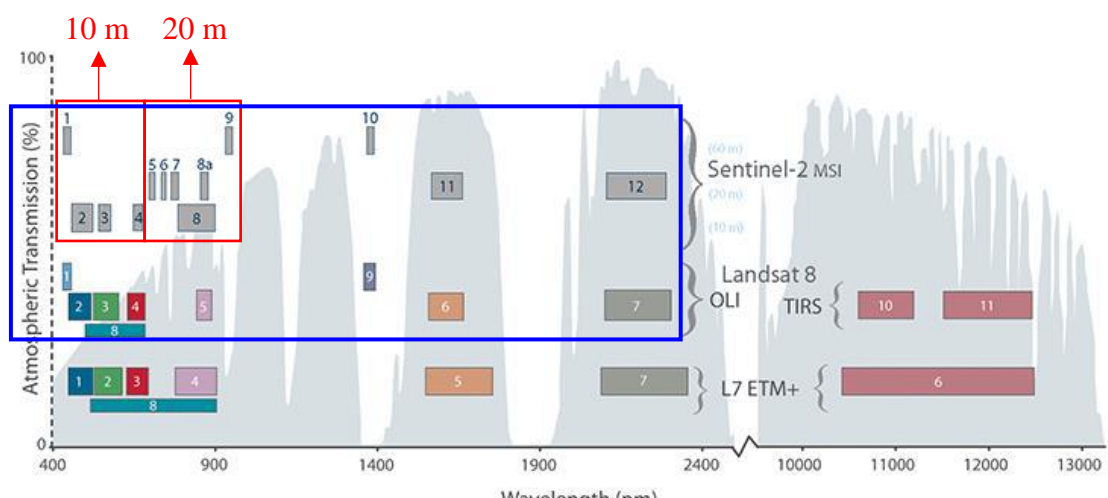
Επιπλέον, πραγματοποιήθηκε σύγκριση μεταξύ της απόδοσης του μοντέλου που δημιουργήθηκε από ένα ανισόρροπο σύνολο δεδομένων εκπαίδευσης (άνισος και τυχαίος επιλεγόμενος αριθμός εικονοστοιχείων για κάθε κλάση) και από ένα ισορροπημένο σύνολο δεδομένων εκπαίδευσης (ίσος αριθμός εικονοστοιχείων σε όλες τις κλάσεις). Οι δύο στρατηγικές ταξινόμησης παρήγαγαν παρόμοια αποτελέσματα (υπεροχή της ταξινόμησης κατά περίπου 2% στα σύνολα δεδομένων Sentinel-2 και Landsat-8), διατυπώνοντας ωστόσο με σαφήνεια τη σημασία της ισορροπημένης κατάστασης μεταξύ των θεματικών κλάσεων.

## 5.2 Σύγκριση αισθητήρων OLI και MSI

Τα νέας γενιάς πολυφασματικά δεδομένα Sentinel-2, θεωρούνται ως η συνέχεια της μακροβιότερης συλλογής δορυφορικών δεδομένων της οικογένειας Landsat. Για αυτό το λόγο, κρίθηκε αναγκαία η σύγκριση των τελικών προϊόντων που παράχθηκαν από τους αισθητήρες των δύο, ούτως ώστε να διερευνηθεί κατά πόσο είναι όμοια.

Οι συνδυασμοί πλατφόρμας και αισθητήρων, διαφέρουν αρχικά ως προς την τροχιακή και χωρική διαμόρφωσή τους. Όσον αφορά τη φασματική διάταξη, οι προδιαγραφές του αισθητήρα MSI, σχεδιάστηκε με τέτοιο τρόπο ώστε να υπάρχει μια σημαντική αντιστοιχία μεταξύ των αντίστοιχων φασματικών ζωνών του αισθητήρα OLI (διάγραμμα 65).

Όπως βλέπουμε, τα τρία κανάλια του οπτικού φάσματος (Blue (490nm), Green (560nm), Red (665nm)) και το κανάλι 4 (NIR: 842nm)) που βρίσκονται στην ανάλυση των 10m, εξασφαλίζουν τη συνέχιση της αποστολής Landsat-8 και απευθύνονται σε εφαρμογές ταξινόμησης βασικών κατηγοριών κάλυψης εδάφους. Παρόλα αυτά, τα κανάλια που βρίσκονται στην ανάλυση των 20 m (Vegetation red edge, Narrow NIR και Shortwave IR), δεν σχετίζονται με τη λειτουργία του δορυφόρου Landsat-8, αφού είναι αφιερωμένα σε εφαρμογές ανίχνευσης χιονιού / πάγου / σύννεφων αλλά και στην εκτίμηση ποσοτήτων υγρασίας σε διάφορους τύπους βλάστησης, υποστηρίζοντας πιο απαιτητικές εφαρμογές ταξινόμησης.



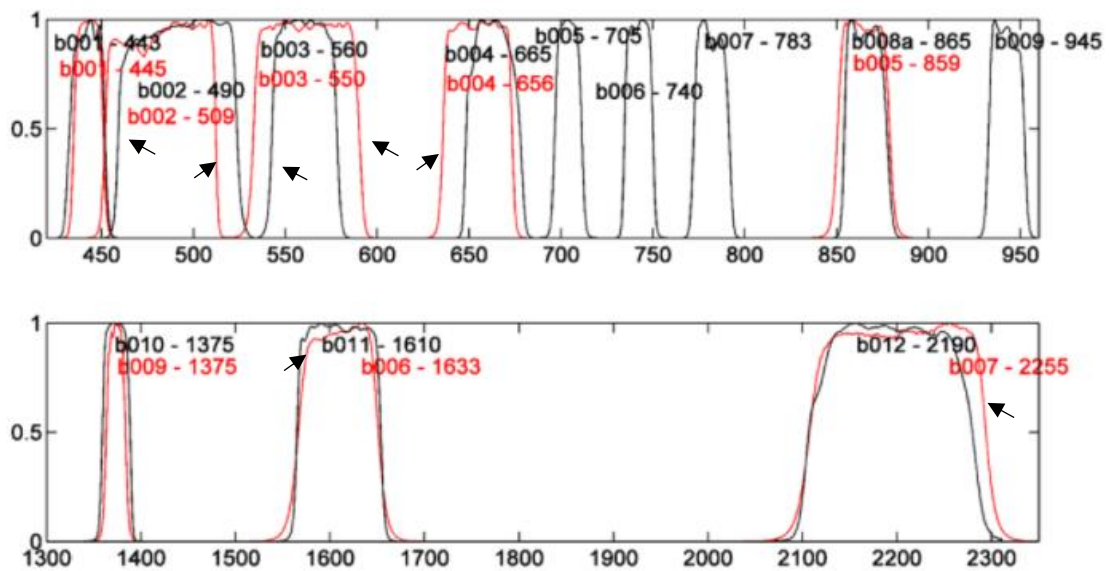
**Διάγραμμα 65:** Φασματικά χαρακτηριστικά Sentinel-2, Landsat-7 και Landsat-8

πηγή: <https://landsat.gsfc.nasa.gov/sentinel-2a-launches-our-compliments-our-complements/>



Ως εκ τούτου, σύμφωνα με τους Mandanici E. and Bitelli G. (2016) τα Relative Spectral Response Filters (τα οποία περιγράφουν την απόδοση των συστημάτων, συμπεριλαμβανομένης της μετάδοσης των φίλτρων και της ευαισθησίας των αισθητήρων) των οργάνων δεν είναι πανομοιότυπα (διάγραμμα 66), έτσι αναμένονται κάποιες διαφορές στις καταγεγραμμένες ραδιομετρικές τιμές.

Η επίδραση των ραδιομετρικών διαφορών μεταξύ των εικόνων που παρουσιάζονται στους δύο αισθητήρες πρέπει να αξιολογηθούν προσεκτικά, προκειμένου να καθοριστεί εάν οι αποκλίσεις στις τιμές ανάκλασης προκαλούν ή όχι διαφορές στα τελικά προϊόντα.



**Διάγραμμα 66:** Φασματική ευαισθησία (Relative Spectral Response Filters) των Sentinel-2 (μαύρες γραμμές) και Landsat-8 (κόκκινες γραμμές), οι περιοχές μη ταύτισης των RSR filters υποδεικνύονται με βελιάκια

Πηγή: <https://hls.gsfc.nasa.gov/algorithms/bandpass-adjustment/>

Τα στατιστικά στοιχεία της ταξινόμησης των τυχαίων δασών, επιβεβαιώνουν ότι τόσο ο αισθητήρας του Landsat-8 (OLI) όσο και ο αισθητήρας του Sentinel-2 (MSI) μπορούν να παράγουν παρόμοια αποτελέσματα ταξινόμησης (με διαφορά OA < 2% και συντελεστή kappa < 3%). Τα δεδομένα Landsat-8, παρουσιάζουν καλύτερες ακρίβειες και αυτό μπορεί να οφείλεται στην αλλοίωση των ραδιομετρικών τιμών των δεδομένων Sentinel-2 μετά από την επανασύσταση που εφαρμόστηκε προκειμένου να είναι γεωμετρικά συγκρίσιμα με αυτά του Landsat.



## 6 Μελλοντικοί ερευνητικοί στόχοι

Υπάρχουν ακόμα πολλά κενά όσο αφορά τις διαδικασίες ανάλυσης των ολοένα και πολυπλοκότερων συνόλων δορυφορικών δεδομένων που είναι διαθέσιμα προς ανάλυση.

Σε αυτή την ενότητα, παρουσιάζονται μερικές προτάσεις τόσο για μελλοντική αντιμετώπιση παρόμοιων θεμάτων όσο και για αξιοποίηση των αποτελεσμάτων που προέκυψαν από τη συγκεκριμένη έρευνα.

### 6.1 Φασματική αποσύνθεση και ταξινόμηση υπό-εικονοστοιχείων

Η δυνατότητα εφαρμογής των πολυφασματικών τηλεπισκοπικών δεδομένων, επηρεάζεται και περιορίζεται από προβλήματα όπως η ύπαρξη μικτών εικονοστοιχείων.

Σε μια πολυφασματική εικόνα, το κάθε εικονοστοιχείο ανάλογα με τη χωρική ανάλυση του αισθητήρα, καλύπτει μια συγκεκριμένη έκταση στο έδαφος. Το σήμα που ανιχνεύεται από τον αισθητήρα σε αυτή την περιοχή, αποτελεί συχνά ένα συνδυασμό πολυάριθμων σημάτων (φασματικό μείγμα), που προέρχονται από τους διαφορετικούς στόχους που εμπίπτουν μέσα στα όρια της έκτασης εδάφους του εικονοστοιχείου. Αυτό το γεγονός, οδηγεί σε λανθασμένη ερμηνεία του προϊόντος ταξινόμησης και κατ' επέκταση στη λήψη λανθασμένων αποφάσεων.

Διάφορες τεχνικές έχουν αναπτυχθεί και χρησιμοποιούνται στην προσπάθεια «αποσύνδεσης» των πληροφοριών που συνθέτουν το μοναδικό σήμα που λαμβάνει ο αισθητήρας για κάθε εικονοστοιχείο. Στην περίπτωση της δικής μας μελέτης, η χωρική ανάλυση στην οποία πραγματοποιήθηκε η ανάλυση και παράχθηκαν τα τελικά προϊόντα, είναι τα 30 μ. Το επόμενο βήμα λοιπόν, θα ήταν η προσπάθεια αναγνώρισης των αναλογιών ανάμειξης των εμπλεκόμενων στόχων στην κάθε περιοχή συνολικής έκτασης 900 τ.μ. (εξίσωση 10), τα οποία θα μας οδηγήσουν στο στάδιο ταξινόμησης των υπό-εικονοστοιχείων που εντοπίζονται.

$$\mathbf{x} = \sum_{i=1}^M a_i \mathbf{s}_i + \mathbf{w}$$

Εξίσωση 13 Μοντέλο φασματικής μίξης

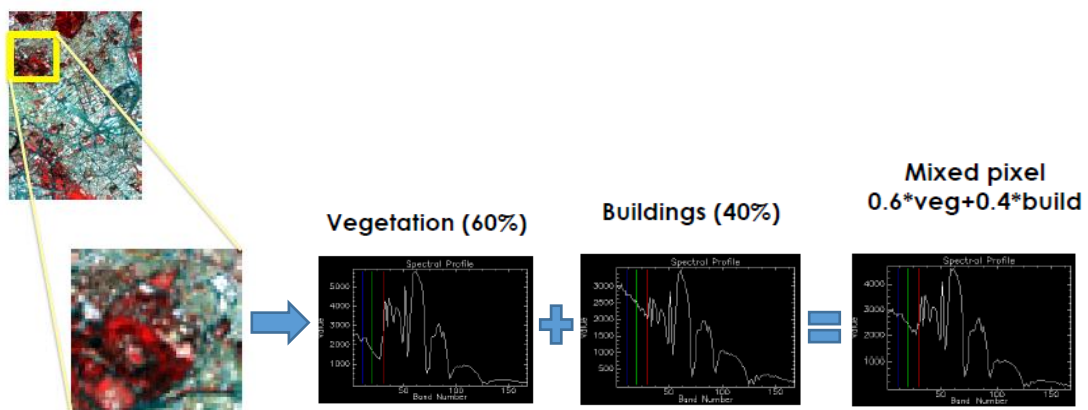
Όπου:

X: το φασματικό διάνυσμα που ανιχνεύει ο αισθητήρας σε κάθε εικονοστοιχείο

S<sub>i</sub>: οι διαφορετικοί τύποι κάλυψης γης, που εμπίπτουν μέσα στα όρια κάθε εικονοστοιχείου

a<sub>i</sub>: τα ποσοστά συμμετοχής του κάθε τύπου κάλυψης γης, στο τελικό φασματικό διάνυσμα που ανιχνεύει ο αισθητήρας σε κάθε εικονοστοιχείο

W: θόρυβος που πιθανώς να εντοπίζεται σε κάθε παρατήρηση



**Διάγραμμα 67:** Παράδειγμα μικτού εικονοστοιχείου και φασματικής μίξης του τελικού φασματικού διανύσματος που εντοπίζει ο αισθητήρας

## 6.2 Εναρμόνιση δεδομένων Landsat και Sentinel-2

Οι δορυφόροι Landsat-8 και Sentinel-2, παρουσιάζουν φασματικές και χωρικές ομοιότητες που καθιστούν δυνατή τη συνδυασμένη χρήση των προϊόντων τους. Το έργο “Harmonized Landsat-Sentinel-2 (HLS)”, είναι μια πρωτοβουλία της NASA για την παραγωγή εναρμονισμένων προϊόντων υψηλότερης χρονικής (2 -3 ημερών) και χωρικής ανάλυσης (< 30m).

Σε αυτό το πλαίσιο, ένα "εναρμονισμένο" προϊόν ανάκλασης συνεπάγει ότι έχουν εφαρμοστεί οι αναγκαίες ραδιομετρικές, φασματικές, γεωμετρικές και χωρικές διορθώσεις/αναγωγές, έτσι ώστε τα προϊόντα που παράγονται ξεχωριστά από τον κάθε δορυφόρο να είναι όσο το δυνατό πιο όμοια. Τα αποτελέσματα που προέκυψαν σε αυτή την μελέτη, μας δίνουν το πράσινο φως προς τον πειραματισμό για την εναρμόνιση των δύο συνόλων δεδομένων, λαμβάνοντας υπόψη τις διαφορές που προκύπτουν από τους αισθητήρες τους.

## ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Βάσιλας Γ. (2013), ‘Διαχρονική Χαρτογράφηση Χρήσεων/Κάλυψης γης της περιοχής Λάρνακας με χρήση δορυφορικών εικόνων υψηλής χωρικής ανάλυσης’, BSc thesis, Τεχνολογικό Πανεπιστήμιο Κύπρου, Λεμεσός.
2. Βλαχάβας Ι κ.α., 2011, Παρουσίαση ‘Τεχνητή Νοημοσύνη’, Εκδόσεις Παν/μίου Μακεδονίας, κεφ.18
3. Δημητρακόπουλος Κ. (2010), ‘Χαρτογράφηση χρήσης/κάλυψης γης με την αντικειμενοστραφή ταξινόμηση εικόνων SPOT’, MSc thesis, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Θεσσαλονίκη.
4. Καρτάλης Κ. και Χ. Φείδας, 2007. Αρχές και εφαρμογές δορυφορικής τηλεπισκόπησης. Β. Γκιούρδας Εκδοτική, Αθήνα.
5. Καρτέρης Μ.Α. & Γιαννακόπουλος Β.Ι., 1998. Περιβαλλοντική Χαρτογραφία, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.
6. Λασπιάς, Ε. (2012), Επιβλεπόμενη & Μη επιβλεπόμενη ταξινόμηση πολυφασματικών εικόνων τηλεπισκόπησης και θεματικές εφαρμογές τους στον Ελλαδικό χώρο : ανάπτυξη σε περιβάλλον Wiki, Ph. D Thesis, Εθνικό Μετσόβιο Πολυτεχνείο.
7. Μερτίκας Π.Σ. (2006), “Τηλεπισκόπηση και Ψηφιακή Ανάλυση Εικόνας”, ΙΩΝ, Περιστερί, κεφ. Ταξινόμηση Εικόνας.
8. Παρχαρίδης Ι. (2015), “Αρχές Δορυφορικής Τηλεπισκόπησης - Θεωρία και Εφαρμογές”, Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκονόμων, Εθνικό Μετσόβιο Πολυτεχνείο, Ταξινόμηση Εικόνας, 88-93
9. Ahlqvist O. (2008), “In search of classification that that supports the dynamics of science: The FAO Land Cover Classification System and proposed modifications”, Environment and Planning B: Planning and Design, 35, 169-186.
10. Anderson J.R. et al. (1976), “A Land Use and Land Cover Classification System For Use With Remote Sensor Data”, Geological Survey Professional paper 964.
11. Belgiu, M. and Drăguta L., (2016). “Random Forest in remote sensing: a review of applications and future directions”. ISPRS J. Photogramm. Remote Sens. 114, 24–31.

12. Briem, G.J., Benediktsson, J.A., Sveinsson, J.R., 2002. "Multiple classifiers applied to multisource remote sensing data.", *IEEE Trans. Geosci. Remote Sens.* 40, 2291–2299.
13. Breiman L., 2001, Random Forests, *Machine Learning*, 45, 5-32.
14. Campell B.J., 2002. *Introduction to Remote Sensing*. 3rd edition. Virginia Polytechnic Institute and State University. The Guilford Publications Press, New York.
15. Colditz, R., 2015. "An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms.", *Remote Sens.* 7, 9655.
16. DeFries, R.S.; Foley, J.A.; Asner, G.P. Land-use choices: Balancing human needs and ecosystem function. *Front. Ecol. Environ.* 2004, 2, 249–257.
17. Drusch, M. et al, 2012. 'Sentinel-2: ESA's optical high-resolution mission for {GMES} operational services.' *Remote Sens. Environ.* 120, 25–36.
18. Eitel, J.U. et al (2011), "Red-edge information from satellites improves early stress detection in a New Mexico conifer woodland." *Remote Sens. Environ.*, 115
19. Friedl M.A. and Brodley C.E. (1997), "Decision Tree Classification of Land Cover from Remotely Sensed Data", *Remote Sensing of Environment*, 61,399-409
20. Galiano R. et al (2012), "An assessment of the effectiveness of a Random Forest classifier for land-cover classification." *ISPRS J. Photogramm. Remote Sens.* 67, 93–104.
21. Hastie T. et al. (2009), "The Elements of Statistical Learning : Data Mining, Inference, and Prediction", Springer, Random Forest ; Ensemble Learning, 587-624
22. Immitzer, M. et al (2016), "First experience with sentinel-2 data for crop and tree species classifications in Central Europe." *Remote Sens.*, 8, 166
23. Lillesand T.M., & Kiefer R.W., (1994). *Remote Sensing and Image Interpretation*, John Wiley & Sons, New York, 3<sup>rd</sup> ed.
24. Lillesand T.M. et al. (2004), "Remote Sensing and Image Interpretation ", John Wiley & Sons, USA, Image Classification,550-552.
25. Mandanici E. and Bitelli G., (2016), "Preliminary Comparison of Sentinel-2 and Landsat-8 Imagery for a Combined Use", *Remote Sens.*, 8, 1014. 402

26. Olofsson et al., 2014, "Good practices for estimating area and assessing accuracy of land change", *Remote Sensing of Environment*, 148,42-57
27. Pelletier C. et al, 2016, 'Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas', *Remote Sensing of Environment* 187 (2016) 156–168.
28. Pesaresi, M. et al (2016). "Assessment of the added-value of sentinel-2 for detecting built-up areas." *Remote Sens.*, 8, 299
29. Pontius, R. G. (2000). "Quantification error versus location error in comparison of categorical maps.", *Photogrammetric Engineering & Remote Sensing*, 66, 1011–1016.
30. Pontius, R. G., & Lippitt, C. D. (2006). "Can error explain map differences over time?", *Cartography and Geographic Information Science*, 33, 159–171
- Rogan, J., and D. Chen. (2004). "Remote Sensing Technology for Mapping and Monitoring Land-Cover and Land-use Change." *Progress and Planning* 61 (4): 301 - 325.
31. Rokach L. (2009), "Ensemble-based classifiers", *Artif Intell Rev* (2010) 33, 1–39.
32. Schuster, C. et al (2012). "Testing the red edge channel for improving land-use classifications based on high-resolution multi-spectral satellite data.", *Int. J. Remote Sens.*, 33, 5583–5599
33. Sharma R. et al (2013), "Decision tree approach for classification of remotely sensed satellite data using open source support", *J. Earth Syst. Sci.* 122, p 1237–1247
34. Simms É.L. and Ward H. (2013), "Multisensor NDVI-Based Monitoring of the Tundra-Taiga Interface (Mealy Mountains, Labrador, Canada)", *Remote Sensing* 2013, 5(3), 1066-1090
35. Stehman, S. V., & Wickham, J.D. (2011). "Pixels, blocks of pixels, and polygons: Choosing a spatial unit for thematic accuracy assessment.", *Remote Sensing of Environment*, 115, 3044–3055.
36. Szuster, B.W., Chen, Q., Borger, M., 2011. 'A comparison of classification techniques to support land cover and land use analysis in tropical coastal zones'. *Applied Geography* 31, 525–532.
37. Tso B. and Mather P.M. (2009), *Classification Methods for Remotely Sensed Data*, Taylor & Francis Group, Decision Trees, Sound Parkway NW,,183-220

38. Verburg, P.H et al (2011), “Challenges in using land use and land cover data for global change studies”. *Glob. Chang. Biol.*, 17, 974–989.
39. Yap B.W. et al (2013), “An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets”, *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, p 13-22
40. Yuan B. and LIU W., (2012), “Measure oriented training: a targeted approach to imbalanced classification problems”, *Front. Comput. Sci.* 6(5), 489–497.

